Designing Genomic Solutions for Abiotic Traits in Flax (*Linum usitatissimum* L.)

by

Nadeem Khan

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements of the degree Doctor of Philosophy

Department of Biology

Faculty of Science

University of Ottawa

© Nadeem Khan, Ottawa, Canada, 2022

Abstract

Flax (*Linum usitatissimum* L.) is a self-pollinated crop widely cultivated for fiber and oil production. Flaxseed is renowned for its health attributes but the presence of compounds, such as the heavy metal cadmium (Cd), is undesirable. Genomic studies in flax have produced large amounts of data in the last 15 years, providing useful resources to improve the genetic of this crop using genomics-based technologies and strategies. The goal of this thesis is therefore to capitalize on these advances to address the Cd problem and to propose solutions to improve breeding efficiencies. To find genomic-based solutions to Cd content, to the currently low breeding efficiency and to abiotic stress resistance in flax, this study utilized four major strategies: (1) genomic cross prediction, (2) gene family identification, (3) genome-wide association study (GWAS) and (4) genomic selection (GS). Characterization of the ATP-binding cassette (ABC) transporter and heavy metal associated (HMA) gene families was performed using the flax genome sequence. A total of 198 ABC transporter and 12 HMA genes were identified in the flax genome, of which nine were orthologous to Cd-associated genes in *Arabidopsis*, rice and maize. A transcriptomic analysis of eight tissues provided some support towards the functional annotation of these genes and confirmed the expression of these ABC transporter and HMA genes in flax seeds and other tissues. A diversity panel of 168 flax accessions was grown in the field at multiple locations and years and the seed content of 24 heavy metals (HMs) was measured. The panel was also sequenced and a single nucleotide polymorphism (SNP) dataset of nearly 43,000 SNPs was defined. A GWAS was conducted using these genotypic and phenotypic data and a total of 355 non-redundant quantitative trait nucleotides (QTNs) were identified for ten of the 24 metal contents. Overall, a total of 24 major and 331 minor effect QTNs were detected, including 11 that were pleiotropic. After allelic tests, 108 non-redundant QTNs were retained for eight of the ten

Abstract

metals and ranging from one for copper (Cu) to 70 for strontium (Sr). A total of 20 candidate genes for HM accumulation were identified at 12 of the 24 major QTN loci, of which five belonged to the ABC transporter family. Many of the metal contents, including Cd, appeared to be controlled by many genes of small effects; hence, GS is better suited than marker-assisted selection for application in breeding. To test this, predictive ability using ten GS statistical models was evaluated using trait-specific QTN and the random genome-wide 43K SNP datasets. Significantly higher predictive abilities were observed from the GS models built with the dataset made of QTNs associated with metal contents (70-80%) compared to that of the 43K dataset (10-25%).

This study showed the feasibility of using GS to improve the predictive ability of polygenic traits such as metal content in seeds. GS can be applied in early generation selection to accelerate the improvement of abiotic stress resistance and either select low-Cd lines or discard high-Cd lines. These findings validate the use of a QTL-based strategy as a highly effective method for improving the efficiency of predictive ability of GS for highly complex traits such as resistance or tolerance to HM accumulation. Identification of both large and minor effect QTNs and/or pleiotropic effects hold potential for flax breeding improvement. Candidate gene functional validation can be performed using methods such as genome editing or targeting induced local lesions in genomes (TILLING).

Keywords: Flax (*Linum usitatissimum* L.) ● Cadmium (Cd) accumulation ● Genome-wide association study (GWAS) ● Quantitative trait loci (QTLs) ● Quantitative trait nucleotides (QTNs) ● Genomic selection (GS) ● Genomic cross prediction ● Candidate genes ● ATP-binding cassette (ABC) transporters ● Heavy metal associated (HMA) genes ● RNASeq

Acknowledgments

First of all, I want to express my gratitude to ALLAH Almighty for giving me courage, strength, and safety while I was in Canada. Then, I would want to sincerely thank my two supervisors, Drs. Sylvie Cloutier and Frank M. You, for their knowledge, persistence, and support. I appreciate you for providing me with this wonderful opportunity and teaching me many other priceless skills, such as how to analyze, plan, and write logically. You will always have my gratitude. My advisory committee members, Dr. Frances R. Pick and Dr. Bahram Samanfar, from the Universities of Ottawa and Carleton, are also to be appreciated for their valuable advice and support in helping me to improve the quality of my research work from my initial proposal to my completed thesis.

I am incredibly grateful for the friendship that has been shown to me by the members of the Cloutier and You teams for the constant care, support, comments, and guidance: Hamna Shazadee, Sridhar Ravichandran, Bosen Jia, Sampurna Bartaula, Madeleine Lévesque-Lemay, Tara Edwards, Akshaya Vasudevan, and Chunfang Zheng.

I would like to express my sincere gratitude to the International Doctoral and Admission Scholarship for providing me with a full scholarship that allowed me to pursue my academic goals, as well as to the University of Ottawa for providing me with teaching assistant positions throughout all of the Fall and Winter terms. I am grateful to AAFC for hiring me as a Research Affiliate Program student. During my time here, I have learned a lot that will benefit both my research career and my personal development. Being a part of such a wonderful team was a great experience, and the memories we made together will live on in my mind forever. I want to thank everyone who contributed to the funding of this research.

Acknowledgments

Thank you to my parents, sister, and brothers for all of your inspiration and support. It would take more than one lifetime to repay what you have done for me. Last but not least, I want to express my gratitude to the frontline personnel who have been risking their lives daily throughout COVID-19 to rescue ours.

Table of Contents

Table of Contents

Table of Contents

Table of Contents

List of Tables

List of Tables

List of Figures

List of Figures

List of Figures

Appendices

Appendices

Acronyms

ABC	ATP-binding cassette
AP2/ERF	Apetala2/ethylene responsive factor
ARD	Root diameter
Ba	Barium
BCV	Biological coefficient of variation
BDLWt	Bundle weight
bHLH	Basic helix-loop-helix protein
BLUP	Best linear unbiased predictor
BRR	Bayesian ridge regression
bZIP	Basic leucine zipper
Cd	Cadmium
CDF	Cation diffusion facilitator
CIM	Composite interval mapping
Co	Cobalt
CS	Cold stress
CT	Canopy temperature
Cu	Copper
CV	Cross-validation
DH	Doubled haploid
DS	Drought stress
DTF	Days to flowering
DTM	Days to maturity
EMG	Emergence
EMMAX	Efficient mixed-model association expedited
EN	Elastic net
FDR	False discovery rate
Fe	Iron
FW	*Fusarium* wilt
GBLUP	Genomic best linear unbiased prediction
GBTS	Genotyping by target sequencing
GCA	General combining ability
GE	Genotype-by-environment
GEBV	Genomic estimated breeding value
GLM	General linear model
GO	Gene ontology
GS	Genomic selection
GWAS	Genome-wide association study/ies
HB	Haplotype block
HM	Heavy metal

Acronyms

HMA	Heavy metal associated
HMM	Hidden Markov model
HS	Heat stress
HSP	Heat-shock protein
HT	Plant height
HTP	High-throughput phenotyping
HTS	High-throughput sequencing
IM	Interval mapping
IR	Irrigated
JTT	Jones, Taylor, and Thornton
Ka	Non-synonymous substitution
Ks	Synonymous substitution
LASSO	Least absolute shrinkage and selection operator
LB	Leaf blotching
LD	Linkage disequilibrium
LEA	Late embryogenesis abundant
LF	Leaf firing
LIN	Linolenic acid content
LM	Linkage map
LOD	Logarithm of the odd
MAPKKK	Mitogen-activated protein kinase kinase kinase
MAS	Marker-assisted selection
MaxR	Maximum number of root
ML	Maximum likelihood
MLM	Mixed linear model
MLSM	Multi-locus statistical model
Mn	Manganese
MRA	Multiple regression association
mrMLM	Multi-locus random-SNP-effect mixed linear model
MTP	Metal tolerance protein
MYA	Million years ago
NBD	Nucleotide-binding domain
Ni	Nickel
NIL	Near-isogenic line
NIR	Non-irrigated
NWA	Network area
NWL	Network length
NWSA	Network surface area
NWV	Network volume
OIL	Oil content
PM	Powdery mildew
Pop	Population
PP2C	Protein phosphatase

Acronyms

QTL	Quantitative trait locus
QTLs	Quantitative trait loci
QTN	Quantitative trait nucleotide
RF	Random forest
RGA	Resistant gene analog
RIL	Recombinant inbred line
RKHS	Reproducing Kernel Hilbert Space
ROS	Reactive oxygen species
RPM	Reads per million
RR	Ridge regression
RR-BLUP	Ridge-regression best linear unbiased prediction
SBP	SQUAMOSA promoter binding protein box gene family
SCA	Specific combining ability
SDWt	Shoot dry weight
SLSM	Single-locus statistical model
SNP	Single nucleotide polymorphism
SOD	Superoxide dismutase
Sr	Strontium
SS	Salt stress
SSA	Selective sweep analysis
SVM	Support vector machine
SWEET	Sugars will eventually be exported transporter
Ti	Titanium
TMD	Transmembrane domain
TSW	Thousand seed weight
WGD	Whole genome duplication
YLD	Yield
Zn	Zinc

Chapter 1: Genomic Solutions to Enhance Abiotic Stress Resistance in Flax

Chapter 1

Designing Genomic Solutions to Enhance Abiotic Stress Resistance in Flax

Nadeem Khan[1,2], Frank M. You[1,*], and Sylvie Cloutier[1,2,*]

[1]Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, 960 Carling Avenue, Ottawa, Ontario K1A 0C6, Canada.

[2]Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa, ON K1N 6N5, Canada.

*Correspondance: frank.you@agr.gc.ca; sylviej.cloutier@agr.gc.ca

Sections 1.1-1.2 describe the context of the thesis and sections 1.3-1.6 are a published book chapter that serve as introductory chapter because it covers a substantial proportion of the literature review of the thesis. The remaining background is covered in chapter 2: Khan N, You FM, Cloutier S (2022) Designing genomic solutions to enhance abiotic stress resistance in flax. In: Kole C (ed) Genomic designing for abiotic stress resistant oilseed crops. Springer, Cham. https://doi.org/10.1007/978-3-030-90044-1_8, chapter 8, pp 251-283.

Authors' contribution

Conceptualization, methodology, formal analysis, investigation, writing—original draft preparation, N.K.; writing—review and editing, supervision, project administration, funding acquisition, FMY and S.C. All authors read and approved the final manuscript.

1.1 Preface of this Thesis: Scope and Purpose

Flax is a self-pollinated crop widely cultivated for fiber and oilseed production. Its genome (v1.0) was first sequenced using the Illumina short-read technology (Wang et al. 2012b) and an improved pseudomolecule-based assembly (v2.0) was produced through the incorporation of physical, genetic and optical maps (You et al. 2018b). The research presented herein takes advantage of these genomic resources to provide genomic solutions to improve abiotic stress tolerance or resistance in flax. Sections 1.1 and 1.2 of chapter 1 and the entirety of chapter 2 are a comprehensive overview of four major strategies, namely, genomic cross prediction, GWAS, GS and gene family identification for the improvement of flax breeding programs. Recent advances in genome sequencing technologies have made it possible to rapidly and cost-effectively generate high throughput genotypic data, rendering GS a viable breeding method. GS uses markers to obtain GEBVs, which serves as proxy for traits and consequently decrease, but does not eliminate, the need for phenotyping. Therefore, GS not only saves time, but also resources often devoted to phenotyping. One of the goals of this study was to develop GS models that predict the GEBVs of individuals using both GWAS-based and random-SNP marker datasets. Breeders can employ this strategy to increase genetic gain and improve GEBV accuracy. The findings of this study will aid breeders in choosing superior individuals to include in the field trials and/or to use as parents or eliminate inferior individuals, thereby freeing resources that can be harnessed to improve breeding efficiency.

Both ABC transporter and HMA genes are known for their role in Cd accumulation in plants; hence, one of the objectives was the annotation of the genome to identify all flax ABC transporters and HMA genes (chapter 3), including candidate gene identification for traits such as seed Cd content. Finally, this study was designed to assess the distribution patterns of flax genetic diversity, and to perform both GWAS and GS for metal content (chapter 4).

To accomplish the aims of my research study and the above outputs, the following objectives were established:

- To compare GWAS models and GS strategies for several traits in flax in order to determine the best strategy;
- To characterize the ABC transporter and HMA gene families potentially responsible for heavy metal (HM) accumulation and identification of the major gene(s) involved in Cd accumulation:
- To perform GWAS analyses using both single- and multi-locus models in order to identify quantitative trait nucleotides (QTNs) associated with HM accumulation in seed;
- To identify candidate genes for all the traits harbored at the major QTN loci;
- To perform GS using several models and to compare their accuracy;
- To perform GS with trait-defined QTN and random genome-wide SNP datasets and to compare their respective accuracy for Cd content;

1.2 Hypotheses

1. GS models using trait-defined QTN are likely to achieve a higher predictive ability than random genome-wide SNP datasets (Chapters 2 and 4).

2. Genomic regions associated with HM contents in flax seed can be identified using a combination of single- and multi-locus GWAS models applied to phenotypic data collected from a mini-core collection of 165 flax accessions grown at multiple locations and years and a SNP dataset obtained from its resequencing (Chapter 4).

3. Whole genome and tandem duplications led to the development of both ABC transporter and HMA gene families in flax (Chapter 3).

Abstract

Flax is a self-pollinated crop widely cultivated for fiber and oilseed production. Poor rooting system architecture and abiotic stresses such as drought, extreme temperature, salt, and cadmium accumulation can significantly affect the growth of flax, resulting in severe production losses. As such, improving the root system of flax cultivars and increasing their resistance or tolerance to abiotic stresses are important to achieve sustainable flax production. In recent years, substantial progress has been made towards the generation of flax genomic resources. Reduced cost of high-throughput sequencing (HTS) has fueled the large-scale production of sequence data, allowing an intensification in the genotyping of numerous populations for quantitative trait locus (QTL) mapping. Advanced genome-wide association studies (GWAS) combined with candidate gene identification through bioinformatics approaches make it possible to identify many large- and small-effect quantitative trait loci (QTLs) and candidate genes associated with traits of agronomic importance. To date, a total of 521 QTLs associated with abiotic stress-related traits have been identified in flax. These QTLs constitute markers for genomic selection (GS) to predict breeding values of populations under selection; the goals being to improve the accuracy and efficiency of selection, reduce cost, and shorten the breeding cycles. Combined with genomic cross prediction, this GS strategy offers a new effective approach to predict the breeding performance of crosses and evaluate parents based on their genotype. Here we provide an overview of genome-wide QTL mapping, gene family identification, and outline the potential of a combined GS strategy with genomic cross prediction for breeding improvement of abiotic stress tolerance in flax.

Abstract

Keywords: Flax ● Genome-wide markers ● Quantitative trait loci (QTLs) ● Genome-wide association studies (GWAS) ● Genomic selection (GS) ● Genomic cross prediction ● Abiotic stress

Chapter 1: General Background

1.3 Introduction

Stress in plants is described as the external conditions that adversely disturb plant growth and productivity, mainly as a consequence of their sessile nature (Zhu 2016). A plant must cope with adverse circumstances such as drought, heat, cold, salt, and metal toxicity. Collectively, these stresses are known as abiotic stresses, and they pose a threat to agricultural productivity worldwide (Wang et al. 2004; Wani et al. 2016). Understanding the mechanisms for coping with abiotic stress in plants will lead to the design of counteracting strategies that ultimately contribute to ensure food security and agricultural sustainability for a growing world population. Global temperatures are predicted to rise by approximately 0.2°C per decade over the next thirty years, and this rise in temperature is forecasted to affect crop productivity (Bailey-Serres et al. 2019). Reduction in freshwater availability and shrinking of biodiversity have already altered crop growth as exemplified by yield reduction in affected regions (Keesing et al. 2010; Brown et al. 2019). Such changes are not uniformly distributed around the world. For instance, the European Mediterranean countries are expected to experience warmer temperatures with regular drought stress (DS), while temperature increases in North America are predicted to affect the rainy season cycle from spring to winter (Hopkin 2005). Climatic changes such as heat stress (HS) can lead to plant death. One study showed that HS and DS caused up to 6.2% reduction in cereal production between 2000 and 2007 (Lesk et al. 2016). Paradoxically, as a result of the global rise in temperature, excessive amounts of rainfall are also a threat to crop production. For example, yield losses in maize totaling US$ 10B from 1989 to 2016 were attributed to heavy rainfall, a figure of comparable magnitude to the losses caused by extreme drought in the USA (Li et al. 2019b). Salt stress (SS) is another growing threat to plant growth because of the increasing amount of salinization worldwide (Munns and Tester 2008). In high salt concentration, i.e., 200 mM, most plants are unable to survive

(Flowers 2008; Zhou et al. 2016). High salinity leads to an increase in ionic toxicity and osmotic pressure that disturbs growth from seedling to flowering (Zhao et al. 2010; Feng et al. 2014; Guo et al. 2018). SS and DS are often associated, thereby compounding the challenge (Ashraf and Foolad 2007; Slama et al. 2015). Further, heavy metal (HM) stress has also become a concern worldwide due to extensive industrialization and because it directly and indirectly affects soil health and crop productivity (Shahid et al. 2015). HMs can cause significant yield losses and disturb various physiological and molecular responses in plants (Panuccio et al. 2009; Hassan et al. 2017). Most HMs, for example zinc, copper, manganese, cobalt, and nickel, are crucial for various biological processes (Salla et al. 2011). In contrast, others, including arsenic, lead, cadmium (Cd), are highly toxic and adversely affect plant growth and productivity (Xiong et al. 2014; Pierart et al. 2015). Several studies have identified candidate genes or loci involved in HM response, particularly for Cd stress (Wu et al. 2015; Chen et al. 2018; Zhao et al. 2018). Overall, it is imperative to develop stress resistant cultivars for two main reasons: i) to respond to the global environmental changes, and ii) to meet the global demand for safe food.

Flax (*Linum usitatissimum* L.) is a self-pollinated crop that is widely grown for fiber and oilseed productions. It is a tap-rooted plant with a shallow root system that depends largely on moisture and nutrient resources mostly located in the top 70 cm of soil (Hocking et al. 1997; Kar et al. 2007; Hall et al. 2016). Drought (Soto-Cerda et al. 2019; Soto-Cerda et al. 2020), extreme temperatures, whether heat or cold stresses (Cross et al. 2003; Tchoumtchoua et al. 2019), salt (Hashem 2011; Yu et al. 2014), and mineral toxicity (Angelova et al. 2004) negatively impact the growth and development of flax. To facilitate abiotic stress studies, tremendous progress has been made towards the development of flax genomic resources including the first assembly of the flax reference genome sequence into 15 pseudomolecules, its BioNano optical map, a consensus

genetic map, and a first-generation haplotype map constructed using 407 widespread accessions and ~1.7 million single nucleotide polymorphisms (SNPs) (Cloutier et al. 2012; You et al. 2018b). These genomic resources lay a solid foundation for flax improvement, including abiotic stress-related traits. Traditional approaches to improve abiotic stress resistance have not been successful in part because these traits are often controlled by many genes. Identification of key candidate genes through genome-wide association studies (GWAS) and other genome-wide investigations can be capitalized upon to address the problems associated with abiotic stress tolerance. Several candidate genes have been identified successfully through GWAS, especially for root traits and DS tolerance (Sertse et al. 2019; Soto-Cerda et al. 2019; Soto-Cerda et al. 2020). These genes can be pyramided into elite genotypes via marker-assisted selection (MAS) to produce abiotic stress tolerant cultivars.

This review summarizes the outcomes of genomic studies and their potential uses in mitigating plants' responses to abiotic stresses. First, the importance of GWAS in flax and other species experiencing a range of abiotic stresses is discussed prior to addressing genome-wide strategies for gene family identification. Recent concepts and novel strategies applied to abiotic stresses such as genomic selection (GS) and genomic cross prediction are outlined in view of their applications in flax breeding. This knowledge holds potential in genomic studies for abiotic stress tolerance, not only in flax but also in other economically important species. Candidate gene mining can and must be performed imminently to benefit the development of abiotic stress resistant or tolerant cultivars in a timely manner and in view of the potential severe consequences associated with climate change.

1.4 Genomics for Crop Improvement

Rapid advances in genomics and the availability of inexpensive and reliable HTS technologies make it possible to sequence billions of fragments of DNA sequences. Nowadays, short- and long-read sequence technologies are amenable to a wide range of genomes compared to first-generation sequencing (Mardis 2008; Amarasinghe et al. 2020). In modern crop genomics, DNA markers such as SNPs are broadly used because of their high polymorphism, ubiquitous nature, and low cost. Large sets of genotypes with thousands to millions of markers are routinely produced in several plant species. Further, whole-genome resequencing data is gaining popularity and provides insights into structural diversity (Voss-Fels and Snowdon 2016). The recent development in HTS and the use of genomic approaches will provide key strategies to promote efficiency and precision in flax breeding research including studies of abiotic stress tolerant and sensitive genotypes. For instance, the reference genome sequence of flax and its first-generation haplotype map comprising 407 accessions and ~1.7 million SNPs have recently been updated (You et al. 2018b; Wang et al. 2019a). A better understanding of fundamental molecular information has the potential to improve flax breeding practices and accuracy towards superior abiotic stress tolerance. Finding the relationship between genotypic and phenotypic variations using GWAS can provide insights to assist flax breeding programs in designing the most efficient breeding strategies. Once improved, genotypes adapted to respond to abiotic stresses like drought, heat, and metal toxicity for instance, offer not only a practical genetic improvement but also a sustainable solution towards bridging the food security gap. In several crops worldwide, yield increase or stability of crops subjected to abiotic stresses are currently major goals of plant breeding programs (Ciesla et al. 2016; Landi et al. 2017). Generally, accurate and reliable phenotyping is labor-intensive and time-consuming. The use of high-throughput phenotyping (HTP), GS, and genomic cross prediction offers

alternative ways of tackling this issue to achieve accuracy and precision in breeding. Currently, a combination of abiotic stress-related indices, methods, and population types have been used to identify QTLs and candidate genes including GWAS, genome-wide identification of candidate genes, doubled haploids (DHs), near-isogenic lines (NILs), bi-parental population, and others (Khadivi-Khub 2014; Xia et al. 2014; Fu et al. 2017; Hao et al. 2018; Khan et al. 2018b; Sukumaran et al. 2018; Soto-Cerda et al. 2019). Here we will discuss studies focusing on the identification of abiotic stress-related QTLs including traits such as root architecture, drought, water efficiency, heat, cold, salt, and Cd stresses, mainly in flax, but also in other crops.

1.4.1 Genome-wide Association Studies (GWAS)

GWAS aims to establish associations between genotype and phenotype and to quantify the contribution of genetic variants across the genomes of many individuals to the measured traits. They are robust methodologies for predicting candidate genes that underpins complex traits (Nicod 2016). GWAS exploit a large set of genetic variants to identify the specific subset linked with a trait of interest (Tam et al. 2019). Such studies can be based on multiple environments, years, and traits. GWAS have been successful in identifying QTLs for abiotic stress-related traits including root architecture and drought tolerance (Sertse et al. 2019; Soto-Cerda et al. 2019; Liu et al. 2020; Soto-Cerda et al. 2020; Yonis et al. 2020), heat and cold tolerances (Chen et al. 2017; Lafarge et al. 2017; Song et al. 2018; Zhang et al. 2018a), and Cd stress (Wu et al. 2015; Chen et al. 2018) in flax, as well as in many other plant species as summarized in **Table 1.1**.

The QTLs identified in early studies were mostly obtained using the general linear model (GLM) or the mixed linear model (MLM) which were the main single-locus methods at the time. The power of GLM and MLM lies in their ability to detect large effect quantitative trait nucleotides (QTNs) or QTLs compared to the multi-locus models that were developed later and that are better

suited to detect smaller effect QTNs or QTLs. An example of the latter models is the multi-locus random-SNP-effect mixed linear model (mrMLM) (Wang et al. 2016b; He et al. 2019b). Using both single- and multi-locus models, a total of 521 QTLs for root traits and drought stress were previously identified in flax (Sertse et al. 2019; Soto-Cerda et al. 2019; Soto-Cerda et al. 2020; Sertse et al. 2021). Multi- and single-locus models are somewhat complementary and their combined use enables the identification of both small and large effect QTNs for complex and low heritability traits. A common concern in GWAS is the difficulties associated with the identification of causal variants and genes associated with abiotic stress-related traits that have relatively low heritability. Such candidate genes must be further validated through functional genomic experiments. This problem may be overcome by increasing the sample size; however, this would greatly increase the cost of the experiment. The other problem typically associated with GWAS is the elimination of markers with low allele frequency, i.e., markers with alleles present in less than 5-10% of the individuals (Hawkesford and Griffiths 2019). Minor alleles can be difficult to ascertain and their statistical power of detection of association is poor because it is derived from few individuals; hence, the minor allele frequency cut-offs of 5-10%, commonly used depending on the population size, precludes genotype-phenotype associations of these rare alleles. Proper design is critical in GWAS, including accounting for the genetic structure of the population to minimize both false-negative and false-positive results. With proper methodology, GWAS has been applied successfully to identify numerous key genes. For example, many of the genes commonly used by the US Food and Drug Administration as molecular targets in several drugs were identified by GWAS, and this success reinforces the continuous efforts to improve the methodology and to consider its application to large-size populations (Altshuler et al. 2008; Hirschhorn 2009). This section summarizes several studies focusing strictly on GWAS for abiotic

stress-related traits including root architecture, water efficiency, and tolerance to drought, heat, cold, salt, and Cd.

Table 1.1 Quantitative trait loci (QTLs) identification in different species in response to abiotic stresses.

Trait	Species	Pop size	No. markers	Method	Statistical model	No. QTL	No. traits	Reference
Root	Flax	115	7,707 & 3,243	GWAS	6 multi- & 1 single-locus	228	16	Sertse et al. (2019)
	Barley	221	6,336	GWAS	MLM	55	6	Jia et al. (2019)
	Maize	384	681,257	GWAS	LM & MLM	268	22	Pace et al. (2015)
	Maize	297	131,271	GWAS	GLM/MLM	34	13	Wang et al. (2019a)
	Wheat	199	12,109	QTL mapping	CIM	18	8	Bhandari et al. (2019)
Drought	Flax	115	12,316	GWAS	6 multi- & 1 single-locus	148	11	Sertse et al. (2021)
	Flax	41	170,534	GWAS	3 multi- & 2 single-locus	118	4	Soto-Cerda et al. (2020)
	Flax	105	394	SSA	MRA	27	8	Soto-Cerda et al. (2019)
	Soybean	259	4,616	GWAS	MLM	15	3	Liu et al. (2020)
	Chrysanthemum	88	92,811	GWAS	GLM & MLM	137 & 14	7	Su et al. (2019)
	Maize	144	45,868	GWAS	GLM & MLM	47	6	Zhang et al. (2013)
	Wheat	100	15,600	GWAS	MLM	75	6	Mathew et al. (2019)
	Barley	156	4,438	GWAS	MLM	14	5	Wehner et al. (2016)
Heat	Sorghum	374	13,987	GWAS	MLM	14	2	Chen et al. (2017)
	Wheat	200	15,574	GWAS	MLM	15	4	Maulana et al. (2018)
	Rice	167	1,3160	GWAS	MLM	14	20	Lafarge et al. (2017)
	Rapeseed	88	37,539	GWAS	GLM & MLM	20	3	Rahaman et al. (2018)
Cold	Rice	115	67,511	GWAS	MLM	26	5	Song et al. (2018)
	Rice	295	44,100	GWAS	MLM	67	3	Wang et al. (2016b)
	Rice	202	157	GWAS	MLM	48	5	Schläppi et al. (2017)
	Rice	249	3,867	GWAS	MLM	47	2	Zhang et al. (2018a)
	Maize	306 & 292	49,585	GWAS	MLM	47 & 4	2	Revilla et al. (2016)
	Barley	184	1,536	GWAS	MLM	12 & 7	2	Visioni et al. (2013)
Salt	Rice	478	162,529	GWAS	6 multi-locus	378	21	Cui et al. (2018)
	Rice	208	395,553	GWAS	mrMLM	20	11	Naveed et al. (2018)
	Rice	708	1,101,404	GWAS	MLM	2255	8	Liu et al. (2019a)
	Soybean	305	37,573	GWAS	EMMAX & MLMM	29	4	Do et al. (2019)
	Barley	350	~24,000	GWAS	MLM	52	4	Mwando et al. (2020)
	Alfalfa	291	7,401	GWAS	MLM	53	5	Liu et al. (2019g)
	Alfalfa	304	6,862	GWAS & GP	MLM	27	4	Medina et al. (2020)

Trait	Species	Pop size	No. markers	Method	Statistical model	No. QTL	No. traits	Reference
Cadmium	Barley	100	1,536	GWAS	MLM	59	4	Wu et al. (2015)
	Rapeseed	419	19,167	GWAS	GLM & MLM	98	3	Chen et al. (2018)
	Rice	312	183,884	GWAS	MLM	14	2	Zhao et al. (2018)
	Maize	269	43,737	GWAS & QTL mapping	GLM & MLM	5	3	Zhao et al. (2018)

Pop: population; GWAS: genome-wide association study; SSA: selective sweep analysis; GP: genomic prediction; QTL: quantitative trait loci; CIM: composite interval mapping; MLM: mixed linear model; LM; linear model; GLM: general linear model; MRA: multiple regression association; EMMAX: efficient mixed-model association expedited; MLMM: multi-locus mixed model; mrMLM: multi-locus random-SNP-effect mixed linear model

1.4.1.1 Root Characters

The root system is a critical tissue implicated in many abiotic stress responses but current knowledge of its role is limited because its phenotyping is challenging. Consequently, improvement of root architecture is a daunting task that must be overcome to improve plant adaptation against drought stress, water efficiency, nutrient supplies, and others. Because of the inherent difficulties associated with accurate phenotyping of root traits, breeding efforts to date have focused on altering aboveground traits (Diederichsen 2006; Heller and Byczyńska 2015). Root traits are, however, vital for the availability of water and nutrients, making them crucial for crop improvement under DS conditions (Narayanan et al. 2014). For example, flax cultivars with deeper and denser root systems, specifically in rain-fed agricultural systems, could more effectively access water from deeper soil layers (Dash et al. 2014; Sertse et al. 2019; Soto-Cerda et al. 2019). Drought tolerance is key to producing plants with high fiber yields with minimum adverse impact on fiber quality (El-Hariri et al. 2005). Therefore, the improvement of both aboveground and root traits has the capability to produce well-adapted cultivars for flax fiber production.

Due to the major threats posed by abiotic stresses, efforts are being made to enhance tolerance in flax. QTNs associated with drought tolerance or root traits have been reported in flax (Sertse et al. 2019; Soto-Cerda et al. 2019; Soto-Cerda et al. 2020). Root development is considered to play a significant role in plant nutrient uptake, and consequently, both morphological and physiological traits are important for abiotic stress tolerance (Jia et al. 2019). Incorporating genetic information of the root architecture system into flax breeding practices would benefit efforts to enhance resource efficiency and/or stress tolerance. Recent developments in image software tools and advances in HTP have enabled more refined studies of root architecture (Furbank and Tester 2011;

Hartmann et al. 2011; Fahlgren et al. 2015). GWAS applications have also enabled the identification of important QTLs for root traits (Hochholdinger et al. 2018), some of which have already been applied to develop elite wheat genotypes (Wasson et al. 2014). In flax, a GWAS of 115 accessions grown hydroponically was phenotyped for 15 root and two shoot traits, as well as for the shoot to root dry weight ratio (Sertse et al. 2019). This study tested seven models and identified 228 QTNs for 16 traits. Candidate genes at the QTN loci encoded GRAS transcription factors, mitogen-activated protein kinases, and auxin-related lateral organ boundary proteins among others.

Plants rely on a wide range of protective and adaptive mechanisms against abiotic stresses. Numerous GWAS studies have revealed that root traits are important for breeding practices in crops such as cassava (Yonis et al. 2020), maize (Pace et al. 2015; Wang et al. 2019a), wheat (Bhandari et al. 2019), and barley (Jia et al. 2019). A GS analysis in cassava suggested that root-related traits could be predicted with higher accuracy than yield (Yonis et al. 2020). A comprehensive knowledge of the root architecture system is crucial to understand its direct function in abiotic stress responses such as drought. Both GWAS and GS offer opportunities to expand our knowledge of the pivotal functions of roots in abiotic stress tolerance and hold potential for indirect selection of the difficult-to-phenotype root system.

1.4.1.2 Drought Stress (DS)

In plant breeding, DS is considered to be one of the most noteworthy abiotic stresses affecting agricultural productivity worldwide (Shao et al. 2009). Drought and temperature fluctuations are major environmental threats that limit agricultural crop productivity and often lead to extensive socio-economic crises (Zhao and Dai 2015). In the year 2060, more than 50% of the crop production from earth's arable lands is anticipated to experience water scarcity (Dhankher and

Foyer 2018). DS causes a significant reduction in both yield and in harvested areas when it is combined with extreme temperature fluctuations (Lesk et al. 2016). For example, between 1961 and 2014, the projected combined impact of drought and temperature stresses in maize, soybean, and wheat resulted in yield losses of 11.6, 12.4, and 9.2%, respectively (Matiu et al. 2017). For effective plant growth, development, and productivity, water availability is crucial, and its deficit can result in lower yields or even plant death. For instance, flax fiber yield is highly dependent on water availability and can be reduced by as much as 35–50% when water deficits are experienced during the vegetative stage as a result of the high transpiration rate experienced during this growth phase (Heller and Byczyńska 2015). So far in flax genetic improvement, few researchers have achieved the development of drought resistant genotypes (Diederichsen 2006; Qi et al. 2010; Sharma JC 2012; Asgarinia et al. 2016). Recently, Sertse et al. (2021) conducted a GWAS for 11 traits for three years in irrigated (IR) and non-irrigated (NIR) fields. Six of the 11 traits showed significant variations between IR and NIR conditions. To identify QTNs associated with DS, seven GWAS models were used and a total of 148 QTNs were found to be associated with at least one trait or stress index. Among these QTNs, 16 were deemed to have major effects because they accounted for more than 15% of the genetic variance of the associated trait. Genotypes such as CN101595, CN98566, and fiber types from China outperformed others under drought conditions and were proposed as good drought-resistant resources in flax. Similarly, under DS and IR conditions, Soto-Cerda et al. (2020) investigated four agronomic and four root traits using 170,534 SNPs from 418 flax accessions. Two single- and three multi-locus GWAS models identified 118 QTNs for drought-related traits. The genes proposed as candidates were involved in drought-responsive pathways, root, and vascular tissue development. Similar types of GWAS analyses for

drought and waterlogging conditions have also been performed in soybean (Liu et al. 2020), chrysanthemum (Su et al. 2019), maize (Zhang et al. 2013), and barley (Wehner et al. 2016).

Although the identification of QTLs for DS-related traits remains a challenge because of the polygenic nature of the traits, the magnitude of the genotype by environment interactions, and the low heritability of the traits, the cited studies are promising. Enhanced research on flax is expected to lead to novel insights towards a better understanding of drought tolerance. Taken together, these studies should greatly facilitate MAS breeding not only in flax but also in other species experiencing drought stress.

1.4.1.3 Heat Stress (HS)

Global average temperatures are forecasted to surge by 1-4°C by the end of the 21st century (Driedonks et al. 2016). This could threaten global crop production because these heightened temperatures may cause HS. The development of cultivars resistant to such abiotic stress is needed to ensure sustainability in agricultural production under unfavourable environmental conditions (Bita and Gerats 2013). Limited literature is available regarding the impact of HS on flax. One study reported that HS does not significantly affect flower production but contributes to a reduction in boll formation and seed set (Cross et al. 2003). To date, no GWAS has been conducted on HS-related traits in flax; this may be due to the lack of appropriate phenotyping methods for such complex traits and the inherent difficulties associated with controlling the HS conditions in field environments, thereby limiting measurements to the small number of individuals that could be handled in controlled environments. Reliable and accurate phenotyping is crucial for successful GWAS (Powell et al. 2012). This is why GWAS has been successfully applied to many more biotic and agronomic traits than to abiotic stress-related traits (Soto-Cerda et al. 2018b; You et al. 2018a; He et al. 2019b). Recent studies have revealed the existence of QTLs associated with HS-related

traits that are promising as molecular markers in breeding programs. GWAS was utilized to delve into heat tolerance in *Brassica napus* and identified key candidate genes associated with flowering, male sterility, pollen abortion, and others (Rahaman et al. 2018). Plants were exposed to different HS regimes during the flowering stage and examined for traits such as pollen sterility, sterile/aborted seeds in siliques, and number of siliques on the main raceme. Using 37,539 SNPs and 88 diverse accessions of *B. napus*, a total of 5, 8, and 7 QTNs were discovered for the above traits, respectively (Rahaman et al. 2018). A number of the candidate genes proposed were tissue-specific and temporally expressed. A GWAS of sorghum was conducted to measure HS response for leaf firing (LF) and leaf blotching (LB) at the vegetative stage (Chen et al. 2017). These traits were assessed for up to three years at three locations. With 339 accessions and 13,987 SNPs, nine QTNs were associated with LF and five with LB. The candidate gene analysis showed that most of them were linked to known plant stress response genes including HS. Other similar GWAS as well as meta analyses also identified QTNs associated with HS tolerance and their underlying candidate genes (Acuña-Galindo et al. 2015; Lafarge et al. 2017; Maulana et al. 2018).

Refinement of the identified QTLs would benefit candidate gene identification by narrowing down the search to more focused and relevant regions. Candidate genes can be used as transgene to validate their functional roles in HS tolerance. This strategy has been applied in several species including *Arabidopsis* (Yokotani et al. 2008; Khurana et al. 2017), wheat (Zang et al. 2017), yeast, and rice (Qin et al. 2015).

1.4.1.4 Cold Stress (CS)

Cold or low temperature stresses is one of the major type of environmental stresses that causes crop production losses in many climatic regions, whether temperate, tropical or subtropical (Sanghera et al. 2011). Prolonged CS disturbs plant growth activities and causes growth

retardation, late flowering, stem elongation, and others effects (Patel and Franklin 2009). For instance, losses of up to 26% were observed in rice exposed to a CS during the reproductive stage (Ye et al. 2009). CSs not only limit crop productivity but also determine crops' geographical distribution (Chinnusamy et al. 2007; Thakur et al. 2010). The intensity of the CSs can vary from chilling (0-15°C) to freezing (sub-zero) (Sanghera et al. 2011).

Under CS, many plants, including grapevine, watermelon, *Arabidopsis*, and flax, can produce certain "counteracting" metabolites such as phenolics and flavonoids (Rivero et al. 2001; Król et al. 2015; Schulz et al. 2016; Tchoumtchoua et al. 2019). Flax is mostly grown in temperate regions and cultivars vary greatly in their ability to respond to CS. Winter-type cultivars are generally more resistant to CS than spring types. The two most abundant phenolic compounds found in winter flax (methylated C-glycosylflavonoid swertisin and swertiajaponin) were not detected in spring flax (Tchoumtchoua et al. 2019).

A GWAS analysis performed on 150 accessions of rice landraces using 67,511 SNPs identified 26 QTNs associated with CS and explaining 26-33% of the phenotypic variation (Song et al. 2018). This study suggested that rice landraces are useful sources of cold tolerance and that they could be capitalized upon in plant breeding. Several other rice GWAS identified between 47 and 67 QTLs each for CS-related traits (Wang et al. 2016a; Schläppi et al. 2017; Zhang et al. 2018a). In maize, 275 QTLs were identified using 49,585 SNPs (Revilla et al. 2016). Interestingly, 47 flint inbreds harbored the favourable alleles for six major QTLs but only four dent-type inbreds had the favourable alleles from the test crosses.

In summary, numerous genes have been associated with CS in different species, including *Oryza sativa* (Yoon et al. 2016), transgenic tobacco (Jin et al. 2016), *Verbena bonariensis* (Wang et al. 2020), *Arabidopsis thaliana* (Dai et al. 2007; Visconti et al. 2019), transgenic cotton (Hao et

al. 2018), *Brassica napus* (Savitch et al. 2005), *Zoyia japonica* (Kim et al. 2020), and *Hordeum vulgare* (Gierczik et al. 2019). The identification of functional polymorphism(s) in these genes remains a daunting task. However, further validation of these genes via orthologous gene identification and GWAS could be conducted for CS-related traits to lend confidence prior to their use in flax breeding.

1.4.1.5 Salt Stress (SS)

SS is another abiotic stress that plagues crop productivity worldwide (Munns and Tester 2008). The annual global losses due to soil salinization in irrigated areas have been estimated to be upward of 27B US$ (Qadir et al. 2014). One study reported that soil salinity affects almost half of the global irrigated land area, and about one-fifth of the arable land (Qiao et al. 2014). SS negatively influences crop growth, indirectly by affecting soil water potential and directly by affecting water uptake. The soil water potential decreases with an increase in ion concentration, thereby effectively reducing water availability to the plant (Medina et al. 2020). Generally, SS affects plant's growth in comparable ways to DS and responses to both stresses share physiological mechanisms such as stomatal closure, loss of turgescence and photosynthetic rate, development of reactive oxygen species (ROS), increase in heat-shock proteins (HSPs), and others (Zhu 2000; Chaves et al. 2008; Rosyara et al. 2016).

Soil salinity is a growing problem in agriculture production worldwide. Several strategies have been employed to understand flax's response to SS but these were met with limited success; they include gene expression analyses, biochemical markers, and biotechnological applications (McHughen and Swartz 1984; McHughen 1987; El-Beltagi et al. 2008; Hashem 2011; Yu et al. 2014). To identify SS responsive QTLs in crops, several GWAS have been performed in rice (Cui et al. 2018; Lekklar et al. 2019), barley (Mwando et al. 2020), soybean (Do et al. 2019),

Arabidopsis (Deolu-Ajayi et al. 2019), and alfalfa (Liu et al. 2019g; Medina et al. 2020). One study evaluated seed-germination-related traits of 350 diverse accessions of barley grown under control and SS conditions (Mwando et al. 2020). Approximately 24K markers were used to identify 19 loci containing 52 significant markers that were associated with salt tolerance. A combined GWAS and GS approach was used in alfalfa (Medina et al. 2020). Using three phenotypic datasets and several GWAS models, 27 QTNs were associated with salt tolerance. Both barley and alfalfa studies illustrate the potential for GWAS and GS, not only to identify QTLs for SS tolerance, but also in practical breeding applications.

Several candidate genes have been hypothesized to play a pivotal role in salt tolerance in plants such as rice (Tang et al. 2019), *Arabidopsis* (Gao et al. 2003; Zhang et al. 2019a), soybean (Nguyen et al. 2019), wheat (Jiang et al. 2014), potato (Wang et al. 2019b), and rapeseed (Yang et al. 2019). These candidate genes may be useful in MAS breeding for developing salt resilient genotypes. More insights into the molecular mechanisms underlying SS response is required because they remain generally obscure and, this knowledge is nearly completely lacking in flax. Likewise, the identification by GWAS of genes or loci involved in salt tolerance will provide additional avenues to address this growing problem.

1.4.1.6 Cadmium Stress (Cd)

Cd is a non-essential toxic metal present in both plants and animals and which is abundant in water and the atmosphere. Cd in soils may come from sources such as industrial emission, the extensive and prolonged use of fertilizers, atmospheric deposition, and public waste (Tanhuanpää et al. 2007; Ismael et al. 2019). Kidney and bone damage-causing *itai-itai* disease in Japan has made consumers and scientists aware of the negative effects of Cd when consumed by humans (Roberts 2014). Soils vary in Cd content due to both natural and anthropogenic processes, with

phosphate fertilizers being one of the main culprit sources of Cd. Cd is toxic to most of the plant cells, even at low leaf concentrations of 5-10 µg/g (White and Brown 2010), although a few species have adapted to Cd toxicity and can tolerate as much as 100 µg/g (Broadley et al. 2001; Verbruggen et al. 2009; Lux et al. 2010). Flax seeds bio-accumulate Cd ranging from 0.23–0.72 ppm (Booker 2019). Among crops grown in Canada, flax and durum wheat are particularly sensitive to soil Cd stress (Jiao et al. 2004). Cd is not evenly distributed in the plant and its partitioning is temporally modulated throughout its life cycle. In flax, for example, Cd accumulates in the following tissues in decreasing order: root > stem > leaf > seed (Angelova et al. 2004). Cd uptake by roots can be translocated to the aboveground organs, including seeds, which can have toxic effects on the consumer (Li et al. 2017). GWAS analyses of Cd accumulation have been performed in several species including rapeseed (Chen et al. 2018), bread wheat (Hussain et al. 2020), rice (Zhao et al. 2018), maize (Zhao et al. 2018), and barley (Wu et al. 2015). Several genes have also been reported to be involved in Cd tolerance, such as *OsMTP1* (Yuan et al. 2012; Das et al. 2016), *OsABCG36*, *AtABCC1*, *AtABCC2* (Park et al. 2012; Fu et al. 2019), *AtHMA3*, *AtHMA4* (Verret et al. 2004; Morel et al. 2009), and *GmWRKY142* (Cai et al. 2020). To date, little is known about flax's response at the molecular level to Cd exposure. The flax orthologues to some of the above Cd-associated genes have been identified in a genome-wide analysis and are suggested as candidate genes (Khan et al. 2020). Following validation, these genes can eventually accelerate flax breeding improvement towards the development of low Cd varieties.

Recently, the Cd accumulation in 418 diverse accessions of *B. napus* was evaluated by GWAS with a total number of 19,167 SNPs (Chen et al. 2018). Twenty-five QTLs identified with 98 SNPs dispersed on 15 chromosomes that accounted for 3.49-7.57% of the variation associated with root Cd concentration. In this study, 32 Cd-related genes were explored in regions of 0.33–497.97 kb

from the QTNs. Likewise in maize, 269 accessions were grown in a Cd-contaminated soil to identify loci controlling Cd accumulation at the seedling and maturity stages (Zhao et al. 2018). The GWAS identified the major QTL *qLCd2* which explained 39.8% of the average phenotypic variance across experiments. The candidate gene analyses from the rapeseed and maize studies included genes encoding a cadmium-sensitivity protein, a cadmium/zinc-transporting ATPase, and an iron-regulated transporter.

