Developing statistical and analytical methods for untargeted analysis of complex environmental matrices

Madison Bell

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Doctor of Philosophy in Biology
(Specializing in Chemical and Environmental Toxicology)

Department of Biology
Faculty of Science
University of Ottawa
Ottawa, ON, Canada

Supervisor:
Dr. Jules Blais[1]

Thesis Advisory Committee:
Dr. David Overy[2]
Dr. Robert Letcher[3]
Dr. Cory Harris[1]

[1]University of Ottawa
[2]Adjunct research professor at Agriculture and Agri-food Canada
[3]Adjunct research professor at Environment and Climate Change Canada

ABSTRACT

The main objective of this thesis was to develop statistical and analytical methods for untargeted analyses of complex environmental matrices like soil and sediment. Untargeted analyses are notoriously difficult to perform in matrices like soil and sediment because of the complexity of organic matter composition within these matrices. This thesis aimed to (1) Develop and compare extraction methods for untargeted analyses of soil and sediment while also developing data handling and quality control protocols; (2) Investigate novel applications of untargeted analyses for environmental classification and monitoring; and (3) Investigate the experimental factors that can influence the organic matter composition of untargeted extractions. CHAPTER TWO is a literature review of metabolomics protocols, and these protocols were incorporated into a proposed workflow for performing untargeted analysis in oil, soil, and sediment. This thesis contains the first application of untargeted analysis to freshwater lake sediment organic matter (i.e. sedimentomics) in CHAPTER THREE, and this has implications for discovering new biomarkers for paleolimnology (APPENDIX ONE). I demonstrated successful extraction methods for both sedimentomics and soil metabolomics studies in CHAPTER THREE and CHAPTER FIVE, respectively, using the proposed workflow from CHAPTER TWO. I also applied sedimentomics to the classification of lake sediments using machine learning and geostatistics based on sediment organic matter compositions in CHAPTER FOUR; this was a novel application of sedimentomics that could have implications for ecosystem classifications and advance our knowledge of organic matter cycling in lake sediments. Lastly, in CHAPTER FIVE I determined microbial activity, extraction method, and soil type can all influence the composition of soil organic matter extracts in soil metabolomics experiments. I also developed novel quality controls and quantitative methods that can help control these influences in CHAPTER FIVE and APPENDIX THREE. APPENDIX TWO was written in collaboration with multiple

researchers and is a review of all “omics” types of analyses that can be performed on soil or sediment, and how methods like the untargeted analysis of soil and sediment organic matter can be linked with metagenomics, metatranscriptomics, and metaproteomics for a comprehensive metaphenomics analysis of soil and sediment ecosystems. In CHAPTER SIX the conclusions and implications for each chapter and overall for this thesis are detailed and I describe future directions for the field. In the end the overall conclusions of this thesis were: 1) Quality controls are necessary for sedimentomics and soil metabolomics studies, 2) Sedimentomics is a valid technique to highlight changes in sediment organic matter, 3) Soil metabolomics and sedimentomics yield more information about carbon cycling than traditional measurements, and 4) Soil metabolomics organic matter extractions are more variable and require more quality controls.

TABLE OF CONTENTS

DEDICATION

To Family,

who are always there holding up

the light at the end of the tunnel

ACKNOWLEDGEMENTS

To Dr. Jules Blais, my supervisor, thank you for your support and guidance through my graduate studies. I am sincerely grateful for the opportunities you have given me, and for the significant time and effort you have put in to make those life changing opportunities happen. Thank you Linda Kimpe for your guidance and your encyclopedic knowledge of all things lab, I have learned a lot from you. To David Eickmeyer, thank you for making lab work more fun, your dedication to teaching, and exchanging nerdy chemistry jokes. To my committee members, Dr. David Overy, Dr. Robert Letcher, and Dr. Cory Harris, thank you for your guidance over the years as I have completed my PhD.

Thank you Dr. David Overy for providing me with the opportunity to work with you at Agriculture and Agri-food Canada (AAFC). Without your mentorship and expertise, this thesis would not have been possible. Thank you to the AAFC team, Ed Gregorich, Amanda Sproule, Tom Witte, Ulrica McKim, Anne Hermans, Ryan Tobalt, and others for all of your expertise, answering my endless questions, and fun conversations. Special thank you to Amanda Sproule, for your mad Orbitrap skills and fulfilling my requests for “just one more” replicate, and to Ulrica McKim, for your knowledge as a scientist and accommodating my many “Why don’t we just…” changes to the SOPs.

To the Blais lab, Lauren Gallant, Mija Azdajic, Sawyer Stoyanovich, Cynthia Cheney, Jennifer Kissinger, Jonathan Seguin, Jennifer Keir, Kirsten Smythe, Alexandre Salat, Jose Rodriguiz-Gil thank you everyone for the brainstorming sessions, inspiration, laughs, and support. I am honored to have worked with all of you, and I truly believe that all of you are incredibly hard working, talented, and committed scientists. Special thank you to Cynthia and Lauren for helping me maintain my sanity and for being there when I needed friends.

To my family, for their unequivocal love, support, and encouragement. Rain or shine they are always a drive or phone call or text away, and ready to share laughs and hugs. To Dad, thank you for teaching me to hold my ground, be resilient, and how to laugh even when everything seems like it is falling down. Thank you Mom for teaching me the meaning of perseverance (aka stubbornness), and how to be "spicy". Josh, thank you(?) for never letting me get full of myself, and your endless sass (I honestly don't know where you keep it all ;)). Thank you to my grandmothers, Barbara and Elizabeth. Barbara – thank you for always being encouraging, for supplying endless care packages (even though I have to share them with Kyle now :O), and for teaching me the meaning of kindheartedness. Elizabeth – thank you for the life chats, and for showing me what passion and zest for life look like. Thank you to my grandfather, Jim, for reading my papers and plying me with endless snacks (P.S. I am not that scary when I am hangry). Grandpa you have been an endless source of support, ideas, and jokes (Yes, even the bad ones like joke 254), and you have been a pillar of support throughout my entire education. Lastly, to Kyle, I met you at the beginning of my journey in Ottawa as a PhD student, but I surely would not have finished it without you. You and Taz are my happiness, and I look forward to our future adventures and being "weird" together.

LIST OF ABBREVIATIONS

AMDIS	Automated Mass Spectral Deconvolution and Identification System
ANOVA	Analysis of Variance
APCI	Atmospheric Pressure Chemical Ionization
APPI	Atmospheric Pressure Ionization
B	Boreal region
CCS	Collisional Cross Section
COLMAR	Complex Mixture Analysis by NMR
CV	Cross-Validation
DBE	Double Bond Equivalents
DIA	Data Independent Acquisition
DOM	Dissolved Organic Matter
EI	Electron Impact
ESI	Electrospray Ionization
FT1	Forest-tundra region relating to the Taiga Shield (TS2)
FT2	Forest-tundra region relating to the Mackenzie Delta Uplands in the Low Arctic (MDU)
FT-ICR	Fourier Transform Ion Cyclotron Resonance
GCMS	Gas Chromatography Mass Spectrometry
GCxGC	Two-dimensional Gas Chromatography
GMD	Golm Metabolome Database
HA	High Arctic
HCA	Hierarchical Clustering
HMDB	Human Metabolome Database
HRMS	High Resolution Mass Spectrometry
HSD	Honestly Significant Difference
IMMS	Ion Mobility Mass Spectrometry
IR	Infrared Spectroscopy
LA	Low Arctic
LCxLC	Two-dimensional Liquid Chromatography
LCMS	Liquid Chromatography Mass Spectrometry
LC-QToF-MS	Ultra performance liquid chromatography quadrupole time of flight mass spectrometry
LRMS	Low Resolution Mass Spectrometry
LSD	Least Significant Difference
MCCV	Monte Carlo Cross-Validation
MDU	Mackenzie Delta Uplands in the forest-tundra region of the Low Arctic ecozone
MMCD	Madison Metabolomics Consortium Database
MS^n	Tandem Mass spectrometry
NEF	National Ecological Framework
NIST	National Institute of Standards and Technology
NMR	Nuclear Magnetic Resonance
OOB	Out-of-bag
OPLS	Orthogonal Partial Least Squares or Projection to Latent Structures

OPLS-DA	Orthogonal Partial Least Squares Discriminant Analysis
OM	Organic Matter
PCA	Principal Component Analysis
PLS	Partial Least Squares or Projection to Latent Structures
PTFE	Polytetrafluoroethylene
QC	Quality Control
RDA	Redundancy Analysis
RF	Random Forest
ROC	Receiver Operating Characteristic
sedOM	Sedimentary Organic Molecules
SOM	Soil Organic Matter
SOMap	Self-organizing Maps
sPLS	Sparse Partial Least Squares or Projection to Latent Structures
SVM	Support Vector Machines
T	Tundra region
TN	Total Nitrogen
TOC	Total Organic Carbon
TOF	Time of Flight
TP	Taiga Plains
TS	Total Sulfur
TS1	Taiga Shield in the boreal region
TS2	Taiga Shield in the forest-tundra region
VIP	Variables of Importance
WCSS	Within Cluster Sum of Squares

LIST OF TABLES

LIST OF FIGURES

PREFACE

The thesis is a collection of manuscripts formatted in accordance with the guidelines provided by the University of Ottawa's Faculty of Graduate and Postdoctoral Studies. CHAPTER ONE is a general introduction to the objectives and background of the thesis. CHAPTERS TWO, THREE, and FOUR were formatted for and published in *Science of the Total Environment*. CHAPTER FIVE is formatted for *Soil Biology and Biochemistry*. The final chapter, CHAPTER SIX, will contain the conclusions from each chapter and the overall conclusions of the thesis.

CHAPTER 1: General Introduction

1. BRIEF OVERVIEW OF THE CARBON CYCLE

Carbon (C) is essential for life and is contained in four major global compartments 1) atmosphere, lithosphere, hydrosphere, and biosphere (Ussiri and Lal, 2017a). C cycling is the process of C exchanging between the different compartments. In order to be exchanged through these compartments, C is transformed; C alternates between inorganic forms (i.e. carbonates, CO_2, CH_4, etc.) and organic forms through biotic and abiotic interactions within the compartments. The flux of C between the compartments can be fast (i.e. years – millennia) or slow (1,000,000+ years). Specifically, the fast C cycle is considered the biological C cycle because C flux is catalyzed by biological activity (Ussiri and Lal, 2017b).

C cycling between these surficial earth compartments is controlled by a number of different abiotic and biotic processes. Abiotic processes include weathering, degassing, adsorption/desorption equilibriums, redox reactions, volcanic eruptions, climate, burial, etc.; while biotic processes include photosynthesis, C transformations mediated by microbial activity, and the growth of biomass (i.e. living organisms). Anthropogenic factors (i.e. fossil fuel emissions, land use changes, etc.) have also disturbed the steady-state equilibrium of the pre-industrial C cycle leading to climate change (Friedlingstein et al., 2019). The interaction of these processes is complex, and requires very detailed and extensive models to predict fluxes between compartments (Dwivedi et al., 2019; Luo et al., 2016; Runge et al., 2019).

The C cycle is critical to understand because the impact human industrialization has had on the C cycle has led to global repercussions like climate change (Lal, 2019). Climate change in turn threatens to further amplify C cycle perturbations through positive feedback mechanisms like climate forcing, permafrost thawing, treeline migration, etc. (Foster et al., 2017; Pearson et al., 2013; Schuur et al.,

2015). There are solutions proposed to reduce the impact humans have had on the C cycle such as: improving global energy efficiency, carbon sequestration, solar radiation management, atmospheric CO_2 removal methods, etc. (Ussiri and Lal, 2017c). However, these methods are theoretical and/or difficult to scale to the global problem that is climate change. Regardless, a stronger understanding of the processes inherent in the C cycle will potentially lead to better solutions to mitigate climate change in the future.

2. BIOLOGICAL CARBON CYCLE

2.1. Processes of the biological carbon cycle

The biological C cycle, or the fast C cycle, is of specific interest to this thesis. The biological C cycle occurs on a decadal scale, as opposed, to the millennial scale of the slow geological C cycle. The biological C cycle involves fewer compartments than the slow C cycle and is limited to Earth's surficial (or near surficial) compartments (i.e. atmosphere, biosphere, and hydrosphere – mainly oceans)). It is estimated the atmosphere contains ~849 Pg C (Ussiri and Lal, 2017d), the terrestrial biosphere contains 450-700 Pg C (Prentice et al., 2001), soils contain 1500-2400 Pg C (Jobbágy and Jackson, 2000), and the oceans contain 39,000 Pg C (Ussiri and Lal, 2017d).

The main driver of the biological C cycle is biological activity. Photosynthesis absorbs CO_2 from the atmosphere for plant and microbial growth; their growth converts inorganic C to organic C that is consumed by other organisms to fuel their growth, or the organic C is deposited and buried in soils and/or sediment. During deposition and burial organic C is modified further by microbes or it is stabilized via soil/sediment geochemical properties (i.e. adsorption, aggregation, etc.). The process where microbes transform organic C is called diagenesis, and it results in stabilized organic C that, if it is

stabilized long enough, undergoes additional diagenesis and eventual catagenesis leading to oil/coal formation (Killops and Killops, 2013a). Diagenesis can also result in mineralization of the organic C where it is degassed from soil-air and water-air interfaces as inorganic C (i.e. CO_2, CH_4, etc.).

Thus, soil and sediment are critical matrices for the biological C cycle. They are complex matrices that vary chemically, physically, and biologically across space, in depth, and over time (Masoom et al., 2016). Alone, soil organic C, or soil organic matter (SOM), contains more C than plant biomass and the atmosphere (Jackson et al., 2017; Schlesinger, 1990); a fraction of SOM also enters freshwater and marine waters to get buried into the sediments as sediment organic matter (sedOM; Figure 1.1). While freshwater sedOM is considered a minor pool of organic C, marine sedOM is large C reservoir for both slow and fast C cycles (Killops and Killops, 2013b; Ussiri and Lal, 2017d).

SOM and freshwater sedOM are of specific interest to this thesis. SOM and sedOM are similar in theory, but with a few important differences. SOM is the result of organic C input from terrestrial biomass (i.e. agriculture, and terrestrial plants, animals, and fires) and subterranean biomass (i.e. microbes, roots systems, animals; Figure 1.1), where most of the organic C is contained within the first metre, and < 840 Pg C is located in the second and third metres combined (Batjes, 1996; Jobbágy and Jackson, 2000). Freshwater sedOM has more diverse inputs than SOM, but is less important to the biological C cycle in terms of scale (Killops and Killops, 2013b). Freshwater sedOM inputs include terrestrial biomass and subterranean biomass directly and from soil erosion into freshwater systems (Figure 1.1); any terrestrial organic C is modified before it is buried in freshwater lake sediments (Meyers and Ishiwatari, 1993). SedOM also contains organic C from aquatic (i.e. algae, fish, aquatic plants, microbes etc.) and benthic biomass (i.e. microbes, invertebrates, animals, etc.). Another critical difference between SOM and sedOM is the rate of C cycling and flux to the atmosphere. The mean

residence time of organic C in SOM is 25 years although this varies depending on the composition of SOM, the microbiota present, climate conditions, and the depth of the organic C (Killops and Killops, 2013b). Thus, organic C is efficiently recycled in soil and organic C sequestration and preservation is largely limited to peat formation in bogs and wetlands (Killops and Killops, 2013b). Rates of organic C burial and residence times in sedOM are less well studied, but freshwater sedOM is theorized to be a significant global long-term C sink because of cool temperatures and anoxic conditions within the sediments (Dean and Gorham, 1998; Mendonça et al., 2017; Sobek et al., 2009). In summary, SOM is abundant and dynamic, while sedOM is diverse and better preserved.

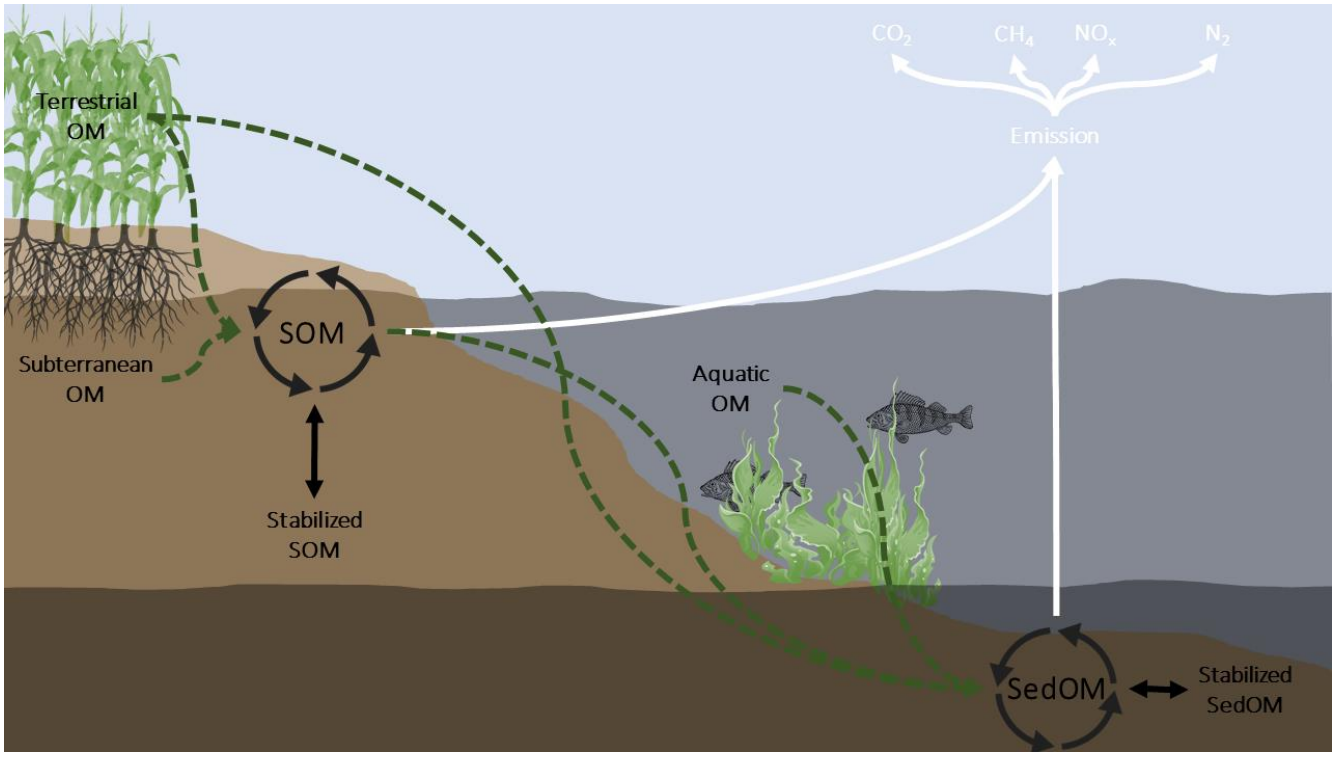


Figure 1.1: Generalized diagram of C cycling for SOM and SedOM. Dotted green lines represent the deposition and burial of organic C from biomass, black lines indicate abiotic and biotic C cycling (mostly microbial) in situ, and white arrows indicate mineralization.

2.2. Knowledge gaps in the biological C cycle

While knowledge of the biological C cycle has grown in leaps in bounds during the last few decades there are still knowledge gaps because soil and sediment matrices and the processes acting on them are complex. Organic C composition is also notoriously difficult to study (Masoom et al., 2016; Young and Crawford, 2004). Recently, new research suggests organic C processes are controlled by microbial activity and mineral-matrix interactions (Lehmann and Kleber, 2015; Schmidt et al., 2011), as opposed to decades old theory of humus formation, based on chemical recalcitrance (i.e. molecules resistant to biotic and abiotic degradation), controlling organic C cycling. Recent research also postulates that microbially derived organic C and necromass are the main components of stable organic C (Kindler et al., 2006; Schweigert et al., 2015; Simpson et al., 2007). However, the interactions between biotic and abiotic processes regulating the biological C cycle in soil and sediment are still poorly constrained (Jackson et al., 2017). Also, the structure-function relationship between soil processes and individual molecules is the subject of future research. These knowledge gaps mean that soil organic matter (SOM) and sediment organic matter (sedOM) C fluxes are poorly represented in global and/or ecosystem scale system models, which are critical for predicting C cycle turnover in a changing climate (Dwivedi et al., 2019; Todd-Brown et al., 2013; Wieder et al., 2018, 2013). While this thesis does not directly address these knowledge gaps in sediment and soil C cycling, it does contribute work toward developing, improving, and critically examining the “Omics” methods that will be used to address these knowledge gaps in future studies.

3. "OMICS"

3.1. Overview of "Omics" methods

"-Omics" is an untargeted approach to studying complex biological systems. The term "-omics" is likely derived from the Sanskrit "OM" meaning completeness (Yadav, 2007), and when combined with the discipling of interest, it implies "all metabolites" for metabolomics, "all genes" for genomics, "all proteins" for proteomics, etc. "Omics" was started in the 80's relating to genomics (Robertson, 2005; Yadav, 2007), and by the late 90's the sister fields of metabolomics (Fiehn, 2001), and proteomics (Yadav, 2007) emerged. This approach incorporates many different disciplines including field-specific knowledge (i.e. cell biology or biochemistry or soil science), analytical chemistry, data mining and management, and statistical modelling and multivariate analyses.

The advantage of using "omics" methods is the depth and breadth of the knowledge generated. "Omics" investigations use untargeted analyses to look at the distributions of many different analytes and thus generate a "global" view of the system. This global perspective allows the assessment of gross patterns, and successive data mining can lead to the identification of potential biomarkers (i.e. variables that drive pattern formation in the system under study). Untargeted "omics" is often used for hypothesis generation, and after "omics" studies the new hypotheses are explored using semi-targeted and targeted analyses to validate the hypothesis and develop them further.

The process of "Omics" studies, specifically regarding small organic molecules, is expounded in CHAPTER TWO of this thesis, but in summary, there are 5 main method components: 1) Sampling, 2) Extraction, 3) Sample analysis, 4) Chemometrics, and 5) Hypothesis validation (Figure 1.2). Sampling refers to the experiment design. "Omics" studies require well defined experiment designs to constrain the inherent variability in untargeted assessments. Experimental questions should be concise and

compare well-defined groups, sampling replication should be large, and attention should be paid to sample collection procedures to minimize contamination and the introduction of artifacts. The extraction methods are typically simple and produce crude extracts. Crude extracts themselves can introduce artifacts, which is a problem that can be controlled with the introduction of appropriate quality controls. Sample analysis for untargeted analyses typically requires nuclear magnetic resonance (NMR) or high resolution mass spectrometry (HRMS); HRMS is used in this thesis. HRMS instruments are ideal for untargeted analyses of complex matrices because of their resolving power. HRMS produces highly accurate and precise mass information relating to different molecules, which is later used as a variable in chemometrics and annotation. Chemometrics is the processing and analysis of the copious amounts of data produced by HRMS analysis. The processing of HRMS data is extensive, and the results can be subjective based on the software used and the methods used to clean and align the data. A variety of statistical analyses can be used to analyze the processed data and the types of analyses used depends on the original research question. Lastly, “Omics” analyses require hypothesis validation, and while this is not typically considered part of the “Omics” method, it is critical to take the hypotheses revealed from the statistical analyses and validate them using more specific and targeted methods. For instance, biomarkers need to be annotated, mechanisms of action and robustness need to be determined before they can be applied as biomarkers in diagnostic tools.

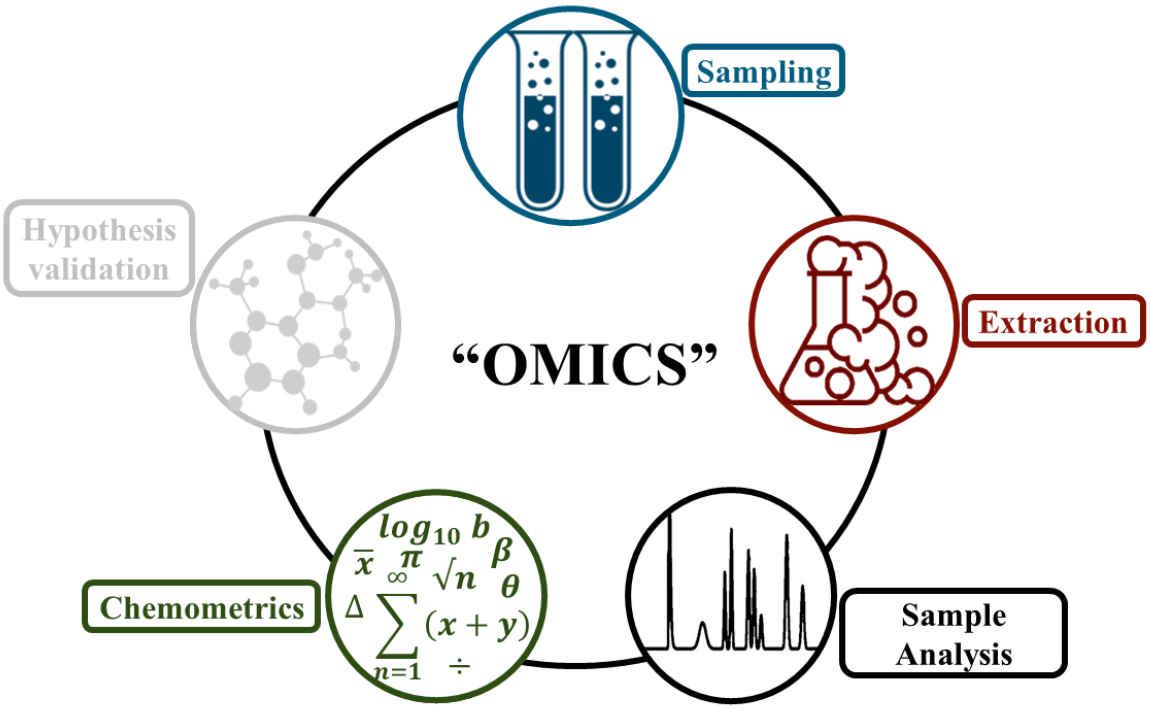


Figure 1.2: Generalized process for soil metabolomics and sedimentomics. Figure modified from graphical abstract to Bell and Blais (2019).

3.2. High resolution mass spectrometry

HRMS instruments are critically important for untargeted analyses because they allow for the separation of many different signals (Resolution > 10,000 full width half maximum) within the instrument with a high degree of accuracy (±0.001 Da) and sensitivity (Hollender et al., 2017; Krauss et al., 2010). HRMS is often combined with soft ionization techniques (i.e. electrospray ionization (ESI), atmospheric pressure photoionization (APPI), or atmospheric pressure chemical ionization (APCI)) so that when the ions enter the detector they will not be extensively fragmented and this often allows for the determination of the parent ion. The parent ion provides key information for structural annotation because the mass of the compound can be derived from *m/z* of the parent ion, and structural annotation is critical for untargeted analyses and hypothesis validation.

Resolution is a critical parameter of HRMS instruments. Resolution is the mass of the ion (m) divided by Δm (Hernández et al., 2012). However, different methods for determining Δm exist, which has led to some ambiguity in the industry. For instance, Δm can be defined using the full width half maximum (FWHM) or the percent valley method, and this can produce very different resolution and/or resolving power results (Murray et al., 2013; Urban et al., 2014); thus, it is important to define which method was used to calculate the instrumental resolution. Using FWHM, the minimum resolution for an HRMS instrument is 10,000 and mass analyzers in this class include: magnetic sector (10,000-80,000 resolution), time of flight (TOF; 20,000-80,000 resolution), orbitrap (up to 500,000 resolution at 200 m/z), and Fourier-transform ion cyclotron resonance (FT-ICR; up to 1,000,000 resolution) (Hernández et al., 2012; Krauss et al., 2010; Špánik and Machyňáková, 2018). Specifically, for this thesis, TOF and orbitrap instruments were used.

TOF and orbitrap HRMS instruments can be used for untargeted analyses of complex matrices; however, they are fundamentally different mass spectrometers. TOF mass analyzers simultaneously accelerate multiple ions after ionization. The ions enter a field free region (also known as the flight tube) and the velocity of the ion in the free field depends on the *m/z* of the ion, whereby smaller ions hit the detector first (Fjeldsted, 2016). The original TOF instruments had linear flight tubes and resolution was poor because resolution is a function of flight tube length. Modern TOF instruments have an orthogonal flight tube with a reflectron to extend the flight tube length and improve resolution (Watson and Sparkman, 2008). Quadrupoles can be added before the TOF mass analyzer, the so called QTOF, to add mass filtering and/or add collision capabilities to fragment ions for annotation (Allen and McWhinney, 2019). The advantages of the QTOF include its theoretically unlimited mass range, simplicity, and ruggedness (Skoog et al., 2016). Orbitaps are essentially ion traps, where the injected ions are injected

between two electrodes (i.e. the outer barrel electrode and the inner spindle electrode) in a static electric potential (Skoog et al., 2016; Watson and Sparkman, 2008). In this system the electric field is inhomogeneous with nonparallel vectors, which causes mass dependent oscillations around the spindle electrode (Watson and Sparkman, 2008). The frequency of this mass dependent oscillation are transformed in mass spectra using fast Fourier transform (Martins et al., 2016; Watson and Sparkman, 2008). Orbitraps can also be paired with quadrupoles before the mass analyzer. An LTQ Orbitrap XL has a collision cell, and the Q Exactive Orbitrap has a mass filter quadrupole and an octapole which drastically increased the mass range of the instrument (Martins et al., 2016). The advantages of orbitraps are their much higher resolution, greater accuracy, and wider linear dynamic range than the TOF family of mass analyzers (Aydoğan, 2020; Skoog et al., 2016).

4. OMICS APPLICATIONS TO ADDRESS KNOWLEDGE GAPS IN SOM AND SedOM C CYCLING

4.1. Soil metabolomics applications

Soil metabolomics is the study of SOM using “Omics” methods. SOM C cycling occurs faster than in sedOM so often the focus of soil metabolomics studies is on the microbially accessible fraction of SOM. The microbially accessible fraction is considered to be the water extractable fraction (WEOM) also known as dissolved organic matter (DOM). The applications of studying the WEOM fraction are myriad in climate change research and in agricultural research. The biggest mystery is the composition of SOM, SOM is very difficult to study and there are many gaps in our theories on the rate of SOM decomposition and how SOM composition (also known as litter quality) affects the mineralization – stabilization equilibrium in SOM. Further, we have even less understanding how changes in climate (i.e. warming weather, land use changes, thawing permafrost, etc.) or agricultural practices (i.e. tilling, farming

intensification, nutrient addition, etc.) affect SOM composition, and then ultimately what impact these changes in SOM composition have on the global biological C cycle. Soil is a major terrestrial source of global C and many initiatives are in place to increase C sequestration in soil. For instance, the "4 per mille" project includes 100+ nations who have committed to increasing the quantity of C sequestered in the soil by 4‰ in order to halt the annual increase in atmospheric CO_2 (Minasny et al., 2017). Whether this initiative is feasible remains under debate (Amundson and Biardeau, 2019, 2018; Loisel et al., 2019; Minasny et al., 2017); however, the development of soil metabolomics and its insight into C cycling in the soil may help facilitate this initiative by building a better understanding of how more C can be sequestered in the soil to help fight climate change.

4.2. Sedimentomics applications

Sedimentomics is the study of SedOM using "Omics" methods. SedOM is a critically understudied component of the biological C cycle. Recent research has indicated that freshwater sedOM is a large C sink and the composition of sedOM and the biotic and abiotic processes influencing sedOM are understudied when compared to SOM. Sedimentomics can be used to address knowledge gaps in sedOM biological C cycling. In addition, the ability of most freshwater lake sediment to store C has another application, the preservation of paleolimnological biomarkers. Paleolimnology is the study of old organic matter that has been stratified into the sediments of lakes. By taking lake sediment cores, the cores can be separated into intervals that can be radioisotopically dated, and specific compounds from the sedOM, also known as biomarkers, can be extracted and quantified to see how they changed over time. These biomarkers can be paleoecological (i.e. track changes in plant productivity over time), paleoclimatogical (i.e. track changes in regional temperatures over time), paleofire, or even

archeological (i.e. track sewage input into lakes over time) proxies. The advantage of this application is to reconstruct the past and to establish baselines of what past environments were like before human influence. This information also provides critical reality checks against simulated models for global and ecosystem scale models (Dwivedi et al., 2019; Heiri et al., 2006; Schwörer et al., 2014; Todd-Brown et al., 2013; Wieder et al., 2018, 2013). Sedimentomics is useful to systematically discover new biomarkers preserved in the sedOM for paleolimnological applications. Thus, sedimentomics is a tool to investigate knowledge gaps in historical and contemporary sedOM C cycling.

4.3. Systematics

The future of "Omics" in soil and sediment lies in systematics. Currently, different "Omics" (i.e. metabolomics, genomics, metatranscriptomics, etc.) studies are performed separately; but to fully address knowledge gaps in soil and sediment C cycling a systems biology approach is required where these methods are integrated together. A systems biology approach, or systematics, focuses in the interaction between SOM/sedOM, microbes, and their surrounding environments (Jansson and Hofmockel, 2018; Schmidt et al., 2011). It integrates this information to investigate how changes in the environment affects soil/sediment microbial communities (i.e. metagenomics), functionality (i.e. metatranscriptomics and metaproteomics), and C cycling dynamics (i.e. soil metabolomics and sedimentomics). However, there are current methodological problems with applying a systematics approach. Integrating the data from multiple "Omics" experiments is computationally expensive and few truly integrative analyses and visualizations have been developed for systematics. In addition, soil "Omics" and sediment "Omics" studies are still young and employ state-of-the-art technology, thus further methodological advances are needed to standardize these methods. This thesis focuses on

improving sedimentomics and soil metabolomics methods so that they can be used in sediment and soil systematics studies in the future.

5. IN SUMMARY

The purpose of this work was to develop statistical and analytical methods for untargeted analysis (i.e. sedimentomics and soil metabolomics) of complex environmental matrices. Specific objectives included: (1) Develop new extraction methods for untargeted analyses for soil metabolomics and sedimentomics that include appropriate data handling and quality control measures; (2) Investigate novel applications of untargeted analyses for environmental classification and monitoring; (3) Investigate the parameters influencing soil metabolomics extractions.

This thesis is a collection of manuscripts. CHAPTER TWO is a comprehensive review and a proposed workflow for future soil metabolomics and sedimentomics studies published in *Science of the Total Environment*. The objective of this review was to present the application of "Omics" to soil scientists, paleolimnologists, etc. with a detailed guide on how to design and implement soil metabolomics and sedimentomics studies. CHAPTER THREE is about the application of sedimentomics to systematically discover new biomarkers for Arctic paleoecology, and it was published in *Science of the Total Environment*. The objective of this research was to test the viability of sedimentomics as a systematic biomarker discovery method, and to discover discriminant biomarkers between boreal and tundra regions. CHAPTER FOUR was published in *Science of the Total Environment* and describes the spatial variation in freshwater sedOM composition and the potential application of sedimentomics to environmental monitoring. I wanted to determine if sedOM composition is a spatially dependent and if this concept can be used to draw ecologically relevant boundaries. CHAPTER FIVE is submitted to *Soil Biology and Biochemistry*. It

describes method development and the novel quality control procedures for soil metabolomics. The objective of CHAPTER FIVE was to explore factors influencing soil metabolomics extractions and establish quality control procedures to control extraction variability and improve data quality. The final chapter, CHAPTER SIX, will contain the conclusions from each chapter and the overall conclusions of the thesis.

In summary, I demonstrated that sedimentomics and soil metabolomics are valid research methods that can be used in the future to address knowledge gaps in the biological C cycle. This work discusses and presents novel quality control methods and applies statistical methods in new ways to soil metabolomics and sedimentomics.

REFERENCES

Allen, D.R., McWhinney, B.C., 2019. Quadrupole Time-of-Flight Mass Spectrometry: A Paradigm Shift in Toxicology Screening Applications. The Clinical Biochemist Reviews 40, 135–146. doi:10.33176/AACB-19-00023

Amundson, R., Biardeau, L., 2019. Reply to Loisel et al.: Soil in climate mitigation and adaptation. Proceedings of the National Academy of Sciences 201905360. doi:10.1073/pnas.1905360116

Amundson, R., Biardeau, L., 2018. Opinion: Soil carbon sequestration is an elusive climate mitigation tool. Proceedings of the National Academy of Sciences 115, 11652–11656. doi:10.1073/pnas.1815901115

Aydoğan, C., 2020. Recent advances and applications in LC-HRMS for food and plant natural products: a critical review. Analytical and Bioanalytical Chemistry 412, 1973–1991. doi:10.1007/s00216-019-02328-6

Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. European Journal of Soil Science 47, 151–163. doi:10.1111/j.1365-2389.1996.tb01386.x

Bell, M., Blais, J.M., 2019. "-Omics" workflow for paleolimnological and geological archives: A review. Science of The Total Environment 672, 438–455. doi:10.1016/j.scitotenv.2019.03.477

Dean, W.E., Gorham, E., 1998. Magnitude and significance of carbon burial in lakes, reservoirs, and peatlands. Geology 26, 535–538. doi:10.1130/0091-7613(1998)026<0535:MASOCB>2.3.CO;2

Dwivedi, D., Tang, J., Bouskill, N., Georgiou, K., Chacon, S.S., Riley, W.J., 2019. Abiotic and Biotic Controls on Soil Organo–Mineral Interactions: Developing Model Structures to Analyze Why Soil Organic Matter Persists. Reviews in Mineralogy and Geochemistry 85, 329–348. doi:10.2138/rmg.2019.85.11

Fjeldsted, J.C., 2016. Chapter 2 - Advances in Time-of-Flight Mass Spectrometry, in: Pérez, S., Eichhorn, P., Barceló, D. (Eds.), Comprehensive Analytical Chemistry, Applications of Time-of-Flight and Orbitrap Mass Spectrometry in Environmental, Food, Doping, and Forensic Analysis. Elsevier, pp. 19–49. doi:10.1016/bs.coac.2016.01.002

Foster, G.L., Royer, D.L., Lunt, D.J., 2017. Future climate forcing potentially without precedent in the last 420 million years. Nature Communications 8, 14845. doi:10.1038/ncomms14845

Friedlingstein, P., Jones, M.W., O'Sullivan, M., Andrew, R.M., Hauck, J., Peters, G.P., Peters, W., Pongratz, J., Sitch, S., Quéré, C.L., Bakker, D.C.E., Canadell, J.G., Ciais, P., Jackson, R.B., Anthoni, P., Barbero, L., Bastos, A., Bastrikov, V., Becker, M., Bopp, L., Buitenhuis, E., Chandra, N., Chevallier, F., Chini, L.P., Currie, K.I., Feely, R.A., Gehlen, M., Gilfillan, D., Gkritzalis, T., Goll, D.S., Gruber, N., Gutekunst, S., Harris, I., Haverd, V., Houghton, R.A., Hurtt, G., Ilyina, T., Jain, A.K., Joetzjer, E., Kaplan, J.O., Kato, E., Klein Goldewijk, K., Korsbakken, J.I., Landschützer, P., Lauvset, S.K., Lefèvre, N., Lenton, A., Lienert, S., Lombardozzi, D., Marland, G., McGuire, P.C., Melton, J.R., Metzl, N., Munro, D.R., Nabel, J.E.M.S., Nakaoka, S.-I., Neill, C., Omar, A.M., Ono, T., Peregon, A., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Séférian, R., Schwinger, J., Smith, N., Tans, P.P., Tian, H., Tilbrook, B., Tubiello, F.N., Werf, G.R. van der, Wiltshire, A.J., Zaehle, S., 2019. Global Carbon Budget 2019. Earth System Science Data 11, 1783–1838. doi:https://doi.org/10.5194/essd-11-1783-2019

Heiri, C., Bugmann, H., Tinner, W., Heiri, O., Lischke, H., 2006. A model-based reconstruction of Holocene treeline dynamics in the Central Swiss Alps. Journal of Ecology 94, 206–216. doi:10.1111/j.1365-2745.2005.01072.x

Hernández, F., Sancho, J.V., Ibáñez, M., Abad, E., Portolés, T., Mattioli, L., 2012. Current use of high-resolution mass spectrometry in the environmental sciences. Analytical and Bioanalytical Chemistry 403, 1251–1264. doi:10.1007/s00216-012-5844-7

Hollender, J., Schymanski, E.L., Singer, H.P., Ferguson, P.L., 2017. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environmental Science & Technology 51, 11505–11512. doi:10.1021/acs.est.7b02184

Jackson, R.B., Lajtha, K., Crow, S.E., Hugelius, G., Kramer, M.G., Piñeiro, G., 2017. The Ecology of Soil Carbon: Pools, Vulnerabilities, and Biotic and Abiotic Controls. Annual Review of Ecology, Evolution, and Systematics 48, 419–445. doi:10.1146/annurev-ecolsys-112414-054234

Jansson, J.K., Hofmockel, K.S., 2018. The soil microbiome — from metagenomics to metaphenomics. Current Opinion in Microbiology 43, 162–168. doi:10.1016/j.mib.2018.01.013

Jobbágy, E.G., Jackson, R.B., 2000. The Vertical Distribution of Soil Organic Carbon and Its Relation to Climate and Vegetation. Ecological Applications 10, 423–436. doi:10.1890/1051-0761(2000)010[0423:TVDOSO]2.0.CO;2

Killops, S., Killops, V., 2013a. Long-Term Fate of Organic Matter in the Geosphere, in: Introduction to Organic Geochemistry. John Wiley & Sons, Ltd, pp. 117–165. doi:10.1002/9781118697214.ch4

Killops, S., Killops, V., 2013b. The Carbon Cycle and Climate, in: Introduction to Organic Geochemistry. John Wiley & Sons, Ltd, pp. 246–294. doi:10.1002/9781118697214.ch6

Kindler, R., Miltner, A., Richnow, H., Kastner, M., 2006. Fate of gram-negative bacterial biomass in soil—mineralization and contribution to SOM. Soil Biology and Biochemistry 38, 2860–2870. doi:10.1016/j.soilbio.2006.04.047

Krauss, M., Singer, H., Hollender, J., 2010. LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. Analytical and Bioanalytical Chemistry 397, 943–951. doi:10.1007/s00216-010-3608-9

Lal, R., 2019. Climate Change and the Global Soil Carbon Stocks, in: Soil and Climate. Taylor & Francis Group, LLC, Boca Raton, FL, pp. 419–426. doi:10.1201/b21225-16

Lehmann, J., Kleber, M., 2015. The contentious nature of soil organic matter. Nature. doi:10.1038/nature16069

Loisel, J., Casellas Connors, J.P., Hugelius, G., Harden, J.W., Morgan, C.L., 2019. Soils can help mitigate CO_2 emissions, despite the challenges. Proceedings of the National Academy of Sciences 201900444. doi:10.1073/pnas.1900444116

Luo, Y., Ahlström, A., Allison, S.D., Batjes, N.H., Brovkin, V., Carvalhais, N., Chappell, A., Ciais, P., Davidson, E.A., Finzi, A., Georgiou, K., Guenet, B., Hararuk, O., Harden, J.W., He, Y., Hopkins, F., Jiang, L., Koven, C., Jackson, R.B., Jones, C.D., Lara, M.J., Liang, J., McGuire, A.D., Parton, W., Peng, C., Randerson, J.T., Salazar, A., Sierra, C.A., Smith, M.J., Tian, H., Todd-Brown, K.E.O., Torn, M., Groenigen, K.J. van, Wang, Y.P., West, T.O., Wei, Y., Wieder, W.R., Xia, J., Xu, Xia, Xu, Xiaofeng, Zhou, T., 2016. Toward more realistic projections of soil carbon dynamics by Earth system models. Global Biogeochemical Cycles 30, 40–56. doi:10.1002/2015GB005239

Martins, C.P.B., Bromirski, M., Prieto Conaway, M.C., Makarov, A.A., 2016. Chapter 1 - Orbitrap Mass Spectrometry: Evolution and Applicability, in: Pérez, S., Eichhorn, P., Barceló, D. (Eds.), Comprehensive Analytical Chemistry, Applications of Time-of-Flight and Orbitrap Mass Spectrometry in Environmental, Food, Doping, and Forensic Analysis. Elsevier, pp. 3–18. doi:10.1016/bs.coac.2016.01.001

Masoom, H., Courtier-Murias, D., Farooq, H., Soong, R., Kelleher, B.P., Zhang, C., Maas, W.E., Fey, M., Kumar, R., Monette, M., 2016. Soil Organic Matter in Its Native State: Unravelling the Most Complex Biomaterial on Earth. Environmental Science & Technology 50, 1670–1680.

Mendonça, R., Müller, R.A., Clow, D., Verpoorter, C., Raymond, P., Tranvik, L.J., Sobek, S., 2017. Organic carbon burial in global lakes and reservoirs. Nature Communications 8, 1694. doi:10.1038/s41467-017-01789-6

Meyers, P.A., Ishiwatari, R., 1993. Lacustrine Organic Geochemistry - An Overview of Indicators of Organic Matter Sources and Diagenesis in Lake Sediments. Organic Geochemistry 20, 867–900. doi:10.1016/0146-6380(93)90100-p

Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B.S., Field, D.J., Gimona, A., Hedley, C.B., Hong, S.Y., Mandal, B., Marchant, B.P., Martin, M., McConkey, B.G., Mulder, V.L., O'Rourke, S., Richer-de-Forges, A.C., Odeh, I., Padarian, J., Paustian, K., Pan, G., Poggio, L., Savin, I., Stolbovoy, V., Stockmann, U., Sulaeman, Y., Tsui, C.-C., Vågen, T.-G., van Wesemael, B., Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma 292, 59–86. doi:10.1016/j.geoderma.2017.01.002

Murray, K.K., Boyd, R.K., Eberlin, M.N., Langley, G.J., Li, L., Naito, Y., 2013. Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). Pure and Applied Chemistry 85, 1515–1609. doi:10.1351/PAC-REC-06-04-06

Pearson, R.G., Phillips, S.J., Loranty, M.M., Beck, P.S.A., Damoulas, T., Knight, S.J., Goetz, S.J., 2013. Shifts in Arctic vegetation and associated feedbacks under climate change. Nature Climate Change 3, 673–677. doi:10.1038/nclimate1858

Prentice, I., Farquhar, G., Fasham, M., Goulden, M., Heimann, M., Jaramillo, V., Kheshgi, H., Quéré, C., Scholes, R., Wallace, D., 2001. The carbon cycle and atmospheric carbon dioxide, in: Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, USA, pp. 183–237.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M.D., Muñoz-Marí, J., van Nes, E.H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., Zscheischler, J., 2019. Inferring causation from time series in Earth system sciences. Nature Communications 10. doi:10.1038/s41467-019-10105-3

Schlesinger, W.H., 1990. Evidence from chronosequence studies for a low carbon-storage potential of soils. Nature 348, 232–234. doi:10.1038/348232a0

Schmidt, M.W.I., Torn, M.S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I.A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D.A.C., Nannipieri, P., Rasse, D.P., Weiner, S., Trumbore, S.E., 2011. Persistence of soil organic matter as an ecosystem property. Nature 478, 49–56. doi:10.1038/nature10386

Schuur, E. a. G., McGuire, A.D., Schädel, C., Grosse, G., Harden, J.W., Hayes, D.J., Hugelius, G., Koven, C.D., Kuhry, P., Lawrence, D.M., Natali, S.M., Olefeldt, D., Romanovsky, V.E., Schaefer, K., Turetsky, M.R., Treat, C.C., Vonk, J.E., 2015. Climate change and the permafrost carbon feedback. Nature 520, 171–179. doi:10.1038/nature14338

Schweigert, M., Herrmann, S., Miltner, A., Fester, T., Kästner, M., 2015. Fate of ectomycorrhizal fungal biomass in a soil bioreactor system and its contribution to soil organic matter formation. Soil Biology and Biochemistry 88, 120–127. doi:10.1016/j.soilbio.2015.05.012

Schwörer, C., Henne, P.D., Tinner, W., 2014. A model-data comparison of Holocene timberline changes in the Swiss Alps reveals past and future drivers of mountain forest dynamics. Global Change Biology 20, 1512–1526. doi:10.1111/gcb.12456

Simpson, A.J., Simpson, M.J., Smith, E., Kelleher, B.P., 2007. Microbially Derived Inputs to Soil Organic Matter: Are Current Estimates Too Low? Environmental Science & Technology 41, 8070–8076. doi:10.1021/es071217x

Skoog, D.A., Holler, F.J., Crouch, S.R., 2016. Chapter 20 - Molecular mass spectrometry, in: Principles of Instrumental Analysis. Cengage learning, Boston, USA, pp. 501–536.

Sobek, S., Durisch-Kaiser, E., Zurbrügg, R., Wongfun, N., Wessels, M., Pasche, N., Wehrli, B., 2009. Organic carbon burial efficiency in lake sediments controlled by oxygen exposure time and sediment source. Limnology and Oceanography 54, 2243–2254. doi:10.4319/lo.2009.54.6.2243

Špánik, I., Machyňáková, A., 2018. Recent applications of gas chromatography with high-resolution mass spectrometry. Journal of Separation Science 41, 163–179. doi:10.1002/jssc.201701016

Todd-Brown, K.E.O., Randerson, J.T., Post, W.M., Hoffman, F.M., Tarnocai, C., Schuur, E. a. G., Allison, S.D., 2013. Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. Biogeosciences 10, 1717–1736. doi:https://doi-org.proxy.bib.uottawa.ca/10.5194/bg-10-1717-2013

Urban, J., Afseth, N.K., Štys, D., 2014. Fundamental definitions and confusions in mass spectrometry about mass assignment, centroiding and resolution. TrAC Trends in Analytical Chemistry 53, 126–136. doi:10.1016/j.trac.2013.07.010

Ussiri, D.A.N., Lal, R., 2017a. Introduction to Global Carbon Cycling: An Overview of the Global Carbon Cycle, in: Ussiri, D.A.N., Lal, R. (Eds.), Carbon Sequestration for Climate Change Mitigation and Adaptation. Springer International Publishing, Cham, pp. 61–76. doi:10.1007/978-3-319-53845-7_3

Ussiri, D.A.N., Lal, R., 2017b. Historical Perspectives of the Global Carbon Cycle, in: Ussiri, D.A.N., Lal, R. (Eds.), Carbon Sequestration for Climate Change Mitigation and Adaptation. Springer International Publishing, Cham, pp. 103–161. doi:10.1007/978-3-319-53845-7_5

Ussiri, D.A.N., Lal, R., 2017c. Mitigation of Climate Change: Introduction, in: Ussiri, D.A.N., Lal, R. (Eds.), Carbon Sequestration for Climate Change Mitigation and Adaptation. Springer International Publishing, Cham, pp. 287–325. doi:10.1007/978-3-319-53845-7_8

Ussiri, D.A.N., Lal, R., 2017d. The Global Carbon Inventory, in: Ussiri, D.A.N., Lal, R. (Eds.), Carbon Sequestration for Climate Change Mitigation and Adaptation. Springer International Publishing, Cham, pp. 77–102. doi:10.1007/978-3-319-53845-7_4

Watson, J.T., Sparkman, O.D., 2008. The Mass Spectrometer, in: Introduction to Mass Spectrometry. John Wiley & Sons, Ltd, pp. 53–172. doi:10.1002/9780470516898.ch2

Wieder, W.R., Bonan, G.B., Allison, S.D., 2013. Global soil carbon projections are improved by modelling microbial processes. Nature Climate Change 3, 909–912. doi:10.1038/nclimate1951

Wieder, W.R., Hartman, M.D., Sulman, B.N., Wang, Y.-P., Koven, C.D., Bonan, G.B., 2018. Carbon cycle confidence and uncertainty: Exploring variation among soil biogeochemical models. Global Change Biology 24, 1563–1579. doi:10.1111/gcb.13979

Young, I.M., Crawford, J.W., 2004. Interactions and Self-Organization in the Soil-Microbe Complex. Science 304, 1634–1637. doi:10.1126/science.1097394

CHAPTER 2: "-Omics" workflow for paleolimnological and geological archives: A review

Madison A. Bell[1], Jules M. Blais[1*]

[1]Laboratory for the Analysis of Natural and Synthetic Environmental Toxicants, Department of Biology, University of Ottawa, Ottawa ON, K1N 6N5, Canada

[*]Corresponding Author

Jules Blais (jules.blais@uOttawa.ca)

STATEMENT OF CONTRIBUTIONS

Madison A. Bell: conceptualization, writing, visualizations, formal analysis, and literature search. Jules M. Blais: supervision, reviewing, funding acquisition, and corresponding author. *Science of the Total Environment* published this chapter:

Bell, M.A., Blais, J.M. (2019) "-Omics" workflow for paleolimnological and geological archives: A review. Science of the Total Environment. 672: 438-455.

https://doi.org/10.1016/j.scitotenv.2019.03.477

ABSTRACT

"-Omics" is a powerful screening method with applications in molecular biology, toxicology, wildlife biology, natural product discovery, and many other fields. Genomics, proteomics, metabolomics, and lipidomics are common examples included under the "-omics" umbrella. This screening method uses combinations of untargeted, semi-targeted, and targeted analyses paired with data mining to facilitate researchers' understanding of the genome, proteins, and small organic molecules in biological systems. Recently, however, the use of "-omics" has expanded into the fields of geology, specifically petrology, and paleolimnology. Specifically, untargeted analyses stand to transform these fields as petroleomics, and sediment-"omics" become more prevalent. "-Omics" facilitates the visualization of small molecule profiles from environmental matrices (i.e. oil and sediment). Small molecule profiles can provide improved understanding of small molecules distributions throughout the environment, and how those compositions can change depending on conditions (i.e. climate change, weathering, etc.). "-Omics" also facilities discovery of next-generation biomarkers that can be used for oil source identification and as proxies for reconstructing past environmental changes. Untargeted analyses paired with data mining and multivariate statistical analyses represents a powerful suite of tools for hypothesis generation, and new method development

1. INTRODUCTION

1.1. "-Omics"

The field of "-omics" research is relatively young, starting in the 80's with genomics (Robertson, 2005; Yadav, 2007). The origin of the term "- omics" may originate from the Sanskrit "OM" meaning completeness (Yadav, 2007). Thus, when combined with the targets under study, it implies "all genes"

for genomics, or “all lipids” for lipidomics. Currently, the field of “-omics” is extensive and includes: genomics, transcriptomics, proteomics, metabolomics, lipidomics, metallomics, and environment-omics applied to environmental samples. The focus of this review is on environmental “-omics”, and how to apply “-omics” techniques to environmental matrices like sediment, soil, peat, oil, etc. “-Omics” is a holistic approach to studying biological systems and refers to an interdisciplinary methodology that incorporates analytical chemistry, statistical modelling and multivariate analyses, and knowledge of the systems under survey. The advantage of using “-omics” is the generation of novel hypotheses. “-Omics” investigations use untargeted analyses to generate a “broad compositional” or “global” view of the system. This untargeted global survey of the targeted molecules (i.e. commonly DNA, proteins, or small molecules) allows for the assessment of broad patterns based on compositional changes. Successive data mining of these patterns can pinpoint where these compositional changes are, which can lead to the identification of potential biomarkers. Biomarkers are organic molecules that are significantly different in either abundance or presence between the sampled groups. Thus, the utility of an untargeted “-omics” approach is to generate new hypotheses regarding these “global” patterns and potential biomarkers. Semi-targeted and targeted analyses explore the new hypotheses to determine if they can stand up to rigorous scientific validation. For instance, a targeted analysis of a potential biomarker would answer more specific questions like: Does this biomarker only appear in oil from this region? Or does this biomarker preserve over time in sediment samples? Ideally, follow-up targeted analyses validate the hypotheses generated with the untargeted “global” analysis. The “-omics” framework of untargeted “global” analyses followed by semi-targeted and targeted analysis is widely applicable to many fields.

1.2. Current applications of "-omics" in environmental sciences

Currently, the use of "-omics" in environmental fields is burgeoning. There are four main areas in which this methodology has been applied to sediment, soil, and oil matrices: (1) Untargeted screening for environmental contaminants; (2) Investigation of soil and sediment bacterial community metabolites and carbon cycling; (3) Untargeted screening of organic matter (OM) composition leading to novel biomarkers for paleolimnological applications; and (4) petroleomics or oil fingerprinting.

Semi-targeted and untargeted screening approaches have been used on sediment or soil samples for pollutants analysis. There are many studies documenting the screening of sediment for flame retardants (Gustavsson et al., 2018), sediment-bound and unbound xenobiotics (Kronimus and Schwarzbauer, 2007), and anthropogenic contaminants in wastewater, sludge and sediment (Grigoriadou and Schwarzbauer, 2011; Rostkowski et al., 2013; Veenaas and Haglund, 2017). Many of these untargeted methods are not considered "-omics" methods, because they do not incorporate discriminant statistical analyses into their studies and are not focussed on hypothesis generation, but rather the confirmation pollutants are present. However, current untargeted and semi-targeted pollutant screening methodologies paired with "-omics" statistical analyses (see *Section 4.3.*) may facilitate pollutant discovery and source identifications with the appropriate study design and extraction techniques to compensate for complex matrices and trace levels of pollutants.

Soil and sediment microbial communities and metabolites can also be investigated with "-omics". Beale et al. (2017) demonstrated a multi-"omics" approach combining genomics and sediment-"omics" to assess the impact of environmental toxicants on bacterial colonies. A later publication compared the effect of changes in nutrients, OM, light and pollutants on bacterial populations, and found precipitation had a major influence on bacterial community structure and linked OM composition

to bacterial metabolic pathways (Beale et al., 2018). Swenson et al. (2018) also linked the impact of wetting events on soil microbial community structures with changes in water extractable metabolites. Furthermore, "-omics" can be paired with stable isotope labelling to investigate microbial carbon cycling (Swenson et al., 2015). They were able to differentiate the isotope-labelled microbially-derived metabolites from endogenous soil OM and determine the extent of microbial activity (Swenson et al., 2015).

Other OM components, aside from microbial metabolites and contaminants, have undergone "-omics". For instance, an untargeted approach was used to investigate changes in soil OM composition after fire-exposures and they were able to isolate the fire affected organic molecules in different particle size fractions (Jiménez-Morillo et al., 2018). Similar methods have also been applied to soil OM (Kujawinski, 2011; X.M. Li et al., 2018; Z. Li et al., 2018; Ward and Cory, 2015). Sedimentary OM have had environmental changes assessed in sediments using "-omics". Untargeted "-omics" led to the discovery of novel proxies for climate reconstructions. Farrés et al. (2015a) used an "-omics" method on marine sediment where they correlated long chain n-alkanes, long and short chain n-alkan-1-ols, alkenols, cholesterol, and squalene with down-core sea surface temperatures.

Lastly, "-omics" has been applied to oil. Petroleomics is useful for both identifying biomarkers specific to an oil deposit, and also for determining sources of oil spills (Wang et al., 2006). Thus far, untargeted "-omics" techniques have differentiated Brazilian crude oils (Kiepper et al., 2014; Laakia et al., 2017; Prata et al., 2016), Columbian oils (Silva et al., 2011), Alberta oil sands (Yang et al., 2011), Jianghan basin, and Nanyan basin oils (Zhang et al., 2015). It can also be applied to oil weathering processes. Hall et al. (2013) used two-dimensional gas chromatography (GCxGC) mass spectrometry to

differentiate Deepwater Horizon oil post- and pre-weathering. They were able to identify the precursors of some of the oxygenated weathering products using partial least squares regression (Hall et al., 2013).

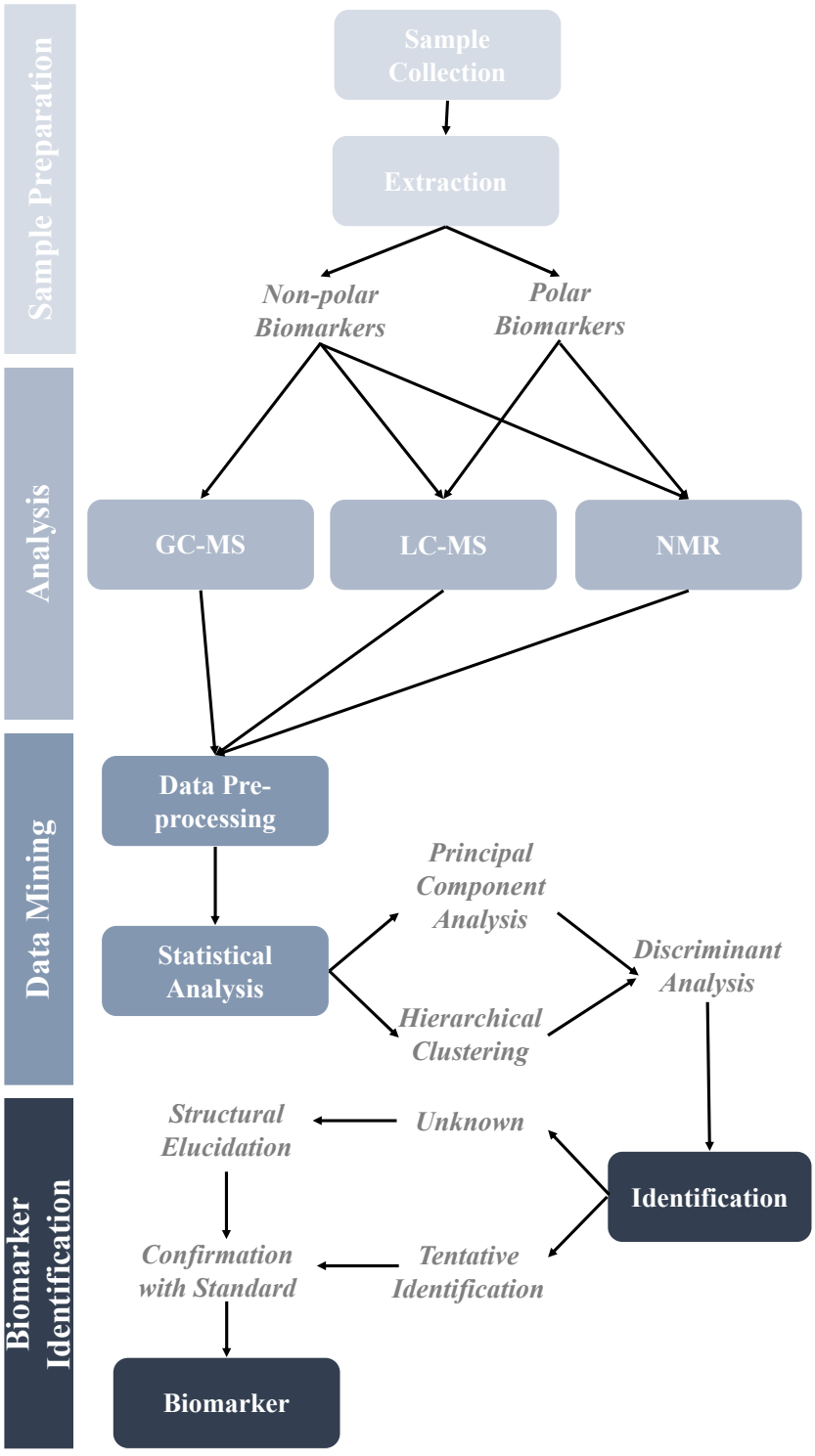


Figure 2.1: A generalized workflow for small molecule untargeted "omics" strategies.

1.3. Purpose

Currently, "-omics" is widely applied in many fields, but we emphasize methodologies for small organic molecules (< 1000 kDa) located in environmental matrices (i.e. soil, sediment, and oil). The purpose of this review is to provide an "-omics" workflow for paleolimnological and geological research. We will highlight important considerations at each step in the "-omics" workflow (Figure 2.1) including sample preparation, chemical analysis, data mining and structure identification. We emphasize the quality control steps required to produce reliable and statistically powerful data. In addition, each section cites more specific reviews for interested readers. The overarching objective is to elaborate on the potential of "-omics" research, and to establish practical protocols for future "-omics" studies on paleolimnological and geological samples.

2. SAMPLE PREPARATION

2.1. Experimental design

2.1.1. Source of experimental variation

"-Omics" can be susceptible to variability at all stages of the workflow (Dudzik et al., 2018; Martins et al., 2018; Szczepańska et al., 2018). Good experimental design requires the consideration of all stages of the workflow (Figure 2.1) before experimentation begins. This review emphasizes each stage of the workflow with the aim to reduce experimental variability.

There are five major types of variation for "-omics" in environmental matrices: (1) Sample variation – is due to samples' spatial relatedness (e.g. distances between sites or within cores), weathering, diagenesis, time between sample collections, environmental conditions during sample

collection and sample collection techniques. (2) Pre-analytical variation – is caused by inconsistent or ill-considered sample extraction procedures, chemical modifications during extraction, pre-extraction sample storage and post-extraction degradation during storage. (3) Analytical variation – is caused by the presence of plasticizers, instrumental variation, batch effects, matrix effects causing ion suppression or enhancement and carryover effects from one sample to the next. (4) Post-analytical variation – includes data handling errors, false discoveries, mis-annotation of compounds, and data preprocessing errors. Lastly, (5) Stochastic variation – variation due to random chance. Untargeted "-omics" techniques are sensitive to many of these errors due to the nature of trying to "see" as much as possible in one sample, so experimental design is critical to minimize variation.

2.1.2. Untargeted, semi-targeted and targeted analytical strategies

There are three types of analysis strategies: (1) untargeted; (2) semi-targeted; and (3) targeted (Figure 2.2). Untargeted strategies, also known as profiling, fingerprinting, or global assessments, are qualitative assessments of the molecular composition. The focus of an untargeted analysis is to see as many compounds as possible in an extract. Data mining identifies similarities and differences among the samples. Examples of relevant untargeted analyses include the work done by Farrés et al. (2015a, 2015b) on the discovery of climate biomarkers, and Alizadeh et al. (2017) who demonstrated a fractionated untargeted analysis of petroleum. This type of analysis is crucial for hypothesis generation and the discovery of potential biomarkers. Afterward, semi-targeted and targeted strategies validate the new hypotheses and biomarkers.

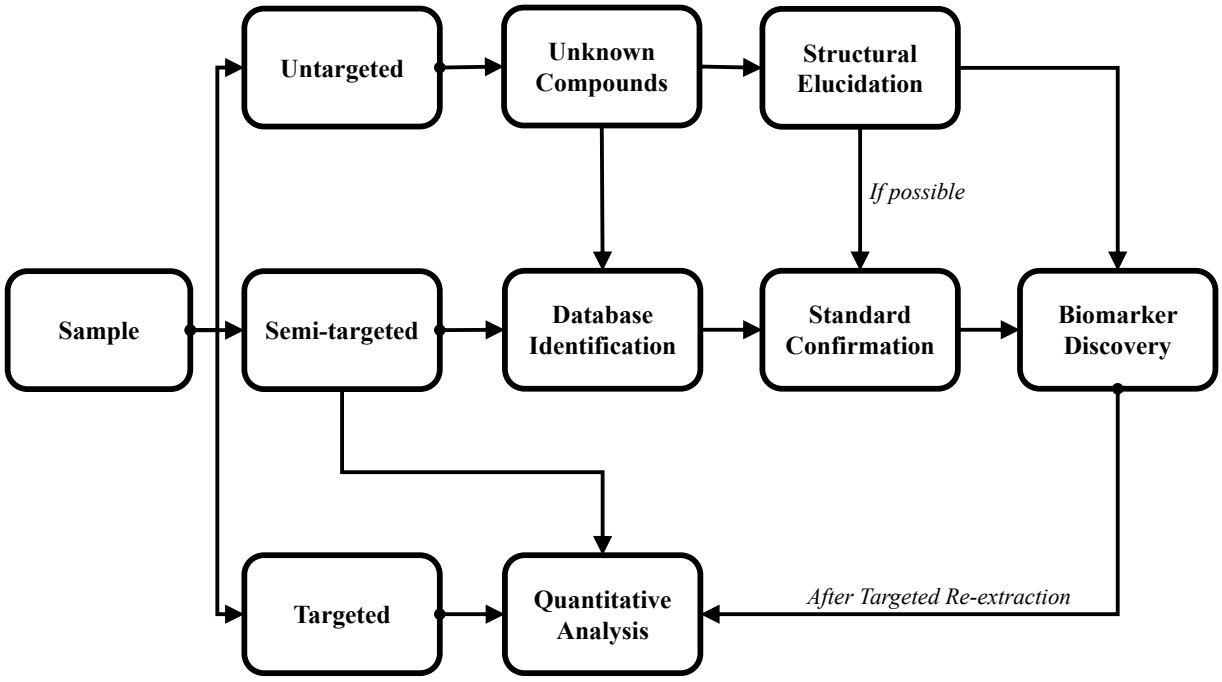


Figure 2.2: Comparison workflow for targeted, semi-targeted, and untargeted strategies.

Semi-targeted strategies can be quantitative, or at least semiquantitative. This strategy requires more sample preparation than untargeted approaches but has fewer analytical artifacts such as matrix effects, carryover, and coelution. In a semi-targeted analysis, the focus is on multiple known compounds, potentially from different compound classes. Examples of semi-targeted strategies include publications on the metabolomics profiles of polar bears (Morris et al., 2018) and screening organic contaminants (Bu et al., 2014; Gustavsson et al., 2018). The advantages of semi-targeted approaches are increased sensitivity, resolution, and quantification. The main disadvantage, when compared to an untargeted strategy, is the absence of novel compound discovery outside of the classes of interest.

Known molecular compounds undergo targeted analyses. Often there are multiple stages of sample clean-up, which will further reduce matrix effects, carryover, and coelution issues. Targeted analyses are quantitative with sensitive, precise, and accurate detection and quantification. These

targeted studies validate biomarkers for future studies. The disadvantage of targeted analyses is that new biomarker discovery does not occur, and they require labour intensive extractions, which increases time to analysis. This review focuses on untargeted strategies, but there are detailed reviews for semi-targeted and targeted workflows specifically (Cajka and Fiehn, 2016; Gorrochategui et al., 2016).

2.1.3. Sample collection

Sample collection is an important consideration of experimental design for untargeted "-omics". A clear question is necessary before sampling. If the question is relevant to the molecular fingerprinting of different samples (see Section 4.3.1.), then contrast between the sampled groups does not need to be as defined (i.e. sampling a gradient of impact). If the question is relevant to biomarker discovery, then maximizing the differences between "Group 1" and "Group 2" (i.e. impacted versus unimpacted) during sample collection is necessary. High sampling contrast also optimizes the ability of discriminant statistical analyses (see Section 4.3.3.) to isolate potential biomarkers from the large amounts of data collected.

For sediment studies, there are additional considerations for sample collection. Sediment OM undergoes diagenesis at the sediment surface, so the OM in the surface layer of a sediment core will not be the same as a layer within the anoxic layer after diagenesis (Meyers and Ishiwatari, 1993). There is also a temporal component as the sediment ages downcore; a single core can represent many years of sediment deposition. Thus, paleolimnologists must consider spatial sampling, but also temporal sampling as well. For example, if the research question addresses spatial changes, then temporal variation needs to be controlled by sampling layers in similar time frames. A proposed solution would be to use a top-bottom approach to account for diagenesis and environmental changes over time where a

surface sediment sample and a down-core sediment sample are simultaneously analyzed. If the research question relates to temporal changes over time (i.e. oil weathering over some duration or OM changes during diagenesis) then the spatial variation needs to be minimized like in any time series or repeated measure experiments (Berk et al., 2011). Farrés et al. (2015a) provides an example of monitoring climatic signals over time, but limited in space, to correlate biomarkers with sea surface temperatures.

In addition, the number of samples requires careful consideration. It is prudent to perform a statistical power analysis before the start of the study. Greater sample sizes result in a reduced standard error and better precision in the results (Dudzik et al., 2018; Krzywinski and Altman, 2013). Of course, practical, budgetary, and resource constraints need consideration, but a statistical power analysis can help balance ideal sample size and prudence. The result should be a statistical power of 0.8, which is where, if the experiment is repeated multiple times, then there will be a significant difference between the two groups 80% of the time (Xia and Wishart, 2016). There are different tools and tests that can be used to perform power analysis for "-omics" studies and for more information see Xia and Wishart (2016), Blaise et al. (2016) and Dudzik et al. (2018).

2.2. Extractions

2.2.1. Non-polar extractions

The choice of extraction method is important, but it is always a compromise. No single extraction will extract all compounds. This section explores non-polar (i.e. lipid, organic contaminants, large terpenes) and polar (i.e. amino acids, small terpenes, carbohydrates) compound extraction methods. Table 2.1 contains examples of extraction methods for "-omics" on environmental matrices.

A non-polar extraction method may be more relevant for paleolimnological and geological biomarkers. Non-polar compounds preserve better in the sediment, especially saturated and aromatic compounds (Meyers and Ishiwatari, 1993) and oil samples largely contain very non-polar compounds. To demonstrate this, the majority of current biomarkers for paleolimnology can be seen in Figure 2.3 in the context of their relative polarity and the solvent likely to extract them.

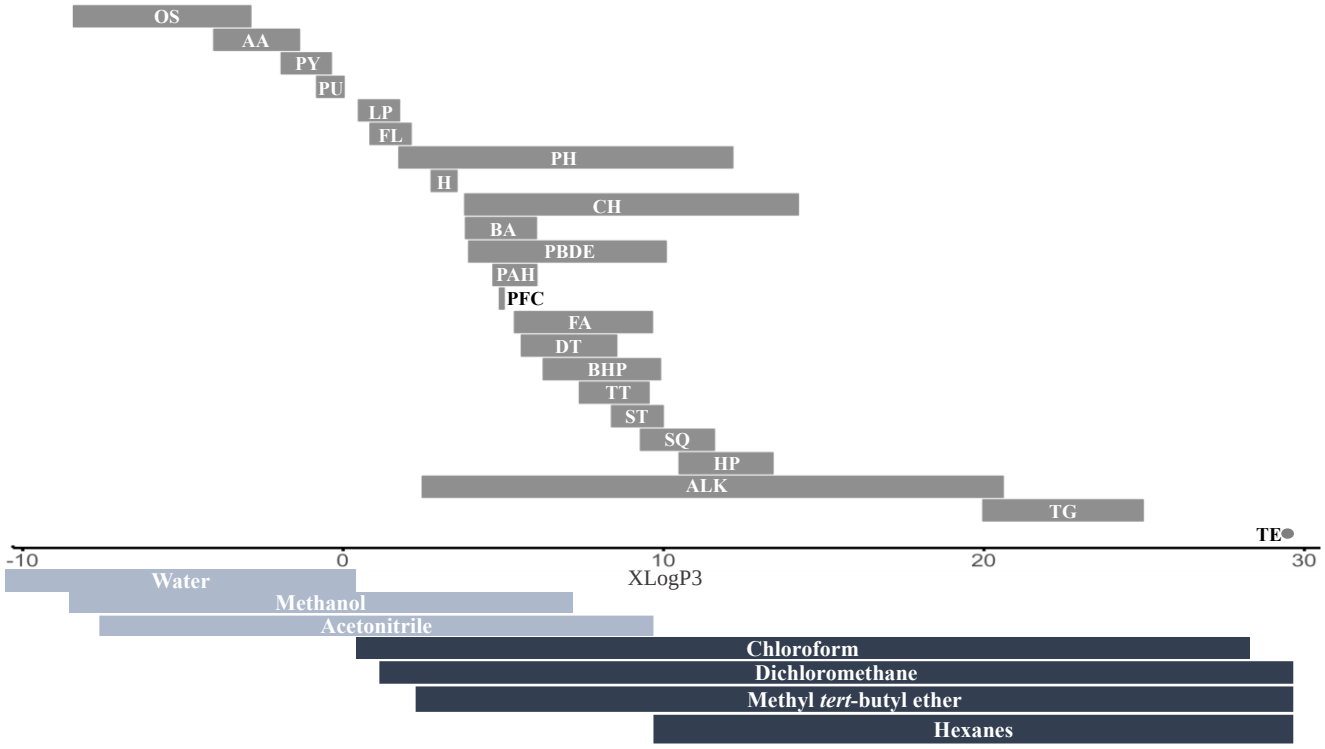


Figure 2.3: Solvent and common biomarkers chart. Each family of compounds is represented by a range of XLogP3 (calculated octanol/water coefficients where <0 is more polar and >0 is more non-polar) values found in the PubMed Database (Kim et al., 2016). The families do not contain all molecules within the family but rather a subsample chosen by relevance and different polarities to maximize range. Below the plot are the common solvents for "-omics" studies, and their approximate range of extractable metabolites. If the solvents overlap, then they are miscible. [a]Compound shortforms are: oligosaccharides (OS), amino acids (AA), pyrimidines (PY), purines (PU), lignin phenols (LP), flavonoids (FL), phthalates (PH), hormones (H), chlorins (CH), bile acids (BA), polybrominated diphenyl ethers (PBDE), polyaromatic hydrocarbons (PAH), polyfluorinated compounds (PFC), fatty acids (FA), diterpenoids (DT), bacteriohopanepolyols (BHP), triterpenoids (TT), sterols (ST), squalene-derivatives (SQ), saturated hopanes (HP), saturated alkanes (ALK), triacylglycerides (TG), and glycerol dialkyl glycerol tetraethers (TE).

Table 2.1: Examples of extraction and analysis protocols for non-polar compounds and polar compounds.

Matrix	Molecular Classes[a]	Extraction Solvent [b]	Resuspension Solvent	Analysis[c]	Reference
		Non-polar Extractions			
Wastewater, sewage, sediment	OC	DCM	Direct injection	GCxGC-TOF & GC-TOF	Rostkowski et al., 2013
Sediment	SH, CE, OH, SQ	DCM; fractionation	Toluene	GC-MS	Farrés et al., 2015a
Soil	OS, AA, P, SH, TN, LN, FA	Parallel H_2O, MeOH, ACN, hexane	Varied	ESI-FTICR/MS	Tfaily et al., 2015
Soil	OS, AA, P, SH, TN, LN, FA	Sequential H_2O, MeOH, ACN, hexane	Varied	ESI-FTICR/MS	Tfaily et al., 2017
Sewage	SH, OC, PAH, PL	4:1 *n*-hexane:DCM	Isooctane	GCxGC-TOF	Veenaas and Haglund, 2017
Feces	GP, FA, SL, GL, BA, CE, PL	1:4:3 HFIP:MTBE:H_2O (MTBE fraction)	65:30:5 ACN:IPA:H_2O	LC-Orbitrap	Gregory et al., 2012
Tissue	CE, GP, GL, SL, FA	1:1 1-butanol:MeOH	Direct injection	LC-ESI-MS/MS	Alshehry et al., 2015
Tissue	AC	8:6:7 MTBE:MeOH:H_2O (MTBE fraction)	65:30:5 ACN:IPA:H_2O	LC-Orbitrap	Chen et al., 2013
Seeds, tissue	FA	2:1 DCM:MeOH (DCM fraction)	Hexane	GC-FID	Cequier-Sánchez et al., 2008
Plant	CE	2:2:1.2 CHF:MeOH:H_2O (CHF fraction)	CDF	NMR	Brasili et al., 2014
Oil	SH, H, ST	Hexane; fractionation	Isooctane or toluene	GC-MS	Alizadeh et al., 2017
Oil	SH, MA, TP, DI	Hexane; fractionation	DCM	GCxGC-TOF	Casilli et al., 2014
Oil	SH, TP	9:1 DCM:MeOH	Direct injection	GCxGC-FID	Gros et al., 2014
Oil	SH, TP	9:1 DCM:MeOH; fractionation	DCM/methanol	GCxGC-TOF	Hall et al., 2013

Oil	Heteroatom	toluene	Direct injection	FT-ICR/MS	Hur et al., 2018
			Polar Extractions		
Wastewater, sewage, sediment	OC	1:1 MeOH:ACN	Direct injection	LC-TOF	Rostkowski et al., 2013
Sediment	Unknown	Cell lysis; cold MeOH	Pyridine	GC-MS	Beale et al., 2017
Dissolved organic matter	LN, OS, TN, SO	MeOH	Direct injection	FT-ICR/MS	D'Andrilli et al., 2013
Dust, Lint	PH, AA, OC, FA	MeOH then ACN	1:1 MeOH:ACN	LCxLC-TOF	Ouyang et al., 2017
Soil	OS, AA, NT	Fumigation, direct CHF-K_2SO_4, cold MeOH:CHF:H_2O, & hot EtOH	100 mM NH_4HCO_2 in 25% ACN	GC-MS	Warren, 2015
Soil	OS, AA, NT	H_2O, 10 mM K_2SO_4, 10 mM NH_4HCO_3, IPA:MeOH:H_2O, & 10-100% MeOH	Pyridine	GC-MS	Swenson et al., 2015
Soil biocrust	OS, AA, NT	H_2O	MeOH	LC-Orbitrap & LC-QTOF	Swenson et al., 2018
Soil	LN, OS, TN, SO	H_2O	MeOH	FT-ICR/MS	Jiménez-Morillo et al., 2018
Tissue	FA	8:06:07 MTBE:MeOH:H_2O (MeOH/ H_2O fraction)	1:4 MeOH:H_2O	LC-QTOF	Chen et al., 2013
Plant	AA, OS	2:2:1.2 CHF:MeOH:H_2O (MeOH/H2O fraction)	1:2 CD_3OD:D_2O	NMR	Brasili et al., 2014

[a]*Classes include: fatty acids (FA), oligosaccharides (OS), glycerophospholipids (GP), glycerolipids (GL), sphingolipids (SL), bile acids (BA), cholesterol-derivatives (CE), phenol lipids (PL), amino acids (AA), proteins (P), acylcarnitines (AC), organic contaminants (OC), saturated hydrocarbons (SH), monoaromatic steroids (MA), thiophenes (TP), diamondoids (DI), alcohols (OH), flavonoid (Fl), cyclitol (CY), phenolic acid (PA), hydroxycinnamic acids (HA), polyaromatic hydrocarbons (PAH), squalene (SQ), biohopanepolyols (BHP), hopanes (H), steranes (ST), chlorins (CHL), nucleotides (NT), tannins (TN), lignin (LN), soot (SO), phthalate (PH)*

[b]*Solvents include: dichloromethane (DCM), methyl-tert-butyl ether (MTBE), methanol (MeOH), ethanol (EtOH), acetonitrile (ACN), hexafluoroisopropanol (HFIP), chloroform (CHF), isopropyalcohol (IPA), water (H_2O), deuterated methanol (CD_3OD), deuterated water (D_2O), deuterated chloroform (CDF)*

[c]*Analytical instrumentation includes: gas chromatography(GC), two-dimensional gas chromatography (GCxGC), liquid chromatography (LC), two-dimensional liquid chromatography (LCxLC), time of flight mass spectrometer (TOF), quadrupole time of flight mass spectrometer (QTOF), nuclear magnetic resonance spectroscopy (NMR), mass spectrometer (MS), tandem mass spectrometer (MS^n), flame ionization detector (FID), fourier transform ion cyclotron resonance mass spectrometer (FT-ICR/MS)*

Currently, there is no review of ideal "-omics" solvent extractions for environmental matrices, but the most commonly used lipid extraction methods in biological matrices are liquid-liquid extractions derived from the Folch (Folch et al., 1956), Bligh-Dyer (Bligh and Dyer, 1959), and modified versions of both (Table 2.1; Alshehry et al., 2015; CequierSánchez et al., 2008; S. Chen et al., 2013; Y. Chen et al., 2013; Löfgren et al., 2016; Matyash et al., 2008). In addition, dichloromethane or *n*-hexanes are the solvents of choice for oil, sewage, sludge, and sediment samples (Table 2.1). In a lipidomics study, Reis et al. (2013) determined solvent composition had little effect on predominant lipids, but was important for less abundant lipids. Sediment-"omics" and petroleomics studies may have the same solvent effects.

There are only a few examples of extraction methods for untargeted analyses in environmental matrices. One example used dichloromethane and fractionation into neutral and acidic groups to identify novel paleoclimatic biomarkers in marine sediments (Farrés et al., 2015a). Only a few studies have compared different extraction methods (Gregory et al., 2012; Mopper et al., 2007; Swenson et al., 2015; Tfaily et al., 2015, 2017; Warren, 2015). Tfaily et al. (2015, 2017) explored parallel and sequential extraction techniques and found the order of solvent extraction matters. For oil samples, Alizadeh et al. (2017) classified 27 different oils from the Southern Persian Gulf Basin and were able to separate the oils into two distinct families after extraction with *n*-hexanes then fractionation. Nizio et al. (2012) provided additional extraction strategies for petroleum, and many of the methods presented there may also be relevant to sediment samples.

Other methods include solid phase extraction (SPE) and solid phase microextraction (SPME; Cajka and Fiehn, 2016; Jurowski et al., 2017). Semi-targeted and targeted approaches typically use SPE and SPME for post-extraction clean-up and/or additional fractionation. There are advantages to using these methods such as reduced ion suppression, removal of interferents, and less solvent. SPME also

excels in situations where the organic content is low or the sample size is small (Jurowski et al., 2017; Zhao et al., 2014). The major disadvantage is SPE and SPME can result in the loss of potentially discriminant compounds through retention and increased sample handling. Fractionation is not typical for traditional untargeted "-omics" methods; however, oil and sediment matrices are very complex, and fractionation may be required in order to avoid damaging sensitive instrumentation. Rostkowski et al. (2013) demonstrated the use of SPE cartridges and dichloromethane to assess different matrices like sewage, wastewater, sediment and sludge.

2.2.2. Polar extractions

Polar extraction methods extract compounds such as amino acids, flavonoids, hormones, and some amphoteric lipids, etc. (Figure 2.3). Many metabolomics experiments are also designed to minimize protein and lipid content (Cajka and Fiehn, 2016; Raterink et al., 2014). However, polar extraction methods can also be applied to environmental matrices, especially if there is interest in linking microbial activity to metabolites excreted into the environment. Water and/or methanol extractable fractions often contain many compounds relevant to microbial communities. Example polar extractions can be found for sediment (Beale et al., 2017, 2018) and soil (Swenson et al., 2015, 2018; Warren, 2015). Also, Rostkowski et al. (2013) used a mid-polar approach (methanol:acetonitrile) on sewage, wastewater, sediment and sludge matrices to find organic contaminants.

2.2.3. Extraction quality control

Quality control (QC) is an important aspect of any untargeted analysis to reduce variability and noise. In order to reduce experimental variability, there has been a unified movement into established

good experimental and reporting practices for "-omics", and there are a number of excellent reviews (Dudzik et al., 2018; Dunn et al., 2012). Every level of the workflow requires QCs. At the extraction level it is useful to account for pre-analytical variability due to extraction technique and operator error. Extraction QC includes:

- Standard Operating Procedures (SOPs) ensures future extractions follow the same protocols as closely as possible. SOPs should contain the information suggested for the minimum reporting standards (Sumner et al., 2007).
- Environmental replicates are parallel measurements from the same area but different locations (i.e. from the same oil field). These replicates account for random and spatial variation (Dudzik et al., 2018). Replicates should be in triplicates (n = 3), but if possible, more should be incorporated (n = 5; Sumner et al., 2007). Study failure can result from too few replicates when sample variance is high; however, these replicates may not always be possible. For instance, sediment samples from the same lake can vary due to sedimentation rates, and bathymetry.
- Sample replicates are from the same sample, but extracted in triplicate (n = 3) and are used to judge sample extraction consistency (Dudzik et al., 2018). When environmental replicates are not possible, then sample replicates are strongly encouraged to account for both extraction consistency and unknown variation.
- Method blanks are "blank samples" that undergo the same extraction protocol, but without sample (Dudzik et al., 2018). They account for contamination incorporated during the extraction and need to be incorporated into the initial experimental design. As a rule of thumb, the number of method blanks should be 10% of the total number of samples including replicates.

- Recovery standards can be used to measure the extraction consistency by adding known amounts of an artificial (i.e. isotope labelled) compound into the sample before extraction to determine how much is lost during extraction.

Aside from QC procedures, sample and extract storage are important to consider. Organic molecules will degrade after sampling due to light, oxygen, and temperature deviations from the samples' native environment. The use of certain solvents, like methanol, can generate artifacts during sample extraction and storage (Sauerschnig et al., 2017). Thus, long-term storage of samples and extracts should not be at room temperature unless freeze-dried or evaporated, and as much as possible, analyses should proceed immediately after extraction.

It is also important to sample and extract with other interferents in mind. Only use high-grade solvents for extraction and analysis to avoid the introduction of contaminants. Avoid plastic as much as possible. Solvents like dichloromethane and chloroform can strip plasticizers from plastic during extraction, and non-polar compounds have a strong adsorption capacity for plastics. Contamination from plastic manufacturing can be highly variable, even within plastic consumables from the same lot (Yao et al., 2016). At a minimum, dichloromethane and/or chloroform resistant plastics should be used, but the use of glass will increase yields and reduce the interference from plasticizers. If plastic use is unavoidable, which it so often is, proper method blanks can account for some of the plastic-derivatives, but not all (Yao et al., 2016). The highly sensitive instrumentation used in untargeted “-omics” studies can detect trace levels of plastic, which can impact later data analysis.

3. ANALYSIS

3.1. Detection

3.1.1. High resolution mass spectrometry

High resolution mass spectrometry (HRMS) detection is the work horse of untargeted "-omics" methods because of its resolving power; but how resolution is defined is important, and it can be defined in different ways (Hernández et al., 2012). Here resolution will be defined using full width half maximum, which is the manufacturing standard (Murray et al., 2013). The minimum resolution for an HRMS instrument is 10,000. Instruments with this ability include: (1) magnetic sector (10,000–80,000 resolution), (2) time of flight (TOF; 20,000–80,000 resolution), (3) orbitrap (up to 500,000 resolution at 200 m/z), and (4) Fourier-transform ion cyclotron resonance (FT-ICR; up to 1,000,000 resolution; Hernández et al., 2012; Krauss et al., 2010; Špánik and Machyňáková, 2018). Most HRMS instruments also have high mass accuracy (±0.001 Da), wide mass ranges (up to 2000 Da), and high sensitivity (Hollender et al., 2017; Krauss et al., 2010). What this translates to for untargeted "-omics" analysis is: (1) high resolution allows the MS to separate more compounds from complex matrices, (2) high mass accuracy allows structural information to be determined, and (3) large mass ranges allows analysis of larger molecules/polymers. While low resolution mass spectrometers (LRMS) have been used in the past, next-generation HRMS instruments are preferred and add unparalleled information to "-omics" studies.

An important consideration of HRMS is the choice of ionization technique. There are several types of ionization such as: (1) electrospray ionization (ESI), (2) atmospheric pressure photoionization (APPI), (3) electron impact (EI), and (4) atmospheric pressure chemical ionization (APCI). Ionization methods are grouped into hard (EI), and soft (ESI, APCI, APPI). Typically, HRMS "-omics" studies use soft

methods because it provides key information for structural identification if the mass of the entire compound is known. However, no single ionization technique is good for all compounds in all situations. Using multiple ionization methods obtains the most comprehensive information, but this is not always possible due to cost restrictions and instrument compatibilities. The advantages and disadvantages of the different ionization methods have been previously discussed (Cai and Syage, 2006; Gutiérrez Sama et al., 2018; Imbert et al., 2012).

A new feature for HRMS instruments is the addition of ion mobility mass spectrometry (IMMS). Typical HRMS instruments only sort by mass and charge, IMMS separates ions by their size, shape, charge, and mass. Recently, there have been a number of reviews regarding IMMS and "-omics" applications (Cajka and Fiehn, 2016; Hinz et al., 2018; Mairinger et al., 2018; Zhang et al., 2018). The application of IMMS could: (1) reduce background noise, (2) increase selectivity, (3) discriminate isomeric molecules - molecules with the same chemical formula but different structures, and (4) add structural identification information (Cajka and Fiehn, 2016; Lanucara et al., 2014). Collision cross-section (CCS) values are of particular importance. CCS values represent the time the ion requires to cross the chamber, which is dependent on ion shape (Lanucara et al., 2014). CCS values are unique to compounds, and have high inter-laboratory reproducibility as setting changes in mass spectrometer or hyphenated chromatography instruments do not influence the value (Lanucara et al., 2014; Paglia et al., 2014, 2015). This kind of technology has been applied to complex mixtures such as petroleum samples (Fernandez-Lima et al., 2009; Gutiérrez Sama et al., 2018), lipidomics (Hinz et al., 2018; Paglia et al., 2015; Paglia and Astarita, 2017; Poad et al., 2018), and metabolomics (Dwivedi et al., 2010; Paglia et al., 2014; Paglia and Astarita, 2017).

3.1.2. Nuclear magnetic resonance

NMR is another routine detection technique for small molecule "- omics" methods. NMR has several advantages over mass spectrometry: (1) it is non-destructive; (2) reproducible; (3) provides more structural information; (4) sample preparation is simple; (5) does not produce matrix effects; (6) discriminates isobaric compounds; and (7) no derivatization is required (Gathungu et al., 2018; Li et al., 2017). Yet despite its many advantages, NMR is less frequently used than mass spectrometry for several reasons: (1) less sensitivity; (2) overlapping signals; (3) larger samples needed especially for carbon NMR (^{13}C NMR); (4) fewer databases for structural annotation; and (5) poor resolution of compounds in the same class (Li et al., 2017).

However, NMR use is expanding, and new technology is being developed to mitigate its disadvantages. Spectroscopic analyses like two-dimensional proton-proton homonuclear total correlation spectroscopy (2D-^{1}H, ^{1}H TOCSY) or two-dimensional proton-carbon heteronuclear single quantum coherence spectroscopy (2D-^{1}H,^{13}C HSQC) can increase compound resolution (Heude et al., 2017). In addition, LC hyphenated NMR techniques are in development, which will hopefully lead to better compound resolution, and spectral decongestion (Gathungu et al., 2018). NMR has been used on soil samples to assess the effects of environmental gradients (Pisani et al., 2016), soil physical states (Masoom et al., 2016), plant tissue (Brasili et al., 2014) and soil organic matter turnover (Wang et al., 2017).

3.2. Hyphenated techniques

3.2.1. Liquid chromatography mass spectrometry

LCMS is the most commonly used hyphenated chromatography technique. LC separates compounds before they enter the mass spectrometer. With LC there is better sensitivity, compound resolution, and reduced ion suppression than direct injection onto the MS alone (Cajka and Fiehn, 2016).

How to separate compounds with LC is an important consideration for method development. There are several separation techniques for LCMS: (1) reverse phase LCMS, (2) normal phase LCMS, and (3) hydrophobic interaction liquid chromatography (HILIC) - a variant of normal phase LCMS (Cajka and Fiehn, 2016). The use of reverse phase LCMS provides separation based on lipophilicity with polar compounds separating best (Cajka and Fiehn, 2014, 2016). HILIC and normal phase LCMS both separate compounds based on hydrophilicity. Normal phase LC has been applied to petroleum samples and is extensively reviewed by Kamínski et al. (2005). HILIC is better known to separate small polar molecules (Buszewski and Noga, 2012; Nováková and Vlčková, 2009). Thus, the choice of extraction method (i.e. polar or non-polar) will influence the type of LC separation technique used.

Recently, the use of two-dimensional liquid chromatography (LCxLC) with HRMS has been increasing. There are two reasons to use LCxLC: (1) Matrices are too complex for one-dimensional liquid chromatography, and (2) isomeric/isobaric species of interest require increased resolution (Cajka and Fiehn, 2016; Stoll and Carr, 2017). LCxLC methods have been demonstrated for untargeted analyses of blood plasma (Holčapek et al., 2015; Wang et al., 2013), dust and dryer lint (Ouyang et al., 2017), wastewater (Haun et al., 2013; Ouyang et al., 2015), and rice (Navarro-Reig et al., 2018).

3.2.2. Gas chromatography mass spectrometry

GCMS is another useful hyphenated technique for "-omics" methods. GC paired with HRMS is a powerful tool in untargeted analyses according to a recent review by Špánik and Machyňáková (2018). GCMS is ideal for volatile and semi-volatile compounds, with the option of modifying non-volatile compounds to increase GCMS compatibility. Thus, GCMS can analyze a wide range of compounds for "-omics" studies.

A caveat of GCMS is the ionization techniques. Commonly, GCMS is paired with EI ionization. Extensive compound libraries exist for EI fragmentation patterns; however, since EI is a hard ionization source, it is rare that the compound mass can be determined. This is a disadvantage when working with unknown compounds. Thus, for untargeted analysis, having a GCMS that is also capable of using soft ionization (i.e. chemical ionization, APCI, APPI, etc.) is useful for structural annotation, but soft ion ionization methods can also be less sensitive (Tranchida et al., 2018). The application of GCMS is extensive and includes the untargeted analysis of petrochemicals (Gros et al., 2014; Rodgers and McKenna, 2011), migrating compounds from packaging (Cherta et al., 2015), and environmental contaminants (Hernández et al., 2012; Serrano et al., 2011).

Recently, the popularity of comprehensive two-dimensional gas chromatography (GCxGC) with HRMS has grown immensely. A number of reviews have been written about GCxGC (Mondello et al., 2008), and to describe its applications to petrochemicals (Pollo et al., 2018), environmental analyses (Pani and Górecki, 2006), and forensic analyses (Gruber et al., 2018). GCxGC with LRMS can provide better sensitivity than GC HRMS alone (Fushimi et al., 2012), but it becomes more powerful when the separation efficiency of GCxGC is combined with HRMS. Parastar et al. (2018) determined that GCxGC-TOF far surpassed GCMS in resolution and sensitivity. GCxGC-HRMS has been used to screen samples for

halogenated compounds (Ieda et al., 2011; Pena-Abaurrea et al., 2014), petrochemicals (Casilli et al., 2014; Hall et al., 2013; Rodgers and McKenna, 2011; Walters et al., 2018), dust constituents (Hilton et al., 2010), and sewage sludge (Veenaas and Haglund, 2017). Thus, due to the outstanding capabilities of GCxGC technology there has been increasing interest in using it for untargeted "-omics" work.

3.3. Analysis quality control

Even with HRMS instruments, analysis quality control (QC) is integral to reducing analytical variability. The purpose of analysis QC is to reduce the effects of: (1) Batch Effect - shifts in retention time and intensity between batches of samples analyzed at different times or days leading to data processing issues (Rusilowicz et al., 2016; Wehrens et al., 2016), (2) Matrix Effect – where co-eluting compounds, buffer additives, inadequate flow rate, poor solvent purity, and an unclean source can lead to ion suppression/enhancement, reduced sensitivity, quantification/qualification problems, and non-detection (Furey et al., 2013; Uclés et al., 2017), (3) Carryover – when analytes from the previous sample appear in the following sample leading to poor quantification/qualification, and loss of accuracy (Dudzik et al., 2018). It is necessary to consider these effects as they can seriously impact data analysis and may prevent the integration of data from multiple batches and/or experiments. Below are the potential analysis QC's, and which effect they mitigate:

- Pooled QCs are samples containing aliquots from multiple samples pooled together and then extracted following the same procedure, or post-extraction aliquots pooled together. Pooled QCs are used to equilibrate the instrument to the sample matrix before running samples and can correct for batch effect (Dudzik et al., 2018; Dunn et al., 2012).

- Randomization of the analysis order is critical to also reduce bias introduced by batch effect and carryover (Dudzik et al., 2018; Dunn et al., 2012).
- Blanks are mobile phase blanks and sample solvent blanks. Running the same extraction/analysis solvent allows the removal of compounds originating from the solvents later during data analysis, and helps monitor for carryover when a set of intermittently placed blanks are throughout the run (i.e. approximately every 10 samples depending on the sample matrix).
- Cleaning Blanks are blanks containing strong solvents (i.e. like isopropyl alcohol) to clean the system during analysis and reduce carryover effect.
- Internal Standard is a compound (often isotopically-labelled) that is spiked into post-extraction samples and pooled QCs. The internal standard, and recovery standard if used, are also analyzed without the matrix, but in the same solvent as the samples. The internal standard controls for variations in solvation and injection volumes between samples. The magnitude of the matrix effect is determined by comparing the intensity of the spiked samples to just the internal standard in the sample solvent. Matrix effect is a percentage (intensity of spiked internal standard in sample/internal standard in blank solvent x 100%). The strength of the matrix effect can be grouped into three categories: (1) No Matrix Effect (|0|–|20|%), 2) Weak Matrix Effect (|20|–|50|%), and 3) Strong Matrix Effect (>|50|%; Uclés et al., 2017).
- Technical Replicates are repeated analyses of the same samples and are not to be confused with environmental or sampling replicates. Each sample should have six replicates which consists of any combination of environmental, sampling, and technical replicates. The replicates should be randomized so that they are all not done in the same batch or one after another. A large number

of replicates mitigates batch effect and carryover influences, which will make data integration easier, and also lend more statistical power to data analyses (Bader et al., 2016).

4. DATA MINING

4.1. Tools for data mining

Data mining is the process of handling and analyzing data in such a manner as to "mine" the information buried within. Currently, there is a large push to move data mining in metabolomics and lipidomics into simple, but flexible, data workflows that increase reproducibility, and are easily accessible. We also advocate such data workflows for sediment-"omics" and petroleomics. Cambiaghi et al. (2016), Gorrochategui et al. (2016), Weber et al. (2017), and Yi et al. (2016) have all reviewed data workflows for "-omics". It should be noted that, we are focussing on untargeted "-omics" data workflows.

There are many tools to data mine "-omics" data. Spicer et al. (2017) wrote a comprehensive review of open-source tools that can be found to process metabolomics data. Websites hosted by OMICtools (https:// omictools.com) and ms-utils (http://www.ms-utils.org) maintain updated lists of available tools along with forums and relevant references. Briefly, common commercial tools include: (1) Progenesis QI (Waters), (2) Compound Discoverer (Thermofisher), (3) XCMS^{plus} (ABSciex), (4) MetaboScape (Bruker), and (5) MassHunter (Agilent). There are also a number of open-source workflows: (1) Workflow4metabolomics (Giacomoni et al., 2015), (2) Galaxy-M (Davidson et al., 2016), (3) XCMS (Tautenhahn et al., 2012b), (4) MetaboAnalyst 3 (Chong et al., 2018), (5) MAVEN (Melamud et al., 2010), (6) MAIT (Fernández-Albert et al., 2014), (7) SECIMTools (Kirpich et al., 2018), and (8) MZmine2 (Pluskal et al., 2010). A research survey hosted by Weber et al. (2017) determined that most

LCMS users use open-source software like XCMS and MZmine2. GCMS users were split between commercial software and open-source. Other useful resources are how-to publications detailing MZmine2 (Verkh et al., 2018), XCMS (Forsberg et al., 2018), and MetaboAnalyst (Xia and Wishart, 2011, 2016) workflows.

There are many different tools, and it is easy to become paralyzed by the number of options available. In addition, there are major difficulties with tool compatibility where the output of one tool is not acceptable to the next tool (Spicer et al., 2017). Thus, there has been a major push to generate open-source tools that contain the entire workflow, but are also flexible to address different datasets and transparent to facilitate reproducibility. Thus, learning with an open-source workflow is often a good place to begin.

4.2. Data preprocessing

4.2.1. File types

The first step in data mining is knowing where your data are coming from, and how your data were acquired. The user can acquire HRMS data in centroid mode or profile mode. Profile mode looks like a Gaussian curve across m/z values but varying in intensity. Centroid mode is a compressed profile mode where only the m/z values associated with the highest intensity are reported. The choice of acquisition mode influences the resulting data. Centroided data files are smaller than profiled data, but result in the loss of noise characteristics, ion signal linearity, interfering ions, and informative isotope features (Wang and Gu, 2010).

Additionally, most of the vendors have proprietary file types. The difficulty with propriety file types arises when a researcher does not have access to the vendor's commercial software or would like

to perform a statistical analysis that is beyond the scope of their software. The researcher then has a difficult to open file. If a researcher is looking to use a specific instrument for "-omics" analysis, then it is prudent to also ask the vendor in question how to convert their files to more open-access file types like mzXML, mxData (Orchard et al., 2007), mzML (Martens et al., 2011), netCDF (Gorrochategui et al., 2016). There are also available open-source software packages, such as Proteowizard (Holman et al., 2015) to convert proprietary file types into open-access files.

4.2.2. Spectral processing

The processing of raw HRMS files takes several steps and each software package will have its own algorithms, or have multiple algorithms to choose from (Gorrochategui et al., 2016). Thus, reviewing the advantages and disadvantages of each algorithm is useful, but often the data dictate which algorithms to use. Reviews of this process are available in Gorrochategui et al. (2016), and Karaman (2017).

Software choice will affect the results (Coble and Fraga, 2014; Gürdeniz et al., 2012). Studies have compared the resulting biomarkers from different software programs and discovered only two of fourteen potential biomarkers were common between them (S. Chen et al., 2013; Y. Chen et al., 2013). A recent paper compared MZmine2 and XCMS and suggested problems within their algorithms (Myers et al., 2017a); however, in a following paper they implemented an improved algorithm that also minimizes false discovery rates (Myers et al., 2017b). This is not to say that open-source software is of lesser quality than commercial software. X.M. Li et al. (2018) and Z. Li et al. (2018) determined MZmine2 outperformed other open-source software and commercial software with the lowest false discovery rate and highest identification of true positives. It is an advantage for open-access software that the

research community can assess their accuracy and precision, which ultimately benefits the field as a whole. Thus, the user should be aware that their choice of commercial or open-source programs will impact their results, but this is a rapidly developing field and open-source offers transparency.

4.2.3. Missing value imputation and data normalization

After spectra processing the dataset then needs to be curated to move forward into the statistical analysis. The following steps are flexible and can be used in any combination thereof; however, there is a purpose to each step and the result should be a dataset following a typical Gaussian distribution. The first technique is missing value imputation. Missing values in a dataset can cause problems during the statistical analysis (Di Guida et al., 2016), and are usually attributed to non-detects, peak picking errors during spectral processing or values below the analytical noise threshold. Some missing values can be addressed by gap filling algorithms to fill in small or missing peaks (Karaman, 2017). Missing value imputation algorithms are applied if missing values persist. Wei et al. (2018) proposed several methods to address missing values. Simple filtering procedures can be applied where only bins (i.e. retention time + *m/z* variables) with >80% detection frequency in any one sample group are kept (Yang et al., 2015) or removing samples/replicates with a high proportion of missing values (Armitage et al., 2015). Imputation algorithms can be applied as well, either in lieu of or after filtering, and studies have compared the impact of the different imputation strategies (Armitage et al., 2015; Di Guida et al., 2016; Gromski et al., 2014b; Wei et al., 2018). The purpose of this strategy is to reduce the number of "zeros" in the dataset, which improves the performance of many different statistical analyses.

After the missing value imputation, row-wise corrections (i.e. within chromatogram corrections), also called normalization, are performed. The purpose of normalization is to correct for systemic bias in ion intensities from sample collection, experimental bias, and analytical variations (Gorrochategui et al., 2016). There are several reviews discussing different normalization strategies (De Livera et al., 2012; Li et al., 2016; Chen et al., 2017). Each normalization strategy has its own advantages and disadvantages (Karaman, 2017). The internal, external, or pooled QC's, and the data structure itself determines the normalization strategy used. Every dataset is different and will require different techniques as demonstrated by Wu and Li (2016) who compared the results of different normalization procedures on two biological matrices. Usually, multiple normalization strategies are compared. The success of the normalization is judged by a Gaussian distribution of intensities, and reduced distance within sample groups combined with increased distances between groups on a principal component analysis (PCA) - described later.

Lastly, data processing requires column-wise corrections (i.e. between chromatograms), which involves data scaling and transformations. The purpose of scaling is to correct for fold changes that could mask important lower intensity peaks (Van den Berg et al., 2006). The most commonly used scaling methods are autoscaling (Kvalhein, 1985), and pareto-scaling (Kasprzak and Lewis, 2001). The purpose of transformations is to correct for heteroscedasticity (also known as unequal variance due to noise) and skewness. Many statistical analyses require that analytical noise is homoscedastic (Karaman, 2017). Van den Berg et al. (2006), Di Guida et al. (2016) and Karaman (2017) compare the advantages and disadvantages of scaling and transformation methods.

4.3. Statistical analysis

The 2D matrices generated from the spectral processing can be analyzed with many different statistical analyses. Typical "omics" workflows use discriminatory analyses to generate hypotheses about which molecular compounds are potential biomarkers. These discriminatory analyses include univariate and multivariate statistical tests such as ANOVAs with multiple pairwise comparisons, volcano plots, partial least squares (PLS), random forest (RF), and support vector machines (SVM) to list a few. In addition, there are many pattern recognition algorithms that provide useful information such as molecular fingerprinting, unsupervised multivariate analyses, and clustering. A single data set does not require all of these analyses, and the choice of test is data dependent. Often pattern recognition algorithms assess data structure and group differences. Discriminatory analyses determine what molecules are driving variance between groups. Here we present commonly used statistical tests for "omics" data analysis.

4.3.1. Molecular fingerprinting analysis

Molecular fingerprinting analyses visualize differences in samples. These analyses require the use of HRMS, typically FT-ICR or Orbitrap. The elemental compositions (i.e. molecular formula) are estimated based on the accurate mass measurements, isotopic ratios and atomic limitations based on the type of molecules expected in the sample (*Section 5.2*; Kind and Fiehn, 2007; Stubbins et al., 2010). Once the molecular formulas have been estimated, then double bond equivalents (DBE) calculations, Kendrick mass defect plots, van Krevelen diagrams or multidimensional stoichiometric compound classifications can be built (Gutiérrez Sama et al., 2018; Rivas-Ubach et al., 2018; van Krevelen, 1950).

Molecular fingerprinting is commonly used in petroinformatics (Gutiérrez Sama et al., 2018; Hur et al., 2017, 2018), but has been applied to soils (D'Andrilli et al., 2013; Jiménez-Morillo et al., 2018).

4.3.2. Univariate analysis

Several univariate analyses can be performed on "-omics" data. Univariate analyses consider only one dependent variable, whereas multivariate considers more than one. Common univariate tests include: Fold change, t-tests, volcano plots, correlation analysis, analysis of variance (ANOVA), Kruskal-Wallis and/or multiple comparison. Univariate analyses require establishing whether parametric (e.g. Pearson correlation, ANOVA, t-tests, Tukey HSD, or Fisher's LSD) or non-parametric (e.g. Mann–Whitney U test, Spearman correlation, Kruskal-Wallis, Dunn's test) tests are appropriate. According to Pinto (2017), non-parametric tests are less powerful, but only when there are normal populations, equal variances (i.e. homoscedastic) and equal sample sizes. Thus, the choice of parametric or non-parametric strategies is entirely dependent on the data.

Fold changes indicate absolute value changes between two group means (i.e. whether one group demonstrates an increase or decrease in magnitude when compared to a base group or control). t-tests demonstrate whether a variable is significantly different between two groups. Volcano plots compare two groups using statistical significance (y-axis) and fold change (x-axis). A "volcano" shaped arrangement of points often appears whereby points further from 0 on the x-axis demonstrate greater fold change, and points higher on the y-axis are more significant (i.e. on a -log(*p*-value) scale y-values > 1.3 indicate *p*-values < 0.05; Hur et al., 2017). The main limitation of fold changes, t-tests, and volcano plots is that they only compare two groups.

Multi-group analyses compare more than two groups and include ANOVA, Kruskal-Wallis, correlation analysis and multiple comparisons. ANOVA (parametric) and Kruskal-Wallis (non-parametric) compare means between more than two groups. They indicate if there are differences between the groups and assign *p*-values. Multiple comparison tests (e.g. Tukey HSD, Fisher's LSD, Dunn's test, etc.) inform about where the differences lie between the groups. (Pinto, 2017; Xi et al., 2014). Lastly, correlation analysis can also be used for both two group, and multi-group analyses. Correlation analyses visualize the degree of association between variables, and can even be combined with molecular fingerprinting analyses performed above (Asuero et al., 2006; Hur et al., 2017). Multi-group analyses are well established methods, however for large data sets or when there are many groups, they can be computationally intensive.

As a side note, it is common in "-omics" for the number of measured variables to be far greater than the number of samples due to the nature of the data (e.g. there are thousands of organic molecules in one sample). This imbalance increases the risk of false discoveries and postanalytical variation because when there are too few samples the variation from the variables is over-emphasized (Martins et al., 2018). Thus, when using multiple comparisons, (i.e. post-hoc tests) the *p*-values often have to be corrected using Bonferroni, Benjamini– Hochberg, or metabolome-wide significance level corrections (Pinto, 2017; Xi et al., 2014). If corrections cannot be applied then the results should be interpreted with caution.

4.3.3. Multivariate analysis

Multivariate analyses are used to determine relationships between multiple dependent variables, and are better at handling batch-to-batch variation and drift (Pinto, 2017). There are two main

categories: (1) Unsupervised - where no assumptions are drawn from group classifications (i.e. Principal component analysis (PCA), hierarchical clustering (HCA) and self-organizing map (SOMap)), or (2) Supervised - where groups and samples are associated together to build the final model (e.g. partial least squares or projection to latent structures (PLS), random forest (RF), and support vector machines (SVM)). Unsupervised models identify outliers, and determine which variables are the most influential without group bias. Often an unsupervised model can determine whether variance from pre-analytical, analytical, and post-analytical artifacts are overwhelming the desired biological/experimental variance. Supervised methods are powerful models that integrate groupings (i.e. impacted or unimpacted) into the generated matrices so that the model correctly associates the research questions with the output (Goodacre et al., 2004). Supervised models isolate candidate biomarkers from the data in discriminant analyses, as demonstrated in a S-plot derived from an orthogonal PLS (Figure 2.4C). Usually, variables of importance (VIP) or selectivity ratios can be displayed to illustrate which compound signals are the most discriminatory (Farrés et al., 2015b). Below, the analyses are briefly explored.

PCA models explain variance within the data set. A PCA generates a scores plot (Figure 2.4A) – comparing separation between samples, a loadings plot – expressing direction of variance for each variable (i.e. detected compounds), and a scree plot – showing the percent variance explained by each principal component. PCA allows for visual interpretation of the variance. Samples grouped close together are more similar than those grouped farther apart. The loadings plot is useful in determining which variables are key to defining the variance in any one principal component. Using a PCA, outliers can be identified as they often do not cluster with the others, they are outside a 95% confidence window, using cumulative measures of distance, or by inspecting model residuals (Pinto, 2017). PCA is the work horse of multivariate data analyses and provides insight into the data structure.

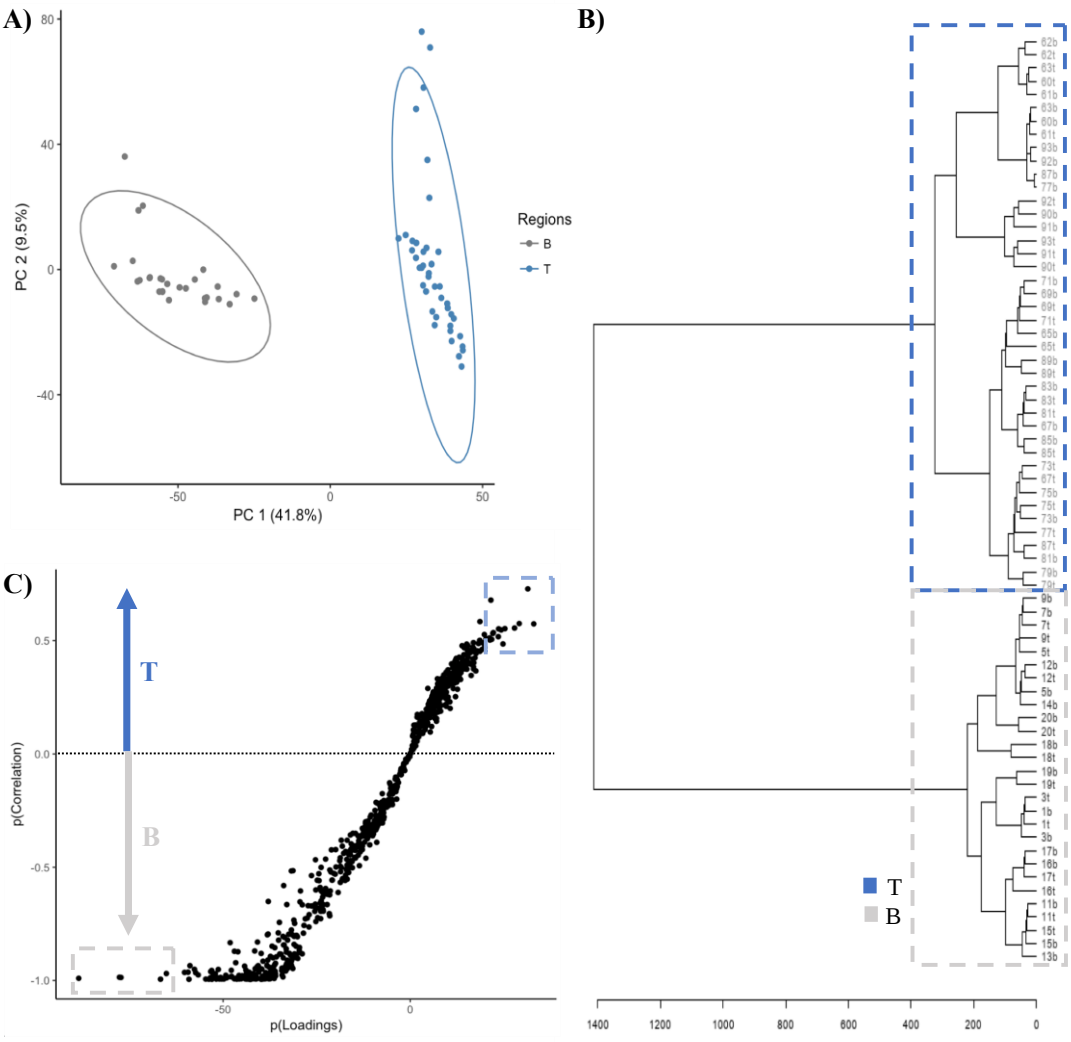


Figure 2.4: Example outputs of the chemical similarity between lakes based on their molecular composition. A) Principal Component Analysis two-dimensional scores plot - similar samples are clustered closer together. Principal component 1 (PC1) explains 41.8% of the total variance and defines the component where groups of the samples are most dissimilar. PC2 explains 9.5% of the variance. Overall this model explains 51.3% of the variance in the first two dimensions. B) Hierarchical Clustering dendrogram based on Euclidean distance and the Ward linkage algorithm. Ideally, the samples classed together should appear closer together based on the linkages in the dendrogram, which indicates similarity (highlighted in boxes). C) S-plot is a loadings scatter plot for a discriminant OPLSDA model. The y-axis represents the correlation where the increasing |+/-| distance from 0 indicates the more likely the variable (i.e. unique mass + retention time signal) occurs in a class (i.e. either T or B in this case). The x-axis represents the covariance and indicates the variable's magnitude of contribution to the model. Thus, the variables that are the most reliable and contribute the most to the model are at the extremes of the ideal "S" shape (highlighted in boxes). These variables are discriminatory for each class (i.e. T or B).

HCA is an unsupervised method, and it is used to visualize natural clustering in the data with dendrograms as demonstrated in Figure 2.4B (Ren et al., 2015). Distance measures (e.g. Euclidean distance, Pearson's correlation, Spearman's rank correlation) and linkage functions (e.g. average linkage, complete linkage, or Ward's linkage) are chosen to build a dendrogram. Useful reviews for deciding on the best combination can be found in Hastie et al. (2009), Jain et al. (1999), and Ren et al. (2015). In addition, there are several papers comparing distance measures, which include methods for assessing the model validity (Giancarlo et al., 2010, 2011; Jaskowiak et al., 2013, 2014). The major advantage of HCA is that it is easy to visualize, but interpretation can be difficult when there are a large number of samples or too many variables (Pinto, 2017).

Another unsupervised analysis is SOMap. SOMap is a neural network algorithm that visualizes multi-dimensional data in several ways including using heatmaps, Umatrix, and HCA. SOMap is a type of artificial neural network that uses neighbourhood functions to arrange “neurons” or nodes in a grid (Ren et al., 2015). Thus far there are only a few examples available in literature involving the use of SOMap (Binder and Wirth, 2015; Franceschi and Wehrens, 2014; Haddad et al., 2009).

PLS is a common statistical analysis for “-omics” work. It is a supervised regression method that can be used to maximize the separation between sample groups (Xi et al., 2014). There are different versions of PLS: orthogonal PLS (OPLS; Thévenot et al., 2015), and sparse PLS (sPLS; Lê Cao et al., 2011). Presently, PLS methods are poorly utilized as many researchers do not optimize or correctly report parameters (Brereton and Lloyd, 2014; Considine et al., 2018). PLS models are also prone to over-fitting where the results cannot be replicated with a new set of data (Westerhuis et al., 2008; Xi et al., 2014). Thus, the models need to be validated to avoid overfitting and proper parameter optimization. There are several ways to validate PLS models including cross-validation, Monte Carlo cross-validation (MCCV),

receiver operating characteristic (ROC) analysis, and permutation tests (Pinto, 2017; Westerhuis et al., 2008; Xi et al., 2014). In addition, a high Q2 value (commonly enlisted to indicate model quality) is not always indicative of a good model, albeit it is generally a good indicator of a poor one, so care should be taken during the interpretation as well (Gromski et al., 2015; Westerhuis et al., 2008).

Random forest (RF) is a supervised learning algorithm that is useful for classification, identification of outliers and discriminant analysis of biomarkers (Breiman, 2001; Cutler et al., 2007). It is adept at handling large datasets, missing values, resistant to outliers and over-fitting, and it can handle noisy data (Gromski et al., 2015). A useful metric for RF analyses are out-of-bag (OOB) errors, where a lower error indicates a better model. OOB errors are calculated during the RF computation so no additional validation methods are required (Kuznetsova, 2014). A current issue with RF is the lack of simplistic visualization (Gromski et al., 2015), although visualization techniques are under development (da Silva et al., 2017; Kuznetsova, 2014).

The last method discussed here, is the supervised SVM method. SVM is a nonparametric machine learning method that can be used for classification and regression (Gromski et al., 2015). This method is not influenced by population distributions, and is resistant to outliers, overfitting, and noise (Gromski et al., 2015; Xu et al., 2006). However, currently, there are also a number of disadvantages such as data visualization, solving only two-group problems (multi-group solutions exist, but require additional algorithms), and that it is computationally expensive (Ren et al., 2015). Despite, these disadvantages, in a two-group dataset, SVM has been shown to outperform PLS in classification accuracy and discriminant analyses (Mahadevan et al., 2008) and is equal to RF (Gromski et al., 2014a).

5. STRUCTURE IDENTIFICATION

5.1. Reporting Standards

Structural identification, or annotation, is a bottleneck in the untargeted "-omics" workflow. A comprehensive review on structural annotation by Dunn et al. (2013) states that in a typical GCMS run, only 50% of the metabolites structures can be identified. In addition, the quality of the identification needs consideration. In 2005, a group of researchers formed the Metabolomics Standards Initiative (MSI: http://www.metabolomics-msi.org), which worked to establish good reporting and data handling methods for metabolomics, but can apply to other "-omics" fields (Fiehn et al., 2007). MSI established the Chemical Analysis Working Group (CAWG). CAWG outlined four levels of identification: (1) Identified compounds – confirmation with an authentic standard with at least two orthogonal approaches; (2) Putative annotated compounds – identification based on physical chemistry and/ or spectral libraries; (3) Putatively identified class level – structure remains unknown but physical chemistry and/or spectra suggests a likely class of molecules; and (4) unknown – little to no knowledge on the compound (Sumner et al., 2007). Currently it is expected to annotate compounds of interest according to their level of identification (Creek et al., 2014). The original CAWG reporting standards are sometimes too simple, so other reporting standards have been developed by Schymanski et al. (2014), EU Guideline 2002/657/EC, and Rochat (2017). Identification and/or annotation of compounds can be done during the data acquisition step or after acquisition using additional targeted methods and purification strategies (Dunn et al., 2013). Below, tools and strategies for structural annotation are listed with the caveat that these techniques are typically used in tandem as confirmation from multiple orthogonal techniques provides more confident annotations of the compounds of interest.

5.2. *In silico* identification tools

There are several different *in silico* tools that facilitate structural annotation. Usually, spectral libraries are integral to the annotation process, but mass fragmenters and elemental compositions calculations are also useful. Below there is a general discussion of tools along with common software.

Common annotation resources are spectral libraries. Spectral libraries provide examples of spectra taken under specific conditions and reference accurate mass, CAS numbers, database specific identifiers, chemical formulas, experimental/predicted physical chemistry data and fragmentation information. There are many databases that are open-access or commercial. Some common open-access databases include:

- LC-ESI-MS databases - METLIN (Guijas et al., 2018; Smith et al., 2005; Tautenhahn et al., 2012a), MassBank (Horai et al., 2010), Human Metabolome Database (HMDB; Wishart et al., 2013), LipidMaps (Fahy et al., 2007, 2018), mzCloud, Chemspider (Pence and Williams, 2010), and MetFrag (Ruttkies et al., 2016)
- NMR databases - Madison Metabolomics Consortium Database (MMCD; Cui et al., 2008), HMDB, and Complex Mixture Analysis by NMR (COLMAR; Bingol et al., 2012, 2015)
- GC-EI-MS databases – Automated Mass Spectral Deconvolution and Identification System (AMDIS), Golm metabolome database (GMD; Kopka et al., 2005).

Useful commercial databases include NIST, and the Wiley Registry. There is also a current movement to develop databases for CCS obtained via IMMS (Zheng et al., 2017). Choose databases with care as the source of the data, the method of acquiring the data, and the focus of the database impacts the quality and/or relevance of the data stored there. Gil de la Fuente et al. (2017) and Vinaixa et al. (2016) provide more extensive reviews of spectral databases.

In silico mass fragmentation is also a handy structural elucidation tool. These tools use algorithms to computationally predict a compound's mass spectrum so that experimental data can be compared with the *in silico* predictions. MetFrag is an example of a reliable open-access fragmenter (Scheubert et al., 2013), and has also been incorporated into the popular METLIN database (Tautenhahn et al., 2012a). For more software and algorithm details please refer to Scheubert et al. (2013).

Elemental composition predictions can be useful for structural annotations and Molecular Fingerprinting Analyses (see *Section 4.3.1*.). When acquiring data for elemental composition, there are two important considerations: (1) the mode of acquiring the data, and (2) instrumental accuracy. Profile mode provides better information for determining elemental compositions. In addition, mass accuracy is important. Typically mass accuracy for HRMS instrumentation is 5 ppm (Gorrochategui et al., 2016), but this can vary by instrument and throughout a batch run. However, according to Kind and Fiehn (2006), mass accuracy alone is not enough for accurate elemental compositions and a low (< 2%) error for isotopic abundance patterns is also required. There are several open-access algorithms to determine elemental composition such as SIRIUS (Böcker et al., 2009), BRAIN (Claesen et al., 2012), and Fourier (Fernandez-de-Cossio Diaz and Fernandez-de-Cossio, 2012), all of which are recommended by Scheubert et al. (2013). Commercially available software includes: SigmaFit (Bruker), Formula Predictor (Shimadzu), iFit (Waters), and MassHunter (Agilent). Scheubert et al. (2013) provides more detailed information on the algorithms of the individual tools.

5.3. Analytical identification tools

To have a better annotation for the compounds of interest, often the sample requires further structural elucidation analyses. Many of the *in silico* tools listed above, are useful, but by themselves can

only lead to putatively annotated compounds. In addition, there are instances, especially for paleolimnologists and geologists, where standards are unavailable and/or the *in silico* tools do not contain information on the compound. Thus, annotation requires further analytical measurements.

MS can acquire additional information. Accurate masses from HRMS can be reported with the annotation. In addition, tandem mass spectrometry (MSn; where n = number of mass analyzers) can determine chemical structures. MS^n is a process linking several mass analyzers together to repeatedly filter and fragment compounds. It generates different mass spectral scans that provide additional structural information. MS^n scans are important when trying to discriminate between compounds with the same molecular formula or similar structures (Dunn et al., 2013). Lately, there has been a push to incorporate both HRMS, and simultaneous acquisition of MS/MS (i.e. MS^2) information. For instance, Wrona et al. (2005) suggested an "all-in-one" analysis with a hybrid quadrupole TOF instrument. This kind of all-in-one strategy is called Data Independent Acquisition (DIA), which acquires MS/MS for all m/z ions in the specified range (Cajka and Fiehn, 2016). If the reader is interested in DIA instrumentation, Cajka and Fiehn (2016) provide a review on merging untargeted and targeted "-omics" methods.

NMR and infrared spectroscopy (IR) provide additional information when MS and *in silico* methods are not enough. Together MS, NMR, and IR spectra can determine complete structures. MS provides information about the molecular formula using accurate mass and isotopic ratios. NMR demonstrates how are the atoms are connected, and IR provides information about functional groups. To acquire spectra from NMR and IR, however, the compound needs to be purified and concentrated. Packed silica columns, or a preparative LC or GC are used for purification. Also, NMR is typically less sensitive than MS, so greater amounts of the compound are required (Gathungu et al., 2018). Obviously,

this process can take a long time, or laboratories may not have access to the instrumentation required. Regardless of data gaps, all the information acquired about an unknown compound should be published. Proper annotation of molecular structures is important to the research community, and high quality spectra of unknown structures can be included in databases, so the information can be made available to other researchers as well.

6. CONCLUSIONS

6.1. Challenges

The application of "-omics is an open field for paleolimnologists and geologists alike, but there are challenges facing the wholesale adoption of "-omics". There are challenges in all aspects of the workflow. Presented below are the current challenges for sample preparation, sample analysis, data mining, and structural elucidation.

