DNA viruses generally show higher phylogenetic divergence in polar regions and remain underrepresented in current databases

Vaibhav Kulkarni

A thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree,
Master of Science in Biology – Specialization in Bioinformatics

Department of Biology
Faculty of Science
University of Ottawa

© Vaibhav Kulkarni, Ottawa, Canada, 2026

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF COMMON ABBREVIATIONS

BH — Benjamini–Hochberg correction

CD-HIT — Cluster Database at High Identity with Tolerance

CoPHSe — Clusters of Putatively Homologous Sequences

FDR — False Discovery Rate

HMM — Hidden Markov Model

IMG/VR — Integrated Microbial Genomes/Virus database

LDG — Latitudinal Diversity Gradient

LG+Γ — Le & Gascuel substitution model + gamma-distributed rate heterogeneity

MAFFT — Multiple Alignment using Fast Fourier Transform

MGM — MetaGeneMark

MPD — Mean Pairwise Patristic Distance

MRCA — Most Recent Common Ancestor

N50 — Contig length N such that 50% of the assembly is in contigs ≥ N

NGS — Next-Generation Sequencing

PE — Paired-end (sequencing reads)

QC — Quality control

SRA — Sequence Read Archive

subs/site — substitutions per site

TSV — Tab-separated values (file format)

URL — Uniform Resource Locator

vOTU — viral Operational Taxonomic Unit

Abstract

Viruses dominate ecosystems worldwide, yet the broader phylogenetic space they occupy remains poorly represented in current reference databases. How environmental viral divergence varies across latitudes and biomes, and to what extent current references capture it, remains largely unquantified. This thesis addresses this research gap through a global-scale comparative analysis of divergence in viral sequences reconstructed from environmental DNA metagenomes, spanning marine and terrestrial biomes from pole to pole. The metagenomes (n = 180) were sampled equally across six biome–latitude treatments (marine and terrestrial × north, equatorial, and south regions; n = 30 each). Two principal patterns emerged in the analysis. First, reconstructed viral sequences exhibited significantly higher divergence than database sequences, indicating that existing references still fail to capture much of the global viral phylogenetic diversity. Second, most polar viromes displayed statistically higher divergence relative to non-polar regions, with Antarctic sequences consistently higher, whilst the Arctic Ocean was a prominent outlier. Collectively, these findings generally identify polar regions as reservoirs of highly divergent viral sequences and underscore the need to refine current databases by expanding sampling efforts. Since all samples were analysed using the same workflow, any methodological biases would have affected all treatments uniformly, rendering artefactual explanations unlikely. This thesis also establishes a reproducible framework alongside a temporal reference for global viral divergence patterns, particularly in the polar regions where rapid climate-change-driven ecological shifts are anticipated to be much more disruptive.

Résumé

Les virus dominent les écosystèmes mondiaux, pourtant le vaste espace phylogénétique qu'ils occupent reste mal représenté dans les bases de données actuelles. La mesure dans laquelle la divergence virale environnementale varie selon les latitudes et biomes, et sa capture par les références, demeure peu quantifiée. Cette thèse comble cette lacune par une analyse comparative mondiale de la divergence de séquences virales reconstruites à partir de métagénomes d'ADN environnemental (n = 180), couvrant les biomes marins et terrestres d'un pôle à l'autre. L'échantillonnage fut réparti sur six traitements biome–latitude (marin/terrestre × régions nord, équatoriale et sud; n = 30 chacun). Deux schémas principaux émergent. Premièrement, les séquences reconstruites présentent une divergence significativement plus élevée que celles de référence, indiquant que les bases actuelles ne captent pas l'essentiel de la diversité phylogénétique virale mondiale. Deuxièmement, la plupart des viromes polaires affichent une divergence statistiquement plus élevée, les virus antarctiques étant les plus divergents, tandis que l'océan Arctique constitue une exception notable. Collectivement, ces résultats identifient les régions polaires comme des réservoirs de séquences hautement divergentes, soulignant la nécessité d'affiner les bases de données par un échantillonnage élargi. Puisque tous les échantillons suivaient le même protocole, tout biais méthodologique aurait affecté les traitements uniformément, rendant une explication artefactuelle peu probable. Cette thèse établit un cadre reproductible et une référence temporelle pour la divergence virale mondiale, particulièrement dans les régions polaires où des changements écologiques rapides dus au climat sont anticipés.

Acknowledgements

First and foremost, I express my sincere gratitude to my supervisor, Dr. Stéphane Aris-Brosou, for his guidance, mentorship, and patience throughout this project. His input shaped both the conceptual framing and the analytical execution. I also thank the members of the Aris-Brosou lab for their collegiality and helpful discussions, especially Audrée Lemieux, whose foundational work and creativity inspired key elements of my analytical pipeline.

I am grateful to my Thesis Advisory Committee members, Dr. Francesco Marchetti and Dr. Xuhua Xia, for their constructive feedback throughout the research process. I also thank the University of Ottawa for providing the academic environment and institutional support that enabled me to complete my graduate studies.

I thank my examiners, Dr. Jonathan Lee and Dr. Theodore Perkins, for their careful evaluation and rigorous, constructive feedback, which materially improved the clarity and interpretive precision of my final thesis.

I gratefully acknowledge the Digital Research Alliance of Canada (formerly Compute Canada) for access to high-performance computing resources, including the Cedar and Graham clusters and their respective successors, Fir and Nibi, which enabled the large-scale analyses conducted in this thesis.

I am also grateful to my family and friends for their support and encouragement throughout this process. To all those who contributed in any way to this thesis, I extend my sincere thanks.

Chapter 1: Introduction

Background and Context

Viruses permeate every ecosystem, outnumbering cellular life by at least an order of magnitude and reaching an estimated 10^{31} particles globally, a scale that dwarfs all other biological entities (Mushegian, 2020; Wommack & Colwell, 2000). Viruses are integral to all ecosystems as they drive horizontal gene transfer and consequently modulate microbial evolution (Irwin et al., 2021). They also influence elemental cycles across marine, freshwater, and terrestrial realms. Viruses exert this influence through infection driven lysis and the resultant "viral shunt," which redistributes organic carbon and nutrients, thereby shaping food web structure and biogeochemical feedbacks that influence the climate system (Suttle, 2005; Tong et al., 2023). Their evolutionary dynamism—via mutation, recombination, and reassortment—generates a distinct reservoir of genetic novelty that continually reshapes host adaptability and ecosystem resilience (Broecker & Moelling, 2019; Hou et al., 2023).

Despite their ubiquity, viral genomes are chronically underrepresented in reference databases. More than 99% of the planet's viral diversity—the so-called "viral dark matter"—remains uncatalogued (Paez-Espino et al., 2016), prompting proposals such as the Global Virome Project which advocated for systematic exploration of unknown viral space (Carroll et al., 2018) and expansive reference databases for uncultivated viruses of unprecedented scale like the IMG/VR database (Camargo, Nayfach, et al., 2023). Nowhere is this knowledge gap more consequential than in high-latitude regions like the rapidly warming High Arctic, where surface air temperatures have risen nearly four times faster than the global average since 1979, a phenomenon termed "Arctic amplification" (Rantanen et al., 2022).

Extreme seasonality, oligotrophic waters, and strong physicochemical gradients in these regions impose selection pressures distinct from those at lower latitudes (Buscaglia et al., 2024; Calayag et al., 2025; Meng et al., 2023), making high-latitude ecosystems natural laboratories for studying the boundaries of viral adaptation. This, coupled with accelerated sea ice loss (Serreze & Stroeve, 2015), shifting hydrological cycles (Yang et al., 2021), and destabilizing permafrost (Vonk et al., 2015) creates highly dynamic niches that may foster the emergence and persistence of distinct viral lineages, while simultaneously releasing relic viruses locked in ice for millennia. Ancient giant viruses revived from Siberian permafrost (Alempic et al., 2023) and novel lineages flourishing in other extreme polar habitats (Heinrichs et al., 2024; Manuel & Snyder, 2024) underscore the urgency of developing a robust, quantitative framework for polar virology (Sommers et al., 2021).

Traditional Ecological Paradigms

Given these harsh environmental constraints and isolated host communities, traditional ecological theory (Hillebrand, 2004) would predict that polar viral diversity, in the richness-based sense, should be correspondingly low. The High Arctic, once perceived as a biological frontier with limited biodiversity (De Cárcer et al., 2015), has recently emerged as a region of unexpected viral richness and genetic distinctness (Labbé et al., 2020; Lemieux et al., 2024; Zhong et al., 2020). This contrast has prompted a reassessment of how viral systems conform to long-standing ecological paradigms, particularly the latitudinal diversity gradient (LDG), which predicts decreasing diversity toward the poles: attributed to declines in temperature, primary productivity, and evolutionary time in high-latitude regions (Hillebrand, 2004). For most free-living organisms, including many potential viral hosts, diversity does indeed peak at lower latitudes and diminishes toward the poles (Mittelbach et al., 2007). Consistent with this

paradigm, ecosystem-scale analyses indicate that viral community structure is strongly shaped by host phylogeny, with deep phylogenetic barriers constraining cross-species transmission (French et al., 2023).

However, recent large-scale environmental surveys have identified the Arctic as a previously underappreciated hotspot for viral diversity (Brum et al., 2015; Gregory et al., 2019; Labbé et al., 2020). Studies of Arctic marine and freshwater systems, such as Lake Hazen (a freshwater lake in the northern part of Ellesmere Island, Nunavut, Canada, north of the Arctic Circle) and the Arctic Ocean, have revealed unexpectedly high viral richness and distinct community structures (Emerson et al., 2018; Gregory et al., 2019; Lemieux et al., 2022, 2024). Notably, the Arctic has been recognized as one of five distinct global ecological zones based on viral community composition and diversity, due to its unique viral phylogenies that differ significantly from those observed in lower-latitude environments (Gregory et al., 2019). Metagenomic analyses consistently show that a substantial proportion of viral sequences recovered from High Arctic environments are novel, with few or no detectable homologs in existing viral databases, as determined by sequence similarity thresholds used in current viral classification tools (Emerson et al., 2018; Gregory et al., 2019; Lemieux et al., 2024). Phylogenetic analyses further indicate that these viruses are highly divergent, often clustering separately from known viral taxa and displaying unique evolutionary lineages (Lemieux et al., 2024). While methodological considerations including database representation bias (Nayfach et al., 2020; Roux et al., 2015) and sequencing artifacts (Rose et al., 2016) could potentially influence these observations, convergent evidence from multiple independent studies suggests that they reflect genuine biological phenomena, that warrants investigation (Calayag et al., 2025; De Cárcer et al., 2015; Labbé et al., 2020; Meng et al., 2023; Wang et al., 2022). Prior reports of

apparent LDG deviations in viral systems, including proposed polar “hotspots” (Anesio & Bellas, 2011), are largely grounded in OTU or population richness estimates and community-composition analyses (De Cárcer et al., 2015; Lemieux et al., 2024). While global surveys typically report standard latitudinal declines in viral richness (Angly et al., 2006; Brum et al., 2015), more recent studies reinforce the broader conclusion that viral ecological patterns (like microdiversity) can decouple from simple host-richness expectations (Gregory et al., 2019; Warwick-Dugdale et al., 2024), rather than constituting a direct contradiction of richness-based formulations of the LDG itself.

Host-Virus Relationship Effects

Host-dependent ecology provides a partial explanation for apparent departures from simple richness-based LDG expectations. In some Arctic systems, viral diversity appears structured more by host population dynamics, local adaptation, and phylogeography than by absolute host richness (Lemieux et al., 2024). Further, the extreme environmental conditions in the High Arctic—characterized by low temperatures, oligotrophy, and seasonal light cycles—exert strong selective pressures, promoting the emergence of unique viral lineages and high endemicity (Emerson et al., 2018; Lemieux et al., 2024). Geographic isolation exacerbated by the aforementioned factors further limits gene flow, facilitating genetic drift and the accumulation of novel mutations (C. Andrews, 2010). Arctic viral communities were found to be dominated by previously uncharacterized viruses, with a prevalence of single-stranded DNA viruses and bacteriophages, as commonly observed in many aquatic microbial communities (De Cárcer et al., 2015; Emerson et al., 2018; Gregory et al., 2019). The composition of viral assemblages varies across different Arctic environments, reflecting both habitat-specific selection and historical contingency, i.e., the influence of past colonization events, glaciation patterns, and evolutionary

trajectories that continue to shape present-day viral diversity. Another point of consideration is that although temperate oceans yield more total viral genomes because of both higher sampling effort and primary productivity (Angly et al., 2006; Gregory et al., 2019), metagenomic studies show that nearly 60% of Arctic viral operational taxonomic units (vOTUs) lack host predictions, exacerbating the knowledge gap (Calayag et al., 2025). Polar lakes, for instance, harbour a disproportionately high share of previously undescribed giant viruses—here defined as double stranded DNA viruses with genomes >300 kbp and capsids >200 nm—compared to temperate lakes (Pitot et al., 2024; Wilhelm et al., 2017). This enrichment may reflect the ability of giant viruses to carry extensive accessory gene repertoires that confer cold-adaptation and metabolic flexibility under oligotrophic, low-temperature conditions (Pitot et al., 2024). However, these unique ecological adaptations and the novel viral lineages they produce pose significant challenges for taxonomic classification and database representation.

Approach and Contribution of this Study

Despite recent advances, polar viral studies remain sparse, fragmented, and often methodologically inconsistent, impeding synthesis across habitats and hampering predictive ecological modelling. Subsequently, considerable gaps remain in our understanding of polar viral diversity (Zayed et al., 2022). A significant methodological gap in this field pertains to the underrepresentation of RNA viruses. Although RNA viruses are ubiquitous and ecologically important (Wu et al., 2022), their genomes degrade rapidly in Antarctic samples—viral RNA yields are vanishingly low (Trivedi et al., 2022), residual RNases remain active at subzero temperatures (Mittal et al., 2023), and prolonged transit on dry ice or in liquid-nitrogen shippers often involves freeze–thaw cycles that fragment RNA (Williams et al., 2024). Moreover, high particulate loads, salts, and inhibitors in ice, snow, and sediment clog concentration filters and

thwart reverse-transcription protocols (Hata et al., 2015; Sorrentino et al., 2025). Resultantly, most polar viral studies have focused on DNA-based metagenomes, leaving RNA virus diversity poorly resolved. Although my thesis was originally conceived to utilize RNA data, the paucity of Antarctic metatranscriptomic and RNA-seq datasets made a rigorous analysis unfeasible. Consequently, I focus here on DNA-based metagenomes to ensure consistent, cross-region comparisons of polar viral communities.

Considering these challenges and knowledge gaps, the study undertook a comparative global analysis of DNA viral metagenomes across the latitudinal gradient in both marine and terrestrial environments. This includes both polar (High Arctic and Antarctic) and non-polar (temperate and tropical) environments to assess the structure, novelty, and biogeographic (the study of how and why organisms and ecosystems are distributed across geographic space and through geological time) patterns of DNA viral communities.

Using a pipeline that reconstructs viral genes, clusters homologues, and quantifies divergence with phylogenetic metrics relative to IMG/VR (Camargo, Nayfach, et al., 2023), this study establishes a standardized framework for assessing database completeness and ecological distinctiveness across biomes. Overall, the study addresses three questions that structure the analysis. Firstly, how do viral community structures and divergence patterns differ across polar and non-polar environments? Secondly, I question to what degree do polar viral sequences diverge from existing reference sequences contained in current databases, with specific comparison to the most robust and extensive viral reference database available at the time of analysis (namely the IMG/VR v4.1)? Thirdly, does viral phylogenetic divergence exhibit coherent latitudinal structuring, and how does such structuring compare to (but neither test nor contradict) more traditional ecological paradigms, particularly the richness-based latitudinal

diversity gradient (LDG) that is ubiquitous among diverse biological systems? While my thesis quantifies these claims through a tree-based phylogenetic analysis, the ecological mechanisms underlying these patterns are interpreted cautiously and addressed as potentially testable hypotheses, backed up primarily through synthesis with existing literature rather than a direct empirical assessment.

Hypotheses and Thesis Statement

Hypothesis 1. The reconstructed viral sequences from environmental DNA samples exhibit greater phylogenetic divergence among themselves relative to their closest homologs identified in current reference databases (here, namely the IMG/VR v4.1).

Hypothesis 2. Reconstructed viral sequences from polar regions exhibit higher phylogenetic divergence among themselves, relative to their non-polar counterparts. Notably, this is independent of richness-based LDG expectations.

Thesis Statement. This thesis provides a latitudinally balanced global comparison of the structure of marine and terrestrial DNA viromes, quantifies patterns of their phylogenetic divergence from established reference databases and further establishes a methodological framework for future polar virology research in the context of accelerating climate change.

Chapter 2: Methods

Data Sourcing and Selection

The raw Next Generation Sequencing (NGS) data were obtained from the Sequence Read Archive (SRA) of the National Institute of Health. The SRA database, as part of the International Nucleotide Sequence Database Collaboration (INSDC) alongside the European Bioinformatics Institute (EBI), and the DNA Database of Japan (DDBJ), is the largest repository of publicly available high-throughput sequencing data (Sayers et al., 2022). Since its creation in 2009, the SRA database has grown significantly, driven by the rapidly decreasing cost of high-throughput sequencing and its widespread adoption across the life sciences. Currently, it boasts over 52 petabases across over 28.6 million sequencing runs (Shiryev & Agarwala, 2024).