Taken together, such findings will facilitate future studies and could be useful for flax breeding. Cd content could become an important trade barrier due to its toxicity, making its presence in food and feed undesirable. In addition, Cd is also toxic to plants and affects their growth. Thus, developing varieties with low Cd accumulation is desirable as it is the most economical and environmentally-friendly way to solve this critical issue.

1.4.2 Identification of Gene Families Associated with Abiotic Stresses

Genome-wide investigation of the gene families associated with abiotic and other types of stresses mainly relies on protein-protein sequences from model species such as *Arabidopsis thaliana,* and/or hidden Markov model (HMM) searches (Khan et al. 2018a; Khan et al. 2018b; Khan et al. 2020). Generally, this approach is combined with gene expression studies and, qRT-PCR is used to verify the functionality of the orthologous genes (Wu et al. 2017; Khan et al. 2018b). Advanced resources in both genomics and bioinformatics have facilitated the study of the evolution of plant gene families. The knowledge developed in one plant species can be leveraged to identify abiotic stress-related gene families across multiple species based on sequence homology and/or synteny. Plant genes belonging to families are often clustered. The characterization of individual family members therefore becomes paramount to understanding their functional

diversity. Both sequence homology and synteny might result in the identification of orthologous candidate genes in the species of interest. In the last decade, a number of studies have shown that transcription factors including AP2/ERF (Xie et al. 2019), MYB (Dubos et al. 2010), NAC (Liu et al. 2019f), WRKY (Jiang et al. 2017), bZIP (Dröge-Laser et al. 2018), homeobox (Khan et al. 2018b), and bHLH (Sun et al. 2018) play a vital role in abiotic stress tolerance. Functional genomic studies in many plant species have also validated candidate genes associated with abiotic stresses. Several examples can be cited such as an improved root architecture system and enhanced SS tolerance in rice (*RCc3*) (Li et al. 2018b), drought and SS tolerances (*OsMYB6)* (Tang et al. 2019), heat tolerance (*TaMBF1c*) (Qin et al. 2015), low temperature tolerance (*CBF1, CBF2,* and *CBF3*) leading to an increase in freezing tolerance (Gilmour et al. 2004), and low Cd accumulation (*OsHMA3* and *OsABCG36*) (Sasaki et al. 2014; Fu et al. 2019). Using such strategy, we recently identified nine candidate genes for Cd accumulation in flax (Khan et al. 2020). These genes hold potential to assist in solving the Cd toxicity problem in flax. As a whole, these functional genomics studies have contributed to the validation of genes and the underlying mechanisms by which plants respond to abiotic stresses.

GWAS and genome-wide gene family analyses are complementary. Studies have shown that bioinformatics and functional investigation of candidate genes associated with abiotic stress were synergistic (Shaban et al. 2018; Agarwal et al. 2019; Yan et al. 2019; Zhang et al. 2020c). Genome-wide gene family studies usually provide insights into evolutionary events such as gene duplication rate of evolution, and an understanding of the phylogenetic relationships for examples (Wu et al. 2007). However, genome-wide gene is unable to provide a direct association with a phenotypic trait such as is accomplished by GWAS. In the last decade, the genomic resources available in public databases have facilitated the evolutionary studies of gene families. As such, publicly

available genomic information has significantly empowered comparative gene family analyses across species to reveal their functional diversity. The outcome provides focus to further the functional genomic characterization of these genes towards the improvement of stress tolerance in crops. A list of gene families which mediate response to abiotic stress tolerances in several species is provided (**Table 1.2**). Finding and mining candidate genes involved in abiotic stress response is an important step to unraveling and manipulating stress tolerance in flax.

Table 1.2 List of abiotic stress-related gene families and their functions.

Abiotic stress-related gene families	Functions	References
Lipoxygenase gene family	Biotic and abiotic stresses	Shaban et al. (2018)
Basic leucine zipper (bZIP) transcription factor	Heat, salinity, and drought	Agarwal et al. (2019)
WRKY transcription factors	Multiple abiotic stresses	Yan et al. (2019)
Mitogen-activated protein kinase kinase kinases (MAPKKKs)	Drought stress	Zhang et al. (2020c)
Superoxide dismutase (SOD)	Drought, heat, cold and salinity	Verma et al. (2019)
Apetala2/Ethylene responsive factor (AP2/ERF) transcription factor	Hormone and abiotic stresses	Xie et al. (2019)
MYB transcription factors	Biotic and abiotic stresses	Dubos et al. (2010)
NAC transcription factor	Abiotic stress	Liu et al. (2019f)
Basic helix-loop-helix protein (bHLH) transcription factor	Drought, salt, and cold stress	Sun et al. (2018)
Sugars will eventually be exported transporters (SWEET)	Biotic and abiotic stresses	Miao et al. (2017)
Heat shock proteins (HSPs)	Biotic and abiotic stresses	Schöffl et al. (1998)
Phospholipase D Gene	Abiotic stress	Lu et al. (2019)
SQUAMOSA promoter binding protein (SBP)-box gene family	Abiotic stress	Zhang et al. (2020a)
Late embryogenesis abundant (LEA)	Abiotic stress	Liu et al. (2019c)
Protein phosphatases (PP2C)	Abiotic stress	Khan et al. (2019b)
ATP-binding cassette (ABC)	Biotic and abiotic stresses	Khan et al. (2020)
Heavy metal ATPase (HMA)	Metals toxicity	Pierart et al. (2015)
Cation diffusion facilitator (CDF)	Metal ion uptake and transport in plants	Li et al. (2018c)
Plant metal tolerance proteins (MTPs)	Metals toxicity	Liu et al. (2019e)

1.4.3 Novel Breeding Strategies Using Genomic Selection (GS) and Genomic Cross Prediction

GS and genomic cross prediction have recently emerged as promising and effective breeding strategies. GS is designed to generate genetic estimated breeding values (GEBVs) for many traits with the objectives of high prediction accuracy, low cost and reduced breeding cycles. Similarly, both GS and genomic cross prediction hold potential in improving the effectiveness in cross prediction and selection of breeding parents. In flax breeding programs, crosses are made to generate segregating populations with favourable genetic variation; however, prediction of cross performance and selection of parents remain challenging. The application of GS and genomic cross prediction can help address some of these challenges due to the remarkable increase in the amount of high quality genomic sequence data and the decrease in the cost of HTS of the last several years (Bevan et al. 2017). Here, we will outline the potential of their applications in flax breeding programs.

1.4.3.1 Genomic Selection (GS)

GS is a form of MAS that is relying either on highly dense random genome-wide markers or on the QTLs identified from GWAS associated with the target trait(s). The main purpose of GS is to improve selection efficiency, precision, and decision-making by predicting GEBVs of breeding populations. GS takes advantage of a dense marker saturation or QTLs to predict phenotypes solely based on genotypic information. The original idea behind GS was introduced in 2001 (Meuwissen et al. 2001). Since then, remarkable progress has been made in improving biotic, abiotic and agronomic traits in species such as flax (Wang et al. 2019a; Lan et al. 2020), wheat (Bassi et al. 2016), maize (Albrecht et al. 2011), and soybean (Matei et al. 2018) to name a few. For abiotic

stress-related traits, GS has been applied to drought (Li et al. 2019a; Velazco et al. 2019; Wang et al. 2019a), salt (Medina et al. 2020), heat (Yuan et al. 2018), and cold stresses (Jähne et al. 2019).

GS success hinges on two main elements: statistical models and the choice of markers. Many GS models have been developed and tested for multiple traits with inconsistent success rates as evaluated based on their accuracy and the genetic gains made from selection. For example, seven GS models were tested for agronomic traits in maize (Shikha et al. 2017). Of those, Bayes B had a higher prediction accuracy than random regression (RR), least absolute shrinkage and selection operator (LASSO), Elastic net (EN), Bayes A, random forest (RF), and reproducing kernel hilbert space (RKHS). Several additional studies investigating the potential for GS models to predict abiotic stress-related traits including DS, SS, HS, and CS are summarized (**Table 1.3**). The overall accuracy and performance of the GS models depend on several factors such as the model performance, the complexity and heritability of the traits, the density of markers, the overall number of SNPs, and the marker selection (Desta and Ortiz 2014). At the moment, the ridge regression best linear unbiased prediction (RR-BLUP) model is commonly used to predict abiotic stress-related traits because it often outperformed most other models (**Table 1.3**). However, it is quite possible that other models would prove superior if the choice of a different subset of markers or other experimental designs was used. For instance, machine learning methods such as support vector machine (SVM) and RF outperformed RR-BLUP for yield in alfalfa (Medina et al. 2020).

Table 1.3 Models commonly used for genomic selection (GS) to predict abiotic stress-related traits in crops.

Models tested	Best model	Trait	Species	Reference
RR, LASSO, EN, Bayes A, Bayes B, RF, RKHS	Bayes B	DS	Maize	Shikha et al. (2017)
RR-BLUP	RR-BLUP	DS	Maize	Wang et al. (2019c)
GBLUP	GBLUP	DS	Maize	Dias et al. (2018)
RR-BLUP	RR-BLUP	DS & HS	Maize	Yuan et al. (2018)
RKHS, GBLUP	RKHS	DS	Rice	Bhandari et al. (2019)

Models tested	Best model	Trait	Species	Reference
RR-BLUP, BRR, LASSO	RR-BLUP	DS	Chickpea	Li et al. (2018a)
GBLUP	GBLUP	DS	Sorghum	Velazco et al. (2019)
RR-BLUP	RR-BLUP	CS	Soybean	Jähne et al. (2019)
RR-BLUP, Bayes A, Bayes B, Bayes C, BRR, LASSO, SVM, RF	SVM & RF	SS	Alfalfa	Medina et al. (2020)

RR: Ridge Regression; LASSO: Least Absolute Shrinkage and Selection Operator; EN: Elastic Net, RKHS: Reproducing Kernel Hilbert Space; RF: Random Forest; RR-BLUP: Ridge-Regression Best Linear Unbiased Prediction; BRR: Bayesian Ridge Regression; GBLUP: Genomic Best Linear Unbiased Prediction; SVM: Support Vector Machine; DS: Drought stress; HS: Heat stress; CS: Cold stress; SS: Salt stress

Selection of markers has proven to be an important factor for increasing prediction accuracy of GEBVs. For instance, GS models based on QTLs as fixed effects consistently outperformed models built on random genome-wide SNP datasets in flax (Wang et al. 2019a; Lan et al. 2020). These QTLs were detected using both single- and multi-locus GWAS statistical models. The GS results obtained with such QTL datasets also demonstrate the robustness and reliability of these QTLs. The high accuracy obtained with the QTLs is hypothesized to be a consequence of the reduction in background noises through the removal of unrelated markers. This noise may arise from genome-wide unrelated markers and through imputation (Rutkoski et al. 2012). Generally, QTL input datasets improve GS accuracy over genome-wide random SNP input datasets (Deshmukh et al. 2014). The combination of single- and multi-locus GWAS models is recommended to identify a comprehensive QTL dataset which in turns will be used as marker input dataset to achieve high accuracy and consistency in GEBVs in GS modeling (Lan et al. 2020) (**Fig. 1.1**).

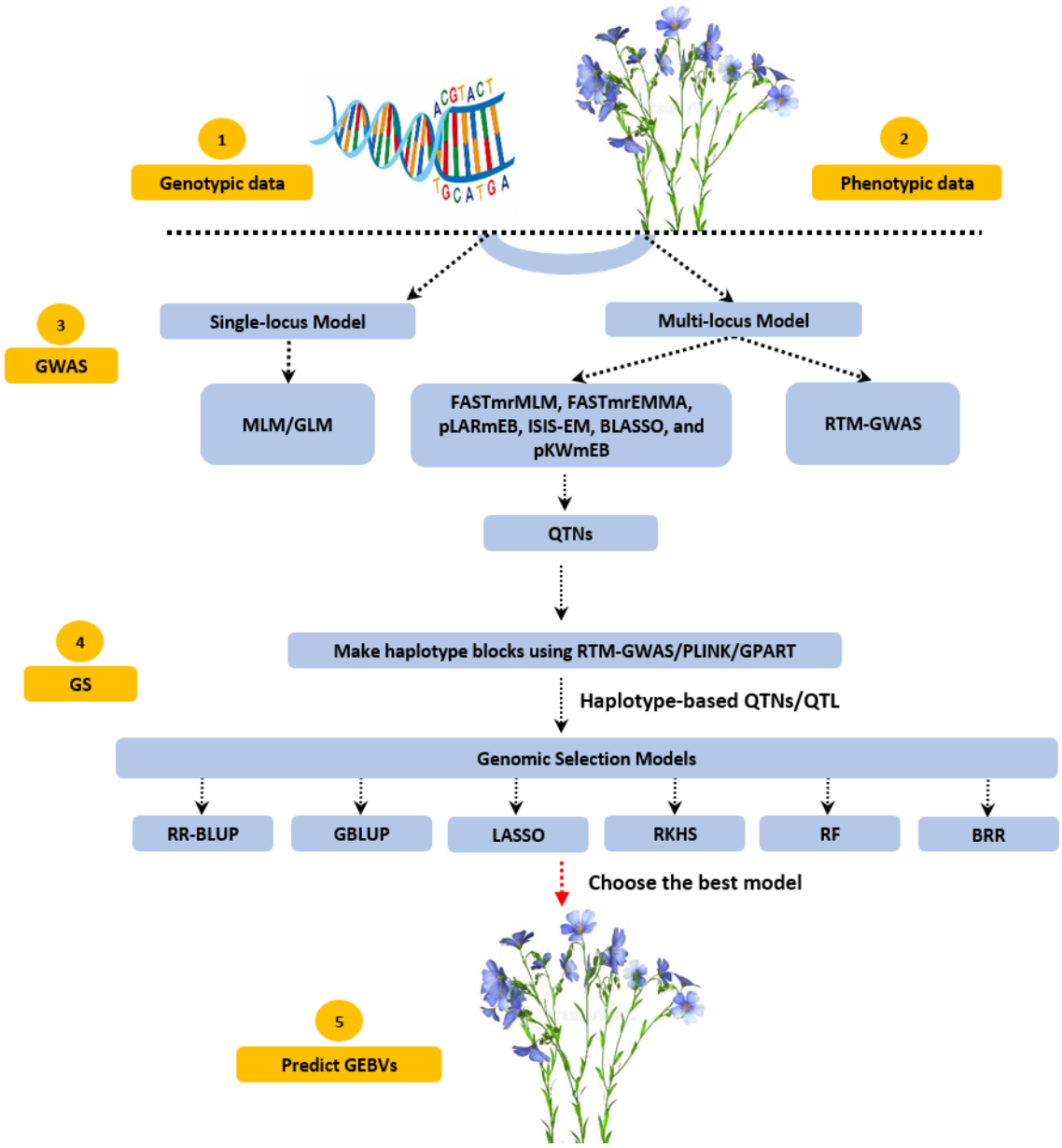


Fig. 1.1 A comprehensive approach for genomic selection (GS) based on genome-wide association studies (GWAS)-derived quantitative trait locus (QTL) as markers for flax breeding.

GWAS in flax have already identified a large number of QTNs associated with root characteristics and drought tolerance traits (Sertse et al. 2019; Soto-Cerda et al. 2019; Soto-Cerda et al. 2020; Sertse et al. 2021). However, no GS studies have been reported for these traits to date. With the current increased frequency and intensity of occurrence of these stresses, the popularity of these methods is predicted to grow in response to the need to create crops that are better able to

withstand these stresses. The GS approach is anticipated to facilitate future abiotic stress studies in flax specifically for drought, Cd toxicity, and other important traits.

1.4.3.2 Strategies for Cross Prediction and Parent Selection

The main objective of a breeding program is to develop superior cultivars for specific target traits that can perform well under a wide range of environmental conditions. Traditional breeding is time consuming and largely relies on the breeder's experience and plant phenotypes, which variably translate into inaccurate predictions and low efficiency. Cross-breeding using single crosses, double crosses or backcrosses, followed by offspring selection such as pedigree selection and single-seed descent, is the most used method in plant breeding. Parent evaluation and selection are the first step of cross-breeding. In order to extend the narrow genetic base of Canadian flax cultivars (You et al. 2016a), a core collection of 407 accessions has been assembled from the world flax collections (Diederichsen et al. 2013; Soto-Cerda et al. 2013). These accessions have been phenotypically evaluated in multiple environments for several years and at multiple locations (You et al. 2017). Although breeders can select genotypes with superior phenotypes as parents for crossing, the accuracy and efficiency of parent selection are impeded by unknown genetic structures and allelic make-up of the potential parents. Making crosses using all accessions of the core collection is impractical due to the extensive resources that would be required. The limited number of crosses possible narrows the probability to recover the best recombinants. Genomic cross prediction can enrich the cross-breeding strategy through effective cross prediction, best parent selection and the generation of hundreds to thousands of virtual crosses in a short time. Genomic cross prediction is applied to breeding whereby genomic data of parents are used to predict the traits of interest and then virtual crosses and their progenies are generated using computer tools. The use of computer simulation tools can assist breeding schemes and allow the

exploration of a wide range of hypotheses in a limited time for complex traits at a low cost. The recent development of extensive genomic datasets and through the use of computer simulations, high accuracy in virtual cross-breeding in a limited time have been made possible. These advances promise to greatly facilitate selection of suitable breeding schemes, allowing enhanced prediction in the selection for complex traits, and increasing efficiency in cross-breeding. As genome sequencing grows to be more cost-effective, genomic cross prediction becomes a useful approach to save time and solve the problems that cannot easily be solved via conventional breeding approaches.

In cross-breeding, parents are crossed to produce segregating populations from which a high-quality inbred offspring is carefully chosen. Many crosses are made and the value of a specific cross depends on the average performance of its progeny (population mean) and the success of its best progeny. A critical issue in a breeding program is the need to evaluate these crosses cost-effectively and efficiently. Thus, the important elements to consider in using genomic cross prediction include i) phenotyping and genotyping of potential parents mostly as training population for the development of GS models and simulation of a progeny population for each cross based on genomic data of parents, and ii) the application of the developed GS models to predict GEBVs of all progenies of a cross. The cross performance is mainly evaluated according to usefulness criteria that are a function of progeny population mean and genetic variance (Zhong and Jannink 2007) (**Fig. 1.2**). This strategy will ensure the best progeny crosses, the most effective breeding methods, and high breeding accuracy and efficiency in flax breeding. Such strategy has been used to predict crosses in crops such as wheat (Lado et al. 2017; Yao et al. 2018), and maize (Bernardo 2014) as well as in theoretical studies (Zhong and Jannink 2007).

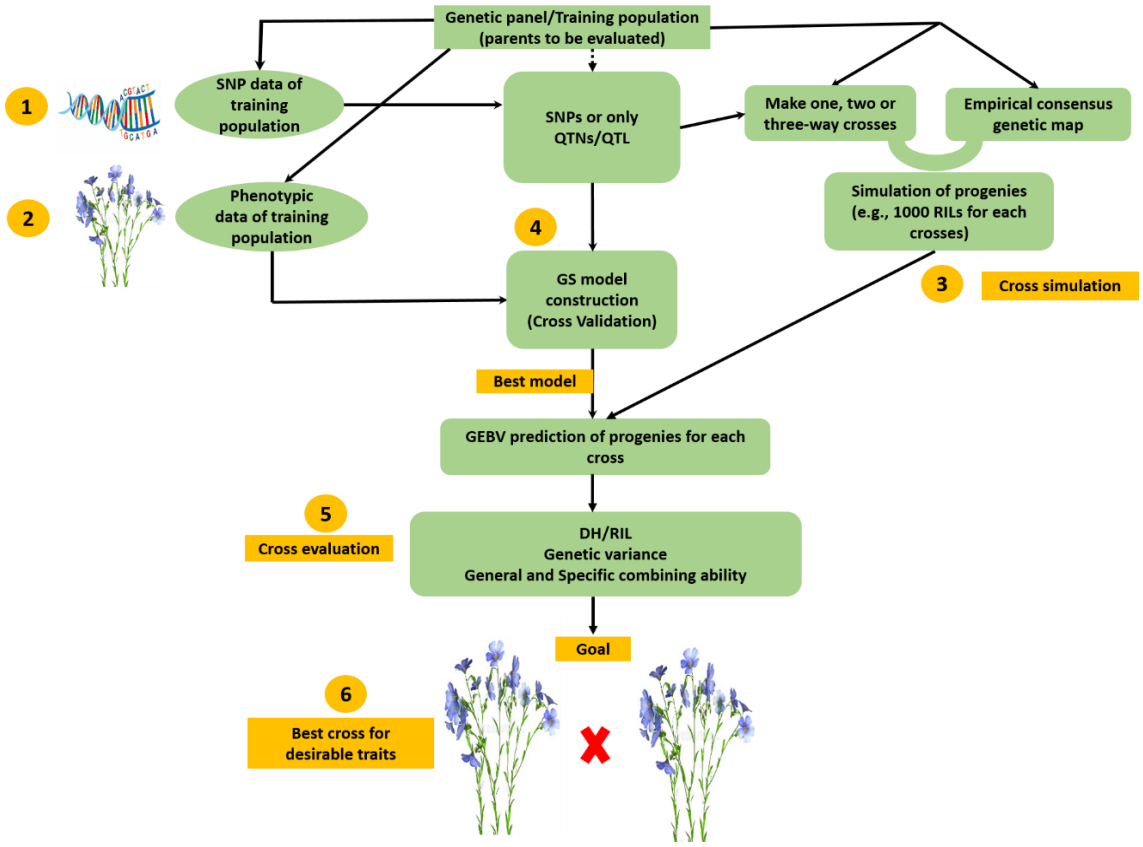


Fig. 1.2 A strategy of cross prediction and parent selection using genomic selection (GS) and genomic cross prediction.

1.4.3.2.1 Software Tools and Their Applications in Flax Breeding Simulation

Numerous tools are available to simulate breeding schemes. Some of the recent tools for breeding scheme simulation are listed in **Table 1.4**. The simulation tool pSBVB for complex phenotypic traits is used to compute the genomic matrix relationships among polyploids (Zingaretti et al. 2019). PedigreeSim is used for both diploid and tetraploid species and predicts pedigrees and cross progenies (Voorrips and Maliepaard 2012). Similarly, the ADAM-plant and QuLinePlus models can predict breeding outcomes of cross- and self-pollinated crops (Hoyos-Villegas et al. 2019; Liu et al. 2019b). The ADAM-plant simulation, effective for predicting self-pollinated crop plants (Liu et al. 2019b), also has the ability to trace significant genetic changes in a set of populations under various scenarios. Simulation of progeny populations requires

recombination rates between all adjacent markers, and MareyMap is a useful tool to estimate them through a consensus genetic map as a training dataset (Siberchicot et al. 2017).

Table 1.4 Main features of different phenotypic-genotypic simulation models.

Model	Feature	Reference
pSBVB	Simulates any number of complex phenotypes	Zingaretti et al. (2019)
PedigreeSim	Simulates both diploid and tetraploid species and predict pedigrees and cross populations	Voorrips and Maliepaard (2012)
PhenotypeSimulator	Predicts multiple traits with multiple underlying genetic loci	Meyer and Birney (2018)
ADAM-plant	A tool that models breeding schemes for both self- and cross-pollinated crop plants	Liu et al. (2019b)
QuLinePlus	A simulation model for cross-pollinated crops	Hoyos-Villegas et al. (2019)
MareyMap	A tool to calculate recombination rates of all markers on a physical map based on a training genetic map	Siberchicot et al. (2017)
AlphaSim	Enables several aspects of breeding programs such as haplotype sequences and pedigrees to be simulated, and generates new data simulation for selection	Faux et al. (2016)
Phenosim	Simulates phenotypes for testing in GWAS	Günther et al. (2011)
SLiM	A simulation framework that enables modeling of a wide variety of complex evolutionary scenarios	Haller and Messer (2019)
G2P	A simulation tool for both genotypes and phenotypes	Tang and Liu (2019)
SimPed	A simulation approach for the development of haplotype and genotype data for pedigree populations	Leal et al. (2005)

1.4.3.2.2 Cross Prediction and Parent Evaluation in Flax Breeding

We recently performed cross prediction and parent evaluation through GS and genomic cross prediction for seed yield (YLD), days to maturity (DTM), oil content (OIL), linolenic acid content (LIN), and powdery mildew resistance (PM) in flax (unpublished data). A total of 290 flax linseed accessions were evaluated, generating 4096 (290 x 289 /2) possible virtual single crosses with a partial diallel cross scheme. Ten GS models were evaluated, in which the RR-BLUP model displayed the highest prediction ability. General combining ability (GCA) and specific combining ability (SCA) were computed for the simulated doubled haploid (DH) and recombinant inbred line (RIL) populations and a significant linear relationship between GCA and SCA (R^2 = 0.93-0.98) was obtained for all five traits (**Fig. 1.3**), indicating that a parent of higher GCA is more likely to generate high performance crosses. Also, a high correlation (R^2 = 0.93-0.98) between the mid-parent GEBVs and SCA was observed (**Fig. 1.4**), suggesting that high mid-parent GEBVs of

crosses are good indicators of their potential in crosses. The findings of this research provide a framework for future cross-breeding studies.

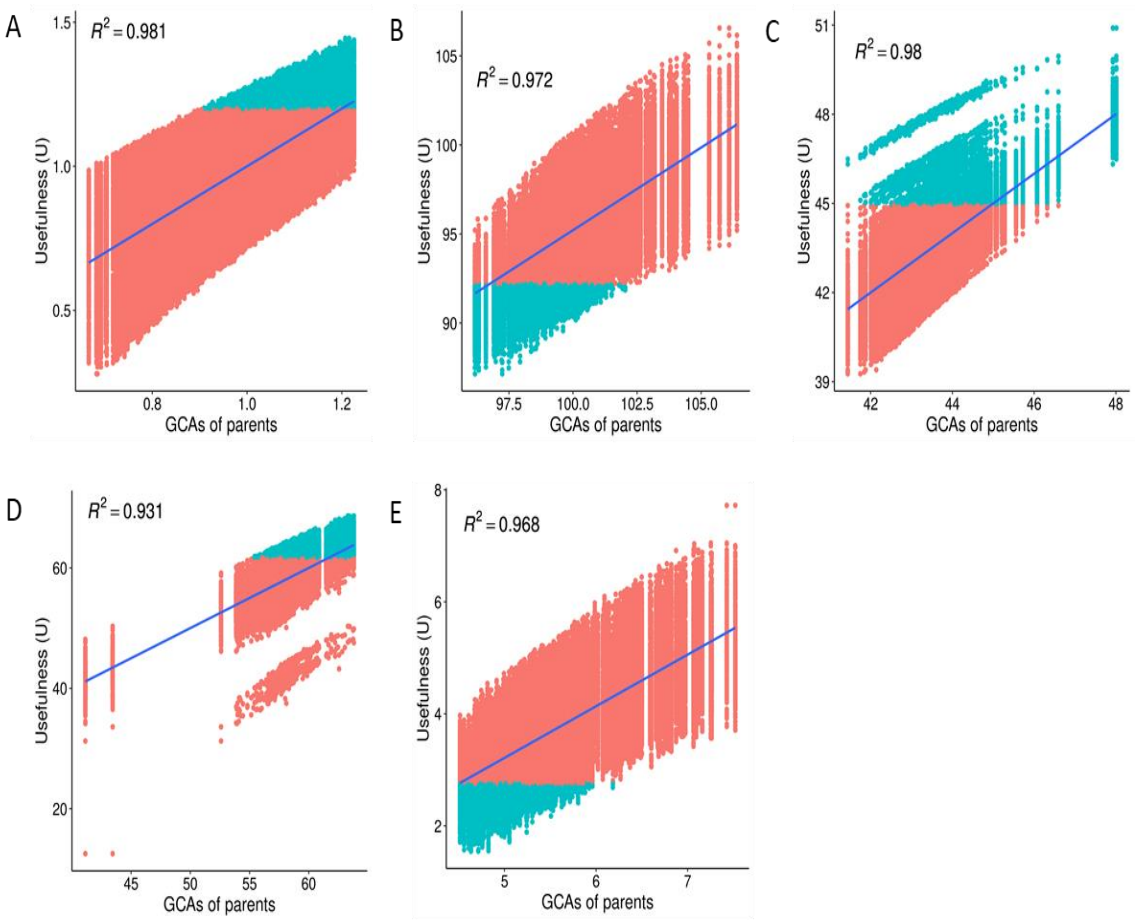


Fig. 1.3 Relationship between general combining abilities (GCAs) of parents and usefulness (U) of crosses for five traits. The blue dots represent the top 10% of crosses based on U of crosses at a 10% selection rate. A. Seed yield (t/ha); B. Days to maturity; C. Oil content (%); D. Linolenic acid content (%); E. Powdery mildew (PM) resistance (0-10 scale) in flax. Source: unpublished data

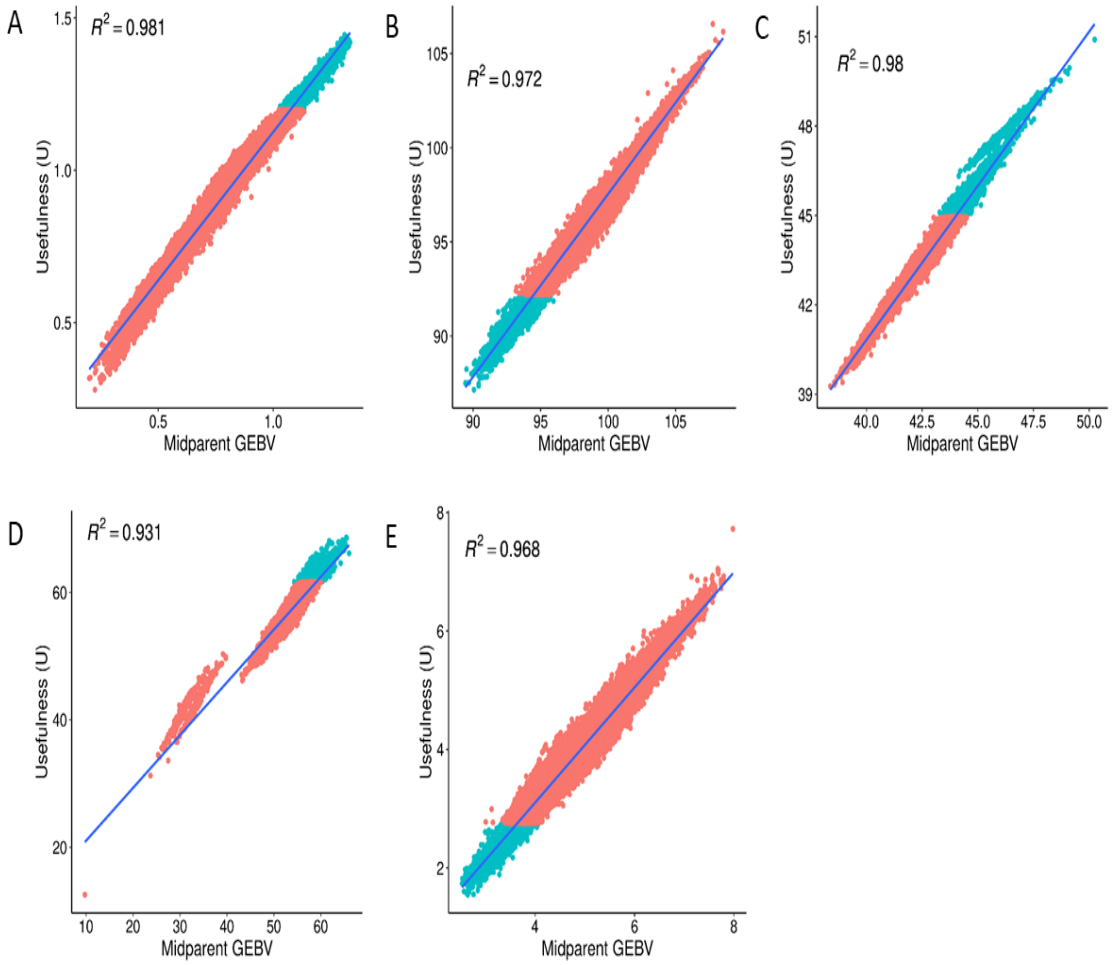


Fig. 1.4 Relationship between the mid-parent genomic estimated breeding values (GEBVs) and usefulness (U) of single crosses for five traits. The blue dots represent the top 10% of crosses based on U of crosses at a 10% selection rate. A. Seed yield (t/ha); B. Days to maturity; C. Oil content (%); D. Linolenic acid content (%); E. Powdery mildew (PM) resistance (0-10 scale) in flax. Source: unpublished data

1.5 Future Perspectives

The recent significant advancements in GS and genomic cross prediction offer new opportunities to further improve abiotic trait-marker associations, cross-breeding accuracy, and breeding selection efficiency. Genomic studies investigating the associations between root traits and drought tolerance have resulted in the discovery of 521 QTLs in flax (Sertse et al. 2019; Soto-

Cerda et al. 2019; Soto-Cerda et al. 2020; Sertse et al. 2021). The data generated in these studies can be capitalized upon in GS and genomic cross prediction to gain insights into the genetic complexities of these traits. For flax genetic improvement, it is vital to expand the portfolio of abiotic stress-related traits to include Cd stress for example. To date, cross-breeding and parent selection remain a major challenge in plant breeding. The combined GS and genomic cross prediction strategy offers opportunities to increase the accuracy and efficiency of cross-breeding in flax by predicting the best crosses. However, the use of genomic cross prediction in flax remains in its infancy and much awaits to be done to turn theoretical demonstrations into practical applications. Further investigations on the identification of significant QTLs, the use of HTP, and accuracy improvement of GS and genomic cross prediction for both biotic and abiotic traits in flax are warranted.

Chapter 2

Genomics Assisted Breeding Strategy in Flax

Nadeem Khan[1,2], Hamna Shazadee[1,2], Frank M. You[1,*], and Sylvie Cloutier[1,2,*]

[1]Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, 960 Carling Avenue, Ottawa, Ontario K1A 0C6, Canada.

[2]Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa, ON K1N 6N5, Canada.

*Correspondance: frank.you@agr.gc.ca; sylviej.cloutier@agr.gc.ca

This chapter has been accepted for publication as a book chapter: Khan N, Shazadee H, You FM, and Cloutier S. Genomics assisted breeding strategy in flax. In: You FM and Fofana B (eds) The flax genome, Springer, chapter 12 (In press).

Authors' contribution

Conceptualization, methodology, formal analysis, investigation, writing—original draft preparation, N.K.; writing—review and editing, H.S.; supervision, project administration, funding acquisition, FMY and S.C. All authors read and approved the final manuscript.

Abstract

Genomic selection (GS) or genomic prediction (GP) is a type of marker-assisted selection that relies on genome-wide markers to predict genomic estimated breeding values (GEBVs) of phenotypes. GS is quickly becoming a conventional approach in both plant and animal breeding to increase selection accuracy, reduce breeding cost, and shorten breeding cycles. The concept of GS models was first developed using genome-wide random markers, with marker density being a key element in estimating the predictive ability in breeding populations. It is currently straightforward to generate high-density marker datasets thanks to the remarkable advances in genotyping technologies. Recent studies showed that high-density genome-wide random markers do not necessarily generate high genomic predictive ability in GS because the vast majority of markers are unrelated to the traits of interest; thus generating background noises and lowering the predictive ability. Alternatively, the use of quantitative trait loci (QTLs), identified through genome-wide association study (GWAS) methods, in GS models can significantly improve genomic predictive ability and reduce the genotyping cost of the test populations. Here, we present recent findings, discuss a few case studies, a QTL-based GS strategy and a genomic cross predictions for flax breeding improvement.

Keywords: Genomic selection (GS) ● Genomic prediction (GP) ● Genome-wide association study (GWAS) ● Molecular breeding ● Molecular marker ● Quantitative trait loci (QTLs) ● Quantitative trait nucleotides (QTNs) ● RAD capture (Rapture) ● Genotyping by target sequencing (GBTS)

Chapter 2: Genomics Assisted Breeding Strategy in Flax

2.1 Introduction

Genomic selection (GS) or genomic prediction (GP) is a form of marker-assisted selection that takes genome-wide markers to predict genomic estimated breeding values (GEBVs) of phenotypes (Meuwissen et al. 2001). A major purpose of plant breeding programs is to maximize genetic gain for desirable traits. The genetic gain can be calculated using "Breeder's equation" in classical quantitative genetics (Lush 1937). In the past, only phenotypic data was normally utilized to calculate genetic gain. GS is becoming a conventional approach in plant and animal breeding. The use of GS over traditional breeding programs offers three major benefits namely increased selection accuracy, shorten breeding cycles, and reducing breeding cost. Cost-effective genotyping on a broad scale has enabled the adoption of low-cost genome sequencing. As a result, more complicated and innovative approaches such as genomic cross prediction using GS methodologies to achieve various breeding goals can now be investigated. With the development of high-throughput genotyping technology, high-density genome-wide molecular markers can be readily obtained and sufficient individuals of breeding populations can be genotyped at a much lower cost. Some of the widely used genotyping methods include genotyping by sequencing (GBS), whole-genome resequencing and genotyping, array-based genotyping (Ali et al. 2015), and target sequences-based genotyping (He et al. 2014). GS also offers three major applications for breeding programs: (1) offspring selection, (2) evaluation of germplasm and parent selection, (3) hybrid prediction based on general combining ability (GCA) and specific combining ability (SCA) (Zhao et al. 2015; You et al. 2022b). One major challenge in hybrid breeding is to efficiently select superior hybrids, however, GS has the ability to resolve such complexities by predicting the best parental material.

Most GS models are constructed based on genome-wide random markers. However, the costs associated with generating large-scale datasets limit the use of GS in a practical breeding program. Recently, several studies (He et al. 2019a; Lan et al. 2020) have discussed the use of QTL-based strategy as markers for improving the predictive ability as well as reducing the cost. Recent studies (He et al. 2019a; He et al. 2019b) revealed that combined use of single- and multi-locus genome-wide association study (GWAS) methods can identify both large and minor effect QTL that can be used to build GS models, significantly improving genomic predictive ability. The use of a QTL-based strategy is highly desirable for the successful implementation of GS in breeding programs to reduce genotyping costs.

Similarly, the advent of GP and the use of computer simulation or genomic cross prediction has opened the door for developing virtual crosses and parent selection for two major reasons: i) utilization of both dense markers or QTL-based strategy for estimating GEBVs, and ii) a large number of virtual crosses such as single, double, three-way or backcross, can be generated cost-effectively in a short time. Thus, a major goal of genomic cross prediction is to provide best parental material and develop virtual crosses and both of them are particularly important for crossbreeding. The ability of plant breeders to test the effects of several possibilities determines which breeding program is selected. In terms of time and effort, comparing different breeding strategies in large-scale field studies can be time-consuming. Genomic cross prediction studies, rather than field tests, can be an effective way to simulate and forecast the efficacy of various breeding strategies. In addition, genomic cross prediction studies can help us understand the impact of various factors on genetic gain and other breeding program variables of interest (e.g. selection accuracy and genetic variation), allowing us to create and select superior breeding programs (Liu et al. 2019b). The major objective of this chapter is to summarize recent findings as a case study

on the use of QTL-based strategy and also discuss genomic cross prediction for their potential use in flax breeding programs.

2.2 Factors Affecting Genomic Prediction (GP) Efficiency

Several key factors influence the predictive ability of GS such as the training population (TRP), the size of the test populations (TPs) and the relationship between the two, the GP models, the markers, etc. Here, we will discuss GP models, TRP, TP and marker types for improving GS accuracy, and we will describe several case studies in flax. Prediction efficiency of GS depends on the predictive ability and the unit cost. There are several methods to evaluate GS predictive ability (de los Campos et al. 2013a). The most popular one is cross-validation (CV), which is important in GS for empirical assessments of predictive ability before the GS models are applied to actual selection. CV is performed when both phenotypic and genotypic data are available for a given germplasm collection. Here, the collection is randomly divided into subsets: TRP and TP sets. The marker effects are then calculated based on the data from the TRP set, followed by a prediction based on the marker effects of the genotypic values of plants in the TP set. A measure of the predictive ability is given by the Pearson correlation (r) or the square root of the average squared error (Daetwyler et al. 2013) between the estimated and observed values of the TP set. The five-fold cross-validation is mostly used. That is, the collection is randomly split into five subsets with equal number of individuals. One of them is in turn used as a TP and the remaining as a TRP to construct GS models and calculate predictive ability. This approach has been used in many crops, including flax (Lan et al. 2020), maize (Albrecht et al. 2011), wheat (Heffner et al. 2011), and barley (Heslot et al. 2012). However, the predictive ability estimated by CV is specific to the population and may not always represent the predictive ability required in a practical

breeding program. This is due to the fact that population structure can have a significant impact on predictive ability estimations. Several studies such as those from Hickey et al. (2014) and Lehermeier et al. (2014) have used stochastic simulation data and real datasets from maize breeding programs, and have shown that predictive ability within and among families can differ significantly in structured populations. To boost the potential of the breeding program, the population structure should be taken into account. This may determine whether or not GS can be applied in a given breeding program. Continuous use of closely-related (TRPs to TPs) populations to improve prediction will likely narrow or reduce genetic variation that could contribute to the future selection response, and consequently slow down the genetic gains required in long-term GS application (Jannink et al. 2010; Moeinizade et al. 2019).

2.2.1 Genomic Prediction (GP) Models

Many GS methodologies, which differ with respect to assumptions about marker effects, have been proposed to obtain GEBVs (Meuwissen et al. 2001). GS models, for example, are generally divided into two groups: parametric and non-parametric (Desta and Ortiz 2014). Parametric models, such as ridge regression best linear unbiased prediction (RR-BLUP) and Bayesian approaches, are the most popular (Endelman 2011; Habier et al. 2011). Only additive effects are accounted for in parametric or linear models, which presume a linear relationship between the phenotype and genotype (Pérez-Rodríguez et al. 2012). The accuracy of GS models varies according to the assumptions and treatments of marker effects. For example, both Bayesian LASSO (BL) and ridge regression models have been found to outperform support vector machines in predicting GEBVs for host-plant resistance to wheat rusts (Ornella et al. 2012). Another study tested six statistical models for GS, namely Bayesian ridge regression, BL, reproducing kernel Hilbert spaces (RKHS), and Bayes A, B and C (Rolling et al. 2020), and found only slight

variations in their predictive potentials. Different GP models, such as RR-BLUP, genomic best linear unbiased prediction (GBLUP), Bayesian regression, partial least squares regression and machine learning, can produce slight variations in predictive ability (Jannink et al. 2010). Although the accuracy of the different methods are often similar (He et al. 2019a), the prediction differences may be due to difference in the genetic architecture of the traits (Spindel et al. 2015). As a result, optimization of GS models varies depending on the trait and the crop of interest, necessitating the testing of numerous models to determine the model with the best fit, i.e., the one with the highest predictive ability. Theoretically, models like RR-BLUP and GBLUP assume that all markers influence trait variation, whereas BayesA and BayesB assume a unique variance for each marker. As a result these models are grouped into two, the first group of models should be more appropriate for complex quantitative traits with a polygenic architecture, whereas the second group of models would likely be better-suited for traits controlled by a small number of genes or QTLs with high impacts (You et al. 2022b). For traits controlled by a few genes with large effects, BayesB has been found to perform better than GBLUP or RR-BLUP in several studies (Daetwyler et al. 2010; Jannink et al. 2010; Thavamanikumar et al. 2015).

A case study for seed yield (YLD), days to flowering (DTF) and *Fusarium* wilt (FW) disease resistance was performed to assess the effect of different models on the predictive ability of GS (**Fig. 2.1**). To evaluate the predictive ability of three datasets comprising 308, 422 and 571 markers, respectively, ten GS models were tested in the R package BGLR (Pérez and de los Campos 2014) or in G2P (Tang and Liu 2019). These encompass both parametric or non-parametric regressions (Desta and Ortiz 2014), including RR-BLUP (Endelman 2011), GBLUP (Clark and van der Werf 2013), BRR (Heffner et al. 2009), BL (Li and Sillanpää 2012), Bayes A, Bayes B, Bayes C (Habier et al. 2011), random forest (RF) (González-Recio and Forni 2011), SVR (Moser et al. 2009) and

RKHS (Gianola et al. 2006). The parametric models RR-BLUP, GBLUP and BL consistently outperformed the RF and SVR models for all traits and marker sets. The RR-BLUP model has been widely used to predict GEBV because it often outperforms the other models and has high computation efficiency. However, it is quite possible that other models could perform superior with different marker sets or experimental designs. For example, in alfalfa, the machine learning methods support-vector machine and RF outperformed RR-BLUP (Medina et al. 2020). Therefore, to implement GS successfully, it is suggested that several models be evaluated for specific populations and traits.

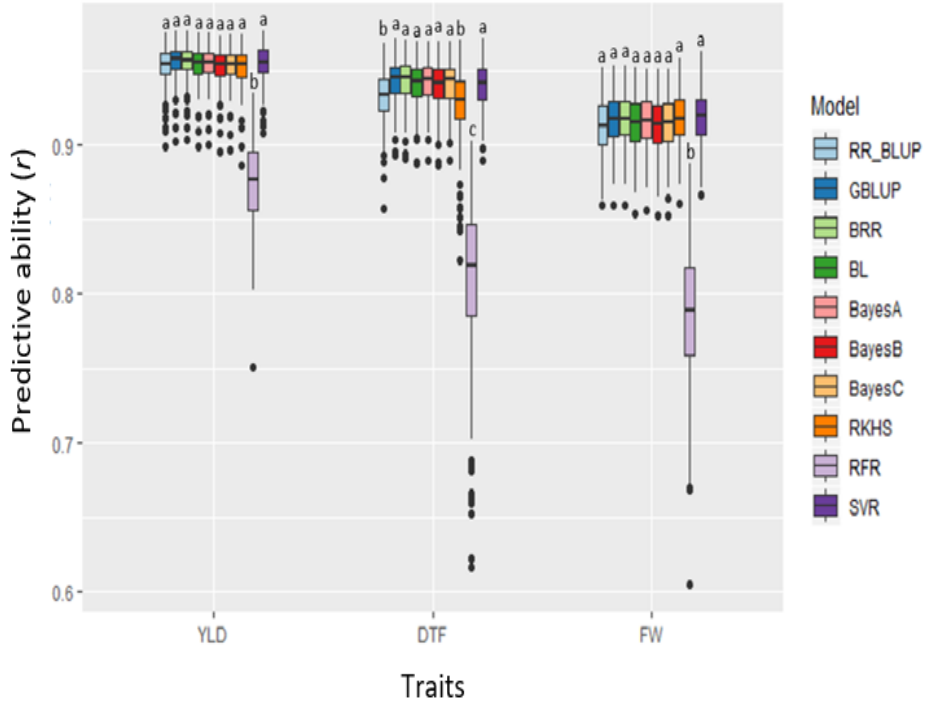


Fig. 2.1 Genomic predictive ability (*r*) obtained using ten genomic selection (GS) statistical models for seed yield (YLD), days to flowering (DTF) and *Fusarium* wilt (FW) disease resistance in the flax core collection consisting of 408 diverse genotypes. A five-fold cross-validation (CV) scheme was used. A total of 422, 571 and 308 quantitative trait loci (QTLs) for YLD, DTF and FW, respectively, were used for the construction of GS models. Statistical significance of predictive ability among these models for each trait is indicated by a letter above each box plot. The letters indicate significant differences at 5% probability level obtained using the Tukey multiple comparison method.