In sample preparation, there is a lack of consensus and literature on comprehensive OM extraction procedures in environmental matrices. There are many ways to extract OM, but which method is best suited for "-omics"? A common question is, are we extracting too much? What are we missing in this extraction method, and is it important? Sediment-"omics" and petroleomics still require extensive method development. The current literature also does not often include extraction QCs. QCs play a significant role in ensuring the quality of results.

Sample analysis challenges include: expensive instrumentation, and the absence of analysis QCs. HRMS instruments are not cheap, and the fields are reluctant to move away from LRMS instruments like single quad GCMS. Analysis QCs are essential, but also poorly reported.

The entrenchment of older univariate statistical analysis combined with the lack of technical expertise in multivariate analyses stymies the adoption of data mining. Often, a major complaint of sediment"- omics" and petroleomics approaches is that they are too data reliant; however, metabolomics, metagenomics, and transcriptomics fields have demonstrated the reliability of these data workflows. Thus, exposure to these new statistical methods may prove useful to paleolimnologists and geologists alike.

Lastly, one of the largest challenges is annotation. There are no relevant, open-access HRMS and/or NMR libraries for geologically relevant compounds. The absence of libraries severely limits the number of possible structural annotations for biomarker identifications. This represents a major bottleneck in sediment-"omics" and petroleomics because annotation then requires additional analyses (if enough sample is available) and/or extensive, time-consuming literature searches. Standards can also be prohibitively expensive or unavailable so high level annotations may not be possible. In this area there is the potential for massive development in geologically relevant annotation infrastructure.

There is a lot of future work for interested parties, and the fields have a high potential for synergism. Thus, the future of sediment-"omics" and petroleomics requires the development of open-access databases, data repositories and, as a community, discussion on transparency, SOPs, QCs, and practical reporting standards.

6.2. Future implications

"-Omics" methodologies could have large implications for paleolimnological and geological research. "-Omics" is not just a tool for molecular biologists, it is a power screening tool for

environmental scientists as well. Untargeted "-omics" presents a comprehensive method for hypothesis generation to add new perspectives to old questions.

There are potentially many applications of "-omics" to paleolimnological and geological samples. Geologists are already beginning to use petroleomics for oil bed discrimination, oil spill reconstructions, and for determining the effects of weathering on crude oil for remediation research. However, petroleomics may benefit from the addition of the data mining tools mentioned here such as additional statistical analyses, QCs, and reporting standards.

In addition, the application of "-omics" to paleolimnology as sediment-"omics", could lead to more biomarkers for paleoecology, archaeology, paleoecotoxicology, and paleoclimatology. For paleolimnologists, this could even lead to a better understanding of diagenesis and the distribution of organic molecules in ecosystems as transfer functions. The addition of new biomarkers would facilitate multi-proxy reconstructions, which would make historical reconstructions more detailed and paleolimnological research an even better resource for paleoclimate models to understand how climate change will affect us in the future.

ACKNOWLEDGEMENTS

This research was supported by a Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN-2018-04248) to JMB.

REFERENCES

Alizadeh, B., Alipour, M., Chehrazi, A., Mirzaie, S., 2017. Chemometric classification and geochemistry of oils in the Iranian sector of the southern Persian Gulf Basin. Org. Geochem. 111, 67–81. https://doi.org/10.1016/j.orggeochem.2017.05.006

Alshehry, Z.H., Barlow, C.K., Weir, J.M., Zhou, Y., McConville, M.J., Meikle, P.J., 2015. An Efficient Single Phase Method for the Extraction of Plasma Lipids. Metabolites 5, 389–403. https://doi.org/10.3390/metabo5020389

Armitage, E.G., Godzein, J., Alonso-Herranz, V., Lopez-Gonzalvez, A., Barbas, C., 2015. Missing value imputation strategies for metabolomics data. Electrophoresis 36, 3050–3060.

Asuero, A.G., Sayago, A., González, A.G., 2006. The Correlation Coefficient: An Overview. Crit. Rev. Anal. Chem. 36, 41–59. https://doi.org/10.1080/10408340500526766

Bader, T., Schulz, W., Kümmerer, K., Winzenbacher, R., 2016. General strategies to increase the repeatability in non-target screening by liquid chromatography-high resolution mass spectrometry. Anal. Chim. Acta 935, 173–186. https://doi.org/10.1016/j.aca.2016.06.030

Beale, D.J., Crosswell, J., Karpe, A.V., Ahmed, W., Williams, M., Morrison, P.D., Metcalfe, S., Staley, C., Sadowsky, M.J., Palombo, E.A., Steven, A.D.L., 2017. A multi-omics based ecological analysis of coastal marine sediments from Gladstone, in Australia's Central Queensland, and Heron Island, a nearby fringing platform reef. Sci. Total Environ. 609, 842–853. https://doi.org/10.1016/j.scitotenv.2017.07.184

Beale, D.J., Crosswell, J., Karpe, A.V., Metcalfe, S.S., Morrison, P.D., Staley, C., Ahmed, W., Sadowsky, M.J., Palombo, E.A., Steven, A.D.L., 2018. Seasonal metabolic analysis of marine sediments collected from Moreton Bay in South East Queensland, Australia, using a multi-omics-based approach. Sci. Total Environ. 631–632, 1328–1341. https://doi.org/10.1016/j.scitotenv.2018.03.106

Berk, M., Ebbels, T., Montana, G., 2011. A statistical framework for biomarker discovery in metabolomic time course data. Bioinformatics 27, 1979–1985. https://doi.org/10.1093/bioinformatics/btr289

Binder, H., Wirth, H., 2015. Analysis of large-scale OMIC data using Self Organizing Maps, in: Encyclopedia of Information Science and Technology. IGI Global, Hershey, PA, pp. 1642–1653.

Bingol, K., Li, D.-W., Bruschweiler-Li, L., Cabrera, O.A., Megraw, T., Zhang, F., Brüschweiler, R., 2015. Unified and Isomer-Specific NMR Metabolomics Database for the Accurate Analysis of 13 C– 1 H HSQC Spectra. ACS Chem. Biol. 10, 452–459. https://doi.org/10.1021/cb5006382

Bingol, K., Zhang, F., Bruschweiler-Li, L., Brüschweiler, R., 2012. TOCCATA: A Customized Carbon Total Correlation Spectroscopy NMR Metabolomics Database. Anal. Chem. 84, 9395–9401. https://doi.org/10.1021/ac302197e

Blaise, B.J., Correia, G., Tin, A., Young, J.H., Vergnaud, A.-C., Lewis, M., Pearce, J.T.M., Elliott, P., Nicholson, J.K., Holmes, E., Ebbels, T.M.D., 2016. Power Analysis and Sample Size Determination in Metabolic Phenotyping. Anal. Chem. 88, 5179–5188. https://doi.org/10.1021/acs.analchem.6b00188

Bligh, E.G., Dyer, W.J., 1959. A rapid method of total lipid extraction and purification. Can. J. Biochem. Physiol. 37, 911–917.

Böcker, S., Letzel, M.C., Lipták, Z., Pervukhin, A., 2009. SIRIUS: decomposing isotope patterns for metabolite identification†. Bioinformatics 25, 218–224. https://doi.org/10.1093/bioinformatics/btn603

Brasili, E., Praticò, G., Marini, F., Valletta, A., Capuani, G., Sciubba, F., Miccheli, A., Pasqua, G., 2014. A non-targeted metabolomics approach to evaluate the effects of biomass growth and chitosan elicitation on primary and secondary metabolism of Hypericum perforatum in vitro roots. Metabolomics 10, 1186–1196. https://doi.org/10.1007/s11306-014-0660-z

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brereton, R.G., Lloyd, G.R., 2014. Partial least squares discriminant analysis: taking the magic away. J. Chemom. 28, 213–225. https://doi.org/10.1002/cem.2609

Bu, Q., Wang, D., Liu, X., Wang, Z., 2014. A high throughout semi-quantification method for screening organic contaminants in river sediments. J. Environ. Manage. 143, 135–139. https://doi.org/10.1016/j.jenvman.2014.05.009

Buszewski, B., Noga, S., 2012. Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique. Anal. Bioanal. Chem. 402, 231–247. https://doi.org/10.1007/s00216-011-5308-5

Cai, S.-S., Syage, J.A., 2006. Comparison of Atmospheric Pressure Photoionization, Atmospheric Pressure Chemical Ionization, and Electrospray Ionization Mass Spectrometry for Analysis of Lipids. Anal. Chem. 78, 1191–1199. https://doi.org/10.1021/ac0515834

Cajka, T., Fiehn, O., 2016. Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. Anal. Chem. 88, 524–545. https://doi.org/10.1021/acs.analchem.5b04491

Cajka, T., Fiehn, O., 2014. Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. TrAC Trends Anal. Chem. 61, 192–206. https://doi.org/10.1016/j.trac.2014.04.017

Cambiaghi, A., Ferrario, M., Masseroli, M., 2016. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. Brief. Bioinform. bbw031. https://doi.org/10.1093/bib/bbw031

Casilli, A., Silva, R.C., Laakia, J., Oliveira, C.J.F., Ferreira, A.A., Loureiro, M.R.B., Azevedo, D.A., Aquino Neto, F.R., 2014. High resolution molecular organic geochemistry assessment of Brazilian lacustrine crude oils. Org. Geochem. 68, 61–70. https://doi.org/10.1016/j.orggeochem.2014.01.009

Cequier-Sánchez, E., Rodríguez, C., Ravelo, Á.G., Zárate, R., 2008. Dichloromethane as a Solvent for Lipid Extraction and Assessment of Lipid Classes and Fatty Acids from Samples of Different Natures. J. Agric. Food Chem. 56, 4297–4303. https://doi.org/10.1021/jf073471e

Chen, S., Hoene, M., Li, J., Li, Y., Zhao, X., Häring, H.-U., Schleicher, E.D., Weigert, C., Xu, G., Lehmann, R., 2013. Simultaneous extraction of metabolome and lipidome with methyl tert-butyl ether from a single small tissue sample for ultra-high performance liquid chromatography/mass spectrometry. J. Chromatogr. A 1298, 9–16. https://doi.org/10.1016/j.chroma.2013.05.019

Chen, Y., Xu, J., Zhang, R., Shen, G., Song, Y., Sun, J., He, J., Zhan, Q., Abliz, Z., 2013. Assessment of data pre-processing methods for LC-MS/MS-based metabolomics of uterine cervix cancer. Analyst 138, 2669–2677. https://doi.org/10.1039/C3AN36818A

Chen, J., Zhang, P., Lv, M., Guo, H., Huang, Y., Zhang, Z., Xu, F., 2017. Influences of normalization method on biomarker discovery in gas chromatography–mass spectrometry based untargeted metabolomics: what should be considered? Anal. Chem. 89, 5342–5348.

Cherta, L., Portolés, T., Pitarch, E., Beltran, J., López, F.J., Calatayud, C., Company, B., Hernández, F., 2015. Analytical strategy based on the combination of gas chromatography coupled to time-of-

flight and hybrid quadrupole time-of-flight mass analyzers for non-target analysis in food packaging. Food Chem. 188, 301–308. https://doi.org/10.1016/j.foodchem.2015.04.141

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D.S., Xia, J., 2018. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res. https://doi.org/10.1093/nar/gky310

Claesen, J., Dittwald, P., Burzykowski, T., Valkenborg, D., 2012. An Efficient Method to Calculate the Aggregated Isotopic Distribution and Exact Center-Masses. J. Am. Soc. Mass Spectrom. 23, 753–763. https://doi.org/10.1007/s13361-011-0326-2

Coble, J.B., Fraga, C.G., 2014. Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery. J. Chromatogr. A 1358, 155–164. https://doi.org/10.1016/j.chroma.2014.06.100

Considine, E.C., Thomas, G., Boulesteix, A.L., Khashan, A.S., Kenny, L.C., 2018. Critical review of reporting of the data analysis step in metabolomics. Metabolomics 14, 7. https://doi.org/10.1007/s11306-017-1299-3

Creek, D.J., Dunn, W.B., Fiehn, O., Griffin, J.L., Hall, R.D., Lei, Z., Mistrik, R., Neumann, S., Schymanski, E.L., Sumner, L.W., Trengove, R., Wolfender, J.-L., 2014. Metabolite identification: are you sure? And how do your peers gauge your confidence? Metabolomics 10, 350–353. https://doi.org/10.1007/s11306-014-0656-8

Cui, Q., Lewis, I.A., Hegeman, A.D., Anderson, M.E., Li, J., Schulte, C.F., Westler, W.M., Eghbalnia, H.R., Sussman, M.R., Markley, J.L., 2008. Metabolite identification via the Madison Metabolomics Consortium Database. Nat. Biotechnol. 26, 162–164. https://doi.org/10.1038/nbt0208-162

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random Forests for Classification in Ecology. Ecology 88, 2783–2792. https://doi.org/10.1890/07-0539.1

D'Andrilli, J., Foreman, C.M., Marshall, A.G., McKnight, D.M., 2013. Characterization of IHSS Pony Lake fulvic acid dissolved organic matter by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry and fluorescence spectroscopy. Org. Geochem. 65, 19–28. https://doi.org/10.1016/j.orggeochem.2013.09.013

Davidson, R.L., Weber, R.J.M., Liu, H., Sharma-Oates, A., Viant, M.R., 2016. Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. GigaScience 5. https://doi.org/10.1186/s13742-016-0115-8

De Livera, A.M., Dias, D.A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., Roessner, U., McConville, M., Speed, T.P., 2012. Normalizing and integrating metabolomics data. Anal. Chem. 84, 10768–10776.

Di Guida, R., Engel, J., Allwood, J.W., Weber, R.J.M., Jones, M.R., Sommer, U., Viant, M.R., Dunn, W.B., 2016. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics 12. https://doi.org/10.1007/s11306-016-1030-9

Dudzik, D., Barbas-Bernardos, C., García, A., Barbas, C., 2018. Quality assurance procedures for mass spectrometry untargeted metabolomics. a review. J. Pharm. Biomed. Anal. 147, 149–173. https://doi.org/10.1016/j.jpba.2017.07.044

Dunn, W.B., Erban, A., Weber, R.J.M., Creek, D.J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., Viant, M.R., 2013. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics 9, 44–66. https://doi.org/10.1007/s11306-012-0434-4

Dunn, W.B., Wilson, I.D., Nicholls, A.W., Broadhurst, D., 2012. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. Bioanalysis 4, 2249–2264. https://doi.org/10.4155/bio.12.204

Dwivedi, P., Schultz, A.J., Jr, H.H.H., 2010. Metabolic profiling of human blood by high-resolution ion mobility mass spectrometry (IM-MS). Int. J. Mass Spectrom. 298, 78–90. https://doi.org/10.1016/j.ijms.2010.02.007

Fahy, E., Alvarez-Jarreta, J., Brasher, C.J., Nguyen, A., Hawksworth, J.I., Rodrigues, P., Meckelmann, S., Allen, S.M., O'Donnell, V.B., 2018. LipidFinder on LIPID MAPS: peak filtering, MS searching and statistical analysis for lipidomics. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty679

Fahy, E., Sud, M., Cotter, D., Subramaniam, S., 2007. LIPID MAPS online tools for lipid research. Nucleic Acids Res. 35, W606–W612. https://doi.org/10.1093/nar/gkm324

Farrés, M., Martrat, B., Mol, B. de, Grimalt, J.O., Tauler, R., 2015a. Extraction of climatic signals from fossil organic compounds in marine sediments up to 11.7Ma old (IODP-U1318). Anal. Chim. Acta 879, 1–9. https://doi.org/10.1016/j.aca.2015.04.051

Farrés, M., Platikanov, S., Tsakovski, S., Tauler, R., 2015b. Comparison of the variable importance inprojection (VIP) and of the selectivity ratio (SR) methods for variable selection andinterpretation. J. Chemom. 29, 528–536.

Fernández-Albert, F., Llorach, R., Andrés-Lacueva, C., Perera, A., 2014. An R package to analyse LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit). Bioinformatics 30, 1937–1939. https://doi.org/10.1093/bioinformatics/btu136

Fernandez-de-Cossio Diaz, J., Fernandez-de-Cossio, J., 2012. Computation of Isotopic Peak Center-Mass Distribution by Fourier Transform. Anal. Chem. 84, 7052–7056. https://doi.org/10.1021/ac301296a

Fernandez-Lima, F.A., Becker, C., McKenna, A.M., Rodgers, R.P., Marshall, A.G., Russell, D.H., 2009. Petroleum Crude Oil Characterization by IMS-MS and FTICR MS. Anal. Chem. 81, 9941–9947. https://doi.org/10.1021/ac901594f

Fiehn, O., Robertson, D., Griffin, J., Werf, M. van der, Nikolau, B., Morrison, N., Sumner, L.W., Goodacre, R., Hardy, N.W., Taylor, C., Fostel, J., Kristal, B., Kaddurah-Daouk, R., Mendes, P., Ommen, B. van, Lindon, J.C., Sansone, S.-A., 2007. The metabolomics standards initiative (MSI). Metabolomics 3, 175–178. https://doi.org/10.1007/s11306-007-0070-6

Folch, J., Lees, M., Stanley, G.H.S., 1956. A simple method for the isolation and purification of total lipides from animal tissues. J. Biol. Chem. 226, 497–509.

Forsberg, E.M., Huan, T., Rinehart, D., Benton, H.P., Warth, B., Hilmers, B., Siuzdak, G., 2018. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. Nat. Protoc. 13, 633–651. https://doi.org/10.1038/nprot.2017.151

Franceschi, P., Wehrens, R., 2014. Self-organizing maps: A versatile tool for the automatic analysis of untargeted imaging datasets. Proteomics 14, 853–861.

Furey, A., Moriarty, M., Bane, V., Kinsella, B., Lehane, M., 2013. Ion suppression; A critical review on causes, evaluation, prevention and applications. Talanta 115, 104–122. https://doi.org/10.1016/j.talanta.2013.03.048

Fushimi, A., Hashimoto, S., Ieda, T., Ochiai, N., Takazawa, Y., Fujitani, Y., Tanabe, K., 2012. Thermal desorption – comprehensive two-dimensional gas chromatography coupled with tandem mass spectrometry for determination of trace polycyclic aromatic hydrocarbons and their derivatives. J. Chromatogr. A 1252, 164–170. https://doi.org/10.1016/j.chroma.2012.06.068

Gathungu, R.M., Kautz, R., Kristal, B.S., Bird, S.S., Vouros, P., 2018. The integration of LC-MS and NMR for the analysis of low molecular weight trace analytes in complex matrices. Mass Spectrom. Rev. 1–19. https://doi.org/10.1002/mas.21575

Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., Duperier, C., Tremblay-Franco, M., Martin, J.-F., Jacob, D., Goulitquer, S., Thévenot, E.A., Caron, C., 2015. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics 31, 1493–1495. https://doi.org/10.1093/bioinformatics/btu813

Giancarlo, R., Bosco, G.L., Pinello, L., Utro, F., 2011. The Three Steps of Clustering in the Post-Genomic Era: A Synopsis, in: Rizzo, R., Lisboa, P.J.G. (Eds.), Computational Intelligence Methods for Bioinformatics and Biostatistics. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 13–30. https://doi.org/10.1007/978-3-642-21946-7_2

Giancarlo, R., Lo Bosco, G., Pinello, L., 2010. Distance Functions, Clustering Algorithms and Microarray Data Analysis, in: Blum, C., Battiti, R. (Eds.), Learning and Intelligent Optimization. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 125–138. https://doi.org/10.1007/978-3-642-13800-3_10

Gil de la Fuente, A., Grace Armitage, E., Otero, A., Barbas, C., Godzien, J., 2017. Differentiating signals to make biological sense - A guide through databases for MS-based non-targeted metabolomics: General. ELECTROPHORESIS 38, 2242–2256. https://doi.org/10.1002/elps.201700070

Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G., Kell, D.B., 2004. Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol. 22, 245–252. https://doi.org/10.1016/j.tibtech.2004.03.007

Gorrochategui, E., Jaumot, J., Lacorte, S., Tauler, R., 2016. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. Trends Anal. Chem. 82, 425–442. https://doi.org/10.1016/j.trac.2016.07.004

Gregory, K.E., Bird, S.S., Gross, V.S., Marur, V.R., Lazarev, A.V., Walker, W.A., Kristal, B.S., 2012. Method development for fecal lipidomics profiling. Anal. Chem. 85, 1114–1123.

Grigoriadou, A., Schwarzbauer, J., 2011. Non-target Screening of Organic Contaminants in Sediments from the Industrial Coastal Area of Kavala City (NE Greece). Water. Air. Soil Pollut. 214, 623–643. https://doi.org/10.1007/s11270-010-0451-8

Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., Goodacre, R., 2015. A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. Anal. Chim. Acta 879, 10–23. https://doi.org/10.1016/j.aca.2015.02.012

Gromski, P.S., Xu, Y., Correa, E., Ellis, D.I., Turner, M.L., Goodacre, R., 2014a. A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. Anal. Chim. Acta 829, 1–8. https://doi.org/10.1016/j.aca.2014.03.039

Gromski, P.S., Xu, Y., Kotze, H.L., Correa, E., Ellis, D.I., Armitage, E.G., Turner, M.L., Goodacre, R., 2014b. Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data. Metabolites 4, 433–452. https://doi.org/10.3390/metabo4020433

Gros, J., Reddy, C.M., Aeppli, C., Carmichael, C.A., Arey, J.S., 2014. Resolving Biodegradation Patterns of Persistent Saturated Hydrocarbons in Weathered Oil Samples from the Deepwater Horizon Disaster. Environ. Sci. Technol. 48, 1628–1637. https://doi.org/10.1021/es4042836

Gruber, B., Weggler, B.A., Jaramillo, R., Murrell, K.A., Piotrowski, P.K., Dorman, F.L., 2018. Comprehensive two-dimensional gas chromatography in forensic science: A critical review of recent trends. TrAC Trends Anal. Chem. https://doi.org/10.1016/j.trac.2018.05.017

Guijas, C., Montenegro-Burke, J.R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., Koellensperger, G., Huan, T., Uritboonthai, W., Aisporna, A.E., Wolan, D.W., Spilker, M.E., Benton, H.P., Siuzdak, G., 2018. METLIN: A Technology Platform for Identifying Knowns and Unknowns. Anal. Chem. 90, 3156–3164. https://doi.org/10.1021/acs.analchem.7b04424

Gürdeniz, G., Kristensen, M., Skov, T., Dragsted, L.O., 2012. The Effect of LC-MS Data Preprocessing Methods on the Selection of Plasma Biomarkers in Fed vs. Fasted Rats. Metabolites 2, 77–99. https://doi.org/10.3390/metabo2010077

Gustavsson, J., Wiberg, K., Ribeli, E., Nguyen, M.A., Josefsson, S., Ahrens, L., 2018. Screening of organic flame retardants in Swedish river water. Sci. Total Environ. 625, 1046–1055. https://doi.org/10.1016/j.scitotenv.2017.12.281

Gutiérrez Sama, S., Farenc, M., Barrère-Mangote, C., Lobinski, R., Afonso, C., Bouyssière, B., Giusti, P., 2018. Molecular Fingerprints and Speciation of Crude Oils and Heavy Fractions Revealed by Molecular and Elemental Mass Spectrometry: Keystone between Petroleomics, Metallopetroleomics, and Petrointeractomics. Energy Fuels 32, 4593–4605. https://doi.org/10.1021/acs.energyfuels.7b03218

Haddad, I., Hiller, K., Frimmersdorf, E., Benkert, B., Schomburg, D., Jahn, D., 2009. An Emergent Self-Organizing Map Based Analysis Pipeline for Comparative Metabolome Studies. In Silico Biol. 9, 163–178. https://doi.org/10.3233/ISB-2009-0396

Hall, G.J., Frysinger, G.S., Aeppli, C., Carmichael, C.A., Gros, J., Lemkau, K.L., Nelson, R.K., Reddy, C.M., 2013. Oxygenated weathering products of Deepwater Horizon oil come from surprising precursors. Mar. Pollut. Bull. 75, 140–149. https://doi.org/10.1016/j.marpolbul.2013.07.048

Hastie, T., Tibshirani, R., Friedman, J., 2009. Unsupervised Learning, in: The Elements of Statistical Learning, Springer Series in Statistics. Springer, New York, NY, pp. 485–585. https://doi.org/10.1007/978-0-387-84858-7_14

Haun, J., Leonhardt, J., Portner, C., Hetzel, T., Tuerk, J., Teutenberg, T., Schmidt, T.C., 2013. Online and Splitless NanoLC × CapillaryLC with Quadrupole/Time-of-Flight Mass Spectrometric Detection for Comprehensive Screening Analysis of Complex Samples. Anal. Chem. 85, 10083–10090. https://doi.org/10.1021/ac402002m

Hernández, F., Sancho, J.V., Ibáñez, M., Abad, E., Portolés, T., Mattioli, L., 2012. Current use of high-resolution mass spectrometry in the environmental sciences. Anal. Bioanal. Chem. 403, 1251–1264. https://doi.org/10.1007/s00216-012-5844-7

Heude, C., Nath, J., Carrigan, J.B., Ludwig, C., 2017. Nuclear Magnetic Resonance Strategies for Metabolic Analysis, in: Metabolomics: From Fundamentals to Clinical Applications, Advances in Experimental Medicine and Biology. Springer, Cham, pp. 45–76. https://doi.org/10.1007/978-3-319-47656-8_3

Hilton, D.C., Jones, R.S., Sjödin, A., 2010. A method for rapid, non-targeted screening for environmental contaminants in household dust. J. Chromatogr. A 1217, 6851–6856. https://doi.org/10.1016/j.chroma.2010.08.039

Hinz, C., Liggi, S., Griffin, J.L., 2018. The potential of Ion Mobility Mass Spectrometry for high-throughput and high-resolution lipidomics. Curr. Opin. Chem. Biol., Omics 42, 42–50. https://doi.org/10.1016/j.cbpa.2017.10.018

Holčapek, M., Ovčačíková, M., Lísa, M., Cífková, E., Hájek, T., 2015. Continuous comprehensive two-dimensional liquid chromatography–electrospray ionization mass spectrometry of complex

lipidomic samples. Anal. Bioanal. Chem. 407, 5033–5043. https://doi.org/10.1007/s00216-015-8528-2

Hollender, J., Schymanski, E.L., Singer, H.P., Ferguson, P.L., 2017. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environ. Sci. Technol. 51, 11505–11512. https://doi.org/10.1021/acs.est.7b02184

Holman, J.D., Tabb, D.L., Mallick, P., 2015. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. Curr. Protoc. Bioinforma. 46, 2–13.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, Kenichi, Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M.Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, Ken, Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., Nishioka, T., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass Spectrom. 45, 703–714. https://doi.org/10.1002/jms.1777

Hur, M., Kim, S., Hsu, C.S., 2017. Petroinformatics, in: Springer Handbook of Petroleum Technology, Springer Handbooks. Springer, Cham, pp. 173–198. https://doi.org/10.1007/978-3-319-49347-3_4

Hur, M., Ware, R.L., Park, J., McKenna, A.M., Rodgers, R.P., Nikolau, B.J., Wurtele, E.S., Marshall, A.G., 2018. Statistically Significant Differences in Composition of Petroleum Crude Oils Revealed by Volcano Plots Generated from Ultrahigh Resolution Fourier Transform Ion Cyclotron Resonance Mass Spectra. Energy Fuels 32, 1206–1212. https://doi.org/10.1021/acs.energyfuels.7b03061

Ieda, T., Ochiai, N., Miyawaki, T., Ohura, T., Horii, Y., 2011. Environmental analysis of chlorinated and brominated polycyclic aromatic hydrocarbons by comprehensive two-dimensional gas chromatography coupled to high-resolution time-of-flight mass spectrometry. J. Chromatogr. A 1218, 3224–3232. https://doi.org/10.1016/j.chroma.2011.01.013

Imbert, L., Gaudin, M., Libong, D., Touboul, D., Abreu, S., Loiseau, P.M., Laprévote, O., Chaminade, P., 2012. Comparison of electrospray ionization, atmospheric pressure chemical ionization and atmospheric pressure photoionization for a lipidomic analysis of Leishmania donovani. J. Chromatogr. A 1242, 75–83. https://doi.org/10.1016/j.chroma.2012.04.035

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31, 264–323. https://doi.org/10.1145/331499.331504

Jaskowiak, P.A., Campello, R.J., Costa, I.G., 2014. On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics 15, S2. https://doi.org/10.1186/1471-2105-15-S2-S2

Jaskowiak, P.A., Campello, R.J.G.B., Costa, I.G., 2013. Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis. IEEE/ACM Trans. Comput. Biol. Bioinform. 10, 845–857. https://doi.org/10.1109/TCBB.2013.9

Jiménez-Morillo, N.T., González-Pérez, J.A., Almendros, G., De la Rosa, J.M., Waggoner, D.C., Jordán, A., Zavala, L.M., González-Vila, F.J., Hatcher, P.G., 2018. Ultra-high resolution mass spectrometry of physical speciation patterns of organic matter in fire-affected soils. J. Environ. Manage. 225, 139–147. https://doi.org/10.1016/j.jenvman.2018.07.069

Jurowski, K., Kochan, K., Walczak, J., Barańska, M., Piekoszewski, W., Buszewski, B., 2017. Comprehensive review of trends and analytical strategies applied for biological samples preparation and storage in modern medical lipidomics: State of the art. TrAC Trends Anal. Chem. 86, 276–289. https://doi.org/10.1016/j.trac.2016.10.014

Kamínski, M., Kartanowicz, R., Gilgenast, E., Namieśnik, J., 2005. High-Performance Liquid Chromatography in Group-Type Separation and Technical or Process Analytics of Petroleum Products. Crit. Rev. Anal. Chem. 35, 193–216. https://doi.org/10.1080/10408340500304024

Karaman, I., 2017. Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis, in: Metabolomics: From Fundamentals to Clinical Applications, Advances in Experimental Medicine and Biology. Springer, Cham, pp. 145–161. https://doi.org/10.1007/978-3-319-47656-8_6

Kasprzak, E.M., Lewis, K.E., 2001. Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method. Struct. Multidiscip. Optim. 22, 208–218. https://doi.org/10.1007/s001580100138

Kiepper, A.P., Casilli, A., Azevedo, D.A., 2014. Depositional paleoenvironment of Brazilian crude oils from unusual biomarkers revealed using comprehensive two dimensional gas chromatography coupled to time of flight mass spectrometry. Org. Geochem. 70, 62–75. https://doi.org/10.1016/j.orggeochem.2014.03.005

Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., Wang, J., Yu, B., Zhang, J., Bryant, S.H., 2016. PubChem Substance and compound databases. Nucleic Acids Res. 44, D1202-D1213. https://doi.org/ 10.1093/nar/gkv951

Kind, T., Fiehn, O., 2007. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics 8, 105. https://doi.org/10.1186/1471-2105-8-105

Kind, T., Fiehn, O., 2006. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. BMC Bioinformatics 7, 234. https://doi.org/10.1186/1471-2105-7-234

Kirpich, A.S., Ibarra, M., Moskalenko, O., Fear, J.M., Gerken, J., Mi, X., Ashrafi, A., Morse, A.M., McIntyre, L.M., 2018. SECIMTools: a suite of metabolomics data analysis tools. BMC Bioinformatics 19. https://doi.org/10.1186/s12859-018-2134-1

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmuller, E., Dormann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A.R., Steinhauser, D., 2005. GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics 21, 1635–1638. https://doi.org/10.1093/bioinformatics/bti236

Krauss, M., Singer, H., Hollender, J., 2010. LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. Anal. Bioanal. Chem. 397, 943–951. https://doi.org/10.1007/s00216-010-3608-9

van Krevelen, D.W., 1950. Graphical-statistical method for the study of structure and reaction processes of coal. Fuel 29, 269–284.

Kronimus, A., Schwarzbauer, J., 2007. Non-target screening of extractable and non-extractable organic xenobiotics in riverine sediments of Ems and Mulde Rivers, Germany. Environ. Pollut. 147, 176–186. https://doi.org/10.1016/j.envpol.2006.08.014

Krzywinski, M., Altman, N., 2013. Power and sample size: Points of significance. Nat. Methods 10, 1139–1140. https://doi.org/10.1038/nmeth.2738

Kujawinski, E.B., 2011. The Impact of Microbial Metabolism on Marine Dissolved Organic Matter. Annu. Rev. Mar. Sci. 3, 567–599. https://doi.org/10.1146/annurev-marine-120308-081003

Kuznetsova, N., 2014. Random Forest Visualization. Eindhoven University of Technology.

Kvalhein, O.M., 1985. Scaling of analytical data. Anal. Chim. Acta 177, 71–79.

Laakia, J., Casilli, A., Araújo, B.Q., Gonçalves, F.T.T., Marotta, E., Oliveira, C.J.F., Carbonezi, C.A., Loureiro, M.R.B., Azevedo, D.A., Aquino Neto, F.R., 2017. Characterization of unusual tetracyclic compounds and possible novel maturity parameters for Brazilian crude oils using comprehensive two-dimensional gas chromatography-time of flight mass spectrometry. Org. Geochem. 106, 93–104. https://doi.org/10.1016/j.orggeochem.2016.10.012

Lanucara, F., Holman, S.W., Gray, C.J., Eyers, C.E., 2014. The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. Nat. Chem. 6, 281–294. https://doi.org/10.1038/nchem.1889

Lê Cao, K.-A., Boitard, S., Besse, P., 2011. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics 12, 253. https://doi.org/10.1186/1471-2105-12-253

Li, B., Tang, J., Yang, Q., Cui, X., Li, S., Chen, S., Cao, Q., Xue, W., Chen, N., Zhu, F., 2016. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. Sci. Rep. 6, 1–13

Li, J., Vosegaard, T., Guo, Z., 2017. Applications of nuclear magnetic resonance in lipid analyses: An emerging powerful tool for lipidomics studies. Prog. Lipid Res. 68, 37–56. https://doi.org/10.1016/j.plipres.2017.09.003

Li, X.M., Sun, G.-X., Chen, S.-C., Fang, Z., Yuan, H.-Y., Shi, Q., Zhu, Y.-G., 2018. Molecular Chemodiversity of Dissolved Organic Matter in Paddy Soils. Environ. Sci. Technol. 52, 963–971. https://doi.org/10.1021/acs.est.7b00377

Li, Z., Lu, Y., Guo, Y., Cao, H., Wang, Q., Shui, W., 2018. Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. Anal. Chim. Acta 1029, 50–57.

Löfgren, L., Forsberg, G.-B., Ståhlman, M., 2016. The BUME method: a new rapid and simple chloroform-free method for total lipid extraction of animal tissue. Sci. Rep. 6, 27688. https://doi.org/10.1038/srep27688

Mahadevan, S., Shah, S.L., Marrie, T.J., Slupsky, C.M., 2008. Analysis of Metabolomic Data Using Support Vector Machines. Anal. Chem. 80, 7562–7570. https://doi.org/10.1021/ac800954c

Mairinger, T., Causon, T.J., Hann, S., 2018. The potential of ion mobility–mass spectrometry for non-targeted metabolomics. Curr. Opin. Chem. Biol., Omics 42, 9–15. https://doi.org/10.1016/j.cbpa.2017.10.015

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Römpp, A., Neumann, S., Pizarro, A.D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., Deutsch, E.W., 2011. mzML—a Community Standard for Mass Spectrometry Data. Mol. Cell. Proteomics 10, R110.000133. https://doi.org/10.1074/mcp.R110.000133

Martins, M.C.M., Caldana, C., Wolf, L.D., Abreu, L.G.F. de, 2018. The Importance of Experimental Design, Quality Assurance, and Control in Plant Metabolomics Experiments, in: Plant Metabolomics, Methods in Molecular Biology. Humana Press, New York, NY, pp. 3–17. https://doi.org/10.1007/978-1-4939-7819-9_1

Masoom, H., Courtier-Murias, D., Farooq, H., Soong, R., Kelleher, B.P., Zhang, C., Maas, W.E., Fey, M., Kumar, R., Monette, M., 2016. Soil Organic Matter in Its Native State: Unravelling the Most Complex Biomaterial on Earth. Environ. Sci. Technol. 50, 1670–1680.

Matyash, V., Liebisch, G., Kurzchalia, T.V., Shevchenko, A., Schwudke, D., 2008. Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. J. Lipid Res. 49, 1137–1146.

Melamud, E., Vastag, L., Rabinowitz, J.D., 2010. Metabolomic Analysis and Visualization Engine for LC−MS Data. Anal. Chem. 82, 9818–9826. https://doi.org/10.1021/ac1021166

Meyers, P.A., Ishiwatari, R., 1993. Lacustrine Organic Geochemistry - An Overview of Indicators of Organic Matter Sources and Diagenesis in Lake Sediments. Org. Geochem. 20, 867–900. https://doi.org/10.1016/0146-6380(93)90100-p

Mondello, L., Tranchida, P.Q., Dugo, P., Dugo, G., 2008. Comprehensive two-dimensional gas chromatography-mass spectrometry: A review. Mass Spectrom. Rev. 27, 101–124. https://doi.org/10.1002/mas.20158

Mopper, K., Stubbins, A., Ritchie, J.D., Bialk, H.M., Hatcher, P.G., 2007. Advanced Instrumental Approaches for Characterization of Marine Dissolved Organic Matter: Extraction Techniques, Mass Spectrometry, and Nuclear Magnetic Resonance Spectroscopy. Chem. Rev. 107, 419–442. https://doi.org/10.1021/cr050359b

Morris, A.D., Letcher, R.J., Dyck, M., Chandramouli, B., Fisk, A.T., Cosgrove, J., 2018. Multivariate statistical analysis of metabolomics profiles in tissues of polar bears (Ursus maritimus) from the Southern and Western Hudson Bay subpopulations. Polar Biol. 41, 433–449.

Murray, K.K., Boyd, R.K., Eberlin, M.N., Langley, G.J., Li, L., Naito, Y., 2013. Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). Pure Appl. Chem. 85, 1515–1609. https://doi.org/10.1351/PAC-REC-06-04-06

Myers, O.D., Sumner, S.J., Li, S., Barnes, S., Du, X., 2017a. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. Anal. Chem. 89, 8689–8695. https://doi.org/10.1021/acs.analchem.7b01069

Myers, O.D., Sumner, S.J., Li, S., Barnes, S., Du, X., 2017b. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. Anal. Chem. 89, 8696–8703. https://doi.org/10.1021/acs.analchem.7b00947

Navarro-Reig, M., Jaumot, J., Tauler, R., 2018. An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis. J. Chromatogr. A. https://doi.org/10.1016/j.chroma.2018.07.017

Nizio, K.D., McGinitie, T.M., Harynuk, J.J., 2012. Comprehensive multidimensional separations for the analysis of petroleum. J. Chromatogr. A 1255, 12–23. https://doi.org/10.1016/j.chroma.2012.01.078

Nováková, L., Vlčková, H., 2009. A review of current trends and advances in modern bio-analytical methods: Chromatography and sample preparation. Anal. Chim. Acta 656, 8–35. https://doi.org/10.1016/j.aca.2009.10.004

Orchard, S., Montechi-Palazzi, L., Deutsch, E.W., Binz, P.-A., Jones, A.R., Paton, N., Pizarro, A., Creasy, D.M., Wojcik, J., Hermjakob, H., 2007. Five years of progress in the Standardization of Proteomics Data 4th Annual Spring Workshop of the HUPO-Proteomics Standards Initiative April 23–25, 2007 Ecole Nationale Supérieure (ENS), Lyon, France. PROTEOMICS 7, 3436–3440. https://doi.org/10.1002/pmic.200700658

Ouyang, X., Leonards, P., Legler, J., van der Oost, R., de Boer, J., Lamoree, M., 2015. Comprehensive two-dimensional liquid chromatography coupled to high resolution time of flight mass spectrometry

for chemical characterization of sewage treatment plant effluents. J. Chromatogr. A 1380, 139–145. https://doi.org/10.1016/j.chroma.2014.12.075

Ouyang, X., Weiss, J.M., de Boer, J., Lamoree, M.H., Leonards, P.E.G., 2017. Non-target analysis of household dust and laundry dryer lint using comprehensive two-dimensional liquid chromatography coupled with time-of-flight mass spectrometry. Chemosphere 166, 431–437. https://doi.org/10.1016/j.chemosphere.2016.09.107

Paglia, G., Angel, P., Williams, J.P., Richardson, K., Olivos, H.J., Thompson, J.W., Menikarachchi, L., Lai, S., Walsh, C., Moseley, A., Plumb, R.S., Grant, D.F., Palsson, B.O., Langridge, J., Geromanos, S., Astarita, G., 2015. Ion Mobility-Derived Collision Cross Section As an Additional Measure for Lipid Fingerprinting and Identification. Anal. Chem. 87, 1137–1144. https://doi.org/10.1021/ac503715v

Paglia, G., Astarita, G., 2017. Metabolomics and lipidomics using traveling-wave ion mobility mass spectrometry. Nat. Protoc. 12, 797–813. https://doi.org/10.1038/nprot.2017.013

Paglia, G., Williams, J.P., Menikarachchi, L., Thompson, J.W., Tyldesley-Worster, R., Halldórsson, S., Rolfsson, O., Moseley, A., Grant, D., Langridge, J., Palsson, B.O., Astarita, G., 2014. Ion Mobility Derived Collision Cross Sections to Support Metabolomics Applications. Anal. Chem. 86, 3985–3993. https://doi.org/10.1021/ac500405x

Pani, O., Górecki, T., 2006. Comprehensive two-dimensional gas chromatography (GC×GC) in environmental analysis and monitoring. Anal. Bioanal. Chem. 386, 1013–1023. https://doi.org/10.1007/s00216-006-0568-1

Parastar, H., Garreta-Lara, E., Campos, B., Barata, C., Lacorte, S., Tauler, R., 2018. Chemometrics comparison of gas chromatography with mass spectrometry and comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry Daphnia magna metabolic profiles exposed to salinity. J. Sep. Sci. 41, 2368–2379. https://doi.org/10.1002/jssc.201701336

Pena-Abaurrea, M., Jobst, K.J., Ruffolo, R., Shen, L., McCrindle, R., Helm, P.A., Reiner, E.J., 2014. Identification of Potential Novel Bioaccumulative and Persistent Chemicals in Sediments from Ontario (Canada) Using Scripting Approaches with GC×GC-TOF MS Analysis. Environ. Sci. Technol. 48, 9591–9599. https://doi.org/10.1021/es5018152

Pence, H.E., Williams, A., 2010. ChemSpider: An Online Chemical Information Resource. J. Chem. Educ. 87, 1123–1124. https://doi.org/10.1021/ed100697w

Pinto, R.C., 2017. Chemometrics Methods and Strategies in Metabolomics, in: Metabolomics: From Fundamentals to Clinical Applications, Advances in Experimental Medicine and Biology. Springer, Cham, pp. 163–190. https://doi.org/10.1007/978-3-319-47656-8_7

Pisani, O., Haddix, M.L., Conant, R.T., Paul, E.A., Simpson, M.J., 2016. Molecular composition of soil organic matter with land-use change along a bi-continental mean annual temperature gradient. Sci. Total Environ. 573, 470–480.

Pluskal, T., Castillo, S., Villar-Briones, A., Orešič, M., 2010. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11, 395. https://doi.org/10.1186/1471-2105-11-395

Poad, B.L.J., Zheng, X., Mitchell, T.W., Smith, R.D., Baker, E.S., Blanksby, S.J., 2018. Online Ozonolysis Combined with Ion Mobility-Mass Spectrometry Provides a New Platform for Lipid Isomer Analyses. Anal. Chem. 90, 1292–1300. https://doi.org/10.1021/acs.analchem.7b04091

Pollo, B.J., Alexandrino, G.L., Augusto, F., Hantao, L.W., 2018. The impact of comprehensive two-dimensional gas chromatography on oil & gas analysis: Recent advances and applications in petroleum industry. TrAC Trends Anal. Chem. 105, 202–217. https://doi.org/10.1016/j.trac.2018.05.007

Prata, P.S., Alexandrino, G.L., Mogollón, N.G.S., Augusto, F., 2016. Discriminating Brazilian crude oils using comprehensive two-dimensional gas chromatography–mass spectrometry and multiway principal component analysis. J. Chromatogr. A 1472, 99–106. https://doi.org/10.1016/j.chroma.2016.10.044

Raterink, R.-J., Lindenburg, P.W., Vreeken, R.J., Ramautar, R., Hankemeier, T., 2014. Recent developments in sample-pretreatment techniques for mass spectrometry-based metabolomics. TrAC Trends Anal. Chem. 61, 157–167. https://doi.org/10.1016/j.trac.2014.06.003

Reis, A., Rudnitskaya, A., Blackburn, G.J., Fauzi, N.M., Pitt, A.R., Spickett, C.M., 2013. A comparison of five lipid extraction solvent systems for lipidomic studies of human LDL. J. Lipid Res. 54, 1812–1824. https://doi.org/10.1194/jlr.M034330

Ren, S., Hinzman, A.A., Kang, E.L., Szczesniak, R.D., Lu, L.J., 2015. Computational and statistical analysis of metabolomics data. Metabolomics 11, 1492–1513. https://doi.org/10.1007/s11306-015-0823-6

Rivas-Ubach, A., Liu, Y., Bianchi, T.S., Tolić, N., Jansson, C., Paša-Tolić, L., 2018. Moving beyond the van Krevelen Diagram: A New Stoichiometric Approach for Compound Classification in Organisms. Anal. Chem. 90, 6152–6160. https://doi.org/10.1021/acs.analchem.8b00529

Robertson, D.G., 2005. Metabonomics in Toxicology: A Review. Toxicol. Sci. 85, 809–822. https://doi.org/10.1093/toxsci/kfi102

Rochat, B., 2017. Proposed Confidence Scale and ID Score in the Identification of Known-Unknown Compounds Using High Resolution MS Data. J. Am. Soc. Mass Spectrom. 28, 709–723. https://doi.org/10.1007/s13361-016-1556-0

Rodgers, R.P., McKenna, A.M., 2011. Petroleum Analysis. Anal. Chem. 83, 4665–4687. https://doi.org/10.1021/ac201080e

Rostkowski, P., Haglund, P., Dye, C., Schlabach, M., 2013. Non-target screening of environmental samples by low and high resolution time of flight mass spectrometry (TOF-MS). Proc. 13th Int. Conf. Envrionmental Sci. Technol. 1–4.

Rusilowicz, M., Dickinson, M., Charlton, A., O'Keefe, S., Wilson, J., 2016. A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples. Metabolomics 12, 56. https://doi.org/10.1007/s11306-016-0972-2

Ruttkies, C., Schymanski, E.L., Wolf, S., Hollender, J., Neumann, S., 2016. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J. Cheminformatics 8. https://doi.org/10.1186/s13321-016-0115-9

Sauerschnig, C., Doppler, M., Bueschl, C., Schuhmacher, R., 2017. Methanol Generates Numerous Artifacts during Sample Extraction and Storage of Extracts in Metabolomics Research. Metabolites 8, 1. https://doi.org/10.3390/metabo8010001

Scheubert, K., Hufsky, F., Böcker, S., 2013. Computational mass spectrometry for small molecules. J. Cheminformatics 5, 12. https://doi.org/10.1186/1758-2946-5-12

Schymanski, E.L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H.P., Hollender, J., 2014. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. Environ. Sci. Technol. 48, 2097–2098. https://doi.org/10.1021/es5002105

Serrano, R., Nácher-Mestre, J., Portolés, T., Amat, F., Hernández, F., 2011. Non-target screening of organic contaminants in marine salts by gas chromatography coupled to high-resolution time-of-flight mass spectrometry. Talanta 85, 877–884. https://doi.org/10.1016/j.talanta.2011.04.055

Silva, R.S.F., Aguiar, H.G.M., Rangel, M.D., Azevedo, D.A., Aquino Neto, F.R., 2011. Comprehensive two-dimensional gas chromatography with time of flight mass spectrometry applied to biomarker analysis of oils from Colombia. Fuel 90, 2694–2699. https://doi.org/10.1016/j.fuel.2011.04.026

da Silva, N., Cook, D., Lee, E.-K., 2017. Interactive Graphics for Visually Diagnosing Forest Classifiers in R. ArXiv170402502 Stat

Smith, C.A., Maille, G.O., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G., 2005. METLIN: A Metabolite Mass Spectral Database. Ther. Drug Monit. 27, 747–751. https://doi.org/10.1097/01.ftd.0000179845.53213.39

Špánik, I., Machyňáková, A., 2018. Recent applications of gas chromatography with high-resolution mass spectrometry. J. Sep. Sci. 41, 163–179. https://doi.org/10.1002/jssc.201701016

Spicer, R.A., Salek, R.M., Moreno, P., Cañueto, D., Steinbeck, C., 2017. Navigating freely-available software tools for metabolomics analysis. Metabolomics 13. https://doi.org/10.1007/s11306-017-1242-7

Stoll, D.R., Carr, P.W., 2017. Two-Dimensional Liquid Chromatography: A State of the Art Tutorial. Anal. Chem. 89, 519–531. https://doi.org/10.1021/acs.analchem.6b03506

Stubbins, A., Spencer, R.G.M., Chen, H., Hatcher, P.G., Mopper, K., Hernes, P.J., Mwamba, V.L., Mangangu, A.M., Wabakanghanzi, J.N., Six, J., 2010. Illuminated darkness: Molecular signatures of Congo River dissolved organic matter and its photochemical alteration as revealed by ultrahigh precision mass spectrometry. Limnol. Oceanogr. 55, 1467–1477. https://doi.org/10.4319/lo.2010.55.4.1467

Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H., Beger, R., Daykin, C.A., Fan, T.W.-M., Fiehn, O., Goodacre, R., Griffin, J.L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A.N., Lindon, J.C., Marriott, P., Nicholls, A.W., Reily, M.D., Thaden, J.J., Viant, M.R., 2007. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics Off. J. Metabolomic Soc. 3, 211–221. https://doi.org/10.1007/s11306-007-0082-2

Swenson, T.L., Jenkins, S., Bowen, B.P., Northen, T.R., 2015. Untargeted soil metabolomics methods for analysis of extractable organic matter. Soil Biol. Biochem. 80, 189–198. https://doi.org/10.1016/j.soilbio.2014.10.007

Swenson, T.L., Karaoz, U., Swenson, J.M., Bowen, B.P., Northen, T.R., 2018. Linking soil biology and chemistry in biological soil crust using isolate exometabolomics. Nat. Commun. 9. https://doi.org/10.1038/s41467-017-02356-9

Szczepańska, N., Rutkowska, M., Owczarek, K., Płotka-Wasylka, J., Namieśnik, J., 2018. Main complications connected with detection, identification and determination of trace organic constituents in complex matrix samples. TrAC Trends Anal. Chem. 105, 173–184. https://doi.org/10.1016/j.trac.2018.05.005

Tautenhahn, R., Cho, K., Uritboonthai, W., Zhu, Z., Patti, G.J., Siuzdak, G., 2012a. An accelerated workflow for untargeted metabolomics using the METLIN database. Nat. Biotechnol. 30, 826–828. https://doi.org/10.1038/nbt.2348

Tautenhahn, R., Patti, G.J., Rinehart, D., Siuzdak, G., 2012b. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. Anal. Chem. 84, 5035–5039. https://doi.org/10.1021/ac300698c

Tfaily, M.M., Chu, R.K., Tolić, N., Roscioli, K.M., Anderton, C.R., Paša-Tolić, L., Robinson, E.W., Hess, N.J., 2015. Advanced Solvent Based Methods for Molecular Characterization of Soil Organic Matter by High-Resolution Mass Spectrometry. Anal. Chem. 87, 5206–5215. https://doi.org/10.1021/acs.analchem.5b00116

Tfaily, M.M., Chu, R.K., Toyoda, J., Tolić, N., Robinson, E.W., Paša-Tolić, L., Hess, N.J., 2017. Sequential extraction protocol for organic matter from soils and sediments using high resolution mass spectrometry. Anal. Chim. Acta 972, 54–61. https://doi.org/10.1016/j.aca.2017.03.031

Thévenot, E.A., Roux, A., Xu, Y., Ezan, E., Junot, C., 2015. Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses. J. Proteome Res. 14, 3322–3335. https://doi.org/10.1021/acs.jproteome.5b00354

Tranchida, P.Q., Aloisi, I., Giocastro, B., Mondello, L., 2018. Current State Of Comprehensive Two-Dimensional Gas Chromatography-Mass Spectrometry With Focus On Processes Of Ionization. TrAC Trends Anal. Chem. https://doi.org/10.1016/j.trac.2018.05.016

Uclés, S., Lozano, A., Sosa, A., Parrilla Vázquez, P., Valverde, A., Fernández-Alba, A.R., 2017. Matrix interference evaluation employing GC and LC coupled to triple quadrupole tandem mass spectrometry. Talanta 174, 72–81. https://doi.org/10.1016/j.talanta.2017.05.068

van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J., 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7, 142–142. https://doi.org/10.1186/1471-2164-7-142

Veenaas, C., Haglund, P., 2017. Methodology for non-target screening of sewage sludge using comprehensive two-dimensional gas chromatography coupled to high-resolution mass spectrometry. Anal. Bioanal. Chem. 409, 4867–4883. https://doi.org/10.1007/s00216-017-0429-0

Verkh, Y., Rozman, M., Petrovic, M., 2018. Extraction and cleansing of data for a non-targeted analysis of high-resolution mass spectrometry data of wastewater. MethodsX 5, 395–402. https://doi.org/10.1016/j.mex.2018.04.008

Vinaixa, M., Schymanski, E.L., Neumann, S., Navarro, M., Salek, R.M., Yanes, O., 2016. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. TrAC Trends Anal. Chem. 78, 23–35. https://doi.org/10.1016/j.trac.2015.09.005

Walters, C.C., Wang, F.C., Higgins, M.B., Madincea, M.E., 2018. Universal biomarker analysis using GC×GC with dual FID and ToF-MS (EI/FI) detection. Org. Geochem. 115, 57–66. https://doi.org/10.1016/j.orggeochem.2017.10.003

Wang, J.-J., Pisani, O., Lin, L.H., Lun, O.O.Y., Bowden, R.D., Lajtha, K., Simpson, A.J., Simpson, M.J., 2017. Long-term litter manipulation alters soil organic matter turnover in a temperate deciduous forest. Sci. Total Environ. 607–608, 865–875. https://doi.org/10.1016/j.scitotenv.2017.07.063

Wang, S., Li, J., Shi, X., Qiao, L., Lu, X., Xu, G., 2013. A novel stop-flow two-dimensional liquid chromatography–mass spectrometry method for lipid analysis. J. Chromatogr. A 1321, 65–72. https://doi.org/10.1016/j.chroma.2013.10.069

Wang, Y., Gu, M., 2010. The Concept of Spectral Accuracy for MS. Anal. Chem. 82, 7055–7062. https://doi.org/10.1021/ac100888b

Wang, Z., Stout, S.A., Fingas, M., 2006. Forensic Fingerprinting of Biomarkers for Oil Spill Characterization and Source Identification. Environ. Forensics 7, 105–146. https://doi.org/10.1080/15275920600667104

Ward, C.P., Cory, R.M., 2015. Chemical composition of dissolved organic matter draining permafrost soils. Geochim. Cosmochim. Acta 167, 63–79. https://doi.org/10.1016/j.gca.2015.07.001

Warren, C.R., 2015. Comparison of methods for extraction of organic N monomers from soil microbial biomass. Soil Biol. Biochem. 81, 67–76. https://doi.org/10.1016/j.soilbio.2014.11.005

Weber, R.J.M., Lawson, T.N., Salek, R.M., Ebbels, T.M.D., Glen, R.C., Goodacre, R., Griffin, J.L., Haug, K., Koulman, A., Moreno, P., Ralser, M., Steinbeck, C., Dunn, W.B., Viant, M.R., 2017. Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy. Metabolomics 13, 12. https://doi.org/10.1007/s11306-016-1147-x

Wehrens, R., Hageman, J.A., van Eeuwijk, F., Kooke, R., Flood, P.J., Wijnker, E., Keurentjes, J.J.B., Lommen, A., van Eekelen, H.D.L.M., Hall, R.D., Mumm, R., de Vos, R.C.H., 2016. Improved batch correction in untargeted MS-based metabolomics. Metabolomics 12. https://doi.org/10.1007/s11306-016-1015-8

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., Ni, Y., 2018. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. Sci. Rep. 8. https://doi.org/10.1038/s41598-017-19120-0

Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J.J., van Duijnhoven, J.P.M., van Dorsten, F.A., 2008. Assessment of PLSDA cross validation. Metabolomics 4, 81–89. https://doi.org/10.1007/s11306-007-0099-6

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., Scalbert, A., 2013. HMDB 3.0—The Human Metabolome Database in 2013. Nucleic Acids Res. 41, D801–D807. https://doi.org/10.1093/nar/gks1065

Wrona, M., Mauriala, T., Bateman, K.P., Mortishire-Smith, R.J., O'Connor, D., 2005. 'All-in-One' analysis for metabolite identification using liquid chromatography/hybrid quadrupole time-of-flight mass spectrometry with collision energy switching. Rapid Commun. Mass Spectrom. 19, 2597–2602. https://doi.org/10.1002/rcm.2101

Wu, Y., Li, L., 2016. Sample normalization methods in quantitative metabolomics. J. Chromatogr. A, Editors' Choice X 1430, 80–95. https://doi.org/10.1016/j.chroma.2015.12.007

Xi, B., Gu, H., Baniasadi, H., Raftery, D., 2014. Statistical Analysis and Modeling of Mass Spectrometry-Based Metabolomics Data. Methods Mol. Biol. Clifton NJ 1198, 333–353. https://doi.org/10.1007/978-1-4939-1258-2_22

Xia, J., Wishart, D.S., 2016. Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis. Curr. Protoc. Bioinforma. 55, 14.10.1-14.10.91. https://doi.org/10.1002/cpbi.11

Xia, J., Wishart, D.S., 2011. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. Nat. Protoc. 6, 743–760. https://doi.org/10.1038/nprot.2011.319

Xu, Y., Zomer, S., Brereton, R.G., 2006. Support Vector Machines: A Recent Method for Classification in Chemometrics. Crit. Rev. Anal. Chem. 36, 177–188. https://doi.org/10.1080/10408340600969486

Yadav, S.P., 2007. The Wholeness in Suffix -omics, -omes, and the Word Om. J. Biomol. Tech. JBT 18, 277.

Yang, C., Wang, Z., Yang, Z., Hollebone, B., Brown, C.E., Landriault, M., Fieldhouse, B., 2011. Chemical Fingerprints of Alberta Oil Sands and Related Petroleum Products. Environ. Forensics 12, 173–188. https://doi.org/10.1080/15275922.2011.574312

Yang, J., Zhao, X., Lu, X., Lin, X., Xu, G., 2015. A data preprocessing strategy for metabolomics to reduce the mask effect in data analysis. Front. Mol. Biosci. 2. https://doi.org/10.3389/fmolb.2015.00004

Yao, C.-H., Liu, G.-Y., Yang, K., Gross, R., Patti, G., 2016. Inaccurate quantitation of palmitate in metabolomics and isotope tracer studies due to plastics. Metabolomics 12, 1–7. https://doi.org/10.1007/s11306-016-1081-y

Yi, L., Dong, N., Yun, Y., Deng, B., Ren, D., Liu, S., Liang, Y., 2016. Chemometric methods in data processing of mass spectrometry-based metabolomics: A review. Anal. Chim. Acta 914, 17–34. https://doi.org/10.1016/j.aca.2016.02.001

Zhang, W., Zhu, S., He, S., Wang, Y., 2015. Screening of oil sources by using comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry and multivariate statistical analysis. J. Chromatogr. A 1380, 162–170. https://doi.org/10.1016/j.chroma.2014.12.068

Zhang, X., Quinn, K., Cruickshank-Quinn, C., Reisdorph, R., Reisdorph, N., 2018. The application of ion mobility mass spectrometry to metabolomics. Curr. Opin. Chem. Biol., Omics 42, 60–66. https://doi.org/10.1016/j.cbpa.2017.11.001

Zhao, Y.-Y., Wu, S.-P., Liu, S., Zhang, Y., Lin, R.-C., 2014. Ultra-performance liquid chromatography–mass spectrometry as a sensitive and powerful technology in lipidomic applications. Chem. Biol. Interact. 220, 181–192.

Zheng, X., Aly, N.A., Zhou, Y., Dupuis, K.T., Bilbao, A., Paurus, V.L., Orton, D.J., Wilson, R., Payne, S.H., Smith, R.D., Baker, E.S., 2017. A structural examination and collision cross section database for over 500 metabolites and xenobiotics using drift tube ion mobility spectrometry. Chem Sci 8, 7724–7736. https://doi.org/10.1039/C7SC03464D

CHAPTER 3: Identifying novel treeline biomarkers in lake sediments using an untargeted screening approach

Ammar Saleem[1], Madison A. Bell[1], Linda E. Kimpe[1], Jennifer B. Korosi[2], John T. Arnason[1], Jules M. Blais[1*]

[1]Department of Biology, University of Ottawa, Ottawa ON, K1N 6N5, Canada

[2]Faculty of Liberal Arts & Professional Studies, Department of Geography, York University. Toronto, ON M3J 1P3, Canada

*Corresponding Author

Jules Blais (jules.blais@uOttawa.ca)

STATEMENT OF CONTRIBUTIONS

Madison A. Bell: manuscript writing (Introduction, method, discussion, conclusion, abstract), statistical analysis (OPLS-DA, ANOVA, t-tests, post-hoc analysis), supplementary information preparation (Table S3.1), figure generation (Figure 3.1, graphical abstract, Figure S3.3), study design, and literature search. Ammar Saleem: manuscript writing (method, results), statistical analysis (PCA, OPLS-DA), figure generation, supplementary information preparation, study design, sample analysis, and peer review. Linda E. Kimpe: study design, sample analysis, sample collection, and peer review. Jennifer B. Korosi: study design, laboratory work, sample collection, and peer review. John T. Arnason: study design, funding, and peer review. Jules M. Blais: study design, sample collection, funding, peer review, and corresponding author. *Science of the Total Environment* published this chapter:

Saleem, A., Bell, M.A., Kimpe, L., Korosi, J., Arnason, J., and Blais, J. (2019) Identifying novel treeline biomarkers in lake sediments using an untargeted screening approach. Science of the Total Environment. 694. https://doi.org/10.1016/j.scitotenv.2019.133684

ABSTRACT

Paleolimnology uses sedimentary biomarkers as proxies to reconstruct long-term changes in environmental conditions from lake sediment cores. This work describes an untargeted metabolomics-based approach and uniquely applies it to the field of paleolimnology to identify novel sediment biomarkers to track long-term patterns in treeline dynamics. We identified new potential biomarkers across the Canadian northern Arctic, non-alpine, treeline using high-resolution accurate mass spectrometry, and pattern recognition analysis. This method was applied to 120 sediment core extracts from 14 boreal, 25 forest-tundra, and 21 tundra lakes to assess long-term fluctuations in treeline position. High resolution accurate mass spectrometry resolved many compounds from complex mixtures with low mass accuracy errors. This generated a large dataset that required metabolomics styled statistical analyses to identify potential biomarkers. In total, 29 potential biomarkers discriminated between boreal and tundra lakes. Tetrapyrrole-type phorbides and squalene derivatives dominated in boreal regions, while biohopane-type lipids were in the tundra regions. Tetrapyrroles were in both surface and subsurface sediments of boreal lakes indicating these compounds can survive long-term burial in sediments. At the ecozone level, tetrapyrroles were more abundant in boreal Taiga Shield, and Taiga Plains. Boreal plant extracts belonging to Pinaceae and Ericaceae also contained tetrapyrroles. Squalene derivatives demonstrated long-term preservation, but wider distribution than tetrapyrroles. Hopanoids were present in tundra and forest-tundra lake regions, specifically the Low Arctic and Taiga

Shield, and were absent in all boreal lake sediments. Herein, we describe a method that can systematically identify new paleolimnological biomarkers. Novel biomarkers would facilitate multi-proxy paleolimnological studies and potentially lead to more accurate paleoenvironmental reconstructions.

1. INTRODUCTION

Climate change is continuous; however, recent anthropogenic warming is a major concern. Currently, the Intergovernmental Panel on Climate Change predicts exceptional large-scale warming of land and ocean temperatures in the Arctic in the coming decades (IPCC, 2018). In addition, rising temperatures will be accompanied with changing precipitation patterns, ocean level rise, glacial retreat, reduced seasonal polar ice extent, and the expansion of the boreal forest into tundra regions (Harsch et al., 2009; MacDonald, 2010; Seidl et al., 2017). Specifically, the North American non-alpine Arctic treeline is projected to reach Arctic coastlines by 2100 CE (MacDonald, 2010), or as soon as 2050 CE in some areas (Pearson et al., 2013). This has major implications for the invaded tundra ecosystems. Encroaching treelines will change regional albedo estimations, soil temperatures, nutrient/carbon cycling, rates of evapotranspiration, and species distribution – factors that can influence ecosystem feedback mechanisms and may ultimately contribute to accelerated regional warming (MacDonald, 2010; Myers-Smith and Hik, 2018; Pearson et al., 2013). In this context, the northern treeline is an important indicator of changing climate.

Treeline expansion rates are difficult to predict. Recent studies suggest current models overestimate treeline's sensitivity to climate change (Fang et al., 2013; MacDonald, 2010; Schibalski et al., 2017). Aside from warming, treeline expansion is influenced by soil composition, permafrost depth, species behaviour, and the native vegetation's resistance to invasion (Dullinger et al., 2004; Harsch et

al., 2009; Pearson et al., 2013). The complex interaction of these factors makes building models difficult; however, these models can be augmented with information about historical treeline changes. Paleo-archive data are used as reality checks against simulated models (Heiri et al., 2006; Schwörer et al., 2014). Historical treelines have been reconstructed using biogenic proxies such as pollen, macrofossils, charcoal, and diatoms (Viau and Gajewski, 2009); however, climatic reconstructions solely based on these factors can be limited because of the varying levels of proxy sensitivity, the lack of modern analogues, and few integrated studies with multiple proxies (Kaufman et al., 2004; MacDonald, 2010). Hence improved paleolimnological reconstructions of treeline dynamics and the simultaneous use of diverse multiple proxies may help build better treeline expansion models.

Sediment biomarkers are small solvent-extractable organic molecules with abundance changes that correlate with changes in an environmental parameter (i.e. temperature, pH, productivity) throughout a sediment core. The adoption of sedimentary biomarkers has been slow because of the complex chemical composition of lake sediments derived from allochthonous and autochthonous sources (Gudasz et al., 2015; Xu et al., 2019). In addition, organic diagenesis occurs as organic matter is oxidized in the water column, and during resuspension if bioturbation or physical mixing is present (Meyers and Ishiwatari, 1993). Microbes and benthic organisms in the sediment metabolize organic matter resulting in further modifications (Meyers and Ishiwatari, 1993). Thus, most small organic compounds in sediment are modified from their original forms, which makes source identification difficult. Yet, despite diagenesis, many sediment biomarkers including lignin-derived phenols and n-alkanes (Korosi et al., 2017), highly branched isoprenoids (Smik et al., 2016), branched glycerol dialkyl glycerol tetraethers (Hopmans et al., 2004; Schouten et al., 2013), sterols and stanols (Hargan et al.,

2018), bile acids (Zocatelli et al., 2017), pigments (Michelutti et al., 2010), and others have proved effective at reconstructing past environments.

To date these biomarkers have not been integrated into treeline reconstruction studies because: 1) most of the treeline studies were performed as biomarkers were being discovered, and before mass spectrometers became more universally available, and 2) biomarker discovery is slow due to the complexity of sediment organic matter. Untargeted chemical analyses represent a potentially powerful approach to discover novel biomarkers. In untargeted analyses, sample preparation strategies are used to maximize the number of compounds extracted from the sample (de Castro and Priego-Capote, 2018). The extracts are then analyzed by high-resolution chromatography coupled with accurate mass spectrometry (Hernández et al., 2012; Hollender et al., 2017). The data are processed and then visualized with appropriate pattern recognition statistical tools to identify compounds that can be assigned as potential biomarkers (Karaman, 2017; Pinto, 2017). These methods are often applied in metabolomics studies when plants are chemically compared to identify biologically active (Shang et al., 2015) or ecologically relevant biomarkers in plants (Brunetti et al., 2018, 2013), marine sediments (Beale et al., 2018, 2017; Farrés et al., 2015), free-living organisms in their natural environment (Lankadurai et al., 2013), and soil (Swenson et al., 2018, 2015).

The primary objective here is to present a unique metabolomics-derived methodology for untargeted analysis and biomarker discovery for paleolimnological reconstructions. This technique has the potential to answer questions about the composition of organic matter, to generate sediment-specific databases on organic molecules, and to discover new biomarkers for the field of paleolimnology. With this goal in mind, we combined high-resolution accurate mass spectrometry, and pattern recognition analyses to chemically compare lake sediments in Canadian boreal, tundra and transitional

forest-tundra regions to identify new potential biomarkers that can be used in future studies to map treeline migration in dated lake sediment cores. We also consider diagenesis as a factor and used a top-bottom sampling approach to assess molecular preservation in sediment.

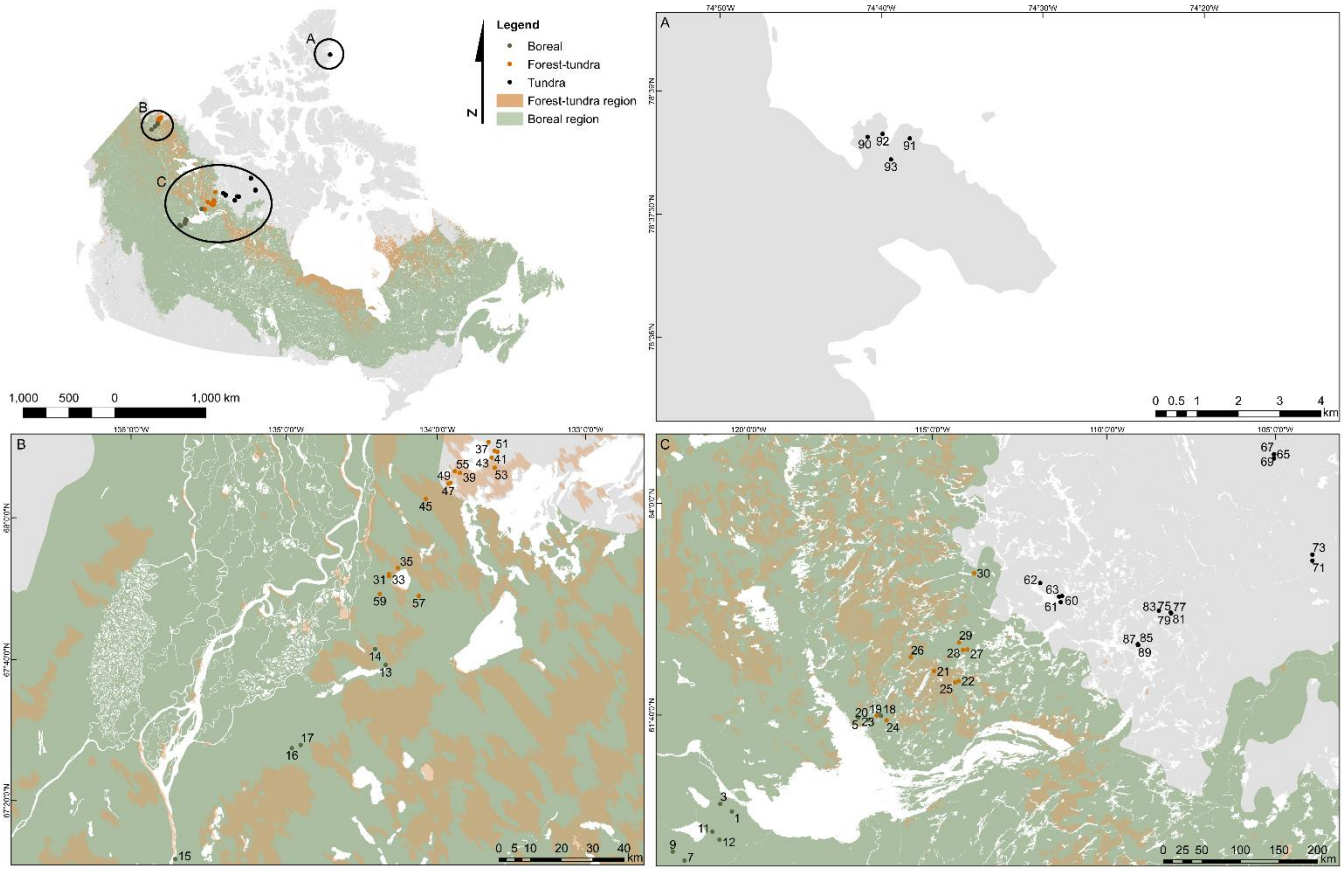


Figure 3.1: Distribution of sampled lakes across northern Canada. Dark green, orange and black dots respectively represent boreal, forest-tundra and tundra sampling sites. The highlighted green area represents the boreal biome (Brandt, 2009) and orange area represents MODIS Vegetation Continuous Fields tree cover data for the forest-tundra region (Ranson et al., 2011).

2. MATERIALS AND METHODS

2.1. Sampling

2.1.1. Sample collection

In this study, 60 lakes were selected from three ecological regions. The regions included boreal, forest-tundra, and tundra (Figure 3.1 and Table S3.1). In Figure 3.1, an approximate range of forest-tundra is defined as to where the tree cover is between 5 and 20%, and where trees are > 5m tall (Ranson et al., 2011). The boreal forest margins outlined by the Canadian forestry Service are based on past publications and current boundaries of the Taiga Shield and Taiga Plains (Brandt, 2009). The lakes were classified into unambiguously boreal or tundra, while the forest-tundra group absorbed any lake with ambiguous designations. These regions can also be sub-divided into 4 different ecozones according to the National Ecological Framework of Canada (Marshall et al., 1999) as follows: 14 boreal lakes were divided into Taiga Plains (TP) and Taiga Shield (TS1), 25 forest-tundra lakes were divided into Mackenzie Delta Uplands (MDU) and Taiga Shield (TS2), and 21 tundra lakes were divided into High Arctic (HA) and Low Arctic (LA). The Low Arctic was further divided into a forest-tundra Low Arctic (named MDU after the Mackenzie Delta Uplands), and a tundra Low Arctic (LA) based on clustering in the principal components analysis (PCA).

Cores were taken with a Glew gravity corer and polycarbonate core tubes. The cores were collected between 1996 and 2014, and were extruded into 0.25–0.5 cm intervals. The intervals were freeze dried and stored in the dark at room temperature. The sedimentation rates and chronologies were estimated in the cores using ^{210}Pb dating (Constant Rate of Supply (CRS) models), or in some cases, based on nearby lakes (Table S3.1). We used a top-bottom approach to account for diagenesis throughout the sediment. To do this, we analyzed a 'surface' (top) interval representing recently

deposited sediments, and a 'subsurface' (bottom) sediment interval from each core. Subsurface sediment intervals were chosen because they were >100 years old based on pre-existing dating profiles from 21 of our 60 lakes. The subsurface sediments from the undated 39 cores were chosen to approximate sediments >100 years old based on regional estimates of lake sedimentation rates (Table S3.1). Ideally, the subsurface sediments were deposited between 100 and 500 years ago which should avoid the climate influences from the Holocene Medieval Warming and part of the Little Ice Age; however only ^{210}Pb and ^{137}Cs dating were applied. In addition, the >100-year mark also often lies in the anoxic region of the sediment where most of the diagenetic changes have already occurred, and only very slow changes occur driven by anaerobic (mainly methanogenic) bacteria (Emerson and Hedges, 2003; Meyers and Ishiwatari, 1993).

Additionally, needle, inner bark, and cone samples were collected from boreal trees of James Bay area, Quebec, Canada to assess the potential sources of the biomarkers (Table S3.2). The sampled plants are ubiquitous to Canadian boreal and forest-tundra regions. It has been shown that organic constituents in plants of a given species can vary with season and latitude, but the molecular composition is largely under genetic control (Rapinski et al., 2014; Saleem et al., 2010). Thus, we expected the chemical profiles to contain the same compounds across regions, though abundances may vary. Additional rationale is included in the supplementary information.

2.1.2. Sample preparation

Figure S3.1 illustrates the analytical workflow for this study. Sediment samples (n = 120), and plant organs (n = 9), were freeze dried for 24 h on a Supermoduylo Freeze Dryer, Fisher Scientific, Brockville ON, Canada. A 4 mL methanol:acetonitrile:water (2:2:1) mixture was added to the freeze-

dried samples (100 mg sediment or 10 mg powdered plant) and method blanks (n = 3). The solvent mixture is a novel trisolvent system for untargeted analyses. It is not possible to optimally extract all organic molecules at once with maximum efficiency. This novel multi-solvent system extracts compounds over a wide polarity range. The mixtures were sonicated for 5 min and centrifuged at 3000 RPM for 10 min at room temperature. Supernatants were collected, and the pellet was re-extracted with the same extraction procedure. The supernatants were pooled and filtered through 0.2 µm PTFE syringe filter

2.2. Analyses

2.2.1. LC-QToF-MS analysis

Both liquid chromatography (LC) and mass spectrometer (MS) conditions were modified from Shang et al. (2015). All solvents for extraction and analyses were LCMS grade, Fisher Scientific, Brockville, ON, Canada. Analyses were performed on an Acquity ultra high performance LC (UPLC) coupled with a quadrupole time of flight (QToF) XevoG2 MS (Waters Inc., Milford, MT, USA) controlled by MassLynx™ software (Version 4.1 SCN918). Autosampler and column compartments were kept at 4 °C and 65 °C. Separations were performed on a BEH C18, 1.7 µm particle size, 2.1 mm × 100 mm column (Waters Inc., Milford, MT, USA) at a flow rate of 0.8 mL/min. Mobile phase was isocratic 95% A (water +0.1% formic acid) + 5% B (acetonitrile + 0.1% formic acid), 0.0–3.0 min, linear gradient 5–20% B, 3.0–5.0 min, 100% B isocratic 5.01–6.0 min. Sediment extracts and method blanks were injected at 5 µL, and plant extracts at 1 µL. MS was operated in resolution mode. The raw data were acquired in positive and negative ionization modes for standards listed in Table S3.4. Positive ionization mode was preferred due to better ionization efficiency and mass accuracy of our quality controls Leucine Enkephalin, and

Catechin standard mix. The electrospray ionization spectra were acquired in centroid mode to generate low energy spectra (6.0 eV), and high energy spectra (ramp 10.0 eV to 50 eV) at a scan speed of 0.08 s within the mass range of 100–1500 Da. Leucine Enkephalin ^{12}C at 556.2615 Da $[M+H]^{+1}$, was infused at 30 s interval at 2 ng/µL through z-spray to control mass accuracy throughout the acquisition. Source and desolvation temperatures were at 150 °C and 500 °C, while cone gas and desolvation gas (N_2) were at 50 L/h and 1200 L/h respectively. Catechin standard mix (Cerilliant Analytical Reference Standards) is a quality control sample containing seven closely related analogues of (+)-catechin – a metabolite commonly found in plants (Lee et al., 2017). The Catechin standard mix was used as a performance check of the LCMS. This mix was injected at the start, in the middle, and at the end of the sequence to monitor chromatographic (i.e. retention time accuracy) and spectrometric (signal intensity and mass accuracy) reproductivity and accuracy. Internal standards were not used because the chemical composition of the extracts was unknown.

2.3. Statistical Analyses

2.3.1. Pattern recognition analysis

The raw data were transferred to EZInfo™ (V8.03) to generate unsupervised (i.e. PCA), and supervised (i.e. orthogonal partial least squares discriminant analysis - OPLS-DA) plots (Pinto, 2017). The peak heights of 748 variables were normalized so that the sum of the intensities for these variables within each sample were the same. The mass exclusion list included 54 ions that were observed in blank injections and were considered background signals. A noise exclusion threshold was set at 1000 counts (i.e. ion intensity cut off), and the data were pareto-scaled. The PCA scores plot visualized the similarity between lake sediment composition. Lakes plotted closer together are more chemically similar than

those further apart, and the overall variance is represented on the axes (Pinto, 2017). Lake #13, and extract 14 t were excluded from the PCA analysis as they produced no signal above noise, and were noticeable outliers. Secondly, an OPLS-DA was performed. It is similar to PCA, but it is supervised, so the classifications (i.e. boreal, forest-tundra, and tundra) are considered during matrix construction (Pinto, 2017). Forest-tundra sediment extracts were excluded from the OPLS-DA analysis in order to maximize separation between boreal and tundra sediment (Figure 3.2B). Discriminant analyses are used to report the importance of a variable (i.e. potential biomarker) to the model, in our case we reported this as Variables of Importance (VIP) where larger numbers indicate more discriminatory biomarkers (Pinto, 2017). Thirdly, a loadings scatter plot (S-plot) from the OPLS-DA model demonstrated the potential discriminatory biomarkers as those with the greatest separation on the x- and y- axes (Figure 3.3). Lastly, trend plots for each potential biomarker depict their abundance in all 3 regions and 4 ecozones (Figures 3.4 and S3.2).

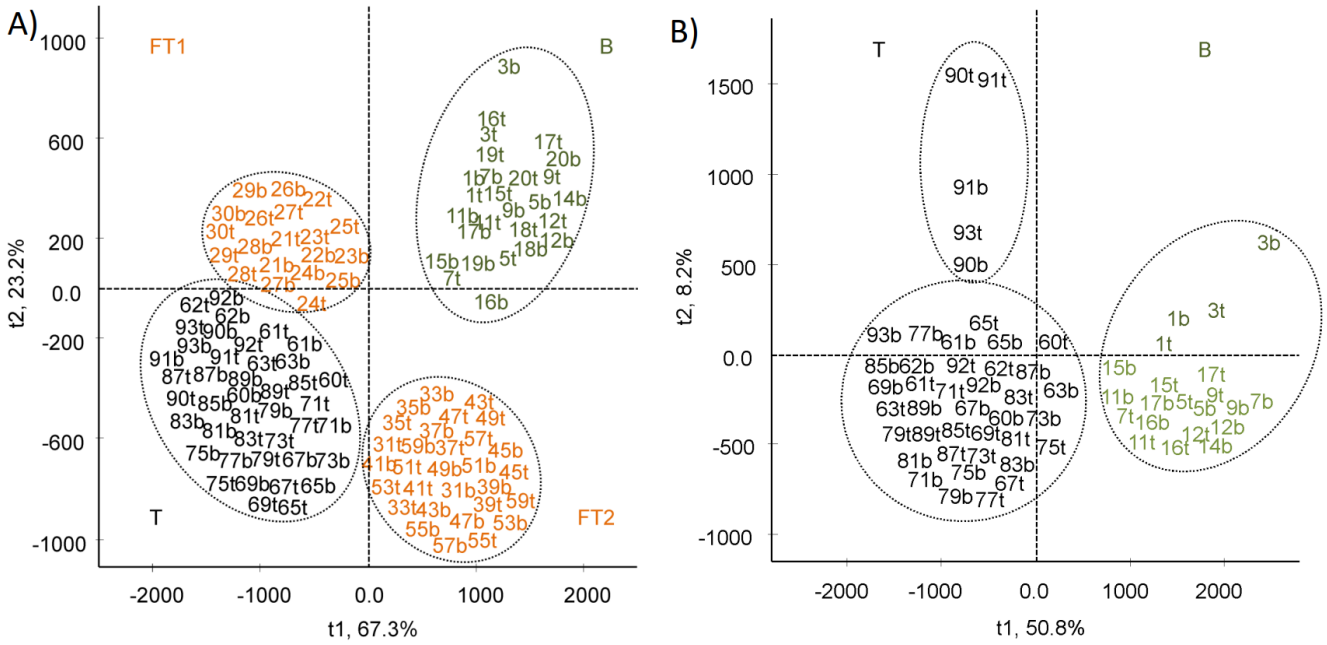


Figure 3.2: A) Pareto-scaled unsupervised PCA scores plot of shows clustering based on retention time and accurate mass; The scores t[1] and t[2] are the two most important new variables in summarizing

the dataset. Each point in the plot corresponds to an observation. Observations near each other in the plot are similar; observations far away from each other are dissimilar. The plot shows the possible presence of atypical observations, groups, similarities, trends and other patterns in the data. Atypical observations are outside the ellipse. B) Supervised OPLS-DA score plot shows the clustering of boreal (green) and tundra (black) lakes. B = boreal, T = tundra, FT = forest-tundra.

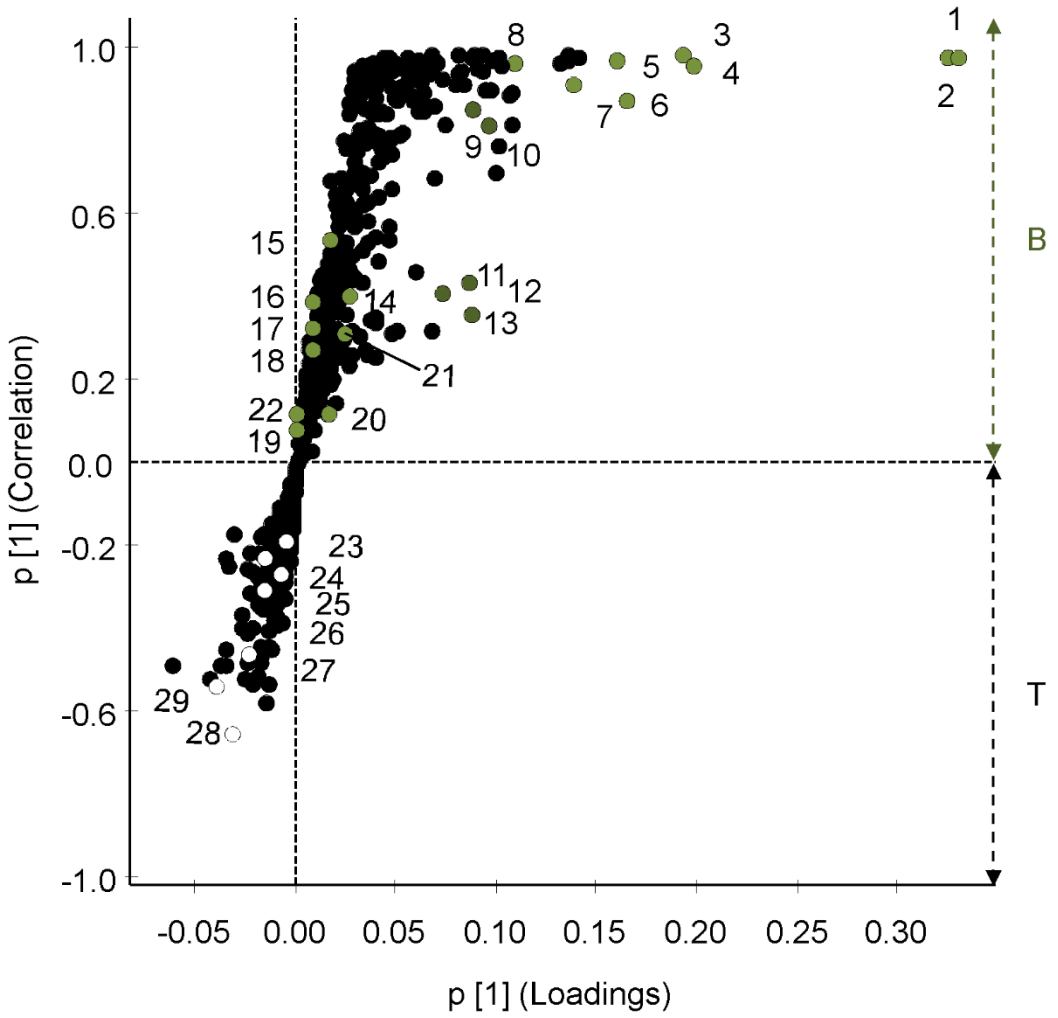


Figure 3.3: Loadings scatter plot showing the assignment of biomarkers boreal (y axis = 0 to +1) and tundra biomarkers (y axis = 0 to -1). Green circles show boreal markers defined in this study, whereas white circles show tundra biomarkers defined in this study. The covariance p[1] and correlation p(corr)[1] loadings from a two class OPLS-DA model (Group 2 vs. Group 1) are shown here in S-Plot format for Group 1. The points are Exact Mass/Retention Time (EMRT) pairs. The upper right quadrant of the S-plot shows those components which are elevated in Group 2, the control group, while the lower left quadrant shows EMRTs elevated in Group 1, the treated group. The farther along the x-axis the greater the contribution to the variance between the groups, while the farther the y-axis the higher the reliability of the analytical result. Some of the most important EMRTs are tabulated and plotted below. The measured intensities and factor of change are based on the average of the measured values for each Exact Mass/Retention Time pairs in the group.

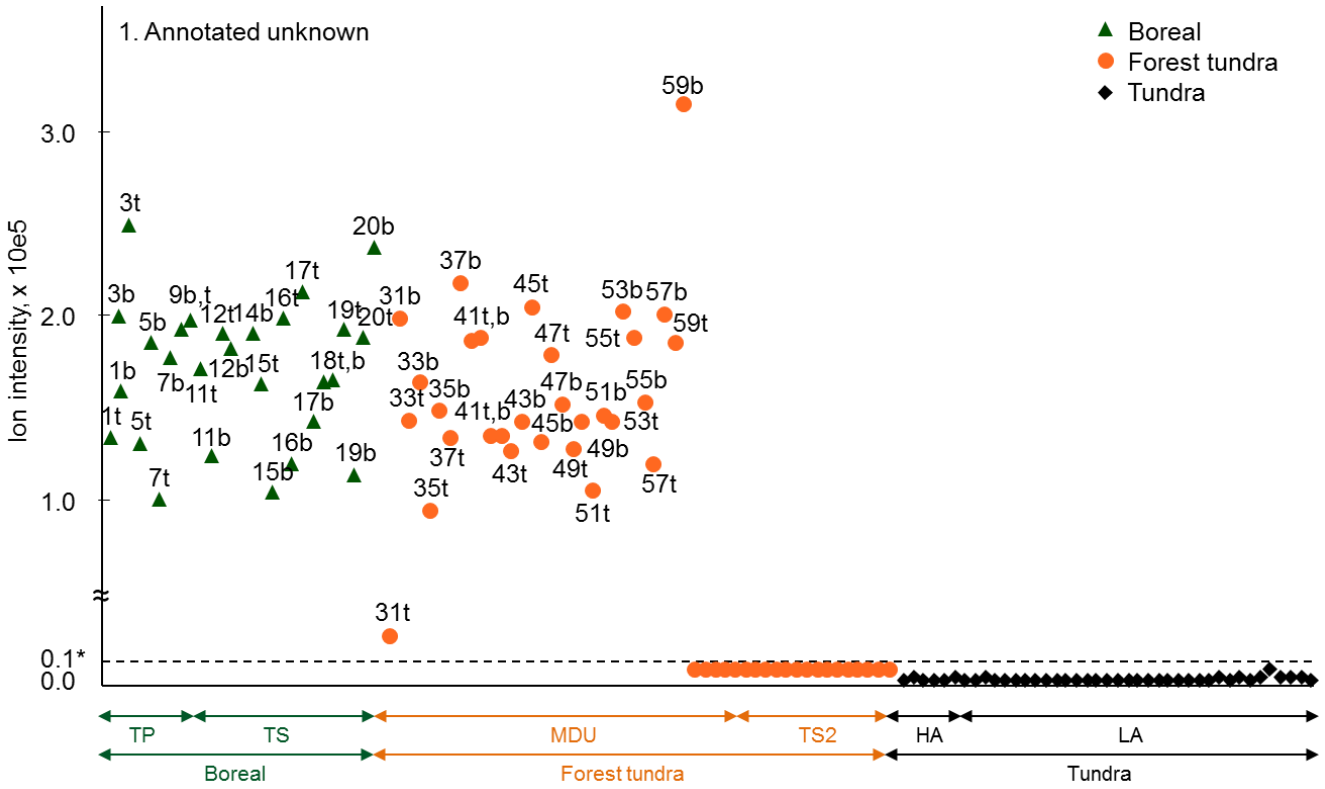


Figure 3.4: Trend plot of Annotated unknown (1) across regions (Boreal, Forest-tundra and Tundra) and ecozones TP (Taiga Plains), TS1 and TS2 (Taiga Shield), MDU (Makenzie Delta Uplands), HA (High Arctic), LA (Low Arctic). Dotted line at 0.1* on y-axis is ion intensity threshold.

2.3.2. Top-bottom comparison statistical analysis

Using R (V3.5.0), the surface and subsurface sediments (>100-years old) pareto-scaled data were compared using linear regression and ANOVA to determine whether the surface and subsurface mean intensities differed while also considering the interaction of lakes, and the accurate masses of all compounds present. The ANOVAs were followed by a TukeyHSD. A paired t-test across all lakes compared the mean intensities each of the potential biomarkers. Figure S3.3 contains figures pertaining to the data dimensionality and the ANOVA regression of surface versus subsurface.

Table 3.1: Annotation of sediment biomarkers as denoted in Fig. 3. Variable Importance Projection (VIP) represents the most influential variable in the discriminant analysis. Biomarkers identified in boreal plants: p1 (rg, rt, ka, ll, ab, pm, pba, pb, pg, ai and sp), p2 (rg and pm), p3 (rg, rt, ka and pb), which are detailed in the supplementary information.

#	Biomarker	VIP	Class	Occurrence
1	Annotated unknown	8.30	unknown	TP, TS1, MDU
2	Annotated unknown	8.20	unknown	TP, TS1, MDU
3	Annotated unknown	4.87	unknown	TP, TS1, MDU
4	Annotated unknown	4.96	unknown	TP, TS1, MDU
5	Annotated unknown	3.52	unknown	TP, TS1, MDU
6	Annotated unknown	4.31	unknown	TP, TS1, MDU
7	Annotated unknown	3.52	unknown	TP, TS1, MDU
8	Annotated unknown	3.61	unknown	TP, TS1, MDU
9	3-Phorbinepropanoic acid	2.90	tetrapyrrole	TP, TS1, MDU
10	Purpurin 18 methyl ester*	0.68	tetrapyrrole	TP, TS1, MDU
11	Chlorophyllone a	2.46	tetrapyrrole	TP, TS1, MDU
12	Pyropheophorbide a methyl ester*	2.17	tetrapyrrole	TP, TS1, MDU, p2
13	Pheophorbide a methyl ester*	2.50	tetrapyrrole	TP, TS1, MDU, p1
14	Hydroxypheophorbide a methyl ester*	0.75	tetrapyrrole	TP, TS1, p3
15	Protoporphyrin IX analogue	1.13	tetrapyrrole	TP, TS1, MDU
16	Bacteriochlorophyll b - Mg	0.60	tetrapyrrole	TS1, MDU
17	Squalene-2,3-epoxide	0.46	lipid	TP, TS1, MDU
18	Dehydrosqualene	0.39	lipid	TP, TS1, MDU, TS2, LA
19	Hydroperoxysqualene	0.07	lipid	MDU, LA
20	Pyropheophorbide a*	0.25	tetrapyrrole	TS1, TS2, LA, p2
21	Pheophorbide a*	0.49	tetrapyrrole	TS2, LA, p1
22	32,35-Anhydrobacteriohopaneterol	0.04	lipid	MDU
23	Bacteriohopane-32,33,34,35-tetrol	0.10	lipid	HA, LA
24	35-aminobacteriohopane-31,32,33,34-tetrol	0.07	lipid	TS2, LA
25	35-Aminobacteriohopane-32,33,34-triol	0.19	lipid	TS2, LA
26	N-tryptophanyl-35-aminobacteriohopane-32,33,34-triol	0.19	lipid	HA, LA
27	Adenosylhopane	0.39	lipid	TS2, LA
28	35-aminobacteriohopane-30,31,32,33,34-pentol	0.11	lipid	TS2, LA
29	Bacteriohopanetetrol cyclitol	0.09	lipid	TS2, LA

* Identification based on LCMS chromatographic and spectral data match with standards.

2.4. Identification of Biomarkers

Expected adducts from ESI ionization were auto-excluded by the MassLynx™ and manually confirmed. All observed diagnostic fragments of the identified compounds are listed in Table S3.3. Low energy spectra and high energy spectra were compared against 303 authentic standards (Table S3.4). Elemental compositions of 29 ions were determined using the elemental composition option of MassLynx™, and compared against authentic standards and Metlin™. The low energy spectra of signals were submitted to Metlin™ search using mass accuracy threshold of 5.0 PPM. Peptides and drugs were excluded from the search. The obtained results were further evaluated further with MassFragment™. MassFragment™ was used to match the parent and product ions within mass accuracies of 5 PPM for parent ions and 10 PPM for product ions (Table S3.3). To perform this step, the candidate structures were drawn in .mol format and spectral matching of the experimental data was performed with the *in silico* fragments generated by MassFragment™. The short-listed compounds were purchased as standards where possible (Table 3.1). The identification of the annotated compounds was confirmed by matching the (1) retention times, (2) elemental compositions, (3) parent ions, and (4) product ions.

3. RESULTS

3.1. Identification of compounds

The high and low energy spectra of 120 sediment extracts and 9 boreal plant extracts were compared against 303 authentic standards (see Table S3.4 for list), *in silico* predictions from Mass Fragment™, and the Metlin™ database to identify the compounds (Tables S3.4 and S3.5). Unfortunately, there is no MS/MS database available for geologically relevant compounds. The compounds identified

by the OPLS-DA model as potential biomarkers further explored using *in silico* predictions and standards when the standards were commercially available. The identified biomarkers can be shortlisted into three major classes: 1) Annotated unknowns, 2) tetrapyrroles, and 3) lipids. Table 3.1 lists the 29 identified biomarkers and their presence at the regional and ecozone levels, and unknowns are annotated according to the Metabolomics Standards Initiative (Sumner et al., 2007).

3.2. Comparison of lake sediments by PCA and OPLS-DA

The PCA scores plot (Figure 3.2A) demonstrates between group variance of principal component 1 on the x-axis and within group variance of principal component 2 on the y-axis. The total amount of variance explained by the PCA model was 90.5%. The clustering of each lake sample into boreal (B), forest-tundra (FT1 and FT2), and tundra (T) can be seen in Figure 3.2A. The B cluster contains Taiga Plains (TP) and Taiga Shield (TS1) ecozone lake sediments, while the T cluster contains Low Arctic (LA) and High Arctic (HA) lake sediments. Forest-tundra lakes clustered into two distinct groups. A cluster on top left quadrant (FT1) corresponds to samples from the Taiga Shield (TS2), and a cluster in the lower right quadrant (FT2) corresponds to lakes in the Low Arctic (MDU). This pattern indicates that the forest-tundra region may be chemically diverse in terms of lake sediments. Additionally, the closeness of FT1 with T and FT2 with B clusters indicate the forest-tundra regions share some chemical similarity with the tundra or boreal regions. In addition, surface and subsurface sediments (i.e. t and b labels for each core) clustered close to each other, indicating similar chemical composition between surface (recent) and subsurface (>100-years-old) sediment.

The OPLS-DA score plot in Figure 3.2B compares 15 boreal with 21 tundra lake sediments. The cumulative variance of 59% and between group variance of 50.8% of principal component 1 on the x-

axis further confirms clear separation of boreal and tundra lake sediments. The lake sediments within the tundra cluster demonstrate ecozone level clustering as well. This indicates that the lake sediments from the High Arctic ecozone may be different in their chemical profiles compared to those in the Low Arctic ecozone. In addition, the OPLS-DA model generated a loadings scatter plot (Figure 3.3) that demonstrates the chemical signals that are responsible for explaining most of the variance from boreal lakes (y-axis correlation 0 to +1) and tundra lakes (y-axis correlation 0 to −1). A VIP is then calculated for each biomarker from the OPLS-DA model (Table 3.1). Higher VIP scores (typically >1) indicate highly discriminant biomarkers between boreal versus tundra regions.

3.3. Top-Bottom statistical comparison

Linear regression and ANOVA showed significant differences between surface and subsurface lake sediments considering the mean intensity of all accurate masses (i.e. the accurate mass reported by the LC-QToF-MS for each signal) over all lakes (p-value < 0.05). The ANOVA containing terms for lakes and depth showed significant correlation (p-value < 0.05). A TukeyHSD post-hoc analysis demonstrated significant differences in the mean intensities of all masses between the surface and subsurface sediments of boreal lakes: 3, 7, 13, 16, 17, 18, 19, 20, forest-tundra lakes: 31, 37, 45, 57, and tundra lakes: 65, 77, 87, 90 and 93. Some lakes (i.e. lakes 16, 17, 18, 19, 45, 65, 77, 87, 90, and 93) experienced greater mean intensities in the surface than the subsurface. The other lakes (i.e. lakes 3, 7, 13, 31, 37, and 57) demonstrated mean intensities that were greater in the subsurface than the surface. An ANOVA considering the interaction of masses and depth showed no significant correlation overall (p-value > 0.05). For each individual biomarker (Table 3.1), a paired t-test comparing the mean intensities across all lakes demonstrated significant differences between surface and subsurface intensities of N-

tryptophanyl-35- aminobacteriohopane-32,33,34-triol (26), adenosylhopane (27), and 35-aminobacteriohopane-30,31,32,33,34-pentol (28).

3.4. Distribution of identified biomarkers in lake sediments

The VIP scores indicated the annotated unknowns are the most discriminant biomarkers within boreal lake sediments. Except an annotated unknown (6), CHNO is the common elemental composition for all annotated unknowns identified. Most predominant in boreal lake sediments were the tetrapyrroles (9–15). Among them, 12–14 and 19–20 were identified in the plant extracts that belong to the boreal Ericaceae, Pinaceae and Betulaceae families in Table S3.2. Unesterified tetrapyrroles including pheophorbide a (19) and pyropheophorbide a (20) are predominant in tundra and forest-tundra regions compared to their esterified form, pheophorbide a methyl ester (13), pyropheophorbide a methyl ester (12), and hydroxypheophorbide a methyl ester, which are almost exclusively prevalent in boreal lakes. Purpurin 18 methyl ester (10) occurs in the boreal region (Figures S3.2-S3.10). Chlorophyllone a (11) is found in the boreal (TS1) region. The occurrence of a protoporphryin IX analogue (15) in the boreal lake sediments followed a similar pattern as the esterified pheophorbides (10, and 12–14). Protoporphryin IX analogue was confirmed with an authentic standard; however, the retention time did not match. Hence this analyte was putatively identified as an isomer of protoporphryin IX. Bacteriochlorophyll b-Mg (16) is distributed widely throughout boreal and forest-tundra regions, and less frequently in tundra regions. Squalene derivatives are widely distributed in boreal and forest-tundra regions, including squalene-2,3-epoxide (17), dehydrosqualene (18), and hydroperoxysqualene (19). The hopanoids have lower VIP scores indicating they are not as discriminant as tetrapyrroles in our statistical model (Table 3.1 and Figure 3.3). Most hopanoids occur in the surface

sediments. Adenosylhopane (**27**) is the most abundant hopanoid and is present in both tundra (LA) and forest-tundra (TS1, TS2) regions. Bacteriohopane-32,33,34,35-tetrol (**23**) is present in the tundra (HA), while its closely related analogue 32,35-anhydrobacteriohopaneterol (**22**) is most abundant in forest-tundra (MDU). N-tryptophanyl-35- aminobacteriohopane-32,33,34-triol (**26**) was almost exclusively found in the tundra region. The remaining 35-aminobacteriohopane- 32,33,34-triol (**25**), 35-aminobacteriohopane-31,32,33,34-tetrol (**24**), 35-aminobacteriohopane-30,31,32,33,34-pentol (**28**), and Bacteriohopanetetrol cyclitol (**29**) were restricted to the forest-tundra (TS1).

4. DISCUSSION

Environmental metabolomics and sedimentomics are growing fields (Bell and Blais, 2019). There are very few examples of "omics" style studies performed in sediment (Beale et al., 2018, 2017), and even fewer specifically targeted toward biomarker identification for historical reconstructions from sediment cores (Farrés et al., 2015). To date there are no sedimentomics studies in lacustrine sediment. Although untargeted methods have been applied to examine organic contaminants in sediments (Grigoriadou and Schwarzbauer, 2011), and sewage sludge (Veenaas and Haglund, 2017), these untargeted methods do not implement the discriminant analyses that are useful for biomarker identifications so they are not considered "omics". In this work, we describe the first sedimentomics study in lacustrine sediment. We chose to study the treeline because the chemical difference between boreal and tundra lakes are well documented (Rühland et al., 2003), and it follows the sediment compositions would vary as well. Thus, application of untargeted metabolomics-derived analysis presents a novel approach to discover novel biomarkers for paleolimnological reconstructions of the Canadian Arctic treeline.

4.1. Ecological heterogeneity based on treeline margins

The study area includes 60 lakes across northern Canada (Figure 3.1). The lakes cover different geographical regions extending from the northwest near the Mackenzie Delta to the Barren Lands in Nunavut. The study lake catchments are diverse and contain boreal forest, tundra, or forest-tundra vegetation. As shown in the PCA loadings plot (Figure 3.2A), the regional classifications are represented as separate clusters. Lake sediments in the boreal forest cluster separately from tundra lake sediments, while forest-tundra lake sediments cluster separately between the boreal and tundra clusters. Thus, forest-tundra lake sediments illustrate various degrees of tundra and/or boreal forest characteristics, as expected. This regional clustering was verified by the trend plots of the potential biomarkers (Figures 3.4 and S3.2). Commonly, a biomarker that is discriminant in either the boreal region or the tundra region, will be seen in the forest-tundra transition region as well. This supports the forest-tundra as a transitional region and that the molecular composition of forest-tundra lake catchment areas exhibit both boreal and tundra lake catchment characteristics.

Separation of lake characteristics based on ecological considerations has been documented previously by Catalan et al. (2009) who defined this as the “lake district concept”. The lake district concept describes the regional grouping of lakes with similar ecological identities. They proposed two reasons lakes may have their own ecological identity: (1) abiotic environmental contributions (i.e. temperature, geology, photoperiod, ice cover, etc.) that provide evolutionary constraints on lake and terrestrial biota; and (2) metapopulation dynamics of biota in the lake. Prior studies have demonstrated the “lake district concept” across different lakes using water geochemistry (Camarero et al., 2009; Pienitz et al., 1997a, 1997b; Rühland et al., 2003), autochthonous biota (Catalan et al., 2009; Kernan et

al., 2009), and pollutants in the sediment (Catalan et al., 2009). Intuitively, this makes sense as lakes sharing similar catchments, underlying geology, and latitudes would also share similar biota and thus sedimentary organic matter compositions. Rühland et al. (2003) investigated 56 lakes spanning the treeline and determined dissolved organic carbon was higher in boreal lakes than tundra lakes further supporting the conclusion that boreal lake sediments should vary from tundra lake sediments, and even forest-tundra sediments. In addition, allochthonous organic matter has been shown to dominate organic matter deposition into lakes (Xu et al., 2019, 2015), and can be more refractory than autochthonous organic material (Gudasz et al., 2015). Thus, as the treeline represents a boundary between different allochthonous sources. It is evident that the lake sediments will reflect the different allochthonous sources and that the lake sediments should also demonstrate the lake district concept based on the treeline.

The forest-tundra subdivisions in the PCA scores plot are also notable (Figure 3.2A). For instance, the forest-tundra region demonstrated clear separation between lake sediments from the MDU (aka FT2; lakes 31–59) and the TS2 (aka FT1; lakes 21–30) (Figure 3.2B). Both the MDU and TS2 areas are located within the forest-tundra transitional region. The separation within the forest-tundra region suggests the forest-tundra is not homogeneous. Payette et al. (2001) has suggested the forest-tundra region is heterogeneous due to steep edaphic and elevation gradients. In addition, the MDU and TS2 are different from their ecozone counterparts. MDU is in the Low Arctic ecozone, but it seems to differ from its Low Arctic counterparts in the tundra region (lakes 60–89). TS2 also differs from its boreal Taiga Shield counterparts (lakes 18–20). It seems the forest-tundra region is heterogeneous, but so too are ecozones.

This heterogeneity may further emphasize the lake district concept where the "lake districts" may represent finer scale demarcations than ecological regions (i.e. boreal, forest-tundra, tundra) and/or ecozones (i.e. Taiga Shield or Low Arctic) due to other geological and hydrological differences. LA lakes are in a continuous permafrost zone with low ground ice and exposed bedrock, while the Low Arctic MDU lakes are in an area of mixed discontinuous and continuous permafrost with high ground ice content according to data from the National Snow and Ice Data Centre (Brown et al., 2002). The TS1 and TS2 lakes also share different permafrost conditions where TS1 permafrost is sporadic and the TS2 permafrost is discontinuous with more ground ice content (Brown et al., 2002). The difference between the forest-tundra MDU and TS2 may be related to permafrost ice content as well; however MDU and TS2 also exist in different geological regions where MDU is within the Arctic Coastal Plain (within the Borderlands), and TS2 is located in the Kazan region (within the Canadian Shield) according to the Geological Survey of Canada (Bostock, 2014). Thus, there are many underlying mechanisms that could be controlling the composition of organic matter in sediments into discrete lake districts. This apparent demonstration of the lake district concept bears further investigation, but may suggest a different way to classify ecological and geological regions aside from the typical ecozone designations.

4.2. Top-Bottom assessment of biomarker preservation

It was essential that the study lakes cover the ecological transition across the treeline; however, diagenesis and past ecological changes recorded in the sediment needed consideration as well. The top-bottom approach was designed to account for changes caused by diagenesis and past ecological changes (i.e. a warmer or cooler period of time). The >100-year mark was chosen to reduce the influence of past ecological changes because the last 100 years have been relatively stable although

changes in lacustrine primary productivity and/or changes in sedimentation rates cannot be completely avoided. Also, at the >100-year mark most of the diagenetic changes have already occurred (Emerson and Hedges, 2003; Meyers and Ishiwatari, 1993). Thus, we can infer if an organic molecule survives early diagenesis in measurable quantities using the top-bottom approach.

When we compared the interaction of surface and subsurface sediments with lakes and compounds identified by the untargeted analysis, there was a significant difference between surface and subsurface lake sediments. Further post-hoc analysis showed 17 of the 20 lakes were significantly different between surface and subsurface sediments. Most of the 17 lakes experienced greater mean intensities in the surface than the subsurface, which is consistent with diagenetic modifications occurring in the sediment due to microbial and/or benthic organisms (Meyers and Ishiwatari, 1993). Greater surface organic matter intensities could also be explained by increased organic matter input into the catchment, increased primary productivity (i.e. benthic organisms), and/or increased sedimentation rates causing increased deposition into the sediment. The rest of the 17 lakes experienced greater mean intensities in the subsurface than the surface, which may suggest environmental changes to the catchment (i.e. previously higher sedimentation rates or decreases in catchment vegetation and/or lake productivity). Thus, it appears that our sampling protocol was able to account for diagenetic and/or ecological related changes in the lake and/or catchment, which is useful for assessing biomarker validity.

Building on this, we aimed to determine if the compounds of interest were preserved in the sediment. Ideally, a compound should be recalcitrant if it is a useful biomarker for paleolimnological reconstructions. Overall the surface and subsurface sediments for many of our biomarkers (e.g. Figures 3.4 and S3.2) suggested good preservation. Only a few hopanoids (26–28) demonstrated significant

differences between the surface and subsurface mean intensities across all lakes suggesting they may not be preserved during deposition, and thus may not be ideal biomarkers.

4.3. Identification and paleolimnological relevance of potential biomarkers

4.3.1. Annotated unknowns

The 29 biomarkers assigned to boreal and tundra lake sediments were broadly classified into three categories: (1) annotated unknowns, (2) tetrapyrroles, and (3) lipids. The annotated unknowns were the most discriminant biomarkers of boreal lake sediments (Table 3.1). The annotated unknowns (1 to 8) are present in most of boreal and forest-tundra sediments except the forest-tundra Taiga Shield (TS2; Figures 3.4 and S3.2). These annotated unknowns did not appear in blanks so they are likely not from plastic contamination. It is notable the seven nitrogen containing biomarkers annotated in this study demonstrated preservation between the surface and subsurface samples.

One annotated unknown (6) contains the elemental composition of CHO, while the remaining seven annotated unknowns contain CHNO as common elements. The exact chemical structures of these annotated unknowns have not been determined yet, but the mass spectra for annotated unknown 1 can be seen in Figure S3.5. The presence of nitrogen in the molecular formula may represent diagenetically modified peptides, alkaloids, and/or nucleic acid derivatives (Buckinhgam et al., 2010; Knicker, 2004), or they could be amines of anoxic bacterial origin (Gale, 1941). Their predicted molecular weights range between 309 and 637 Da suggesting larger alkaloids or long chain amines/amides. When the spectra in Figure S3.5 was compared with the spectra of long chain amines/amides database Metlin™ they did not appear to match based on fragmentation patterns. Further work is needed to assign their correct chemical identity using structural elucidation techniques such as nuclear magnetic resonance

spectroscopy. The absence of a geologically relevant database of molecular structures and structural information (i.e. MS/MS, NMR, *in silico* predictions) makes the annotation of these compounds difficult, especially when sediment quantities are limited. Structural elucidation analyses for nuclear magnetic resonance spectroscopy require larger amounts of material which is not always available.

4.3.2. Tetrapyrroles

Tetrapyrroles, listed in Table 3.1, are predominant in boreal lakes. Tetrapyrrole is the name of a chemical structural family that includes plant pigments, also known as chlorins. Sedimentary tetrapyrroles are degradation products of chlorophyll and are produced by phototrophic organisms such as phytoplankton, algae, bacteria, and higher plants (Leavitt and Hodgson, 2001). Algae and bacteria are considered the major sources of lake pigments (Keely, 2006; Leavitt and Hodgson, 2001). Terrestrial chlorophyll is typically catabolized during leaf senescence and these products are exposed to further degradation during deposition and burial, whereas aquatic chlorophyll is not subject to leaf senescence and the path to sedimentation is shorter (Keely, 2006). In addition, the degradation of tetrapyrroles is also influenced by temperature and light exposure (Szymczak-Żyla et al., 2008). Thus, the accumulation of aquatically sourced chlorophyll is typically assumed to be greater than the accumulation of terrestrial chlorophyll (Keely, 2006).

Thus, sedimentary tetrapyrroles have been used to reconstruct historical lake primary productivity (Florian et al., 2015; Michelutti et al., 2007; Rühland et al., 2013), trophic status (Antoniades et al., 2011; Pienitz et al., 2000), biovector inputs (Michelutti and Smol, 2016; Stewart et al., 2015), and anthropogenic impacts on the catchment (Savage et al., 2010). The presence of sedimentary tetrapyrroles increases during periods when the aquatic environment was more productive (i.e. more

algae/bacterial growth). Thus, in regards to the treeline, it is expected that tetrapyrroles would be more prevalent in boreal regions. Boreal lakes are more productive because of greater phosphorous, dissolved organic carbon, and total nitrogen input from catchment-derived organic carbon than arctic tundra lakes (Rühland et al., 2003).

Of the tetrapyrroles present (**9–16** and **20–21**), 3- phorbinepropanoic acid (**9**), and purpurin 18 methyl ester (**10**) demonstrated the best correlation with boreal lakes. Purpurin 18 methyl ester (**10**) is an oxidation product of chlorophyll a and has been examined in Antarctic lake sediments (Keely, 2006; Squier et al., 2002). The tetrapyrroles also demonstrated lesser abundances in the forest-tundra Taiga Shield (TS2), as opposed to the boreal Taiga Shield (TS; Figure S3.2), which follows with the demonstrated ecological heterogeneity seen in the PCA (Figure 3.2A). Chlorophyllone a (**11**) is in the chlorophyll a degradation pathway and has also been identified in lake and marine sediments (Aydin et al., 2003; Keely, 2006). Bacteriochlorophyll b-Mg (**16**) is widely distributed throughout the boreal and forest-tundra regions. Its occurrence indicates the presence of photosynthetic purple or green sulphur bacteria where these organisms require dysaerobic or anoxic conditions, respectively (Keely, 2006). The presence of bacterial tetrapyrroles can be indicative of highly productive lakes, which would be consistent with boreal lakes compared to tundra lakes (Keely, 2006). Pyropheophorbide a (**20**) and pheophorbide a (**21**) share common biosynthetic origins in bacteria (Chew and Bryant, 2007), and in higher plants (Hörtensteiner and Kräutler, 2011). We confirmed their presence in the needle, cone, and inner bark extracts from the boreal plants that are listed in Table 3.1, however as pigments from aquatic environments tend to dominate the sediments, it is likely they may originate from bacteria.

Potential biomarkers **20** and **21** are distributed in forest-tundra (TS2) and the tundra (LA), however, their esterified forms (**12** and **13**) were present in boreal (TP), and forest-tundra (MDU)

regions. Interestingly, unesterified tetrapyrroles (20–21) demonstrated more selectivity for tundra regions. This seems contradictory as tundra lakes are typically less productive (Rühland et al., 2003), and the lower productivity would support fewer benthic organisms and bacteria, which would theoretically decrease the rate and/or degree of diagenesis. Tetrapyrroles often have the ester removed, or switched to another alkyl group during diagenesis (Aydin et al., 2003; Keely, 2006). Thus, tundra regions would have more esterified tetrapyrroles rather than unesterified since diagenesis is supposedly slower. However, Arctic tundra soils have demonstrated greater functional gene diversity in microbial communities than boreal soils, although the alpha diversity may be lower in tundra soils due to more acidic soils (Fierer et al., 2012). There are no studies comparing microbial communities in boreal and tundra lake sediment; however, a high functional diversity in soils may be transferred to lake sediments, which could impact diagenetic modifications of organic matter resulting in the differences between unesterified and esterified tetrapyrroles. Regardless, this could represent a future avenue of research between collaborating microbiologists and paleolimnologists.

4.3.3. Squalene derivatives

Squalene derivatives (i.e. squalene-2,3-epoxide (17), dehydrosqualene (18), and hydroperoxysqualene (19)) are widely distributed in boreal and forest-tundra regions. Squalenes are precursors to cyclic terpenes including biohopanes in bacteria, and sterols in plants, animals, algae, and fungi (Volkman, 2005). An untargeted analysis of millions of years old marine sediments correlated squalene abundance with increasing sea surface temperatures (Farrés et al., 2015). This indicates that squalene derivatives tend to preserve over time and thus may be good potential biomarkers, although the preservation potentials are likely greater for more substituted and less reactive squalene derivatives.

Squalene derivatives may also be source specific. For instance, sterols are formed from oxidosqualenes, while hopanes are formed from non-oxidized squalenes (Volkman, 2005). Thus, the presence of oxidosqualenes would indicate inputs from plants, animals, algae, and fungi; non-oxygenated squalene derivatives would then indicate bacterial sources. In addition, our results indicate squalene derivatives are preserved in the sediment. The implications for paleolimnological reconstructions of the treeline are that they may be useful biomarkers for tracking changes in productivity from both the surrounding catchment (i.e. increased oxidosqualenes), and within the lake (i.e. increased oxidosqualenes and bacterial squalenes).

4.3.4. Biohopanoids derivatives

All hopanoids are triterpenes (Ourisson and Rohmer, 1992) and biohopanes are hopanoids of direct biological origin. Biohopanes are synthesized by bacteria (Blumenberg et al., 2012; Pearson et al., 2007), and plants (Talapatra and Talapatra, 2015) where they influence cell membrane fluidity and stability (Kannenberg and Poralla, 1999; Sáenz et al., 2012). Hopanoids are found in soils (Cooke et al., 2008; Shunthirasingham and Simpson, 2006), peat bogs (Inglis et al., 2018; Talbot et al., 2016), marine and lacustrine sediments (Cooke et al., 2009; Doğrul Selver et al., 2015; Taylor and Rodger Harvey, 2011).

They have been used as paleo-environmental proxies for methane cycling (Talbot et al., 2014), pH (Inglis et al., 2018), terrestrial organic matter input to marine environments (Doğrul Selver et al., 2012), and bacterial community characterization (Höfle et al., 2015). However, the application of biohopanoids as proxies has been constrained by biohopane production from anaerobic bacteria in the buried sediments, and by the rapid diagenetic transformations of biohopanes to geohopanes (Sinninghe

Damsté et al., 2004; Talbot et al., 2016). Although, there are studies documenting biohopanes dating >2 million years old (Spencer-Jones et al., 2015). In this study our top-bottom approach suggested a lack of diagenetic preservation of biohopanoids between the surface and subsurface sediments. Specifically, biohopanoids **26–28** demonstrated a greater abundance in the surface sediment. This could be suggesting biohopanoid loss to diagenetic modifications, an assumption supported by the lack of significant difference with our other biomarkers. Alternatively, the production of biohopanes by aerobic heterotrophic bacteria may be greater in the sediment surface (Talbot et al., 2016). Thus, while biohopanes are shown to be potential biomarkers, further targeted validation studies are required to investigate biohopanes' diagenesis and production in Arctic lacustrine sediments.

In our study, the biohopanoids are limited to the forest-tundra region (TS2) and tundra (LA) lakes. The presence of hopanoids such as 35-aminobacteriohopane-31,32,33,34-tetrol (**24**) and 35-aminobacteriohopane-30,31,32,33,34-pentol (**28**) are associated with aerobic methanotrophic bacteria, which are commonly found in different soils and vary regionally (Spencer-Jones et al., 2015). Other hopanoids, such as 32,35-anhydrobacteriohopaneterol (**22**), may result from the diagenetic transformation of other hopanoids in the sediment (Schaeffer et al., 2010). Adenosylhopane (**27**) is abundantly distributed in the tundra (LA) and forest-tundra (TS2) regions (Figure S3.2-**27**), and its presence is typically associated with bacteriohopane-32,33,34,35-tetrol (**23**), and 35-aminobacteriohopane-32,33,34-triol (**25**) (Spencer-Jones et al., 2015). Adenosylhopane is frequently reported in Arctic soils and has been used to measure terrestrial soil input into Arctic marine environments (Cooke et al., 2009; Höfle et al., 2015; Talbot and Farrimond, 2007).

There could be several reasons why biohopanes seem to be more abundant in forest-tundra and tundra regions. The decreased terrestrial input into tundra lakes could make the biohopanes more

prevalent, or, as mentioned for the tetrapyrroles, boreal and tundra soils have demonstrated different microbial communities, in terms of taxa, and functional diversity (Fierer et al., 2012). Biohopane distributions are closely linked with bacterial communities and demonstrate good correlation with 16S RNA analyses (Höfle et al., 2015). The difference in microbial functional diversity may also demonstrate an environmental adaptation strategy. Bacteria seem to produce more biohopanes with acidic or alkaline conditions (Welander et al., 2009). Tundra lakes are often more acidic than boreal lakes (Rühland et al., 2003), which may increase biohopanoid production in arctic lakes. Thus, biohopanes may also be a paleolimnological biomarker for the treeline although additional validation studies are required to test this hypothesis and to also investigate the effects of diagenesis.

5. CONCLUSIONS

Our study demonstrates an untargeted metabolomics-derived approach can identify potential biomarkers that can be used to discriminate between boreal forest and tundra lake sediments. This analytical approach is useful for hypothesis generation, and as a result we have identified potential novel biomarkers that could be used as environmental proxies to reconstruct the Arctic treeline in dated lake sediment cores. The trend plots demonstrate that annotated unknowns, tetrapyrroles, and lipids can be regionally specific. In addition, tetrapyrroles or squalene derivatives could be used to reconstruct treeline movement based on primary productivity. We hypothesize lake catchments that have transitioned from tundra to boreal over time would be expected to have higher abundances of tetrapyrroles or squalene derivatives closer to the sediment surface due the increased productivity associated with increased temperature, and organic matter deposition. Biohopanes may be suitable biomarkers to ratiometrically reconstruct tundra to boreal transitions based on pH changes between

tundra and boreal sediments and changing bacterial communities. The tendency for unesterified tetrapyrrole biomarkers to be more abundant in tundra regions may suggest they would be a useful ratiometric analysis for down core analyses.

This untargeted "omics" style analysis of lake sediments will need to be combined with targeted analyses of these compounds in the future to validate the potential biomarkers we have suggested in this study. A good starting point would be to collect a core from a lake that has experienced a tundra to boreal (or vice versa) transition so that these biomarkers can be assessed. The application of untargeted analysis of organic matter in sediments is a new direction with the potential to identify new biomarkers that could be used in paleolimnological reconstructions. In addition, untargeted analysis of lake sediments can also be used to generate novel hypotheses to better understand diagenetic modifications, sediment organic molecule distributions and to build sediment specific databases on organic molecules. Small organic molecules can provide important insights for paleoecological reconstructions and also provide additional context and mechanistic understanding of ecosystem changes through time (Last and Smol, 2006), which will be very useful going forward as our climate continues to change, especially in our Arctic regions.

ACKNOWLEDGEMENTS

Funding for this study was provided by a Natural Sciences and Engineering Research Council (Canada) Discovery (RGPIN 217112-2013) and Northern Supplement (RGPNS 444180-2013) grants to JMB, and grants from the Polar Continental Shelf Program to JMB and J. Smol. We thank J. Smol for providing sediment samples from Cape Herschel (Ellesmere Island) and the TK region. We also thank J. Thienpont, D. Eickmeyer, and K. Rühland for sample collection.

SUPPLEMENTARY INFORMATION

Sampling Methods in detail

Sampling

Sample collection

In this study, 60 lakes were selected from three ecological regions. The regions included boreal, forest-tundra, and tundra (Figure 3.1 and Table S3.1). The Canadian forest-tundra region is where the boreal gradually transitions to tundra and is a complex mixture of both (Payette et al., 2001; Payette and Lavoie, 1994). The northern-most edge of the forest-tundra region is considered the treeline (Clements, 1936). In Figure 3.1, an approximate range of forest-tundra is defined as to where the tree cover is between 5-20%, and where trees are >5m tall (Ranson et al., 2011). The boreal forest margins outlined by the Canadian forestry Service are based on past publications and current boundaries of the Taiga Shield and Taiga Plains (Brandt, 2009), but these margins do not always agree with forest-tundra maps, although overall there is good correlation (Ranson et al., 2011). Thus, in light of these discrepancies, the lakes were classified as having boreal, tundra, or forest-tundra catchments according the Canadian Forestry Service (Brandt, 2009), MODIS Vegetation Continuous Fields tree cover data (Ranson et al., 2011), and field reports. The lakes were classified into unambiguously boreal forest or tundra, while the forest-tundra group absorbed any lake with ambiguous designations. These regions can also be sub-divided into 4 different ecozones according to the National Ecological Framework of Canada (Marshall et al., 1999) as follows: 14 boreal lakes were divided into Taiga Plains (TP) and Taiga Shield (TS1), 25 forest-tundra lakes were divided into Mackenzie Delta Uplands (MDU) and Taiga Shield (TS2), and 21 tundra lakes were divided into High Arctic (HA) and Low Arctic (LA). The Low Arctic was further divided into a

forest-tundra Low Arctic (named MDU after the Mackenzie Delta Uplands), and a tundra Low Arctic (LA) based on clustering in the principal components analysis (PCA).

Cores were taken with a Glew gravity corer and polycarbonate core tubes. The cores were collected between 1996 and 2014, and were extruded into 0.25-0.5 cm intervals. The intervals were freeze dried and stored in the dark at room temperature. The sedimentation rates and chronologies were estimated in the cores using ^{210}Pb dating (Constant Rate of Supply (CRS) models), or in some cases, based on nearby lakes (Table S3.1). We used a top-bottom approach to account for diagenesis throughout the sediment. The modification of non-refractory organic matter occurs faster at the surface as burial occurs (Gälman et al., 2009; Sobek et al., 2014). To do this, we analyzed a 'surface' (top) interval representing recently deposited sediments, and a 'subsurface' (bottom) sediment interval from each core. Subsurface sediment intervals were chosen because they were >100 years old based on pre-existing dating profiles from 21 of our 60 lakes. The subsurface sediments from the undated 39 cores were chosen to approximate sediments >100 years old based on regional estimates of lake sedimentation rates (Table S3.1). Ideally, the subsurface sediments were deposited between 100-500 years ago which should avoid the climate influences from the Holocene Medieval Warming and the Little Ice Age; however only ^{210}Pb and ^{137}Cs dating were applied. In addition, the >100-year mark also often lies in the anoxic region of the sediment where most of the diagenetic changes have already occurred, and only very slow changes occur driven by anaerobic (mainly methanogenic) bacteria (Emerson and Hedges, 2003; Meyers and Ishiwatari, 1993). So any small organic molecules preserved for 100 years are likely resistant to diagenesis. Resistance to diagenesis suggests the potential biomarkers are ideal candidates for long-term reconstructions spanning millennia as opposed to decades.

Additionally, needle, inner bark, and cone samples were collected from boreal trees of James Bay area, Quebec, Canada to assess the potential sources of the biomarkers (Table S3.2). The sampled plants (*Rhododendron groenlandicum, Rhododendron tomentosum, Kalmia angustifolia, Larix laricina, Abies balsamea, Picea mariana, Pinus banksiana, Pinus balsamifera, Picea glauca, Alnus incana,* and *Salix planifolia*) are ubiquitous to Canadian boreal and forest-tundra regions. It has been shown that organic constituents in plants of a given species can vary with season and latitude, but the molecular composition is largely under genetic control (Rapinski et al., 2014; Saleem et al., 2010). Thus, we expected the chemical profiles to contain the same compounds across regions, though abundances may vary.