The environmental DNA metagenomes were sampled and downloaded as paired-end (PE) reads from the SRA database. To test my hypotheses, the dataset was organised along two experimental axes. First, by global biome type: with (n = 90) samples across both marine and terrestrial biomes, allowing a direct and equitable comparison between oceanic and land-based environments. Second, by latitude: spanning the two polar extremes across through the non-polar tropics. North Polar (> 66.5° N), Equatorial/Temperate (between 66.5° N and 66.5° S), and South Polar (> 66.5° S) zones. Thirty samples were selected for every ecosystem-latitude zone. Within each such zone, geographic separation was maximized to capture broad environmental variability while keeping climate and biome representation comparable. The geographical coordinates of the sampling sites of the metagenomic assemblies that ultimately passed all quality control (QC) measures (n = 120) are shown in Figure 1. This initially balanced factorial design aimed to minimize regional biases.

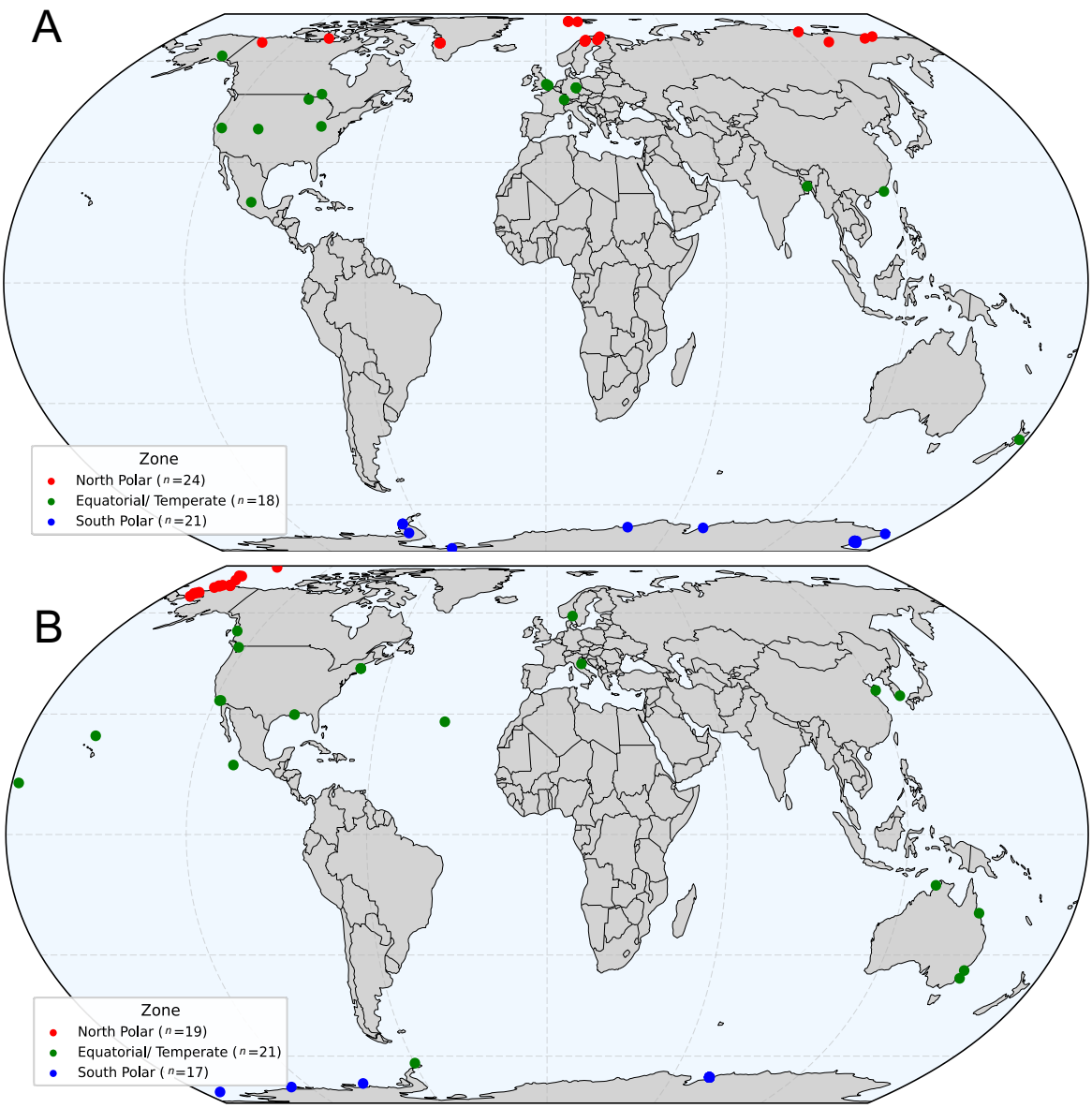


Figure 1. Global distribution of analyzed viral metagenomic samples. Free-earth projection maps illustrating the geographic locations of samples that passed assembly and quality control procedures. (A) Terrestrial (n = 63) and (B) Marine samples (n = 57) are shown separately, highlighting the latitudinal coverage and spatial representation of the datasets used in downstream analyses.

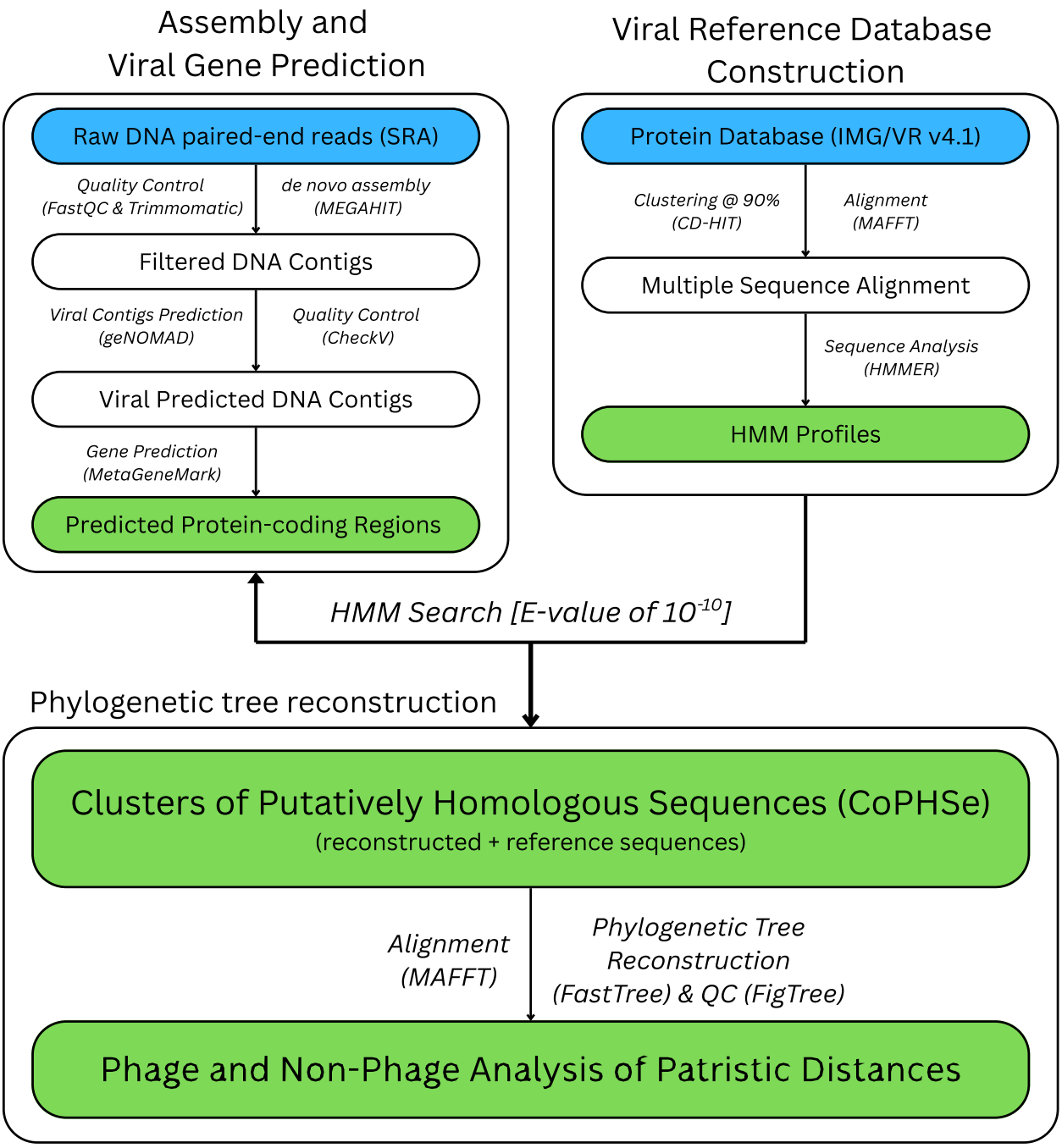


Figure 2. Viral metagenomic analysis pipeline. Raw DNA paired-end reads (SRA) were assembled and protein-coding regions predicted. Together with IMG/VR v4.1 reference proteins, HMM profiles were built to generate clusters of putatively homologous sequences (CoPHSe). Patristic distances were then computed for phage and non-phage groups.

Generating Clusters of Putatively Homologous Sequences

The overview of the data analysis pipeline is shown in Figure 2. Raw sequencing reads were downloaded using the SRA Toolkit v3.0.0 and underwent quality control using FastQC v0.12.1 to evaluate base quality profiles and identify sequence composition anomalies consistent with potential contaminants or sequencing artifacts (Andrews, 2010; SRA Toolkit Development Team, 2016). Adapter trimming and low-quality read removal were performed using Trimmomatic v0.39 under default parameters optimized for Illumina paired-end reads: (TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36) (Bolger et al., 2014). The cleaned reads were then *de novo* assembled using MEGAHIT v1.2.9 with default settings, generating contigs for downstream analysis (Li et al., 2015). The contigs were checked for broad comparability by verifying that their N50 values were not significantly different (Supplementary Figure 1). Viral contigs that passed this QC were identified using the 'end-to-end' command of geNomad v1.11.1 (Camargo, Roux, et al., 2023). The 'end-to-end' command within geNomad v1.11.1 orchestrates, in sequence, all analytical modules and writes a log and subdirectory for each stage of its analysis. This pipeline performs: (i) gene prediction and marker annotation; (ii) provirus boundary detection; (iii) two complementary classifiers (marker-based and sequence-only neural network); (iv) score aggregation; and (v) post-classification filtering plus reporting alongside the FASTA format exports that are used further downstream (Camargo, Roux, et al., 2023). The resulting outputs were further quality controlled with CheckV v1.0.3, keeping only medium and high quality viral contigs (Nayfach et al., 2020). The quality filtering steps had reduced the number of valid samples to 120, down from the original 180, ranging from 17-24 samples in each of the six treatments. Protein-coding regions within these contigs were predicted using MetaGeneMark (MGM) v3.38 (Zhu et al., 2010).

To frame these genes in a comparative context, a reference set of high-confidence viral proteins, the largest and most extensive of its kind (IMG/VR v4.1, release 2022-09-20_7.1) (Camargo, Nayfach, et al., 2023) was employed. This database was dereplicated with CD-HIT v4.8.1 at 90% identity to reduce within cluster redundancy whilst preserving clade structure (Fu et al., 2012; Li & Godzik, 2006). MAFFT v7.471 produced multiple-sequence alignments for these clusters (Katoh et al., 2002; Katoh & Standley, 2013), and HMMER v3.2.1 (Finn et al., 2011) converted alignments containing at least ten sequences into Hidden Markov Models (HMMs). Clusters with <10 sequences were excluded from HMM construction, as profiles derived from very small alignments risk overfitting and provide insufficient statistical power to capture conserved evolutionary patterns, thereby reducing reliability in downstream homology detection. These HMMs were then searched back against the MGM annotated proteins at an *E*-value threshold of 1×10^{-10} limiting spurious matches. These matches were clustered into "Clusters of Putatively Homologous Sequences" (CoPHSe) i.e., locus-level groups that are alignable by construction and therefore suitable for phylogenetic reconstruction (Lemieux et al., 2024). These CoPHSe were further aligned with MAFFT using the "--auto" option. Alignments containing at least one predicted gene and one or more reference sequences were filtered for quality using trimAL v1.4 with the "-gappyout" option to remove poorly aligned sites (Capella-Gutiérrez et al., 2009). Phylogenetic trees for each CoPHSe were constructed with FastTree v2.1.11, employing the LG +Γ substitution model which provides consistent, model-based branch length estimates at scale (Le & Gascuel, 2008; Price et al., 2010).

Assessment of Phylogenetic Trees

In most phylogenetic trees, the predicted viral sequences formed a single, easily identifiable clade (i.e., they were monophyletic), so no explicit rerooting was needed for them. The few trees

that were non-monophyletic were omitted from downstream analysis (see Supplementary Figure 2). Each tree file was read into R with the "ape" package function 'ape::read.tree()' (Paradis & Schliep, 2019). The trees containing at least two (>2) reconstructed sequences, and 2-400 IMG/VR sequences were parsed to generate two sub-trees: one containing only reconstructed sequence derived tips (identified by the string "gene") and one containing only reference database derived tips (identified by the string "IMGVR") (see Supplementary Figure 4). Computing patristic distances scales quadratically with the number of tips, because the pairwise distance matrix has $n(n-1)/2$ entries. In R this entails $O(n^2)$ time and $O(n^2)$ memory, with additional overhead from intermediate copies during garbage collection. To keep the analysis tractable across millions of trees, I imposed a uniform cap of 400 IMG/VR references per tree, which lies at the "elbow" or threshold point where computational costs begin to grow sharply due to the pathological nature of the calculations. The consistent cap at 400 reference tips therefore prevents a small number of extremely reference-rich trees from dominating compute time, while preserving the overwhelming majority of the dataset. For each sub-tree, i.e. the reconstructed and reference sequence trees, the mean pairwise patristic distance (MPD) was calculated. Patristic distance is defined as the sum of branch lengths along the unique path connecting two tips in a phylogenetic tree (Fourment & Gibbs, 2006). The MPD metric quantifies the average evolutionary separation among tips based on summed branch lengths and is therefore a measure of phylogenetic divergence. For each subtree, the most recent common ancestor (MRCA) was calculated using the getMRCA() function. Thus, the average bridging edge–which is the MRCA-to-MRCA path length between the reconstructed and reference subtrees–was calculated (see Supplementary Figure 4). Unlike MPD, this metric captures divergence between reconstructed and reference subtrees rather than within-clade structure.

Furthermore, the IMG/VR sequences were cross-referenced with a “lookup table” based on the metadata TSV file (available at the Joint Genome Institute’s portal: https://genome.jgi.doe.gov/portal/IMG_VR/IMG_VR.home.html) that contains host predictions and taxonomic predictions for the sequences and thus could be used to assign the trees as either “phage”, “non-phage” or “unknown”. Host predictions (7.2% of database) and taxonomy (96.7% of database) were not available for all sequences (Camargo, Nayfach, et al., 2023), and sequences lacking this data were marked as “unknown” and filtered out from downstream analysis. This workflow guaranteed consistent handling of well-behaved and problematic (unknown) trees alike, while providing a robust and internally consistent measure of phylogenetic divergence between assembled environmental contigs and their closest database homologues. The patristic distances at the biome level thus obtained were further subject to the Kruskal-Wallis test followed by pairwise Dunn tests for multiple comparisons with the Benjamini-Hochberg correction to control the false discovery rate (FDR) and ascertain the statistical significance of their phylogenetic divergence ($\alpha = 0.01$) (Benjamini & Hochberg, 1995).

To complement rank-based significance testing and to quantify the magnitude and direction of regional contrasts, non-parametric effect sizes were calculated. Specifically, the Hodges–Lehmann estimator was used to estimate the median shift between groups, providing a robust, distribution-free measure of location difference that is consistent with rank-based inference (Hodges & Lehmann, 1963). In parallel, Cliff’s delta (δ) was computed as an ordinal effect size describing the probability that a randomly selected value from one group exceeds a randomly selected value from another group, minus the reverse probability (Cliff, 1993). Cliff’s δ ranges from -1 to $+1$, with values near zero indicating substantial overlap between

distributions and larger absolute values indicating stronger stochastic dominance. These effect sizes were calculated specifically for the reconstructed sequence MPD values and are reported alongside sample sizes for each contrast. Together, the Hodges–Lehmann estimator and Cliff's δ provide complementary, distribution-free summaries of effect size that remain valid under skewed, heteroscedastic, and unequal-variance conditions, and are therefore well suited to large, non-Gaussian datasets like the metagenomes I use in this thesis (Cliff, 1993; Hodges & Lehmann, 1963; Wilcox, 2012). Non-phage datasets were sufficiently small to permit effect-size estimation using all available per-tree MPD values. Because the Hodges–Lehmann estimator and Cliff's δ quantify distributional displacement rather than inferential significance (Cliff, 1993, 1996; Hodges & Lehmann, 1963), representative subsampling preserves effect-size interpretation while avoiding sample size inflation effects associated with very large datasets (Wilcox, 2012). Thus, for my phage datasets, which comprise of a very large numbers of per-tree reconstructed sequence MPD estimates, effect sizes were computed using a fixed-size, representative subsample of MPD values per region to ensure computational tractability. Subsamples consisted of 1,333 MPD values per region (from a total budget of 4,000 points distributed equally across three metrics) remaining well below the n = 5,000 threshold at which dominance-matrix computation becomes prohibitively slow as noted in recent literature (Meissel & Yao, 2024). Subsamples were drawn using a fixed random seed (seed = 123) to ensure future reproducibility; these values correspond to the jittered points shown in the boxplots o Figures 3 and 4. This subsampling was applied exclusively for effect-size estimation and visualization.