2.2.2 Training Populations (TRPs) and Relatedness to the Test Populations (TPs)

The development of both TRP and TP datasets is also a vital step toward the implementation of GS. In GS, a TRP is used to train a prediction model, which consists of germplasm that have been phenotyped for target traits and genotyped, generally with genome-wide markers. After being trained, this model is used to calculate the GEBVs of the individuals of the TP, strictly on the basis of its genotypic information. To achieve high GS accuracy, the TRP and TP should be closely related. For example, the predictive ability in maize was remarkably improved when the two populations had a close relationship (Schulz-Streeck et al. 2012; Zhang et al. 2017). Similarly, GS for grain yield, performed by Windhausen et al. (2012) using diverse panels of maize lines achieved lower predictive ability between groups than within groups, and more accurate predictions were obtained with closely related populations (Heslot et al. 2015). In summary, TRP optimization has attracted the attention of the breeding community for a variety of reasons. First, GS predictions are highly dependent on the markers and optimization of the TRP is key to improve GS efficiency and efficacy. Second, phenotyping is expensive but accurate phenotyping is paramount to breeding and the success of GS. To better allocate resources within plant breeding programs and improve GS' predictive ability, it is preferable to increase the size of the TRP. In breeding programs, phenotyping has always been a bottleneck because selection progress is directly proportional to the number of genotypes that can be phenotypically analyzed. Genotyping prices have decreased considerably in the genomic era, whereas phenotyping costs have remained roughly constant. In this regard, one of the key goals of TRP optimization is to improve the "selected phenotyping" process, which aims to reduce phenotyping costs while retaining high predictive ability models.

There are a number of models (**Table 2.1**) that appropriately address the statistical issues in GS, the most prominent being the RR-BLUP model (Endelman 2011). Several factors, however, may have an impact on predictive ability. They occur at various levels and are affected by a combination of genetic, environmental, and statistical factors. In GS, some of the major issues include: i) diverse and multiple genotype-by-environment (GE) interaction, ii) the size of the TRP and its relationship with the TP, and iii) genotypic complexity and heritability of traits. The most significant challenge in GS is the influence of GE and its interaction, which are connected to the density of markers and accuracy of phenotypic measurements. These complications necessitate the use of parametric and non-parametric statistical models, particularly Bayesian estimations and more recently deep machine learning approaches that can handle massive datasets (Crossa et al. 2017). The large amount of data that needs to be handled in GS protocols has created computational and statistical hurdles, such as model fitting and parameter optimization. As a result, efficient implementation of GS, requires the creation of complete and simple computer programs to estimate the GEBVs of the individuals to be selected under these complex scenarios.

Table 2.1 Summary of major R packages for genomic selection (GS).

R package	Major features	References
lme4GS	Genomic-based prediction of GS that can fit mixed models with multiple variance–covariance matrices	Caamal-Pat et al. (2021)
Sommer	GS for single and multi-environments and heritability estimation	Covarrubias-Pazaran (2016)
learnMET	Deep learning model for GP using multi-environment trial data	Westhues et al. (2021)
rrBLUP	Fast maximum-likelihood algorithm for mixed models	Endelman (2011)
IBCF.MTME	GP for multi-traits and multi-environment data	Montesinos-López et al. (2018)
TrainSel	Selection of TRP	Akdemir et al. (2021)
GAPIT	Integrated tool for both GWAS and GS	Lipka et al. (2012)
GSelection	GP with an integrated model	Guha Majumdar et al. (2020)

2.2.3 Markers

Given a practical TRP, the choice of markers is a critical factor for improving predictive ability (Meuwissen et al. 2001). In general, increasing the marker density can help to improve predictive ability until it reaches a point when it plateaus and where additional markers no longer improve the predictive ability of the models (Xu et al. 2018; Juliana et al. 2019). The required marker density will differ depending on the plant species, as well as the type and size of the populations. Generally, cross-pollinated species require a higher density of markers than self-pollinated species (Liu et al. 2018; Juliana et al. 2019). For instance, the optimization of genomic predictive ability was explored by using a single QTL effect and a combination of all traits in flax (He et al. 2019a; Lan et al. 2020). Here single trait effect had a better genomic predictive ability compared to the combined QTL effect and random SNPs. The accuracy of prediction explicitly tests the efficiency of a marker set in GS. The genomic predictive ability is more likely to increase for a particular trait when using a trait-specific QTL set identified from GWAS than a random set of markers, even if the latter is of a higher density (Hoffstetter et al. 2016; He et al. 2019a). Ultimately, accurate genotyping and phenotyping remain critical to the success of any large-scale analysis of genetic association and GS, as the number of false-positive and false-negative associations can be inflated by systematic biases caused by even small sources of error. The strategy outlined below as a case study is based on the sequential use of GWAS and GS, and is efficient and powerful even for complex traits such as FW, YLD and pasmo resistance (He et al. 2019a). QTL effect from a single or multiple related traits cannot only significantly improve predictive ability, but can also decrease the number of markers needed in GS models, thereby potentially reducing the genotyping costs through the use of a QTL-defined genotyping platform for example. This input cost reduction combined with the simplification of the computation associated with using few markers can tip the

scale towards adoption of GS versus the *status quo* of phenotypic selection where cost reduction strategy such as phenomics are still at the developmental stages for many important traits. Along with the constant reduction in sequencing costs *per se*, these additional cost-saving strategies make GS a viable option for many breeding programs, and we anticipate that this will translate into its increased adoption in the next few years.

2.3 Improving Predictive Ability by QTL Markers

The implementation of GP as a breeding tool is limited by the cost of sequencing and the challenge of laborious phenotyping of complex traits such as those related to drought, root traits, YLD and resistance to diseases. The application of GP depends on the production of high-throughput genotyping data of large breeding populations. Here, we discuss the impact of a sequential approach that relies first on a suite of GWAS models to identify QTLs, which then become the input dataset for GS implementation. The combination of GWAS and GS is highly effective, and is more likely to increase predictive ability when using a QTL-based strategy set identified from GWAS.

2.3.1 QTL Identification by Single- and Multi-locus GWAS

GWAS is a hypothetical study involving an association between genotype and phenotype that estimates the contribution of genetic variants across the genomes of many individuals to the measured traits. GWAS is used to identify QTLs or quantitative trait nucleotides (QTNs) that are statistically related to a target trait, such as YLD, DTF or FW. GWAS analysis requires three sets of input data: genome-wide markers, population structure, and trait measurements. The use of GWAS is based on two types of models: SNP- or haplotype-based. SNP-based statistical models include single-locus (SLSMs) and multi-locus statistical models (MLSMs). GLM and MLM

models are examples of SLSMs and both tend to detect major QTLs. MLSMs are a more recent addition to GWAS statistical methods, e.g., the multi-locus random-SNP-effect Mixed Linear Models (Wang et al. 2016b). The goal of MLSMs is to identify both major and minor effect QTLs or QTNs. In GWAS, the use of models such as SLSMs and MLSMs is somewhat complementary and their combined use enables the identification of both small and large effect QTNs for complex and low heritability traits. The primary distinction between SLSMs and MLSMs is that SLSMs evaluate the relationship between each marker and the trait in a sequential manner, whereas all MLMs use two-step algorithms. SLSM GWAS approach is used to scan the complete genome in the first step, and putative QTLs are found using a less rigorous critical value, such as $P < 0.005$ or $P < 1/n$, where n is the number of markers. In the second stage, an MLSM examines all putative QTLs to determine their significance using a logarithm of the odds (LOD) statistics. As a result, more possible QTLs/QTNs related to target traits can be discovered. In the case of haplotype-based, the RTM-GWAS model firstly groups SNPs into linkage disequilibrium blocks and then uses a two-stage analysis for QTL identification, where markers are preselected by a SLSM followed by a multi-locus multi-allele model stepwise regression (He et al. 2017). These models do not require Bonferroni correction, have a higher power for QTN detection and were found superior to SLSMs in identifying small-effect loci for complex associations (Cui et al. 2018; Zhang et al. 2018b).

2.3.2 Genomic Heritability (h^2)

The advantages of GS are more pronounced for traits that are complex, time-consuming and/or costly to measure, particularly when it is important to analyze multiple environments. In practical GS applications, the marker effects are calculated based on the TRP from which the lines are genotyped and phenotyped, and subsequently used to predict GEBVs. Four random SNP datasets

of different sizes and one QTL dataset were used to evaluate the effects of the marker density and marker type on the predictive ability of three traits (**Table 2.2**). The predictive ability of the QTL-based dataset was 0.96 for YLD, 0.94 for DTF and 0.92 for FW. The random SNP datasets had significantly lower predictive abilities with 0.78 for YLD using 5,000 SNPs and 0.60 for FW using 258,870 SNPs for examples (**Table 2.2**). The QTL set from the traits *per se* achieved higher genomic predictive ability than the random SNP datasets. The high accuracy obtained with the QTL is hypothesized to be a consequence of the reduction in background noise through the removal of unrelated markers (He et al. 2019a).

In breeding, high heritability means that the trait is less affected by environment and thus the phenotypic selection to the trait is more effective. High heritability of a trait implies high accuracy in phenotypic selection and also likely translates into high predictive ability through GS. Genomic heritability (h^2) measures the extent with which the phenotypic performance are explained by currently available markers. High h^2 implies the trait performance is more accurately predicted by the markers. Among the three traits considered in the study described above, DTF had the highest h^2 of 0.83 and the highest predictive ability (r) of 0.94. YLD had the lowest h^2 value of 0.33 but a moderate predictive ability of 0.78 when using a genome-wide random dataset of 258,870 SNPs. For all traits, the genomic heritability and predictive ability were highest when calculated from the QTL datasets. Even though the QTLs and the highest-density SNP datasets had the same h^2 (0.83) for DTF, the predictive ability of the QTL dataset far exceeded that of the random SNP datasets with 0.04 versus 0.62-0.63, respectively (**Table 2.2)**. In summary, a steep increase in genomic predictive ability was observed for all traits when using a trait-specific QTL input dataset as was previously reported for pasmo resistant by (He et al. 2019a). In conclusion, the combined use of GWAS and GS gives the breeding community a new way to estimate breeding values by using

QTLs/QTNs as markers instead of genome-wide random markers. This technique has been tested in flax for several traits (He et al. 2019a; Lan et al. 2020) and has become a viable tool for improving genomic predictive ability and reducing the associated cost.

Table 2.2 Genomic heritability ($h^2 \pm s$) and genomic predictive ability ($r \pm s$) for different types and number of markers for three traits in the flax core collection.

Trait	Marker set	Genomic heritability ($h^2 \pm s$) [1]	Predictive ability ($r \pm s$) [2]
Seed yield	422 QTLs	0.77 ± 0.04	0.96 ± 0.01a
	5,000 SNPs	0.43 ± 0.09	0.78 ± 0.05b
	10,000 SNPs	0.41 ± 0.09	0.78 ± 0.05b
	50,000 SNPs	0.35 ± 0.10	0.78 ± 0.05b
	258,870 SNPs	0.33 ± 0.10	0.78 ± 0.05b
Days to flowering	571 QTLs	0.83 ± 0.04	0.94 ± 0.01a
	5,000 SNPs	0.75 ± 0.07	0.59 ± 0.12c
	10,000 SNPs	0.79 ± 0.07	0.60 ± 0.11c
	50,000 SNPs	0.83 ± 0.07	0.63 ± 0.09b
	258,870 SNPs	0.83 ± 0.07	0.62 ± 0.10b
Fusarium wilt resistance	308 QTLs	0.70 ± 0.05	0.92 ± 0.02a
	5,000 SNPs	0.38 ± 0.09	0.57 ± 0.08c
	10,000 SNPs	0.46 ± 0.10	0.60 ± 0.07b
	50,000 SNPs	0.43 ± 0.10	0.59 ± 0.07b
	258,870 SNPs	0.43 ± 0.10	0.60 ± 0.08b
Pasmo resistance (He et al. 2019a)	500 QTLs	0.72 ± 0.04	0.92 ± 0.02a
	52,347 SNPs	0.54 ± 0.07	0.67 ± 0.07b

[1]Genomic heritability is the proportion of additive genetic variance in the total phenotypic variation and estimated using the R package "sommer". [2]Predictive ability is defined as the Pearson correlation coefficient (*r)* of observed values with predicted values using a five-fold cross-validation scheme and the GBLUP model. The predictive ability reported is the average value of 250 iterations. The letters on the right of the predictive ability values represent significant difference among different marker sets at 5% probability level using Tukey multiple comparison statistical test.

2.3.3 A Case Study for Drought and Root Traits

A mini-core collection consisting of 101 and 106 flax accessions were assessed for drought-related and root traits using 258,708 SNP markers to test their predictive ability. To identify

putative QTNs associated with both drought-stress related and root traits, two types of GWAS models were used: the SLSMs included GLM (Price et al. 2006) and MLM (Yu et al. 2006) and the MLSMs included pLARmEB, FASTmrMLM, FASTmrEMMA, ISIS EM-BLASSO and mrMLM from the R packages mrMLM (https://cran.r-project.org/web/packages/mrMLM/index.html), and rMVP (https://github.com/xiaolei-lab/rMVP). In addition, another MLSM called FarmCPU was also used (Liu et al. 2016).

Six drought tolerance-related traits, including bundle weight (BDLWt), canopy temperature (CT), plant height (HT), seeds per boll (SPB), thousand seed weight (TSW) and grain yield (YLD) were evaluated under non-stress (irritated) and stress (non-irrigated) condition as previously described by Sertse et al. (2021). The six respective indices were used for GWAS with the high-density 250K SNP dataset. The six indices of drought susceptibility index, drought tolerance efficiency, stress susceptibility index, Schneider's stress severity index, tolerance against stress and stress tolerance index (STI) was also calculated for each of the above six traits as previously described (Sertse et al. 2021). The six traits measured under the drought stress condition (S) and their STI traits used for GS analyses were BDLWt_S and BDLWt_STI, CT_S and CT_STI, HT_S and HT_STI, SPB_S and SPB_STI, TSW_S and TSW_STI, and YLD_S and YLD_STI. The correlations have been previously described (Sertse et al. 2021) and based on high correlation values traits were selected for the remaining GS analyses, and STI was chosen as an important indicator for evaluating both yield and stress tolerance (Fernandez et al. 1993). A similar procedure of using correlation values was adopted for the root traits based on previously reported study (Sertse et al. 2019), and the following seven traits were carefully chosen for GS analyses: average root diameter (ARD), maximum number of roots (MaxR), network area (NWA), network length (NWL), network surface area (NWSA), network volume (NWV), and shoot dry weight (SDWt).

A total of 1,057 QTNs were identified for six drought-related traits and 528 QTNs were associated with the sixteen root traits. All QTNs were grouped into haplotype blocks (HBs) and, the QTNs within a HB with the highest R^2 for the traits were chosen as tag QTNs to represent QTL. A five-fold random cross-validation was used to test the accuracy of the ten GS models. Both the genomic and phenotypic datasets for drought-related and root traits obtained from 106 and 101 accessions, respectively, were randomly divided into five subsets. Each subset was sequentially used as test data for a given partition, while the remaining four subsets constituted the training dataset. This partitioning method was replicated 50 times. Then, the Pearson correlation coefficient (*r*) between the GS-predicted GEBVs and the measured phenotypic values was calculated to determine the accuracy of the genomic predictions. Further, a joint analysis of variance with Tukey's multiple pairwise-comparisons (HSD.test function) was performed to compare GS models constructed from QTL marker sets to test the statistical significance of differences in *r* values using the R agricolae package (https://cran.r-project.org/web/packages/agricolae/index.html). Tag QTNs or QTL identified by GWAS were used as markers to construct GS models. Two sets of QTN markers were accessed for GS of each trait: the QTL set from the traits *per se*, and the combined QTL set from all the traits. The positive-effect allele of the tag QTN was encoded "1" and the alternative allele "-1" for model construction. Missing marker data was imputed using the EM algorithm implemented in the R package RR-BLUP (Endelman 2011)

Three marker sets based on GWAS were used to evaluate GS models for six drought-related traits: (1) tag QTNs for the drought-related traits *per se* (**Table 2.3**), (2) tag QTNs for their respective stress indices, and (3) the overall set of 1,057 QTNs identified for all drought-related traits and indices. A total of ten GS statistical models were evaluated for genomic prediction

performance: seven parametric (RR-BLUP, GBLUP, BRR, BL, Bayes A, Bayes B, and Bayes C) and three non-parametric (RF, SVR and RKHS) models.

Table 2.3 Number of tag quantitative trait nucleotides (QTNs) for drought-stress related and root traits used for downstream genomic selection (GS).

Drought-related traits	No. of QTNs	No. of Tag QTNs	Drought-related stress index traits	No. of QTNs	No. of Tag QTNs	Root traits	No. of QTNs	No. of Tag QTNs
BDLWt_S	34	29	BDLWt_STI	55	42	ARD	127	95
CT_S	32	24	CT_STI	32	18	MaxR	48	39
HT_S	30	25	HT_STI	63	57	NWA	27	18
SPB_S	41	31	SPB_STI	39	29	NWL	25	23
TSW_S	42	37	TSW_STI	38	29	NWSA	31	22
YLD_S	33	26	YLD_STI	42	33	NWV	76	60
						SDWt	85	68
Total	212	172	Total	269	208	Total	419	325

The parametric models RR-BLUP, GBLUP and BL consistently outperformed the RF and SVR models across all traits and sets of markers (**Fig. 2.2a-f**). The highest predictions were observed using tag QTNs specific to the traits *per se* compared to the overall set of 1,057 QTNs, and the highest predictive ability range of 0.88 to 0.92 was observed for HT_S and the lowest of 0.60 to 0.75 for SPB_S. Likewise, the predictive ability for the stress indices varied from 0.88 to 0.93 for HT_STI, and 0.64 to 0.87 for CT_STI as illustrated in **Fig. 2.2a-f**. A slightly higher predictive ability was obtained for all index-based traits when compared to the drought-related traits determined under water-limiting stress conditions, using either the tag QTNs for the drought-related traits *per se* or the overall set of 1,057 QTNs.

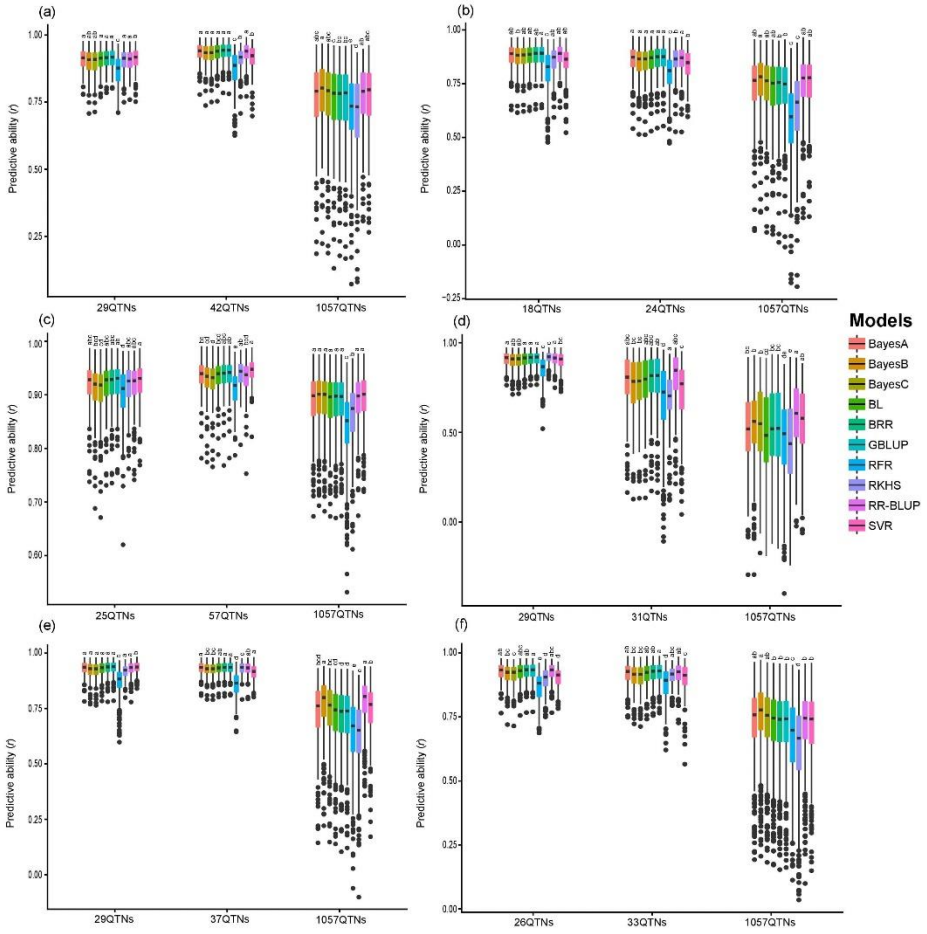


Fig. 2.2 Comparisons of genomic predictive ability (*r*) of ten genomic selection (GS) models for six drought-related traits *per se* and their respective indices: (a) BDLWt_S and BDLWt_STI; (b) CT_STI and CT_S; (c) HT_S and HT_STI; (d) SPB_STI and SPB_S; (e) TSW_STI and TSW_S; and (f) YLD_S and YLD_STI. Each panel displays the *r* values obtained from the ten GS models using three quantitative trait nucleotide (QTN) marker sets: the tag QTNs identified by all genome-wide association studies (GWAS) models for each trait *per se* (left), for their corresponding stress indices (middle), and the overall set of 1,057 QTNs detected for all drought-related traits (right). The number of QTNs used in the models are on the x-axis. A five-fold cross-validation was used to estimate *r* values. The different letters on the top of boxes represent statistical significance at the 5% probability level obtained using the Tukey multiple comparison statistical test.

For root traits, two marker sets were used for genomic prediction: (1) tag QTNs for the root traits listed in **Table 2.3** and (2) the overall set of 528 QTNs identified for all root traits combined. Most root traits had a significantly higher predictive ability using tag QTNs specific to the root traits compared to the overall set of 528 QTNs, except for NWA and NWL where the predictive ability difference of the two marker sets for these traits was only 0.01 and 0.04 across all models.

For the overall marker set, the highest predictive ability of 0.90 was observed for NWSA, whereas the lowest was 0.72 for SDWt. Similarly, using the tag QTNs specific to the individual traits as markers, the highest predictive abilities were obtained for NWSA, ARD and NWV, with ranges of 0.92 to 0.93, as shown in **Fig. 2.3a-g**.

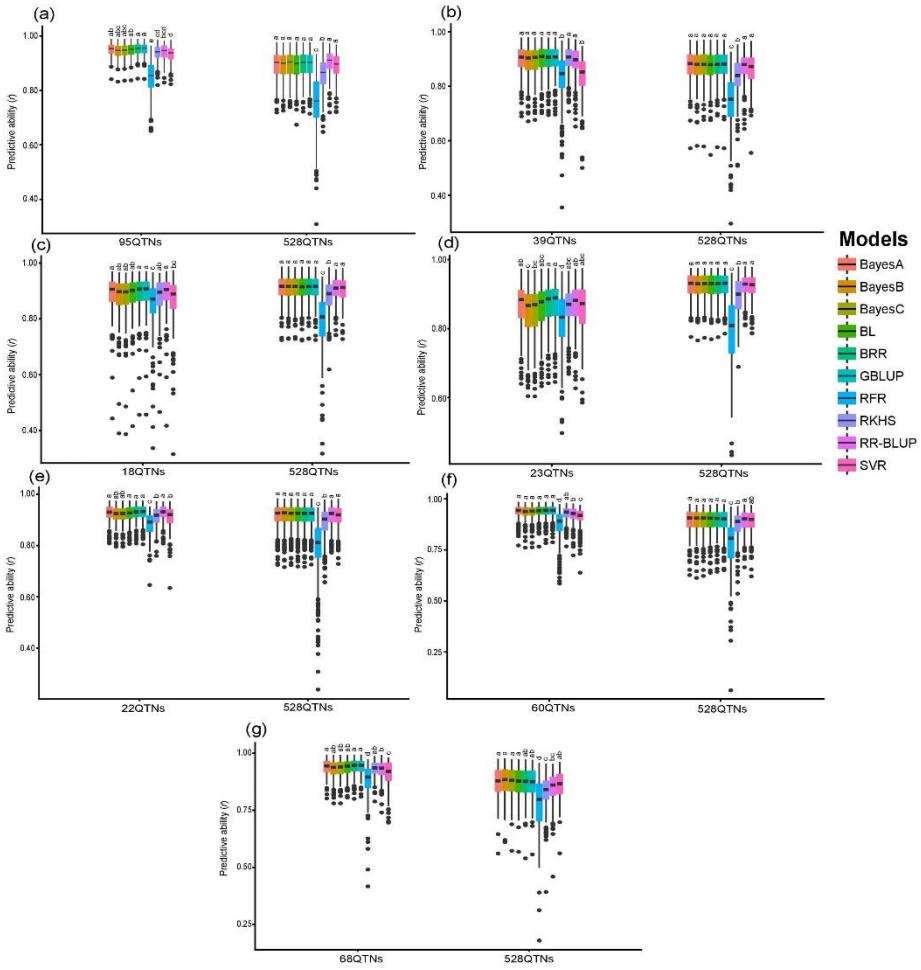


Fig. 2.3 Comparisons of genomic predictive ability (*r*) of ten genomic selection (GS) models for seven root traits: (a) ARD; (b) MaxR; (c) NWA; (d) NWL; (e) NWSA; (f) NWV; and (g) SDWt. Each panel displays the *r* values obtained from the ten GS models using two quantitative trait nucleotides (QTN) marker sets: the tag QTNs identified by all genome-wide association studies (GWAS) models for the root traits *per se*, and the overall set of 528 QTNs detected for all root traits. The number of QTNs used in the GS models are on the x-axis. A five-fold cross-validation was used to estimate *r*. The different letters on the top of boxes represent statistical significance at the 5% probability level obtained using the Tukey multiple comparison statistical test.

GP has revolutionized many breeding programs by accelerating the selection process and the rate of genetic gains for highly complex traits such as drought resistance (Shikha et al. 2017; Ben Hassen et al. 2018; Velazco et al. 2019). The implementation of GP as a breeding tool is limited by the cost of sequencing and the many challenges of phenotyping the TRP for complex traits such as those related to drought and roots for examples. An attempt at optimizing genomic predictive ability was performed by comparing the predictive ability of a trait-specific QTN marker set to that of a combined QTN marker set from related traits. In all cases, the trait-specific QTN marker-based GS models achieved higher genomic predictive ability.

2.4 A QTL-based Genomic Selection (GS) Strategy

Based on the above-described findings, we propose the QTL-based GS strategy depicted in **Fig. 2.4**. This strategy incorporates GS modeling with QTL identification by SLSM and MLSM GWAS from TRP followed by GP from TP, and optional GS evaluation in breeding programs.

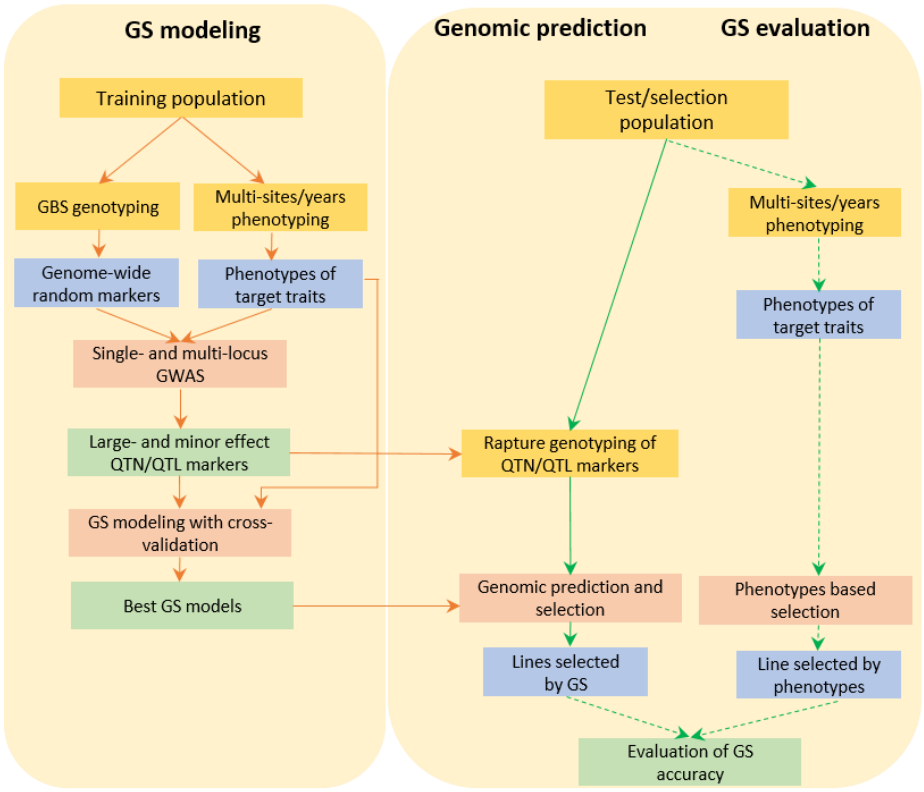


Fig. 2.4 A quantitative trait locus (QTL)-based genomic selection (GS) strategy that incorporates GS modeling with QTL identification by single- and multi-locus genome-wide association studies (GWAS) from training population (TRP), genomic prediction (GP) from test population (TP), and an optional GS evaluation in breeding programs.

2.4.1 Genomic Selection (GS) Modeling

In a breeding program using GS, the first step is to choose a suitable TRP to build GS models. A good TRP should be closely related with the TP that is usually comprised of breeding lines developed from multiple crosses. Consequently, the TRP should contain the parental material or its ancestors. The TRP should also contain as many of the target traits' favourable alleles or genes. Breeders aim to accumulate as many potential favourable alleles as possible. Thus, a sufficient number of diverse genotypes are needed. In a practical GS-based breeding program, TRPs often include historical breeding lines and cultivars, where some could be part of regional or national trials which are evaluated over multiple years, locations and with replications. These provide useful data for target traits, such as yield, seed quality and disease resistance.

2.4.2 Test Population (TP) and Genomic Selection (GS)

The use of TP is also one of the important factors for improving GS accuracy. The optimal trade-off for the genetic diversity between the TP and TRP must be determined case by case such as their correlation and genetic distance. For instance, to successfully deploy GS in flax breeding programs, we propose an integrated approach that includes GS, genomic cross prediction, and phenotypic evaluation, as shown in **Fig. 2.5**. In the flax breeding program, every year from Yr1-Yr5 additional lines of both TP and TRP are added to optimise diversity and increase selection accuracy. For this purpose, each line needs to be evaluated using the best GS models, traditional selection, and their performance comparison to assess the progress of successful implementation of GS programs.

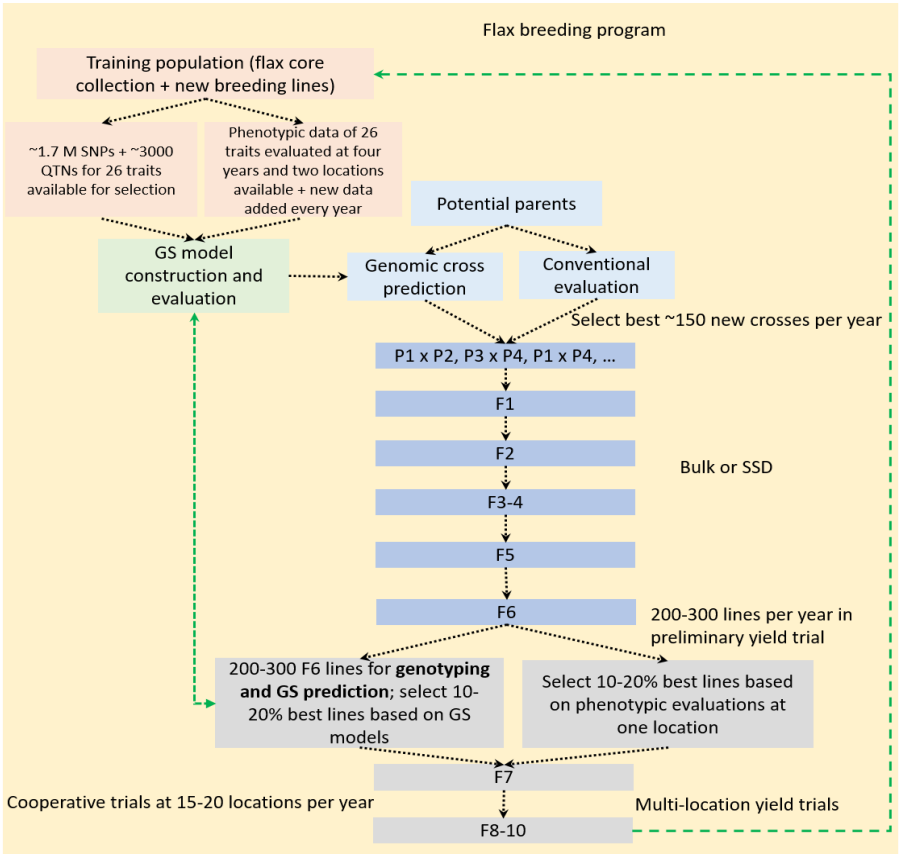


Fig. 2.5 A proposed approach for of the integration of genomic selection (GS) in a flax breeding program, including cross prediction and its evaluation through comparisons with conventional phenotypic evaluations.

2.5 Genomic Selection (GS) Evaluation

2.5.1 Cost of Genomic Selection (GS)

The rapid advances in HTS technologies have significantly reduced the cost of marker data points. Marker technologies that result in even greater cost-effective genotype characterization continue to be welcomed. The availability of such technology would promote the use of GS in practical breeding schemes. Genotyping-by-sequencing (GBS) is one of the most cost-effective genotyping platforms that can be employed in large populations and is extensively used in breeding programs. GBS generate thousands of SNPs at a reasonable cost per sample that can be used for GS of species regardless of the existence of a reference genome (Poland and Rife 2012). Some of the recent breakthroughs in GBS include skim-based GBS (Bayer et al. 2015), genotyping-in-thousands by sequencing (GT-seq) (Campbell et al. 2015), and rAMpSeq (Buckler et al. 2016). Skim-based GBS allows for high-resolution genotyping while requiring low-coverage sequencing (Bayer et al. 2015). GT-seq dramatically reduces the cost of genotyping hundreds of targeted SNPs compared to existing methods (Campbell et al. 2015). Similarly, rAmpSeq in maize has been reported (Buckler et al. 2016) to cost as little as $2 per sample and has the potential of transforming breeding programs.

GBS markers adopted in an oat breeding program ran smoothly and predictably at a lower cost per sample than any other whole-genome marker system available at the time (Mellers et al. 2020). The cost per sample of this approach and that of a few other variations are outlined in **Table 2.4**. These costs are similar to those reported by CIMMYT, Cornell University and Kansas State University (Poland and Rife 2012; Bassi et al. 2016). The sequencing cost vary depending on the read depth and number of markers targeted. For GBS, the sequencing depth is lower than we might need for an application such as GWAS, but adequate for GS (Bekele et al. 2018). The sequencing

cost can also be reduced by moving to a higher-capacity sequencing system where samples are multiplexed using distinct adaptors. A subset of useful GBS tags has been assembled to develop a bait-based tag capture system that is designed specifically for Canadian oat breeding programs, similar to the procedure known as 'Rapture'(Ali et al. 2015). This method provides the possibility for 384 samples to be combined per sequencing lane, which reduces the sequencing cost to $21/sample and/or provide a higher sequencing depth that improves the accuracy of genotyping. Another advantage is the possibility to combine this dataset with more than ten thousand lines that have been previously genotyped. Rapture translates into low-cost and high-read coverage at target loci, while allowing previous TRP genotype data based on non-captured GBS data to be fully compatible with the added Rapture data. Given the high proportion of homozygous sites within these lines, even low coverage can provide sufficient data for a probabilistic genotyping approach (Nielsen et al. 2011).

Phenotyping costs vary from $40-45 per plot. In the oat program, phenotyping is replicated twice. Thus, the total cost may reach $80-90 per genotype. In Brandon, Manitoba phenotyping for each plot (plot size: 1.5 m x 7.5 m trimmed to a 6-m harvest area, three replicates, with data collected for heading, height, maturity, lodging, yield, test weight, seed moisture content at harvest) was estimated at $125 per plot or $375 per genotype. Quality analysis extra: *Fusarium* head blight/deoxynivalenol testing adds another ~ $75 per plot or $225 per genotype. In summary, the phenotypic cost is high and cost-reduction options are limited because field phenotyping and quality analyses rely heavily on labor. On the other hand, the cost of genotyping continues to drop through either higher throughputs or cheaper chemistries and platforms which, assuming that genomic accuracy is achieved, gives the edge to genotyping over phenotyping. However, in order

to maintain accurate prediction equations, phenotyping accuracy remains paramount and efforts must be made to achieve the best quality phenotyping of the TRP.

Table 2.4 Cost (US dollar per sample) of genotyping by target sequencing (GBTS), genotyping by sequencing (GBS), and RAD capture (Rapture) for test population (TP).

Crops	**Maize (2.3 Gb) Guo et al. (2019)**		**Oat (12.3 Gb) Bekele et al. (2018)**	
Genotyping protocol	GBTS	GBTS	GBS	Rapture
Number of markers	5 K	1 K	~10K	5K
DNA extraction	0.50	0.50	3.08	3.08
Construction of library	6.61	6.61	8.46	8.46
Probe hybridization	3.00	3.00	-	-
Sequencing	2.25	0.75	12.31	4.62
Total (per sample)	12.36	10.86	23.85	16.15

2.6 Parent Evaluation and Cross Prediction

Plant breeding, or genetic improvement of plant species, has played an essential part in the evolution of human societies (Spiertz 2014). Farmers must improve productivity and plant tolerance to biotic and/or abiotic stresses to provide a steady food supply. Plant breeding has traditionally relied on a large number of crosses for selection of desirable traits, which is a lengthy and laborious approach that requires a lot of resources. However, to boost yields and feed a fast-growing global population (Tester and Langridge 2010), to meet the needs of customers with different dietary requirements and to produce high-quality and safe food, more effective breeding methods are needed. It is commonly acknowledged that genetic data will be used to design effective breeding methods (Bernardo 2008). The advances in next-generation sequencing technologies have resulted in the availability of a large number of low-cost molecular markers (Davey et al. 2011). This has paved the way for the widespread use of molecular markers in plant breeding. For instance, the use of genomic cross predictions in flax breeding is a novel approach

to assist quickly the evaluation and optimize breeding programs (Khan et al. 2022). Plant breeders make many crosses each year and analyse their progeny in fields or greenhouses. However, generally only few progeny will outperform the parents. The use of genomic cross prediction in breeding can help with regards to three aspects: i) to increase accuracy and efficiency, ii) to shorten the cycle and iii) to select parents and progeny.

Genomic cross prediction methods that allows plant breeders to accelerate the selection of superior parents based on genotypic and phenotypic data promise to improve the efficiency of breeding programs, not only in flax but also in other plant species (You et al. 2022b). Genomic cross prediction has indeed evolved into a vital tool in scientific research, serving as both a preliminary validation of hypotheses and as a guide for empirical studies (You et al. 2022b). Many software packages are available for the genomic cross prediction. They can be used to simulate a wide range of crossing and progeny scenarios using genomic data. For instance, QuLinePlus (Hoyos-Villegas et al. 2019), QuHybrid (Wang et al. 2003), ADAM-plant (Liu et al. 2019b), MareyMap (Siberchicot et al. 2017) and other simulation tools are useful for the purpose of evaluating the efficiency of different selections strategies.

2.6.1 Genomic Cross Prediction for Flax Improvement

To date, crossbreeding remains the corner stone of breeding programs of its success relies heavily on having access to comprehensively evaluated germplasm for parent selection which is a major challenge because such knowledge is usually limited to a narrow set of genotypes. In flax, You et al. (2022b) have compiled 290 linseed accessions and identified a total of 1,811 QTNs for five major traits, i.e., YLD, DTM, oil content, linolenic acid, and powdery mildew (PM) resistance. The major objective of this study is to use this 290-accession TRP to evaluate the accessions' potential as parents through the development of GS models and simulation of progeny populations

from intercross using the available genomic data (SNPs). In this scheme, the developed GS models are subsequently used to predict GEBVs of all progenies of a cross (You et al. 2022b). Cross performance is evaluated according to usefulness criteria which are a function of progeny population mean and genetic variance. This study highlights that the use of genomic cross prediction in flax can be highly effective for two major reasons: i) a large number of virtual crosses can be easily simulated, ii) the use of GWAS for the identification of QTNs and that of optimized GS models can advantageously predict the GEBVs of progenies. In the future, genomic cross predictions will need to be evaluated on a variety of traits, including biotic and/or abiotic stress, not just in flax but also in maize, barley, and wheat, to ensure their accuracy. Finally, the use of genomic cross prediction in breeding programs has the potential to provide an effective and low-cost breeding tool.

2.6.2 Future Based: Integrated Flax Breeding Improvement Strategy

For the flax breeding program, we propose the use of an integrated approach that comprises GS, genomic cross prediction, and conventional phenotypic evaluation. In this strategy, GS models for target traits such as YLD, DTF, and FW using SNPs and QTNs identified from the flax core collection are first developed. The flax core collection (~290 linseed and 90 fiber lines) serves as the basic TRP, banking on high-quality phenotypic data already collected in multiple environments (years and locations), on a curated ~1.7 M SNP dataset and on previously identified QTNs for all major traits. This TRP should be genetically highly correlated with the breeding populations because most ancestors of present-day flax cultivars are included in this core collection. This primary TRP is supplemented with new cultivars and breeding lines each year. The predictive GS models from multi-year and multi-environment is integrated or applied independently using common criteria. Each year, multiple models are assessed from the new TRP iteration and its

incremental phenotypic data. The genome-wide random SNP and the trait-specific QTN marker sets are used to evaluate models for their predictive abilities. The best models are then selected based on the CV procedure and the target environments' correlations.

Secondly, every year the best GS models are applied to the flax breeding program, including the test progenies. TPs and additional training genotypes are genotyped using GBS. Every year, 300 F_6 lines are chosen from different populations. The performance of these selected lines are compared by GS vs. those chosen through traditional selection in the following generations to assess the progress and success of GS. Every year, breeding progress for each mode of selection i.e., overall breeding progress measured in the same multi-environment tests and with GS selections integrated in the second year and beyond. Multiple comparisons are performed through the assessment of common checks across locations and years such as an overall trajectory-analysis of GS vs traditionally-selected lines is feasible.

Thirdly, GS models are used to predict progenies from which the putative best parents are selected for crossing (**Fig. 2.5**). Both linseed and fiber accessions of the core collection and newly-added cultivars and breeding lines are used to make virtual single crosses and evaluate their progeny. For every cross, 500 recombinant inbred lines (RILs) are simulated based on an additive model. Their genotypes are generated based on the QTN/SNP markers of the two parents and the genetic recombination within individuals of the progeny populations based on the previous genetic maps (Cloutier et al. 2012). The optimal GS models for each trait are used to predict the GEBVs of the RILs for each cross and trait. The crosses and their parents thus evaluated are recommended to the breeding program based on their general and specific combining abilities. We anticipate that this integrated approach will boost up flax breeding progress through accelerating the genetic gains for yield and other complex traits.

2.7 Conclusions and Future Prospects

Employing GS with a QTL-based method not only improves predictive ability, but also aids in the implementation of successful genomic cross prediction in flax breeding for two primary reasons. First, it helps breeders in assessing parents and choosing the potentially best crosses, and second, it offers an effective and low-cost breeding assisting tool. The predictive ability of a GS program is determined by multiple factors, such as the TRP and its relatedness with the TP, the statistical GS models and the molecular marker type and density. The use of large- and minor effect QTL identified by single- and multi-locus GWAS models significantly increase GS models' predictive ability. As previously discussed, the advantage of QTL-based GS modelling over genome-wide random markers has been demonstrated in few case studies. In summary, the two main goals of GS are to improve the accuracy of GEBVs and to improve breeding program outcomes. Another goal may be to trade part of the phenotyping effort for genotyping in order to reduce cost and/or shorten the breeding cycle, keeping in mind that the incorporation of target-loci based genotyping techniques, such as Rapture and GBTS, is likely to reduce cost in flax or other crops.

Chapter 3

Genome-wide Identification of ATP Binding Cassette (ABC) Transporter and Heavy Metal Associated (HMA) Gene Families in Flax (*Linum usitatissimum* L.)

Nadeem Khan[1, 2], Frank M. You[1, *], Raju Datla[3], Sridhar Ravichandran[1], Bosen Jia[1, 2] and Sylvie Cloutier[1, 2, *]

[1]Ottawa Research and Development Center, Agriculture and Agri-Food Canada, 960 Carling Avenue, Ottawa

[2]Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa, ON K1N 6N5, Canada

[3]Aquatic and Crop Resource Development, National Research Council Canada, 110 Gymnasium Place, Saskatoon, SK S7N 0W9, Canada

*****Correspondence: frank.you@agr.gc.ca (F.M.Y.) and Sylviej.Cloutier@agr.gc.ca (S.C.)

This chapter has been published in a peer-reviewed journal: Khan N, You FM, Datla R, Ravichandran S, Jia B, Cloutier S (2020) Genome-wide identification of ATP binding cassette (ABC) transporter and heavy metal associated (HMA) gene families in flax (*Linum usitatissimum* L.). BMC Genomics. 21:722

Authors' contribution

Conceptualization, methodology, formal analysis, investigation, writing—original draft preparation, N.K.; writing—review and editing, supervision, project administration, funding acquisition, FMY and S.C.; design and performing the RNA-seq work R.D.; the creation of new software and pipeline used in the work, S.R and B.J. All authors read and approved the final manuscript.