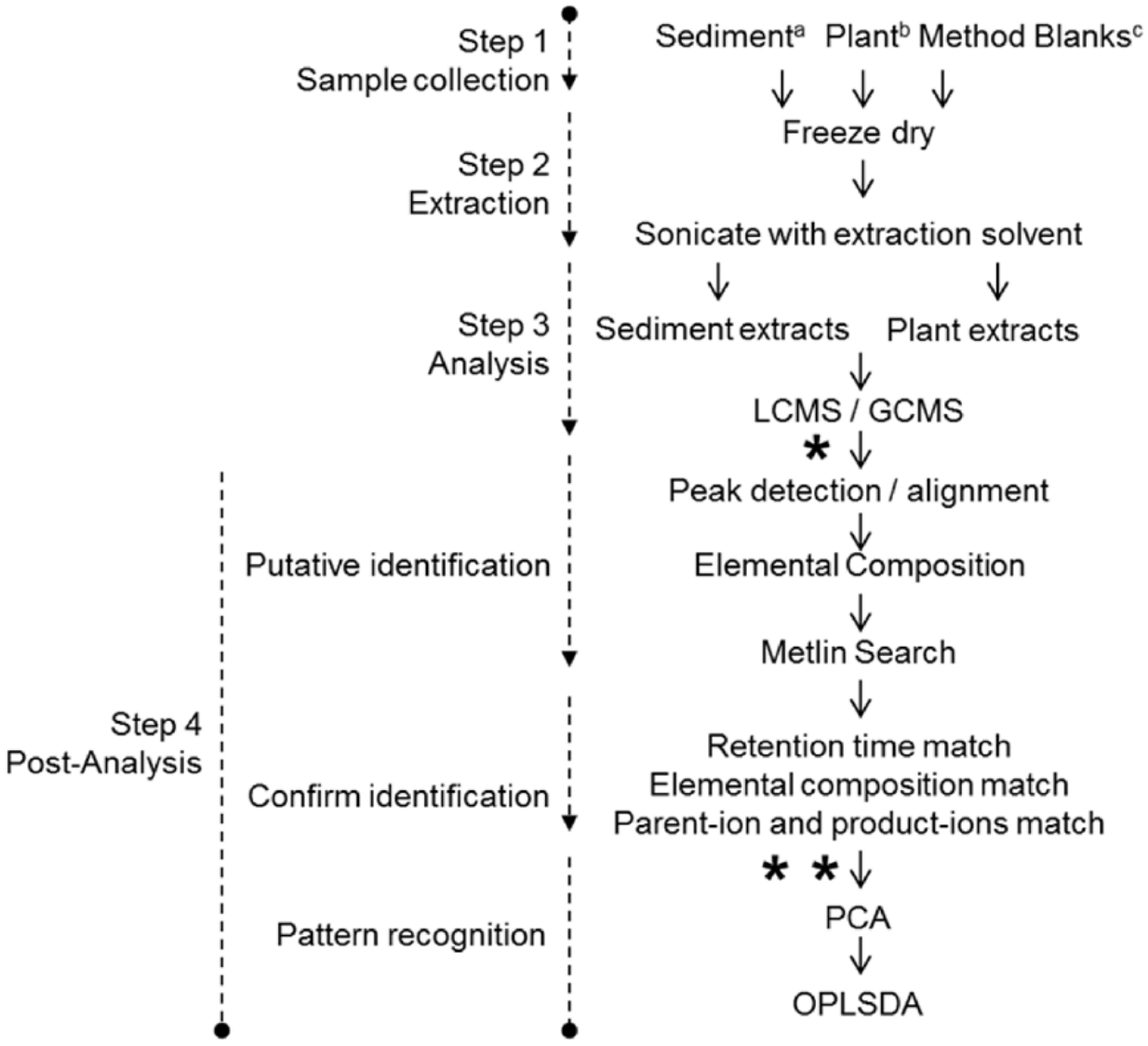
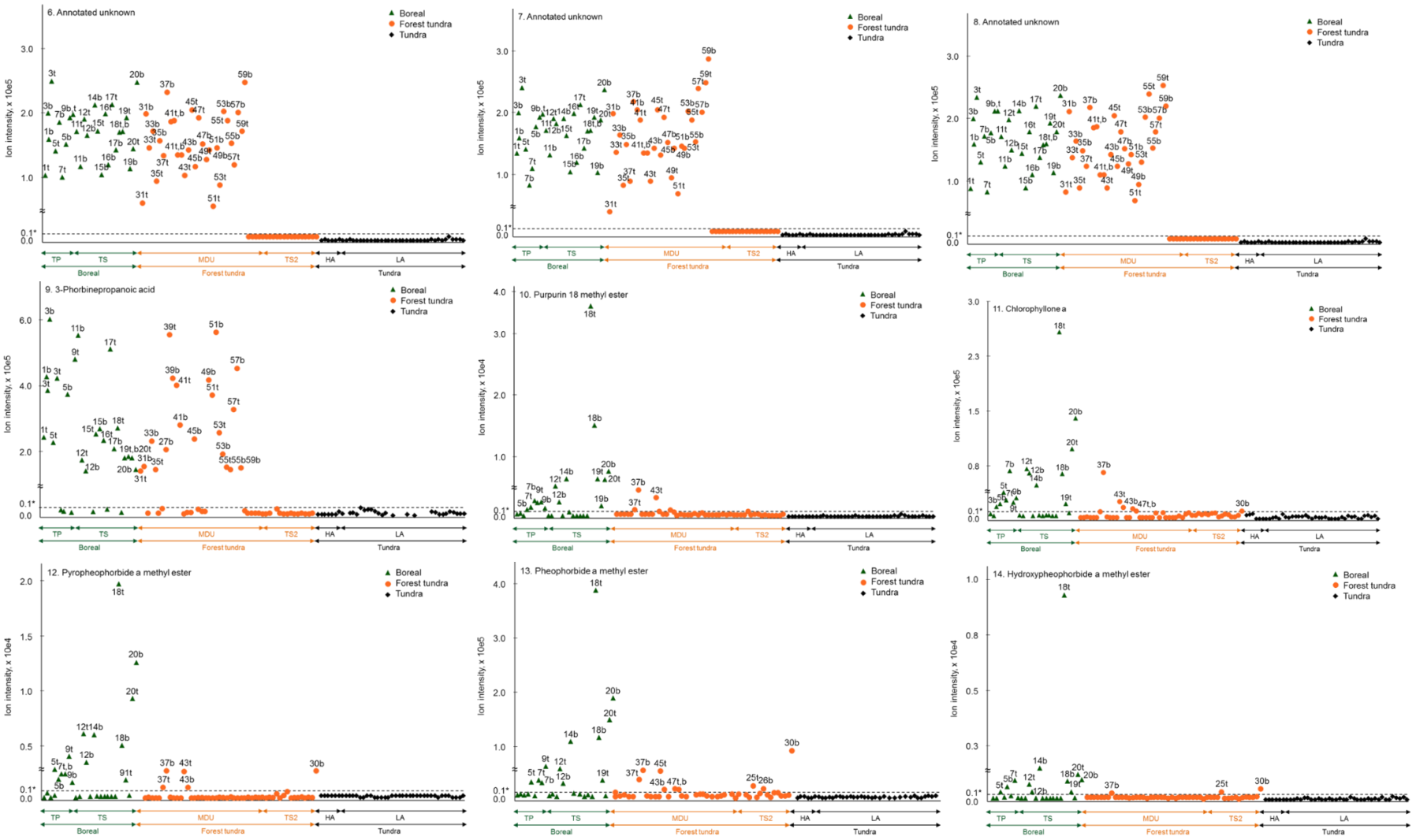
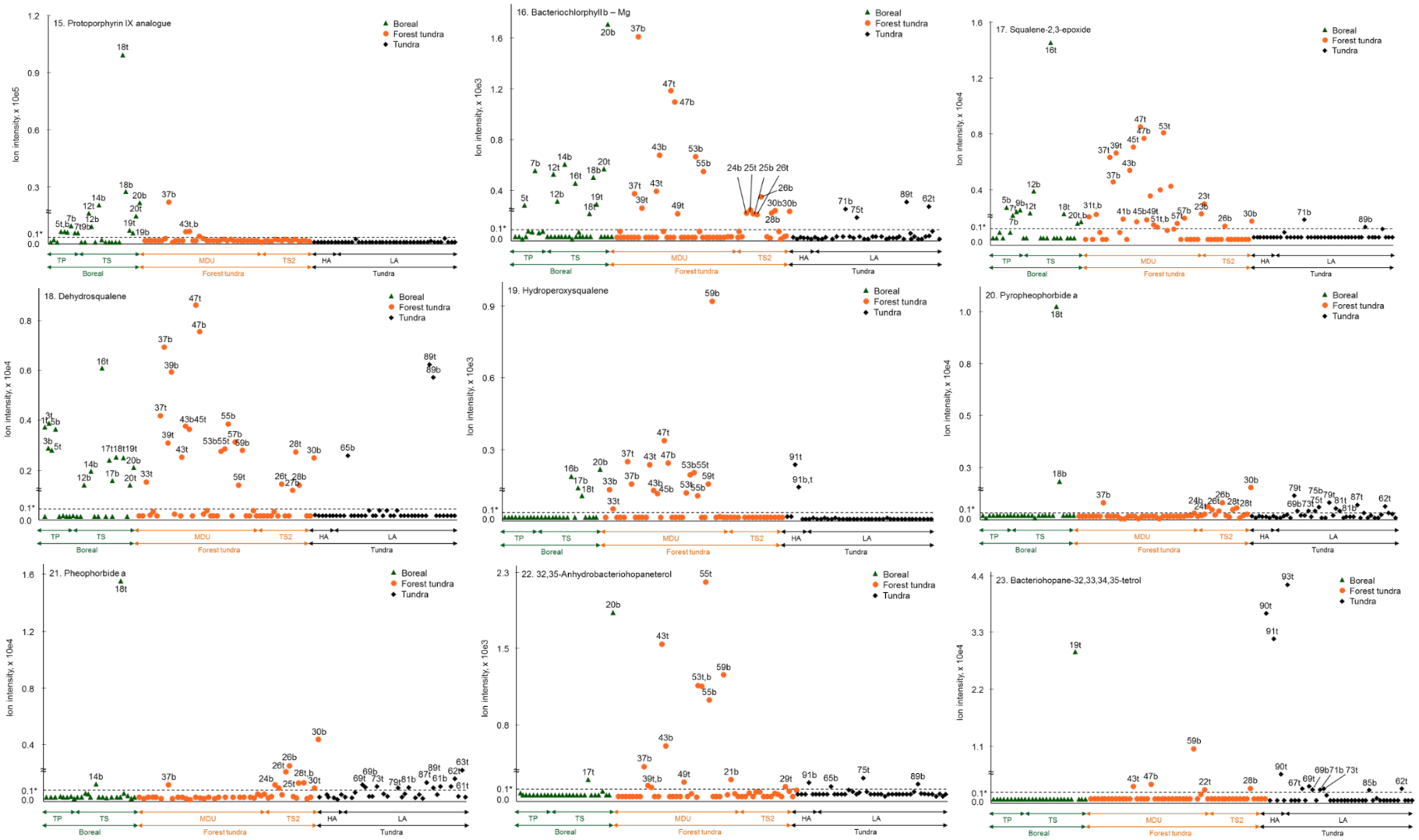


Figure S3.1: Analytical workflow.


6. Annotated unknown
7. Annotated unknown
8. Annotated unknown
9. 3-Phorbinepropanoic acid
10. Purpurin 18 methyl ester
11. Chlorophyllone a
12. Pyropheophorbide a methyl ester
13. Pheophorbide a methyl ester
14. Hydroxypheophorbide a methyl ester
Boreal
Forest tundra
Tundra
Ion intensity, x 10e5
Ion intensity, x 10e4
Ion intensity, x 10e6
TP
TS
MDU
TS2
HA
LA


15. Protoporphyrin IX analogue
16. Bacteriochlorphyll b – Mg
17. Squalene-2,3-epoxide
18. Dehydrosqualene
19. Hydroperoxysqualene
20. Pyropheophorbide a
21. Pheophorbide a
22. 32,35-Anhydrobacteriohopaneterol
23. Bacteriohopane-32,33,34,35-tetrol
Boreal
Forest tundra
Tundra
Ion intensity, x 10e5
Ion intensity, x 10e4
Ion intensity, x 10e3
TP
TS
MDU
TS2
HA
LA

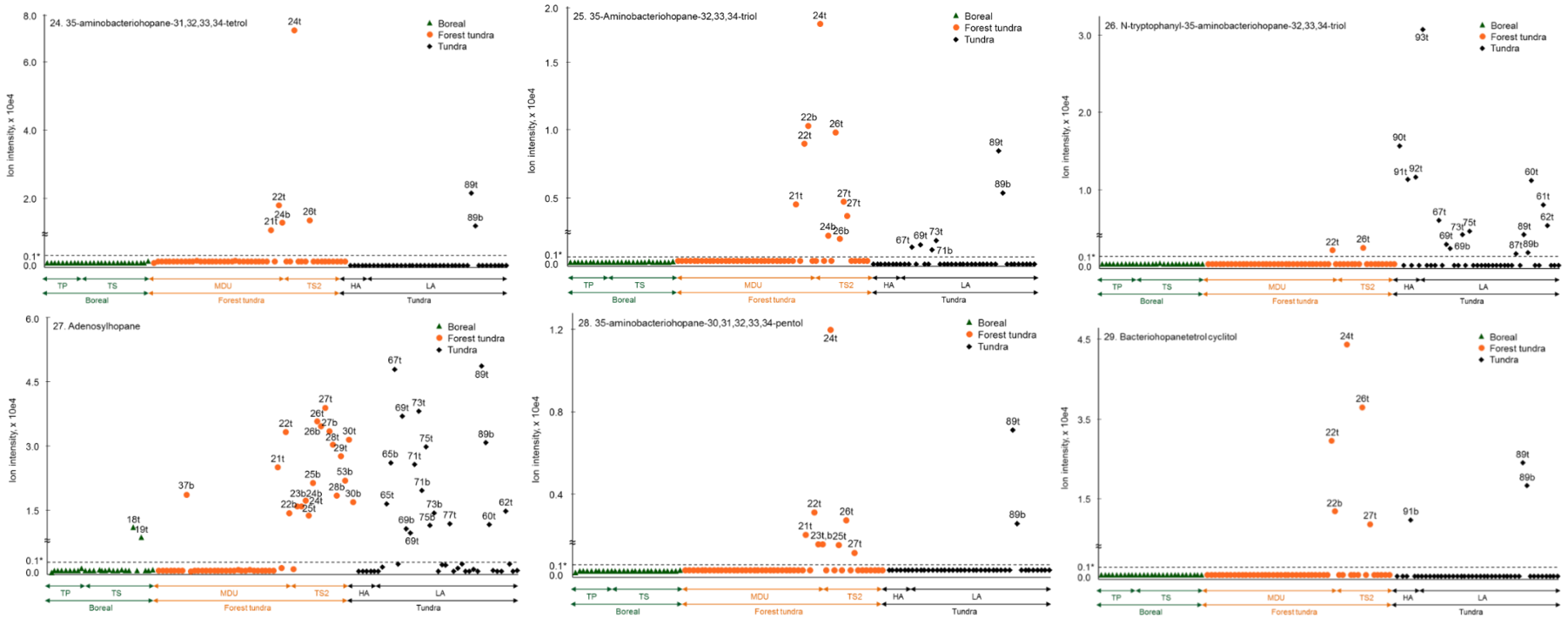


Figure S3.2: Trend plots showing the distribution of biomarkers **2-29** in regions (i.e. Boreal, Forest-tundra and Tundra) and ecozones TP (Taiga Plains), TS1 and TS2 (Taiga Shield), MDU (Makenzie Delta Uplands), HA (High Arctic), LA (Low Arctic). Dotted line at 0.1* on y-axis is ion intensity threshold cut off.

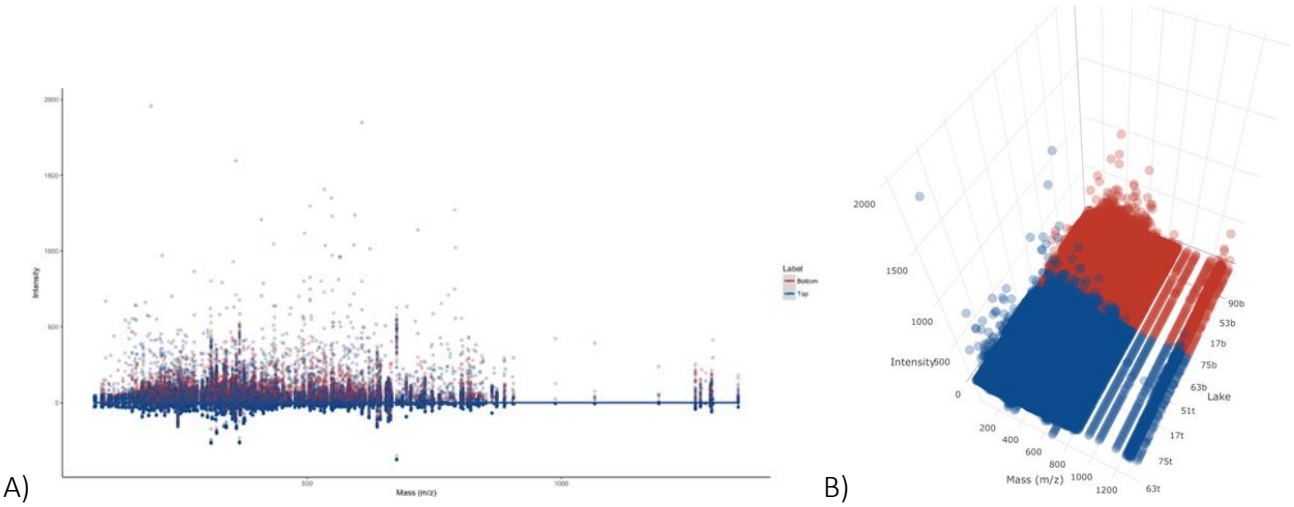


Figure S3.3: Scatterplots of data. A) 3D scatterplot of Mass (m/z) vs. Lakes vs. Intensities with coloring according to top and bottom sediment intervals; B) 2D scatterplot of Intensity vs. Mass (m/z) including regression lines for top vs. bottom.

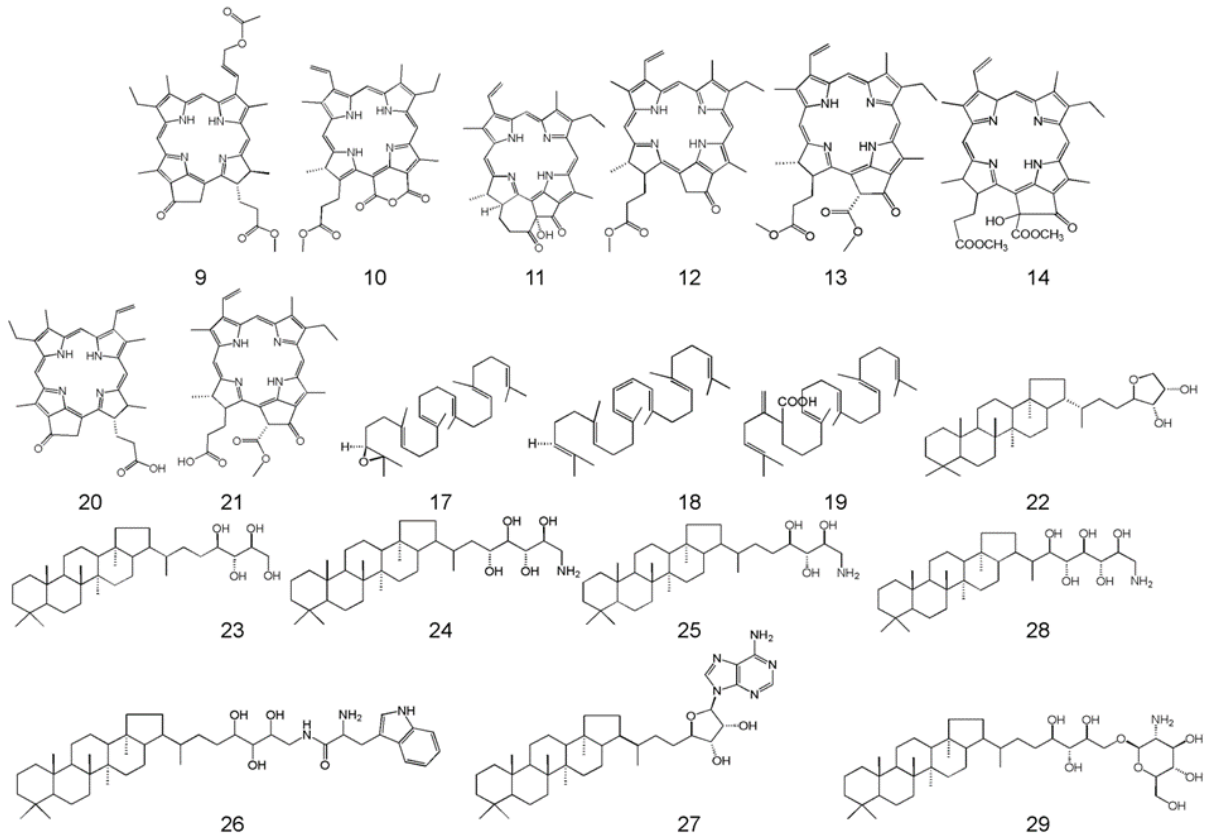


Figure S3.4: Chemical structures of identified markers.

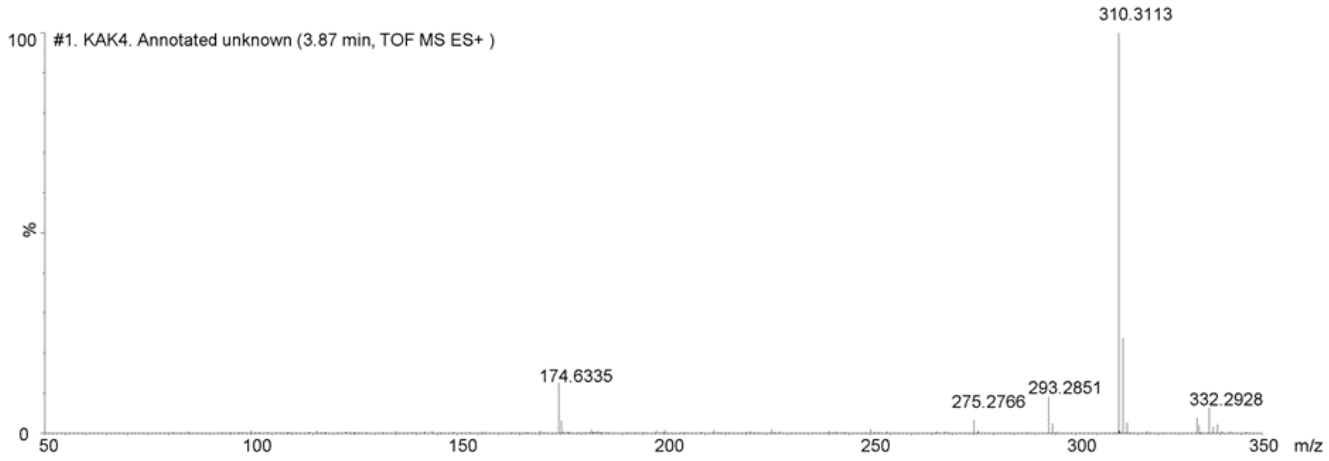


Figure S3.5: Accurate mass spectrum of annotated unknown #1.

Table S3.1: Regional and ecozone assignments and dating of the lakes.

I.D.	Latitude N	Longitude W	Core depth (cm)**	Date Cored	Ecozone*	Sedimentation Rate (g/cm^2/year)	Comments$^{\$}$
				Boreal Region			
1	60.91025	116.82242	6-6.5, 34-35	2012	TP	0.001	sporadic permafrost, dating for area suggests 1850's for ~13.5 cm for Kakisa lake area
3	60.94997	117.15683	3.5-4, 28-30	2012	TP	0.001	sporadic permafrost, dating for area suggests 1850's for ~13.5 cm for Kakisa lake area
5	62.4612	114.55822	1.5-2, 38-40	2012	TP	0.0029$^{\$}$	discontinuous permafrost, dating for area suggests 1850's for ~13.5 cm for Yellowknife area
7	60.16722	117.5	1.5-2, 24-26	2012	TP	0.0085	sporadic permafrost, dating for area suggests 1850's for ~8-15 cm for Cameron Hills
9	60.21758	117.83989	1.5-2, 34-36	2012	TP	0.0085	sporadic permafrost, dating for area suggests 1850's for ~8-15 cm for Cameron Hills
11	60.60667	117.10333	1.5-2, 12-13	2012	TP	0.0842	sporadic permafrost, dating for area suggests 1850's for ~14 cm for Tathlina lake area
12	60.54161	116.87644	1.5-2, 20-20.5	2012	TP	0.0842	sporadic permafrost, dating for area suggests 1850's for ~14 cm for Tathlina lake area
13^{x}	68.29547	133.28209	1-2, 10-11	2009	TP	no data	continuous permafrost
14^{x}	68.31683	133.42075	25-26	2009	TP	no data	discontinuous permafrost
15	67.47382	133.77003	1-2, 7-8	2009	TP	no data	continuous permafrost
16	67.93813	133.5176	0-1, 15-16	2009	TP	no data	continuous permafrost
17	67.96019	133.47489	1-2, 10-11	2009	TP	no data	continuous permafrost
18	62.56532	114.01318	3.5-4, 20-22	2012	TS1	no data	discontinuous permafrost
19	62.56122	114.00532	6-6.5, 32-34	2012	TS1	no data	discontinuous permafrost
20	62.47525	114.29353	2.5-3, 40-42	2013	TS1	no data	discontinuous permafrost
				Forest-tundra Region			
21	63.25861	113.00611	0-1, 16-17	1996	TS2	low	discontinuous permafrost
22	63.22694	112.30472	0-1, 23-24	1996	TS2	low	discontinuous permafrost
23	62.55333	114.12028	0-1, 12-13	1998	TS2	low	discontinuous permafrost
24	62.53028	113.83889	0-0.5, 29.5-30	1998	TS2	low	discontinuous permafrost

25	63.19417	112.38972	0.5-2, 27.5-28.5	1998	TS2	low	discontinuous permafrost
26	63.33917	113.70833	0.5-1.5, 13-14	1998	TS2	low	discontinuous permafrost
27	63.6175	112.30472	0-0.5, 14-15	1996	TS2	low	discontinuous permafrost
28	63.60028	112.42	1-1.5, 9.5-10.5	1998	TS2	low	discontinuous permafrost
29	63.6715	112.57778	0.5-1.0, 16-17	1998	TS2	low	discontinuous permafrost
30	64.51889	112.69639	0-0.25, 23-23.2	1998	TS2	0.0065[$]	continuous permafrost, dating for area suggests 1850's for ~10 cm
31	68.52074	133.67177	2.8-3, 24-24.4	2013	MDU	0.0124[$]	discontinuous permafrost, dating for area suggests 1850's for ~9 cm for Noel lake area
33	68.51609	133.66103	1.5-2, 18.5-19	2011	MDU	0.0091[$]	discontinuous permafrost, dating for area suggests 1850's for ~9 cm for Noel Lake area
35	68.55117	133.63983	1.5-2.25, 35-35.5	2013	MDU	0.0157[$]	discontinuous permafrost, dating for area suggests 1850's for ~10 cm for Noel Lake area
37	69.006	133.61962	0-1, 20-21	2009	MDU	low	continuous permafrost
39	68.88382	133.66661	1.0-2.0, 15-16	2009	MDU	low	continuous permafrost
41	68.99774	133.51292	0-1, 15-16	2009	MDU	0.008[$]	continuous permafrost, dating for area suggests 1850's for ~8 cm for Parsons Lake area
43	68.97418	133.52522	0-1, 20-21	2009	MDU	0.004[$]	continuous permafrost, dating for area suggests 1850's for ~10 cm for Parsons Lake area
45	68.7635	133.77346	0-1, 20-21	2009	MDU	low	continuous permafrost
47	68.83861	133.69388	0-1, 15-16	2009	MDU	low	continuous permafrost
49	68.84477	133.68295	0-1, 20-21	2009	MDU	low	continuous permafrost
51	68.99532	133.53802	1-2, 15-16	2009	MDU	0.01[$]	continuous permafrost, dating for area suggests 1850's for ~14 cm for Parsons Lake area
53	68.95486	133.45729	1-2, 10-11	2009	MDU	0.01[$]	continuous permafrost, dating for area suggests 1850's for ~9 cm for Parsons Lake area
55	68.87971	133.70848	1-2, 20-21	2009	MDU	0.009[$]	continuous permafrost, dating for area suggests 1850's for ~8 cm for Parsons Lake area

57	68.51843	133.37436	1.5-2, 20.5-21	2013	MDU	0.0842[$]	discontinuous permafrost, dating for area suggests 1850's for ~17 cm for Noel Lake area
59	68.45779	133.64032	2-2.5, 20.5-21	2013	MDU	0.0124[$]	discontinuous permafrost, dating for area suggests 1850's for ~9 cm for Noel lake area
					Tundra Region		
60	64.525	110.1634	0.25-0.5, 16.25-16.75	2014	LA	0.0138[$]	continuous permafrost, dating for area suggests 1850's for ~4.5 cm for the Lac de Gras area
61	64.4484	110.3251	0-0.5, 16-16.5	2014	LA	0.0038[$]	continuous permafrost, dating for area suggests 1850's for ~3.5 cm for the Lac de Gras area
62	64.61333	110.84972	0.25-0.5, 16-16.5	1997	LA	low	continuous permafrost, dating for area suggests 1850's for ~4.5 cm
63	64.50977	110.2424	0.25-0.5, 16-16.5	2014	LA	0.0138[$]	continuous permafrost, dating for area suggests 1850's for ~6 cm for the Lac de Gras area
65	66.64056	104.92278	0.5-1, 6-7	1996	LA	low	continuous permafrost
67	66.69194	104.94056	0-1, 14-15	1996	LA	low	continuous permafrost
69	66.68972	104.94056	0-1, 13-14	1996	LA	low	continuous permafrost
71	65.495	103.36694	0-0.5, 11-12	1996	LA	low	continuous permafrost
73	65.56722	103.39028	0.5-1, 12-13	1996	LA	low	continuous permafrost
75	64.6025	107.09167	22-23, 0-1	1996	LA	low	continuous permafrost
77	64.6	107.09167	24-25, 0-1	1996	LA	low	continuous permafrost
79	64.60389	107.09694	8.0-9, 0-1	1996	LA	low	continuous permafrost
81	64.60833	107.10917	15-16, 0-1	1996	LA	low	continuous permafrost
83	64.60417	107.43972	12.0-13, 0-0.5	1996	LA	low	continuous permafrost
85	64.16083	107.83	0-1, 7.5-8	1996	LA	low	continuous permafrost
87	64.15	107.81667	0-0.5, 14-15	1996	LA	low	continuous permafrost, dating for area suggests 1850's for ~5 cm
89	64.15	107.81667	0-1, 10-10.5	1996	LA	low	continuous permafrost
90	78.62663	74.7233	0-1.25, 17-17.5	2006	HA	very low	continuous permafrost
91	78.62349	74.6804	0-0.25, 14.5-15	2006	HA	very low	continuous permafrost
92	78.62623	74.7067	0-0.25, 12-12.5	2006	HA	very low	continuous permafrost
93	78.62048	74.707	0-0.25, 15-15.5	2006	HA	very low	continuous permafrost

* Ecozones: High Arctic (HA), Low Arctic (LA), Mackenzie Delta uplands (MDU), Taiga shield (TS1 and TS2), and Taiga plains.
** Core depth is referenced in the text and figures as top (t) for intervals sampled close to the surface, and bottom (b) for intervals sampled below the surface interval.
[$]Sedimentation rates and dating estimated from ^{210}Pb and ^{137}Cs dating models (Constant Flux Constant Sedimentation Rate models). Since not all cores were dated, other sedimentation rates were estimated based on modelling data from lakes in the same region, and from literature.
[x] Samples were removed from analysis due to absence of signal above the noise threshold.

Table S3.2: Boreal plants under study.

#	Species	Family	Part	Accession*
rg	*Rhododendron groenlandicum*	Ericaceae	Needle	MoB-91025
rt	*Rhododendron tomentosum*	Ericaceae	Needle	MoB-15000
ka	*Kalmia angustifolia*	Ericaceae	Inner bark	MoB-62663
ll	*Larix laricina*	Pinaceae	Inner bark	MoB-62664
ab	*Abies balsamea*	Pinaceae	Cone	MoB-62665
pm	*Picea mariana*	Pinaceae	Cone	MoB-62666
pba	*Pinus banksiana*	Pinaceae	Cone	MoB-62667
pb	*Pinus balsamifera*	Pinaceae	Cone	MoB-62668
pg	*Picea glauca*	Pinaceae	Needle	MoB-62669
ai	*Alnus incana*	Betulaceae	Inner bark	MoB-62663
sp	*Salix planifolia*	Sailaceae	Inner bark	MoB-51601
rg	*Rhododendron groenlandicum*	Ericaceae	Needle	MoB-91022

* Voucher specimens are deposited at the Montreal Botanical Garde

Table S3.3: LCMS and GCMS data of biomarkers whose molecular ion $[M+H]^+$ and mass fragments were identified in the sediment and plant extracts. Entries in bold match standards.

#	Biomarker	R.t.*	$[M+H]^{+\,a}$	Formula	Accurate mass fragments
1	Annotated unknown	3.86	310.3112	$C_{20}H_{39}NO$	275.2782, 268.2995, 248.1278, 174.6332
2	Annotated unknown	4.27	366.3736	$C_{24}H_{48}NO$	349.3462, 331.3380, 202.6650
3	Annotated unknown	3.94	336.3267	$C_{22}H_{41}NO$	319.3021, 301.3005, 187.6421
4	Annotated unknown	4.23	340.3576	$C_{22}H_{46}NO$	102.0914, 88.0764
5	Annotated unknown	3.47	376.3192	$C_{24}H_{42}NO_2$	374.2711, 109.0993, 95.0863, 81.0694
6	Annotated unknown	4.02	637.3046	$C_{29}H_{48}O_{15}$	581.2437, 525.1811, 487.0965, 469.1023, 393.0859, 377.0240, 312.3242, 147.1174
7	Annotated unknown	3.58	352.3215	$C_{22}H_{42}NO_2$	355.2997, 195.6390
8	Annotated unknown	3.69	376.3189	$C_{24}H_{42}NO_2$	156.8887
9	3-Phorbinepropanoic acid	4.40	621.3111	$C_{37}H_{40}N_4O_5$	565.2472 -C_4H_8, 509.1861 -C_8H_{16}, 415.1532 -$C_{13}H_{20}NO$, 335.1791 $C_{16}H_{18}N_2O_3$, 297.1257 -$C_{17}H_{26}NO_5$, 279.1166 $C_{20}H_{26}N_2O_3$, 223.0530 -$C_{24}H_{34}N_2O_3$, 215.1443 -$C_{24}H_{24}NO_5$, 205.0407 $C_{24}H_{36}N_2O_4$, 191.1452 -$C_{26}H_{24}NO_5$, 119.0867 $C_{28}H_{30}N_4O_5$, 97.0983 -$C_{30}H_{28}N_4O_5$, 83.0488 -$C_{32}H_{34}N_4O_4$
10	Purpurin 18 methyl ester	4.23	579.2601	$C_{34}H_{34}N_4O_5$	563.2664 -O, 547.2706 -O_2, 538.2332 -C_2HN, 535.2720 -CO_2, 532.2468 -CHO_2, 519.2736 -CO_3, 503.2445 -$C_2H_2O_3$, 487.2473 -C_5NO, 476.2196 -C_7H_3N, 473.1996 -$C_4H_8O_3$
11	Chlorophyllone a	3.90	533.2548	$C_{35}H_{32}N_4O_3$	517.2584 -O, 515.2432 -H_2O, 505.2567 -CO, 503.2454 -CH_2O, 500.2213 -CH_5O, 487.2591 -NO_2, 109.1022 -$C_{25}H_{20}N_4O_3$
12	Pyropheophorbide a methyl ester	**4.17**	**549.2856**	**$C_{34}H_{36}N_4O_3$**	**521.2913 -CO, 517.2653 -CH_4O, 461.2334 -$C_4H_8O_2$, 447.2190 -$C_5H_{10}O_2$, 435.2546 -$C_5H_6O_3$, 433.2072 -$C_6H_{12}O_2$, 420.2308 -$C_6H_9O_3$**
13	Pheophorbide a methyl ester	4.16	607.2918	$C_{36}H_{38}N_4O_5$	563.2665 -C_2H_4O, 549.2875 -$C_2H_2O_2$, 547.2715 -$C_2H_4O_2$, 535.2722 -$C_3H_4O_2$, 521.2926 -$C_3H_2O_3$, 503.2449 -$C_4H_8O_3$, 476.2226 -$C_6H_{11}O_3$, 461.2342 -$C_9H_8NO_2$, 447.2215 -$C_7H_{12}O_4$, 435.2551 -$C_7H_8O_5$, 421.2391 -$C_8H_{10}O_5$, 301.1125 -$C_{19}H_{22}N_4$, 173.1329 -$C_{23}H_{22}N_4O_5$, 149.1282 -$C_{25}H_{22}N_4O_5$, 147.1175 -$C_{25}H_{24}N_4O_5$, 145.1033 -$C_{25}H_{26}N_4O_5$, 135.1147 -$C_{26}H_{24}N_4O_5$, 133.1035 -$C_{26}H_{26}N_4O_5$, 125.0986 -$C_{27}H_{28}N_4O_5$, 123.1158 -$C_{27}H_{24}N_4O_5$, 121.1021 -$C_{27}H_{26}N_4O_5$, 119.0862 -$C_{27}H_{28}N_4O_5$, 114.0913 -$C_{30}H_{27}N_3O_4$, 109.1016 -$C_{28}H_{26}N_4O_5$, 105.0703 -

					$C_{28}H_{30}N_4O_5$, 97.1015 -$C_{29}H_{26}N_4O_5$, 95.0885 -$C_{29}H_{28}N_4O_5$, 93.0703 -$C_{29}H_{30}N_4O_5$, 83.0862 -$C_{30}H_{28}N_4O_5$, 81.0703-$C_{30}H_{30}N_4O_5$
14	Hydroxypheophorbide a methyl ester	4.05	623.2869	$C_{36}H_{38}N_4O_6$	605.2755 -H_2O, 573.2501 -CH_6O_2, 563.2656 -$C_2H_4O_2$, 545.2556 -$C_2H_6O_3$, 149.1315 -$C_{25}H_{22}N_4O_6$, 135.1164 -$C_{26}H_{24}N_4O_6$, 121.1007 -$C_{27}H_{26}N_4O_6$, 114.0909 -$C_{30}H_{27}N_3O_5$
15	Protoporphyrin IX analogue	4.18	**563.2653**	**$C_{34}H_{34}N_4O_4$**	547.2703 -O, 535.2692 -CO, **517.2627 -CH_2O_2**, **504.2514 -$C_2H_3O_2$**, **503.2456 -$C_2H_4O_2$**, **475.2219 -$C_3H_6NO_2$**, 145.1038 -$C_{23}H_{22}N_4O_4$, 135.1160 -$C_{25}H_{20}N_4O_4$, 133.1025 -$C_{24}H_{22}N_4O_4$, 121.1010 -$C_{25}H_{22}N_4O_4$, 97.1021 -$C_{27}H_{22}N_4O_4$, 95.0860 -$C_{27}H_{24}N_4O_4$
16	Bacteriochlorophyll b -Mg	4.96	887.5666	$C_{55}H_{74}N_4O_6$	829.5624 -$C_2H_2O_2$, 725.4860 -$C_9H_{10}N_2O$, 715.5533 -$C_{10}H_4O_3$, 671.4058 -$C_{14}H_{20}N_2$, 124.0875 -$C_{47}H_{63}N_4O_5$
17	Squalene-2,3-epoxide	4.63	427.3933	$C_{30}H_{50}O$	**411.3616 -CH_4**, 205.1967 -$C_{15}H_{26}O$, **191.1811 [c] -$C_{16}H_{28}O$**, 163.1457 -$C_{18}H_{32}O$, 149.1308 -$C_{19}H_{34}O$, 135.1164 -$C_{20}H_{36}O$, 123.1170 -$C_{21}H_{36}O$, 109.1011 -$C_{22}H_{38}O$
18	Dehydrosqualene	4.75	409.3846	$C_{30}H_{48}$	241.1981 -$C_{15}H_{24}$, 199.1443 -$C_{15}H_{30}$, 189.1619 -$C_{16}H_{28}$, **161.1339 -$C_{18}H_{32}$, 147.1160 -$C_{19}H_{34}$, 109.1011 -$C_{22}H_{36}$**
19	Hydroperoxysqualene	4.35	443.3861	$C_{30}H_{50}O_2$	215.1770 -$C_{14}H_{28}O_2$, **191.1804 -$C_{16}H_{28}O_2$, 161.1322 -$C_{18}H_{34}O_2$, 149.1335 -$C_{19}H_{34}O2$, 147.1186 -$C_{19}H_{36}O_2$, 137.1340 -$C_{20}H_{34}O_2$, 135.1166 -$C_{20}H_{36}O_2$, 121.1001 -$C_{21}H_{38}O_2$, 109.1021 -$C_{22}H_{38}O_2$**
20	Pyropheophorbide a	**3.88**	**535.2695**	**$C_{33}H_{34}N_4O_3$**	**533.2559** -H_2, 521.2566 -CH_2, **507.2770 -CO**, **505.2561 -CH_2O**, 503.2464 -CH_4O, 500.2216 -CH_7O, **487.2488 -CH_4O_2**, 477.2250 -C_3H_6O, 476.2231 -C_3H_7O, **461.2365 -$C_3H_6O_2$**, **447.2204 -$C_4H_8O_2$**, 435.2551 -$C_4H_4O_3$, 237.1382 -$C_{17}H_{18}N_2O_3$, 149.1301 -$C_{22}H_{18}N_4O_3$, 147.1196 -$C_{22}H_{20}N_4O_3$, 145.0979 -$C_{22}H_{22}N_4O_3$, 123.1172 -$C_{24}H_{22}N_4O_3$, 121.1002 -$C_{24}H_{22}N_4O_3$, 119.0863 -$C_{24}H_{24}N_4O_3$, 109.1011-$C_{25}H_{22}N_4O_3$, 107.0855 -$C_{25}H_{24}N_4O_3$, 105.0700 -$C_{25}H_{26}N_4O_3$, 100.0767 -$C_{28}H_{25}N_3O_2$, 97.1013 -$C_{26}H_{22}N_4O_3$, 97.0646 -$C_{27}H_{26}N_4O_2$, 95.0847 -$C_{26}H_{24}N_4O_3$, 97.0646 -$C_{27}H_{26}N_4O_2$, 95.0847 -$C_{26}H_{24}N_4O_3$, 83.0857 -$C_{27}H_{24}N_4O_2$, 81.0699 -$C_{27}H_{26}N_4O_3$, 69.0713 -$C_{28}H_{26}N_4O_3$
21	Pheophorbide a	**3.87**	**593.2759**	**$C_{35}H_{36}N_4O_5$**	**565.2780 -CO**, **549.2514 -C_2H_4O**, **549.2514 -C_2H_4O**, 539.2455 -C_3H_4N, **533.2565 -$C_2H_4O_2$**, **515.2458 -$C_2H_6O_3$**, **505.2605 -$C_3H_4O_3$**, **461.2269 -C_8H_6NO**, **447.2157 -$C_6H_{10}O_4$**, **431.1892 -$C_7H_{14}O_4$**, **91.0537 -$C_{28}H_{30}N_4O_5$**
22	32,35-Anhydrobacteriohopaneterol	4.56	529.4557	$C_{35}H_{60}O_3$	451.3590 -$C_4H_{12}O$, 429.3685 -$C_6H_{10}O$, 409.3504 -$C_6H_{14}O$, 407.3312 -$C_6H_{16}O_2$, 397.3559 -$C_7H_{14}O_2$, 395.3652 -$C_6H_{12}O_3$, 385.3465 -$C_8H_{14}O_2$, 124.0882 -$C_{27}H_{47}O_2$
23	Bacteriohopane-32,33,34,35-tetrol	4.51	547.4711	$C_{35}H_{62}O_4$	311.2509 -$C_{16}H_{28}O$, 293.1778 -$C_{18}H_{38}$, 184.0735 -$C_{28}H_{50}O_2$, 163.1479 -$C_{23}H_{44}O_4$

24	35-aminobacteriohopane-31,32,33,34-tetrol	3.30	562.4833	$C_{35}H_{63}NO_4$	233.1523 -$C_{20}H_{43}NO_2$, 133.1003 -$C_{25}H_{51}NO_4$, 81.0696 -$C_{29}H_{55}NO_4$
25	35-Aminobacteriohopane-32,33,34-triol	3.41	546.4888	$C_{35}H_{63}NO_3$	529.4861 -HO, 492.4371 -H_8NO_2, 191.1833 -$C_{21}H_{41}NO_3$, 163.1508 -$C_{23}H_{45}NO_3$, 161.1326 -$C_{23}H_{47}NO_3$, 149.1336 -$C_{24}H_{47}NO_3$, 137.1344 -$C_{25}H_{47}NO_3$, 133.0996 -$C_{25}H_{51}NO_3$, 121.0983 -$C_{26}H_{51}NO_3$, 109.1010 -$C_{27}H_{51}NO_3$, 107.0859 -$C_{27}H_{53}NO_3$, 86.0955 -$C_{30}H_{52}NO_3$
26	N-tryptophanyl-35-aminobacteriohopane-32,33,34-triol	4.42	732.5722	$C_{46}H_{73}N_3O_4$	645.5042 -C_4H_9NO, 566.4482 -$C_{10}H_{16}NO$, 562.4370 -$C_{10}H1_8O_2$, 549.4734 -$C_{12}H_{11}N_2$, 335.2607 -$C_{25}H_{39}N_3O$, 317.1825 -$C_{29}H_{51}O$, 285.2410 -$C_{29}H_{41}N_3O$, 184.0733 -$C_{34}H_{64}N_2O_3$, 123.1180 -$C_{37}H_{59}N_2O_4$, 109.1028 -$C_{38}H_{61}N_3O_4$
27	Adenosylhopane	4.22	662.4999	$C_{40}H_{63}N_5O_3$	555.3619 -$C_6H_{19}O$, 413.2661 -$C_{16}H_{29}N_2$, 177.0529 $C_{32}H_{57}N_2O$, 169.0956 -$C_{34}H_{53}O_2$
28	35-aminobacteriohopane-30,31,32,33,34-pentol	3.19	578.4766	$C_{35}H_{63}NO_5$	560.4658 -H_2O, 543.4417 -H_5NO, 524.4485 -H_6O_3, 471.3521 -$C_5H_{17}NO$, 175.1459 -$C_{22}H_{45}NO_5$, 161.1352 -$C_{23}H_{47}NO_5$, 135.1132 -$C_{25}H_{49}NO_5$
29	Bacteriohopanetetrol cyclitol	3.32	708.5440	$C_{41}H_{73}NO_8$	550.3350 -$C_{11}H_{26}$, 355.2150 -$C_{22}H_{43}NO_2$, 345.2277 -$C_{23}H_{44}NO_2$, 331.2163 -$C_{24}H_{43}NO_2$, 162.0780 -$C_{35}H_{62}O_4$

* Retention time in min.

[a] Positive ionization MS^e acquisition in resolution mode: Low energy cone voltage to produce $[M+H]^+$ = 6.0 eV and high energy cone voltage ramp to produce accurate mass fragments = 10.0-50.0 eV

Table S3.4: LCMS data of the plant metabolites standards including molecular ion $[M+H]^+$ and mass fragments.

#	Compound	Source	$[M+H]^{+}$ [a]	Formula	Accurate mass fragments*		
1	(-)-alpha-Santonin	Sigma	247.1354	$C_{15}H_{18}O_3$	173.0970	158.0740	145.1033
2	(-)-Arctigenin	Sigma	373.1697	$C_{21}H_{24}O_6$	355.1567	177.0927	137.0624
3	(-)-Epicatechin	Extrasynthese	291.0949	$C_{15}H_{14}O_6$	139.0421	123.0474	91.0550
4	(-)-Epicatechin gallate	Extrasynthese	443.1113	$C_{22}H_{18}O_{10}$	273.0829	139.0421	123.0474
5	(-)-Epigallocatechin	Chromadex	307.0829	$C_{15}H_{14}O_7$	/	/	/
6	(-)-Epinephrine	Sigma	184.0943	$C_9H_{13}NO_3$	166.0855	135.0422	107.0488
7	(-)-Pinocembrin	Sigma	257.0855	$C_{15}H_{12}O_4$	153.0205	131.0530	103.0568
8	(+)-13-Hydroxylupanine	Toronto Research Chemicals	265.1928	$C_{15}H_{24}N_2O_2$	152.1065	112.0761	84.0810
9	(+)-Butaclamol	Sigma	362.2521	$C_{25}H_{31}NO$	344.2380	234.1273	202.0786
10	(+)-Catechin	Chromadex	291.0810	$C_{15}H_{14}O_6$	139.0373	123.0428	91.0550
11	11,13-Dihydroparthenolide	Sigma	251.1685	$C_{15}H_{22}O_3$	245.1217	187.1482	159.1176
12	17α(H),21β(H)- 22S-homohopane	Sigma	427.4304	$C_{31}H_{54}$	191.1797	/	/
13	17α(H),21β(H)-22R-homohopane	Sigma	427.4304	$C_{31}H_{54}$	191.1792	/	/
14	17α(H),21β(H)-30-norhopane	Sigma	399.3991	$C_{29}H_{50}$	191.1794	/	/
15	17α(H),21β(H)-hopane	Sigma	413.4147	$C_{30}H_{52}$	191.1797	/	/
16	17α(H)-22,29,30-trisnorhopane	Sigma	371.3678	$C_{27}H_{46}$	191.1797	/	/
17	1-Methyluric acid	Sigma	183.0526	$C_6H_6N_4O_3$	155.0598	126.0293	96.0165
18	1-Methylxanthine	Sigma	167.0588	$C_6H_6N_4O_2$	149.0491	110.0372	82.0398
19	2,4-dihydroxybenzoic acid	Sigma	155.0335	$C_7H_6O_4$	137.0217	109.0278	81.0341
20	20-Hydroxyecdysone	Extrasynthese	481.3203	$C_{27}H_{44}O_7$	445.2977	371.2205	301.1821
21	2-Desmethylcolchicine glucoside	Extrasynthese	549.2228	$C_{27}H_{34}NO_{11}$	408.1466	386.1617	181.0664
22	3,3',4'-Hydroxyflavone	Sigma	271.0568	$C_{15}H_{10}O_5$	197.0564	141.0676	121.0255
23	3,4',7-trihydroxyflavone	Extrasynthese	271.0568	$C_{15}H_{10}O_5$	197.0564	137.0217	121.0277
24	3-Amino-4-hydroxybenzoic acid	Sigma	154.0460	$C_7H_7NO_3$	136.0361	108.0424	79.0196
25	3-Hydroxybenzoic acid	Sigma	139.0373	$C_7H_6O_3$	/	/	/
26	3-Oxograndolide	Extrasynthese	265.1475	$C_{15}H_{20}O_4$	173.0992	145.1026	115.0550
27	4-Aminosalicylic acid	Sigma	154.0460	$C_7H_7NO_3$	119.0092	91.0159	80.0475

28	4-Bromomethyl-7-methoxycoumarin	Sigma	268.9845	$C_{11}H_9BrO_3$	190.0669	162.0706	147.0481
29	4-Hydroxybenzoic acid	Chromadex	139.0421	$C_7H_6O_3$	121.0300	93.0348	77.0394
30	4-Hydroxyquinoline	Chromadex	146.0624	C_9H_7NO	128.0483	118.0645	104.0519
31	5,7-Dimethyoxycoumarin	Sigma	207.0637	$C_{11}H_{10}O_4$	192.0393	164.0444	121.0615
32	6-Benzylaminopurine	Sigma	226.1149	$C_{12}H_{11}N_5$	184.0900	148.0635	91.0550
33	6-Hydroxymelatonin	Extrasynthese	249.1241	$C_{13}H_{16}N_2O_3$	217.0701	158.0594	130.0648
34	6-Methoxyharmalan	Extrasynthese	215.1259	$C_{13}H_{14}N_2O$	200.0986	172.1029	131.0754
35	8-alpha-Hydroxyparthenolide	Synthetic	265.1475	$C_{15}H_{20}O_4$	201.1263	183.1202	128.0656
36	Acetovanillone	Extrasynthese	167.0710	$C_9H_{10}O_3$	/	/	/
37	Acetyl-L-tryptophan	Extrasynthese	247.1021	$C_{13}H_{14}N_2O_3$	241.0634	142.0620	115.0541
38	Aconine	Toronto Research Chemicals	500.2885	$C_{25}H_{41}NO_9$	450.2466	233.0947	165.0696
39	Acroptilin	Sigma	399.1114	$C_{19}H_{23}ClO_7$	261.1191	243.1034	153.0712
40	Acycloguanosine	Sigma	226.0903	$C_8H_{11}N_5O_3$	174.0367	152.0535	135.0285
41	Adenine	Sigma	136.0629	$C_5H_5N_5$	119.0363	109.0516	92.0246
43	Allicin	Sigma	163.0251	$C_6H_{10}OS_2$	/	/	/
44	Alliin	Sigma	178.0558	$C_6H_{11}NO_3S$	161.0272	120.0141	74.0255
45	Aloe Emodin	Sigma	271.0536	$C_{15}H_{10}O_5$	253.0439	225.0491	121.0277
46	alpha-Chaconine	Sigma	852.4910	$C_{45}H_{73}NO_{14}$	560.3836	398.3353	98.0945
47	alpha-Ketoglutaric acid	Sigma	147.0293	$C_5H_6O_5$	/	/	/
48	alpha-Methyl-D-mannopyranoside	Sigma	217.0610	$C_7H_{14}O_6Na$	162.9668	98.9738	/
49	Amentoflavone	Sigma	539.1022	$C_{30}H_{18}O_{10}$	403.0546	377.0714	153.0211
50	Angelicin	Sigma	187.0449	$C_{11}H_6O_3$	131.0538	115.0560	77.0412
51	Anhalinine	1717 CheMall	224.1305	$C_{12}H_{17}NO_3$	195.1027	161.0826	146.0599
52	Anhalonine	1717 CheMall	222.1152	$C_{12}H_{15}NO_3$	205.0873	147.0811	115.0541
53	Anthrarobin	Sigma	227.0653	$C_{14}H_{10}O_3$	209.0540	181.0597	152.0578
54	Apigenin	Sigma	271.0635	$C_{15}H_{10}O_5$	153.0211	119.0516	91.0589
55	Apigetrin	Extrasynthese	433.1184	$C_{21}H_{20}O_{10}$	271.0635	153.0211	91.0550
56	Apoatropine	Boc Sciences	272.1694	$C_{17}H_{21}NO_2$	226.9500	124.1129	93.0704
57	Arecoline	Sigma	156.1017	$C_8H_{13}NO_2$	135.0428	106.9982	84.9599
58	Ascorbic acid	Sigma	177.0437	$C_6H_8O_6$	141.9506	114.9506	95.0120

59	Asperilin	Sigma	249.1512	$C_{15}H_{20}O_3$	231.1371	185.1315	128.0632
60	Baicalein	Sigma	271.0534	$C_{15}H_{10}O_5$	169.0111	123.0065	95.0120
61	Baicalin	Extrasynthese	447.1066	$C_{21}H_{18}O_{11}$	271.0703	253.0569	123.0110
62	Belladonnine	Boc Sciences	543.3277	$C_{34}H_{42}N_2O_4$	420.2209	272.1694	124.1129
63	Benzamide	Sigma	122.0596	C_7H_7NO	105.0336	77.0376	/
64	Berberine	Sigma	337.1264	$C_{20}H_{18}NO_4$	263.0567	191.0716	177.0584
65	Bergapten	Sigma	217.0442	$C_{12}H_8O_4$	202.0218	174.0275	146.0320
66	beta-D-Allose	Sigma	203.0460	$C_6H_{12}O_6$	158.9900	140.0633	98.9718
67	beta-D-Glucose 6-phosphate	Sigma	283.0090	$C_6H_{13}O_9PNa$	243.0191	201.0004	127.0333
68	beta-Maltose	Sigma	365.0944	$C_{12}H_{22}O_{11}Na$	281.0596	158.9926	98.9718
69	Britanin	Sigma	367.1805	$C_{19}H_{26}O_7$	285.0842	168.0962	128.0609
70	Burrodin	1717 CheMall	265.1441	$C_{15}H_{20}O_4$	247.1356	183.1174	/
71	Caffeic acid	Sigma	181.0459	$C_9H_8O_4$	163.0364	135.0422	117.0320
72	Caffeoylputrescine	Synthetic	251.1380	$C_{13}H_{18}N_2O_3$	234.1148	135.0428	89.0385
73	Callistephin	Sigma	433.1099	$C_{21}H_{21}O_{10}$	271.0568	140.0675	113.9606
74	Canadine	Boc Sciences	340.1592	$C_{20}H_{21}NO_4$	292.1011	176.0709	149.0591
75	Cephaeline	1717 CheMall	467.2949	$C_{28}H_{38}N_2O_4$	438.2680	246.1504	165.0906
76	Chlorogenic acid	Chromadex	355.1066	$C_{16}H_{18}O_9$	163.0390	135.0422	89.0385
77	Chloroquine	Chromadex	320.1921	$C_{18}H_{26}ClN_3$	247.1021	179.0390	142.1597
78	Chrysin	Sigma	255.0703	$C_{15}H_{10}O_4$	153.0211	129.0367	103.0579
79	Chymostatin	Sigma	608.3218	$C_{31}H_{41}N_7O_6$	465.2216	266.1605	127.0980
80	Cinchonine	Chromadex	295.1848	$C_{19}H_{22}N_2O$	277.1706	158.0594	130.0648
81	Cinerenin	Synthetic	321.1360	$C_{17}H_{20}O_6$	226.9543	141.0717	128.0632
82	Cinnamoyltropine	Synthetic	272.1694	$C_{17}H_{21}NO_2$	124.1106	93.0685	77.0376
83	Colchicine	Sigma	400.1703	$C_{22}H_{25}NO_6$	358.1568	165.0669	141.0682
84	Confertiflorin	Sigma	307.1556	$C_{17}H_{22}O_5$	269.1160	247.1356	201.1292
85	Coronopilin	1717 CheMall	265.1475	$C_{15}H_{20}O_4$	247.1356	183.1174	115.0572
86	Coumarin	Sigma	147.0407	$C_9H_6O_2$	103.0538	91.0530	77.0394
87	Coumestrol	Sigma	269.0483	$C_{15}H_8O_5$	213.0573	157.0692	115.0560
88	Cyanidin	Extrasynthese	287.0538	$C_{15}H_{11}O_6$	213.0513	137.0217	109.0278
89	Cynarin	Extrasynthese	517.1553	$C_{25}H_{24}O_{12}$	499.1392	359.0836	163.0443

90	Cynaroside	Extrasynthese	449.1088	$C_{21}H_{20}O_{11}$	287.0538	213.0483	153.0161
91	Cytisine	Sigma	191.1197	$C_{11}H_{14}N_2O$	162.0950	148.0760	133.0517
92	D-(-)-Fructose	Sigma	203.0430	$C_6H_{12}O_6Na$	170.9820	158.9900	98.9698
93	D-(-)-Lyxose	Sigma	173.0361	$C_5H_{10}O_5Na$	98.9718	/	/
94	D-(-)-Ribose	Sigma	173.0361	$C_5H_{10}O_5Na$	158.9936	110.9734	98.9718
95	D-(+)-Fucose	Sigma	187.0518	$C_6H_{12}O_5$	158.9926	110.9713	98.9738
96	D-(+)-Galactose	Sigma	203.0489	$C_6H_{12}O_6Na$	158.9900	110.9734	98.9698
97	D-(+)-Glucosamine	Sigma	202.0640	$C_6H_{13}NO_5Na$	/	/	/
98	D-(+)-Glucose	Sigma	203.0401	$C_6H_{12}O_6Na$	174.9874	158.9874	98.9698
99	D-(+)-Mannose	Sigma	203.0460	$C_6H_{12}O_6Na$	174.9643	158.9874	98.9718
100	D-(+)-Trehalose	Sigma	365.0944	$C_{12}H_{22}O_{11}Na$	281.0630	203.0430	98.9718
101	D-(+)-Xylose	Sigma	173.0361	$C_5H_{10}O_5Na$	110.9713	98.9738	/
102	Daidzein	Sigma	255.0605	$C_{15}H_{10}O_4$	199.0733	181.0624	137.0217
103	Daphnoretin	Sigma	353.0648	$C_{19}H_{12}O_7$	338.0409	179.0323	146.0345
104	Deacetyl neotenulin	Synthetic	265.1441	$C_{15}H_{20}O_4$	247.1356	201.1292	135.0819
105	Delphinidin	Chromadex	303.0423	$C_{15}H_{11}O_7$	177.0164	153.0161	149.0209
106	DIBOA	Extrasynthese	182.0429	$C_8H_7NO_4$	180.0046	133.9719	80.0493
107	Dihydroreynosin	Sigma	251.1653	$C_{15}H_{22}O_3$	233.1491	159.1176	105.0700
108	DIM2BOA	1717 CheMall	242.0626	$C_{10}H_{11}NO_6$	240.0188	179.0582	109.0281
109	DIMBOA	Sigma	212.0531	$C_9H_9NO_5$	177.9586	149.0467	110.0609
110	Diosmetinidin	Extrasynthese	285.0904	$C_{16}H_{13}O_5+$	270.0667	242.0700	213.0662
111	DL-Tryptophan	Sigma	205.0961	$C_{11}H_{12}N_2O_2$	143.0716	118.0645	91.0531
112	DL-Tyrosine	Sigma	182.0788	$C_9H_{11}NO_3$	107.0003	91.0531	77.0376
113	Dopamine	Sigma	154.0892	$C_8H_{11}NO_2$	137.0624	119.0516	91.0569
114	D-Tyrosine	Sigma	182.0761	$C_9H_{11}NO_3$	174.0151	119.0498	91.0531
115	Ellagic acid	Extrasynthese	303.0138	$C_{14}H_6O_8$	257.0096	201.0162	173.0242
116	Emetine	Sigma	481.3081	$C_{29}H_{40}N_2O_4$	452.2819	246.1472	165.0906
117	Enhydrin	Sigma	465.1805	$C_{23}H_{28}O_{10}$	371.1140	289.1081	229.0874
118	Epigallocatechin gallate	Extrasynthese	459.1079	$C_{22}H_{18}O_{11}$	289.0795	153.0262	139.0445
119	Eriodictyol	Sigma	289.0691	$C_{15}H_{12}O_6$	163.0390	153.0186	117.0298
120	Esculetin	Sigma	179.0323	$C_9H_6O_4$	133.0276	123.0451	105.0334

121	Ethylvanillin	Sigma	167.0710	$C_9H_{10}O_3$	111.0465	93.0348	65.0404
122	Ferulic acid	Extrasynthese	195.0612	$C_{10}H_{10}O_4$	177.0491	145.0244	117.0298
123	Flumazenil	Sigma	304.1189	$C_{15}H_{14}FN_3O_3$	258.0732	217.0460	134.0430
124	Folic acid	Chromadex	442.1486	$C_{19}H_{19}N_7O_6$	295.0911	176.0555	120.0443
125	Formononetin	Sigma	269.0785	$C_{16}H_{12}O_4$	253.0471	237.0508	197.0564
126	Furaneol	Sigma	129.0530	$C_6H_8O_3$	100.9654	90.9729	70.9578
127	Fusaric acid	Sigma	180.1007	$C_{10}H_{13}NO_2$	152.1065	134.0952	92.0482
128	Galegine	1717 CheMall	158.1178	$C_6H_{13}N_3$	121.9804	106.9516	90.9769
129	Gallic acid	Sigma	171.0265	$C_7H_6O_5$	153.0161	125.0219	107.0107
130	gamma-Aminobutyric acid	Sigma	104.0684	$C_4H_9NO_2$	/	/	/
131	Geigerinin	1717 CheMall	267.1615	$C_{15}H_{22}O4$	231.1402	185.1315	128.0656
132	Gelsemine	Chromadex	323.1774	$C_{20}H_{22}N_2O_2$	236.1082	195.0655	149.0599
133	Genistein	Sigma	271.0703	$C_{15}H_{10}O_5$	243.0744	153.0237	91.0569
134	Genostrychnine	1717 CheMall	351.1744	$C_{21}H_{22}N_2O_3$	334.1671	220.0757	120.0818
135	Gentisic acid	Chromadex	155.0335	$C_7H_6O_4$	137.0217	118.0841	81.0323
136	Germerine	Sigma	694.4205	$C_{37}H_{59}NO_{11}$	676.4064	558.3428	456.2745
137	Germitrine	Sigma	736.4276	$C_{39}H_{61}NO_{12}$	718.4107	540.3350	162.1289
138	Glaucolide A	Sigma	465.1805	$C_{23}H_{28}O_{10}$	337.1207	319.1176	259.1017
139	Glutathione reduced	Sigma	308.0846	$C_{10}H_{17}N_3O_6S$	233.0521	162.0185	76.0203
140	Glycitein	Sigma	285.0807	$C_{16}H_{12}O_5$	242.0607	197.0619	118.0431
141	Gnoscopine	1717 CheMall	414.1618	$C_{22}H_{23}NO_7$	353.1030	220.0969	165.0696
142	Gramine	Boc Sciences	175.1187	$C_{11}H_{14}N_2$	103.0519	77.0376	/
143	Harmaline	Chromadex	215.1259	$C_{13}H_{14}N_2O$	200.0986	174.0934	131.0754
144	Harmalol	Toronto Research Chemicals	201.1048	$C_{12}H_{12}N_2O$	184.0761	160.0760	115.0541
145	Harmane	Extrasynthese	183.0996	$C_{12}H_{10}N_2$	168.0712	142.0669	115.0563
146	Harmine	Extrasynthese	213.1098	$C_{13}H_{12}N_2O$	198.0823	170.0864	144.0799
147	Harmol	Sigma	199.0921	$C_{12}H_{10}N_2O$	171.0945	131.0520	103.0560
148	HBOA	Sigma	166.0469	$C_8H_7NO_3$	157.9925	120.0436	93.0349
149	Hesperetin	Sigma	303.0957	$C_{16}H_{14}O_6$	177.0600	153.0237	117.0386
150	Hippuryl-L-lysine	Sigma	308.1631	$C_{15}H_{21}N_3O_4$	130.0881	105.0336	84.0810

151	Homatropine	Extrasynthese	276.1660	$C_{16}H_{21}NO_3$	142.1255	124.1129	93.0724
152	Hydroxyvalerenic acid	Sigma	251.1621	$C_{15}H_{22}O_3$	233.1522	187.1482	105.0700
153	Hyperoside	Extrasynthese	465.1064	$C_{21}H_{20}O_{12}$	303.0529	229.0502	185.0427
154	Ideain	Chromadex	449.1044	$C_{21}H_{21}O_{11}$	287.0538	241.0455	109.0278
155	Indole	Sigma	118.0645	C_8H_7N	106.9453	90.9730	84.9599
156	Indole-3-acetic acid	Extrasynthese	176.0709	$C_{10}H_9NO_2$	170.0303	130.0648	103.0560
157	Inositol	Sigma	181.0679	$C_6H_{12}O_6$	169.9626	143.9490	113.9409
158	Isoorientin	Sigma	449.1305	$C_{21}H_{20}O_{11}$	299.0709	283.0753	165.0292
159	Isorhamnetin	Sigma	317.0790	$C_{16}H_{12}O_7$	302.0523	229.0564	153.0237
160	Isotenulin	Sigma	307.1591	$C_{17}H_{22}O_5$	247.1356	201.1263	119.0864
161	Isovitexin	Extrasynthese	433.1141	$C_{21}H_{20}O_{10}$	337.0704	313.0708	283.0615
162	Jervine	Sigma	426.3072	$C_{27}H_{39}NO_3$	313.2224	145.1015	114.0942
163	Kaempferol	Sigma	287.0503	$C_{15}H_{10}O_6$	165.0160	153.0161	121.0255
164	Keracyanin	Chromadex	595.1579	$C_{27}H_{31}O_{15}$+	287.0538	213.0483	137.0217
165	Ketoconazole	Sigma	531.1657	$C_{26}H_{28}Cl_2N_4O_4$	489.1550	255.0105	158.9797
166	Khellin	Sigma	261.0724	$C_{14}H_{12}O_5$	246.0487	231.0281	163.0024
167	Kuromanin	Sigma	449.1044	$C_{21}H_{21}O_{11}$+	287.0538	140.0675	113.9628
168	L-(-)-Methionine	Sigma	150.0581	$C_5H_{11}NO_2S$	143.9546	106.9982	87.0235
169	L-(-)-Tyrosine	Sigma	182.0816	$C_9H_{11}NO_3$	119.0475	91.0550	77.0393
170	L-(+)-Alanine	Sigma	90.0556	$C_3H_7NO_2$	84.9562	59.9117	/
171	L-(+)-Arabinose	Sigma	173.0388	$C_5H_{10}O_5Na$	110.9756	98.9718	/
172	L-(+)-Cysteine	Sigma	122.0257	$C_3H_7NO_2S$	120.0122	89.9857	74.0254
173	L-(+)-Glutamic acid	Sigma	148.0585	$C_5H_9NO_4$	127.8932	104.9623	90.9808
174	L-Aspartic acid	Sigma	134.0454	$C_4H_7NO_4$	117.9844	87.4787	76.5120
175	Leptocladine	Atomax Chemicals	201.1396	$C_{13}H_{16}N_2$	158.0981	143.0741	115.0541
176	L-Glutamine	Sigma	147.0762	$C_5H_{10}N_2O_3$	124.0080	97.9925	84.0434
177	L-Histidine	Sigma	156.0761	$C_6H_9N_3O_2$	148.0162	93.0448	83.0610
178	Linifolin A	Sigma	305.1388	$C_{17}H_{20}O_5$	245.1185	199.1135	135.0819
179	L-Leucine	Sigma	132.1017	$C_6H_{13}NO_2$	106.9982	84.9618	58.4861
180	L-Ornithine	Sigma	133.0942	$C_5H_{12}N_2O_2$	125.0360	95.9764	70.0655

181	L-Phenylalanine	Sigma	166.0865	$C_9H_{11}NO_2$	148.9992	103.0539	77.0376
182	L-Rhamnose	Sigma	187.0490	$C_6H_{12}O_5Na$	174.9616	158.9874	98.9698
183	L-Tryptophan	Sigma	205.0961	$C_{11}H_{12}N_2O_2$	170.0597	143.0716	118.0645
184	Luteolin	Extrasynthese	287.0642	$C_{15}H_{10}O_6$	241.0582	153.0262	135.0493
185	Malvidin	Extrasynthese	331.0833	$C_{17}H_{15}O_7$+	177.0219	153.0566	121.0300
186	Malvidin-3,5-diglucoside	Chromadex	655.1745	$C_{29}H_{35}O_{17}$	353.0610	331.0759	533.1200
187	Malvidin-3-O-galactoside	Extrasynthese	493.1335	$C_{23}H_{25}O_{12}$	331.0796	315.0481	287.0538
188	MBOA	Extrasynthese	166.0495	$C_8H_7NO_3$	161.9830	110.0587	95.0480
189	Melampodin A	Sigma	421.1515	$C_{21}H_{24}O_9$	403.1433	152.0647	/
190	Melatonin	Sigma	233.1353	$C_{13}H_{16}N_2O_2$	174.0934	159.0700	130.0695
191	meta-Coumaric acid	Extrasynthese	165.0607	$C_9H_8O_3$	147.0481	119.0516	91.0569
192	Methoxsalen	Sigma	217.0503	$C_{12}H_8O_4$	202.0276	174.0302	161.0610
193	Methyl salicylate	Sigma	153.0642	$C_8H_8O_3$	121.0345	93.0388	65.0420
194	Microcystin LA	Sigma	910.4775	$C_{46}H_{67}N_7O_{12}$	347.1972	/	/
195	Microcystin LR	Sigma	995.5336	$C_{49}H_{74}N_{10}O_{12}$	772.4109	163.1114	91.0516
196	Mitraphylline	Extrasynthese	369.1740	$C_{21}H_{24}N_2O_4$	337.1490	160.0735	117.0567
197	Morin	Sigma	303.0529	$C_{15}H_{10}O_7$	257.0458	165.0186	153.0186
198	Myricetin-3-O-galactoside	Extrasynthese	481.1010	$C_{21}H_{20}O_{13}$	319.0434	165.0160	153.0161
199	Myricitrin	Extrasynthese	465.1241	$C_{21}H_{20}O_{12}$	319.0544	245.0542	153.0237
200	Naringenin	Extrasynthese	273.0796	$C_{15}H_{12}O_5$	153.0211	119.0516	91.0550
201	N-Decanoyl-L-homoserine lactone	Sigma	256.1888	$C_{14}H_{25}NO_3$	226.9500	158.9642	90.9789
202	Neogermbudine	1717 CheMall	710.4160	$C_{37}H_{59}NO_{12}$	692.4042	558.3477	456.2789
203	Neogermitrine	Sigma	678.3939	$C_{36}H_{55}NO_{11}$	600.3598	540.3350	438.2680
204	N-Hexanoyl-L-homoserine lactone	Sigma	200.1276	$C_{10}H_{17}NO_3$	194.0831	154.0544	110.9734
205	Nicotiflorin	Extrasynthese	595.1678	$C_{27}H_{30}O_{15}$	525.2880	287.0538	153.0211
206	Nicotinic acid	Sigma	124.0372	$C_6H_5NO_2$	106.0282	80.0493	78.0327
207	Nodularin	Sigma	825.4392	$C_{41}H_{60}N_8O_{10}$	/	/	/
208	Nordihydroguaiaretic acid	Sigma	303.1635	$C_{18}H_{22}O_4$	193.1303	123.0474	77.0412
209	Noreleagnine	Chromadex	173.1088	$C_{11}H_{12}N_2$	143.0716	128.0483	115.0541
210	Norharman	Chromadex	169.0813	$C_{11}H_8N_2$	142.0669	115.0563	89.0404
211	Norlobelanine	Sigma	322.1808	$C_{21}H_{23}NO_2$	202.1222	105.0336	82.0657

212	Oenin	Extrasynthese	493.1517	$C_{23}H_{25}O_{12}$	331.0945	140.0724	113.9693
213	Olivetol	Sigma	181.1203	$C_{11}H_{16}O_2$	158.9633	111.0443	90.9768
214	ortho-Coumaric acid	Extrasynthese	165.0528	$C_9H_8O_3$	147.0432	103.0538	91.0530
215	para-Aminobenzoic acid	Sigma	138.0541	$C_7H_7NO_2$	120.0433	94.0645	77.0376
216	para-Anisic acid	Sigma	153.0591	$C_8H_8O_3$	135.0470	109.0684	77.0412
217	para-Coumaric acid	Extrasynthese	165.0607	$C_9H_8O_3$	147.0481	119.0516	91.0550
218	para-Coumaroylputrescine	Synthetic	235.1434	$C_{13}H_{18}N_2O_2$	226.9500	119.0498	91.0550
219	Parthenin	1717 CheMall	263.1311	$C_{15}H_{18}O_4$	227.1085	199.1135	166.0790
220	Parthenolide	Sigma	249.1512	$C_{15}H_{20}O_3$	185.1371	128.0656	105.0700
221	Pelletierine	Sigma	142.1206	$C_8H_{15}NO$	123.4364	116.0025	98.0959
222	Peonidin-3-O-glucoside	Extrasynthese	463.1152	$C_{22}H_{23}O_{11}$	301.0575	286.0423	258.0480
223	Phenethyl glucosinolate	Chromadex	424.0740	$C_{15}H_{21}NO_9S_2$	320.9619	142.9345	91.0531
224	Phenethyl isothiocyanate	Sigma	164.0541	C_9H_9NS	140.0665	105.0699	88.0200
227	Picrotoxinin	Extrasynthese	293.0992	$C_{15}H_{16}O_6$	287.0530	226.9500	115.0519
228	Pilocarpine	Sigma	209.1326	$C_{11}H_{16}N_2O_2$	163.1262	121.0754	95.0620
229	Piperine	Sigma	286.1484	$C_{17}H_{19}NO_3$	201.0554	143.0496	115.0541
230	Piperlongumine	Sigma	318.1366	$C_{17}H_{19}NO_5$	221.0821	163.0400	147.0439
231	Procyanidin A2	Extrasynthese	577.1266	$C_{30}H_{24}O_{12}$	425.0736	287.0503	137.0217
232	Procyanidin B2	Chromadex	579.1793	$C_{30}H_{26}O_{12}$	579.1793	287.0711	139.0469
233	Prostaglandin B2	Sigma	335.2134	$C_{20}H_{30}O_4$	317.2001	299.1954	185.0894
234	Protocatechuic acid	Extrasynthese	155.0284	$C_7H_6O_4$	137.0193	109.0257	81.0305
235	Protoporphyrin IX	Cambridge Chemicals	563.2653	$C_{34}H_{34}N_4O_4$	517.2627	504.2514	457.2219
236	Protoveratrine A	Sigma	794.4291	$C_{41}H_{63}NO_{14}$	658.3557	436.2516	162.1289
237	Psilostachyin A	1717 CheMall	281.1399	$C_{15}H_{20}O_5$	245.1185	203.1084	128.0632
242	Quercetin	Sigma	303.0500	$C_{15}H_{10}O_7$	303.0601	/	/
243	Quercetin-3,7-di-O-beta-D-glucoside	Chromadex	627.1467	$C_{27}H_{30}O_{17}$	487.0738	465.0931	303.0458
244	Quercetin-3-arabinoglucoside	Sigma	357.5072	$C_{26}H_{28}O_{16}$	303.0601	/	/
245	Quercetin-3-O-glucoside	Sigma	465.1033	$C_{21}H_{20}O_{12}$	303.0601	/	/
246	Quercetin-3-O-glucoside-2-gallate	Sigma	587.4619	$C_{27}H_{22}O_{15}$	303.0601	/	/
247	Quercetin-3-rhamnoside	Extrasynthese	449.1218	$C_{21}H_{20}O_{11}$	303.0601	229.0564	153.0211

248	Quercetin-7-O-glucoside	Extrasynthese	465.0931	$C_{21}H_{20}O_{12}$	465.0976	303.0458	153.0161
249	Quinidine	Chromadex	325.1937	$C_{20}H_{24}N_2O_2$	307.1791	160.0760	117.0567
250	Quinine	Sigma	325.1937	$C_{20}H_{24}N_2O_2$	174.0556	160.0760	117.0567
252	Repin	Sigma	363.1470	$C_{19}H_{22}O_7$	243.1066	147.0822	90.9794
253	Reserpic acid	Chromadex	401.2030	$C_{22}H_{28}N_2O_5$	240.1203	174.0880	131.0707
254	Reynosine	Sigma	249.1480	$C_{15}H_{20}O_3$	231.1371	185.1343	128.0609
255	Rhein	Sigma	285.0316	$C_{15}H_8O_6$	241.0423	183.0375	155.0437
256	Robinin	Extrasynthese	741.2443	$C_{33}H_{40}O_{19}$	617.1636	433.1226	287.0607
257	Rubijervine	1717 CheMall	414.3368	$C_{27}H_{43}NO_2$	226.9530	140.0681	98.0959
258	Rutin	Extrasynthese	611.1463	$C_{27}H_{30}O_{16}$	303.0529	/	/
259	Salicylic acid	Extrasynthese	139.0373	$C_7H_6O_3$	121.0277	93.0328	72.9374
260	Sanguinarine	Sigma	333.1013	$C_{20}H_{14}NO_4$	274.0878	216.0818	189.0704
261	Santamarine	Sigma	249.1480	$C_{15}H_{20}O_3$	231.1402	185.1343	141.0741
262	Schkuhriolide	Atomax Chemicals	263.1311	$C_{15}H_{18}O_4$	245.1217	199.1135	128.0632
263	Scoparone	Extrasynthese	207.0637	$C_{11}H_{10}O_4$	191.0334	163.0390	151.0748
264	Scopolamine	Chromadex	304.1546	$C_{17}H_{21}NO_4$	156.1017	138.0908	103.0539
265	Scopoletin	Extrasynthese	193.0478	$C_{10}H_8O_4$	178.0257	150.0298	133.0276
266	Scopoline	Sigma	156.1042	$C_8H_{13}NO_2$	138.0932	110.0974	94.0685
267	Serotonin	Sigma	177.1048	$C_{10}H_{12}N_2O$	160.0786	132.0829	115.0563
268	Serpentine	Boc Sciences	349.1555	$C_{21}H_{20}N_2O_3$	317.1259	206.0824	151.0529
269	Silibinin	Sigma	483.1239	$C_{25}H_{22}O_{10}$	465.1152	195.0269	153.0186
270	Sinapic acid	Extrasynthese	225.0799	$C_{11}H_{12}O_5$	207.0695	175.0444	119.0516
271	Solanine	Sigma	868.5105	$C_{45}H_{73}NO_{15}$	560.3958	398.3458	98.0979
272	Sucrose	Sigma	365.0866	$C_{12}H_{22}O_{11}Na$	203.0460	185.0300	98.9718
273	Syringetin-3-glucoside	Extrasynthese	509.1323	$C_{23}H_{24}O_{13}$	347.0778	287.0572	258.0544
274	Syringic acid	Extrasynthese	199.0617	$C9H_{10}O_5$	140.0482	125.0219	97.0301
275	Tatridin B	Extrasynthese	265.1475	$C_{15}H_{20}O_4$	247.1356	229.1246	133.1023
276	Taxifolin	Extrasynthese	305.0842	$C_{15}H_{12}O_7$	259.0750	195.0384	153.0262
277	Telekin	Sigma	249.1512	$C_{15}H_{20}O_3$	185.1315	155.0889	128.0656
278	Terfenadine	Sigma	472.3181	$C_{32}H_{41}NO_2$	436.2944	262.1543	129.0674

279	Thiamine	Sigma	266.1137	$C_{12}H_{17}N_4OS$	133.0564	122.0175	81.0452
280	trans-Cinnamic acid	Sigma	149.0609	$C_9H_8O_2$	131.0491	103.0559	77.0412
281	trans-Piceatannol	Extrasynthese	245.0863	$C_{14}H_{12}O_4$	199.0762	135.0470	107.0531
282	trans-Resveratrol	Sigma	229.0936	$C_{14}H_{12}O_3$	165.0739	135.0493	107.0531
283	Tropacocain	1717 CheMall	246.1504	$C_{15}H_{19}NO_2$	124.1106	105.0315	96.0807
284	Tyramine	Sigma	138.0926	$C_8H_{11}NO$	121.0638	103.0538	77.0394
285	Umbelliferone	Sigma	163.0390	$C_9H_6O_3$	119.0472	107.0488	91.0530
286	Vanillic acid	Chromadex	169.0483	$C_8H_8O_4$	151.0370	93.0309	65.0387
287	Vanillin	Chromadex	153.0540	$C_8H_8O_3$	125.0585	110.0349	93.0328
288	Veratramine	Sigma	410.3044	$C_{27}H_{39}NO_2$	392.2930	295.2059	157.0983
289	Veratrine	Sigma	592.3406	$C_{32}H_{49}NO_9$	456.2701	332.2188	188.1426
290	Vigabatrin	Sigma	130.0858	$C_6H_{11}NO_2$	107.0003	97.9946	84.9580
291	Vitexin	Extrasynthese	433.1099	$C_{21}H_{20}O_{10}$	337.0704	313.0671	283.0580
292	Warfarin	Sigma	309.1127	$C_{19}H_{16}O_4$	/	/	/
293	Xanthaline	Sigma	354.1386	$C_{20}H_{19}NO_5$	173.0469	157.0521	144.0455
294	Xanthine	Sigma	153.0397	$C_5H_4N_4O_2$	136.0152	106.9982	97.9925
295	Xerantholide	Sigma	247.1356	$C_{15}H_{18}O_3$	229.1215	201.1263	91.0575
296	Yohimbine	Extrasynthese	355.2066	$C_{21}H_{26}N_2O_3$	212.1306	144.0823	117.0700
297	Zinniol	Sigma	267.1677	$C_{15}H_{22}O_4$	191.1804	/	/
298	ααα-20R-24R-ethylcholestane	Sigma	401.4147	$C_{29}H_{52}$	191.1804	/	/
299	ααα-20R-cholestane	Sigma	373.3834	$C_{27}H_{48}$	191.1804	/	/
300	αββ-20R-24R-ethylcholestane	Sigma	401.4147	$C_{29}H_{52}$	191.1806	/	/
301	αββ-20R-24R-ethylcholestane	Sigma	401.4147	$C_{29}H_{52}$	191.1805	/	/
302	αββ-20R-24S-methylcholestane	Sigma	387.3991	$C_{28}H_{50}$	191.1801	/	/
303	αββ-20R-cholestane	Sigma	373.3834	$C_{27}H_{48}$	191.1804	/	/

[a] Positive ionization MS^e acquisition in resolution mode: Low energy cone voltage to produce $[M+H]^+$ = 6.0 eV, high energy cone voltage ramp to produce accurate mass fragments = 10.0-50.0 eV with LCMS.

*An "/" indicates the absence of accurate mass fragments using LC-QToF-MS with the method described above

REFERENCES

Antoniades, D., Michelutti, N., Quinlan, R., Blais, J.M., Bonilla, S., Douglas, M.S.V., Pienitz, R., Smol, J.P., Vincent, W.F., 2011. Cultural eutrophication, anoxia, and ecosystem recovery in Meretta Lake, High Arctic Canada. Limnol. Oceanogr. 56, 639–650. https://doi.org/10.4319/lo.2011.56.2.0639

Aydin, N., Daher, S., Gülaçar, F.O., 2003. On the sedimentary occurrence of chlorophyllone a. Chemosphere 52, 937–942. https://doi.org/10.1016/S0045-6535(03)00296-0

Beale, D.J., Crosswell, J., Karpe, A.V., Ahmed, W., Williams, M., Morrison, P.D., Metcalfe, S., Staley, C., Sadowsky, M.J., Palombo, E.A., Steven, A.D.L., 2017. A multi-omics based ecological analysis of coastal marine sediments from Gladstone, in Australia's Central Queensland, and Heron Island, a nearby fringing platform reef. Sci. Total Environ. 609, 842–853. https://doi.org/10.1016/j.scitotenv.2017.07.184

Beale, D.J., Crosswell, J., Karpe, A.V., Metcalfe, S.S., Morrison, P.D., Staley, C., Ahmed, W., Sadowsky, M.J., Palombo, E.A., Steven, A.D.L., 2018. Seasonal metabolic analysis of marine sediments collected from Moreton Bay in South East Queensland, Australia, using a multi-omics-based approach. Sci. Total Environ. 631–632, 1328–1341. https://doi.org/10.1016/j.scitotenv.2018.03.106

Bell, M., Blais, J.M., 2019. "-Omics" workflow for paleolimnological and geological archives: A review. Sci. Total Environ. 672, 438–455. https://doi.org/10.1016/j.scitotenv.2019.03.477

Blumenberg, M., Hoppert, M., Kruger, M., Dreier, A., Thiel, V., 2012. Novel findings on hopanoid occurrences among sulfate reducing bacteria: Is there a direct link to nitrogen fixation? Org. Geochem. 49, 1–5.

Bostock, H.S., 2014. Geology, Physiographic Regions of Canada.

Brandt, J.P., 2009. The extent of the North American boreal zone. Environ. Rev. 17, 101–161.

Brown, J., Ferrians, O., Heginbottom, J.A., Melnikov, E., 2002. Circum-Arctic Map of Permafrost and Ground-Ice Conditions, Version 2.

Brunetti, A., Neto, F.C., C. Vera, M., Taboada, C., P. Pavarini, D., Bauermeister, A., P. Lopes, N., 2018. An integrative omics perspective for the analysis of chemical signals in ecological interactions. Chem. Soc. Rev. 47, 1574–1591. https://doi.org/10.1039/C7CS00368D

Brunetti, C., George, R.M., Tattini, M., Field, K., Davey, M.P., 2013. Metabolomics in plant environmental physiology. J. Exp. Bot. 64, 4011–4020. https://doi.org/10.1093/jxb/ert244

Buckinhgam, J., Baggaley, K.H., Roberts, A.D., Szabó, L., 2010. Dictionary of Alkaloids, Second Edition. ed. CRC Press, Taylor & Francis Group, Boca Raton, FL.

Camarero, Lluís, Rogora, Michela, Mosello, Rosario, Anderson, Nicholas J., Barbieri, Alberto, Botev, I., Kernan, M., Kopáček, J., Korhola, A., Lotter, A.F., Muri, G., Postolache, C., Stuchlík, E., Thies, H., Wright, R.F., 2009. Regionalisation of chemical variability in European mountain lakes. Freshw. Biol. 54, 2452–2469. https://doi.org/10.1111/j.1365-2427.2009.02296.x

Catalan, J., Curtis, C.J., Kernan, M., 2009. Remote European mountain lake ecosystems: regionalisation and ecological status. Freshw. Biol. 54, 2419–2432. https://doi.org/10.1111/j.1365-2427.2009.02326.x

Cooke, M.P., Talbot, H.M., Farrimond, P., 2008. Bacterial populations recorded in bacteriohopanepolyol distributions in soils from Northern England. Org. Geochem. 39, 1347–1358. https://doi.org/10.1016/j.orggeochem.2008.05.003

Cooke, M.P., van Dongen, B.E., Talbot, H.M., Semiletov, I., Shakhova, N., Guo, L., Gustafsson, Ö., 2009. Bacteriohopanepolyol biomarker composition of organic matter exported to the Arctic Ocean by seven of the major Arctic rivers. Org. Geochem. 40, 1151–1159. https://doi.org/10.1016/j.orggeochem.2009.07.014

Chew, A.G.M., Bryant, D.A., 2007. Chlorophyll Biosynthesis in Bacteria: The Origins of Structural and Functional Diversity. Annu. Rev. Microbiol. 61, 113–129

de Castro, M.D.L., Priego-Capote, F., 2018. The analytical process to search for metabolomics biomarkers. J. Pharm. Biomed. Anal. 147, 341–349. https://doi.org/10.1016/j.jpba.2017.06.073

Doğrul Selver, A., Sparkes, R.B., Bischoff, J., Talbot, H.M., Gustafsson, Ö., Semiletov, I.P., Dudarev, O.V., Boult, S., van Dongen, B.E., 2015. Distributions of bacterial and archaeal membrane lipids in surface sediments reflect differences in input and loss of terrestrial organic carbon along a cross-shelf Arctic transect. Org. Geochem. 83–84, 16–26. https://doi.org/10.1016/j.orggeochem.2015.01.005

Doğrul Selver, A., Talbot, H.M., Gustafsson, Ö., Boult, S., van Dongen, B.E., 2012. Soil organic matter transport along an sub-Arctic river–sea transect. Org. Geochem. 51, 63–72. https://doi.org/10.1016/j.orggeochem.2012.08.002

Dullinger, S., Dirnböck, T., Grabherr, G., 2004. Modelling climate change-driven treeline shifts: relative effects of temperature increase, dispersal and invasibility. J. Ecol. 92, 241–252. https://doi.org/10.1111/j.0022-0477.2004.00872.x

Emerson, S., Hedges, J., 2003. Sediment diagenesis and benthic flux. Treatise Geochem. 6, 625.

Fang, K., Morris, J.L., Salonen, J.S., Miller, P.A., Renssen, H., Sykes, M.T., Seppä, H., 2013. How robust are Holocene treeline simulations? A model–data comparison in the European Arctic treeline region. J. Quat. Sci. 28, 595–604. https://doi.org/10.1002/jqs.2654

Farrés, M., Martrat, B., Mol, B. de, Grimalt, J.O., Tauler, R., 2015. Extraction of climatic signals from fossil organic compounds in marine sediments up to 11.7Ma old (IODP-U1318). Anal. Chim. Acta 879, 1–9. https://doi.org/10.1016/j.aca.2015.04.051

Fierer, N., Leff, J.W., Adams, B.J., Nielsen, U.N., Bates, S.T., Lauber, C.L., Owens, S., Gilbert, J.A., Wall, D.H., Caporaso, J.G., 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc. Natl. Acad. Sci. 109, 21390–21395. https://doi.org/10.1073/pnas.1215210110

Florian, C., Miller, G., Fogel, M., Wolfe, A., Vinebrooke, R., Geirsdóttir, Á., 2015. Algal pigments in Arctic lake sediments record biogeochemical changes due to Holocene climate variability and anthropogenic global change. J. Paleolimnol. 54, 53–69. https://doi.org/10.1007/s10933-015-9835-5

Gale, E.F., 1941. Production of amines by bacteria. Biochem. J. 35, 66–80.

Grigoriadou, A., Schwarzbauer, J., 2011. Non-target Screening of Organic Contaminants in Sediments from the Industrial Coastal Area of Kavala City (NE Greece). Water. Air. Soil Pollut. 214, 623–643. https://doi.org/10.1007/s11270-010-0451-8

Gudasz, C., Sobek, S., Bastviken, D., Koehler, B., Tranvik, L.J., 2015. Temperature sensitivity of organic carbon mineralization in contrasting lake sediments. J. Geophys. Res. Biogeosciences 120, 1215–1225. https://doi.org/10.1002/2015JG002928

Hargan, K.E., Stewart, E.M., Michelutti, N., Grooms, C., Kimpe, L.E., Mallory, M.L., Smol, J.P., Blais, J.M., 2018. Sterols and stanols as novel tracers of waterbird population dynamics in freshwater ponds. Proc R Soc B 285, 20180631. https://doi.org/10.1098/rspb.2018.0631

Harsch, M.A., Hulme, P.E., McGlone, M.S., Duncan, R.P., 2009. Are treelines advancing? A global meta-analysis of treeline response to climate warming. Ecol. Lett. 12, 1040–1049. https://doi.org/10.1111/j.1461-0248.2009.01355.x

Heiri, C., Bugmann, H., Tinner, W., Heiri, O., Lischke, H., 2006. A model-based reconstruction of Holocene treeline dynamics in the Central Swiss Alps. J. Ecol. 94, 206–216. https://doi.org/10.1111/j.1365-2745.2005.01072.x

Hernández, F., Sancho, J.V., Ibáñez, M., Abad, E., Portolés, T., Mattioli, L., 2012. Current use of high-resolution mass spectrometry in the environmental sciences. Anal. Bioanal. Chem. 403, 1251–1264. https://doi.org/10.1007/s00216-012-5844-7

Höfle, S.T., Kusch, S., Talbot, H.M., Mollenhauer, G., Zubrzycki, S., Burghardt, S., Rethemeyer, J., 2015. Characterisation of bacterial populations in Arctic permafrost soils using bacteriohopanepolyols. Org. Geochem. 88, 1–16. https://doi.org/10.1016/j.orggeochem.2015.08.002

Hollender, J., Schymanski, E.L., Singer, H.P., Ferguson, P.L., 2017. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environ. Sci. Technol. 51, 11505–11512. https://doi.org/10.1021/acs.est.7b02184

Hopmans, E.C., Weijers, J.W., Schefuß, E., Herfort, L., Damsté, J.S.S., Schouten, S., 2004. A novel proxy for terrestrial organic matter in sediments based on branched and isoprenoid tetraether lipids. Earth Planet. Sci. Lett. 224, 107–116.

Hörtensteiner, S., Kräutler, B., 2011. Chlorophyll breakdown in higher plants. Biochim. Biophys. Acta 1807, 977–988. https://doi.org/10.1016/j.bbabio.2010.12.007

Inglis, G.N., Naafs, B.D.A., Zheng, Y., McClymont, E.L., Evershed, R.P., Pancost, R.D., 2018. Distributions of geohopanoids in peat: Implications for the use of hopanoid-based proxies in natural archives. Geochim. Cosmochim. Acta 224, 249–261. https://doi.org/10.1016/j.gca.2017.12.029

IPCC, 2018. Special Report on Global Warming of 1.5 °C (SR15). Intergovernmental Panel on Climate Change, Incheon, Republic of Korea.

Kannenberg, E.L., Poralla, K., 1999. Hopanoid Biosynthesis and Function in Bacteria. Naturwissenschaften 86, 168–176. https://doi.org/10.1007/s001140050592

Karaman, I., 2017. Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis, in: Metabolomics: From Fundamentals to Clinical Applications, Advances in Experimental Medicine and Biology. Springer, Cham, pp. 145–161. https://doi.org/10.1007/978-3-319-47656-8_6

Kaufman, D.S., Ager, T.A., Anderson, N.J., Anderson, P.M., Andrews, J.T., Bartlein, P.J., Brubaker, L.B., Coats, L.L., Cwynar, L.C., Duvall, M.L., Dyke, A.S., Edwards, M.E., Eisner, M.R., Gajewski, K., Geirsdóttiro, O., Hu, F.S., Jennings, A.E., Kaplan, M.R., Kerwin, M.W., Lozhkin, L.V., MacDonald, G.M., Miller, G.H., Mock, C.J., Oswald, W.W., Otto-Bliesner, B.L., Porinchu, D.F., Rühland, K., Smol, J.P., Steig, E.J., Wolfe, B.B., 2004. Holocene thermal maximum in the western Arctic (0–180 W). Quat. Sci. Rev. 23, 529–560.

Keely, B.J., 2006. Geochemistry of Chlorophylls, in: Chlorophylls and Bacteriochlorophylls: Biochemitsry, Biophysics, Functions and Applications, Advances in Photosynthesis and Respiration. Springer, Dordrecht, Netherlands, pp. 535–561.

Kernan, M., Ventura, M., Bitušík, P., Brancelj, A., Clarke, G., Velle, G., Raddum, G.G., Stuchlík, E., Catalan, J., 2009. Regionalisation of remote European mountain lake ecosystems according to their biota: environmental versus geographical patterns. Freshw. Biol. 54, 2470–2493. https://doi.org/10.1111/j.1365-2427.2009.02284.x

Knicker, H., 2004. Stabilization of N-compounds in soil and organic-matter-rich sediments—what is the difference? Mar. Chem. 92, 167–195.
Korosi, J.B., Thienpont, J.R., Pisaric, M.F., 2017. Broad-scale lake expansion and flooding inundates essential wood bison habitat. Nat. Commun. 8.
Lankadurai, B.P., Nagato, E.G., Simpson, M.J., 2013. Environmental metabolomics: an emerging approach to study organism responses to environmental stressors. Environ. Rev. 21, 180–205. https://doi.org/10.1139/er-2013-0011
Last, W.M., Smol, J.P., 2006. Tracking Environmental Change Using Lake Sediments: Volume 2: Physical and Geochemical Methods. Springer Netherlands.
Leavitt, P.R., Hodgson, D.A., 2001. Sedimentary pigments., in: Tracking Environmental Change Using Lake Sediments. Volume 3: Terrestrial, Algal, and Siliceous Indicators. Kluwer Academic Publishers, Dordrect.
Lee, K.-M., Jeon, J.-Y., Lee, B.-J., Lee, H., Choi, H.-K., 2017. Application of Metabolomics to Quality Control of Natural Product Derived Medicines. Biomol. Ther. 25, 559–568. https://doi.org/10.4062/biomolther.2016.249
MacDonald, G.M., 2010. Some Holocene palaeoclimatic and palaeoenvironmental perspectives on Arctic/Subarctic climate warming and the IPCC 4th Assessment Report. J. Quat. Sci. 25, 39–47.
Marshall, I.B., Schut, P.H., Ballard, M., 1999. A National Ecological Framework for Canada: Attribute Data. Agriculture and Agri-Food Canada, Research Branch, Centre for Land and Biological Resources Research and Environment Canada, State of the Environment Directorate, Ecozone Analysis Branch, Ottawa/Hull.
Meyers, P.A., Ishiwatari, R., 1993. Lacustrine Organic Geochemistry - An Overview of Indicators of Organic Matter Sources and Diagenesis in Lake Sediments. Org. Geochem. 20, 867–900. https://doi.org/10.1016/0146-6380(93)90100-p
Michelutti, N., Blais, J.M., Cumming, B.F., Paterson, A.M., Rühland, K., Wolfe, A.P., Smol, J.P., 2010. Do spectrally inferred determinations of chlorophyll a reflect trends in lake trophic status? J. Paleolimnol. 43, 205–217.
Michelutti, N., Smol, J., 2016. Visible spectroscopy reliably tracks trends in paleo-production. J. Paleolimnol. 56, 253–265. https://doi.org/10.1007/s10933-016-9921-3
Michelutti, N., Wolfe, A.P., Briner, J.P., Miller, G.H., 2007. Climatically controlled chemical and biological development in Arctic lakes. J. Geophys. Res. Biogeosciences 112. https://doi.org/10.1029/2006JG000396
Myers-Smith, I.H., Hik, D.S., 2018. Climate warming as a driver of tundra shrubline advance. J. Ecol. 106, 547–560. https://doi.org/10.1111/1365-2745.12817
Ourisson, G., Rohmer, M., 1992. Hopanoids. 2. Biohopanoids: a novel class of bacterial lipids. Acc. Chem. Res. 25, 403–408. https://doi.org/10.1021/ar00021a004
Payette, S., Fortin, M.J., Gamache, I., 2001. The subarctic forest-tundra: the structure of a biome in a changing climate. Bioscience 51, 709–718
Pearson, A., Flood Page, S.R., Jorgenson, T.L., Fischer, W.W., Higgins, M.B., 2007. Novel hopanoid cyclases from the environment. Environ. Microbiol. 9, 2175–2188. https://doi.org/10.1111/j.1462-2920.2007.01331.x
Pearson, R.G., Phillips, S.J., Loranty, M.M., Beck, P.S.A., Damoulas, T., Knight, S.J., Goetz, S.J., 2013. Shifts in Arctic vegetation and associated feedbacks under climate change. Nat. Clim. Change 3, 673–677. https://doi.org/10.1038/nclimate1858

Pienitz, R., Smol, J.P., Last, W.M., Leavitt, P.R., Cumming, B.F., 2000. Multi-proxy Holocene palaeoclimatic record from a saline lake in the Canadian Subarctic. The Holocene 10, 673–686.
Pienitz, R., Smol, J.P., Lean, D.R., 1997a. Physical and chemical limnology of 59 lakes located between the southern Yukon and the Tuktoyaktuk Peninsula, Northwest Territories (Canada). Can. J. Fish. Aquat. Sci. 54, 330–346.
Pienitz, R., Smol, J.P., Lean, D.R., 1997b. Physical and chemical limnology of 24 lakes located between Yellowknife and Contwoyto Lake, Northwest Territories (Canada). Can. J. Fish. Aquat. Sci. 54, 347–358.
Pinto, R.C., 2017. Chemometrics Methods and Strategies in Metabolomics, in: Metabolomics: From Fundamentals to Clinical Applications, Advances in Experimental Medicine and Biology. Springer, Cham, pp. 163–190. https://doi.org/10.1007/978-3-319-47656-8_7
Ranson, K.J., Montesano, P.M., Nelson, R., 2011. Object-based mapping of the circumpolar taiga–tundra ecotone with MODIS tree cover. Remote Sens. Environ. 115, 3670–3680. https://doi.org/10.1016/j.rse.2011.09.006
Rapinski, M., Liu, R., Saleem, A., Arnason, J.T., Cuerrier, A., 2014. Environmental trends in the variation of biologically active phenolic compounds in Labrador tea, Rhododendron groenlandicum, from northern Quebec, Canada. Botany 92, 783–794.
Rühland, K.M., Paterson, A.M., Keller, W., Michelutti, N., Smol, J.P., 2013. Global warming triggers the loss of a key Arctic refugium. Proc R Soc B 280, 20131887. https://doi.org/10.1098/rspb.2013.1887
Rühland, K.M., Smol, J.P., Wang, X., Muir, D.C.G., 2003. Limnological characteristics of 56 lakes in the Central Canadian Arctic Treeline Region. J. Limnol. 62, 9–27.
Sáenz, J.P., Sezgin, E., Schwille, P., Simons, K., 2012. Functional convergence of hopanoids and sterols in membrane ordering. Proc. Natl. Acad. Sci. 109, 14236–14240. https://doi.org/10.1073/pnas.1212141109
Saleem, A., Harris, C.S., Asim, M., Cuerrier, A., Martineau, L., Haddad, P.S., Arnason, J.T., 2010. A RP-HPLC-DAD-APCI/MSD Method for the Characterisation of Medicinal Ericaceae used by the Eeyou Istchee Cree First Nations. Phytochem. Anal. 21, 328–339.
Savage, C., Leavitt, P.R., Elmgren, R., 2010. Effects of land use, urbanization, and climate variability on coastal eutrophication in the Baltic Sea. Limnol. Oceanogr. 55, 1033–1046. https://doi.org/10.4319/lo.2010.55.3.1033
Schaeffer, P., Schmitt, G., Adam, P., Rohmer, M., 2010. Abiotic formation of 32,35-anhydrobacteriohopanetetrol: A geomimetic approach. Org. Geochem., Advances in Organic Geochemistry 2009 41, 1005–1008. https://doi.org/10.1016/j.orggeochem.2010.04.013
Schibalski, A., Lehtonen, A., Hickler, T., Schröder, B., 2017. Comparing correlative and process-based modelling approaches in a boreal forest identifies important areas for model development. Silva Fenn. 51. https://doi.org/10.14214/sf.6977
Schouten, S., Hopmans, E.C., Damste, J.S.S., 2013. The organic geochemistry of glycerol dialkyl glycerol tetraether lipids: A review. Org. Geochem. 54, 19–61. https://doi.org/10.1016/j.orggeochem.2012.09.006
Schwörer, C., Henne, P.D., Tinner, W., 2014. A model-data comparison of Holocene timberline changes in the Swiss Alps reveals past and future drivers of mountain forest dynamics. Glob. Change Biol. 20, 1512–1526. https://doi.org/10.1111/gcb.12456

Seidl, R., Thom, D., Kautz, M., Martin-Benito, D., Peltoniemi, M., Vacchiano, G., Wild, J., Ascoli, D., Petr, M., Honkaniemi, J., 2017. Forest disturbances under climate change. Nat. Clim. Change 7, 395–402.

Shang, N., Saleem, A., Musallam, L., Walshe-Roussel, B., Badawi, A., Cuerrier, A., Arnason, J.T., Haddad, P.S., 2015. Novel Approach to Identify Potential Bioactive Plant Metabolites: Pharmacological and Metabolomics Analyses of Ethanol and Hot Water Extracts of Several Canadian Medicinal Plants of the Cree of Eeyou Istchee. PLOS ONE 10, e0135721. https://doi.org/10.1371/journal.pone.0135721

Shunthirasingham, C., Simpson, M.J., 2006. Investigation of bacterial hopanoid inputs to soils from Western Canada. Appl. Geochem. 21, 964–976. https://doi.org/10.1016/j.apgeochem.2006.03.007

Sinninghe Damsté, J.S., Rijpstra, W.I.C., Schouten, S., Fuerst, J.A., Jetten, M.S.M., Strous, M., 2004. The occurrence of hopanoids in planctomycetes: implications for the sedimentary biomarker record. Org. Geochem. 35, 561–566. https://doi.org/10.1016/j.orggeochem.2004.01.013

Smik, L., Cabedo-Sanz, P., Belt, S.T., 2016. Semi-quantitative estimates of paleo Arctic sea ice concentration based on source-specific highly branched isoprenoid alkenes: a further development of the PIP 25 index. Org. Geochem. 92, 63–69.

Spencer-Jones, C.L., Wagner, T., Dinga, B.J., Schefuß, E., Mann, P.J., Poulsen, J.R., Spencer, R.G.M., Wabakanghanzi, J.N., Talbot, H.M., 2015. Bacteriohopanepolyols in tropical soils and sediments from the Congo River catchment area. Org. Geochem. 89–90, 1–13. https://doi.org/10.1016/j.orggeochem.2015.09.003

Squier, A.H., Hodgson, D.A., Keely, B.J., 2002. Sedimentary pigments as markers for environmental change in an Antarctic lake. Org. Geochem. 33, 1655–1665. https://doi.org/10.1016/S0146-6380(02)00177-8

Stewart, E.M., Michelutti, N., Shenstone-Harris, S., Grooms, C., Weseloh, C., Kimpe, L.E., Blais, J.M., Smol, J.P., 2015. Tracking the History and Ecological Changes of Rising Double-Crested Cormorant Populations Using Pond Sediments from Islands in Eastern Lake Ontario. PLoS ONE 10. https://doi.org/10.1371/journal.pone.0134167

Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H., Beger, R., Daykin, C.A., Fan, T.W.-M., Fiehn, O., Goodacre, R., Griffin, J.L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A.N., Lindon, J.C., Marriott, P., Nicholls, A.W., Reily, M.D., Thaden, J.J., Viant, M.R., 2007. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics Off. J. Metabolomic Soc. 3, 211–221. https://doi.org/10.1007/s11306-007-0082-2

Swenson, T.L., Jenkins, S., Bowen, B.P., Northen, T.R., 2015. Untargeted soil metabolomics methods for analysis of extractable organic matter. Soil Biol. Biochem. 80, 189–198. https://doi.org/10.1016/j.soilbio.2014.10.007

Swenson, T.L., Karaoz, U., Swenson, J.M., Bowen, B.P., Northen, T.R., 2018. Linking soil biology and chemistry in biological soil crust using isolate exometabolomics. Nat. Commun. 9. https://doi.org/10.1038/s41467-017-02356-9

Szymczak-Żyla, M., Kowalewska, G., Louda, J.W., 2008. The influence of microorganisms on chlorophyll a degradation in the marine environment. Limnol. Oceanogr. 53, 851–862. https://doi.org/10.4319/lo.2008.53.2.0851

Talapatra, S.K., Talapatra, B., 2015. Chemisry of Natural Products: Stereochemistry, Conformation, Synthesis, Biology and Medicine. Springer-Verlag Berlin Heidelberg, London.

Talbot, H.M., Farrimond, P., 2007. Bacterial populations recorded in diverse sedimentary biohopanoid distributions. Org. Geochem. 38, 1212–1225. https://doi.org/10.1016/j.orggeochem.2007.04.006

Talbot, H.M., Handley, L., Spencer-Jones, C.L., Dinga, B.J., Schefuß, E., Mann, P.J., Poulsen, J.R., Spencer, R.G.M., Wabakanghanzi, J.N., Wagner, T., 2014. Variability in aerobic methane oxidation over the past 1.2Myrs recorded in microbial biomarker signatures from Congo fan sediments. Geochim. Cosmochim. Acta 133, 387–401. https://doi.org/10.1016/j.gca.2014.02.035

Talbot, H.M., McClymont, E.L., Inglis, G.N., Evershed, R.P., Pancost, R.D., 2016. Origin and preservation of bacteriohopanepolyol signatures in Sphagnum peat from Bissendorfer Moor (Germany). Org. Geochem. 97, 95–110. https://doi.org/10.1016/j.orggeochem.2016.04.011

Taylor, K.A., Rodger Harvey, H., 2011. Bacterial hopanoids as tracers of organic carbon sources and processing across the western Arctic continental shelf. Org. Geochem. 42, 487–497. https://doi.org/10.1016/j.orggeochem.2011.03.012

Veenaas, C., Haglund, P., 2017. Methodology for non-target screening of sewage sludge using comprehensive two-dimensional gas chromatography coupled to high-resolution mass spectrometry. Anal. Bioanal. Chem. 409, 4867–4883. https://doi.org/10.1007/s00216-017-0429-0

Viau, A.E., Gajewski, K., 2009. Reconstructing millennial-scale, regional paleoclimates of boreal Canada during the Holocene. J. Clim. 22, 316–330.

Volkman, J.K., 2005. Sterols and other triterpenoids: source specificity and evolution of biosynthetic pathways. Org. Geochem. 36, 139–159. https://doi.org/10.1016/j.orggeochem.2004.06.013

Welander, P.V., Hunter, R.C., Zhang, L., Sessions, A.L., Summons, R.E., Newman, D.K., 2009. Hopanoids Play a Role in Membrane Integrity and pH Homeostasis in Rhodopseudomonas palustris TIE-1. J. Bacteriol. 191, 6145–6156. https://doi.org/10.1128/JB.00460-09

Xu, X., Li, W., Deng, H., Fujibayashi, M., Nomura, M., Nishimura, O., Wang, G., 2019. Approach deliberation for source identification of sedimentary organic matters via comparing freshwater lakes with multi-ecotypes. Sci. Total Environ. 649, 327–334. https://doi.org/10.1016/j.scitotenv.2018.08.236

Xu, X., Li, W., Fujibayashi, M., Nomura, M., Nishimura, O., Li, X., 2015. Predominance of terrestrial organic matter in sediments from a cyanobacteria- blooming hypereutrophic lake. Ecol. Indic. 50, 35–43. https://doi.org/10.1016/j.ecolind.2014.10.020

Zocatelli, R., Lavrieux, M., Guillemot, T., Chassiot, L., Le Milbeau, C., Jacob, J., 2017. Fecal biomarker imprints as indicators of past human land uses: Source distinction and preservation potential in archaeological and natural archives. J. Archaeol. Sci. 81, 79–89.

CHAPTER 4: A continental scale spatial investigation of lake sediment organic compositions using sedimentomics

Madison A. Bell[1*], David P. Overy[2], Jules M. Blais[1]

[1]Laboratory for the Analysis of Natural and Synthetic Environmental Toxicants, Department of Biology, University of Ottawa, Ottawa ON, K1N 6N5, Canada

[2]Agriculture and Agri-food Canada, Ottawa Research and Development Centre, Ottawa, ON, K1A 0C6, Canada

*Corresponding Author

Madison Bell

STATEMENT OF CONTRIBUTIONS

Madison A. Bell: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing – Original Draft, Visualization, Corresponding author. **David P. Overy:** Writing, Visualization, Supervision. **Jules M. Blais:** Resources, Methodology, Supervision, Writing, Project administration, Funding acquisition. *Science of the Total Environment* published this chapter:

Bell, M.A., Overy, D.P., and Blais, J.M. (2020) A continental scale spatial investigation of lake sediment organic compositions using sedimentomics. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.137746

ABSTRACT

Sedimentomics is a new method used to investigate carbon cycling in sediment organic matter. This untargeted method, based on metabolomics workflows, was used to investigate the molecular composition of sediment organic matter across northern Canada (Nunavut and Northwest Territories). Unique "lake districts" were defined using unsupervised clustering based on changes in sediment organic carbon compositions across space. Supervised machine learning analyses were used to compare the "lake districts" to commonly used regional classification systems like the treeline, ecozones, and/or georegions. Treeline was the best model to explain the compositional variance of sediment organic carbon from lakes across Canada, closely followed by the georegions model. A novel sediment metaphenomics analysis was also applied to determine how well environmental constraints explain the variation of sediment organic matter composition across a continent. We determined that sedimentomics is more informative than traditional measurements (such as total organic carbon) and can be integrated with other "omics" techniques.

1. INTRODUCTION

Organic carbon (C) refers to the pool of organic molecules deposited into soil and sediment environments. Total organic C (TOC) is a measurement of the sum of all organic C molecules in an environmental sample. Studying how TOC varies in molecular composition over time via biotic and abiotic factors is essential to further our understanding of organic C stabilization and decomposition, and in turn, improve CO_2 emission models, which are currently highly variable, but are essential to future climate change predictions (Wieder et al., 2018). The balance of emission versus organic C sequestration is spatially dynamic and dependent on a number of different environmental constraints

(i.e. climate, altitude, catchment, underlying geology, etc.) (McGowan et al., 2018). Recent publications have demonstrated there are many inconsistencies in the spatial dynamics of organic C cycling models across different systems; therefore, more investigations into the spatial dynamics of organic C cycling are required (Graham et al., 2018; Krause et al., 2017). Currently, there are few studies investigating the variance of organic C across large spatial gradients (Rossel et al., 2019; Zhao et al., 2019).

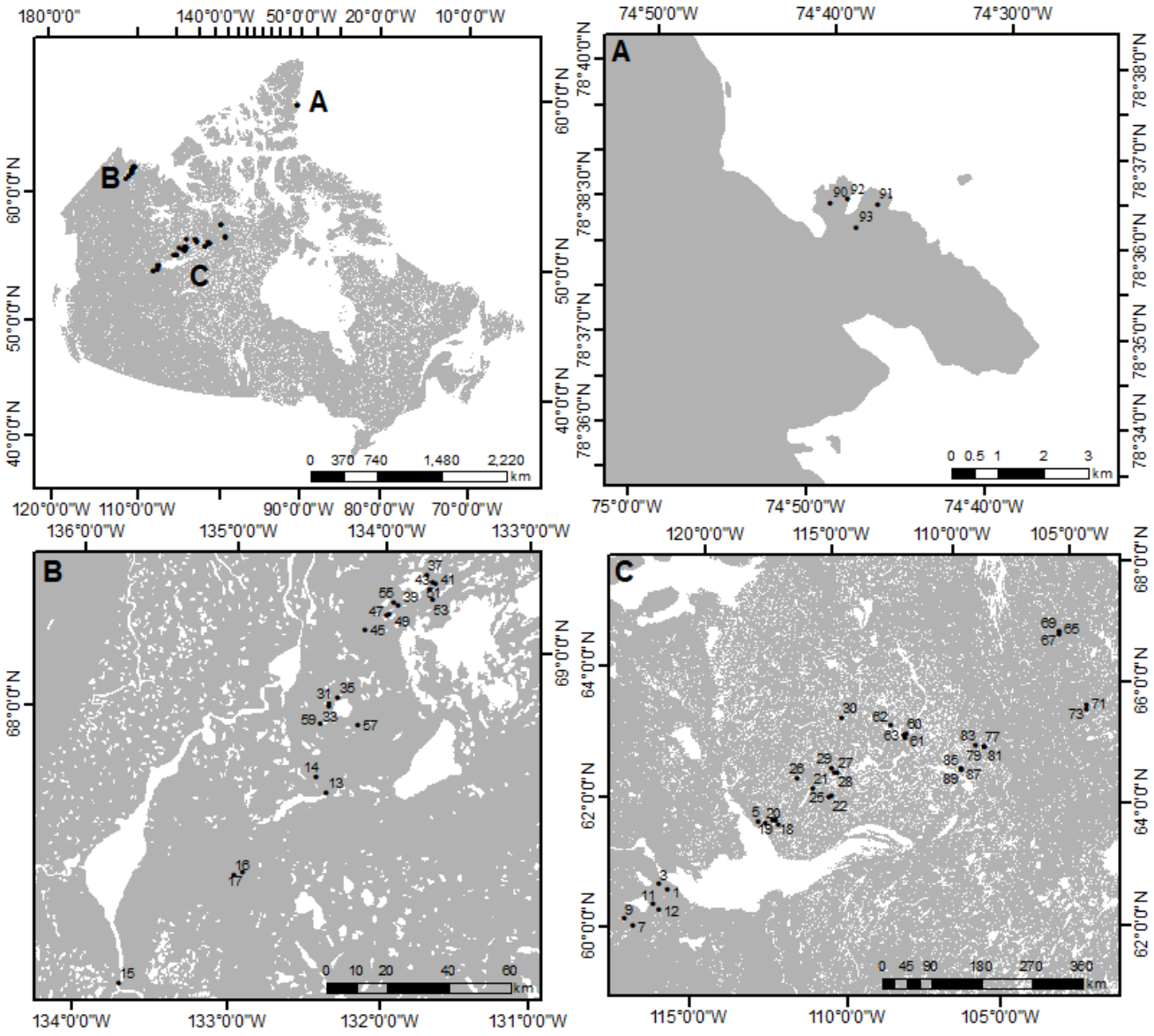


Figure 4.1: Distribution of sampled lakes across Canada including naming designations.

The Canadian North is vast and diverse (Figure 4.1) with regional boundaries defined by climate, geographic landforms, surface and subsurface geology, and biological adaptability. Canada's North contains 70% of Canada's coastline and major continental drainage basins (i.e. Mackenzie Delta) that border various large geological boundaries (i.e. Borderland versus Shield geology) (Bostock, 2014). At higher latitudes, mean solar radiation and temperature gradients influence the distribution of plants and animals (Harsch et al., 2009; Myers-Smith and Hik, 2018; Post et al., 2009). Combined, the northern landscape is spatially varied and subject to the influence of many environmental abiotic and biotic factors, which all have a direct effect upon organic C deposition and composition.

There are several spatial classification models based on a variety of environmental descriptors for northern regions (i.e. bedrock, permafrost, plants, thermal radiation, etc. (Figure 4.2)). Some models are based solely on the composition of terrestrial vegetation – i.e. the treeline model, which is defined by the boreal forest continuous tree line shifting into a transitional forest-tundra ecotone, and ultimately into the tundra region (Brandt, 2009; Ranson et al., 2011). Other models are limited to geological constraints such as permafrost conditions (Brown et al., 2002), or the underlying geology (i.e. georegions or geodistricts) (Bostock, 2014). While other classification models, for example the models proposed by the National Ecological Framework (NEF), account for both abiotic (i.e. elevation, landform, surface form, surface material, temperature, precipitation, windspeed, solar radiation, etc.) and biotic conditions (i.e. land cover, soil development, growing season, etc.) (Marshall et al., 1999). The NEF ecozones model contains the broadest zones and other NEF models subdivide ecozones into ecoprovinces, ecoregions, and ecodistricts using an increasingly finer resolution of ecological and geological features (Marshall et al., 1999).

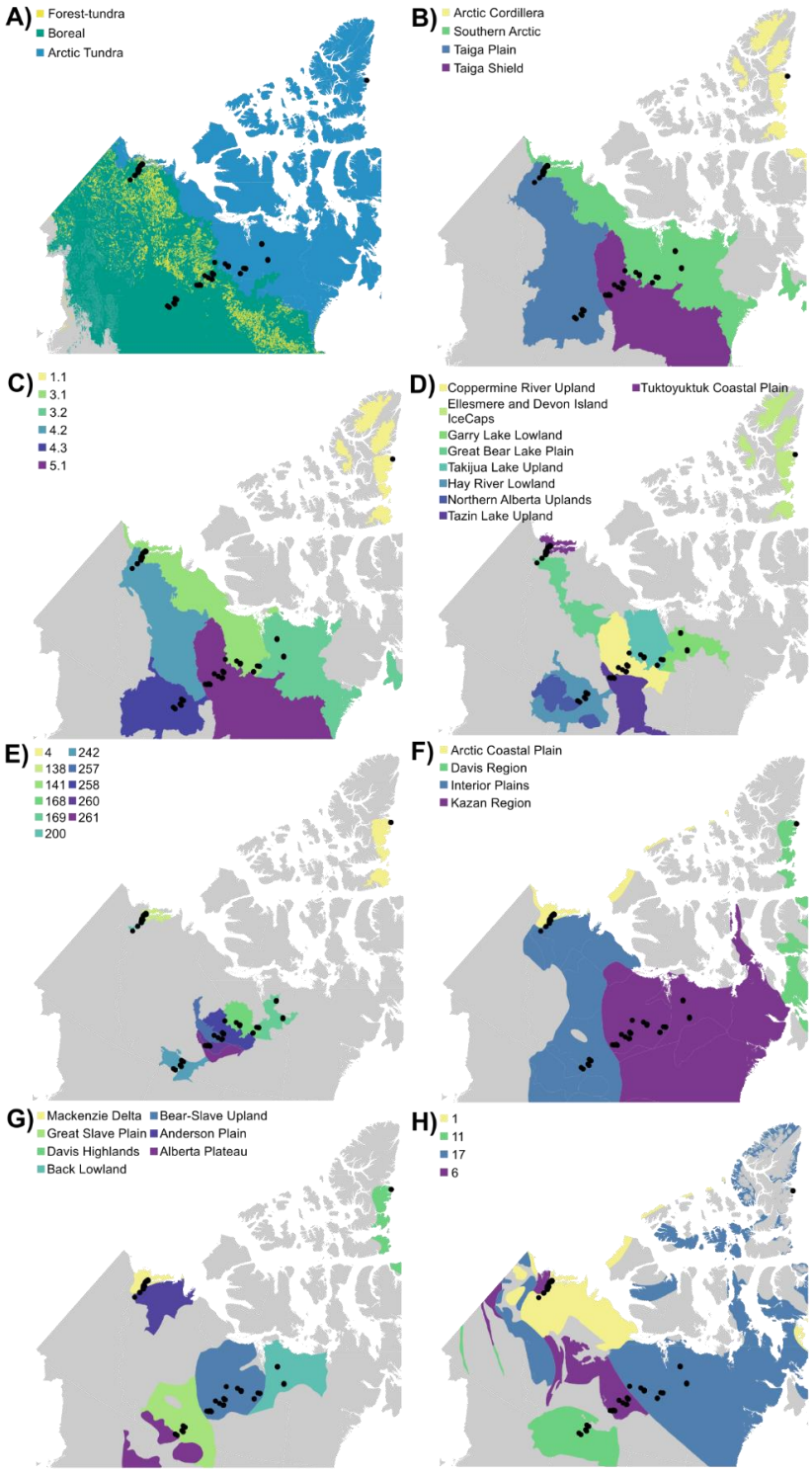


Figure 4.2: Distribution of sampled lakes superimposed over different environmental spatial classifications. From top left to right, and then bottom left to right the classification systems are: A) Treeline regions, B) Ecozones, C) Ecoprovinces, D) Ecoregions, E) Ecodistricts, F) Georegions, G) Geodistricts, and H) Permafrost.

Catalan et al. (2009) defined the “lake district concept” based on abiotic conditions providing constraints on catchment/lake biota, and on competitive population level interactions in the lake and surrounding catchment. These authors defined regional groupings between lakes with similar ecological identities. Other studies have demonstrated this concept through lake geochemistry (Camarero et al., 2009; Osburn et al., 2017; Pienitz et al., 1997a, 1997b; Rühland et al., 2003; Sobek et al., 2007), autochthonous biota (Catalan et al., 2009; Kernan et al., 2009), and microbes (Wang et al., 2016). Lakes during sedimentation actively modify environmental organic C and influence regional C budgets (Cole et al., 2007). Like soil, microbial processes within lake sediments actively emit CO_2, but unlike soil, they also sequester large amounts of C, primarily because they are continuously anaerobic (Kellerman et al., 2014; McGowan et al., 2018). As such, lake sediments preserve more molecular organic C than soil, acting as a natural archive or repository for recent C input/C deposition within a given geographical region and therefore are ideal for C cycle modelling. Theoretically, the molecular organic C composition of lake sediment can also be used to define regional/spatial variation, or lake districts.

Sedimentomics is an untargeted qualitative analysis method based on metabolomics to model compositional variation of sedimentary organic molecules (sedOM) (Bell and Blais, 2019). The term sedOM encompasses microbial metabolites, geometabolites (i.e. organic molecules formed through abiotic transformations during deposition), and other naturally derived compounds formed by abiotic processes (i.e. fire derived small molecules) and/or are synthetically produced (i.e. pollutants, plastics, etc.). To date, “omics” approaches have been applied to soil organic molecules/matter (SOM) (Jiménez-Morillo et al., 2018; Swenson et al., 2015b; Li et al., 2018; Swenson et al., 2018; Ward and Cory, 2015) and marine sedOM (Beale et al., 2018, 2017; Farrés et al., 2015; Oni et al., 2015), but only a few studies have looked at freshwater sedOM (Graham et al., 2018; Wu et al., 2018). Recently, our research group

applied sedimentomics to determine biomarkers for use in the paleolimnological reconstruction of the Canadian Arctic treeline from lake sediments (Saleem et al., 2019). Metaphenomics is an extension of sedimentomics that incorporates environmental constraints (i.e. temperature, solar radiation, geochemistry, etc.) with “omics” techniques to provide unprecedented insights into the interface between the environment, the organic C cycle, and microbial activity (Jansson and Hofmockel, 2018).

Our study integrates a freshwater sedimentomics dataset (Saleem et al., 2019), derived from lake sediment cores spanning Canada's North, with environmental constraints in a metaphenomic approach to explore whether organic C, in the form of individual organic carbon molecules, can be used to classify regions, or lake districts, in Canada's North. We hypothesized the sedOM composition from northern Canadian lakes will vary according to different environmental conditions, where lakes with similar environmental conditions share similar sedOM compositions (i.e. the lake district concept). To the best of our knowledge, this is the first study incorporating sedimentomics and metaphenomics to model how the molecular composition of freshwater sedOM varies across a large continental gradient. Insight into the spatial variance/dynamics of sedOM molecular composition is critical for the development of continental scale organic C diagenesis models that will prove essential when forecasting the temporal impact of climate change on lake sediments, their catchments and surrounding geographical regions.

2. METHODS AND MATERIALS

2.1. Sample collection

Sediment samples from 60 archived lake sediment cores were selected. The lake sediment samples described in this study are widely distributed across northern Canada (Figure 4.1). They span a longitudinal range of 134°W to 74°4′W, and a latitudinal range of 78°39′N to 60°N. The lake locations are

in the Canadian Northwest Territories and Nunavut. In addition, most of the lakes are north of the Arctic Circle boundary (66°33′47.6″ N).

The samples were collected using gravity cores between 1996 and 2014 (Table S4.1), and they were extruded into 0.25 or 0.5 cm intervals. Only the data from the surface sediments (within 6.25 cm of the core surface) were used in this study. The core depth (cm) of each surface sample is presented in Table S4.1. Surface and sub-surface sediments were analyzed previously (Saleem et al., 2019).