Chapter 3: Results

Dataset Quality Control and Filtering

A total of 180 environmental metagenomes were assembled using MEGAHIT v1.2.9, 30 in each treatment. Assembly quality was evaluated using the N50 statistic, which represents the contig length at which half of the assembly is contained in contigs of equal or greater size. Using the study-wide significance threshold ($\alpha = 0.01$), N50 values were checked for comparability at the regional scale i.e. no statistically significant differences among them. This was done by ensuring that Kruskal–Wallis tests indicated no significant differences (Terrestrial: $H = 5.12$, $p = 0.077$; Marine: $H = 6.54$, $p = 0.038$; $\alpha = 0.01$), ensuring that downstream comparisons were not biased by assembly contiguity (Supplementary Figure 1).

Of the initial 180 assemblies, 1 terrestrial and 16 marine samples failed to meet minimum quality thresholds and were excluded, leaving 163 assemblies for viral screening. The filtered assemblies were then processed with geNomad v1.11.1 to identify viral contigs and remove non-viral sequences. Quality control measures were applied with CheckV v1.0.3, which excluded incomplete or low-quality viral genomes. After these filtering steps, the final dataset consisted of 120 samples with high quality viral sequences, distributed across biomes and regions. The summary of this refined dataset is presented in Supplementary Tables 1 and 2 where the SRA sample accessions IDs are listed alongside their geographic coordinates. The geospatial distribution of this finalized dataset which formed the basis of all downstream tree generation and phylogenetic analyses can be seen in the Equal Earth projection of world maps in Figure 1.

Overview of Dataset and Analyses

The 120 samples were distributed evenly across regions, with 17–24 assemblies per region spanning both marine and terrestrial biomes. The reconstructed viral genes from these samples

were then organized into homologous clusters with their closest database matches (CoPHSe) represented on phylogenetic trees. Globally, 5,924,864 gene-family trees were initially inferred. As described in the Methods, trees that were non-monophyletic were filtered out; Supplementary Figure 2 shows that this step removed <1% of trees globally and consistently <1.5% for each biome-region subset. To maintain computational tractability of patristic-distance calculations, I imposed a uniform cap of 400 IMG/VR references per tree (i.e., reference subsampling was applied only in extremely reference-rich cases). This threshold affected a very small fraction of the total tree pool: 20,775 trees (0.35%) exceeded the cap and were excluded, leaving 5,904,089 trees (99.65%) for downstream filtering steps (Supplementary Figure 3). From this reduced pool, additional tree inclusion criteria were applied to ensure consistent patristic-distance estimation, namely the presence of a minimum of two (>2 tips) of each reconstructed and IMG/VR reference sequences per tree (to form a clade). These constraints reduced the initial tree pool by approximately two-thirds (68.8% excluded). A complete quantitative breakdown of tree counts retained and excluded at each filtering stage is provided in Supplementary Table 3. Ultimately, a total of 1,846,393 trees were assessed for viral phylogenetic divergence. Three complementary metrics were derived from the branch lengths within each tree, which included the MPD of the reconstructed sequences, the MPD of the reference sequences and the "bridging edge" connecting the MRCA of the two clusters (as seen in Table 1). A schematic of the tree-based analytical framework is shown in Supplementary Figure 4. Each comparison involved large sample sizes, often numbering hundreds of thousands of tree-level divergence estimates, ensuring robust statistical power. Hereafter, their divergence unit of “substitutions/site” is abbreviated as “subs/site” for brevity.

Table 1. The median patristic distances calculated across the regions and biomes.

Biome	Taxonomy	Region	Reconstructed sequences	Bridging edge	IMG/VR sequences	Total trees (*n*)
			Patristic distance (substitutions/site)			
Terrestrial	Phage	North	2.176	0.844	0.027	175,182
		Equator	1.971	0.861	0.027	267,462
		South	2.223	0.834	0.028	115,839
	Non-phage	North	1.790	1.076	0.024	925
		Equator	1.578	1.231	0.026	587
		South	2.270	0.834	0.033	648
Marine	Phage	North	1.320	0.652	0.037	116,226
		Equator	1.458	0.655	0.032	710,703
		South	1.912	0.710	0.030	454,700
	Non-phage	North	0.126	1.530	0.024	190
		Equator	1.690	0.779	0.029	2,354
		South	1.967	0.873	0.029	1,577

Table 2. Non-parametric effect sizes for regional contrasts in reconstructed-sequence MPD (subs/site) across marine and terrestrial viromes, stratified by phage and non-phage trees. Values report Hodges–Lehmann median shifts and Cliff's δ, with sample sizes (n_1, n_2) and corresponding Figure 3–4.

Biome	Taxonomy	Regional Contrast	Hodges–Lehmann shift (subs/site)	n_1	n_2	Cliff's δ	Figure panel
Terrestrial	Phage	Equator vs North	-0.149	1333	1333	-0.116	3A
		Equator vs South	-0.224	1333	1333	-0.174	3A
		North vs South	-0.077	1333	1333	-0.062	3A
	Non-phage	Equator vs North	-0.101	587	925	-0.07	3B
		Equator vs South	-0.58	587	648	-0.406	3B
		North vs South	-0.486	925	648	-0.367	3B
Marine	Phage	Equator vs North	0.176	1333	1333	0.131	4A
		Equator vs South	-0.384	1333	1333	-0.278	4A
		North vs South	-0.545	1333	1333	-0.355	4A
	Non-phage	Equator vs North	1.059	2354	190	0.684	4B
		Equator vs South	-0.264	2354	1577	-0.176	4B
		North vs South	-1.568	190	1577	-0.88	4B

Global Patterns of Viral Genetic Divergence

Across all phylogenetic trees, reconstructed environmental sequences exhibited substantially greater divergence than their IMG/VR reference counterparts. This result held across both biomes and all tree types, confirming a strong and consistent global pattern (Table 1; Figures 3 and 4). Median values across the dataset showed a consistent hierarchy of values. Reconstructed sequences had the highest medians, ranging typically between 1.3 and 2.3 subs/site (with the notable exception of the marine North Polar non-phage partition at 0.13 subs/site, discussed below). Bridging edges were generally between 0.6 and 1.2 subs/site, though two non-phage values exceeded this range. IMG/VR references were consistently the lowest at ~0.02–0.04 subs/site across the board.

In marine equatorial phages for example, reconstructed genes reached a median of 1.46 subs/site, while IMG/VR homologues were only 0.03 subs/site, with bridging edges at 0.66 subs/site. Similarly, in terrestrial equatorial phages, reconstructed medians were higher (~1.97 subs/site), compared to IMG/VR references at 0.03 subs/site, again with bridging edges being intermediate (~0.86 subs/site). Thus, across all datasets, environmental reconstructions occupied a broader and more divergent genetic space, while reference homologues remained tightly clustered. Bridging edges consistently linked these two extremes, reinforcing that novel environmental clades are distantly but detectably related to known viruses.

These results provide strong support for Hypothesis 1, which predicted that environmental reconstructions would display greater divergence than database sequences. They also highlight the incompleteness of current viral references, since even well-sampled databases like IMG/VR captured only a narrow subset of the total genetic diversity observed.

Terrestrial Systems

Patterns of viral divergence in terrestrial environments paralleled those observed in marine systems but displayed even clearer amplification at high latitudes. Across Equatorial, North Polar, and South Polar zones, reconstructed sequences consistently exhibited greater divergence among them than IMG/VR references, with bridging edges occupying intermediate positions (Table 1; Figure 3).

Terrestrial phages were represented by n = 267,462 trees in the Equatorial/Temperate region, n = 175,182 in the North Polar region, and n = 115,839 in the South Polar region. In the equatorial zone, reconstructed sequences had median divergences near 2.0 subs/site, compared to IMG/VR references at 0.03 subs/site and bridging edges at ~0.86 subs/site. Divergence intensified toward the poles, with reconstructed medians exceeding 2.1 subs/site in the North Polar region and reaching ~2.2 subs/site in the South Polar region. Bridging edges rose accordingly (to ~0.8–0.85 subs/site), while IMG/VR references remained narrowly constrained near 0.03 subs/site. These differences in reconstructed sequences and bridging edges were highly significant across all regions (Kruskal–Wallis, $\alpha = 0.01$), with Dunn post-hoc tests confirming pairwise significance between Equatorial, North Polar, and South Polar assemblages (all adjusted $P < 0.01$).

Non-phage trees were represented by smaller but statistically robust sample sizes: n = 587 in the Equatorial/Temperate region, n = 925 in the North Polar region, and n = 648 in the South Polar region. Despite these lower totals, the same pattern was observed: reconstructed medians approached ~1.6 subs/site in the equatorial zone, approached 1.8 subs/site in the North Polar region, and were highest (2.27 subs/site) in the South Polar zone. Bridging edge values consistently decreased polewards (Equator 1.231 → North 1.076 → South 0.834), while

IMG/VR reference trees consistently remained the lowest around ~0.03 subs/site. These differences in South Polar reconstructed sequences and bridging edges were highly significant compared to the North Polar and Equatorial regions (Kruskal–Wallis, $\alpha = 0.01$), with Dunn post-hoc tests confirming pairwise significance between Equatorial, North Polar, and South Polar assemblages (all adjusted $P < 0.01$).

A small number of trees could not be confidently classified as phage or non-phage. These groups showed more variability due to limited sample sizes ($n < 1{,}000$), but the overall pattern remained consistent: reconstructed sequences were most divergent, bridging edges intermediate, and IMG/VR homologues least divergent. For example, in South Polar unclassified trees, reconstructed sequences had a median of 2.38 subs/site, bridging 1.21 subs/site, and IMG/VR references 0.003 subs/site.

Non-parametric effect sizes for latitudinal contrasts in reconstructed sequence MPD are reported in Table 2 and visualized in Figures 3 and 4. In terrestrial phages (Figure 3A), pole-to-equator shifts were directionally consistent but modest: Equatorial/Temperate vs North Polar shows a Hodges–Lehmann shift of −0.149 subs/site with Cliff's $\delta = -0.116$, and Equatorial/Temperate vs South Polar shows −0.224 subs/site with $\delta = -0.174$; North vs South differences are smaller still (−0.077 subs/site; $\delta = -0.062$). These values indicate a subtle latitudinal displacement of the reconstructed divergence distribution rather than strong group separation. In terrestrial non-phages (Fig. 3B), the equator-to-south contrast is substantially larger (Eq vs South: −0.580 subs/site; $\delta = -0.406$), and the North vs South contrast is similarly pronounced (−0.486 subs/site; $\delta = -0.367$), whereas Eq vs North remains comparatively small (−0.101 subs/site; $\delta = -0.070$). Overall, terrestrial effect sizes therefore indicate that the strongest

latitudinal displacement is found in the South Polar contrast for non-phage trees, while phage shifts are relatively weaker in magnitude.

Marine Systems

Patterns of viral divergence within marine systems were generalizable with global trends but also revealed clear latitudinal structure. Across Equatorial, North Polar, and South Polar zones, reconstructed environmental sequences were markedly more divergent than IMG/VR references, while bridging edges consistently occupied intermediate values (Table 1; Figure 4).

Marine phages were represented by very large sample sizes of trees: n = 710,703 in the equatorial zone, n = 116,226 in the North Polar zone, and n = 454,700 in the South Polar zone. In the equatorial zone, reconstructed sequences had a median divergence of 1.46 subs/site, compared to 0.03 subs/site for IMG/VR references and 0.66 subs/site for bridging edges. In the North Polar zone, reconstructed medians were 1.32 subs/site, with bridging edges at 0.65 subs/site and IMG/VR references at 0.037 subs/site. The South Polar zone showed the greatest divergence, with reconstructed medians of 1.91 subs/site, bridging edges at 0.71 subs/site, and IMG/VR references again limited to 0.03 subs/site. These results indicate a pronounced polar amplification of divergence in marine phages, with Antarctic assemblages the most distinct. These differences in South Polar reconstructed sequences were highly significant compared to both the North Polar and Equatorial regions, which also differed from one another (Kruskal–Wallis, $\alpha = 0.01$). Dunn post-hoc tests confirmed pairwise significance among all three regions for reconstructed sequences (adjusted $P < 0.01$). For bridging edges, South Polar values were significantly higher than North Polar and Equatorial values, while North and Equator were statistically comparable. IMG/VR references remained uniformly low but were also significantly distinct across regions (adjusted $P < 0.01$).

Non-phage trees were represented by smaller but sufficient sample sizes: n = 2,354 in the equatorial zone, n = 190 in the North Polar zone, and n = 1,577 in the South Polar zone. In the equatorial zone, reconstructed sequences had a median divergence of 1.69 subs/site, bridging edges 0.78 subs/site, and IMG/VR references only 0.03 subs/site. In the North Polar zone, reconstructed medians were markedly lower at 0.13 subs/site, while bridging edges were much higher at 1.53 subs/site and IMG/VR references remained low at 0.024 subs/site. As in phages, the South Polar zone yielded the highest divergence among non-phages, with reconstructed sequences at 1.97 subs/site, bridging edges at 0.87 subs/site, and IMG/VR references at 0.029 subs/site. For non-phage datasets, South Polar reconstructed sequences were significantly higher than both North Polar and Equatorial regions, which also differed from each other (Kruskal–Wallis, $\alpha = 0.01$; Dunn post-hoc adjusted $P < 0.01$). In bridging edges, Equatorial and South Polar values were statistically comparable, but both were significantly elevated relative to North Polar. IMG/VR references again remained uniformly low with no significant pairwise differences among regions (same significance letter “e”).

Similar to the terrestrial data, a small number of trees could not be confidently classified as phage or non-phage. These groups showed more variability due to limited sample sizes (n < 1,500), but the general overall pattern remained consistent: reconstructed sequences were most divergent, bridging edges intermediate, and IMG/VR homologues least divergent. For example, in South Polar unclassified trees, reconstructed sequences had a median of 1.57 subs/site, bridging of 0.72 subs/site, and IMG/VR references at 0.002 subs/site.

An exception to the otherwise consistent latitudinal structuring of divergence was observed in the marine north region. Using the same non-parametric magnitude estimates on the representative jitter subsample like their terrestrial counterparts, marine phages (Figure 4A)

showed Equatorial/Temperate divergence exceeds North Polar divergence (Eq − North Hodges–Lehmann = +0.176 subs/site; δ = +0.131), while South Polar divergence is elevated relative to both Equatorial/Temperate (Eq − South = −0.384 subs/site; δ = −0.278) and North Polar regions (North − South = −0.545 subs/site; δ = −0.355). In contrast, marine non-phages (Figure 4B) show a much stronger separation involving the North Polar region, consistent with the markedly low North Polar reconstructed median in Table 1: Eq − North is +1.059 subs/site with δ = +0.684, and North − South is −1.568 subs/site with δ = −0.880, whereas Eq − South is comparatively modest (−0.264 subs/site; δ = −0.176).

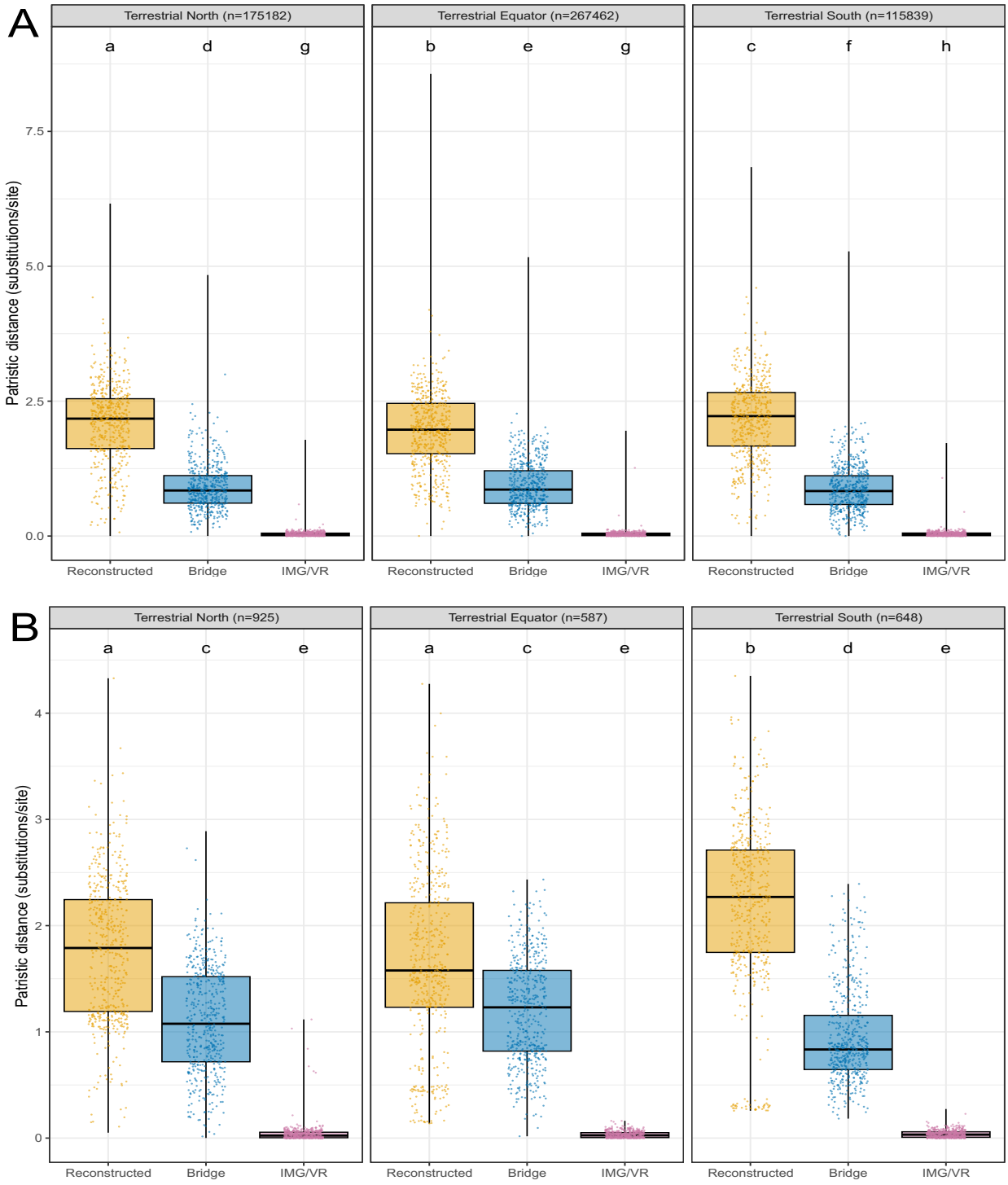


Figure 3. Patristic distance distributions in terrestrial viral communities. Boxplots show phylogenetic divergence for (A) phage sequences and (B) non-phage sequences in terrestrial samples. Distances are derived from reconstructed environmental sequences compared with IMG/VR v4.1 references, highlighting differences in evolutionary divergence across groups. Statistical significance was assessed using Kruskal–Wallis tests followed by Dunn post hoc tests with Benjamini–Hochberg correction. Different letters above the boxplots denote statistically significant differences between groups at $\alpha = 0.01$, df = 8, $P < 0.01$ (phage: $\chi^2 = 1{,}351{,}608.14$; non-phage: $\chi^2 = 4898.15$).