Abstract

The recent release of the reference genome sequence assembly of flax, a self-pollinated crop with 15 chromosome pairs, into chromosome-scale pseudomolecules enables the characterization of gene families. The ABC transporter and HMA gene families are important in the control of cadmium (Cd) accumulation in crops. To date, the genome-wide analysis of these two gene families has been successfully conducted in some plant species, but no systematic evolutionary analysis is available for the flax genome. Here we describe the ABC transporter and HMA gene families in flax to provide a comprehensive overview of its evolution and some support towards the functional annotation of its members. The 198 ABC transporter and 12 HMA genes identified in the flax genome were classified into eight ABC transporter and four HMA subfamilies based on their phylogenetic analysis and domains' composition. Nine of these genes, i.e., *LuABCC9*, *LuABCC10*, *LuABCG58*, *LuABCG59*, *LuABCG71*, *LuABCG72*, *LuABCG73*, *LuHMA3*, and *LuHMA4*, were orthologous with the Cd associated genes in *Arabidopsis*, rice and maize. Ten motifs were identified from all ABC transporter and HMA genes. Also, several motifs were conserved among genes of similar length, but subfamilies each had their own motif structures. Both the ABC transporter and HMA gene families were highly conserved among subfamilies of flax and with those of *Arabidopsis*. While four types of gene duplication were observed at different frequencies, whole-genome or segmental duplications were the most frequent with 162 genes, followed by 29 dispersed, 14 tandem and 4 proximal duplications, suggesting that segmental duplications contributed the most to the expansion of both gene families in flax. The rates of non-synonymous to synonymous (*Ka/Ks*) mutations of paired duplicated genes were for the most part lower than one, indicative of a predominant purifying selection. Only five pairs of genes clearly

Abstract

exhibited positive selection with a *Ka/Ks* ratio greater than one. Gene ontology analyses suggested that most flax ABC transporter and HMA genes had a role in ATP binding, transport, catalytic activity, ATPase activity, and metal ion binding. The RNA-Seq analysis of eight organs demonstrated diversified expression profiling patterns of the genes and revealed their functional or sub-functional conservation and neo-functionalization. Characterization of the ABC transporter and HMA gene families will help in the functional analysis of candidate genes in flax and other crop species.

Keywords: Flax ● ABC transporter ● HMA ● Gene duplication ● Expression profiling

Chapter 3: Identification of ABC and HMA Gene Families in Flax

3.1 Introduction

ATP binding cassette (ABC) transporter genes are ubiquitous across the three domains of life: Eukarya, Eubacteria and Archaea (Jones and George 2004; Rees et al. 2009). Plant genomes harbor more than 100 ABC transporters which are involved in a broad range of biological functions (Kang et al. 2011). ABC transporters comprise at least four domains: two transmembrane domains (TMDs) embedded in the membrane bilayer, and two nucleotide-binding domains (NBDs) located in the cytoplasm (Rees et al. 2009). The structure of the TMDs is highly diverse and varies in the number of transmembrane helices, whereas the NBDs have highly conserved helices (Schneider and Hunke 1998). The ABC transporters are further categorized into full-size transporters with two NBDs and two TMDs and half-size transporters with only one of each domain (Hollenstein et al. 2007). Therefore, two half-size transporters must form either a homodimer or a heterodimer to be functionally active.

In plant genomes, ABC transporters are categorized into eight subfamilies (ABCA-ABCG and ABCI) (Kang et al. 2011). Proteins belonging to the ABCA-ABCD subfamilies have a forward domain organization (TMD-NBD) whereas ABCG and ABCI subfamilies have an inverse domain organization (NBD-TMD) (Mishra et al. 2019). ABCE and ABCF possess only two NBDs and are designated as soluble proteins (Mishra et al. 2019). In *Arabidopsis*, 130 ABC transporter genes have been identified but few have been functionally characterized (Yan et al. 2017). Previous studies have shown that ABC transporters participate in a wide range of processes including the transport of ions, carbohydrates, lipids, xenobiotics, antibiotics, drugs, and heavy metals (Pighin et al. 2004; Gadsby et al. 2006; Sipos and Kuchler 2006). The two members of the ABCB gene family in *Arabidopsis* (*AtABCB1* and *AtABCB2*) are auxin transporters and the overexpression of *AtABCB1* causes the elongation of hypocotyl cells (Sidler et al. 1998; Noh et al. 2001). Several

members of the ABCC subfamily are responsible for phytate transport as exemplified in *Arabidopsis* (*AtABCC5*) (Nagy et al. 2009), maize (*ZmABCC4*) (Badone et al. 2010) and rice (*OsABCC13)* (Tagashira et al. 2015). Two other ABCC transporters (*AtABCC1* and *AtABCC2*) mediate tolerance to both cadmium (Cd) and mercury by vacuolar sequestration (Park et al. 2012). The ABCF subfamily member *AtABCF3* in *Arabidopsis* is involved in root growth and development (Kato et al. 2009). ABCG subfamily members were reported to be involved in cuticle formation and Cd tolerance such as in *Arabidopsis* (*AtABCG32*) (Bessire et al. 2011) and rice (*OsABCG31* and *OsABCG36*) (Chen et al. 2011; Fu et al. 2019). Also, the ABC transporter *AtABCG36* in *Arabidopsis* was shown to mediate Cd uptake in the epidermal cells of roots (Kim et al. 2007) and to be up-regulated by a Cd treatment (Bovet et al. 2003).

Heavy metal (HM) pollution in food, water, and soil is hazardous to human health. HMs have become one of the major concerns across the globe due to the extensive industrialization and because of their direct and indirect effects on soil and crop productivity (Emamverdian et al. 2015). HMs such as zinc, copper, manganese, cobalt, and nickel are essential for various biological processes (Salla et al. 2011). In contrast, other HMs such as arsenic, lead, and Cd are highly toxic to plants and negatively affect crop productivity (Emamverdian et al. 2015). Many HMA genes have been shown to play specific functions in different plant species. For example, *OsHMA2* is associated with vascular tissue loading of zinc and tonoplast localization in rice (Yamaji et al. 2013). *OsHMA3,* localized in the tonoplasts, translocates Cd to the roots while *OsHMA4* transports copper to the roots (Huang et al. 2016). *HvHMA1* is involved in zinc and Cd translocation into barley grain (Mikkelsen et al. 2012). In wheat, HMA genes also play an important role in Cd translocation and are localized in the plasma membrane (Tan et al. 2013). Overexpression of *AtHMA3* in *Arabidopsis* resulted in a 2- to 3-fold increase in Cd accumulation when compared to

wild-type plants (Morel et al. 2009). These cited studies provide an overview of the importance of both ABC transporter and HMA genes in various plant species, but no systemic studies have been reported in flax.

The initial draft of the flax genome sequence was produced using whole-genome shotgun (WGS) sequencing with short reads obtained on the Illumina sequencing platform (Wang et al. 2012b). A *de novo* assembly generated 88,384 scaffolds, totaling 318 Mb and representing ~81% of the estimated ~370 Mb flax reference genome (You et al. 2018b). Thus, the availability of this recent update of the flax genome (version 2.0) constitutes a genomic resource that allows the identification of gene families, evolutionary relationships, and structural analyses. To date, ABC transporter and HMA gene families have been studied in several plant species including *Oryza sativa* and *Arabidopsis thaliana* (Jasinski et al. 2003), *Zea mays* (Pang et al. 2013; Cao et al. 2019), *Brassica rapa* (Yan et al. 2017), *Brassica napus* (Li et al. 2018c), *Triticum aestivum* (Bhati et al. 2015), and *Vitis vinifera* (Çakır and Kılıçkaya 2013). A previous report on ABC transporters in flax has been published (Lane et al. 2016), but classification patterns, gene duplications, evolutionary and collinear relationships, structural and correlation analyses of paralogous pairs, and the identification of Cd associated genes covered herein were not explored. Several other gene families have also been identified in flax, based on bioinformatics analysis such as aquaporin (Shivaraj et al. 2017), dirigent (Corbin et al. 2018), chalcone (Eom and Hyun 2016), detoxification efflux carriers (Monostori et al. 2017), and ubiquitous glycosyltransferase (Barvkar et al. 2012).

In this research work, we hypothesized that either whole genome duplications (WGDs) or tandem events contributed to the expansion of the ABC transporter and HMA gene families in flax. Therefore, we studied the phylogenetic relationships, gene annotation, physicochemical properties, chromosomal distribution, gene synteny, protein-protein interactions (PPI), and gene

duplications of all predicted ABC transporter and HMA genes of the flax genome to understand their evolution and hypothesize their putative functions. Finally, we examined the gene ontology (GO) and expression profiling of ABC transporter and HMA genes in eight tissues, namely root, seed, ovary, and embryo at five different stages (heart, globular, torpedo, mature and cotyledon). This comprehensive analysis is the first report on ABC transporter and HMA genes in flax, providing gene candidate information for future marker association studies on heavy metal accumulation including Cd.

3.2 Materials and Methods

3.2.1 Identification of ABC Transporter and HMA Genes, and their Duplications

Two methods were used to identify ABC transporter genes in flax. Firstly, the ABC transporters were identified based on the 129 reference sequences (Verrier et al. 2008) from the *Arabidopsis* genome (version 10.0, http://www.arabidopsis.org/) by BLASTP search method using an E-value cut-off of 1.0E-10 against the flax genome (Wang et al. 2012b). Secondly, we performed a Hidden Markov Model (HMM) search against the flax genome to confirm the presence of ABC transporter genes using HMMER (version 3.2.1) with the default options (Finn et al. 2011). The ABC transporter domains included ABC transporter (PF00005), ABC-2 transporter (PF01061), ABC transporter transmembrane region (PF00664), cytochrome c polymerization (CYT) (PF01458) or mammalian cell entry (mce) related protein (PF02470). These domains were downloaded from the Pfam (version 32.0) database (http://pfam.xfam.org/) (El-Gebali et al. 2018). Similarly, the *HMA* genes were identified based on the eight reference gene sequences (**Appendix 3.1**) of *Arabidopsis* using BLASTP method as discussed above.

After merging the results, ABC transporter and HMA genes were further screened on the basis of their domain composition. The duplicated results between the two methods were eliminated. A total of 745 ABC transporter and HMA proteins were identified among all species other than those of *Arabidopsis thaliana*.

Sequences of the 15 chromosomes of flax (version 2.0) were obtained from NCBI under GenomeProject ID no. 68161 (accession numbers from CP027619–CP027633) (You et al. 2018b) and protein sequences were downloaded from Phytozome (version 12.1) (Goodstein et al. 2012). Genomic sequences of *Populus trichocarpa* (version 3.1), *Vitis vinifera* (version Genoscope12X), and *Brachypodium distachyon* (version 1.2) were obtained from Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html) (Goodstein et al. 2012). The obtained protein sequences of ABC transporter and HMA genes were further verified for ABC/HMA domain compositions in the NCBI-Conserved Domain database (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) (Marchler-Bauer et al. 2017) and SMART (http://smart.embl-heidelberg.de/) (Letunic and Bork 2017). Protein sequences with errors, short (<100 aa), or without ABC or HMA domains were removed.

For the identification of flax Cd associated genes, several ABC transporter and HMA responsive genes were used as a reference. These genes included *AtABCC1*, *AtABCC2*, *AtABCG36*, *AtHMA3* and *AtHMA4* in *Arabidopsis* (Verret et al. 2004; Kim et al. 2007; Morel et al. 2009; Park et al. 2012), *OsABCG31*, *OsABCG36* and *OsHMA2* in rice (Takahashi et al. 2012b; Fu et al. 2019), *ZmHMA2* and *ZmHMA3* in maize (Cao et al. 2019). Nine Cd associated genes were identified by performing the BLASTP method using the same criteria as discussed above for ABC and HMA genes.

3.2.2 Phylogenetic Characterization of ABC Transporter and HMA Genes, and Synonymous (Ks) and Non-synonymous (Ka) Substitution Rates for Duplicated Genes

Multiple sequence alignments for ABC transporter or HMA protein sequences from *Linum usitatissimum*, *Arabidopsis thaliana*, *Populus trichocarpa*, *Vitis vinifera*, and *Brachypodium distachyon* were performed using MUSCLE (version 3.8.1551). Phylogenetic trees were then constructed using the MEGA (version 7.0) software (Kumar et al. 2016) with the maximum likelihood (ML) method and the Jones, Taylor, and Thornton amino acid substitution model (JTT model). The JTT model was chosen based on the results using the option provided in MEGA to find the best fit model for evolution of ABC/HMA genes.

The different types of gene duplications in the flax genome were identified using MCScanX (Wang et al. 2012a). Synonymous (*Ks*) and non-synonymous substitution (*Ka*) rates were also calculated for duplicated gene pairs as previously described (Khan et al. 2019a). Also, a substitution rate of 1.5×10^{-8} substitutions per synonymous site per year (Koch et al. 2000) was used to estimate the divergence time of duplicated genes.

3.2.3 Conserved Motifs, Gene Structure, and Physicochemical Parameters

Multiple Em for Motif Elicitation (version 5.1.0) (Bailey et al. 2009) were used for scanning the conserved motifs of LuABC and LuHMA proteins. The maximum number of motifs of 10, with a minimum width of 50 aa and a maximum of 100 aa, were set as parameters. TBtools (version 0.66) (Chen et al. 2020) was used to visualize both the motif composition and gene structure. The ExPASY PROTPARAM tool (http://web.expasy.org/protparam/) was accessed to determine the following physicochemical properties: molecular weight (MW), isoelectronic points (pI), and GRAVY values for each gene. The subcellular localization of both ABC and HMA genes was

predicted using the web applications of WOLF PSORT (http://wolfpsort.hgc.jp/) (Horton et al. 2007) and BUSCA (http://busca.biocomp.unibo.it/) (Savojardo et al. 2018).

3.2.4 Chromosomal Location, Gene Synteny Analysis, and Protein-Protein Interaction (PPI) Analysis

The gene synteny between flax and *Arabidopsis* was analyzed based on the gene annotation data of both species and illustrated using shinyCircos (Yu et al. 2017). The PPI analysis for all the ABC transporter and HMA proteins was carried out using the online STRING server (version 11.0) (http://string-db.org/) (Szklarczyk et al. 2018) with the following parameters: medium score of 0.400, number of K means clustering of 3, and default values for the remaining options. The results of the interaction network were visualized using Cytoscape (version 3.4.0) (Shannon et al. 2003).

3.2.5 Plant Materials, RNA Sequencing and Read Data Analysis

Flax cultivar CDC Bethune was planted in the greenhouse under growth conditions previously described (Venglat et al. 2011). Tissues from root, seed, anther, and ovary at five embryonic developmental stages (heart, globular, torpedo, mature and cotyledon embryo) were collected for RNA extraction with two biological replicates for each tissue. Total RNA was extracted from each sample following the RNAqueous kit protocol (Ambion, Catalog #1912) and RNAqueous-Micro kit protocol (Ambion, Catalog #1931). The samples were homogenized in lysis buffer with polypropylene pestles in 1.5 ml Eppendorf tubes on ice. For RNA-Seq profile analysis, Illumina mRNA-Seq libraries were prepared using the TruSeq RNA kit (ver. 1, rev. A) according to the manufacturer's instructions. An Agilent 2100 Bioanalyzer was used for quantification and quality assessment of sample libraries. For Illumina HiSeq 2000 sequencing, four indexed libraries were pooled per sequencing lane and 100-paired-end sequencing was performed.

The raw reads were initially trimmed by trimmomatic (Bolger et al. 2014) and the trimmed reads were aligned to the flax reference genome sequence using kallisto (Wang et al. 2012b; Bray et al. 2016). The reads with fewer than 5 reads per million (RPM) in at least one library were filtered out. Normalization was performed at trimmed mean of M-values (TMM) using edgeR (Robinson et al. 2010). A general linear model (GLM) was used with glmLRT function to identify differentially expressed genes with false discovery rate (FDR) less than 0.05 (Robinson et al. 2010). The anther results were used as a reference for expression analysis of all tissues. Thus, expression results were presented for the eight tissues by comparing them to those of the anther.

Heat maps were drawn based on normalized Log_2 scale read counts using ClustVis (Metsalu and Vilo 2015). Bidirectional cluster analysis was conducted using maximum distance and complete linkage method**.** Pearson correlation coefficient (*r*) based on expression data among the syntenic pairs of flax was calculated using Rstudio (version 3.6.0).

The GO analysis for ABC transporter and HMA genes in flax was conducted using the Phytozome database (http://phytozome.jgi.doe.gov/pz/portal.html) with keyword search options against the flax genome.

3.2.3 Results

3.2.3.1 ABC Transporter and HMA Genes in Flax and Their Physicochemical Properties

A total of 198 ABC transporter and 12 HMA genes were identified in the flax genome reference sequence of CDC Bethune (You et al. 2018b). The ABC transporter and the HMA genes were classified into eight and four subfamilies, respectively. These genes were denoted as *LuABCA1-LuABCA8*, *LuABCB1-LuABCB48*, *LuABCC1-LuABCC19*, *LuABCD1-LuABCD5*, *LuABCE1-LuABCE2*, *LuABCF1-LuABCF9*, *LuABCG1-LuABCG85*, *LuABCI1-LuABCI22*, and *HMA1-*

HMA12. The basic information on these genes based on their subfamilies, including the protein identifier, coding sequence (CDS) length (bp), and protein properties such as the number of amino acid (aa) residues, molecular weight (kDa), isoelectric point (pIs), and grand average of hydropathicity (GRAVY), is listed in **Appendix 3.2**. The CDS length ranged from 663 to 7,313 bp. The proteins had from 220 to 2,438 aa and a molecular weight of 25.54 to 273.69 kDa. The pIs ranged from 4.93 to 11.67 while the GRAVY values varied from -0.606 to 0.619. Most genes (132/210) had positive GRAVY values, indicating hydrophobic properties.

LuABC and LuHMA proteins were predicted to localize to many subcellular compartments such as the plasma membrane, vacuole, endoplasmic reticulum, nucleus, cytoplasm, chloroplast, Golgi apparatus and mitochondrion. Further, several studies have also confirmed almost an identical localization of ABC transporter and HMA genes (Yan et al. 2017; Fan et al. 2018).

3.2.3.2 Annotation and Phylogenetic Analysis of ABC Transporter and HMA Genes in Plant Species

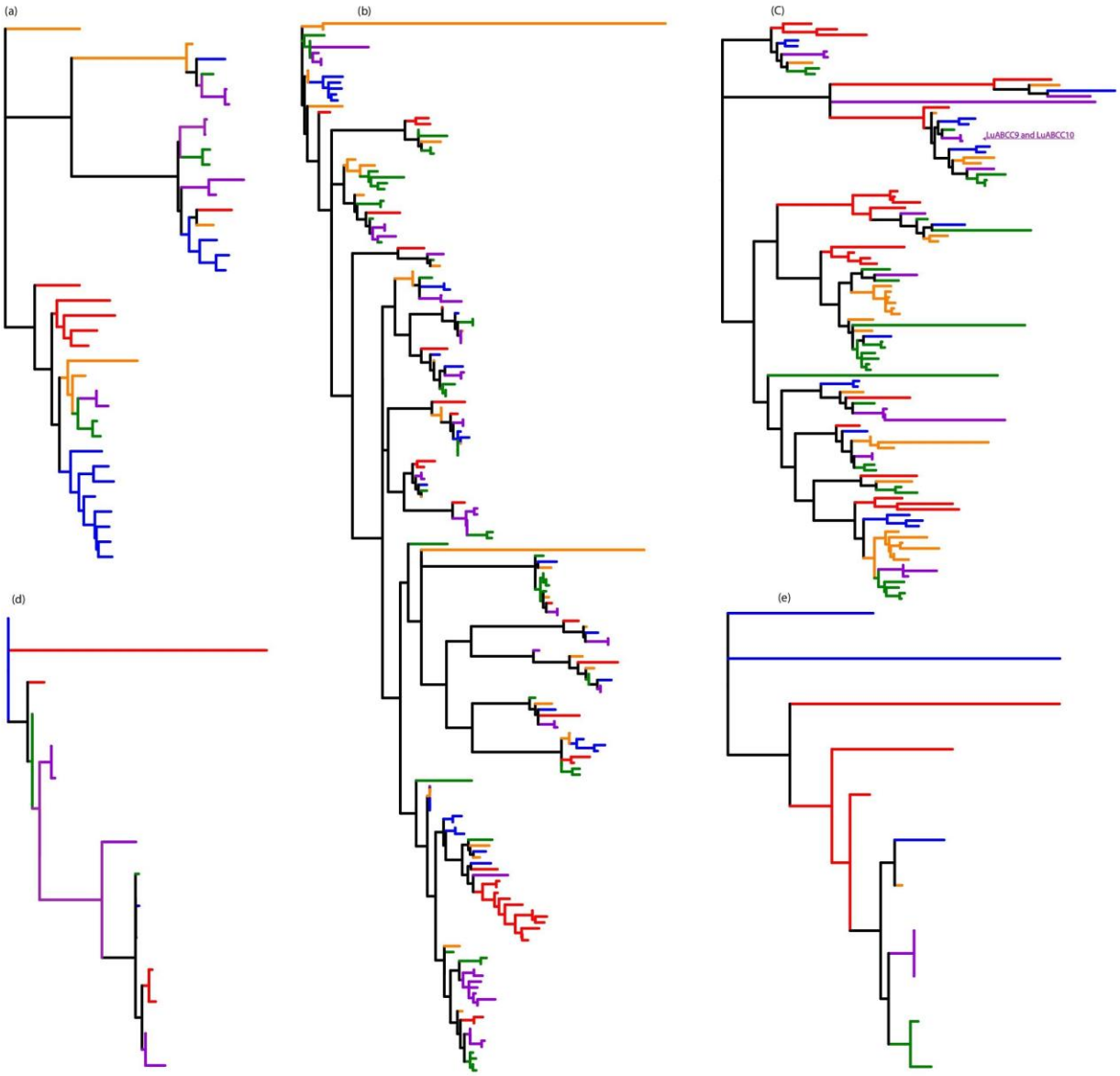
The gene annotation analysis of LuABC and LuHMA provides putative function(s) for each gene products (**Appendix 3.3**). In brief, based on the predicted function(s), most of the LuABC and LuHMA genes confirmed the presence of ABC transporter, ATP binding, and heavy metal ATPase domain functions. In addition, the LuABC subfamily were also involved in other important functions. For example, the subfamily LuABCC had multidrug resistance functions; the LuABCE encode to RNAse l inhibitor protein, whereas LuABCG were involved in regulating pleiotropic drug resistance. Other functions are also assumed to be assisted by LuABCs (**Appendix 3.3**). Thus, the annotation clearly shows the functional diversity of ABC transporters and HMAs in flax.

Unrooted phylogenetic trees were constructed using the protein sequences for each of the eight ABC transporter and four HMA gene subfamilies of *Linum usitatissimum*, *Arabidopsis thaliana*, *Populus trichocarpa*, *Vitis vinifera*, and *Brachypodium distachyon* (**Fig. 3.1** and **Appendix 3.4**). The phylogenetic relationships within LuABC, AtABC, PtABC, VvABC, and BdABC were highly conserved. Based on the phylogenetic relationships of flax and other species, the ABC transporter proteins were divided into eight subfamilies: ABCA-ABCG and ABCI (**Fig. 3.1a-h**). Subfamilies ABCB and ABCG were the largest across all species, while ABCD and ABCE were the smallest based on the number of gene members per subfamily. With 81 members, ABCG was the largest subfamily and the dominant ABC transporter gene subfamily in flax.

The HMA genes of different species were, like flax, divided into four subfamilies based on their phylogenetic relationships (**Fig. 3.1i** and **Appendix 3.1**). Subfamily IV was the largest one with 20 members, of which five belonged to flax. Subfamily I was the smallest with only six members across all the species studied. In short, the ABCG and HMA IV subfamilies had the highest number of genes in flax compared to other species except *Populus trichocarpa* in HMA. The distribution patterns of both ABC transporter and HMA genes and their subfamilies in five species are given in **Table 3.1**.

Based on previous reports on *Arabidopsis*, rice, and maize, several ABC transporter and HMA genes are associated with Cd tolerance, including *AtABCC1*, *AtABCC2*,, *AtABCG36*, *AtHMA3* and *AtHMA4* in *Arabidopsis* (Verret et al. 2004; Kim et al. 2007; Morel et al. 2009; Park et al. 2012) , *OsABCG31, OsABCG36* and *OsHMA2* in rice (Takahashi et al. 2012b; Fu et al. 2019), as well as *ZmHMA2* and *ZmHMA3* in maize (Cao et al. 2019). We identified seven ABC transporter genes (*LuABCC9-LuABCC10*, *LuABCG58-LuABCG59*, and *LuABCG71-LuABCG73)*, and two HMA

genes (*LuHMA3-LuHMA4*) which were orthologous with the above Cd-related genes (**Fig. 3.1c, g** and **h**). These genes are the most likely candidates for Cd accumulation in flax.



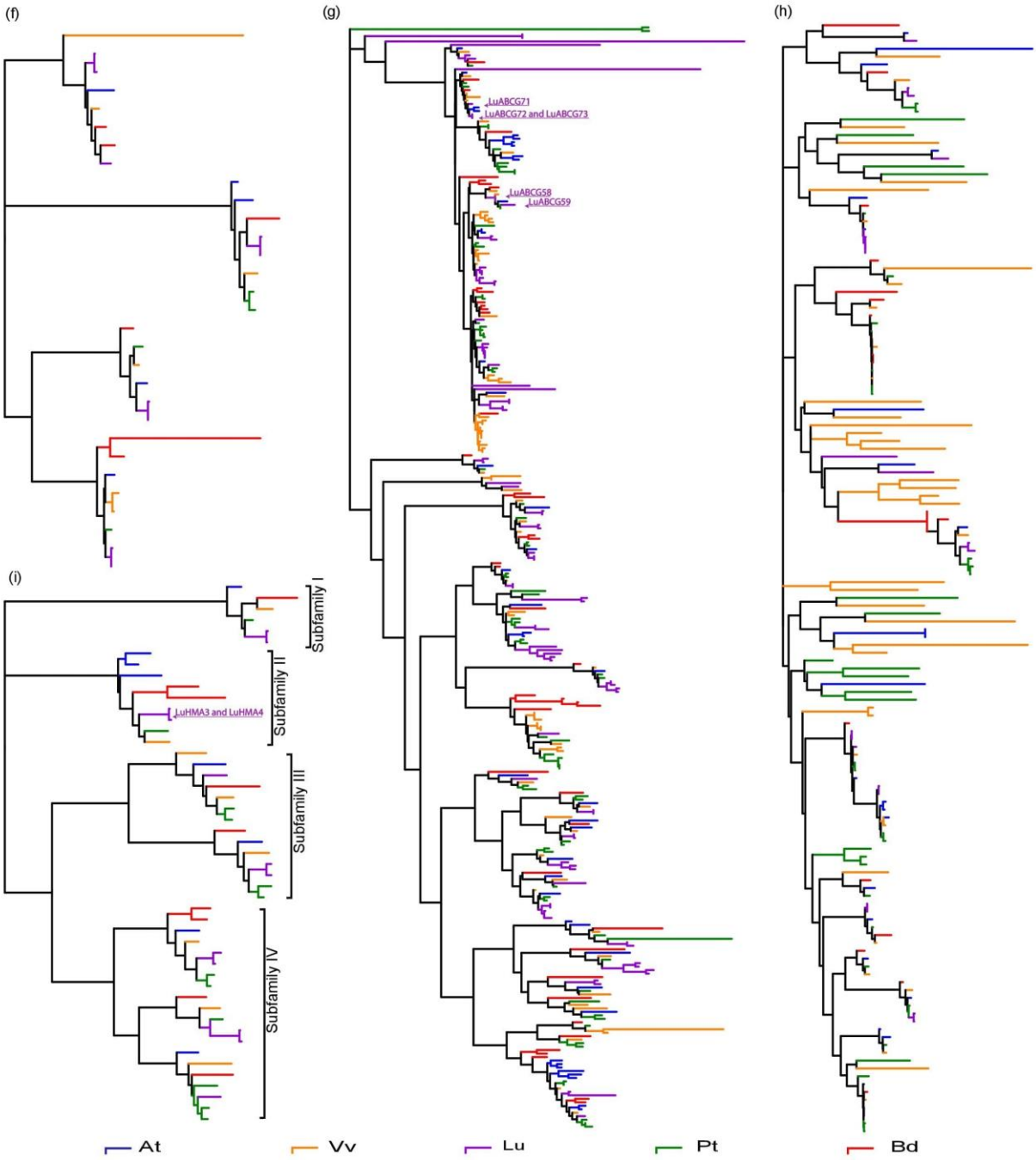


Fig. 3.1 Phylogenetic relationships of eight subfamilies of the ATP binding cassette (ABC) transporter proteins (a-h) and four subfamilies of heavy metal associated (HMA) proteins (i) in five species. *Arabidopsis thaliana* has 129 ABC transporter and 8 HMA proteins (AtABC and AtHMA), *Vitis vinifera* has 181 and 8 (VvABC and VvHMA), *Linum usitatissimum* has 198 and 12 (LuABC and LuHMA), *Populus trichocarpa* has 192 and 12 (PtABC and PtHMA), and *Brachypodium distachyon* has 133 and 9 (BdABC and LuHMA). The nine flax cadmium (Cd) candidate genes are indicated in purple (Fig. 3.1c, g and i). Note: the taxa are available in Appendix 3.3 and Appendix 3.4.

3.2.3.3 Chromosomal Localization, Syntenic Relationships, and Duplication of ABC Transporter and HMA Genes

A total of 196 LuABC and 11 LuHMA genes were located on the 15 chromosomes of flax and three of these genes (*LusABCG10*, *LusABCG15*, and *LuHMA5*) were found on scaffolds that have not been positioned onto any of the chromosomes (**Appendix 3.2**). LuABC gene subfamilies and genes *per se* are distributed unevenly across flax chromosomes. The largest number of ABC transporter and HMA genes was on Chr3 (23), followed by Chr11 (18), and Chr1 (17). The nine predicted Cd-accumulation candidate genes were scattered on multiple flax chromosomes: *LuABCG71* on Chr1, *LuABCG73* on Chr3, *LuABCG59* on Chr6, *LuABCC9* and *LuHMA3* on Chr7, *LuABCC10, LuABCG58* and *LuHMA4* on Chr12, and *LuABCG72* on Chr14 (**Fig. 3.2**). The gene collinearity analysis revealed high conservation among subfamilies of both ABC transporters and HMA with *Arabidopsis* orthologues (**Fig. 3.2**).

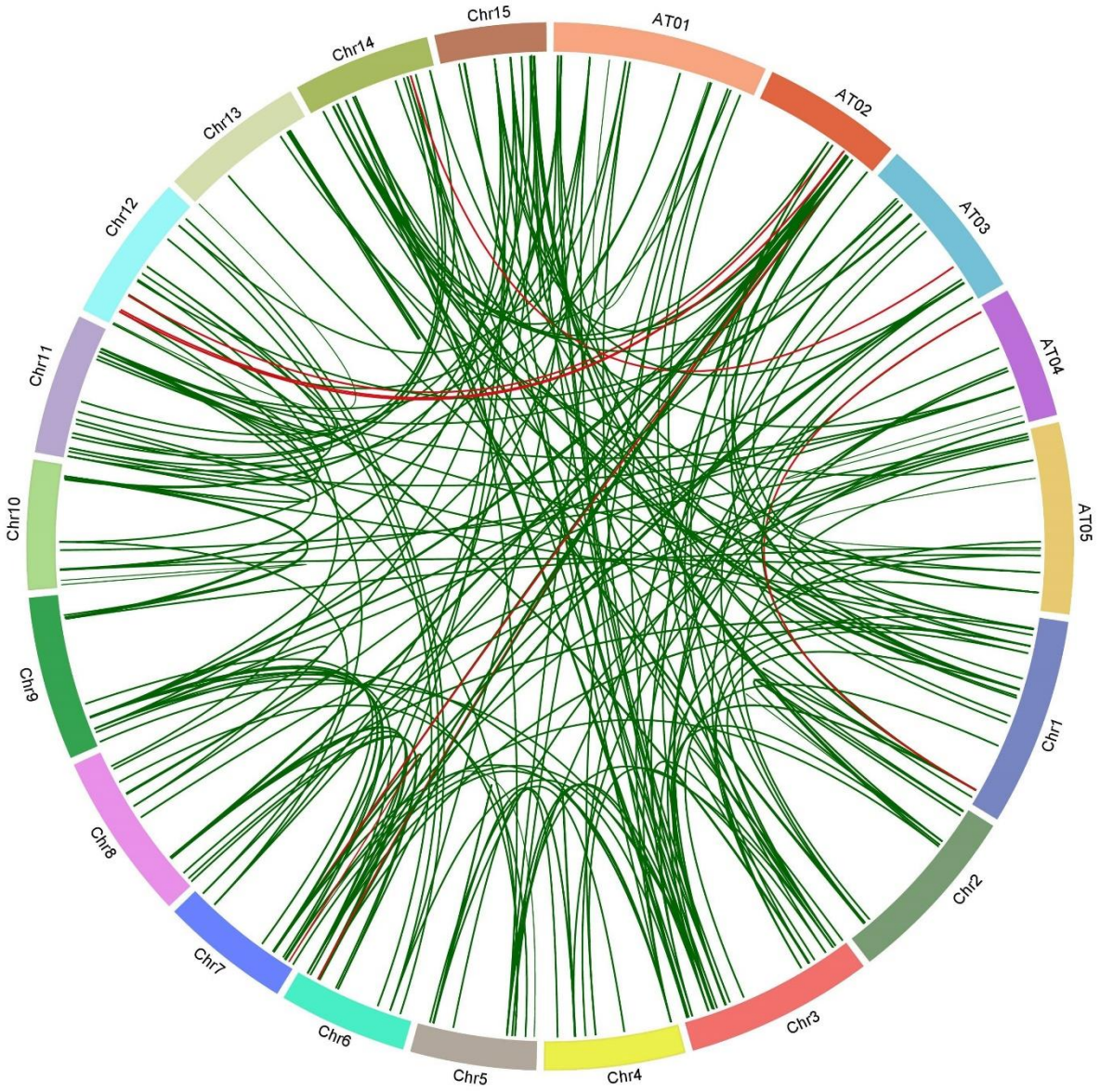


Fig. 3.2 Chromosomal locations of the orthologous ATP binding cassette (ABC) transporter and heavy metal associated (HMA) genes of flax and *Arabidopsis*. The 196 ABC transporter and 11 HMA genes on the 15 chromosomes of flax (Chr1-15) and the 129 AtABC transporter genes on the five chromosomes of *Arabidopsis* (AT01-05) are illustrated and orthologous relationships are indicated by green lines. The nine cadmium (Cd) associated genes are marked in red.

Four types of gene duplications were observed from the identified ABC transporter and HMA genes, including 162 WGD (segmental), 29 dispersed, 14 tandem, and 4 proximal duplications. Only one ABC transporter gene (*LuABCI16*) was a singleton. Eight of the nine flax Cd candidate

genes were of the segmental type and one (*LuHMA4*) was a tandem duplication. Thus, segmental duplications played a dominant role in the expansion of the ABC transporter and HMA gene families in flax and confirmed our hypothesis.

Table 3.1 The distribution patterns of ATP binding cassette (ABC) transporter and heavy metal associated (HMA) genes in five plant species.

Gene Families	Subfamilies	*Lu*	*At*	*Pt*	*Vv*	*Bd*
ABC transporter	ABCA	8	12	5	5	6
	ABCB	48	28	40	30	32
	ABCC	19	15	25	26	19
	ABCD	5	2	3	1	4
	ABCE	2	3	2	1	3
	ABCF	9	5	4	6	6
	ABCG	85	43	74	71	44
	ABCI	22	21	39	41	19
	Sub total	198	129	192	181	133
HMA	I	2	1	1	1	1
	II	2	3	1	1	2
	III	3	2	4	3	2
	IV	5	2	6	3	4
	Sub total	12	8	12	8	9
	Total	210	137	204	189	142

Lu= *Linum usitatissimum*, At= *Arabidopsis thaliana*, Pt= *Populus trichocarpa*, Vv=*Vitis vinifera*, and Bd= *Brachypodium distachyon*.

3.2.3.4 Synonymous and Non-synonymous Substitution Rates, Gene Structure Analysis and Motif Composition

The synonymous (*Ks*) and non-synonymous (*Ka*) values were estimated based on the duplicated pairs of genes across the flax genome. The *Ka*/*Ks* ratios of five pairs (*LuABCG71/LuABCG72, LuABCG61/LuABCG64, LuABCG80/LuABCG69, LuABCG4/LuABCG3,* and *LuHMA6/LuHMA8)* exceeded one, suggesting positive selection. The

remaining gene pairs underwent purifying selection with a *Ka/Ks* ratio of less than one. The estimated duplication time of LuABC and LuHMA gene pairs ranged from 1.53 to 28.27 million years ago (MYA), with an average of 8.59 MYA (**Appendix 3.5**).

Conserved motifs and gene structure organization of LuABCs and LuHMAs were analyzed to better understand the global conservation and diversification of these two gene families. A total of ten distinct conserved motifs were identified. Several motifs were highly conserved; for instance, motifs 2 and 5 commonly occurred among subfamilies LuABCA-LuABCI members as well as in HMA proteins (**Fig. 3.3a**). Of the ten motifs, motif 6 was prevalent in both ABC transporter and HMA proteins except in ABCB, ABCD, and ABCF subfamilies. Of the nine flax Cd candidate genes, three (*LuABCG71, LuABCG72,* and *LuABCG73*) consistently exhibited 9-10 of these motifs and similar gene lengths. However, distinct motif compositions existed among most of the subfamilies.

The exon structure analysis based on the coding sequences of LuABCs and LuHMAs showed diversification between and within subfamilies. Specifically, 19 ABC transporter genes, including *LuABCC7*, *LuABCC9*-11, *LuABCD1*, *LuABCF4*, *LuABCG59-62*, *LuABCG70-71*, *LuABCG75*, *LuABCG78-81*, *LuABCG88,* and *LuABCG85,* contain a variable number of exons ranging from 20-31 (**Fig. 3.3b**). The number of exons was highly conserved within ABC transporter gene subfamilies. The large number of exons observed in *LuABCA6* (38) and *LuABCA5* (40) was unique because the remaining ABC transporters and HMA had 1-19 exon(s).

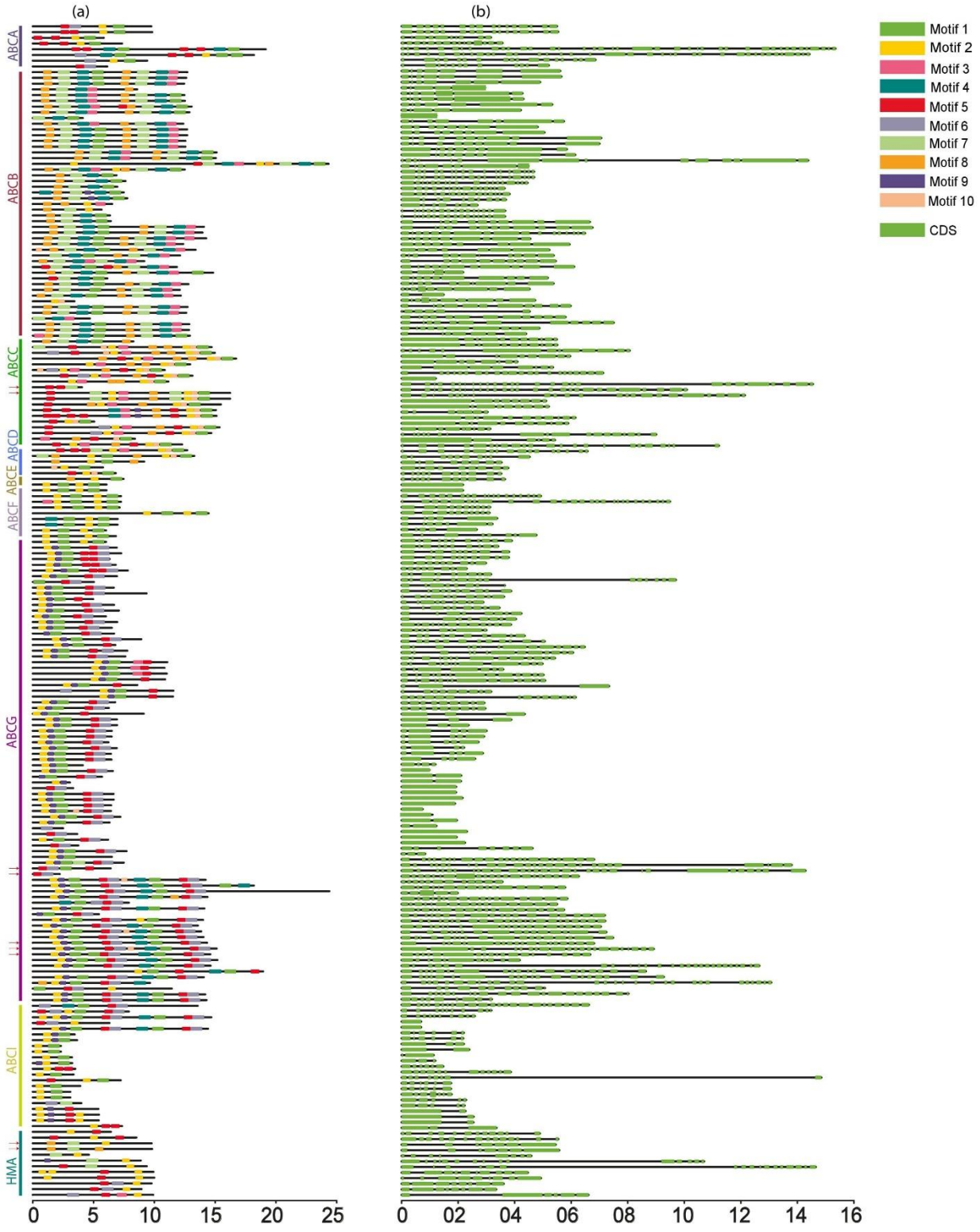


Fig. 3.3 Motif structures (a) and gene structures (b) of subfamilies LuABCA-LuABCG, LuABCI and LuHMA in flax. The nine potential cadmium (Cd) candidate genes are marked with red arrows. Motif 1-10 are displayed in different colors. The gene structure of LuABCA-LuABCG and LuABCI are based on the coding sequences (CDS) presented in green.

3.2.3.5 Gene Ontology (GO) and Expression Profiling

To predict the regulatory functions of the LuABC and LuHMA genes in flax, we performed GO analyses. The GO terms were categorized into their three subgroups: molecular function (MF), cellular component (CC), and biological process (BP), as described in **Appendix 3.6**. The LuABC and LuHMA proteins were enriched for MF such as ATP binding (GO:0005524), ATPase activity (GO:0016887), transporter activity (GO:0005215), catalytic activity (GO:0003824), kinase activity (GO:0016301), and metal ion binding (GO:0046872). CC GO terms associated with LuABC included the integral component of membrane (GO:0016021), intracellular (GO:0005622), membrane (GO:0016020), and integral component membrane (GO:0016021). BP terms comprised transport (GO:0006810), transmembrane transport (GO:0055085), signal transduction (GO:0007165), GTPase-mediated signal transduction (GO:0007264), and cation transport (GO:0006812). Taken together, GO terms indicated roles in central processes involving ATP binding and metal ion transport but also a wide range of other processes and activities in flax.

The expression patterns of the LuABC and LuHMA genes in the root, seed, ovary, and five stages of embryo development (heart, globular, torpedo, mature, and cotyledon) from RNA-Seq data are presented in a heatmap (**Fig. 3.4**). In general, the highest number of genes up-regulated for both LuABC and LuHMA genes was in seed (84/152), followed by root (75/152), and ovary (70/152) tissues. Among the five embryo development stages, both LuABC and LuHMA genes showed a relatively weak expression considering that only 47, 41, 39, 37 and 36 of the 152 expressed genes were up-regulated in mature, cotyledon, torpedo, globular, and heart stages of embryo development, respectively. The remaining LuABC and LuHMA genes were not expressed or displayed a low-level expression in these different organs (**Fig. 3.4a** and **b**). The high expression

level of LuABC and LuHMA genes in root and seed suggests that these genes might play a ubiquitous (housekeeping) transport role in these tissues in flax.

Eight of the nine flax Cd candidate genes were highly expressed in different tissues, including *LuABCG71*, *LuABCG72*, and *LuABCG73* in root and seed (**Fig. 3.4c**). Additionally, multidimensional scaling (MDS) also revealed differential gene expression between organs and high consistency of expression data between biological replicates (**Appendix 3.7**).