2.2. Sample extraction

The detailed extraction procedure was described previously (Saleem et al., 2019). In summary, sediment samples were freeze dried, and 4 mL of a methanol:acetonitrile:water (2:2:1) mixture was added to 100 mg of sample. Method blanks were prepared and processed with the samples. The samples were sonicated (5 min), and centrifuged (3000 RPM; 10 min). The supernatant was collected, and the process repeated. The pooled supernatants were filtered through 0.2 µm PTFE syringe filters.

2.3. Sample analysis

Extract analyses were performed at the University of Ottawa on an Acquity ultra high performance liquid chromatography (UPLC) coupled with a quadrupole time of flight (QTOF) XevoG2 mass spectrometer (MS; Waters Inc.) with a BEH C18 column (1.7 µm particle size; 2.1 mm × 100 mm column; Waters Inc.). The mobile phase was 95% A (water + 0.1% formic acid) + 5% B (acetonitrile + 0.1% formic acid) for 0.0–3.0 min, a linear gradient from 5 to 20% B was applied from 3.0–5.0 min, and then 100% B from 5.01–6.0 min. The flow rate was 0.8 mL/min and 5 µL of the extracts was injected.

The MS was operated in resolution mode with positive electrospray ionization. The spectra were acquired in centroid mode with mass range of 100–1500 Da. Lock mass was performed with leucine enkephalin ^{12}C (2 ng/µL) at 30 s infusion intervals. The temperatures for the source (150 °C) and desolvation (500 °C) were set as well as flow rates for the cone gas (50 L/h), and desolvation gas (N_2; 1200 L/h). Catechin standard mix (Cerilliant Analytical Reference Standards) was used to ensure reproducibility and accuracy of retention time and mass accuracy at ± 0.03 min, 5.0 ppm for parent ions, and 10 ppm for product ions.

Additional sediment was prepared for total organic carbon (TOC), nitrogen (TN), and sulfur (TS) analysis. 50 mg subsamples were acid desiccated with HCl and washed with deionized water. After the acid digestion, the samples were freeze dried, weighed into tin capsules, and analyzed at the GG Hatch Isotope Facility using an elemental analyzer interfaced to an isotope ratio MS.

2.4. Data preparation and statistical analysis

2.4.1. Variable matrix preparation

The raw spectral data from the MS were transferred to EZInfo™ (V8.03) for data processing. The noise level excluded intensities below 1000. The peak heights were normalized to the sum of the intensities. The resulting matrix had 620 variables (unique bins of the m/z values detected at specific retention times) with the normalized peak heights of each of the variables across the 60 samples. Further data processing was performed using Metaboanalyst (Chong et al., 2019). Missing values were replaced by a small number (half of the smallest minimum positive value in the matrix), and the data were filtered using interquartile range. Lastly, the data were pareto-scaled and cube root transformed

to account for large fold changes between variables and heteroscedasticity, respectively. Samples 13 and 14 were outliers and excluded from all analyses.

Molecular formulas were determined using MassLynx™ software (Version 4.1) and by consensus with Metlin, LipidMaps, and HMDB databases. The formulas were constrained with $^{12}C_{0–50}$, $^{1}H_{5–100}$, $^{14}N_{0–6}$, $^{16}O_{0–30}$, $^{32}S_{0–2}$, and $^{31}P_{0–2}$ and O/C ≤ 1.2., 0.3 ≤ H/C ≤ 2.25, N/C ≤ 0.5, S/ C ≤ 0.2, and double bond equivalents (DBE) ≥ 0 (Jiménez-Morillo et al., 2018; Stubbins et al., 2010). Molecular class assignments (i.e. lipid, tannin, etc.) were based on atomic ratios and DBE limitations (Jiménez-Morillo et al., 2018; Kim et al., 2003) (see Figure S4.10). Figure 4.5A incorporates a presence-absence variable matrix prepared by assigning 1 to any value >1000. The data were aggregated into a contingency matrix based on boreal, forest-tundra, and tundra classifications. The difference between boreal and tundra was calculated to evaluate changing compositions for each molecular species.

2.4.2. Constraining environmental matrix preparation

An environmental data table of constraining variables was prepared for each lake (Table S4.1). There were 25 constraining environmental variables. Permafrost data were encoded to an ordinal scale according to permafrost continuity and the ground ice extent. Georegions and geodistricts were represented as binary variables with the exclusion of geodistrict Alberta Plateau and georegion Interior Plains to avoid collinearity. Historical downward thermal infrared (longwave) radiative fluxes, temperature, and rainfall conditions were collected from different databases, and the data points collected were from the time of core collection to reduce the effect of environmental changes (i.e. warming temperatures, changed precipitation, etc.) at a field site since the core was collected. C/N was

calculated from the TC/TN ratio. Missing values were estimated using the R packages Hmisc (*aregImpute* with nk = 0) (Harrell, 2019). R code is included in supplementary information.

Regional designations were determined by overlaying the sample locations in ArcGIS (V 10.6.1) with maps from open-source data repositories. For the treeline model, the tundra region was designated as north of the boreal region (Brandt, 2009), and the forest-tundra spans both regions (Ranson et al., 2011). Ecozones, ecodistricts, and ecoprovinces were defined according to the National Ecological Framework (NEF) by Agriculture and Agri-food Canada and Environment Canada (Marshall et al., 1999). The Global Geocryological Data System hosted by the National Snow and Ice Data Center designated the permafrost regions (Brown et al., 2002). Georegions and geodistricts were designated by the Geological Survey of Canada (Bostock, 2014).

The temperature and precipitation data were collected using an online database of Historical Climate Data hosted by the Government of Canada (http://climate.weather.gc.ca/historical_data/search_historic_ data_e.html). The mean daily temperature (°C) and average daily precipitation (mm) amounts were exported from the closest weather station (6 km to 270 km away from the site) for the year the sample was collected. In the case of missing data, missing values were estimated by averaging data from years before or after the missing information, and by referencing adjacent stations.

Lastly, downward thermal infrared (longwave) radiative flux ($kWh/m^2/day$) was determined for each site at the time of sample collection. The radiative flux data were collected from the NASA Langley Research Centre (LaRC) POWER project funded by the NASA Earth Science/Applied Science program (Stackhouse et al., 2018). The longwave radiative flux data are based on observational data from satellites, and the flux is calculated from radiative transfer models.

2.4.3. Statistical analysis

Unsupervised and supervised analyses were performed using *MetaboAnalystR* (Chong et al., 2019) in RStudio (R 3.6.0). Using *MetaboAnalystR* principal components analysis (PCA), partial least squares discriminant analysis (PLSDA), hierarchical clustering (HCA), and random forest (RF) were performed. *MetaboAnalystR* also allows cross validation (CV) and permutation tests of PLSDA plots to assess overfitting. The supervised analyses (PLSDA and RF) were constrained by regional designations (Figure 4.2). K means clustering on the variable data matrix was performed using the *kmeans* package. Lastly, *vegan* (Dixon, 2003) was used to perform the redundancy analysis (RDA) on a reduced environmental data table containing: longitude, latitude, TOC, TN, TS, C/N, temperature variables (mean July and January temperatures and mean/minimum/maximum annual temperatures), radiative flux variables (mean July and January fluxes and mean annual flux), and total precipitation. All ordination and clustering plots were drawn using the R packages: *tidyverse, ggplot2* (Wickham, 2011), *factoextra* (Kassambara and Mundt, 2017), *ggpubr* (Kassambara, 2019), and *viridis* (Garnier et al., 2018).

3. RESULTS AND DISCUSSION

3.1. Unsupervised Kmeans clustering of lakes in lake districts

Due to the molecular complexity of each lake sediment sample, the resulting sedimentomics data matrix consisted of a considerable number of pseudomolecular ion variables (representing bins of unique accurate mass (*m/z*: mass to charge ratio) and retention time (min) combinations). Only molecules compatible with the extraction technique and chromatography parameters used were

resolved by the instrument so the results were interpreted with the understanding that not all organic molecules present in the sample were accounted for in the data model.

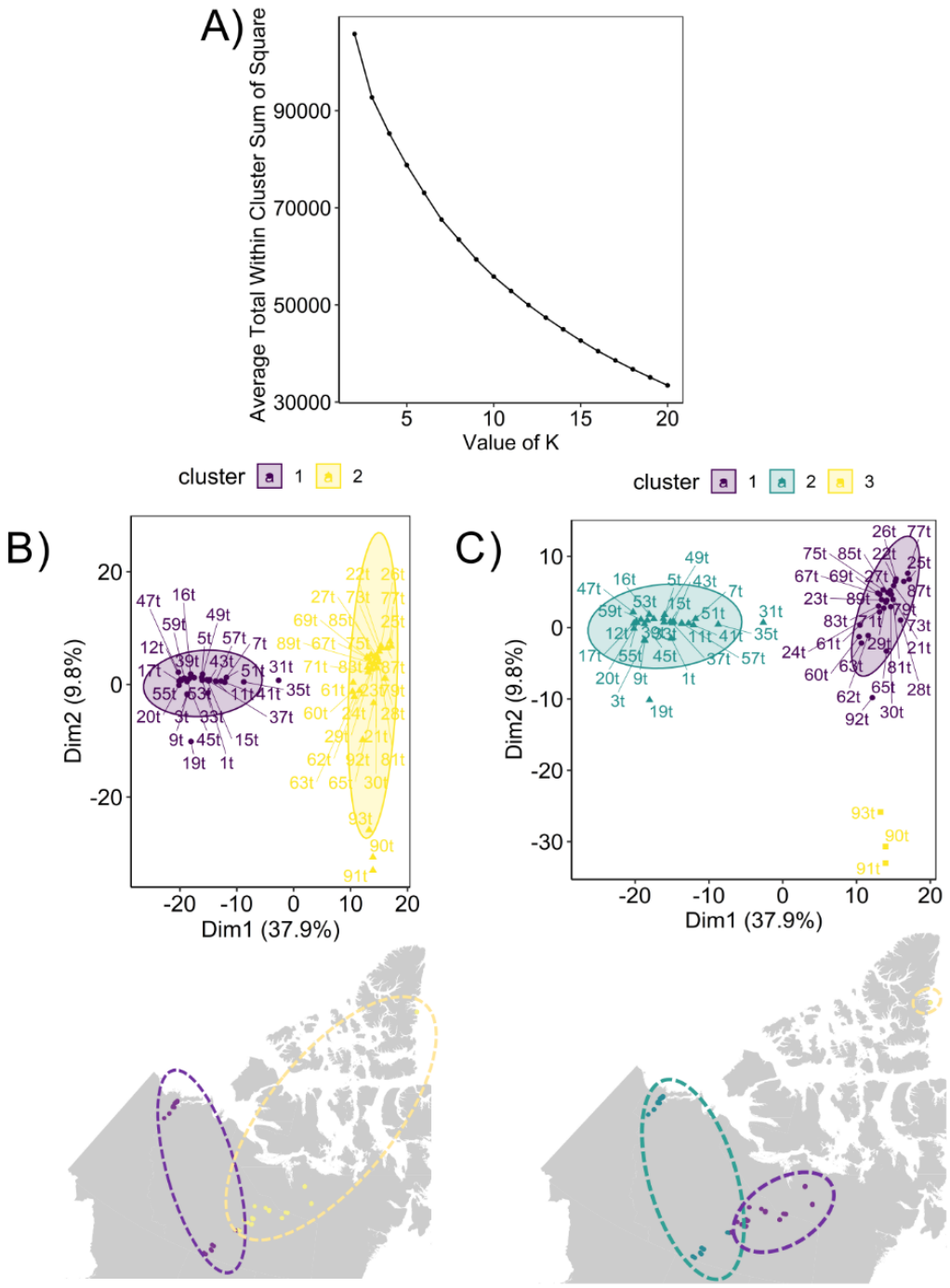


Figure 4.3: Unsupervised cluster analysis with kmeans. The figure contains A) elbow plot of Total Within Cluster Sum of Squares, B) kmeans cluster plot K = 2, and C) kmeans cluster plot K = 3.

Kmeans clustering of the data matrix was used as an unsupervised method to group sample sites together, to predict "lake districts" (Catalan et al., 2009; Kernan et al., 2009) based on sedOM composition prior to the application of additional constraints from overlying models and classification systems. An elbow plot displaying the Total Within Cluster Sum of Squares (WCSS) versus the value of K (i.e. the number of clusters; Figure 4.3A) was used to determine the optimal number of clusters. The arc of the elbow plot curve was gradual, with only a minor inflection point between K = 2 and K = 3, indicating that the optimal number of clusters for the model was likely two to three clusters. The gradual arc of the curve also indicated a high degree of variability within the dataset suggesting that the composition of organic C was spatially heterogeneous, which was observed when the parameter K was greater than three (K > 3; see Figure S4.9).

The two clusters in Figure 4.3B (K = 2) explain 47.7% of the total variance with 37.9% explained in the first dimension and 9.8% in the second dimension. The lakes in the two clusters include: 11t, 12t, 15t, 16t, 17t, 19t, 1t, 20t, 3t, 5t, 7t, 9t, 31t, 33t, 35t, 37t, 39t, 41t, 43t, 45t, 47t, 49t, 51t, 53t, 55t, 57t, 59t (Cluster 1), and 21t, 22t, 23t, 24t, 25t, 26t, 27t, 28t, 29t, 30t, 60t, 61t, 62t, 63t, 65t, 67t, 69t, 71t, 73t, 75t, 77t, 79t, 81t, 83t, 85t, 87t, 89t, 90t, 91t, 92t, 93t (Cluster 2). There were outliers (i.e. lake sediment samples outside the multivariate normal distribution) indicating they are farther from the true mean of the multivariate normal distribution estimated for that cluster. The outliers of cluster 1 were 19t and 31t, and the outliers of cluster 2 were 90t and 91t. However, when three clusters were specified (K = 3) then the outliers from cluster 2 formed their own "grouping" (the limited number of samples in this group prevented the calculation of an ellipse; Figure 4.3C); consisting of High Arctic lake sediments (with the exception of 92t). These 2–3 clusters could be likely "lake districts" defined by the composition

of organic C molecules from lake sediments. Unsupervised PCA and HCA (Figures S4.1A and S4.1B) corroborated these “lake districts” in their clustering patterns.

The main clusters in the K = 3 model consisted of lake sediment samples that generally span the Borderlands versus Shield geological bedrock paradigm or the Boreal versus Tundra paradigm. Cluster 1, or lake district 1, contains lakes is in the Borderlands/Boreal regions. Cluster 2, or lake district 2, consists of lakes in the Shield/tundra regions, while Grouping 3 could be considered a sub-group of lake district 2, or lake district 3, where the lakes are within the Shield/tundra regions but are located farther north in the High Arctic. Lake district 1 lakes in the Borderland regions are influenced by catchment primary productivity (as they are all within the treeline) and have areas with less permafrost (i.e. sporadic and discontinuous). Permafrost is known to control ecological processes and organic C stores in the North (Grosse et al., 2011; Jorgenson and Grosse, 2016; Schuur et al., 2015). Lake district 2 lakes are located in areas of continuous or discontinuous permafrost, in the forest-tundra or tundra regions. Lake district 3 contains only continuous permafrost, and is located in the High Arctic where the temperature is colder, the growing season shorter, and there is more exposed bedrock. Consequently, the lake sediments within the lake districts 1 and 2 share different and/or lower primary productivity properties than lake district 1 suggesting the sedOM compositions are strongly influenced by their catchments, and that unsupervised clustering predicts two to three “lake districts”.

3.2. Ecosystem classification systems

As sedimentomics datasets contain considerably more variables than samples, supervised multivariate models can be susceptible to overfitting (Gromski et al., 2015; Westerhuis et al., 2008). Overfitting occurs when the constraining information is too complicated for the dataset, and the model

begins to describe random noise or minor changes in variance. In order to detect overfitting cross-validation (CV) and permutation tests were performed with each PLSDA model (Westerhuis et al., 2008). Overfitting of PLSDA models occurs when $Q^2 < 0$, and when the permutation test reports p-values > 0.05. An orthogonal classification test, supervised random forest (RF), was also performed as this modelling algorithm has proven robust against overfitting (Gromski et al., 2015). The combination of PLSDA and RF supervised classifiers indicate which classification systems best fits the dataset (Table 4.1). In order to compare the best models, the PLSDA CV Q^2, R^2, and accuracy values were considered with the random forest out-of-bag (OOB) error. Ideal PLSDA models have the highest CV Q^2, R^2, and accuracy values. Q^2 is the value most frequently used to assess PLSDA model suitability as Q^2 represents a ratio between the prediction error and the total sum of squares (Szymańska et al., 2012). RF models are assessed based on the OOB error rate (i.e. the mean prediction error), where the smaller the error the more accurate the classifier is, and OOB error can be compared between different models of the same underlying data.

Table 4.1: Model fit estimates for PLSDA (%Variance, Q^2, R^2, and accuracy) and random forest models (Out-of-bag (OOB) error).

Model	# of Groups	%Var PLSDA	Q^2	R^2	Accuracy	Permutation p-value	RF OOB
Treeline	3	54.5	0.75499	0.84486	0.84667	0.00050	0.190
Ecozone	4	55.3	0.59659	0.75222	0.65517	0.00700	0.310
Ecoprovince	6	55.3	0.54972	0.70597	0.53448	0.00115	0.276
Ecoregion	9	53.0	0.12706	0.58554	0.43103	0.00250	0.328
Ecodistrict	11	55.5	0.56476	0.70502	0.41379	0.00050	0.483
Georegion	4	55.3	0.69915	0.80317	0.72414	0.00050	0.190
Geodistrict	7	54.5	0.20109	0.48415	0.65517	0.00050	0.310
Permafrost	4	52.8	0.21039	0.54648	0.67241	0.03000	0.345

In summary, when all the models and model parameters were compared (Table 4.1), the treeline model had the highest CV Q^2, R^2, and accuracy values. Both the treeline model and the georegions model share the lowest OOB error in the RF models, making georegions the next best model after the treeline. The third best model to describe the distribution of organic molecules in lake sediments was the ecozone model. The least suitable models were ecodistricts and permafrost, although neither of these models overfitted, they had the highest RF OOB errors. The three best models (i.e. treeline, georegions, and ecozones) are described further below, the remaining models are described in the supplementary information.

Van Krevelen plots were generated to illustrate compositional differences in organic C molecules from boreal and tundra lake sediments of the treeline model. Van Krevelen plots compare the *in silico* predicted molecular formulas for each observed pseudomolecular ion depending on set molecular criteria derived from all known primary and secondary metabolites (Jiménez-Morillo et al., 2018; Kim et al., 2003; Kind and Fiehn, 2007; Stubbins et al., 2010). The molecular formulas are partitioned into ratios comparing number of C, H, N, O, S, and P atoms. Van Krevelen plots often compare H/C (an indicator of the degree of chemical saturation) to O/C (an indicator of the degree of oxygenation). Based on H/C and O/C ratios, particular regions of the plot relate to different molecule classes (Figure S4.10). Figure 4.5A illustrates that a compositional difference of organic C molecules was observed to be associated with boreal and tundra lake sediments. Chlorin-type tetrapyrroles (such as 3-phorbinepropanoic acid and purpurin 18 methyl ester) and squalene derivatives dominated in boreal lake sediments, while biohopanoids (such as adenosylhopane and 35-aminobacteriohopane-31,32,33,34- tetrol) were predominant in the tundra lake sediments (Saleem et al., 2019).

3.2.1. Treeline regions

The treeline regions of the Canadian Arctic include: boreal (Brandt, 2009), forest-tundra (Ranson et al., 2011), and tundra. The boreal region represented continuous boreal forest. The forest-tundra is an ecotone (a transitional region between two ecozones) between the continuous boreal forest and tundra, and is a complex mixture of both (Payette et al., 2001). The tundra region represented all exposed land north of the treeline (Figure 4.2A). The lakes used in this study were classified based on their location and field reports. Lakes that were ambiguous (i.e. a mix of boreal and tundra characteristics) were considered to be forest-tundra. In this model the boreal classification contained 12 lakes, the forest-tundra ecotone had 25 lakes, and tundra contained 21 lakes.

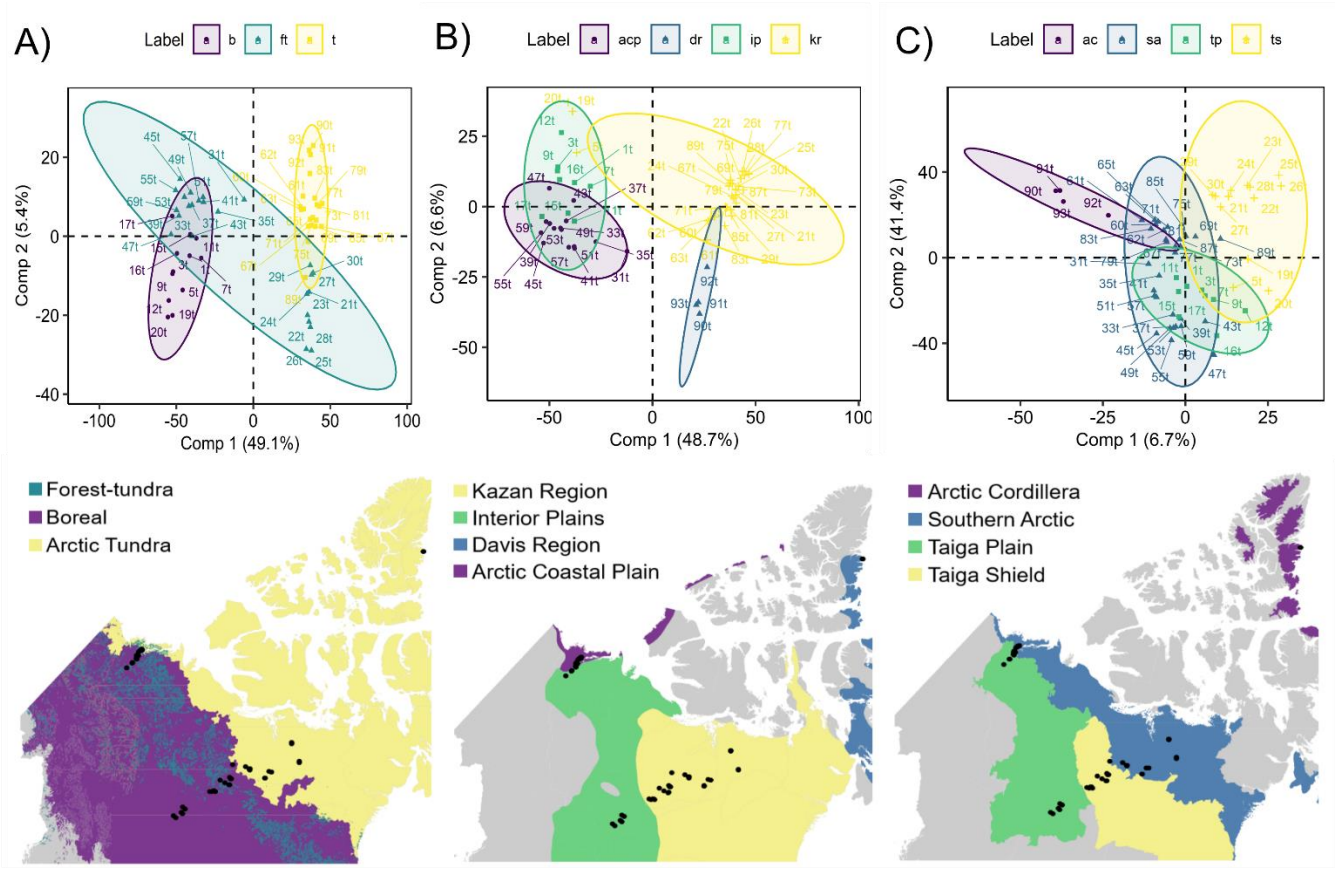


Figure 4.4: Supervised PLSDA scores plots and regional maps with spatial classification systems for the A) Treeline region, B) Georegions, and C) Ecozones models.

The PLSDA (Figure 4.4A) for the treeline model was the most robust and did not overfit. The PLSDA validation tests reported a Q^2 value of 0.75499 in the first two components, an accuracy of 0.84667, and an R^2 value of 0.84486 from the CV analysis and the permutation test reported a p-value of 0.0005 (Table 4.1). The random forest classification reported an OOB error of 0.190. In the treeline PLSDA model, boreal and tundra lake sediment groups clearly separated along the first principal component and explained 49.1% of the variance in the data model. Forest-tundra lake sediments separated, either clustering within the boreal group or the tundra group.

Lake sediments originating from across a boreal to tundra gradient have very distinct biota and thus sedimentary organic matter compositions. Boreal lakes are known to have higher DOM content than tundra lakes because boreal lakes are typically more productive (i.e. more biological activity) than tundra lakes (Rühland et al., 2003). Boreal catchments are also expected to have more primary production and since allochthonous organic matter often dominates organic matter deposition (Xu et al., 2019, 2015), the higher rates of catchment primary production and decomposition result in increased DOM abundance, and thus sedOM abundance between boreal and tundra lakes. Treeline regions not only demarcates a large shift in primary productivity, but also a major species shift. Boreal vegetation consists of coniferous forests (i.e. larch, spruce, firs, pines, etc.), while the tundra vegetation consists of shrubs, lichens, saxifrage, and arctic poppies, etc. The microbial community also varies across this ecological boundary in terms of functional beta diversity and taxonomic beta diversity (Fierer et al., 2012). The shift in species would suggest there would also be a compositional difference in the organic molecules that are deposited into the lake sediments as allochthonous organic material is thought to be more recalcitrant than autochthonous organic material (Gudasz et al., 2015; MarínSpiotta et al., 2014); although, recent research suggests microbial necromass, is a larger component of SOM and sedOM than

previously predicted (Lehmann and Kleber, 2015; Marín-Spiotta et al., 2014; Schimel and Schaeffer, 2012; Schmidt et al., 2011; Simpson et al., 2007). Thus, the combination of different productivity rates, organic input compositions, and microbial taxonomic and functional diversity play a significant role in the compositional diversity of organic molecules in lake sediments across the treeline (Kellerman et al., 2014). Compositional differences in sedOM can be used to discover novel paleolimnological biomarkers and play an important role in historical treeline reconstructions from sediment (Saleem et al., 2019).

3.2.2. Georegions

Georegions are defined by the geological bedrock underlying Canada. There are two major divisions, the Shield and the Borderlands (Bostock, 2014). The Shield consists of old Precambrian crystalline rock, which is surrounded by younger geology found in the Borderlands. Further divisions of the Shield and Borderland regions are based on topography and differences in unique geology (Figure 4.2F). The Arctic Coastal Plain (acp; n = 15) and the Interior Plains (ip; n = 9) are in the Borderlands, and the Davis Region (dr; n = 4) and Kazan Region (kr; n = 30) are in Shield Region.

The georegion PLSDA model explained 55.3% of the total variance in the first two PCs (Q^2 0.69915; Accuracy 0.72414; R^2 0.80317; Figure 4.4B). The permutation value was 0.0005 so there was no over-fitting. The RF OOB error rate was 0.190. The clustering for the Georegions model (Figure 4.4B) subdivided the lake sediments into the two broader underlying geologies Borderlands (i.e. Arctic Coastal Plains (acp) and Interior Plains (ip)) versus Shield (i.e. Kazan Region (kr) and Davis Region (dr)), which agrees with the unsupervised kmeans “lake districts”.

Geological conditions are known to be important controls on C cycling and microbial dynamics (Heckman et al., 2009). Abiotic conditions such as soil/sediment pH, exchangeable cations, aggregate

formation, and adsorption influence microbial communities and organic diagenesis (Heckman et al., 2009; Marín-Spiotta et al., 2014; Swenson et al., 2015a), so it is not surprising the underlying geology influences the distribution of organic molecules in the sediment.

There are three outliers affecting the Georegions model, which are the lake sediments from lakes 5t, 19t, and 20t. These lakes are close to Yellowknife (Northwest Territories), and they are all located in the Kazan Region on Shield bedrock but are not within the same cluster as neighboring lake sediments (i.e. 23t and 24t). These lakes were sampled at different times. Lakes 5t and 19t were sampled in 2012, 20t was sampled in 2013, and lakes 23t and 24t were sampled in 1998. The area around Yellowknife has changed between 1998 and 2012/2013. Regional climate changes could explain why these lakes are outliers. Previous literature described evidence of lake expansion in the Mackenzie Bison Sanctuary within the Great Slave lake area over the last 25 years (Korosi et al., 2017). In addition, the regional annual mean air temperature has increased by 0.6 °C per decade since 1970's (i.e. +0.9 °C since 1998 to 2013) (Wolfe et al., 2017), and since the area is underlain by discontinuous permafrost, permafrost thaw is expected to significantly alter boreal forest margins and wetlands, which in turn would impact the compositional patterns of the sedOM in these lakes (Morse et al., 2016; Wolfe et al., 2017).

3.2.3. Ecozones

The ecozones PLSDA (Figure 4.4C) described 55.3% of the variance of which 6.6% was explained in the first component, and 48.7% in the second. The PLSDA validation tests reported a Q^2 value of 0.59659 in the first two components, an R^2 of 0.75222, and an accuracy of 0.65517 from the CV analysis and the permutation test reported a p-value of 0.007 (Table 4.1); therefore the PLDA model did not

overfit. The random forest classification reported an OOB error of 0.310. Compared to the other two previous models, obvious clusters were less apparent (Figure 4.4C). The two most distinct clusters in the ecozone PLSDA model were the Arctic Cordillera (ac; n = 4) (defined largely in the first component), and the Taiga Shield (ts; n = 13; with the same samples 5t, 19t, and 20t separating from the main cluster). The Southern Arctic (sa; n = 32) sediment samples were varied and failed to form a distinct cluster, but rather grouped with the Taiga Plains (tp; n = 9), and between the Arctic Cordillera and Taiga Shield. Surprisingly, the ecozone PLSDA model did not perform as well as the Treeline and Georegions PLSDA models, despite the ecozones classifications considering both abiotic (i.e. elevation, landform, surface form, surface material, temperature, precipitation, windspeed, solar radiation, etc.) and biotic factors (i.e. land cover, soil development, growing season, etc.) (Marshall et al., 1999).

3.2.4. Correlation between geographical and ordinational spatial adjacencies

The distribution of molecular organic C in lake sediments shows a strong relationship between regional/zonal geographical adjacency and adjacency in the coordinate space of PLSDA models. The farther the regions are from each other the less likely the 95% confidence intervals overlap in the PLSDA models, and this is demonstrated in the three best models: treeline, georegions, and ecozones. In Figure 4.4A, the boreal and tundra regions are separated in geographical space and in PLSDA coordinate space, but the forest-tundra, which is adjacent to the boreal and the tundra, has an ellipse in the PLSDA that overlaps both boreal and tundra ellipses. This correlation between geographical and mathematical spatial adjacencies is also reflected in the georegions model where if the PLSDA score plot was flipped, rotated, and then overlaid on the map in Figure 4.4B, the georegions and the ellipses share the same distribution. For example, the Davis region is closest to the Kazan region in geographical space, and

these two ellipses overlap on the PLSDA, but the Davis region ellipse does not overlap with either the Arctic Coastal Plain or the Interior Plains – regions that are farthest from the Davis region. The Kazan region also overlaps with the Interior Plains, but not the Arctic Coastal Plain because the Kazan region is not adjacent to the Arctic Coastal Plain. The same pattern of spatial adjacencies between the PLSDA coordinate space and the geographical space can also be seen in the ecozones model (Figure 4.4C). The relationship between the adjacency in coordinate PLSDA space and the geographical orientation of the different regions/zones provides strong evidence that the molecular composition of organic C in lake sediments is spatially diverse and that they can be used to define different regions across space. It would be interesting to study if this relationship is maintained with a greater number of samples from other neighboring regions/zones, and if regional biomarkers could be determined using this approach.

3.3. Environmental constraints

The organic molecular C composition in sediment is dependent on many different variables and represents a complex mixture of organic matter from animal, microbial, vegetation, and other sources. In addition, diagenesis - the transformation and burial of organic matter in the sediment, is dependent on both abiotic and biotic conditions. In metaphenomics, environmental variables (i.e. temperature, solar radiation, etc.) are modelled with “omics” data to determine how environmental constraints affect variation in the “omics” dataset. Thus, an RDA was used to determine which environmental constraints influenced the distribution of organic molecules in lake sediments. An RDA model is a combination of PCA and multiple linear regression where two matrices, the response matrix (i.e. samples x bins) and an explanatory matrix (i.e. samples x environmental constraint variables), are used to model how changes in the response matrix relate to changes in the environmental variables for the same set of samples. A

total of 25 constraining environment variables were initially selected for RDA modelling. Georegions and geodistricts as binary variables were included to account for geochemical parameters and landform characteristics. The treeline model parameters were not included because the factors influencing treeline dynamics were represented directly by the constraining environmental variables (i.e. radiative flux, TOC, temperature, and precipitation). Permafrost information was included as the treeline is directly influenced by permafrost conditions (Morse et al., 2016; Wolfe et al., 2017), and therefore represents another geological parameter.

Table 4.2: RDA model fit parameters for the full model after redundancy analysis, the moderate model and the minimum model.

	Models		
	A	B	C
Parameters			
# of Environmental Variables	16	12	9
%Variance RDA1	41.99	41.49	39.20
%Variance RDA2	6.74	6.62	6.55
Total % Variance Explained	61.30	57.25	52.42
p-value	0.001*	0.001*	0.001*
Environmental Variables			
Permafrost (ordinal scale)	0.426		
%TOC	0.048*	0.055^	0.043*
%N	0.051^	0.070^	
%S	0.018*	0.010*	0.015*
C/N	0.224		
Mean January Radiative Flux	0.678		
Mean July Radiative Flux	0.001*	0.001*	0.001*
Mean Annual Radiative Flux	0.001*	0.001*	0.001*
Mean July Temperature	0.001*	0.001*	0.001*
Total Precipitation	0.004*	0.003*	0.001*
Geology Division (ap)	0.224		
Geology Division (bl)	0.054^	0.051^	
Geology Division (bsu)	0.001*	0.001*	0.001*
Geology Division (dh)	0.001*	0.001*	0.001*
Geology Division (gsp)	0.076^	0.056^	
Geology Division (md)	0.001*	0.001*	0.001*

^Near significant variables (0.1 > p-value > 0.05)
**Significant variables (p-value < 0.005)*

A *Hmisc* redundancy analysis (*redun*) was performed to remove collinear variables. Linearity was assumed and an R^2 cutoff of 0.95 was applied (where correlating variables with an $R^2 > 0.95$ were removed); after which 16 variables remained (permafrost_ordinal_scale, percent_c, TN, TS, C/N, mean January radiative flux, mean July radiative flux, mean annual radiative flux, mean July temperature, total precipitation, and the geological divisions - Anderson Plain, Back Lowland, BearSlave Upland, Davis Highland, Great Slave Plain, and Mackenzie Delta). The RDA was performed using all 16 remaining variables and designated as the full model (Model A; Figure S4.11A). Model results are presented in Table 4.2. The full model was significant (p-value < 0.05), but the *envfit* function revealed insignificant variables. Insignificant variables (p-value > 0.1; permafrost, C/N, mean January radiative flux, and the geological division Anderson Plain) were removed, leaving the near significant (0.1 > p-value > 0.05) and significant (p-value < 0.05) variables in the model. Model B (Figure S4.11B) consisted of the remaining 12 variables and was found to be significant (p-value < 0.05). The total variance of Model B decreased compared to Model A, as the number of constraining variables between models also decreased. To generate the minimum adequate model (Model C; Figure 4.5B), all insignificant variables were removed, with the exception of TOC because the total explained variance in the final model decreased if TOC was absent (as removal of the other variables caused TOC to become more significant to the model). Model C contained 9 environmental variables (TOC, TS, mean July radiative flux, mean annual radiative flux, mean July temperature, total precipitation, and only three geological divisions (Bear-Slave Uplands, Davis Highlands, and the Mackenzie Delta)) and explained 52.42% of the total variance in the first two components (RDA1 39.20% and RDA2 6.55%).

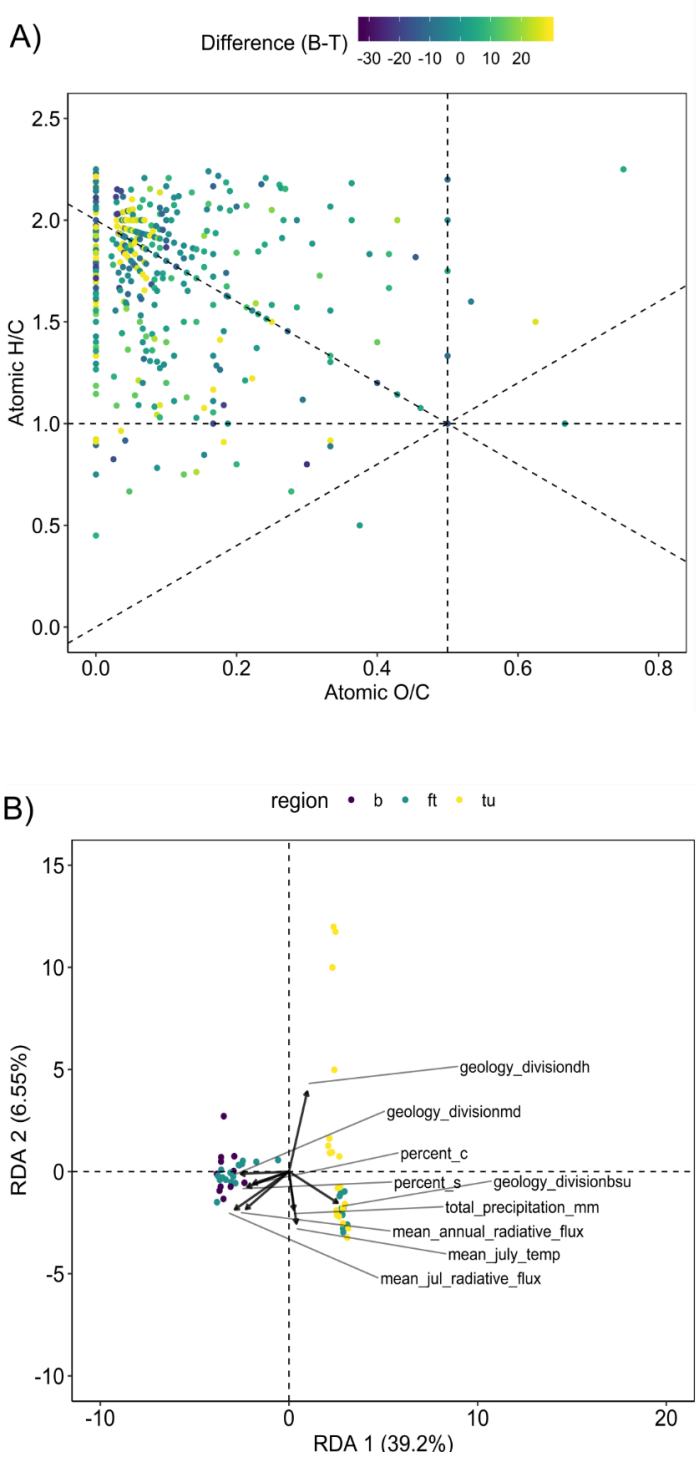


Figure 4.5: A) Van Krevelen plot where yellow indicates increased presence (or exclusive presence) in boreal sedOM, purple indicates increased presence (or exclusive presence) in tundra sedOM, while green shading to blue indicates it is present in both regions but to varying amounts. B) RDA biplot of the Minimum Adequate Model (Model C) overlaying the treeline classifications (boreal = purple, forest-tundra = green, and yellow = tundra).

The minimum adequate model, Model C, demonstrated interesting trends when overlaid with lake assignments from the most robust PLSDA ecosystem classification model - the treeline. In general, environmental variables such as mean radiative flux, mean annual radiative flux, TS, TOC, and the geological division Mackenzie Delta describe the cluster containing the boreal (and some forest-tundra) lake sediments. Environmental variables such as the geological divisions (Bear-Slave Uplands and Davis Highlands) describe the clustering containing the tundra lake sediments (and some forest-tundra). Mean July temperature and total precipitation seem to explain variation along the second RDA axis indicating the large separation between the tundra lakes could be explained by these two parameters. The lake sediments located at the bottom of the RDA2 axis were the lake sediments sampled from the High Arctic (93t, 91t, and 90t).

Climate has a large influence on aquatic ecosystem structure and function (Fritz and Anderson, 2013), and also influences the growth and distribution of vegetation. Arguably, climatic factors will also influence sedOM deposition, composition, and mineralization because of the climate's strong influence on vegetation, aquatic productivity, and microbial communities (Beale et al., 2018; Fierer et al., 2012; Gudasz et al., 2015; Johnston et al., 2019; Swenson et al., 2018). Previous results using particulate, humus, and resistant organic C fractions in soil from across the continent of Australia agreed with this study as they also correlated with climate factors such as mean annual temperature and precipitation, potential evapotranspiration, and solar radiation (Rossel et al., 2019). Thus, the climatic conditions used in the RDA model, such as radiative flux, mean July temperature, and total precipitation, are significant and will influence the composition of organic C in lake sediments.

Geological factors affect soil/sediment pH, exchangeable cations, aggregate formation and thus C cycling through microbial community demographics (Heckman et al., 2009; Marín-Spiotta et al., 2014;

Swenson et al., 2015a). Geology also influences permafrost type (i.e. overburden) and other hydrological dynamics (Jorgenson and Grosse, 2016), which in turn affects organic matter distribution in rivers during freeze-thaw cycles or permafrost loss (Grewer et al., 2015). The most important geological regions for Model C were the Bear-Slave Uplands (bsu), Davis Highlands (dh), and the Mackenzie Delta (md), which corroborates well with the "lake district" unsupervised kmeans clustering (when K = 3) as the Bear-Slave Uplands contained most of the lakes located in the Shield lake district 2, except for the High Arctic lakes (minus 92t), which are present in the Davis Highlands (the other significant geological division) – lake district 3. The Mackenzie Delta lakes represent the Borderlands lake district 1 and were most significantly different from the Davis Highlands and Bear-Slave Uplands sediments, suggesting that geological controls on sediment organic C are more influential than the biotic in more arctic regions.

Geochemical factors such as TS and TOC were found to influence the spatial distribution of organic C composition in the lake sediments (Table 4.2). Surprisingly, TOC was not significant in Models A and B as boreal lake sediments are expected to have higher TOC than the tundra lake sediments (Rühland et al., 2003; Sobek et al., 2007). TOC is a bulk measurement and provides no further information regarding organic molecular composition in lake sediments, but organic C composition of soil and sediment is expected to vary over space as proven by the supervised and unsupervised models, the Van Krevelen plot, and also from previous literature (Rossel et al., 2019) and many environmental factors influence the composition (Marín-Spiotta et al., 2014; Rossel et al., 2019). A landmark study combined sedimentomics with genomics to document changes in DOM chemistry over 50 days (Wu et al., 2018). Large scale changes in the TOC value were absent but large compositional differences in the organic C were observed, demonstrating that TOC has only a limited use as a descriptor of lake sediment and that the molecular composition of organic C sediments is far more informative. In addition, previous

studies found that inclusion of bulk TOC or TN measurements failed to improve their riverine hyporheic sedOM models, rather, the molecular organic C compositions of the sediment were more useful predictors of sediment function (Graham et al., 2018).

The other geochemical factor significant to the model was TS; where sedimentary TS was higher in the boreal than the tundra regions. Lake sulphate concentrations depend on the bedrock (i.e. where limestone– dolomite, sandstone, siltstone, and conglomerate prevail), and are higher in boreal versus tundra ecozones (Pienitz et al., 1997a; Rühland et al., 2003). The influence of sulfur on organic C cycling is understudied in lake sediments. Sulfur is spatially dynamic and influences the anoxic mineralization of organic C (Holmer and Storkholm, 2001). Here, our model suggests TS influences the distribution of organic C and should be considered in future sedOM models.

4. CONCLUSIONS

We hypothesized northern Canadian lake sedOM compositions vary according to different environmental conditions. For instance, we expected lakes with similar environmental conditions will share similar sedOM compositions (i.e. the lake district concept). To test this hypothesis, unsupervised (kmeans) and supervised (i.e. PLSDA and random forest) multivariate classification analyses were used to cluster lake sediment samples based on sedOM composition, to determine if current models of land classification reflected “lake districts” across Canada's North. In a metaphenomic approach, an RDA was performed on the sedimentomics dataset to determine which factors were the most influential on the distribution of organic C molecules in lake sediments over a large spatial area.

There is substantial correlation between the treeline and georegions model with the 2–3 “lake districts” suggested by this dataset. The large productivity shift between the boreal and tundra or

Borderland and Shield regions were the drivers responsible for the majority of the model variance. However, higher order distinctions of “lake districts” were poorly resolved (i.e. geodistricts, ecoprovinces, ecodistricts, etc.), as the limited sample size and the uneven sample distribution reduced the potential resolving power of the spatial models; likely, smaller “lake districts” only become apparent when comparing within a smaller geographical space such as single region or ecozone. Unfortunately, due to the limited size and uneven sample distribution of this dataset, we are unable to draw strong conclusions regarding the confounding influence of productivity across the subregions. Improved sedimentomics analytical strategies (i.e. merged datasets from multiple parallel or sequential extractions (Tfaily et al., 2017) using appropriate quality controls (Bell and Blais, 2019)) and collaborative sampling efforts would provide for a more comprehensive analysis and potentially increase the resolving power of the molecular organic C spatial models.

The purpose of attempting to define “lake districts” in Canada's North is multi-faceted. Lake sediments are the ideal “record” of catchment vegetation, aquatic productivity, and microbial activity. By using sedimentomics to analyze the sedOM compositions, researchers are no longer just monitoring a single bulk value (i.e. TOC), but actual molecules. The molecular compositions can then help define regional classification systems, which is important for partitioning spatial variability for legislation (Jorgenson and Grosse, 2016), and when paired with lake chemistry, could be used to monitor lakes for climate change. The Arctic Coastal Plain in Alaska has already been impacted by climate change (Lara et al., 2018). Permafrost is expected to release large stores of organic C, some of which will drain into lakes (Grewer et al., 2018; Grosse et al., 2011; Plaza et al., 2019; Schuur et al., 2015). Permafrost loss and other climate changes will also impact boreal forest margins and vegetation boundaries (Carpino et al., 2018; Myers-Smith and Hik, 2018). Previous statistical modelling predicted half of the current vegetative

areas will shift vegetation types, and woody plant abundance will increase by >50% (Pearson et al., 2013). Rising Arctic temperatures will also impact microbial diversity and nutrient cycling, although the full effect of this has yet to be determined (Colby et al., 2019). Additionally, there are numerous other key indicators of Arctic climate change: increased movement of organic C and nutrients, changing pollination/ flowering periods, increased plant vulnerability to insects, increased shrub density, more wildfires, increased CO_2 uptake but also increased CO_2 emissions in the winter, more C cycling, and shifting animal demographics (Box et al., 2019). All of these key indicators have the potential to influence organic C deposition into lake sediments and change ecological boundaries. Thus, sedimentomics could be used to monitor these changes over time and monitor how the "lake district" defined ecological boundaries shift as well.

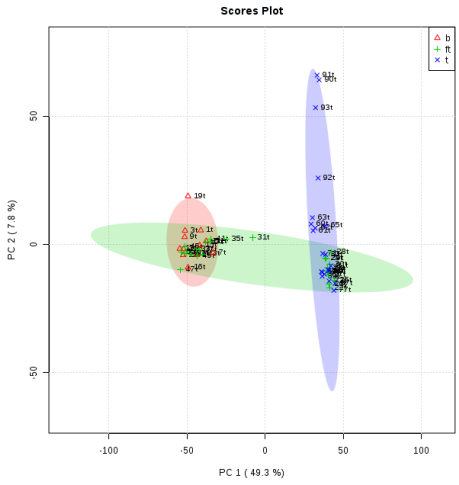
This paper demonstrates the versatility of "omics" untargeted analyses when combined with machine learning algorithms to study the distribution of sedimentary organic C molecules across northern Canada. Sedimentomics is a burgeoning field and has the potential to provide novel information about the dynamics of organic C across space. This study presents a novel method to classify lakes across large areas that could be integrated with current land cover classifications/mapping systems and also to define baselines for the molecular organic C compositions across space. Baseline information is also useful for recording molecular organic C compositions as the climate changes. We also used sediment metaphenomics, where we demonstrated the organic compositional data is influenced by the same values constraining the commonly applied bulk TOC measurements. Here, we have "opened the black box" by demonstrating that untargeted analyses not only provide similar utility to that of TOC, but also add a wealth of information and a more a detailed view of sedOM.

ACKNOWLEDGEMENTS

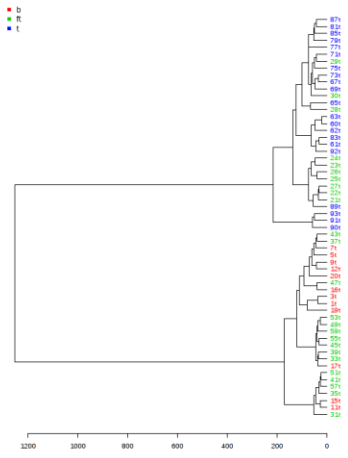
Funding for this study was provided by a Natural Sciences and Engineering Research Council of Canada (RGPIN 217112-2013 and RGPIN- 2018-04248) and Northern Supplement (RGPNS 444180-2013) grants to JMB, and grants from the Polar Continental Shelf Program to JMB and J. Smol. We thank J. Smol for providing sediment samples from Cape Herschel (Ellesmere Island) and the TK region. We also thank J. Thienpont, J. Korosi, D. Eickmeyer, L. Kimpe, and K. Rühland for sample collection and J. Korosi, A. Saleem, and L. Kimpe for method development.

SUPPLEMENTARY INFORMATION

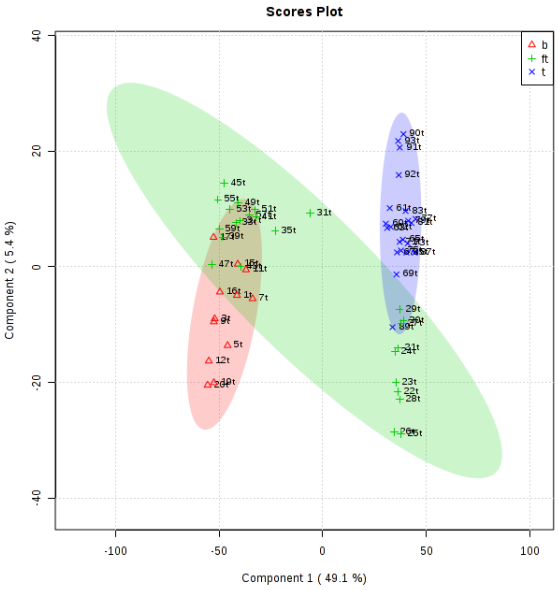
Treeline model



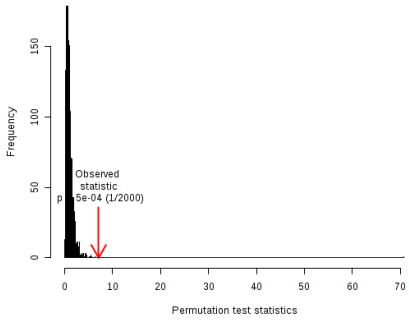
A)



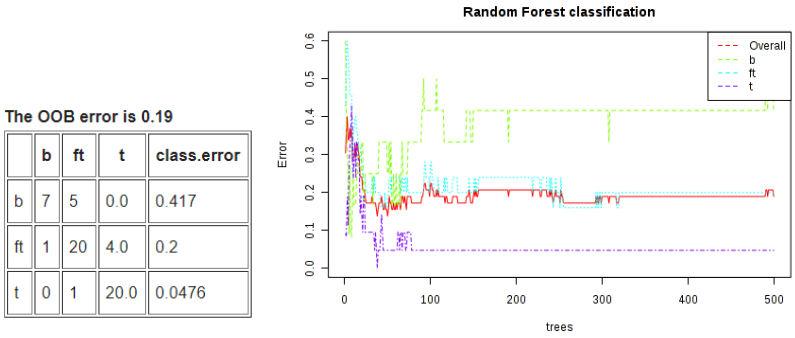
B)



C)



D)



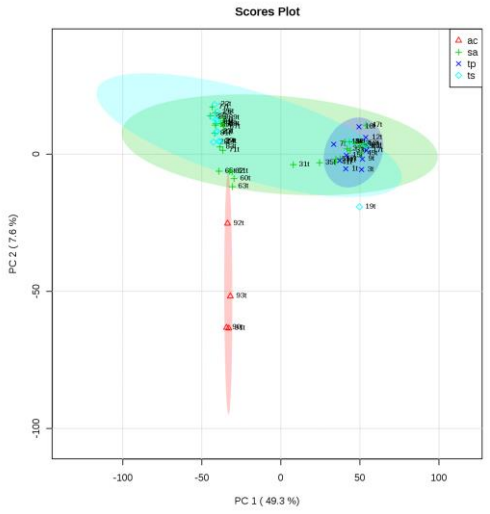
	b	ft	t	class.error
b	7	5	0.0	0.417
ft	1	20	4.0	0.2
t	0	1	20.0	0.0476

E)

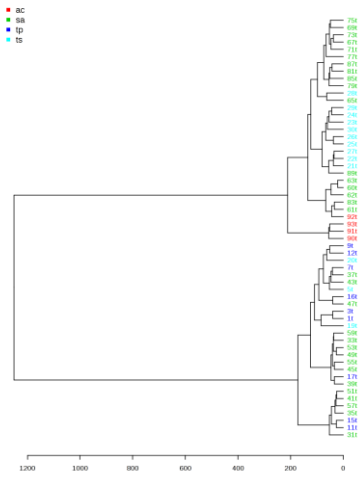
Figure S4.1: Treeline regions A) PCA, B) Dendrogram, C) PLSDA, D) PLSDA Permutation plot, and E) Random Forest classification plot. Group names are b = boreal forest, ft = Forest-tundra, t = tundra. All figures generated with MetaboAnalyst.

Ecozones model

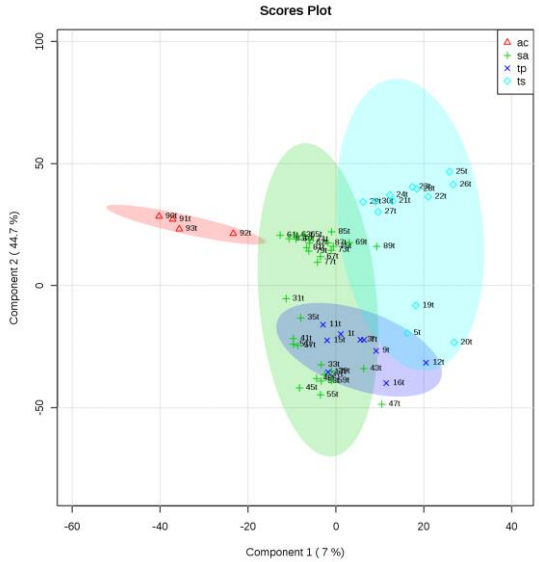
A)



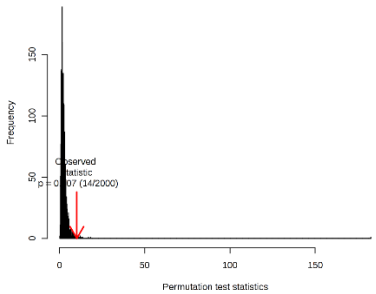
B)



C)



D)



E)

The OOB error is 0.31

	ac	sa	tp	ts	class.error
ac	3	1	0.0	0.0	0.25
sa	0	29	2.0	1.0	0.0938
tp	0	6	2.0	1.0	0.778
ts	0	4	3.0	6.0	0.538

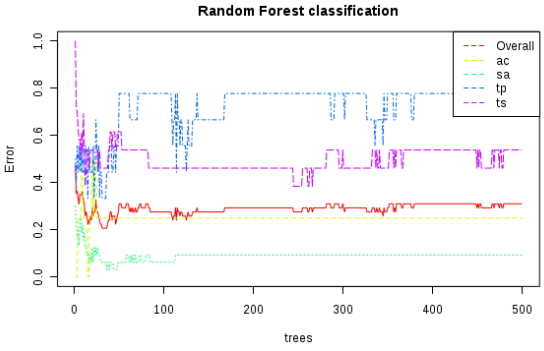
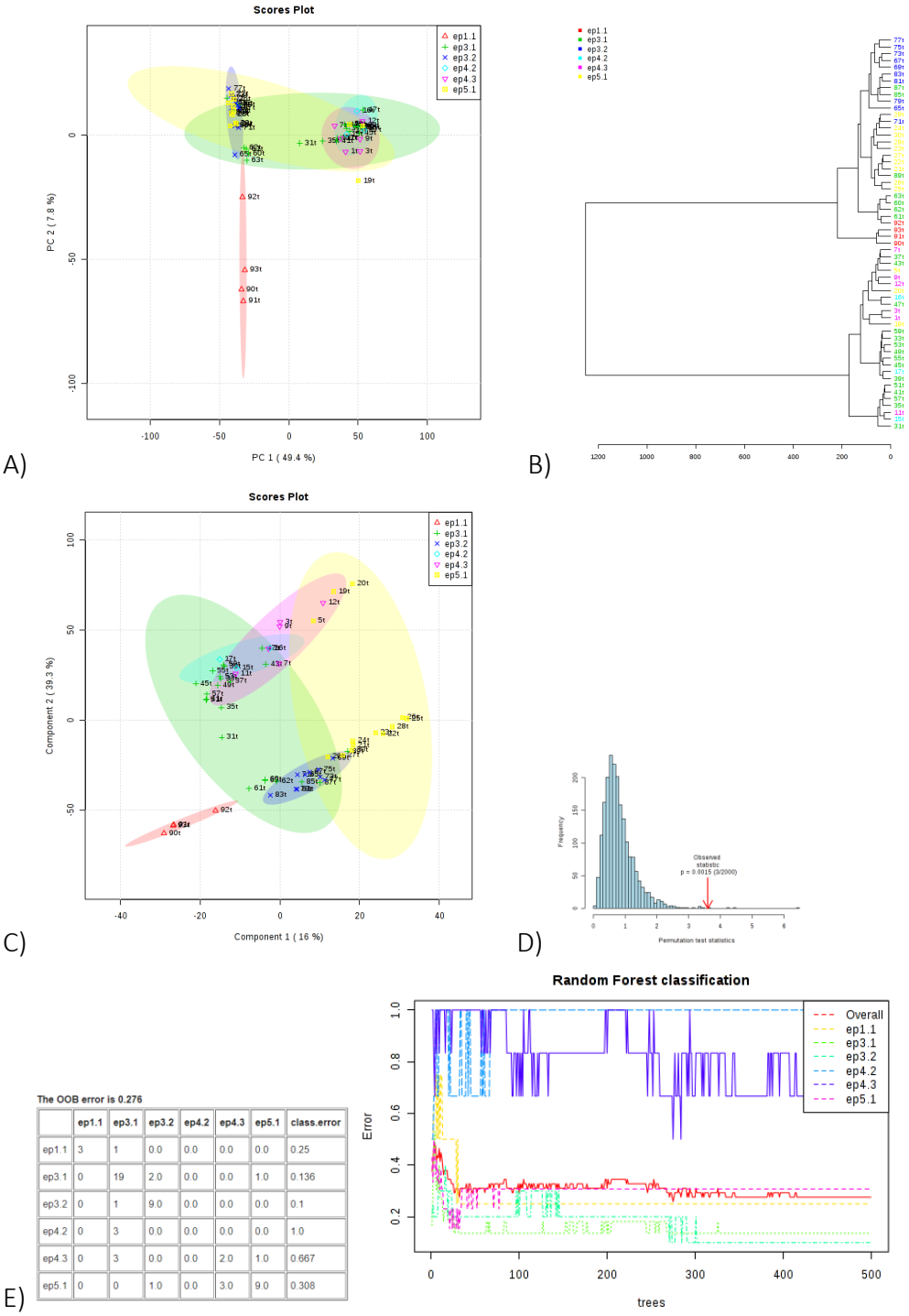


Figure S4.2: Ecozones A) PCA, B) Dendrogram, C) PLSDA, D) PLSDA Permutation plot, and E) Random Forest classification plot. Group names are ac = Arctic Cordillera, sa = Southern Arctic, tp = Taiga Plains, and ts = Taiga Shield. All figures generated with MetaboAnalyst.

Ecoprovinces model



	ep1.1	ep3.1	ep3.2	ep4.2	ep4.3	ep5.1	class.error
ep1.1	3	1	0.0	0.0	0.0	0.0	0.25
ep3.1	0	19	2.0	0.0	0.0	1.0	0.136
ep3.2	0	1	9.0	0.0	0.0	0.0	0.1
ep4.2	0	3	0.0	0.0	0.0	0.0	1.0
ep4.3	0	3	0.0	0.0	2.0	1.0	0.667
ep5.1	0	0	1.0	0.0	3.0	9.0	0.308

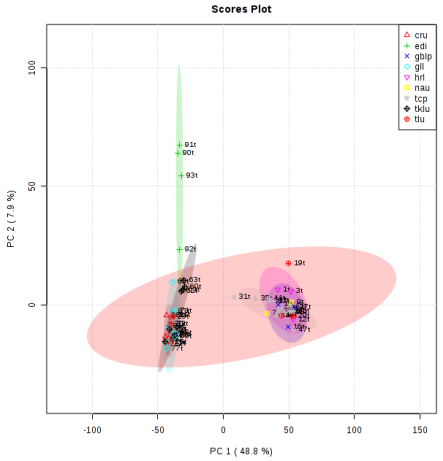
E)

Figure S4.3: Ecoprovinces A) PCA, B) Dendrogram, C) PLSDA, D) PLSDA Permutation plot, and E) Random Forest classification plot. All figures generated with MetaboAnalyst.

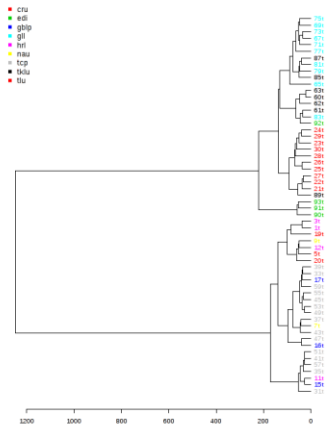
The ecoprovince designation was also defined in the National Ecological Framework (Marshall et al., 1999). Ecoprovinces are a subdivision of ecozones based on structural/surface forms (i.e. dissected, hummocky, inclined, level, terraced, etc.), the realms of terrestrial fauna and vegetation, hydrology, soil (i.e. cryosol, brunisol, podzol, rock/ice, etc.), and regional climates. Ecoprovinces were designed with conservation and resource needs in mind to make it easier to address environmental concerns across Canada, Mexico, and the USA. There are 53 ecoprovinces in Canada; only 6 of which are relevant to our sampling distribution (Figure 2C). The relevant ecoprovinces are: 1.1 (n=4), 3.1 (n=22), 3.2 (n=10), 4.2 (n=3), 4.3 (n=6), and 5.1 (n=13).

The ecozones PLSDA (Figure S4.3C) described 55.3% of the variance. 16% was explained in the first component, and 39.3% in the second. The PLSDA CV Q^2 value was 0.54972, an accuracy of 0.53448, and a an R^2 of 0.70597 in the first two components (Table 4.1). There are three broad clusters of points and these clusters consist of ecoprovinces 3.2 in cluster 1, cluster 2 contains 4.3 and 4.2, and cluster 3 has 1.1. The ecoprovinces 3.1 and 5.1 are spread across clusters 2 and 3. The permutation test reported a p-value of 0.00115. Thus, there is no overfitting in the ecoprovinces PLSDA model. The random forest classification reported an OOB error of 0.276 (Figure S4.3E).

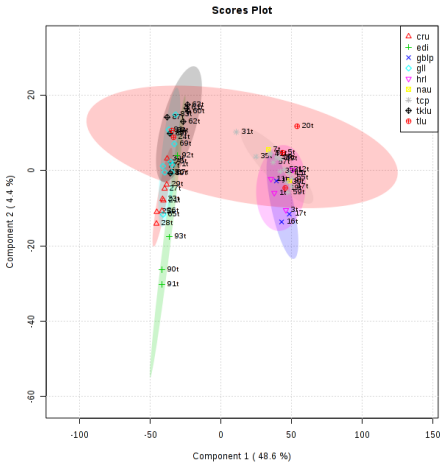
Ecoregions model



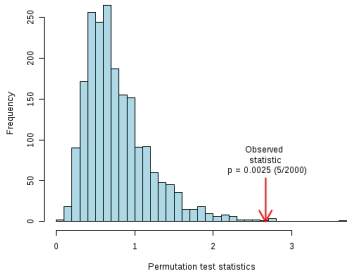
A)



B)



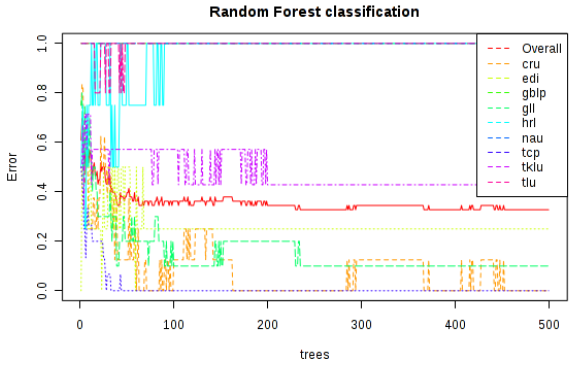
C)



D)

The OOB error is 0.328

	cru	edi	gblp	gll	hrl	nau	tcp	tklu	tlu	class.error
cru	8	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
edi	0	3	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.25
gblp	0	0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	1.0
gll	0	0	0.0	9.0	0.0	0.0	0.0	1.0	0.0	0.1
hrl	0	0	0.0	0.0	0.0	0.0	3.0	0.0	1.0	1.0
nau	0	0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	1.0
tcp	0	0	0.0	0.0	0.0	0.0	15.0	0.0	0.0	0.0
tklu	1	0	0.0	2.0	0.0	0.0	0.0	4.0	0.0	0.429
tlu	2	0	0.0	0.0	1.0	0.0	2.0	0.0	0.0	1.0



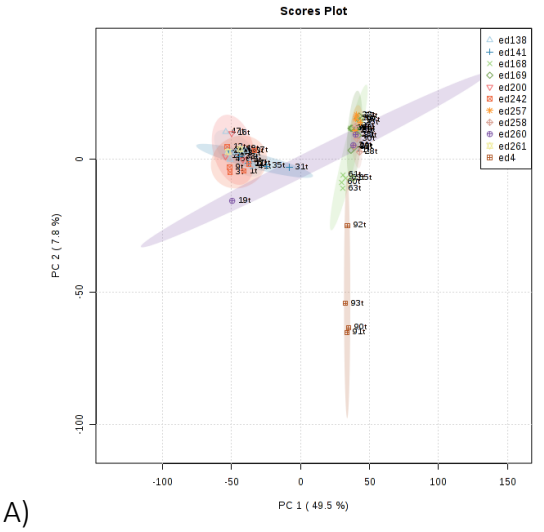
E)

Figure S4.4: Ecoregions A) PCA, B) Dendrogram, C) PLSDA, D) PLSDA Permutation plot, and E) Random Forest classification plot. Group names are cru = Coppermine River Upland, edi = Ellesmere and Devon Ice Caps, gblp = Great Bear Lake Plain, gll = Garry Lake Lowland, hrl = Hay River Lowland, nau = Northern Alberta Uplands, tcp = Tuktoyuktuk Coastal Plain, tklu = Takijua Lake Upland, tlu = Tazin Lake Upland. All figures generated with MetaboAnalyst.

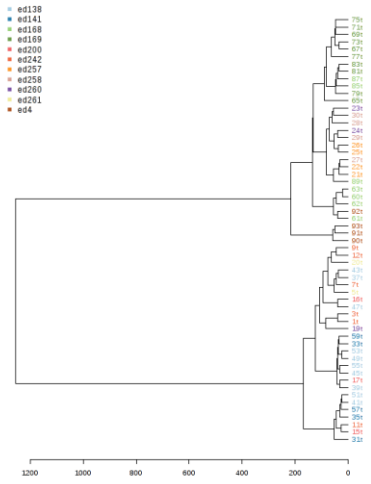
Ecoregions are subdivisions of ecoprovinces. There are 194 ecoregions in Canada as specified in the National Ecological Framework (Marshall et al., 1999). This subdivision is decided by distinctive regional conditions (i.e. ecological factors, climate, physiography, etc.). The relevant ecoregions for this sample set are the Coppermine River Upland (cru; n=8), Ellesmere and Devon Ice Caps (edi; n=4), Great Bear Lake Plain (gblp; n=3), Garry Lake Lowland (gll; n=10), Hay River Lowland (hrl; n=4), Northern Alberta Uplands (nau; n=2), Tuktoyuktuk Coastal Plain (tcp; n=15), Takijua Lake Upland (tklu; n=7), and Tazin Lake Upland (tlu; n=5; Figure 4.2D).

The PLSDA explains 53% (48.6% component 1 and 4.4% component 2) of the variance. The reported CV Q^2, R^2 and accuracy for the first two components are 0.12706, 0.58554, and 0.43103, respectively. The permutation test also reported no overfitting (p-value 0.0025). The most distinct cluster of lakes were located in the Ellesmere and Devon Ice caps ecoregion. The most ambiguous group was the Tazin Lake Upland, which had lake sediments classify into other clusters. The Garry Lake Lowland was the most similar to most of the lake sediments within the Takijua Lake Upland, and the Coppermine River Upland. The ecoregions Tuktoyuktuk Coastal Plain, Hay River Lowland, Northern Alberta Uplands, and Great Bear Lake Plain are clustered together. The RF analysis OOB error rate was 0.328, and a number of groups had class errors of 1 (gblp, hrl, nau, and tlu), which is likely due to small group sizes.

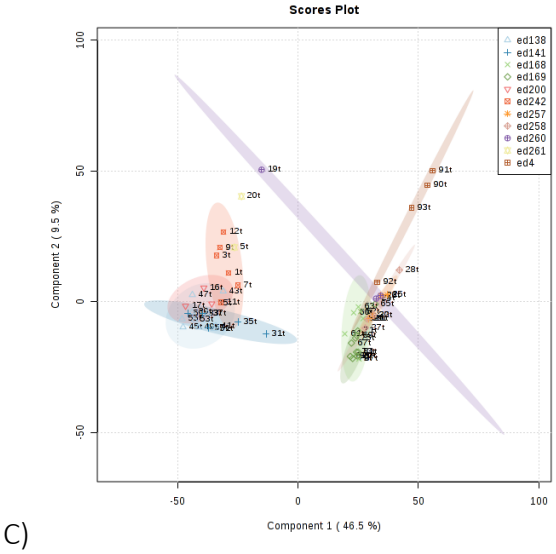
Ecodistricts model



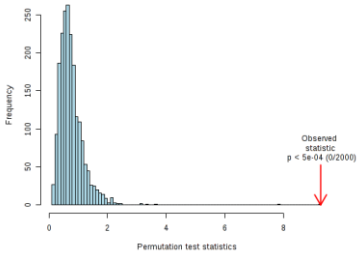
A)



B)



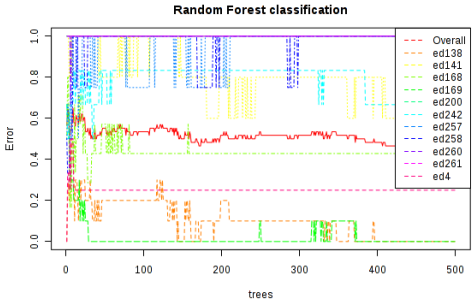
C)



D)

The OOB error is 0.483

	ed138	ed141	ed168	ed169	ed200	ed242	ed257	ed258	ed260	ed261	ed4	class.error
ed138	10	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ed141	3	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6
ed168	0	0	4.0	2.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.429
ed169	0	0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ed200	2	0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
ed242	2	0	0.0	0.0	2.0	1.0	0.0	0.0	0.0	1.0	0.0	0.833
ed257	0	0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	1.0
ed258	0	0	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
ed260	0	0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
ed261	0	0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0
ed4	0	0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.25



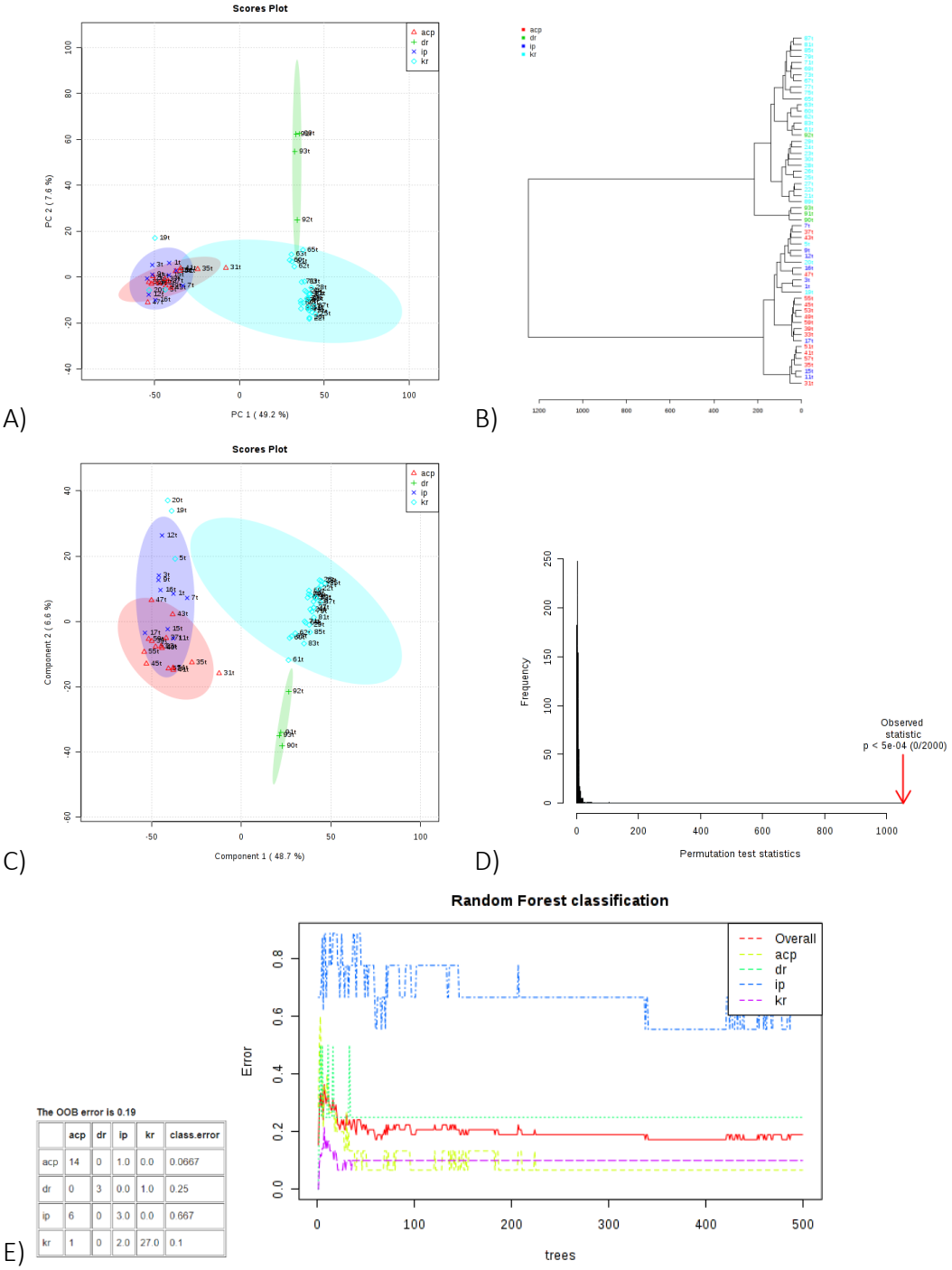
E)

Figure S4.5: Ecodistricts A) PCA, B) Dendrogram, C) PLSDA, D) PLSDA Permutation plot, and E) Random Forest classification plot. All figures generated with MetaboAnalyst.

The smallest subdivision in the National Ecological Framework is the ecodistricts. In Canada, there are 1021 ecodistricts, which are subdivisions of ecoregions based on even smaller graduations of landforms, relief, water bodies, etc (Marshall et al., 1999). There are 11 relevant ecoregions: 138 (n=10), 141 (n=5), 168 (n=7), 169 (n=10), 200 (n=3), 242 (n=6), 257 (n=4), 258 (n=4), 260 (n=3), 261 (n=2), and 4 (n=4; Figure 4.2E).

The ecodistrict PLSDA explains 55.5% of the total variance in the first two PCs (Q^2 0.56476; Accuracy 0.41379; R^2 0.70502). Largely, it seems like there are two main clusters that the ecodistricts cluster into. In the first cluster the ecodistricts are: 200, 138, 242, 261, and 141. The second cluster ecodistricts are: 169, 257, 168, and 4. Only the ecodistrict 260 seems to span the two main clusters. The permutation value was 0.0005. The RF OOB error rate was 0.483, and this was the highest reported error rate.

Georegions model

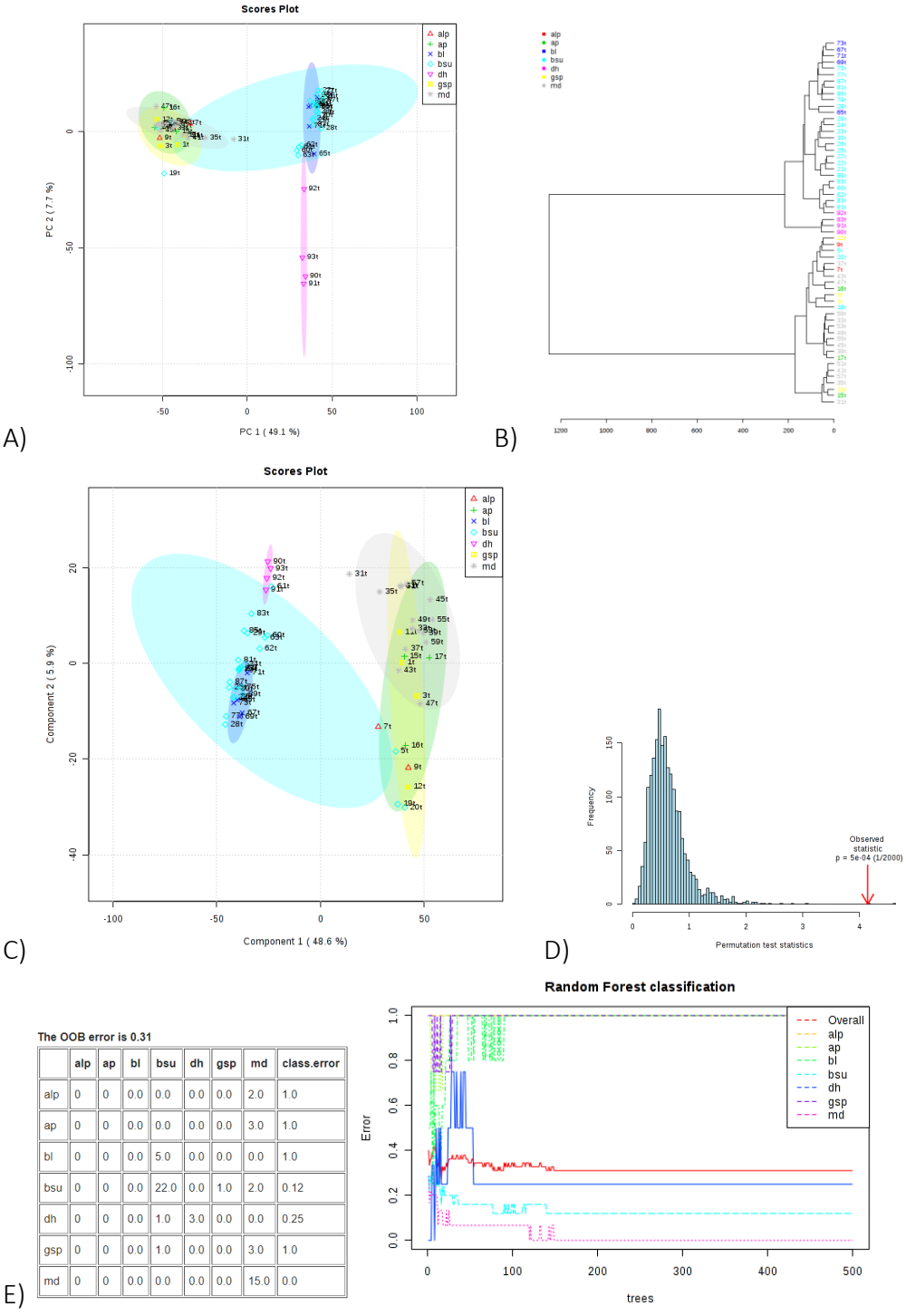


The OOB error is 0.19

	acp	dr	ip	kr	class.error
acp	14	0	1.0	0.0	0.0667
dr	0	3	0.0	1.0	0.25
ip	6	0	3.0	0.0	0.667
kr	1	0	2.0	27.0	0.1

Figure S4.6: Georegions A) PCA, B) Dendrogram, C) PLSDA, D) PLSDA Permutation plot, and E) Random Forest classification plot. Group names are acp = Arctic Coastal Plain, dr = Davis Region, ip = Interior Plains, and kr = Kazan Region. All figures generated with MetaboAnalyst.

Geodistricts model



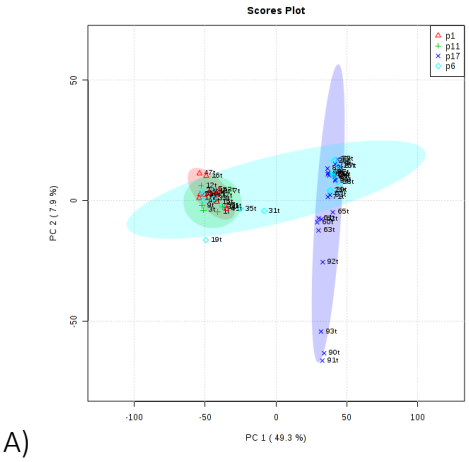
	alp	ap	bl	bsu	dh	gsp	md	class.error
alp	0	0	0.0	0.0	0.0	0.0	2.0	1.0
ap	0	0	0.0	0.0	0.0	0.0	3.0	1.0
bl	0	0	0.0	5.0	0.0	0.0	0.0	1.0
bsu	0	0	0.0	22.0	0.0	1.0	2.0	0.12
dh	0	0	0.0	1.0	3.0	0.0	0.0	0.25
gsp	0	0	0.0	1.0	0.0	0.0	3.0	1.0
md	0	0	0.0	0.0	0.0	0.0	15.0	0.0

Figure S4.7: Geodistricts A) PCA, B) Dendrogram, C) PLSDA, D) PLSDA Permutation plot, and E) Random Forest classification plot. Group names are alp = Alberta Plateau, ap = Anderson Plains, bl = Back Lowland, bsu = Blear-Slave Upland, dh = Davis Highlands, gsp = Great Slave Plain, and md = Mackenzie Delta. All figures generated with MetaboAnalyst.

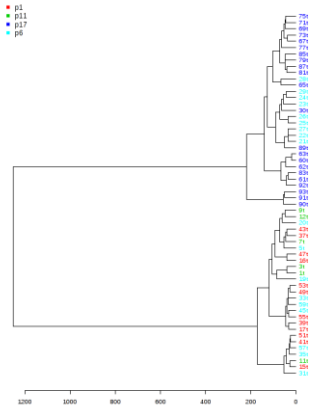
The geodistricts are further divisions of the georegions defined by Bostock (2014). There are 7 relevant geodistricts for this sampling set (Figure 4.2G). In the Borderlands the relevant Arctic Coastal Plain subdivision is the Mackenzie Delta (md; n=15), and the Interior Plains are divided into the Anderson Plain (ap; n=3), Alberta Plateau (alp; n=2), and the Great Slave Plain (gsp; n=4). In the Shield, the Kazan region is divided into the Bear-Slave Upland (bsu; n=25) and the Back Lowland (bl; n=5); the Davis region of the Shield contains the Davis Highlands (dh; n=4).

The PLSDA model explains 54.5% of the total variance (48.6% component 1 and 5.9% component 2). The geodistricts model broadly follows the same patterns are the georegions PLSDA, where the broad clusters seems to follow the Borderland vs Shield Division except for the Bear-Slave Upland. The Bear-Slave Uplands is in the Kazan region, and the same lakes (5t, 19t, and 20t) are in the Borderlands cluster. The CV test produced values 0.20109, 0.65517, 0.48415 for Q^2, accuracy, and R^2, respectively. The permutation p-value was 0.0005. The RF OOB was 0.310.

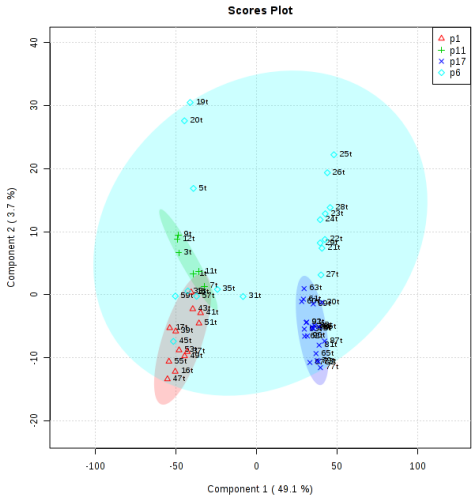
Permafrost model



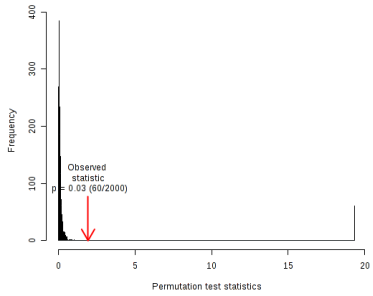
A)



B)



C)



D)

E)

The OOB error is 0.345

	p1	p11	p17	p6	class.error
p1	9	1	0.0	2.0	0.25
p11	3	0	0.0	3.0	1.0
p17	0	0	21.0	1.0	0.0455
p6	5	2	3.0	8.0	0.556

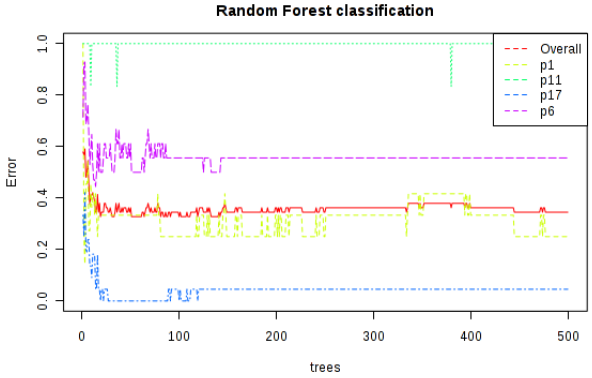


Figure S4.8: Permafrost regions A) PCA, B) Dendrogram, C) PLSDA, D) PLSDA Permutation plot, and E) Random Forest classification plot. Classifications are: 1 = Continuous permafrost with high ground ice extent and thick overburden, 17 = Continuous permafrost with low ground ice extent and exposed bedrock, 6 = Discontinuous permafrost with medium ground ice extent and thick overburden, and 11 = Sporadic permafrost with low ground ice extent and thick overburden. All figures generated with MetaboAnalyst.

Permafrost from the Global Geocryological Data System indicates the continuity of the permafrost, the ground ice content, and the thickness of the overburden (Brown et al., 2002). There are 4 relevant classifications of permafrost: 1 = Continuous permafrost with high ground ice extent and thick overburden (n=12), 17 = Continuous permafrost with low ground ice extent and exposed bedrock (n=22), 6 = Discontinuous permafrost with medium ground ice extent and thick overburden (n=18), and 11 = Sporadic permafrost with low ground ice extent and thick overburden (n=6; Figure 4.2H).

The PLSDA explains 52.8% of the total variance in component 1 (49.1%) and component 2 (3.7%). There is no overfitting because the permutation p-value is >0.05 and Q2 is positive (but it is one of the smallest values so the PLSDA may be close to overfitting). Group 1 clusters best with 11, while 6 spans both main clusters. Group 17 defines most of the second cluster, with some Group 6 lake sediments. It becomes very apparent group 6 does not cluster well with either of the two main groupings. The RF model OOB was 0.345.

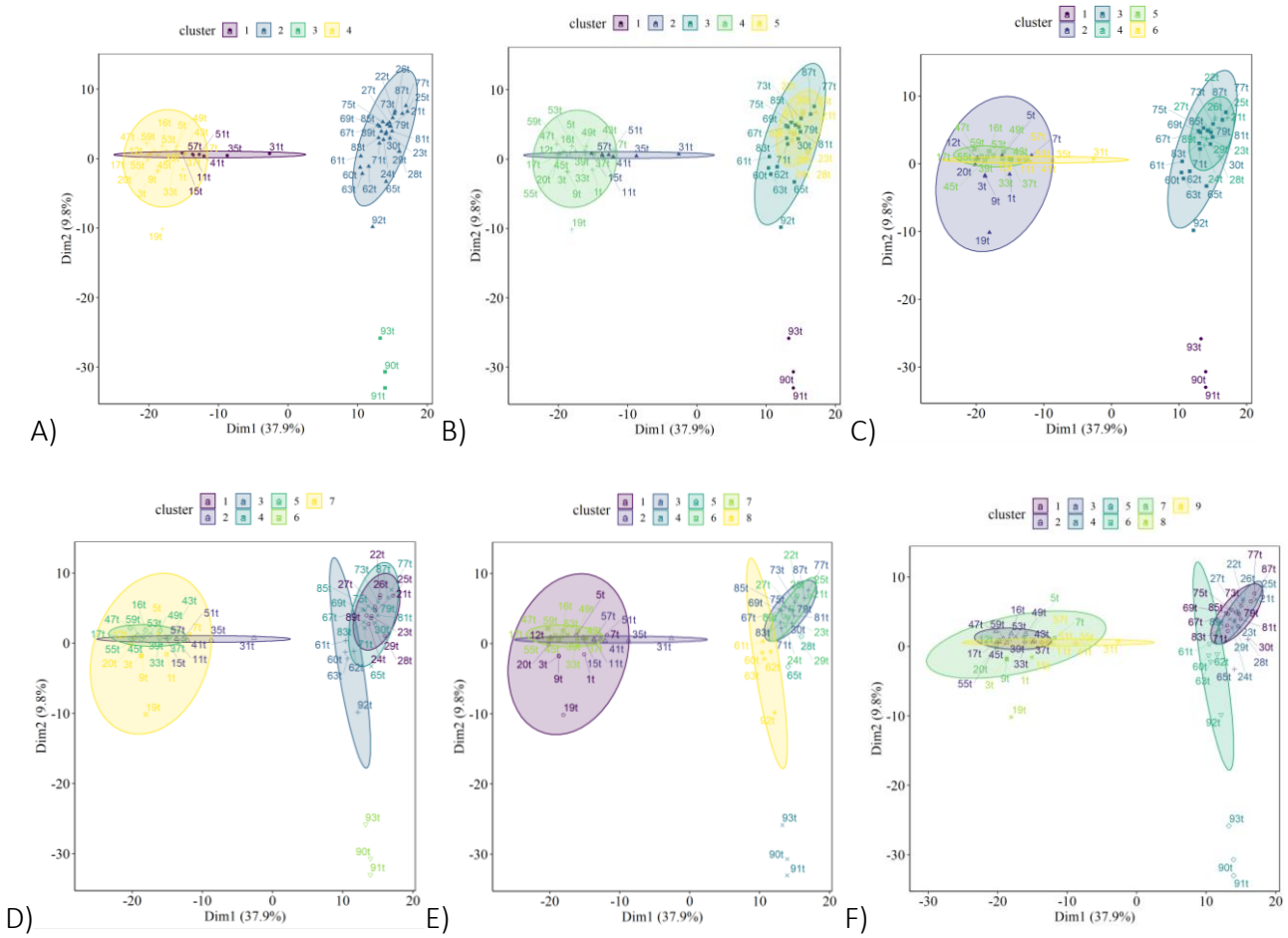


Figure S4.9: Additonal kmeans cluster plots for A) K=4, B) K=5, C) K=6, D) K=7, E) K=8, and F) K=9.

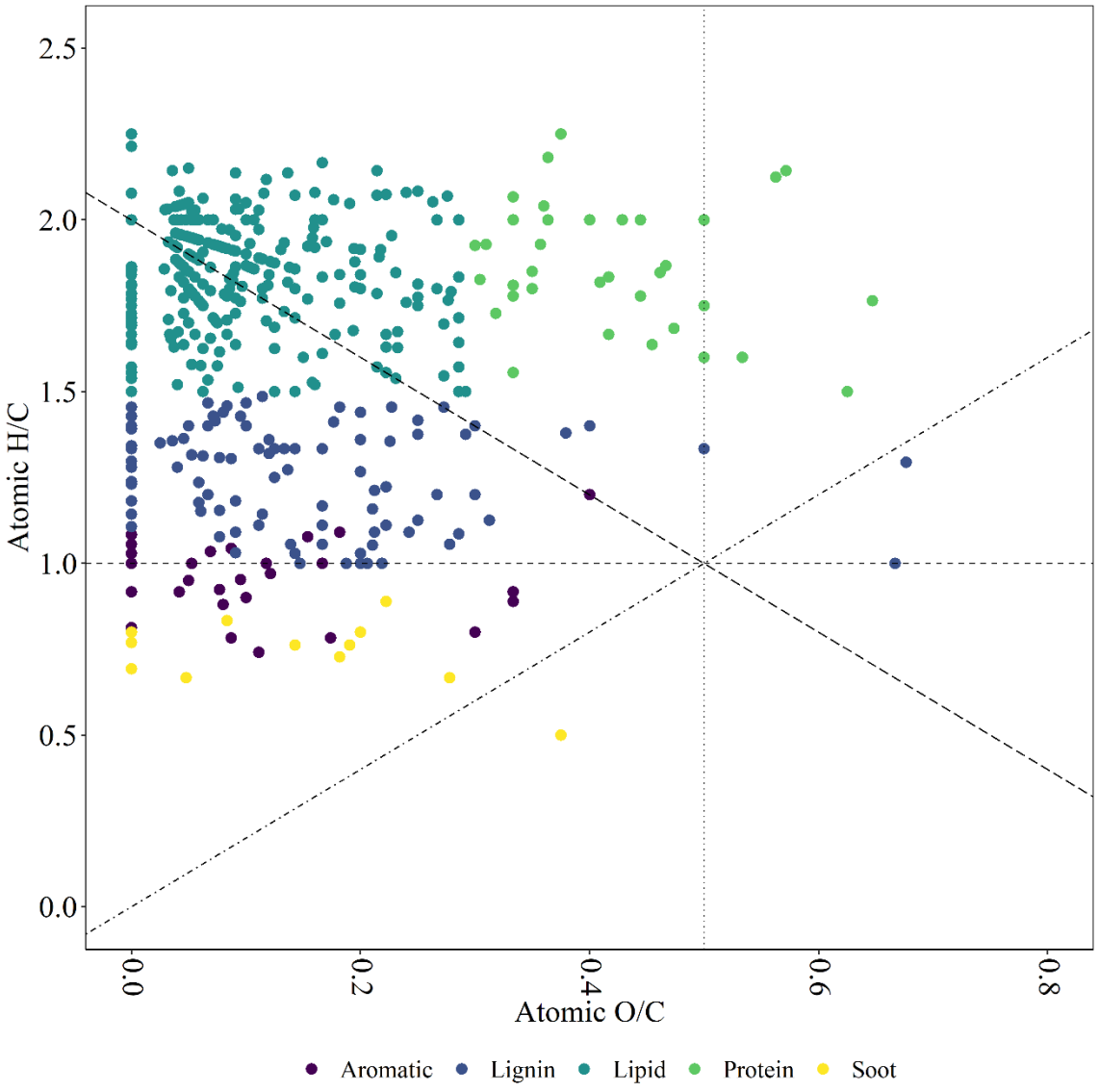
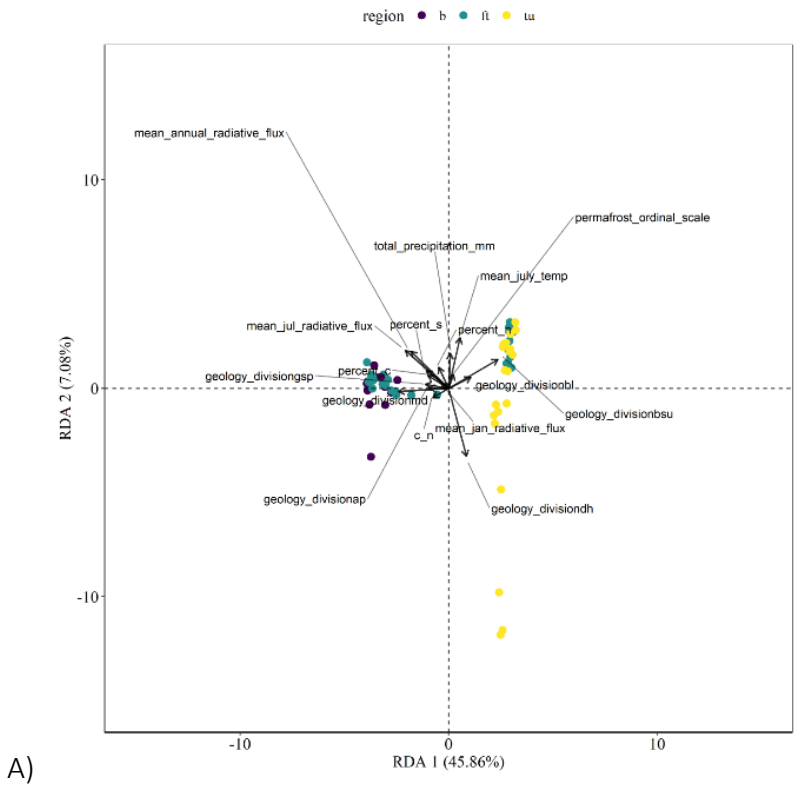
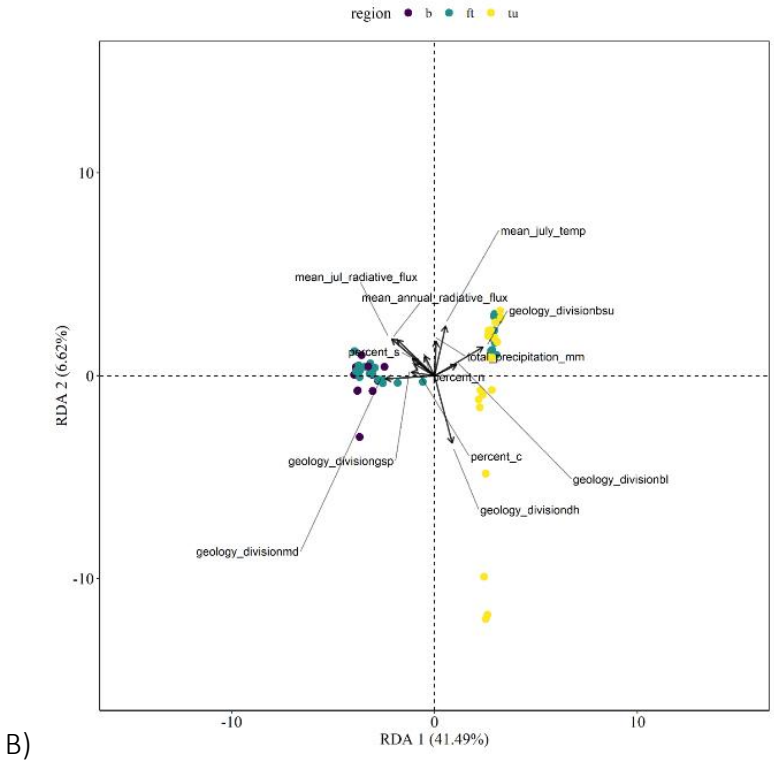


Figure S4.10: Van Krevelen diagram of the molecular sedOM.



A)



B)

Figure S4.11: RDA of A) Model 1 – no redundant variables, and B) Model 2 – only significant variables, and close to significant variables.

Table S4.1: Lake data and constraining data used for RDA.

ID	Treeline Region	Ecozone	Eco-province	Ecoregion	Ecodistrict	Geology Regions	Geology Division	Permafrost	Permafrost (Ordinal Scale)	Latitude	Longitude	Sample Interval Depth (Median cm)	TOC	TN	TS	C/N	Date Cored	Sedimentation Rate (g/cm²/year)*	Mean Jan Radiative Flux (kWh/m²/day)	Mean Jul Radiative Flux (kWh/m²/day)	Mean Annual Radiative Flux (kWh/m²/day)	Mean Jul Temp (°C)	Mean Jan Temp (°C)	Mean Annual Temp (°C)	Highest Annual Temp (°C)	Lowest Annual Temp (°C)	Total Rain Fall (mm)	Total Snow Fall (cm)	Total Precipitation (mm)	Weather Station (Year)	Distance from weather station (km)
61t	tu	sa	3.1	tklu	168	kr	bsu	17	2	64.4484	-110.325	0.25	2.4000	0.1000	NA	24.0000	2014	0.0038	3.8	7.9	5.6	15.6	-29.1	-9.2	23	-36.5	154.6	216.0	370.5	Ekati A (2014)	34.9
75t	tu	sa	3.2	gll	169	kr	bsu	17	2	64.6025	-107.092	0.50	13.5299	1.2379	0.4073	10.9299	1996	low	3.9	7.7	5.7	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	168.5
77t	tu	sa	3.2	gll	169	kr	bsu	17	2	64.6	-107.092	0.50	15.5992	1.5140	0.4238	10.3035	1996	low	3.9	7.7	5.7	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	168.6
79t	tu	sa	3.2	gll	169	kr	bsu	17	2	64.60389	-107.097	0.50	6.5740	0.7089	0.1970	9.2739	1996	low	3.9	7.7	5.7	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	168.3
81t	tu	sa	3.2	gll	169	kr	bsu	17	2	64.60833	-107.109	0.50	12.8041	1.2444	0.3267	10.2894	1996	low	3.9	7.7	5.7	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	167.7
83t	tu	sa	3.2	gll	169	kr	bsu	17	2	64.60417	-107.44	0.25	8.2866	0.7044	0.2992	11.7637	1996	low	3.9	7.7	5.7	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	151.9
11t	b	tp	4.3	hrl	242	ip	gsp	11	4	60.60667	-117.103	1.75	3.5955	0.2073	0.3656	17.3441	2012	0.0842	4.5	8.7	6.4	19.2	-20.0	-1.8	25.4	-32.9	NA	NA	370.3	Hayriver (2012)	76.8
12t	b	tp	4.3	hrl	242	ip	gsp	11	4	60.54161	-116.876	1.75	44.5195	3.6190	1.0319	12.3015	2012	0.0842	5.1	8.6	6.7	19.2	-20.0	-1.8	25.4	-32.9	NA	NA	370.3	Hayriver (2012)	68.6
13t	b	sa	3.1	tcp	141	ip	ap	1	1	68.29547	-133.282	1.50	11.1216	0.5486	0.2684	20.2740	2009	NA	3.9	8.1	5.9	12.4	-24.9	-7.6	23.7	-37.8	NA	NA	240.2	Inuvik (2009)	10.0
15t	b	tp	4.2	gblp	200	ip	ap	1	1	67.47382	-133.77	1.50	3.7110	0.1747	0.0686	21.2383	2009	NA	3.9	8.2	5.9	12.4	-24.9	-7.6	23.7	-37.8	NA	NA	240.2	Inuvik (2009)	95.0
16t	b	tp	4.2	gblp	200	ip	ap	1	1	67.93813	-133.518	0.50	27.4511	1.6147	0.7321	17.0007	2009	NA	3.9	8.2	5.9	12.4	-24.9	-7.6	23.7	-37.8	NA	NA	240.2	Inuvik (2009)	42.5
17t	b	tp	4.2	gblp	200	ip	ap	1	1	67.96019	-133.475	1.50	17.7924	1.4058	0.5991	12.6568	2009	NA	3.9	8.2	5.9	12.4	-24.9	-7.6	23.7	-37.8	NA	NA	240.2	Inuvik (2009)	40.0
18t	b	ts	5.1	tlu	260	kr	bsu	6	3	62.56532	-114.013	3.75	34.2505	3.5947	0.9411	9.5280	2012	NA	4.2	8.8	6.2	19.4	-22.4	-2.9	25.2	-37.5	150.8	140.8	263.4	Yellowknife A (2012)	25.2
19t	b	ts	5.1	tlu	260	kr	bsu	6	3	62.56122	-114.005	6.25	40.9604	4.0657	0.5814	10.0747	2012	NA	4.2	8.8	6.2	19.4	-22.4	-2.9	25.2	-37.5	150.8	140.8	263.4	Yellowknife A (2012)	25.3
1t	b	tp	4.3	hrl	242	ip	gsp	11	4	60.91025	-116.822	6.25	24.1042	1.5543	0.4937	15.5077	2012	0.001	4.5	8.7	6.4	19.2	-20.0	-1.8	25.4	-32.9	NA	NA	370.3	Hayriver (2012)	57.1
20t	b	ts	5.1	tlu	261	kr	bsu	6	3	62.47525	-114.294	2.75	27.9475	2.8473	0.8279	9.8154	2013	NA	3.8	8.2	6.0	15.1	-28.9	-5.0	23.5	-41.1	93.0	96.5	171.4	Yellowknife A (2013)	7.9
21t	ft	ts	5.1	cru	257	kr	bsu	6	3	63.25861	-113.006	0.50	20.0272	1.9123	0.4430	10.4729	1996	low	4.3	7.9	5.9	18.1	-29.7	-4.9	23.9	-38.9	180.5	129.5	261.2	Yellowknife A (1996)	114.5
22t	ft	ts	5.1	cru	257	kr	bsu	6	3	63.22694	-112.305	0.50	27.0583	2.3106	0.3310	11.7105	1996	low	4.3	7.9	5.9	18.1	-29.7	-4.9	23.9	-38.9	180.5	129.5	261.2	Yellowknife A (1996)	139.2
23t	ft	ts	5.1	tlu	260	kr	bsu	6	3	62.55333	-114.12	0.50	26.4564	2.3757	0.5217	11.1362	1998	low	4.1	8.1	6.2	19.1	-28.2	-1.1	25.2	-37.8	195.0	119.8	310.6	YellowKnife A (1998)	20.0
24t	ft	ts	5.1	tlu	260	kr	bsu	6	3	62.53028	-113.839	0.25	21.1584	2.2374	0.3921	9.4567	1998	low	4.1	8.1	6.1	19.1	-28.2	-1.1	25.2	-37.8	195.0	119.8	310.6	YellowKnife A (1998)	32.2
25t	ft	ts	5.1	cru	257	kr	bsu	6	3	63.19417	-112.39	1.25	18.5553	1.7528	0.3425	10.5860	1998	low	4.1	8.0	6.0	19.1	-28.2	-1.1	25.2	-37.8	195.0	119.8	310.6	YellowKnife A (1998)	132.8
26t	ft	ts	5.1	cru	257	kr	bsu	6	3	63.33917	-113.708	1.00	20.7101	1.7291	0.2498	11.9773	1998	low	4.1	8.0	6.0	19.1	-28.2	-1.1	25.2	-37.8	195.0	119.8	310.6	YellowKnife A (1998)	106.9
27t	ft	ts	5.1	cru	258	kr	bsu	6	3	63.6175	-112.305	0.25	14.0642	0.4568	0.1862	30.7910	1996	low	4.3	7.9	5.9	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	145.6
28t	ft	ts	5.1	cru	258	kr	bsu	6	3	63.60028	-112.42	1.25	NA	NA	NA	NA	1998	low	4.1	8.0	6.0	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	150.1
29t	ft	ts	5.1	cru	258	kr	bsu	6	3	63.6715	-112.578	0.75	6.4313	0.5468	0.0524	11.7614	1998	low	4.1	8.0	6.0	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	148.8
30t	ft	ts	5.1	cru	258	kr	bsu	17	2	64.51889	-112.696	0.13	15.9611	1.5282	0.2964	10.4447	1998	0.0065	3.9	7.8	5.9	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	101.6
31t	ft	sa	3.1	tcp	141	acp	md	6	3	68.52074	-133.672	2.90	5.5537	0.5918	0.1438	9.3847	2013	0.0124	3.9	8.0	5.9	13.9	-27.6	-7.5	24.1	-37.6	NA	NA	254.5	Inuvik (2013)	23.8
33t	ft	sa	3.1	tcp	141	acp	md	6	3	68.51609	-133.661	1.75	17.5269	1.3003	0.5094	13.4791	2011	0.0091	4.0	8.4	5.9	15.4	-23.9	-6.6	25.5	-38.3	NA	NA	200.1	Inuvik (2011)	23.0
35t	ft	sa	3.1	tcp	141	acp	md	6	3	68.55117	-133.64	1.88	7.4824	0.6289	0.0977	11.8973	2013	0.0157	3.9	8.0	5.9	13.9	-27.6	-7.5	24.1	-37.6	NA	NA	254.5	Inuvik (2013)	26.5
37t	ft	sa	3.1	tcp	138	acp	md	1	1	69.006	-133.62	0.50	31.2706	2.3513	0.4701	13.2991	2009	low	4.0	7.7	5.8	8.6	-25.1	-9.6	21.4	-36.4	78.2	127.5	158.4	Tuktoyaktuk (2009)	54.5

39t	ft	sa	3.1	tcp	138	acp	md	1	1	68.88382	-133.667	1.50	12.6945	0.0000	0.1905	NA	2009	low	3.9	8.1	5.9	12.4	-24.9	-7.6	23.7	-37.8	NA	NA	240.2	Inuvik (2009)	63.6
3t	b	tp	4.3	hrl	242	ip	gsp	11	4	60.94997	-117.157	3.75	17.6691	0.3744	0.2893	47.1878	2012	0.001	4.5	8.7	6.4	19.2	-20.0	-1.8	25.4	-32.9	NA	NA	370.3	Hayriver (2012)	75.3
41t	ft	sa	3.1	tcp	138	acp	md	1	1	68.99774	-133.513	0.50	12.8599	0.0000	0.1969	NA	2009	0.008	3.9	8.1	5.9	8.6	-25.1	-9.6	21.4	-36.4	78.2	127.5	158.4	Tuktoyaktuk (2009)	53.6
43t	ft	sa	3.1	tcp	138	acp	md	1	1	68.97418	-133.525	0.50	28.9613	2.3815	0.5669	12.1609	2009	0.004	3.9	8.1	5.9	8.6	-25.1	-9.6	21.4	-36.4	78.2	127.5	158.4	Tuktoyaktuk (2009)	56.4
45t	ft	sa	3.1	tcp	138	acp	md	6	3	68.7635	-133.773	0.50	16.7553	1.4080	0.2391	11.8999	2009	low	3.9	8.1	5.9	12.4	-24.9	-7.6	23.7	-37.8	NA	NA	240.2	Inuvik (2009)	51.0
47t	ft	sa	3.1	tcp	138	acp	md	1	1	68.83861	-133.694	0.50	25.8687	1.7049	0.5861	15.1734	2009	low	3.9	8.1	5.9	12.4	-24.9	-7.6	23.7	-37.8	NA	NA	240.2	Inuvik (2009)	59.2
49t	ft	sa	3.1	tcp	138	acp	md	1	1	68.84477	-133.683	0.50	17.4325	1.5060	0.3313	11.5752	2009	low	3.9	8.1	5.9	12.4	-24.9	-7.6	23.7	-37.8	NA	NA	240.2	Inuvik (2009)	59.2
51t	ft	sa	3.1	tcp	138	acp	md	1	1	68.99532	-133.538	1.50	9.8530	0.0949	0.1805	103.8237	2009	0.01	3.9	8.1	5.9	8.6	-25.1	-9.6	21.4	-36.4	78.2	127.5	158.4	Tuktoyaktuk (2009)	54.3
53t	ft	sa	3.1	tcp	138	acp	md	1	1	68.95486	-133.457	1.50	20.3754	1.6568	0.3589	12.2983	2009	0.01	3.9	8.1	5.9	8.6	-25.1	-9.6	21.4	-36.4	78.2	127.5	158.4	Tuktoyaktuk (2009)	57.6
55t	ft	sa	3.1	tcp	138	acp	md	1	1	68.87971	-133.708	1.50	18.8365	1.4859	0.1988	12.6766	2009	0.009	3.9	8.1	5.9	12.4	-24.9	-7.6	23.7	-37.8	NA	NA	240.2	Inuvik (2009)	63.2
57t	ft	sa	3.1	tcp	141	acp	md	6	3	68.51843	-133.374	1.75	10.1412	0.6500	0.1649	15.6028	2013	0.0842	3.9	8.0	5.9	13.9	-27.6	-7.5	24.1	-37.6	NA	NA	254.5	Inuvik (2013)	23.0
59t	ft	sa	3.1	tcp	141	acp	md	6	3	68.45779	-133.64	2.25	19.2772	1.6721	0.4978	11.5286	2013	0.0124	3.9	8.0	5.9	13.9	-27.6	-7.5	24.1	-37.6	NA	NA	254.5	Inuvik (2013)	16.5
5t	b	ts	5.1	tlu	261	kr	bsu	6	3	62.4612	-114.558	1.75	38.8891	3.0405	0.6859	12.7902	2012	0.0029	4.2	8.8	6.2	19.4	-22.4	-2.9	25.2	-37.5	150.8	140.8	263.4	Yellowknife A (2012)	6.2
60t	tu	sa	3.1	tklu	168	kr	bsu	17	2	64.525	-110.163	0.38	4.1000	0.4000	NA	10.2500	2014	0.0138	3.8	7.9	5.6	15.6	-29.1	-9.2	23	-36.5	154.6	216.0	370.5	Ekati A (2014)	28.5
62t	tu	sa	3.1	tklu	168	kr	bsu	17	2	64.61333	-110.85	0.38	9.3000	0.8300	NA	11.2048	1997	low	4.6	7.8	5.7	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	14.5
63t	tu	sa	3.1	tklu	168	kr	bsu	17	2	64.50977	-110.242	0.38	4.8000	0.4900	NA	9.7959	2014	0.0138	3.8	7.9	5.6	15.6	-29.1	-9.2	23	-36.5	154.6	216.0	370.5	Ekati A (2014)	27.0
65t	tu	sa	3.2	gll	169	kr	bl	17	2	66.64056	-104.923	0.75	13.3514	0.7401	0.0897	18.0388	1996	low	3.9	7.8	5.6	14.5	-31.5	-10.3	21.6	-40.3	NA	NA	435.0	Ellice River (1996)	121.8
67t	tu	sa	3.2	gll	169	kr	bl	17	2	66.69194	-104.941	0.50	13.2162	0.5023	0.3495	26.3135	1996	low	3.9	7.8	5.6	14.5	-31.5	-10.3	21.6	-40.3	NA	NA	435.0	Ellice River (1996)	116.1
69t	tu	sa	3.2	gll	169	kr	bl	17	2	66.68972	-104.941	0.50	7.8980	0.7476	0.1770	10.5647	1996	low	3.9	7.8	5.6	14.5	-31.5	-10.3	21.6	-40.3	NA	NA	435.0	Ellice River (1996)	116.5
71t	tu	sa	3.2	gll	169	kr	bl	17	2	65.495	-103.367	0.25	3.7899	0.2472	0.0487	15.3313	1996	low	4.0	7.8	5.7	15.1	-34.6	-10.5	22.2	-44.5	NA	NA	96.7	Robertson Lake (AUT)(1996)	62.1
73t	tu	sa	3.2	gll	169	kr	bl	17	2	65.56722	-103.39	0.75	9.4741	0.6957	0.2174	13.6174	1996	low	4.0	7.8	5.7	15.1	-34.6	-10.5	22.2	-44.5	NA	NA	96.7	Robertson Lake (AUT)(1996)	68.9
7t	b	tp	4.3	nau	242	ip	alp	11	4	60.16722	-117.5	1.75	35.7877	2.8220	0.4323	12.6819	2012	0.0085	4.5	8.7	6.4	19.2	-20.0	-1.8	25.4	-32.9	NA	NA	370.3	Hayriver (2012)	120.8
85t	tu	sa	3.1	tklu	168	kr	bsu	17	2	64.16083	-107.83	0.50	13.3385	0.4140	0.1390	32.2157	1996	low	3.9	7.7	5.7	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	146.0
87t	tu	sa	3.1	tklu	168	kr	bsu	17	2	64.15	-107.817	0.25	5.7668	0.5553	0.0844	10.3849	1996	low	3.9	7.7	5.7	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	146.3
89t	tu	sa	3.1	tklu	168	kr	bsu	17	2	64.15	-107.817	0.50	32.4485	1.9204	0.2631	16.8971	1996	low	3.9	7.7	5.7	15.3	-33.0	-6.5	21.5	-39.2	182.8	163.6	267.3	Ekati A (1998)	147.5
90t	tu	ac	1.1	edi	4	dr	dh	17	2	78.62663	-74.7233	0.63	12.3622	0.0000	0.2457	NA	2006	very low	4.0	7.0	5.3	6.4	-34.5	-16.5	11.2	-44.6	55.0	49.3	94.5	Eureka (2009)	276.0
91t	tu	ac	1.1	edi	4	dr	dh	17	2	78.62349	-74.6804	0.13	16.5324	1.2389	0.2361	13.3444	2006	very low	4.0	7.0	5.3	6.4	-34.5	-16.5	11.2	-44.6	55.0	49.3	94.5	Eureka (2009)	276.0
92t	tu	ac	1.1	edi	4	dr	dh	17	2	78.62623	-74.7067	0.13	2.1593	0.1737	0.0656	12.4334	2006	very low	4.0	7.0	5.3	6.4	-34.5	-16.5	11.2	-44.6	55.0	49.3	94.5	Eureka (2009)	276.0
93t	tu	ac	1.1	edi	4	dr	dh	17	2	78.62048	-74.707	0.13	10.9520	0.5754	0.1372	19.0337	2006	very low	4.0	7.0	5.3	6.4	-34.5	-16.5	11.2	-44.6	55.0	49.3	94.5	Eureka (2009)	276.0
9t	b	tp	4.3	nau	242	ip	alp	11	4	60.21758	-117.84	1.75	37.5178	3.2500	0.3692	11.5439	2012	0.0085	4.5	8.7	6.4	19.2	-20.0	-1.8	25.4	-32.9	NA	NA	370.3	Hayriver (2012)	132.4

**Sedimentation rates and dating estimated from ^{210}Pb and ^{137}Cs dating models (Constant Flux Constant Sedimentation Rate models). Since not all cores were dated, other sedimentation rates were estimated based on modelling data from lakes in the same region, and from literature.*

***Lake 15t was in Fort MacPhearson Plain ecoregion, but was close to the border so it was moved to gblp; Lake 7t was in Ecodistrict 250, but was close to border so moved to 242; Lake 9t was moved from ecodistrict 251 to 242 for the same reason; and Lake 15t was moved from ecodistrict 210 to 200 for the same reason.*

Table S4.2: %TOC, %TN, and %TS min, max, and mean values across the treeline model.

	min	max	mean
%C			
Boreal	3.595532	44.5195	26.09411
FT	5.553684	31.27058	17.72131
Tundra	2.159264	32.44849	10.79895
%N			
Boreal	0.17473	4.065669	2.079934
FT	0	2.381518	1.345089
Tundra	0	1.920361	0.720191
%S			
Boreal	0.068644	1.03194	0.54905
FT	0.052448	0.58614	0.314489
Tundra	0.04865	0.359621	0.208246