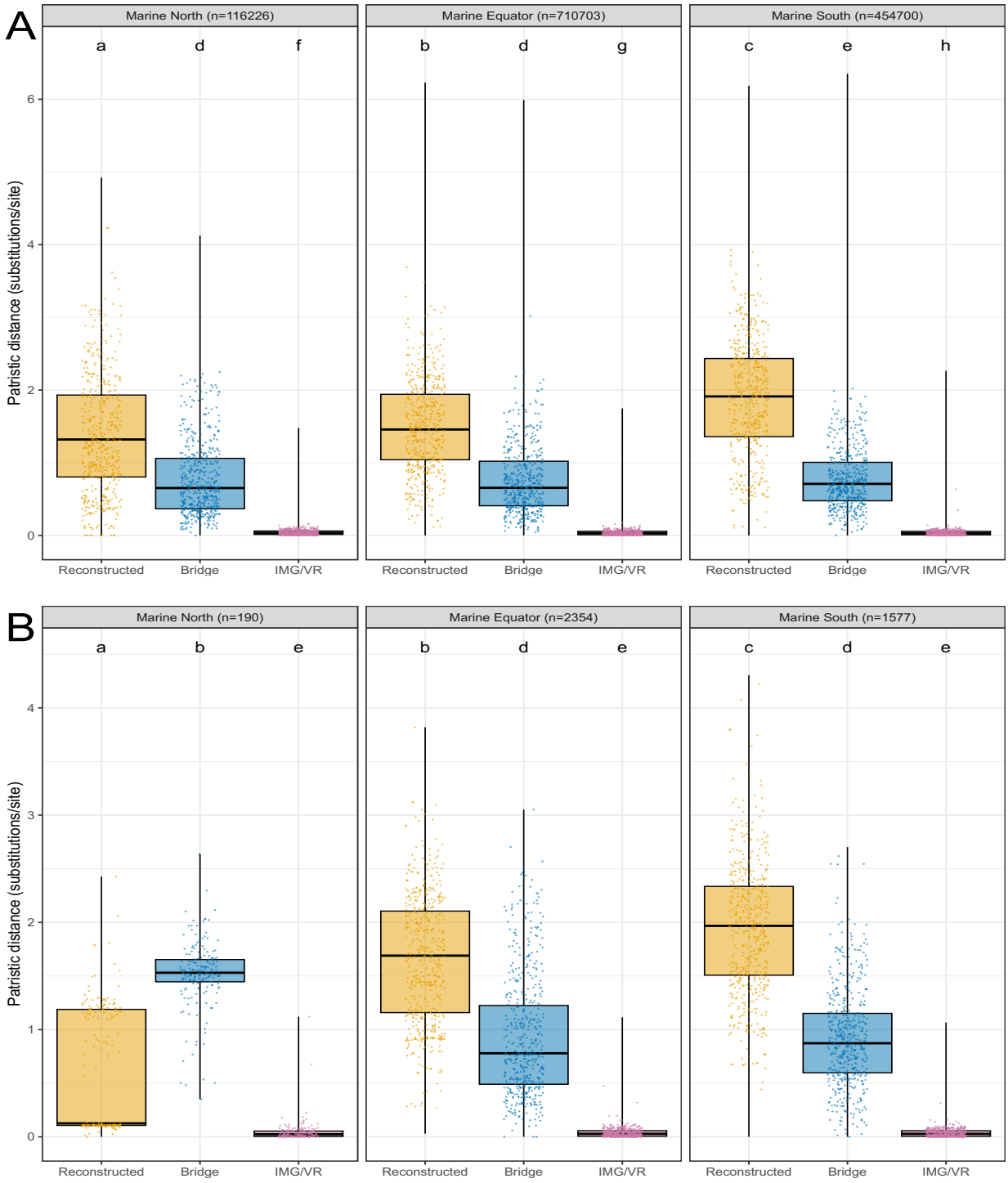


Figure 4. Patristic distance distributions in marine viral communities. Boxplots show phylogenetic divergence for (A) phage sequences and (B) non-phage sequences in marine samples. Distances were calculated from reconstructed environmental sequences relative to IMG/VR v4.1 references. Statistical significance was assessed using Kruskal–Wallis tests followed by Dunn post hoc tests with Benjamini–Hochberg correction. Different letters above the boxplots denote statistically significant differences between groups at $\alpha = 0.01$, df = 8, $P < 0.01$ (phage: $\chi^2 = 2{,}944{,}648.14$; non-phage: $\chi^2 = 9{,}576.11$).

Chapter 4: Discussion

Key Findings

This study provides a global comparison of phylogenetic divergence across marine and terrestrial ecosystems for DNA viruses. Two central patterns and one important qualifier emerge that refine how viral phylogenetic divergence and database representation should be interpreted at a global scale. Firstly, reconstructed environmental sequences consistently exhibited higher MPDs than their closest homologues in the IMG/VR v4.1 database. The average bridging edges generally occupied an intermediate position, with the sole exception occurring in marine non-phages in the north. This supports Hypothesis 1 and adds quantitative evidence for the pervasive "viral dark matter" not adequately represented in current reference databases (Krishnamurthy & Wang, 2017; Paez-Espino et al., 2016; Santiago-Rodriguez & Hollister, 2022). The closer clustering of IMG/VR reference sequences compared to the reconstructed ones (Figures 3 and 4) underscores the enduring incompleteness of current viral databases. These findings are consistent with more recent reports of viral dark matter, especially from polar environments. For instance, a study in Prydz Bay in Antarctica found that 85–90% of viral contigs in their environmental samples were unaffiliated with any known viruses (Gong et al., 2018). Similarly, Wang et al. (2022) found that the dominant viral community of *Microviridae* sequences from Arctic soils showed 58.62%–63.46% identity to known members of *Microviridae*, indicating substantial sequence divergence of this viral community in Arctic polar environments.

Secondly, reconstructed polar sequences generally exhibited higher phylogenetic divergence among them relative to non-polar datasets within the same biome. Quantification of non-parametric effect sizes revealed that the strength of this elevation varied substantially across taxonomic partitions, with modest shifts in phage datasets and larger distributional separations in

non-phage datasets, particularly in South Polar regions. This finding lends qualified support to Hypothesis 2: although a poleward increase appears across the majority of partitions (≈75%), it is not uniformly expressed across all biome-region combinations with the notable exception of the marine north. These findings suggest that high-latitude viral lineages, particularly in the Antarctic regions, may occupy undocumented regions of phylogenetic space (characterized by long branch lengths and sparse reference representation). This pattern is potentially shaped by a combination of long-term geographic isolation, uneven dispersal (Meng et al., 2023), and persistent database representation biases (Lemieux et al., 2024). Importantly, the effect sizes of these regional contrasts must be interpreted in light of the analytical scale and data structure of environmental metagenome datasets. Here, divergence is estimated within CoPHSe families that already span broad evolutionary depth, such that substantial overlap among regional distributions is expected (Roux et al., 2015; Sullivan, 2015), allowing large absolute differences in patristic distances to coexist with small rank-based effect sizes. Rank-based effect sizes can remain small even when differences are consistent, particularly in our datasets characterized by high baseline variance (due to spatiotemporal heterogeneity) and extensive dispersal regimes (Brum et al., 2015; Gregory et al., 2019). Consistent with this expectation, most regional contrasts exhibited small to moderate Cliff's δ values alongside strong statistical support, indicating modest yet systematic shifts in evolutionary separation that are biologically meaningful (Cliff, 1993; Meissel & Yao, 2024). Taken in isolation, no single terrestrial phage contrast would constitute strong evidence for a latitudinal trend; the interpretive weight of the pattern derives instead from its directional consistency across independent gene families and across both phage and non-phage partitions, which collectively reduces the probability that the signal reflects a systematic but trivial bias.

A key point of consideration is that richness-based diversity and phylogenetic divergence are related but non-equivalent descriptors of biological variation: classical formulations of the latitudinal diversity gradient (LDG) primarily concern taxonomic richness (and its correlates), whereas phylogenetic divergence metrics quantify evolutionary separation among lineages (Hillebrand, 2004; Tucker et al., 2017). In phylogenetic terms, diversity can be represented by branch length based measures of evolutionary history (for example, Faith's phylogenetic diversity, which is defined as the sum of the branch lengths connecting all taxa in the sample) and by distance-based summaries of relatedness (Faith, 1992; Kembel et al., 2010; Tucker et al., 2017), but high species richness does not inherently imply high divergence because communities dominated by recent radiations can appear more densely clustered together (spatially clustered together), thereby reducing average pairwise distances even as richness rises (Cavender-Bares et al., 2009; Webb et al., 2002). Conversely, communities with fewer taxa can exhibit elevated divergence when lineages are old, isolated, or evolutionarily dispersed (Tucker et al., 2017; Webb et al., 2002). For this reason, the present thesis quantifies divergence explicitly as patristic distance-based MPD on homologous protein phylogenies, which is a measure of evolutionary separation and should not be interpreted as a proxy for richness or community diversity in the conventional ecological sense (Kembel et al., 2010; Webb et al., 2002).

In principle, divergence in coding and non-coding regions can be assessed separately when complete, consistently annotated viral genomes are available. In practice, this separation is not generally tractable for metagenome-assembled viral datasets. This is because many recovered viral sequences are incomplete genome fragments (García-López et al., 2015; Roux et al., 2015; Smits et al., 2015; Tao et al., 2022), and their completeness varies substantially across samples (Nayfach et al., 2020; Roux et al., 2018). Additionally, DNA virus genomes, on average, only

contain ≈10% non-coding regions (Mahmoudabadi & Phillips, 2018). Under these conditions, genome architecture features outside conserved coding loci, including the delineation and comparison of intergenic intervals, are often not interpretable in a uniform manner across distantly related viruses (Nayfach et al., 2020; Worobey, 2000). Accordingly, the present study restricts phylogenetic analysis to homologous protein-coding regions, for which multiple sequence alignment and model-based branch length estimation are well defined within locus-level gene families (Balinda et al., 2010; Roux et al., 2018). This approach reflects a key difference between viral and cellular phylogenetics: whereas cellular lineages can be anchored by universally conserved marker genes, viruses lack a single gene shared across all lineages (Brüssow, 2009; Khot et al., 2020; Nayfach et al., 2020; Paez-Espino et al., 2016), making protein-family-based tree reconstruction the most coherent method of phylogenetic inference for diverse environmental viromes (Harris & Hill, 2021; Roux et al., 2015; Sullivan, 2015).

A related point of consideration is why the analysis produces many phylogenetic trees rather than a singular global tree per site or per biome. In this pipeline, each tree corresponds to one CoPHSe, i.e., one putatively homologous protein family for which a multiple sequence alignment and branch lengths are meaningful (Lemieux et al., 2024). Additionally, viral evolution is frequently shaped by modular genomes and pervasive gene exchange (Harris & Hill, 2021; Khot et al., 2020), making a single species-like phylogeny spanning all viruses difficult to define in the same sense as cellular ribosomal-gene phylogenies (Brüssow, 2009; Harris & Hill, 2021). Constructing one very large tree per biome would therefore require forcing many sequences into a common alignment despite lacking shared homologous characters, thereby conflating distinct gene histories and rendering branch lengths uninterpretable. Because phylogenetic inference is fundamentally alignment-dependent, metagenomic viral datasets are

commonly analyzed using locus- or family-level units where homology is supported, rather than as a single whole-genome tree (Czech et al., 2022; Darling et al., 2014). Under this framework, viruses appearing in different CoPHSe trees are not being excluded from valid pairwise comparisons; rather, their pairwise patristic distances are undefined because such distances are only interpretable within a single phylogeny inferred from a homologous alignment. Accordingly, the appropriate analysis is statistical aggregation across a large "forest" of gene-family trees, which has been explicitly advocated as a conceptually coherent representation of viral evolutionary history (Harris & Hill, 2021). This is complementary to genome-wide gene-sharing network analyses that have been developed precisely because a universal, scalable tree framework is not available for uncultivated viral diversity (Bin Jang et al., 2019; Simmonds et al., 2023).

A reasonable interpretive question raised by these results is why biome context appears to modulate phylogenetic divergence, particularly given that both marine and terrestrial viromes exhibit strong novelty relative to reference databases. One parsimonious hypothesis is that the marine–terrestrial contrast reflects differences in habitat connectivity, dispersal regimes, and the spatial scale at which environmental filtering operates (Brum et al., 2015; Graham et al., 2024; Santos-Medellin et al., 2022). In broad terms, terrestrial environments are physically discontinuous and compartmentalized, whereas marine systems are embedded in a comparatively continuous fluid medium, a distinction recognized as a primary driver of microbial biogeography (Martiny et al., 2006; Vos et al., 2013). These differences are expected to shift gene flow, and thus the balance between drift and homogenization, for both viruses and their microbial hosts, thereby altering the distribution of phylogenetic divergence among reconstructed viral lineages (Hanson et al., 2012; Nemergut et al., 2013). In terrestrial systems, dispersal limitation can be

strong even across small geographic distances because soils and sediments are highly structured at fine spatial scales, and because viral particles interact directly with the soil matrix. Empirical studies of soil viromes consistently report pronounced distance–decay relationships and strong spatial structuring, patterns that are compatible with restricted dispersal and localized community assembly (Santos-Medellin et al., 2022; Trubl et al., 2018). This structuring is further reinforced by the heterogeneous physical and physicochemical architecture of the soil matrix—the primary substrate for terrestrial assemblages—which promotes high spatial turnover in viral assemblages over short distances and creates semi-isolated microhabitats in which local evolutionary trajectories can persist (Graham et al., 2024; Vos et al., 2013). Mechanistically, retention processes can also reduce viral mobility: viral particles carry surface charges and commonly adsorb to mineral and clay surfaces, with adsorption strength modulated by pH, ionic strength, and surface electrostatics, thereby limiting effective transport through soil pore networks (Florent et al., 2022; Murray & Laband, 1979). Collectively, these features plausibly reduce effective connectivity among terrestrial viral populations and increase the opportunity for divergence to accumulate within and among regions, even when analyses are aggregated across heterogeneous viral families. In contrast to terrestrial fragmentation, marine systems operate under a markedly different physical regime. Viral assemblages are routinely transported by advection and mixing, such that dispersal can be comparatively high and gene flow can counteract strong local drift across broad spatial scales (Brum et al., 2015; Sunagawa et al., 2015). Large-scale ocean virome surveys explicitly emphasize that viral communities are shaped by the joint action of passive transport and environmental selection, with water mass properties exerting a dominant organizing influence (Brum et al., 2015; Gregory et al., 2019). Importantly, however, marine connectivity is not unlimited: water masses, density gradients, and oceanic fronts act as semi-

permeable barriers that structure microbial and viral communities, while depth and water mass identity are major discriminating variables in marine viral biogeography (Bauer et al., 2018; Sunagawa et al., 2015). Under this framework, marine divergence is expected to reflect a balance between homogenizing transport and stratifying oceanographic structure, potentially yielding weaker or more variable regional divergence patterns than in terrestrial systems, particularly when spatial sampling within a region is limited.

Within the present dataset, this ecological framing provides a coherent context for the observed biome-specific divergence patterns. Terrestrial systems display a clearer and more consistent poleward amplification of phylogenetic divergence, as reflected in higher reconstructed sequence MPD values and directionally consistent non-parametric shifts across both phage and non-phage trees (Table 1; Table 2; Figure 3). Whereas marine systems exhibit a more heterogeneous pattern in which Antarctic assemblages are consistently more divergent but Arctic marine divergence is comparatively muted. In terrestrial viromes, equator-to-pole contrasts are uniformly negative, with modest but consistent Hodges–Lehmann shifts and small Cliff's δ values in phages ($\delta \approx -0.12$ to -0.17) and substantially larger effects in non-phage trees (δ up to -0.41), indicating directional but uneven amplification of divergence with latitude (Table 2; Figure 3). These quantitative patterns are consistent with strong dispersal limitation and localized lineage sorting in terrestrial systems, where restricted connectivity can amplify regional divergence even when absolute effect sizes remain modest (Bauer et al., 2018; Hanson et al., 2012; Nemergut et al., 2013; Sunagawa et al., 2015). By contrast, marine phage trees show weaker or more variable latitudinal contrasts (Table 2; Figure 4A), consistent with partial homogenization via background transport, while marine non-phage trees exhibit pronounced Antarctic divergence but a muted Arctic signal, reflected in asymmetric North–South shifts ($\delta \approx$

−0.88) driven by low Arctic divergence values and limited sample sizes (Table 1; Table 2; Figure 4B). Accordingly, the muted Arctic marine signal is consistent with, but not exclusively attributable to, a regime in which background transport reduces the strength of regional differentiation, compounded by limited circumpolar sampling that further constrains the detectable breadth of within-region heterogeneity, particularly for non-phage trees. Whether the low Arctic marine divergence reflects genuine biological attenuation, a sampling artefact, or some combination of the two cannot be resolved with the present data.