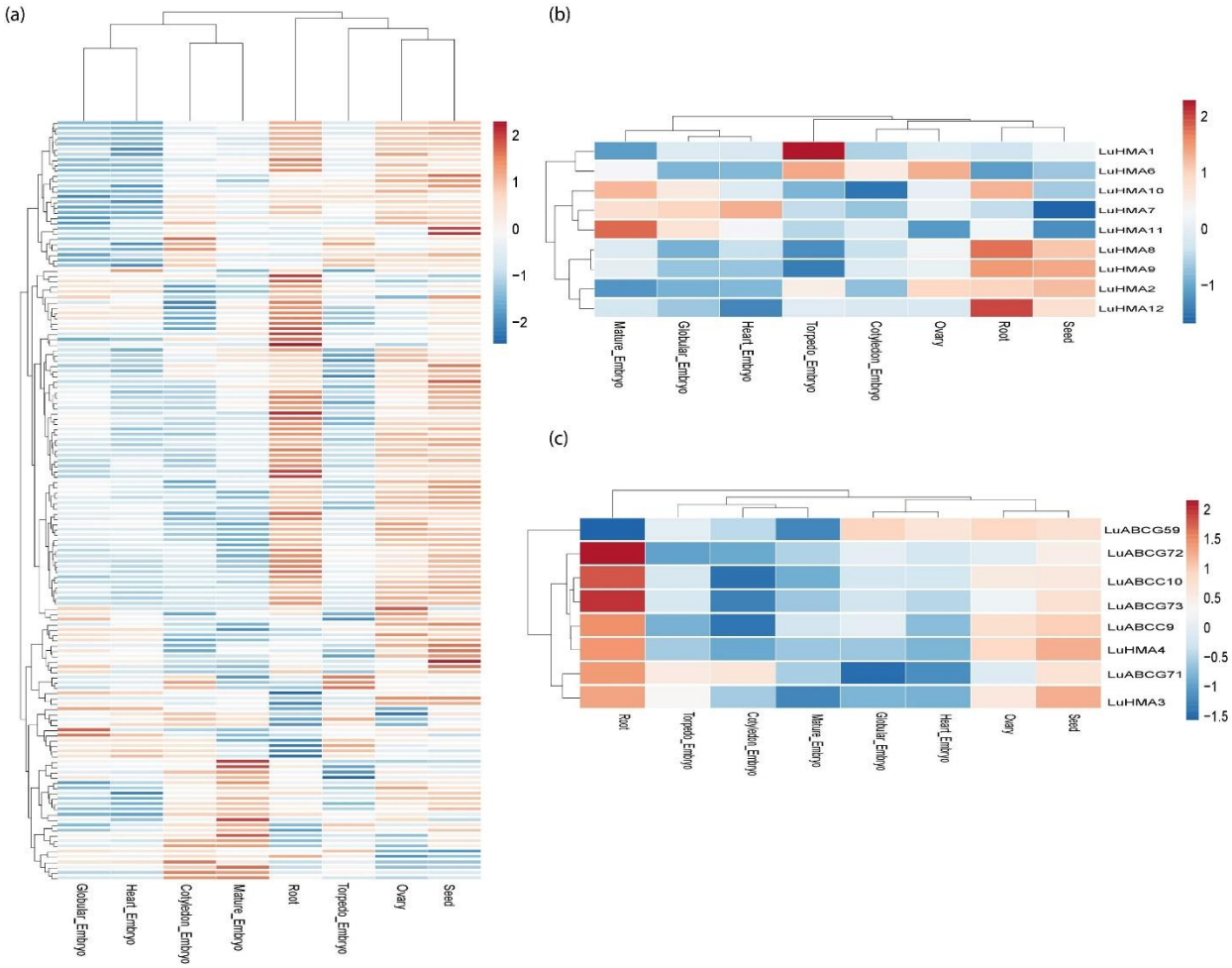


Fig. 3.4 Expression profiling of the total 160 differentially expressed genes in eight tissues based on log fold-changes, including 143 LuABC genes (a), nine LuHMA genes (b), and eight potential cadmium (Cd) candidate genes (c). The red represents up-regulated genes and blue is for down-regulated ones. The remaining 48 LuABC and one each HMA (LuHMA5) and Cd (LuABCG58) genes were discarded because they were represented by less than 5 RPM (reads per million) after normalization of the data. The anther was used as a reference for expression analysis.

3.2.3.6 Functional Evolution of Duplicated Genes and Interaction Network

Expression profiling can be mined to predict the functional fate of genes, and here, investigation of the mode and tempo of duplicated genes was performed to assess their functional evolution. We utilized and took advantage of RNA-Seq data to calculate the Pearson correlation coefficient (r) of the syntenic pairs across the eight tissues used in our research. The significance of each expression level was tested based on the r values. Positive expression correlation was inferred when r values exceeded 0.156 (at a significance level of α=0.05). Using this cut-off value, 42 of the 51 pairs had positive expression correlations, likely indicating functional conservation or sub-functionalization after duplication. The remaining nine pairs had correlation values below 0.156 or a negative correlation, suggesting putative neo-functionalization of at least one of the syntenic pairs (**Appendix 3.8**).

The protein-to-protein interaction networks of LuABC and LuHMA were examined and a dense network formed among the protein subfamilies (**Appendix 3.9a** and **b**). However, a few of the ABC proteins did not interact: LuABCG6, LuABCG66, LuABCG68, LuABCG80, and LuABCG83 (**Appendix 3.9a**). When the different subfamilies were compared, the LuABCG genes were preferentially retained in flax throughout evolution.

3.2.4 Discussion

Here, we identified 198 ABC transporter and 12 HMA genes in flax, accounting for 0.484% of the total 43,384 annotated genes of its reference sequence (Wang et al. 2012b). The observed sequence divergence and variations in the physicochemical properties of LuABC and LuHMA genes is in line with the broad diversity of their biological functions. For example, based on the variations in GRAVY and pIs values among subfamilies of LuABC and LuHMA, we speculate

that the members of these two gene families have the ability to respond to a variety of environmental cues at the micro- or macro-environment levels.

The domain composition analysis and phylogenetic tree validated the eight subfamilies LuABCA-LuABCG and LuABCI and the four subfamilies of LuHMA proteins. The results of the phylogenetic relationships between the genes were consistent with previous findings in *Arabidopsis thaliana, Brassica rapa,* and *Brassica napus* (Yan et al. 2017; Li et al. 2018c). Flax had more ABC transporter and HMA genes than any of the other five species investigated which included the dicots *Arabidopsis thaliana*, *Vitis vinifera*, *Populus trichocarpa*, and the monocot *Brachypodium distachyon*, despite having a smaller genome than all of them except *Arabidopsis*. In general, the LuABC and LuHMA genes were scattered on the phylogenetic tree, suggesting that the expansion of these gene families occurred before evolutionary divergence of the common ancestor.

Previous studies indicated that two ABCC, two ABCG, and two HMA genes in *Arabidopsis*, rice and maize are involved in Cd accumulation. The flax genes orthologous to these genes are potential candidates for Cd tolerance. Therefore, we identified nine flax Cd candidate genes, namely *LuABCC9-10*, *LuABCG58-59*, *LuABCG71-73*, and *LuHMA3-4*. We intend to validate these genes by genome-wide association study. However, the expression profiling of these genes provide us with possible functional roles. Several studies have demonstrated that ABC transporters and HMAs participate in various plants' growth activities and stress tolerance. For instance, *HvABCG31* is essential for the retention of leaf water and *ABCB1* participates in auxin transport, and its overexpression regulates hypocotyl cell elongation (Sidler et al. 1998; Noh et al. 2001; Chen et al. 2011). Similarly, two members of ABCC i.e., *OsABCC1* and *AtBCG25* participate in arsenic accumulation and abscisic acid transport (Kuromori et al. 2010; Song et al. 2014), and its

other member *AtABCG36* improves drought tolerance (Kim et al. 2010). Further, *ABCG13* has been reported to be involved in petal elongation in *Arabidopsis* (Takeda et al. 2014). Metal transporters are known to play pivotal roles in numerous aspects of plants' metabolism including essential and toxic metals' distributions (Acuña-Galindo et al. 2015). Thus, we hypothesized the possible functions of LuABC transporters and LuHMAs by examining their gene annotation, gene ontology, and gene expression in multiple tissues. Taken together, LuABCs and LuHMAs seem to play particular roles in ATP binding, transport and, metal ion binding activities. The gene expression results also revealed that the majority of the genes were highly expressed in one or more of the eight tissues evaluated, thereby confirming tissue-specific expression. Of the nine Cd gene candidates proposed, the four (*LuABCG71-73* and *LuABCC10*) up-regulated in root and seed tissues had conserved gene structure, suggesting their putative redundant functions in these developmental tissues in flax.

Protein sequence analyses of gene families are needed to understand neo-functionalization and divergence (Gu et al. 2013). Most angiosperms have undergone at least two WGD events (Jiao et al. 2011) which are frequently associated with significant evolutionary switches that can contribute to the adaptability of species to a range of environments (Clark and Donoghue 2018). WGDs are associated with the development of distinct plant species, and gene duplications are a vital force in genomic evolution and functional divergence (Segraves 2017). Similarly, in evolutionary history, most higher plants underwent polyploidization, a vital event in shaping plant genome (Moghe and Shiu 2014). Not surprisingly, flax has also undergone a palaeopolyploidization (23-44 MYA) and a mesopolyploidization (3.7-9 MYA) events (Wang et al. 2012b; You et al. 2018b). Our findings suggest an average estimated duplication divergence time of 8.59 MYA for LuABC and LuHMA genes, consistent with the most recent (3.7-9 MYA) WGD of the flax genome.

Segmental and tandem duplications are the predominant mode of expansion in *Arabidopsis* (Cannon et al. 2004). In flax, four types of gene duplications were observed in 210 ABC transporter and HMA genes. Segmental (WGD) duplications (77.14%) contributed by far the most to the expansion of the two gene families in flax, a common mode of expansion of gene families across various plant species (Die et al. 2018; Khan et al. 2019a; Shazadee et al. 2019). The selection pressure analysis of duplicated gene pairs based on three categories (i.e., purifying, positive, and neutral selection) tends to provide valuable evolutionary information (Juretic et al. 2005). *Ka/Ks* ratio values of less than, equal to or greater than one signify purifying, neutral or positive selections, respectively (Lynch and Conery 2000; Li et al. 2009b). Most of the LuABC and LuHMA gene pairs underwent purifying selection. Our findings suggest that these pairs of genes largely contributed to growth and development, while the strong positive selection in a few genes is indicative of functional differentiation. The duplicated genes can lead to one of several fates such as functional or sub-functionalization, neo-functionalization and pseudogenization (Lynch and Conery 2000). The correlation analysis based on the expression of syntenic pairs across tissues indicated only two fates: functional or sub-functionalization and neo-functionalization. Most genes of pairs likely maintain the same function, thereby showing functional conservation. However, nine pairs exhibited neo-functionalization, indicating new function(s) for the two genes of the pairs. The structural diversity mainly contributed to the evolution of the gene families as indicated by evolutionary studies (Mercereau-Puijalon et al. 2002). The observed diversity in a few LuABC and LuHMA genes may have been lost throughout evolution, and that might have contributed to their functional divergence after their loss and birth.

Taken together, these analyses suggest that the ABC transporter and HMA gene families in flax expanded over evolutionary time through gene duplication events. Among the nine putative

Cd responsive genes, we observed consistencies of several properties for *LuABCG71* and *LUABCG72*. Both showed conserved gene structure with similar number of introns (20 and 16, respectively). The gene duplication analysis for this pair indicated a segmental duplication origin and both underwent positive selection. Their expression was up-regulated in root and seed tissues, and functional conservation was revealed through their PCC (**Appendix 3.10**). Also, the remaining four pairs exhibiting positive selection (i.e., *LuABCG61/LuABCG64, LuABCG80/LuABCG69, LuABCG4/LuABCG3,* and *LuHMA6/LuHMA8*) did not show as much consistency when compared with the two Cd genes based on their expression patterns and motif structures.

In summary, the major intention of this study was to provide a comprehensive analysis of ABC and HMA genes which are associated with abiotic stress resistance, and more specifically resistance to Cd. Our intention was not only the identification of genes putatively involved in Cd accumulation in flax but also to understand their evolution. Further functional characterization is needed to validate the Cd-associated genes and to define which syntenic paralogous pairs underwent functional changes during evolution, either as a consequence of structural changes of the CDS or through expression profile changes. Both epigenetic and structural modifications in cis- or trans-elements have an impact on gene expression.

3.2.5 Conclusion

A comprehensive sequence analysis of the ABC transporter and HMA gene families in flax was performed. We identified 198 LuABC and 12 LuHMA genes that clustered into eight ABC transporter (ABCA-ABCG and ABCI) and four HMA subfamilies. Among them, nine were predicted to be potentially involved Cd accumulation *in planta* based on homology with previously characterized genes in *Arabidopsis*, rice and maize. Their phylogenetic relationships, gene

annotation, motif composition, gene structure, syntenic relationships, cis-regulatory elements, and gene ontology are reported. The gene duplication analysis suggested that four types of duplications occurred among LuABC and LuHMA genes, namely WGD or segmental, tandem, dispersed, and proximal. WGD in flax contributed the most to the expansion of LuABC and LuHMA genes. The divergence estimates indicated that recent duplications (mesopolyploidization) occurred within these two gene families. The expression data illustrated the high degree of diversification, and the evolutionary fate of syntenic gene pairs, thereby showing their functional or sub-functional conservation and neo-functionalization. Our results provide insights into the evolution and divergence of LuABC and LuHMA genes in flax. These analyses will be foundational to future investigations into the biological functions of ABC transporter and HMA genes in flax, and will be especially helpful in conjunction with a marker association study for Cd accumulation in flax.

Chapter 4

Genome-wide Association Analysis and Genomic Prediction of Heavy Metal Contents in Flax (*Linum usitatissimum* L.) Seed

Nadeem Khan[1,2], Hamna Shazadee[1,2], Bosen Jia[1,2], Oscar Molina[3], Megan House[4], Bunyamin Ta'ran[4], Axel Diederichsen[5], Frank M. You[1,*], and Sylvie Cloutier[1,2,*]

[1]Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, 960 Carling Avenue, Ottawa, Ontario K1A 0C6, Canada.
[2]Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa, ON K1N 6N5, Canada.
[3]Morden Research and Development Centre, Agriculture and Agri-Food Canada Morden, MB R6M 1Y5 Canada
[4]Department of Plant Sciences, University of Saskatchewan, 51 Campus Drive, Saskatoon, SK S7N 5A8, Canada.
[5]Plant Gene Resources of Canada, 107 Science Place, Saskatoon, S7N 0X2,

*Correspondance: frank.you@agr.gc.ca; sylviej.cloutier@agr.gc.ca

This chapter will be submitted to a peer-reviewed journal.

Authors' contribution

Conceptualization, methodology, formal analysis, investigation, writing—original draft preparation, N.K.; writing—review, editing, H.S.; ANOVA and BLUP analysis, B.J.; Field experiment and phenotyping, B.T., O.M. and M.H; Germplasm collection, conceptualization, project administration and funding acquisition, A.D.; Conceptualization, methodology, SNP data, supervision, project administration, funding acquisition, review and editing, F.M.Y and S.C.

Abstract

Heavy metals (HMs) are toxic and can negatively impact human and animal health. Flax is known to uptake soil's HMs and to translocate them to the seeds to levels that could be a health concern. To understand the genetics of HM accumulation in flax seeds, we conducted a genome-wide association study (GWAS) of the content of several metals. A mini-core collection of the base collection preserved at Plant Gene Resources of Canada consisting of 168 flax accessions was grown in the field at three locations in 2020 and one in 2021. The metal content of 24 elements was measured on the harvested seeds and significant variations were observed for the following ten: barium (Ba), cadmium (Cd), cobalt (Co), copper (Cu), iron (Fe), manganese (Mn), nickel (Ni), strontium (Sr), titanium (Ti) and zinc (Zn). Marker-trait association was performed using single- and multi-locus GWAS models with 42,959 genome-wide single nucleotide polymorphism (SNP) markers. A total of 355 non-redundant quantitative trait nucleotides (QTNs) were identified for ten of the metal contents measured. The post-GWAS Kruskal-Wallis non-parametric test yielded 108 QTNs with a significant allelic effect as follows: Ba (2), Cd (3), Co (10), Cu (1), Fe (10), Ni (5), Sr (70), and Ti (7). Of these, 24 were large effect QTNs (R^2>5%) from which 20 candidate genes were identified for 12 of them, including five belonging to the ATP-binding cassette (ABC) transporter gene family previously reported. Genomic selection (GS) models were tested to predict the genomic estimated breeding values (GEBVs) using the random genome-wide SNP dataset and HM-specific QTN-based datasets derived from the GWAS. The ten metal contents were evaluated using seven parametric and three non-parametric GS models. A joint analysis of variance of the models with both marker input datasets showed significant differences ($p < 0.05$) in predictive ability among GS models, traits and marker input datasets. Significantly superior predictive

Abstract

abilities ranging from 70-80% versus 10-25% were obtained with the trait-specific QTN datasets compared to the random genome-wide 43K SNP dataset. This QTL-based GS strategy seems effective for improving the predictive ability of metal contents, and has the potential to improve breeding efficiency compared to traditional methods requiring wet chemistry measurements. Taken together, these results represent a body of information that can be mined for GS, gene cloning and marker development, and to pyramid favourable alleles into cultivars.

Keywords: Flax ● ATP-binding cassette (ABC) transporter ● Heavy metal associated (HMA) ● Genome-wide association study (GWAS) ● Genomic selection (GS) ● Heavy metals (HMs)

Chapter 4: GWAS and GS for Metal Contents

4.1 Introduction

Heavy metals (HMs) naturally occur in soils, usually in low concentrations; they frequently come from eroded rocks and volcanic soils. However, some of the main sources of HM contamination are from mines, in both manufacturing and industrial facilities, and are frequently linked to extremely high levels of contamination in the soils (Alloway 2012). Other sources contributing to HM accumulation in agricultural soils include sewage sludge, intensive use of soil additives, livestock manures and pesticides (Peng et al. 2019). These HMs can concentrate in the aerial parts of plants, such as leaves and seeds. Plants cultivated on contaminated soils may accumulate HMs and, their presence in the edible parts could have serious effects on the health of both animals and humans consuming them (Vallee and Ulmer 1972; Tembo et al. 2006).

HMs can persist in soil for a long time and cannot be degraded by biological or physical activities. This, can cause long-term harm to the ecosystem (Suman et al. 2018). HMs can be described as essential and non-essential based on their physiological functions in the plant. Essential HMs, including copper (Cu), iron (Fe), manganese (Mn), nickel (Ni) and zinc (Zn), are required for both physiological and metabolic activities (Cempel and Nikel 2006); however, in excess amounts, they can become toxic to plants. Non-essential HMs, such as lead (Pb), cadmium (Cd), arsenic (As) and mercury (Hg), are extremely toxic to plants (Fasani et al. 2018). Given that HMs are found in numerous environmental matrices in amounts ranging from parts per billion (ppb) to less than ten parts per million (ppm), they are considered trace elements (Kabata-Pendias 2000). HMs, such as Cd, are non-essential ions that can be taken up by root transporters and then translocated into the aerial parts of plants (Clemens 2006). The presence of HMs in leaves can cause chlorosis, damage to the photosynthetic pathways and trigger the breakdown of fundamental metabolic processes. Taken together, these have such a negative impact on plant development that

it may lead to death (Clemens 2006). Several genetic loci for integrated transport and tissue detoxification that function at the cellular and molecular levels are used by plants to manage HM toxicity (Paape et al. 2022). The mechanisms of resistance and tolerance to HM accumulation in plants are therefore predicted to be polygenic. For the sake of both human health and agriculture, it is crucial to identify the physiological processes and genetic mechanisms employed by plants to prevent the accumulation of harmful HMs in the plant parts that are consumed without interfering with the desirable ones such as zinc (Zn) and iron (Fe).

Limited literature is available for HM accumulation in flax (*Linum usitatissimum* L.) with the exception of its bio-accumulation of Cd, leading to measurable Cd concentrations in seed. Flax has been reported to bio-accumulate Cd in the seeds ranging from 0.23–0.72 ppm (Booker 2019). A recent report on the physiological and developmental differences in Cd content in the flax cultivars AC Emerson, Flanders, CDC Bethune and AC McDuff revealed higher concentrations of Cd in AC Emerson compared to AC McDuff across tissues and ages, including seeds, whereas the Cd concentration in Flanders was higher than that of AC McDuff in seeds and other reproductive organs, but comparable in roots and leaves (House et al. 2020). The Cd level in the shoot tips between AC Emerson and Flanders was identical, and both were higher than in CDC Bethune and AC McDuff. These findings indicate that mechanisms of Cd accumulation may differ depending on the tissue.

In addition, a general awareness of healthy food choices and health-related benefits is increasing worldwide, and this is of primary importance for flax because flax seed (linseed) is considered a healthy food (Goyal et al. 2014; Parikh et al. 2019). Flax producers would benefit from the identification of varieties with low content of harmful metals because they would contribute to the branding of high-quality flax and its health benefits. Therefore, the identification

of genomic regions associated with HMs such as quantitative trait nucleotides (QTNs), candidate genes and estimating genomic predictive ability is necessary for understanding the genetic mechanisms for HM accumulation. Plant breeding could tackle the problem by using a combinatorial approach of genome-wide association studies (GWAS) and genomic selection (GS).

GWAS is a robust substitute for the conventional linkage-based quantitative trait locus (QTL) mapping that allows for precise QTL positioning (You et al. 2022a). It has been an effective method for identifying QTNs linked to a wide range of traits in flax (He et al. 2019b; You et al. 2022a) and other crops (Hwang et al. 2014; Fatima et al. 2020). Several GWAS analyses for HMs have been performed in plant species such as *Brassica napus* (Chen et al. 2018; Zhang et al. 2020b), *Medicago truncatula* (Paape et al. 2022), *Triticum aestivum* (Safdar et al. 2020), *Oryza sativa* (Yu et al. 2021) and *Zea mays* (Baseggio et al. 2021) but none in flax to date. For most of the HM traits in these other crop studies, candidate genes have been identified including some encoding ATP-binding cassette (ABC) transporters, heavy metal associated (HMA) proteins, magnesium transporters, calcium-binding proteins, vacuolar iron transporters and others. Both ABC transporters and HMA proteins are known for their role in HM accumulation, particularly Cd (Kim et al. 2006; Morel et al. 2009; Fu et al. 2019).

GS is a type of marker-assisted-selection (MAS) often based on a random genome-wide high-density marker set but it can alternatively be performed other types of datasets such as QTNs or QTLs associated with the target trait. GS offers a possible solution to the challenging phenotyping of complex traits, a problem that can be particularly acute when attempting to do so for a large number of populations. In standard GS methods, genome-estimated breeding values (GEBVs) are obtained from genome-wide markers; hence, their ability to predict complex traits such as drought stress and yield (Dias et al. 2018; Velazco et al. 2019). Appropriate GS methods reliably predict

non-phenotyped germplasm based on genotypic information alone, thereby improving breeding programs' efficiency, in part through decreasing the number of lines to be phenotyped in the field, and consequently reducing cost (Krchov and Bernardo 2015). The advantages of GS are more apparent when traits are polygenic with many genes contributing small amount of the genetic variance, when phenotyping is time-consuming and/or costly, or when multiple environments are needed to adequately evaluate the trait's performance. In GS, the marker effects, calculated based on models derived from a training population that is both genotyped and phenotyped, are subsequently used to predict the GEBVs of the non-phenotyped test populations (Heffner et al. 2009). Here we phenotyped a collection of 168 flax accessions grown in the field at multiple locations and over multiple years for metal contents in seed. Using a marker set of 42,959 SNPs, we performed GWAS and GS with the following objectives: (1) to exploit both single- and multi-locus GWAS models to identify QTNs associated with metal contents in seed, (2) to identify candidate genes associated with metal contents, (3) to test statistical GS models for their predictive abilities and, (4) to evaluate the potential of the tested methods and combination thereof for breeding applications.

4.2. Materials and Methods

4.2.1 Plant Material and Experimental Design

In this study, a total of 168 flax accessions (**Appendix 4.1**), including the check cultivar CDC Bethune (main plot control) and AAC Bravo (sub-plot control), and 166 test entries comprised of diverse germplasm and Canadian cultivars, were used. In year one of the project, a single plant of each accession was grown in a greenhouse and seeds of that plant were seed-increased in a

contraseason nursery in New Zealand to produce the pure-line seeds for all field trials. The accessions include a mini-core collection assembled from the flax core collection preserved by Plant Gene Resources of Canada (Diederichsen et al. 2013). All entries were pure-lined and resemble single plant off-spring ensuring complete genotypic homogeneity within the accessions, while there is a great heterogeneity/diversity among the accessions. The pure lines are preserved at Plant Gene Resources of Canada.

All accessions were grown in three environments, namely Saskatoon (Saskatchewan), Melita and Morden (Manitoba) in 2020 (non-irrigated) and Morden (Manitoba) in 2021, which was irrigated with 1.8 cm of water per week from May 28 to July 28, 2021. A type-2 modified augmented design (MAD2) (Lin and Poushinsky 1985) was used for all field trials. In the MAD2 design, test lines were not replicated and soil heterogeneity-related variation was controlled by the replications of the main plot control line CDC Bethune and the subplot control line AAC Bravo. The main plots were arranged in a 6x5 grid and each main plot was divided into seven parallel subplots where the 30 centre subplots contained the main plot control cultivar CDC Bethune. The subplot control AAC Bravo was randomly assigned to a random row location of 14 of the 30 main plots. The row length was 2 m at all locations except Melita 2020 where two 1-m row subplots were used. All locations used 0.3 m distance between rows and 2-m pathways between blocks of five main plots.

Seeding was performed during the second or third week of May at all locations and years. Phenotypic data were adjusted in accordance with the MAD pipeline as previously described (You et al. 2013). Analyses of variance (ANOVA) were performed on MAD-adjusted data, and only traits that showed significant variability ($p < 0.05$) were further analyzed.

4.2.2 Genotyping and SNP Datasets

Approximately 75 mg of young leaf tissue of each of the 168 accessions was sampled from the same single plants used for seed production and immediately flash-frozen in liquid nitrogen prior to being lyophilized until complete dryness. The DNA was extracted using the DNeasy Plant Kit (Qiagen, Toronto, ON, Canada) and quantified by Quant-iT™ PicoGreen™ dsDNA assay kit (Thermo Fisher Scientific, Burlington, ON, Canada); both procedures followed the manufacturer's instructions. The DNA was diluted to 50 ng/µl and 2 µg samples were sent to the Centre d'expertise et de services Génome Québec (Montreal, QC, Canada) for library construction and sequencing. Libraries were generated from 700 ng of genomic DNA for each of the 168 samples using the NxSeq® AmpFREE Low DNA Library Preparation Kit (Lucigen, Middleton, WI, USA) according to the manufacturer's recommendations. Dual-index adaptors were used (IDT, Coralville, IA, USA). Libraries were quantified using the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems, Wilmington, MA, USA) and average size fragment was determined using a LabChip GX (PerkinElmer, Waltham, MA, USA) instrument.

The libraries were normalized and pooled. The library pool was denatured in 0.05N NaOH and neutralized using HT1 buffer. A proprietary clustering method known as exclusion amplification was added to the mix following the manufacturer's instructions. The pool was loaded at 225 pM on an Illumina cBot and the flowcell was run on a HiSeq X using the paired-end mode for 2x150 cycles (Illumina, San Diego, CA, USA). A phiX library was used as a control and mixed with library pool at a 1% level.

The raw reads were mapped and converted to the chromosome-based flax pseudomolecules (You et al. 2018b) using BWA V.0.7.17 (Li and Durbin 2009), and variant calling was performed with BCFtools V.1.9 (Danecek and McCarthy 2017) and samtools V.1.9 (Li et al. 2009a). All steps implemented herein were as previously described (Kumar et al. 2012; You et al. 2012). The

missing SNPs (average 14.13%) were imputed using Beagle V.4.2 with default parameters (Browning et al. 2021). Filtering with minor allele frequency (MAF) > 0.05 and call rate > 60% was subsequently performed to produce a genome-wide dataset of 42,959 SNPs, called the 43K dataset, used in downstream analyses.

4.2.3 Phenotyping

Approximately 1g of clean seeds from each genotype was dispensed in 5-ml plastic tubes containing two acid-washed 8-mm zirconium beads prior to grinding in a GenoGrinder for 90 seconds at 1200 rpm. The seed content of the following 24 elements was measured by the Toxicology Centre (University of Saskatchewan, Saskatoon, Canada): boron (B), aluminium (Al), titanium (Ti), vanadium (V), chromium (Cr), manganese (Mn), iron (Fe), cobalt (Co), nickel (Ni), copper (Cu), zinc (Zn), arsenic (As), selenium (Se), strontium (Sr), molybdenum (Mo), silver (Ag), cadmium (Cd), tin (Sn), stibium (Sb), barium (Ba), mercury (Hg), thallium (Tl), lead (Pb) and uranium (U). Quantification was performed using inductively-coupled plasma-mass spectrometry (ICP-MS) as previously reported (House et al. 2020). An analysis of variance was conducted to determine the population variation for each of the 24 elements measured. Here, location (L) and genotype (G) were considered as fixed effects and year (Y) as a random effect. Only the ten elements that showed significant variability ($p < 0.05$) were further analyzed: Ba, Cd, Co, Cu, Fe, Mn, Ni, Sr, Ti and Zn. Further, for the GWAS and the GS analyses, 165/168 accessions were included because there was insufficient SNP coverage for three accessions.

In addition, days to maturity (DTM) was recorded in all field trials, which is when approximately 80% of the bolls are brown and rattling, indicating that the majority of the plants had reached physiological maturity.

4.2.4 Best Linear Unbiased Predictor (BLUP) Analysis of Metal Contents

The R package Lme4 (Bates et al. 2015) was used to estimate the BLUP values for each metal content across environments. The BLUP value was derived from the MAD-adjusted metal contents. The following R command with the mixed linear model equation was used: lmer(Trait ~ Location + (1| Genotype) + (1| Year) + (1| Genotype:Location) + (1| Genotype:Year) + (1| Genotype:Year:Location)) to determine the BLUP values and calculate the variance for each trait across locations and years where Location was treated as fixed effect and Year and Genotype were considered as random effects. The BLUPs were used for both GWAS and GS analyses.

4.2.5 Population Structure Analysis and Genome-Wide Association Studies (GWAS)

To identify the QTNs associated with metal contents, SNP-based statistical models, including single-locus (SLSMs) and multi-locus models (MLSMs), were used for GWAS. The SLSMs included GLM (Price et al. 2006) and MLM (Yu et al. 2006), and the MLSMs included FarmCPU (Liu et al. 2016) from the R package MVP (https://github.com/xiaolei-lab/rMVP) and pLARmEB, FASTmrMLM, FASTmrEMMA, ISIS EM-BLASSO and mrMLM from the R package mrMLM (https://cran.r-project.org/web/packages/mrMLM/index.html).

To control for the confounding effects of genetic relatedness, the population structure and kinship matrix were computed. The R package landscape and ecological association (LEA) were used to estimate the number of populations (K) by testing $K = 2$ to 20 using the sparse non-negative factorization (sNMF) (Frichot and François 2015). To define the appropriate number of sub-populations, the cross-validation technique described in Alexander and Lange (2011) was employed. A kinship matrix was generated using the built-in algorithm in the mrMLM package.

In addition, Pearson correlation coefficients were calculated for all the traits using the R package ggcorplot (https://cran.r-project.org/web/packages/ggcorrplot/index.html).

GWAS was performed for each model and each trait using the 43K SNPs and the MAD-adjusted BLUP values of traits. The thresholds for significant marker-trait associations were calculated for GLM, MLM and FarmCPU using a critical *p*-value (α = 0.05) subjected to the Bonferroni correction, i.e., the corrected *p*-value = 1.16×10^{-6} (0.05/42,959 SNPs). A log of odds (LOD) score of 3.0 was used for the five models implemented in the R package mrMLM (Zhang et al. 2019b). These unfiltered trait-specific QTNs were summarized and used as input into GS models. To remove potentially false-positive QTNs, the non-parametric Kruskal-Wallis allelic test ($p < 0.05$) was performed. This post-GWAS validation produced a more stringent QTN dataset where only QTNs with a significant allelic effect were retained.

4.2.6 Candidate Gene Prediction

To predict candidate genes co-localized with major QTNs, regions of 100 Kb and linkage disequilibrium (LD) in flax decayed to $r^2 = 0.10$ at a physical distance of ~100 kb (Soto-Cerda et al. 2018a; You et al. 2022a), on either side of the major QTNs ($R^2 > 5\%$) were investigated using the chromosome-scale flax reference genome assembly V.2.0 (You et al. 2018b). The generic file format (GFF3) of the annotated genes present in the QTN windows were extracted to predict gene function(s) based on their *Arabidopsis thaliana* orthologs.

The candidate gene list for metal contents was filtered for genes belonging to the ABC transporter and HMA gene families (Khan et al. 2020). In addition, key search terms associated with HM function(s), such as cation diffusion facilitator (Li et al. 2018c), vacuolar iron transporter (VIT) (Cao 2019), plant metal tolerance proteins (Liu et al. 2019d), ZRT and IRT like protein

(ZIP) (Shuting et al. 2022), natural resistance associated macrophage protein (Pottier et al. 2015), were used for the identification of candidate genes.

4.2.7 Genomic Selection (GS) Analysis

A custom pipeline for GS (GSPipeline v1.0) that includes seven parametric and three non-parametric GS models implemented in the R packages RR-BLUP v4.6.1 (Endelman 2011), BGLR v1.0.9 (Pérez and de los Campos 2014), BLR v1.6 (Pérez et al. 2010), randomForest v4.7-1 (González-Recio and Forni 2011) and sommer v4.1.6 (Covarrubias-Pazaran 2016) was used to evaluate the models' accuracy through cross-validation. The models tested were RR-BLUP (Endelman 2011), GBLUP (Clark and van der Werf 2013), BRR (Heffner et al. 2009), BL (Ornella et al. 2012), Bayes A, Bayes B, Bayes C (Habier et al. 2011), RF (González-Recio and Forni 2011), SVR (Moser et al. 2009) and RKHS (Gianola et al. 2006). For each metal content and GS model, two marker datasets were evaluated: 1) the trait-defined set of unfiltered QTNs (all QTNs detected for the trait regardless of the significance of their allelic effect) for each metal that were previously identified using the above-mentioned GWAS models, and 2) the random genome-wide 43K SNP dataset. The BLUP values as described in 4.2.4 were used as phenotypic datasets for GS.

The predictive ability of the ten GS models was evaluated using a five-fold random cross-validation and a total of 50 iterations were computed. The Pearson correlation coefficients (r) between the GS-predicted GEBVs and the measured phenotypic values were calculated to represent the predictive ability of the GS models. Further, a joint ANOVA with Tukey's multiple pairwise-comparisons (HSD.test function) was performed to compare the predictive ability of the two input datasets (trait-defined QTNs vs 43K) for each model using the R agricolae package

(https://cran.r-project.org/web/packages/agricolae/index.html). Such multiple comparison was also performed for each trait to compare the models analyzed by GWAS.

4.2.8 Broad-sense Heritability (H^2) Estimation

Broad-sense heritability (H^2) was estimated for all metal content traits using the inter-environment correlation approach using the pipeline described previously by You et al. (2016b).

4.3 Results

4.3.1 Trait Correlation

Pearson correlation coefficients were calculated for the ten metal contents with significant variation based on the MAD-adjusted data, and 21 pairs were significant ($p < 0.05$). The highest trait correlations were between Zn and Cu ($r = 0.61$), and Ti and Fe ($r = 0.61$), followed by Cu and Ni ($r = 0.47$), while few of the pairs showed a negative correlation (**Fig. 4.1)**.

Additionally, Pearson correlation coefficients for all traits between all environments (year + location combinations) were used to identify significant correlations ($r > 0.2$, $p < 0.05$) between several pairs of HM and environment combinations (**Appendix 4.2)**.

The overall average and the coefficient of variation (CV) were calculated for all traits over the four site-years datasets (**Table 4.1**). The highest average content of 65.79 ± 33.23 (mg/kg) was observed for Fe and the lowest 0.93 ± 0.47 (mg/kg) for Cd. The highest CV for metal contents was: 65.52% for Sr, and the lowest 24.62% was observed for Zn.

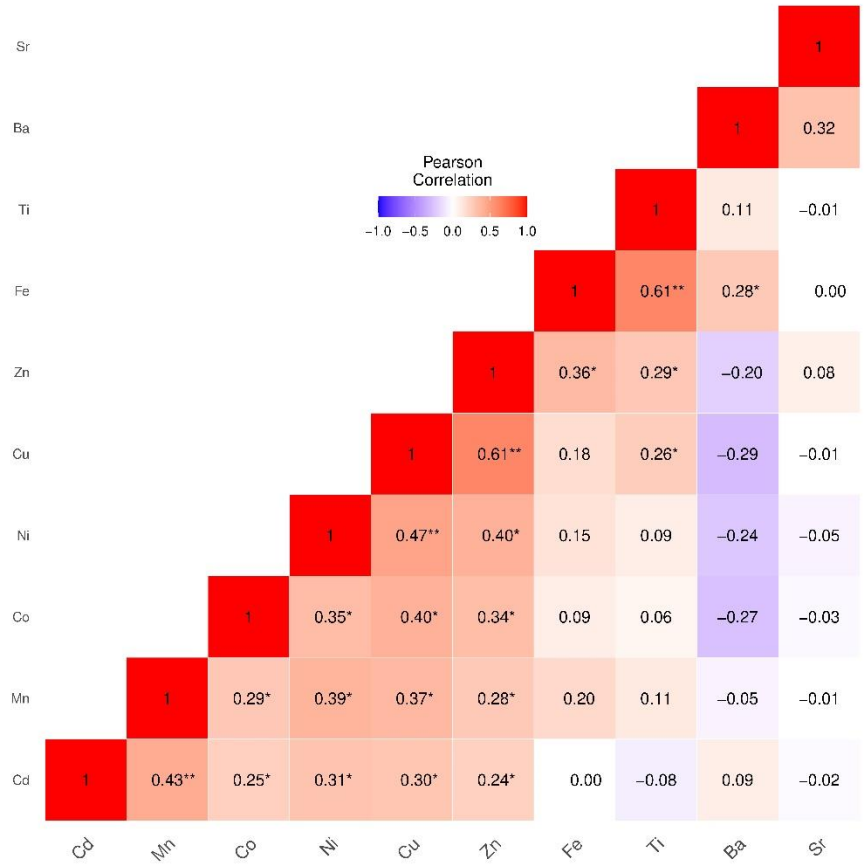


Fig. 4.1 Pearson correlation coefficients (*r*) for ten heavy metal (HM) contents and using the P-value for the level of significance i.e., 0.05 and 0.01 respectively.

Table 4.1 Metal content (mg/kg) statistics of a flax core collection grown in the field at four site-years.

Traits	Year	Sample size	Average metal content ± *s*	Max[1]	Min[2]	Median	CV (%)
Ba	2020_MD	165	1.15 ± 0.52	3.85	0.03	1.12	45.22
Ba	2020_SK	165	2.40 ± 1.24	9.16	0.17	2.20	51.67
Ba	2020_ML	165	1.11 ± 0.52	3.08	0.19	1.07	46.85
Ba	2021_MD	165	1.21 ± 0.52	3.24	0.09	1.17	42.98
Ba	Mean	165	1.47 ± 0.94	9.19	0.03	1.24	63.95
Cd	2020_MD	165	1.14 ± 0.41	2.31	0.10	1.09	35.96
Cd	2020_SK	165	1.03 ± 0.25	1.84	0.41	1.01	24.27
Cd	2020_ML	165	0.35 ± 0.12	0.69	0.10	0.34	34.29
Cd	2021_MD	165	1.22 ± 0.38	2.89	0.09	1.17	31.15
Cd	Mean	165	0.93 ± 0.47	2.89	0.09	0.94	50.54
Co	2020_MD	165	62.31 ± 17.6	110.56	6.81	61.42	28.25
Co	2020_SK	165	0.19 ± 0.06	0.44	0.06	0.18	31.58
Co	2020_ML	165	0.32 ± 0.12	0.69	0.07	0.29	37.50
Co	2021_MD	165	0.12 ± 0.04	0.30	0.01	0.11	33.33
Co	Mean	165	15.73 ± 28.31	110.56	0.01	0.23	179.97
Cu	2020_MD	165	8.91 ± 2.53	18.49	0.32	8.50	28.40
Cu	2020_SK	165	9.74 ± 1.80	14.67	6.70	9.38	18.48
Cu	2020_ML	165	11.40 ± 2.49	22.26	4.58	11.37	21.84
Cu	2021_MD	165	8.11 ± 2.04	13.89	0.77	8.19	25.15
Cu	Mean	165	9.54 ± 2.54	22.26	0.32	9.27	26.62
Fe	2020_MD	165	72.06 ± 20.21	247.60	10.55	69.22	28.05
Fe	2020_SK	165	71.61 ± 55.18	517.03	36.53	57.90	77.06
Fe	2020_ML	165	55.52 ± 10.86	112.84	32.32	54.40	19.56
Fe	2021_MD	165	63.96 ± 26.15	305.15	6.49	60.76	40.88
Fe	Mean	165	65.79 ± 33.23	517.03	6.49	60.93	50.51
Mn	2020_MD	165	20.71 ± 3.81	40.71	2.61	20.78	18.40
Mn	2020_SK	165	24.93 ± 5.92	43.53	16.11	23.47	23.75
Mn	2020_ML	165	28.62 ± 5.63	51.97	17.64	27.80	19.67
Mn	2021_MD	165	33.65 ± 9.08	90.51	3.20	33.07	26.98
Mn	Mean	165	26.98 ± 7.97	90.51	2.61	25.40	29.55
Ni	2020_MD	165	2.16 ± 0.61	3.74	0.28	2.13	28.24
Ni	2020_SK	165	2.67 ± 0.92	11.06	1.32	2.55	34.46
Ni	2020_ML	165	1.13 ± 0.90	11.35	0.49	0.99	79.65
Ni	2021_MD	165	3.28 ± 1.11	7.40	0.14	3.10	33.84
Ni	Mean	165	2.31 ± 1.19	11.35	0.14	2.22	51.52
Sr	2020_MD	165	3.30 ± 1.05	5.79	0.03	3.41	31.82
Sr	2020_SK	165	1.29 ± 0.42	2.81	0.26	1.27	32.56
Sr	2020_ML	165	1.29 ± 0.42	2.81	0.26	1.27	32.56
Sr	2021_MD	165	3.20 ± 2.75	35.96	0.48	3.04	85.94
Sr	Mean	165	3.19 ± 2.09	35.68	0.03	3.14	65.52
Ti	2020_MD	165	0.55 ± 0.22	2.81	0.02	0.52	40.00
Ti	2020_SK	165	3.77 ± 4.30	14.83	0.39	0.72	114.06
Ti	2020_ML	165	0.50 ± 0.10	0.86	0.25	0.50	20.00
Ti	2021_MD	165	0.65 ± 0.32	3.36	0.08	0.59	49.23
Ti	Mean	165	1.37 ± 2.56	14.83	0.02	0.56	186.86
Zn	2020_MD	165	53.91 ± 8.42	78.94	5.90	54.01	15.62
Zn	2020_SK	165	37.37 ± 4.70	50.46	26.32	36.80	12.58
Zn	2020_ML	165	42.60 ± 9.14	124.12	22.92	42.42	21.46
Zn	2021_MD	165	58.34 ± 10.03	84.29	6.21	58.24	17.19
Zn	Mean	165	48.05 ± 11.83	124.12	5.9	47.21	24.62

[1]Max: maximum; [2]Min: minimum; *s*: standard deviation; CV: coefficient of variation. ML represents Melita, SK; Saskatoon, and MD; Morden

4.3.2 Genetic Architecture of the Population

A total of 165 accessions were used for population structure. The population genetic structure analysis revealed the lowest cross-validation error, and *K* was determined to be six sub-populations (**Fig. 4.2a-b)**.

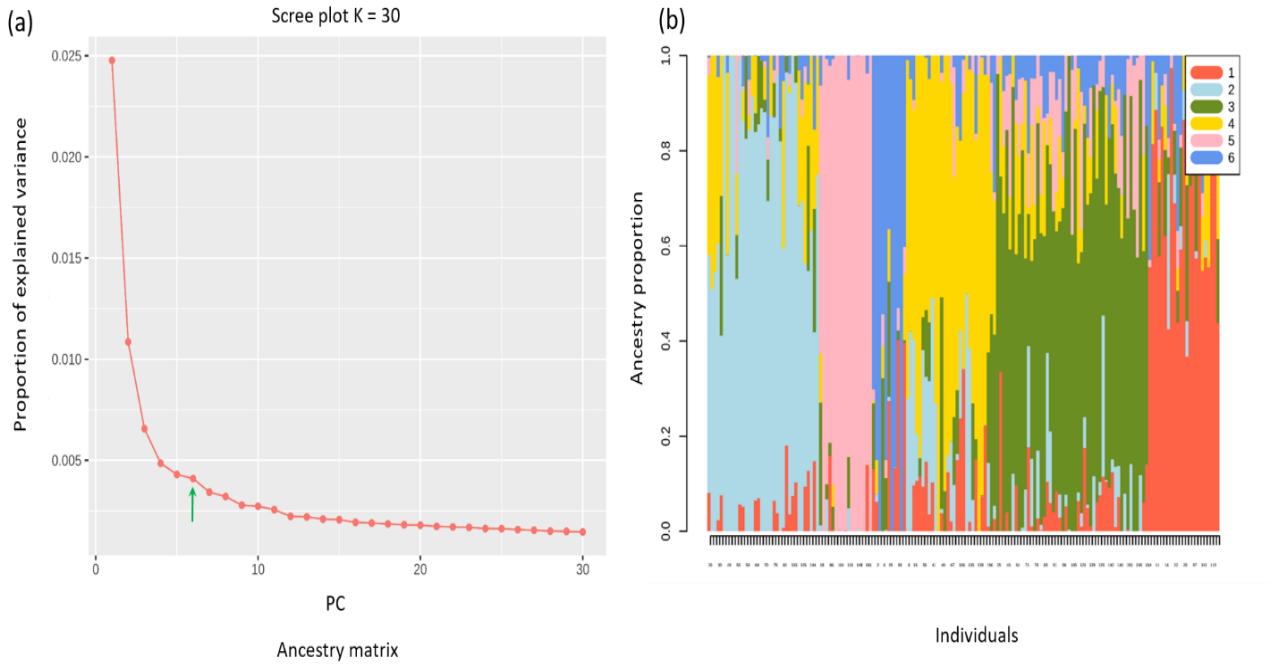


Fig. 4.2 Population genetic structure analysis of a 165-accession flax collection. (a) Plot indicating the lowest cross-validation error (CVE) suggests $K = 6$ as the number of ancestral populations; (b) a graph based on population genetic structure of $K = 6$ for all the accessions.