Appendix A1: R Code

```
##########################################################################
#####################    MetaboAnalyst in R  ##############################
##########################################################################
##########################################################################

## Let's start with a black slate - Let's re-start the R session

# Ctrl + Shift + Fn + F10  (PC / Linux)
# Command + Shift + Fn+ F10 (Mac OS)

#AUTHOR: Madison Bell
#Date updated: 2019-08-22

#================= Packages needed
==========================================================================

library(MetaboAnalystR)
library(ggplot2)
library(tidyverse)


#=============== Import Data
==========================================================================
#Import data as columns as samples, the second row has groups, and the first column contains the
masses/compounds

mSet<-InitDataObjects("pktable", "stat", FALSE)
mSet<-Read.TextData(mSet, "data_raw/MetaboanalystEcozones.csv", "rowu", "disc")

#=============== Sanity Check
==========================================================================

mSet<-SanityCheckData(mSet)


#=============== Replace Mising values
==========================================================================

#Replaces missing or zero values by half of the smallest positive value in the original dataset and
estimate missing values by replacing with a small value
mSet<-RemoveMissingPercent(mSet, percent=0.5)
mSet<-ImputeVar(mSet, method="min")

#=============== Filtering small values
==========================================================================

#Filters the dataset based on relative standard deviation (rsd), non-parametric rsd (nrsd), mean, sd,
median absolute deviation (mad), interquartile range (iqr)
#Final dataset should contain no more than 5000 variables for effective computing
#T or F = filter based on QC samples; 25 = cut off for rsd where any variable with rsd > 25 only
necessary when T
mSet<-FilterVariable(mSet, "iqr", "F", 25)
mSet<-PreparePrenormData(mSet)

#=============== Remove Samples
==========================================================================

#Remove features
feature.nm.vec <- c("")

#remove samples
smpl.nm.vec <-
c("90b","91b","92b","93b","31b","33b","35b","37b","39b","41b","43b","45b","47b","49b","51b","53b",

"55b","57b","59b","13b","14b","65b","67b","69b","71b","73b","75b","77b","79b","81b","83b","85b",

"87b","89b","60b","61b","62b","63b","1b","3b","7b","9b","11b","12b","15b","16b","17b","5b","18b",
                "19b","20b","21b","22b","23b","24b","25b","26b","27b","28b","29b","30b","18t","13t")
```

```
#Remove Groups
grp.nm.vec <- c("")

mSet<-MetaboAnalystR:::UpdateData(mSet)
mSet<-PreparePrenormData(mSet)

#=============== Data Normalization, transformation and scaling
================================================================================
#Can only run once, and then need to reset in order to see different distributions
#mSet<-Normalization(mSet, "NULL", "NULL", "ParetoNorm", ratio=FALSE, ratioNum=20)
#mSet<-Normalization(mSet, "NULL", "LogNorm", "AutoNorm", ratio=FALSE, ratioNum=20)

mSet<-Normalization(mSet, "NULL", "CrNorm", "ParetoNorm", ratio=FALSE, ratioNum=20)

SaveTransformedData(mSetObj = mSet)  #Exports transformed/normalized datasets

#####Plotting using Metaboanalyst
mSet<-PlotNormSummary(mSet, "norm_0_", "png", 72, width=NA)
mSet<-PlotSampleNormSummary(mSet, "snorm_0_", "png", 72, width=NA)

#=============== PCA Analysis
================================================================================

#Run the unsupervised PCA analysis
mSet<-PCA.Anal(mSet)

#Metaboanalyst Plots
mSet<-PlotPCAPairSummary(mSet, "pca_pair_0_", "png", 72, width=NA, 5)
mSet<-PlotPCAScree(mSet, "pca_scree_0_", "png", 72, width=NA, 5)
mSet<-PlotPCA2DScore(mSet, "pca_score2d_0_", "png", 72, width=NA, 1,2,0.95,0,0)
mSet<-PlotPCALoading(mSet, "pca_loading_0_", "png", 72, width=NA, 1,2);
mSet<-PlotPCABiplot(mSet, "pca_biplot_0_", "png", 72, width=NA, 1,2)
mSet<-PlotPCA3DScoreImg(mSet, "pca_score3d_0_", "png", 72, width=NA, 1,2,3, 40)
mSet<-PlotPCA3DLoading(mSet, "pca_loading3d_0_", "json", 1,2,3)

#=============== PLS-DA Analysis
================================================================================

mSet<-PLSR.Anal(mSet, reg=TRUE)

#Metaboanalyst Plots
mSet<-PlotPLSPairSummary(mSet, "pls_pair_0_", "png", 72, width=NA, 5)
mSet<-PlotPLS2DScore(mSet, "pls_score2d_0_", "png", 72, width=NA, 1,2,0.95,0,0)
mSet<-PlotPLS3DScoreImg(mSet, "pls_score3d_0_", "png", 72, width=NA, 1,2,3, 40)
mSet<-PlotPLSLoading(mSet, "pls_loading_0_", "png", 72, width=NA, 1, 2);
mSet<-PlotPLS3DLoading(mSet, "pls_loading3d_0_", "json", 1,2,3)
mSet<-PLSDA.CV(mSet, "T",5, "Q2")
mSet<-PlotPLS.Classification(mSet, "pls_cv_0_", "png", 72, width=NA)
mSet<-PlotPLS.Imp(mSet, "pls_imp_0_", "png", 72, width=NA, "vip", "Comp. 1", 15,FALSE)

#Validation of PLSDA model
View(mSet[["analSet"]][["plsda"]][["fit.info"]]) #To see the model fit from 10-fold Cross Validation
using Q2 as a performance measure
#PERMUTATION TESTS NOT FUCKING WORKING
mSet<-PLSDA.Permut(mSet, 2000, "bw")
mSet<-PlotPLS.Permutation(mSet, "pls_perm_1_", "png", 72, width=NA)

#=============== Hierarchical Clustering
================================================================================

#Dendrogram
#Distance measure options = Euclidean, Spearman, Pearson
#Clustering algorithm = Ward, Average, Complete, Single
mSet<-PlotHCTree(mSet, "tree_0_", "png", 72, width=NA, "euclidean", "ward.D")

#=============== Random Forest
================================================================================

#Random Forest Analysis (Different VIP values can be obtained here)
mSet<-RF.Anal(mSet, 500,7,1)
mSet<-PlotRF.Classify(mSet, "rf_cls_0_", "png", 72, width=NA)
```

```
mSet<-PlotRF.VIP(mSet, "rf_imp_0_", "png", 72, width=NA)
mSet<-PlotRF.Outlier(mSet, "rf_outlier_0_", "png", 72, width=NA)

#Get OOB
errors = mSet$analSet$rf$err.rate;
nrow = dim(errors)[1];
signif(errors[nrow, 1],3)

#Get Classification Performance Table
View(mSet$analSet$rf$confusion)









##############################################################################
#####################   Recreating Metaboanalyst PCA/PLSDA  ###################
##############################################################################
##############################################################################

## Let's start with a black slate - Let's re-start the R session

# Ctrl + Shift +  F10  (PC / Linux)
# Command + Shift + Fn + F10 (Mac OS)

#AUTHOR: Madison Bell
#Date updated: 2019-08-22


#================= Packages needed
==============================================================================

library(tidyverse)
library(pls)
library(factoextra)
library(ggrepel)
library(ggpubr)
library(viridis)
library(extrafont)

#=============== Import Data
================================================================================

norm <-read_csv("data_raw/MetaboanalystTreeline_nobottom_IQR_CubePareto_no13T18T.csv") #Data in the
format of first column = sample names; second columns = labels; rest of columns are variable; data has
already been normalized and scaled


            #ID      Label   Variable1   Variable2
          #-------------------------------------------
            #lake1   group1  intensity   intensity
            #lake2   group1  intensity   intensity

#=============== Clean & Reogrganize Data
================================================================================

###Collect the categories from the original dataframe

category <- norm[,c("ID", "Label")]

###Reorganize the data frame for pca
row.names(norm) <- norm$ID              #Make your sample names into rownames
```

```
norm.pca <-norm %>%
  select(-c(ID, Label)) %>%             #Remove samples names and Labels
  as.data.frame()

str(norm.pca)                           #Check matric to make sure class assignments are correct


#=============== PCA ================================================================================

#Run PCA using prcomp (same package MetaboanalystR uses)
pca <- prcomp(norm.pca, center = TRUE, scale = FALSE)
summary(pca)

#Plot PCA (with labels)
tiff("figures/PCA_RecreatePCAscript_Treeline_Labels.tiff", units="in", width=7, height=7, res=300)

fviz_pca_ind(pca, axes = c(1, 2), habillage=norm$Label, addEllipses = TRUE, ellipse.level=0.95,
             ellipse.alpha = 0.1, repel = T, pointsize = 2, invisible="quali") +
  theme_pubr() +
  theme(panel.grid = element_blank(),
        panel.border = element_rect(fill= "transparent"),
        text = element_text(size=15,  family="Times New Roman")) +
  coord_fixed(ratio = 1) +
  scale_fill_viridis(discrete = TRUE, option="viridis") +
  scale_color_viridis(discrete = TRUE, option="viridis")

dev.off()

#Plot PCA (without labels)
tiff("figures/PCA_RecreatePCAscript_Treeline_noLabels.tiff", units="in", width=7, height=7, res=300)

fviz_pca_ind(pca, axes = c(1, 2), habillage=norm$Label, addEllipses = TRUE, ellipse.level=0.95,
             ellipse.alpha = 0.05,  pointsize = 2, invisible="quali", label="none") +
  theme_pubr() +
  theme(panel.grid = element_blank(),
        panel.border = element_rect(fill= "transparent"),
        text = element_text(size=15,  family="Times New Roman")) +
  coord_fixed(ratio = 1) +
  scale_fill_viridis(discrete = TRUE, option="viridis") +
  scale_color_viridis(discrete = TRUE, option="viridis")

dev.off()

#=============== PLSDA
================================================================================
#Run PLSDA using pls (same package as MetaboanalystR uses)

datmat <- as.matrix(norm.pca)

comp.num <- dim(norm.pca)[1]-1;
if(comp.num > 8) {
  comp.num <- 8;
}

C <- norm$Label
C2 <- as.factor(C)

if(TRUE){                                                   #Make the labels into categorical binary
operators (i.e. 0 or 1)
  C2 <- scale(as.numeric(C2))[,1];
}else{
  C2 <- model.matrix(~C2-1);
}

plsr.analysis <- pls::plsr(C2~datmat, method='oscorespls', ncomp=comp.num)

write.csv(signif(plsr.analysis$scores,5), row.names=rownames(norm.pca), file="plsda_score.csv")

#Plot PLSDA (with labels)
##############Extract information from plsr object
plsr2<-plsr.analysis$scores
```

```
comp1a<-plsr2[,1]
comp2a<-plsr2[,2]
plsr2<-as.data.frame(cbind(comp1a, comp2a, category))

var_comp1 <- as.data.frame(plsr.analysis$Xvar)                  #Determine the variation on the
principle components
round(100*var_comp1[1:1, "plsr.analysis$Xvar"]/plsr.analysis$Xtotvar, 1)   #PC1 variance percent
round(100*var_comp1[2:2, "plsr.analysis$Xvar"]/plsr.analysis$Xtotvar, 1)   #PC2 variance percent

tiff("figures/PLSDA_RecreatePCAscript_Treeline_Labels.tiff", units="in", width=7, height=7, res=300)


ggplot(data=plsr2, aes(comp1a,comp2a, color = Label, shape = Label))+
  geom_point(size=3) +
  geom_text_repel(aes(label=ID)) +
  geom_hline(yintercept=0, linetype="dashed") +
  geom_vline(xintercept=0, linetype="dashed") +
  ylab("Comp 2 (5.4%)") +
  xlab("Comp 1 (49.1%)") +
  theme_pubr() +
  theme(panel.grid = element_blank(),
        panel.border = element_rect(fill= "transparent"),
        text = element_text(size=15,  family="Times New Roman")) +
  coord_fixed(ratio = 1) +
  stat_ellipse(type = "norm", geom = "polygon", alpha = 0.1, aes(fill = Label)) +
  scale_fill_viridis(discrete = TRUE, option="viridis") +
  scale_color_viridis(discrete = TRUE, option="viridis")

dev.off()

tiff("figures/PLSDA_RecreatePCAscript_Treeline_Labels_Coordunfixed .tiff", units="in", width=7,
height=7, res=300)


ggplot(data=plsr2, aes(comp1a,comp2a, color = Label, shape = Label))+
  geom_point(size=3) +
  geom_text_repel(aes(label=ID)) +
  geom_hline(yintercept=0, linetype="dashed") +
  geom_vline(xintercept=0, linetype="dashed") +
  ylab("Comp 2 (5.4%)") +
  xlab("Comp 1 (49.1%)") +
  theme_pubr() +
  theme(panel.grid = element_blank(),
        panel.border = element_rect(fill= "transparent"),
        text = element_text(size=15,  family="Times New Roman")) +
  #coord_fixed(ratio = 1) +
  stat_ellipse(type = "norm", geom = "polygon", alpha = 0.1, aes(fill = Label)) +
  scale_fill_viridis(discrete = TRUE, option="viridis") +
  scale_color_viridis(discrete = TRUE, option="viridis")

dev.off()

#=============== Kmeans Clustering ========
================================================================================
#Run kmeans clustering using the base R stats package

#Determining the breakpoint for the number of clustters to use (code found online at
http://www.learnbymarketing.com/tutorials/k-means-clustering-in-r-example/)
rng<-2:20 #K from 2 to 20
tries <-100 #Run the K Means algorithm 100 times
avg.totw.ss <-integer(length(rng)) #Set up an empty vector to hold all of points
for(v in rng){ # For each value of the range variable
  v.totw.ss <-integer(tries) #Set up an empty vector to hold the 100 tries
  for(i in 1:tries){
    k.temp <-kmeans(norm.pca, 3, nstart=100, centers=v) #Run kmeans
    v.totw.ss[i] <-k.temp$tot.withinss#Store the total withinss
  }
  avg.totw.ss[v-1] <-mean(v.totw.ss) #Average the 100 total withinss
}
```

```
elbow <- cbind(rng, avg.totw.ss)       #Merge the two lists together

tiff("figures/Kmeans_RecreatePCAscript_Treeline_elbowplot.tiff", units="in", width=7, height=7, res=300)

ggplot(data=elbow, aes(rng,avg.totw.ss))+
  geom_point(size=3) +
  geom_line() +
  ggtitle("Total WCSS by Various K") +
  ylab("Average Total Within Cluster Sum of Squares") +
  xlab("Value of K") +
  theme_pubr() +
  scale_fill_viridis(discrete = TRUE, option="viridis") +
  scale_color_viridis(discrete = TRUE, option="viridis")

dev.off()

#Set the number of clusters to the breakpoint (if there is one)
k <- kmeans(norm.pca, 2, nstart=100)      #Specifies three clusters

tiff("figures/Kmeans_RecreatePCAscript_Treeline_n2.tiff", units="in", width=7, height=7, res=300)

fviz_cluster(k, norm.pca, ellipse.type = "norm", addEllipses = TRUE, repel = T, show.clust.cent = F) +
  theme_pubr() +
  theme(panel.grid = element_blank(),
        panel.border = element_rect(fill= "transparent"),
        text = element_text(size=15,  family="Times New Roman")) +
  coord_fixed(ratio = 1) +
  scale_fill_viridis(discrete = TRUE, option="viridis") +
  scale_color_viridis(discrete = TRUE, option="viridis")

dev.off()









################################################################################
#####################   RDA analysis  ######################
################################################################################
################################################################################

## Let's start with a black slate - Let's re-start the R session

# Ctrl + Shift +  F10  (PC / Linux)
# Command + Shift + Fn + F10 (Mac OS)

#AUTHOR:Madison Bell
#Date updated: 2019-08-22


#================= Packages needed
================================================================================

library(vegan)
library(tidyverse)
library(janitor)
library(missForest)
library(Hmisc)
library(ggfortify)
library(ggvegan)
library(ggrepel)
library(ggpubr)
library(viridis)
library(extrafont)
library(caret)
```

```
#=============== Import Data
================================================================================

#Import excel file (sample names in Rows and variables in columns)
tree <-read_csv("data_raw/MetaboanalystTreeline_nobottom_IQR_CubePareto_RDA_no13T18T.csv")
str(tree)


#ID      Variable1   Variable2
#-----------------------------------------
#lake1   intensity   intensity
#lake2   intensity   intensity

desctree <-read_csv("data_raw/Lakes.csv")
desctree <- clean_names(desctree)
str(desctree)

#ID      Variable1   Variable2
#-----------------------------------------
#lake1   value       value
#lake2   value       value

#=============== Clean & Reogrganize Data
================================================================================


#Enode Dummy Variables if necessary

dmy <- desctree %>%
  select(c("identifier", "geology_regions","geology_division"))  ###Select the columns of the variables
to be converted to dummy variables + one variable that is an unique identifer for all samples

dmy2 <- dummyVars(" ~ .", data = dmy)     #Encode the Dummy variables
trsf <- data.frame(predict(dmy2, newdata = dmy))

desctree <- desctree %>%
  left_join(trsf, by=c("identifier")) %>%  ###Join the two data frames
  select(-c("geology_regionsip", "geology_divisionalp" ))  ###Remove one variable from each categorical
dataset to avoid collinearity

#Filter and Subset the dataframe
desctree2 <- desctree %>%
  filter(!str_detect(id, 'b')) %>% #Remove rows containing "b" which indicates bottom
  filter(!id %in% c("13t", "18t")) %>%  #Remove sample lakes based on name in ID column
  select(-c("lake", "identifier", "region", "ecozone",
            "ecoprovince", "ecoregion", "ecodistrict", "permafrost",
            "top_1", "core_depth_median","date_cored", "geology_regions", "geology_division",
            "sedimentation_rate", "weather_station_year",
            "distance_from_weather_station_km", "total_rain_fall_mm",
            "total_snow_fall_cm","median_july_temp", "median_january_temp",
            "median_annual_temp", "x39")) %>%
  as.data.frame

str(desctree2)


tree3 <- tree[-1]   #turn the first column (the ID column) into a vector
row.names(tree3) <- tree$ID  #make the rownames become the first column and match them to the first
column


desctree3 <- desctree2[-1]   #turn the first column (the ID column) into a vector
row.names(desctree3) <- desctree2$id  #make the rownames become the first column and match them to the
first column

#=============== Impute missing data
================================================================================
#Recommended packages are missForest or Hmisc; useful resource
https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/

#Using missForest
```

```
impute_mF <- missForest(desctree3) #impute missing values, using all parameters as default values
impute_mF2 <- impute_mF$ximp    #Save the impute values
impute_mF$OOBerror                 #Error estimation; lower is better

tiff("figures/RDAscript_DistributionScatterplotAfterImputation_missForest_cn.tiff", units="in", width=7,
height=7, res=300)

ggplot() + #Overlay of scatterplots to compare distribution of original dataframe vs. after imputation
  geom_point(data=impute_mF2, aes(latitude, c_n), color='dodgerblue4', size = 3) +
  geom_point(data=desctree3, aes(latitude, c_n), color='firebrick4', size = 3) +
  theme_classic()

dev.off()

tiff("figures/RDAscript_DistributionDensityplotAfterImputation_missForest_cn", units="in", width=7,
height=7, res=300)

ggplot() + #Overlay of density plots to compare distribution of original dataframe vs. after imputation
  geom_density(data=impute_mF2, aes(x = c_n), color ='dodgerblue4') +
  geom_density(data=desctree3, aes(x = c_n), color ='firebrick4') +
  theme_classic()

dev.off()


#Using Hmisc

impute_arg <- aregImpute(~ latitude + longitide + percent_c + percent_n + percent_s +
                           c_n + mean_jan_radiative_flux + mean_jul_radiative_flux +
                           mean_annual_radiative_flux + mean_july_temp + mean_january_temp +
                           mean_annual_temp +  highest_annual_temp + lowest_annual_temp +
                           total_precipitation_mm + geology_regionsacp + geology_regionsdr +
                           geology_regionskr + geology_divisionap + geology_divisionbl +
                           geology_divisionbsu + geology_divisiondh + geology_divisiongsp +
                           geology_divisionmd,  data = desctree3, n.impute = 1)
###impute_arg nk = 0 means to force linearity for categorical variables.
###THERE ARE NO CATEGORICAL VARIABLES IN THIS CASE OR BECASE TOO MANY MISSING VALUES

impute_arg

impute_arg$imputed$percent_c
impute_arg$imputed$percent_n
impute_arg$imputed$percent_s
impute_arg$imputed$c_n


# insert imputed values
impute_Hmisc <- desctree3

impute_Hmisc$percent_c[is.na(impute_Hmisc$percent_c)] <- impute_arg$imputed$percent_c
impute_Hmisc$percent_n[is.na(impute_Hmisc$percent_n)] <- impute_arg$imputed$percent_n
impute_Hmisc$percent_s[is.na(impute_Hmisc$percent_s)] <- impute_arg$imputed$percent_s
impute_Hmisc$c_n[is.na(impute_Hmisc$c_n)] <- impute_arg$imputed$c_n

tiff("figures/RDAscript_DistributionScatterplotAfterImputation_Hmisc_cn.tiff", units="in", width=7,
height=7, res=300)

ggplot() + #Overlay of scatterplots to compare distribution of original dataframe vs. after imputation
  geom_point(data=impute_Hmisc, aes(latitude, c_n), color='dodgerblue4', size = 3) +
  geom_point(data=desctree3, aes(latitude, c_n), color='firebrick4', size = 3) +
  theme_classic()

dev.off()

tiff("figures/RDAscript_DistributionDensityplotAfterImputation_Hmisc_cn.tiff", units="in", width=7,
height=7, res=300)

ggplot() + #Overlay of density plots to compare distribution of original dataframe vs. after imputation
  geom_density(data=impute_Hmisc, aes(x = c_n), color ='dodgerblue4') +
  geom_density(data=desctree3, aes(x = c_n), color ='firebrick4') +
  theme_classic()
```

```
dev.off()

#Comparing imputation methods

tiff("figures/RDAscript_DistributionDensityplotAfterImputation_All_percents.tiff", units="in", width=7,
height=7, res=300)

ggplot() + #Overlay of density plots to compare distribution of original dataframe vs. after imputation
  geom_density(data=impute_Hmisc, aes(x = percent_s), color ='dodgerblue4') +
  geom_density(data=impute_mF2, aes(x =  percent_s), color ='forestgreen') +
  geom_density(data=desctree3, aes(x =  percent_s), color ='firebrick4') +
  theme_classic()

dev.off()
######Choose based on how similar the distributions are to the original data; in this case percent_s and
c_n varies a lot so Hmisc method chosen


#=============== Redundancy Analysis
================================================================================
####Use this to determine if there are redundant variables in your explanatory dataset
V <- redun(~., data = impute_Hmisc, r2 = 0.95, nk=0)
V$In
V$Out

#=============== PCA ============================================================================================
##PCA using the prcomp package which is used by Metaboanalyst
pca <- prcomp(tree3, center=TRUE, scale=F)

tiff("figures/RDAscript_PCA_prcomp.tiff", units="in", width=7, height=7, res=300)

autoplot(prcomp(tree3, center=TRUE, scale=F))  #Uses ggfortifies autoplot

dev.off()

#PCA using the vegan package
tree.pca <- rda(tree3, scale=F)

tiff("figures/RDAscript_PCA_vegan.tiff", units="in", width=7, height=7, res=300)
autoplot(tree.pca)
dev.off()

summary(tree.pca, display=NULL)




#=============== RDA ============================================================================================

vare.rda <- rda(tree3,scale(impute_Hmisc), scale=F)
vare.rda  #shows the output of your ordination
summary(vare.rda) #shows how much is explained by your constraining variables
anova(vare.rda) #even though the variance explained is high, the model
plot(vare.rda)
envfit(vare.rda, scale(impute_Hmisc)) # identify those variables whose vectors are significantly
correlated with the site loadings


impute_Hmisc_sub <- impute_Hmisc %>%
  select(V$In) ##Select variables from the redundancy analysis
vare.rda.envfit <- rda(tree3, scale(impute_Hmisc_sub), scale=F)
vare.rda.envfit
summary(vare.rda.envfit)
anova(vare.rda.envfit)
plot(vare.rda.envfit)
envfit(vare.rda, scale(impute_Hmisc_sub))



#Automatic model selection
```

```
impute_Hmisc_scaled<- impute_Hmisc_sub %>%
  scale() %>%
  as.data.frame             #Scale the imputed_data set since the tree3 is already scaled

######Method one based on AIC criteria (ordination doesn''t really have AIC values so operate with grain
of salt)
m0 <- rda(tree3 ~ 1, impute_Hmisc_scaled)
m1 <- rda(tree3 ~ ., impute_Hmisc_scaled)
step.res <- step(m0, scope=formula(m1), test="perm")
step.res
anova(step.res)
envfit(step.res, impute_Hmisc_scaled)

######Method two based on p-value criteria
m <- ordistep(m0, scope = formula(m1),  Pin = 0.05, Pout = 0.1)
m$anova
m
anova(m)
envfit(m, impute_Hmisc_scaled)

######Method three based on maximizing R2
m2 <- ordiR2step(m0, m1, perm.max = 200,  Pin = 0.05, Pout = 0.1)
m2
m2$anova
anova(m2)
envfit(m2, impute_Hmisc_scaled)



#####Continue to subset models from the original RDA after redundancy analysis and automatic model
selection
impute_Hmisc_sub2 <- impute_Hmisc %>%
  select('percent_c','percent_n','percent_s', "mean_jul_radiative_flux", "mean_annual_radiative_flux",
         "mean_july_temp", "total_precipitation_mm", "geology_divisionbl", "geology_divisionbsu",
         "geology_divisiondh", "geology_divisiongsp", "geology_divisionmd") #Select columns manually
vare.rda.envfit2 <- rda(tree3, scale(impute_Hmisc_sub2), scale=F)
vare.rda.envfit2
summary(vare.rda.envfit2)
anova(vare.rda.envfit2)
plot(vare.rda.envfit2)
envfit(vare.rda, scale(impute_Hmisc_sub2))


impute_Hmisc_sub3 <- impute_Hmisc %>%
  select('percent_c','percent_s', "mean_jul_radiative_flux", "mean_annual_radiative_flux",
         "mean_july_temp", "total_precipitation_mm", "geology_divisionbsu",
         "geology_divisiondh", "geology_divisionmd")
vare.rda.envfit3 <- rda(tree3, scale(impute_Hmisc_sub3), scale=F)
vare.rda.envfit3
summary(vare.rda.envfit3)
anova(vare.rda.envfit3)
plot(vare.rda.envfit3)
envfit(vare.rda, scale(impute_Hmisc_sub3))




#========== Replotting in ggplot2 ====================================================
########Reorganizing so categories can be inlcuded (may not be necessary)
categories <- desctree %>%
  filter(!str_detect(id, 'b')) %>% #Remove rows containing "b" which indicates bottom
  filter(!id %in% c("13t", "18t")) %>%  #Remove sample lakes based on name in ID column
  select('id', 'region', 'geology_regions', 'ecozone') %>%
  rename(Label = id) %>%
  as.data.frame


##############Setting up the data for RDA plot
```

```
ford <- fortify(vare.rda.envfit3, axes = 1:2) %>%
  merge(categories, by = "Label", all=TRUE)  # fortify the ordination and merge the categories in

take <- c('RDA1', 'RDA2')  # make a list of which columns contain the scores we want

arrows <- subset(ford, Score == 'biplot')  # take only biplot arrow scores

mul <- ggvegan:::arrowMul(arrows[, take],
                          subset(ford, select = take, Score == 'sites')) ## multiplier for arrows to
scale them to the plot range

arrows[, take] <- arrows[, take] * mul  # scale biplot arrows


#########Use ggplot2 to make a more customizable plot (split into RDA plot and biplot with arrows)
rda.plot <- ggplot() +
  geom_point(data = subset(ford, Score == 'sites'),
             mapping = aes(x = RDA1, y = RDA2, color = region), size=3) +
  geom_hline(yintercept=0, linetype="dashed") +
  geom_vline(xintercept=0, linetype="dashed") +
  #(data = subset(ford, Score == 'sites'),
                # mapping = aes(x = RDA1, y = RDA2, label = Label), force = 80
                # , segment.alpha = 0.5, direction = "both")+
  ylim(-15,15) +
  xlim(-15,15) +
  labs(x = "RDA 1 (39.2%)", y="RDA 2 (6.55%)") +
  coord_fixed(ratio = 1) +
  theme_pubr() +
  scale_color_viridis(discrete = TRUE, option="viridis") +
  theme(panel.grid = element_blank(),
        panel.border = element_rect(fill= "transparent"),
        text = element_text(size=15,  family="Times New Roman"))

rda.biplot <- rda.plot +
  geom_segment(data = arrows,
               mapping = aes(x = 0, y = 0, xend = RDA1, yend = RDA2),
               arrow = arrow(length = unit(0.01, "npc")), size = 0.8, alpha=0.8) +
  geom_text_repel(data = arrows,
            mapping = aes(label = Label, x = RDA1 * 1.1, y = RDA2 * 1.1), force = 80
            , segment.alpha = 0.5, direction = "both")

tiff("figures/RDAscript_RDA_MinModel.tiff", units="in", width=10, height=10, res=400)

rda.biplot

dev.off()
```

REFERENCES

Beale, D.J., Crosswell, J., Karpe, A.V., Ahmed, W., Williams, M., Morrison, P.D., Metcalfe, S., Staley, C., Sadowsky, M.J., Palombo, E.A., Steven, A.D.L., 2017. A multi-omics based ecological analysis of coastal marine sediments from Gladstone, in Australia's Central Queensland, and Heron Island, a nearby fringing platform reef. Sci. Total Environ. 609, 842–853. https://doi.org/10.1016/j.scitotenv.2017.07.184

Beale, D.J., Crosswell, J., Karpe, A.V., Metcalfe, S.S., Morrison, P.D., Staley, C., Ahmed, W., Sadowsky, M.J., Palombo, E.A., Steven, A.D.L., 2018. Seasonal metabolic analysis of marine sediments collected from Moreton Bay in South East Queensland, Australia, using a multi-omics-based approach. Sci. Total Environ. 631–632, 1328–1341. https://doi.org/10.1016/j.scitotenv.2018.03.106

Bell, M., Blais, J.M., 2019. "-Omics" workflow for paleolimnological and geological archives: A review. Sci. Total Environ. 672, 438–455. https://doi.org/10.1016/j.scitotenv.2019.03.477

Bostock, H.S., 2014. Geology, Physiographic Regions of Canada.

Box, J.E., Colgan, W.T., Christensen, T.R., Schmidt, N.M., Lund, M., Parmentier, F.-J.W., Brown, R., Bhatt, U.S., Euskirchen, E.S., Romanovsky, V.E., Walsh, J.E., Overland, J.E., Wang, M., Corell, R.W., Meier, W.N., Wouters, B., Mernild, S., Mård, J., Pawlak, J., Olsen, M.S., 2019. Key indicators of Arctic climate change: 1971–2017. Environ. Res. Lett. 14, 045010. https://doi.org/10.1088/1748-9326/aafc1b

Brandt, J.P., 2009. The extent of the North American boreal zone. Environ. Rev. 17, 101–161.

Brown, J., Ferrians, O., Heginbottom, J.A., Melnikov, E., 2002. Circum-Arctic Map of Permafrost and Ground-Ice Conditions, Version 2.

Camarero, L., Rogora, M., Mosello, R., Anderson, N.J., Barbieri, A., Botev, I., Kernan, M., Kopáček, J., Korhola, A., Lotter, A.F., Muri, G., Postolache, C., Stuchlík, E., Thies, H., Wright, R.F., 2009. Regionalisation of chemical variability in European mountain lakes. Freshw. Biol. 54, 2452–2469. https://doi.org/10.1111/j.1365-2427.2009.02296.x

Carpino, O.A., Berg, A.A., Quinton, W.L., Adams, J.R., 2018. Climate change and permafrost thaw-induced boreal forest loss in northwestern Canada. Environ. Res. Lett. 13, 084018. https://doi.org/10.1088/1748-9326/aad74e

Catalan, J., Curtis, C.J., Kernan, M., 2009. Remote European mountain lake ecosystems: regionalisation and ecological status. Freshw. Biol. 54, 2419–2432. https://doi.org/10.1111/j.1365-2427.2009.02326.x

Chong, J., Yamamoto, M., Xia, J., 2019. MetaboAnalystR 2.0: From Raw Spectra to Biological Insights. Metabolites 9, 57. https://doi.org/10.3390/metabo9030057

Colby, G.A., Ruuskanen, M.O., St. Pierre, K.A., St. Louis, V.L., Poulain, A.J., Aris-Brosou, S., 2019. Climate change lowers diversity and functional potential of microbes in Canada's high Arctic. bioRxiv. https://doi.org/10.1101/705178

Cole, J.J., Prairie, Y.T., Caraco, N.F., McDowell, W.H., Tranvik, L.J., Striegl, R.G., Duarte, C.M., Kortelainen, P., Downing, J.A., Middelburg, J.J., Melack, J., 2007. Plumbing the Global Carbon Cycle: Integrating Inland Waters into the Terrestrial Carbon Budget. Ecosystems 10, 172–185. https://doi.org/10.1007/s10021-006-9013-8

Dixon, P., 2003. VEGAN, a package of R functions for community ecology. J. Veg. Sci. 14, 927–930. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x

Farrés, M., Martrat, B., Mol, B. de, Grimalt, J.O., Tauler, R., 2015. Extraction of climatic signals from fossil organic compounds in marine sediments up to 11.7Ma old (IODP-U1318). Anal. Chim. Acta 879, 1–9. https://doi.org/10.1016/j.aca.2015.04.051

Fierer, N., Leff, J.W., Adams, B.J., Nielsen, U.N., Bates, S.T., Lauber, C.L., Owens, S., Gilbert, J.A., Wall, D.H., Caporaso, J.G., 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc. Natl. Acad. Sci. 109, 21390–21395. https://doi.org/10.1073/pnas.1215210110

Fritz, S.C., Anderson, N.J., 2013. The relative influences of climate and catchment processes on Holocene lake development in glaciated regions. J. Paleolimnol. 49, 349–362. https://doi.org/10.1007/s10933-013-9684-z

Garnier, S., Ross, N., Rudis, B., Sciaini, M., Scherer, C., 2018. viridis: Default Color Maps from "matplotlib."

Graham, E.B., Crump, A.R., Kennedy, D.W., Arntzen, E., Fansler, S., Purvine, S.O., Nicora, C.D., Nelson, W., Tfaily, M.M., Stegen, J.C., 2018. Multi 'omics comparison reveals metabolome biochemistry, not microbiome composition or gene expression, corresponds to elevated biogeochemical function in the hyporheic zone. Sci. Total Environ. 642, 742–753. https://doi.org/10.1016/j.scitotenv.2018.05.256

Grewer, D.M., Lafrenière, M.J., Lamoureux, S.F., Simpson, M.J., 2018. Spatial and temporal shifts in fluvial sedimentary organic matter composition from a High Arctic watershed impacted by localized slope disturbances. Org. Geochem. 123, 113–125. https://doi.org/10.1016/j.orggeochem.2018.07.004

Grewer, D.M., Lafrenière, M.J., Lamoureux, S.F., Simpson, M.J., 2015. Potential shifts in Canadian High Arctic sedimentary organic matter composition with permafrost active layer detachments. Org. Geochem. 79, 1–13.

Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., Goodacre, R., 2015. A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. Anal. Chim. Acta 879, 10–23. https://doi.org/10.1016/j.aca.2015.02.012

Grosse, G., Harden, J., Turetsky, M., McGuire, A.D., Camill, P., Tarnocai, C., Frolking, S., Schuur, E.A.G., Jorgenson, T., Marchenko, S., Romanovsky, V., Wickland, K.P., French, N., Waldrop, M., Bourgeau-Chavez, L., Striegl, R.G., 2011. Vulnerability of high-latitude soil organic carbon in North America to disturbance. J. Geophys. Res. Biogeosciences 116. https://doi.org/10.1029/2010JG001507

Gudasz, C., Sobek, S., Bastviken, D., Koehler, B., Tranvik, L.J., 2015. Temperature sensitivity of organic carbon mineralization in contrasting lake sediments. J. Geophys. Res. Biogeosciences 120, 1215–1225. https://doi.org/10.1002/2015JG002928

Harrell, F.E.Jr., 2019. Hmisc.

Harsch, M.A., Hulme, P.E., McGlone, M.S., Duncan, R.P., 2009. Are treelines advancing? A global meta-analysis of treeline response to climate warming. Ecol. Lett. 12, 1040–1049. https://doi.org/10.1111/j.1461-0248.2009.01355.x

Heckman, K., Welty-Bernard, A., Rasmussen, C., Schwartz, E., 2009. Geologic controls of soil carbon cycling and microbial dynamics in temperate conifer forests. Chem. Geol., Combined Ecological and Geologic Perspectives in Ecosystem Studies 267, 12–23. https://doi.org/10.1016/j.chemgeo.2009.01.004

Holmer, M., Storkholm, P., 2001. Sulphate reduction and sulphur cycling in lake sediments: a review. Freshw. Biol. 46, 431–451. https://doi.org/10.1046/j.1365-2427.2001.00687.x
Jansson, J.K., Hofmockel, K.S., 2018. The soil microbiome — from metagenomics to metaphenomics. Curr. Opin. Microbiol. 43, 162–168. https://doi.org/10.1016/j.mib.2018.01.013
Jiménez-Morillo, N.T., González-Pérez, J.A., Almendros, G., De la Rosa, J.M., Waggoner, D.C., Jordán, A., Zavala, L.M., González-Vila, F.J., Hatcher, P.G., 2018. Ultra-high resolution mass spectrometry of physical speciation patterns of organic matter in fire-affected soils. J. Environ. Manage. 225, 139–147. https://doi.org/10.1016/j.jenvman.2018.07.069
Johnston, E.R., Hatt, J.K., He, Z., Wu, L., Guo, X., Luo, Y., Schuur, E.A.G., Tiedje, J.M., Zhou, J., Konstantinidis, K.T., 2019. Responses of tundra soil microbial communities to half a decade of experimental warming at two critical depths. Proc. Natl. Acad. Sci. 201901307. https://doi.org/10.1073/pnas.1901307116
Jorgenson, M.T., Grosse, G., 2016. Remote Sensing of Landscape Change in Permafrost Regions. Permafr. Periglac. Process. 27, 324–338. https://doi.org/10.1002/ppp.1914
Kassambara, A., 2019. ggpubr: “ggplot2” Based Publication Ready Plots.
Kassambara, A., Mundt, F., 2017. factoextra: Extract and Visualize the Results of Multivariate Data Analysis.
Kellerman, A.M., Dittmar, T., Kothawala, D.N., Tranvik, L.J., 2014. Chemodiversity of dissolved organic matter in lakes driven by climate and hydrology. Nat. Commun. 5. https://doi.org/10.1038/ncomms4804
Kernan, M., Ventura, M., Bitušík, P., Brancelj, A., Clarke, G., Velle, G., Raddum, G.G., Stuchlík, E., Catalan, J., 2009. Regionalisation of remote European mountain lake ecosystems according to their biota: environmental versus geographical patterns. Freshw. Biol. 54, 2470–2493. https://doi.org/10.1111/j.1365-2427.2009.02284.x
Kim, S., Kramer, R.W., Hatcher, P.G., 2003. Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram. Anal. Chem. 75, 5336–5344. https://doi.org/10.1021/ac034415p
Kind, T., Fiehn, O., 2007. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics 8, 105. https://doi.org/10.1186/1471-2105-8-105
Korosi, J.B., Thienpont, J.R., Pisaric, M.F., 2017. Broad-scale lake expansion and flooding inundates essential wood bison habitat. Nat. Commun. 8.
Krause, S., Lewandowski, J., Grimm, N.B., Hannah, D.M., Pinay, G., McDonald, K., Martí, E., Argerich, A., Pfister, L., Klaus, J., Battin, T., Larned, S.T., Schelker, J., Fleckenstein, J., Schmidt, C., Rivett, M.O., Watts, G., Sabater, F., Sorolla, A., Turk, V., 2017. Ecohydrological interfaces as hot spots of ecosystem processes. Water Resour. Res. 53, 6359–6376. https://doi.org/10.1002/2016WR019516
Lara, M.J., Nitze, I., Grosse, G., Martin, P., McGuire, A.D., 2018. Reduced arctic tundra productivity linked with landform and climate change interactions. Sci. Rep. 8. https://doi.org/10.1038/s41598-018-20692-8
Lehmann, J., Kleber, M., 2015. The contentious nature of soil organic matter. Nature. https://doi.org/10.1038/nature16069
Li, H.-Y., Wang, H., Wang, H.-T., Xin, P.-Y., Xu, X.-H., Ma, Y., Liu, W.-P., Teng, C.-Y., Jiang, C.-L., Lou, L.-P., Arnold, W., Cralle, L., Zhu, Y.-G., Chu, J.-F., Gilbert, J.A., Zhang, Z.-J., 2018. The chemodiversity of

paddy soil dissolved organic matter correlates with microbial community at continental scales. Microbiome 6. https://doi.org/10.1186/s40168-018-0561-x

Marín-Spiotta, E., Gruley, K.E., Crawford, J., Atkinson, E.E., Miesel, J.R., Greene, S., Cardona-Correa, C., Spencer, R.G.M., 2014. Paradigm shifts in soil organic matter research affect interpretations of aquatic carbon cycling: transcending disciplinary and ecosystem boundaries. Biogeochemistry 117, 279–297. https://doi.org/10.1007/s10533-013-9949-7

Marshall, I.B., Schut, P.H., Ballard, M., 1999. A National Ecological Framework for Canada: Attribute Data. Agriculture and Agri-Food Canada, Research Branch, Centre for Land and Biological Resources Research and Environment Canada, State of the Environment Directorate, Ecozone Analysis Branch, Ottawa/Hull.

McGowan, S., Gunn, H.V., Whiteford, E.J., John Anderson, N., Jones, V.J., Law, A.C., 2018. Functional attributes of epilithic diatoms for palaeoenvironmental interpretations in South-West Greenland lakes. J. Paleolimnol. 60, 273–298. https://doi.org/10.1007/s10933-017-9968-9

Morse, P.D., Wolfe, S.A., Kokelj, S.V., Gaanderse, A.J.R., 2016. The Occurrence and Thermal Disequilibrium State of Permafrost in Forest Ecotopes of the Great Slave Region, Northwest Territories, Canada. Permafr. Periglac. Process. 27, 145–162. https://doi.org/10.1002/ppp.1858

Myers-Smith, I.H., Hik, D.S., 2018. Climate warming as a driver of tundra shrubline advance. J. Ecol. 106, 547–560. https://doi.org/10.1111/1365-2745.12817

Oni, O.E., Schmidt, F., Miyatake, T., Kasten, S., Witt, M., Hinrichs, K.-U., Friedrich, M.W., 2015. Microbial Communities and Organic Matter Composition in Surface and Subsurface Sediments of the Helgoland Mud Area, North Sea. Front. Microbiol. 6. https://doi.org/10.3389/fmicb.2015.01290

Osburn, C.L., Anderson, N.J., Stedmon, C.A., Giles, M.E., Whiteford, E.J., McGenity, T.J., Dumbrell, A.J., Underwood, G.J.C., 2017. Shifts in the Source and Composition of Dissolved Organic Matter in Southwest Greenland Lakes Along a Regional Hydro-climatic Gradient: DOM Quality in Greenland Lakes. J. Geophys. Res. Biogeosciences 122, 3431–3445. https://doi.org/10.1002/2017JG003999

Payette, S., Fortin, M.J., Gamache, I., 2001. The subarctic forest-tundra: the structure of a biome in a changing climate. Bioscience 51, 709–718. https://doi.org/10.1641/0006-3568(2001)051[0709:tsftts]2.0.co;2

Pearson, R.G., Phillips, S.J., Loranty, M.M., Beck, P.S.A., Damoulas, T., Knight, S.J., Goetz, S.J., 2013. Shifts in Arctic vegetation and associated feedbacks under climate change. Nat. Clim. Change 3, 673–677. https://doi.org/10.1038/nclimate1858

Pienitz, R., Smol, J.P., Lean, D.R., 1997a. Physical and chemical limnology of 59 lakes located between the southern Yukon and the Tuktoyaktuk Peninsula, Northwest Territories (Canada). Can. J. Fish. Aquat. Sci. 54, 330–346.

Pienitz, R., Smol, J.P., Lean, D.R., 1997b. Physical and chemical limnology of 24 lakes located between Yellowknife and Contwoyto Lake, Northwest Territories (Canada). Can. J. Fish. Aquat. Sci. 54, 347–358.

Plaza, C., Pegoraro, E., Bracho, R., Celis, G., Crummer, K.G., Hutchings, J.A., Hicks Pries, C.E., Mauritz, M., Natali, S.M., Salmon, V.G., Schädel, C., Webb, E.E., Schuur, E.A.G., 2019. Direct observation of permafrost degradation and rapid soil carbon loss in tundra. Nat. Geosci. 12, 627–631. https://doi.org/10.1038/s41561-019-0387-6

Post, E., Forchhammer, M.C., Bret-Harte, M.S., Callaghan, T.V., Christensen, T.R., Elberling, B., Fox, A.D., Gilg, O., Hik, D.S., Hoye, T.T., Ims, R.A., Jeppesen, E., Klein, D.R., Madsen, J., McGuire, A.D., Rysgaard, S., Schindler, D.E., Stirling, I., Tamstorf, M.P., Tyler, N.J.C., van der Wal, R., Welker, J.,

Wookey, P.A., Schmidt, N.M., Aastrup, P., 2009. Ecological Dynamics Across the Arctic Associated with Recent Climate Change. Science 325, 1355–1358. https://doi.org/10.1126/science.1173113
Ranson, K.J., Montesano, P.M., Nelson, R., 2011. Object-based mapping of the circumpolar taiga–tundra ecotone with MODIS tree cover. Remote Sens. Environ. 115, 3670–3680. https://doi.org/10.1016/j.rse.2011.09.006
Rossel, R.A.V., Lee, J., Behrens, T., Luo, Z., Baldock, J., Richards, A., 2019. Continental-scale soil carbon composition and vulnerability modulated by regional environmental controls. Nat. Geosci. 12, 547. https://doi.org/10.1038/s41561-019-0373-z
Rühland, K.M., Smol, J.P., Wang, X., Muir, D.C.G., 2003. Limnological characteristics of 56 lakes in the Central Canadian Arctic Treeline Region. J. Limnol. 62, 9–27.
Saleem, A., Bell, M.A., Kimpe, L.E., Korosi, J.B., Arnason, J.T., Blais, J.M., 2019. Identifying novel treeline biomarkers in lake sediments using an untargeted screening approach. Sci. Total Environ. 694, 1–11. https://doi.org/10.1016/j.scitotenv.2019.133684
Schimel, J.P., Schaeffer, S.M., 2012. Microbial control over carbon cycling in soil. Front. Microbiol. 3. https://doi.org/10.3389/fmicb.2012.00348
Schmidt, M.W.I., Torn, M.S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I.A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D.A.C., Nannipieri, P., Rasse, D.P., Weiner, S., Trumbore, S.E., 2011. Persistence of soil organic matter as an ecosystem property. Nature 478, 49–56. https://doi.org/10.1038/nature10386
Schuur, E. a. G., McGuire, A.D., Schädel, C., Grosse, G., Harden, J.W., Hayes, D.J., Hugelius, G., Koven, C.D., Kuhry, P., Lawrence, D.M., Natali, S.M., Olefeldt, D., Romanovsky, V.E., Schaefer, K., Turetsky, M.R., Treat, C.C., Vonk, J.E., 2015. Climate change and the permafrost carbon feedback. Nature 520, 171–179. https://doi.org/10.1038/nature14338
Simpson, A.J., Simpson, M.J., Smith, E., Kelleher, B.P., 2007. Microbially Derived Inputs to Soil Organic Matter: Are Current Estimates Too Low? Environ. Sci. Technol. 41, 8070–8076. https://doi.org/10.1021/es071217x
Sobek, S., Tranvik, L.J., Prairie, Y.T., Kortelainen, P., Cole, J.J., 2007. Patterns and regulation of dissolved organic carbon: An analysis of 7,500 widely distributed lakes. Limnol. Oceanogr. 52, 1208–1219. https://doi.org/10.4319/lo.2007.52.3.1208
Stackhouse, P.W.Jr., Zhang, T., Westberg, D., Barnett, A.J., Bristow, T., Macpherson, B., Hoell, J.M., 2018. POWER Release 8.0.1 (with GIS Applications) Methodology (Data Parameters, Sources, & Validation). NASA Langley Research Centre.
Stubbins, A., Spencer, R.G.M., Chen, H., Hatcher, P.G., Mopper, K., Hernes, P.J., Mwamba, V.L., Mangangu, A.M., Wabakanghanzi, J.N., Six, J., 2010. Illuminated darkness: Molecular signatures of Congo River dissolved organic matter and its photochemical alteration as revealed by ultrahigh precision mass spectrometry. Limnol. Oceanogr. 55, 1467–1477. https://doi.org/10.4319/lo.2010.55.4.1467
Swenson, T.L., Bowen, B.P., Nico, P.S., Northen, T.R., 2015a. Competitive sorption of microbial metabolites on an iron oxide mineral. Soil Biol. Biochem. 90, 34–41. https://doi.org/10.1016/j.soilbio.2015.07.022
Swenson, T.L., Jenkins, S., Bowen, B.P., Northen, T.R., 2015b. Untargeted soil metabolomics methods for analysis of extractable organic matter. Soil Biol. Biochem. 80, 189–198. https://doi.org/10.1016/j.soilbio.2014.10.007

Swenson, T.L., Karaoz, U., Swenson, J.M., Bowen, B.P., Northen, T.R., 2018. Linking soil biology and chemistry in biological soil crust using isolate exometabolomics. Nat. Commun. 9. https://doi.org/10.1038/s41467-017-02356-9

Szymańska, E., Saccenti, E., Smilde, A.K., Westerhuis, J.A., 2012. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics 8, 3–16. https://doi.org/10.1007/s11306-011-0330-3

Tfaily, M.M., Chu, R.K., Toyoda, J., Tolić, N., Robinson, E.W., Paša-Tolić, L., Hess, N.J., 2017. Sequential extraction protocol for organic matter from soils and sediments using high resolution mass spectrometry. Anal. Chim. Acta 972, 54–61. https://doi.org/10.1016/j.aca.2017.03.031

Wang, N.F., Zhang, T., Yang, X., Wang, S., Yu, Y., Dong, L.L., Guo, Y.D., Ma, Y.X., Zang, J.Y., 2016. Diversity and Composition of Bacterial Community in Soils and Lake Sediments from an Arctic Lake Area. Front. Microbiol. 7. https://doi.org/10.3389/fmicb.2016.01170

Ward, C.P., Cory, R.M., 2015. Chemical composition of dissolved organic matter draining permafrost soils. Geochim. Cosmochim. Acta 167, 63–79. https://doi.org/10.1016/j.gca.2015.07.001

Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J.J., van Duijnhoven, J.P.M., van Dorsten, F.A., 2008. Assessment of PLSDA cross validation. Metabolomics 4, 81–89. https://doi.org/10.1007/s11306-007-0099-6

Wickham, H., 2011. ggplot2. Wiley Interdiscip. Rev. Comput. Stat. 3, 180–185.
https://doi.org/10.1002/wics.147

Wieder, W.R., Hartman, M.D., Sulman, B.N., Wang, Y.-P., Koven, C.D., Bonan, G.B., 2018. Carbon cycle confidence and uncertainty: Exploring variation among soil biogeochemical models. Glob. Change Biol. 24, 1563–1579. https://doi.org/10.1111/gcb.13979

Wolfe, S.A., Morse, P.D., Kokelj, S.V., Gaanderse, A.J., 2017. Great Slave Lowland: The Legacy of Glacial Lake McConnell, in: Slaymaker, O. (Ed.), Landscapes and Landforms of Western Canada. Springer International Publishing, Cham, pp. 87–96. https://doi.org/10.1007/978-3-319-44595-3_5

Wu, X., Wu, L., Liu, Y., Zhang, P., Li, Q., Zhou, J., Hess, N.J., Hazen, T.C., Yang, W., Chakraborty, R., 2018. Microbial Interactions With Dissolved Organic Matter Drive Carbon Dynamics and Community Succession. Front. Microbiol. 9, 1234. https://doi.org/10.3389/fmicb.2018.01234

Xu, X., Li, W., Deng, H., Fujibayashi, M., Nomura, M., Nishimura, O., Wang, G., 2019. Approach deliberation for source identification of sedimentary organic matters via comparing freshwater lakes with multi-ecotypes. Sci. Total Environ. 649, 327–334. https://doi.org/10.1016/j.scitotenv.2018.08.236

Xu, X., Li, W., Fujibayashi, M., Nomura, M., Nishimura, O., Li, X., 2015. Predominance of terrestrial organic matter in sediments from a cyanobacteria- blooming hypereutrophic lake. Ecol. Indic. 50, 35–43. https://doi.org/10.1016/j.ecolind.2014.10.020

Zhao, X., Yang, Y., Shen, H., Geng, X., Fang, J., 2019. Global soil–climate–biome diagram: linking surface soil properties to climate and biota. Biogeosciences 16, 2857–2871. https://doi-org.proxy.bib.uottawa.ca/10.5194/bg-16-2857-2019

CHAPTER 5: Evaluation of quality controls, extraction method, and soil type on soil metabolomics studies

Madison A. Bell[1,2], Ulrica McKim[1], Amanda Sproule[1], Ryan Tobalt[1], Ed Gregorich[1*], David P. Overy[1*]

[1]Agriculture and Agri-food Canada, Ottawa Research and Development Centre, Ottawa, ON, K1A 0C6, Canada

[2]Laboratory for the Analysis of Natural and Synthetic Environmental Toxicants, Department of Biology, University of Ottawa, Ottawa ON, K1N 6N5, Canada

*Corresponding Author(s):

David Overy (david.overy@canada.ca) & Ed Gregorich (ed.gregorich@canada.ca)

STATEMENT OF CONTRIBUTIONS

Madison A. Bell: Conceptualization, Methodology, Software - Scripting, Formal analysis, Data curation, Visualization, Writing – Original draft preparation. **Ulrica McKim:** Methodology, Investigation. **Amanda Sproule:** Methodology, Investigation. **Ryan Tobalt:** Investigation. **Ed Gregorich:** Conceptualization, Methodology, Funding acquisition, Project administration, Supervision, Writing – Review & Editing. **David P. Overy:** Conceptualization, Methodology, Supervision, Funding acquisition, Project administration, Writing – Review & Editing. This chapter has been submitted to Soil Biology and Biochemistry:

Bell, M.A., McKim, U., Sproule, A., Tobalt, R., Gregorich, E., Overy, D.P. (2020) Effects of extraction method and soil type on soil metabolomics studies. Soil biology and biochemistry.

ABSTRACT

Soil metabolomics, the study of low molecular weight organic compounds in soil, is a developing field of research. To date, there has been little research on quality controls and standardization of soil organic matter (SOM) extraction methods in soil metabolomics studies. The purpose of this study was to evaluate the use of quality controls and several SOM extraction methods in diverse soils. Autoclaving and baking of the soils did not adequately reduce microbial activity and losses of water extractable SOM within 24 hr after extraction were about ~1% hr^{-1}. To further understand losses attributed to microbial activity, we introduced two quality controls: 1) microbially-sensitive recovery standards, and 2) pure clays as control matrices. The control matrices allowed us to partition the loss of the recovery standards to adsorption, extraction procedure, and other processes (i.e. biological degradation). The two quality controls were then trialed in three SOM extraction methods (involving chloroform fumigation, shaking, and ultrasonication) on three different soils. From our results, we recommend including quality controls like recovery standards, control matrices, and method blanks to compensate for significant differences in recoveries across different soil types. Compared to ultrasonication and chloroform fumigation, the shaking method of extraction demonstrated the greatest recoveries of SOM with the smallest losses of SOM to adsorption and other processes.

1. INTRODUCTION

Recent theories suggest that the stabilization of soil organic matter (SOM) is dominated by an equilibrium between microbial activity and SOM-mineral interactions (Lehmann and Kleber, 2015; Schmidt et al., 2011) and that the main constituents of stable SOM are microbially-derived C and necromass (Kindler et al., 2006; Schweigert et al., 2015; Simpson et al., 2007). Still, the specific mechanisms of the biotic and abiotic processes regulating SOM stabilization are poorly understood (Jackson et al., 2017; Lehmann and Kleber, 2015; Zhang et al., 2020). Knowledge gaps in these SOM processes can be addressed with a systems biology approach. In systems biology, interactions between individual molecules, microbes, and their surrounding environments are studied to derive a holistic understanding of complex biological processes (Jansson and Hofmockel, 2018; Schmidt et al., 2011). Application of systems biology to investigate the processes involved in SOM stabilization requires the development of soil-optimized untargeted "omics" methods to investigate microbial communities (i.e. metagenomics), microbial functionality (i.e. metatranscriptomics and metaproteomics), and SOM composition (i.e. soil metabolomics). Soil-optimized method development has undergone more extensive evaluation for metagenomics, metatranscriptomics, and metaproteomics compared to metabolomics as many techniques and technical infrastructure from molecular biology are directly transferrable to soil science. SOM extraction has not yet been optimized or standardized for soil metabolomics research.

Despite decades of research, a comprehensive knowledge of the true composition of SOM remains elusive due to challenges in extracting and analyzing this chemically diverse and dynamic matrix. New analytical technologies like high resolution mass spectrometry (HRMS) have helped to overcome some technical problems; however, a distinct lack of information characterizing the structural

diversity of organic C molecules contained in soils makes molecular identification difficult (Bell and Blais, 2019). SOM is an integral part of the soil physical structure. Thus, solvent extracts of SOM offer only a partial representation of its native form and the extraction process can introduce chemical artifacts. Although there are several examples of using soil metabolomics to study SOM composition (Fox et al., 2017; Li et al., 2018; Tfaily et al., 2017, 2015), little study of how soil metabolomics extraction protocols influence the extracted SOM composition has been undertaken. For instance, although Tfaily et al. (2017, 2015) demonstrated that the choice of solvent has a significant impact on the chemical composition of SOM extracted from soil, most researchers still use a single extraction solvent. Other potential methodological problems have had limited discussion in soil metabolomics literature (Bell and Blais, 2019; Swenson et al., 2015); these problems include: absence of common quality controls, microbial activity during extraction, experimental replication, and the consistency of extraction methods across different soil types. These are critical for developing robust methods to extract representative samples of SOM across diverse soils varying in their physical, chemical, and biological properties.

Our purpose in this study was to investigate potential issues in SOM extraction methods used in soil metabolomics research. Specifically, we aimed to evaluate the effects of soil microbial activity, experimental replication, and the inclusion of microbially-sensitive recovery standards in a soil metabolomics experiment. We also compared the effects of different soil types and extraction methods on the chemical composition of SOM extracts. Incubations were performed to understand the impact of autoclaving and baking of soil on microbial respiration. Three extraction methods were used with quality control procedures (i.e. replication, microbially-sensitive recovery standards, control matrices, and method blanks). All of these experiments were performed on three diverse soils varying in their texture, surface area, and SOM content to assess the efficacy of extraction methods across different soil types.

2. METHODS AND MATERIALS

2.1. Soil collection and characterization

Two mined clays and three soils were used in this study. Low-defect kaolinite (KGa-1b, Georgia, USA) and Ca-montmorillonite (STx-1b, Texas, USA) were obtained from the Clay Minerals Society (USA). Three soils were collected at the Central Experimental Farm, Agriculture and Agri-Food Canada, Ottawa, Canada (Table 5.1). The soils were selected because they spanned a range of texture (from 10 to 30% clay), organic C contents (1.17 to 2.55 %), specific surface area (3.3 to 16.3 m^2/g) and cation exchange capacity (7.6 to 19 meq/100g). They were classified as Brunisols and Gleysols in the Canadian System of Soil Classification and were from the Grenville, Rideau, and Granby soil series. Bulk samples were collected from the surface 15 cm, sieved (< 4mm), air-dried, and stored at room temperature until analyzed. The soils were analyzed for particle size distribution using the pipette method (Kroetsch and Wang, 2008). Soil pH was determined using a 1:1 ratio of water to soil. Organic C (OC) and total N (TN) were analyzed (using a subsample <250 µm) on a vario El cube CN elemental analyzer (Elementar, Germany). Cation exchange capacity was determined using the $BaCl_2$ method (Hendershot et al., 2008). Specific surface area measurements were performed on air-dried soil (sieved <1mm; degassed with He at room temperature), using a TriStar 3000 analyzer (Micromeritics, GA, USA) and the Brunauer-Emmett-Teller (BET) method (Whitfield and Reid, 2013).

Table 5.1: Characteristics of the soils and clay controls used in this study.

	Grenville	Rideau	Granby	Kaolinite	Montmorillonite
Soil Classification	Orthic Melanic Brunisol	Gleyed Melanic Brunisol	Orthic Humic Gleysol	Clay (low-defect)	Clay
Sand (%)	64	50	73	0	0
Silt (%)	26	20	17	0	0
Clay (%)	10	30	10	100	100
% Organic C	2.55	1.58	1.17	0.02	0.01
% Total N	0.21	0.13	0.1	0.02	0.04
pH (1:1 DH_2O)	6.27	6.57	6.21	NA	NA
Specific surface area (m^2/g)	4.04	16.31	3.3	10.05	83.79
CEC (meq/100g)	13	19	7.63	2	84.4

2.2. Experiment outlines

2.2.1. Evaluating microbial activity

The three soils were collected on the Central Experimental Farm in Ottawa, Ontario in July 2019, subsampled into 150g batches, and refrigerated (4°C) until required. Five different treatments (i.e. fresh field moist soil, air drying at room temperature (~20°C) for one week, two autoclaving cycles for 30 min at 121°C and 16 psi, two baking cycles at 120°C for 2 hours, and cold soils incubated at 4°C) were applied all three soil types.

Soil incubations were conducted to measure soil CO_2 respiration (Hopkins, 2007) and evaluate microbial growth and activity. Final autoclaving and baking cycles were performed immediately before incubation. 10g of autoclaved LCMS grade H_2O was added to re-wet each treated soil (150g). The 150g treated soils were subsampled (25g) into pre-sterilized (120°C for 2 hours) 1L wide-mouth mason jars with four replicates per treatment and per soil type. The remaining soil was used to determine the water content. The jars were sealed with pre-sterilized lids (soaked in 70% ethanol and dried overnight),

flushed with CO_2 free air (4 min) and the internal jar pressures adjusted to ~0.6kPa. Jars were placed in incubation chambers set at 25°C, except for the cold treatment samples which were incubated at 4°C.

Respiration of soil CO_2 was measured at 0.75, 2, 3, 4, 7, and 14 days after being placed in the jars using Teflon syringes and injection into a LI-7000 CO_2/H_2O Analyzer (LI-COR). Calibrations were performed daily. After CO_2 measurements, the jars were flushed (4 min), adjusted to ~0.6 kPa, and placed back into the incubation chambers. Calibration curves were used to calculate the µg CO_2-C/ g of soil respired, the mineralization rate (µg C/g of soil/day), and the cumulative µg CO_2-C / g of soil.

2.2.2. Efficacy testing of microbially-sensitive recovery standards using shaking extraction

The microbially-sensitive recovery standards for this experiment were: 4-Methylumbelliferyl phosphate (MUBph; Sigma Aldrich), 2,6-Dichloroindophenol sodium salt hydrate (DCIP; Sigma Aldrich), and 3,4-Dihydroxy-L-phenylalanine (LDOPA; Sigma Aldrich). Efficacy testing experiments included recovery standards spiked into method blanks, kaolinite clay, montmorillonite clay, and each of the three soil types. Soil organic matter was extracted using the shaking protocol (*section 2.3.2.)* with modifications. Recovery standards were dissolved in LCMS grade H_2O and 250µL (2mg/mL) of the solution was spiked into the soil using a glass 1 mL Hamilton syringe. There were three shaking durations (i.e. 1, 2, and 4 hours) and each group contained three sampling replicates. Only H_2O extractions were performed for these experiments and the extraction solvent was maintained at room temperature. All samples were analyzed in duplicate on the UPLC-HRMS/DAD, resulting in 6 replicates per time point per group.

Table 5.2: Extract total organic carbon (TOC) and total nitrogen (TN) results from the method comparison experiment.

Sample Type	Extraction Technique	Mean TOC (StDev) [mg/kg dry soil]	Mean TN (StDev) [mg/kg dry soil]
Granby soil	Chloroform Fumigation	183.5 (8.2)	21.4 (1.3)
	Shaking	163.6 (8.5)	18.2 (0.2)
	Ultrasonication	170.5 (10.6)	19.9 (0.5)
Grenville loam	Chloroform Fumigation	251.4 (3.7)	42.8 (2.8)
	Shaking	220.0 (11)	33.7 (0.6)
	Ultrasonication	264.3 (2.3)	39.1 (0.8)
Rideau clay	Chloroform Fumigation	232.4 (12.5)	20.0 (1.2)
	Shaking	207.6 (6.4)	17.9 (0.1)
	Ultrasonication	184.6 (9.5)	17.7 (0.3)

2.2.3. SOM extraction method comparison

Three different SOM extraction methods were chosen for comparison (i.e. shaking, chloroform fumigation, and ultrasonication). The methods were performed on the 6 different sample types (i.e. method blanks; kaolinite and montmorillonite clays (control matrices); and Grenville, Granby, and Rideau soils) with both H_2O and MeOH LCMS grade extraction solvents. There were 3 sampling replicates for each sample type (except H_2O extraction method blanks had 6 replicates). The extraction method protocols were performed as described in *section 2.3.* with the following modifications. All extraction solvents were chilled (4°C) and samples were kept on ice during the extraction. The samples were spiked with fresh recovery standards (250µL) containing LCMS grade H_2O, LDOPA (2mg/mL), and MUBph (1mg/mL) using a glass 1mL Hamilton syringe (DCIP was not stable in solution with LDOPA and MUBph).

A parallel experiment was performed to determine the H_2O-extracted total organic C (TOC) and N (TN) across the different extraction methods (Table 5.2). The same extraction protocol was followed as in *section 2.3.*, except 2g of soil and 20mL of LCMS grade H_2O were used with three sampling

replicates. H_2O extracts were filtered through 25mm 0.45µm (PALL Acrodisc PTFE HPLC Certified) syringe filters, acidified (50µL; 2M HCl), and analyzed with a Shimadzu TOC-VCPN & TNM-1 (Mandel Scientific, USA). Results were method blank subtracted.

2.3. SOM extraction methods

2.3.1. Soil preparation

Air-dried soils were homogenized using a mortar and pestle and sieving (<1 mm) to improve replicate reproducibility (data not shown). An extraction ratio of 1 g dry weight soil: 10mL solvent was used in all experiments. Polypropylene Falcon tubes (50 mL) were used for water extractions (except chloroform fumigation) due to the high centrifugation speeds required to separate clays from the supernatant during extraction. Polypropylene is not ideal for SOM extractions because of the potential for leaching of plastic additives into the extraction solvent (Yao et al., 2019), but we mitigated this by adding extra method blank replicates (x6) so plastic contaminants could be removed during data processing. Extractions with non-polar solvents (i.e MeOH) were performed in 20ml vials with PTFE-lined caps. Method blanks were prepared for all extractions (minimum 3 replicates).

2.3.2. Shaking extraction

Soil samples were weighed into appropriate tubes/vials and LCMS grade solvent (10mL) was added to each vial using glass serological pipettes. Samples were vortexed, shaken in the dark (1 hr), and centrifuged. H_2O extracts were centrifuged at 4800rpm (10 min), and MeOH extracts at 2800rpm (10 min). Supernatants were transferred into a new tube/vial using glass Pasteur pipettes.

After supernatant removal, H_2O extracts required a freeze-thaw step to precipitate clay particles that remained suspended despite centrifugation. H_2O extracts were frozen overnight (-20°C), thawed in the dark, and centrifuged at 2500rpm (5 min). This freeze-thaw step allowed for easier filtering, without changing the overall composition of the SOM extract (data not shown), which is consistent with results from Pokrovsky et al. (2018). The H_2O extracts were filtered through 25mm 0.45µm (PALL Acrodisc PTFE HPLC Certified) syringe filters using disposable needles and 20mL Teflon syringes. Tubes were rinsed with LCMS grade H_2O and the filtrate was added back into the rinsed tube. The samples were refrozen (-20°C), lyophilized, and stored at -20°C until analysis. MeOH extracts were immediately dried under vacuum (Genevac; low BP; 30°C) after supernatant removal and were not filtered. MeOH pellets were stored at -20°C until analysis.

2.3.3. Chloroform fumigation and shaking

All samples (regardless of extraction solvent) were prepared for fumigation (2007) in 20ml vials with PTFE-lined caps. Samples were placed inside a desiccator along with 50mL ethanol free $CHCl_3$. The desiccator was sealed, vented (3-5 min), and left in the dark at room temperature for 24 hours. After fumigation, the desiccator was evacuated to remove $CHCl_3$, the samples removed, and immediately extracted.

The SOM extraction procedure was the same shaking procedure described in *section 2.3.2.,* except that the H_2O extracts were centrifuged at 2800 rpm (10 min) because the extracts were in scintillation vials. After centrifugation, the H_2O supernatant was removed and placed into 50mL falcon tubes and processed as previously described.

2.3.4. Ultrasonication-assisted extraction

Soil samples were weighed into appropriate tube/vials and LCMS grade solvent (2.5mL) was added using glass serological pipettes. Samples were vortexed, placed in a Brason Ultrasonics digital ultrasonic bath (20.8L 120V; Canada) on HIGH for 10 min, and then centrifuged. H_2O extracts were centrifuged at 4800rpm (10 min) and the non-polar extracts at 2800rpm (10 min). Supernatants were transferred into a new tube/vial using glass Pasteur pipettes. Pellets were re-extracted three more times (i.e. total 4 extractions at 2.5mL each equals 10mL total extraction solvent) and the supernatants pooled. The number of extractions was optimized (data not shown) where a single extraction was found to be too limiting and no significant gains were obtained after 4-5 repeated extractions.

After extraction, the pooled supernatants from MeOH extracts were dried under vacuum using a GeneVac (low BP; 30°C) and stored at -20°C. Pooled H_2O supernatants were frozen (-20°C overnight), thawed in the dark, and centrifuged at 2500rpm (5 min). H_2O extracts were filtered through 25mm 0.45µm (PALL Acrodisc PTFE HPLC Certified) syringe filters using disposable needles and 20mL Teflon syringes, the tubes were rinsed with LCMS grade H_2O, and then the filtrate was added back into the rinsed tube. Samples were refrozen (-20°C), lyophilized, and stored at -20°C until analysis.

2.4. Sample analysis

2.4.1. Sample analysis preparation

Pellets were resolubilized in 0.5mL of their extraction solvent, vortexed, placed in an ultrasonic bath for 10 min, and centrifuged at 2500rpm (10 min). Samples were transferred to 1.5mL HPLC vials after filtering through 13mm 0.2µm (PALL Acrodisc PTFE HPLC Certified) filters with disposable needles and 3mL Teflon syringes. Calibration curves for recovery standards were prepared fresh.

2.4.2. Batch effect and recovery standard stability tests

The triplicate machine replicates for efficacy testing of microbially-sensitive recovery standards using shaking extraction *(section 2.2.2.)* and SOM extraction method comparison (*section 2.2.3.*) were used to assess the stability of the recovery standards during UPLC-HRMS/DAD analysis and monitor for batch effect (Bell and Blais, 2019). Two different randomization procedures were applied: 1) Randomized batches, 2) Randomized reinjections. Randomized batching was used for the recovery standard efficacy testing, where all samples (with triplicate machine replicates) were randomized into three batches, and all three machine replicates of one sample were in one batch. One batch was removed from the freezer at a time, so the samples in the batch were only exposed to ~1.5 days at room temperature. Randomized injection was used for the SOM extract machine replication, all samples were randomized, and then that randomized injection list was reinjected twice more without refrigeration in between. Thus, all samples were exposed for 6 days at room temperature during UPLC-HRMS/DAD analysis.

2.4.3. UPLC-HRMS/DAD

H_2O extracts were analyzed using Hydrophobic Interaction Chromatography (HILIC) Ultra pressure liquid chromatography (UPLC)-HRMS/DAD. 2µL was injected into a Thermo Scientific Dionex Ultimate 3000 UHPLC system coupled to a Thermo LTQ Orbitrap XL high resolution mass spectrometer (HRMS) and a photo-diode array detector (DAD). Acetonitrile (ACN) blanks were interspersed throughout the run (every 10 samples) to monitor for carry over. Injection order of samples was randomized. HILIC was performed using a SeQuant ZIC®-cHILIC 100 Å column (2.1 x 100mm, 3µm) and a

20 x 2.1mm guard column (Merk/EMD Millipore). Flow rate was 0.35mL/min. The mobile phase was H_2O with 5 mM acetic acid (solvent A) and ACN (solvent B). The mobile phase was held at 90% B for 0.5 min and the gradient decreased from 90% to 40% B over 20 min and then held at 40% B for 2 min. The mobile phase was returned to starting conditions over 0.5 min, and allowed to equilibrate for 3 min. Duration of data collection was 24 min.

MeOH extracts were analyzed using reverse phase (RP) chromatography with the same UPLC-HRMS/DAD system. Sample injection and organization were the same. RP chromatography was performed with a Phenomenex Kinetex C_{18} 100 Å column (2.1 x 50mm, 1.7µm) with a SecurityGuard ULTRA C_{18} guard. The flow rate was 0.35mL/min. The mobile phase was water with 0.1% formic acid (solvent A) and ACN with 0.1% formic acid (solvent B). The gradient increased from 5% solvent B to 95% over 4.5 min and then held at 95% for 3.5 min. Mobile phase was returned to starting conditions over 0.5 min, and allowed to equilibrate for 5 min. Duration of data collection was 10 min.

The HRMS was operated in ESI+ mode monitoring m/z 100-2,000 Da and the data was acquired in profile mode. HRMS parameters were: sheath gas (40), auxiliary gas (5), sweep gas (2), spray voltage (4.0kV), capillary temperature (320 ˚C), capillary voltage (35V), tube lens (100V), maximum injection time (500ms), and microscans (1).

2.5. Data analysis

All data were analyzed using RStudio (v1.2.1335) and R (v3.6.0) unless otherwise stated. The data were cleaned and processed using the R packages *tidyverse* (Wickham et al., 2019) and *janitor* (Firke, 2020). Plots were produced using *ggplot2* (Wickham, 2016), *ggrepel* (Slowikowski, 2020), and *ggpubr* (Kassambara, 2020a).

2.5.1. Recovery standard quantification

Recovery standards were quantified for both HRMS and photodiode array (PDA) detection using Thermo Xcalibur™ 2.2 SP1.48. Wavelengths for quantification were 280nm (LDOPA), 315nm (MUBph), and 270nm (DCIP). External five-point (2mg/mL to 0.0008mg/mL) calibration curves were assumed to be linear. Limit of detections were 3 times the absolute value of the standard error of the curve divided by the slope. Limit of quantifications were 10 times the absolute value of the standard error of the curve divided by the slope. DAD data were used due to the lower limit of quantification for most recovery standards (Table S5.1). Non-detects and concentrations < limit of quantification were replaced by the limit of detection divided by the square root of 2. Recoveries were calculated by converting the adjusted concentrations to moles and dividing by the moles in the original spiked solution after adjusting for dilutions.

2.5.2. Analysis of variance and multiple pairwise comparisons

Analysis of variance and multiple pairwise comparisons were analyzed using R packages *car* (Fox and Weisberg, 2019), *rstatix* (Kassambara, 2020b), and *moderndive* (Kim and Ismay, 2019). Simple linear regressions were performed and model residuals for cumulative soil respiration (μg CO_2-C / g of soil), mineralization rate (μg C/g of soil/day), and recovery standard recoveries (%R) were all heteroscedastic and non-normal even after transformations so non-parametric Kruskal-Wallis and Wilcoxon Signed Rank tests on untransformed data were used for analyses of variance and multiple pairwise comparisons, respectively. The term “significant” used in this work indicates p-values < 0.05 from Kruskal-Wallis and Wilcoxon Signed Rank tests.

2.5.3. Loss partitioning

Loss partitioning was determined using the following logic. Percent recoveries (%R) represents the proportion of recovery standard in the sample. 1 - %R yields total percent loss ($\%L_{Total}$). Percent extraction loss ($\%L_{Ex}$) was estimated from the method blanks where it was assumed 1 - %R = $\%L_{Total}$ = $\%L_{Ex}$ and $\%L_{Ex} < 0$ were assumed to be 0. Percent loss due to adsorption ($\%L_{Ad}$) were estimated using kaolinite and montmorillonite because they are assumed to have limited microbial activity and their cation exchange capacity and specific surface area values span the values in the representative soil types. Thus, assuming $\%L_{Ex}$ is already known from the method blanks, $\%L_{Ad} = \%L_{Total} - \%L_{Ex}$. $\%L_{Ad}$ estimates for kaolinite and montmorillonite were then linearly scaled with cation exchange capacity so that $\%L_{Ad}$ were scaled for each soil type using a two-point calibration curve. Any $\%L_{Ad} < 0$ were assumed to be 0. The remaining percent loss ($\%L_{Other}$; i.e. to microbial activity, soil chemistry) was estimated as $\%L_{Other} = \%L_{Total} - \%L_{Ex} - \%L_{Ad}$, and any $\%L_{Other} < 0$ were assumed to be 0.

2.5.4. Omics data analysis

UPLC-HRMS *.raw* data files for both H_2O and MeOH extracts were processed using MZmine (V2.53) (Pluskal et al., 2010). H_2O and MeOH extracted samples were processed separately with parameters reported in SI. Exported data were further curated where variables with <70% group detection frequencies were removed. Group detection frequencies are the number of samples with variable intensities greater than a specified threshold (i.e. H_2O extracts = noise level/2 = 2.5E4; MeOH extracts = 2.5E4) divided by the total number of samples in that group times 100%. Variables with >20% group detection frequencies in method blanks and analytical blanks were removed. Number of unique

features (i.e. m/z-retention time bins) were counted per soil type and displayed in a euler diagram using the online *eulerr* (Larsson, 2020). The data were also analyzed using Metaboanalyst to produce principal component analyses (PCA) and hierarchical cluster analyses (HCA) (Chong et al., 2019). Missing values were replaced by a small number (half of the smallest minimum positive value in the matrix), and were filtered using interquartile range, median normalized, and pareto-scaled.

The *vegan* package (Dixon, 2003) was used to execute a Redundancy Analysis (RDA). Soil descriptor variables pH, specific surface area, cation exchange capacity, %TOC, and H_2O extract TOC were chosen because they are independent measurements. Soil descriptor variables were centred and scaled, and a *Hmisc* (Harrell and Dupont, 2020) redundancy analysis was performed to remove collinear variables. Linearity was assumed and an R^2 cutoff of 0.95 was applied. The three remaining variables (i.e. specific surface area, cation exchange capacity, and water extract TOC) were used for the RDA. Model fits were assessed using R^2 values and minimized variance inflation factor (VIF) scores (Table S5.2). RDA and PCA plots were generated using *ggfortify* (Tang et al., 2016) and *ggvegan* (Simpson, 2019).

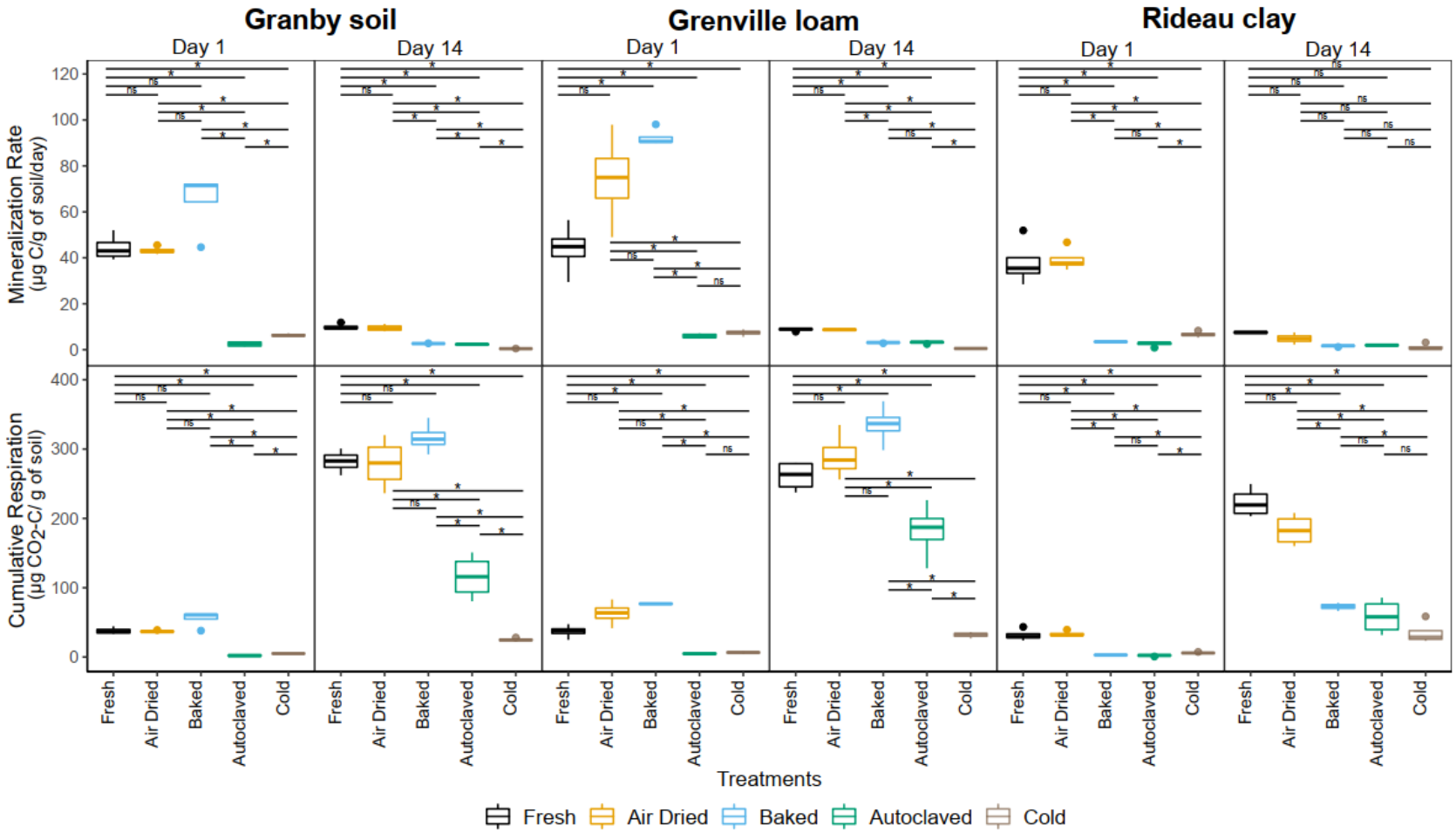


Figure 5.1: The mineralization rates and cumulative respirations after 24 hours and 14 days of the incubation. Significance indicated with * where p-value cut-off points were: 0.0001(****), 0.001(***), 0.01(**), 0.05(*), and not significant(ns).

3. RESULTS

3.1. CO_2 respiration after soil deactivation/sterilization

Statistical analyses indicated that autoclaving and baking the soil affected soil respiration of CO_2 and mineralization rates in the first 24 hours (Figure 5.1; Kruskal-Wallis p-values < 0.0025) and after 14 days (Kruskal-Wallis p-values < 0.0055) in all three soil types. Pairwise comparisons of mineralization rates and cumulative soil respiration determined that fresh and air-dried soils (i.e. controls) were not significantly different in any soil type on Day 1 or 14. However, there was noticeable microbial activity in the first 24 hours where all three fresh soils all respired similarly (37.8 ± 4.7, 36.9 ± 9.3, and 31.6 ± 8.3 µg CO_2-C/g released respectively). This corresponds to organic C (OC) losses of 0.3% (Granby), 0.2% (Grenville), and 0.14% (Rideau), and typically only ~1% OC was extracted using the shaking extraction method (averaged from all soil types and replicates). Thus, ~20% of extractable OC was respired in the first 24 hours (not adjusted for inorganic CO_2 emission) at a rate of ~0.8% OC/hour.

Baking and autoclaving the soils were ineffective at reducing microbial and extracellular enzymatic activity, and, in some cases, significantly increased soil respiration and mineralization. When compared to fresh soil on Day 1, pairwise comparisons indicated respiration in baked soils was significantly greater (Grenville soil), not significantly different (Granby soil), or significantly lower (Rideau soils). On Day 14, the mineralization rates in baked soils were significantly lower (Grenville and Granby soils) or not different (Rideau soil) than in fresh soils. In autoclaved soils mineralization rates (except in Rideau soil) and cumulative soil respiration on Day 1 and 14 were significantly lower compared to fresh soils across all soil types. However, by Day 14, cumulative soil respiration were significantly greater than Day 1 cumulative soil respiration.

Our deactivation procedure (i.e. cold temperature incubation) was effective at reducing soil respiration. Pairwise comparisons demonstrated significantly lower mineralization rates and cumulative respiration in all soil types for Days 1 and 14.

3.2. Microbially-sensitive recovery standard efficacy tests

Experimental quality checks were performed to test the stability of the recovery standards during UPLC-HRMS/DAD analysis and assess the impact of batch effect (Bell and Blais, 2019), and if the duration of extraction affected the results. The term "batch effect" refers to observed differences between replicates that are caused as a result of sample handling, processing, etc., that occur between different time points (batches) rather than inherent differences between samples. Using Kruskal-Wallis and Wilcoxon Signed Rank tests, we determined that Random blocking produced no batch effect between instrument replications during HRMS analysis, and recovery standards were stable over ~1.5 days at room temperature (Figure S5.1). The same tests were used to assess whether shaking duration affected the proportion of recoveries (%R). The dyes tested as microbially-sensitive recovery standard were: 4-Methylumbelliferyl phosphate (MUBph), 2,6-Dichloroindophenol sodium salt hydrate (DCIP), and 3,4-Dihydroxy-L-phenylalanine (LDOPA) and it was determined the duration of shaking had no effect on the recovery of the standards (Figure S5.2). Thus, all %R were merged to compare different soil types (Figure 5.2A).

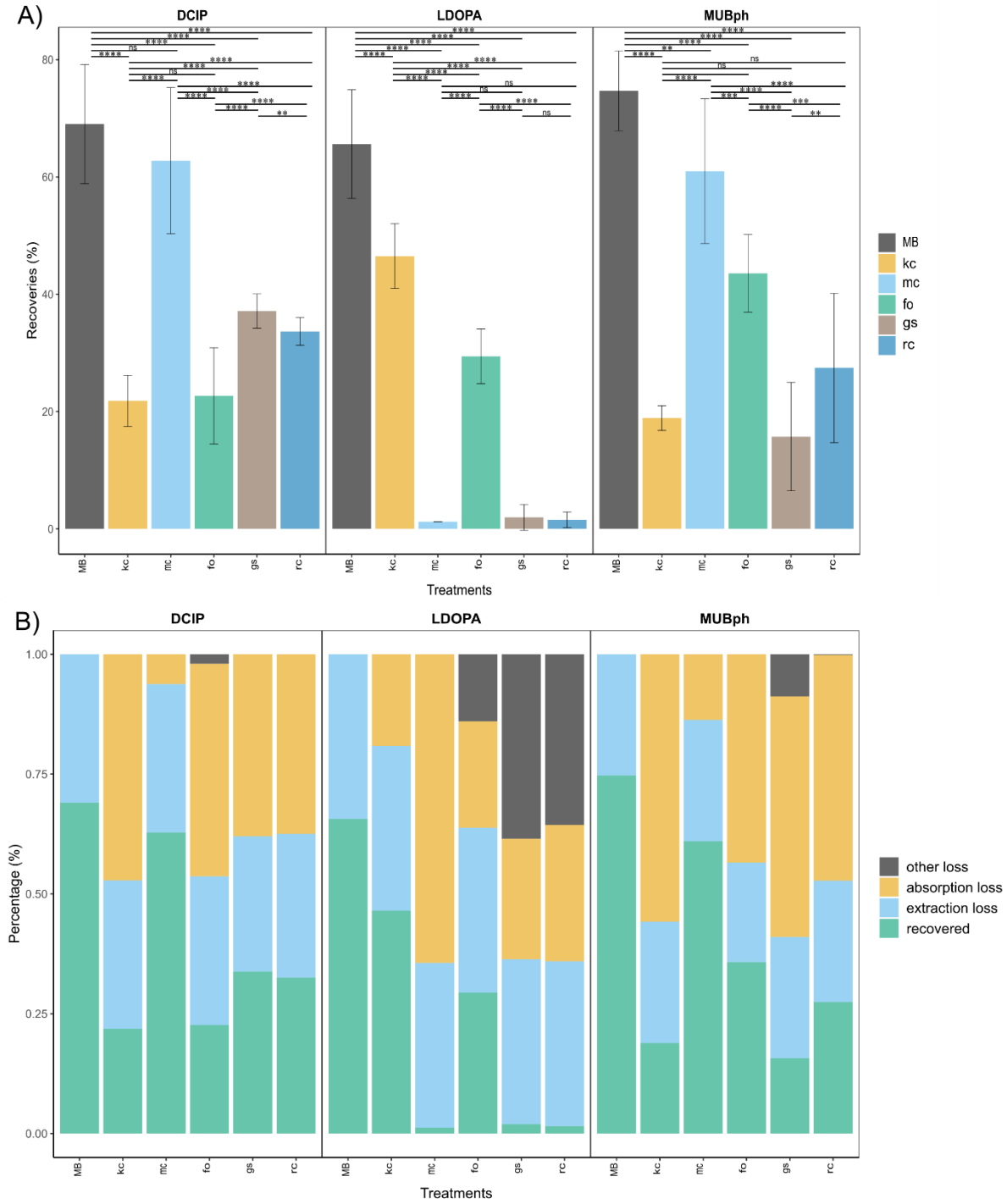


Figure 5.2: A) The % recoveries calculated for each experimental matrix. B) Filled bar plot of the estimated loss of the recovery standard to extraction procedure, adsorption to soil/clay, and other (i.e. biological activity). Treatments were MB = Method blank (no soil/clay), kc = Kaolin Clay, mc = Montmorillonite clay, fo = Granby soil, gs = Grenville loam, and rc = Rideau clay. Significance indicated with * where p-value cut-off points were: 0.0001(****), 0.001(***), 0.01(**), 0.05(*), and not significant(ns).

Soil/clay type had a significant effect on the %R of the recovery standards tested (Kruskal-Wallis p-values: DCIP = 5.90E-19, LDOPA = 1.03E-21, and MUBph = 2.87E-17). All soil/clay %R were significantly less than that observed from the method blanks for all recovery standards. In the clay controls, MUBph and DCIP %R were greater in montmorillonite than kaolinite whereas LDOPA behaviour was opposite. Observed trends in %R of both DCIP and MUBph were significantly different between soil types with the DCIP %R in Grenville soil > Rideau soil > Granby soil and the MUBph %R in Granby soil > Rideau soil > Grenville soil. The LDOPA %R were significantly greater in Granby soil than Grenville and Rideau soil, with no significant difference between the Grenville and Rideau soils.

Total recovery standard loss ($\%L_{Total}$) was partitioned between extraction loss ($\%L_{Ex}$), adsorption loss ($\%L_{Ad}$), and other losses ($\%L_{Other}$; i.e. microbial activity and error relating to $\%L_{Ad}$, etc.). The clay control matrices were used to linearly scale $\%L_{Ad}$ to cation exchange capacity. Both MUBph and DCIP had a negative relationship whereas LDOPA had a positive relationship between $\%L_{Ad}$ and cation exchange capacity. Loss partitioning indicated LDOPA experienced the greatest $\%L_{Other}$, followed by MUBph and DCIP in the fresh soils.

3.3. SOM extraction method comparison

3.3.1. MeOH extracts

Overall, MUBph %R in MeOH extracts were low in the clay controls and soils (Figure 5.3B), and LDOPA was not soluble in MeOH. First, we observed that Randomized reinjections produced no batch effect when comparing between the various metabolite profiles derived from the same MeOH extracts (Figure S5.3; Kruskal-Wallis p-values > 0.005) and that MUBph in MeOH extracts was stable for over 6 days at room temperature. MUBph %R were significantly different across soil types and extraction

methods (Kruskal-Wallis p-value 3.54E-86). MUBph %R were in the order: ultrasonication = shaking > chloroform fumigation, where pairwise comparisons indicated chloroform fumigation MUBph %R were significantly less than both ultrasonication and shaking. All methods had significantly different %R in kaolinite that occurred in the order, shaking > ultrasonication > chloroform fumigation. In montmorillonite MUBph %R were ultrasonication > chloroform fumigation > shaking and all were significantly different. All three soils' MUBph %R were at or near the limit of quantification. Loss partitioning showed $\%L_{Other}$ in the three fresh soils was very similar (Figure 5.3D).

Multivariate analyses were performed on a UPLC-HRMS data matrix [126 samples x 1150 variables] to assess SOM compositional differences between MeOH extraction treatments for each soil type (Figure S5.4; variables represent HRMS mass features with unique retention time and mass-to-charge information). A principal component analysis (PCA) explained 76.5% of the variance in SOM composition between the samples in the first two principal components (PC 1: 57.8%; PC 2: 18.7%). Soil type had a greater influence on differences in SOM composition observed between the three soils tested compared to variations observed in SOM composition arising from the different extraction methodologies used. The various MeOH treatment extracts formed distinct clusters in PC1/PC2 space, due to differences in SOM composition between the soil types: Granby soil extracts separated from Grenville and Rideau soil extracts along PC 1 and Grenville and Rideau soil extracts clustered separately from each other along PC 2. An independent hierarchical clustering analysis (HCA) corroborated the results of the PCA, where MeOH SOM extracts clustered by soil type and not extraction method (Figure S5.4).

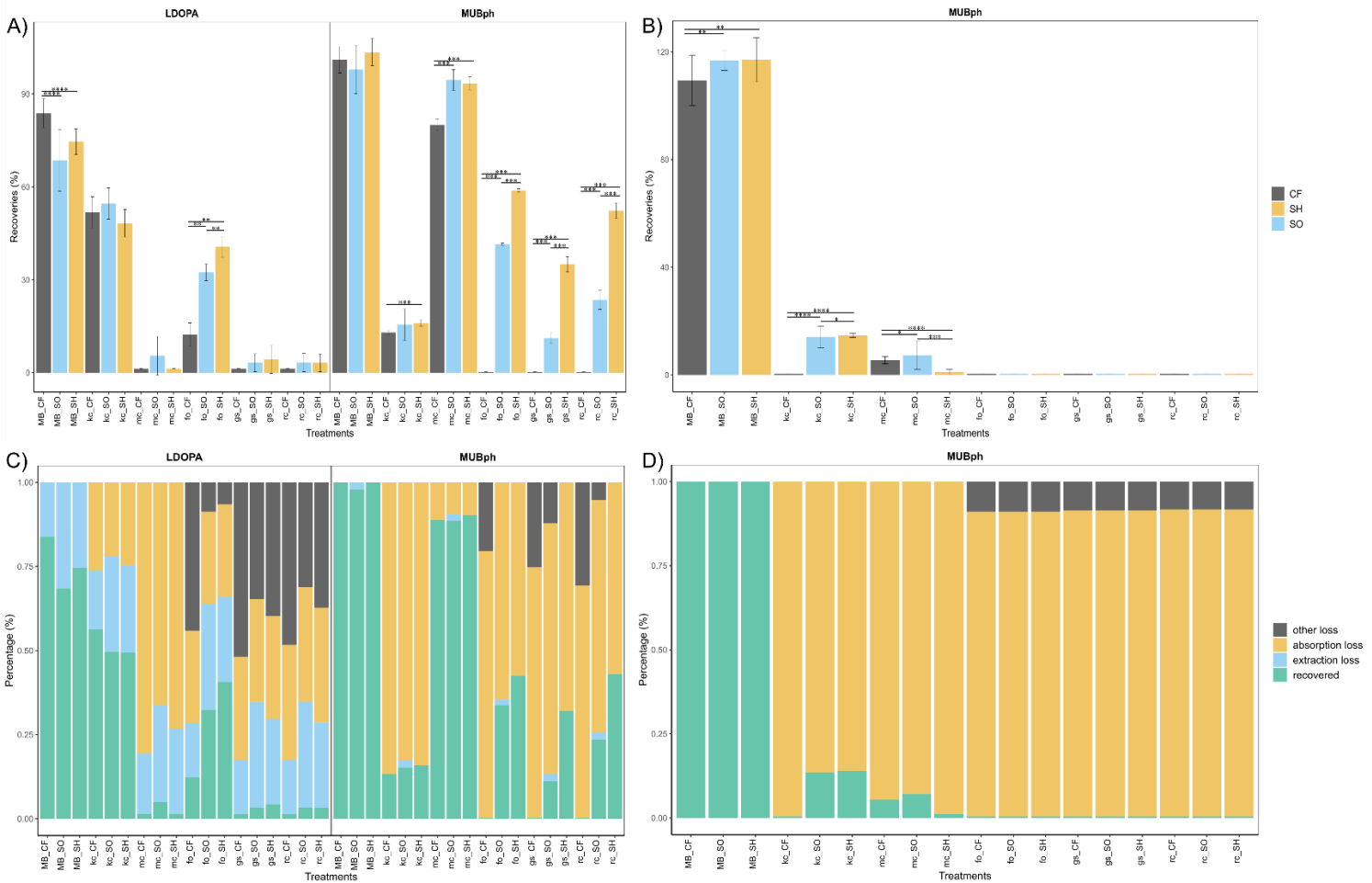


Figure 5.3: The % recoveries calculated for each method in the different experimental matrices from the A) water extraction and HILIC analysis B) methanol extraction and RP analysis. Filled bar plot of the estimated loss of the recovery standard to extraction procedure, adsorption to soil/clay, and other (i.e. biological activity) from the C) water extraction and HILIC analysis D) methanol extraction and RP analysis. CF = chloroform fumigation, SH = shaking, SO = sonication, kc = kaolinite clay, mc = montmorillonite clay, fo = Granby soil, gs = Grenville loam, rc = Rideau clay, and MB = method blank. For the water extraction method, only data from the first replicate run on the instrument were used. Significance indicated with * where p-value cut-off points were: 0.0001(****), 0.001(***), 0.01(**), 0.05(*), and not significant(ns).

3.3.2. H_2O extracts

Contrary to MeOH treatment extracts, randomized reinjections resulted in significant batch effects (a reduction in LDOPA %R (Kruskal-Wallis p-value 0.000845) and MUBph %R (Kruskal-Wallis p-value 8.61E-6) between each of the UPLC-HRMS replicates from the different H_2O extractions (Figure S5.5). In this particular case, the observed batch effects are likely due to an instability of the recovery standards in H_2O at room temperature over time as 3 separate replicate UPLC-HRMS sequences were carried out over 6 days at room temperature to generate the replicates runs. To reduce the impact of the batch effect, we recommend using the Batch randomization method for designing sample runs so that the duration at which samples are held at room temperature between replicates is minimized and where possible to use a temperature-regulated autosampler set to 4°C. Thus, only treatment replicates (rather than treatment and machine replicates combined) were used to calculate the H_2O extract %R data.

Trends in LDOPA %R were consistent with those observed in the LDOPA recovery tests; with a greatest %R from method blanks, followed by kaolinite and Granby soil, while %R from montmorillonite and Grenville and Rideau soils were near the limit of quantification (Figure 5.3A). No significant differences in LDOPA %R were observed between H_2O extraction treatments with the exception of Granby soil (%R shaking > %R ultrasonication > %R chloroform fumigation). LDOPA %R from method blanks was slightly (but significantly) greater in chloroform fumigation compared to both shaking and ultrasonication treatments.

Although trends observed in MUBph %R were also similar to MUBph recovery tests, differences in MUBph %R was more pronounced between H_2O extractions treatments (Figure 5.3A; Kruskal-Wallis p-value 9.56E-101). In all extraction treatments significantly less MUBph was recovered from chloroform treated soils and clay controls compared to both shaking and

ultrasonication treatments; for all three chloroform fumigated soils, MUBph %R was at the limit of quantification. For each of the three soils, MUBph %R was greatest using the shaking method. MUBph %R was greatest from montmorillonite clay, with %R being slightly less to those observed from method blanks (MUBph %R from the three treatments of the method blank samples were not significantly different).

Loss partitioning of MUBph and LDOPA indicated $\%L_{Other}$ is greatest for chloroform fumigation extractions then ultrasonication (especially for MUBph), and lowest for shaking extractions (Figure 5.3C). The greatest differences between $\%L_{Other}$ for shaking and chloroform fumigation were observed in Grenville, Granby, and Rideau soils, not in the method blanks or the clay controls.

To assess the extent to which SOM H_2O extracts differed in molecular composition, multivariate analyses were performed upon the H_2O extract UPLC-HRMS derived data matrix [126 samples x 512 variables]. The PCA model explained 47.7% of sample variance in the first two principal components (PC 1: 25.6%; PC 2: 22.1%). Replicates from the same soil type/extraction method clustered together and separated from each other with minimal overlapping of the 95% confidence bubbles indicating significant SOM compositional differences between all groups (Figure 5.4A). Thus, the SOM composition of H_2O extracts are strongly affected by both soil type and by extraction method. An independent HCA corroborated this observation as major branching events clustered extraction methods before soil types, especially for chloroform fumigation (Figure S5.5). Based on the clustering analyses, the composition of SOM in chloroform fumigation H_2O extracts are more similar between soil types, compared to H_2O extracts from shaking and ultrasonic treated soils. In addition, Grenville and Granby soils were more strongly influenced by extraction method than Rideau soil, based on the distance

between the clusters of the same soil type. The Euler diagrams (Figure 5.5) indicate the number of unique mass features that were specific to an extract or were shared between extracts. The majority of all features were common between all extracts for all soils; however, chloroform fumigation, in support of evidence from the PCA and HCA, had the largest number of unique mass features in all three soils, followed by shaking and then ultrasonication. Chloroform fumigation shared more features with ultrasonication than shaking in all soils. Ultrasonication shared more features with chloroform fumigation than shaking in Grenville and Granby loams, but not in Rideau clay.

A redundancy analysis (RDA) model incorporated soil characteristics (i.e. pH, cation exchange capacity, specific surface area, extractable OC and TN) into the PCA (Figure 5.4B). The RDA model determined that 36.85% of the variance could be constrained by cation exchange capacity, specific surface area, and extractable OC (adjusted R^2: 0.4406; model significance p-value: 0.001; Table S5.2). Cation exchange capacity and specific surface area are soil properties (Table 5.1), while extractable OC is an extraction property (Table 5.2). The length of the arrows is similar indicating they have similar degrees of influence on the variation between groups, and the arrows point in the direction of the group with the largest values.

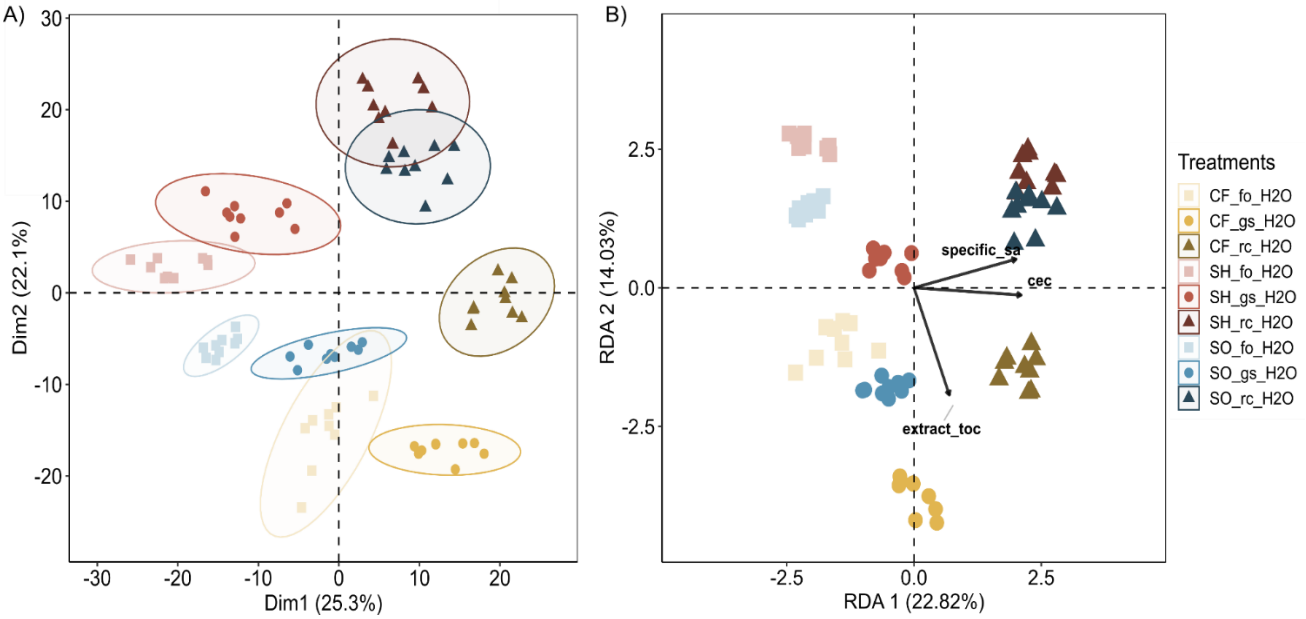


Figure 5.4: A) PCA of the method comparison experiment, where the symbols represent biological and machine replicates. B) RDA of the method comparison experiment incorporating soil parameters. CF = chloroform fumigation, SH = shaking, SO = sonication, fo = Granby soil, gs = Grenville loam, and rc = Rideau clay. Colour families represent the same extraction method (i.e. CF = yellows, SH = reds, and SO = blues). Shapes represent soil types (i.e. fo = squares, gs = circles, rc = triangles).

4. DISCUSSION

4.1. Microbial activity during extractions in soil metabolomics

Soil respiration is linked to microbial activity and growth (Sinsabaugh et al., 2013; Wang et al., 2016), so we used soil incubation studies to determine the impact autoclaving, baking and deactivation (i.e. cold) procedures have on microbial processes when compared to controls (e.g. fresh and air-dried soils). Any microbial activity occurring during SOM extraction would likely affect both the quantity (yield) and quality (composition) of the SOM analyzed; if microbial activity is a factor, then treatments to minimize/reduce microbial activity during the extraction process could be incorporated in untargeted soil metabolomics methods.

Our results indicated that soil microbial activity reduced SOM yield at a rate of ~0.8%/hr; however, none of the procedures were completely effective at reducing soil respiration; baking soils actually increased soil respiration. The observed enhanced microbial activity in baked soils may be due to increased thermal conductance in Granby and Grenville soils because of their relatively high sand content (Abu-Hamdeh and Reeder, 2000; Tarnawski and Leong, 2000) (Table 5.1). Greater thermal conductance and heating can increase the oxidation of SOM (Certini, 2005; O'Brien et al., 2018) and release more labile material for microbial growth (Kiersch et al., 2012). Autoclaving was more effective at reducing microbial activity in the soil than baking across all soil types in the first 24 hours, but the microbial community recovered over time suggesting incomplete sterilization of heat-resistant spores even after two autoclaving cycles (Wolf and Skipper, 2018), which is consistent with past literature (Bach et al., 2013; Lees et al., 2018). Moreover, autoclaving is associated with modifications to SOM (Berns et al., 2008; Lees et al., 2018; Wolf and Skipper, 2018). Thus, both baking and autoclaving were ineffective at completely sterilizing the soils but they likely modified the SOM composition. These procedures are not recommended to be used prior to soil extraction to reduce microbial activity in soil metabolomics experiments.

Our soil deactivation procedure with cold exposure was more effective than baking and autoclaving. The observed reduction of soil respiration in cold soil was consistent with previous literature (Yanni et al., 2018), and is facilitated by a reduction in microbial growth (Rousk and Bååth, 2011) and enzymatic conversion of metabolites (Sitnikov et al., 2016). Cold exposure at 4°C is also not associated with SOM modification; however, the solubility of organic molecules changes with temperatures which can affect extraction efficiency. Cold extractions are associated with protein precipitation (Sitnikov et al., 2016; Warren, 2015), and room

temperature water extractions often extracted less (Hamkalo and Bedernichek, 2014; Landgraf et al., 2006) and different (Landgraf et al., 2006; Zhao et al., 2013) SOM than hot water extractions; thus, it is likely that SOM water extractions at 4°C would extract less SOM than either room temperature or hot water extractions.

In summary, microbial activity can influence extractable SOM yields and in turn influence results in soil metabolomics studies of SOM. The procedures assessed here were not effective at halting microbial activity and will likely modify SOM composition. Cold temperatures effectively reduced respiration, but at the expense of lower SOM extraction yields. However, cold treatments were superior to autoclaving and baking at reducing microbial activity. Hence we suggest reducing the impact of microbial modification of SOM during extractions by: 1) reducing the duration of the extraction, 2) using cold solvents or keeping samples on ice, and/or 3) tracking the loss of SOM to microbial activity through the use of microbially-sensitive recovery standards.

4.2. Incorporation of microbially-sensitive recovery standards and control matrices

Microbial activity modifying SOM during the extraction process is an important consideration in soil metabolomics studies. We propose that the extent of this activity can be quantified by using microbially-sensitive recovery standards spiked into soils before SOM extraction. These microbially-sensitive recovery standards are dyes/substrates (Burns et al., 2013; Sinsabaugh, 2010) that are used to quantify soil microbial and/or exoenzyme activity (Jian et al., 2016; Stone et al., 2014; Tapia-Torres et al., 2015; Zhou et al., 2013). Three dyes/substrates were chosen because they were relatively inexpensive, used in soil experiments before, and sensitive to microbial activities in the soil. DCIP tracks electron transport chain

activity (Tang et al., 1991) and has been used in soil extractions (Crawford et al., 2002). MUBph measures soil phosphatase exoenzyme activity (Saiya-Cork et al., 2002; Sinsabaugh et al., 2003; Sorensen et al., 2018), and LDOPA measures soil phenol oxidase and peroxidase exoenzyme activities (Bach et al., 2013; Sinsabaugh, 2010; Wiedermann et al., 2017).

We also introduce the use of mined clays as control (i.e. microbially inactive) matrices to estimate the loss of recovery standards to adsorption. Pure kaolinite and montmorillonite clays were chosen as control matrices because their cation exchange capacities and specific surface areas (Table 5.1) encompassed those of the three soil types (Grandby, Grenville, and Rideau soils). Both clays had near zero quantities of C, OC, and N (Table 5.1), so we assumed negligible SOM content and a minimal capacity for microbial activity compared to the other soils, which is similar to assumptions from previous studies (Kallenbach et al., 2016).

The results of the microbially-sensitive recovery standard efficacy tests indicated that soil/clay type affected the proportion of standards recovered (%R). In addition, loss partitioning was used to breakdown total losses $\%L_{Total}$ ascribed to extraction ($\%L_{Ex}$), adsorption ($\%L_{Ad}$), and other losses ($\%L_{Other}$; i.e. microbial activity, etc.) to assess the interactions between minerology, SOM, and microbial activity. The clay control matrices were used to linearly scale $\%L_{Ad}$ to cation exchange capacity, where MUBph and DCIP behaved opposite to LDOPA, probably because of differences in clay structure. Montmorillonite (compared to kaolinite) has more surface area and ion-exchange sites on the clay surface to intercalate small amphoteric molecules like LDOPA (Hedges and Hare, 1987). However, larger, more hydrophobic molecules, such as MUBph and DCIP, would form outer-surface complexes and therefore have a lower bonding energy on montmorillonite compared to kaolinite (Zhang et al., 2014). Loss partitioning indicated LDOPA experienced the greatest $\%L_{Other}$, followed by MUBph and DCIP in the fresh soils. According to

the soil incubation experiment, Grenville soil had the greatest soil respiration of the three soils in 24 hours and the greatest $\%L_{Other}$ in LDOPA and MUBph (Figure 5.1). Soil respiration was greater in Granby soil compared to Rideau soil; however, Rideau soil experienced a greater $\%L_{Other}$ with LDOPA and MUBph than Granby soil. This discrepancy could be due to differential SOM degradation rates resulting from differences in microbial community composition (Wu et al., 2018), differences in SOM composition and lability, or to errors associated with estimating $\%L_{Ad}$.

The $\%L_{Ad}$ was dependent on the chemical structure of the recovery standards, the proportion of clay in the soil, as well as the amount and composition of SOM (i.e. OC; (Abate et al., 2004; Wen et al., 2007; Zhang et al., 2017, 2014)), which was not accounted for when estimating $\%L_{Ad}$ from the clay controls. DCIP and MUBph did not demonstrate high $\%L_{Other}$ suggesting that they are less susceptible to degradation during the extraction process than LDOPA. This resistance to degradation is further supported by the absence of a significant difference between different extraction durations (Figure S5.2). There is also evidence that the estimated $\%L_{Ad}$ of DCIP and MUBph may be biased. Assuming that only cation exchange capacity explained changes in %R, then DCIP and MUBph %R would follow the order: Rideau soil > Grenville soil > Granby soil. However, DCIP %R in Rideau soil was significantly less than Grenville soil indicating differences in soil OC content may have influenced DCIP adsorption. Grenville soil had 1.5 times more OC than Rideau soil so OC could be blocking adsorption sites on Grenville soil resulting in an increased recovery of DCIP, similar to previous observations of perfluorooctane sulfonate adsorption on kaolinite and montmorillonite (Zhang et al., 2014). A similar effect was observed with MUBph $\%L_{Ad,}$ where MUBph %R followed the order: Granby soil > Rideau soil> Grenville soil, which correlates with OC (Table 5.1), suggesting that MUBph adsorption increased

with increasing OC, similar to that of other small aromatic molecules such as phenanthrene and 1,2-dichlorobenzene adsorption by different clays (Zhang et al., 2017).

To our knowledge, this work represents the first soil metabolomics study that incorporates recovery standards, and it is the first to use microbially-sensitive recovery standards in any metabolomics study. The recovery standards were extractable using a common SOM extraction technique and were easily detectable using both HRMS and a photo-diode array detector (DAD). Some recovery standards were more sensitive to microbial activity than others, but all recovery standards indicated significant differences in SOM extraction efficiencies among soil types. Based on these findings, we suggest that caution be used when comparing different soil types because extraction recoveries can vary widely depending on the soil chemical properties, and recommend that recovery standards be used to evaluate these differences. This work is also the first to use clay control matrices for loss partitioning and with some refinement, loss partitioning could be a critical tool to use for assessing differences in SOM behaviour across different soil types

4.3. Comparison of different extraction methods

There is no universally accepted approach for extracting SOM from soil. Our purpose in this study was to determine how well commonly used SOM extraction techniques performed when using different soils. Chloroform fumigation prior to solvent extraction has been used for soil metabolomics studies (Swenson et al., 2015), and is usually used to determine soil microbial biomass (Blankinship et al., 2014; von Lützow et al., 2007). Shaking is used frequently in SOM extraction protocols (Fox et al., 2017; Swenson et al., 2015; Tfaily et al., 2017, 2015), and ultrasonication extraction is commonly applied in organismal metabolomics studies (Esclapez et

al., 2011; Gong et al., 2017) and sediment-omics (Farrés et al., 2015; Saleem et al., 2019). To compare the methods, the experiment was performed separately for H_2O (to assess the water extractable SOM) and MeOH (to assess the hydrophobic SOM fraction), and an LDOPA and MUBph mix was spiked into method blanks, control matrices (kaolinite, montmorillonite), and Granby, Grenville and Rideau soils.

Overall, we determined that extraction method had a significant impact on the %R of recovery standards in MeOH and H_2O extracts, but H_2O extracts were more affected by different SOM extraction methods. For both extracts, typically shaking extractions had greater %R than either ultrasonication or chloroform fumigation. Importantly, the greatest differences in extraction %R was in the three soils, not method blanks or clay controls, indicating soil OC and microbial activity were critical factors influencing %R and ultimately, the yield and composition of SOM extracted as well.

We used loss partitioning to investigate these changes. In MeOH extracts, $\%L_{Other}$ does not seem to capture microbial activity likely because 100% MeOH may reduce/inhibit microbial activity during extraction via dehydration of cell membranes, coagulation of enzymes, and protein precipitation. Thus, a larger quantity of recovery standard may need to be spiked into MeOH samples, and/or other types of recovery standards should be developed for MeOH extractions. In H_2O extracts, loss partitioning indicated $\%L_{Other}$ is greatest for chloroform fumigation extractions, followed by ultrasonication, then shaking (especially in MUBph; Figure 5.3C); where the greatest differences between $\%L_{Other}$ for shaking and chloroform fumigation were observed in Grenville, Granby, and Rideau soils, not in the method blanks or the clay controls. The Euler diagrams (Figure 5.5) also indicated more unique features in chloroform fumigation than those found in the sonication and shaking methods. These results suggest that

SOM extracted by chloroform fumigation is more susceptible to microbial/exoenzymatic activity during extraction than in ultrasonication or shaking extractions. There is some evidence for this because chloroform fumigation lyses microbial cells, but it does not inhibit exoenzyme activity (Blankinship et al., 2014; Klose and Tabatabai, 2002a, 2002b, 2002c; Renella et al., 2002). Specifically, phosphatases, phenol oxidases, and peroxidases remain active in the soil after chloroform fumigation (Blankinship et al., 2014). Ultrasonication also lyses cells through cavitation (Esclapez et al., 2011; Gong et al., 2017), although water-bath ultrasonication lyses fewer cells than probe sonication. The similarity between ultrasonication and chloroform fumigation is also highlighted in Figure 5.5, where more features are shared between the two methods that are known to lyse cells, than shaking. Thus, the lyses of microbial cells in the soil may stimulate increased losses of recovery standards (and labile SOM) through digestion by exoenzymes and/or change adsorption characteristics because the quantity and composition of the OC changed after lysing the cells (Blankinship et al., 2014; von Lützow et al., 2007). Shaking, which does not lyse cells, did not result in increased loss of SOM to exoenzymes and produced higher %R.

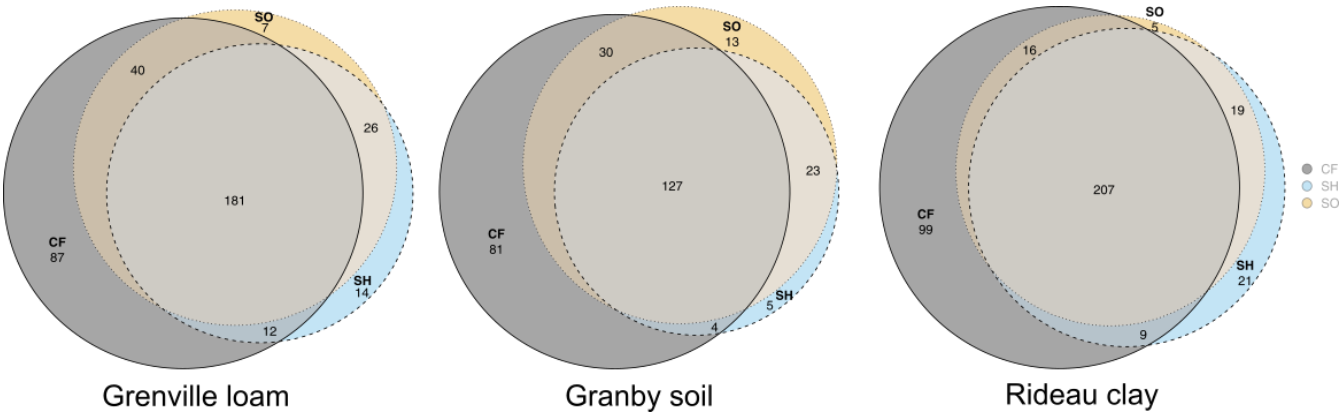


Figure 5.5: Euler diagrams depicting the number of features (i.e. unique m/z-retention time bins) detected by the UPLC-HRMS for each soil type. CF = chloroform fumigation, SH = shaking, SO = sonication.

When we assessed the composition of SOM in MeOH and H_2O extracts using clustering analyses, we determined that the SOM composition of H_2O extracts were more strongly influenced by extraction method than MeOH extracts. H_2O SOM extracts have proven to be susceptible to extraction method in the past (Fox et al., 2017; Guigue et al., 2014; Hamkalo and Bedernichek, 2014; Landgraf et al., 2006; Zhao et al., 2013), but to our knowledge this is the first study to compare soil metabolomics extraction procedures across different soil types. Specifically, we determined the composition of SOM extracted by H_2O is most different between the shaking and chloroform fumigation extractions, suggesting that microbial cell lysis strongly influences SOM composition. In addition, Grenville and Rideau soils were more strongly influenced by extraction method than Rideau soil, based on the distance between the clusters of the same soil type. To explore these differences, we used an RDA to determine which soil (i.e. pH, cation exchange capacity, specific surface area) and extract (i.e. extractable OC and TN) characteristics explain these compositional differences. Cation exchange capacity and specific surface area were important for explaining the differences in SOM composition between soil types, whereas extractable OC explained most of the SOM compositional difference between extraction methods. Extractable OC is influenced by cell lysis, where cell lysis typically produces greater SOM yields (Table 5.2). These findings indicate that cell lysis plays a major role in the SOM composition of the extracts, but at the cost of increased microbial modification of SOM compounds during the extraction, as demonstrated by greater $\%L_{Other}$ indicated by the recovery standards.

5. CONCLUSIONS

Our purpose in this work was to address some methodological issues associated with the extraction of SOM that may affect accurate characterization of its chemical composition. Our findings show that even after autoclaving and baking there is enough microbial activity during extraction to significantly affect the amount and composition of extractable SOM. Reducing soil activity with cold treatments is more effective than autoclaving and baking procedures, but one drawback is that it can reduce the quantity of SOM extracted. To compensate for loss of SOM to microbial activity we used microbially-sensitive recovery standards to track microbial activity during extraction and pure clays as control matrices to partition losses of the standards. We found that soil type strongly influences recovery standard and SOM recovery. The use of recovery standards also indicated that extraction method strongly influenced recovery standard and SOM recoveries, especially in H_2O extracts. Extraction by chloroform fumigation had the greatest loss of recovery standards, due to microbial activity and/or adsorption. Extraction by shaking had the smallest loss of recovery standards and SOM; we therefore recommend it as the extraction method of choice for soil metabolomics studies.

ACKNOWLEDGEMENTS

Funding for this study was provided by the Science & Technology Branch of Agriculture & Agri-Food Canada [Project J-001756 “Biological Soil Carbon Stabilization”] to EG and DPO.

SUPPLEMENTARY INFORMATION

Table S5.1: RS quantification results from the mass spectrometer (m/z) and photodiode array (nm). Where m= slope, y0 = y-intercept, LOD = limit of detection, and LOQ = limit of quantification.

RS	Experiment	Solvent	Quant peak	Retention Time (min)	m	R^2	LOD	LOQ
DCIP	RS Tests	H_2O	267.99 m/z	0.86	7.40E+08	0.969	1.75E-02	5.83E-02
DCIP	RS Tests	H_2O	270 nm	0.92	7.79E+06	0.998	3.76E-05	1.25E-04
LDOPA	RS Tests	H_2O	198.08 m/z	9.69	6.77E+08	0.961	4.81E-04	1.60E-03
LDOPA	RS Tests	H_2O	280 nm	9.72	3.16E+06	0.999	1.95E-02	6.49E-02
MUBph	RS Tests	H_2O	257.02 m/z	4.51	1.27E+09	0.951	2.05E-03	6.84E-03
MUBph	RS Tests	H_2O	315 nm	5.06	1.12E+07	1.000	1.06E-03	3.52E-03
LDOPA	Method Comp	H_2O	198.08 m/z	9.69	6.94E+08	0.982	6.73E-04	2.24E-03
LDOPA	Method Comp	H_2O	280 nm	9.72	3.17E+06	0.999	1.87E-02	6.25E-02
MUBph	Method Comp	H_2O	257.02 m/z	4.51	1.40E+09	0.978	2.37E-03	7.91E-03
MUBph	Method Comp	H_2O	315 nm	5.06	1.01E+07	0.998	1.96E-03	6.52E-03
MUBph	Method Comp	MeOH	257.02 m/z	1.07	9.86E+08	0.956	3.35E-03	1.12E-02

Table S5.2: RDA model estimates where Model A contains all of the soil descriptor variables, and Model B contains only the non-redundant soil descriptor variables according to the Hmisc *redun* function.

Parameters	A	B
Number of Soil descriptor variables	5	3
% RDA1 Variance	22.82	22.82
% RDA2 Variance	14.03	14.03
% Total Variance	36.85	36.85
P-value	0.001	0.001
Adjusted R^2	0.4406	0.4406
pH	0.001	X
Specific_sa	0.001	0.001
Cec	Aliased	0.001
Soil_p_toc	Aliased	X
Extract_toc	0.001	0.001

MzMine 2.53 Data Analysis Parameters

H2O extracts:

1. Mass detection (Wavelet transform; noise level = 5E4, scale level = 6, and wavelet window = 30%)
2. ADAP chromatogram builder (group size = 5, group intensity threshold = 5E4, min. highest intensity = 6E4, m/z tolerance = 0.005)
3. Deconvolution (Wavelets ADAP; S/N threshold = 10, min. feature height = 5E4, area threshold = 45, peak duration range = 0.02-3, and RT wavelet range = 0.01-0.2)
4. Deisotoping (Isotope peaks grouper; m/z tolerance 60 ppm, retention time tolerance = 0.1, max charge = 1)
5. Alignment (Ransac aligner; m/z tolerance = 5 ppm, retention time tolerance = 2, retention time tolerance after correction = 1, RANSAC iterations = 0, minimum number of points = 60%, threshold value = 1, not assumed to be linear)
6. Gap-filling (Same RT and m/z range gap filler; m/z tolerance = 10 ppm)
7. Deduplicated (Duplicate peak filter; m/z tolerance = 0.01 m/z, retention time tolerance = 2)
8. Deadduct (Adduct search; retention time tolerance = 0.1, m/z tolerance = 5 ppm, max relative peak height = 50%)
9. export peak heights.

MeOH extracts:

1. Mass detection (Wavelet transform; noise level = 1E4, scale level = 6, and wavelet window = 30%)
2. ADAP chromatogram builder (group size = 5, group intensity threshold = 5E4, min. highest intensity = 6E4, m/z tolerance = 0.005)
3. Deconvolution (Wavelets ADAP; S/N threshold = 10, min. feature height = 5E4, area threshold = 45, peak duration range = 0.02-1, and RT wavelet range = 0.01-0.1)
4. Deisotoping (Isotope peaks grouper; m/z tolerance 60 ppm, retention time tolerance = 0.1, max charge = 1)
5. Alignment (Ransac aligner; m/z tolerance = 5 ppm, retention time tolerance = 1, retention time tolerance after correction = 0.5, RANSAC iterations = 0, minimum number of points = 60%, threshold value = 1, not assumed to be linear)
6. Gap-filling (Same RT and m/z range gap filler; m/z tolerance = 10 ppm)
7. Deduplicated (Duplicate peak filter; m/z tolerance = 0.01 m/z, retention time tolerance = 0.1)
8. Deadduct (Adduct search; retention time tolerance = 0.1, m/z tolerance = 5 ppm, max relative peak height = 50%)
9. 9) Export peak heights.

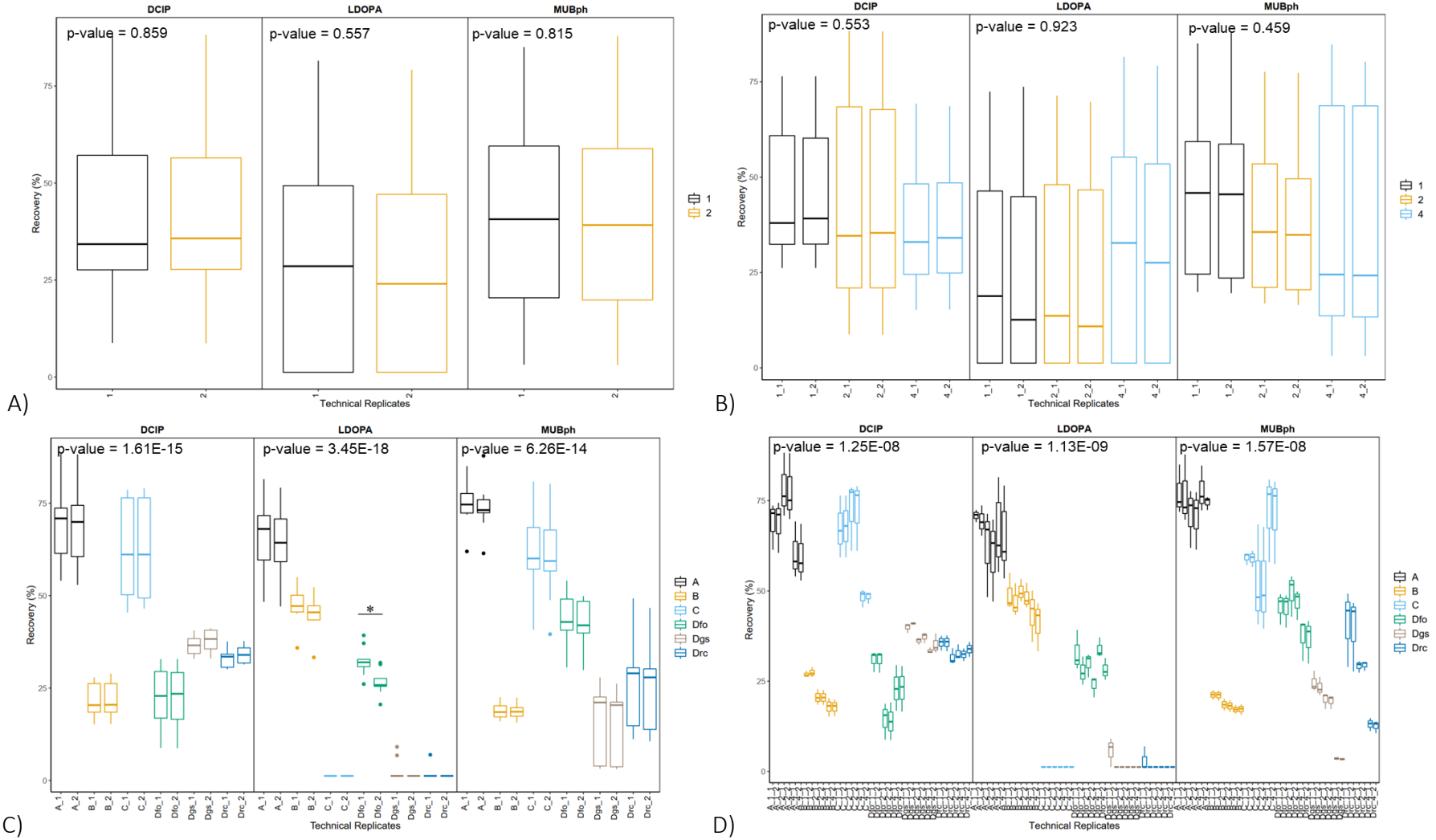


Figure S5.1: Boxplots comparing the %R between A) machine replicates for all msRS, B) machine replicates for each extraction duration (i.e. 1_1 = extraction duration of 1 hour and machine replicate 1) for all msRS, C) machine replicates for each treatment, and D) machine replicates for treatment and extraction duration. Samples injected using the Randomized block method. Significance indicated with * where p-value cutoff-points were: 0.0001(****), 0.001(***), 0.01(**), 0.05(*), and not significant(ns).

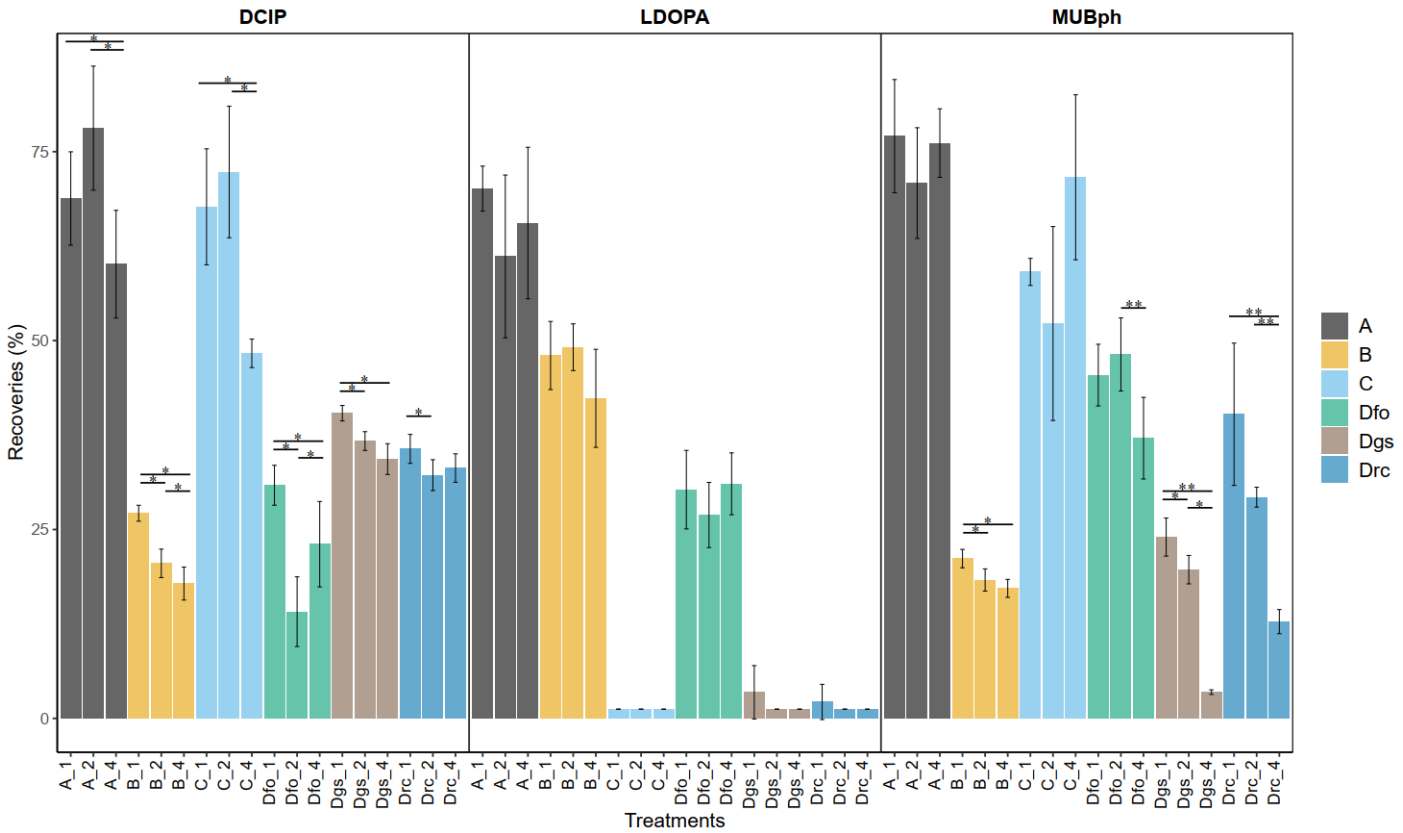


Figure S5.2: The % recoveries calculated for each experimental matrix and each extraction duration (i.e. 1 hours, 2 hours, and 4 hours). Treatments were A = Method blank (no soil/clay), B = Kaolin Clay, C = Montmorillonite clay, Dfo = Granby soil, Dgs = Grenville soil, and Drc = Rideau Clay. Significance indicated with * where p-value cutoff-points were: 0.0001(****), 0.001(***), 0.01(**), 0.05(*), and not significant(ns). KW p-values: DCIP = 0.14, LDOPA = 0.621, and MUBph = 0.1.