Novel Contributions

This study provides the first systematic, quantitative comparison of phylogenetic divergence across both Arctic and Antarctic regions using a standardized methodology. Earlier work focused on single regions (Calayag et al., 2025; De Cárcer et al., 2015) or specific viral groups (Meng et al., 2023). Moreover, many global surveys have relied on clustering thresholds or marker gene inventories rather than tree-based, quantitative divergence metrics (Brum et al., 2015). By applying patristic distances derived under the LG +Γ substitution model—which provides consistent, model-based branch length estimates across diverse protein families while remaining computationally tractable at scale— across marine and terrestrial systems simultaneously on a global scale, this thesis builds upon prior work with a unified phylogenetic framework (Le & Gascuel, 2008).

Importantly, the results demonstrate that database limitations extend beyond underrepresentation of novel viruses. Even with substantial improvements in IMG/VR v4.1 which led to a ~6-fold increase in database size over its last revision (now containing over 15 million virus genomes and genome fragments coming from 25,603 metagenomes, 6,755 metatranscriptomes and 62,342 isolate prokaryotic genomes among others) (Camargo, Nayfach,

et al., 2023), environmental sequences, particularly from polar regions, occupy phylogenetic space not captured by its reference collections. This highlights the need for targeted sampling of the neglected and under-sampled habitats and host ranges rather than relying on expansion from already sampled regions alone.

Finally, the consistently higher internal divergence in South Polar samples compared to North Polar ones indicates that polar viral communities are not evolutionarily uniform across hemispheres. This asymmetry is further supported by non-parametric effect size estimates, which show consistently larger south–equator and south–north shifts compared to their north-equator shift, across most partitions. This likely reflects the greater geographic isolation of Antarctica, suggesting that viral communities, like other polar organisms, are shaped by asymmetric biogeographic pressures. Unlike the Arctic, which maintains broad north-south connections across continental land masses and ecoclimatic zones facilitating species dispersal through wind, water currents, and migrating animals, Antarctica is both oceanographically and meteorologically isolated from the rest of the globe (Kleinteich et al., 2017). A recent study by Piedade et al. (2024) found that Antarctic Ocean communities harbour predominantly endemic viruses adapted specifically to polar environments and that 75% of Antarctic marine viral species are novel and not found elsewhere. Their study further found complex seasonal patterns in viral community composition that was attributed to their specific host interactions.

Building on Methodological and Ecological Frameworks

This phylogenetic framework builds upon established phylogenetic approaches used in viral ecology but applies them in a comparative context globally that aims to resolve inconsistencies noted in prior polar studies, namely the lack of close homolog matches and variable proportions of viral sequences across samples (Santiago-Rodriguez & Hollister, 2022; Wang et al., 2022). By

standardizing comparisons across ecosystems, it aims to provide a template for global analysis of viral phylogenetic divergence.

Ecologically, my results are aligned with findings that polar environments foster unique viral adaptations. While Meng et al. (2023) documented repeated adaptation in giant viruses to polar conditions over 100 times throughout evolutionary history, with polar-adapted viruses scattered across their phylogeny rather than forming monophyletic clades. They identified numerous viral genes specifically associated with polar adaptation that differ from polar-adaptive genes in their eukaryotic hosts, indicating distinct viral evolutionary strategies for cold adaptation. The high divergence that was observed suggests that polar environments might harbour novel functional capabilities, consistent with discoveries of ancient viruses revived from permafrost, which similarly contain high proportions of orphan genes with unknown functions (Alempic et al., 2023). In this study, the researchers successfully revived 13 infectious viruses from Siberian permafrost samples dated up to 48,500 years old, representing five different viral clades: Pandoravirus, Cedratvirus, Megavirus, and Pacmanvirus families, plus additional Pithovirus strains. These discoveries revealed key novel functional capabilities: (1) remarkable preservation of infectivity and structural integrity over tens of thousands of years, with viruses maintaining characteristic replication cycles and morphological features; and (2) discovery of highly divergent viral strains like *Pacmanvirus lupus*, which contained 43.7% orphan genes with no known homologs, suggesting vast undiscovered functional diversity in ancient viral reservoirs.

Limitations and Considerations

While this analysis captures large-scale patterns across biomes and latitudes, several limitations are to be acknowledged and merit consideration. The thesis focused exclusively on DNA viruses

because RNA viral datasets remain sparse and methodologically challenging to incorporate on a global scale. As brought to light earlier, the study by Wu et al. (2022) noted the importance of RNA viruses in polar ecosystems, particularly in thawed permafrost, while Trivedi et al. (2022) documented the technical challenges of preserving viral RNA in polar samples.

The temporal dimension highlighted by Calayag et al. (2025), who documented strong seasonality in Arctic viral communities, provides important context for my findings. Their observation that Arctic viral assemblages shift dramatically between seasons suggests that my current analyses may only capture a fraction of the total temporal viral diversity in polar regions. This limitation arises because collection dates were not consistently available across all sequencing runs, precluding a robust seasonal partitioning of the data. As a result, my analyses represent single "snapshots" of viral community structure rather than seasonally resolved profiles. This temporal variability may partially explain the high within-region divergence that was observed and suggests that future studies incorporating seasonal dynamics could reveal even greater viral diversity in polar systems.

Another important source of uncertainty, distinct from methodological variability, arises from uneven geographic representation, particularly in the marine North Polar region. Following quality control and filtering, most North Polar marine samples were derived from only two bioprojects within the SRA. Namely, bioprojects PRJNA681031 and PRJEB14154—the former of which accounts for the overwhelming majority (~95%, 18 out of 19) of the Arctic samples. Interestingly, they are both associated with the University of Alaska Fairbanks and appear clustered together. As can be seen in Figure 1, the samples in the North Polar marine region are densely clustered along a narrow spatial corridor rather than broadly distributed across. Such spatial clustering is a well-recognized limitation in global marine viromics, where sampling

effort is often logistically constrained and unevenly distributed across regions (Brum et al., 2015; Gregory et al., 2019; Ladau & Eloe-Fadrosh, 2019). This restricted spatial coverage likely limits environmental heterogeneity and may attenuate detectable phylogenetic divergence, as reduced habitat and host variability can compress evolutionary signal in comparative analyses (Nayfach et al., 2020; Roux et al., 2015). However, a genuine biological attenuation of divergence in the Arctic Ocean, driven for instance by stronger oceanographic mixing or younger post-glacial colonization histories, cannot be excluded and remains an open question. This interpretation is, consistent with prior cautions advised against over-interpretation of metagenomic analyses of sparsely sampled regions of the earth (Ladau & Eloe-Fadrosh, 2019). Accordingly, the absence of strong polar signal in the marine Arctic cannot be taken as concrete evidence against it. Considering all these factors, the North Polar marine estimates can be regarded as comparatively less generalizable and treated as conservative until broader circumpolar sampling becomes available (Brum et al., 2015; Gregory et al., 2019). In contrast, Equatorial and South Polar datasets benefit from wider spatial coverage, reinforcing the importance of geographically distributed sampling for robust inference of large-scale viral divergence patterns (Nayfach et al., 2020; Roux et al., 2015). More generally, although non-parametric effect sizes are robust to unequal group sizes (Hess & Kromrey, 2004), extreme imbalances can indeed increase sampling variability, and the observed contrasts involving relatively small sample sizes should be interpreted cautiously (Romano et al., 2006).

This thesis aggregates phylogenetic divergence across diverse viral families, hosts, habitats, and independently generated metagenomic datasets, introducing substantial within-group heterogeneity. Latitude functions as a coarse organizing variable, and reconstructed-sequence MPD integrates signals from both deeply isolated and cosmopolitan lineages. Under

these conditions, the statistically robust shifts observed—modest in phage datasets but larger in non-phage datasets—remain vulnerable to residual biases in sampling, assembly, and database representation. The patterns observed are thus interpreted cautiously. It is important to acknowledge that some of the absolute values can be artefactual yet remain biologically meaningful. Assembly heuristics, particularly in low-complexity regions, can inflate divergence estimates by generating fragmented or chimeric contigs (Peona et al., 2020). Likewise, classification tools such as geNomad and CheckV, while state-of-the-art, rely on probabilistic models that may misassign boundaries or miscategorize contigs (Camargo, Roux, et al., 2023), thereby exaggerating branch lengths. Database underrepresentation compounds these issues, as sparse reference coverage can artificially extend inferred distances, with saturation of extreme branch lengths further inflating some absolute values. Artefactual inflation should act uniformly across samples, as all were processed with the same pipeline, assembly parameters, and reference database. Yet divergence was not random. It showed coherent latitudinal structure (with the prior noted exception of the marine north partitions). This structured pattern was reproducible across most biome partitions and persisted across phage and non-phage datasets. Moreover, bridging edge distances followed a similar same, demonstrating that reconstructed clades are not disconnected outliers, but form evolutionarily separated yet traceable lineages relative to the reference space. These consistent rank-ordered differences across independent datasets are unlikely to be attributable to random technical noise alone. Instead, they are more parsimoniously consistent with genuinely greater biological divergence of polar viral lineages, particularly pronounced in both Antarctic biomes and terrestrial high-latitude systems, likely reflecting the combined effect of long-term geographic isolation, reduced gene flow, and unique selective pressures characteristic of extreme polar ecosystems.

Significance

These findings underscore the need to broaden the representation of polar viruses in reference databases and to refine approaches for viral community characterization and phylogenetic divergence assessment. The persistent gap between reconstructed sequences and IMG/VR references indicates that current databases incompletely capture global viral evolutionary diversity, creating risk if they are used uncritically for ecological inference and predictive modelling. High-latitude datasets, most consistently those from the South Polar region, showed elevated reconstructed-sequence patristic distances relative to reference sequences (Figures 3–4; Table 1), suggesting that polar assemblages contribute disproportionately to undersampled and phylogenetically divergent viral lineages. Although the drivers of this divergence cannot be resolved here, the observed heterogeneity is compatible with prior work showing that Arctic viral communities are strongly correlated with prokaryotic community composition and associated physicochemical gradients, implying that polar ecological context may shape patterns of viral divergence and turnover (Calayag et al., 2025).

As polar regions continue to warm rapidly (Rantanen et al., 2022), the persistence and distribution of these distinct polar viral lineages may be altered, increasing the importance of establishing present-day baselines. Moreover, as the permafrost thaws, it may release previously dormant or isolated viral communities into the polar environments (Wu et al., 2022), making these findings a valuable reference point for monitoring change through time. Beyond polar systems, the methodological framework offers a standardized approach for assessing database completeness and phylogenetic divergence across ecosystems. By providing a replicable template for cross-study comparison, it enables more consistent benchmarking of reference database coverage and more interpretable global syntheses in viral ecology.

Conclusions and Future Directions

This thesis establishes a global, latitudinal comparison of viral eDNA phylogenetic divergence across marine and terrestrial systems, with particular emphasis on polar regions that remain underrepresented in reference databases. The analysis demonstrates that environmental viral sequences consistently occupy broader phylogenetic space than existing references, exposing persistent gaps in viral databases. It also reveals that polar viromes diverge from one another, with Antarctic assemblages showing the strongest distinctiveness across biomes. This distinctiveness is reflected in both higher median divergence and broader divergence distributions relative to lower latitudes, most prominently in non-phage viral assemblages. These findings do not contradict richness-based formulations of the latitudinal diversity gradient. Rather, they indicate that polar environments harbour deeply divergent and under-sampled viral lineages whose evolutionary histories are not captured by existing references. By providing a standardized tree-based framework for quantifying phylogenetic divergence at a global scale, this work establishes a foundation for advancing polar virology and refining broader ecological theory.

Future progress will depend on coordinated efforts along three complementary axes: improving temporal resolution through richer metadata and repeated sampling; expanding analytical scope to include functional and host-associated dimensions; and refining phylogenetic models to better capture viral evolutionary processes. Prioritizing the sampling of RNA viral datasets remains critical, as RNA viruses are markedly underrepresented in public repositories, especially in polar environments. Such efforts are essential not only for improving viral reference databases but also for understanding how microbial ecosystems, especially in rapidly warming polar regions, respond and adapt to environmental change.

Future analyses of patristic distances in viruses can be enhanced by building on and expanding the current tree-based phylogenetic framework developed here. I propose that the revised methodological framework should aim to incorporate a more accurate codon-based substitution model or a comparably accurate yet computationally efficient "hybrid" substitution model in lieu of the current amino acid-based LG model. Building on the analytical framework developed here, future analyses of viral patristic distances can be further strengthened through methodological refinement. In this study, I analyzed viral proteins using the LG +Γ substitution model—an empirically derived amino-acid substitution matrix with among-site rate heterogeneity—which is efficient and widely used but carries assumptions that warrant caution when interpreting viral evolutionary dynamics (Le & Gascuel, 2008). First, its time-reversibility (and, by extension, stationarity and homogeneity) can underrepresent directional processes that arise from host–pathogen arms races (Sironi et al., 2015) or strand-specific mutational biases (Sianga-Mete et al., 2022), both of which can yield non-reversible substitution patterns in viruses. Second, the assumption of site independence cannot capture interactions and correlations like epistasis that are pervasive in viral proteins (Sanjuán et al., 2004). Third, the LG model was estimated primarily from alignments of cellular proteins from Pfam (Le & Gascuel, 2008), and this bias in training data could mean its exchangeability parameters may not accurately reflect the distinct selection pressures faced by viruses—such as immune evasion, host-switching, and rapid ecological changes (Rochman et al., 2022). These factors can drive path-dependent and complex evolutionary trajectories in viral proteins (Dickinson et al., 2013), which the LG model might not capture well. These caveats imply that absolute branch lengths may be locally biased under model misspecification; however, applying a single, consistent model across datasets preserves internal comparability, so relative contrasts remain informative for global analysis even

when perfection in absolute rate estimation is unattainable. Future work could instead incorporate codon substitution models in their phylogenetic analysis as they can also account for positive (or Darwinian) selection as opposed to amino acid substitution models that are constrained to estimating negative (or purifying) selection (Doron-Faigenboim & Pupko, 2007). In their publication Zaheri et al. (2014) proposed a generalized mechanistic codon model that generated more accurate phylogenetic trees in comparison to amino acid substitution models, even in highly divergent species. However, the authors also noted that generalized codon-based models are much more computationally intensive as they must compute 61 x 61 substitution matrices (upto 3721 parameters) as opposed to 20 x 20 matrices (upto 400 parameters) for amino acids. This creates a 9.3-fold increase in matrix size (3,721 vs 400 entries), but the computational impact scales much worse than linearly. Matrix operations scale as $O(n^3)$ for standard algorithms, meaning codon model likelihood calculations are theoretically ~28 times more computationally expensive per operation than amino acid models (61^3 vs 20^3). In practice, the performance penalty is substantial even with optimizations and thus was not utilized in my analysis computing millions of trees. Finally, while patristic distances effectively capture divergence, they do not resolve any ecological interactions. Complementary approaches, such as functional gene analysis (to identify environment specific adaptations) and virus–host network reconstruction that leverages co-occurrence patterns and phylogenetic relationships to identify putative viral-host associations in complex microbial communities (Calayag et al., 2025; Lemieux et al., 2022; Meng et al., 2023), could provide finer resolution.