4.3.3 QTN Identification and Their Pleiotropic Effects

Using eight GWAS models, a total of 355 QTNs for metal content were identified (**Appendix 4.3 and 4.4**). Of these 11 were pleiotropic. The highest number of QTNs (59) were located on chromosome 1, followed by 48 on chromosome 9, and the lowest number of six on chromosome 5.

For metal content, the SLSM GLM detected the most QTNs with 115 and the fewest (10) was obtained by FarmCPU, and the average R^2 values were 1.19 ± 1.65% and 2.91 ± 3.30%, respectively (**Appendix 4.5)**.

Here, major QTNs were defined as those with $R^2 \geq 5\%$ and 24/355 met the criteria, explaining on average 6.89 ± 1.56% of the variance (**Appendix 4.6**). The joint ANOVA based on R^2 of QTNs for metal content suggested significant differences ($p < 0.05$) among GWAS models and traits (**Appendix 4.7**).

Of the 355 non-redundant QTNs, 108 had a significant allelic effect based on the Kruskal-Wallis non-parametric test (**Appendix 4.8**). These were distributed as follows: Ba (2), Cd (3), Co (10), Cu (1), Fe (10), Ni (5), Sr (70), and Ti (7). Of the 11 initial pleiotropic QTNs, only *Lu8_60629* had a significant allelic effect for two metals, i.e., Sr and Ti.

4.3.4 Candidate Genes

A total of 20 candidate genes were identified at the loci of 12 of the 24 major QTNs ($R^2 > 5\%$) for metal contents. These genes belonged to several gene families and some of the most intriguing were predicted to encode ABC transporters, magnesium transporters, calcium -dependent proteins, zinc finger transcription factors, ZIP and copper ion binding (**Table 4.2**).

In particular, the QTN *Lu3_7487534* for Cd content, which explained the highest proportion of the phenotypic variance for this trait at 7.15%, harbored the genes *Lus10004605* and *Lus10004583* that encode a magnesium transporter and a P-glycoprotein. We also discovered QTN *Lu1_1673460* (R^2=5.24%) associated with Cd that harbors *Lus10025824* which encode an ABC transporter. Of the 20 candidate genes, five (**Appendix 4.9**) encoded ABC transporters as previously described in the genome-wide study of these gene families in flax (Khan et al. 2020).

Other candidate genes included calcium-binding family proteins, multidrug resistance-associated protein, heavy metal transport, ATP binding cassette subfamily, zinc-finger domain protein and calcium-dependent protein kinase.

Table 4.2 List of candidate genes identified at major (R^2 > 5%) quantitative trait nucleotide (QTN) loci associated with heavy metal (HM) contents in flax.

Flax Gene ID	Elements	QTN (chr_position)	R^2 (%)	Functional Annotation	*Arabidopsis* ortholog	Models
Lus10025824	Cd	*Lu1_1673460*	5.24	ABC-2 and Plant PDR ABC-type transporter family protein	*AT1G59870.1*	FASTmrMLM
Lus10004605	Cd	*Lu3_7487534*	7.15	Magnesium transporter CorA-like family protein	*AT1G29820.1*	ISIS EM-BLASSO,mrMLM,FASTmrMLM
Lus10004583	Cd	*Lu3_7487534*	7.15	P-glycoprotein 9	*AT4G18050.1*	ISIS EM-BLASSO,mrMLM,FASTmrMLM
Lus10022624	Co	*Lu1_10413121*	5.20	Calcium-dependent protein kinase 23	*AT4G04740.1*	pLARmEB
Lus10022617	Co	*Lu1_10413121*	5.20	Heavy metal transport/detoxification superfamily protein	*AT2G36950.1*	pLARmEB
Lus10004896	Cu	*Lu9_9585748*	6.21	Multidrug resistance-associated protein 6	*AT3G21250.2*	mrMLM,FASTmrMLM
Lus10026301	Fe	*Lu11_4586460*	10.36	Calcium-binding EF-hand family protein	*AT5G39670.1*	FarmCPU
Lus10026291	Fe	*Lu11_4586460*	10.36	Calcium-binding EF-hand family protein	*AT4G38810.2*	FarmCPU
Lus10009389	Sr	*Lu1_11906143*	5.16	Zinc finger C-x8-C-x5-C-x3-H type family protein	*AT5G42820.1*	MLM,GLM
Lus10020624	Sr	*Lu1_12668673*	5.31	Copper ion binding	*AT2G27730.1*	pLARmEB
Lus10031306	Sr	*Lu11_19131603*	5.43	ZIP metal ion transporter family	*AT2G30080.1*	FASTmrMLM
Lus10031307	Sr	*Lu11_19131603*	5.43	ZIP metal ion transporter family	*AT2G30080.1*	FASTmrMLM
Lus10031331	Sr	*Lu11_19131603*	5.43	Stabilizer of iron transporter SufD/Polynucleotidyl transferase	*AT1G48410.1*	FASTmrMLM
Lus10000318	Sr	*Lu12_7583472*	6.06	Pleiotropic drug resistance 12	*AT1G15520.1*	MLM,GLM
Lus10025570	Sr	*Lu13_17334670*	5.40	Calcium-dependent protein kinase 24	*AT2G31500.1*	MLM,GLM
Lus10005249	Sr	*Lu13_17334670*	5.40	ATP binding cassette subfamily B19	*AT3G28860.1*	MLM,GLM
Lus10033487	Ti	*Lu3_17173441*	5.66	Zinc finger C-x8-C-x5-C-x3-H type family protein	*AT3G18640.1*	pLARmEB
Lus10033469	Ti	*Lu3_17173441*	5.66	Heavy metal transport/detoxification superfamily protein	*AT5G27690.1*	pLARmEB
Lus10033470	Ti	*Lu3_17173441*	5.66	ATP binding cassette subfamily B19	*AT3G28860.1*	pLARmEB
Lus10000845	Ti	*Lu13_4884368*	6.67	Zinc-finger domain of monoamine-oxidase A repressor R1	*AT2G23530.1*	pLARmEB

R^2: represent the proportion of phenotypic variation explained by the QTNs; Cd: cadmium; Co: cobalt; Cu: copper; Fe: iron; Sr: strontium; Ti: titanium; Zn: zinc

4.3.5 GS Models and Predictive Ability

Two marker sets were used to evaluate GS models for ten metal contents: (1) the trait-defined QTNs for each of the ten metal content and (2) the random set of 43K SNP markers. The trait-defined marker set was obtained by combining the results of all GWAS models and were compared with the random 43K SNP dataset. A joint ANOVA of predictive ability showed were found to be significant ($p < 0.05$) among traits, GS models, marker sets, and their interactions for metal contents (**Appendix 4.10**).

The two parametric models RR-BLUP and GBLUP consistently outperformed the RF and SVR models across all traits and sets of markers. The highest predictive ability was observed using the trait-defined QTNs compared to the overall set of 43K SNP dataset. The highest predictive abilities of 0.82 ± 0.17 (GBLUP) and 0.84 ± 0.13 (RR-BLUP) were observed for Ba and the lowest of 0.56 ± 0.32 (RKHS) and 0.58 ± 0.34 (RFR) were for Sr as illustrated in **Fig. 4.3**, **Appendix 4.11** and **4.12**. For the 43K SNP dataset, the predictive ability was drastically decreased for all traits; the highest being 0.22 ± 0.15 for Ba, followed by 0.20 ± 0.17 for Cd, and the lowest was -0.07 ± 0.14 for Mn (**Fig. 4.3** and **Appendix 4.11 and 4.13**).

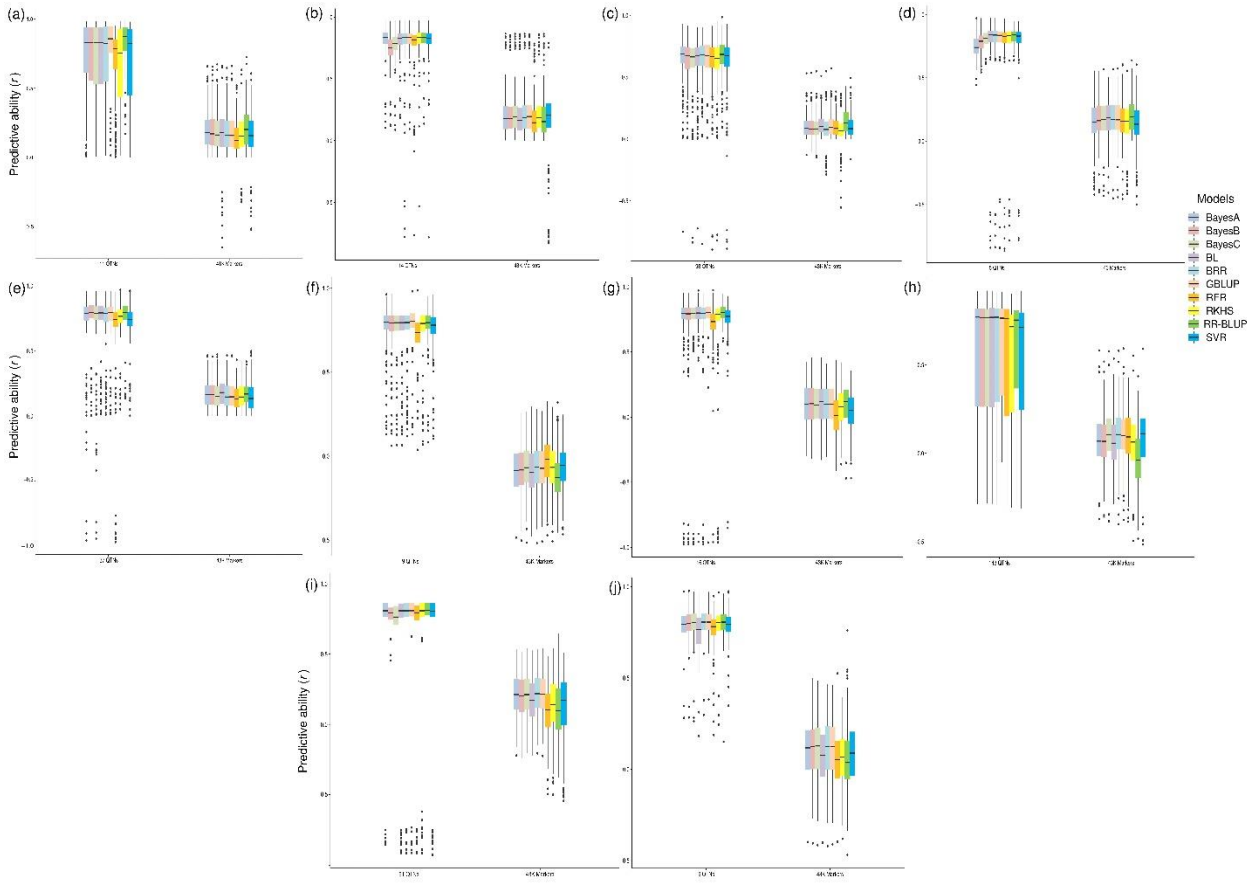


Fig. 4.3 Box plots showing the predictive ability (*r*) for ten genomic selection (GS) models for ten heavy metal (HM) contents *per se* using the trait-defined quantitative trait nucleotides (QTNs) and the random 43K single nucleotide polymorphism (SNP) datasets. (a) 11 Ba; (b) 14 Cd; (c) 93 Co; (d) 5 Cu; (e) 37 Fe; (f) 9 MN; (g) 16 Ni; (h) 148 Sr; (i) 51 Ti; and (j) 6 Zn. Each panel displays the *r* values obtained from the ten GS models comparing the trait-defined QTN datasets to the random 43K SNP dataset. The number of QTNs used in the models are on the x-axis. Five-fold cross-validation was used to estimate the *r* values.

4.3.6 Heritability and its Relationship with the Relative Efficiency of Metal Contents

Broad-sense heritabilities (H^2) of metal contents were estimated (**Table 4.3**). Most of metal contents had intermediate H^2 estimates with a maximum of 0.61 for Ba but some had low heritability such as Ti at 0.04. The average predictive ability ranged from 0.33 (Sr) to 0.47 (Cu and Ti). Notably, Ti was observed to have the lowest H^2 of 0.04 but the average GS model for both sets of data generated the highest predictive ability of 0.47.

Table 4.3 Broad-sense heritability (H^2), average predictive ability (r) from both datasets and ten GS models, and their relative efficiency of ten heavy metal (HM) contents.

Elements	Broad-sense heritability (H^2)	Predictive ability (r)	Relative efficiency (r/H^2)
Ba	0.61	0.45 ± 0.20	0.74
Cd	0.46	0.47 ± 0.25	1.02
Co	0.47	0.37 ± 0.14	0.79
Cu	0.36	0.42 ± 0.17	1.17
Fe	0.15	0.45 ± 0.15	3.00
Mn	0.15	0.35 ± 0.13	2.33
Ni	0.25	0.42 ± 0.18	1.68
Sr	0.42	0.33 ± 0.24	0.79
Ti	0.04	0.47 ± 0.21	11.75
Zn	0.29	0.44 ± 0.12	1.52

Ba: barium; Cd: cadmium; Co: cobalt; Cu: copper; Fe: iron; Mn: manganese; Ni: nickel; Ti: titanium; Sr: strontium; Zn: zinc

However, taking into account the relative efficiency (RE) of genomic selection over phenotypic selection, defined as r/H^2 (Dekkers 2007), the low H^2 traits resulted in high RE values and, to a large extent, displayed an inverse relationship. Hence, based on RE performance for GS, traits such as Cd, Cu, Fe, Mn, Ni, Ti and Zn outperformed phenotypic selection when compared to other traits.

4.3.7 Relationship between Days to Maturity (DTM) and Cadmium (Cd) content in flax

A scatter plot was constructed to illustrate the relationship between cadmium (Cd) content and days to maturity (DTM) **(Fig. 4.4)**. The results were found to be highly significant and negatively correlated ($p < 0.05$ and $r = -0.51$) and the longer the DTM the lower the Cd content. For instance, the Cd concentration exceeded 0.77 for plants maturing at 72 days, but dropped to 0.20 after 98 DTM using the equation: $y = -19x+90$. Thus, these results indicate that early-maturing germplasm seem to accumulate Cd in the seed to a greater extent.

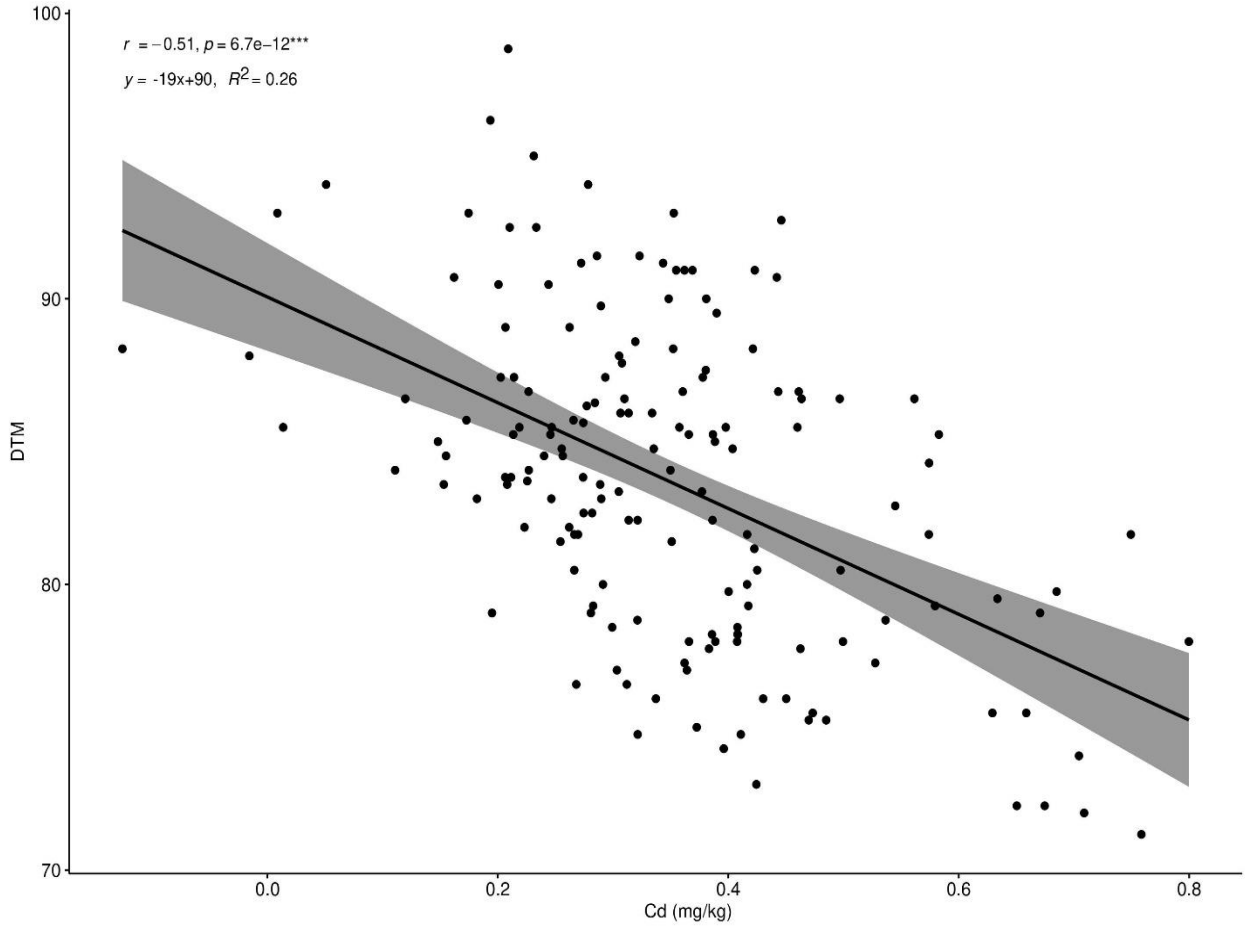


Fig. 4.4 Scatter plot showing the relationship between days to maturity (DTM) and cadmium (Cd) content in flax seeds for 165 genotypes of flax. The BLUP value of the Cd content (x-axis) and the average value of DTM across years and locations (y-axis) were used.

4.4 Discussion

This is the first study using both GWAS and GS to investigate genomics of heavy metal contents in flax seeds. Here, we took advantage of a sizable genetic panel grown in multiple environments to phenotype ten metal contents and genotyped with 42,959 SNPs to explore genomic solutions for improvement of these traits in flax breeding programs. In brief, we employed two single- and six multi-locus GWAS models to identify both large- and small-effect QTNs, resulting in a total of 355 non-redundant QTNs and the 20 candidate genes identified at 12 of the 24 major QTN loci. This strategy successfully leveraged the strengths of each models (He

et al. 2019a; Lan et al. 2020). Of the 355 QTNs associated with metal accumulation in flax seed, only 24 were major and the total R^2 across the ten metal contents was only 6.89 ± 1.56%, indicating that accumulation of these ten elements is polygenic in nature. However, the post-GWAS validation of the QTNs using an allelic test yielded only 108 non-redundant QTNs, of which only one was pleiotropic; hence not providing strong indication of shared pathways across the different elements. GWAS is a reliable method to identify QTNs and its efficiency depends on numbers of factors: (1) the number of genome-wide markers, (2) the size of the genetic panels, and (3) the relationship between QTNs and the traits or their effects all contribute to the reliability of the approach (You et al. 2022a). The ten metal contents were analyzed using eight GWAS models (**Appendix 4.6**), and the most QTNs were identified for Sr (133), followed by Co (74), Ti (54), Fe (33), Ni (15), Cd (14), Ba (11), Mn (9), and Zn (7). However, many explained very little of the variance for the trait and the post-GWAS allelic test removed nearly 70% of the QTNs: Sr (70), Co (10), Fe (10), Ti (7), Ni (5), Cd (3), Ba (2) and Cu (1) as given in **Appendix 4.8**. The usefulness of GWAS for metal content in flax is three-fold: 1) molecular markers can be designed specifically for the 24 large effect QTNs and applied in marker-assisted selection, 2) both small and large effect QTNs can be used as trait-defined QTN datasets in GS models for predicting the GEBVs, and 3) the predicted candidate genes can be functionally characterized using genome editing, targeting induced local lesions in genomes (TILLING) or other similar approaches.

Candidate gene search is a separate method and it relies on the linkage of the QTNs and the annotated genes from a reference genome. Annotation of genes located in the vicinity of QTNs is key to identifying candidate genes. Based on linkage disequilibrium (LD) decay, an ideal window size can be established (Remington et al. 2001; Sun et al. 2016). For this study, we used a 200-Kb window (100 Kb on either side of the QTNs) based on previously reported studies in flax (Kumar

et al. 2015; He et al. 2019b). For instance, in flax the LD was previously calculated with r^2 of 0.1 and physical distances between two SNPs =100kb (Sertse et al. 2021). Based on proximity and annotation, 20 candidate genes putatively associated with HM accumulation in seeds were identified, of which five were ABC transporter genes including some previously predicted based on their orthology with functionally-characterized *Arabidopsis* genes (Khan et al. 2020). Three types of transporters are known to be involved in the sequestration of Cd into plant vacuoles: HMA, ABC and CAX-type antiporter. Here, we discovered the QTNs *Lu3_7487534* and *Lu1_1673460* (R^2= 7.15 and 5.24%) associated with Cd was identified within genes *Lus10004605*, *Lus10004583* and *Lus10025824*, that encode magnesium transporter, P-glycoprotein and ABC transporter protein. The accumulation of Cd in tobacco leaves was enhanced by heterologous expression of *AtCAX2* and *AtCAX4* (Korenkov et al. 2007) and decreased by simply expressing *AtCAX4* in the roots (Korenkov et al. 2009). Our study did not identify any loci harboring CAX-type antiporter.

Several studies have shown that *HMA3* orthologs sequester HMs into the vacuole of a wide range of plant species; however, the metal selectivity and their tissue-specific function with regards to Cd accumulation tend to vary. The role of HMA transporters has been shown for Co, Pb and Zn in *O. sativa* and *A. thaliana* (Chao et al. 2012; Takahashi et al. 2012a). Ueno et al. (2010) identified a single *HMA* gene (*OsHMA3*) responsible for Cd accumulation in rice. This gene encodes a transporter that permits its translocation from root to shoot via the xylem. Deleterious mutations of this gene do not prevent root Cd uptake but result in sequestration in the root, hence reducing its accumulation in the above-ground plant parts, including the seeds (Ueno et al. 2011). One could simplistically assume that finding the flax orthologs of the above would identify the most likely candidate genes responsible for Cd accumulation in flax.

GS has been incorporated into many breeding programs to accelerate the selection process and the rate of genetic gain for complex traits such as agronomic traits and drought resistance (Shikha et al. 2017; Ben Hassen et al. 2018; Velazco et al. 2019; Lan et al. 2020). The implementation of GS as a breeding tool is limited by the cost of sequencing and the cost and difficulties of phenotyping training populations for complex traits such as metal contents in seeds. To date, applications of GS to large test (breeding) populations have mostly relied on the production of low-cost high-throughput genotyping data and the ability to timely analyze it in time for use in selection. For practical breeding purposes, determining a balance between predictive ability and effective marker density is critical to the successful implementation of GS. This study investigated the impact of a sequential approach that relies first on applying a suite of GWAS models to identify trait-specific QTNs that were subsequently used as an input dataset for GS. This method was compared to the more traditional method which uses a genome-wide random SNP dataset for GS implementation. The trait-defined marker-based GS models achieved higher genomic predictive abilities compared to the 43K SNP dataset. To date, GS models have been developed based on the use of genome-wide random markers where the marker density was deemed to be an important factor for predictive ability (Meuwissen et al. 2001). With the development of genotyping techniques, it is relatively easy to generate high-density marker datasets; nevertheless more is not necessarily better, because the vast majority of the markers are unrelated to the traits, which cause background noise and reduce predictive ability (He et al. 2019a; Lan et al. 2020). For this reason, the impact of the marker set was evaluated for their predictive ability of ten metal contents. High genomic predictive ability was obtained for all traits when the trait-defined QTNs were used for GS model construction (Lan et al. 2020). Notably, the predictive ability of GS was significantly reduced from 70-80% to 10-25% for all the traits when using the 43K SNP dataset. The QTN

marker sets are not only more predictive but also considerably smaller, consequently simplifying and speeding up the downstream computation of GS models; however, this is not a net gain because additional computations are required for the preliminary GWAS implementation.

The usefulness of GEBVs derived from parametric and non-parametric based GS models has been determined using multiple approaches (Hayes et al. 2009; Heslot et al. 2012; Desta and Ortiz 2014). This is the first study to identify QTNs associated with metal contents in flax using these two methods. In our analysis, we consistently found that parametric models produced higher prediction accuracies for all traits. These differences were further confirmed and validated by the joint ANOVA, which revealed significant differences ($p < 0.05$) between GS models, traits and marker sets. Several studies have also identified significant predictive ability differences between GS models (Technow et al. 2012; Spindel et al. 2015). In general, no single GS model consistently outperformed others across all traits. It therefore remains imperative to test several statistical models for each trait to select the best-performing one. This study illustrates the potential of GS as an early generation selection method to facilitate the improvement of metal content in flax. Additionally, our findings suggest that GS can be used as a substitute for phenotypic selection in order to forecast an individual's GEBVs for metal contents. GS uses all markers including those with minor effects allowing breeders to obtain GEBVs and to carry out parent selection for breeding purposes. This is important for quantitative traits like metal contents, which are influenced by many genes of minor effects.

Trait heritability is also important in phenotypic selection. The heritability of a trait is a factor that can significantly influence the performance of GS, as it does in traditional phenotypic selection, because it is easier to achieve high predictive ability for high-heritability traits (Ali et al. 2020). However, GS may be better at predicting low-heritability traits (Goddard 2009; Li et al.

2015). GS's advantage over typical phenotypic selection lies in the fact that a trait with a low heritability can have a higher relative efficiency (RE). For instance, all metal contents with lower H^2 had higher RE. Similar trends were also observed in flax for agronomic traits (Lan et al. 2020). In brief, yield with H^2 of 0.2 had as high as 4.45 times selection efficiency over phenotypic selection, demonstrating the benefits of GS for low heritability traits. Thus, based on RE performance, GS for traits such as Cd, Cu, Fe, Mn, Ni, Ti and Zn outperformed phenotypic selection.

Finally, we observed negative correlation between DTM and Cd where early maturing accessions had higher Cd content than late maturing ones (**Fig. 4.4**). The findings of our study were corroborated by a previous study (Gutsch et al. 2018). In *Medicago sativa*, plants grown on high-Cd soil were stunted in the early developmental stages but the plants appeared to recover at later stage because no differences in biomass were observed between the plant grown in high-Cd and low-Cd soils.

In summary, accurate genotyping and phenotyping remain critical to successful GWAS of metal contents in flax because they contribute to reducing the number of false-positive and false-negative associations. Because our GS strategy proposes to build the models on a GWAS-derived QTN dataset, any factor that contributes to the quality of the QTN dataset(s) will likely translate to superior predictive ability of the GS models. The strategy outlined herein, which is based on the sequential use of GWAS and GS, is efficient and powerful even for complex traits such as metal contents. Trait-defined QTNs can potentially reduce genotyping costs through the use of a QTN-defined genotyping platform but such platforms also have their drawbacks (e.g. low flexibility) and are not always more affordable. The pros and cons remain to be evaluated carefully. Regardless of the genotyping method, the advantage of the simplification of the computation

associated with the reduced number of markers can tip the scale towards adoption of GS versus the *status quo* of phenotypic selection. Phenomics approaches to phenotyping are promising for cost reduction but such methodologies remain at the developmental stages for many important traits and crops. Along with the constant reduction in sequencing costs *per se*, these additional advantages make GS a viable option for many breeding programs and, we anticipate that this approach will be increasingly adopted in the next few years.

4.5 Conclusion

In this study, a total of 355 non-redundant QTNs for metal contents, and 20 candidate genes were identified. A set of 42,959 high-density markers and the phenotypic data of ten metal contents were used to identify QTNs associated with them, which in turn were used to optimize the genomic predictive ability of GS models; all with the optics of providing a superior approach to breeding for these traits. The sequential GWAS-GS approach illustrated the effectiveness of the trait-defined QTN datasets to predict metal contents in flax. Collectively, significant differences in predictive ability ($p < 0.05$) were observed between GS models, traits, marker sets and their interactions. We propose the use of trait-defined QTN datasets for the practical implementation of GS in flax for their power, cost reduction and computational simplicity. Such an approach has successfully been demonstrated in flax for the complex pasmo disease (He et al. 2019a) and for several agronomic traits (Lan et al. 2020), and now we extended it to metal contents in seeds.

5: General Discussion and Conclusion

5.1 General Discussion and Conclusion

5.1.1 Major Highlights of the Thesis

This thesis capitalized on the recently published version 2.0 of the flax reference genome and its annotation (You et al. 2018b) to identify novel genomic regions and candidate genes associated with abiotic stress conditions, with an emphasis on heavy metals (HM), particularly cadmium (Cd) which can negatively affect plant development. This improved reference genome sequence has provided a foundation for comparing and validating newly discovered quantitative trait nucleotides (QTNs) associated with metal content in flax seeds and was invaluable in performing candidate gene searches. This thesis, which focused on (1) genomic cross prediction, (2) gene family identification, (3) genome-wide association study (GWAS) and (4) genomic selection (GS), aims to propose genomic solutions for the improvement of flax breeding programs.

Genomic cross prediction in flax can be highly effective for two major reasons: i) a large number of virtual crosses can be easily simulated, ii) the use of GWAS for the identification of QTNs and that of optimized GS models can advantageously predict the genomic estimated breeding values (GEBVs) of progenies. I suggest that an integrated approach that comprises GS, genomic cross prediction, and conventional phenotyping has potential for improving flax breeding program efficiencies.

Based on a literature search, 521 QTNs/QTLs associated with abiotic stress related traits have been reported. These are linked and not causal. They are also limited to the germplasm and the trait studied. Precision of the trait measurement, environmental effect, significant genotype by environment, nature and size of the genotypic dataset, statistical models and many other factors impact the discovery of QTNs and QTLs; hence, these are not exhaustive but these QTN/QTL

datasets constitute a freely available and important input dataset for GS. To increase the precision and efficacy of selection, lower the costs and shorten the breeding cycles, these QTNs can serve as marker input for GS to predict breeding values of test populations. If used in combination with genomic cross prediction, this GS strategy presents a novel and efficient method for predicting the breeding performance of crosses and for selecting parents based on their genotype. Several case studies presented in this thesis demonstrated that GS predictive ability can be improved through trait-defined QTN marker set. Current findings revealed that high-density genome-wide random markers may not necessarily produce high genomic predictive ability in GS because the vast majority of the markers are unrelated to the traits of interest, causing background noise and lowering the predictive ability. Alternatively, trait-defined QTNs, identified through single- and multi-locus GWAS methods, can significantly improve the genomic predictive ability of GS models, reduce the computational time and possibly reduce the genotyping cost of the test populations.

Both GWAS and genome-wide identification of the ATP-binding cassette (ABC) transporters and heavy metal associated (HMA) gene families are integral in identifying candidate genes for traits such as metal content, and more specifically here for cadmium (Cd) content in seeds which was a main objective of this thesis. ABC transporter and HMA genes had previously been characterized for their function in Cd accumulation in other plant species, and here they were identified as candidate genes in flax for the first time. A total of 198 ABC transporter and 12 HMA genes were identified in the flax genome, nine of them, i.e., *LuABCC9*, *LuABCC10*, *LuABCG58*, *LuABCG59*, *LuABCG71*, *LuABCG72*, *LuABCG73*, *LuHMA3* and *LuHMA4*, were proposed as the most likely candidate genes responsible for Cd accumulation in seeds based on their homology and phylogenetic relationships with the *Arabidopsis*, rice and maize orthologs.

GWAS based on single- and multi-locus models were used to identify QTNs associated with ten metal contents. The metal contents included were barium (Ba), cadmium (Cd), cobalt (Co), copper (Cu), iron (Fe), manganese (Mn), nickel (Ni), strontium (Sr), titanium (Ti) and zinc (Zn). A total of 355 non-redundant QTNs were identified for these metal contents, of which 108 had a significant allelic effect. A total of 20 candidate genes associated with metal content were identified, among which five were ABC and HMA transporters reported in the gene family study (chapter 3). Some of the 24 large (R^2 > 5%) and one pleiotropic effect QTNs can be pyramided using marker-assisted selection to develop low-HM flax. The parametric GS models RR-BLUP and GBLUP outperformed the non-parametric models RF and SVR across traits and marker sets. The use of a trait-defined input dataset improved the predictive ability of GS for the complex seed metal contents evaluated.

5.2 Limitation and Suggested Future Studies

This study has provided a comprehensive overview of four major breeding strategies, including genomic cross prediction, GS, GWAS and gene family identification of ABC transporter and HMA genes, offering a wide range of genomic solutions for the improvement of flax breeding programs. In the “gene family identification” strategy, the solution is specific to HM accumulation in plants where the outcome of the genome-wide analyses of ABC transporter and HMA genes was the identification of the most likely gene candidates which was based on their homology to orthologs from other species that had previously been functionally-characterized for their role(s) in HM accumulation. Keeping in mind this four-prong strategy, it is important to also stress the research’s limitations and the benefits that would arise from additional validations in some cases.

5.2.1 Genomic Cross Prediction

A breeding program's principal goal is to create superior cultivars that can thrive in a variety of environmental conditions. Traditional breeding methods tend to be labor-intensive and heavily dependent on the breeder's knowledge, expertise and experience. Cross-breeding is achieved by single crosses, double crosses, multi-way crosses or backcrosses, followed by population development and offspring selection via pedigree or bulk selection methods, for example. The initial step in cross-breeding is the assessment and selection of the parents. In order to improve the flax breeding program, this study investigated genomic approaches to perform cross prediction and produced a case study. The cross-breeding method can be improved through efficient cross prediction, best parent selection and the quick generation of hundreds to thousands of virtual crosses. In breeding, computer simulations are used to predict the desired qualities from the parents' genomic data. Then, using computer tools, virtual crosses and their offspring are generated. Genomic cross prediction is still in its infancy and has only been reported in a limited number of studies (Zhong and Jannink 2007; Bernardo 2014). The QTNs identified for metal content may also be used for genomic cross prediction to forecast the genetic performance of crosses (You et al. 2022b).

High-throughput sequencing is becoming quite affordable, thus increasing access to genotyping which has the potential to enhance breeding accuracy, dependability and speed up decision-making. The findings described herein offer important information for improving flax breeding for metal content and the principles can be extended to other traits including abiotic and biotic stress-related traits. Breeders have significant hurdles when predicting the outcomes of crosses because they rely heavily on the phenotyping of a large number of plants, which is associated with high costs, various degrees of efficiency and errors. The simulation of hundreds to

thousands of crosses has the potential to aid in decreasing error and cost and in improving efficiency through genomic cross prediction. By examining the breeding potential and genetic variants of the segregating populations arising from virtual crosses, this strategy may be able to improve the likelihood of success in cross breeding. In order to assess the effectiveness of GS, future studies must focus on empirical studies to validate predictive abilities. With the availability of marker data, the potential breeding value of crosses through the application of GS approaches may be evaluated, providing a computational tool to simulate segregating populations and estimate the genetic performance of a range of crosses.

Favourable allele analysis for the detected QTNs is also useful for evaluating the parents' or the germplasm quality. The favourable allele of a QTN is defined as the allele that contributes to the performance of the trait. Breeding crosses can also be assessed by directly counting and comparing favourable alleles accumulated in parents, as well as by evaluating the complementation of favourable alleles between parents.

It would be interesting to perform a few case studies of genomic cross prediction using seed content data of other metals in flax and in other crops to verify the generalizability of these findings across elements and species but such extensive analyses were beyond the scope of this thesis.

5.2.2 Genome-wide Association Studies (GWAS)

GWAS are potent statistical genetics methods for detecting marker-trait associations, while candidate gene search is a separate method that links the location of the associated markers and the annotated genes from a reference genome. Eight statistical models were employed to perform the GWAS, which resulted in the discovery of 355 non-redundant QTNs for metal content, which was narrowed down to 108 high-confidence QTNs upon performing the post-GWAS Kruskal-

Wallis allelic test. Recent developments in multi-locus statistical models have made it possible to identify large- and small-effect QTNs for complex traits with excellent accuracy and detection power. Recent GWAS based methods in flax (He et al. 2019b; Soto-Cerda et al. 2021; You et al. 2022a) and other crops (Fang et al. 2020; Fatima et al. 2020; Zhong et al. 2021) have shown the benefits of employing multiple models, particularly multi-locus models to detect QTNs with small effects associated with complex traits. The combined use of many statistical models is a solid method to capitalize on the various genetic or statistical assumptions in each model. Even though several statistical methods were applied to both QTNs and candidate genes, validation of the QTNs using an independent germplasm would be informative. The statistical models used in GWAS have the drawback of identifying several false positives (Ioannidis et al. 2009). Practical solutions to reduce this issue include using a vast and diversified population, high-density markers, accurate trait phenotyping across multiple environments, population structure control methods, stricter multiple test correction criteria, advanced statistical tools and cross-validation with several models. Genome-wide high-density single nucleotide polymorphisms (SNPs) can be quickly and cheaply identified thanks to the rapid development of high-throughput genotyping technology, particularly for crops with small genomes like flax (~370 Mb). However, high-density marker set means that many markers could co-locate within haplotype blocks, and this may result in a multi-collinearity problem when using single-locus statistical models. This would likely decrease the effectiveness and precision of these statistical methods. The elimination of redundant genomic markers can improve computational performance and eliminate the undesirable effects of multi-collinearity.

Using multiple GWAS statistical models with phenotypic datasets frequently results in the identification of sets of QTNs from several environments. To eliminate redundant information and false-positive QTNs, while also identifying stable and environment-specific QTNs, post-QTN

identification analyses are an effective quality control method. Thus, a post-GWAS analysis to remove false-positive QTNs has been applied as previously reported in flax (You et al. 2022a) to ensure that only the most reliable QTNs are considered in downstream applications such as marker-assisted selection. The meta-analysis described by Kang et al. (2014) is another method that strengthens the identification of QTNs that are specific to a certain environment and provides a better understanding of genotype by environment interaction (G x E) (Lo et al. 2019). Such meta-analysis cannot be performed for metal content in flax yet because this study is the first and only study reported to date on these traits in flax.

Another major issue stems from the difficulties associated with comparisons across plant species. Direct comparisons are actually impossible because statistical models and characteristics such as the level of significance and probability correction techniques (Bonferroni or FDR-based corrections) vary between studies. However, through the use of meta-GWAS or meta-quantitative trait loci (QTL) analyses, it may be possible to get around these problems and improve the detection power. Meta-GWAS or meta-QTL investigations have the potential to identify consensus QTLs/QTNs or meta-QTLs. Although meta-GWAS and meta-QTL studies have been used in several crops (Sun et al. 2020), flax has not yet benefited from their use. A meta-analysis of GWAS or QTLs is warranted, given the expanding number of independent association studies with shared traits. With the availability of high-density and genome-wide molecular markers, QTLs associated with polygenic traits can be positioned onto chromosomes, and their effects on phenotypes can be estimated using statistical methods. A QTL may represent a region that is genetically-defined based on recombination or physically-defined based on a nucleotide position(s) on a chromosome, where the extreme of the latter is a single nucleotide position called a QTN.

Complete experimental designs with replications are difficult to apply in the early phases of breeding projects due to the large number of test lines and the limited seed supply. A possible answer to this issue is a class of experimental designs known as "augmented designs" (Federer et al. 1975; Federer and Raghavarao 1975). Control lines in the augmented design are typically organised in soil-homogeneous blocks in several replications of a conventional form, like a Latin square. A modified augmented design (MAD) with two subtypes was proposed by Lin and Poushinsky (1983). The type 1 MAD is employed for square plots (Lin and Poushinsky 1983), while type 2 MAD (MAD2) is designed for rectangular plots (Lin and Poushinsky 1985). This modified design outperforms the standard augmented design in systematic placement of control and test genotypes within blocks to improve adjustment for soil heterogeneity (Lin and Poushinsky 1983). The MAD2 is primarily used for error control in field experiments because it can be used to adjust for soil heterogeneity which was an important factor here considering that mineral concentration in soils can vary greatly within a field. When the same set of test lines and control cultivars with the same MAD2 desgin are used across sites and years, a joint analysis of variance (ANOVA) can be performed. Genetic parameters such as genetic variance, broad-sense heritability for a genetic population can be also estimated by using replicated controls to estimate error variances (You et al. 2016c).

The MAD design is commonly employed in Canadian flax breeding projects (Soto-Cerda et al. 2014; Kumar et al. 2015; Sertse et al. 2021). Breeders frequently apply an upgraded design to a large number of lines that are intended to be planted in a field with a limited area. In unreplicated experimental designs like MAD2, error variance can be calculated from replicated controls. In the current investigation, the phenotypic data were adjusted in accordance with the MAD pipeline as previously described by You et al. (2013). The main reason for using a MAD is to reduce the test

size but still retaining the ability to adjust the data for soil heterogeneity effect which is crucial in case of metal content.

5.2.3 Genomic Selection (GS)

GS offers a significant opportunity for increasing crop yield and improving the efficiency of breeding programs by reducing breeding cycles, selecting superior genotypes and early removal of subpar genotypes from progeny.

An effective technique in GS is the use of QTNs associated with desirable traits as input markers rather than genome-wide random SNPs for model construction. Breeding selection is made possible by the use of multiple models, which has led to a significant number of large- and small-effect QTNs. Prior research demonstrated that trait-defined GS had a higher predictive ability than random SNPs for pasmo, powdery mildew and agronomic traits in flax (He et al. 2019b; Lan et al. 2020; You et al. 2022b). The application of GS depends on the production of high-throughput genotyping data of large breeding populations. For practical breeding purposes, determining a balance between predictive ability and effective marker density is critical to the successful implementation of GS. In this thesis, I examined the effects of a sequential strategy that uses a number of GWAS models to first identify QTNs and then use them as the input dataset for GS implementation. The results show that the GS models derived from the trait-defined QTN marker set outperformed the 43K random SNP set as measured by their genomic predictive abilities for the ten traits evaluated. The combined QTN marker sets are not only significantly smaller, but also more predictive, which facilitates the downstream computation of GS models and could reduce the cost of genotyping test populations. Our GS results may improve practical breeding efficiency through early selection of promising lines, although for Cd the predicted RE

performance of 1.02 was only marginally superior. In general though, all metal contents with lower H^2 had RE greater than one. Similar trends were also observed in flax for agronomic traits (Lan et al. 2020). In brief, yield with H^2 of 0.2 had as high as 4.45 times selection efficiency over phenotypic selection, demonstrating the benefits of GS for low heritability traits. These findings are encouraging from the perspective of practical breeding because selection based on GEBVs potentially reduce the time and costs associated with phenotyping, while still identifying the low-Cd lines.

Small or modest numbers of distant relatives cannot make accurate predictions since they only share a small fraction of the selection candidates' genome, making their predictions less accurate (de Los Campos et al. 2013b). Recent studies in both cattle (Jenko et al. 2017) and barley (Neyhart et al. 2017) provide evidence that the training set should also include diverse individuals in order to get accurate predictions. Continuous use of closely-related i.e., training to test populations to improve prediction will likely narrow or reduce genetic variation that could contribute to the future selection response, and consequently slow down the genetic gains required in long-term GS application (Jannink et al. 2010; Moeinizade et al. 2019). The high accuracy obtained with the QTLs is hypothesized to be a consequence of the reduction in background noises through the removal of unrelated markers. It is possible that the predictive ability might be inflated due to the confounding effect of the QTNs. I independently tested this inflation hypothesis briefly and showed that the QTN-derived input sets generated superior predictive abilities even when the confounding inflation was removed (data not shown). This limited analysis will need to be expanded upon but the preliminary results are promising and support the reported findings. Thus, further validation experiments are needed to test various types of datasets, keeping in mind that

the number of false-positive and false-negative associations can be easily inflated by systematic biases caused by even small sources of error.

5.2.4 Validation of QTNs and Candidate Genes

The detection of QTNs and the prediction of candidate genes mostly depend on the absence of false positive results. The identification of candidate genes is further dependent on a number of other criteria, including the number of markers employed and the density of markers on chromosomes. The most common and straightforward method currently used for determining candidate genes is to scan the annotated genes surrounding the QTNs, such as a window size of 200 Kb (Sertse et al. 2021; You et al. 2022a). Candidate gene prediction of QTNs based on chromosomal position in conjunction with *a priori* knowledge of gene functions produce a list of likely or putative genes that must ultimately be functionally validated.

Herein, eight statistical GWAS models were utilized, resulting in the discovery of 355 non-redundant QTNs, of which 108 has significant allelic effect for metal content. A total of 20 candidate genes were identified, of which five were previously identified through their gene family study (chapter 3) (Khan et al. 2020). Thus, both GWAS and genome-wide gene family analyses were found to be complementary in terms of gene identifications. The outcome provides focus to further the functional genomic characterization of these genes towards the improvement of stress tolerance, particularly for Cd. In this thesis, the QTNs identified for metal content were linked to novel candidate genes for mediation or tolerance. It is anticipated that the findings of this study will facilitate the pre-breeding choices for the transfer of favourable ABC and HMA genes for flax improvement. Overall, the 24 major QTNs explained on average 6.89 ± 1.56% of the phenotypic

variation (R^2) of metal content, highlighting the quantitative nature of these traits and the likely importance of the minor effect QTNs.

The penultimate goal of QTL mapping is to identify the causal gene(s) that control the trait(s) of interest. Candidate genes can be predicted through QTL mapping but the causality must be established through functional validation. The biological role of important candidate genes highlighted in this thesis, along with nine ABC transporter and HMA predicted genes (Khan et al. 2020) would benefit from further explorations using functional approaches such as genome editing and TILLING for example.

5.3 The Practical and Scientific Implications of Results

The knowledge created in this study can be used in practical breeding programs such as the integrated approach described above. I anticipate that such an approach will likely boost flax breeding through accelerating the genetic gains for complex traits. Several QTNs associated with different attributes have been found through GWAS. All QTNs and their associated potential genes can be cross-referenced with future results on flax or other crops. The large-effect loci identified by several models and containing plausible candidate gene(s) can be deployed right away upon validation through marker assisted selection, especially those associated with Cd. The flax seed accumulate Cd into its tissues and thus it is important to tackle this serious issue. While, the flax seeds can be fractionated into oil and protein components for use in specialized food applications. Thus, a better understanding and validation of the GS is the best approach to developing low-Cd lines. Overall, this study has produced significant data and offers a variety of genomic solutions that will be a valuable resource in the future for the improvement of flax varieties.