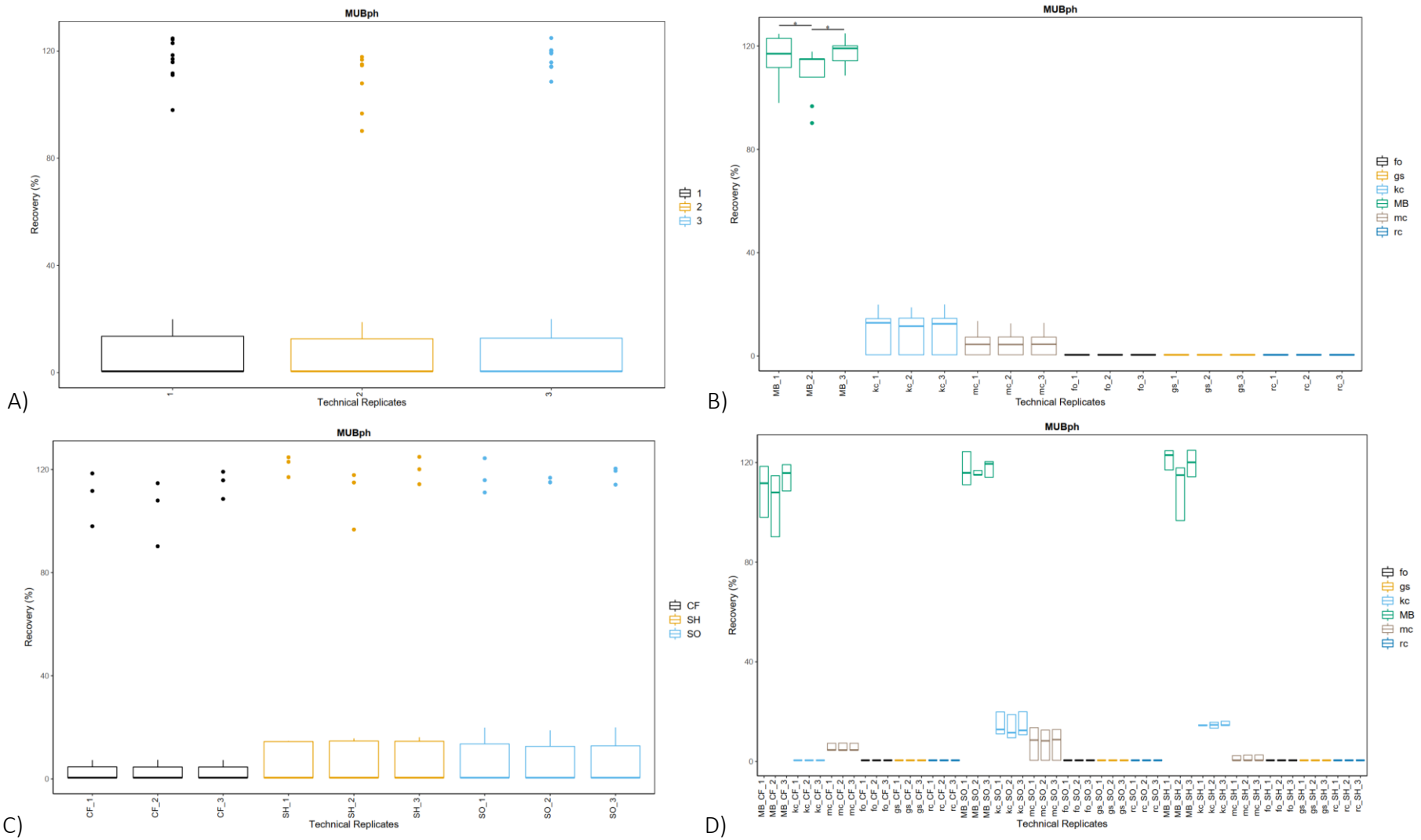
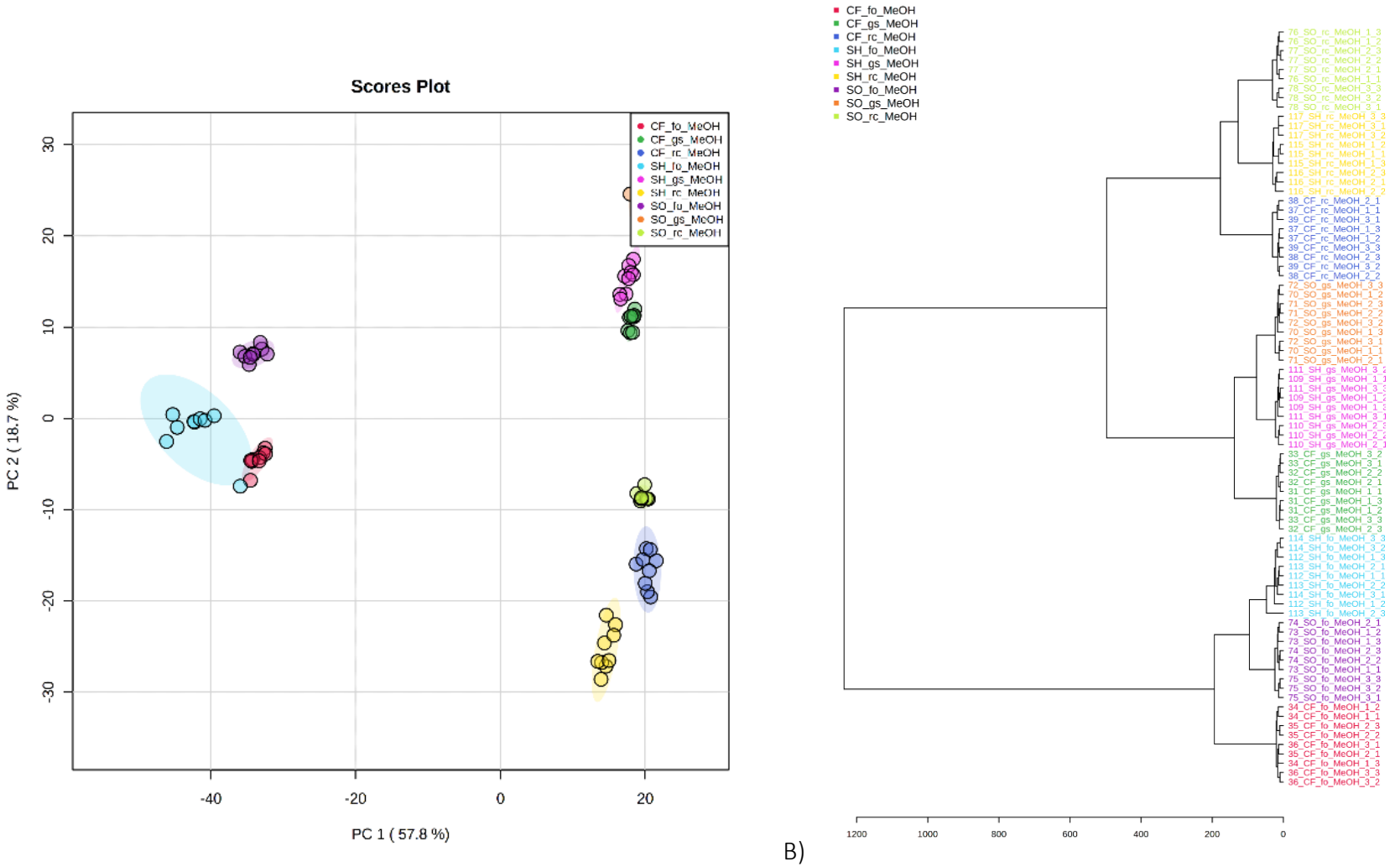


Figure S5.3: Boxplots from MeOH extracted method comparison extraction data comparing the % recoveries between machine replicates generated using the Randomized reinjection method: A) for all data, B) for each soil type, C) for each extraction method, D) for each soil type and extraction method. Significance indicated with * where p-value cutoff-points were: 0.0001(****), 0.001(***), 0.01(**), 0.05(*), and not significant(ns).



A) B)

Figure S5.4: A) PCA of the MeOH extracts from the method comparison experiment, where the symbols represent biological and machine replicates. B) Hierarchical clustering analysis of extracted SOM from MeOH extracted soils. CF = chloroform fumigation, SH = shaking, SO = sonication, fo = Field One Granby soil, gs = Grenville soil, and rc = Rideau clay.

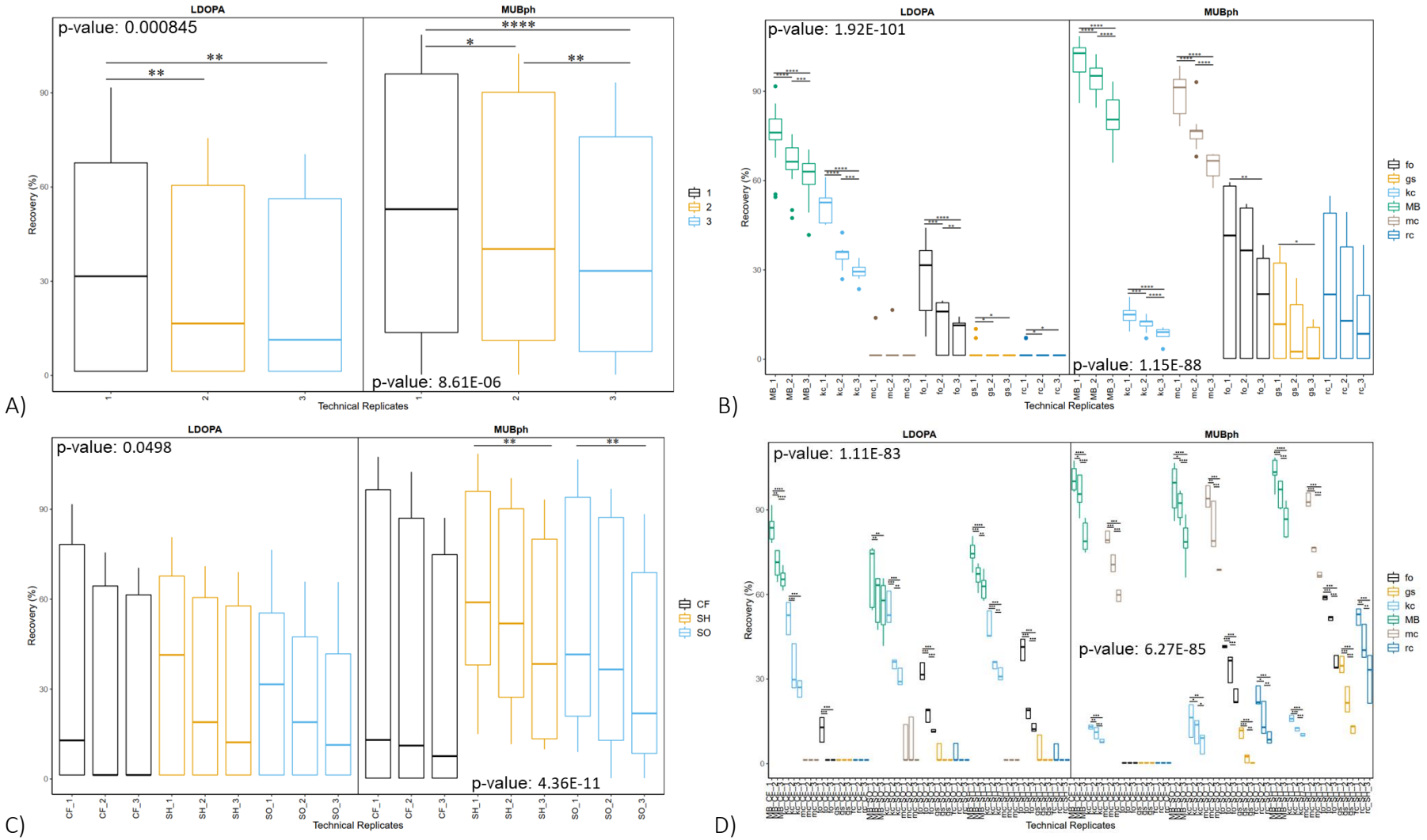


Figure S5.5: Boxplots from H_20 extracted method comparison extraction data comparing the % recoveries between machine replicates generated using the Randomized reinjection method A) for all data, B) for each soil type, C) for each extraction method, D) for each soil type and extraction method. Significance indicated with * where p-value cutoff-points were: 0.0001(****), 0.001(***), 0.01(**), 0.05(*), and not significant(ns).

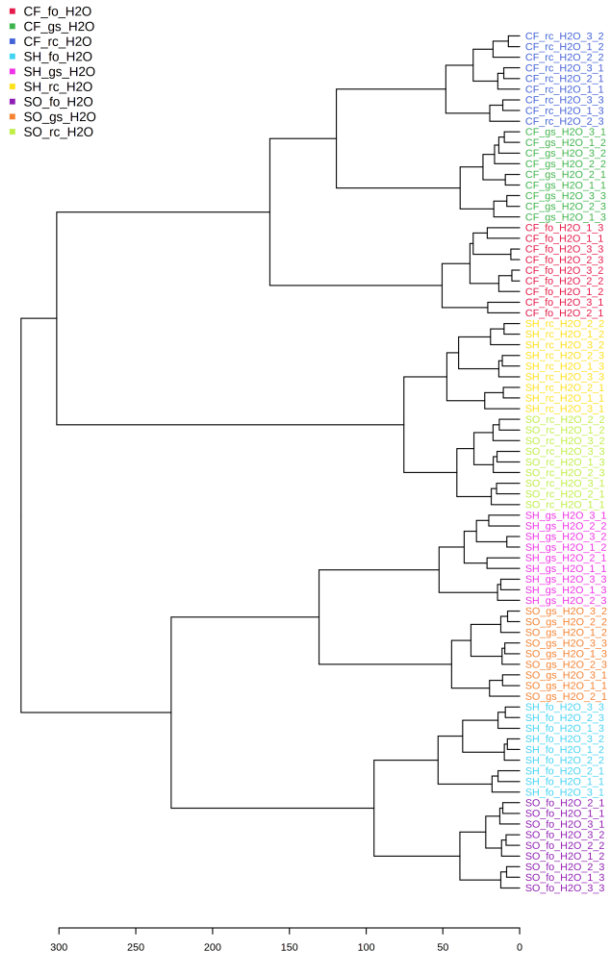


Figure S5.6: Hierarchical clustering analysis of extracted SOM from H_2O extracted soils. CF = chloroform fumigation, SH = shaking, SO = sonication, fo = Field One Granby soil, gs = Grenville soil, and rc = Rideau clay.

REFERENCES

Abate, G., Penteado, J.C., Cuzzi, J.D., Vitti, G.C., Lichtig, J., Masini, J.C., 2004. Influence of Humic Acid on Adsorption and Desorption of Atrazine, Hydroxyatrazine, Deethylatrazine, and Deisopropylatrazine onto a Clay-Rich Soil Sample. Journal of Agricultural and Food Chemistry 52, 6747–6754. doi:10.1021/jf049229e

Abu-Hamdeh, N.H., Reeder, R.C., 2000. Soil Thermal Conductivity Effects of Density, Moisture, Salt Concentration, and Organic Matter. Soil Science Society of America Journal 64, 1285–1290. doi:10.2136/sssaj2000.6441285x

Bach, C.E., Warnock, D.D., Van Horn, D.J., Weintraub, M.N., Sinsabaugh, R.L., Allison, S.D., German, D.P., 2013. Measuring phenol oxidase and peroxidase activities with pyrogallol, l-DOPA, and ABTS: Effect of assay conditions and soil type. Soil Biology and Biochemistry 67, 183–191. doi:10.1016/j.soilbio.2013.08.022

Bell, M., Blais, J.M., 2019. "-Omics" workflow for paleolimnological and geological archives: A review. Science of The Total Environment 672, 438–455. doi:10.1016/j.scitotenv.2019.03.477

Berns, A.E., Philipp, H., Narres, H.-D., Burauel, P., Vereecken, H., Tappe, W., 2008. Effect of gamma-sterilization and autoclaving on soil organic matter structure as studied by solid state NMR, UV and fluorescence spectroscopy. European Journal of Soil Science 59, 540–550. doi:10.1111/j.1365-2389.2008.01016.x

Blankinship, J.C., Becerra, C.A., Schaeffer, S.M., Schimel, J.P., 2014. Separating cellular metabolism from exoenzyme activity in soil organic matter decomposition. Soil Biology and Biochemistry 71, 68–75. doi:10.1016/j.soilbio.2014.01.010

Burns, R.G., DeForest, J.L., Marxsen, J., Sinsabaugh, R.L., Stromberger, M.E., Wallenstein, M.D., Weintraub, M.N., Zoppini, A., 2013. Soil enzymes in a changing environment: Current knowledge and future directions. Soil Biology and Biochemistry 58, 216–234. doi:10.1016/j.soilbio.2012.11.009

Certini, G., 2005. Effects of fire on properties of forest soils: a review. Oecologia 143, 1–10. doi:10.1007/s00442-004-1788-8

Chong, J., Yamamoto, M., Xia, J., 2019. MetaboAnalystR 2.0: From Raw Spectra to Biological Insights. Metabolites 9, 57. doi:10.3390/metabo9030057

Crawford, R.L., Paszczynski, A., Lang, Q., Erwin, D.P., Allenbach, L., Corti, G., Anderson, T.J., Cheng, F., Wai, C., Barnes, B., Wells, R., Assefi, T., Mojarradi, M., 2002. Measurement of microbial activity in soil by colorimetric observation of in situ dye reduction: an approach to detection of extraterrestrial life. BMC Microbiology 8.

Dixon, P., 2003. VEGAN, a package of R functions for community ecology. Journal of Vegetation Science 14, 927–930. doi:10.1111/j.1654-1103.2003.tb02228.x

Esclapez, M.D., García-Pérez, J.V., Mulet, A., Cárcel, J.A., 2011. Ultrasound-Assisted Extraction of Natural Products. Food Engineering Reviews 3, 108. doi:10.1007/s12393-011-9036-6

Farrés, M., Martrat, B., Mol, B. de, Grimalt, J.O., Tauler, R., 2015. Extraction of climatic signals from fossil organic compounds in marine sediments up to 11.7Ma old (IODP-U1318). Analytica Chimica Acta 879, 1–9. doi:10.1016/j.aca.2015.04.051

Firke, S., 2020. janitor: Simple Tools for Examining and Cleaning Dirty Data.

Fox, J., Weisberg, S., 2019. An R Companion to Applied Regression, Third. ed. Sage, Thousand Oaks CA.

Fox, P.M., Nico, P.S., Tfaily, M.M., Heckman, K., Davis, J.A., 2017. Characterization of natural organic matter in low-carbon sediments: Extraction and analytical approaches. Organic Geochemistry 114, 12–22. doi:10.1016/j.orggeochem.2017.08.009

Gong, Z.-G., Hu, J., Wu, X., Xu, Y.-J., 2017. The Recent Developments in Sample Preparation for Mass Spectrometry-Based Metabolomics. Critical Reviews in Analytical Chemistry 47, 325–331. doi:10.1080/10408347.2017.1289836

Guigue, J., Mathieu, O., Lévêque, J., Mounier, S., Laffont, R., Maron, P.A., Navarro, N., Chateau, C., Amiotte-Suchet, P., Lucas, Y., 2014. A comparison of extraction procedures for water-extractable organic matter in soils: Comparison of WEOM extraction procedures. European Journal of Soil Science 65, 520–530. doi:10.1111/ejss.12156

Hamkalo, Z., Bedernichek, T., 2014. Total, cold and hot water extractable organic carbon in soil profile: impact of land-use change. Zemdirbyste-Agriculture 101, 125–132. doi:10.13080/z-a.2014.101.016

Harrell, F.E.H., Dupont, C., 2020. Hmisc: Harrell Miscellaneous.

Hedges, J.I., Hare, P.E., 1987. Amino acid adsorption by clay minerals in distilled water. Geochimica et Cosmochimica Acta 51, 255–259. doi:10.1016/0016-7037(87)90237-7

Hendershot, W.H., Lalande, H., Duquette, M., 2008. Ion Exchange and Exchangeable Cations, in: Sampling and Methods of Analysis. CRC Press, Taylor & Francis Group, Oxford, UK, pp. 197–206.

Hopkins, D.W., 2007. Carbon mineralization, in: Soil Sampling and Methods of Analysis. Taylor & Francis Group, LLC, Boca Raton, FL, pp. 589–598.

Jackson, R.B., Lajtha, K., Crow, S.E., Hugelius, G., Kramer, M.G., Piñeiro, G., 2017. The Ecology of Soil Carbon: Pools, Vulnerabilities, and Biotic and Abiotic Controls. Annual Review of Ecology, Evolution, and Systematics 48, 419–445. doi:10.1146/annurev-ecolsys-112414-054234

Jansson, J.K., Hofmockel, K.S., 2018. The soil microbiome — from metagenomics to metaphenomics. Current Opinion in Microbiology 43, 162–168. doi:10.1016/j.mib.2018.01.013

Jian, S., Li, J., Chen, J., Wang, G., Mayes, M.A., Dzantor, K.E., Hui, D., Luo, Y., 2016. Soil extracellular enzyme activities, soil carbon and nitrogen storage under nitrogen fertilization: A meta-analysis. Soil Biology and Biochemistry 101, 32–43. doi:10.1016/j.soilbio.2016.07.003

Kallenbach, C.M., Frey, S.D., Grandy, A.S., 2016. Direct evidence for microbial-derived soil organic matter formation and its ecophysiological controls. Nature Communications 7. doi:10.1038/ncomms13630

Kassambara, A., 2020a. ggpubr: “ggplot2” Based Publication Ready Plots.

Kassambara, A., 2020b. rstatix: Pipe-Friendly Framework for Basic Statistical Tests.

Kiersch, K., Kruse, J., Regier, T.Z., Leinweber, P., 2012. Temperature resolved alteration of soil organic matter composition during laboratory heating as revealed by C and N XANES spectroscopy and Py-FIMS. Thermochimica Acta 537, 36–43. doi:10.1016/j.tca.2012.02.034

Kim, A.Y., Ismay, C., 2019. moderndive: Tidyverse-Friendly Introductory Linear Regression.

Kindler, R., Miltner, A., Richnow, H., Kastner, M., 2006. Fate of gram-negative bacterial biomass in soil—mineralization and contribution to SOM. Soil Biology and Biochemistry 38, 2860–2870. doi:10.1016/j.soilbio.2006.04.047

Klose, S., Tabatabai, M., 2002a. Response of glycosidases in soils to chloroform fumigation. Biology and Fertility of Soils 35, 262–269. doi:10.1007/s00374-002-0463-7

Klose, S., Tabatabai, M.A., 2002b. Response of amidohydrolases in soils to chloroform fumigation. Journal of Plant Nutrition and Soil Science 165, 125–132. doi:10.1002/1522-2624(200204)165:2<125::AID-JPLN125>3.0.CO;2-P
Klose, S., Tabatabai, M.A., 2002c. Response of phosphomonoesterases in soils to chloroform fumigation. Journal of Plant Nutrition and Soil Science 165, 429–434. doi:10.1002/1522-2624(200208)165:4<429::AID-JPLN429>3.0.CO;2-S
Kroetsch, D., Wang, C., 2008. Particle Size Distribution, in: Soil Sampling and Methods of Analysis. CRC Press, Taylor & Francis Group, Oxford, UK, pp. 713–725.
Landgraf, D., Leinweber, P., Makeschin, F., 2006. Cold and hot water–extractable organic matter as indicators of litter decomposition in forest soils. Journal of Plant Nutrition and Soil Science 169, 76–82. doi:10.1002/jpln.200521711
Larsson, J., 2020. eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses.
Lees, K., Fitzsimons, M., Snape, J., Tappin, A., Comber, S., 2018. Soil sterilisation methods for use in OECD 106: How effective are they? Chemosphere 209, 61–67. doi:10.1016/j.chemosphere.2018.06.073
Lehmann, J., Kleber, M., 2015. The contentious nature of soil organic matter. Nature. doi:10.1038/nature16069
Li, X.M., Sun, G.-X., Chen, S.-C., Fang, Z., Yuan, H.-Y., Shi, Q., Zhu, Y.-G., 2018. Molecular Chemodiversity of Dissolved Organic Matter in Paddy Soils. Environmental Science & Technology 52, 963–971. doi:10.1021/acs.est.7b00377
O'Brien, P.L., DeSutter, T.M., Casey, F.X.M., Khan, E., Wick, A.F., 2018. Thermal remediation alters soil properties – a review. Journal of Environmental Management 206, 826–835. doi:10.1016/j.jenvman.2017.11.052
Pluskal, T., Castillo, S., Villar-Briones, A., Orešič, M., 2010. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11, 395. doi:10.1186/1471-2105-11-395
Pokrovsky, O.S., Karlsson, J., Giesler, R., 2018. Freeze-thaw cycles of Arctic thaw ponds remove colloidal metals and generate low-molecular-weight organic matter. Biogeochemistry 137, 321–336. doi:10.1007/s10533-018-0421-6
Renella, G., Landi, L., Nannipieri, P., 2002. Hydrolase activities during and after the chloroform fumigation of soil as affected by protease activity. Soil Biology and Biochemistry 34, 51–60. doi:10.1016/S0038-0717(01)00152-3
Rousk, J., Bååth, E., 2011. Growth of saprotrophic fungi and bacteria in soil: Growth of saprotrophic fungi and bacteria in soil. FEMS Microbiology Ecology 78, 17–30. doi:10.1111/j.1574-6941.2011.01106.x
Saiya-Cork, K.R., Sinsabaugh, R.L., Zak, D.R., 2002. The effects of long term nitrogen deposition on extracellular enzyme activity in an Acer saccharum forest soil. Soil Biology and Biochemistry 34, 1309–1315. doi:10.1016/S0038-0717(02)00074-3
Saleem, A., Bell, M.A., Kimpe, L.E., Korosi, J.B., Arnason, J.T., Blais, J.M., 2019. Identifying novel treeline biomarkers in lake sediments using an untargeted screening approach. Science of The Total Environment 694, 1–11. doi:https://doi.org/10.1016/j.scitotenv.2019.133684
Schmidt, M.W.I., Torn, M.S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I.A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D.A.C., Nannipieri, P., Rasse, D.P., Weiner, S., Trumbore, S.E.,

2011. Persistence of soil organic matter as an ecosystem property. Nature 478, 49–56. doi:10.1038/nature10386
Schweigert, M., Herrmann, S., Miltner, A., Fester, T., Kästner, M., 2015. Fate of ectomycorrhizal fungal biomass in a soil bioreactor system and its contribution to soil organic matter formation. Soil Biology and Biochemistry 88, 120–127. doi:10.1016/j.soilbio.2015.05.012
Simpson, A.J., Simpson, M.J., Smith, E., Kelleher, B.P., 2007. Microbially Derived Inputs to Soil Organic Matter: Are Current Estimates Too Low? Environmental Science & Technology 41, 8070–8076. doi:10.1021/es071217x
Simpson, G.L., 2019. ggvegan: "ggplot2" Plots for the "vegan" Package.
Sinsabaugh, R.L., 2010. Phenol oxidase, peroxidase and organic matter dynamics of soil. Soil Biology and Biochemistry 42, 391–404. doi:10.1016/j.soilbio.2009.10.014
Sinsabaugh, R.L., Manzoni, S., Moorhead, D.L., Richter, A., 2013. Carbon use efficiency of microbial communities: stoichiometry, methodology and modelling. Ecology Letters 16, 930–939. doi:10.1111/ele.12113
Sinsabaugh, R.L., Saiya-Cork, K., Long, T., Osgood, M.P., Neher, D.A., Zak, D.R., Norby, R.J., 2003. Soil microbial activity in a Liquidambar plantation unresponsive to CO2-driven increases in primary production. Applied Soil Ecology 24, 263–271. doi:10.1016/S0929-1393(03)00002-7
Sitnikov, D.G., Monnin, C.S., Vuckovic, D., 2016. Systematic Assessment of Seven Solvent and Solid-Phase Extraction Methods for Metabolomics Analysis of Human Plasma by LC-MS. Scientific Reports 6. doi:10.1038/srep38885
Slowikowski, K., 2020. ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2."
Sorensen, P.O., Finzi, A.C., Giasson, M.-A., Reinmann, A.B., Sanders-DeMott, R., Templer, P.H., 2018. Winter soil freeze-thaw cycles lead to reductions in soil microbial biomass and activity not compensated for by soil warming. Soil Biology and Biochemistry 116, 39–47. doi:10.1016/j.soilbio.2017.09.026
Stone, M.M., DeForest, J.L., Plante, A.F., 2014. Changes in extracellular enzyme activity and microbial community structure with soil depth at the Luquillo Critical Zone Observatory. Soil Biology and Biochemistry 75, 237–247. doi:10.1016/j.soilbio.2014.04.017
Swenson, T.L., Jenkins, S., Bowen, B.P., Northen, T.R., 2015. Untargeted soil metabolomics methods for analysis of extractable organic matter. Soil Biology and Biochemistry 80, 189–198. doi:10.1016/j.soilbio.2014.10.007
Tang, H.T., Hajizadeh, K., Halsall, H.B., Heineman, W.R., 1991. Flow-injection analysis with electrochemical detection of reduced nicotinamide adenine dinucleotide using 2,6-dichloroindophenol as a redox coupling agent. Analytical Biochemistry 192, 243–250. doi:10.1016/0003-2697(91)90215-F
Tang, Y., Horikoshi, M., Li, W., 2016. ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages. The R Journal 8.
Tapia-Torres, Y., Elser, J.J., Souza, V., García-Oliva, F., 2015. Ecoenzymatic stoichiometry at the extremes: How microbes cope in an ultra-oligotrophic desert soil. Soil Biology and Biochemistry 87, 34–42. doi:10.1016/j.soilbio.2015.04.007
Tarnawski, V.-R., Leong, W.H., 2000. Thermal Conductivity of Soils at Very Low Moisture Content and Moderate Temperatures. Transport in Porous Media 41, 137–147. doi:10.1023/A:1006738727206

Tfaily, M.M., Chu, R.K., Tolić, N., Roscioli, K.M., Anderton, C.R., Paša-Tolić, L., Robinson, E.W., Hess, N.J., 2015. Advanced Solvent Based Methods for Molecular Characterization of Soil Organic Matter by High-Resolution Mass Spectrometry. Analytical Chemistry 87, 5206–5215. doi:10.1021/acs.analchem.5b00116

Tfaily, M.M., Chu, R.K., Toyoda, J., Tolić, N., Robinson, E.W., Paša-Tolić, L., Hess, N.J., 2017. Sequential extraction protocol for organic matter from soils and sediments using high resolution mass spectrometry. Analytica Chimica Acta 972, 54–61. doi:10.1016/j.aca.2017.03.031

von Lützow, M., Kögel-Knabner, I., Ekschmitt, K., Flessa, H., Guggenberger, G., Matzner, E., Marschner, B., 2007. SOM fractionation methods: Relevance to functional pools and to stabilization mechanisms. Soil Biology and Biochemistry 39, 2183–2207. doi:10.1016/j.soilbio.2007.03.007

Voroney, R.P., Brookes, P.C., Beyaert, R.P., 2007. Soil Microbial Biomass C, N, P, and S, in: Soil Sampling and Methods of Analysis. CRC Press, Boca Raton, FL, pp. 637–651. doi:10.1201/9781420005271-59

Wang, X., Helgason, B., Westbrook, C., Bedard-Haughn, A., 2016. Effect of mineral sediments on carbon mineralization, organic matter composition and microbial community dynamics in a mountain peatland. Soil Biology and Biochemistry 103, 16–27. doi:10.1016/j.soilbio.2016.07.025

Warren, C.R., 2015. Comparison of methods for extraction of organic N monomers from soil microbial biomass. Soil Biology and Biochemistry 81, 67–76. doi:10.1016/j.soilbio.2014.11.005

Wen, B., Zhang, J., Zhang, S., Shan, X., Khan, S.U., Xing, B., 2007. Phenanthrene Sorption to Soil Humic Acid and Different Humin Fractions. Environmental Science & Technology 41, 3165–3171. doi:10.1021/es062262s

Whitfield, C.J., Reid, C., 2013. Predicting surface area of coarse-textured soils: Implications for weathering rates. Canadian Journal of Soil Science. doi:10.1139/CJSS2012-009

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. Journal of Open Source Software 4, 1686. doi:10.21105/joss.01686

Wiedermann, M.M., Kane, E.S., Veverica, T.J., Lilleskov, E.A., 2017. Are colorimetric assays appropriate for measuring phenol oxidase activity in peat soils? Soil Biology and Biochemistry 105, 108–110. doi:10.1016/j.soilbio.2016.11.019

Wolf, D.C., Skipper, H.D., 2018. Soil Sterilization, in: Methods of Soil Analysis. John Wiley & Sons, Ltd, pp. 41–51. doi:10.2136/sssabookser5.2.c3

Wu, X., Wu, L., Liu, Y., Zhang, P., Li, Q., Zhou, J., Hess, N.J., Hazen, T.C., Yang, W., Chakraborty, R., 2018. Microbial Interactions With Dissolved Organic Matter Drive Carbon Dynamics and Community Succession. Frontiers in Microbiology 9, 1234. doi:10.3389/fmicb.2018.01234

Yanni, S.F., Diochon, A., Helgason, B.L., Ellert, B.H., Gregorich, E.G., 2018. Temperature response of plant residue and soil organic matter decomposition in soil from different depths. European Journal of Soil Science 69, 325–335. doi:10.1111/ejss.12508

Yao, S.-H., Zhang, Y.-L., Han, Y., Han, X.-Z., Mao, J.-D., Zhang, B., 2019. Labile and recalcitrant components of organic matter of a Mollisol changed with land use and plant litter management: An advanced 13C NMR study. Science of The Total Environment 660, 1–10. doi:10.1016/j.scitotenv.2018.12.403

Zhang, H., Goll, D.S., Wang, Y.-P., Ciais, P., Wieder, W.R., Abramoff, R., Huang, Y., Guenet, B., Prescher, A.-K., Rossel, R.A.V., Barré, P., Chenu, C., Zhou, G., Tang, X., 2020. Microbial dynamics and soil physicochemical properties explain large-scale variations in soil organic carbon. Global Change Biology 26, 2668–2685. doi:10.1111/gcb.14994

Zhang, L., Xiang, P., Bao, X., Xiong, M., Liu, F., 2017. The Influence of Humic Substances on the Sorption of Three Organic Contaminants with Different Structure and Polarity to Clay Minerals. Water, Air, & Soil Pollution 228, 1–7. doi:10.1007/s11270-017-3380-y

Zhang, R., Yan, W., Jing, C., 2014. Mechanistic study of PFOS adsorption on kaolinite and montmorillonite. Colloids and Surfaces A: Physicochemical and Engineering Aspects 462, 252–258. doi:10.1016/j.colsurfa.2014.09.019

Zhao, A., Zhang, M., He, Z., 2013. Spectroscopic Characteristics and Biodegradability of Cold and Hot Water–Extractable Soil Organic Matter under Different Land Uses in Subarctic Alaska. Communications in Soil Science and Plant Analysis 44, 3030–3048. doi:10.1080/00103624.2013.829086

Zhou, X., Chen, C., Wang, Y., Xu, Z., Han, H., Li, L., Wan, S., 2013. Warming and increased precipitation have differential effects on soil extracellular enzyme activities in a temperate grassland. Science of The Total Environment 444, 552–558. doi:10.1016/j.scitotenv.2012.12.023

CHAPTER 6: General conclusions and implications

STUDY OBJECTIVES AND IMPLICATIONS

Chapter 2

The objective of CHAPTER TWO was to present a guide for soil metabolomics and sedimentomics methods. The implications of this literature review are that there is a pre-existing body of research in metabolomics that can be used to bolster and aid development of sedimentomics and soil metabolomics. While not all of the metabolomics methods will be transferable to sedimentomics and soil metabolomics, the body of metabolomics research provides: 1) information on methods/concepts that did and did not work, 2) software infrastructure that can be used as models for sedimentomics and soil metabolomics , and 3) a model of a unified community working toward developing standard operating procedures, quality controls, software, and quality assurance protocols. Sedimentomics and soil metabolomics would benefit from the lessons learned in metabolomics.

Chapter 3

The objective of this research was to determine if there are discriminant biomarkers between boreal and tundra regions across the Canadian Arctic treeline using sedimentomics. To our knowledge, this work is the first application of untargeted “Omics” analysis in freshwater lake sediments to discover paleolimnological biomarkers, and this CHAPTER validates the application of sedimentomics as a systematic paleolimnological biomarker discovery method. The implications of CHAPTER THREE are that paleolimnology, and other paleogeology fields, have a system to discovery new biomarkers, and that these fields are uniquely situated to apply an “Omics” style analysis. Currently, biomarker discovery is difficult, due to the complex nature of sedOM and that there is no current method to deliberately

discover biomarkers for environmental and climate reconstructions from sediment. Most biomarkers are discovered via a targeted approach where scientists theorize a molecule may be a potential biomarker, and then they focus on that specific molecule and test if it is present in the sediment core. The alternate biomarker discovery method is accidental discovery via the recovery of coextracts. Thus, having a rigorous and controlled systematic biomarker discovery method will facilitate the discovery of more biomarkers. Ultimately having more biomarkers in paleolimnology and paleogeology will lead to more multi-biomarker (also known as multi-proxy) reconstructions, which leads to more robust environmental reconstructions, and more knowledge about historical climates/environments and improved climate models. Lastly, paleolimnology and paleogeology have a strong body of literature in existing biomarkers (and proxies) that are uniquely situated to validate potential biomarkers discovered via sedimentomics, this a massive advantage compared to other metabolomics fields that do not have pre-existing biomarkers to compare against.

Chapter 4

The objective of CHAPTER FOUR was to determine if sediment organic matter composition is spatially dependent and if this concept can be used to draw ecologically relevant boundaries. We demonstrated that sedOM changes across predicted ecological boundaries and, potentially, with climate changes. We also demonstrated that determining the sedOM composition is more informative than traditional TOC analyses. Thus, this study acts as a proof-of-concept for using sedOM composition to define unique lake districts, and could have implications for ecological, contaminant, and climate monitoring. For instance, if the sedOM composition is assessed from the same lake repeatedly over years and/or before onset of industrial activities, changes in sedOM composition could function as an

early warning system of ecological changes. FT-ICR HRMS would be uniquely suited to this task because of its ultra-high resolution capabilities, which allows for improved elemental composition determination and better separation between co-eluting compounds. Of course, sedimentomics would need to be rigorously tested before this idea could become a valid monitoring technique, but this area of research is far from exhausted.

Chapter 5

The objective of CHAPTER FIVE was to explore the factors influencing soil metabolomics extractions and establish quality control procedures to control extraction variability and improve data quality. This CHAPTER has the most important implications for the future of soil metabolomics and sedimentomics. In this CHAPTER, we determined that microbial activity mineralizes ~1%/hr of water extractable soil organic molecules in the first 24 hours, physical characteristics of soil (i.e. cation exchange capacity, surface area, etc.) impact the quantity of SOM extracted, and different extraction methods affect the quality and quantity of SOM extracted even within the same soil.

The first conclusion that microbial activity removes ~1%/hr of water extractable soil organic molecules in the first 24 hours has implications for the quality of the SOM extract. Longer extraction methods will result in more of the soil molecules being modified so the extract becomes less representative of the native SOM extract, which in turn will affect the quality of downstream results. Thus, traditional soil (and sediment) OM extraction procedures that require 24 hours of shaking would not be appropriate for use in soil metabolomics or sedimentomics because of the extensive microbial modifications to ~ 24% of the extractable organic molecules.

The second conclusion, that the physical characteristics of soil (i.e. cation exchange capacity, surface area, etc.) impact the quantity of SOM extracted, is critical for studies that want to compare soil or sediment across space (and even downcore in some cases). Spatial changes in soil/sediment geophysical characteristics will affect which molecules are protected by adsorption or intercalation with soil or sediment mineral structures. This implies that studies having many different soil and sediment types may experience different SOM, or sedOM extraction compositions even if the initial compositions were the same. It also implies that comparing SOM or sedOM compositions across studies may prove to be very difficult and lead to inaccurate conclusions. While this may seem like a major drawback, there is the potential for adaption. For instance, the recovery standards introduced in CHAPTER FIVE, could be used to project or adjust the losses that may have occurred so that the resulting data matrix could be mathematically adjusted. This would be difficult and require the development of new techniques that would have to be rigorously tested, but it could be one of the answers to this difficult challenge. Also, this conclusion has stronger implications for some types of research questions, but not all. For instance, this would have strong implications for carbon flux analyses where the extraction of the biotransformed products would not be consistent between soil types, but for biomarker research, where you are looking for the molecule that is selective for your group regardless of soil characteristics, this conclusion may have a lesser impact and may advantageously improve the robustness of the biomarker selected.

Lastly, the final conclusion of this work, that different extraction methods affect the quality and quantity of SOM extracted even within the same soil, has implications for the current state of the fields of sedimentomics and soil metabolomics. Currently, there are many different methods being applied, and as shown in CHAPTER FIVE, this results in different OM extraction compositions even within the same soil. The implication is that, until the research community has established standard operating

procedures and settled on specific extraction methods, studies using different extraction methods should not be directly compared because of the potentially vastly different extraction yields. Some methods, like chloroform fumigation, also change the quality of the OM being extracted because it lyses cells. While this could be an advantage in some cases, it should still be remembered that the extractability of some OM will change because of the influx of new molecules, and that the influx of more material may also lead to more ion suppression and matrix effects. Thus, the community must start establishing quality control procedures and standard operating procedures so that interoperability and study-to-study comparisons are improved.

OVERALL CONCLUSIONS AND IMPLICATIONS

Quality controls are necessary for sedimentomics and soil metabolomics studies

To date, few common quality controls are employed by sedimentomics and soil metabolomics studies; however, this is not typical for other "Omics" studies. For instance, metabolomic studies in biological organisms have many quality control procedures and international working groups focus on developing reporting standards and quality control procedures (Dudzik et al., 2018; Fiehn et al., 2007). The reason quality controls procedures are fewer in sedimentomics and soil metabolomics is likely because these methods are new, with small, isolated research communities, and some quality control procedures from biological metabolomics do not translate into sedimentomics or soil metabolomics studies. Regardless, conclusions from CHAPTERS THREE, FOUR, and FIVE of this research suggest soil metabolomics and sedimentomics specific-quality controls are critical for the accurate determination of SOM and sedOM compositions. These quality controls must be included in the experimental design from the beginning and used at every stage of the process including sample collection, sample storage,

extraction, instrumental analysis, data processing, and statistical analysis. Many of these quality controls are listed in CHAPTER TWO such as: standard operating procedures, method blanks, sampling and environmental replication, recovery standards, analytical blanks, analysis randomization, cleaning blanks, internal standards, technical/machine replicates, and pooled quality controls. CHAPTER FIVE also applied two new quality controls to soil metabolomics: 1) microbially-sensitive recovery standards, and 2) pure clay as control matrices. Thus, future research should focus on building on the work presented here so that, eventually, more quality controls can be developed to improve the current state of sedimentomics and soil metabolomics.

Sedimentomics is a valid technique to highlight changes in sedOM

SedOM is well preserved in freshwater lake sediments and can act as an archive for environmental OM over time. Archived sedOM can be used to reconstruct the past environments that surrounded the freshwater lakes. This thesis demonstrated the first application of sedimentomics to freshwater lake sediments in CHAPTER THREE and FOUR, these proof-of-concept methods require further development but have implications for sedOM research moving forward. The advancement of technology has made analyzing sedOM, a critically understudied component of the C cycle, easier and more accurate. Paleolimnologists, geologists, soil scientists, etc. should embrace these technological advances to improve on our current knowledge of sedOM cycling and how it impacts the broader global C cycle.

Soil metabolomics and sedimentomics yield more information about carbon cycling than traditional measurements

Traditional measurements (i.e. bulk measurements like TS, TOC, TN) were fundamental to building our current knowledge, however, the advent of newer technologies, like HRMS and machine learning, allow new techniques like soil metabolomics and sedimentomics to provide similar information but in much greater detail. CHAPTER FOUR demonstrated that traditional bulk measurements were not as significant to models as sedOM compositional information in explaining the clustering of lakes. This presents an opportunity for research scientists to expand on this research and add sedimentomics and soil metabolomics into the formidable arsenal of tools used to investigate soil and sediment C cycling. The additional information provided by sedimentomics and metabolomics could be used to support C cycling models, to investigate the here-to-for elusive mechanisms and biotic-abiotic interactions involved in perturbation response, and ultimately how these mechanisms could scale in a global context.

Soil metabolomics SOM extractions are more variable and require more quality controls

Soil metabolomics is more established than sedimentomics with more literature; however, to improve the transparency, interoperability, and integrative ability of soil metabolomics studies more work is needed to be done to compare different SOM extraction methods and parameters. This work, specifically CHAPTER FIVE, determined that SOM extracted with water is especially susceptible to changes in extraction methods and is inherently more variable than methanol SOM extracts and even the sedOM extracts. This difference is likely rooted in the underlying mechanisms driving pore water transportation and adsorption dynamics where smaller water-soluble molecules are present in the pore water and are less likely to be adsorbed onto a surface. The water-soluble fraction of OM is also more

dynamic since it is the fraction most easily used by soil microbiota for nutrients. Lastly, the residence times are significantly shorter in soil than in sediment so more variation is expected because SOM is more dynamic than sedOM. The implication of this work is that quality controls need to be adjusted based on the environmental matrix so that the results are reliable and robust despite additional variation.

FUTURE DIRECTIONS

This thesis has demonstrated the validity of using "Omics" untargeted analysis of small molecules in soil and sediment matrices. I used untargeted analysis on a variety of different soils and sediments and still produced coherent results, I also proposed workflows and quality controls for future studies. However, there is still much to do in this field, and there are methodological improvements to be made at all stages.

To date there are no standardized sample collection, storage/processing, and/or extraction protocols for sedimentomics or soil metabolomics. The wrong sample collection strategy (i.e. too little soil/sediment, no environmental replicates, the overuse of plastic during collection, etc.) and/or improper storage or sample processing (i.e. moist at room temperature, autoclaving, baking, etc.) can introduce experimental artifacts that can alter the extracted composition to the detriment of the entire study. Artifacts can also be introduced in the extraction method if careful consideration and quality controls are not applied. The future of this field depends on further research into standardizing procedures so that studies have increased transparency, interoperability, integration with other studies in meta-analyses, and to scale the research beyond local/regional contexts to broader global applications.

Data processing in any “Omics” field is complex and often ill-reported and/or subjective. For instance, the processing software used, and the operator, strongly influence the results of data processing. In my experience, current open-source and/or proprietary software struggle to adapt to the much greater complexity in sedOM and SOM versus the biological tissues they were designed for. Paleolimnologists, geologists, and soil scientists should begin to foster collaborations with computer scientists and developers to improve the software used to process “Omics” data and increase the software’s utility to sedimentomics and soil metabolomics. This would require studies comparing how the results change after different processing strategies are applied on the same SOM and sedOM datasets for HRMS and NMR. Also, the design and sharing of new open-source software specifically for soil metabolomics and sedimentomics has not be undertaken and may represent a productive low-hanging fruit for a collaboration of interested scientists.

The most significant bottleneck in furthering soil metabolomics and sedimentomics is annotation. Annotation of molecules from complex mixtures is difficult, and not only because of the number of compounds present. Instrumental capabilities and settings influence the quality of the data produced, and very specific parameters need to be met (i.e. acquisition in profile mode, mass accuracy < 5 ppm, < 2% error for isotopic abundance patterns, MS/MS information, standard collision energy) before elemental compositions can be determined accurately. In addition, to date, there are no SOM or sedOM HRMS structural databases. These databases are critical for annotation because they contain high quality MS spectra, and can also contain adduct information, retention times, collisional cross-sections, fragmentation patterns, etc. Molecules of interest can be matched against the compounds present in a database for identification, and this is especially important when: 1) there is not enough soil/sediment for orthogonal analyses with NMR or IR, and/or 2) there are no standards available for

structural confirmations (as is often the case). Future work in this field needs the development of these types of databases in order to progress. This will require massive collaborative efforts to create the database and/or build-off current metabolomics databases. Interested researchers will need to set standards for these databases as well (i.e. common analysis parameters, common instrumentation, etc.) and a network of collaborating laboratories should be able to send out isolated samples of confirmed structures for analysis to laboratories with the capabilities to meet the analytical standards for submission into the database. There is an immense body of work here in developing new databases, but such a database would be critical infrastructure for sedimentomics and soil metabolomics.

Lastly, metaphenomics (also known as multi-omics) is the ultimate goal of sedimentomics and soil metabolomics. Metaphenomics is the integration of small molecule compositions (as determined via soil metabolomics and sedimentomics), with metagenomics, metatransciptomics, metaproteomics, and environmental information (i.e. bathymetry, geochemical parameters, weather conditions, latitude, etc.). The integration of all of the "Omics" techniques is the future and ultimate goal for understanding C cycling in the environment because that data from all of these "Omics" experiments provide highly correlated and inter-connected information (i.e. Metagenomics = community presence; metaproteomics/metatranscriptomics = functional information; sedimentomics/soil metabolomics = feedback mechanisms and cycling) that will shed light on the highly elusive mechanisms driving C cycling. However, this work is not easy. There is much research to be done in first optimizing all "Omics" methods in soil and sediment, but also in integrating the data. To date, most metaphenomics research is done by doing separate processing and statistical analysis for each "Omics" experiment and then making manual observations. There are very few truly integrative analyses, and all "Omics" field could benefit from new methods to merge all of this complex data.

CLOSING STATEMENT

This thesis aimed to (1) Develop new extraction methods for untargeted analyses of soil and sediment matrices that include appropriate data handling and quality control measures; (2) Investigate novel applications of untargeted analyses for environmental classification and monitoring; and, (3) Investigate the experimental parameters that influence soil metabolomics extractions. I demonstrated that there are successful extraction methods for both sedimentomics and soil metabolomics studies in CHAPTER THREE and CHAPTER FIVE, respectively. I also determined that sedimentomics can be used for the classification of lake sediments based on sedOM compositions in CHAPTER FOUR. Lastly, I determined that there are many factors in soil metabolomics experiments that influence the SOM composition in the extracts, but proper quality controls can help control these influences in CHAPTER FIVE.

REFERENCES

Chong, J., Yamamoto, M., Xia, J., 2019. MetaboAnalystR 2.0: From Raw Spectra to Biological Insights. Metabolites 9, 57. https://doi.org/10.3390/metabo9030057

Dudzik, D., Barbas-Bernardos, C., García, A., Barbas, C., 2018. Quality assurance procedures for mass spectrometry untargeted metabolomics. a review. J. Pharm. Biomed. Anal. 147, 149–173. https://doi.org/10.1016/j.jpba.2017.07.044

Fiehn, O., Robertson, D., Griffin, J., Werf, M. van der, Nikolau, B., Morrison, N., Sumner, L.W., Goodacre, R., Hardy, N.W., Taylor, C., Fostel, J., Kristal, B., Kaddurah-Daouk, R., Mendes, P., Ommen, B. van, Lindon, J.C., Sansone, S.-A., 2007. The metabolomics standards initiative (MSI). Metabolomics 3, 175–178. https://doi.org/10.1007/s11306-007-0070-6

APPENDIX 1: Paleolimnology in support of archeology: a review of past investigations and a proposed framework for future study design

Madison A. Bell[1*], Jules M. Blais[1]

[1]Laboratory for the Analysis of Natural and Synthetic Environmental Toxicants, Department of Biology, University of Ottawa, Ottawa ON, K1N 6N5, Canada

[*]Corresponding Author

Madison Bell

STATEMENT OF CONTRIBUTIONS

Madison A. Bell: conceptualization, writing – original draft, visualizations, formal analysis, literature search, and corresponding author. Jules M. Blais: supervision, writing – review & editing, funding acquisition. *Journal of Paleolimnology* published this chapter:

Bell, M.A., and Blais, J.M. (2020) Paleolimnology in support of archeology: a review of past investigations and a proposed framework for future study design. Journal of Paleolimnology. https://doi.org/10.1007/s10933-020-00156-8

ABSTRACT

We conducted a systematic review of 89 paleolimnological studies applied to archeological questions. Where we discuss the physical, chemical and biological sediment variables used in these studies in terms of their advantages and disadvantages as paleolimnological proxies for archeological studies. We make four key observations: 1) This field is rapidly growing, 2) More research is needed, 3) More standardization is required for future integrative analyses, and 4) More robust studies with multiple proxies are needed as the field grows. To address these challenges, we developed a framework to help researchers design paleolimnological studies in support of archeology. The framework includes standardized terminology of proxy characteristics and definition of a new term: orthogonality. This framework was then integrated with decision matrix analysis to build study scores that can be used to help researchers optimize their study design. This approach will help future researchers build more robust paleolimnology studies to more effectively complement independent archeological work. We also summarized new areas of archeology and chemistry that could be integrated with paleolimnology in the future.

INTRODUCTION

Natural archives of paleoenvironmental information (e.g. tree rings, ice, peat, and marine and lake sediments) can reveal how past humans altered their environment (Dubois and Jacob 2016; Mills et al. 2017). The environmental information preserved is interpreted using chemical, physical, and biological proxies. These chemical and biological remains are referred to as "proxies" because they can be used to indirectly infer changes in past environmental and ecological conditions (i.e. temperature, pH, past species diversity, etc.), and these conditions can then be related to actual historical events or trends (e.g. agriculture, soil erosion, etc.) that cannot be measured directly. Many types of proxies are used in natural archives, and they can be classified into three main groups: 1) physical, 2) biological, and 3) chemical (Figure A1.1).

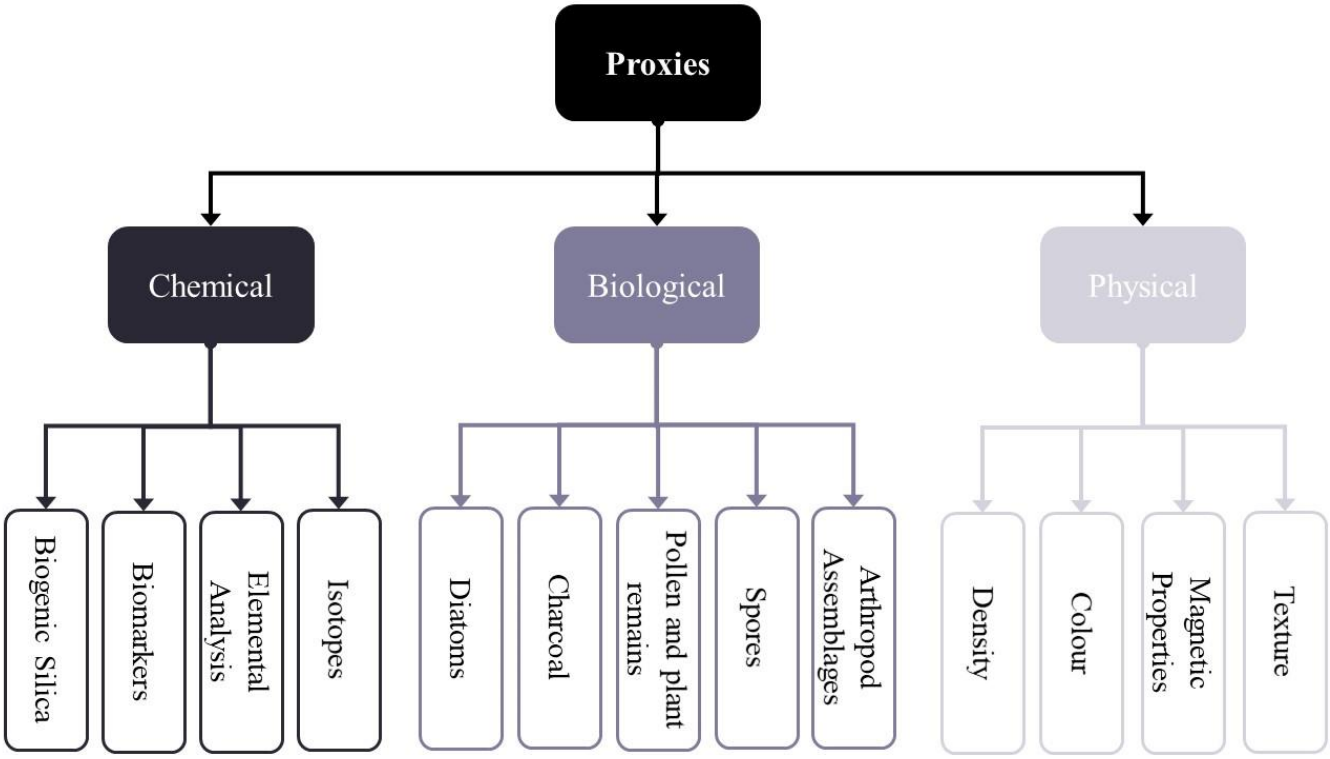


Figure A1.1: Flow chart of proxy categories proposed by U.S. Geological Survey (USGS 2016).

Physical proxies in lake sediments include descriptive characteristics of the natural archive like color, texture, density, magnetic properties, and particle size (Last and Smol 2001a; USGS 2016). They can provide information about past changes in hydrology, soil erosion, and cultural eutrophication (Mills et al. 2017). Physical proxies are rarely used alone, and are most useful in a multi-proxy approach (Mills et al. 2017).

Biological proxies are the remains of living organisms. Biological community assemblages enable inferences about past changes in species diversity, temperature, moisture, fire frequency/intensity, and water column variables like pH, dissolved oxygen, ion concentrations, and nutrient status (USGS 2016; Mills et al. 2017). Examples of biological proxies include pollen, fungal spores, plant and animal remains, and charcoal. Such organisms, whose biological remains are used as proxies, have ranges of tolerance for conditions related to population maintenance and growth. The relative abundance of these biological remains can serve as substitute measures for past changes in abiotic or biotic conditions that influenced the organism's population (Korosi et al. 2017b). For instance, diatoms are used widely to document the effects of acid rain on lakewater pH because: 1) diatoms are numerous, 2) the optimal pH ranges are well established for many species, and 3) diatom communities respond quickly to shifts in pH (Korosi et al. 2017b; Mills et al. 2017).

The last group, chemical proxies, is a large group that includes metals (e.g. aluminum, iron, titanium, manganese, etc.), non-metals (e.g. carbon, nitrogen, sulfur, phosphorus, etc.), stable isotopes (e.g. $\delta^{13}C$, $\delta^{18}O$, $\delta^{15}N$ and multiple Pb isotopes [^{204}Pb, ^{206}Pb, ^{207}Pb, ^{208}Pb]), and molecular compounds such as minerals (e.g. carbonates, gypsum, and halite), and biomarkers (Meyers and Teranes 2001; USGS 2016). Chemical proxies are used to infer many different environmental changes. For instance, biogenic metals (e.g. calcium) can be linked to biological activity, whereas other metals (e.g. lead) can indicate

human activities like mining and smelting. Macro-nutrient elements, such as carbon, phosphorus, and nitrogen, can be linked with bulk biological contributions from plant matter, humans, and animals (Last and Smol 2001b; Mills et al. 2017). Biomarkers are organic compounds from specific biological sources (e.g. microbes, plants, animals), environmental sources (e.g. fire, abiotic oxidation or reduction), or anthropogenic sources (e.g. agricultural and animal husbandry products, pesticides, plastics, sewage). Chemical proxies from multiple sources can be studied in combination with other proxies to infer past climate, pre-human-disturbance conditions, and anthropogenic influences on the environment.

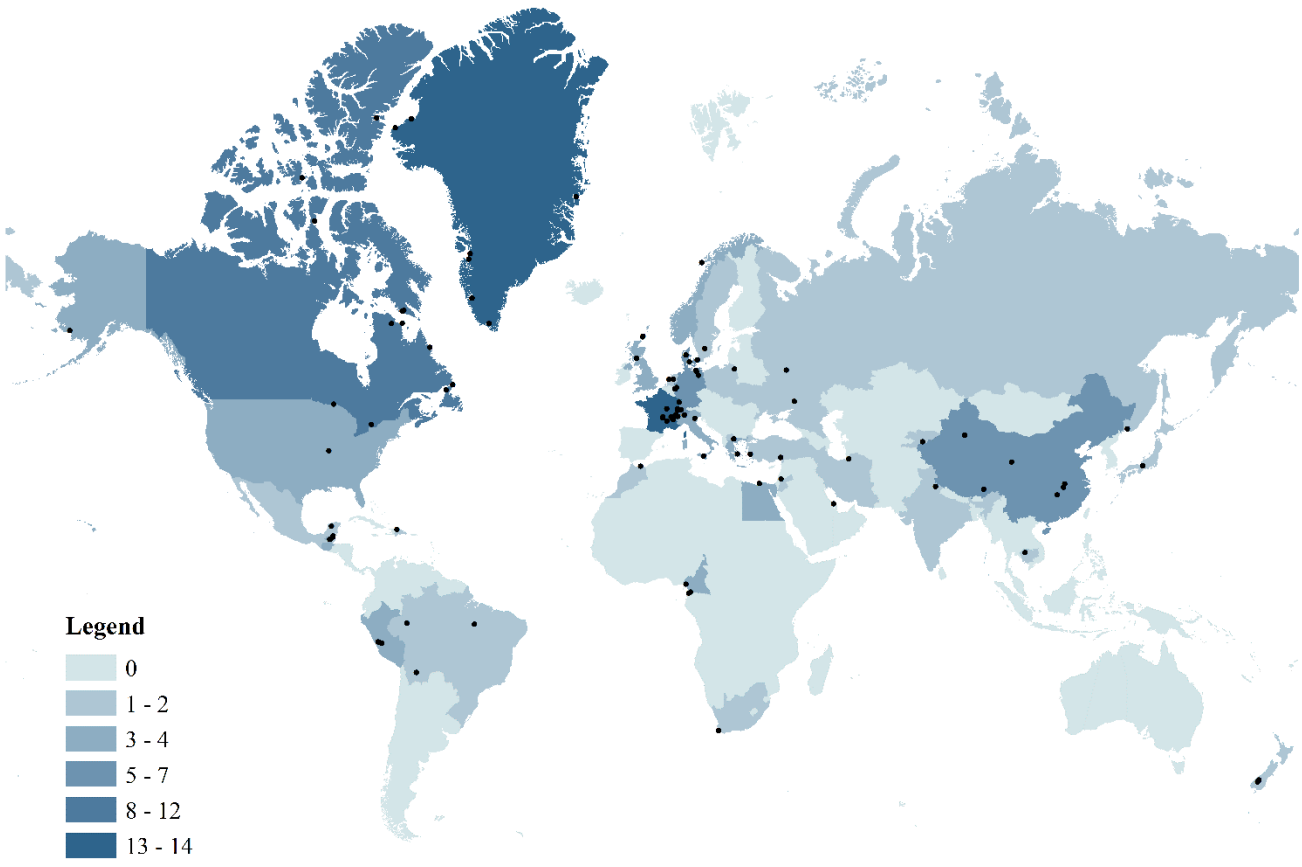


Figure A1.2: Distribution of the archeological study sites included in this review. The coloring indicates the activity of paleoenvironmetal research in archeology, by country.

Proxies in lake sediments, whether physical, chemical, or biological, can provide unique insights for archeological studies. In this review we provide a framework to help researchers develop more robust study designs. In the first section, we review proxies that are useful to infer human influence at pre-industrial (pre-1760s) archeological sites. We chose proxies to review from 89 archeological studies of chemical and biological proxies in lake sediment cores and a few paleosols (Table A1.1; Figure A1.2). These proxies guided our generation of a framework to help researchers decide which proxies to select and when to use them, using key proxy characteristics defined in this review and the concept of orthogonality. This framework for designing multi-proxy paleolimnological studies to address archeological questions is novel because we integrated World Health Organization (WHO) definitions for biomarker characteristics with the United States Geological Survey (USGS) definitions for proxies. In addition, we introduce the concept of orthogonality to improve the inferential strength in these studies, where possible. Following our proposed framework, we include a section on study design that integrates the proposed framework with multiple proxies to illustrate a method for choosing which proxies to use.

Table A1.1: Paleolimnological studies at archeological sites referenced in this review. The terms multi-proxy, orthogonality (V=vegetative, F=Fire, and A=Animal/Human), and selectivity indicate whether the chosen proxies were specific for human occupation (e.g. fecal sterols, bile acids, agriculturally related crop pollens, biomarkers (cannabinol or miliacin) or eDNA, mining-related metals, or corprophilous fungi)

Country	Culture	Oldest occupied period detected (years ago)	Matrix	# of proxies used	Proxies Used	Multi-proxy (3+)	Orthogonality (V,F,A)[#]	Specificity with to human occupation	Reference(s)
United Kingdom	Roman	800	paleosol	3	sterols; bile acids; pollen	Yes	V,A	Yes	Knights et al. 1983
Switzerland	Roman	10,590	peat	1	metals	No	A	Yes	Shotyk et al. 1998
United Kingdom	Orcadian Bronze age	3400	paleosol	3	fatty acids; sterols; bile acids	Yes	V,A	Yes	Bull et al. 1999
United Kingdom	Norse	800	paleosol	3	n-alkyls; sterols; bile acids	Yes	V,A	Yes	Simpson et al. 1999
Brazil	Carajás Indians	700	sediment	2	MA; charcoal	No	F	No	Elias et al. 2001
Bolivia	Tiwanaku Empire and Incan	1000	sediment	1	metal	No	A	Yes	Abbott and Wolfe 2003
Greece	Roman	1400	sediment	2	sterols; bile acids	No	V,A	Yes	Bull et al. 2003
Netherlands	Roman period	unreported	paleosol	3	fungal spores; pollen; arthropod assemblages	Yes	V,A	Yes	van Geel et al. 2003
Switzerland	Bell Beaker	3800	sediment	2	Pollen; stable isotopes δ^{18}O)	No	V	Yes	Tinner et al. 2003
Canada	Thule	800	sediment	2	stable isotopes δ^{15}N); diatom	No	V	No	Douglas et al. 2004
Turkey	Çatalhöyük	9000	paleosol	2	sterols; bile acids	No	V,A	Yes	Bull et al. 2005
Germany	Neolithic settlers	5800	sediment	2	metals; pollen	No	V	No	Selig et al. 2007
Peru	Wari Empire	1000	sediment	2	metal; stable isotopes ^{206}Pb, ^{207}Pb)	No	A	Yes	Cooke et al. 2007
Italy	Bronze and Iron Age farming	3000	sediment	2	pollen; charcoal	No	V,F	No	Sadori and Giardini 2007
Egypt	Rhakotis Egyptian	2300	sediment	2	metal; stable isotopes ^{206}Pb, ^{208}Pb)	No	A	Yes	Stanley et al. 2007
Greenland	Norse	600	paleosol	1	arthropod assemblages	No	V,A	No	Panagiotakopulu et al. 2007

Peru, Bolivia, Brazil	pre-Incan Tiwanaku Empire and Wari Empire	1000	sediment	1	metal	No	A	Yes	Cooke et al. 2008
France	Neolithic farmers	3500	sediment	1	terpenoids PTMEs)	No	A	Yes	Jacob et al. 2008a, 2008b
China	Han Dynasty	5000	sediment	2	metal; stable isotopes ^{206}Pb, ^{207}Pb, ^{208}Pb)	No	A	Yes	Lee et al. 2008
Greenland	Norse	1000	paleosol	3	arthropod assemblages; pollen; fungal spores	Yes	V,A	Yes	Buckland et al. 2009
France	Bronze age farming	3700	sediment	1	pollen	No	V	Yes	Gauthier and Richard 2009
France	Neolithic farming	3500	sediment	1	terpenoids PTMEs)	No	A	Yes	Jacob et al. 2009
Canada	Thule	800	sediment	3	stable isotopes δ^{15}N); diatoms; pigments	No	V	No	Hadley et al. 2010a, 2010b
Greenland	Norse	1100	sediment	2	fungal spores; pollen	No	V,A	Yes	Gauthier et al. 2010
China	Ming-Qing Dynasty and late Xizhou to early Dongzhou Dynasty	400	paleosol	2	PAHs, charcoal	No	F	No	Zou et al. 2010
Egypt	Egyptian New Kingdom	3000	sediment	2	pollen; charcoal	No	V,F	No	Stanley and Bernhardt 2010
Peru	Inca	3500	sediment	2	arthropod assemblages; pollen	No	V,A	No	Chepstow-Lusty 2011
Norway	Stone age farming	2500	sediment	3	sterols; n-alkanes; PAHs	Yes	V,F,A	Yes	D'Anjou et al. 2012
Greenland	Norse	2500	sediment	2	diatoms; pollen; metals	Yes	V	No	Massa et al. 2012
France	Neolithic farrming	7000	sediment	3	pollen; charcoal; metals	No	V,F	Yes	Doyen et al. 2013
France	Neolithic farrming	7000	peat	3	pollen; charcoal; fungal spores	No	V,F	Yes	Jouffroy-Bapicot et al. 2013
France	Medieval farming	600	sediment	2	cannabinols; pollen	No	V	Yes	Lavrieux et al. 2013

Canada	Dorset	1600	sediment	2	stable isotopes δ^{15}N); diatom	No	V	No	Michelutti et al. 2013
France	Neolithic farming	3500	sediment	2	terpenoids PTMEs); plant fossils	No	V	Yes	Simonneau et al. 2013
Canada	Paleoindian	9500	sediment	2	testate amoeba; pollen	No	V	No	Sonnenburg et al. 2013
United States	Old Copper culture	8000	sediment	1	metals	No	A	Yes	Pompeani et al. 2013
Egypt	Rakhotis Egyptian	5800	sediment	1	metals	No	A	Yes	Veron et al. 2013
France	Iron age farming	5000	sediment	1	eDNA	No	V,A	Yes	Giguet-Covex et al. 2014
Greenland	Thule	800	paleosol	1	arthropod assemblages	No	V	No	Dussault et al. 2014
Greenland	Norse	1000	sediment	2	fungal spores; bile acids	No	A	Yes	Guillemot et al. 2015
Morocco	Neolithic farming	7600	paleosol	1	BPCA	No	F	No	Lehndorff et al. 2015
Canada	Thule	600	sediment	2	pollen; charcoal	No	V,F	No	Roy et al. 2015
Guatemala	Maya	3700	sediment	2	MA; charcoal; pollen	Yes	V,F	Yes	Schüpbach et al. 2015
Norway	Viking settlement	1000	paleosol	1	parasite eDNA	No	A	No	Søe et al. 2015
Canada	Dorset, Thule	800	sediment	2	pollen; plant macrofossils; charcoal	No	V,F	No	Bhiry et al. 2016
India	Prehistorical farming	4600	sediment	1	pollen	No	V	Yes	Demske et al. 2016
Cameroon	Hunter gathers and farmers	2300	paleosol	2	charcoal; plant macrofossils	No	V,F	No	Morin-Rivat et al. 2016
Ukraine	Iron age farming	unreported	paleosol	1	terpenoids PTME)	No	V	Yes	Motuzaite-Matuzeviciute et al. 2016
Greenland	Thule, Saqqaq, Dorset, Norse	4000	sediment	1	eDNA	No	V,A	Yes	Seersholm et al. 2016
Iran	Prehistorical farming	6000	sediment	5	pollen; fungal spores; arthropod assemblages; charcoal; n-alkanes	Yes	V,F,A	Yes	Shumilovskikh et al. 2016

France	Prehistorical farming	6000	paleosol	1	Terpenoids PTMEs)	No	V	Yes	Courel et al. 2017
Canada	Dorset	1990	paleosol	1	arthropod assemblages	No	V	No	Dussault et al. 2017
Greenland	Norse	1000	sediment	7	bile acids; fungal spores; n-alkanes; Terpenoids PTMEs); pollen; diatoms; metals	Yes	V,A	Yes	Guillemot et al. 2017
Italy	Prehistorical farming	6000	sediment	3	pollen, fungal spores, charcoal	Yes	V,F,A	Yes	Pini et al. 2017
Germany	Linearbandkeramik, Urnfield period / Bronze age, Iron age, and Roman age periods	7300	paleosol	2	sterols; bile acids	No	V,A	Yes	Prost et al. 2017
Turkey	Roman age harbor	2100	sediment	4	n-alkyls; n-alkanes; PAHs; tepenoids TDs & Caro)	Yes	V,F,A	Yes	Schwarzbauer et al. 2017
Greece	Bronze age farming	4000	sediment	6	n-alkanes; sterols; PAHs; GDGT; pollen; microcharcoal	Yes	V,F,A	Yes	Thienemann 2017
Greenland, France	Norse and Middle Age farming	2000	sediment	2	sterols; bile acids	No	V,A	Yes	Zocatelli et al. 2017
New Zealand	Mãori	650	sediment	5	sterols; MA; PAH; charcoal; pollen	Yes	V,F,A	Yes	Argiriadis et al. 2018
France	Romen period farming	2000	sediment	3	pollen; arthropod assemblages; fungal spores	Yes	V,A	Yes	Bajard et al. 2018
Norway	Iron age farming	2000	sediment	3	sterols; n-alkanes	No	V,A	Yes	Balascio and Wickler 2018*
Dominican Republic	Pre-colonial indigenous and colonials	900	sediment	3	fungal spores; pollen; charcoal	Yes	V,F,A	Yes	Castilla-Beltrán et al. 2018
China	Ancient China	1300	sediment	5	n-alkanes; PAHs; sterols; MA; pollen	Yes	V,F,A	Yes	Callegaro et al. 2018*

Mexico, Guatemala	Mayan agriculture	3500	sediment	2	fatty acids; pollen	No	V,F	No	Douglas et al. 2018*
Cameroon	Ancient rainforest clearing	3000	sediment	1	n-alkanes	No	V	No	Garcin et al. 2018
Denmark	Medievel settlement	400	paleosol	3	pollen; arthropod assemblages; eDNA	Yes	V,A	No	Hald et al. 2018*
United States	Yup'ik	500	sediment	3	fungal spores; pollen; charcoal	Yes	V,F,A	Yes	Ledger 2018
Greece	Neolithic farming	5000	sediment	3	pollen; charcoal; stable isotopes δ^{18}O)	Yes	V,F	Yes	Masi et al. 2018
Italy	Baptistery of Padua	1400	paleosol	1	sterols	No	V,A	Yes	Nicosia et al. 2018
Greenland	Thule	600	sediment; paleosol	4	arthropod assemblages; pollen; fungal spores; charcoal	Yes	V,F,A	Yes	Panagiotakopulu et al. 2018
Norway, Netherlands, Denmark, Bahrain, Jordan	Viking, medival, Hellenistic, and Achaemenian settlements	1000	paleosol	1	parasite eDNA	No	V,A	No	Søe et al. 2018
China	Tang Dynasty	1300	sediment	1	charcoal	No	F	No	Tan et al. 2018
South Africa	Pre-european pastoralism	600	sediment	5	pollens; diatoms; coprophilous fungal spores; stable isotopes δ^{18}O); charcoal	Yes	V,F,A	Yes	Cordova et al. 2019
France	Neolithic farming	7700	peat	3	pollen; plant macrofossils; charcoal; fungal spores	Yes	V,F,A	Yes	Dendievel et al. 2019
China	Loulan Kingdom collapse	1800	sediment	2	diatoms; stable isotopes δ^{18}O)	No	V	No	Fontana et al. 2019
Sweden	Iron age farming	1300	sediment	2	pollen; charcoal	No	V,F	Yes	Fredh et al. 2019
Switzerland	Roman farming	2000	sediment	1	stable isotopes δ^{15}N, δ^{13}C)	No	V	No	Haas et al. 2019
Canada	Norse	800	peat	2	pollen; fungal spores	No	V,A	Yes	Ledger et al. 2019

Poland	Neolithic farming	2900	sediment	3	pollen; diatoms; stable isotopes δ^{15}N, δ^{13}C)	Yes	V	Yes	Kinder et al. 2019
Belize	Mayan agriculture	3000	sediment	2	pollen; charcoal	No	V,F	Yes	Krause et al. 2019
Japan	Yayoi period agriculture	1700	paleosol	2	pollen; charcoal	No	V,F	Yes	Naito et al. 2019
Russia	Early Iron Age farming	2000	peat	1	pollen; plant macrofossils	No	V	Yes	Novenko et al. 2019
Cambodia	Angkor collapse	1900	sediment	2	pollen; charcoal	No	V,F	No	Penny et al. 2019
Russia	Zaisanovskaya culture	2050	paleosol	2	diatoms; pollen	No	V	Yes	Razjigaeva et al. 2019
United States	Cahokia decline	900	sediment	2	stable isotopes δ^{18}O); sterols	No	V,A	Yes	White et al. 2019*
France	Bronze age mining	2500	sediment	1	metals	No	A	Yes	Elbaz-Poulichet et al. 2020
Switzerland	Neolithic pile dwellers	5000	sediment	5	pollen; charcoal; n-alkanes; fatty acids; fungal spores	Yes	V,F,A	Yes	Haas et al. 2020
China	Neolithic pastoralism	6000	sediment	3	pollen; fungal spores; stable isotopes δ^{18}O)	Yes	V,A	Yes	Huang et al. 2020
Kyrgyzstan	Bronze age pastoralism	5900	sediment	3	pollen; sterols; n-alkanes	Yes	V,A	Yes	Schroeter et al. 2020

*Synthesis of previously published results

[#]V=Vegetation; F = Fire; A = Animal/Human

PALEOLIMNOLOGICAL PROXIES USED IN ARCHEOLOGICAL STUDIES

Many proxies are used to infer the presence of human settlements and their impacts on the environment (Mills et al. 2017; Dubois et al. 2018). We do not include all such proxies, but rather highlight biological and chemical proxies (Figure A1.1), because they are of greatest utility for interpreting human impacts associated with archeological sites, and because they bring many new opportunities to the field. Only proxies in the 89 studies cited in this review are discussed.

Search strategies and study selection

We used Scopus and Google Scholar to identify studies that used paleolimnology to investigate archaeological sites. The human-limited search terms included: “Paleo*,” “archeo*,” and “archeo” combined with OR. The matrix-limiting search terms included: “soil,” “sediment,” and “peat” using the OR operator. Lastly, the proxy-limited search terms included: “prox*” OR “biomarker.” Unfortunately, the language used for keywords and in abstracts is not consistent in this field, and this presented a major challenge to finding relevant papers. Additional searches were performed by tracking down citations. We excluded studies not published in English and those that used paleolimnology to reconstruct post-industrial-age (after 1760s) anthropogenic impacts. Reviews were excluded, but are mentioned in the main text. Only biological and chemical proxies are included in this review and we limited studies to sediment, peat, and soil paleoarchives. The data extracted from each study included the biological and chemical proxies used, the name of the culture investigated, and the number of proxies used. We assessed orthogonality of the proxies in each study. Eighty-nine studies met the criteria (Table 1).

Biological proxies

Charcoal

Charcoals are the remains of incompletely combusted biological material and are a common proxy for fire occurrence and intensity. Charcoal is described as microscopic (<100 µm) or macroscopic (>100 µm), and such particle size classes are deposited regionally or locally, respectively (Whitlock and Millspaugh 1996; Clark and Patterson 1997; Whitlock and Larsen 2001). Charcoal records cannot be used to determine whether a fire was natural or anthropogenic, if it occurred within the catchment, or, in the case of microscopic charcoal, if it was blown in from afar. Thus, charcoal is rarely used alone in paleolimnological studies that investigate prehistoric human occupation, and requires other proxies to link the charcoal in the sediment to the archeological record. Fire was used by many cultures for slash-and-burn agriculture. Vegetation was felled, allowed to dry, and then burned to transfer nutrients to local soils. Slash-and-burn practices generated abundant charcoal preserved in lake deposits. For example, Thienemann (2017) used charcoal in sediments from Lake Dojran (Macedonia/Greece) to document Bronze Age slash-and-burn agriculture. Many other studies of charcoal in lake and peat deposits have documented the use of fire near archeological settlements (Table 1).

Pollen

Pollen is a commonly used paleolimnological proxy in archeological studies. Pollen comes from many plant taxa, but pollen counts can be biased. Biases in pollen counts are attributable to multiple factors (Albert et al. 2016) including taxonomic differences in pollen production, means of pollination (aerophilous versus entomophilous), and dispersal and preservation of grains (Bennett and Willis 2001). Pollen grains of some species disperse widely, whereas others are deposited locally (Ritchie 1995;

Bourgeois et al. 2001). Pollen of insect- and other animal-pollinated plants can be rare or absent in sediments (Ritchie 1995). Some regions (e.g. the High Arctic) have few local pollen sources, so pollen records from such areas are hard to construct. Nevertheless, in temperate regions, pollen from agricultural sources (e.g. *Cannabis* sp., *Chenopodium quinoa, C. pallidicaule, Zea mays,* etc.) can indicate human land use, whereas pollen from native vegetation can be used to explore changes in catchment vegetation that may or may not derive from human activity. Shumilovskikh et al. (2016) studied pollen, fungal spores, charcoal, *n*-alkanes, and insect fossils in a 6000-year record from Kongor Lake, Iran. They documented the presence of pollen from multiple crops (i.e. *Sambucus ebulus*, *Chenopodium glaucum*, *C. album, Sambucus ebulus, Valerianella dentata, Solanum nigrum, Sonchus arvensis, Polygonum aviculare* agg.) and linked the rise and fall of ruderal plants (i.e. taxa that colonize disturbed areas) to other proxies that they related to human activity (i.e. fire occurrence, fungal spores, and insect fossils). Ledger (2018) demonstrated that circumpolar hunter-gatherers also left a record of their presence. In peat archives from a Yup'ik village in Nunalleq (Alaska, USA) shifts in ruderal plants (e.g. apophyte *Montia fontana*), fungal spores and charcoal, provided evidence for the influence of hunter-gatherers on the paleoenvironment. Ledger (2018) suggested that a shift in apophytes 360-560 years ago marked the beginning of peat accretion, and the development of a meadow-like environment was linked to the construction of drainage systems, for sod-cutting.

Diatoms

Diatoms are excellent proxies for tracking changes in lake productivity, pH, temperature, dissolved organic carbon, and other physical and chemical variables (Smol and Cumming 2000; Smol et al. 2006). They are ubiquitous in lake environments and are autochthonous, so they are typically

associated with local catchment changes (though local catchment changes can also be related to broader climate-induced changes). Transfer functions enable the quantification of environmental conditions (e.g. lake pH or salinity) from the relative abundances of sensitive diatom species. This approach, however, is not without pitfalls, as transfer functions can be unreliable in certain circumstances (Juggins 2013). Transfer functions do not always need to be used, however, because general changes in relative species abundance and distribution can be informative. For example, the diversity of diatom assemblages from lake sediments was used to gain insights into Thule activities in the Canadian High Arctic. Diatom assemblages in the sediments changed as a result of nutrient enrichment from Thule hunting practices (Douglas et al. 2004; Hadley et al. 2010a, 2010b). Lakes of the High Arctic are typically very unproductive (i.e. nutrient-poor), however in areas occupied by Thule people, ponds often display higher productivity, with expansion of moss substrates and algal growth, and more zooplankton. Hadley et al. (2010a; 2010b) attributed this to greater nutrient input from the hunting of whales, seals, caribou, and from human waste, with those nutrients still cycling through the local ecosystem centuries later. Areas occupied by the Thule became hubs of environmental productivity that is evident in many sites to this day.

Spores of coprophilous fungi

Obligate coprophilous fungi commonly grow on animal, especially herbivore, dung (van Geel et al. 2003; Gauthier et al. 2010), and indicate the presence of animal populations. Interpretation of spores from such fungi in lake sediments, however, is not straightforward, because fungal growth is affected by humidity, temperature, and species competition (Baker et al. 2013). Few studies have applied coprophilous fungi spores and spore data are often correlated with other proxies to increase the

reliability of interpretations. Fungal remains in lake sediments and paleosols informed studies of Norse occupation in Greenland (van Geel et al. 2003; Gauthier et al. 2010; Guillemot et al. 2015, 2017). Guillemot et al. (2017) demonstrated the utility of coprophilous fungi spores by correlating multiple proxies for fecal input in a sediment core. They reported positive stratigraphic correlations among bile acid, sterols, and fungal spores, which corroborated with the estimated timeline of Norse occupation.

Arthropod assemblages

A variety of arthropod remains found in lake sediments are associated with human activities. Abundances of some taxa are associated with fecal matter or specific vegetation types (i.e. crops or native vegetation). Identification of arthropod remains in sediments is difficult. Limited spatial sampling means one core is not representative of all the arthropods existing in an entire region and the remains can lose identifying features over time (Kenward 1976; Panagiotakopulu and Buchan 2015). Arthropod assemblages found in sediments come from the most recalcitrant species and represent only a fraction of the historical species diversity (Kenward 1976; Panagiotakopulu and Buchan 2015). Nevetheless, arthropod assemblages in lake sediments have been used in archaeological studies (Elias 2010). Orbatid mites (Arachnida) are associated with camelid feces, and their presence in lake sediments in the Andes was linked to increased pastoralism and travelling caravans at the time of the Incan Empire (Chepstow-Lusty 2011). The authors reported increasing orbatid mite remnant adundances with the rise of the Incan Empire, then decreasing numbers with the empire's eventual decline. A study in Greenland reviewed the arthropod assemblages related to Norse sites (Panagiotakopulu and Buchan 2015). In particular, they noted changes in Coleoptera (beetles) and Hemiptera (true bugs) species associated with Norse haymaking (Panagiotakopulu and Buchan 2015). The presence of parasites, like lice

assemblages, was also associated with Thule winter houses (Dussault et al. 2014) and Norse settlements in Greenland (Barlow et al. 1997; Panagiotakopulu et al. 2007).

Chemical proxies

Radioisotopes, stable isotopes, and metals

Radioisotopes (^{14}C, ^{210}Pb, ^{137}Cs) are used to date sediment profiles. Radiocarbon can also be utilized as a tracer for carbon storage and mobilization in watersheds when the ^{14}C of specific organic components (e.g. fatty acids, *n*-alkanes, miliacin) is determined (Diefendorf and Freimuth 2017). Such information can be used to date when specific organic compounds in the sediment were deposited and explore watershed soil erosion in the context of archeology (Courel et al. 2017; Douglas et al. 2018; Haas et al. 2019).

Stable isotopes are proxies that can be used to reconstruct human-driven environmental changes in archeological studies. Stable isotopes of carbon ($\delta^{13}C$) and nitrogen ($\delta^{15}N$) in radiocarbon-dated cereal grains, along with pollen, can provide information on ancient crop watering and field conditions (Vignola et al. 2017). $\delta^{15}N$ was used to document Thule occupation in the High Arctic (Douglas et al. 2004; Hadley et al. 2010a, 2010b). Douglas et al. (2004) used the $\delta^{15}N$ values from pond sediment cores to identify when the area was occupied by the Thule because they subsisted primarily on marine animals (e.g. bowhead whales or seals) that are enriched in ^{15}N (Fry 1988; Hobson and Schell 1998; Hirons et al. 2001). Greater values of $\delta^{15}N$ coincided with archeologically estimated dates of Thule occupation and they concluded that $\delta^{15}N$ could be used to estimate dates of Thule occupation in sediment cores that had insufficient ^{14}C from sedimented terrestrial vegetation to generate reliable dates, a common problem in the High Arctic (Hadley et al. 2010a, 2010b). Lastly, changes in $\delta^{15}N$ can

also be indicative of eutrophication. Haas et al. (2019) reported that elevated δ^{15}N values in lake sediments in Switzerland coincided with greater human occupation and suggested the increase reflected early eutrophication from sewage and manuring practices.

Stable isotopes can also provide insights into past climate (Grossman and Ku 1986; Sachse et al. 2012; Steinman and Abbott 2013), and although commonly used paleoclimate proxies (e.g. δ^{18}O and δ^{2}H) cannot be used directly to infer the presence of ancient human influence, they can provide critical insights into the effect of climate changes on archeological settlements. Studies have used stable isotopes to explore paleoclimate in archeological contexts, but only in a few regions, e.g the Maya area (Brenner et al. 2002; Douglas et al. 2016). Tinner et al. (2003) used δ^{18}O and pollen from sediment to evaluate the influence of climate (i.e. temperature) on human land-use phases in western and central Europe. They found temperature correlated strongly with prehistoric agriculture yields, and linked land abandonment with cereal crop losses caused by lower temperatures (Tinner et al. 2003). White et al. (2019) compiled existing δ^{18}O and sterol data from sediment to evaluate past climate changes and human responses at the Cahokia Mounds site (Illinois, USA). They determined that warm-season droughts, inferred using δ^{18}O, reduced their agricultural capacity and the population size shrank in response, according to sterol proxies (White et al. 2019). The study emphasized the utility of using paleoclimate proxies in conjunction with indicators of human presence to demonstrate the intimate link between climate and humans.

Metals, especially lead (Pb), have been used to gain insights into the history of mining and smelting by ancient cultures. Increased Pb deposition in sediments has been correlated with many cultures (Boyle 2002; Bindler 2011; Cooke and Bindler 2015). Pompeani et al. (2013) analyzed lake sediments and demonstrated simultaneous Pb increases across multiple sediment cores as a

consequence of copper smelting by prehistoric North Americans (Michigan, USA) ~8000 to 7000 years ago. Changes in lake sediment Pb have also been attributed to prehistoric silver mining (Shotyk et al. 1998; Abbott and Wolfe 2003; Cooke et al. 2008), and copper and Pb mining to make bronze (Cooke et al. 2007; Stanley et al. 2007; Lee et al. 2008; Veron et al. 2013). Ratios of Pb isotopes differ according to the origin of mined ores, which enables the determination of Pb sources in sediment archives (Cooke and Bindler 2015). Other metals are proxies for soil erosion, which reveal prehistoric agricultural activities that caused soil loss. Metals like titanium (Ti), iron (Fe), lead (Pb), and aluminum (Al) are used as proxies for soil erosion (Shotyk et al. 1998; Massa et al. 2012; Doyen et al. 2013; Guillemot et al. 2017).

Polyaromatic hydrocarbons

Polyaromatic hydrocarbons (PAHs) are combustion products derived from organic matter. Fires incompletely combust organic matter, resulting in aeolian PAHs that are later deposited locally and regionally (Meyers and Ishiwatari 1993; Lima et al. 2005). PAH sources can be difficult to determine because they can travel long distances through the air, and originate from both natural and anthropogenic fires. Thus, PAHs alone do not provide definitive evidence of human presence. Recent studies by D'Anjou et al. (2012) and Thienemann (2017), however, demonstrated the utility of PAHs at archeological sites in Norway and Greece, respectively. Slash-and-burn farming techniques increased PAHs and charcoal in lake sediments and were correlated with decreases in *n*-alkanes and pollen, indicating fewer woody plants on the surrounding landscape (D'Anjou et al. 2012; Thienemann 2017). PAHs in sediments were later used in archeological research by Argiriadis et al. (2018), who linked them to fires associated with human occupation in New Zealand. They combined data from PAHs,

monosaccharide anhydrates, charcoal, pollen, and fecal sterols, to show the simultaneous occurrence of large fire events and Mãori settlements around two lakes (Argiriadis et al. 2018). PAHs in the sediments correlated well with other proxies for fire, like charcoal and monosaccharide anhydrates. They also correlated with a decrease in woody plants inferred from pollen, and greater human/animal presence inferred from fecal sterols. Thus, when used in conjunction with other proxies in lake sediments, PAHs contribute to evidence for human occupation.

Monosaccharide anhydrates

Monosaccharide anhydrates (MAs) are also combustion products of organic material, often cellulose (Oros and Simoneit 2001a, 2001b; Elias et al. 2001). The major cellulose derivative is levoglucosan, but other minor MAs include mannosan and galactosan (Oros and Simoneit 2001a). Like PAHs, MAs can be transported long distances before they are deposited and do not directly indicate human presence (Zennaro et al. 2014). Evidence, however, exists that MA ratios reveal MA source material from fires. For instance, levoglucosan/mannosan or levoglucosan/(mannosan + galactosan) ratios tend to be higher for grass fires than for woody fires (Oros and Simoneit 2001a, 2001b; Engling et al. 2006; Kuo et al. 2011). Three studies thus far have used such sediment records for archeological investigation. Schüpbach et al. (2015) used pollen and charcoal, along with MAs, to explore ancient Maya agricultural practices in Guatemala. Argiriadis et al. (2018) used MAs to link Mãori settlement to the regional fire history in New Zealand, and Callegaro et al. (2018) used MAs, along with *n*-alkanes, PAHs, sterols, and pollen to investigate human settlement around Paru Co, in China.

Benzene polycarboxylic acids

Black carbon is the sum of organic fire residues derived from charcoal/soot particles (Glaser et al. 1998), and can be oxidized to benzene polycarboxylic acids (BPCAs), which are used to infer fire temperature (Brodowski et al. 2005; Lehndorff et al. 2015). Lehndorff et al. (2015) pioneered an approach using BPCAs to link prehistoric fires to human settlement. They converted the black carbon in soil samples from Morocco to BPCAs, and linked fire temperature to increased grass fire frequency during the Neolithic period (Lehndorff et al. 2015). This method, however, can only be applied to paleosols because mellitic acid, a BPCA, is preferentially reduced chemically in marine and lacustrine sediments, leading to erroneous inferred fire temperatures (Lehndorff et al. 2015). Furthermore, the diagenesis of black carbon is still poorly understood (Masiello and Louchouarn 2013; Jaffe et al. 2013).

n-alkanes

n-alkanes are derived from the epicuticular leaf waxes of terrestrial plants, and are useful to investigate vegetation changes caused by humans (Eglinton and Hamilton 1967). The *n*-alkane chain length can be used to distinguish algal (short chain) from higher plant (long chain) sources, which can be used to infer past temperature and moisture changes (Howard et al. 2018). Like pollen, *n*-alkanes can be transported over long distances and thus represent both regional and catchment-derived vegetation. *n*-alkanes were first used in the 1990s for archeological study (Bull et al. 1999; Simpson et al. 1999), but were not used extensively until recently. A number of recent studies used *n*-alkanes in sediments, along with other proxies, like pollen and sterols, to infer past vegetation changes related to human occupation (D'Anjou et al. 2012; Shumilovskikh et al. 2016; Thienemann 2017; Guillemot et al. 2017; Schwarzbauer et al. 2017; Garcin et al. 2018; Callegaro et al. 2018). *n*-alkane concentrations correlate with fire

occurrence when ancient cultures employed slash-and-burn agriculture. For instance, Thienemann (2017) and D'Anjou et al. (2012) reported a shift from mid-molecular-weight *n*-alkanes (C_{25}, C_{27} and C_{29}) which come from woody plants, to high-molecular-weight *n*-alkanes (C_{29} and C_{31}), which come from grasses/herbaceous plants. The change was associated with the "slash" (i.e. removal) of woody plants, and correlated with "burn," i.e. other fire proxies.

Fatty acids

Fatty acids, or *n*-alkenoic acids, are environmental indicators produced by microbes and animals (Cranwell 1973). The fatty acid carbon chain length reveals its origin. The C_{26}-C_{32} fatty acids are derived from higher plants, whereas the C_{16}-C_{18} fatty acids come from algae and aquatic macrophytes (Meyers and Ishiwatari 1993; Thienemann 2017). Like *n*-alkanes, fatty acids in sediment contexts have been used in archeological studies since the 1990s (Bull et al. 1999; Simpson et al. 1999). Two recent studies used fatty acid biomarkers in archeological research. Schwarzbauer et al. (2017) examined fatty acids to reconstruct environmental changes in the Roman harbor of Ephesus (Turkey). Fatty acid concentrations increased when the Roman harbor was active, and authors suggested that the increase was caused by greater waste input into the harbor (Schwarzbauer et al. 2017). Douglas et al. (2018) measured the radiocarbon (^{14}C) in fatty acids from lake sediments in southern Mexico and Guatemala to estimate mean soil transit times of these plant waxes (MTT_{wax}) from the watershed to the lake. Shorter MTT_{wax}, along with pollen data, was interpreted to indicate ancient Maya deforestation and consequent accelerated soil erosion. The authors also used this technique to provide evidence that the tropical dry forest soils of the region have not yet fully recovered to their pre-deforestation condition.

Cannabinol

Cannabinol is a biomarker produced by hemp (*Cannabis* sp.) (Lavrieux et al. 2013). Hemp was important to many early cultures for food, medicine, and fiber (Russo 2007). It was one of the earliest cultivated plants (Russo 2007). Lavrieux et al. (2013) were able to detect cannabinol in lake sediments in France, and found that cannabinol tracked the development of hemp agriculture. Lavrieux et al. (2013), however, found amounts of hemp pollen and cannabinol concentrations diverged at times. Cannabinol was not detected at some depths in the core where hemp production was documented by pollen. They came up with two hypotheses to explain the lack of agreement between the two proxies for hemp: 1) pollen was transported from a distant region, whereas cannabinol was produced locally, so the two measures reflect different hemp sources; 2) in the absence of evidence for changing sedimentation rates, they suggested changes in environmental conditions (e.g. light intensity) lowered phytocannabinol in the plants, thereby reducing the cannabinol in the sediment.

Terpenoids

Terpenoids are a large class of molecules synthesized from squalenes; they are structurally diverse and produced by all life forms, but terpenoids can be source-specific. Here we focus on terpenoids produced by plants. Other terpenoids, like sterols/stanols, chlorophyll-related pigments, and bile acids, are addressed in later sections because they are more established paleolimnological proxies. Plant-derived terpenoids used in archeological studies are pentacyclic triterpene methyl ethers (PTMEs), tricyclic diterpenoids (TDs), and carotenoids (aka tetraterpenoids).

PTMEs are large lipid molecules that often originate from higher plants. Example PTMEs commonly used in archeological studies include miliacin (Jacob et al. 2008a, 2008b, 2009; Bossard et al.

2013; Simonneau et al. 2013; Motuzaite-Matuzeviciute et al. 2016; Courel et al. 2017), and trimethyl tetrahydrochrysene (TTHC) (Jacob et al. 2009; Guillemot et al. 2017). Jacob et al. (2005) linked PTMEs to a range of plant taxa, and Trendel et al. (2010) linked different PTMEs to other angiosperms. Work by Jacob et al. (2008a; 2008b) and Bossard et al. (2013) on lake sediment cores from the French Alps validated the use of miliacin, a PTME produced by broomcorn millet, as a biomarker for prehistoric agriculture. In the same region, Jacob et al. (2009) used sedimentary miliacin and TTHC to investigate Neolithic-age farming. TTHC is a diagenetic product of leaf litter and the authors used it as a proxy for soil erosion into lakes. They found a strong correlation between miliacin and TTHC, suggesting soil erosion increased because of millet production around 1700 BC. In addition, isotope measurements of miliacin have proven useful for archeological studies. Jacob et al. (2008b) used δ^{13}C of sedimentary miliacin to link *Panicum miliaceum*, the variety of millet used in Bronze Age farming in the French Alps, to agriculture in the area. Courel et al. (2017) used radiocarbon (^{14}C) measurements of miliacin to establish when miliacin was deposited in paleosols at Obernai, in France. They demonstrated that millet production started earlier than expected, and provided new insights into Bronze Age farming in the area.

TDs and carotenoids are also in the terpenoid family of molecules. They are typically associated with higher plants and algae and thus can be used as proxies for vegetative input into natural archives (Diefendorf et al. 2012; Giri et al. 2015; Rees-Owen et al. 2018). TDs and carotenoids, however, have not been used in many archeological studies. Schwarzbauer et al. (2017) used TDs (abietane and retene) and carotenoids (loliolide and β-cyclocitral) to track changes in vegetation at the Roman harbor of Ephesus in Turkey. Abietane and retene were especially informative about increased inputs from conifers when

the harbor was active because resins were used to coat ship hulls, ropes, and barrels (Schwarzbauer et al. 2017).

Chlorophyll-related pigments

Chlorophyll-related pigments are also terpenoids produced by phototrophic organisms such as phytoplankton, algae, bacteria, and higher plants, and they, along with their degradation products, have an extensive history of being measured in sediments (Leavitt and Hodgson 2001). Algae and bacteria are thought to be the major sources of pigments in most lake sediments, because it is generally assumed that accumulation of aquatically derived pigments on the lake bottom occurs at rates greater than the accumulation of pigments from terrestrial plants (Leavitt and Hodgson 2001; Keely 2006). Pigments can be used to infer past lake trophic state (Pienitz et al. 2000; Michelutti et al. 2007; Antoniades et al. 2011; Rühland et al. 2013; Florian et al. 2015). Hadley et al. (2010a; 2010b) used pigments, in conjunction with diatoms, to infer changes in lake primary productivity caused by Thule occupation in the High Arctic, Canada.

Sterols

Sterols are ubiquitous in microbial, animal, and plant cell membranes and control membrane fluidity (Mouritsen and Zuckermann 2004). Stanols are the diagenetic reductive products of sterols (Mackenzie et al. 1982). Individual sterol/stanol compounds can be highly source-specific, like coprostanol, which is indicative of animal fecal input, and a commonly used proxy for fecal input in paleolimnoloigical studies at archeology sites (Goodfellow et al. 1977; Grimalt et al. 1990; Chan et al. 1998). Several studies have used sterols and stanols to infer past human occupation. D'Anjou et al.

(2012) used different sterols (i.e. total fecal 5β-stanols, 5β-stigmastanol, and coprostanol), along with PAHs and *n*-alkanes, to explore human occupation at an archeological site in Norway over the last 2500 years. They hypothesized that changes in biomarker concentrations correlated with socio-economic events of the time (e.g. Merovingian Depression), climate changes (e.g. Little Ice Age), and even a pandemic (i.e. plague) (D'Anjou et al. 2012). There are, however, limitations to using stanols and sterols in paleoenvironmental archives because these proxies originate from other animals, like migratory birds (Cheng et al. 2016; Thienemann 2017), which can confound interpretation of results. Stanol production by anaerobic bacteria can also occur within a core (Shah et al. 2007; Bataglion et al. 2015). To account for *in situ* production of stanols, Thienemann (2017) used the ratio of total organic carbon to total sulfur (TOC/TS) to track changes in redox conditions. A lower ratio indicates more reducing conditions, which the author used to infer more *in situ* stanol production.

Bile acids

Bile acids are indicators of fecal input from animals because they are byproducts of microbial activity in animal digestive systems (Kuhajda et al. 2006). Bull et al. (2003) suggested bile acids are more recalcitrant than sterols and may be more sensitive biomarkers for fecal input detection, in agreement with a study on the degradation of modern sewage (Elhmmali et al. 1997). Zocatelli et al. (2017) used bile acids in sediments to investigate European cultures of the Middle Ages in France, and the Norse in southern Greenland. They combined bile acid and sterol analysis in multiple paleoarchives (middens, soils, sediments) to reconstruct past land use at both sites. In addition, they used recent fecal samples to determine if bile acids were species-specific, and found that the bile acid signal can be used to

discriminate between fecal input from omnivores (e.g. humans and pigs) and herbivores (e.g. cows, horses, and sheep).

Environmental DNA

Environmental DNA (eDNA) in lake cores, also known as sedimentary ancient DNA (sedaDNA), is derived from animal, microbial, and plant sources, and can remain source-specific for many years after deposition. DNA degradation can, however, occur rapidly under some environmental conditions. Long-term DNA preservation in sediment requires adsorptive materials (e.g. clays, sands, humic substances, etc.), cold temperatures, and anoxic conditions (Pietramellara et al. 2009). Interpretation of eDNA can be challenging, and appropriate care must be taken at all analytical (e.g. sampling, extraction, and measurement) and data processing steps to prevent contamination (Cooper and Poinar 2000; Llamas et al. 2017; Key et al. 2017; Lan and Lindqvist 2019).

Several recent studies used eDNA to explore human pastoralism. Seersholm et al. (2016) used eDNA from 4000-year-old midden deposits in Greenland to demonstrate bowhead whale hunting by the paleo-Inuit (Saqqaq). The authors provided a comprehensive record of animal and plant eDNA (ringed seal, harp seal, caribou, walrus, parasite, canine, and flora) and inferred what subsistence foods were used by the Thule, Saqqaq, Dorset, and Norse (Seersholm et al. 2016). They also used a multivariate analysis to distinguish Dorset, Thule, and Saqqaq cultures from one another using the midden DNA contents (Seersholm et al. 2016). Other studies focused on eDNA from different sources and linked the presence of those sources with human occupation. For instance, Giguet-Covex et al. (2014) used eDNA to infer pastoralism by detecting the presence of cattle and sheep, whereas other studies inferred

pastoralism by detecting the eDNA of farming/human-specific parasites (Søe et al. 2015, 2018; Hald et al. 2018).

PROXY FRAMEWORK

Why do we need a framework?

Our systematic review of 89 studies published from 1983 to 2020 highlighted several key observations: 1) 56% (50) of the studies were published since 2015 (Table 1), indicating that the integration of archeology and paleolimnology is a rapidly growing field. 2) These studies are being conducted worldwide (Figure A1.2), but the distribution of research activity is uneven and many under-investigated areas exist. These gaps indicate numerous opportunities for future researchers to expand into these areas. 3) The absence of consistent terminology made typical search methods difficult. This lack of coherence is a missed opportunity for the field because it complicates the integration of archeological data with other broader climatological and environmental data from paleolimnology. These integrative studies are key for the critical appraisal of research studies, and to synthesize findings across cultures and larger regions. 4) There is increasing need for more robust studies with multiple proxies. These studies are becoming more common, but they are resource demanding and difficult to design. Thus, to address the need for more integration and more robust studies, we designed a framework that can be used to facilitate study designs and standardize the field. The standardized proxy characteristics described here are not new concepts. They are chosen based on the discussion of the advantages and disadvantages of the different proxies, and are also usually described in specialized reviews. To use the framework, these characteristics must be standardized and defined in a paleolimnology context. Many of the terms in the following list of “ideal characteristics” for proxies are

modified from the "ideal traits" used for clinical and risk assessment biomarkers in a list published by the World Health Organization (WHO 1993). In addition, we introduce the term orthogonality, as a method to guide researchers in proxy selection for more robust studies

Defining proxy characteristics

Characteristics that describe an informative and useful proxy include: 1) Specificity, 2) Resistance to degradation, 3) Detectability, 4) Accuracy, 5) Sensitivity, 6) Reproducibility, 7) Dispersibility, and 8) Mobility. This list is not all inclusive, but helps define a meaningful proxy for paleolimnology. This list of terms can be used to decide which proxies complement one another for multi-proxy studies. In Figure A1.3, a schematic of a proxy from source-to-sink is depicted, with the list of characteristics included to illustrate where they are relevant in the process.

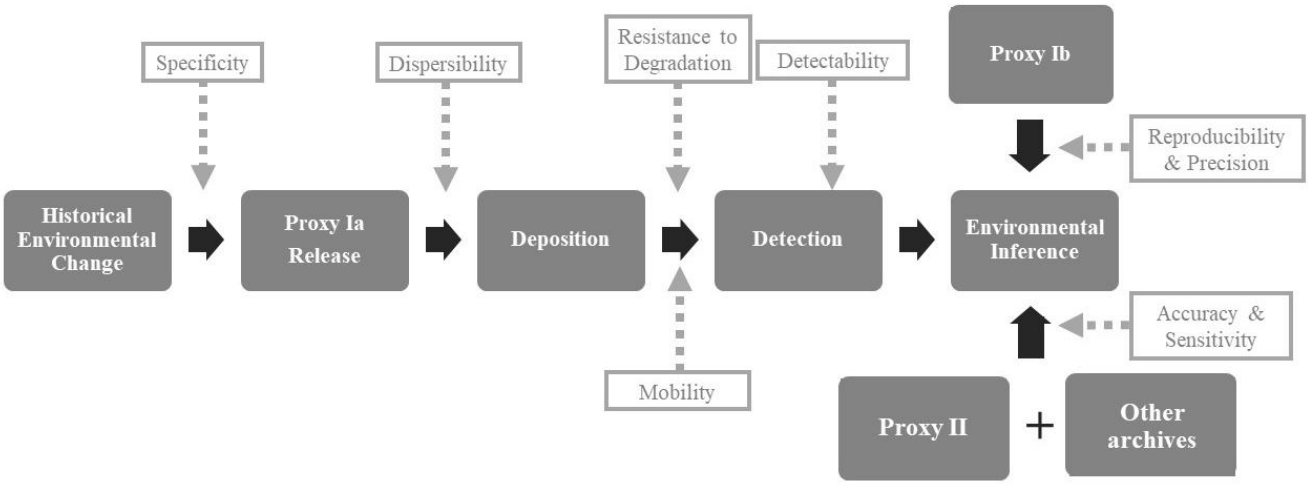


Figure A1.3: A schematic of proxy deposition and detection, illustrating the relevance of proxy characteristics.

Specificity

Specificity describes the particular source, or sources, of a proxy. Simply, if a proxy is produced by multiple sources or environmental conditions, it has low specificity. Bulk measurements of variables

like biogenic silica and total organic carbon (TOC) have low specificity because they are generated by multiple sources. For instance, both diatoms and sponge spicules contribute to biogenic silica, and all organisms can contribute to the sediment TOC pool. Proxies with low specificity are useful only to infer general shifts in past environmental conditions. An example of a highly specific proxy is bile acid, which can be used to identify sources of fecal input from animals (Tyagi et al. 2008).

Dispersibility

Dispersibility refers to whether a proxy is deposited locally, regionally, or even globally. Proxies subject to aeolian transport can be transported widely, depending on physical properties like mass, and their dispersibility can be high. Lead, a proxy for mining and smelting activities, is highly dispersible because it can adsorb to aerosols and be carried to distant regions (Cooke and Bindler 2015). For such highly dispersed proxies, there can be a lag between the time of production and deposition. If that lag period is long, it can lead to a mismatch between archeological dates and the modelled dates of proxy accumulation in lake sediments. Highly dispersed proxies are also susceptible to "blanket" deposition events (i.e. proxies deposited over large areas), which causes the proxy to be detected in sediments far from where the environmental source existed or condition prevailed (e.g. conifer pollen detected in the High Arctic). If a proxy has low dispersibility, it is considered to derive from local sources. Macrocharcoal in lake sediments is an example of a proxy with low dispersibility (Spear 1993). Large proxies like macrocharcoal, however, can extend over a considerable vertical span in the sediment core and confound dating of the profile. Furthermore, proxies with low dispersibility may display very local accumulation and may not appear in all lakes from an area, or even in multiple cores from a single lake.

Resistance to degradation

Once proxies are released from their source they can degrade (Blais et al. 2015; Mills et al. 2017). For instance, molecular biomarkers are susceptible to diagenesis (Meyers and Ishiwatari 1993), which refers to structural changes in organic molecules that occur after deposition (Emerson and Hedges 2003). Many organic compounds are broken down to secondary degradation compounds in the sediment (Meyers et al. 1984), but the degree of diagenesis is situationally dependent. Biological proxies (e.g. siliceous organisms, pollen) are not exempt from degradation. Species-dependent diatom dissolution can occur in the water column and sediment (Smol et al. 2006), and pollen can be susceptible to oxidation (Emerson and Hedges 2003). An ideal proxy is resistant to most degradation, but if the original proxy is degraded in well-defined conditions or a known secondary product forms, it can still be useful for paleolimnological studies. For instance, retene and abietane are secondary products derived from tricyclic diterpenoids in woody plants (Schwarzbauer et al. 2017).

Mobility

Proxies can be mobile within sediments after deposition. Mobilily depends on the proxy, geochemical properties of the sediment (e.g. pH, redox condition, permeability), bioturbation and physical mixing, concentration gradients, and other factors (Zoumis et al. 2001). Mobility can be related to the chemical and physical properties of the substance of interest. For instance, hydrophobic biomarkers with high octanol-water partition coefficients are more likely to associate with solids, which protects them from diffusion and diagenesis (Emerson and Hedges 2003). Hydrophilic biomarkers, on the other hand, are susceptible to diffusion in sediment pore water (Korosi et al. 2015). Metals can also be mobile in sediments, depending on the metal's valence and solubility, and the mineralogy of the

sediment. Many paleolimnological studies assume metals are fixed in the sediment profile, but this is not always true (Kuzyk et al. 2015). In some cases, large-scale mixing events (i.e. large storms, floods) occur that disrupt the stratigraphy.

Detectability

Proxies must be detectable (WHO 1993). Detectability largely pertains to chemical proxies and the analyte of interest must be extractable from the sediment matrix. Molecules that involve many steps for isolation, or that must be isolated under special conditions, require more time and money to analyze, but can yield more informative results. For example, metal species (Blais et al. 2015), polyaromatic hydrocarbons (D'Anjou et al. 2012; Blais et al. 2015), and sedimentary DNA (Seersholm et al. 2016; Niemeyer et al. 2017) are proxies that can require multiple or carefully considered isolation steps, but are very informative for paleoenvironmental inference. Another consideration is the environmental and human health impacts of extraction techniques. Green chemistry strives to reduce the use of toxic and harmful extraction chemicals. Some proxies, like pigments, can be analyzed without extraction using near-infrared spectroscopy (Michelutti and Smol 2016; Schneider et al. 2018), which enables rapid and environmentally friendly analysis. Pollen extraction often requires hydrofluoric acid, which is very hazardous (Smol et al. 2006); however, there are alternate pollen extraction techniques that use density separation (van den Bos et al. 2020). Lastly, detectability should also consider whether a proxy can be measured quantitatively, or only qualitatively. Most chemical proxies can be evaluated quantitatively, enabling calculation of errors.

Reproducibility

Analyses of proxies in lake sediments should be reproducible, which is analogous to the WHO (1993) definition of precision. Proxies used to infer the same environmental condition should yield similar results from different local or regional paleoenvironmental archives. If local and regional landscapes are similar (i.e. geology, soil types, etc.) one can expect to generate reproducible results when the same proxy is measured in sediment cores from a single lake or nearby lakes. Pollen analyses have been shown to yield reproducible results across many sites in a region (Lamb 1985; Ritchie 1987; Gajewski et al. 1995).

Accuracy

Proxies in lake sediments must be able to track historical environmental changes accurately. The ability of a proxy to respond quantitatively to the magnitude of an environmental change (stressor) is an expression of the proxy's accuracy (i.e. a correlation exists between the proxy signal intensity and the magnitude of environmental change). Accuracy should be validated using modern calibration datasets or by comparison with other historical reference archives (e.g. satellite information, historical records, other established proxies, etc.). For instance, Korosi et al. (2017a) used historical Landsat imagery to confirm that lignins and TOC/TN ratios in lake sediments accurately reflected lake expansion events. To understand the accuracy of a proxy, it is essential to know how it was produced and released, and about subsequent transport and/or diagenetic changes (Ritchie 1995; Smol et al. 2006). Accuracy can also be influenced by geographic location. For example, pollen from most plant taxa in the Canadian Arctic are dispersed by wind and are thus found in lake deposits, and can be used to infer past climate. In tropical Africa, however, most plant species are insect-pollinated and thus "invisible" in lake sediments, making

African pollen assemblages less informative about past forest dynamics, and hence paleoclimate (Ritchie 1995).

Sensitivity

Sensitivity is the ability of the proxy to track increasingly small fluctuations in the targeted environmental variable of interest. Sensitivity is dependent on a proxy's temporal resolution in the sediment, which depends on both experimental design (e.g. sampling interval) and proxy characteristics. Innate sensitivity is a function of the source of the proxy and its deposition mechanisms. How well a proxy tracks an environmental variable depends on how reactive the source of the proxy is to environmental change. A sensitive proxy source reacts to the environmental change by releasing a large amount of the proxy, but releases little to no proxy when the environmental trigger ceases. The proxy's sensitivity is also affected by its dispersibility. If dispersibility is high and there is a lag time to deposition, then the signal in the core will be "diluted," i.e. it will not appear as a discrete peak, and the sensitivity will thus be lower. Proxy sensitivity is also dependent on experimental design, in that uncertainties in sediment dating (e.g. ^{210}Pb, ^{14}C, varves) and sampling resolution of the core influence whether past minor fluctuations in the environmental variable of interest will be detectable in the archive. For instance, if a core is sampled at broad intervals, it means that fewer sample depths in the core are analyzed, and they are done so at lower temporal resolution. This reduces the effective sensitivity of the proxy in the core, regardless of its innate sensitivity. Other factors also influence the accuracy of the core chronology, such as the effects of hard water lake error and long water residence times on radiocarbon dates, and the veracity of dating model assumptions (Kaufman et al. 2004; Blais et al. 2015). The greater the age model uncertainty, the lower the functional sensitivity.

Orthogonality

Proxy selection is important in paleolimnology studies, and we promote the use of orthogonal, as opposed to complementary, proxies to provide the strongest evidence for inferring environmental and ecological changes caused by humans (Figure A1.4). Orthogonality refers to two methods or variables that are mechanistically different so that results generated by them are independent. We propose grouping proxies into three different groups based on their sources to achieve orthogonality: environmental, animal/human, and fire. Proxies derived from these three sources, when combined, provide the most information about past human occupation.

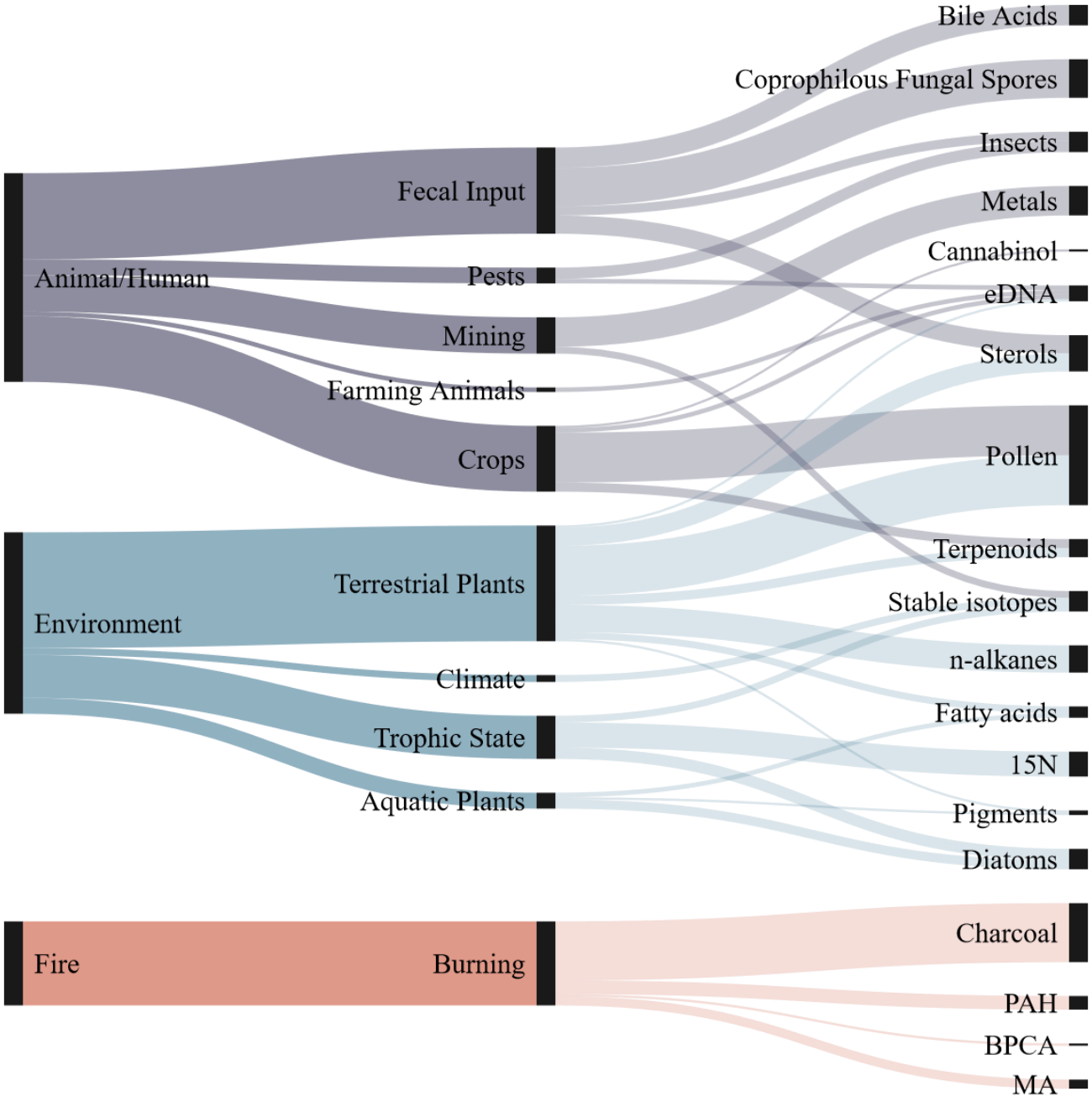


Figure A1.4: Sankey diagram depicting orthogonal proxy groups (left), what they indicate (middle) and proxy used to infer them (right). Relative size of the final nodes (black boxes on the right) indicates number of times the proxies were used across the 89 studies in this review. Diagram made using SankeyMATIC online (http://sankeymatic.com/build/).

Fire is linked to past human activity over many millennia (Glikson 2013), but fires can also have natural causes. Thus, fire occurrence alone is not sufficient to document past human presence. Distinguishing between natural and anthropogenic fires in lake sediment records can be difficult, but some fires have been definitively associated with pre-industrial human settlements.

Environmental proxies (a broad category) are also less specific indicators of human occupation because they can track catchment, ecological, or climate changes, but such shifts are not necessarily human-driven. Humans have a long history of farming and pastoralism, and some vegetation-derived proxies are indicative of human presence, such as certain agricultural pollen types and cannabinol, which would otherwise be absent, if not for human activities. Proxies can also be used to infer past agriculture/pastoralism indirectly, because slash-and-burn agricultural techniques and grazing change the vegetation cover. Climate proxies (e.g. δ^{18}O and δ^{2}H) do not directly track changes caused by pre-industrial human cultures, but are useful to explain shifts in ancient human populations, when used in conjunction with other proxies.

Some proxies are indicators of human and other animal presence. Many such proxies are related to waste products and amounts in the sediment can sometimes be related to past population size. There are, however, instances in which large groups of animals introduced these biomolecules (e.g. sterols, bile acids, eDNA), even in the absence of humans (Haile et al. 2009; Machado et al. 2018; Hargan et al. 2018).

Orthogonal proxies in lake sediments provide information from the same site, but come from different, independent sources. For example, coprostanol, a proxy for human/animal fecal matter, can be paired with proxies for fires and/or vegetation, to explore whether the presence of animals coincided with increased fire frequency from slash-and-burn agriculture, and/or a shift in pollen that indicates

crop cultivation. Use of a proxy or multiple proxies from orthogonal groups provides the strongest evidence for inferring environmental changes caused by humans.

MULTI-PROXY STUDY DESIGN

Choosing proxies

Choosing which proxies to consider in sediment archives is a challenge for archeological investigations. Many proxies have been used, others have potential, and still others remain to be discovered. Past paleolimnological studies in support of archeology have, in some cases, suffered from use of too few and/or poorly chosen proxies.

Proxy choice is determined by several factors: cost, available instrumentation, expertise, and relevance to the culture being studied. As interest in this type of cross-disciplinary research grows, and the costs of analyses drop and/or the ease of analysis increases with new technologies, multi-proxy studies are becoming more common. This is already evident in the studies we have cited here where a few studies have embraced the orthogonal multi-proxy approach. Examples include: D'Anjou et al. (2012), Shumilovskikh et al. (2016), Pini et al. (2017), Schwarzbauer et al. (2017), Thienemann (2017), Argiriadis et al. (2018), Callegaro et al. (2018), Ledger (2018), Panagiotakopulu et al. (2018), and Castilla-Beltrán et al. (2018). Thus, only 15% (13) of the 89 studies presented in this review embraced the orthogonal multi-proxy approach. Of those publications, one was published in 2012, and the other 12 were published since 2016. Future studies should consider employing this emerging orthogonal multi-proxy approach, and proxy choice is critical for this type of study design.

One way to choose proxies is to combine the terms presented in the framework above with decision matrix analysis. Decision matrix analysis is a qualitative technique used to rank options. It is

commonly used in engineering and business to decide which object/material is optimal, based on common criteria (e.g. cost effectiveness, durability, etc.). Decision matrix analysis can be applied to defined proxy characteristics and orthogonality, as criteria. The first step in decision matrix analysis is to set the context for one's decision. In the case of archeological studies, one may want to know: 1) Did the culture of interest use agriculture? Or were they hunter gatherers? 2) Did they use fire for agriculture or mining? Or was it only for local hearth fires? and 3) Where were they located? Next, based on the researcher's capabilities, and those of potential collaborators, a number of proxies can be selected. Once one has an idea of the culture of interest and potential proxies, one can begin to build a matrix. Such a matrix is formatted as in Table A1.2, with criteria in rows and the potential proxies in columns. Each proxy can be scored using a scale (-2, -1, 0, 1, 2), where a score of -2 indicates poor performance in the context of the location and the culture under study, and 2 is the best score. After all proxies have been scored for each category, the scores are summed, and the best-performing proxies will have the highest scores. The last step is to calculate a study score. The matrix proxy sum alone does not inform the researcher about which proxies to select, because the objective is to build an orthogonal multi-proxy study. The researcher can then pick three or more of the scored proxies, and sum their scores. The summed score is multiplied by one if there is no orthogonality, by two if a fire or vegetation proxy is paired, for example, with a human/animal indicator, and by three if selected proxies come from all three orthogonal groups (e.g. vegetation, fire, and human/animal). The objective is to select the combination of scored proxies that has the highest study score, while also taking into account practical considerations such as cost, logistics and the researcher's capabilities.

Table A1.2: Example decision matrix for the case study presented in the text

Criteria	Charcoal*	Plant Terpenoids	Arthropod Assemblages	Coprophilous fungi spores	PAHs	Sterols
Specificity	0	1	1	2	-1	2
Dispersibility	1	0	-1	1	-1	1
Resistance to degradation	1	1	1	1	1	1
Mobility	1	1	1	1	1	1
Detectability	-1	0	1	1	0	0
Reproducibility	0	0	-1	0	0	1
Accuracy	1	1	1	0	1	1
Sensitivity	1	1	1	1	1	1
SUM	4	5	5	7	2	8

* Charcoal example: Charcoal is specific to fires, but not specific to human fires specificity 0). Charcoal can be deposited locally, which is important for study of a culture that had only small localized fires dispersibility 1). Charcoal is resistant to diagenesis, and its preservation is facilitated by the cold conditions resistant to degradation 1). It is not mobile in the sediment mobility 1). Charcoal is extracted using hydrofluoric acid and charcoal particles can be broken during this process, resulting in artificially high charcoal numbers Whitlock and Larsen 2001), and hydrofluoric acid is very toxic. Both are negatives detectability -1). Charcoal is somewhat reproducible, but the analytical method may influence charcoal numbers reproducibility 0). Charcoal has been shown to be a reliable indicator for fires accuracy 1). Lastly, larger charcoal pieces can reduce sensitivity, but because we are using microscopic charcoal, not macroscopic charcoal, our resolution will depend on the number of intervals analyzed, not the size of the charcoal particles sensitivity 1).

To demonstrate this concept, we present an example. In this case study, we are researching Culture X. They were semi-nomadic hunter-gatherers and hunted large marine mammals, e.g. whales, seals, etc. They had small localized hearth fires for cooking and warmth. We have laboratories with access to instrumentation to measure PAHs and sterols. We must, however, choose between working with two potential collaborators because of limited sediment quantities. One collaborator (A) can analyze microscopic charcoal and plant terpenoids, and the other (B) can analyze insect fossils and coprophilous fungi spores. We begin filling in our matrix (Table A1.2). Once the table is completed, scores can be calculated for each proxy. The scores are then used to calculate a study score. The highest

possible score, using all of the proxies, is 31, and since proxies are present from all orthogonal groups (i.e. Fire indicators = charcoal and PAHs; Vegetation indicators = plant terpenoids, insect fossils, and sterols; Human/animal indicators = cannabinol, sterols, insect fossils, and coprophilous fungi spores) the score can be multiplied by three (31 x 3) and equals 93. Using all the proxies, however, is not feasible for several reasons. PAHs have the lowest scores because Culture X had only small localized hearths, which likely did not produce enough PAHs to track the fire record. So PAHs were removed from the list.

Now we have to choose between collaborators A and B, as to who can best complement our sterol analysis. If we choose collaborator A who analyzes plant terpenoids and charcoal, our study score would be 51 (4 *(Charcoal)* + 5 *(Plant terpenoids)* + 8 *(Sterols)* = 17 x 3 = 51). If we choose collaborator B who analyzes insect assemblages and coprophilous fungi spores, our study score would be 40 (5 *(Arthropod assemblages)* + 7 *(coprophilous fungi spores)* + 8 *(Sterols)* = 20 x 2 = 40). Thus, the best choice for an orthogonal multi-proxy study would be collaborator A, because even though there is little vegetation in the region, the areas adjacent to Culture X archeological sites have abundant algae, moss, and lichen which still produce terpenoids that have been detected in geological samples (Romero-Sarmiento et al. 2011; Tian et al. 2017). What, however, if an orthogonal study design is not feasible? For example, Culture X's small hearth fires were not large enough to leave a record in the sediment. If there is no fire proxy, and full orthogonality cannot be reasonably achieved, the decision matrix approach is still useful because a study with a greater number of proxies from two orthogonal groups can have a higher study score than a study with proxies from all three orthogonal groups. Thus, although complete orthogonality is best, study scores are useful for cases where orthogonality is not possible, and our framework, used with decision matrix analysis, can help researchers choose proxies and design better paleolimnological studies of archeology sites.

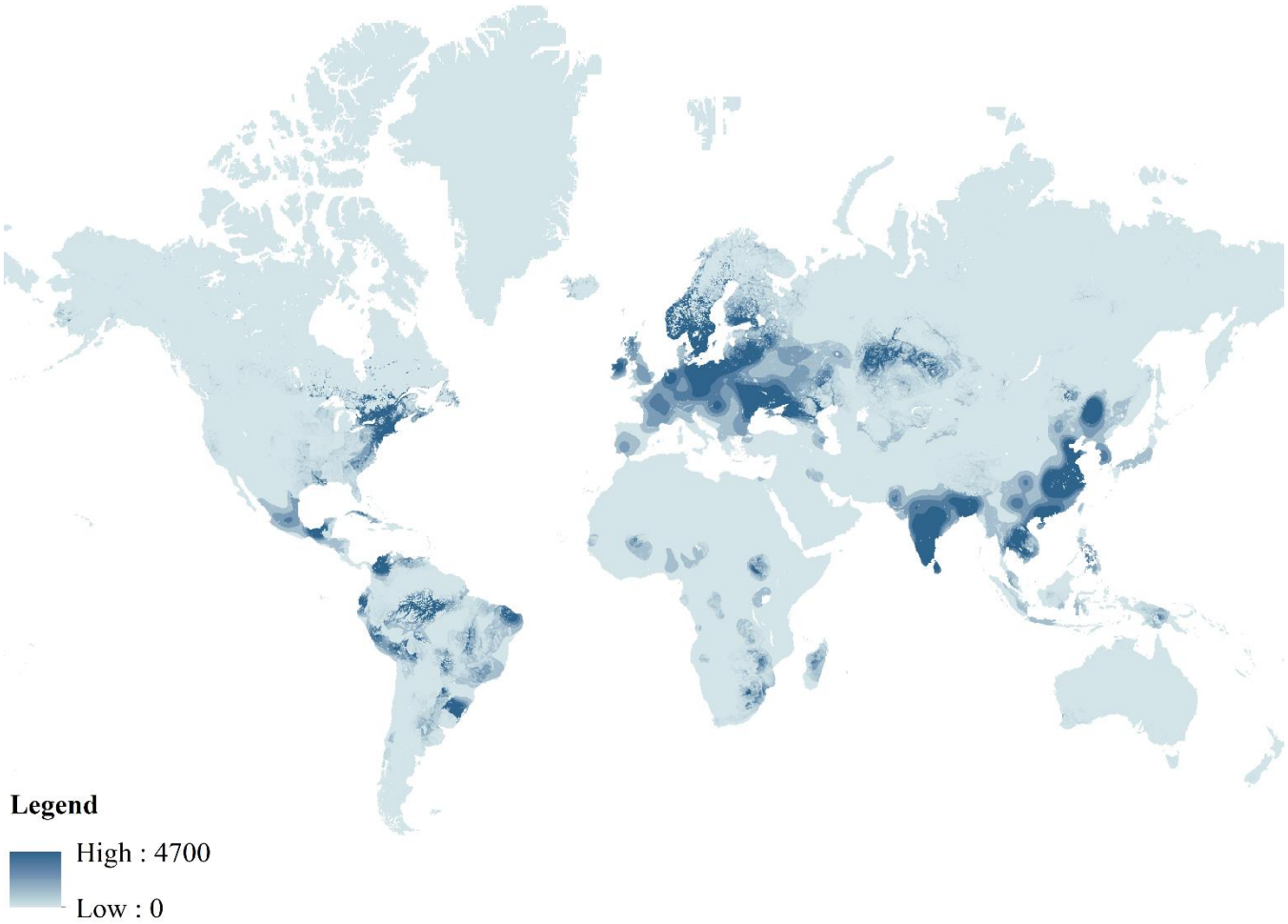


Figure A1.5: A global estimate of likely areas for paleolimnological study in support of archeology produced using ArcGIS (v10.6.1) and fuzzy overlay (product). Base maps were population counts of pre-industrial civilizations from 10,000 BCE to CE 1760 (HYDE 3.2 database; (Goldewijk et al. 2017)) and global lake maps from the McGill HydroLab (Messager et al. 2016).