Also, the data curation from the SRA for subsequent analyses should involve greater scrutiny in its QC filtering steps, specifically by accounting for the availability of the sample collection dates (or a seasonal timeframe) in their associated metadata. This allows future

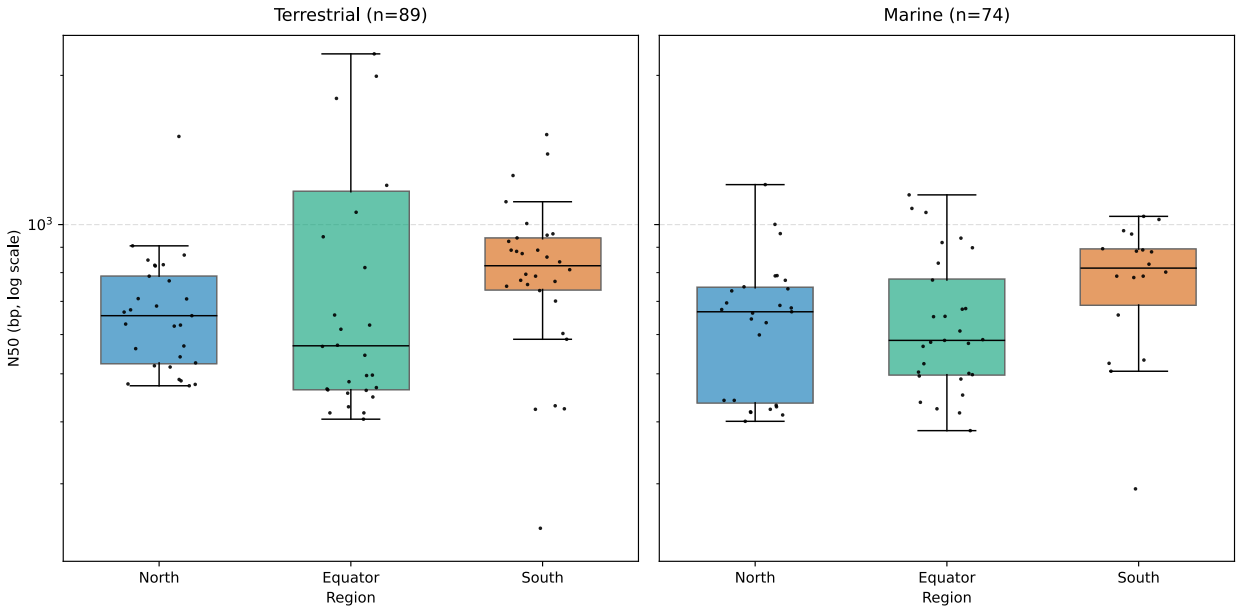
research to add a temporal dimension that can infer seasonal variations, like the ones observed in viral community composition—highlighted earlier—could also be present for genetic divergence values at the regional level. Repeated sampling from the same locations on a seasonal basis would provide the most comparable results if feasible. Moreover, greater analytical resolution can be achieved by coupling divergence analyses with functional gene profiling. Functional gene signatures often reveal environment-specific adaptations—such as stress response, nutrient acquisition, or host-interaction mechanisms—that provide ecological context beyond taxonomic structure. Integrating these signatures with virus–host network reconstruction, which combines co-occurrence patterns with phylogenetic associations, can help infer candidate host linkages within complex microbial assemblages. This integrative approach would yield a more mechanistic and ecologically grounded understanding of how viral divergence interacts with both temporal variability and functional landscape structure across ecosystems.

Finally, applying this enhanced analytical pipeline across global ecosystems will sharpen the resolution and interpretability of viral evolutionary patterns and establish a conceptual and methodological scaffold for exploring the forces shaping viral diversification at planetary scale. By integrating phylogenetic divergence, temporal metadata, functional gene signatures, and virus–host ecological networks, future work can move beyond static diversity inventories toward a dynamic understanding of how viral communities are assembled, structured, and reshaped through time. Such integration will allow detection of seasonal rhythms, ecological contingencies, and evolutionary innovations across biomes, with polar regions serving as natural laboratories for observing viral responses to rapid environmental change. A globally unified framework will enable researchers to quantify how warming, shifting host ranges, and altered biogeochemical cycles influence viral diversification and ecological roles. This multi-scale

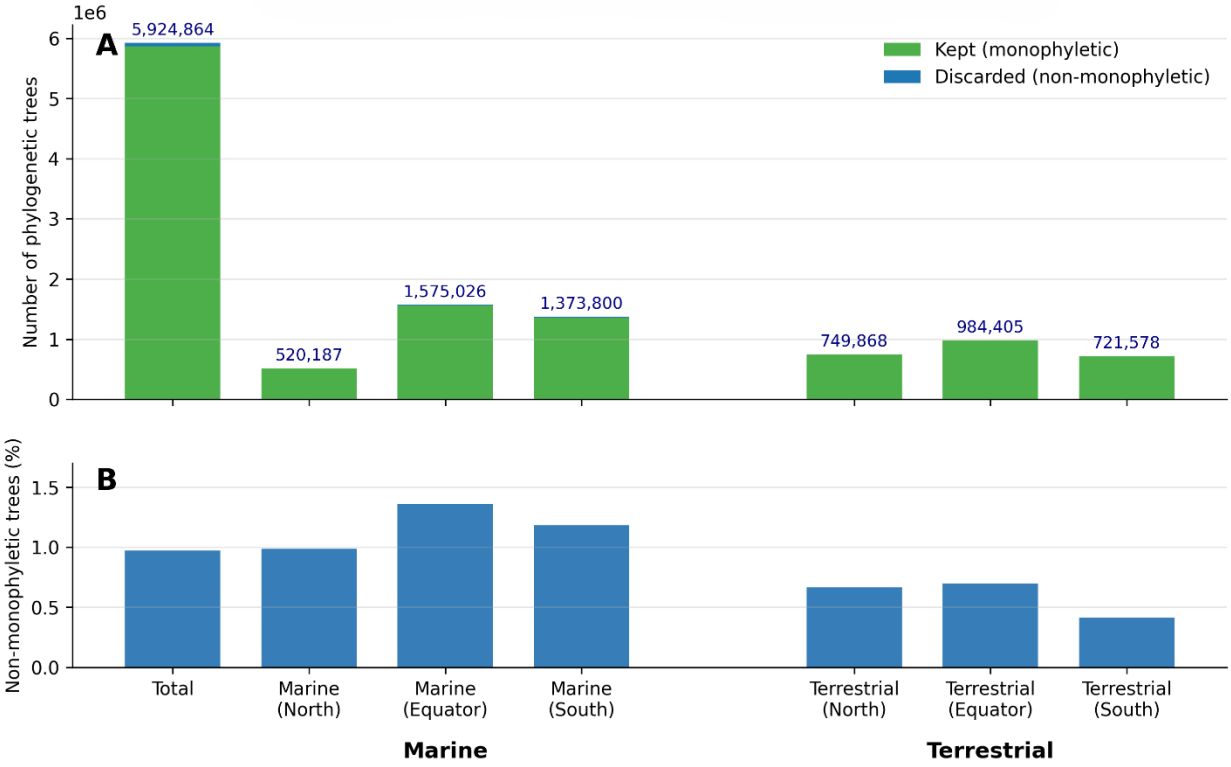
approach transforms viral ecology from a fragmented descriptive field into a predictive science, linking evolutionary trajectories to ecological processes and environmental change. In doing so, it will provide critical insight into how viruses shape—and are shaped by—ecosystem dynamics. Polar virology, once peripheral to global syntheses, will become a strategic testbed for forecasting virosphere responses to climate perturbation. Ultimately, integrating evolutionary, temporal, and functional dimensions will allow the scientific community to not only map the contours of viral diversity but also anticipate its future pathways in a rapidly changing world. These proposed extensions are beyond the scope of the present thesis but represent natural and tractable next steps building on its framework.

Appendices

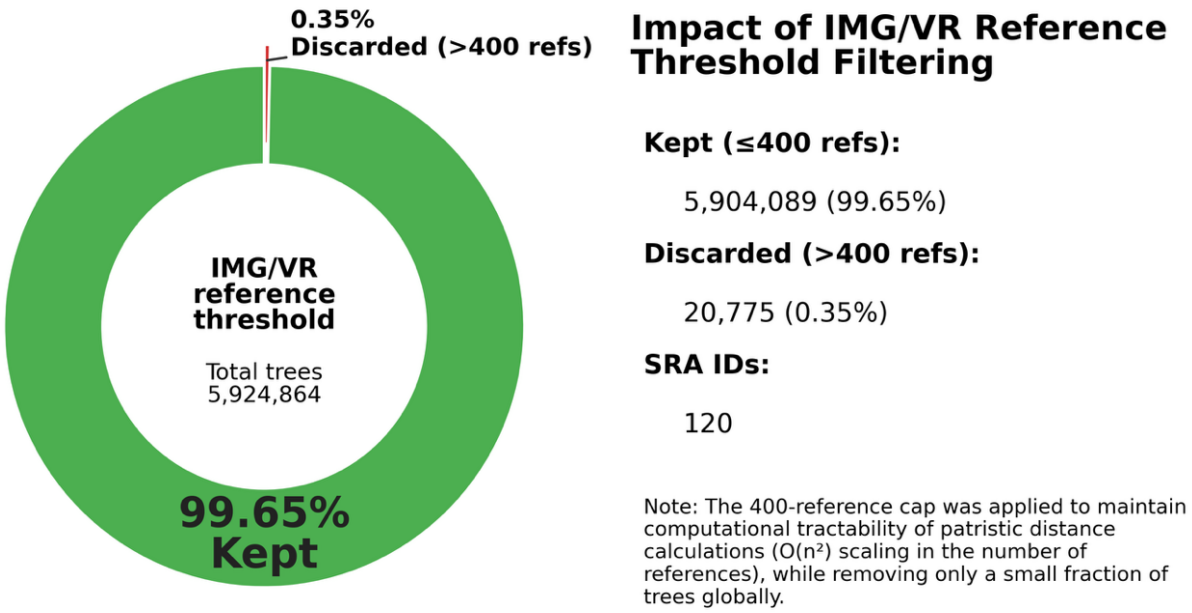
Supplementary Figures



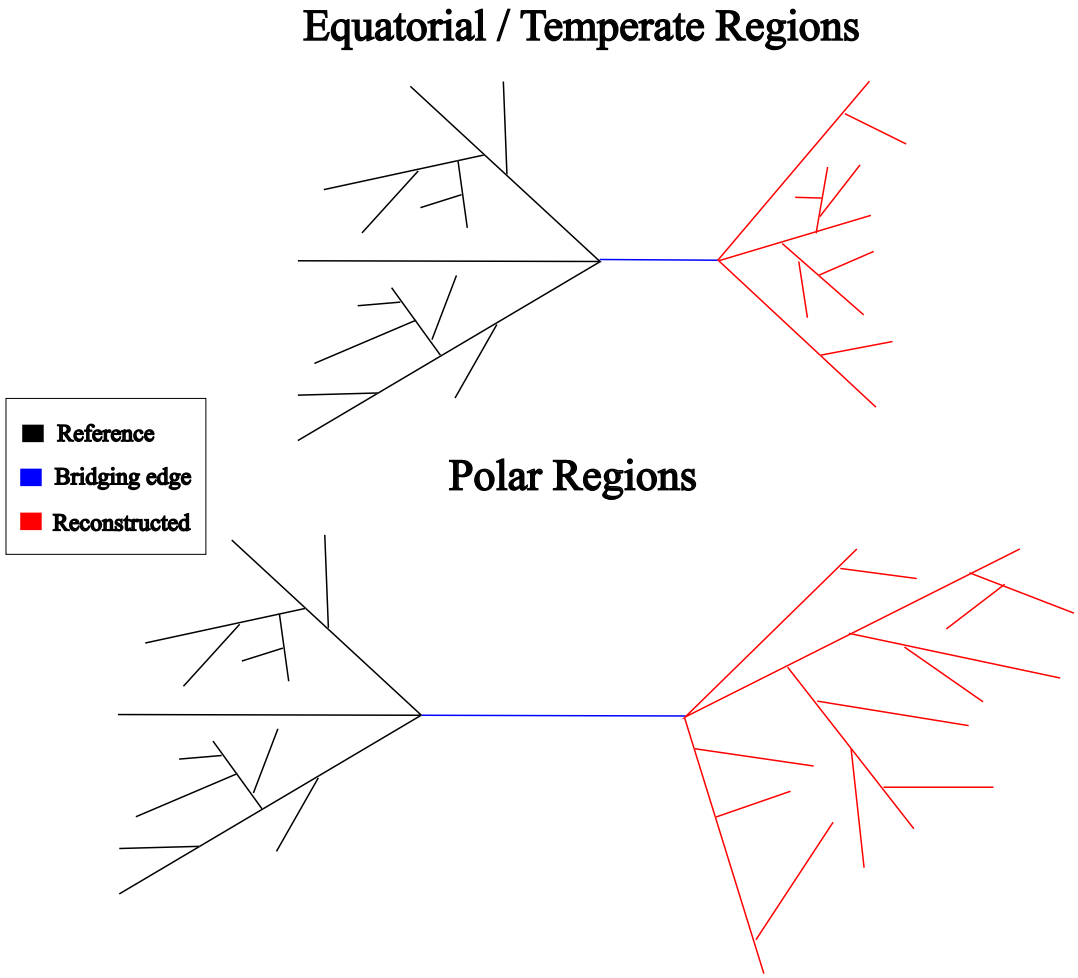
Supplementary Figure 1. Box-plots of contig N50 values (log scale) across region within terrestrial and marine environments. Regions are ordered North (blue), Equator (green), and South (vermillion). Kruskal–Wallis tests yielded no statistically significant differences at $\alpha = 0.01$ (Terrestrial: $H = 5.12$, $p = 0.077$; Marine: $H = 6.54$, $p = 0.038$).



Supplementary Figure 2. Impact of monophyly filtering on viral gene phylogenies across regions. Phylogenetic trees inferred for homologous viral protein-coding gene families reconstructed from environmental metagenomes were filtered to retain only trees in which reconstructed sequences formed monophyletic clades relative to reference sequences. (A) Absolute numbers of phylogenetic trees retained (kept) or excluded (discarded) following filtering, shown for the global dataset and for Marine and Terrestrial subsets partitioned by latitude (North Polar, Equatorial, South Polar). Values above bars indicate the total number of trees evaluated per subset. (B) The same filtering outcome expressed as the percentage of trees discarded due to non-monophyly. Despite substantial variation in total tree availability among regions, monophyly filtering removes <1% of trees overall and consistently <1.5% across all subsets.



Supplementary Figure 3. Impact of IMG/VR reference-tip threshold filtering on global phylogenetic tree retention. A maximum limit of 400 IMG/VR reference sequences per tree was imposed to ensure computational tractability of patristic distance calculations, which scale quadratically with the number of references *(O(n²))*. Of the 5,924,864 total phylogenetic trees constructed across 120 SRA datasets, <0.5% exceeded this threshold and were excluded from subsequent analytical filtering.



Supplementary Figure 4. Conceptual schematic illustrating the tree-based analytical framework used to compare phylogenetic branch length structure between Equatorial/Temperate and Polar viral assemblages. The schematic depicts a simplified representation of gene-level phylogenetic trees in which Reconstructed environmental sequences (red) are connected to Reference sequences (black) via a Bridging edge (blue). Branch lengths associated with reconstructed sequences and bridging edges form the basis for patristic distance–based comparisons of phylogenetic divergence among regions. This figure is illustrative only and is intended to clarify the analytical approach rather than to represent empirical patterns.

Supplementary Tables

Supplementary Table 1. Summary of the SRA run accession IDs of all surviving terrestrial environmental viral metagenomic DNA runs with their approximate geospatial coordinates (n = 63).

	Sample ID	Coordinates	Sample ID	Coordinates
Region	North polar (n = 24)		South polar (n = 21)	
Polar	ERR4837088	67.15,-50.05	SRR11719863	-76.44,161.01
	ERR4837089	67.15,-50.05	SRR11719871	-76.47,161.27
	SRR13106731	70.05,159.55	SRR11719875	-77.00,162.24
	ERR4183329	72.37,126.47	SRR13165308	-71.88,-68.25
	ERR9690361	67.58,134.77	ERR10836094	-69.03,39.52
	ERR10836080	78.64,16.86	SRR18933239	-67.59,-68.23
	SRR18933227	68.49,24.70	SRR11719866	-76.39,161.06
	ERR4837100	67.08,-50.32	SRR11719876	-77.65,163.12
	ERR4837138	67.90,18.61	SRR11719872	-77.01,161.46
	ERR3890093	67.37,-134.85	ERR10836092	-69.38,76.30
	ERR4837098	67.16,-50.08	ERR10836093	-72.35,169.92
	SRR18933223	78.91,11.88	SRR11719870	-76.49,162.06
	ERR11584945	70.00,26.27	SRR18933251	-77.57,163.52
	ERR4837131	67.90,18.61	SRR18933246	-77.02,161.74
	SRR8842248	78.90,11.83	SRR18933241	-67.59,-68.25
	SRR13557795	68.35,19.05	SRR11719864	-77.02,161.41
	ERR4837103	67.06,-50.46	SRR18933250	-77.57,163.52
	SRR8188253	69.20,154.59	SRR18933240	-67.60,-68.25
	SRR13557794	68.35,19.05	ERR10836091	-82.45,-51.35
	ERR4837142	67.90,18.61	SRR11719862	-76.44,161.01
	SRR18933221	69.13,-105.06	SRR11719865	-77.01,161.43
	ERR10878209	78.90,12.07		
	ERR4183325	72.37,126.47		
	ERR4837143	67.90,18.61		
	Equatorial or temperate (n = 18)			
Non-polar	ERR4863013	51.07,12.20	SRR7224129	22.53,116.44
	ERR4603509	23.81,90.42	ERR1346830	39.26,-120.78
	ERR3890111	52.21,0.12	ERR1346888	49.08,-89.38
	ERR1811647	51.79,0.86	ERR4863034	51.07,12.20
	SRR8863434	39.70,-83.89	SRR7072777	38.96,-106.99
	SRR17356570	19.82,-100.67	ERR1811635	51.80,0.92
	ERR4863020	51.07,12.20	SRR18729194	-39.77,176.93
	SRR1914421	47.38,7.16	ERR4863016	51.07,12.20
	SRR5824314	47.50,-93.48	ERR9118890	65.03,-147.71

Supplementary Table 2. Summary of the SRA run accession IDs of all surviving marine environmental viral metagenomic DNA runs with their approximate geospatial coordinates (n = 57).