5.4 Conclusion

Flax is a multipurpose plant and is a historical source of fiber and oil. Despite the fact that the crop has recently attracted attention for its health benefits and nutritional qualities related to its high levels of plant-based omega-3 and other essential fatty acids, its production has been limited by low yield, which has prompted growers to switch to crops that perform better. However, this thesis provides several genomic strategies for improvement of flax breeding programs. Recent important advances in GS and genomic cross prediction provide new potential to facilitate breeding selection efficiency (chapter 1). Selecting the best parental material and the best crosses, and the combined use of GS and genomic cross prediction provide opportunities to improve the accuracy of cross-breeding in flax. However, the use of genomic cross prediction in flax is still in its early stages, and considerable work needs to be done before theoretical examples can be implemented into the practical breeding programs.

Two key factors contribute to the success of GS in flax breeding, and both are improved by the use of trait-defined methodologies. First, these methodologies assist breeders in evaluating parents and selecting the probable best crosses. Secondly, they provide a useful and affordable tool for assisting in breeding programs (chapter 2). A number of factors affect the GS predictive ability, including the training population (TRP) and how it relates to the testing populations (TPs), the statistical GS models, as well as the marker type and density. The predictive ability of GS models is greatly increased by the inclusion of both minor- and large-effect QTNs discovered by single- and multi-locus GWAS models.

Several factors are necessary for successful QTL mapping, such as the ability to accurately phenotype traits across many environments, and the power of the statistical tools. In particular, for crops with small genomes like flax, genome-wide high-density SNPs can now be produced quickly

and inexpensively thanks to the rapid development of high-throughput genotyping technologies like SNP chips and genotyping-by-sequencing (GBS). Several GWAS studies in flax (He et al. 2019b; Soto-Cerda et al. 2021; You et al. 2022a) and other crops (Fang et al. 2020; Fatima et al. 2020; Zhong et al. 2021) have shown the benefits of employing a variety of statistical techniques, particularly multi-locus models, to detect QTNs with minor effects for highly complex traits. The combined use of many statistical models is a solid method to capitalize on each model's advantages, while minimizing their drawbacks caused by the genetic or statistical assumptions of each model.

The ABC transporter and HMA gene family study and GWAS provided complementarity with regards to metal content, particularly Cd, which was the main focus of the research at its onset due to its negative impact on health. The roles of ABC transporter and HMA in Cd accumulation or tolerance have been characterized in some species (chapter 3) and they were analysed by bioinformatics approaches. A total of 198 LuABC and 12 LuHMA genes, which grouped into eight ABC transporter (ABCA-ABCG and ABCI) and four HMA subfamilies, were discovered in the flax genome. Based on homology and phylogenetic relationships to previously identified genes in *Arabidopsis*, rice, and maize, nine of them were predicted to be possibly involved in Cd accumulation in flax.

For ten metal contents, a total of 355 non-redundant QTNs including 108 with significant allelic effect were discovered and 20 candidate genes were proposed at 12 of the 24 major QTN loci (chapter 4). Validation of the candidate genes can be accomplished through functional approaches such as genome editing or targeting induced local lesions in genomes (TILLING) to confirm their role in abiotic stress. The combined GWAS and GS strategy demonstrated the efficiency of the trait-defined QTN datasets in predicting metal content in flax. I suggest the

implementation of this approach to GS in flax because of its robustness, ability to reduce costs and ease of computation. This recommendation stems from the data on Cd content presented herein but could easily be extended to other polygenic traits pursuant testing similar to what is described within this thesis. I believe that the research described provides evidence to convince flax breeders of the benefits of GS, describes its methodology clearly and, I hope that together these two factors will translate into adoption and implementation of GS in flax breeding.

Chapter 6: References

6.0 References

Acuña-Galindo MA, Mason RE, Subramanian NK, et al. (2015) Meta-analysis of wheat QTL regions associated with adaptation to drought and heat stress. Crop Sci 55:477-492

Agarwal P, Baranwal VK, Khurana P (2019) Genome-wide analysis of bZIP transcription factors in wheat and functional characterization of a *TabZIP* under abiotic stress. Sci Rep 9:4608

Akdemir D, Rio S, Isidro y Sánchez J (2021) TrainSel: an r package for selection of training populations. Front Plant Sci 12:655287

Albrecht T, Wimmer V, Auinger H-J, et al. (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123:339

Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12:246

Ali M, Zhang Y, Rasheed A, et al. (2020) Genomic prediction for grain yield and yield-related traits in chinese winter wheat. Int J Mol Sci 21:1342

Ali OA, O'Rourke SM, Amish SJ, et al. (2015) RAD capture (Rapture): flexible and efficient sequence-based genotyping. Genetics 202:389-400

Alloway BJ (2012) Heavy metals in soils: trace metals and metalloids in soils and their bioavailability. Springer Science & Business Media

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881-888

Amarasinghe SL, Su S, Dong X, et al. (2020) Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21:30

Angelova V, Ivanova R, Delibaltova V, et al. (2004) Bio-accumulation and distribution of heavy metals in fibre crops (flax, cotton and hemp). Ind Crops Prod 19:197-205

Chapter 6: References

Asgarinia P, Mirlohi A, Saeidi G, et al. (2016) Selection criteria for assessing drought tolerance in a segregating population of flax (*Linum usitatissimum* L.). Can J Plant Sci 97:424-437

Ashraf M, Foolad MR (2007) Roles of glycine betaine and proline in improving plant abiotic stress resistance. Environ Exp Bot 59:206-216

Badone FC, Cassani E, Landoni M, et al. (2010) The low phytic acid1-241 (lpa1-241) maize mutation alters the accumulation of anthocyanin pigment in the kernel. Planta 231:1189-1199

Bailey-Serres J, Parker JE, Ainsworth EA, et al. (2019) Genetic strategies for improving crop yields. Nature 575:109-118

Bailey TL, Boden M, Buske FA, et al. (2009) MEME suite: tools for motif discovery and searching. Nucleic Acids Res 37:10.1093/nar/gkp1335

Barvkar VT, Pardeshi VC, Kale SM, et al. (2012) Phylogenomic analysis of UDP glycosyltransferase 1 multigene family in *Linum usitatissimum* identified genes with varied expression patterns. BMC Genomics 13:175-175

Baseggio M, Murray M, Wu D, et al. (2021) Genome-wide association study suggests an independent genetic basis of zinc and cadmium concentrations in fresh sweet corn kernels. G3: Genes Genomes Genet 11:8

Bassi FM, Bentley AR, Charmet G, et al. (2016) Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). Plant Sci 242:23-36

Bates D, Mächler M, Bolker B, et al. (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67:1-48

Chapter 6: References

Bayer PE, Ruperao P, Mason AS, et al. (2015) High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in *Cicer arietinum* and *Brassica napus*. Theor Appl Genet 128:1039-1047

Bekele WA, Wight CP, Chao S, et al. (2018) Haplotype-based genotyping-by-sequencing in oat genome research. Plant Biotechnol J 16:1452-1463

Ben Hassen M, Bartholomé J, Valè G, et al. (2018) Genomic prediction accounting for genotype by environment interaction offers an effective framework for breeding simultaneously for adaptation to an abiotic stress and performance under normal cropping conditions in rice. G3: Genes Genet Genom 8:2319-2332

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48:1649-1664

Bernardo R (2014) Genomewide selection of parental inbreds: classes of loci and virtual biparental populations. Crop Sci 54:2586-2595

Bessire M, Borel S, Fabre G, et al. (2011) A member of the pleiotropic drug resistance family of ATP binding cassette transporters is required for the formation of a functional cuticle in *Arabidopsis*. Plant Cell 23:1958-1970

Bevan MW, Uauy C, Wulff BBH, et al. (2017) Genomic innovation for crop improvement. Nature 543:346-354

Bhandari A, Bartholomé J, Cao-Hamadoun T-V, et al. (2019) Selection of trait-specific markers and multi-environment models improve genomic predictive ability in rice. PLoS One 14:e0208871

Bhati KK, Sharma S, Aggarwal S, et al. (2015) Genome-wide identification and expression characterization of ABCC-MRP transporters in hexaploid wheat. Front Plant Sci 6:488

Chapter 6: References

Bita C, Gerats T (2013) Plant tolerance to high temperature in a changing environment: scientific fundamentals and production of heat stress-tolerant crops. Front Plant Sci 4:273

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30:2114-2120

Booker HM (2019) Flax breeding and cultivar registration in canada. In: Cullis CA (ed) Genetics and Genomics of Linum. Springer International Publishing, Cham, Switzerland, pp 39-62

Bovet l, Eggmann t, Meylan-bettex m, et al. (2003) Transcript levels of *ATMRPS* after cadmium treatment: induction of *ATMRP3*. Plant Cell Environ 26:371-381

Bray NL, Pimentel H, Melsted P, et al. (2016) Near-optimal probabilistic RNA-Seq quantification. Nat Biotechnol 34:525-527

Broadley MR, Willey NJ, Wilkins JC, et al. (2001) Phylogenetic variation in heavy metal accumulation in angiosperms. New Phytol 152:9-27

Brown C, Alexander P, Arneth A, et al. (2019) Achievement of Paris climate goals unlikely due to time lags in the land system. Nat Clim Chang 9:203-208

Browning BL, Tian X, Zhou Y, et al. (2021) Fast two-stage phasing of large-scale sequence data. Am J Hum Genet 108:1880-1890

Buckler ES, Ilut DC, Wang X, et al. (2016) rAmpSeq: using repetitive sequences for robust genotyping. bioRxiv:096628

Caamal-Pat D, Pérez-Rodríguez P, Crossa J, et al. (2021) lme4GS: An r-package for genomic selection. Front Plant Sci 12:680569

Cai Z, Xian P, Wang H, et al. (2020) Transcription factor *GmWRKY142* confers cadmium resistance by up-regulating the cadmium tolerance 1-like genes. Front Plant Sci 11:724

Chapter 6: References

Çakır B, Kılıçkaya O (2013) Whole-genome survey of the putative ATP-binding cassette transporter family genes in *Vitis Vinifera*. PLoS One 8:e78860-e78860

Campbell NR, Harmon SA, Narum SR (2015) Genotyping-in-thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. Mol Ecol Resour 15:855-867

Cannon SB, Mitra A, Baumgarten A, et al. (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biol 4:10-10

Cao J (2019) Molecular evolution of the vacuolar iron transporter (VIT) family genes in 14 plant species. Genes 10:2

Cao Y, Zhao X, Liu Y, et al. (2019) Genome-wide identification of ZmHMAs and association of natural variation in *ZmHMA2* and *ZmHMA3* with leaf cadmium accumulation in maize. PeerJ 7:e7877-e7877

Cempel M, Nikel G (2006) Nickel: a review of its sources and environmental toxicology. Pol J Environ Stud 15:3

Chao D-Y, Silva A, Baxter I, et al. (2012) Genome-wide association studies identify heavy metal ATPase3 as the primary determinant of natural variation in leaf cadmium in *Arabidopsis thaliana*. PLoS Genet 8:e1002923

Chaves MM, Flexas J, Pinheiro C (2008) Photosynthesis under drought and salt stress: regulation mechanisms from whole plant to cell. Ann Bot 103:551-560

Chen C, Chen H, Zhang Y, et al. (2020) TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant 13:1194-1202

Chapter 6: References

Chen G, Komatsuda T, Ma JF, et al. (2011) An ATP-binding cassette subfamily G full transporter is essential for the retention of leaf water in both wild barley and rice. Proc Natl Acad Sci USA 108:12354-12359

Chen J, Chopra R, Hayes C, et al. (2017) Genome-wide association study of developing leaves' heat tolerance during vegetative growth stages in a sorghum association panel. Plant Genome 10:10.3835/plantgenome2016.3809.0091

Chen L, Wan H, Qian J, et al. (2018) Genome-wide association study of cadmium accumulation at the seedling stage in rapeseed (*Brassica napus* L.). Front Plant Sci 9:375

Chinnusamy V, Zhu J, Zhu JK (2007) Cold stress regulation of gene expression in plants. Trends Plant Sci 12:444-451

Ciesla A, Mitula F, Misztal L, et al. (2016) A role for barley calcium-dependent protein kinase CPK2a in the response to drought. Front Plant Sci 7:1550

Clark JW, Donoghue PCJ (2018) Whole-genome duplication and plant macroevolution. Trends Plant Sci 23:933-945

Clark SA, van der Werf J (2013) Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. Methods Mol Biol 1019:321-330

Clemens S (2006) Toxic metal accumulation, responses to exposure and mechanisms of tolerance in plants. Biochimie 88:1707-1719

Cloutier S, Ragupathy R, Miranda E, et al. (2012) Integrated consensus genetic and physical maps of flax (*Linum usitatissimum* L.). Theor Appl Genet 125:1783-1795

Corbin C, Drouet S, Markulin L, et al. (2018) A genome-wide analysis of the flax (*Linum usitatissimum* L.) dirigent protein family: from gene identification and evolution to differential regulation. Plant Mol Biol 97:73-101

Chapter 6: References

Covarrubias-Pazaran G (2016) Genome-assisted prediction of quantitative traits using the R package sommer. PLoS One 11:e0156744

Cross RH, Mckay SAB, G. Mchughen A, et al. (2003) Heat-stress effects on reproduction and seed set in *Linum usitatissimum* L. (flax). Plant Cell Environ 26:1013-1020

Crossa J, Pérez-Rodríguez P, Cuevas J, et al. (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961-975

Cui Y, Zhang F, Zhou Y (2018) The application of multi-locus GWAS for the detection of salt-tolerance loci in rice. Front Plant Sci 9:1464

Daetwyler HD, Calus MP, Pong-Wong R, et al. (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347-365

Daetwyler HD, Pong-Wong R, Villanueva B, et al. (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021-1031

Dai X, Xu Y, Ma Q, et al. (2007) Overexpression of an *R1R2R3* MYB gene, *OsMYB3R*, increases tolerance to freezing, drought, and salt stress in transgenic *Arabidopsis*. Plant Physiol 143:1739-1751

Danecek P, McCarthy SA (2017) BCFtools/csq: haplotype-aware variant consequences. Bioinformatics 33:2037-2039

Das N, Bhattacharya S, Maiti MK (2016) Enhanced cadmium accumulation and tolerance in transgenic tobacco overexpressing rice metal tolerance protein gene *OsMTP1* is promising for phytoremediation. Plant Physiol Biochem 105:297-309

Dash PK, Cao Y, Jailani AK, et al. (2014) Genome-wide analysis of drought induced gene expression changes in flax (*Linum usitatissimum*). GM Crops Food 5:106-119

Chapter 6: References

Davey JW, Hohenlohe PA, Etter PD, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499-510

de los Campos G, Hickey JM, Pong-Wong R, et al. (2013a) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327-345

de Los Campos G, Vazquez AI, Fernando R, et al. (2013b) Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet 9:e1003608

Dekkers JCM (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet 124:331-341

Deolu-Ajayi AO, Meyer AJ, Haring MA, et al. (2019) Genetic loci associated with early salt stress responses of roots. iScience 21:458-473

Deshmukh R, Sonah H, Patil G, et al. (2014) Integrating omic approaches for abiotic stress tolerance in soybean. Front Plant Sci 5:244

Desta ZA, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci 19:592-601

Dhankher OP, Foyer CH (2018) Climate resilient crops for improving global food security and safety. Plant Cell Environ 41:877-884

Dias KODG, Gezan SA, Guimarães CT, et al. (2018) Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity 121:24-37

Die JV, Gil J, Millan T (2018) Genome-wide identification of the auxin response factor gene family in *Cicer arietinum*. BMC Genomics 19:301

Chapter 6: References

Diederichsen A, Kusters PM, Kessler D, et al. (2013) Assembling a core collection from the flax world collection maintained by plant gene resources of Canada. Genet Resour Crop Evol 60:1479-1485

Diederichsen A, Rozhmina, RT. Zhuchenko, AA, et al. (2006) Screening for broad adaptation in 96 flax (*Linum usitatissimum* L.) accessions under dry and warm conditions in Canada and Russia. Plant Genet Resour Newsl 146:9–16

Do TD, Vuong TD, Dunn D, et al. (2019) Identification of new loci for salt tolerance in soybean by high-resolution genome-wide association mapping. BMC Genomics 20:318

Driedonks N, Rieu I, Vriezen WH (2016) Breeding for plant heat tolerance at vegetative and reproductive stages. Plant Reprod 29:67-79

Dröge-Laser W, Snoek BL, Snel B, et al. (2018) The *Arabidopsis* bZIP transcription factor family—an update. Curr Opin Plant Biol 45:36-49

Dubos C, Stracke R, Grotewold E, et al. (2010) MYB transcription factors in *Arabidopsis*. Trends Plant Sci 15:573-581

El-Beltagi HS, Salama ZA, El Hariri DM (2008) Some biochemical markers for evaluation of flax cultivars under salt stress conditions. J Nat Fibers 5:316-330

El-Gebali S, Mistry J, Bateman A, et al. (2018) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427-D432

El-Hariri DM, Al-Kordy MA, Hassanein MS, et al. (2005) Partition of photosynthates and energy production in different flax cultivars. J Nat Fibers 1:1-15

Emamverdian A, Ding Y, Mokhberdoran F, et al. (2015) Heavy metal stress and some mechanisms of plant defense response. Sci World J 2015:756120

Chapter 6: References

Endelman JB (2011) Ridge regression and other kernels for genomic selection with r package rrBLUP. Plant Genome 4:https://doi.org/10.3835/plantgenome2011.08.0024.

Eom SH, Hyun TK (2016) Genome-wide identification and transcriptional expression analysis of chalcone synthase in flax (*Linum usitatissimum* L.). Gene Rep 5:51-56

Fahlgren N, Gehan MA, Baxter I (2015) Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Curr Opin Plant Biol 24:93-99

Fan W, Liu C, Cao B, et al. (2018) Genome-wide identification and characterization of four gene families putatively involved in cadmium uptake, translocation and sequestration in mulberry. Front Plant Sci 9:879-879

Fang Y, Liu S, Dong Q, et al. (2020) Linkage analysis and multi-locus genome-wide association studies identify qtns controlling soybean plant height. Front Plant Sci 11:00009

Fasani E, Manara A, Martini F, et al. (2018) The potential of genetic engineering of plants for the remediation of soils contaminated with heavy metals. Plant Cell Environ 41:1201-1232

Fatima F, McCallum BD, Pozniak CJ, et al. (2020) Identification of new leaf rust resistance loci in wheat and wild relatives by array-based snp genotyping and association genetics. Front Plant Sci 11:583738

Faux AM, Gorjanc G, Gaynor RC, et al. (2016) AlphaSim: software for breeding program simulation. Plant Genome 9:10.3835/plantgenome2016.3802.0013

Federer W, Nair R, Raghavarao D (1975) Some augmented row-column designs. Biometrics:361-373

Federer WT, Raghavarao D (1975) On augmented designs. Biometrics:29-35

Feng ZT, Deng YQ, Fan H, et al. (2014) Effects of NaCl stress on the growth and photosynthetic characteristics of *Ulmus pumila* L. seedlings in sand culture. Photosynthetica 52:313-320

Chapter 6: References

Fernandez GCJ, Asian Vegetable R, Development C, et al. (1993) Effective selection criteria for assessing plant stress tolerance. International symposium, adaptation of food crops to temperature and water stress, 410 edn. AVRDC;, Taipei, Taiwan, Taipei, [unconfirmed], pp 257-270

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29-W37

Flowers TJC, T.D. (2008) Salinity tolerance in halophytes. New Phytol 179:945-963

Frichot E, François O (2015) LEA: an R package for landscape and ecological association studies. Methods Ecol Evol 6:925-929

Fu S, Lu Y, Zhang X, et al. (2019) The ABC transporter *ABCG36* is required for cadmium tolerance in rice. J Exp Bot 70:5909-5918

Fu YB, Yang MH, Zeng F, et al. (2017) Searching for an accurate marker-based prediction of an individual quantitative trait in molecular plant breeding. Front Plant Sci 8:1182

Furbank RT, Tester M (2011) Phenomics – technologies to relieve the phenotyping bottleneck. Trends Plant Sci 16:635-644

Gadsby DC, Vergani P, Csanády L (2006) The ABC protein turned chloride channel whose failure causes cystic fibrosis. Nature 440:477-483

Gao X, Ren Z, Zhao Y, et al. (2003) Overexpression of *SOD2* increases salt tolerance of *Arabidopsis*. Plant Physiol 133:1873-1881

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761-1776

Chapter 6: References

Gierczik K, Székely A, Ahres M, et al. (2019) Overexpression of two upstream phospholipid signaling genes improves cold stress response and hypoxia tolerance, but leads to developmental abnormalities in barley. Plant Mol Biol Rep 37:314-326

Gilmour SJ, Fowler SG, Thomashow MF (2004) *Arabidopsis* transcriptional activators *CBF1*, *CBF2*, and *CBF3* have matching functional activities. Plant Mol Biol 54:767-781

Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257

González-Recio O, Forni S (2011) Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genet Sel Evol 43:7

Goodstein DM, Shu S, Howson R, et al. (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178-D1186

Goyal A, Sharma V, Upadhyay N, et al. (2014) Flax and flaxseed oil: an ancient medicine & modern functional food. J Food Sci Technol 51:1633-1653

Gu X, Zou Y, Su Z, et al. (2013) An update of diverge software for functional divergence analysis of protein family. Mol Biol Evol 30:1713-1719

Guha Majumdar S, Rai A, Mishra DC (2020) Integrated framework for selection of additive and nonadditive genetic markers for genomic selection. J Comput Biol 27:845-855

Günther T, Gawenda I, Schmid KJ (2011) phenosim - a software to simulate phenotypes for testing in genome-wide association studies. BMC Bioinform 12:265

Guo J, Li Y, Han G, et al. (2018) NaCl markedly improved the reproductive capacity of the euhalophyte *Suaeda salsa*. Funct Plant Biol 45:350-361

Guo Z, Wang H, Tao J, et al. (2019) Development of multiple SNP marker panels affordable to breeders through genotyping by target sequencing (GBTS) in maize. Mol Breed 39:37

Chapter 6: References

Gutsch A, Keunen E, Guerriero G, et al. (2018) Long-term cadmium exposure influences the abundance of proteins that impact the cell wall structure in *Medicago sativa* stems. Plant Biol 20:1023-1035

Habier D, Fernando RL, Kizilkaya K, et al. (2011) Extension of the bayesian alphabet for genomic selection. BMC Bioinform 12:186

Hall LM, Booker H, Siloto RMP, et al. (2016) Chapter 6 - Flax (*Linum usitatissimum* L.). In: McKeon TA, et al. (eds) Industrial Oil Crops. AOCS Press, pp 157-194

Haller BC, Messer PW (2019) SLiM 3: forward genetic simulations beyond the wright–fisher model. Mol Biol Evol 36:632-637

Hao Y-Q, Lu G-Q, Wang L-H, et al. (2018) Overexpression of *AmDUF1517* enhanced tolerance to salinity, drought, and cold stress in transgenic cotton. J Integr Agric 17:2204-2214

Hartmann A, Czauderna T, Hoffmann R, et al. (2011) HTPheno: an image analysis pipeline for high-throughput plant phenotyping. BMC Bioinform 12:148

Hashem HB, FM; Hassanein, RA; Baraka, DM and Khalil, RR. (2011) Stigmasterol seed treatment alleviates the drastic effect of NaCL and improves quality and yield in flax plants. Aust J Crop Sci 5:1858-1867

Hassan TU, Bano A, Naz I (2017) Alleviation of heavy metals toxicity by the application of plant growth promoting rhizobacteria and effects on wheat grown in saline sodic field. Int J Phytoremediation 19:522-529

Hawkesford MJ, Griffiths S (2019) Exploiting genetic variation in nitrogen use efficiency for cereal crop improvement. Curr Opin Plant Biol 49:35-42

Hayes BJ, Bowman PJ, Chamberlain AJ, et al. (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433-443

Chapter 6: References

He J, Meng S, Zhao T, et al. (2017) An innovative procedure of genome-wide association analysis fits studies on germplasm population and plant breeding. Theor Appl Genet 130:2327-2343

He J, Zhao X, Laroche A, et al. (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci 5:00484

He L, Xiao J, Rashid KY, et al. (2019a) Evaluation of genomic prediction for pasmo resistance in flax. Int J Mol Sci 20:2

He L, Xiao J, Rashid KY, et al. (2019b) Genome-wide association studies for pasmo resistance in flax (*Linum usitatissimum* L.). Front Plant Sci 9:1982

Heffner EL, Jannink J-L, Iwata H, et al. (2011) Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci 51:2597-2606

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci 49:1-12

Heller K, Byczyńska M (2015) The impact of environmental factors and applied agronomy on quantitative and qualitative traits of flax fiber. J Nat Fibers 12:26-38

Heslot N, Jannink J-L, Sorrells ME (2015) Perspectives for genomic selection applications and research in plants. Crop Sci 55:1-12

Heslot N, Yang H-P, Sorrells ME, et al. (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146-160

Hickey JM, Dreisigacker S, Crossa J, et al. (2014) Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci 54:1476-1488

Hirschhorn JN (2009) Genomewide association studies--illuminating biologic pathways. N Engl J Med 360:1699-1701

Chapter 6: References

Hochholdinger F, Yu P, Marcon C (2018) Genetic control of root system development in maize. Trends Plant Sci 23:79-88

Hocking PJ, Kirkegaard JA, Angus JF, et al. (1997) Comparison of canola, Indian mustard and Linola in two contrasting environments. Effects of nitrogen fertilizer on dry-matter production, seed yield and seed quality. Field Crops Res 49:107-125

Hoffstetter A, Cabrera A, Huang M, et al. (2016) Optimizing training population data and validation of genomic selection for economic traits in soft winter wheat. G3: Genes Genom Genet 6:2919-2928

Hollenstein K, Frei DC, Locher KP (2007) Structure of an ABC transporter in complex with its binding protein. Nature 446:213-216

Hopkin M (2005) Climate change: world round-up. Nature:10.1038/news051114-051110

Horton P, Park K-J, Obayashi T, et al. (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res 35:W585-W587

House MA, Young LW, Liu X, et al. (2020) Comparative analysis of cadmium uptake and distribution in contrasting canadian flax cultivars. BMC Res Notes 13:424

Hoyos-Villegas V, Arief VN, Yang W-H, et al. (2019) QuLinePlus: extending plant breeding strategy and genetic model simulation to cross-pollinated populations—case studies in forage breeding. Heredity 122:684-695

Huang X-Y, Deng F, Yamaji N, et al. (2016) A heavy metal P-type ATPase *OsHMA4* prevents copper accumulation in rice grain. Nat Commun 7:12138

Hussain W, Campbell MT, Jarquin D, et al. (2020) Variance heterogeneity genome-wide mapping for cadmium in bread wheat reveals novel genomic loci and epistatic interactions. Plant Genome 13:10.1002/tpg1002.20011

Chapter 6: References

Hwang E-Y, Song Q, Jia G, et al. (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genomics 15:1

Ioannidis JP, Thomas G, Daly MJ (2009) Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 10:318-329

Ismael MA, Elyamine AM, Moussa MG, et al. (2019) Cadmium in plants: uptake, toxicity, and its interactions with selenium fertilizers. Metallomics 11:255-277

Jähne F, Balko C, Hahn V, et al. (2019) Cold stress tolerance of soybeans during flowering: QTL mapping and efficient selection strategies under controlled conditions. Plant Breed 138:708-720

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166-177

Jasinski M, Ducos E, Martinoia E, et al. (2003) The ATP-binding cassette transporters: structure, function, and gene family comparison between rice and *Arabidopsis*. Plant Physiol 131:1169-1177

Jenko J, Wiggans GR, Cooper TA, et al. (2017) Cow genotyping strategies for genomic selection in a small dairy cattle population. J Dairy Sci 100:439-452

Jia Z, Liu Y, Gruber BD, et al. (2019) Genetic dissection of root system architectural traits in spring barley. Front Plant Sci 10:400

Jiang J, Ma S, Ye N, et al. (2017) WRKY transcription factors in plant responses to stresses. J Integr Plant Biol 59:86-101

Jiang Q, Hu Z, Zhang H, et al. (2014) Overexpression of *GmDREB1* improves salt tolerance in transgenic wheat and leaf protein response to high salinity. Crop J 2:120-131

Chapter 6: References

Jiao Y, Grant CA, Bailey LD (2004) Effects of phosphorus and zinc fertilizer on cadmium uptake and distribution in flax and durum wheat. J Sci Food Agric 84:777-785

Jiao Y, Wickett NJ, Ayyampalayam S, et al. (2011) Ancestral polyploidy in seed plants and angiosperms. Nature 473:97-100

Jin C, Huang X-S, Li K-Q, et al. (2016) Overexpression of a bHLH1 transcription factor of *Pyrus ussuriensis* confers enhanced cold tolerance and increases expression of stress-responsive genes. Front Plant Sci 7:441

Jones PM, George AM (2004) The ABC transporter structure and mechanism: perspectives on recent research. Cell Mol Life Sci 61:682-699

Juliana P, Poland J, Huerta-Espino J, et al. (2019) Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nat Genet 51:1530-1539

Juretic N, Hoen DR, Huynh ML, et al. (2005) The evolutionary fate of mule-mediated duplications of host gene fragments in rice. Genome Res 15:1292-1297

Kabata-Pendias A (2000) Trace elements in soils and plants. CRC press

Kang EY, Han B, Furlotte N, et al. (2014) Meta-analysis identifies gene-by-environment interactions as demonstrated in a study of 4,965 mice. PLoS Genet 10:e1004022

Kang J, Park J, Choi H, et al. (2011) Plant ABC transporters. *Arabidopsis* Book 9:e0153-e0153

Kar G, Kumar A, Martha M (2007) Water use efficiency and crop coefficients of dry season oilseed crops. Agric Water Manage 87:73-82

Kato T, Tabata S, Sato S (2009) Analyses of expression and phenotypes of knockout lines for *Arabidopsis* ABCF subfamily members. Plant Biotech 26:409-414

Keesing F, Belden LK, Daszak P, et al. (2010) Impacts of biodiversity on the emergence and transmission of infectious diseases. Nature 468:647-652

Chapter 6: References

Khadivi-Khub A (2014) Regression association analysis of fruit traits with molecular markers in cherries. Plant Syst Evol 300:1163-1173

Khan N, Fatima F, Haider MS, et al. (2019a) Genome-wide identification and expression profiling of the polygalacturonase (PG) and pectin methylesterase (PME) genes in grapevine (*Vitis vinifera* L.). Int J Mol Sci 20:3180

Khan N, Hu C-m, Amjad Khan W, et al. (2018a) Genome-wide identification, classification, and expression divergence of glutathione-transferase family in *Brassica rapa* under multiple hormone treatments. Biomed Res Int 2018:6023457

Khan N, Hu C-m, Khan WA, et al. (2018b) Genome-wide identification, classification, and expression pattern of homeobox gene family in *Brassica rapa* under various stresses. Sci Rep 8:16265

Khan N, Ke H, Hu C-M, et al. (2019b) Genome-wide identification, evolution, and transcriptional profiling of *PP2C* gene family in *Brassica rapa*. Biomed Res Int 2019:2965035

Khan N, You FM, Cloutier S (2022) Designing genomic solutions to enhance abiotic stress resistance in flax. In: Kole C (ed) Genomic Designing for Abiotic Stress Resistant Oilseed Crops. Springer International Publishing, Cham, pp 251-283

Khan N, You FM, Datla R, et al. (2020) Genome-wide identification of ATP binding cassette (ABC) transporter and heavy metal associated (HMA) gene families in flax (*Linum usitatissimum* L.). BMC Genomics 21:722

Khurana N, Sharma N, Khurana P (2017) Overexpression of a heat stress inducible, wheat myo-inositol-1-phosphate synthase 2 (*TaMIPS2*) confers tolerance to various abiotic stresses in *Arabidopsis thaliana*. Agri Gene 6:24-30

Chapter 6: References

Kim D-Y, Bovet L, Kushnir S, et al. (2006) AtATM3 is involved in heavy metal resistance in *Arabidopsis*. Plant Physiol 140:922

Kim D-Y, Bovet L, Maeshima M, et al. (2007) The ABC transporter *AtPDR8* is a cadmium extrusion pump conferring heavy metal resistance. Plant J 50:207-218

Kim D-Y, Jin J-Y, Alejandro S, et al. (2010) Overexpression of *AtABCG36* improves drought and salt stress resistance in *Arabidopsis*. Physiol Plant 139:170-180

Kim Y-J, Yang D-H, Park M-Y, et al. (2020) Overexpression of *Zoysia ZjCIGR1* gene confers cold stress resistance to zoysiagrass. Plant Biotechnol Rep 14:21-31

Koch MA, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (*Brassicaceae*). Mol Biol Evol 17:1483-1498

Korenkov V, King B, Hirschi K, et al. (2009) Root-selective expression of *AtCAX4* and *AtCAX2* results in reduced lamina cadmium in field-grown *Nicotiana tabacum* L. Plant Biotechnol J 7:219-226

Korenkov V, Park S, Cheng N-H, et al. (2007) Enhanced Cd^{2+}-selective root-tonoplast-transport in tobaccos expressing *Arabidopsis* cation exchangers. Planta 225:403-411

Krchov L-M, Bernardo R (2015) Relative efficiency of genomewide selection for testcross performance of doubled haploid lines in a maize breeding program. Crop Sci 55:2091-2099

Król A, Amarowicz R, Weidner S (2015) The effects of cold stress on the phenolic compounds and antioxidant capacity of grapevine (*Vitis vinifera* L.) leaves. J Plant Physiol 189:97-104

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870-1874

Chapter 6: References

Kumar S, You FM, Cloutier S (2012) Genome wide SNP discovery in flax through next generation sequencing of reduced representation libraries. BMC Genomics 13:684

Kumar S, You FM, Duguid S, et al. (2015) QTL for fatty acid composition and yield in linseed (*Linum usitatissimum* L.). Theor Appl Genet 128:965-984

Kuromori T, Miyaji T, Yabuuchi H, et al. (2010) ABC transporter *AtABCG25* is involved in abscisic acid transport and responses. Proc Natl Acad Sci USA 107:2361-2366

Lado B, Battenfield S, Guzmán C, et al. (2017) Strategies for selecting crosses using genomic prediction in two wheat breeding programs. Plant Genome 10:10.3835/plantgenome2016.3812.0128

Lafarge T, Bueno C, Frouin J, et al. (2017) Genome-wide association analysis for heat tolerance at flowering detected a large set of genes involved in adaptation to thermal and other stresses. PLoS One 12:e0171254

Lan S, Zheng C, Hauck K, et al. (2020) Genomic prediction accuracy of seven breeding selection traits improved by QTL identification in flax. Int J Mol Sci 21:1577

Landi S, Hausman J-F, Guerriero G, et al. (2017) Poaceae vs. abiotic stress: focus on drought and salt stress, recent insights and perspectives. Front Plant Sci 8:1214

Lane TS, Rempe CS, Davitt J, et al. (2016) Diversity of ABC transporter genes across the plant kingdom and their potential utility in biotechnology. BMC Biotechnol 16:47

Leal SM, Yan K, Müller-Myhsok B (2005) SimPed: a simulation program to generate haplotype and genotype data for pedigree structures. Hum Hered 60:119-122

Lehermeier C, Krämer N, Bauer E, et al. (2014) Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction. Genetics 198:3-16

Chapter 6: References

Lekklar C, Pongpanich M, Suriya-arunroj D, et al. (2019) Genome-wide association study for salinity tolerance at the flowering stage in a panel of rice accessions from Thailand. BMC Genomics 20:76

Lesk C, Rowhani P, Ramankutty N (2016) Influence of extreme weather disasters on global crop production. Nature 529:84-87

Letunic I, Bork P (2017) 20 years of the SMART protein domain annotation resource. Nucleic Acids Res 46:D493-D496

Li H, Durbin R (2009) Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics 25:1754-1760

Li H, Handsaker B, Wysoker A, et al. (2009a) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079

Li H, Rasheed A, Hickey LT, et al. (2018a) Fast-forwarding genetic gain. Trends Plant Sci 23:184-186

Li J, Zhang Z, Vang S, et al. (2009b) Correlation between *ka/ks* and *ks* is related to substitution model and evolutionary lineage. J Mol Evol 68:414-423

Li K, Yu H, Li T, et al. (2017) Cadmium accumulation characteristics of low-cadmium rice (*Oryza sativa* L.) line and F(1) hybrids grown in cadmium-contaminated soils. Environ Sci Pollut Res Int 24:17566-17576

Li T, Ma J, Zou Y, et al. (2019a) Quantitative trait loci for seeding root traits and the relationships between root and agronomic traits in common wheat. Genome 63:27-36

Li X, Chen R, Chu Y, et al. (2018b) Overexpression of RCc3 improves root system architecture and enhances salt tolerance in rice. Plant Physiol Biochem 130:566-576

Chapter 6: References

Li X, Wu Y, Li B, et al. (2018c) Genome-wide identification and expression analysis of the cation diffusion facilitator gene family in turnip under diverse metal ion stresses. Front Genet 9:103

Li Y, Guan K, Schnitkey GD, et al. (2019b) Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States. Glob Chang Biol 25:2325-2337

Li Y, Telfer E, Wilcox PL (2015) New Zealand forestry enters the genomics era - applications of genomics in tree breeding. N Z J For 60:23-25

Li Z, Sillanpää MJ (2012) Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor Appl Genet 125:419-435

Lin C-S, Poushinsky G (1983) A modified augmented design for an early stage of plant selection involving a large number of test lines without replication. Biometrics:553-561

Lin C-S, Poushinsky G (1985) A modified augmented design (type 2) for rectangular plots. Can J Plant Sci 65:743-749

Lipka AE, Tian F, Wang Q, et al. (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397-2399

Liu C, Chen K, Zhao X, et al. (2019a) Identification of genes for salt tolerance and yield-related traits in rice plants grown hydroponically and under saline field conditions by genome-wide association study. Rice 12:88

Liu H, Tessema BB, Jensen J, et al. (2019b) ADAM-Plant: a software for stochastic simulations of plant breeding from molecular to phenotypic level and from simple selection to complex speed breeding programs. Front Plant Sci 9:1926

Chapter 6: References

Liu H, Xing M, Yang W, et al. (2019c) Genome-wide identification of and functional insights into the late embryogenesis abundant (LEA) gene family in bread wheat (*Triticum aestivum*). Sci Rep 9:13375

Liu J, Gao Y, Tang Y, et al. (2019d) Genome-wide identification, comprehensive gene feature, evolution, and expression analysis of plant metal tolerance proteins in tobacco under heavy metal toxicity. Front Genet 10:345

Liu J, Gao Y, Tang Y, et al. (2019e) Genome-wide identification, comprehensive gene feature, evolution, and expression analysis of plant metal tolerance proteins in tobacco under heavy metal toxicity. Front Genet 10:345

Liu M, Ma Z, Sun W, et al. (2019f) Genome-wide analysis of the NAC transcription factor family in Tartary buckwheat (*Fagopyrum tataricum*). BMC Genomics 20:113

Liu X, Huang M, Fan B, et al. (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet 12:e1005767

Liu X, Wang H, Wang H, et al. (2018) Factors affecting genomic selection revealed by empirical evidence in maize. Crop J 6:341-352

Liu XP, Hawkins C, Peel MD, et al. (2019g) Genetic loci associated with salt tolerance in advanced breeding populations of tetraploid alfalfa using genome-wide association studies. Plant Genome 12:10.3835/plantgenome2018.3805.0026

Liu Z, Li H, Gou Z, et al. (2020) Genome-wide association study of soybean seed germination under drought stress. Mol Genet Genom 295:661-673

Lo S, Munoz-Amatriain M, Hokin SA, et al. (2019) A genome-wide association and meta-analysis reveal regions associated with seed size in cowpea [*Vigna unguiculata* (L.) Walp]. Theor Appl Genet 132:3079-3087

Chapter 6: References

Lu S, Fadlalla T, Tang S, et al. (2019) Genome-wide analysis of phospholipase d gene family and profiling of phospholipids under abiotic stresses in *Brassica napus*. Plant Cell Physiol 60:1556

Lush JL (1937) Animal breeding plans, Ames, Iowa : Collegiate Press, 1937

Lux A, Martinka M, Vaculík M, et al. (2010) Root responses to cadmium in the rhizosphere: a review. J Exp Bot 62:21-37

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155

Marchler-Bauer A, Bo Y, Han L, et al. (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res 45:D200-D203

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133-141

Matei G, Woyann LG, Milioli AS, et al. (2018) Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. Mol Breed 38:117

Mathew I, Shimelis H, Shayanowako AIT, et al. (2019) Genome-wide association study of drought tolerance and biomass allocation in wheat. PLoS One 14:e0225383

Matiu M, Ankerst DP, Menzel A (2017) Interactions between temperature and drought in global and regional crop yield variability during 1961-2014. PLoS One 12:e0178339

Maulana F, Ayalew H, Anderson JD, et al. (2018) Genome-wide association mapping of seedling heat tolerance in winter wheat. Front Plant Sci 9:1272

McHughen A (1987) Salt tolerance through increased vigor in a flax line (STS-II) selected for salt tolerance *in vitro*. Theor Appl Genet 74:727-732

Chapter 6: References

McHughen A, Swartz M (1984) A tissue-culture derived salt-tolerant line of flax (*Linum usitatissimum*). J Plant Physiol 117:109-117

Medina CA, Hawkins C, Liu X-P, et al. (2020) Genome-wide association and prediction of traits related to salt tolerance in autotetraploid alfalfa (*Medicago sativa* L.). Int J Mol Sci 21:3361

Mellers G, Mackay I, Cowan S, et al. (2020) Implementing within-cross genomic prediction to reduce oat breeding costs. Plant Genome 13:e20004

Mercereau-Puijalon O, Barale JC, Bischoff E (2002) Three multigene families in *Plasmodium* parasites: facts and questions. Int J Parasitol 32:1323-1344

Metsalu T, Vilo J (2015) ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. Nucleic Acids Res 43:W566-570

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829

Meyer HV, Birney E (2018) PhenotypeSimulator: a comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. Bioinformatics 34:2951-2956

Miao H, Sun P, Liu Q, et al. (2017) Genome-wide analyses of SWEET family proteins reveal involvement in fruit development and abiotic/biotic stress responses in banana. Sci Rep 7:3536

Mikkelsen MD, Pedas P, Schiller M, et al. (2012) Barley *HvHMA1* is a heavy metal pump involved in mobilizing organellar Zn and Cu and plays a role in metal loading into grains. PLoS One 7:e49027-e49027

Mishra AK, Choi J, Rabbee MF, et al. (2019) In silico genome-wide analysis of the ATP-binding cassette transporter gene family in soybean (*Glycine max* L.) and their expression profiling. Biomed Res Int 2019:14

Chapter 6: References

Moeinizade S, Hu G, Wang L, et al. (2019) Optimizing selection and mating in genomic selection with a look-ahead approach: an operations research framework. G3: Genes Genom Genet 9:2123-2133

Moghe GD, Shiu SH (2014) The causes and molecular consequences of polyploidy in flowering plants. Ann NY Acad Sci 1320:16-34

Monostori I, Szira F, Tondelli A, et al. (2017) Genome-wide association study and genetic diversity analysis on nitrogen use efficiency in a Central European winter wheat (*Triticum aestivum* L.) collection. PLoS One 12:e0189265

Montesinos-López OA, Luna-Vázquez FJ, Montesinos-López A, et al. (2018) An r package for multitrait and multienvironment data with the item-based collaborative filtering algorithm. Plant Genome 11:10.3835/plantgenome2018.02.0013

Morel M, Crouzet J, Gravot A, et al. (2009) *AtHMA3*, a P_{1B}-ATPase allowing Cd/Zn/Co/Pb vacuolar storage in *Arabidopsis*. Plant Physiol 149:894-904

Moser G, Tier B, Crump RE, et al. (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet Sel Evol 41:56

Munns R, Tester M (2008) Mechanisms of salinity tolerance. Annu Rev Plant Biol 59:651-681

Mwando E, Han Y, Angessa TT, et al. (2020) Genome-wide association study of salinity tolerance during germination in barley (*Hordeum vulgare* L.). Front Plant Sci 11:118

Nagy R, Grob H, Weder B, et al. (2009) The *Arabidopsis* ATP-binding cassette protein *AtMRP5/AtABCC5* is a high affinity inositol hexakisphosphate transporter involved in guard cell signaling and phytate storage. J Biol Chem 284:33614-33622

Narayanan S, Mohan A, Gill KS, et al. (2014) Variability of root traits in spring wheat germplasm. PLoS One 9:e100317

Chapter 6: References

Naveed SA, Zhang F, Zhang J, et al. (2018) Identification of QTN and candidate genes for salinity tolerance at the germination and seedling stages in rice by genome-wide association analyses. Sci Rep 8:6505

Neyhart JL, Tiede T, Lorenz AJ, et al. (2017) Evaluating methods of updating training data in long-term genomewide selection. G3: Genes Genomes Genet 7:1499-1510

Nguyen QH, Vu LTK, Nguyen LTN, et al. (2019) Overexpression of the *GmDREB6* gene enhances proline accumulation and salt tolerance in genetically modified soybean plants. Sci Rep 9:19663

Nicod J, Davies, R., Cai, N. et al. (2016) Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nat Genet 48:912-918

Nielsen R, Paul JS, Albrechtsen A, et al. (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443-451

Noh B, Murphy AS, Spalding EP (2001) Multidrug resistance–like genes of *Arabidopsis* required for auxin transport and auxin-mediated development. Plant Cell 13:2441-2454

Ornella L, Singh S, Perez P, et al. (2012) Genomic prediction of genetic values for resistance to wheat rusts. Plant Genome 5:https://doi.org/10.3835/plantgenome2012.07.0017.