FUTURE DIRECTIONS

Many directions have yet to be explored in paleoenvironmental science as a means to inform archeology. Devos et al. (2019) pointed out the need for synthesis studies that combine multiple lines of evidence into a comprehensive picture of the ancient culture under study. These holistic paleoenvironmental syntheses could integrate paleoclimate data (e.g. from tree rings or $\delta^{18}O$) to determine the role that climate played in cultural development. Digital geoarcheology (Siart et al. 2018) can also be integrated with paleolimnological studies. There is a growing trend in digitizing archeological

information (i.e. digital archeology), and online databases of archeological sites are developing, although they often have restricted access or fuzzy coordinates to protect the sites from theft (e.g. CARD [Martindale et al. 2016]), Nunniffiit (Greenland National Museum & Archive 2015), and the Nunavut *Archeological Sites Information Management System (permit access only).* With this information, digital geoarcheology (Figure A1.5) can be used to identify potential areas for future paleolimnological investigations of archeological sites. For example, Figure A1.5 was produced using ArcGIS (v10.6.1) and fuzzy overlay (product) analysis. This analysis overlays different map layers and highlights areas where both layers have co-occurring features. Base maps that were overlaid for this analysis were rasters from the HYDE 3.2 dataset (Goldewijk et al. 2017), which depicts global population estimates from 10,000 BCE to CE 1760, and lake densities produced from a base map of global lakes (Messager et al. 2016). This analysis highlights areas where a high density of lakes existed concurrently with a large human population, over the last ~12,000 years. Dark blue indicates potential research sites for future environmental archeology investigations. This analysis, however, is only as good as the base maps. For instance, the global lakes base map only includes lakes > 10 ha (Messager et al. 2016). Smaller lakes are also very useful for paleolimnological studies because proxies that enter the lake from the watershed are not "diluted" by the large size of the lake as much. Small lakes, however, are often close to areas with many larger lakes, so the correlation between this lake base map and archeological sites may still hold in this case. The HYDE 3.2 dataset is comprehensive, but does not contain many High Arctic cultures. These cultures had very low population densities, so potential field sites for studying these cultures do not appear on the map. Digital archeology and paleolimnology could also be used to support the discovery of archeological sites by Airbourne and Spacebourne Remote Sensing (ASRS; (Luo et al. 2019)). Paleolimnological studies do not disturb sites and require limited human labor to collect samples

for analysis; thus, paleolimnologists are in a unique position to validate proposed archeological sites detected by ASRS, without disturbing the site and before archeologists commit large amounts of resources to excavating.

Another future direction for paleolimnological investigations into archeology is discovery and development of new biomarkers. There have been many recent advances in analytical chemistry, and these new technologies and methods should lead to the discovery of new biomarkers. Recently, we have developed an untargeted analysis workflow to discover new environmental biomarkers using sedimentomics (Bell and Blais 2019) and this approach could be applied to archeological paleoarchives to identify relevant biomarkers. The discovery of new biomarkers would also support the use of multi-proxy orthogonal study designs in paleolimnological studies of archeological sites because there would be more biomarkers, and ideally more specific biomarkers from which to choose.

CONCLUSIONS

In conclusion, we presented a systematic review of 89 studies (Figure A1.2; Table A1.1) of paleolimnological studies for archeology. This review identified that the field is rapidly growing, but there is still need for more studies, especially in areas with low densities of research, but high densities of archeological sites. More study standardization is required for future integrative analyses, and there is increasing need for more robust studies with multiple proxies as the field grows. To address these observations from the review we designed a framework that standardized terminology of proxy characteristics and defined orthogonality so that they could be used with weighted decision matrices to design robust future studies. We also summarized the potential synergy between digital archeology and

paleolimnology, and how new analytical chemistry methods could facilitate paleolimnological biomarker discovery in support of archeological research.

ACKNOWLEDGEMENTS

This research was supported by a Natural Sciences and Engineering Research Council (Canada) Discovery Grant (RGPIN-2018-04248) to JMB.

REFERENCES

Abbott MB, Wolfe AP (2003) Intensive pre-Incan metallurgy recorded by lake sediments from the Bolivian Andes. Science 301:1893–1895. https://doi.org/10.1126/science.1087806

Albert B, Innes J, Blackford J, Taylor B, Conneller C, Milner N (2016) Degradation of the wetland sediment archive at Star Carr: An assessment of current palynological preservation. J Archaeol Sci Rep 6:488–495. https://doi.org/10.1016/j.jasrep.2016.03.010

Antoniades D, Michelutti N, Quinlan R, Blais JM, Bonilla S, Douglas MSV, Pienitz R, Smol JP, Vincent WF (2011) Cultural eutrophication, anoxia, and ecosystem recovery in Meretta Lake, High Arctic Canada. Limnol Oceanogr 56:639–650. https://doi.org/10.4319/lo.2011.56.2.0639

Argiriadis E, Battistel D, McWethy DB, Vecchiato M, Kirchgeorg T, Kehrwald NM, Whitlock C, Wilmshurst JM, Barbante C (2018) Lake sediment fecal and biomass burning biomarkers provide direct evidence for prehistoric human-lit fires in New Zealand. Sci Rep 8:1-9. https://doi.org/10.1038/s41598-018-30606-3

Bajard M, Etienne D, Quinsac S, Dambrine E, Sabatier P, Frossard V, Gaillard J, Develle AL, Poulenard J, Arnaud F, Dorioz JM (2018) Legacy of early anthropogenic effects on recent lake eutrophication (Lake Bénit, northern French Alps). Anthropocene 24:72–87. https://doi.org/10.1016/j.ancene.2018.11.005

Baker AG, Bhagwat SA, Willis KJ (2013) Do dung fungal spores make a good proxy for past distribution of large herbivores? Quat Sci Rev 62:21–31. https://doi.org/10.1016/j.quascirev.2012.11.018

Balascio NL, Wickler S (2018) Human–environment dynamics during the Iron Age in the Lofoten Islands, Norway. Nor Geogr Tidsskr - Nor J Geogr 72:146–160. https://doi.org/10.1080/00291951.2018.1466831

Barlow LK, Sadler JP, Ogilvie AEJ, Buckland PC, Amorosi T, Ingimundarson JH, Skidmore P, Dugmore AJ, McGovern TH (1997) Interdisciplinary investigations of the end of the Norse western settlement in Greenland. Holocene 7:489–499. https://doi.org/10.1177/095968369700700411

Bataglion GA, Meurer E, de Albergaria-Barbosa ACR, Bícego MC, Weber RR, Eberlin MN (2015) Determination of geochemically important sterols and triterpenols in sediments using ultrahigh-performance liquid chromatography tandem mass spectrometry (UHPLC–MS/MS). Anal Chem 87:7771–7778. https://doi.org/10.1021/acs.analchem.5b01517

Bell M, Blais JM (2019) "-Omics" workflow for paleolimnological and geological archives: a review. Sci Total Environ 672:438–455. https://doi.org/10.1016/j.scitotenv.2019.03.477

Bennett KD, Willis KJ (2001) Pollen. In: Smol JP, Birks HJB, Last WM, Bradley RS, Alverson K (eds) Tracking Environmental Change Using Lake Sediments: Terrestrial, Algal, and Siliceous Indicators. Springer, Dordrecht, pp 5–32

Bhiry N, Marguerie D, Lofthouse S (2016) Paleoenvironmental reconstruction and timeline of a Dorset-Thule Ssettlement at Quaqtaq (Nunavik, Canada). Arct Antarct Alp Res 48:293–313. https://doi.org/10.1657/AAAR0015-045

Bindler R (2011) Contaminated lead environments of man: reviewing the lead isotopic evidence in sediments, peat, and soils for the temporal and spatial patterns of atmospheric lead pollution in Sweden. Environ Geochem Health 33:311–329. https://doi.org/10.1007/s10653-011-9381-7

Blais JM, Rosen MR, Smol JP (eds) (2015) Environmental Contaminants: Using Natural Archives to Track Sources and Long-Term Trends of Pollution. Springer, Dordrecht

Bossard N, Jacob J, Le Milbeau C, Sauze J, Terwilliger V, Poissonnier B, Vergès E (2013) Distribution of miliacin (olean-18-en-3β-ol methyl ether) and reputed sources: Implications for the use of

sedimentary miliacin as a tracer of millet. Org Geochem 63:48–55. http://dx.doi.org/10.1016/j.orggeochem.2013.07.012
Bourgeois JC, Gajewski K, Koerner RM (2001) Spatial patterns of pollen deposition in arctic snow. J Geophys Res Atmospheres 106:5255–5265. https://doi.org/10.1029/2000JD900708
Boyle JF (2002) Inorganic geochemical methods in palaeolimnology. In: Blais JM, Rosen MR, Smol JP (eds) Tracking Environmental Change Using Lake Sediments. Springer, Dordrecht, pp 83–141
Brenner M, Rosenmeier MF, Hodell DA, Curtis JH (2002) Paleolimnology of the Maya Lowlands: long–term perspectives on interactions among climate, environment, and humans. Anc Mesoam 13:141–157
Brodowski S, Rodionov A, Haumaier L, Glaser B, Amelung W (2005) Revised black carbon assessment using benzene polycarboxylic acids. Org Geochem 36:1299–1310. https://doi.org/10.1016/j.orggeochem.2005.03.011
Buckland PC, Edwards KJ, Panagiotakopulu E, Schofield JE (2009) Palaeoecological and historical evidence for manuring and irrigation at Garðar (Igaliku), Norse eastern settlement, Greenland. Holocene 19:105–116. https://doi.org/10.1177/0959683608096602
Bull ID, Elhmmali MM, Perret V, Matthews W, Roberts DJ, Evershed RP (2005) Biomarker evidence of faecal deposition in archaeological sediments in Çatalhöyük. In: Inhabiting Çatalhöyük: Reports from the 1995-99 Seasons. McDonald Institute Monographs, Cambridge, pp 415–420
Bull ID, Elhmmali MM, Roberts DJ, Evershed RP (2003) The application of steroidal biomarkers to track the abandonment of a Roman wastewater course at the Agora (Athens, Greece). Archaeometry 45:149–161
Bull ID, Simpson IA, Dockrill SJ, Evershed RP (1999) Organic geochemical evidence for the origin of ancient anthropogenic soil deposits at Tofts Ness, Sanday, Orkney. Org Geochem 30:535–556
Callegaro A, Battistel D, Kehrwald NM, Matsubara Pereira F, Kirchgeorg T, Villoslada Hidalgo M del C, Bird BW, Barbante C (2018) Fire, vegetation, and Holocene climate in a southeastern Tibetan lake: a multi-biomarker reconstruction from Paru Co. Clim Past 14:1543–1563. https://doi.org/10.5194/cp-14-1543-2018
Castilla-Beltrán A, Hooghiemstra H, Hoogland MLP, Pagán-Jiménez J, van Geel B, Field MH, Prins M, Donders T, Herrera Malatesta E, Ulloa Hung J, McMichael CH, Gosling WD, Hofman CL (2018) Columbus' footprint in Hispaniola: A paleoenvironmental record of indigenous and colonial impacts on the landscape of the central Cibao Valley, northern Dominican Republic. Anthropocene 22:66–80. https://doi.org/10.1016/j.ancene.2018.05.003
Chan KH, Lam MHW, Poon K-F, Yeung H-Y, Chiu TKT (1998) Application of sedimentary fecal stanols and sterols in tracing sewage pollution in coastal waters. Water Res 32:225–235 https://doi.org/10.1016/S0043-135497)00175-9
Cheng W, Sun L, Kimpe LE, Mallory ML, Smol JP, Gallant LR, Li J, Blais JM (2016) Sterols and stanols preserved in pond sediments track seabird biovectors in a High Arctic environment. Environ Sci Technol 50:9351–9360
Chepstow-Lusty A (2011) Agro-pastoralism and social change in the Cuzco heartland of Peru: a brief history using environmental proxies. Antiquity 85:570–582
Clark JS, Patterson WA (1997) Background and local charcoal in sediments: scales of fire evidence in the paleorecord. In: Clark JS, Cachier H, Goldammer JG, Stocks BJ (Eds.) Sediment Records of Biomass Burning and Global Change. Springer, Berlin, pp 23–48

Cooke CA, Abbott MB, Wolfe AP (2008) Late-Holocene atmospheric lead deposition in the Peruvian and Bolivian Andes. Holocene 18:353–359. https://doi.org/10.1177/0959683607085134
Cooke CA, Abbott MB, Wolfe AP, Kittleson JL (2007) A millennium of metallurgy recorded by lake sediments from Morococha, Peruvian Andes. Environ Sci Technol 41:3469–3474
Cooke CA, Bindler R (2015) Lake sediment records of preindustrial metal pollution. In: Blais JM, Rosen MR, Smol JP (eds) Environmental Contaminants. Springer, Dordrecht, pp 101–119
Cooper A, Poinar HN (2000) Ancient DNA: do it right or not at all. Science 289:1139–1139
Cordova CE, Kirsten KL, Scott L, Meadows M, Lücke A (2019) Multi-proxy evidence of late-Holocene paleoenvironmental change at Princessvlei, South Africa: the effects of fire, herbivores, and humans. Quat Sci Rev 221:105896. https://doi.org/10.1016/j.quascirev.2019.105896
Courel B, Schaeffer P, Adam P, Motsch E, Ebert Q, Moser E, Féliu C, Bernasconi SM, Hajdas I, Ertlen D, Schwartz D (2017) Molecular, isotopic and radiocarbon evidence for broomcorn millet cropping in Northeast France since the Bronze Age. Org Geochem 110:13–24. https://doi.org/10.1016/j.orggeochem.2017.03.002
Cranwell PA (1973) Branched-chain and cyclopropanoid acids in a recent sediment. Chem Geol 11:307–313
D'Anjou RM, Bradley RS, Balascio NL, Finkelstein DB (2012) Climate impacts on human settlement and agricultural activities in northern Norway revealed through sediment biogeochemistry. Proc Natl Acad Sci USA 109:20332–20337
Demske D, Tarasov PE, Leipe C, Kotlia BS, Joshi LM, Long T (2016) Record of vegetation, climate change, human impact and retting of hemp in Garhwal Himalaya (India) during the past 4600 years. Holocene 26:1661–1675. https://doi.org/10.1177/0959683616650267
Dendievel A-M, Dietre B, Cubizolle H, Hajdas I, Kofler W, Oberlin C, Haas JN (2019) Holocene paleoecological changes and agro-pastoral impact on the La Narce du Béage mire (Massif Central, France). Holocene 29:992–1010. https://doi.org/10.1177/0959683619831416
Devos Y, Nicosia C, Wouters B (2019) Urban geoarchaeology in Belgium: experiences and innovations. Geoarchaeology 2019:1-15. https://doi.org/10.1002/gea.21755
Diefendorf AF, Freeman KH, Wing SL (2012) Distribution and carbon isotope patterns of diterpenoids and triterpenoids in modern temperate C3 trees and their geochemical significance. Geochim Cosmochim Acta 85:342–356. https://doi.org/10.1016/j.gca.2012.02.016
Diefendorf AF, Freimuth EJ (2017) Extracting the most from terrestrial plant-derived n-alkyl lipids and their carbon isotopes from the sedimentary record: a review. Org Geochem 103:1–21. https://doi.org/10.1016/j.orggeochem.2016.10.016
Douglas MSV, Smol JP, Savelle JM, Blais JM (2004) Prehistoric Inuit whalers affected Arctic freshwater ecosystems. Proc Natl Acad Sci USA 101:1613–1617. https://doi.org/10.1073/pnas.0307570100
Douglas PMJ, Brenner M, Curtis JH (2016) Methods and future directions for paleoclimatology in the Maya Lowlands. Glob Planet Change 138:3–24. https://doi.org/10.1016/j.gloplacha.2015.07.008
Douglas PMJ, Pagani M, Eglinton TI, Brenner M, Curtis JH, Breckenridge A, Johnston K (2018) A long-term decrease in the persistence of soil carbon caused by ancient Maya land use. Nat Geosci 11:645–649. https://doi.org/10.1038/s41561-018-0192-7
Doyen É, Vannière B, Berger J-F, Arnaud F, Tachikawa K, Bard E (2013) Land-use changes and environmental dynamics in the upper Rhone valley since Neolithic times inferred from sediments in Lac Moras. Holocene 23:961–973. https://doi.org/10.1177/0959683612475142

Dubois N, Jacob J (2016) Molecular biomarkers of anthropic impacts in natural archives: a review. Front Ecol Evol 4:4-16. https://doi.org/10.3389/fevo.2016.00092

Dubois N, Saulnier-Talbot É, Mills K, Gell P, Battarbee R, Bennion H, Chawchai S, Dong X, Francus P, Flower R, Gomes DF, Gregory-Eaves I, Humane S, Kattel G, Jenny J, Langdon P, Massaferro J, McGowan S, Mikomägi A, Ngoc NTM, Ratnayake AS, Reid M, Rose N, Saros J, Schillereff D, Tolotti M, Valero-Garcés B (2018) First human impacts and responses of aquatic systems: A review of palaeolimnological records from around the world. Anthr Rev 5:28–68. https://doi.org/10.1177/2053019617740365

Dussault F, Bain A, LeMoine G (2014) Early Thule winter houses: an archaeoentomological analysis. Arct Anthropol 51:101–117. https://doi.org/10.3368/aa.51.1.101

Dussault F, Bell TJ, Grimes V (2017) Archaeoentomological perspectives on Dorset occupations in Newfoundland: a case study from the site of Phillip's Garden (EeBi-1). Arctic 69:1. https://doi.org/10.14430/arctic4660

Eglinton G, Hamilton RJ (1967) Leaf epicuticular waxes. Science 156:1322–1335

Elbaz-Poulichet F, Guédron S, Anne-Lise D, Freydier R, Perrot V, Rossi M, Piot C, Delpoux S, Sabatier P (2020) A 10,000-year record of trace metal and metalloid (Cu, Hg, Sb, Pb) deposition in a western Alpine lake (Lake Robert, France): Deciphering local and regional mining contamination. Quat Sci Rev 228:106076. https://doi.org/10.1016/j.quascirev.2019.106076

Elhmmali MM, Roberts DJ, Evershed RP (1997) Bile acids as a new class of sewage pollution indicator. Environ Sci Technol 31:3663–3668. https://doi.org/10.1021/es9704040

Elias SA (2010) The use of insect fossils in archeology. Dev Quarternary Sci 12:89–121

Elias VO, Simoneit BRT, Cordeiro RC, Turcq B (2001) Evaluating levoglucosan as an indicator of biomass burning in Carajas, Amazonia: A comparison to the charcoal record. Geochim Cosmochim Acta 65:267–272. https://doi.org/10.1016/s0016-703700)00522-6

Emerson S, Hedges J (2003) Sediment diagenesis and benthic flux. Treatise Geochem 6:625

Engling G, Carrico CM, Kreidenweis SM, Collett Jr. JL, Day DE, Malm WC, Lincoln E, Min Hao W, Iinuma Y, Herrmann H (2006) Determination of levoglucosan in biomass combustion aerosol by high-performance anion-exchange chromatography with pulsed amperometric detection. Atmos Environ 40:299–311. https://doi.org/10.1016/j.atmosenv.2005.12.069

Florian C, Miller G, Fogel M, Wolfe A, Vinebrooke R, Geirsdóttir Á (2015) Algal pigments in Arctic lake sediments record biogeochemical changes due to Holocene climate variability and anthropogenic global change. J Paleolimnol 54:53–69. https://doi.org/10.1007/s10933-015-9835-5

Fontana L, Sun M, Huang X, Xiang L (2019) The impact of climate change and human activity on the ecological status of Bosten Lake, NW China, revealed by a diatom record for the last 2000 years. Holocene 29:1871–1884. https://doi.org/10.1177/0959683619865586

Fredh ED, Lagerås P, Mazier F, Björkman L, Lindbladh M, Broström A (2019) Farm establishment, abandonment and agricultural practices during the last 1,300 years: a case study from southern Sweden based on pollen records and the LOVE model. Veg Hist Archaeobotany 28:529–544. https://doi.org/10.1007/s00334-019-00712-x

Fry B (1988) Food web structure on Georges Bank from stable C, N, and S isotopic compositions. Limnol Oceanogr 33:1182–1190. https://doi.org/10.4319/lo.1988.33.5.1182

Gajewski K, Garneau M, Bourgeois JC (1995) Paleoenvironments of the Canadian High Arctic derived from pollen and plant macrofossils: problems and potentials. Quat Sci Rev 14:609–629

Garcin Y, Deschamps P, Ménot G, de Saulieu G, Schefuß E, Sebag D, Dupont LM, Oslisly R, Brademann B, Mbusnum KG, Onana J-M, Ako AA, Epp LS, Tjallingii R, Strecker MR, Brauer A, Sachse D (2018) Early anthropogenic impact on western central African rainforests 2,600 y ago. Proc Natl Acad Sci USA 115:3261–3266. https://doi.org/10.1073/pnas.1715336115

Gauthier E, Bichet V, Massa C, Petit C, Vannière B, Richard H (2010) Pollen and non-pollen palynomorph evidence of medieval farming activities in southwestern Greenland. Veg Hist Archaeobotany 19:427–438. https://doi.org/10.1007/s00334-010-0251-5

Gauthier E, Richard H (2009) Bronze Age at Lake Bourget (NW Alps, France): Vegetation, human impact and climatic change. Quat Int 200:111–119. https://doi.org/10.1016/j.quaint.2008.10.004

Giguet-Covex C, Pansu J, Arnaud F, Rey PJ, Griggo C, Gielly L, Domaizon I, Coissac E, David F, Choler P, Poulenard J, Taberlet P (2014) Long livestock farming history and human landscape shaping revealed by lake sediment DNA. Nat Commun 5:7. https://doi.org/10.1038/ncomms4211

Giri SJ, Diefendorf AF, Lowell TV (2015) Origin and sedimentary fate of plant-derived terpenoids in a small river catchment and implications for terpenoids as quantitative paleovegetation proxies. Org Geochem 82:22–32. https://doi.org/10.1016/j.orggeochem.2015.02.002

Glaser B, Haumaier L, Guggenberger G, Zech W (1998) Black carbon in soils: the use of benzenecarboxylic acids as specific markers. Org Geochem 29:811–819

Glikson A (2013) Fire and human evolution: The deep-time blueprints of the Anthropocene. Anthropocene 3:89–92. https://doi.org/10.1016/j.ancene.2014.02.002

Goldewijk KK, Beusen A, Doelman J, Stehfest E (2017) Anthropogenic land use estimates for the Holocene – HYDE 3.2. Earth Syst Sci Data 9:927–953. https://doi.org/10.5194/essd-9-927-2017

Goodfellow RM, Cardoso J, Eglinton G, Dawson JP, Best GA (1977) A faecal sterol survey in the Clyde Estuary. Mar Pollut Bull 8:272–276. https://doi.org/10.1016/0025-326X77)90168-0

Greenland National Museum & Archive (2015) Nunniffiit (v1.03). http://nunniffiit.natmus.gl/cbkort. Accessed 30 Oct 2019

Grimalt JO, Fernandez P, Bayona JM, Albaiges J (1990) Assessment of fecal sterols and ketones as indicators of urban sewage inputs to coastal waters. Environ Sci Technol 24:357–363. https://doi.org/10.1021/es00073a011

Grossman EL, Ku TL (1986) Oxygen and carbon isotope fractionation in biogenic aragonite: Temperature effects. Chem Geol Isot Geosci Sect 59:59–74. https://doi.org/10.1016/0168-962286)90057-6

Guillemot T, Bichet V, Gauthier E, Zocatelli R, Massa C, Richard H (2017) Environmental responses of past and recent agropastoral activities on south Greenlandic ecosystems through molecular biomarkers. Holocene 27:783–795. https://doi.org/10.1177/0959683616675811

Guillemot T, Zocatelli R, Bichet V, Jacob J, Massa C, Le Milbeau C, Richard H, Gauthier E (2015) Evolution of pastoralism in Southern Greenland during the last two millennia reconstructed from bile acids and coprophilous fungal spores in lacustrine sediments. Org Geochem 81:40–44

Haas M, Baumann F, Castella D, Haghipour N, Reusch A, Strasser M, Eglinton TI, Dubois N (2019) Roman-driven cultural eutrophication of Lake Murten, Switzerland. Earth Planet Sci Lett 505:110–117. https://doi.org/10.1016/j.epsl.2018.10.027

Haas M, Kaltenrieder P, Ladd SN, Welte C, Strasser M, Eglinton TI, Dubois N (2020) Land-use evolution in the catchment of Lake Murten, Switzerland. Quat Sci Rev 230:106154. https://doi.org/10.1016/j.quascirev.2019.106154

Hadley KR, Douglas MS, Blais JM, Smol JP (2010a) Nutrient enrichment in the High Arctic associated with Thule Inuit whalers: a paleolimnological investigation from Ellesmere Island (Nunavut, Canada). Hydrobiologia 649:129–138

Hadley KR, Douglas MS, McGhee R, Blais JM, Smol JP (2010b) Ecological influences of Thule Inuit whalers on high Arctic pond ecosystems: a comparative paleolimnological study from Bathurst Island (Nunavut, Canada). J Paleolimnol 44:85–93

Haile J, Froese DG, MacPhee RDE, Roberts RG, Arnold LJ, Reyes AV, Rasmussen M, Nielsen R, Brook BW, Robinson S, Demuro M, Gilbert MTP, Munch K, Austin JJ, Cooper A, Barnes I, Möller P, Willerslev E (2009) Ancient DNA reveals late survival of mammoth and horse in interior Alaska. Proc Natl Acad Sci USA 106:22352–22357. https://doi.org/10.1073/pnas.0912510106

Hald MM, Mosekilde J, Magnussen B, Søe MJ, Hansen CH, Mortensen MF (2018) Tales from the barrels: Results from a multi-proxy analysis of a latrine from Renaissance Copenhagen, Denmark. J Archaeol Sci Rep 20:602–610. https://doi.org/10.1016/j.jasrep.2018.06.006

Hargan KE, Stewart EM, Michelutti N, Grooms C, Kimpe LE, Mallory ML, Smol JP, Blais JM (2018) Sterols and stanols as novel tracers of waterbird population dynamics in freshwater ponds. Proc R Soc B 285:20180631. https://doi.org/10.1098/rspb.2018.0631

Hirons AC, Schell DM, Finney BP (2001) Temporal records of $\delta^{13}C$ and $\delta^{15}N$ in North Pacific pinnipeds: inferences regarding environmental change and diet. Oecologia 129:591–601. https://doi.org/10.1007/s004420100756

Hobson KA, Schell DM (1998) Stable carbon and nitrogen isotope patterns in baleen from eastern Arctic bowhead whales. Can J Fish Aquat Sci 55: 2601-2607

Howard S, McInerney FA, Caddy-Retalic S, Hall PA, Andrae JW (2018) Modelling leaf wax n-alkane inputs to soils along a latitudinal transect across Australia. Org Geochem 121:126–137. https://doi.org/10.1016/j.orggeochem.2018.03.013

Huang X, Zhang J, Storozum M, Liu S, Gill JL, Xiang L, Ren X, Wang J, Qiang M, Chen F, Grimm EC (2020) Long-term herbivore population dynamics in the northeastern Qinghai-Tibetan Plateau and its implications for early human impacts. Rev Palaeobot Palynol 275:104171. https://doi.org/10.1016/j.revpalbo.2020.104171

Jacob J, Disnar J-R, Arnaud F, Chapron E, Debret M, Lallier-Vergès E, Desmet M, Revel-Rolland M (2008a) Millet cultivation history in the French Alps as evidenced by a sedimentary molecule. J Archaeol Sci 35:814–820. https://doi.org/10.1016/j.jas.2007.06.006

Jacob J, Disnar J-R, Arnaud F, Gauthier E, Billaud Y, Chapron E, Bardoux G (2009) Impacts of new agricultural practices on soil erosion during the Bronze Age in the French Prealps. Holocene 19:241–249 . https://doi.org/10.1177/0959683608100568

Jacob J, Disnar J-R, Bardoux G (2008b) Carbon isotope evidence for sedimentary miliacin as a tracer of Panicum miliaceum (broomcorn millet) in the sediments of Lake le Bourget (French Alps). Org Geochem 39:1077–1080. https://doi.org/10.1016/j.orggeochem.2008.04.003

Jacob J, Disnar J-R, Boussafir M, Spadano Albuquerque AL, Sifeddine A, Turcq B (2005) Pentacyclic triterpene methyl ethers in recent lacustrine sediments (Lagoa do Caçó, Brazil). Org Geochem 36:449–461. https://doi.org/10.1016/j.orggeochem.2004.09.005

Jaffe R, Ding Y, Niggemann J, Vahatalo AV, Stubbins A, Spencer RGM, Campbell J, Dittmar T (2013) Global charcoal mobilization from soils via dissolution and riverine transport to the oceans. Science 340:345–347. https://doi.org/10.1126/science.1231476

Jouffroy-Bapicot I, Vannière B, Gauthier É, Richard H, Monna F, Petit C (2013) 7000 years of vegetation history and land-use changes in the Morvan Mountains (France): a regional synthesis. Holocene 23:1888–1902. https://doi.org/10.1177/0959683613508161

Juggins S (2013) Quantitative reconstructions in palaeolimnology: new paradigm or sick science? Quat Sci Rev 64:20–32. https://doi.org/10.1016/j.quascirev.2012.12.014

Kaufman DS, Ager TA, Anderson NJ, Anderson PM, Andrews JT, Bartlein PJ, Brubaker LB, Coats LL, Cwynar LC, Duvall ML, Dyke AS, Edwards ME, Eisner MR, Gajewski K, Geirsdóttiro O, Hu FS, Jennings AE, Kaplan MR, Kerwin MW, Lozhkin LV, MacDonald GM, Miller GH, Mock CJ, Oswald WW, Otto-Bliesner BL, Porinchu DF, Rühland K, Smol JP, Steig EJ, Wolfe BB (2004) Holocene thermal maximum in the western Arctic (0–180 W). Quat Sci Rev 23:529–560

Keely BJ (2006) Geochemistry of chlorophylls. In: Grimm B, Porra RJ, Rüdiger W, Scheer H (eds) Chlorophylls and Bacteriochlorophylls: Biochemistry, Biophysics, Functions and Applications. Springer, Dordrecht, pp 535–561

Kenward HK (1976) Reconstructing ancient ecological conditions from insect remains; some problems and an experimental approach. Ecol Entomol 1:7–17. https://doi.org/10.1111/j.1365-2311.1976.tb01200.x

Key FM, Posth C, Krause J, Herbig A, Bos KI (2017) Mining metagenomic data sets for ancient DNA: recommended protocols for authentication. Trends Genet 33:508–520. https://doi.org/10.1016/j.tig.2017.05.005

Kinder M, Tylmann W, Bubak I, Fiłoc M, Gąsiorowski M, Kupryjanowicz M, Mayr C, Sauer L, Voellering U, Zolitschka B (2019) Holocene history of human impacts inferred from annually laminated sediments in Lake Szurpiły, northeast Poland. J Paleolimnol 61:419–435. https://doi.org/10.1007/s10933-019-00068-2

Knights BA, Dickson CA, Dickson JH, Breeze DJ (1983) Evidence concerning the Roman military diet at Bearsden, Scotland, in the 2nd century AD. J Archaeol Sci 10:139–152

Korosi JB, Cheng W, Blais JM (2015) Organic pollutants in sediment core archives. In: : Blais JM, Rosen MR, Smol JP (eds) Environmental Contaminants. Springer, Dordrecht, pp 161–185

Korosi JB, Thienpont JR, Pisaric MF (2017a) Broad-scale lake expansion and flooding inundates essential wood bison habitat. Nat Commun 8:1-8

Korosi JB, Thienpont JR, Smol JP, Blais JM (2017b) Paleo-ecotoxicology: what can lake sediments tell us about ecosystem responses to environmental pollutants? Environ Sci Technol 51:9446–9457

Krause S, Beach T, Luzzadder-Beach S, Guderjan TH, Valdez F, Eshleman S, Doyle C, Bozarth SR (2019) Ancient Maya wetland management in two watersheds in Belize: soils, water, and paleoenvironmental change. Quat Int 502:280–295. https://doi.org/10.1016/j.quaint.2018.10.029

Kuhajda K, Kandrac J, Kevresan S, Mikov M, Fawcett JP (2006) Structure and origin of bile acids: an overview. Eur J Drug Metab Pharmacokinet 31:135–143

Kuo LJ, Louchouarn P, Herbert BE (2011) Influence of combustion conditions on yields of solvent-extractable anhydrosugars and lignin phenols in chars: implications for characterizations of biomass combustion residues. Chemosphere 85:797–805. https://doi.org/10.1016/j.chemosphere.2011.06.074

Kuzyk ZZA, Macdonald RW, Johannessen SC (2015) Calculating rates and dates and interpreting contaminant profiles in biomixed sediments. In: Blais JM, Rosen MR, Smol JP (eds) Environmental Contaminants. Springer, Dordrecht, pp 61–87

Lamb HF (1985) Palynological evidence for postglacial change in the position of tree limit in Labrador. Ecol Monogr 55:241–258. https://doi.org/10.2307/1942559
Lan T, Lindqvist C (2019) Technical advances and challenges in genome-scale analysis of ancient DNA. In: Lindqvist C, Rajora OP (eds) Paleogenomics: Genome-Scale Analysis of Ancient DNA. Springer International Publishing, Cham, pp 3–29
Last WM, Smol JP (2001a) Tracking Environmental Change Using Lake Sediments: Volume 1: Basin Analysis, Coring, and Chronological Techniques. Springer, Dordrecht
Last WM, Smol JP (2001b) Tracking Environmental Change Using Lake Sediments: Volume 2: Physical and Geochemical Methods. Springer, Dordrecht
Lavrieux M, Jacob J, Disnar JR, Bréheret JG, Le Milbeau C, Miras Y, Andrieu-Ponel V (2013) Sedimentary cannabinol tracks the history of hemp retting. Geology 41:751–754. https://doi.org/10.1130/G34073.1
Leavitt PR, Hodgson DA (2001) Sedimentary pigments. In: Last WM, Smol JP (eds) Tracking Environmental Change Using Lake Sediments. Volume 3: Terrestrial, Algal, and Siliceous Indicators. Kluwer Academic Publishers, Dordrect, pp 295-326
Ledger PM (2018) Are circumpolar hunter-gatherers visible in the palaeoenvironmental record? Pollen-analytical evidence from Nunalleq, southwestern Alaska. Holocene 28:415–426. https://doi.org/10.1177/0959683617729447
Ledger PM, Girdland-Flink L, Forbes V (2019) New horizons at L'Anse aux Meadows. Proc Natl Acad Sci USA 116:15341–15343. https://doi.org/10.1073/pnas.1907986116
Lee CSL, Qi S, Zhang G, Luo C, Zhao LYL, Li X (2008) Seven thousand years of records on the mining and utilization of metals from lake sediments in central China. Environ Sci Technol 42:4732–4738. https://doi.org/10.1021/es702990n
Lehndorff E, Linstädter J, Kehl M, Weniger GC (2015) Fire history reconstruction from black carbon analysis in Holocene cave sediments at Ifri Oudadane, northeastern Morocco. Holocene 25:398–402 . https://doi.org/10.1177/0959683614558651
Lima ALC, Farrington JW, Reddy CM (2005) Combustion-derived polycyclic aromatic hydrocarbons in the environment—a review. Environ Forensics 6:109–131. https://doi.org/10.1080/15275920590952739
Llamas B, Valverde G, Fehren-Schmitz L, Weyrich LS, Cooper A, Haak W (2017) From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. STAR Sci Technol Archaeol Res 3:1–14. https://doi.org/10.1080/20548923.2016.1258824
Luo L, Wang X, Guo H, Lasaponara R, Zong X, Masini N, Wang G, Shi P, Khatteli H, Chen F, Tariq S, Shao J, Bachagha N, Yang R, Yao Y (2019) Airborne and spaceborne remote sensing for archaeological and cultural heritage applications: A review of the century 1907–2017). Remote Sens Environ 232:111280. https://doi.org/10.1016/j.rse.2019.111280
Machado KS, Froehner S, Rizzi J, Torres M (2018) Tracking capybara (*Hydrochoerus hydrochaeris*) feces contribution method in aquatic environments using sterols: capybara feces tracking in the aquatic environment. Environ Toxicol Chem 37:353–361. https://doi.org/10.1002/etc.4004
Mackenzie AS, Brassell SC, Eglinton G, Maxwell JR (1982) Chemical fossils: the geological fate of steroids. Science 217:491–504

Martindale A, Morlan R, Betts M, Blake M, Gajewski K, Chaput MA, Mason A, Vermeersch P (2016) Canadian Archaeological Radiocarbon Database (CARD 2.1). https://www.canadianarchaeology.ca. Accessed 30 Oct 2019

Masi A, Francke A, Pepe C, Thienemann M, Wagner B, Sadori L (2018) Vegetation history and paleoclimate at Lake Dojran (FYROM/Greece) during the Late Glacial and Holocene. Clim Past 14:351–367. https://doi.org/10.5194/cp-14-351-2018

Masiello CA, Louchouarn P (2013) Fire in the ocean. Science 340:287–288. https://doi.org/10.1126/science.1237688

Massa C, Bichet V, Gauthier É, Perren BB, Mathieu O, Petit C, Monna F, Giraudeau J, Losno R, Richard H (2012) A 2500 year record of natural and anthropogenic soil erosion in South Greenland. Quat Sci Rev 32:119–130. https://doi.org/10.1016/j.quascirev.2011.11.014

Messager ML, Lehner B, Grill G, Nedeva I, Schmitt O (2016) Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nat Commun 7: 1-11. https://doi.org/10.1038/ncomms13603

Meyers PA, Ishiwatari R (1993) Lacustrine organic geochemistry - an overview of indicators of organic matter sources and diagenesis in lake sediments. Org Geochem 20:867–900. https://doi.org/10.1016/0146-638093)90100-p

Meyers PA, Leenheer MJ, Eadie BJ, Maule SJ (1984) Organic geochemistry of suspended and settling particulate matter in Lake Michigan. Geochim Cosmochim Acta 48:443–452. https://doi.org/10.1016/0016-703784)90273-4

Meyers PA, Teranes JL (2001) Sediment organic matter. In: Last WM, Smol JP (eds) Tracking Environmental Change Using Lake Sediments: Physical and Geochemical Methods. Springer, Dordrecht, pp 239–269

Michelutti N, McCleary KM, Antoniades D, Sutherland P, Blais JM, Douglas MS, Smol JP (2013) Using paleolimnology to track the impacts of early Arctic peoples on freshwater ecosystems from southern Baffin Island, Nunavut. Quat Sci Rev 76:82–95

Michelutti N, Smol J (2016) Visible spectroscopy reliably tracks trends in paleo-production. J Paleolimnol 56:253–265. https://doi.org/10.1007/s10933-016-9921-3

Michelutti N, Wolfe AP, Briner JP, Miller GH (2007) Climatically controlled chemical and biological development in Arctic lakes. J Geophys Res Biogeosciences 112:1-10. https://doi.org/10.1029/2006JG000396

Mills K, Schillereff D, Saulnier-Talbot É, Gell P, Anderson NJ, Arnaud F, Dong X, Jones M, McGowan S, Massaferro J (2017) Deciphering long-term records of natural variability and human impact as recorded in lake sediments: a palaeolimnological puzzle. Wiley Interdiscip Rev Water 4:1–29

Morin-Rivat J, Biwolé A, Gorel AP, Vleminckx J, Gillet JF, Bourland N, Hardy OJ, Smith AL, Daïnou K, Dedry L, Beeckman H, Doucet JL (2016) High spatial resolution of late-Holocene human activities in the moist forests of central Africa using soil charcoal and charred botanical remains. Holocene 26:1954–1967. https://doi.org/10.1177/0959683616646184

Motuzaite-Matuzeviciute G, Jacob J, Telizhenko S, Jones MK (2016) Miliacin in palaeosols from an Early Iron Age in Ukraine reveal in situ cultivation of broomcorn millet. Archaeol Anthropol Sci 8:43–50. https://doi.org/10.1007/s12520-013-0142-7

Mouritsen OG, Zuckermann MJ (2004) What's so special about cholesterol? Lipids 39:1101–1113. https://doi.org/10.1007/s11745-004-1336-x

Naito S, Katsuta N, Kawakami S, Koido Y, Shimono H (2019) Late Holocene climatic impact on vegetation and human activity in central Japan, recorded in sediment at Arao-Minami archaeological site, northwestern Nobi Plain. Quat Int 519:144–155. https://doi.org/10.1016/j.quaint.2019.04.019
Nicosia C, Ertani A, Vianello A, Nardi S, Brogiolo GP, Arnau AC, Becherini F (2018) Heart of darkness: an interdisciplinary investigation of the urban anthropic deposits of the Baptistery of Padua (Italy). Archaeol Anthropol Sci 11: 1977–1993. https://doi.org/10.1007/s12520-018-0646-2
Niemeyer B, Epp LS, Stoof-Leichsenring KR, Pestryakova LA, Herzschuh U (2017) A comparison of sedimentary DNA and pollen from lake sediments in recording vegetation composition at the Siberian treeline. Mol Ecol Resour 17:e46–e62. https://doi.org/10.1111/1755-0998.12689
Novenko EY, Zyuganova IS, Volkova EM, Dyuzhova KV (2019) A 7000-year pollen and plant macrofossil record from the Mid-Russian Upland, European Russia: vegetation history and human impact. Quat Int 504:70–79. https://doi.org/10.1016/j.quaint.2017.11.025
Oros DR, Simoneit BRT (2001a) Identification and emission factors of molecular tracers in organic aerosols from biomass burning Part 2. Deciduous trees. Appl Geochem 16: 1545-1565
Oros DR, Simoneit BRT (2001b) Identification and emission factors of molecular tracers in organic aerosols from biomass burning Part 1. Temperate climate conifers. Appl Geochem 16: 1513-1544
Panagiotakopulu E, Buchan AL (2015) Present and Norse Greenlandic hayfields – Insect assemblages and human impact in southern Greenland. Holocene 25:921–931. https://doi.org/10.1177/0959683615574585
Panagiotakopulu E, Schofield JE, Vickers K, Edwards KJ, Buckland PC (2018) Thule Inuit environmental impacts on Kangeq, southwest Greenland. Quat Int 549: 176-190. https://doi.org/10.1016/j.quaint.2018.09.011
Panagiotakopulu E, Skidmore P, Buckland P (2007) Fossil insect evidence for the end of the Western Settlement in Norse Greenland. Naturwissenschaften 94:300–306. https://doi.org/10.1007/s00114-006-0199-6
Penny D, Hall T, Evans D, Polkinghorne M (2019) Geoarchaeological evidence from Angkor, Cambodia, reveals a gradual decline rather than a catastrophic 15th-century collapse. Proc Natl Acad Sci USA 116:4871–4876 . https://doi.org/10.1073/pnas.1821460116
Pienitz R, Smol JP, Last WM, Leavitt PR, Cumming BF (2000) Multi-proxy Holocene palaeoclimatic record from a saline lake in the Canadian Subarctic. Holocene 10:673–686
Pietramellara G, Ascher J, Borgogni F, Ceccherini MT, Guerri G, Nannipieri P (2009) Extracellular DNA in soil and sediment: fate and ecological relevance. Biol Fertil Soils 45:219–235. https://doi.org/10.1007/s00374-008-0345-8
Pini R, Ravazzi C, Raiteri L, Guerreschi A, Castellano L, Comolli R (2017) From pristine forests to high-altitude pastures: an ecological approach to prehistoric human impact on vegetation and landscapes in the western Italian Alps. J Ecol 105:1580–1597. https://doi.org/10.1111/1365-2745.12767
Pompeani DP, Abbott MB, Steinman BA, Bain DJ (2013) Lake sediments record prehistoric lead pollution related to early copper production in North America. Environ Sci Technol 47:5545–5552. https://doi.org/10.1021/es304499c
Prost K, Birk JJ, Lehndorff E, Gerlach R, Amelung W (2017) Steroid biomarkers revisited – improved source identification of faecal remains in archaeological soil material. PLOS ONE 1–30
Razjigaeva NG, Ganzey LA, Lyaschevskaya MS, Makarova TR, Kudryavtseva EP, Grebennikova TA, Panichev AM, Arslanov KhA, Maksimov FE, Petrov AY, Malkov SS (2019) Climatic and human

impacts on landscape development of the Murav'ev Amursky Peninsula (Russian South Far East) in the middle/late Holocene and historical time. Quat Int 516:127–140. https://doi.org/10.1016/j.quaint.2017.12.007
Rees-Owen RL, Gill FL, Newton RJ, Ivanovic RF, Francis JE, Riding JB, Vane CH, Lopes dos Santos RA (2018) The last forests on Antarctica: reconstructing flora and temperature from the Neogene Sirius Group, Transantarctic Mountains. Org Geochem 118:4–14
Ritchie J (1987) Postglacial Vegetation of Canada. Cambridge University Press, New York
Ritchie JC (1995) Tansley Review 83. Current trends in studies of long-term plant community dynamics. New Phytol 130:469–494. https://doi.org/10.1111/j.1469-8137.1995.tb04325.x
Romero-Sarmiento MF, Riboulleau A, Vecoli M, Laggoun-Défarge F, Versteegh GJM (2011) Aliphatic and aromatic biomarkers from Carboniferous coal deposits at Dunbar (East Lothian, Scotland): Palaeobotanical and palaeoenvironmental significance. Palaeogeogr Palaeoclimatol Palaeoecol 309:309–326. https://doi.org/10.1016/j.palaeo.2011.06.015
Roy N, Woollett J, Bhiry N (2015) Paleoecological perspectives on landscape history and anthropogenic impacts at Uivak Point, Labrador, since AD 1400. Holocene 25:1742–1755. https://doi.org/10.1177/0959683615591350
Rühland KM, Paterson AM, Keller W, Michelutti N, Smol JP (2013) Global warming triggers the loss of a key Arctic refugium. Proc R Soc B 280:20131887. https://doi.org/10.1098/rspb.2013.1887
Russo EB (2007) History of Cannabis and its preparations in saga, science, and sobriquet. Chem Biodivers 4:1614–1648. https://doi.org/10.1002/cbdv.200790144
Sachse D, Billault I, Bowen GJ, Chikaraishi Y, Dawson TE, Feakins SJ, Freeman KH, Magill CR, McInerney FA, van der Meer MTJ, Polissar P, Robins RJ, Sachs JP, Schmidt H-L, Sessions AL, White JWC, West JB, Kahmen A (2012) Molecular paleohydrology: interpreting the hydrogen-isotopic composition of lipid biomarkers from photosynthesizing organisms. Annu Rev Earth Planet Sci 40:221–249. https://doi.org/10.1146/annurev-earth-042711-105535
Sadori L, Giardini M (2007) Charcoal analysis, a method to study vegetation and climate of the Holocene: The case of Lago di Pergusa (Sicily, Italy). Geobios 40:173–180. https://doi.org/10.1016/j.geobios.2006.04.002
Schneider T, Rimer D, Butz C, Grosjean M (2018) A high-resolution pigment and productivity record from the varved Ponte Tresa basin (Lake Lugano, Switzerland) since 1919: insight from an approach that combines hyperspectral imaging and high-performance liquid chromatography. J Paleolimnol 60:381–398. https://doi.org/10.1007/s10933-018-0028-x
Schroeter N, Lauterbach S, Stebich M, Kalanke J, Mingram J, Yildiz C, Schouten S, Gleixner G (2020) Biomolecular evidence of early human occupation of a high-Altitude Site in Western Central Asia During the Holocene. Front Earth Sci 8:1-13. https://doi.org/10.3389/feart.2020.00020
Schüpbach S, Kirchgeorg T, Colombaroli D, Beffa G, Radaelli M, Kehrwald NM, Barbante C (2015) Combining charcoal sediment and molecular markers to infer a Holocene fire history in the Maya Lowlands of Petén, Guatemala. Quat Sci Rev 115:123–131. https://doi.org/10.1016/j.quascirev.2015.03.004
Schwarzbauer J, Stock F, Brückner H, Dsikowitzky L, Krichel M (2017) Molecular organic indicators for human activities in the Roman harbor of Ephesus, Turkey. Geoarchaeology 33:498–509. https://doi.org/10.1002/gea.21669
Seersholm FV, Pedersen MW, Soe MJ, Shokry H, Mak SST, Ruter A, Raghavan M, Fitzhugh W, Kjaer KH, Willerslev E, Meldgaard M, Kapel CMO, Hansen AJ (2016) DNA evidence of bowhead whale

exploitation by Greenlandic Paleo-Inuit 4,000 years ago. Nat Commun 7:9. https://doi.org/10.1038/ncomms13389
Selig U, Leipe T, Dorfler W (2007) Paleolimnological records of nutrient and metal profiles in prehistoric, historic and modern sediments of three lakes in north-eastern Germany. Water Air Soil Pollut 184:183–194. https://doi.org/10.1007/s11270-007-9407-z
Shah VG, Hugh Dunstan R, Geary PM, Coombes P, Roberts TK, Von Nagy-Felsobuki E (2007) Evaluating potential applications of faecal sterols in distinguishing sources of faecal contamination from mixed faecal samples. Water Res 41:3691–3700. https://doi.org/10.1016/j.watres.2007.04.006
Shotyk W, Weiss D, Appleby PG, Cheburkin AK, Frei R, Gloor M, Kramers JD, Reese S, Knaap WOVD (1998) History of atmospheric lead deposition since 12,370 ^{14}C yr BP from a peat bog, Jura Mountains, Switzerland. Science 281:1635–1640. https://doi.org/10.1126/science.281.5383.1635
Shumilovskikh LS, Hopper K, Djamali M, Ponel P, Demory F, Rostek F, Tachikawa K, Bittmann F, Golyeva A, Guibal F, Talon B, Wang LC, Nezamabadi M, Bard E, Lahijani H, Nokandeh J, Omrani Rekavandi H, de Beaulieu JL, Sauer E, Andrieu-Ponel V (2016) Landscape evolution and agro-sylvo-pastoral activities on the Gorgan Plain (NE Iran) in the last 6000 years. Holocene 26:1676–1691. https://doi.org/10.1177/0959683616646841
Siart C, Forbriger M, Bubenzer O (2018) Digital Geoarchaeology: Bridging the gap between archaeology, geosciences and computer sciences. In: Siart C, Forbriger M, Bubenzer O (eds) Digital Geoarchaeology: New Techniques for Interdisciplinary Human-Environmental Research. Springer International Publishing, Cham, pp 1–7
Simonneau A, Doyen E, Chapron E, Millet L, Vannière B, Di Giovanni C, Bossard N, Tachikawa K, Bard E, Albéric P, Desmet M, Roux G, Lajeunesse P, Berger JF, Arnaud F (2013) Holocene land-use evolution and associated soil erosion in the French Prealps inferred from Lake Paladru sediments and archaeological evidences. J Archaeol Sci 40:1636–1645. https://doi.org/10.1016/j.jas.2012.12.002
Simpson IA, van Bergen PF, Perret V, Elhmmali MM, Roberts DJ, Evershed RP (1999) Lipid biomarkers of manuring practice in relict anthropogenic soils. Holocene 9:223–229
Smol JP, Birks HJ, Last WM (2006) Tracking Environmental Change Using Lake Sediments: Volume 3: Terrestrial, Algal, and Siliceous Indicators. Springer, Dordrecht
Smol JP, Cumming BF (2000) Tracking long-term changes in climate using algal indicators in lake sediments. J Phycol 36:986–1011
Søe MJ, Nejsum P, Fredensborg BL, Kapel CMO (2015) DNA typing of ancient parasite eggs from environmental samples identifies human and animal worm infections in Viking-age settlement. J Parasitol 101:57. https://doi.org/10.1645/14-650.1
Søe MJ, Nejsum P, Seersholm FV, Fredensborg BL, Habraken R, Haase K, Hald MM, Simonsen R, Højlund F, Blanke L, Merkyte I, Willerslev E, Kapel CMO (2018) Ancient DNA from latrines in northern Europe and the Middle East (500 BC–1700 AD) reveals past parasites and diet. PLOS ONE 13:e0195481. https://doi.org/10.1371/journal.pone.0195481
Sonnenburg EP, Boyce JI, Reinhardt EG (2013) Multi-proxy lake sediment record of prehistoric (Paleoindian-Archaic) archaeological paleoenvironments at Rice Lake, Ontario, Canada. Quat Sci Rev 73:77–92. https://doi.org/10.1016/j.quascirev.2013.05.012
Spear RW (1993) The palynological record of the Late-Quarternary Arctic tree-line in northwest Canada. Rev Palaeobot Palynol 79:99–111. https://doi.org/10.1016/0034-666793)90040-2

Stanley J-D, Bernhardt CE (2010) Alexandria's eastern harbor, Egypt: pollen, microscopic charcoal, and the transition from natural to human-modified basin. J Coast Res 261:67–79. https://doi.org/10.2112/JCOASTRES-D-09-00089.1

Stanley J-D, Carlson RW, Van Beek G, Jorstad TF, Landau EA (2007) Alexandria, Egypt, before Alexander the Great: a multidisciplinary approach yields rich discoveries. GSA Today 17:4. https://doi.org/10.1130/GSAT01708A.1

Steinman BA, Abbott MB (2013) Isotopic and hydrologic responses of small, closed lakes to climate variability: hydroclimate reconstructions from lake sediment oxygen isotope records and mass balance models. Geochim Cosmochim Acta 105:342–359. https://doi.org/10.1016/j.gca.2012.11.027

Tan Z, Mao L, Han Y, Mo D, Gu H, Liu Z, Long Y, An Z (2018) Black carbon and charcoal records of fire and human land use over the past 1300 years at the Tongguan Kiln archaeological site, China. Palaeogeogr Palaeoclimatol Palaeoecol 504:162–169. https://doi.org/10.1016/j.palaeo.2018.05.022

Thienemann, M (2017). Reconstructing Holocene climatic and environmental change using molecular and isotopic proxies from lake sedimentary records. PhD thesis, Universität zu Köln https://kups.ub.uni-koeln.de/7836/

Tian Y, Li YL, Zhao FC (2017) Secondary metabolites from polar organisms. Mar Drugs 15:28. https://doi.org/10.3390/md15030028

Tinner W, Lotter AF, Ammann B, Conedera M, Hubschmid P, van Leeuwen JFN, Wehrli M (2003) Climatic change and contemporaneous land-use phases north and south of the Alps 2300 BC to 800 AD. Quat Sci Rev 22:1447–1460. https://doi.org/10.1016/S0277-379103)00083-0

Trendel JM, Schaeffer P, Adam P, Ertlen D, Schwartz D (2010) Molecular characterisation of soil surface horizons with different vegetation in the Vosges (Massif France). Org Geochem 41:1036–1039. https://doi.org/10.1016/j.orggeochem.2010.04.014

Tyagi P, Edwards DR, Coyne MS (2008) Use of sterol and bile acid biomarkers to identify domesticated animal sources of fecal pollution. Water Air Soil Pollut 187:263–274

USGS (2016) Proxies. In: Proxies. https://www2.usgs.gov/landresources/lcs/paleoclimate/proxies.asp. Accessed 24 Oct 2019

van den Bos V, Newnham R, Rees A, Woods L (2020) Density separation in pollen preparation: How low can you go? J Paleolimnol 63:225–234. https://doi.org/10.1007/s10933-020-00112-6

van Geel B, Buurman J, Brinkkemper O, Schelvis J, Aptroot A, van Reenen G, Hakbijl T (2003) Environmental reconstruction of a Roman Period settlement site in Uitgeest (The Netherlands), with special reference to coprophilous fungi. J Archaeol Sci 30:873–883. https://doi.org/10.1016/S0305-440302)00265-0

Veron AJ, Flaux C, Marriner N, Poirier A, Rigaud S, Morhange C, Empereur JY (2013) A 6000-year geochemical record of human activities from Alexandria (Egypt). Quat Sci Rev 81:138–147. https://doi.org/10.1016/j.quascirev.2013.09.029

Vignola C, Masi A, Balossi Restelli F, Frangipane M, Marzaioli F, Passariello I, Stellato L, Terrasi F, Sadori L (2017) $\delta^{13}C$ and $\delta^{15}N$ from ^{14}C-AMS dated cereal grains reveal agricultural practices during 4300–2000 BC at Arslantepe (Turkey). Rev Palaeobot Palynol 247:164–174. https://doi.org/10.1016/j.revpalbo.2017.09.001

White AJ, Stevens LR, Lorenzi V, Munoz SE, Schroeder S, Cao A, Bogdanovich T (2019) Fecal stanols show simultaneous flooding and seasonal precipitation change correlate with Cahokia's population decline. Proc Natl Acad Sci USA 116:5461–5466. https://doi.org/10.1073/pnas.1809400116

Whitlock C, Larsen C (2001) Charcoal as a fire proxy. In: Smol JP, Birks HJB, Last WM, Bradley RS, Alverson K (eds) Tracking Environmental Change Using Lake Sediments: Terrestrial, Algal, and Siliceous Indicators. Springer, Dordrecht, pp 75–97

Whitlock C, Millspaugh SH (1996) Testing the assumptions of fire-history studies: an examination of modern charcoal accumulation in Yellowstone National Park, USA. Holocene 6:7–15

WHO (1993) Biomarkers and risk assessment: concepts and principles. World Health Organization

Zennaro P, Kehrwald N, McConnell JR, Schüpbach S, Maselli OJ, Marlon J, Vallelonga P, Leuenberger D, Zangrando R, Spolaor A, Borrotti M, Barbaro E, Gambaro A, Barbante C (2014) Fire in ice: two millennia of boreal forest fire history from the Greenland NEEM ice core. Clim Past 10:1905–1924. https://doi.org/10.5194/cp-10-1905-2014

Zocatelli R, Lavrieux M, Guillemot T, Chassiot L, Le Milbeau C, Jacob J (2017) Fecal biomarker imprints as indicators of past human land uses: Source distinction and preservation potential in archaeological and natural archives. J Archaeol Sci 81:79–89

Zou S, Li R, Xie S, Zhu J, Wang X, Huang J (2010) Paleofire indicated by polycyclic aromatic hydrocarbons in soil of Jinluojia archaeological site, Hubei, China. J Earth Sci 21:247–256. https://doi.org/10.1007/s12583-010-0089-x

Zoumis T, Schmidt A, Grigorova L, Calmano W (2001) Contaminants in sediments: remobilisation and demobilisation. Sci Total Environ 266:195–202. https://doi.org/10.1016/S0048-969700)00740-3

APPENDIX 2: Linking 'omics' technologies to study soil carbon stabilization: A review

David P. Overy[1], Madison A. Bell[1], Jermaneh Habtewold[1], Bobbi Helgason[2], Ed G. Gregorich[1]

[1]Agriculture and Agri-food Canada, Ottawa Research and Development Centre, Ottawa, Ontario, K1A 0C6, Canada

[2]Department of Soil Science, University of Saskatchewan, Saskatoon, Saskatchewan, S7N 5A8, Canada

[*]Corresponding Author

David Overy (david.overy@canada.ca)

STATEMENT OF CONTRIBUTIONS

David P. Overy: conceptualization, writing – original draft, writing – review & editing, supervision, funding acquisition, and corresponding author. Madison A. Bell: writing – original draft, writing – review & editing, visualizations, formal analysis, data curation, and literature search. Jermaneh Habtewold: writing – original draft, writing – review & editing, literature search, and data curation. Bobbi Helgason: writing – review & editing, and data curation. Ed Gregorich: conceptualization, supervision, writing – original draft, writing – review & editing, and funding acquisition. This work is submitted to *Frontiers in Environmental Science*:

> Overy, D.P., Bell, M.A., Habtewold, J., Helgason, B., and Gregorich, E. Linking 'omics' technologies to study soil carbon stabilization: A review. Frontiers in Environmental Science. (Submitted).

ABSTRACT

Evidence-based decisions governing sustainable agricultural land management practices require a mechanistic understanding of SOM transformations and C-stabilization in soil. Large amounts of carbon from organic fertilizers, root exudates, and crop residues are input into agricultural soils. Microbes then catalyze soil biogeochemical processes including C extracellular transformation, mineralization and assimilation of resources that are later returned to the soil as metabolites and necromass. A systems biology approach for a holistic study of the transformation of C inputs into stable soil organic matter requires the use of soil 'omics' platforms (metagenomics, metatranscriptomics, metaproteomics, and metabolomics). Linking these various platforms to data derived from the metabolic pathways and gene expression will enhance our understanding of structure-function relationships involved in soil C cycling and stabilization. In this review we discuss the application, potential and suitability of different 'omics' approaches (independently and in combination) for elucidating the C transformation processes in soil. We also highlight potential methodological biases associated with these approaches from limitations of the methods to experimental design and soil sampling to data analysis and interpretation.

INTRODUCTION

Soil organic matter (SOM) underpins the health and productivity of soil. Representing the largest carbon (C) pool in terrestrial ecosystems, SOM plays a pivotal role in the global C cycle and climate regulation. Invested with the sun's energy by photosynthesis, the continual influx of C-rich plant litter and exudates to the soil drives the C cycle as nutrients are released during decay and eventually

absorbed by growing plants or released/sequestered to the environment (Amundson et al., 2015; Haberl et al., 2014; Pelletier et al., 2011). The ability of the soil microbiome to access this energy source depends on its nutritional context and by the degree to which soil physically protects it (within small pores or by chemical bonding to metals and mineral surfaces). Therefore, soil type, cropping sequence and land management and the nutritional complexity of inputs to soil affect the fate and turnover of C in soil. The turnover rate of this cycle affects many ecosystem processes and properties: soil biodiversity – by delivery of solar energy; plant growth – by providing soil nutrients; water quality – by release of soluble nitrogen and phosphorus; climate change – by exchange of CO_2 and other greenhouse gases; and soil resilience – by effects on SOM stocks (Handa et al., 2014). Many of these influences are critical, especially now in the face of global stresses such as growing food demands, changing climate, and loss of biodiversity (Lin, 2014). Improved knowledge of the processes that govern SOM stability and long-term C storage under different land use, management practices and climates will be important to identify practices that enhance and preserve soil and ecosystem health and productivity and the ongoing recycling and supply of nutrients.

The persistence and stability of soil organic C depends on its biotic and abiotic environment. In terrestrial ecosystems, SOM represents a continuum of organic compounds, from fresh inputs like plant litter to progressively decomposed compounds. Biological, physical and chemical transformation processes convert plant exudates and intact residues into derivative products that form complex and intimate associations with soil minerals (Lehmann and Kleber, 2015). These transformative processes of SOM are determined by interdependent factors that include: compound chemistry, spatial arrangement and interaction with mineral surfaces, temperature and moisture conditions, soil acidity and redox state, and the proximity, biomass and community composition of microbial degraders (Lehmann and Kleber,

2015; Schmidt et al., 2011). SOM plays a key role in forming a stable physical structure (aggregates and biopores) within the soil matrix. This structure promotes aeration and the infiltration and holding capacity of plant-available water as nutrients and energy are released during decomposition, to promote soil fertility (Liang et al., 2017; Janzen, 2015; Lehmann and Kleber, 2015).

In soil, the majority of plant residue inputs decay within the first year of deposition (Gregorich et al., 2017; Jenkinson and Rayner, 1977). This rapid decay is, for the most part, interdependent upon the molecular structure/composition of the plant residues. Under suitable conditions, adapted microbes can decompose materials that were previously presumed to be recalcitrant (such as polycondensed aromatics, alkanes and fire-derived carbon) more rapidly than expected (Schmidt et al., 2011). Inputs of energy-rich substrates, like those originating from the rhizosphere, strongly influence the microbial community composition and activity, and promote the decomposition of compounds that may otherwise be selectively avoided by microorganisms (Kuzyakov, 2010). Microbial exoenzymes are released into the soil matrix to actively degrade polymeric plant residues, leading to distinct patterns of C use and stabilization (Liang et al., 2017). Fungi preferentially degrade fresh plant litter inputs as they release a broad suite of exoenzymes associated with the breakdown of complex polymers (Poll et al., 2006). Decomposition of complex organic structures like lignins are fastest at early stages of decomposition and compound-specific isotopic analysis has been used to show that these structures undergo rapid degradation with comparable turnover rates to bulk organic matter, when in proximity to microbes and small organic molecules to “prime” microbial growth (Amelung et al., 2008; Grandy and Neff, 2008; Marschner et al., 2008).

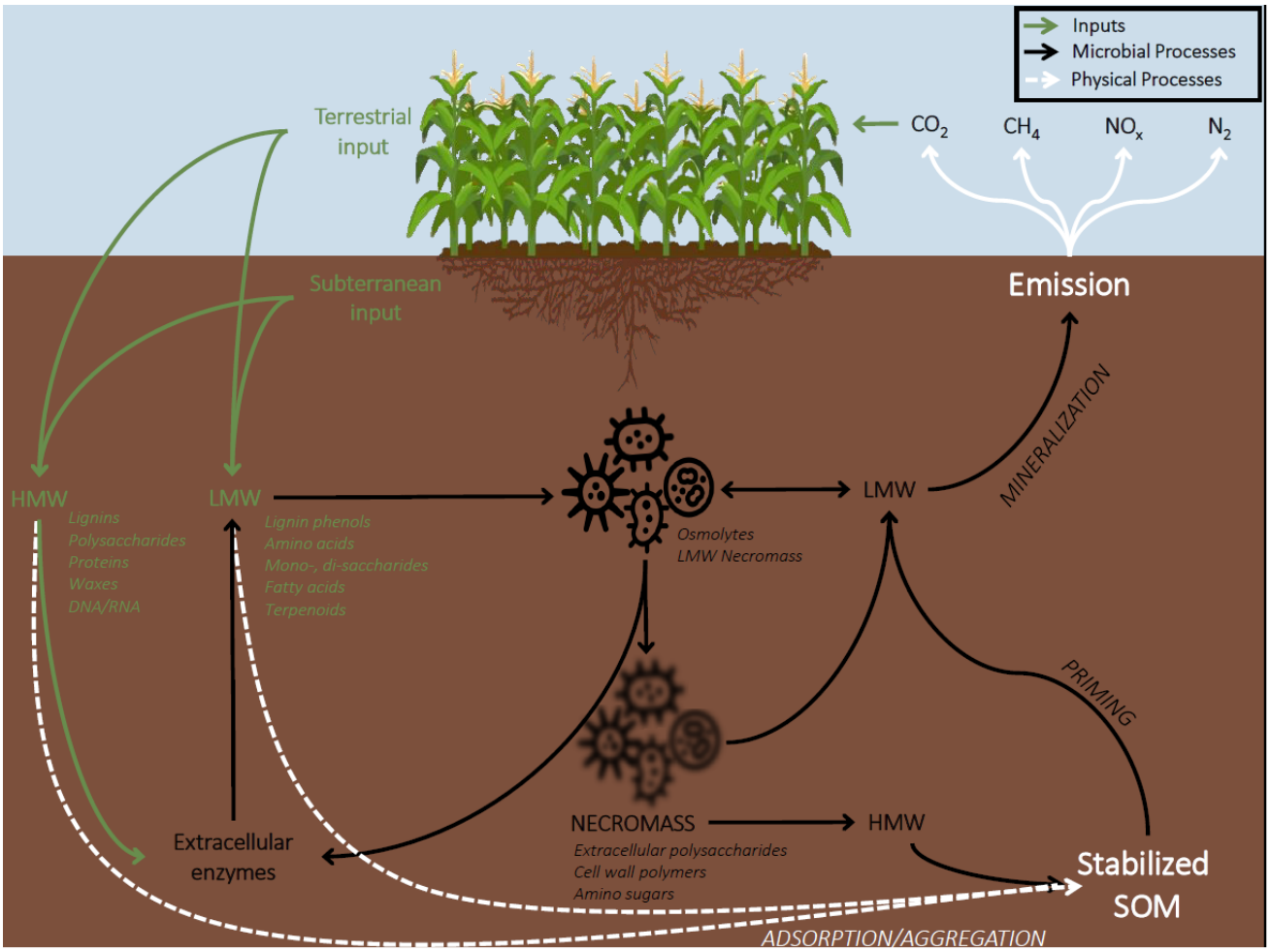


Figure A2.1: Nutrient cycling schematic demonstrating pathways for stabilization of soil organic matter (SOM). Plant inputs in the form of litter and exudates condition the physiology of the soil microbial community. Soil microbes directly assimilate some low molecular weight (LMW) molecules and produce extracellular enzymes to degrade high molecular weight (HMW) plant inputs, undergoing respiration (and mineralization of SOM) and biomass accumulation, ultimately resulting in a net increase of stabilized C in the form of microbial necromass.

Within soil, microbial degradation of plant residues promotes the release of C to the atmosphere through catabolism and at the same time, stabilizes C into forms of SOM that are not easily decomposed (Liang et al., 2017; Schimel and Schaeffer, 2012)(Figure A2.1). Fungal and bacterial necromass accumulation via biomass turnover is one of the main drivers of C-stabilization of the SOM reservoir (Kallenbach et al., 2016; Schweigert et al., 2015; Kindler et al., 2006). Through microbial anabolism,

accessible organic compounds are re-synthesized into molecules that are relatively more chemically stable (i.e. microbial cells, cell debris, biofilms); where fungal residues are thought to be more persistent in soils compared to bacterial residues (Liang et al., 2017; Six et al., 2006). In time, over the course of decomposition, the distinct chemistries of initial litter types (organic inputs) slowly converge following assimilation into microbial biomass and subsequent microbial turnover, into chemical compounds that serve as indicators of plant inputs (e.g., lignin phenols) and microbial metabolites and necromass (e.g., amino sugars) (Liang et al., 2017; Kallenbach et al., 2016; Wickings et al., 2011). Physiochemical stabilization also occurs when organic compounds interact with metals, charged soil minerals and organo-mineral complexes, adsorbing onto particle surfaces, or encapsulate into micro- or macro-aggregates through hydrogen bonding and Van der Wall interactions (Hernandez-Soriano et al., 2018; Sarkar et al., 2018; Liang et al., 2017; Lehmann and Kleber, 2015; Schmidt et al., 2011; Ekschmitt et al., 2008; Kleber et al., 2007). Thus, microbes transform organic matter, and the transformed organic matter can influence soil structure by controlling the composition of SOM, which, in turn facilitates formation of macro- and micro-aggregates – major components of soil architecture (Baldock and Broos, 2012; Six et al., 2004).

Evidence-based decisions governing sustainable agricultural land management practices require a mechanistic understanding of SOM transformations and C-stabilization in soil. Our current knowledge of microbial population diversity and dynamics of soil communities is due, in part, through the application of modern molecular biological tools. Significant gaps in our understanding of microbial mediated stabilization of soil C include comprehensive knowledge of the specific compounds involved, their turnover rates and the nature of their stabilization mechanisms (Liang et al., 2017). Application of a systems biology approach has been proposed to study the soil environment as whole, to link metabolic

pathways and gene expression into structure-function relationships regarding soil C cycling and stabilization (Schmidt et al., 2011). Our aim in this review is to discuss the suitability of different 'omics' approaches (independently and in combination) in soil C transformation studies. We also highlight potential methodological biases associated with these approaches, from limitations of the methods to experimental design and soil sampling to data analysis and interpretation.

SOIL METAGENOMICS

Typical culture-based approaches substantially underestimate the diversity of the soil microbial community, the majority of which remain uncharacterized (Fierer, 2017). Although most soil microbes cannot be cultured or grown in the lab, application of culture-independent molecular approaches such as metagenomics have uncovered an extraordinary diversity of microorganisms associated with the soil environment (Nesme et al., 2016). Metagenomics approaches provide insight into the genomic potential of an entire microbial community by cloning and analyzing community DNA directly from environmental samples (Prosser, 2015; Morales and Holben, 2011) (Figure A2.2). Targeted gene sequencing or shotgun sequencing of gene fragments are common approaches used in soil metagenomics studies (Nesme et al., 2016). Using phylogenetic barcoding genes, targeted soil metagenomics applications examine microbial population composition at a resolution of taxonomic order or higher (some studies attempt to annotate to the genus or species levels). But in this context, primer and PCR biases need to be considered because they can often skew population predictions (Sipos et al., 2007) and indeed important soil taxa (e.g. Verrucomicrobia) have been previously overlooked by commonly used 'universal' 16S rRNA gene primers (Bergmann et al., 2011). Targeted sequencing analyses also focus on particular gene sequences encoding for enzymes and other proteins involved with functional ecosystem

processes. Deep sequencing efforts can also be done at the full genome level and used to predict potential functional capability of a microbial population. However, reconstruction of 'population genomes' from biologically complex samples such as soil is difficult and so genome closure remains a challenge (Prosser, 2015; Sharon and Banfield, 2013).

Soil metagenomics has proven to be useful for characterizing the soil microbiome, but like all tools, there are inherent limitations and biases Several of these limitations and biases are common to all molecular techniques associated with cell lysis, efficient nucleic acid extraction and stability, and sequencing errors that affect reliable gene annotation and quantification of shotgun environmental gene sequences; moreover, the assignment of observed sequences to specific taxa is difficult and not always possible (Prosser, 2015). To address some of these limitations, data analysis pipelines (e.g. MEGAN, Mothur, Qiime, DADA2, Deblur) are available that can selectively remove or correct erroneous sequences (Amir et al., 2017; Callahan et al., 2016; Caporaso et al., 2010; Schloss et al., 2009; Huson et al., 2007). Also common to high-throughput sequencing approaches is the challenge that the data generated are compositional. That is, the total number of sequences generated is arbitrary and sequencing instrument, which imposes constraints for data analysis and interpretation (Gloor et al., 2017).

Other limitations involved in using metagenomics in soil research are related to working with a physically heterogeneous material. For example, efficient extraction of DNA requires complete dispersion of soil aggregates and detachment of cells from a broad range of soil types and textures. Hence, repeated extractions of a given soil sample maybe necessary to maximize cell yield (Williamson et al., 2011). The cells obtained through repeated extractions may capture unique assemblages of microbes that would otherwise have been missed in single-pass extractions. Despite homogenization of

soil samples prior to nucleic acid extraction, most DNA extractions are performed on a small sub-sample (less than one gram). Even with the use of technical replicates, it is possible that bias occurs as a result of the small sample mass.

Soil microbial communities are dynamic and active members of a population fluctuate, where at any one point in time, a significant proportion of the community may be dead or dormant (Blagodatskaya and Kuzyakov, 2013). As amplicon-based or shotgun metagenomics cannot differentiate between active, dead/dormant and relic DNA (Carini et al., 2016) (qualitatively, the presence of a functional gene is not evidence of its activity (Prosser, 2015)), a complete portrayal of active soil functions at the time of sampling is not possible (rather such techniques provide a prediction of potential soil function – past, present, and future). Caution is required when assuming links between microbial activity and relative gene abundance (Prosser, 2015). Accurate separation of functional variation from taxonomic variation within metagenomics data is also necessary, if deeper insight into microbial community assembly and mechanistic soil microbial ecology is to be achieved (Louca et al., 2016). However, despite these limitations, metagenomics approaches can be used to make hypotheses regarding how a microbial population may respond as environmental and soil management conditions change.

Metagenomics applications for soil carbon stabilization

Through sequence libraries generated by soil metagenomics, researchers have demonstrated how vast and under sampled soil microbiomes are globally. Bahram and coworkers (2018) sequenced approximately 160 million unique genes/gene fragments (with only 0.5% overlap with published genomes) from 189 topsoil samples collected from representative regions/biomes across the globe.

Bacterial taxonomic diversity, composition, richness and biomass correlate with soil pH, nutrient concentration and to a lesser extent climatic variables (Bahram et al., 2018). In comparison, similar characteristics in fungi were more strongly correlated with soil C/N ratios as a predictor for fungal biomass and relative abundance and composition of gene functions; compared to bacteria, fungal global distribution was influenced by substrate/resource availability and energy demand variables (Bahram et al., 2018). Shotgun soil metagenomics have also been used to predict microbial genes involved in biogeochemical transformations in the soil environment. For example, Howe et al. (2016) used shotgun metagenomics to detect the presence of genes that encoded for carbon-cycling enzymes involved in biopolymer degradation (such as glycosyltransferase, glycoside hydrolase, carbohydrate esterase, carbohydrate-binding module, and polysaccharide lyase) on different soil types from a fertilized tallgrass prairie. Although differences in the relative abundance of these genes were observed, they were detected in multiple soil samples, indicating the ubiquity of microbial potential to drive C cycling in soils (Howe et al., 2016). Using 454 pyrosequencing and network analysis of bacterial 16S rRNA and fungal ITS genes in a time-course microcosm experiment, Banerjee et al. (2016) demonstrated that nutrient addition significantly altered microbial abundance, composition and co-occurrence. Decomposition was negatively related to microbial richness, evenness and diversity and microbial communities were interpreted to exhibit functional redundancy and keystone taxa were proposed to have strong positive associations during C cycling (Banerjee et al., 2016). Metagenomic approaches have also been applied to investigate C-cycling genes associated with management practices (e.g., addition of organic matter and/or fertilizers, and tillage practices) to predict C loss or storage in agricultural soils (Nivelle et al., 2016; Zhao et al., 2016; Cline and Zak, 2015). Field experiments in this area of research have shown that shifts of the copiotrophic communities within the soil microbiome reflect changes in soil biochemical

processes in agricultural land with frequent fertilizer and organic material additions (Carbonetto et al., 2014). Shotgun metagenomics has also been applied to investigate how microbial communities adapt to different management practices, where an increase in abundance of carbohydrate metabolism related gene fragments was associated with conventional tillage and crop rotation (Souza et al., 2015).

Modern sequencing techniques have dramatically improved our taxonomic understanding of soil microorganisms and have helped answer the question: 'Who is there?', but the next critical question of 'Who's doing what?' has not been resolved. Combining phylogenetic markers with functional measurements begins to address this problem (Cadotte et al., 2009), but this has rarely been done for microbial ecology (Morrissey et al., 2016). New approaches employing stable isotopes (e.g. ^{13}C and ^{18}O) to simultaneously evaluate soil processes and microbial community structures are beginning to resolve the relationships between phylogeny and function. In contrast to previous concepts of resource partitioning among broad microbial groups, fungi are known to be important in early stage decomposition of plant litter (Arcand et al., 2017; Müller et al., 2017; Kramer et al., 2016). Further, Morrisey et al. (2017) recently showed widespread bacterial plasticity of C use during decomposition with 69% of taxa using both fresh C and indigenous SOC. This new evidence challenges previous concepts of distinct, phylogenetically-clustered decomposers that exhibit consistent successional patterns of C use by showing that soil microbial resource use is more cosmopolitan than previously thought.

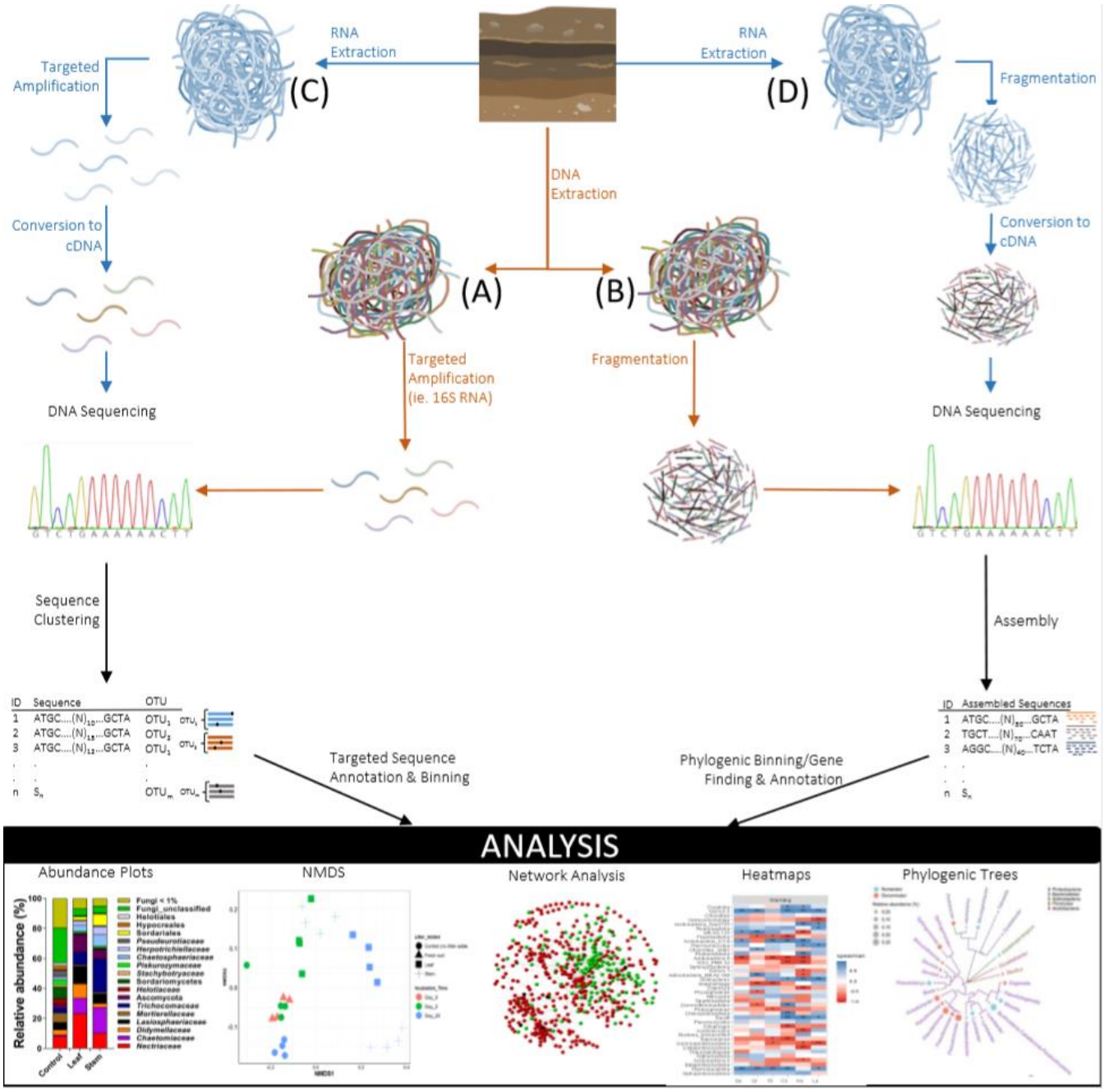


Figure A2.2: Soil metagenomics and metatranscriptomic studies provide insight into population diversity and functional dynamics of soil microbial communities. Targeted gene sequencing (A) or shotgun sequencing of gene fragments (B) are common soil metagenomics approaches used to examine microbial population composition and can focus on particular gene sequences encoding for enzymes and other proteins involved with functional ecosystem processes. Metatranscriptomics involves the extraction of RNA, targeted amplification (C) or fragmentation (D), and conversion to cDNA prior to DNA sequencing, to provide contemporaneous information of community functional processes (such as SOM stabilization) actively occurring in soil.

SOIL METATRANSCRIPTOMICS

Unlike metagenomics, soil metatranscriptomics provides contemporaneous community functional information about processes that are actively occurring in soil. Soil metatranscriptomics strives to represent the total RNA in a soil sample and is typically an indicator of the live and active members of a microbial community (Blagodatskaya and Kuzyakov, 2013). Growth and metabolism of soil microbial communities involve the transcription of genes into ribonucleic acids (RNA); molecules that are physically involved with protein assembly in the cell (e.g. rRNA and tRNA) or carry specific template information for protein translation (mRNA). However, as RNA gene transcripts are transient intermediate molecules, transcript profiles can vary over short periods of time, particularly in dynamic environments such as agricultural soils. Stable isotope probing with 18O recently showed that most soil bacteria produce new rRNA, indicating that they are metabolically active (Papp et al., 2018). Although soil metatranscriptomics is inherently challenging, well thought out experiments can yield strong, direct evidence of microbial activities in the soil from which associated soil biochemical processes can be extrapolated (Levy-Booth et al., 2019; Banerjee et al., 2016).

Metatranscriptomics involves the extraction of RNA from a microbial community (Figure A2.2), followed by sequencing (or the synthesis of cDNA, amplification and then sequencing). Ribosomal RNA (rRNA) accounts for the largest proportion of soil RNA (Orellana et al., 2018; Turner et al., 2013; Urich et al., 2008); whereas messenger RNA (mRNA) represents only a very small fraction (<5%) of the soil transcriptome (He et al., 2010). Owing to the low proportion of mRNA in the total soil transcriptome and the interests of many studies to investigate specific soil functions (e.g. carbon and/or nitrogen transformations), enrichment of mRNA prior to downstream analysis is often conducted (Carvalhais et al., 2012; He et al., 2010). Annotation of gene transcripts generally requires the use of a metagenome

scaffold as the identification of transcripts benefits from the availability of genome sequences obtained from pure cultures of microorganisms that are present in the soil sample (Prosser, 2015). Interpretation of metatranscriptomics data requires a link to be made between a transcript and the activity of its associated enzyme often requiring assumptions to be made about transcript and protein stability and turnover.

Similar technical issues that are encountered in metagenomic studies are associated with metatranscriptomics, with further potential bias introduced during the construction of cDNA libraries and a requirement for rapid inactivation of samples to prevent mRNA turnover (Prosser, 2015). The short half-life of RNA molecules requires RNA stabilizing reagents or flash-freezing samples at the site using liquid nitrogen. This can lead to significant challenges in applying this transcriptomics in field-based research where experiments are far from research infrastructure. Commercially available stabilization protocols are used as alternatives to the use of liquid nitrogen (e.g. LifeGuard Soil Preservation Solution, Qiagen); however, caution should be noted as stabilization protocol induced biases have been reported (McCarthy et al., 2015; Tatangelo et al., 2014). Some of the major challenges in extracting RNA from soil include the presence of coextractants (mainly organic acids; (Peršoh et al., 2008)), RNA adsorption to soil minerals, low yield of RNA, and the predominance of rRNA in the total extract (Wang et al., 2012). Although there are bioinformatics tools to specifically identify mRNA and rRNA sequences, low-expressed mRNA genes may be absent in the final sequence reads due to various biases (e.g. incomplete coverage as primers are database-dependent, preferential amplifications, and variable efficiency due to varied annealing temperature for different organisms (Brooks et al., 2015; García-Ortega and Martínez, 2015).

Metatranscriptomics applications for soil carbon stabilization

Metatranscriptomes of soil microbial communities provides valuable specific information on community population and metabolic response to environmental change and potential consequences for ecosystem functions such as SOM stabilization (Prosser, 2015). There are considerably fewer examples of the application of metatranscriptomics to investigate agricultural soils than to other substrates such as marine sediments and forest and grassland soils (Bailly et al., 2007). Tveit et al. (2014) sampled layers at different depths from two high-Arctic peat soils and carried out a total RNA metatranscriptomics comparison (using 16S rRNA and mRNA) to elucidate the patterns of microbial activity in peat soils. Transcript patterns revealed a gradual shift from aerobic to anaerobic microbiotas and metabolisms in the deep peat layers as polysaccharide compositions changed, corresponding to a transition from CO_2- to CH_4-dominated gas emissions with depth (Tveit et al., 2014). Kim and Liesack (2015) compared microbiome structure and functional succession in the oxic and anoxic layers of paddy soils using metatranscriptomics. Their results suggested that the taxonomic composition of paddy soil microbiomes is governed by the prevailing habitat conditions (such as oxygen availability); differences in the composition and succession of the microbiome between the two oxygen zones was related to the differential expression of transcripts of highly conserved and ubiquitous genes essential for metabolic activity in the respective zones. Fungi and Xanthomonadales appeared to cooperate in decomposing plant materials in the oxic zone whereas most of the dominant groups in the anoxic zone collaboratively transformed organic matter. Another study using shotgun sequencing of total extracted RNA from rice paddy soils showed that anaerobic biodegradation of SOM and methanogenesis were the predominate C transformations (Masuda et al., 2018). They also provided strong evidence that Anaeromyxobacter plays a key role in not only in the C and N cycles, but also in Fe reduction in the paddy soil environment.

SOIL METAPROTEOMICS

Soil metaproteomics studies the total protein in a sample to provide information of environmental processes. Proteins such as enzymes are related to microbial community functionality and are persistent in the environment, unlike RNA which is transient. Environmental metaproteomics has been applied to characterize microbial function in different environments (e.g. soil, activated sludge, acid mine drainage biofilms, water, and human gut (Starke et al., 2019; Bastida et al., 2009)). Many of the studies that characterized the soil metaproteome of different environments used similar, or closely related, analytical methods and instruments (Liu et al., 2017; Grob et al., 2015; Hultman et al., 2015; Keiblinger et al., 2012; Wang et al., 2011; Williams et al., 2010; Benndorf et al., 2007). Prior to analysis, soil-extracted proteins are usually separated by using 1D or 2D SDS polyacrylamide gel electrophoresis. Then, separated gel bands of proteins are digested (usually overnight with trypsin) into smaller peptides. Peptides are further separated using different liquid chromatographic systems and spectral features are collected using mass spectrometry. Fine tuning of protein extraction and sample preparation methods, and technological advances in mass spectrometers are important for the advancement of soil metaproteomics.

The unbiased extraction of proteins directly from soil is difficult because soil is one of the most complex and heterogeneous of all microbial environments. There are numerous methods reported for direct protein extraction from the soil; all generally involve cell lysis (using detergent, boiling, or sonication), stabilization of proteins (i.e. to prevent disulfide bond formation in cysteine residues of proteins) often by using reducing agents (e.g. dithiothreitol), precipitation of proteins (e.g. by using TCA), and cleanup (e.g. by using ice-cold acetone) (Hultman et al., 2015; Chourey et al., 2010; Benndorf

et al., 2007). Major challenges associated with the direct isolation of proteins from soil include: co-extraction of organic acids (which impact the downstream analysis), binding of proteins to metal ions, soil particles and organic matter, denaturation of proteins by the extraction conditions (e.g. chemicals, temperature, ionic strength, and pH), the low abundance of some important proteins, and the heterogeneous nature of the soil matrix (Bastida et al., 2009). Protein recovery is highly dependent on extractant and soil type, suggesting that no single or universal extractant is capable of complete protein recovery. This may limit direct comparison of studies using different recovery methods, especially if no extraction controls are used (Greenfield et al., 2018). Finally, both the high diversity of microorganisms in soil and the interaction of environmental conditions likely plays a role in the specific response of the microbial community to drivers of microbial functioning, such as land use and management practices.

These challenges make characterizing the soil metaproteome a daunting task. A study by Nicora et al. (2013) demonstrated enhanced recovery of proteins from sediment soil samples by blocking protein binding sites in the soil using a mixture of polar positive amino acids prior to in situ lysis and by enhancing the release of proteins from the soil using an optimized desorption buffer. As these amino acids may pose interferences on the MS analysis, the authors optimized the concentrations to be small enough to avoid detection by the MS. Parallel or sequential extraction of soil proteins from a soil sample using different extraction methods and combining the spectral data may improve the detection capability. For instance, Mattarozzi et al. (2017) combined protein spectral features obtained from three different extraction methods, and showed that up to 1.5% of the total proteins (~ 294 unique proteins) from each sample could be detected by using the methods. Extraction efficiency of proteins may also be improved following separation and isolation of microorganisms from the soil by differential

centrifugation; however, in such circumstances, extra-cellular proteins (e.g. polymer-degrading enzymes) in the soil solution are lost (Chourey et al., 2010).

Soil microbial communities are heterogeneous, dominated in density by a few species, while the remainder of the microbial diversity are present in low abundances. Deciphering the presence of specialized functional proteins from complex protein profiles dominated by proteins of ubiquitous function (such as 'house-keeping' proteins involved with cellular respiration, energy metabolism, DNA replication, etc.) is therefore a considerable challenge, especially when produced by low abundance community members (Starke et al., 2019). Further complications arise during the annotation stage; where protein identification relies upon successful database matching. Multiple protein databases are available; however, these databases favor intensively-investigated and well characterized model organisms (i.e. Escherichia coli, Saccharomyces cerevisiae, and Caenorhabditis elegans) and therefore annotation of functional proteins from community members of "unknown" identity is difficult and tenuous (Starke et al., 2019).

Metaproteomics applications for soil carbon stabilization

The large-scale identification and quantification of proteins from soil microbial communities has provided novel insights into the phenotypes of microorganisms on the molecular level. Liu et al. (2017) used soil metaproteomics to evaluate the effects of artificial soil heating (4 °C above ambient) on the structure and function of the microbial-derived enzymatic pool in forest soil. Significant increases in soil respiration with soil warming and a reduction in dissolved organic carbon levels coincided with increases in some bacterial and fungal taxa abundances (e.g. Actinobacteria and Agaricomycetes). However, increases in the abundances were observed at higher taxonomic levels where members can have

significant metabolic differences, and hence assignment of proteins to a specific taxon is difficult. Williams et al. (2010) characterized the microbial community proteome in a soil and a culture inoculated with soil microbes, both amended with toluene. Although a total of 47 proteins were identified, the authors suggested that this was likely a small subset of the proteins expressed by the soil microbial communities. The proteins included glutamine synthetase, ATP-binding cassette (ABC) transporters, extracellular solute-binding proteins, and outer membrane proteins.

Aiming to investigate if the addition of recalcitrant suberin to the soil can enrich specific microbial communities, Sidibe et al. (2016) conducted an incubation study and characterized the soil metaproteome using quadrupole TOF MS (Qq-TOF MS). Counts of protein spectral features indicated a decline in the proportion of fast-growing bacteria and enrichment of microbes that can degrade these substrates. Soil metaproteomics have also been successfully applied to characterize the structure of metabolically-active microorganisms in rhizosphere soils (Bona et al., 2019; Mattarozzi et al., 2017; Wang et al., 2011). Wang et al. (2011) detected numerous proteins with potential functions in energy metabolism, protein turnover and amino acid biosynthesis, secondary metabolism, nucleotide metabolism, signal transduction, and resistance mechanisms in crop rhizosphere soils. Although 43% of the proteins were unknown, about 23% of the remaining proteins were from the rhizosphere, and from microorganisms in which Proteobacteria have been identified as the most predominant in the metaproteomics (44%) and T-RFLP (48%) libraries. Another study by Mattarozzi et al. (2017) successfully applied soil metaproteomics to rhizosphere soils obtained from plants that were metal-tolerant and metal hyperaccumulators. The authors identified up to 294 unique bacterial proteins using liquid chromatography (LC)-high resolution mass spectrometry (HRMS) analysis of the soil metaproteome.

An ideal tool to help unravel important abiotic and biotic interactions in the carbon cycle would be to trace carbon from the source input through individual species into a stable form in soil using a marker. Tracing carbon using stable isotope-labelling of the soil metaproteome may help to link species to specific substrate in situ (Grob et al., 2015; von Bergen et al., 2013; Seifert et al., 2012). Kleiner et al. (2018) developed a direct protein stable isotope fingerprint method that links microbial species in communities to the carbon source they consume by determining their stable carbon isotope signature. The method involves preparation of peptide mixtures, measuring their $^{12}C/^{13}C$ or $^{14}N/^{15}N$ ratios using 1D or 2D LC-MS/MS, database searching of the peptide spectra, and using the spectral scores and raw MS data as inputs to the software which has been shown to provide single, robust, averaged $^{12}C/^{13}C$ or $^{14}N/^{15}N$ ratios per species.

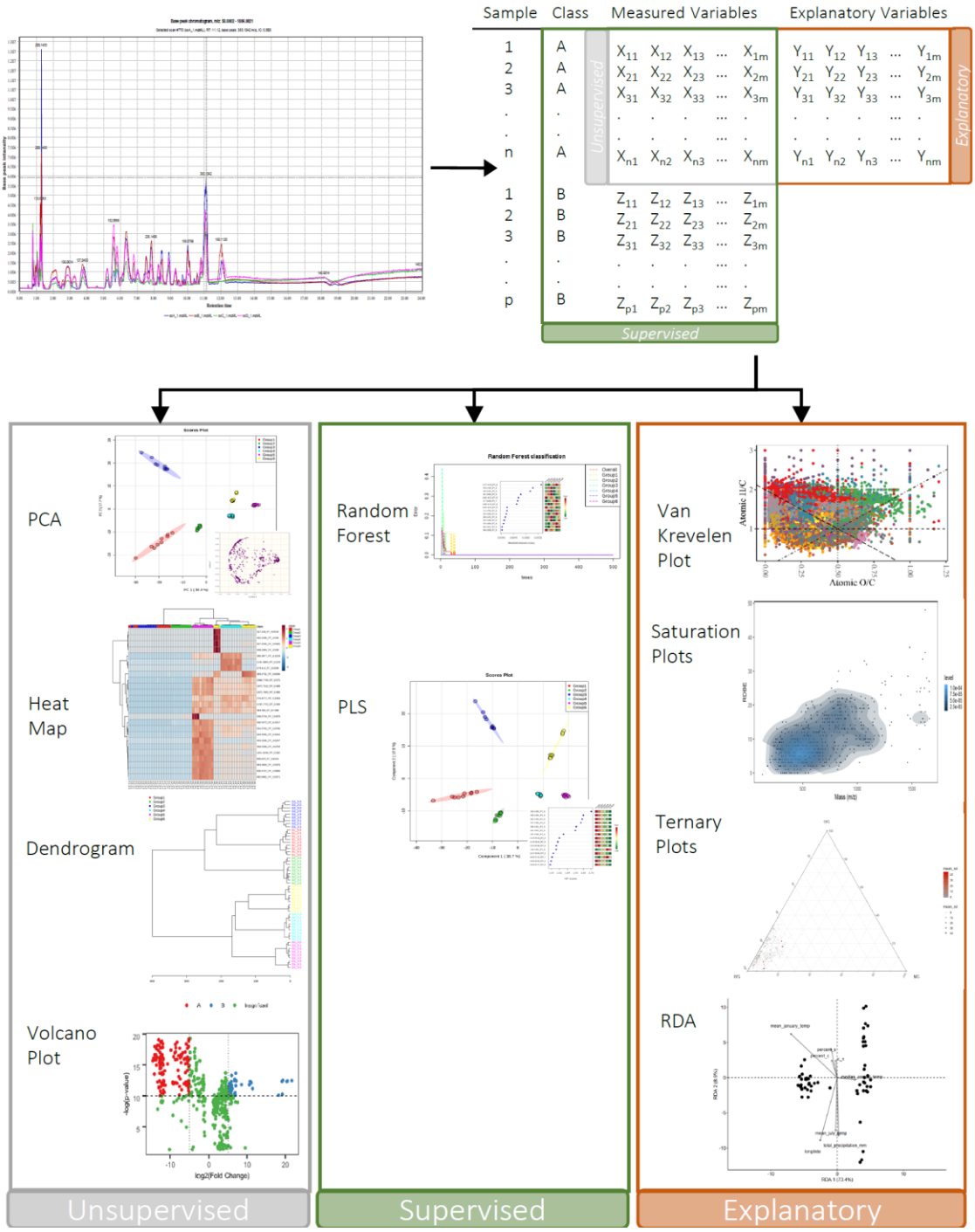


Figure A2.3: Application of untargeted or semi-targeted metabolomics analysis of SOM provides insight into physiochemical and biological drivers of SOM stabilization and mineralization. Untargeted metabolite profiling is intended to have maximal coverage of chemical space; whereas semi-targeted analyses focus on a specific molecular class or range of classes (e.g., N containing compounds, or fatty acids). During data processing (i.e. of a LC-MS dataset), mass spectral chromatograms (3 dimensional data: retention time, m/z range (50-2000 Da), relative signal abundance) are converted into 2 dimensional data matrices through data preprocessing software, followed by multivariate & univariate statistical analysis (unsupervised, supervised and molecular fingerprinting methods).

SOIL METABOLOMICS

When compared with fields such as toxicology, pharmacology, and medicine, application of metabolomics to the soil environment is still in its infancy (Swenson et al., 2018). Soil metabolomics characterizes the composition of small metabolites in the soil (by-products, intermediates, and end products of cellular processes), provides information directly on carbon transformation and stabilization as well as the constituents of SOM therein (Figure A2.3). Historically, microbial metabolites in soil are classified as endometabolites (i.e., intracellular metabolites) and exometabolites (i.e., extracellular metabolites) (Silva and Northen, 2015); however, endometabolites may also be released to the soil solution from dead and lysed microbial cells. Therefore, although exometabolites may predominate the soil solution and are presumed to be more accessible to the soil microbiome, metabolites extracted from the soil solution (without using cell lysing agents such as chloroform fumigation and organic solvents) are not exclusively exometabolites. Application of untargeted or semi-targeted metabolomics analysis of SOM may be used improve our understanding of the physiochemical and biological drivers of SOM stabilization and mineralization. There are a number of excellent untargeted analyses of small molecules and SOM composition of soil by soil chemists and geologists (Grewer et al., 2018; Li et al., 2018; Wu et al., 2018; Pisani et al., 2016; Seifert et al., 2016; Pautler et al., 2013); in which the term 'molecular characterization' has been used rather than 'metabolomics' – as the term 'metabolomics' infers small molecules associated with metabolic processes. This discrepancy in semantics is an important consideration, in that, the soil environment is complex and SOM chemistry is a reflection of both biotic and abiotic effects upon small organic molecules, wherein not all small molecules in SOM are directly derived from metabolism.

Metabolomics is applied in research as a tool to characterize the metabolite or small molecule (100 – 2000 Da range) content of a representative extract and is used to compare differences in composition between sample types. Untargeted metabolomics (metabolite profiling) is a global analysis that characterizes as much of small molecule chemistry of the SOM as possible; whereas semi-targeted analyses focus on a specific molecular class or range of classes (e.g., N containing compounds, or fatty acids). Untargeted metabolite profiling is intended to have maximal coverage of chemical space; however, regardless of the extraction method chosen, some molecules will be missed because they are present in trace amounts, not soluble with the extract used, and/or incompatible with the instrument methods used for profiling. Semi-targeted extractions are comparably less complex and easier to analyze, but important cross-class metabolite interactions may be missed. Untargeted metabolite profiling aims to generate hypotheses to explain difference in molecular patterns observed while targeted analyses are often used to test hypotheses as they are can be more quantitative and are therefore used to answer questions related to specific classes of compounds.

Selection of an appropriate extraction system is critical for unbiased characterization of soil metabolites. The diverse nature of chemical constituents within soil limits the use of a single solvent to extract and obtain all available molecules. Thus, global-untargeted studies require parallel or sequential extraction techniques (using different extractants) on a single soil sample (Tfaily et al., 2015, 2017), and to combine data obtained from different analytical tools (Jenkins et al., 2017). The terms ‘soil dissolved organic matter’ (DOM), ‘water extractable organic matter’ (WEOM), or ‘water extractable organic carbon’ (WEOC) are commonly used to describe water extracts from soil. Swenson et al. (2015) profiled metabolites using different solvents and demonstrated that water ranked as the best extractant for soil exometabolites because no significant differences were observed when compared with other

extractants which included methanol, K_2SO_4, and NH_4HCO_3. By using water as an extractant, potential artifacts created by the other chemical extractants may be avoided (Sauerschnig et al., 2017), but non-polar metabolites such as lipophilic molecules (that are also important components of SOM stabilization) are excluded. Chloroform fumigation is commonly used to lyse cells prior to extraction in order to increase both the diversity and abundance of extractable soil metabolites (with a particular bias toward nucleosides, nucleotides, and amino acids (Warren, 2014)); however, chloroform fumigation does not inhibit all exoenzyme activity, and care should be taken following cell lysis during sample processing as active enzymatic degradation of organic matter can continue for the duration of the extraction process (Blankinship et al., 2014).

Metabolites of various sources are usually a mixture of different molecular compounds of varying abundance and chemical properties (e.g. carbohydrates, amino acids, peptides, lipids, nucleic acids, and organic acids). Different analytical tools are available for the identification and quantification of soil metabolites (Alonso et al., 2015; Zhang et al., 2012); however, Nuclear Magnetic Resonance (NMR) spectroscopy and Mass Spectrometry (MS) are the most commonly used. Both tools have advantages and drawbacks. NMR provides direct relationships to the amount of each molecular constituent in a mixture and yields interpretable information on molecular structures; this technique is considerably less sensitive in comparison to mass spectrometry, which is more ideally suited for the detection of low-abundance metabolites (Zhang et al., 2012). Annotation of the individual constituents of an extract is difficult and often requires some degree of separation prior to detection for proper annotation. For this reason, MS is typically coupled with separation tools including high-performance liquid chromatography (HPLC), ultra high-performance liquid chromatography (UHPLC), gas chromatography (GC), and capillary electrophoresis (CE) that can reduce matrix complexity through

separation of metabolites prior to MS detection. However separation comes at a cost, as only metabolites that are soluble in solvent systems used for injection and elution (in liquid chromatography) and derivatization and thermostability (in gas chromatography) are observed; introducing a bias or complexity reduction of metabolic coverage in global-untargeted metabolomics analyses. In such a case, often multiple chromatographic systems or technologies will be used to profile soil extracts to maximize molecular coverage (i.e. GCMS and LCMS profiling, or reverse phase and HPLIC chromatography).

Equally important to a proper study design (e.g. with a sufficient number of biological replicates and including control experiments and method blanks), is a well-planned sampling and sample preparation. Sample collection and storage duration can influence the results as the passage of time can shift the quantity and quality of metabolites available in the soil of improperly stored samples. Thus. unbiased characterization of metabolites at a specific time of the study requires an immediate quenching of the microbial/enzymatic activity in the soil samples, and reduced exposure to light and oxygen. The easiest method of quenching microbial metabolism is using liquid nitrogen and flash freezing the soil; this is particularly important when samples need to be transported from a study site to the lab. If the soil samples are collected from a lab study, it may be possible to immediately store in a -80°C freezer or flash freeze with liquid N2 prior to -80 °C storage. Properly quenched soil samples can then be used to extract endo and/or exometabolites, depending on the research question.

Metabolomics applications for soil carbon stabilization

Metabolomics has been successfully applied in several experimental models to explore SOM. Clemente et al. (2013) compared the composition of SOM after incubation with different parts of the maize plant. They found that after 36-weeks of incubation, soil amended with maize roots had the

greatest microbial SOM contribution, compared to soil amended with stems or leaves. This is consistent with other studies showing that the 'quality' of plant inputs affects the soil microbial community. Rochfort et al. (2015) used NMR-based soil metabolomics and compared metabolite profiles in soils under different land use and found that lipid- and sugar-related metabolites were distinctly abundant in an agricultural soil, whereas soil from a remnant site was dominated by terpene metabolites. The abundance of easily degradable carbon substrates in the agricultural soils may be related to the larger inputs of organic carbon. Pisani et al. (2015) assessed the SOM composition across a large temperature gradient with changes in land use. Using NMR they were able to determine that soil cultivation resulted in increased microbial contributions at higher temperatures, had lower plant contributions, and increased suberin and lignin degradation. They concluded land-use changes associated with cultivation may become sources of atmospheric CO_2. Similarly, Seifert et al. (2016) compared soil DOM from managed meadows, deciduous forest, mixed-coniferous forest, and agricultural lands by electrospray ionization (ESI) FTICR MS to link changes in environmental parameters (i.e. pH and nitrate concentrations) to changes in SOM composition across the different land uses. Li et al. (2017) also used FTICR MS to assess the DOM in paddy soils. They observed that pH, iron redox and carbon to nitrogen ratios controlled DOM compositions, where soil DOM was more recalcitrant at higher pH and with larger carbon to nitrogen ratios.

Untargeted SOM analyses have attempted to link together SOM compositional knowledge with microbial analyses to investigate SOM stabilization. Linked soil metabolomics has been used in a few instances to determine soil metabolite profiles during wetting/drying (Swenson et al., 2018), in ecological assessments (Beale et al., 2017, 2018), to study permafrost carbon storage (Ward and Cory, 2015), with soil amendments (Mitchell et al., 2015), and for developing improved growth media for the

soil microbiome (Jenkins et al., 2017). Some studies link to microbial activity more generally. Jenkins et al. (2017) used LC-MS and GC-MS to characterize soil metabolites to formulate soil-defined media for the cultivation of previously uncultured microorganisms. Ward and Cory (2015) demonstrated linking soil DOM to characterize bacterial growth and activity parameters such as respiration, production, and growth efficiency in permafrost soils. They found DOM from surficial organic mats contained more high-molecular weight compounds that were more oxidized with a higher degree of unsaturation compared DOM from the mineral layer in permafrost. They also found the bacterial production rates and efficiencies were greater in permafrost than in the surficial organic mat, that may have implications for DOM cycling as the permafrost melts. Warren (2014) proposed a method using LC-MS to target fatty acids as indicators of microbial activities, to infer specific microbial contributions to the formation of SOM. Similarly, Mitchell et al. (2015) used untargeted NMR paired with a targeted analysis of phospholipid fatty acids (PLFAs) to link the effects of biochar amendment to impacts on the soil microbial community and the WEOM fraction of SOM. PLFAs can be used to monitor changes among Gram-positive bacteria, Gram-negative bacteria, actinomycetes, and fungi communities in soil. This technique, paired with untargeted NMR of the DOM suggested that the microbial community shifted from fungal-dominated to more bacterial dominated and from Gram-negative bacteria to more Gram-positive bacteria within 16 weeks after biochar amendment. In the same time frame they also saw an increase in WEOM concentrations of short-chain carboxylic acids but decreases in WEOM carbohydrates and peptides. Their conclusion was that biochar amendments increased soil microbial activity and thus CO_2 emissions.

SOIL SAMPLING FOR OMICS

Soils vary laterally across the landscape and vertically through the profile and therefore comprise diverse microhabitats, even at fine (cm) scales (O'Brien et al., 2016). The degree of heterogeneity strongly depends on: (i) texture, which influences extraction efficiency; ii) soil structure and related properties (e.g. aggregation, drainage); iii) vegetation type and proximity to growing plants (e.g. the rhizosphere); (iv) season; and (v) specific site characteristics such as topographic location (slope, depressions) and groundwater table. Taking this heterogeneity into account, the typical 0.5 to 2 g soil sample used for an 'omics' extraction does not reflect a single microsite, but a mixture of different microhabitats with different chemical, physical, and biological properties. The effects of spatial heterogeneity can be reduced by increasing the amount of soil and number of samples used for an extraction. For example, regarding DNA extraction from soil, Penton et al. (2016) observed that sample size significantly affected overall bacterial and fungal community structure, and the number of operational taxonomic units retrieved. They found that recovery of bacterial and fungal diversity improved substantially as the mass of soil samples increased from 0.25 to 10 g.

The inherent heterogeneity of soil is a primary driving factor when developing a sampling strategy for a soil 'omics' experiment. A preliminary assessment of the spatial variation of soil properties is a prime requisite for developing a sound soil sampling strategy. In agricultural soils, where factors such as tillage practices influence the distribution of resources with depth in the profile, the vertical heterogeneity may be of interest and must be accounted for (Sun et al., 2018). In this case the sampling strategy should consider the depth of plowing as well as the stratification of horizons with varying depths, to avoid the mixing of different soil horizons. One soil sample taken from a given site is an inadequate representation; therefore, true soil sample replicates (multiple individual samples taken,

rather than a single sample split into ‘replicates’) need to be collected and analyzed. Moreover, to limit the influence of soil heterogeneity within a given sample, a representative sampling strategy, involving the harvesting and subsequent pooling of subsamples, should be used.

Overlaid on these spatial factors are temporal factors that need to be taken into account because soil microhabitats and organisms therein change in response to management practices like fertilization, tillage, harvesting, plant growth stage, weather, and season. Likewise, the stabilization and accumulation of proteins and metabolites are strongly influenced by the local conditions prior to sampling. Therefore. sampling at a time point is a ‘snapshot’ of nucleic acids, proteins and metabolites, each of which may change at different time scales. Monitoring and evaluating antecedent local conditions (e.g., soil temperature and water content) is therefore necessary to interpret omics data. A snapshot for metagenomic data may require evaluation of temperature and moisture conditions over a period of days, weeks or months. The issue of spatial and temporal heterogeneity is more pronounced when analyzing RNA because the stability of mRNA in cells is often in the order of minutes to hours. Thus one sampling reflects only a brief snapshot that depends very much on the actual environmental conditions at the time of sampling. By taking account of the strong spatial and temporal dynamics of soil microbial communities in both time and space, a given sampling strategy needs to be driven by a clear research hypothesis.

As abiotic soil properties are a major driver of soil microbiomes, a minimum dataset is required, which needs to be analyzed and implemented independently from the research question(s). A basic suite of these should consist of soil texture, pH, organic C and N. Appropriate contextual information is critical to interpret most omics information. For example, at agricultural sites, current and past management practices should be listed. These include time of fertilization, tillage, crop growth and

yield, crop sequence and biomass should be given. For unmanaged sites the land use and/or vegetation type should be characterized.

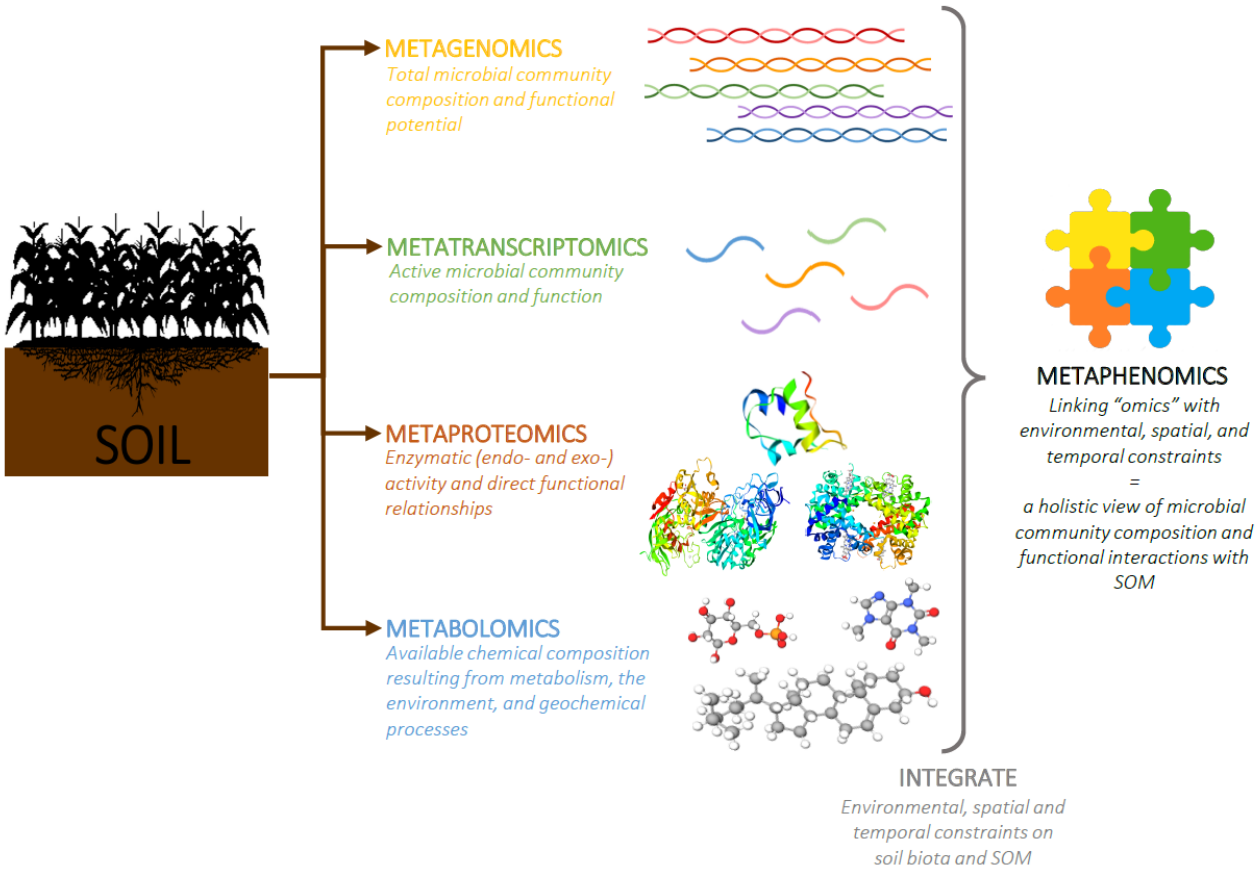


Figure A2.4: Linking data resulting from parallel analysis of soil samples using various 'omics' platforms (metagenomics, metatranscriptomics, metaproteomics, metabolomics) provides a more holistic view (metaphenomics) of microbial community composition and understanding of functional interactions associated with biogeochemical processes involved in SOM stabilization.

LINKING OMICS TECHNOLOGIES – SOIL METAPHENOMICS

Soil biogeochemical processes are intimately interconnected with their associated soil microbiome. Jansson and Hofmockel (2018) recently defined the term soil metaphenome "as the product of expressed functions encoded in microbial genomes (metagenome) and the environment

(resources available; spatial, biotic and abiotic constraints)". To evaluate the biological stabilization of carbon in soil, the interconnectivity within this network has multiple overlaps between functional response and microbial community composition. Many common microbial enzymes and metabolites are produced from ubiquitous gene clusters, all playing specific roles in network interconnectivity as inputs/outputs from one type of microbe are actively used by a different group of the soil microbial community. Current 'omics' approaches have successfully modeled various aspects of these processes; however, a single 'omics' tool only allows for a single perspective of this heterogenous environment, generating a partial (or fragmented) representation of a complex network of interactions. Only by combining observations from multiple 'omics' platforms together, to link genetic potential and community structure into functional relationships with gene, protein, and metabolite, can a more holistic vision of soil metaphenomics begin to be resolved (Figure A2.4).

Going forward, soil microbial metagenomics needs to move beyond simple descriptions of community diversity (Fierer, 2017) to characterize and evaluate important soil processes (such as C cycling). Community metabolic function is strongly influenced by nutritional, energetic and stoichiometric constraints, while taxonomic variation within these communities is only poorly explained by environmental conditions (Louca et al., 2016; Nelson et al., 2016; Raes et al., 2011). Linking 'omics' experiments through genome-centric metagenomics and metatranscriptomics will provide insight into the metabolic basis and mechanism for phylotype involvement (Prosser, 2015); however, consideration of temporal dynamics when setting up an experiment of merged 'omics' studies is necessary to link gene cluster potential with transcript abundance and metabolic activity.

In field studies these temporal dynamics embody changes in the soil 'climate', i.e., temperature and water content, environmental variables that continually change in soil. Soil is in a continual state of

flux with wetting/drying, warming/cooling or freezing sometimes occurring hourly. In agricultural soils these rapid changes can be influenced by crop growth (temporal input of C and competition for N and P resources) and soil management (such as tillage). Overlaid on these rapid changes are diurnal and seasonal changes in temperature and water content. Thus, characterizing and understanding antecedent soil biotic conditions before the time of soil sampling is a prerequisite for interpreting metagenomics data in field studies. In contrast, laboratory studies are often conducted where temperature and water content is kept static to either avoid fluctuations or to slow them down to allow for specific sampling times to capture the dominant process(es). For these reasons, selection of an appropriate soil sampling strategy and the collection of appropriate and well curated metadata is essential for data interpretation.

Environmental DNA can represent dead, dormant, or relic members of the microbial community (Levy-Booth et al., 2007). Carini et al. (2016) found that 40% of DNA extracted from soils was extracellular or belonged to cells that were no longer intact, potentially skewing interpretations of the relative abundance of different taxa. Effects of relic DNA were shown by Lennon et al. (2018) to represent 33 to 80% of extracted bacterial DNA in sediments. RNA transcripts are transient and rapidly degraded following their transcription making them less subject to similar bias. Changes in cellular metabolism and exoenzymatic degradation of SOM can only be detected after transcription of their associated proteins have occurred and metabolite products have accumulated, decoupling SOM breakdown from population estimates based on nucleic acid analysis. Without factoring in temporal dynamics, erroneous assumptions can be made regarding community structure (genomic predictions of functional potential) and the expression of a functional response. In experiments factoring in temporal change by repeated sampling over various time points, trends in gene transcription, translation and

metabolite fluxes become decipherable and stronger associations with genome potential and functional expression of a microbial population can be made.

An excellent example of a controlled and temporally monitored experiment that would have been an ideal candidate for a linked 'omics' approach was published by Kallenbach et al. (2016). In this study, soil derived microbial seed inoculum was added to different sterile, mineral soils (kaolinite or montmorillonite clay mixed with quartz sand – containing little to no SOM) and each mineral soil mesocosm was provided with weekly additions of differing C-substrate inputs (glucose, cellobiose, syringol, or soil DOC) over a 15 month period and then left without additions for an additional 3 months. Temporal changes in microbial population composition were characterized using PLFA biomarkers (cursorily classified in terms of Gram Negative bacterial, Gram Positive bacterial, Actinobacterial, and fungal groupings). Proxy measurements of carbon use efficiency were also made to estimate microbial residue production. The authors used a metabolomics approach to characterize changes in SOM chemistry in terms of major functional classes of metabolites. Over an 18 month period, convergence in SOM chemistry (and in turn SOM stabilization) was observed regardless of C-input or clay soil type as a direct result of microbial activity, biomass accumulation and stabilization into microbial necromass. Inclusion in this experiment of one or more of the other 'omics' platforms would have provided relevant functional information in terms of which particular transformative processes were initiated and when (metatranscriptomics), what enzymatic reactions were occurring (metaproteomics), and how microbial population dynamics (metagenomics) fluctuated over time to result in a convergence in SOM chemistry between the different mesocosms investigated.

There are a few examples of research to link soil 'omics' techniques in the scientific literature. Recently, Beale et al. (2017, 2018) combined untargeted GC-MS metabolomics with bacterial

metagenomics. Beale et al. (2018) characterized metabolite profiles and microbial community structure in sediments during dry and wet seasonal conditions to account for biological and physicochemical variance due to non-rainfall-based abiotic stresses. The authors identified significant metabolic features (e.g. 3,5-dihydroxyphenylglycine, beta-alanine, L-allothreonine, and numerous unknown metabolites) that increase (2.2 to 13.2-fold) during the dry season that were linked with a transition in bacterial community composition towards organisms that utilize more complex organic energy sources, such as carbohydrates and fatty acids, and anaerobic redox processes (Beale et al., 2018). Swenson et al. (2018) also combined metagenomics with GC-MS metabolomics to provide a high resolution assessment of the soil microbial community. They used biocrusts as a model system to demonstrate the feasibility of integrating metagenomics with soil metabolomics to generate food webs that develop after wetting dry soils. They also discussed building databases that can contribute to more sophisticated models for carbon flux predictions. Such models would be invaluable for future data miners and modellers interested in climate change or modelling state changes in agricultural soils following a change in cropping history or soil management practices.

CONCLUSION AND FUTURE PERSPECTIVE

'Omics' technologies encompass a broad suite of techniques that are currently being applied to diverse scientific fields to understand complex environmental processes. These technologies provide opportunities to advance our understanding of how and which microorganisms transform soil organic matter inputs into stable compounds that enhance soil health and productivity. The real potential of applying 'omics' in this area of soil science will be realized by linking 'omics' technologies together in well-designed experiments (taking spatial/temporal changes into consideration) to generate relevant

and testable hypotheses. Integration of linked 'omics' approaches with traditional methods used in soil science are relatively recent but will become more prevalent in the near future. Modern technological advances are constantly occurring in all fields of 'omics'. With increased sequencing depth and decreasing costs, sequencing of environmental DNA and RNA are making the assembly of 'population genomes' feasible. Application of high resolution mass spectrometers in environmental studies are providing greater accuracy and resolution for peak deconvolution and annotation in metabolomics and proteomic studies. These technological advances aid to provide greater resolution and insight into microbial populations and how these populations change and interact with their environment. 'Omics' platforms are multi-disciplinary and require expensive technological infrastructure. Often a single lab is not able to integrate more than one 'omics' tool from their toolbox; therefore, more collaboration will likely be required to achieve a linked 'omics' approach (e.g. soil metaphenomics approach) to studying C-dynamics/stabilization in soil, as well as in other areas of environmental sciences.

ACKNOWLEDGEMENTS

This work was supported by the Canadian Agricultural Partnership – AgriInnovate Program from Agriculture and AgriFood Canada [Project ID# J-001756: "Understanding mechanisms in the biological stabilization of carbon in soil"].

REFERENCES

Alonso, A., Marsal, S., Julià, A., 2015. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. Frontiers in Bioengineering and Biotechnology 3. doi:10.3389/fbioe.2015.00023

Amelung, W., Brodowski, S., Sandhage-Hofmann, A., Bol, R., 2008. Chapter 6 Combining Biomarker with Stable Isotope Analyses for Assessing the Transformation and Turnover of Soil Organic Matter, in: Advances in Agronomy. Elsevier, pp. 155–250. doi:10.1016/S0065-2113(08)00606-8

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Zech Xu, Z., Kightley, E.P., Thompson, L.R., Hyde, E.R., Gonzalez, A., Knight, R., 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. MSystems 2, e00191-16, /msys/2/2/e00191-16.atom. doi:10.1128/mSystems.00191-16

Amundson, R., Berhe, A.A., Hopmans, J.W., Olson, C., Sztein, A.E., Sparks, D.L., 2015. Soil and human security in the 21st century. Science 348, 1261071–1261071. doi:10.1126/science.1261071

Arcand, M.M., Levy-Booth, D.J., Helgason, B.L., 2017. Resource Legacies of Organic and Conventional Management Differentiate Soil Microbial Carbon Use. Frontiers in Microbiology 8. doi:10.3389/fmicb.2017.02293

Bahram, M., Hildebrand, F., Forslund, S.K., Anderson, J.L., Soudzilovskaia, N.A., Bodegom, P.M., Bengtsson-Palme, J., Anslan, S., Coelho, L.P., Harend, H., Huerta-Cepas, J., Medema, M.H., Maltz, M.R., Mundra, S., Olsson, P.A., Pent, M., Põlme, S., Sunagawa, S., Ryberg, M., Tedersoo, L., Bork, P., 2018. Structure and function of the global topsoil microbiome. Nature 560, 233–237. doi:10.1038/s41586-018-0386-6

Bailly, J., Fraissinet-Tachet, L., Verner, M.-C., Debaud, J.-C., Lemaire, M., Wésolowski-Louvel, M., Marmeisse, R., 2007. Soil eukaryotic functional diversity, a metatranscriptomic approach. The ISME Journal 1, 632–642. doi:10.1038/ismej.2007.68

Baldock, J., Broos, K., 2012. Soil organic matter, in: Handbook of Soil Sciences: Properties and Processes. Taylor & Francis Group, LLC, Boca Raton, FL, pp. 11-1-11–52.