	Sample ID	Coordinates	Sample ID	Coordinates
Region	North polar (n = 19)		South polar (n = 17)	
Polar	SRR13153527	74.32,-159.44	SRR10156263	-76.20,-170.51
	SRR13153377	76.86,-160.11	SRR10156266	-73.30,-129.56
	SRR13153333	77.07,-161.81	SRR10156275	-71.42,-91.43
	ERR2169009	85.16,-149.97	SRR7529760	-68.47,78.19
	SRR13153473	71.62,-157.90	SRR7529410	-68.47,78.19
	SRR13153437	71.36,-163.00	SRR7526679	-68.47,78.19
	SRR13153475	71.58,-157.81	SRR7529180	-68.47,78.19
	SRR13153393	71.60,-161.54	SRR5581653	-68.47,78.19
	SRR13153419	66.52,-168.42	SRR5458846	-68.47,78.19
	SRR13153461	67.67,-168.73	SRR7529089	-68.47,78.19
	SRR13153260	67.90,-168.23	SRR7528420	-68.47,78.19
	SRR13153443	68.18,-167.31	SRR7528416	-68.47,78.19
	SRR13153257	67.78,-168.60	SRR7527863	-68.47,78.19
	SRR13153411	70.47,-163.75	SRR7528417	-68.47,78.19
	SRR13153438	71.10,-162.25	SRR7529027	-68.47,78.19
	SRR13153550	71.33,-157.31	SRR7528920	-68.47,78.19
	SRR13153304	71.46,-157.57	SRR7528806	-68.47,78.19
	SRR13153483	71.50,-157.65		
	SRR13153401	68.24,-167.13		
	Equatorial or temperate (n = 21)			
Non-polar	SRR17983481	-62.61,-59.83	SRR20241727	42.42,-70.91
	SRR18243841	58.88,11.12	SRR6370751	28.00,-36.00
	SRR7479780	53.80,-128.83	SRR30952025	29.87,-89.68
	SRR28362950	48.68,-122.88	SRR6403353	-36.23,151.27
	ERR4674065	43.76,13.21	SRR16323548	-34.12,151.22
	SRR20242041	42.42,-70.91	SRR8812858	-19.31,147.62
	SRR20241651	42.42,-70.91	SRR26945621	12.60,-177.72
	SRR18489176	34.86,128.42	SRR4465026	17.03,-106.53
	SRR32280785	33.61,-117.93	SRR5131906	24.39,-156.77
	SRR9643251	33.55,-118.40	SRR7597704	-12.40,130.77
	SRR24091367	36.30,120.55		

Supplementary Table 3. Hierarchical filtering workflow for viral gene-family phylogenies. This table summarizes the sequential and concurrent filtering steps used to refine an initial set of 5,924,864 inferred gene-family trees for downstream phylogenetic analyses. Asterisks (∗) indicate filters applied concurrently to the initial inference pool. Monophyly filtering removed trees in which reconstructed sequences were non-monophyletic (0.97% globally), while the reference-tips cap excluded trees containing more than 400 IMG/VR reference sequences (0.35% globally). Final tree inclusion requirements enforced a minimum of ≥2 reconstructed and ≥2 reference tips per tree to ensure stable distance estimation, yielding a final analysis set of 1,846,393 trees (31.2% of the initial pool).

Filtering stage	Description	Trees passing filter (n)	Trees excluded by filter (n)	Trees retained (%)
Initial inference	All inferred viral gene-family phylogenies	5,924,864	–	100
*Monophyly filtering	Removal of trees in which reconstructed sequences were non-monophyletic	5,867,219	57,645	99.03
*Reference tips cap	Exclusion of trees exceeding 400 IMG/VR reference sequences	5,904,089	20,775	99.65
Tree inclusion requirements	Minimum of ≥2 reconstructed and ≥2 IMG/VR reference tips per tree	1,846,393	4,020,826	31.2
Final analysis set	Trees used for MPD and bridging-edge analyses	1,846,393	–	31.2

Data Availability Statement

All raw sequencing reads analysed in this thesis are publicly available in the NCBIs Sequence Read Archive (SRA) at the URL: https://www.ncbi.nlm.nih.gov/sra

All analysis scripts used in this thesis are available on GitHub at:

https://github.com/vkulk094/viral_phylo

Bibliography

Alempic, J. M., Lartigue, A., Goncharov, A. E., Grosse, G., Strauss, J., Tikhonov, A. N., Fedorov, A. N., Poirot, O., Legendre, M., Santini, S., Abergel, C., & Claverie, J. M. (2023). An update on eukaryotic viruses revived from ancient permafrost. *Viruses*, *15*(2). https://doi.org/10.3390/V15020564

Andrews, C. A. (2010). Natural selection, genetic drift, and gene flow do not act in isolation in natural populations. *Nature Education Knowledge*, *3*(10), 5.

Andrews, S. (2010). *FastQC: A quality control tool for high-throughput sequence data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Anesio, A. M., & Bellas, C. M. (2011). Are low temperature habitats hot spots of microbial evolution driven by viruses? *Trends in Microbiology*, *19*(2), 52–57. https://doi.org/10.1016/J.TIM.2010.11.002

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J. M., Mueller, J. E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C. A., & Rohwer, F. (2006). The marine viromes of four oceanic regions. *PLoS Biology*, *4*(11), 2121–2131. https://doi.org/10.1371/JOURNAL.PBIO.0040368

Balinda, S. N., Siegismund, H. R., Muwanika, V. B., Sangula, A. K., Masembe, C., Ayebazibwe, C., Normann, P., & Belsham, G. J. (2010). Phylogenetic analyses of the polyprotein coding sequences of serotype O foot-and-mouth disease viruses in East Africa: Evidence for interserotypic recombination. *Virology Journal*, *7*. https://doi.org/10.1186/1743-422X-7-199

Bauer, S. L. M., Stensland, A., Daae, F. L., Sandaa, R. A., Thorseth, I. H., Steen, I. H., & Dahle, H. (2018). Water masses and depth structure prokaryotic and T4-like viral communities around hydrothermal systems of the Nordic Seas. *Frontiers in Microbiology*, *9*, 1002. https://doi.org/10.3389/FMICB.2018.01002

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing [Article]. *Journal of the Royal Statistical Society. Series B, Methodological*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., Brister, J. R., Kropinski, A. M., Krupovic, M., Lavigne, R., Turner, D., & Sullivan, M. B. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology*, *37*(6), 632–639. https://doi.org/10.1038/S41587-019-0100-8

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Broecker, F., & Moelling, K. (2019). What viruses tell us about evolution and immunity: Beyond Darwin? *Annals of the New York Academy of Sciences*, *1447*(1). https://doi.org/10.1111/NYAS.14097

Brum, J. R., Cesar Ignacio-Espinoza, J., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., Chaffron, S., Cruaud, C., De Vargas, C., Gasol, J. M., Gorsky, G., Gregory, A. C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B. T., … Weissenbach, J. (2015). Patterns and ecological drivers of ocean viral communities. *Science*, *348*(6237). https://doi.org/10.1126/SCIENCE.1261498

Brüssow, H. (2009). The not so universal tree of life or the place of viruses in the living world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1527), 2263. https://doi.org/10.1098/RSTB.2009.0036

Buscaglia, M., Iriarte, J. L., Schulz, F., & Díez, B. (2024). Adaptation strategies of giant viruses to low-temperature marine ecosystems. *The ISME Journal*, *18*(1), 162. https://doi.org/10.1093/ISMEJO/WRAE162

Calayag, A. M., Priest, T., Oldenburg, E., Muschiol, J., Popa, O., Wietz, M., & Needham, D. M. (2025). Arctic Ocean virus communities and their seasonality, bipolarity, and prokaryotic associations. *Nature Communications*, *16*(1), 1–15. https://doi.org/10.1038/s41467-025-61568-6

Camargo, A. P., Nayfach, S., Chen, I. M. A., Palaniappan, K., Ratner, A., Chu, K., Ritter, S. J., Reddy, T. B. K., Mukherjee, S., Schulz, F., Call, L., Neches, R. Y., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A., Kyrpides, N. C., & Roux, S. (2023). IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Research*, *51*(D1), D733–D743. https://doi.org/10.1093/NAR/GKAC1037

Camargo, A. P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P. S. G., Nayfach, S., & Kyrpides, N. C. (2023). Identification of mobile genetic elements with geNomad. *Nature Biotechnology*, *42*(8), 1303–1312. https://doi.org/10.1038/s41587-023-01953-y

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, *25*(15), 1972–1973. https://doi.org/10.1093/BIOINFORMATICS/BTP348

Carroll, D., Daszak, P., Wolfe, N. D., Gao, G. F., Morel, C. M., Morzaria, S., Pablos-Méndez, A., Tomori, O., & Mazet, J. A. K. (2018). The Global Virome Project. *Science*, *359*(6378), 872–874. https://doi.org/10.1126/SCIENCE.AAP7463

Cavender-Bares, J., Kozak, K. H., Fine, P. V. A., & Kembel, S. W. (2009). The merging of community ecology and phylogenetic biology. *Ecology Letters*, *12*(7), 693–715. https://doi.org/10.1111/J.1461-0248.2009.01314.X

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. https://doi.org/10.1037/0033-2909.114.3.494

Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics [Article]. *Multivariate Behavioral Research*, *31*(3), 331–350. https://doi.org/10.1207/s15327906mbr3103_4

Czech, L., Stamatakis, A., Dunthorn, M., & Barbera, P. (2022). Metagenomic analysis using phylogenetic placement—A review of the first decade. *Frontiers in Bioinformatics*, *2*, 871393. https://doi.org/10.3389/FBINF.2022.871393

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. A., Bik, H. M., & Eisen, J. A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, *2*(1), e243. https://doi.org/10.7717/PEERJ.243

De Cárcer, D. A., López-Bueno, A., Pearce, D. A., & Alcamí, A. (2015). Biodiversity and distribution of polar freshwater DNA viruses. *Science Advances*, *1*(5), e1400127. https://doi.org/10.1126/SCIADV.1400127

Dickinson, B. C., Leconte, A. M., Allen, B., Esvelt, K. M., & Liu, D. R. (2013). Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(22), 9007–9012. https://doi.org/10.1073/PNAS.1220670110

Doron-Faigenboim, A., & Pupko, T. (2007). A combined empirical and mechanistic codon model. *Molecular Biology and Evolution*, *24*(2), 388–397. https://doi.org/10.1093/MOLBEV/MSL175

Emerson, J. B., Roux, S., Brum, J. R., Bolduc, B., Woodcroft, B. J., Jang, H. Bin, Singleton, C. M., Solden, L. M., Naas, A. E., Boyd, J. A., Hodgkins, S. B., Wilson, R. M., Trubl, G., Li, C., Frolking, S., Pope, P. B., Wrighton, K. C., Crill, P. M., Chanton, J. P., … Sullivan, M. B. (2018). Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology*, *3*(8), 870–880. https://doi.org/10.1038/s41564-018-0190-y

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity [Article]. *Biological Conservation*, *61*(1), 1–10. https://doi.org/10.1016/0006-3207(92)91201-3

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(Web Server issue), W29–W37. https://doi.org/10.1093/NAR/GKR367

Florent, P., Cauchie, H. M., Herold, M., & Ogorzaly, L. (2022). Bacteriophages pass through candle-shaped porous ceramic filters: Application for the collection of viruses in soil water. *MicrobiologyOpen*, *11*(5), e1314. https://doi.org/10.1002/MBO3.1314

Fourment, M., & Gibbs, M. J. (2006). PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. *BMC Evolutionary Biology*, *6*, 1. https://doi.org/10.1186/1471-2148-6-1

French, R. K., Anderson, S. H., Cain, K. E., Greene, T. C., Minor, M., Miskelly, C. M., Montoya, J. M., Wille, M., Muller, C. G., Taylor, M. W., Digby, A., Crane, J., Davitt, G., Eason, D., Hedman, P., Jeynes, B., Latimer, S., Little, S., Mitchell, M., … Holmes, E. C. (2023). Host phylogeny shapes viral transmission networks in an island ecosystem. *Nature Ecology & Evolution*, *7*(11), 1834–1843. https://doi.org/10.1038/s41559-023-02192-9

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/BIOINFORMATICS/BTS565

García-López, R., Vázquez-Castellanos, J. F., & Moya, A. (2015). Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. *Frontiers in Bioengineering and Biotechnology*, *3*, 141. https://doi.org/10.3389/FBIOE.2015.00141

Gong, Z., Liang, Y., Wang, M., Jiang, Y., Yang, Q., Xia, J., Zhou, X., You, S., Gao, C., Wang, J., He, J., Shao, H., & McMinn, A. (2018). Viral diversity and its relationship with environmental factors at the surface and deep sea of Prydz Bay, Antarctica. *Frontiers in Microbiology*, *9*, 411673. https://doi.org/10.3389/FMICB.2018.02981

Graham, E. B., Camargo, A. P., Wu, R., Neches, R. Y., Nolan, M., Paez-Espino, D., Kyrpides, N. C., Jansson, J. K., McDermott, J. E., Hofmockel, K. S., Blanchard, J. L., Liu, X. J. A., Rodrigues, J. L. M., Freedman, Z. B., Baldrian, P., Stursova, M., DeAngelis, K. M., Lee, S., Godoy-Vitorino, F., … Pietrasiak, N. (2024). A global atlas of soil viruses reveals unexplored biodiversity and potential biogeochemical impacts. *Nature Microbiology*, *9*(7), 1873–1883. https://doi.org/10.1038/s41564-024-01686-x

Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Domínguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., Pisarev, S., … Roux, S. (2019). Marine DNA viral macro- and microdiversity from pole to pole. *Cell*, *177*(5), 1109-1123.e14. https://doi.org/10.1016/J.CELL.2019.03.040

Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C., & Martiny, J. B. H. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology*, *10*(7), 497–506. https://doi.org/10.1038/nrmicro2795

Harris, H. M. B., & Hill, C. (2021). A place for viruses on the tree of life [Article]. *Frontiers in Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.604048

Hata, A., Katayama, H., & Furumai, H. (2015). Organic substances interfere with reverse transcription-quantitative PCR-based virus detection in water samples. *Applied and Environmental Microbiology*, *81*(5), 1585. https://doi.org/10.1128/AEM.03082-14

Heinrichs, M. E., Piedade, G. J., Popa, O., Sommers, P., Trubl, G., Weissenbach, J., & Rahlff, J. (2024). Breaking the ice: A review of phages in polar ecosystems. *Methods in Molecular Biology (Clifton, N.J.)*, *2738*, 31–71. https://doi.org/10.1007/978-1-0716-3549-0_3

Hess, M., & Kromrey, J. D. (2004). Robust confidence intervals for effect sizes: A comparative study of Cohen's d and Cliff's delta under nonnormality and heterogeneous variances. *Paper Presented at the Annual Meeting of the American Educational Research Association, San Diego, CA*. https://www.academia.edu/34057137/Robust_Confidence_Intervals_for_Effect_Sizes_a_Comparative_Study_of_Cohens_D_and_Cliffs_Delta_Under_Non_Normality_and_Heterogeneous_Variances

Hillebrand, H. (2004). On the generality of the latitudinal diversity gradient. *American Naturalist*, *163*(2), 192–211. https://doi.org/10.1086/381004

Hodges, J. L., & Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, *34*(2), 598–611.

Hou, X., He, Y., Fang, P., Mei, S.-Q., Xu, Z., Wu, W.-C., Zhang, S., Zeng, Z.-Y., Gou, Q.-Y., Xin, G.-Y., Le, S.-J., Xia, Y.-Y., Zhou, Y.-L., Hui, F.-M., Pan, Y.-F., Eden, J.-S., Yang, Z.-H., Han, C., Shu, Y.-L., … Laboratory, G. (2023). Artificial intelligence redefines RNA virus discovery. *BioRxiv*. https://doi.org/10.1101/2023.04.18.537342

Irwin, N. A. T., Pittis, A. A., Richards, T. A., & Keeling, P. J. (2021). Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nature Microbiology*, *7*(2), 327–336. https://doi.org/10.1038/s41564-021-01026-3

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/NAR/GKF436

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/MOLBEV/MST010

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., Blomberg, S. P., & Webb, C. O. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, *26*(11), 1463–1464. https://doi.org/10.1093/BIOINFORMATICS/BTQ166

Khot, V., Strous, M., & Hawley, A. K. (2020). Computational approaches in viral ecology. *Computational and Structural Biotechnology Journal*, *18*, 1605–1612. https://doi.org/10.1016/J.CSBJ.2020.06.019

Kleinteich, J., Hildebrand, F., Bahram, M., Voigt, A. Y., Wood, S. A., Jungblut, A. D., Küpper, F. C., Quesada, A., Camacho, A., Pearce, D. A., Convey, P., Vincent, W. F., Zarfl, C., Bork, P., & Dietrich, D. R. (2017). Pole-to-Pole connections: Similarities between arctic and antarctic microbiomes and their vulnerability to environmental change. *Frontiers in Ecology and Evolution*, *5*(NOV), 300617. https://doi.org/10.3389/FEVO.2017.00137

Krishnamurthy, S. R., & Wang, D. (2017). Origins and challenges of viral dark matter. In *Virus Research* (Vol. 239, pp. 136–142). https://doi.org/10.1016/j.virusres.2017.02.002

Labbé, M., Girard, C., Vincent, W. F., & Culley, A. I. (2020). Extreme viral partitioning in a marine-derived High Arctic lake. *MSphere*, *5*(3). https://doi.org/10.1128/MSPHERE.00334-20

Ladau, J., & Eloe-Fadrosh, E. A. (2019). Spatial, temporal, and phylogenetic scales of microbial ecology. *Trends in Microbiology*, *27*(8), 662–669. https://doi.org/10.1016/J.TIM.2019.03.003

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320. https://doi.org/10.1093/MOLBEV/MSN067

Lemieux, A., Colby, G. A., Poulain, A. J., & Aris-Brosou, S. (2022). Viral spillover risk increases with climate change in High Arctic lake sediments. *Proceedings of the Royal Society B*, *289*(1985). https://doi.org/10.1098/RSPB.2022.1073

Lemieux, A., Poulain, A. J., & Aris-Brosou, S. (2024). The High Arctic is dominated by uncharacterized, genetically highly diverse bacteriophages. *BioRxiv*, 2024.09.10.612304. https://doi.org/10.1101/2024.09.10.612304

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)*, *31*(10), 1674–1676. https://doi.org/10.1093/BIOINFORMATICS/BTV033

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/BIOINFORMATICS/BTL158

Mahmoudabadi, G., & Phillips, R. (2018). A comprehensive and quantitative exploration of thousands of viral genomes. *ELife*, *7*. https://doi.org/10.7554/ELIFE.31955

Manuel, R. D., & Snyder, J. C. (2024). The expanding diversity of viruses from extreme environments. *International Journal of Molecular Sciences*, *25*(6), 3137. https://doi.org/10.3390/IJMS25063137

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Øvreås, L., Reysenbach, A. L., Smith, V. H., & Staley, J. T. (2006). Microbial biogeography: putting

microorganisms on the map. *Nature Reviews Microbiology*, *4*(2), 102–112. https://doi.org/10.1038/nrmicro1341

Meissel, K., & Yao, E. S. (2024). Using Cliff's delta as a non-parametric effect size measure: An accessible web app and R tutorial. *Practical Assessment, Research & Evaluation*, *29*(2). https://doi.org/10.7275/pare.1977

Meng, L., Delmont, T. O., Gaïa, M., Pelletier, E., Fernàndez-Guerra, A., Chaffron, S., Neches, R. Y., Wu, J., Kaneko, H., Endo, H., & Ogata, H. (2023). Genomic adaptation of giant viruses in polar oceans. *Nature Communications*, *14*(1), 1–12. https://doi.org/10.1038/s41467-023-41910-6

Mittal, P., Sipani, R., Pandiyan, A., Sulthana, S., Sinha, A. K., Hussain, A., Ray, M. K., & Pavankumar, T. L. (2023). Exoribonuclease RNase R protects Antarctic Pseudomonas syringae Lz4W from DNA damage and oxidative stress. *Applied and Environmental Microbiology*, *89*(11), e01168-23. https://doi.org/10.1128/AEM.01168-23

Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush, M. B., Harrison, S. P., Hurlbert, A. H., Knowlton, N., Lessios, H. A., McCain, C. M., McCune, A. R., McDade, L. A., McPeek, M. A., Near, T. J., Price, T. D., Ricklefs, R. E., Roy, K., Sax, D. F., … Turelli, M. (2007). Evolution and the latitudinal diversity gradient: Speciation, extinction and biogeography. *Ecology Letters*, *10*(4), 315–331. https://doi.org/10.1111/J.1461-0248.2007.01020.X

Murray, J. P., & Laband, S. J. (1979). Degradation of poliovirus by adsorption on inorganic surfaces. *Applied and Environmental Microbiology*, *37*(3), 480–486. https://doi.org/10.1128/AEM.37.3.480-486.1979

Mushegian, A. R. (2020). Are there 1031 virus particles on Earth, or more, or fewer? *Journal of Bacteriology*, *202*(9). https://doi.org/10.1128/JB.00052-20

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2020). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology 2020 39:5*, *39*(5), 578–585. https://doi.org/10.1038/s41587-020-00774-7

Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Bilinski, T. M., Stanish, L. F., Knelman, J. E., Darcy, J. L., Lynch, R. C., Wickey, P., & Ferrenberg, S. (2013). Patterns and processes of microbial community assembly. *Microbiology and Molecular Biology Reviews*, *77*(3), 342–356. https://doi.org/10.1128/MMBR.00051-12

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., & Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature 2016*, *536*(7617), 425–430. https://doi.org/10.1038/nature19094

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/BIOINFORMATICS/BTY633

Peona, V., Blom, M. P. K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., Liachko, I., Haryoko, T., Jønsson, K. A., Zhou, Q., Irestedt, M., & Suh, A. (2020). Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology Resources*, *21*(1), 263. https://doi.org/10.1111/1755-0998.13252

Piedade, G. J., Schön, M. E., Lood, C., Fofanov, M. V., Wesdorp, E. M., Biggs, T. E. G., Wu, L., Bolhuis, H., Fischer, M. G., Yutin, N., Dutilh, B. E., & Brussaard, C. P. D. (2024). Seasonal dynamics and diversity of Antarctic marine viruses reveal a novel viral seascape. *Nature Communications*, *15*(1), 1–16. https://doi.org/10.1038/s41467-024-53317-y

Pitot, T. M., Rapp, J. Z., Schulz, F., Girard, C., Roux, S., & Culley, A. I. (2024). Distinct and rich assemblages of giant viruses in Arctic and Antarctic lakes. *ISME Communications*, *4*(1), ycae048. https://doi.org/10.1093/ISMECO/YCAE048

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3). https://doi.org/10.1371/JOURNAL.PONE.0009490

Rantanen, M., Karpechko, A. Y., Lipponen, A., Nordling, K., Hyvärinen, O., Ruosteenoja, K., Vihma, T., & Laaksonen, A. (2022). The Arctic has warmed nearly four times faster than the globe since 1979. *Communications Earth & Environment*, *3*(1), 1–10. https://doi.org/10.1038/s43247-022-00498-3

Rochman, N. D., Wolf, Y. I., & Koonin, E. V. (2022). Molecular adaptations during viral epidemics. *EMBO Reports*, *23*(8). https://doi.org/10.15252/EMBR.202255393

Romano, J., Kromrey, J. D., Coraggio, J., & Skowronek, J. (2006). Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen'sd for evaluating group differences on the NSSE and other surveys. *In Annual Meeting of the Florida Association of Institutional Research*, *177*(34).

Rose, R., Constantinides, B., Tapinos, A., Robertson, D. L., & Prosperi, M. (2016). Challenges in the analysis of viral metagenomes. *Virus Evolution*, *2*(2), vew022. https://doi.org/10.1093/VE/VEW022

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., Amid, C., Aziz, R. K., Bordenstein, S. R., Bork, P., Breitbart, M., Cochrane, G. R., Daly, R. A., Desnues, C., Duhaime, M. B., … Eloe-Fadrosh, E. A. (2018). Minimum information about an uncultivated virus genome (MIUViG). *Nature Biotechnology*, *37*(1), 29–37. https://doi.org/10.1038/nbt.4306

Roux, S., Hallam, S. J., Woyke, T., & Sullivan, M. B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *ELife*, *4*(JULY2015). https://doi.org/10.7554/ELIFE.08490

Sanjuán, R., Moya, A., & Elena, S. F. (2004). The contribution of epistasis to the architecture of fitness in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(43), 15376–15379. https://doi.org/10.1073/PNAS.0404125101

Santiago-Rodriguez, T. M., & Hollister, E. B. (2022). Unraveling the viral dark matter through viral metagenomics. *Frontiers in Immunology*, *13*. https://doi.org/10.3389/fimmu.2022.1005107

Santos-Medellin, C., Estera-Molina, K., Yuan, M., Pett-Ridge, J., Firestone, M. K., & Emerson, J. B. (2022). Spatial turnover of soil viral populations and genotypes overlain by cohesive responses to moisture in grasslands [Article]. *Proceedings of the National Academy of Sciences - PNAS*, *119*(45), 1–11. https://doi.org/10.1073/pnas.2209132119

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Serreze, M. C., & Stroeve, J. (2015). Arctic sea ice trends, variability and implications for seasonal ice forecasting. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *373*(2045). https://doi.org/10.1098/RSTA.2014.0159

Shiryev, S. A., & Agarwala, R. (2024). Pebblescout is an easy-to-use tool for fast sequence search in petabase-scale nucleotide resources. *Nature Methods*, *21*(6), 938–939. https://doi.org/10.1038/S41592-024-02281-Y

Sianga-Mete, R., Hartnady, P., Mandikumba, W. C., Rutherford, K., Currin, C. B., Phelanyane, F., Stefan, S., Pond, S. L. K., & Martin, D. P. (2022). Viral genome sequence datasets display pervasive evidence of strand-specific substitution biases that are best described using non-reversible nucleotide substitution models. *Research Square*, rs.3.rs-2407778. https://doi.org/10.21203/RS.3.RS-2407778

Simmonds, P., Adriaenssens, E. M., Murilo Zerbini, F., Abrescia, N. G. A., Aiewsakun, P., Alfenas-Zerbini, P., Bao, Y., Barylski, J., Drosten, C., Duffy, S., Paul Duprex, W., Dutilh, B. E., Elena, S. F., García, M. L., Junglen, S., Katzourakis, A., Koonin, E. V., Krupovic, M., Kuhn, J. H., … Vasilakis, N. (2023). Four principles to establish a universal virus taxonomy. *PLOS Biology*, *21*(2), e3001922. https://doi.org/10.1371/JOURNAL.PBIO.3001922

Sironi, M., Cagliani, R., Forni, D., & Clerici, M. (2015). Evolutionary insights into host–pathogen interactions from mammalian sequence data. *Nature Reviews. Genetics*, *16*(4), 224. https://doi.org/10.1038/NRG3905

Smits, S. L., Bodewes, R., Ruiz-González, A., Baumgärtner, W., Koopmans, M. P., Osterhaus, A. D. M. E., & Schürch, A. C. (2015). Recovering full-length viral genomes from metagenomes. *Frontiers in Microbiology*, *6*, 1069. https://doi.org/10.3389/FMICB.2015.01069

Sommers, P., Chatterjee, A., Varsani, A., & Trubl, G. (2021). Integrating viral metagenomics into an ecological framework. *Annual Review of Virology*, *8*(1), 133–158. https://doi.org/10.1146/ANNUREV-VIROLOGY-010421-053015

Sorrentino, G., Chellappah, K., & Biscontin, G. (2025). Understanding clogging mechanisms in filter media: An integration of laboratory findings and theoretical perspectives. *Separation and Purification Technology*, *359*, 130602. https://doi.org/10.1016/J.SEPPUR.2024.130602

SRA Toolkit Development Team. (2016, April 20). *SRA Toolkit (v3.0.0)*. NCBI. https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

Sullivan, M. B. (2015). Viromes, not gene markers, for studying double-stranded DNA virus communities. *Journal of Virology*, *89*(5), 2459–2461. https://doi.org/10.1128/JVI.03289-14

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., D'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., … Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237). https://doi.org/10.1126/SCIENCE.1261359

Suttle, C. A. (2005). Viruses in the sea. *Nature*, *437*(7057), 356–361. https://doi.org/10.1038/nature04160

Tao, Y., Xun, F., Zhao, C., Mao, Z., Li, B., Xing, P., & Wu, Q. L. (2022). Improved assembly of metagenome-assembled genomes and viruses in Tibetan saline lake sediment by HiFi metagenomic sequencing. *Microbiology Spectrum*, *11*(1), e03328-22. https://doi.org/10.1128/SPECTRUM.03328-22

Tong, D., Wang, Y., Yu, H., Shen, H., Dahlgren, R. A., & Xu, J. (2023). Viral lysing can alleviate microbial nutrient limitations and accumulate recalcitrant dissolved organic matter components in soil. *The ISME Journal*, *17*(8), 1247. https://doi.org/10.1038/S41396-023-01438-5

Trivedi, C. B., Keuschnig, C., Larose, C., Rissi, D. V., Mourot, R., Bradley, J. A., Winkel, M., & Benning, L. G. (2022). DNA/RNA preservation in glacial snow and ice samples. *Frontiers in Microbiology*, *13*, 894893. https://doi.org/10.3389/FMICB.2022.894893

Trubl, G., Jang, H. Bin, Roux, S., Emerson, J. B., Solonenko, N., Vik, D. R., Solden, L., Ellenbogen, J., Runyon, A. T., Bolduc, B., Woodcroft, B. J., Saleska, S. R., Tyson, G. W., Wrighton, K. C., Sullivan, M. B., & Rich, V. I. (2018). Soil viruses are underexplored players in ecosystem carbon processing. *MSystems*, *3*(5). https://doi.org/10.1128/MSYSTEMS.00076-18

Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Jonathan Davies, T., Ferrier, S., Fritz, S. A., Grenyer, R., Helmus, M. R., Jin, L. S., Mooers, A. O., Pavoine, S., Purschke, O., Redding, D. W., Rosauer, D. F., Winter, M., & Mazel, F. (2017). A guide to phylogenetic metrics for conservation,

community ecology and macroecology. *Biological Reviews*, *92*(2), 698–715. https://doi.org/10.1111/BRV.12252

Vonk, J. E., Tank, S. E., Bowden, W. B., Laurion, I., Vincent, W. F., Alekseychik, P., Amyot, M., Billet, M. F., Canário, J., Cory, R. M., Deshpande, B. N., Helbig, M., Jammet, M., Karlsson, J., Larouche, J., Macmillan, G., Rautio, M., Anthony, K. M. W., & Wickland, K. P. (2015). Reviews and syntheses: Effects of permafrost thaw on Arctic aquatic ecosystems. *Biogeosciences*, *12*, 7129–7167. https://doi.org/10.5194/bg-12-7129-2015

Vos, M., Wolf, A. B., Jennings, S. J., & Kowalchuk, G. A. (2013). Micro-scale determinants of bacterial diversity in soil. *FEMS Microbiology Reviews*, *37*(6), 936–954. https://doi.org/10.1111/1574-6976.12023

Wang, J., Xiao, J., Zhu, Z., Wang, S., Zhang, L., Fan, Z., Deng, Y., Hu, Z., Peng, F., Shen, S., & Deng, F. (2022). Diverse viromes in polar regions: A retrospective study of metagenomic data from Antarctic animal feces and Arctic frozen soil in 2012–2014. *Virologica Sinica*, *37*(6), 883. https://doi.org/10.1016/J.VIRS.2022.08.006

Warwick-Dugdale, J., Tian, F., Michelsen, M. L., Cronin, D. R., Moore, K., Farbos, A., Chittick, L., Bell, A., Zayed, A. A., Buchholz, H. H., Bolanos, L. M., Parsons, R. J., Allen, M. J., Sullivan, M. B., & Temperton, B. (2024). Long-read powered viral metagenomics in the oligotrophic Sargasso Sea. *Nature Communications*, *15*(1). https://doi.org/10.1038/S41467-024-48300-6

Webb, C. O., Ackerly, D. D., McPeek, M. A., & Donoghue, M. J. (2002). Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, *33*(Volume 33, 2002), 475–505. https://doi.org/10.1146/ANNUREV.ECOLSYS.33.010802.150448

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.) [Book]. Elsevier/Academic Press. https://doi.org/10.1016/C2010-0-67044-1

Wilhelm, S. W., Bird, J. T., Bonifer, K. S., Calfee, B. C., Chen, T., Coy, S. R., Jackson Gainer, P., Gann, E. R., Heatherly, H. T., Lee, J., Liang, X., Liu, J., Armes, A. C., Moniruzzaman, M., Hunter Rice, J., Stough, J. M. A., Tams, R. N., Williams, E. P., & Lecleir, G. R. (2017). A student's guide to giant viruses infecting small eukaryotes: From Acanthamoeba to zooxanthellae. *Viruses*, *9*(3), 46. https://doi.org/10.3390/V9030046

Williams, R. C., Perry, W. B., Lambert-Slosarska, K., Futcher, B., Pellett, C., Richardson-O'Neill, I., Paterson, S., Grimsley, J. M. S., Wade, M. J., Weightman, A. J., Farkas, K., & Jones, D. L. (2024). Examining the stability of viral RNA and DNA in wastewater: Effects of storage time, temperature, and freeze-thaw cycles. *Water Research*, *259*. https://doi.org/10.1016/J.WATRES.2024.121879

Wommack, K. E., & Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews : MMBR*, *64*(1), 69–114. https://doi.org/10.1128/MMBR.64.1.69-114.2000

Worobey, M. (2000). Extensive homologous recombination among widely divergent TT viruses. *Journal of Virology*, *74*(16), 7666–7670. https://doi.org/10.1128/JVI.74.16.7666-7670.2000

Wu, R., Bottos, E. M., Danna, V. G., Stegen, J. C., Jansson, J. K., & Davison, M. R. (2022). RNA viruses linked to eukaryotic hosts in thawed permafrost. *MSystems*, *7*(6), e00582-22. https://doi.org/10.1128/MSYSTEMS.00582-22

Yang, D., Yang, Y., & Xia, J. (2021). Hydrological cycle and water resources in a changing world: A review. *Geography and Sustainability*, *2*(2), 115–122. https://doi.org/10.1016/J.GEOSUS.2021.05.003

Zaheri, M., Dib, L., & Salamin, N. (2014). A generalized mechanistic codon model. *Molecular Biology and Evolution*, *31*(9), 2528–2541. https://doi.org/10.1093/MOLBEV/MSU196

Zayed, A. A., Wainaina, J. M., Dominguez-Huerta, G., Pelletier, E., Guo, J., Mohssen, M., Tian, F., Pratama, A. A., Bolduc, B., Zablocki, O., Cronin, D., Solden, L., Delage, E., Alberti, A., Aury, J. M., Carradec, Q., da Silva, C., Labadie, K., Poulain, J., … Sullivan, M. B. (2022). Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science*, *376*(6589), 156–162. https://doi.org/10.1126/SCIENCE.ABM5847

Zhong, Z.-P., Rapp, J. Z., Wainaina, J. M., Solonenko, N. E., Maughan, H., Carpenter, S. D., Cooper, Z. S., Jang, H. Bin, Bolduc, B., Deming, J. W., & Sullivan, M. B. (2020). Viral ecogenomics of Arctic cryopeg brine and sea ice. *MSystems*, *5*(3). https://doi.org/10.1128/MSYSTEMS.00246-20

Zhu, W., Lomsadze, A., & Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research*, *38*(12), e132–e132. https://doi.org/10.1093/NAR/GKQ275