Banerjee, S., Kirkby, C.A., Schmutter, D., Bissett, A., Kirkegaard, J.A., Richardson, A.E., 2016. Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. Soil Biology and Biochemistry 97, 188–198. doi:10.1016/j.soilbio.2016.03.017

Bastida, F., Moreno, J.L., Nicolás, C., Hernández, T., García, C., 2009. Soil metaproteomics: a review of an emerging environmental science. Significance, methodology and perspectives. European Journal of Soil Science 60, 845–859. doi:10.1111/j.1365-2389.2009.01184.x

Beale, D.J., Crosswell, J., Karpe, A.V., Ahmed, W., Williams, M., Morrison, P.D., Metcalfe, S., Staley, C., Sadowsky, M.J., Palombo, E.A., Steven, A.D.L., 2017. A multi-omics based ecological analysis of coastal marine sediments from Gladstone, in Australia's Central Queensland, and Heron Island, a nearby fringing platform reef. Science of The Total Environment 609, 842–853. doi:10.1016/j.scitotenv.2017.07.184

Beale, D.J., Crosswell, J., Karpe, A.V., Metcalfe, S.S., Morrison, P.D., Staley, C., Ahmed, W., Sadowsky, M.J., Palombo, E.A., Steven, A.D.L., 2018. Seasonal metabolic analysis of marine sediments collected from Moreton Bay in South East Queensland, Australia, using a multi-omics-based approach. Science of The Total Environment 631–632, 1328–1341. doi:10.1016/j.scitotenv.2018.03.106

Benndorf, D., Balcke, G.U., Harms, H., von Bergen, M., 2007. Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. The ISME Journal 1, 224–234. doi:10.1038/ismej.2007.39

Bergmann, G.T., Bates, S.T., Eilers, K.G., Lauber, C.L., Caporaso, J.G., Walters, W.A., Knight, R., Fierer, N., 2011. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biology and Biochemistry 43, 1450–1455. doi:10.1016/j.soilbio.2011.03.012

Blagodatskaya, E., Kuzyakov, Y., 2013. Active microorganisms in soil: Critical review of estimation criteria and approaches. Soil Biology and Biochemistry 67, 192–211. doi:10.1016/j.soilbio.2013.08.024

Blankinship, J.C., Becerra, C.A., Schaeffer, S.M., Schimel, J.P., 2014. Separating cellular metabolism from exoenzyme activity in soil organic matter decomposition. Soil Biology and Biochemistry 71, 68–75. doi:10.1016/j.soilbio.2014.01.010

Bona, E., Massa, N., Novello, G., Boatti, L., Cesaro, P., Todeschini, V., Magnelli, V., Manfredi, M., Marengo, E., Mignone, F., Berta, G., Lingua, G., Gamalero, E., 2019. Metaproteomic characterization of Vitis vinifera rhizosphere. FEMS Microbiology Ecology. doi:10.1093/femsec/fiy204

Brooks, J.P., Edwards, D.J., Harwich, M.D., Rivera, M.C., Fettweis, J.M., Serrano, M.G., Reris, R.A., Sheth, N.U., Huang, B., Girerd, P., Vaginal Microbiome Consortium (additional members), Strauss, J.F., Jefferson, K.K., Buck, G.A., 2015. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiology 15, 66. doi:10.1186/s12866-015-0351-6

Cadotte, M.W., Cavender-Bares, J., Tilman, D., Oakley, T.H., 2009. Using Phylogenetic, Functional and Trait Diversity to Understand Patterns of Plant Community Productivity. PLoS ONE 4. doi:10.1371/journal.pone.0005695

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13, 581–583. doi:10.1038/nmeth.3869

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R., 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7, 335–336. doi:10.1038/nmeth.f.303

Carbonetto, B., Rascovan, N., Álvarez, R., Mentaberry, A., Vázquez, M.P., 2014. Structure, Composition and Metagenomic Profile of Soil Microbiomes Associated to Agricultural Land Use and Tillage Systems in Argentine Pampas. PLoS ONE 9, e99949. doi:10.1371/journal.pone.0099949

Carini, P., Marsden, P.J., Leff, J.W., Morgan, E.E., Strickland, M.S., Fierer, N., 2016. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. Nature Microbiology 2, 1–6. doi:10.1038/nmicrobiol.2016.242

Carvalhais, L.C., Dennis, P.G., Tyson, G.W., Schenk, P.M., 2012. Application of metatranscriptomics to soil environments. Journal of Microbiological Methods 91, 246–251. doi:10.1016/j.mimet.2012.08.011

Chourey, K., Jansson, J., VerBerkmoes, N., Shah, M., Chavarria, K.L., Tom, L.M., Brodie, E.L., Hettich, R.L., 2010. Direct Cellular Lysis/Protein Extraction Protocol for Soil Metaproteomics. Journal of Proteome Research 9, 6615–6622. doi:10.1021/pr100787q

Clemente, J.S., Simpson, M.J., Simpson, A.J., Yanni, S.F., Whalen, J.K., 2013. Comparison of soil organic matter composition after incubation with maize leaves, roots, and stems. Geoderma 192, 86–96. doi:10.1016/j.geoderma.2012.08.007

Cline, L.C., Zak, D.R., 2015. Soil microbial communities are shaped by plant-driven changes in resource availability during secondary succession. Ecology 96, 3374–3385. doi:10.1890/15-0184.1

Ekschmitt, K., Kandeler, E., Poll, C., Brune, A., Buscot, F., Friedrich, M., Gleixner, G., Hartmann, A., Kästner, M., Marhan, S., Miltner, A., Scheu, S., Wolters, V., 2008. Soil-carbon preservation through habitat constraints and biological limitations on decomposer activity. Journal of Plant Nutrition and Soil Science 171, 27–35. doi:10.1002/jpln.200700051

Fierer, N., 2017. Embracing the unknown: disentangling the complexities of the soil microbiome. Nature Reviews Microbiology 15, 579–590. doi:10.1038/nrmicro.2017.87

García-Ortega, L.F., Martínez, O., 2015. How Many Genes Are Expressed in a Transcriptome? Estimation and Results for RNA-Seq. PLOS ONE 10, e0130262. doi:10.1371/journal.pone.0130262

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J., 2017. Microbiome Datasets Are Compositional: And This Is Not Optional. Frontiers in Microbiology 8. doi:10.3389/fmicb.2017.02224

Grandy, A.S., Neff, J.C., 2008. Molecular C dynamics downstream: The biochemical decomposition sequence and its impact on soil organic matter structure and function. Science of The Total Environment 404, 297–307. doi:10.1016/j.scitotenv.2007.11.013

Greenfield, L.M., Hill, P.W., Paterson, E., Baggs, E.M., Jones, D.L., 2018. Methodological bias associated with soluble protein recovery from soil. Scientific Reports 8. doi:10.1038/s41598-018-29559-4

Gregorich, E.G., Janzen, H., Ellert, B.H., Helgason, B.L., Qian, B., Zebarth, B.J., Angers, D.A., Beyaert, R.P., Drury, C.F., Duguid, S.D., May, W.E., McConkey, B.G., Dyck, M.F., 2017. Litter decay controlled by temperature, not soil properties, affecting future soil carbon. Global Change Biology 23, 1725–1734. doi:10.1111/gcb.13502

Grewer, D.M., Lafrenière, M.J., Lamoureux, S.F., Simpson, M.J., 2018. Spatial and temporal shifts in fluvial sedimentary organic matter composition from a High Arctic watershed impacted by localized slope disturbances. Organic Geochemistry 123, 113–125. doi:10.1016/j.orggeochem.2018.07.004

Grob, C., Taubert, M., Howat, A.M., Burns, O.J., Dixon, J.L., Richnow, H.H., Jehmlich, N., von Bergen, M., Chen, Y., Murrell, J.C., 2015. Combining metagenomics with metaproteomics and stable isotope probing reveals metabolic pathways used by a naturally occurring marine methylotroph: Metagenomics, metaproteomics and methylotrophy. Environmental Microbiology 17, 4007–4018. doi:10.1111/1462-2920.12935

Haberl, H., Erb, K.-H., Krausmann, F., 2014. Human Appropriation of Net Primary Production: Patterns, Trends, and Planetary Boundaries. Annual Review of Environment and Resources 39, 363–391. doi:10.1146/annurev-environ-121912-094620

Handa, I.T., Aerts, R., Berendse, F., Berg, M.P., Bruder, A., Butenschoen, O., Chauvet, E., Gessner, M.O., Jabiol, J., Makkonen, M., McKie, B.G., Malmqvist, B., Peeters, E.T.H.M., Scheu, S., Schmid, B., van Ruijven, J., Vos, V.C.A., Hättenschwiler, S., 2014. Consequences of biodiversity loss for litter decomposition across biomes. Nature 509, 218–221. doi:10.1038/nature13247

He, S., Wurtzel, O., Singh, K., Froula, J.L., Yilmaz, S., Tringe, S.G., Wang, Z., Chen, F., Lindquist, E.A., Sorek, R., Hugenholtz, P., 2010. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nature Methods 7, 807–812. doi:10.1038/nmeth.1507

Hernandez-Soriano, M.C., Dalal, R.C., Warren, F.J., Wang, P., Green, K., Tobin, M.J., Menzies, N.W., Kopittke, P.M., 2018. Soil Organic Carbon Stabilization: Mapping Carbon Speciation from Intact Microaggregates. Environmental Science & Technology 52, 12275–12284. doi:10.1021/acs.est.8b03095

Howe, A., Yang, F., Williams, R.J., Meyer, F., Hofmockel, K.S., 2016. Identification of the Core Set of Carbon-Associated Genes in a Bioenergy Grassland Soil. PLOS ONE 11, e0166578. doi:10.1371/journal.pone.0166578

Hultman, J., Waldrop, M.P., Mackelprang, R., David, M.M., McFarland, J., Blazewicz, S.J., Harden, J., Turetsky, M.R., McGuire, A.D., Shah, M.B., VerBerkmoes, N.C., Lee, L.H., Mavrommatis, K., Jansson, J.K., 2015. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature 521, 208–212. doi:10.1038/nature14238

Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C., 2007. MEGAN analysis of metagenomic data. Genome Research 17, 377–386. doi:10.1101/gr.5969107

Jansson, J.K., Hofmockel, K.S., 2018. The soil microbiome — from metagenomics to metaphenomics. Current Opinion in Microbiology 43, 162–168. doi:10.1016/j.mib.2018.01.013

Janzen, H.H., 2015. Beyond carbon sequestration: soil as conduit of solar energy: Soil carbon and energy. European Journal of Soil Science 66, 19–32. doi:10.1111/ejss.12194

Jenkins, S., Swenson, T.L., Lau, R., Rocha, A., Aaring, A., Hazen, T.C., Chakraborty, R., Northen, T.R., 2017. Construction of soil defined media using quantitative exometabolomic analysis of soil metabolites. BioRxiv 151282. doi:10.1101/151282

Jenkinson, D.S., Rayner, J.H., 1977. The turnover of soil organic matter in some of the Rothamsted classical experiments. Soil Science 171, S130–S137.

Kallenbach, C.M., Frey, S.D., Grandy, A.S., 2016. Direct evidence for microbial-derived soil organic matter formation and its ecophysiological controls. Nature Communications 7. doi:10.1038/ncomms13630

Keiblinger, K.M., Wilhartitz, I.C., Schneider, T., Roschitzki, B., Schmid, E., Eberl, L., Riedel, K., Zechmeister-Boltenstern, S., 2012. Soil metaproteomics – Comparative evaluation of protein extraction protocols. Soil Biology and Biochemistry 54, 14–24. doi:10.1016/j.soilbio.2012.05.014

Kim, Y., Liesack, W., 2015. Differential Assemblage of Functional Units in Paddy Soil Microbiomes. PLOS ONE 10, e0122221. doi:10.1371/journal.pone.0122221

Kindler, R., Miltner, A., Richnow, H., Kastner, M., 2006. Fate of gram-negative bacterial biomass in soil—mineralization and contribution to SOM. Soil Biology and Biochemistry 38, 2860–2870. doi:10.1016/j.soilbio.2006.04.047

Kleber, M., Sollins, P., Sutton, R., 2007. A conceptual model of organo-mineral interactions in soils: self-assembly of organic molecular fragments into zonal structures on mineral surfaces. Biogeochemistry 85, 9–24. doi:10.1007/s10533-007-9103-5

Kleiner, M., Dong, X., Hinzke, T., Wippler, J., Thorson, E., Mayer, B., Strous, M., 2018. Metaproteomics method to determine carbon sources and assimilation pathways of species in microbial communities. Proceedings of the National Academy of Sciences of the United States of America 115, E5576–E5584. doi:10.1073/pnas.1722325115

Kramer, S., Dibbern, D., Moll, J., Huenninghaus, M., Koller, R., Krueger, D., Marhan, S., Urich, T., Wubet, T., Bonkowski, M., Buscot, F., Lueders, T., Kandeler, E., 2016. Resource Partitioning between Bacteria, Fungi, and Protists in the Detritusphere of an Agricultural Soil. Frontiers in Microbiology 7. doi:10.3389/fmicb.2016.01524

Kuzyakov, Y., 2010. Priming effects: Interactions between living and dead organic matter. Soil Biology and Biochemistry 42, 1363–1371. doi:10.1016/j.soilbio.2010.04.003

Lehmann, J., Kleber, M., 2015. The contentious nature of soil organic matter. Nature 528, 60–68. doi:10.1038/nature16069

Lennon, J.T., Muscarella, M.E., Placella, S.A., Lehmkuhl, B.K., 2018. How, When, and Where Relic DNA Affects Microbial Diversity. MBio 9. doi:10.1128/mBio.00637-18

Levy-Booth, D.J., Campbell, R.G., Gulden, R.H., Hart, M.M., Powell, J.R., Klironomos, J.N., Pauls, K.P., Swanton, C.J., Trevors, J.T., Dunfield, K.E., 2007. Cycling of extracellular DNA in the soil environment. Soil Biology and Biochemistry 39, 2977–2991. doi:10.1016/j.soilbio.2007.06.020

Levy-Booth, D.J., Giesbrecht, I.J.W., Kellogg, C.T.E., Heger, T.J., D'Amore, D.V., Keeling, P.J., Hallam, S.J., Mohn, W.W., 2019. Seasonal and ecohydrological regulation of active microbial populations involved in DOC, CO_2, and CH_4 fluxes in temperate rainforest soil. The ISME Journal 13, 950–963. doi:10.1038/s41396-018-0334-3

Li, H.-Y., Wang, H., Wang, H.-T., Xin, P.-Y., Xu, X.-H., Ma, Y., Liu, W.-P., Teng, C.-Y., Jiang, C.-L., Lou, L.-P., Arnold, W., Cralle, L., Zhu, Y.-G., Chu, J.-F., Gilbert, J.A., Zhang, Z.-J., 2018. The chemodiversity of paddy soil dissolved organic matter correlates with microbial community at continental scales. Microbiome 6. doi:10.1186/s40168-018-0561-x

Li, J., Vosegaard, T., Guo, Z., 2017. Applications of nuclear magnetic resonance in lipid analyses: An emerging powerful tool for lipidomics studies. Progress in Lipid Research 68, 37–56. doi:10.1016/j.plipres.2017.09.003

Liang, C., Schimel, J.P., Jastrow, J.D., 2017. The importance of anabolism in microbial control over soil carbon storage. Nature Microbiology 2. doi:10.1038/nmicrobiol.2017.105

Lin, H., 2014. A New Worldview of Soils. Soil Science Society of America Journal 78, 1831. doi:10.2136/sssaj2014.04.0162

Liu, D., Keiblinger, K.M., Schindlbacher, A., Wegner, U., Sun, H., Fuchs, S., Lassek, C., Riedel, K., Zechmeister-Boltenstern, S., 2017. Microbial functionality as affected by experimental warming of a temperate mountain forest soil—A metaproteomics survey. Applied Soil Ecology 117–118, 196–202. doi:10.1016/j.apsoil.2017.04.021

Louca, S., Jacques, S.M.S., Pires, A.P.F., Leal, J.S., Srivastava, D.S., Parfrey, L.W., Farjalla, V.F., Doebeli, M., 2016. High taxonomic variability despite stable functional structure across microbial communities. Nature Ecology & Evolution 1, 0015. doi:10.1038/s41559-016-0015

Marschner, B., Brodowski, S., Dreves, A., Gleixner, G., Gude, A., Grootes, P.M., Hamer, U., Heim, A., Jandl, G., Ji, R., Kaiser, K., Kalbitz, K., Kramer, C., Leinweber, P., Rethemeyer, J., Schäffer, A., Schmidt, M.W.I., Schwark, L., Wiesenberg, G.L.B., 2008. How relevant is recalcitrance for the stabilization of organic matter in soils? Journal of Plant Nutrition and Soil Science 171, 91–110. doi:10.1002/jpln.200700049

Masuda, Y., Itoh, H., Shiratori, Y., Senoo, K., 2018. Metatranscriptomic insights into microbial consortia driving methane metabolism in paddy soils. Soil Science and Plant Nutrition 64, 455–464. doi:10.1080/00380768.2018.1457409

Mattarozzi, M., Manfredi, M., Montanini, B., Gosetti, F., Sanangelantoni, A.M., Marengo, E., Careri, M., Visioli, G., 2017. A metaproteomic approach dissecting major bacterial functions in the rhizosphere of plants living in serpentine soil. Analytical and Bioanalytical Chemistry 409, 2327–2339. doi:10.1007/s00216-016-0175-8

McCarthy, A., Chiang, E., Schmidt, M.L., Denef, V.J., 2015. RNA preservation agents and nucleic acid extraction method bias perceived bacterial community composition. PLoS ONE 10, e0121659. doi:10.1371/journal.pone.0121659

Mitchell, P.J., Simpson, A.J., Soong, R., Simpson, M.J., 2015. Shifts in microbial community and water-extractable organic matter composition with biochar amendment in a temperate forest soil. Soil Biology and Biochemistry 81, 244–254. doi:10.1016/j.soilbio.2014.11.017

Morales, S.E., Holben, W.E., 2011. Linking bacterial identities and ecosystem processes: can 'omic' analyses be more than the sum of their parts?: Toward linking bacterial identities and ecosystem processes. FEMS Microbiology Ecology 75, 2–16. doi:10.1111/j.1574-6941.2010.00938.x

Morrissey, E.M., Mau, R.L., Schwartz, E., Caporaso, J.G., Dijkstra, P., Gestel, N. van, Koch, B.J., Liu, C.M., Hayer, M., McHugh, T.A., Marks, J.C., Price, L.B., Hungate, B.A., 2016. Phylogenetic organization of bacterial activity. The ISME Journal 10, 2336–2340. doi:10.1038/ismej.2016.28

Morrissey, E.M., Mau, R.L., Schwartz, E., McHugh, T.A., Dijkstra, P., Koch, B.J., Marks, J.C., Hungate, B.A., 2017. Bacterial carbon use plasticity, phylogenetic diversity and the priming of soil organic matter. The ISME Journal 11, 1890–1899. doi:10.1038/ismej.2017.43

Müller, K., Marhan, S., Kandeler, E., Poll, C., 2017. Carbon flow from litter through soil microorganisms: From incorporation rates to mean residence times in bacteria and fungi. Soil Biology and Biochemistry 115, 187–196. doi:10.1016/j.soilbio.2017.08.017

Nelson, M.B., Martiny, A.C., Martiny, J.B.H., 2016. Global biogeography of microbial nitrogen-cycling traits in soil. Proceedings of the National Academy of Sciences 113, 8033–8040. doi:10.1073/pnas.1601070113

Nesme, J., Achouak, W., Agathos, S.N., Bailey, M., Baldrian, P., Brunel, D., Frostegård, Å., Heulin, T., Jansson, J.K., Jurkevitch, E., Kruus, K.L., Kowalchuk, G.A., Lagares, A., Lappin-Scott, H.M., Lemanceau, P., Le Paslier, D., Mandic-Mulec, I., Murrell, J.C., Myrold, D.D., Nalin, R., Nannipieri, P., Neufeld, J.D., O'Gara, F., Parnell, J.J., Pühler, A., Pylro, V., Ramos, J.L., Roesch, L.F.W., Schloter, M., Schleper, C., Sczyrba, A., Sessitsch, A., Sjöling, S., Sørensen, J., Sørensen, S.J., Tebbe, C.C., Topp, E., Tsiamis, G., van Elsas, J.D., van Keulen, G., Widmer, F., Wagner, M., Zhang, T., Zhang, X., Zhao, L., Zhu, Y.-G., Vogel, T.M., Simonet, P., 2016. Back to the Future of Soil Metagenomics. Frontiers in Microbiology 7. doi:10.3389/fmicb.2016.00073

Nicora, C.D., Anderson, B.J., Callister, S.J., Norbeck, A.D., Purvine, S.O., Jansson, J.K., Mason, O.U., David, M.M., Jurelevicius, D., Smith, R.D., Lipton, M.S., 2013. Amino acid treatment enhances protein recovery from sediment and soils for metaproteomic studies. PROTEOMICS n/a-n/a. doi:10.1002/pmic.201300003

Nivelle, E., Verzeaux, J., Habbib, H., Kuzyakov, Y., Decocq, G., Roger, D., Lacoux, J., Duclercq, J., Spicher, F., Nava-Saucedo, J.-E., Catterou, M., Dubois, F., Tetu, T., 2016. Functional response of soil microbial communities to tillage, cover crops and nitrogen fertilization. Applied Soil Ecology 108, 147–155. doi:10.1016/j.apsoil.2016.08.004

O'Brien, S.L., Gibbons, S.M., Owens, S.M., Hampton-Marcell, J., Johnston, E.R., Jastrow, J.D., Gilbert, J.A., Meyer, F., Antonopoulos, D.A., 2016. Spatial scale drives patterns in soil bacterial diversity. Environmental Microbiology 18, 2039–2051. doi:10.1111/1462-2920.13231

Orellana, L.H., Chee-Sanford, J.C., Sanford, R.A., Löffler, F.E., Konstantinidis, K.T., 2018. Year-Round Shotgun Metagenomes Reveal Stable Microbial Communities in Agricultural Soils and Novel Ammonia Oxidizers Responding to Fertilization. Applied and Environmental Microbiology 84, e01646-17. doi:10.1128/AEM.01646-17

Papp, K., Mau, R.L., Hayer, M., Koch, B.J., Hungate, B.A., Schwartz, E., 2018. Quantitative stable isotope probing with H 2 18 O reveals that most bacterial taxa in soil synthesize new ribosomal RNA. The ISME Journal 12, 3043–3045. doi:10.1038/s41396-018-0233-7

Pautler, B.G., Sanborn, P.T., Simpson, A.J., Simpson, M.J., 2013. Molecular characterization of organic matter in Canadian Arctic paleosols for paleoecological applications. Organic Geochemistry 63, 122–138.

Pelletier, N., Audsley, E., Brodt, S., Garnett, T., Henriksson, P., Kendall, A., Kramer, K.J., Murphy, D., Nemecek, T., Troell, M., 2011. Energy Intensity of Agriculture and Food Systems. Annual Review of Environment and Resources 36, 223–246. doi:10.1146/annurev-environ-081710-161014

Penton, C.R., Gupta, V.V.S.R., Yu, J., Tiedje, J.M., 2016. Size Matters: Assessing Optimum Soil Sample Size for Fungal and Bacterial Community Structure Analyses Using High Throughput Sequencing of rRNA Gene Amplicons. Frontiers in Microbiology 7. doi:10.3389/fmicb.2016.00824

Peršoh, D., Theuerl, S., Buscot, F., Rambold, G., 2008. Towards a universally adaptable method for quantitative extraction of high-purity nucleic acids from soil. Journal of Microbiological Methods 75, 19–24. doi:10.1016/j.mimet.2008.04.009

Pisani, O., Frey, S.D., Simpson, A.J., Simpson, M.J., 2015. Soil warming and nitrogen deposition alter soil organic matter composition at the molecular-level. Biogeochemistry 123, 391–409.

Pisani, O., Haddix, M.L., Conant, R.T., Paul, E.A., Simpson, M.J., 2016. Molecular composition of soil organic matter with land-use change along a bi-continental mean annual temperature gradient. Science of The Total Environment 573, 470–480.

Poll, C., Ingwersen, J., Stemmer, M., Gerzabek, M.H., Kandeler, E., 2006. Mechanisms of solute transport affect small-scale abundance and function of soil microorganisms in the detritusphere. European Journal of Soil Science 57, 583–595. doi:10.1111/j.1365-2389.2006.00835.x

Prosser, J.I., 2015. Dispersing misconceptions and identifying opportunities for the use of “omics” in soil microbial ecology. Nature Reviews Microbiology 13, 439–446. doi:10.1038/nrmicro3468

Raes, J., Letunic, I., Yamada, T., Jensen, L.J., Bork, P., 2011. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. Molecular Systems Biology 7, 473–473. doi:10.1038/msb.2011.6

Rochfort, S., Ezernieks, V., Mele, P., Kitching, M., 2015. NMR metabolomics for soil analysis provide complementary, orthogonal data to MIR and traditional soil chemistry approaches – a land use study. Magnetic Resonance in Chemistry 53, 719–725. doi:10.1002/mrc.4187

Sarkar, B., Singh, M., Mandal, S., Churchman, G.J., Bolan, N.S., 2018. Clay Minerals—Organic Matter Interactions in Relation to Carbon Stabilization in Soils, in: The Future of Soil Carbon. Elsevier Inc., London, United Kingdom, pp. 71–83.

Sauerschnig, C., Doppler, M., Bueschl, C., Schuhmacher, R., 2017. Methanol Generates Numerous Artifacts during Sample Extraction and Storage of Extracts in Metabolomics Research. Metabolites 8, 1. doi:10.3390/metabo8010001

Schimel, J.P., Schaeffer, S.M., 2012. Microbial control over carbon cycling in soil. Frontiers in Microbiology 3. doi:10.3389/fmicb.2012.00348

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F., 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Applied and Environmental Microbiology 75, 7537–7541. doi:10.1128/AEM.01541-09

Schmidt, M.W.I., Torn, M.S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I.A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D.A.C., Nannipieri, P., Rasse, D.P., Weiner, S., Trumbore, S.E., 2011. Persistence of soil organic matter as an ecosystem property. Nature 478, 49–56. doi:10.1038/nature10386

Schweigert, M., Herrmann, S., Miltner, A., Fester, T., Kästner, M., 2015. Fate of ectomycorrhizal fungal biomass in a soil bioreactor system and its contribution to soil organic matter formation. Soil Biology and Biochemistry 88, 120–127. doi:10.1016/j.soilbio.2015.05.012

Seifert, A.-G., Roth, V.-N., Dittmar, T., Gleixner, G., Breuer, L., Houska, T., Marxsen, J., 2016. Comparing molecular composition of dissolved organic matter in soil and stream water: Influence of land use and chemical characteristics. Science of The Total Environment 571, 142–152. doi:10.1016/j.scitotenv.2016.07.033

Seifert, J., Taubert, M., Jehmlich, N., Schmidt, F., Völker, U., Vogt, C., Richnow, H., von Bergen, M., 2012. Protein-based stable isotope probing (protein-SIP) in functional metaproteomics. Mass Spectrometry Reviews 31, 683–697. doi:10.1002/mas.21346

Sharon, I., Banfield, J.F.B., 2013. Genomes from Metagenomics. Science 342, 1057–1058. doi:10.1126/science.1247023

Sidibé, A., Simao-Beaunoir, A.-M., Lerat, S., Giroux, L., Toussaint, V., Beaulieu, C., 2016. Proteome Analyses of Soil Bacteria Grown in the Presence of Potato Suberin, a Recalcitrant Biopolymer. Microbes and Environments 31, 418–426. doi:10.1264/jsme2.ME15195

Silva, L.P., Northen, T.R., 2015. Exometabolomics and MSI: deconstructing how cells interact to transform their small molecule environment. Current Opinion in Biotechnology 34, 209–216. doi:10.1016/j.copbio.2015.03.015

Sipos, R., Székely, A.J., Palatinszky, M., Révész, S., Márialigeti, K., Nikolausz, M., 2007. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiology Ecology 60, 341–350. doi:10.1111/j.1574-6941.2007.00283.x

Six, J., Bossuyt, H., Degryze, S., Denef, K., 2004. A history of research on the link between (micro)aggregates, soil biota, and soil organic matter dynamics. Soil and Tillage Research 79, 7–31. doi:10.1016/j.still.2004.03.008

Six, J., Frey, S.D., Thiet, R.K., Batten, K.M., 2006. Bacterial and Fungal Contributions to Carbon Sequestration in Agroecosystems. Soil Science Society of America Journal 70, 555. doi:10.2136/sssaj2004.0347

Souza, R.C., Hungria, M., Cantão, M.E., Vasconcelos, A.T.R., Nogueira, M.A., Vicente, V.A., 2015. Metagenomic analysis reveals microbial functional redundancies and specificities in a soil under different tillage and crop-management regimes. Applied Soil Ecology 86, 106–112. doi:10.1016/j.apsoil.2014.10.010

Starke, R., Jehmlich, N., Bastida, F., 2019. Using proteins to study how microbes contribute to soil ecosystem services: The current state and future perspectives of soil metaproteomics. Journal of Proteomics, 10 Year Anniversary of Proteomics 198, 50–58. doi:10.1016/j.jprot.2018.11.011

Sun, R., Li, W., Dong, W., Tian, Y., Hu, C., Liu, B., 2018. Tillage Changes Vertical Distribution of Soil Bacterial and Fungal Communities. Frontiers in Microbiology 9. doi:10.3389/fmicb.2018.00699

Swenson, T.L., Jenkins, S., Bowen, B.P., Northen, T.R., 2015. Untargeted soil metabolomics methods for analysis of extractable organic matter. Soil Biology and Biochemistry 80, 189–198. doi:10.1016/j.soilbio.2014.10.007

Swenson, T.L., Karaoz, U., Swenson, J.M., Bowen, B.P., Northen, T.R., 2018. Linking soil biology and chemistry in biological soil crust using isolate exometabolomics. Nature Communications 9. doi:10.1038/s41467-017-02356-9

Tatangelo, V., Franzetti, A., Gandolfi, I., Bestetti, G., Ambrosini, R., 2014. Effect of preservation method on the assessment of bacterial community structure in soil and water samples. FEMS Microbiology Letters 356, 32–38. doi:10.1111/1574-6968.12475

Tfaily, M.M., Chu, R.K., Tolić, N., Roscioli, K.M., Anderton, C.R., Paša-Tolić, L., Robinson, E.W., Hess, N.J., 2015. Advanced Solvent Based Methods for Molecular Characterization of Soil Organic Matter by High-Resolution Mass Spectrometry. Analytical Chemistry 87, 5206–5215. doi:10.1021/acs.analchem.5b00116

Tfaily, M.M., Chu, R.K., Toyoda, J., Tolić, N., Robinson, E.W., Paša-Tolić, L., Hess, N.J., 2017. Sequential extraction protocol for organic matter from soils and sediments using high resolution mass spectrometry. Analytica Chimica Acta 972, 54–61. doi:10.1016/j.aca.2017.03.031

Turner, T.R., Ramakrishnan, K., Walshaw, J., Heavens, D., Alston, M., Swarbreck, D., Osbourn, A., Grant, A., Poole, P.S., 2013. Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. The ISME Journal 7, 2248–2258. doi:10.1038/ismej.2013.119

Tveit, A.T., Urich, T., Svenning, M.M., 2014. Metatranscriptomic Analysis of Arctic Peat Soil Microbiota. Applied and Environmental Microbiology 80, 5761–5772. doi:10.1128/AEM.01030-14

Urich, T., Lanzén, A., Qi, J., Huson, D.H., Schleper, C., Schuster, S.C., 2008. Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome. PLoS ONE 3, e2527. doi:10.1371/journal.pone.0002527

von Bergen, M., Jehmlich, N., Taubert, M., Vogt, C., Bastida, F., Herbst, F.-A., Schmidt, F., Richnow, H.-H., Seifert, J., 2013. Insights from quantitative metaproteomics and protein-stable isotope probing into microbial ecology. The ISME Journal 7, 1877–1885. doi:10.1038/ismej.2013.78

Wang, H.-B., Zhang, Z.-X., Li, H., He, H.-B., Fang, C.-X., Zhang, A.-J., Li, Q.-S., Chen, R.-S., Guo, X.-K., Lin, H.-F., Wu, L.-K., Lin, S., Chen, T., Lin, R.-Y., Peng, X.-X., Lin, W.-X., 2011. Characterization of Metaproteomics in Crop Rhizospheric Soil. Journal of Proteome Research 10, 932–940. doi:10.1021/pr100981r

Wang, Y., Hayatsu, M., Fujii, T., 2012. Extraction of Bacterial RNA from Soil: Challenges and Solutions. Microbes and Environments 27, 111–121. doi:10.1264/jsme2.ME11304

Ward, C.P., Cory, R.M., 2015. Chemical composition of dissolved organic matter draining permafrost soils. Geochimica et Cosmochimica Acta 167, 63–79. doi:10.1016/j.gca.2015.07.001

Warren, C.R., 2014. Response of osmolytes in soil to drying and rewetting. Soil Biology and Biochemistry 70, 22–32. doi:10.1016/j.soilbio.2013.12.008

Wickings, K., Stuart Grandy, A., Reed, S., Cleveland, C., 2011. Management intensity alters decomposition via biological pathways. Biogeochemistry 104, 365–379. doi:10.1007/s10533-010-9510-x

Williams, M.A., Taylor, E.B., Mula, H.P., 2010. Metaproteomic characterization of a soil microbial community following carbon amendment. Soil Biology and Biochemistry 42, 1148–1156. doi:10.1016/j.soilbio.2010.03.021

Williamson, Kurt.E., Kan, J., Polson, S.W., Williamson, S.J., 2011. Optimizing the indirect extraction of prokaryotic DNA from soils. Soil Biology and Biochemistry 43, 736–748. doi:10.1016/j.soilbio.2010.04.017

Wu, X., Wu, L., Liu, Y., Zhang, P., Li, Q., Zhou, J., Hess, N.J., Hazen, T.C., Yang, W., Chakraborty, R., 2018. Microbial Interactions With Dissolved Organic Matter Drive Carbon Dynamics and Community Succession. Frontiers in Microbiology 9, 1234. doi:10.3389/fmicb.2018.01234

Zhang, A., Sun, H., Wang, P., Han, Y., Wang, X., 2012. Modern analytical techniques in metabolomics analysis. The Analyst 137, 293–300. doi:10.1039/C1AN15605E

Zhao, J., Ni, T., Li, J., Lu, Q., Fang, Z., Huang, Q., Zhang, R., Li, R., Shen, B., Shen, Q., 2016. Effects of organic–inorganic compound fertilizer with reduced chemical fertilizer application on crop yields, soil biological activity and bacterial community structure in a rice–wheat cropping system. Applied Soil Ecology 99, 1–12. doi:10.1016/j.apsoil.2015.11.006

APPENDIX 3: A comparison of alternate modelling methods to assess short-term soil respiration incubation data

Madison A. Bell[1,2], Ulrica McKim[1], Amanda Sproule[1], Ryan Tobalt, Ed G. Gregorich[1*], David P. Overy[1*]

[1]Agriculture and Agri-food Canada, Ottawa Research and Development Centre, Ottawa, ON, K1A 0C6, Canada

[2]Laboratory for the Analysis of Natural and Synthetic Environmental Toxicants, Department of Biology, University of Ottawa, Ottawa, ON, K1N 6N5, Canada

*Corresponding Author(s)

David Overy (david.overy@canada.ca) & Ed Gregorich (ed.gregorich@canada.ca)

STATEMENT OF CONTRIBUTIONS

Madison A. Bell: Conceptualization, Methodology, Software - Scripting, Formal analysis, Data curation, Visualization, Writing – Original draft preparation. Ulrica McKim: Methodology, Investigation. Amanda Sproule: Methodology, Investigation. Ryan Tobalt: Investigation. Ed G. Gregorich: Methodology, Funding acquisition, Project administration, Supervision, Writing – Review & Editing. David P. Overy: Methodology, Supervision, Funding acquisition, Project administration, Writing – Review & Editing. This chapter consists of extra data from CHAPTER FIVE and this work has not been published or submitted:

Bell, M.A., McKim, U., Sproule, A., Tobalt, R., Gregorich, E., Overy, D.P. A comparison of alternate modelling methods to assess short-term soil respiration incubation data

ABSTRACT

Kinetic models are used to model short-term CO_2 respiration data produced in soil incubation studies, but use of these models can result in inaccurate rate estimates. Our objective in this study was to compare alternate modelling methods for CO_2 respiration data that are more robust than kinetic models and which produce coherent parameters for the investigation of soil CO_2 respiration. We modelled data from a pre-existing dataset from a short-term incubation study and found that a weighted Gompertz model was the best performing model that produced sensical soil accumulation rates and saturation constants when comparing different soil types and treatments. Dimensionality reduction in the form of a PCA was also tested on the same dataset, and it was determined that a PCA analysis of incubation data could also be coherently interpreted and used to illustrate soil CO_2 accumulation and mineralization data. This work is significant because it provides a method for using alternative modelling strategies of soil incubation data that can be applied to large datasets, they are more robust, and the parameters generated from these models could be used in future soil dynamics modelling.

1. INTRODUCTION

Soil respiration of greenhouse gases (i.e. CO_2, etc.) is influenced by many abiotic and biotic factors (Conant et al., 2011; Schmidt et al., 2011). The interactions between these abiotic and biotic factors are poorly understood despite more than half a century of research, and as a result soil

respiration rates are often misrepresented in current soil dynamics models based on soil incubation experiments (Dwivedi et al., 2019; Schädel et al., 2020; Sierra et al., 2015; Weihermüller et al., 2018).

Many different approaches are by researchers to model data derived from soil incubation experiments; kinetic models are the most commonly applied model (Dwivedi et al., 2019; Johnsen et al., 2013; Schädel et al., 2020). Kinetic models allow the simulation of more than one pool of carbon (i.e. labile carbon pools and refractory pools), and permit the derivation of rate constants or half-lives for each pool, which can be used in complex soil dynamics models (Dwivedi et al., 2019; Weihermüller et al., 2018). However, Weihermüller et al. (2018) identified several problems with kinetic models, such as: poor interoperability (i.e. results cannot be compared between models and other data), over-fitting (i.e. where a simper model with fewer parameters describes measured data as well as a more complex model), non-unique models (i.e. where the model may have multiple solutions for the same data), and the generation of unrealistic rate constants. First-order kinetic models require that half-life is independent of time and concentration. If microbial density changes over time then first-order kinetic models may fail to model nonlinear responses (Johnsen et al., 2013).

Along with kinetic models, soil respiration data can also be modelled using fitted nonlinear functions and multivariate dimensionality reduction. Fitted nonlinear functions are more dynamic than linear functions or kinetic models, and are therefore more able to fit soil respiration curves resulting from diverse conditions while making fewer assumptions (Paine et al., 2012). Dimensionality reduction methods like principal components analysis (PCA), are not commonly applied to soil respiration data (Johnsen et al., 2013), but have proven useful for multivariate modelling of time series data in climate, physiology, and finance applications (Daffertshofer et al., 2004; Hargan et al., 2018; Vitanov et al., 2008).

The following work compares the performance of several different saturating sigmoidal functions (Baranyi, Richards, Gompertz, modified Gompertz, Huang, and Logistic) to model trends in soil respiration data following perturbation of a microbial population in a short-term laboratory incubation study. The best model was chosen based on goodness-of-fit estimates, parameter standard errors and p-values, and the avoidance of over-fitting. Estimated model parameters are discussed in terms of mineralization parameters in the context of the three different soil types and several soil sterilization/deactivation treatments (i.e. fresh, air-dried, autoclaved, baked, and cold-incubated). Lastly, a principal components analysis (PCA) was performed as an independent method to describe and visualize short term soil incubation data. This work is significant as the applied modelling techniques are robust and adaptable methods for investigating soil respiration and support the generation of accurate short-term soil dynamics models.

2. METHODS

2.1. Soil collection and descriptions

2.1. Soil Incubations

Three soil types (Grenville loam, Granby soil, and Rideau clay) were used with five different treatments (i.e. fresh untreated soil, air-dried soil, autoclaved soil, baked soil, and cold incubated soil), which were chosen to determine the effect different perturbations have on cumulative, short-term soil respiration and mineralization rates. The top 15 cm of soil were collected from different plots at the Central Experimental Farm (CEF) at Agriculture and Agri-food Canada (AAFC; Ottawa, Canada) in 2019 (Table SA3.1). Bulk samples were sieved (4 mm), subsampled, and refrigerated until required. Air-dried soil was dried at room temperature until incubation (1 week). Soil was autoclaved at 121°C for 30 min

(liquid cycle; 16-17psi), then left in a dark biosafety cabinet for 60 hours, mixed, and autoclaved again to kill any germinated microbes. Soil was baked as bulk samples in glass jars covered with aluminum foil at 120°C for 2 hours, left in the oven for 60 hours in the oven, mixed, and then baked again.

Wide-mouth mason jars (1L) were sterilized by baking at 120°C for 2 hours. Lids were sterilized with 70% ethanol. Soil subsamples were activated with ~10 g of autoclaved LCMS grade H_2O. Soils treatments were further subsampled into mason jars (25 g soil) with four replicates per treatment per soil type. The jars were sealed, flushed with CO_2 free air (4 min), and internal pressures adjusted to ~0.6 kPa. All jars were incubated at 25°C, except for cold treated samples, which were incubated at 4°C. Soil CO_2 respiration was measured at 0.75, 2, 3, 4, 7, and 14 days using a LI-7000 CO_2/H_2O Analyzer (LI-COR). Calibration curves were used to calculate the μg CO_2-C/ g of soil evolved, the mineralization rate (μg C/g of soil/day), and the cumulative μg CO_2-C / g of soil.

2.2. Data analysis

All data were analyzed using RStudio (v1.2.1335) and R (v3.6.0) unless otherwise stated. The data were cleaned and processed using the R packages *tidyverse* (Wickham et al., 2019), and *janitor* (Firke, 2020). Plots were produced using *ggplot2* (Wickham, 2016)*, ggpubr* (Kassambara, 2020), and a colorblind palette provided from *RColorBrewer* .

CO_2 respiration was used to calculate mineralization rate (μg C/g of soil/day), (Figure SA3.1), and the cumulative μg CO_2-C / g of soil. The cumulative CO_2 curves were visually nonlinear. Nonlinear regressions were performed using R's *growthrates* (Petzoldt, 2019) and *minpack.lm* (Elzhov et al., 2016). R *growthrates* (Petzoldt, 2019) was used to compare the fit of different growth rate models with asymptotes and provide initial starting values. Several growth rate models were applied, and formulas

are listed in Table SA3.3. After the initial fitting with *growthrates*, data were determined to be heteroscedastic (i.e. samples in each time point had increasing variance over time) and further modelling was done in *minpack.lm* to accommodate weighted nonlinear regressions. Weights were the cumulative μg CO_2-C / g of soil values divided by their variance at each time interval for all treatments and soil types. Model fits were assessed by comparing AIC, BIC, sigma, and parameter estimate errors and significance (Table SA3.4 and Figures A3.1, S A3.2). The best model was further analyzed using the R package *propagate* (Spiess, 2018) to generate confidence regions for each model using second-order (hessian) Taylor expansion. Overlapping confidence regions were then used to determine whether there are significant differences between treatments, and the final weighted Gompertz model parameters with standard errors are in Table SA3.2.

The principal components analysis (PCA) was performed similarly to the analysis performed by Hargan et al. (2018). The PCA was implemented with the function *prcomp* from the base R package *stats*. Replicates were averaged and the data were centred and scaled. Plots were generated using *factoextra* (Kassambara and Mundt, 2017) and *geom_path()* from the *ggplot2* package. *Geom_path()* was used to connect sample points starting from Day 1 sampling to Day 14 sampling in order to monitor how the data variance changed over time. Biplots had *coord_fixed(1)* to maintain correct distances.

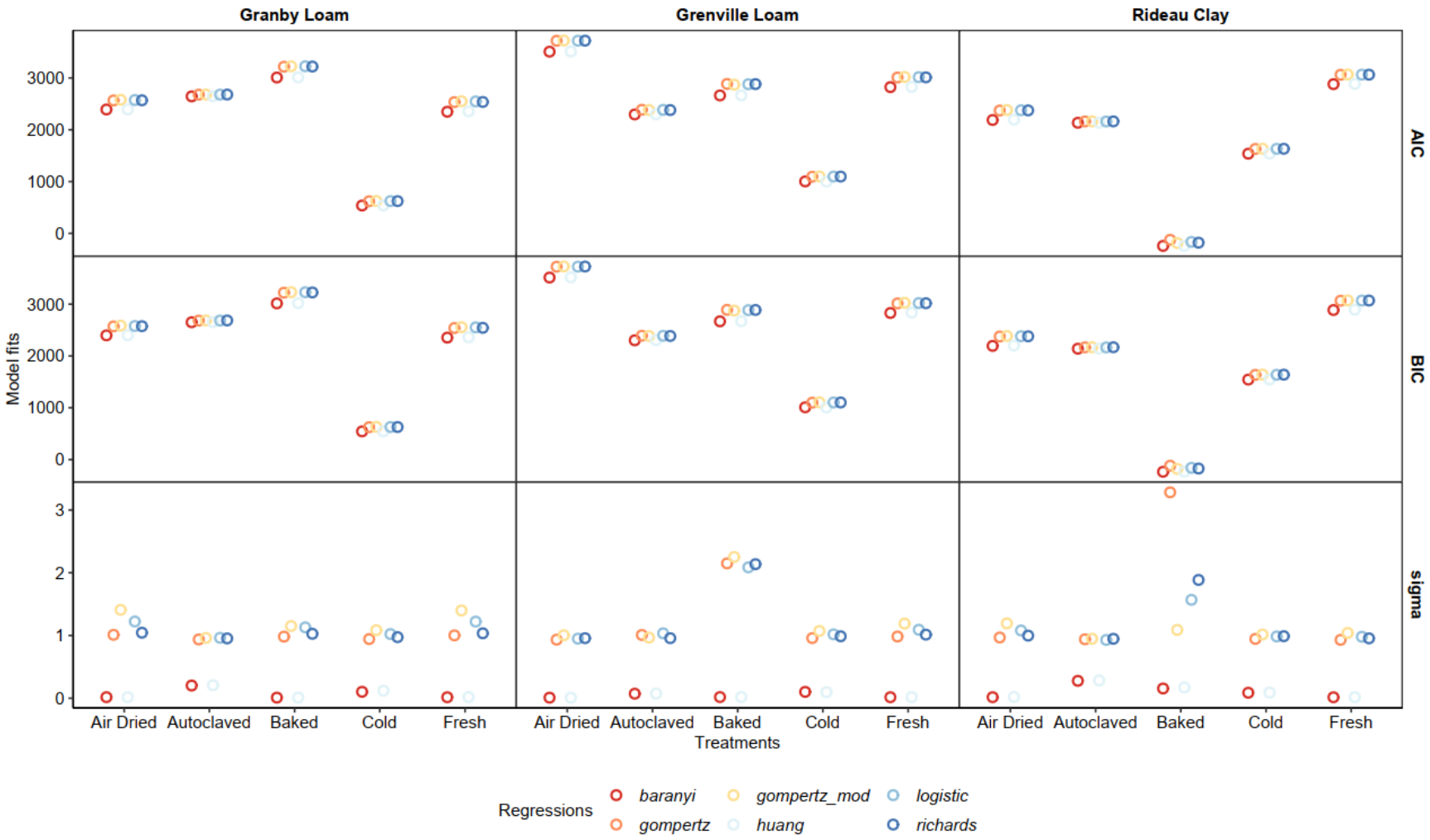


Figure A3.1: The model fit parameters (i.e. AIC, BIC, and sigma) for all models, treatments, and soil types.

3. RESULTS

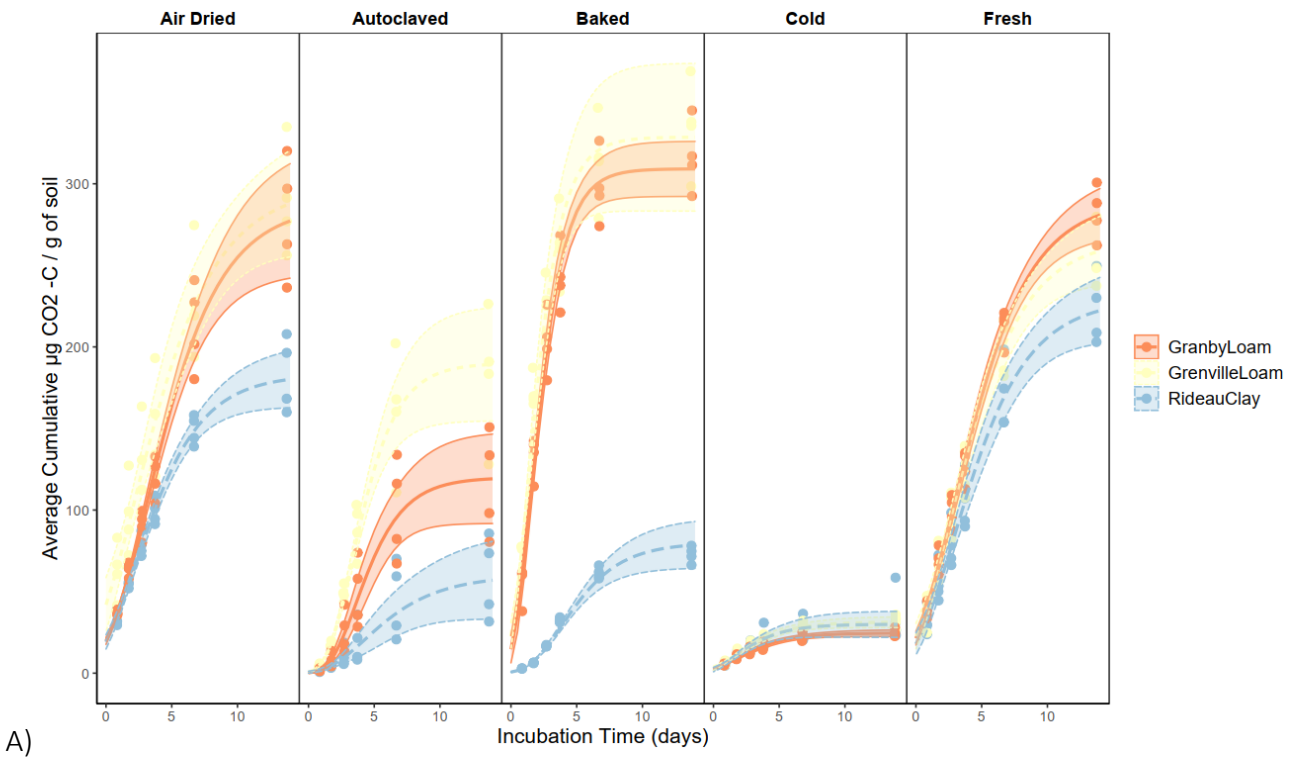
3.1. Model comparison

Several sigmoidal functions were chosen to model cumulative respiration: 1) Logistic 2) Gompertz (Tsoularis and Wallace, 2002), 3) modified Gompertz (Tsoularis and Wallace, 2002; Zwietering et al., 1990), 4) Richards (Richards, 1959; Tsoularis and Wallace, 2002), 5) Baranyi (Baranyi and Roberts, 1995, 1994), and 6) Huang (Huang, 2011, 2008). The common parameters define the y-intercept (y0), the growth coefficient (mumax), and the asymptote (K) (Tjørve and Tjørve, 2017). The modified Gompertz, Richards, Baranyi, and Huang models have varying additional parameters that define lag times (lambda, H0), and curve shape modifiers (alpha, beta). These sigmoidal functions were chosen because they are commonly applied to bacterial growth curves (Zwietering et al., 1990), and the Gompertz and Richards function has been applied to soil respiration data previously in the context of pesticide degradation studies (Johnsen et al., 2013). All of these models assume that growth is asymptotic (i.e. will reach a steady state) (Paine et al., 2012).

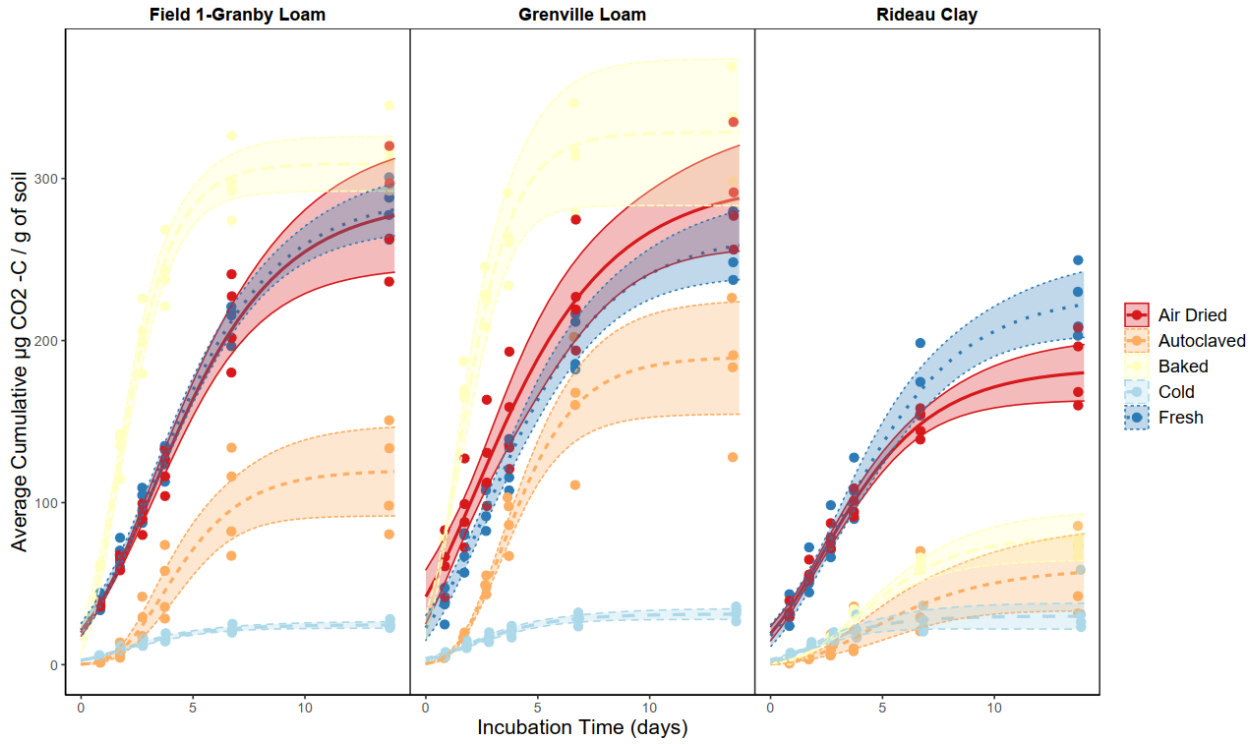
Initial model residuals indicated heteroscedasticity, a common problem for repeated measures growth data (Paine et al., 2012), so weighted non-linear models were built to increase the accuracy of the model estimates (Al turk, 2018). After weighting, model fits were assessed using AIC, BIC, sigma, and parameter p-values for each treatment and soil type (Figure A3.1). Typically, Baranyi and Huang had the lowest AIC, BIC, and sigma values indicating the best model fits. However, when the parameter p-values were assessed many of the additional parameters for Baranyi (H0), Huang (alpha and lambda), Richards (beta), and the modified gompertz (lambda) were insignificant to the models. Thus, in order to avoid over-fitting models, models without significant p-values were excluded, leaving only Gompertz and logistic models for comparison because all of their parameters were significant for most treatments and

soil types. When comparing only Gompertz and Logistic models they performed similarly, however the Gompertz model had lower sigma values for some treatment/soil type groups (Figure A3.1), and also smaller standard errors on parameter estimates (Figure SA3.2). The Gompertz model was chosen as the best model for modelling cumulative soil respiration.

Error propagation on the modelled data was used to calculate the confidence intervals for the fitted values with second-order Taylor expansion (Mekid and Vaja, 2008; Spiess, 2018; Wang and Iyer, 2005), which are visualized in Figure A3.2, and can be used to determine if there are statistically significant differences. The confidence regions provide a visual representation of increasing error as the incubation, which is common in repeated measure experiments, because each jar will have different respiration rates, and this variation is compounded over time when assessing cumulative respirations. The confidence regions also allow for the comparison of the curves and overlapping regions indicate an insignificant difference between treatments in a soil type.



A)



B)

Figure A3.2: Cumulative respiration plots over time (days). Points represent the actual values, while the lines are the modelled weighted Gompertz curve. Confidence intervals were predicted using Taylor expansion. A) Compares soil types within each treatment, while B) compares treatments within each soil type.

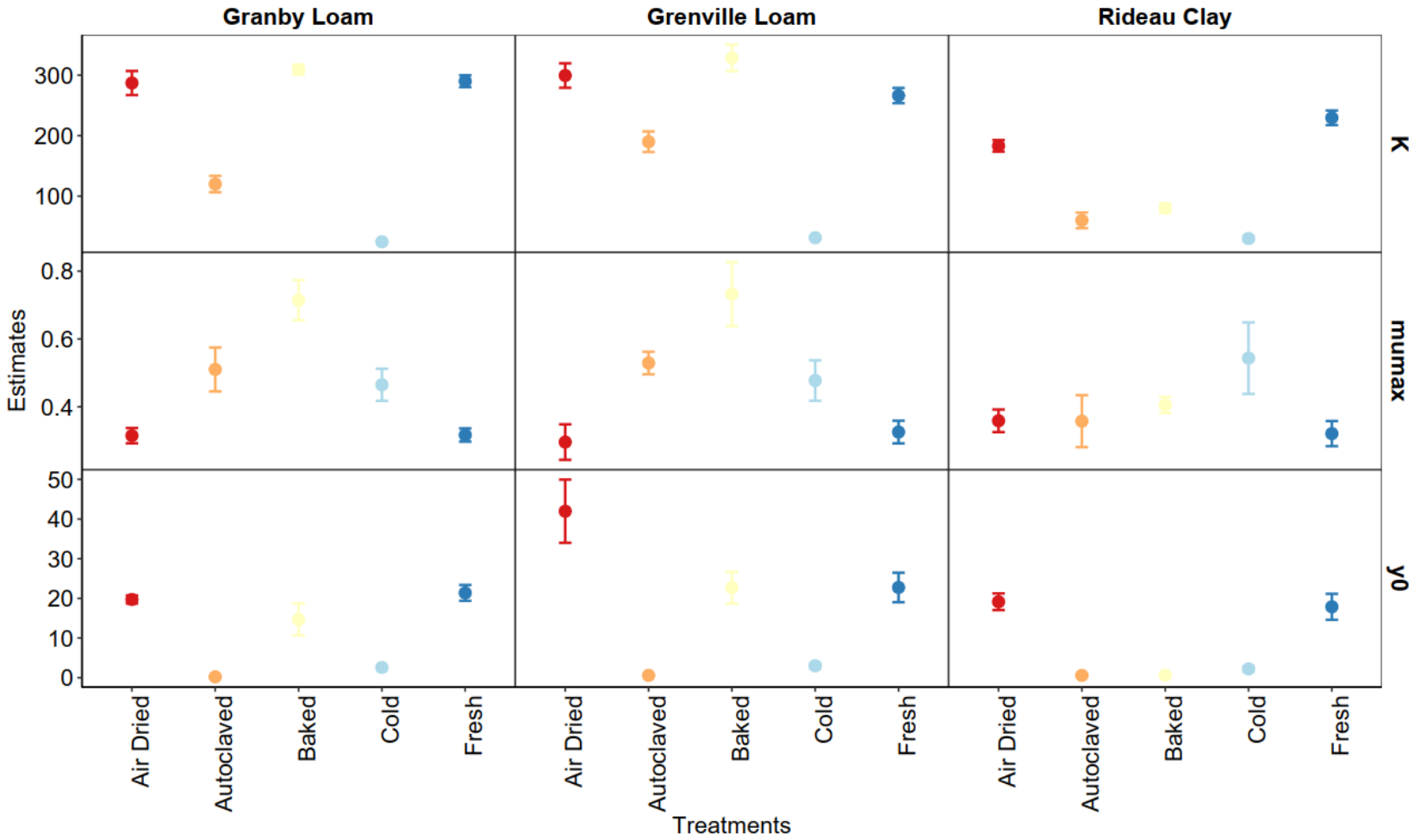


Figure A3.3: Scatter plot of predicted estimate values for K, mumax, and y0 for all treatments and soil types. Error bars were the standard error of the estimate reported by the weighted nonlinear Gompertz model.

3.2. Comparison of soil types

The three soils were chosen because they are broadly representative of different soil types. Grenville loam (gs) is an orthic melanic brunisol (64% sand, 26% silt, and 10% clay) and has the most organic carbon and nitrogen by percent. Rideau clay (rc) is a gleyed melanic brunisol (50% sand, 20% silt, and 30% clay) and has the highest pH and specific surface area (m^2/g). Granby soil (fo) is an orthic humic gleysol (73% sand, 17% silt, and 10% clay) and has the lowest organic carbon and nitrogen by percent. Fo also has the lowest pH and specific surface area.

Soil type had a significant impact on model estimates in the control (untreated) soil. In Figure A3.2A the fresh soils (i.e. control soils) for Granby Loam and Rideau Clay have non-overlapping confidence intervals indicating that, without treatment, their soil respiration differed. Figure A3.3 indicates soil type impacted the cumulative respiration maximum (K) because there is no overlap of standard errors in fresh soil. The maximum respiration was highest in Granby Loam followed by Grenville Loam then Rideau clay. The other parameters, growth coefficient (mumax) and y-intercept (y0), were similar in fresh soils.

The PCA biplots also indicate that soil respiration differed between the three soil types. Figure A3.4A is a PCA biplot where 99.5% of the variance in the cumulative respiration data were explained in two dimensions (Dimension 1: 84.1%, Dimension 2: 15.4%). The colored paths indicate each treatment and each path starts on Day 1 and follows the time series through to Day 14. The black arrows indicate the PCA loadings, in the which the size and direction of the arrows indicate the magnitude and direction each soil type had on the individual points. In Figure A3.4A the loadings arrows indicate Granby Loam and Grenville Loam behaved similarly, while Rideau clay was different. Figure A3.4B is a PCA biplot of the mineralization rates (Figure SA3.1) where 98% of the variance was explained in two dimensions

(Dimension 1: 72.2%, Dimension 2: 25.8%). The loadings arrows in Figure A3.4B also indicates that the mineralization rates over time was the most similar between Grenville Loam and Granby Loam when compared to Rideau clay.

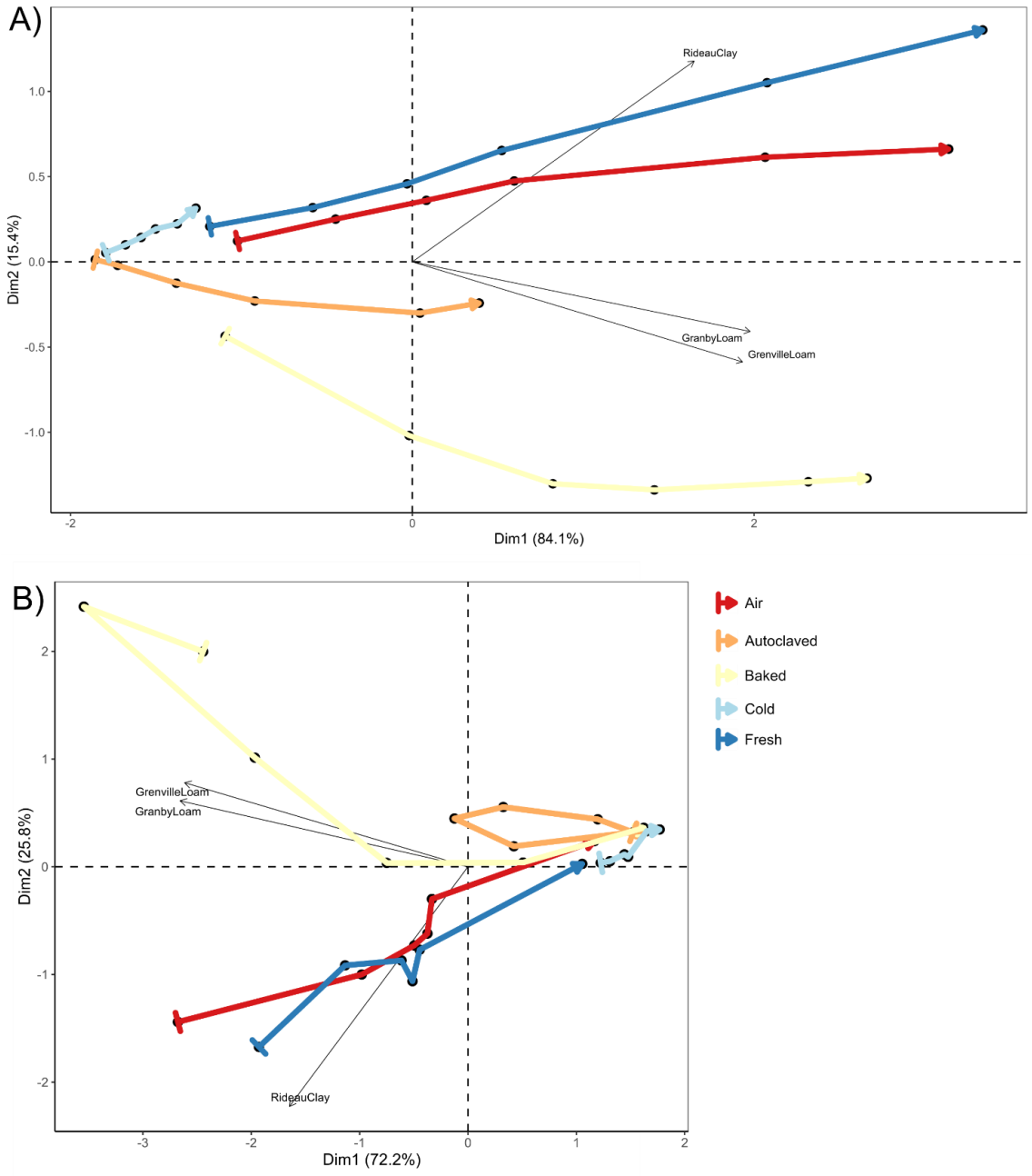


Figure A3.4: A) Cumulative soil respiration B) Mineralization rate represented in an biplot where overall soil trends over time are colored ("|" marker denotes the first measurement on Day 1; the ">" marks the end of the experiment on Day 14) and the respective effects of each treatment are indicated by the black arrows.

3.3. Comparison of soil sterilization/deactivation methods

The Gompertz function successfully modelled cumulative soil respiration from different soils subjected to different treatments. The Gompertz model indicated there are significant differences between treatments in the same soil (Figure A3.2B). Specifically, different soil sterilization (i.e. autoclaving, baking) and deactivation (i.e. cold) procedures significantly affected all model parameters K, mumax, and y0 (Figure A3.3). The cumulative respiration maximum (K) was the most influences by treatment where cold incubated samples had significantly lower cumulative respiration maximums than the other treatments, and this was consistent in all three soil types. Baked soils had the greatest cumulative respiration maximums in Granby Loam and Grenville Loam, but not in Rideau Clay where it was most similar to the cold incubated soils. Autoclaved soil cumulative respiration maximums were typically lower than in Fresh soils, but greater than cold incubated soils. Autoclaved cumulative respiration maximums were also more variable than was indicated in the fresh soil controls. Air dried soils cumulative respiration maximums were similar to Fresh soils except in Rideau clay where it was significant lower.

The growth coefficient parameter (mumax) is the expected growth rate of the curve. As a caveat, growth rates can be difficult to interpret when all of the cumulative respiration maximums are different, and this could be corrected by normalize the cumulative soil respiration data to 1 for each group. The K values would be the same, and differences in growth rate trends would be easier to interpret. However, in these models, most growth rates were not significantly impacted soil type except for Rideau clay in autoclaved and baked soils where the growth rates were lower than in Grenville Loam or Granby Loam. Generally, fresh soils had the lowest growth rates, which were similar to air dried soils. Cold and

autoclaved soils had similar growth rates, while baked soils had the greatest especially in Grenville Loam and Granby Loam.

The y-intercept also varied between treatments. The y-intercept indicates the initial soil respiration on Day 1. There was no Day 0 measurement. The y-intercept was typically greater in fresh, air dried, and baked soils; except for the Rideau clay baked soils which were more similar to cold and autoclave treated soils.

The PCA biplots of the cumulative soil respiration and mineralization rates also indicated the soil treatments had a significant influence on soil mineralization. In Figure A3.4A the colored arrows had a general trend of proceeding from the left to the right following the generally increasing trend of cumulative soil respiration. All of the origins are the left side, and the trend where the y-intercept was typically greater in fresh, air dried, and baked soils is represented by the right shift of the line origins. The lines then spread out from the originating area which correlates with the growth rates, where autoclaved and baked soil treatments had similar impacts on growth rates in Grenville and Granby Loams. Autoclaved soils also typically had lower cumulative respiration maximums than baked soils, and this is represented by a shorter line. Fresh and air dried soils had similar growth rates and behaved similarly in all soil types, the length of the lines indicates they had similar cumulative respiration maximums to baked soils. The cold treated soils were the most different with the lowest cumulative respiration maximums in all three soils, and growth rates similar to those in fresh and air dried soils.

The PCA biplot of the mineralization rates (Figure A3.4B) it can be seen that the mineralization rates are converging on a common mineralization rate over time. Cold, fresh, and air dried treated soil samples proceed from left to right over time, while baked and autoclaved soils do not. Baked and autoclaved soils have spikes in mineralization rates on Days 2 and 4 respectively (Figure SA3.1), and this

is represented by the deviation of the lines away from the common origin. This deviation is especially apparent in autoclave treated soils where the line starts near the origin, deviates from the origin, and then returns to the origin area likely indicating some change in the microbial community. Thus, the PCA biplot provides a unique way to monitor for system perturbations in time series data.

4. DISCUSSION

Incubation studies are critical to build our understanding of the mechanics behind soil respiration of greenhouse gases. A better understanding of greenhouse gas emissions from soil – the largest terrestrial pool of C (Jackson et al., 2017; Schlesinger, 1990) – will lead to better predictive climate simulation models and improving our abilities to fight anthropogenic climate change. To that end we have implemented a variety of different nonlinear models to evaluate their performance at modelling the cumulative soil respiration in three different soils and five different soil treatments designed to sterilize and/or deactivate soil microbes.

The most parsimonious model, Gompertz, was chosen as the best function for modelling soil respiration data. Gompertz models have been used in this capacity previously and were applied to the degradation (Fan et al., 2004; Wang et al., 2008, 2004) and mineralization (Johnsen et al., 2013; LeFevre et al., 2012; Soulas, 1993; Zablotowicz et al., 2007) of different organic pollutants. While Gompertz did not have the best fits compared to Baranyi and Huang models, all of its model parameters were significant demonstrating the trade-off between accuracy (model fit) and simplicity (parsimony).

The Gompertz model was able to generate sensical parameters to analyze the cumulative soil respiration data from a variety of different study groups. When comparing the three soil types, the parameter K or the cumulative respiration maximum, was the most affected by the different soil types,

while growth rates (mumax) and the y-intercept were unaffected. This is consistent with past mineralization studies where clay content was determined to have no effect on the CO_2 production rate (i.e. mumax), but the protective effect of clay became more important as substrate availability and microbial activity reached an equilibrium – or K (Wang et al., 2003). This was indicated when the more clayed soil – Rideau Clay – had similar growth rates (mumax) in fresh soil when compared to Grenville Loam or Granby Loam, but significantly lower cumulative respiration maximums (K). This effect was acerbated in air dried soil – an alternate control – where the differences between cumulative respiration maximums (K) was more apparent, possibly because the composition of stored air dried soil can change over time (Sun et al., 2015), which may enhance the protective effects in clay and reduce the overall SOM availability at equilibrium.

When the Gompertz model was also able to model different soil respiration behaviours after soil sterilization (i.e. autoclaving and baking) and deactivation (i.e. cold incubation) treatments. Soil respiration can act as a proxy for microbial activity because the transformation of SOM to the greenhouse gas (i.e. CO_2) is controlled by soil microbes. Past studies have used similar incubation studies to relate carbon mineralization to microbial community dynamics and carbon use efficiencies (Sinsabaugh et al., 2013; Wang et al., 2016). Thus, since the treatments used here are known to effect microbial growth through sterilization or deactivation, we would expect to see changes in soil respiration. These changes were apparent in parameters estimated by the Gompertz model. Autoclaving caused significantly lower cumulative respiration maximums (K) when compared to fresh soils, as did the cold incubation. This indicates that sterilization and deactivation do have an impact on soil respiration via decreased microbial activities. Baking had a significantly higher cumulative respiration maximum (K) in Granby and Grenville Loams, but a lower maximum in Rideau Clay. This could be due to variances in

the soils composition (i.e. more sand) effecting thermal conductivities (Abu-Hamdeh and Reeder, 2000; Tarnawski and Leong, 2000), and thus the efficiency of the heat treatment leading to more labile material for microbial activity (Kiersch et al., 2012). Indeed the theory of increased lability is supported by greater growth rates (mumax) in baked than fresh soils. However, autoclaved and cold soils also demonstrated greater growth rates than fresh soils. For autoclaved soils, autoclaving is known to also modify SOM producing more labile material (Berns et al., 2008; Lees et al., 2018; Trevors, 1996; Wolf and Skipper, 2018), it is also known to more effective at sterilizing soil microbes will having a lesser impact on SOM than baking, but both baking and autoclaving do not always completely sterilize the soils are known to miss heat resistant spores (Lees et al., 2018; Wolf and Skipper, 2018). So, any heat tolerant microbes that survived had a greater abundance of labile organic C to degrade, but autoclaving may have transformed less SOM and may have reduced the microbial community more than baking, thus it makes sense then that autoclaving would have a lower cumulative respiration maximum (i.e. less labile SOM) and lower growth rates (i.e. smaller microbial community) than baking. Cold having similar growth rates to autoclaving even though it had significantly lower cumulative respiration maximum was puzzling. The cold treatment was not expected to sterilize the soil, but to slow soil respiration (Yanni et al., 2018) by reducing microbial growth (Rousk and Bååth, 2011), and the enzymatic conversion of metabolites (Sitnikov et al., 2016) so the lower cumulative respiration maximum was expected. However, a possible explanation for the increased growth rates compared to fresh soil is that during measurement the cold jars would be removed from the cold incubation chamber and as a result their internal temperatures would increase briefly. While cold, the microbes were nutrient limited, so when the soils were warmed briefly there was increased microbial activity, which could result in an overall

greater growth rate than in fresh soils, with an over all lower cumulative respiration maximum because the soils was cold most of the time.

Aside from the Gompertz model, this work also applied a PCA model to time series data. These PCA models are modelled after paleoecological analyses performed in down-core analyses of sediment profiles (Hargan et al., 2018). This analysis is a unique way to display soil respiration data, and it supports the evidence provided from the Gompertz models. For instance, when assessing the PCA biplot (Figure A3.4A) of the cumulative respiration data is immediately apparent from the loadings arrows that Rideau Clay soil respiration is different from Granby and Grenville Loams. The colored arrows follow the time series, and after comparison with the results from the Gompertz model, this PCA also predicted changes in the cumulative respiration maximum, growth rates, and y-intercepts according to treatment. In addition, when the mineralization rates are also analyzed using PCA, we find support for the sterilization treatments producing more labile material for a temporary increase in growth rates because the colored lines deviate from the common origin. This deviation indicates a temporary spike in mineralization rates where more material was mineralized, and then once consumed, and the progress to the steady state. Lastly, this representation of dimensionality reduction statistical methods is a simplistic example because of the short duration of the incubation. These methods are more commonly applied to larger and longer time series data with more variables, and it has been effectively applied to complex spatio-temporal data (Bueso et al., 2020; Kumar and Duffy, 2009).

5. CONCLUSIONS

The work compared different nonlinear functions for a weighted nonlinear regression of soil cumulative respiration data. The final model, Gompertz, was successfully implemented in R and

produced parameters that coherently described the nonlinear curves, and we were able to translate these parameters in the context of changes to soil respiration between different soil types, and different soil sterilization and deactivation procedures. We also demonstrated a simple application of PCA on small multivariate time series data and discuss the future applications of PCA analysis to longer term multivariate soil respiration time series.

ACKNOWLEDGEMENTS

Funding for this study was provided by the Science & Technology Branch of Agriculture & Agri-Food Canada [Project J-001756 “Biological Soil Carbon Stabilization”] to EG and DPO.

SUPPLEMENTARY INFORMATION

Table SA3.1: Characteristics of the soils and clays used in this study.

Soil Series	Grenville Loam (gs)	Rideau Clay (rc)	Granby soil (fo)
Soil Classification*	Orthic Melanic Brunisol	Gleyed Melanic Brunisol	Orthic Humic Gleysol
Location	CEF Ottawa	CEF Ottawa	CEF Ottawa
Sampling	20 kg bulk (0-15cm)	20 kg bulk (0-15cm)	20 kg bulk (0-15cm)
Treatment/Storage	Sieved <4mm, Air-dried	Sieved <4mm, Air-dried	Sieved <4mm, Air-dried
Post Treatment-subsample	Ground & Sieved <1mm	Ground & Sieived <1mm	Ground & Sieived <1mm
Soil Texture Class	Fine sandy loam	clay loam	Fine sandy loam
Sand (%)	64	50	73
Silt (%)	26	20	17
Clay (%)	10	30	10
% C	2.50	1.64	1.19
% OC	2.55	1.58	1.17
% N	0.21	0.13	0.10
pH (1:1 DH2O)	6.27	6.57	6.21
Bulk Density (g/cm3)	1.34	1.28	1.47
Specific surface area (m^2/g)	4.04	16.31	3.30
CEC (meq/100g)	13.00	19.00	7.63

Table SA3.2: Parameter estimates for weighted non-linear Gompertz models. SE = Standard error

Soil Type	Treatment	Sigma	AIC	BIC	K (SE)	mumax (SE)	y0 (SE)
Granby Loam	Air Dried	1.01	2566	2571	287.02 (19.89)	0.32 (0.02)	19.72 (1.01)
	Fresh	1.00	2534	2539	290.00 (9.78)	0.32 (0.02)	21.38 (2.00)
	Autoclaved	0.94	2678	2682	120.00 (13.56)	0.51 (0.06)	0.22 (0.12)
	Baked	0.98	3217	3222	309.14 (8.13)	0.71 (0.06)	14.67 (4.02)
	Cold	0.94	620	625	24.61 (0.99)	0.46 (0.05)	2.57 (0.31)
Grenville Loam	Air Dried	0.93	3718	3722	299.38 (20.21)	0.30 (0.05)	41.97 (7.97)
	Fresh	0.98	3009	3014	266.30 (12.55)	0.33 (0.03)	22.75 (3.72)
	Autoclaved	1.01	2385	2390	190.00 (16.98)	0.53 (0.03)	0.58 (0.14)
	Baked	2.15	2885	2890	328.57 (21.79)	0.73 (0.09)	22.68 (4.00)
	Cold	0.96	1094	1099	31.32 (1.63)	0.48 (0.06)	3.00 (0.61)
Rideau Clay	Air Dried	0.97	2371	2376	183.14 (9.52)	0.36 (0.03)	19.15 (2.10)
	Fresh	0.93	3061	3066	229.56 (12.07)	0.32 (0.04)	17.87 (3.28)
	Autoclaved	0.94	2161	2166	60.00 (12.92)	0.36 (0.08)	0.57 (0.28)
	Baked	3.28	-124	-119	80.00 (7.39)	0.41 (0.02)	0.66 (0.08)
	Cold	0.95	1630	1635	30.00 (3.89)	0.54 (0.11)	2.21 (0.64)

*Indicates parameter was significant to the model

^Sigma on 21 degrees of freedom

Table SA3.3: Growth rate model formulas. Where y0 = y-intercept, mumax = intrinsic growth rate (1/time), and K = max concentration = carrying capacity

Model	Formula
Logistic	y = (K*y0)/(y0+(K-y0)*exp(-mumax*time))
Gompertz	y = K*exp(log(y0/K)*exp(-mumax*time))
Modified Gompertz	y = y0*(K/y0)^(exp(-exp((exp(1)*mumax *(lambda - time))/log(K/y0)+1)))
Richards	y = K*(1-exp(-beta*mumax*time)*(1(y0/K) ^-beta))^(-1/beta)
Baranyi	A = time+1/mumax*log(exp(-mumax *time)+exp(-h0)-exp(-mumax*time-h0)) log(y) = log(y0)+mumax*A-log(1+ (exp(mumax*A)-1) / exp(log(K) - log(y0)))
Huang	B = time+1/alpha*log((1+exp(-alpha* (time-lambda)))/(1+ exp(alpha*lambda))) log(y) = log(y0)+log(K)-log(y0+(K-y0)*exp(-mumax*B))

Table SA3.4: Parameter estimates with standard errors, p-values, AIC, BIC, and sigma for all models.

sample	treatment	regression	sigma	AIC	BIC	alpha (SE)	Alpha (p.value)	beta (SE)	Beta (p.value)	h0 (SE)	h0 (p.value)	K (SE)	K (p.value)	lambda (SE)	Lambda (p.value)	mumax (SE)	Mumax (p.value)	y0 (SE)	y0 (p.value)
Granby Loam	Air Dried	baranyi	0.017	2391	2397	NA	NA	NA	NA	0 (0.82)	1.00	295 (76.35)	9.67E-04	NA	NA	0.64 (0.22)	9.35E-03	22.71 (3.66)	4.70E-06
Granby Loam	Air Dried	gompertz	1.011	2566	2571	NA	NA	NA	NA	NA	NA	287.02 (19.89)	2.26E-12	NA	NA	0.32 (0.02)	3.98E-12	19.72 (1.01)	5.77E-15
Granby Loam	Air Dried	gompertz_mod	1.409	2579	2585	NA	NA	NA	NA	NA	NA	275 (23.8)	2.64E-10	0 (0.88)	1.00	0.49 (0.03)	2.97E-13	24.17 (9.22)	1.63E-02
Granby Loam	Air Dried	huang	0.017	2393	2400	0 (0.19)	1.00	NA	NA	NA	NA	275 (59.55)	1.88E-04	0 (0.03)	1.00	1.29 (0.42)	6.38E-03	22.6 (3.34)	1.84E-06
Granby Loam	Air Dried	logistic	1.223	2573	2578	NA	NA	NA	NA	NA	NA	275 (18.67)	1.52E-12	NA	NA	0.6 (0.03)	6.33E-14	23.71 (0.98)	8.26E-17
Granby Loam	Air Dried	richards	1.045	2568	2574	NA	NA	0.07 (0.41)	0.87	NA	NA	284.5 (26.3)	8.28E-10	NA	NA	5 (29.26)	8.66E-01	20.12 (2.48)	9.18E-08
Granby Loam	Autoclaved	baranyi	0.203	2645	2651	NA	NA	NA	NA	0 (9.58)	1.00	120 (264.17)	6.55E-01	NA	NA	1.5 (2.07)	4.79E-01	0.55 (2.85)	8.49E-01
Granby Loam	Autoclaved	gompertz	0.937	2678	2682	NA	NA	NA	NA	NA	NA	120 (13.56)	1.57E-08	NA	NA	0.51 (0.06)	1.03E-07	0.22 (0.12)	8.04E-02
Granby Loam	Autoclaved	gompertz_mod	0.960	2679	2685	NA	NA	NA	NA	NA	NA	112.29 (13.07)	3.82E-08	0 (1.49)	1.00	1.43 (0.29)	7.62E-05	0.63 (1.44)	6.66E-01
Granby Loam	Autoclaved	huang	0.208	2647	2654	2 (300)	0.99	NA	NA	NA	NA	120 (275.73)	6.68E-01	0 (153.11)	1.00	1.61 (6.6)	8.10E-01	0.76 (43.24)	9.86E-01
Granby Loam	Autoclaved	logistic	0.963	2678	2683	NA	NA	NA	NA	NA	NA	107.54 (10.99)	2.83E-09	NA	NA	1.33 (0.14)	3.18E-09	0.74 (0.2)	1.34E-03
Granby Loam	Autoclaved	richards	0.954	2679	2685	NA	NA	0.2 (0.43)	0.65	NA	NA	115.22 (14.55)	1.37E-07	NA	NA	3.29 (5.55)	5.60E-01	0.37 (0.35)	2.99E-01
Granby Loam	Baked	baranyi	0.010	3009	3015	NA	NA	NA	NA	0 (2.06)	1.00	300 (22.55)	2.15E-11	NA	NA	1.26 (0.49)	1.79E-02	21.81 (19.96)	2.88E-01
Granby Loam	Baked	gompertz	0.982	3217	3222	NA	NA	NA	NA	NA	NA	309.14 (8.13)	7.57E-21	NA	NA	0.71 (0.06)	6.56E-11	14.67 (4.02)	1.51E-03
Granby Loam	Baked	gompertz_mod	1.151	3225	3231	NA	NA	NA	NA	NA	NA	301.66 (9.29)	8.91E-19	0 (0.96)	1.00	0.87 (0.26)	3.56E-03	28.12 (30.81)	3.72E-01
Granby Loam	Baked	huang	0.011	3012	3020	0.53 (12.78)	0.97	NA	NA	NA	NA	300 (25.36)	3.29E-10	0 (19.84)	1.00	1.82 (16.29)	9.12E-01	26.67 (44.32)	5.54E-01
Granby Loam	Baked	logistic	1.129	3225	3230	NA	NA	NA	NA	NA	NA	303.53 (8.72)	4.64E-20	NA	NA	1.05 (0.1)	4.14E-10	29.85 (4.72)	2.87E-06
Granby Loam	Baked	richards	1.027	3220	3226	NA	NA	0.15 (0.6)	0.80	NA	NA	308.23 (9.21)	4.90E-19	NA	NA	5 (18.44)	7.89E-01	17.9 (11.85)	1.47E-01
Granby Loam	Cold	baranyi	0.103	538	544	NA	NA	NA	NA	0 (1.43)	1.00	24.1 (2.34)	1.89E-09	NA	NA	0.78 (0.36)	4.05E-02	2.93 (1.06)	1.18E-02
Granby Loam	Cold	gompertz	0.943	620	625	NA	NA	NA	NA	NA	NA	24.61 (0.99)	4.92E-17	NA	NA	0.46 (0.05)	2.70E-09	2.57 (0.31)	4.00E-08
Granby Loam	Cold	gompertz_mod	1.084	625	631	NA	NA	NA	NA	NA	NA	23.76 (1.07)	1.44E-15	0 (1.33)	1.00	0.52 (0.07)	6.38E-07	3.4 (2.36)	1.64E-01

Granby Loam	Cold	huang	0.120	539	546	0 (0.09)	1.00	NA	NA	NA	NA	30 (3.37)	3.27E-08	2 (1367.08)	1.00	1.32 (0.25)	4.83E-05	3.28 (0.55)	9.55E-06
Granby Loam	Cold	logistic	1.022	623	628	NA	NA	NA	NA	NA	NA	23.93 (0.95)	3.89E-17	NA	NA	0.73 (0.07)	9.30E-10	3.27 (0.28)	1.46E-10
Granby Loam	Cold	richards	0.973	623	628	NA	NA	0.1 (0.81)	0.91	NA	NA	24.54 (1.18)	4.81E-15	NA	NA	5 (39.68)	9.01E-01	2.68 (0.83)	4.35E-03
Granby Loam	Fresh	baranyi	0.016	2348	2354	NA	NA	NA	NA	0 (0.88)	1.00	275 (31.92)	3.65E-08	NA	NA	0.62 (0.19)	4.37E-03	24.28 (5.15)	1.32E-04
Granby Loam	Fresh	gompertz	1.002	2534	2539	NA	NA	NA	NA	NA	NA	290 (9.78)	1.28E-18	NA	NA	0.32 (0.02)	2.33E-13	21.38 (2)	5.83E-10
Granby Loam	Fresh	gompertz_mod	1.399	2550	2556	NA	NA	NA	NA	NA	NA	275 (12.11)	9.46E-16	0 (0.94)	1.00	0.42 (0.04)	5.02E-10	28.04 (11.3)	2.21E-02
Granby Loam	Fresh	huang	0.018	2354	2361	0.17 (2.97)	0.95	NA	NA	NA	NA	275 (41.98)	2.84E-06	0 (113.98)	1.00	0.99 (10.49)	9.26E-01	27.01 (6.52)	5.54E-04
Granby Loam	Fresh	logistic	1.222	2547	2552	NA	NA	NA	NA	NA	NA	275.9 (9.8)	3.70E-18	NA	NA	0.52 (0.03)	3.69E-13	29.19 (2.19)	1.03E-11
Granby Loam	Fresh	richards	1.035	2537	2543	NA	NA	0.07 (0.34)	0.85	NA	NA	289.66 (12.83)	1.06E-15	NA	NA	5 (24.87)	8.43E-01	22.18 (3.91)	1.52E-05
Grenville Loam	Air Dried	baranyi	0.009	3507	3513	NA	NA	NA	NA	0 (1.76)	1.00	279.74 (43.96)	3.29E-06	NA	NA	0.51 (0.36)	1.74E-01	44.56 (14.79)	6.89E-03
Grenville Loam	Air Dried	gompertz	0.934	3718	3722	NA	NA	NA	NA	NA	NA	299.38 (20.21)	1.37E-12	NA	NA	0.3 (0.05)	1.31E-05	41.97 (7.97)	3.21E-05
Grenville Loam	Air Dried	gompertz_mod	1.004	3722	3728	NA	NA	NA	NA	NA	NA	285.1 (18.2)	1.08E-12	0 (2.1)	1.00	0.31 (0.06)	3.78E-05	50 (32.61)	1.41E-01
Grenville Loam	Air Dried	huang	0.010	3511	3518	0.11 (9.22)	0.99	NA	NA	NA	NA	281.89 (57.05)	9.06E-05	0 (1132.18)	1.00	0.82 (51.11)	9.87E-01	50 (21.1)	2.85E-02
Grenville Loam	Air Dried	logistic	0.950	3719	3724	NA	NA	NA	NA	NA	NA	289.89 (16.86)	7.51E-14	NA	NA	0.45 (0.07)	1.85E-06	50 (6.92)	4.02E-07
Grenville Loam	Air Dried	richards	0.957	3720	3726	NA	NA	0.06 (1.36)	0.96	NA	NA	298.53 (28.55)	1.49E-09	NA	NA	5 (108.21)	9.64E-01	42.62 (15.64)	1.31E-02
Grenville Loam	Autoclaved	baranyi	0.072	2296	2302	NA	NA	NA	NA	0 (2.01)	1.00	190 (164.39)	2.61E-01	NA	NA	1.46 (0.51)	9.73E-03	1.48 (1.5)	3.34E-01
Grenville Loam	Autoclaved	gompertz	1.009	2385	2390	NA	NA	NA	NA	NA	NA	190 (16.98)	2.61E-10	NA	NA	0.53 (0.03)	3.26E-13	0.58 (0.14)	5.27E-04
Grenville Loam	Autoclaved	gompertz_mod	0.967	2381	2386	NA	NA	NA	NA	NA	NA	171.51 (15)	3.17E-10	0 (0.52)	1.00	1.4 (0.11)	1.04E-10	1.54 (1.25)	2.32E-01
Grenville Loam	Autoclaved	huang	0.074	2298	2305	0 (0)	0.99	NA	NA	NA	NA	190 (136.07)	1.79E-01	2 (1095.66)	1.00	2.91 (0.24)	1.62E-10	1.49 (0.16)	1.89E-08
Grenville Loam	Autoclaved	logistic	1.035	2384	2388	NA	NA	NA	NA	NA	NA	161.28 (13.62)	9.34E-11	NA	NA	1.35 (0.06)	7.91E-16	1.77 (0.21)	3.34E-08
Grenville Loam	Autoclaved	richards	0.957	2380	2386	NA	NA	0.36 (0.23)	0.14	NA	NA	176.74 (15.91)	5.24E-10	NA	NA	2.21 (0.92)	2.66E-02	1.16 (0.36)	4.51E-03
Grenville Loam	Baked	baranyi	0.019	2663	2668	NA	NA	NA	NA	1.42 (1.48)	0.35	310 (47.41)	2.26E-06	NA	NA	1.65 (0.67)	2.31E-02	50 (13.97)	1.88E-03
Grenville Loam	Baked	gompertz	2.150	2885	2890	NA	NA	NA	NA	NA	NA	328.57 (21.79)	9.76E-13	NA	NA	0.73 (0.09)	1.39E-07	22.68 (4)	1.27E-05

Grenville Loam	Baked	gompertz_mod	2.251	2870	2876	NA	NA	NA	NA	NA	NA	350 (22.61)	1.34E-12	0.27 (0.72)	0.71	0.75 (0.07)	1.89E-09	50 (25.97)	6.86E-02
Grenville Loam	Baked	huang	0.019	2664	2671	1.51 (5.84)	0.80	NA	NA	NA	NA	310 (55.69)	2.28E-05	0.7 (1.83)	0.70	1.63 (2.57)	5.35E-01	50 (28.31)	9.35E-02
Grenville Loam	Baked	logistic	2.086	2879	2884	NA	NA	NA	NA	NA	NA	312.96 (18.7)	1.28E-13	NA	NA	1.2 (0.12)	3.72E-09	33.38 (2.94)	2.00E-10
Grenville Loam	Baked	richards	2.137	2881	2887	NA	NA	0.88 (1.19)	0.47	NA	NA	314.58 (21.63)	4.23E-12	NA	NA	1.3 (1.13)	2.66E-01	32.56 (10.38)	5.20E-03
Grenville Loam	Cold	baranyi	0.102	1003	1009	NA	NA	NA	NA	0 (1.88)	1.00	32.96 (5.38)	5.54E-06	NA	NA	0.72 (0.43)	1.04E-01	3.38 (1.72)	6.29E-02
Grenville Loam	Cold	gompertz	0.958	1094	1099	NA	NA	NA	NA	NA	NA	31.32 (1.63)	8.11E-15	NA	NA	0.48 (0.06)	8.24E-08	3 (0.61)	7.00E-05
Grenville Loam	Cold	gompertz_mod	1.072	1099	1105	NA	NA	NA	NA	NA	NA	30.15 (1.7)	1.08E-13	0 (1.62)	1.00	0.54 (0.12)	2.41E-04	4.13 (3.91)	3.03E-01
Grenville Loam	Cold	huang	0.097	999	1007	0 (0)	1.00	NA	NA	NA	NA	38 (5.74)	2.48E-06	2 (521.75)	1.00	1.38 (0.45)	6.38E-03	3.89 (1.64)	2.87E-02
Grenville Loam	Cold	logistic	1.018	1098	1102	NA	NA	NA	NA	NA	NA	30.43 (1.53)	4.32E-15	NA	NA	0.73 (0.09)	3.43E-08	4.08 (0.57)	4.52E-07
Grenville Loam	Cold	richards	0.988	1097	1102	NA	NA	0.1 (1)	0.92	NA	NA	31.23 (1.94)	6.52E-13	NA	NA	5 (47.83)	9.18E-01	3.16 (1.55)	5.43E-02
Grenville Loam	Fresh	baranyi	0.016	2823	2829	NA	NA	NA	NA	0 (1.35)	1.00	270 (44.56)	6.37E-06	NA	NA	0.6 (0.29)	4.96E-02	25.06 (8.36)	7.14E-03
Grenville Loam	Fresh	gompertz	0.984	3009	3014	NA	NA	NA	NA	NA	NA	266.3 (12.55)	1.14E-15	NA	NA	0.33 (0.03)	2.96E-09	22.75 (3.72)	4.48E-06
Grenville Loam	Fresh	gompertz_mod	1.189	3020	3025	NA	NA	NA	NA	NA	NA	252.24 (13.45)	3.69E-14	0 (1.38)	1.00	0.4 (0.06)	1.09E-06	30.76 (17.68)	9.73E-02
Grenville Loam	Fresh	huang	0.017	2828	2835	0 (0)	0.98	NA	NA	NA	NA	270 (47.65)	1.84E-05	2 (32.98)	0.95	1.07 (0.3)	1.97E-03	28.02 (9.77)	9.88E-03
Grenville Loam	Fresh	logistic	1.092	3016	3020	NA	NA	NA	NA	NA	NA	255.98 (11.72)	6.40E-16	NA	NA	0.52 (0.05)	1.79E-09	31.24 (3.65)	2.79E-08
Grenville Loam	Fresh	richards	1.014	3012	3017	NA	NA	0.07 (0.61)	0.91	NA	NA	265.35 (16.18)	4.61E-13	NA	NA	5 (43.4)	9.09E-01	23.56 (7.32)	4.30E-03
Rideau Clay	Air Dried	baranyi	0.017	2188	2194	NA	NA	NA	NA	0 (1.02)	1.00	160 (17.7)	1.69E-08	NA	NA	0.68 (0.25)	1.37E-02	21.07 (5.04)	4.59E-04
Rideau Clay	Air Dried	gompertz	0.968	2371	2376	NA	NA	NA	NA	NA	NA	183.14 (9.52)	8.17E-15	NA	NA	0.36 (0.03)	5.03E-10	19.15 (2.1)	9.36E-09
Rideau Clay	Air Dried	gompertz_mod	1.194	2381	2386	NA	NA	NA	NA	NA	NA	164.11 (9.57)	1.97E-13	0 (1.17)	1.00	0.44 (0.05)	3.28E-08	23.94 (12.63)	7.25E-02
Rideau Clay	Air Dried	huang	0.020	2195	2202	0 (0)	0.99	NA	NA	NA	NA	200 (26.86)	4.79E-07	2 (848)	1.00	1.15 (0.15)	1.84E-07	22.74 (2.05)	9.51E-10
Rideau Clay	Air Dried	logistic	1.080	2377	2381	NA	NA	NA	NA	NA	NA	169.09 (7.81)	7.65E-16	NA	NA	0.59 (0.05)	1.42E-10	23.83 (2.08)	1.78E-10
Rideau Clay	Air Dried	richards	0.998	2373	2379	NA	NA	0.07 (0.55)	0.89	NA	NA	181.91 (13.76)	2.42E-11	NA	NA	5 (35.38)	8.89E-01	19.68 (4.08)	1.04E-04
Rideau Clay	Autoclaved	baranyi	0.277	2135	2140	NA	NA	NA	NA	0 (5.44)	1.00	60 (158.32)	7.09E-01	NA	NA	0.99 (1.53)	5.24E-01	0.83 (1.54)	5.97E-01

Rideau Clay	Autoclaved	gompertz	0.940	2161	2166	NA	NA	NA	NA	NA	NA	60 (12.92)	1.39E-04	NA	NA	0.36 (0.08)	1.30E-04	0.57 (0.28)	5.31E-02
Rideau Clay	Autoclaved	gompertz_mod	0.948	2162	2168	NA	NA	NA	NA	NA	NA	57.21 (11.5)	7.26E-05	0 (1.84)	1.00	0.88 (0.16)	2.97E-05	0.98 (1.56)	5.35E-01
Rideau Clay	Autoclaved	huang	0.284	2137	2144	0 (3.29)	1.00	NA	NA	NA	NA	60 (171.88)	7.31E-01	2 (36042.28)	1.00	1.99 (8.32)	8.13E-01	0.82 (0.99)	4.16E-01
Rideau Clay	Autoclaved	logistic	0.929	2160	2165	NA	NA	NA	NA	NA	NA	55.14 (9.72)	1.25E-05	NA	NA	0.89 (0.15)	9.67E-06	1.04 (0.28)	1.49E-03
Rideau Clay	Autoclaved	richards	0.948	2162	2168	NA	NA	0.53 (0.87)	0.55	NA	NA	58.1 (12.4)	1.42E-04	NA	NA	1.16 (1.03)	2.73E-01	0.9 (0.47)	7.06E-02
Rideau Clay	Baked	baranyi	0.156	-242	-236	NA	NA	NA	NA	1.06 (0.31)	0.003	60 (12.05)	7.19E-05	NA	NA	1.32 (0.12)	3.87E-10	1.72 (0.13)	1.25E-11
Rideau Clay	Baked	gompertz	3.282	-124	-119	NA	NA	NA	NA	NA	NA	80 (7.39)	4.70E-10	NA	NA	0.41 (0.02)	9.88E-14	0.66 (0.08)	7.24E-08
Rideau Clay	Baked	gompertz_mod	1.089	-186	-180	NA	NA	NA	NA	NA	NA	68.84 (1.92)	1.22E-19	0.5 (0.1)	5.41E-05	1.03 (0.01)	1.75E-25	1.84 (0.15)	1.19E-10
Rideau Clay	Baked	huang	0.173	-239	-232	0.27 (1.98)	0.89	NA	NA	NA	NA	80 (22.13)	1.84E-03	2 (57.21)	0.97	2.13 (16.33)	8.98E-01	1.44 (0.26)	2.60E-05
Rideau Clay	Baked	logistic	1.569	-165	-160	NA	NA	NA	NA	NA	NA	67.63 (2.41)	4.00E-18	NA	NA	1.07 (0.02)	2.41E-24	1.17 (0.04)	1.55E-18
Rideau Clay	Baked	richards	1.886	-179	-173	NA	NA	2 (0.52)	0.00	NA	NA	63.77 (2.81)	9.55E-16	NA	NA	0.93 (0.03)	1.66E-17	1.32 (0.06)	6.25E-15
Rideau Clay	Cold	baranyi	0.087	1538	1544	NA	NA	NA	NA	0 (2.13)	1.00	38 (13.04)	8.56E-03	NA	NA	0.82 (0.58)	1.72E-01	3.12 (1.84)	1.06E-01
Rideau Clay	Cold	gompertz	0.946	1630	1635	NA	NA	NA	NA	NA	NA	30 (3.89)	1.46E-07	NA	NA	0.54 (0.11)	4.14E-05	2.21 (0.64)	2.42E-03
Rideau Clay	Cold	gompertz_mod	1.015	1635	1641	NA	NA	NA	NA	NA	NA	30 (3.96)	2.67E-07	0 (1.81)	1.00	0.64 (0.15)	3.98E-04	3.47 (4.31)	4.31E-01
Rideau Clay	Cold	huang	0.090	1540	1547	0 (0.42)	1.00	NA	NA	NA	NA	38 (12.56)	6.98E-03	2 (5991.4)	1.00	1.66 (0.62)	1.49E-02	3.08 (0.58)	3.93E-05
Rideau Clay	Cold	logistic	0.985	1632	1637	NA	NA	NA	NA	NA	NA	30 (3.31)	1.08E-08	NA	NA	0.89 (0.13)	1.16E-06	3.11 (0.55)	1.27E-05
Rideau Clay	Cold	richards	0.989	1633	1639	NA	NA	0.57 (1.61)	0.73	NA	NA	30 (4.58)	2.18E-06	NA	NA	1.28 (2.59)	6.27E-01	2.85 (1.4)	5.49E-02
Rideau Clay	Fresh	baranyi	0.016	2880	2886	NA	NA	NA	NA	0 (1.42)	1.00	210 (32.56)	2.73E-06	NA	NA	0.63 (0.31)	5.74E-02	20.11 (6.86)	8.24E-03
Rideau Clay	Fresh	gompertz	0.930	3061	3066	NA	NA	NA	NA	NA	NA	229.56 (12.07)	1.01E-14	NA	NA	0.32 (0.04)	2.22E-08	17.87 (3.28)	2.08E-05
Rideau Clay	Fresh	gompertz_mod	1.040	3067	3073	NA	NA	NA	NA	NA	NA	216.84 (11.65)	4.27E-14	0 (1.45)	1.00	0.43 (0.06)	1.10E-06	23.27 (14.82)	1.32E-01
Rideau Clay	Fresh	huang	0.017	2885	2892	0 (0.06)	1.00	NA	NA	NA	NA	240 (40.8)	1.15E-05	2 (1666.31)	1.00	1.02 (0.24)	4.34E-04	23.18 (4.05)	1.61E-05
Rideau Clay	Fresh	logistic	0.980	3065	3069	NA	NA	NA	NA	NA	NA	220.25 (10.49)	1.42E-15	NA	NA	0.53 (0.06)	5.07E-09	24.12 (3.07)	1.10E-07
Rideau Clay	Fresh	richards	0.955	3063	3069	NA	NA	0.07 (0.69)	0.92	NA	NA	228.59 (15.84)	4.88E-12	NA	NA	5 (49.34)	9.20E-01	18.44 (6.38)	9.05E-03

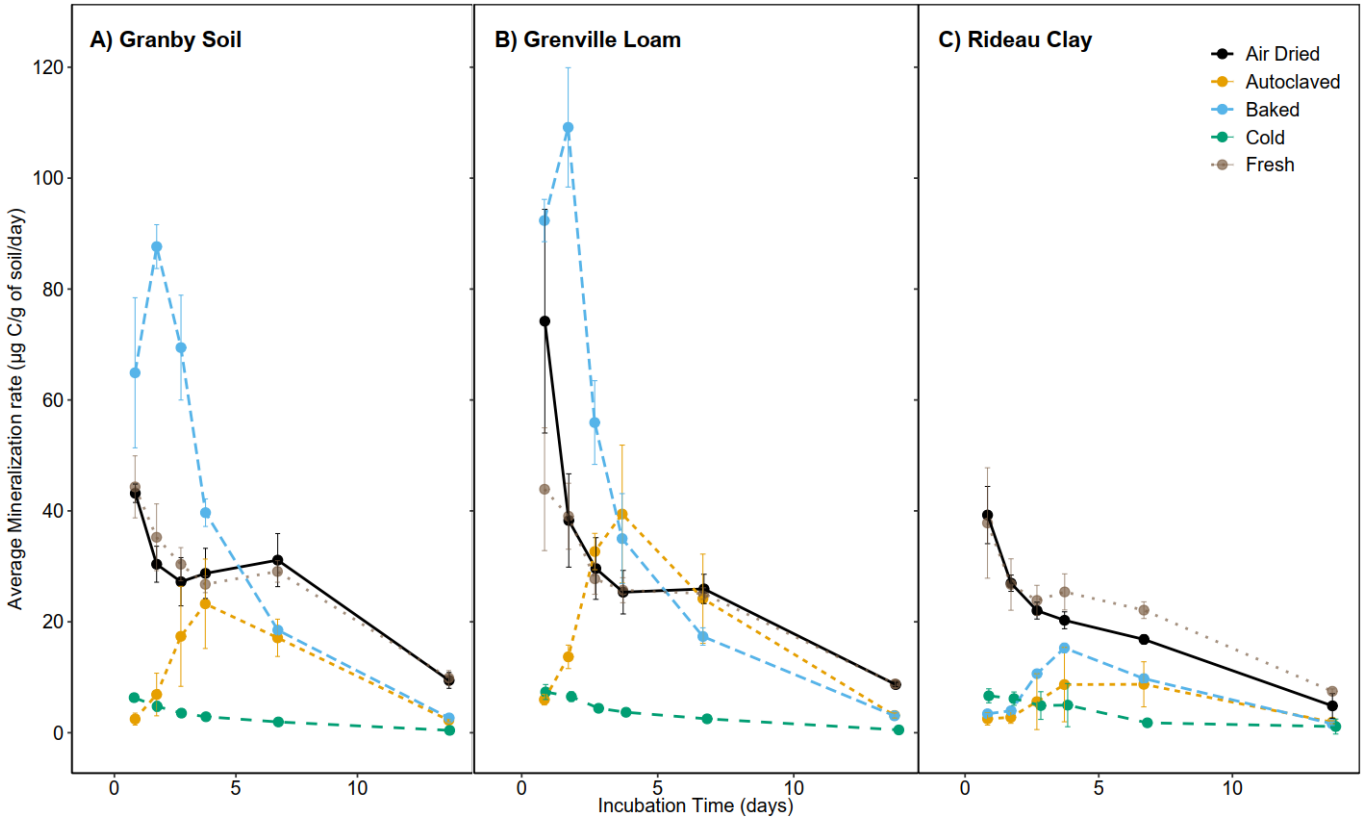


Figure SA3.1: Average mineralization rates (µg C/g of soil/day) of each of the 5 treatments over 14 days across the three soil types.

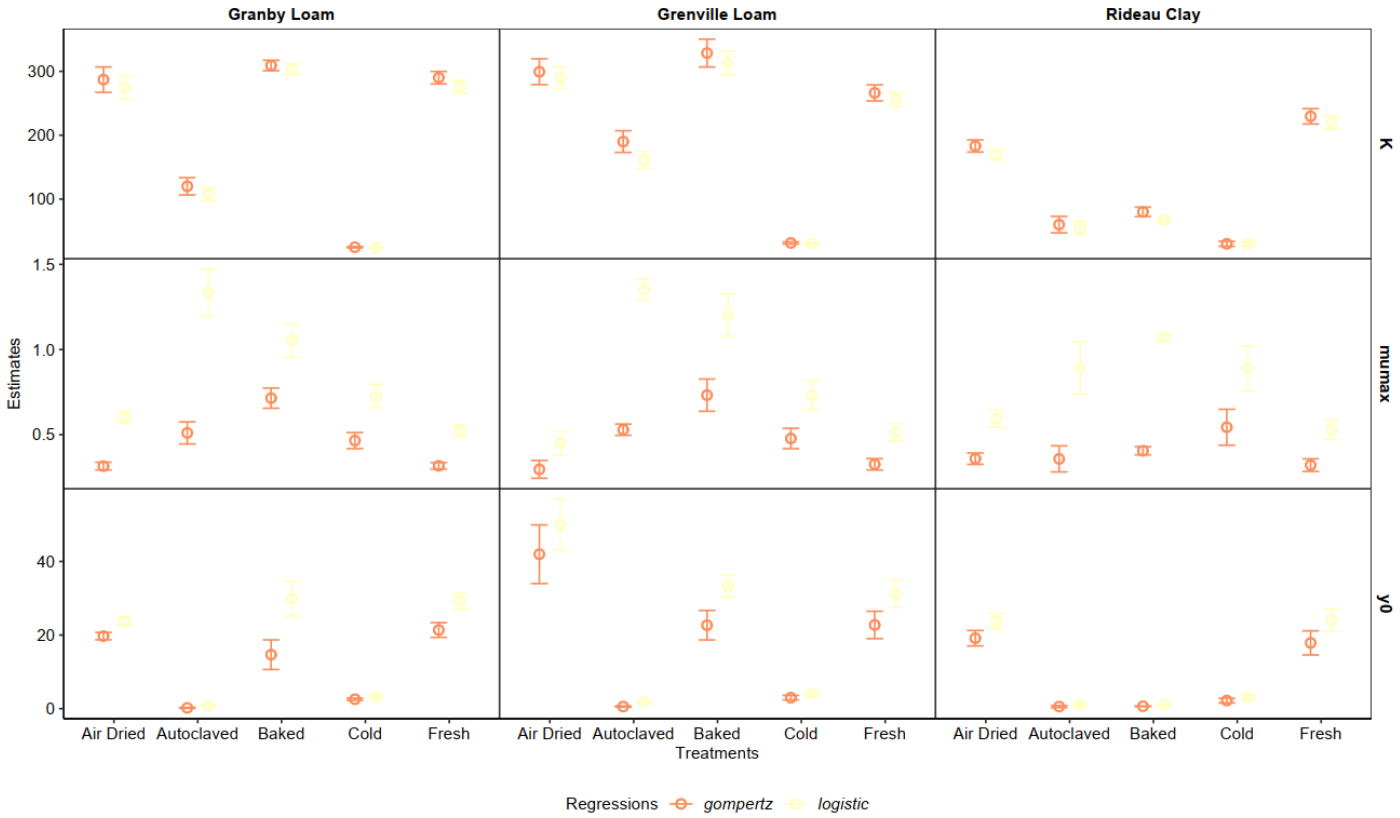
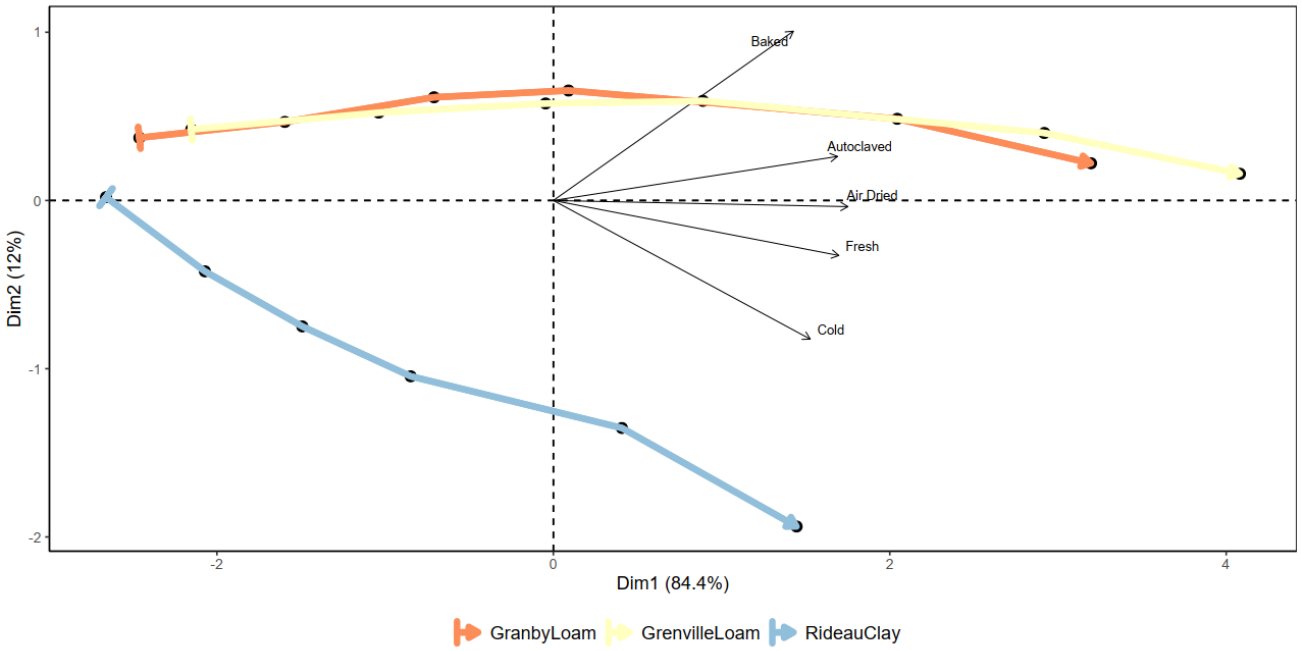
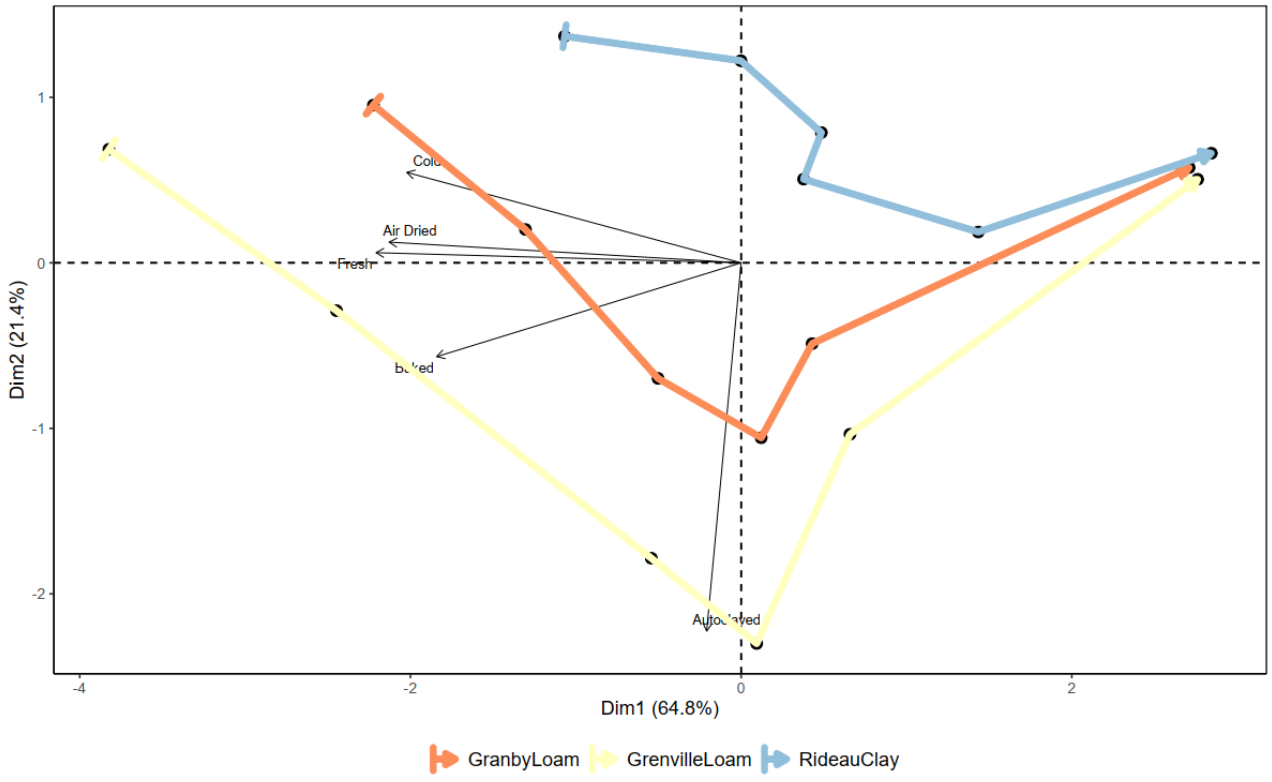


Figure SA3.2: Parameter estimates with standard errors for logistic and gompertz models.



A)



B)

Figure SA3.3: A) Cumulative soil respiration and B) mineralization rate represented in an biplot where overall soil trends over time are colored ("|" marker denotes the first measurement on Day 1; the ">" marks the end of the experiment on Day 14) and the respective effects of each treatment are indicated by the black arrows.

REFERENCES

Abu-Hamdeh, N.H., Reeder, R.C., 2000. Soil Thermal Conductivity Effects of Density, Moisture, Salt Concentration, and Organic Matter. Soil Sci. Soc. Am. J. 64, 1285–1290. https://doi.org/10.2136/sssaj2000.6441285x

Al turk, L.I., 2018. Weighted Nonlinear Least Squares Technique for Parameters Estimation of the NHPP Gompertz Model. J. Inf. Technol. Softw. Eng. 08. https://doi.org/10.4172/2175-7866.1000231

Baranyi, J., Roberts, T.A., 1995. Mathematics of predictive food microbiology. Int. J. Food Microbiol. 26, 199–218. https://doi.org/10.1016/0168-1605(94)00121-L

Baranyi, J., Roberts, T.A., 1994. A dynamic approach to predicting bacterial growth in food. Int. J. Food Microbiol. 23, 277–294. https://doi.org/10.1016/0168-1605(94)90157-0

Berns, A.E., Philipp, H., Narres, H.-D., Burauel, P., Vereecken, H., Tappe, W., 2008. Effect of gamma-sterilization and autoclaving on soil organic matter structure as studied by solid state NMR, UV and fluorescence spectroscopy. Eur. J. Soil Sci. 59, 540–550. https://doi.org/10.1111/j.1365-2389.2008.01016.x

Bueso, D., Piles, M., Camps-Valls, G., 2020. Nonlinear PCA for Spatio-Temporal Analysis of Earth Observation Data. IEEE Trans. Geosci. Remote Sens. 58, 5752–5763. https://doi.org/10.1109/TGRS.2020.2969813

Conant, R.T., Ryan, M.G., Ågren, G.I., Birge, H.E., Davidson, E.A., Eliasson, P.E., Evans, S.E., Frey, S.D., Giardina, C.P., Hopkins, F.M., Hyvönen, R., Kirschbaum, M.U.F., Lavallee, J.M., Leifeld, J., Parton, W.J., Megan Steinweg, J., Wallenstein, M.D., Martin Wetterstedt, J.Å., Bradford, M.A., 2011. Temperature and soil organic matter decomposition rates - synthesis of current knowledge and a way forward. Glob. Change Biol. 17, 3392–3404. https://doi.org/10.1111/j.1365-2486.2011.02496.x

Daffertshofer, A., Lamoth, C.J.C., Meijer, O.G., Beek, P.J., 2004. PCA in studying coordination and variability: a tutorial. Clin. Biomech. 19, 415–428. https://doi.org/10.1016/j.clinbiomech.2004.01.005

Dwivedi, D., Tang, J., Bouskill, N., Georgiou, K., Chacon, S.S., Riley, W.J., 2019. Abiotic and Biotic Controls on Soil Organo–Mineral Interactions: Developing Model Structures to Analyze Why Soil Organic Matter Persists. Rev. Mineral. Geochem. 85, 329–348. https://doi.org/10.2138/rmg.2019.85.11

Elzhov, T.V., Mullen, K.M., Spiess, A.-N., Bolker, B., 2016. minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds.

Fan, Y., Wang, Y., Qian, P.-Y., Gu, J.-D., 2004. Optimization of phthalic acid batch biodegradation and the use of modified Richards model for modelling degradation. Int. Biodeterior. Biodegrad. 53, 57–63. https://doi.org/10.1016/j.ibiod.2003.10.001

Firke, S., 2020. janitor: Simple Tools for Examining and Cleaning Dirty Data.

Hargan, K.E., Stewart, E.M., Michelutti, N., Grooms, C., Kimpe, L.E., Mallory, M.L., Smol, J.P., Blais, J.M., 2018. Sterols and stanols as novel tracers of waterbird population dynamics in freshwater ponds. Proc R Soc B 285, 20180631. https://doi.org/10.1098/rspb.2018.0631

Hopkins, D.W., 2007. Carbon mineralization, in: Soil Sampling and Methods of Analysis. Taylor & Francis Group, LLC, Boca Raton, FL, pp. 589–598.

Huang, L., 2011. A new mechanistic growth model for simultaneous determination of lag phase duration and exponential growth rate and a new Bělehdrádek-type model for evaluating the effect of temperature on growth rate. Food Microbiol., Predictive Modeling in Foods 28, 770–776. https://doi.org/10.1016/j.fm.2010.05.019

Huang, L., 2008. Growth Kinetics of Listeria monocytogenes in Broth and Beef Frankfurters—Determination of Lag Phase Duration and Exponential Growth Rate under Isothermal Conditions. J. Food Sci. 73, E235–E242. https://doi.org/10.1111/j.1750-3841.2008.00785.x

Jackson, R.B., Lajtha, K., Crow, S.E., Hugelius, G., Kramer, M.G., Piñeiro, G., 2017. The Ecology of Soil Carbon: Pools, Vulnerabilities, and Biotic and Abiotic Controls. Annu. Rev. Ecol. Evol. Syst. 48, 419–445. https://doi.org/10.1146/annurev-ecolsys-112414-054234

Johnsen, A.R., Binning, P.J., Aamand, J., Badawi, N., Rosenbom, A.E., 2013. The Gompertz Function Can Coherently Describe Microbial Mineralization of Growth-Sustaining Pesticides. Environ. Sci. Technol. 130711162351002. https://doi.org/10.1021/es400861v

Kassambara, A., 2020. ggpubr: "ggplot2" Based Publication Ready Plots.

Kassambara, A., Mundt, F., 2017. factoextra: Extract and Visualize the Results of Multivariate Data Analysis.

Kiersch, K., Kruse, J., Regier, T.Z., Leinweber, P., 2012. Temperature resolved alteration of soil organic matter composition during laboratory heating as revealed by C and N XANES spectroscopy and Py-FIMS. Thermochim. Acta 537, 36–43. https://doi.org/10.1016/j.tca.2012.02.034

Kumar, M., Duffy, C.J., 2009. Detecting hydroclimatic change using spatio-temporal analysis of time series in Colorado River Basin. J. Hydrol. 374, 1–15. https://doi.org/10.1016/j.jhydrol.2009.03.039

Lees, K., Fitzsimons, M., Snape, J., Tappin, A., Comber, S., 2018. Soil sterilisation methods for use in OECD 106: How effective are they? Chemosphere 209, 61–67. https://doi.org/10.1016/j.chemosphere.2018.06.073

LeFevre, G.H., Novak, P.J., Hozalski, R.M., 2012. Fate of Naphthalene in Laboratory-Scale Bioretention Cells: Implications for Sustainable Stormwater Management. Environ. Sci. Technol. 46, 995–1002. https://doi.org/10.1021/es202266z

Mekid, S., Vaja, D., 2008. Propagation of uncertainty: Expressions of second and third order uncertainty with third and fourth moments. Measurement 41, 600–609. https://doi.org/10.1016/j.measurement.2007.07.004

Paine, C.E.T., Marthews, T.R., Vogt, D.R., Purves, D., Rees, M., Hector, A., Turnbull, L.A., 2012. How to fit nonlinear plant growth models and calculate growth rates: an update for ecologists: Nonlinear plant growth models. Methods Ecol. Evol. 3, 245–256. https://doi.org/10.1111/j.2041-210X.2011.00155.x

Petzoldt, T., 2019. growthrates: Estimate Growth Rates from Experimental Data.

Richards, F.J., 1959. A Flexible Growth Function for Empirical Use. J. Exp. Bot. 10, 290–301. https://doi.org/10.1093/jxb/10.2.290

Rousk, J., Bååth, E., 2011. Growth of saprotrophic fungi and bacteria in soil: Growth of saprotrophic fungi and bacteria in soil. FEMS Microbiol. Ecol. 78, 17–30. https://doi.org/10.1111/j.1574-6941.2011.01106.x

Schädel, C., Beem-Miller, J., Aziz Rad, M., Crow, S.E., Hicks Pries, C.E., Ernakovich, J., Hoyt, A.M., Plante, A., Stoner, S., Treat, C.C., Sierra, C.A., 2020. Decomposability of soil organic matter over time: the Soil Incubation Database (SIDb, version 1.0) and guidance for incubation procedures. Earth Syst. Sci. Data 12, 1511–1524. https://doi.org/10.5194/essd-12-1511-2020

Schlesinger, W.H., 1990. Evidence from chronosequence studies for a low carbon-storage potential of soils. Nature 348, 232–234. https://doi.org/10.1038/348232a0

Schmidt, M.W.I., Torn, M.S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I.A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D.A.C., Nannipieri, P., Rasse, D.P., Weiner, S., Trumbore, S.E., 2011. Persistence of soil organic matter as an ecosystem property. Nature 478, 49–56. https://doi.org/10.1038/nature10386

Sierra, C.A., Trumbore, S.E., Davidson, E.A., Vicca, S., Janssens, I., 2015. Sensitivity of decomposition rates of soil organic matter with respect to simultaneous changes in temperature and moisture. J. Adv. Model. Earth Syst. 7, 335–356. https://doi.org/10.1002/2014MS000358

Sinsabaugh, R.L., Manzoni, S., Moorhead, D.L., Richter, A., 2013. Carbon use efficiency of microbial communities: stoichiometry, methodology and modelling. Ecol. Lett. 16, 930–939. https://doi.org/10.1111/ele.12113

Sitnikov, D.G., Monnin, C.S., Vuckovic, D., 2016. Systematic Assessment of Seven Solvent and Solid-Phase Extraction Methods for Metabolomics Analysis of Human Plasma by LC-MS. Sci. Rep. 6. https://doi.org/10.1038/srep38885

Soulas, G., 1993. Evidence for the existence of different physiological groups in the microbial community responsible for 2,4-d mineralization in soil. Soil Biol. Biochem. 25, 443–449. https://doi.org/10.1016/0038-0717(93)90069-N

Spiess, A.-N., 2018. propagate: Propagation of Uncertainty.

Sun, S.-Q., Cai, H.-Y., Chang, S.X., Bhatti, J.S., 2015. Sample storage-induced changes in the quantity and quality of soil labile organic carbon. Sci. Rep. 5, 1–10. https://doi.org/10.1038/srep17496

Tarnawski, V.-R., Leong, W.H., 2000. Thermal Conductivity of Soils at Very Low Moisture Content and Moderate Temperatures. Transp. Porous Media 41, 137–147. https://doi.org/10.1023/A:1006738727206

Tjørve, K.M.C., Tjørve, E., 2017. The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family. PLOS ONE 12, e0178691. https://doi.org/10.1371/journal.pone.0178691

Trevors, J.T., 1996. Sterilization and inhibition of microbial activity in soil. J. Microbiol. Methods 26, 53–59. https://doi.org/10.1016/0167-7012(96)00843-3

Tsoularis, A., Wallace, J., 2002. Analysis of logistic growth models. Math. Biosci. 179, 21–55. https://doi.org/10.1016/S0025-5564(02)00096-2

Vitanov, N.K., Sakai, K., Dimitrova, Z.I., 2008. SSA, PCA, TDPSC, ACFA: Useful combination of methods for analysis of short and nonstationary time series. Chaos Solitons Fractals 37, 187–202. https://doi.org/10.1016/j.chaos.2006.08.043

Wang, C.M., Iyer, H.K., 2005. On higher-order corrections for propagating uncertainties. Metrologia 42, 406–410. https://doi.org/10.1088/0026-1394/42/5/011

Wang, W.J., Dalal, R.C., Moody, P.W., Smith, C.J., 2003. Relationships of soil respiration to microbial biomass, substrate availability and clay content. Soil Biol. Biochem. 35, 273–284. https://doi.org/10.1016/S0038-0717(02)00274-2

Wang, X., Helgason, B., Westbrook, C., Bedard-Haughn, A., 2016. Effect of mineral sediments on carbon mineralization, organic matter composition and microbial community dynamics in a mountain peatland. Soil Biol. Biochem. 103, 16–27. https://doi.org/10.1016/j.soilbio.2016.07.025

Wang, Y., Fan, Y., Gu, J.-D., 2004. Dimethyl phthalate ester degradation by two planktonic and immobilized bacterial consortia. Int. Biodeterior. Biodegrad. 53, 93–101. https://doi.org/10.1016/j.ibiod.2003.10.005

Wang, Y., Yin, B., Hong, Y., Yan, Y., Gu, J.-D., 2008. Degradation of dimethyl carboxylic phthalate ester by Burkholderiacepacia DA2 isolated from marine sediment of South China Sea. Ecotoxicology 17, 845. https://doi.org/10.1007/s10646-008-0247-4

Weihermüller, L., Neuser, A., Herbst, M., Vereecken, H., 2018. Problems associated to kinetic fitting of incubation data. Soil Biol. Biochem. 120, 260–271. https://doi.org/10.1016/j.soilbio.2018.01.017

Whitfield, C.J., Reid, C., 2013. Predicting surface area of coarse-textured soils: Implications for weathering rates. Can. J. Soil Sci. https://doi.org/10.1139/CJSS2012-009

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Source Softw. 4, 1686. https://doi.org/10.21105/joss.01686

Wolf, D.C., Skipper, H.D., 2018. Soil Sterilization, in: Methods of Soil Analysis. John Wiley & Sons, Ltd, pp. 41–51. https://doi.org/10.2136/sssabookser5.2.c3

Yanni, S.F., Diochon, A., Helgason, B.L., Ellert, B.H., Gregorich, E.G., 2018. Temperature response of plant residue and soil organic matter decomposition in soil from different depths. Eur. J. Soil Sci. 69, 325–335. https://doi.org/10.1111/ejss.12508

Zablotowicz, R.M., Krutz, L.J., Reddy, K.N., Weaver, M.A., Koger, C.H., Locke, M.A., 2007. Rapid Development of Enhanced Atrazine Degradation in a Dundee Silt Loam Soil under Continuous Corn and in Rotation with Cotton. J. Agric. Food Chem. 55, 852–859. https://doi.org/10.1021/jf0620923

Zwietering, M.H., Jongenburger, I., Rombouts, F.M., van 't Riet, K., 1990. Modeling of the Bacterial Growth Curve. Appl. Environ. Microbiol. 56, 1875–1881.