Paape T, Heiniger B, Santo Domingo M, et al. (2022) Genome-wide association study reveals complex genetic architecture of cadmium and mercury accumulation and tolerance traits in *Medicago truncatula*. Front Plant Sci 12:806949

Pace J, Gardner C, Romay C, et al. (2015) Genome-wide association analysis of seedling root development in maize (*Zea mays* L.). BMC Genomics 16:47-47

Pang K, Li Y, Liu M, et al. (2013) Inventory and general analysis of the ATP-binding cassette (*ABC*) gene superfamily in maize (*Zea mays* L.). Gene 526:411-428

Chapter 6: References

Panuccio MR, Sorgonà A, Rizzo M, et al. (2009) Cadmium adsorption on vermiculite, zeolite and pumice: batch experimental studies. J Environ Manage 90:364-374

Parikh M, Maddaford TG, Austria JA, et al. (2019) Dietary flaxseed as a strategy for improving human health. Nutrients 11:5

Park J, Song WY, Ko D, et al. (2012) The phytochelatin transporters *AtABCC1* and *AtABCC2* mediate tolerance to cadmium and mercury. Plant J 69:278-288

Patel D, Franklin KA (2009) Temperature-regulation of plant architecture. Plant Signal Behav 4:577-579

Peng H, Chen Y, Weng L, et al. (2019) Comparisons of heavy metal input inventory in agricultural soils in North and South China: a review. Sci Total Environ 660:776-786

Pérez-Rodríguez P, Gianola D, González-Camacho JM, et al. (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3:Genes Genom Genet 2:1595-1605

Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483-495

Pérez P, de los Campos G, Crossa J, et al. (2010) Genomic-enabled prediction based on molecular markers and pedigree using the bayesian linear regression package in R. Plant Genome 3:2:https://doi.org/10.3835/plantgenome2010.3804.0005

Pierart A, Shahid M, Séjalon-Delmas N, et al. (2015) Antimony bioavailability: knowledge and research perspectives for sustainable agricultures. J Hazard Mater 289:219-234

Pighin JA, Zheng H, Balakshin LJ, et al. (2004) Plant cuticular lipid export requires an ABC transporter. Science 306:702-704

Chapter 6: References

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:https://doi.org/10.3835/plantgenome2012.05.0005.

Pottier M, Oomen R, Picco C, et al. (2015) Identification of mutations allowing natural resistance associated macrophage proteins (NRAMP) to discriminate against cadmium. Plant J 83:625-637

Powell N, Ji X, Ravash R, et al. (2012) Yield stability for cereals in a changing climate. Funct Plant Biol 39:539-552

Price AL, Patterson NJ, Plenge RM, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904-909

Qadir M, Quillérou E, Nangia V, et al. (2014) Economics of salt-induced land degradation and restoration. Nat Resour Forum 38:282-295

Qi X, Wang X, Xu J, et al. (2010) Drought-resistance evaluation of flax germplasm at adult plant stage. Sci Agric Sin 43:3076-3087

Qiao G, Zhang X, Jiang J, et al. (2014) Comparative proteomic analysis of responses to salt stress in chinese willow (*Salix matsudana* Koidz). Plant Mol Biol Rep 32:814-827

Qin D, Wang F, Geng X, et al. (2015) Overexpression of heat stress-responsive *TaMBF1c*, a wheat (*Triticum aestivum* L.) multiprotein bridging factor, confers heat tolerance in both yeast and rice. Plant Mol Biol 87:31-45

Rahaman M, Mamidi S, Rahman M (2018) Genome-wide association study of heat stress-tolerance traits in spring-type *Brassica napus* L. under controlled conditions. Crop J 6:115-125

Rees DC, Johnson E, Lewinson O (2009) ABC transporters: the power to change. Nat Rev Mol Cell Biol 10:218-227

Chapter 6: References

Remington DL, Thornsberry JM, Matsuoka Y, et al. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98:11479-11484

Revilla P, Rodríguez VM, Ordás A, et al. (2016) Association mapping for cold tolerance in two large maize inbred panels. BMC Plant Biol 16:127

Rivero RM, Ruiz JM, García PC, et al. (2001) Resistance to cold and heat stress: accumulation of phenolic compounds in tomato and watermelon plants. Plant Sci 160:315-321

Roberts TL (2014) Cadmium and phosphorous fertilizers: the issues and the science. Procedia Eng 83:52-59

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140

Rolling WR, Dorrance AE, McHale LK (2020) Testing methods and statistical models of genomic prediction for quantitative disease resistance to *Phytophthora sojae* in soybean [*Glycine max* (L.) Merr] germplasm collections. Theor Appl Genet 133:3441-3454

Rosyara UR, De Jong WS, Douches DS, et al. (2016) Software for genome-wide association studies in autopolyploids and its application to potato. Plant Genome 9:10.3835/plantgenome2015.3808.0073

Rutkoski J, Benson J, Jia Y, et al. (2012) Evaluation of genomic prediction methods for *Fusarium* head blight resistance in wheat. Plant Genome 5:10.3835/plantgenome2012.3802.0001

Safdar LB, Almas F, Sarfraz S, et al. (2020) Genome-wide association study identifies five new cadmium uptake loci in wheat. Plant Genome 13:2:https://doi.org/10.1002/tpg1002.20030

Chapter 6: References

Salla V, Hardaway CJ, Sneddon J (2011) Preliminary investigation of *Spartina alterniflora* for phytoextraction of selected heavy metals in soils from Southwest Louisiana. Microchem J 97:207-212

Sanghera GS, Wani SH, Hussain W, et al. (2011) Engineering cold stress tolerance in crop plants. Curr Genomics 12:30-43

Sasaki A, Yamaji N, Ma JF (2014) Overexpression of *OsHMA3* enhances Cd tolerance and expression of Zn transporter genes in rice. J Exp Bot 65:6013-6021

Savitch LV, Allard G, Seki M, et al. (2005) The effect of overexpression of two brassica CBF/DREB1-like transcription factors on photosynthetic capacity and freezing tolerance in *Brassica napus*. Plant Cell Physiol 46:1525-1539

Savojardo C, Martelli Pier L, Fariselli P, et al. (2018) BUSCA: an integrative web server to predict subcellular localization of proteins. Nucleic Acids Res 46:W459-W466

Schläppi MR, Jackson AK, Eizenga GC, et al. (2017) Assessment of five chilling tolerance traits and GWAS mapping in rice using the USDA mini-core collection. Front Plant Sci 8:957

Schneider E, Hunke S (1998) ATP-binding-cassette (ABC) transport systems: functional and structural aspects of the ATP-hydrolyzing subunits/domains. FEMS Microbiol Rev 22:1-20

Schöffl F, Prändl R, Reindl A (1998) Regulation of the heat-shock response. Plant Physiol 117:1135-1141

Schulz-Streeck T, Ogutu JO, Karaman Z, et al. (2012) Genomic selection using multiple populations. Crop Sci 52:2453-2461

Schulz E, Tohge T, Zuther E, et al. (2016) Flavonoids are determinants of freezing tolerance and cold acclimation in *Arabidopsis thaliana*. Sci Rep 6:34027

Chapter 6: References

Segraves KA (2017) The effects of genome duplications in a community context. New Phytol 215:57-69

Sertse D, You FM, Ravichandran S, et al. (2019) The complex genetic architecture of early root and shoot traits in flax revealed by genome-wide association analyses. Front Plant Sci 10:1483

Sertse D, You FM, Ravichandran S, et al. (2021) Loci harboring genes with important role in drought and related abiotic stress responses in flax revealed by multiple GWAS models. Theor Appl Genet 134:191-212

Shaban M, Ahmed MM, Sun H, et al. (2018) Genome-wide identification of lipoxygenase gene family in cotton and functional characterization in response to abiotic stresses. BMC Genomics 19:599

Shahid M, Khalid S, Abbas G, et al. (2015) Heavy metal stress and crop productivity. In: Hakeem KR (ed) Crop production and global environmental issues. Springer International Publishing, Cham, pp 1-25

Shannon P, Markiel A, Ozier O, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498-2504

Shao H-B, Chu L-Y, Jaleel CA, et al. (2009) Understanding water deficit stress-induced changes in the basic metabolism of higher plants – biotechnologically and sustainably improving agriculture and the ecoenvironment in arid regions of the globe. Crit Rev Biotech 29:131-151

Sharma JC TS, Shivran RK, Chandra P (2012) Water requirement water use efficiency consumptive use yield and quality parameters of linseed (*Linum usitatissimum* L.) varieties as influenced by fertility levels irrigation scheduling. Adv Life Sci 1:180–182

Chapter 6: References

Shazadee H, Khan N, Wang J, et al. (2019) Identification and expression profiling of protein phosphatases (*PP2C*) gene family in *Gossypium hirsutum* L. Int J Mol Sci 20:1395

Shikha M, Kanika A, Rao AR, et al. (2017) Genomic selection for drought tolerance using genome-wide SNPs in maize. Front Plant Sci 8:550

Shivaraj SM, Deshmukh RK, Rai R, et al. (2017) Genome-wide identification, characterization, and expression profile of aquaporin gene family in flax (*Linum usitatissimum*). Sci Rep 7:46137

Shuting Z, Hongwei D, Qing M, et al. (2022) Identification and expression analysis of the ZRT, IRT-like protein (ZIP) gene family in *Camellia sinensis* (L.) O. Kuntze. Plant Physiol and Biochem 172:87-100

Siberchicot A, Bessy A, Guéguen L, et al. (2017) MareyMap online: a user-friendly web application and database service for estimating recombination rates using physical and genetic maps. Genome Biol Evol 9:2506-2509

Sidler M, Hassa P, Hasan S, et al. (1998) Involvement of an ABC transporter in a developmental pathway regulating hypocotyl cell elongation in the light. Plant Cell 10:1623-1636

Sipos G, Kuchler K (2006) Fungal ATP-binding cassette (ABC) transporters in drug resistance & detoxification. Curr Drug Targets 7:471-481

Slama I, Abdelly C, Bouchereau A, et al. (2015) Diversity, distribution and roles of osmoprotective compounds accumulated in halophytes under abiotic stress. Ann Bot 115:433-447

Song J, Li J, Sun J, et al. (2018) Genome-wide association mapping for cold tolerance in a core collection of rice (*Oryza sativa* L.) landraces by using high-density single nucleotide polymorphism markers from specific-locus amplified fragment sequencing. Front Plant Sci 9:875

Chapter 6: References

Song W-Y, Yamaki T, Yamaji N, et al. (2014) A rice ABC transporter, *OsABCC1*, reduces arsenic accumulation in the grain. Proc Natl Acad Sci USA 111:15699-15704

Soto-Cerda BJ, Aravena G, Cloutier S (2021) Genetic dissection of flowering time in flax (*Linum usitatissimum* L.) through single- and multi-locus genome-wide association studies. Mol Genet Genomics 296:877-891

Soto-Cerda BJ, Cloutier S, Gajardo HA, et al. (2019) Identifying drought-resilient flax genotypes and related-candidate genes based on stress indices, root traits and selective sweep. Euphytica 215:41

Soto-Cerda BJ, Cloutier S, Gajardo HA, et al. (2020) Drought response of flax accessions and identification of quantitative trait nucleotides (QTNs) governing agronomic and root traits by genome-wide association analysis. Mol Breed 40:15

Soto-Cerda BJ, Cloutier S, Quian R, et al. (2018a) Genome-wide association analysis of mucilage and hull content in flax (*Linum usitatissimum* L.) seeds. Int J Mol Sci 19

Soto-Cerda BJ, Cloutier S, Quian R, et al. (2018b) Genome-wide association analysis of mucilage and hull content in flax (*Linum usitatissimum* L.) seeds. Int J Mol Sci 19:2870

Soto-Cerda BJ, Diederichsen A, Ragupathy R, et al. (2013) Genetic characterization of a core collection of flax (*Linum usitatissimum* L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. BMC Plant Biol 13:78

Soto-Cerda BJ, Duguid S, Booker H, et al. (2014) Association mapping of seed quality traits using the Canadian flax (*Linum usitatissimum* L.) core collection. Theor Appl Genet 127:881-896

Spiertz H (2014) Agricultural sciences in transition from 1800 to 2020: exploring knowledge and creating impact. Eur J Agron 59:96-106

Chapter 6: References

Spindel J, Begum H, Akdemir D, et al. (2015) Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet 11:e1004982

Su J, Zhang F, Chong X, et al. (2019) Genome-wide association study identifies favorable SNP alleles and candidate genes for waterlogging tolerance in chrysanthemums. Hort Res 6:21

Sukumaran S, Reynolds MP, Sansaloni C (2018) Genome-wide association analyses identify QTL hotspots for yield and component traits in durum wheat grown under yield potential, drought, and heat stress environments. Front Plant Sci 9:81

Suman J, Uhlik O, Viktorova J, et al. (2018) Phytoextraction of heavy metals: a promising tool for clean-up of polluted environment? Front Plant Sci 9:1476

Sun C, Wang B, Wang X, et al. (2016) Genome-wide association study dissecting the genetic architecture underlying the branch angle trait in rapeseed (*Brassica napus* L.). Sci Rep 6:1-11

Sun X, Wang Y, Sui N (2018) Transcriptional regulation of bHLH during plant response to stress. Biochem Biophys Res Commun 503:397-401

Sun Y, Wang C, Chen HYH, et al. (2020) Response of plants to water stress: a meta-analysis. Front Plant Sci 11:00978

Szklarczyk D, Gable AL, Lyon D, et al. (2018) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47:D607-D613

Chapter 6: References

Tagashira Y, Shimizu T, Miyamoto M, et al. (2015) Overexpression of a gene involved in phytic acid biosynthesis substantially increases phytic acid and total phosphorus in rice seeds. Plants 4:196-208

Takahashi R, Bashir K, Ishimaru Y, et al. (2012a) The role of heavy-metal ATPases, HMAs, in zinc and cadmium transport in rice. Plant Signal Behav 7:1605-1607

Takahashi R, Ishimaru Y, Shimo H, et al. (2012b) The *OsHMA2* transporter is involved in root-to-shoot translocation of Zn and Cd in rice. Plant Cell Environ 35:1948-1957

Takeda S, Iwasaki A, Tatematsu K, et al. (2014) The half-size abc transporter folded petals 2/ABCG13 is involved in petal elongation through narrow spaces in *Arabidopsis thaliana* floral buds. Plants 3:348-358

Tam V, Patel N, Turcotte M, et al. (2019) Benefits and limitations of genome-wide association studies. Nat Rev Genet 20:467-484

Tan J, Wang J, Chai T, et al. (2013) Functional analyses of *TaHMA2*, a P_{1B}-type ATPase in wheat. Plant Biotechnol J 11:420-431

Tang Y, Bao X, Zhi Y, et al. (2019) Overexpression of a MYB family gene, *OsMYB6*, increases drought and salinity stress tolerance in transgenic rice. Front Plant Sci 10:168

Tang Y, Liu X (2019) G2P: a genome-wide-association-study simulation tool for genotype simulation, phenotype simulation and power evaluation. Bioinformatics 35:3852-3854

Tanhuanpää P, Kalendar R, Schulman AH, et al. (2007) A major gene for grain cadmium accumulation in oat (*Avena sativa* L.). Genome 50:588-594

Tchoumtchoua J, Mathiron D, Pontarin N, et al. (2019) Phenolic profiling of flax highlights contrasting patterns in winter and spring varieties. Molecules 24:4303

Chapter 6: References

Technow F, Riedelsheimer C, Schrag TA, et al. (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125:1181-1194

Tembo BD, Sichilongo K, Cernak J (2006) Distribution of copper, lead, cadmium and zinc concentrations in soils around Kabwe town in Zambia. Chemosphere 63:497-501

Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science (New York, NY) 327:818-822

Thakur P, Kumar S, Malik JA, et al. (2010) Cold stress effects on reproductive development in grain crops: an overview. Environ Exp 67:429-443

Thavamanikumar S, Dolferus R, Thumma BR (2015) Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. G3: Genes Genom Genet 5:1991-1998

Ueno D, Koyama E, Yamaji N, et al. (2011) Physiological, genetic, and molecular characterization of a high-Cd-accumulating rice cultivar, Jarjan. J Exp Bot 62:2265-2272

Ueno D, Yamaji N, Kono I, et al. (2010) Gene limiting cadmium accumulation in rice. Proc Natl Acad Sci USA 107:16500-16505

Vallee BL, Ulmer DD (1972) Biochemical effects of mercury, cadmium, and lead. Annu Rev Biochem 41:91-128

Velazco JG, Jordan DR, Mace ES, et al. (2019) Genomic prediction of grain yield and drought-adaptation capacity in sorghum is enhanced by multi-trait analysis. Front Plant Sci 10:997

Venglat P, Xiang D, Qiu S, et al. (2011) Gene expression analysis of flax seed development. BMC Plant Biol 11:74

Chapter 6: References

Verbruggen N, Hermans C, Schat H (2009) Mechanisms to cope with arsenic or cadmium excess in plants. Curr Opin Plant Biol 12:364-372

Verma D, Lakhanpal N, Singh K (2019) Genome-wide identification and characterization of abiotic-stress responsive SOD (superoxide dismutase) gene family in *Brassica juncea* and *B. rapa*. BMC Genomics 20:227

Verret F, Gravot A, Auroy P, et al. (2004) Overexpression of *AtHMA4* enhances root-to-shoot translocation of zinc and cadmium and plant metal tolerance. FEBS Lett 576:306-312

Verrier PJ, Bird D, Burla B, et al. (2008) Plant ABC proteins – a unified nomenclature and updated inventory. Trends Plant Sci 13:151-159

Visconti S, D'Ambrosio C, Fiorillo A, et al. (2019) Overexpression of 14-3-3 proteins enhances cold tolerance and increases levels of stress-responsive proteins of *Arabidopsis* plants. Plant Sci 289:110215

Visioni A, Tondelli A, Francia E, et al. (2013) Genome-wide association mapping of frost tolerance in barley (*Hordeum vulgare* L.). BMC Genomics 14:424

Voorrips RE, Maliepaard CA (2012) The simulation of meiosis in diploid and tetraploid organisms using various genetic models. BMC Bioinform 13:248

Voss-Fels K, Snowdon RJ (2016) Understanding and utilizing crop genome diversity via high-resolution genotyping. Plant Biotech J 14:1086-1094

Wang D, Liu J, Li C, et al. (2016a) Genome-wide association mapping of cold tolerance genes at the seedling stage in rice. Rice 9:61

Wang H, Wei J, Li P, et al. (2019a) Integrating GWAS and gene expression analysis identifies candidate genes for root morphology traits in maize at the seedling stage. Genes 10:773

Chapter 6: References

Wang J, Van Ginkel M, Podlich D, et al. (2003) Comparison of two breeding strategies by computer simulation. Crop Sci 43:1764-1773

Wang L, Liu Y, Li D, et al. (2019b) Improving salt tolerance in potato through overexpression of *AtHKT1* gene. BMC Plant Biol 19:357

Wang M-q, Huang Q-x, Lin P, et al. (2020) The overexpression of a transcription factor gene *VbWRKY32* enhances the cold tolerance in *Verbena bonariensis*. Front Plant Sci 10:1746

Wang N, Liu B, Liang X, et al. (2019c) Genome-wide association study and genomic prediction analyses of drought stress tolerance in China in a collection of off-PVP maize inbred lines. Mol Breed 39:113

Wang S-B, Feng J-Y, Ren W-L, et al. (2016b) Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Sci Rep 6:19444

Wang W, Vinocur B, Shoseyov O, et al. (2004) Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. Trends Plant Sci 9:244-252

Wang Y, Tang H, Debarry JD, et al. (2012a) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40:e49-e49

Wang Z, Hobson N, Galindo L, et al. (2012b) The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. Plant J 72:461-473

Wani SH, Kumar V, Shriram V, et al. (2016) Phytohormones and their metabolic engineering for abiotic stress tolerance in crop plants. Crop J 4:162-176

Wasson AP, Rebetzke GJ, Kirkegaard JA, et al. (2014) Soil coring at multiple field environments can directly quantify variation in deep root traits to select wheat genotypes for breeding. J Exp Bot 65:6231-6249

Chapter 6: References

Wehner G, Balko C, Ordon F (2016) QTL for water use related traits in juvenile barley. Agron 6:62

Westhues CC, Simianer H, Beissinger TM (2021) learnMET: an r package to apply machine learning methods for genomic prediction using multi-environment trial data. bioRxiv:2021.2012.2013.472185

White PJ, Brown PH (2010) Plant nutrition for sustainable development and global health. Ann Bot 105:1073-1080

Windhausen VS, Atlin GN, Hickey JM, et al. (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3: Genes Genom Genet 2:1427-1436

Wu D, Sato K, Ma JF (2015) Genome-wide association mapping of cadmium accumulation in different organs of barley. New Phytol 208:817-829

Wu P, Wang W, Li Y, et al. (2017) Divergent evolutionary patterns of the MAPK cascade genes in *Brassica rapa* and plant phylogenetics. Hort Res 4:17079

Wu R, Ma C, Casella G (2007) Statistical genetics of quantitative traits: linkage, maps and QTL. Springer Science & Business Media

Xia H, Zheng X, Chen L, et al. (2014) Genetic differentiation revealed by selective loci of drought-responding EST-SSRs between upland and lowland rice in China. PLoS One 9:e106352

Xie Z, Nolan TM, Jiang H, et al. (2019) AP2/ERF transcription factor regulatory networks in hormone and abiotic stress responses in *Arabidopsis*. Front Plant Sci 10:228

Xiong T, Leveque T, Shahid M, et al. (2014) Lead and cadmium phytoavailability and human bioaccessibility for vegetables exposed to soil or atmospheric pollution by process ultrafine particles. J Environ Qual 43:1593-1600

Chapter 6: References

Xu Y, Wang X, Ding X, et al. (2018) Genomic selection of agronomic traits in hybrid rice using an NCII population. Rice 11:32

Yamaji N, Xia J, Mitani-Ueno N, et al. (2013) Preferential delivery of zinc to developing tissues in rice is mediated by P-type heavy metal ATPase *OsHMA2*. Plant Physiol 162:927-939

Yan C, Duan W, Lyu S, et al. (2017) Genome-wide identification, evolution, and expression analysis of the ATP-binding cassette transporter gene family in *Brassica rapa*. Front Plant Sci 8:349

Yan H, Li M, Xiong Y, et al. (2019) Genome-wide characterization, expression profile analysis of WRKY family genes in *Santalum album* and functional identification of their role in abiotic stress. Int J Mol Sci 20:5676

Yang H, Deng L, Liu H, et al. (2019) Overexpression of *BnaAOX1b* confers tolerance to osmotic and salt stress in rapeseed. G3: Genes Genom Genet 9:3501-3511

Yao J, Zhao D, Chen X, et al. (2018) Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (*Triticum aestivum* L.). Crop J 6:353-365

Ye H, Du H, Tang N, et al. (2009) Identification and expression profiling analysis of TIFY family genes involved in stress and phytohormone responses in rice. Plant Mol Biol 71:291-305

Yokotani N, Ichikawa T, Kondou Y, et al. (2008) Expression of rice heat stress transcription factor *OsHsfA2e* enhances tolerance to environmental stresses in transgenic *Arabidopsis*. Planta 227:957-967

Yonis BO, Pino Del Carpio D, Wolfe M, et al. (2020) Improving root characterisation for genomic prediction in cassava. Sci Rep 10:8003

Chapter 6: References

Yoon DH, Lee SS, Park HJ, et al. (2016) Overexpression of *OsCYP19-4* increases tolerance to cold stress and enhances grain yield in rice (*Oryza sativa*). J Exp Bot 67:69-82

You F, Duguid S, Thambugala D, et al. (2013) Statistical analysis and field evaluation of the type 2 modified augmented design (MAD) in phenotyping of flax (*Linum usitatissimum*) germplasm in multiple environments. Aust J Crop Sci 7:1789-1800

You FM, Deal KR, Wang J, et al. (2012) Genome-wide SNP discovery in walnut with an AGSNP pipeline updated for SNP discovery in allogamous organisms. BMC Genomics 13:354

You FM, Duguid SD, Lam I, et al. (2016a) Pedigrees and genetic base of flax cultivars registered in Canada. Can J Plant Sci 96:837-852

You FM, Gaofeng J, Helen MB, et al. (2016b) A method of estimating broad-sense heritability for quantitative traits in the type 2 modified augmented design. J Plant Breed Crop Sci 8:257-272

You FM, Jia G, Xiao J, et al. (2017) Genetic variability of 27 traits in a core collection of flax (*Linum usitatissimum* L.). Front Plant Sci 8:1636

You FM, Rashid KY, Zheng C, et al. (2022a) Insights into the genetic architecture and genomic prediction of powdery mildew resistance in flax (*Linum usitatissimum* L.). Int J Mol Sci 23:9

You FM, Song Q, Jia G, et al. (2016c) Estimation of genetic parameters and their sampling variances for quantitative traits in the type 2 modified augmented design. Crop J 4:107-118

You FM, Xiao J, Li P, et al. (2018a) Genome-wide association study and selection signatures detect genomic regions associated with seed yield and oil quality in flax. Int J Mol Sci 19:8

You FM, Xiao J, Li P, et al. (2018b) Chromosome-scale pseudomolecules refined by optical, physical and genetic maps in flax. Plant J 95:371-384

Chapter 6: References

You FM, Zheng C, Bartaula S, et al. (2022b) Genomic cross prediction for linseed improvement. In: Gosal SS, Wani SH (eds) Accelerated Plant Breeding, Volume 4: Oil Crops. Springer International Publishing, Cham, pp 451-480

Yu J, Liu C, Lin H, et al. (2021) Loci and natural alleles for cadmium-mediated growth responses revealed by a genome wide association study and transcriptome analysis in rice. BMC Plant Biol 21:374

Yu J, Pressoir G, Briggs WH, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203-208

Yu Y, Huang W, Chen H, et al. (2014) Identification of differentially expressed genes in flax (*Linum usitatissimum* L.) under saline–alkaline stress by digital gene expression. Gene 549:113-122

Yu Y, Ouyang Y, Yao W (2017) shinyCircos: an R/Shiny application for interactive creation of circos plot. Bioinformatics 34:1229-1231

Yuan L, Yang S, Liu B, et al. (2012) Molecular characterization of a rice metal tolerance protein, *OsMTP1*. Plant Cell Rep 31:67-79

Yuan Y, Cairns JE, Babu R, et al. (2018) Genome-wide association mapping and genomic prediction analyses reveal the genetic architecture of grain yield and flowering time under drought and heat stress conditions in maize. Front Plant Sci 9:1919

Zang X, Geng X, Wang F, et al. (2017) Overexpression of wheat ferritin gene *TaFER-5B* enhances tolerance to heat stress and other abiotic stresses associated with the ROS scavenging. BMC Plant Biol 17:14

Chapter 6: References

Zhang A, Wang H, Beyene Y, et al. (2017) Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. Front Plant Sci 8:01916

Zhang D, Han Z, Li J, et al. (2020a) Genome-wide analysis of the SBP-box gene family transcription factors and their responses to abiotic stresses in tea (*Camellia sinensis*). Genomics 112:2194-2202

Zhang F, Xiao X, Xu K, et al. (2020b) Genome-wide association study (GWAS) reveals genetic loci of lead (Pb) tolerance during seedling establishment in rapeseed (*Brassica napus* L.). BMC Genomics 21:139

Zhang J-B, Wang X-P, Wang Y-C, et al. (2020c) Genome-wide identification and functional characterization of cotton (*Gossypium hirsutum*) *MAPKKK* gene family in response to drought stress. BMC Plant Biol 20:217

Zhang K, Cui H, Cao S, et al. (2019a) Overexpression of *CrCOMT* from *Carex rigescens* increases salt stress and modulates melatonin synthesis in *Arabidopsis thaliana*. Plant Cell Rep 38:1501-1514

Zhang M, Ye J, Xu Q, et al. (2018a) Genome-wide association study of cold tolerance of Chinese *Indica* rice varieties at the bud burst stage. Plant Cell Rep 37:529-539

Zhang X, Tang B, Yu F, et al. (2013) Identification of major QTL for waterlogging tolerance using genome-wide association and linkage mapping of maize seedlings. Plant Mol Biol Rep 31:594-606

Zhang Y, Liu P, Zhang X, et al. (2018b) Multi-locus genome-wide association study reveals the genetic architecture of stalk lodging resistance-related traits in maize. Front Plant Sci 9:10.3389/fpls.2018.00611

Chapter 6: References

Zhang YM, Jia Z, Dunwell JM (2019b) Editorial: The applications of new multi-locus gwas methodologies in the genetic dissection of complex traits. Front Plant Sci 10:100

Zhao J, Yang W, Zhang S, et al. (2018) Genome-wide association study and candidate gene analysis of rice cadmium accumulation in grain in a diverse rice collection. Rice 11:61

Zhao K-F, Song J, Fan H, et al. (2010) Growth response to ionic and osmotic stress of NaCl in salt-tolerant and salt-sensitive maize. J Integr Plant Biol 52:468-475

Zhao T, Dai AJJoc (2015) The magnitude and causes of global drought changes in the twenty-first century under a low–moderate emissions scenario. J Clim 28:4490-4512

Zhao Y, Mette MF, Reif JC (2015) Genomic selection in hybrid breeding. Plant Breed 134:1-10

Zhong H, Liu S, Sun T, et al. (2021) Multi-locus genome-wide association studies for five yield-related traits in rice. BMC Plant Biol 21:364

Zhong S, Jannink J-L (2007) Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. Genetics 177:567-576

Zhou JC, Fu TT, Sui N, et al. (2016) The role of salinity in seed maturation of the euhalophyte *Suaeda salsa*. Plant Biosyst 150:83-90

Zhu JK (2000) Genetic analysis of plant salt tolerance using *Arabidopsis*. Plant Physiol 124:941-948

Zhu JK (2016) Abiotic stress signaling and responses in plants. Cell 167:313-324

Zingaretti ML, Monfort A, Pérez-Enciso M (2019) pSBVB: a versatile simulation tool to evaluate genomic selection in polyploid species. G3: Genes Genom Genet 9:327-334

Chapter 7: Appendices

7.0 Appendices

7.1 Supplementary Materials for Chapter 3

Appendix 3.1 List of heavy metal associated (HMA) genes identified in different species.

Species	Subfamilies	Gene ID
Linum usitatissimum	I	*Lus10011510*
	I	*Lus10019313*
	II	*Lus10038364*
	II	*Lus10036225*
	II	*Lus10031273*
	II	*Lus10031840*
	II	*Lus10008309*
	III	*Lus10009789*
	III	*Lus10038070*
	III	*Lus10023777*
	IV	*Lus10025467*
	IV	*Lus10006955*
Arabidopsis thaliana	I	*AT4G37270*
	II	*AT1G63440*
	II	*AT5G44790*
	III	*AT4G33520*
	III	*AT5G21930*
	IV	*AT4G30110*
	IV	*AT4G30120*
	IV	*AT2G19110*
Populus trichocarpa	I	*Potri 007G049000*
	II	*Potri 003G075700*
	II	*Potri 001G158900*
	II	*Potri 001G019100*
	II	*Potri 003G125700*
	II	*Potri 003G125600*
	II	*Potri 001G105800*
	III	*Potri 006G220201*
	III	*Potri 018G047800*
	III	*Potri 003G024000*
	III	*Potri 001G205400*
	IV	*Potri 006G076900*
	I	*GSVIVT01000217001*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	II	*GSVIVT01012131001*
Vitis vinifera	II	*GSVIVT01025299001*
	II	*GSVIVT01019769001*
	III	*GSVIVT01035474001*
	III	*GSVIVT01038740001*
	III	*GSVIVT01038742001*
	IV	*GSVIVT01010896001*
	I	*Bradi1g33347*
	II	*Bradi1g31987*
Brachypodium distachyon	II	*Bradi3g05340*
	II	*Bradi3g07110*
	II	*Bradi5g17990*
	III	*Bradi1g72790*
	III	*Bradi3g38790*
	IV	*Bradi1g34140*
	IV	*Bradi1g53670*

Appendix 3.2 is available in the supplementary materials (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-020-07121-9/MediaObjects/12864_2020_7121_MOESM2_ESM.xlsx).

Appendix 3.3 is available in the supplementary materials (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-020-07121-9/MediaObjects/12864_2020_7121_MOESM3_ESM.xlsx).

Appendix 3.4 List of ATP binding cassette (ABC) transporter genes identified in different species.

Species	Subfamilies	Gene ID
	ABCA	*Lus10012755*
	ABCA	*Lus10023157*
	ABCA	*Lus10023613*
	ABCA	*Lus10024244*
	ABCA	*Lus10004262*
	ABCA	*Lus10042172*
	ABCA	*Lus10023612*
	ABCA	*Lus10024243*
	ABCB	*Lus10038049*
	ABCB	*Lus10009988*
	ABCB	*Lus10010068*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCB	*Lus10010067*
	ABCB	*Lus10009989*
Linum usitatissimum	ABCB	*Lus10038050*
	ABCB	*Lus10004583*
	ABCB	*Lus10010074*
	ABCB	*Lus10010012*
	ABCB	*Lus10004520*
	ABCB	*Lus10005249*
	ABCB	*Lus10030674*
	ABCB	*Lus10023929*
	ABCB	*Lus10017270*
	ABCB	*Lus10017271*
	ABCB	*Lus10005839*
	ABCB	*Lus10041565*
	ABCB	*Lus10039458*
	ABCB	*Lus10020905*
	ABCB	*Lus10033470*
	ABCB	*Lus10008139*
	ABCB	*Lus10024162*
	ABCB	*Lus10039533*
	ABCB	*Lus10004571*
	ABCB	*Lus10012959*
	ABCB	*Lus10035834*
	ABCB	*Lus10043069*
	ABCB	*Lus10014427*
	ABCB	*Lus10032911*
	ABCB	*Lus10015595*
	ABCB	*Lus10013178*
	ABCB	*Lus10004530*
	ABCB	*Lus10004519*
	ABCB	*Lus10036617*
	ABCB	*Lus10040315*
	ABCB	*Lus10023437*
	ABCB	*Lus10025032*
	ABCB	*Lus10000910*
	ABCB	*Lus10036616*
	ABCB	*Lus10036857*
	ABCB	*Lus10006209*
	ABCB	*Lus10022270*
	ABCB	*Lus10007987*
	ABCB	*Lus10038886*
	ABCB	*Lus10015007*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCB	*Lus10017024*
	ABCB	*Lus10021348*
	ABCB	*Lus10021145*
	ABCC	*Lus10022858*
	ABCC	*Lus10024965*
	ABCC	*Lus10018027*
	ABCC	*Lus10042023*
	ABCC	*Lus10013201*
	ABCC	*Lus10027567*
	ABCC	*Lus10039320*
	ABCC	*Lus10004896*
	ABCC	*Lus10011498*
	ABCC	*Lus10017548*
	ABCC	*Lus10023139*
	ABCC	*Lus10030716*
	ABCC	*Lus10030718*
	ABCC	*Lus10023275*
	ABCC	*Lus10038530*
	ABCC	*Lus10036153*
	ABCC	*Lus10041661*
	ABCC	*Lus10022808*
	ABCC	*Lus10030717*
	ABCD	*Lus10001758*
	ABCD	*Lus10001839*
	ABCD	*Lus10024056*
	ABCD	*Lus10041690*
	ABCD	*Lus10041691*
	ABCE	*Lus10034771*
	ABCE	*Lus10033309*
	ABCF	*Lus10023043*
	ABCF	*Lus10032431*
	ABCF	*Lus10038681*
	ABCF	*Lus10029363*
	ABCF	*Lus10016184*
	ABCF	*Lus10028810*
	ABCF	*Lus10005798*
	ABCF	*Lus10006804*
	ABCF	*Lus10037947*
	ABCG	*Lus10025824*
	ABCG	*Lus10014150*
	ABCG	*Lus10037562*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCG	*Lus10041668*
	ABCG	*Lus10024061*
	ABCG	*Lus10008090*
	ABCG	*Lus10024060*
	ABCG	*Lus10019453*
	ABCG	*Lus10034020*
	ABCG	*Lus10041683*
	ABCG	*Lus10008984*
	ABCG	*Lus10024062*
	ABCG	*Lus10018230*
	ABCG	*Lus10020550*
	ABCG	*Lus10006277*
	ABCG	*Lus10020568*
	ABCG	*Lus10041680*
	ABCG	*Lus10024063*
	ABCG	*Lus10014384*
	ABCG	*Lus10028841*
	ABCG	*Lus10023879*
	ABCG	*Lus10041685*
	ABCG	*Lus10041681*
	ABCG	*Lus10027725*
	ABCG	*Lus10013123*
	ABCG	*Lus10027727*
	ABCG	*Lus10041684*
	ABCG	*Lus10023878*
	ABCG	*Lus10012508*
	ABCG	*Lus10000318*
	ABCG	*Lus10040644*
	ABCG	*Lus10018268*
	ABCG	*Lus10009961*
	ABCG	*Lus10027354*
	ABCG	*Lus10023154*
	ABCG	*Lus10009959*
	ABCG	*Lus10009960*
	ABCG	*Lus10035445*
	ABCG	*Lus10031063*
	ABCG	*Lus10000338*
	ABCG	*Lus10037384*
	ABCG	*Lus10021440*
	ABCG	*Lus10034775*
	ABCG	*Lus10035692*
	ABCG	*Lus10037283*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCG	*Lus10033314*
	ABCG	*Lus10034387*
	ABCG	*Lus10025597*
	ABCG	*Lus10025596*
	ABCG	*Lus10029635*
	ABCG	*Lus10015994*
	ABCG	*Lus10028704*
	ABCG	*Lus10033315*
	ABCG	*Lus10016649*
	ABCG	*Lus10022564*
	ABCG	*Lus10012284*
	ABCG	*Lus10000339*
	ABCG	*Lus10027353*
	ABCG	*Lus10029905*
	ABCG	*Lus10000737*
	ABCG	*Lus10004496*
	ABCG	*Lus10008500*
	ABCG	*Lus10017683*
	ABCG	*Lus10033637*
	ABCG	*Lus10027545*
	ABCG	*Lus10038714*
	ABCG	*Lus10037981*
	ABCG	*Lus10016115*
	ABCG	*Lus10041325*
	ABCG	*Lus10023020*
	ABCG	*Lus10021448*
	ABCG	*Lus10023199*
	ABCG	*Lus10009400*
	ABCG	*Lus10006046*
	ABCG	*Lus10042949*
	ABCG	*Lus10032449*
	ABCG	*Lus10023465*
	ABCG	*Lus10043305*
	ABCG	*Lus10016125*
	ABCG	*Lus10041333*
	ABCG	*Lus10025281*
	ABCG	*Lus10030306*
	ABCG	*Lus10001972*
	ABCG	*Lus10040341*
	ABCG	*Lus10039304*
	ABCI	*Lus10033908*
	ABCI	*Lus10024914*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCI	*Lus10022909*
	ABCI	*Lus10020258*
	ABCI	*Lus10002636*
	ABCI	*Lus10017374*
	ABCI	*Lus10010168*
	ABCI	*Lus10006727*
	ABCI	*Lus10022692*
	ABCI	*Lus10020188*
	ABCI	*Lus10026992*
	ABCI	*Lus10034152*
	ABCI	*Lus10043429*
	ABCI	*Lus10007420*
	ABCI	*Lus10019360*
	ABCI	*Lus10001117*
	ABCI	*Lus10025788*
	ABCI	*Lus10035874*
	ABCI	*Lus10034750*
	ABCI	*Lus10003971*
	ABCI	*Lus10023792*
	ABCI	*Lus10010838*
	ABCI	*Lus10035874*
	ABCA	*AT2G41700*
	ABCA	*AT3G47730*
	ABCA	*AT3G47740*
	ABCA	*AT3G47750*
	ABCA	*AT3G47760*
	ABCA	*AT3G47770*
	ABCA	*AT3G47780*
	ABCA	*AT3G47790*
	ABCA	*AT5G61730*
	ABCA	*AT5G61740*
	ABCA	*AT5G61690*
	ABCA	*AT5G61700*
	ABCB	*AT2G36910*
	ABCB	*AT4G25960*
	ABCB	*AT4G01820*
Arabidopsis thaliana	ABCB	*AT2G47000*
	ABCB	*AT4G01830*
	ABCB	*AT2G39480*
	ABCB	*AT5G46540*
	ABCB	*AT4G18050*
	ABCB	*AT1G10680*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCB	*AT1G02520*
	ABCB	*AT1G02530*
	ABCB	*AT1G27940*
	ABCB	*AT1G28010*
	ABCB	*AT3G28345*
	ABCB	*AT3G28360*
	ABCB	*AT3G28380*
	ABCB	*AT3G28390*
	ABCB	*AT3G28860*
	ABCB	*AT3G55320*
	ABCB	*AT3G62150*
	ABCB	*AT3G28415*
	ABCB	*AT4G28630*
	ABCB	*AT4G28620*
	ABCB	*AT5G58270*
	ABCB	*AT1G70610*
	ABCB	*AT5G39040*
	ABCB	*AT4G25450*
	ABCB	*AT5G03910*
	ABCC	*AT1G30400*
	ABCC	*AT2G34660*
	ABCC	*AT3G13080*
	ABCC	*AT2G47800*
	ABCC	*AT1G04120*
	ABCC	*AT3G13090*
	ABCC	*AT3G13100*
	ABCC	*AT3G21250*
	ABCC	*AT3G60160*
	ABCC	*AT3G59140*
	ABCC	*AT1G30420*
	ABCC	*AT1G30410*
	ABCC	*AT2G07680*
	ABCC	*AT3G62700*
	ABCC	*AT3G60970*
	ABCD	*AT1G54350*
	ABCD	*AT4G39850*
	ABCE	*AT4G19210*
	ABCE	*AT3G13640*
	ABCE	*AT4G30300*
	ABCF	*AT5G60790*
	ABCF	*AT5G09930*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCF	*AT1G64550*
	ABCF	*AT3G54540*
	ABCF	*AT5G64840*
	ABCG	*AT2G39350*
	ABCG	*AT2G37360*
	ABCG	*AT2G28070*
	ABCG	*AT4G25750*
	ABCG	*AT2G13610*
	ABCG	*AT5G13580*
	ABCG	*AT2G01320*
	ABCG	*AT5G52860*
	ABCG	*AT4G27420*
	ABCG	*AT1G53270*
	ABCG	*AT1G17840*
	ABCG	*AT1G51500*
	ABCG	*AT1G51460*
	ABCG	*AT1G31770*
	ABCG	*AT3G21090*
	ABCG	*AT3G55090*
	ABCG	*AT3G55100*
	ABCG	*AT3G55110*
	ABCG	*AT3G55130*
	ABCG	*AT3G53510*
	ABCG	*AT3G25620*
	ABCG	*AT5G06530*
	ABCG	*AT5G19410*
	ABCG	*AT1G53390*
	ABCG	*AT1G71960*
	ABCG	*AT3G13220*
	ABCG	*AT3G52310*
	ABCG	*AT5G60740*
	ABCG	*AT3G16340*
	ABCG	*AT4G15230*
	ABCG	*AT2G29940*
	ABCG	*AT2G26910*
	ABCG	*AT2G37280*
	ABCG	*AT2G36380*
	ABCG	*AT1G15210*
	ABCG	*AT1G59870*
	ABCG	*AT3G53480*
	ABCG	*AT3G30842*
	ABCG	*AT1G66950*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCG	*AT1G15520*
	ABCG	*AT4G15215*
	ABCG	*AT4G15233*
	ABCG	*AT4G15236*
	ABCI	*AT1G63270*
	ABCI	*AT5G14100*
	ABCI	*AT4G33460*
	ABCI	*AT1G65410*
	ABCI	*AT1G67940*
	ABCI	*AT5G02270*
	ABCI	*AT1G03905*
	ABCI	*AT5G44110*
	ABCI	*AT3G10670*
	ABCI	*AT1G03900*
	ABCI	*AT3G21580*
	ABCI	*AT3G20320*
	ABCI	*AT2G07771*
	ABCI	*ATMG00900*
	ABCI	*AT2G07681*
	ABCI	*ATMG00110*
	ABCI	*AT1G19800*
	ABCI	*AT2G37300*
	ABCI	*AT1G32500*
	ABCI	*AT4G04770*
	ABCI	*AT5G44316*
	ABCA	*Potri.006G049300*
	ABCA	*Potri.012G069800*
	ABCA	*Potri.015G063500*
	ABCA	*Potri.012G069700*
	ABCA	*Potri.015G063400*
	ABCB	*Potri.013G160500*
	ABCB	*Potri.019G132800*
	ABCB	*Potri.006G211700*
	ABCB	*Potri.015G059700*
	ABCB	*Potri.015G059800*
	ABCB	*Potri.015G059600*
	ABCB	*Potri.018G051800*
	ABCB	*Potri.017G097200*
	ABCB	*Potri.015G000200*
	ABCB	*Potri.004G123000*
	ABCB	*Potri.012G004200*
	ABCB	*Potri.010G045900*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCB	*Potri.017G074000*
	ABCB	*Potri.017G085000*
Populus trichocarpa	ABCB	*Potri.018G141300*
	ABCB	*Potri.006G074000*
	ABCB	*Potri.006G074400*
	ABCB	*Potri.018G141400*
	ABCB	*Potri.018G087100*
	ABCB	*Potri.001G139600*
	ABCB	*Potri.003G094400*
	ABCB	*Potri.008G050400*
	ABCB	*Potri.010G210100*
	ABCB	*Potri.001G417600*
	ABCB	*Potri.011G133800*
	ABCB	*Potri.006G074100*
	ABCB	*Potri.002G019600*
	ABCB	*Potri.017G081100*
	ABCB	*Potri.006G123900*
	ABCB	*Potri.016G093600*
	ABCB	*Potri.001G165000*
	ABCB	*Potri.001G139700*
	ABCB	*Potri.001G354900*
	ABCB	*Potri.002G187500*
	ABCB	*Potri.014G113100*
	ABCB	*Potri.014G113200*
	ABCB	*Potri.014G113500*
	ABCB	*Potri.010G003000*
	ABCB	*Potri.002G187400*
	ABCB	*Potri.014G113000*
	ABCC	*Potri.003G197300*
	ABCC	*Potri.003G198400*
	ABCC	*Potri.001G362300*
	ABCC	*Potri.001G362200*
	ABCC	*Potri.002G255800*
	ABCC	*Potri.014G193400*
	ABCC	*Potri.008G179500*
	ABCC	*Potri.010G055200*
	ABCC	*Potri.018G149600*
	ABCC	*Potri.014G180100*
	ABCC	*Potri.003G135800*
	ABCC	*Potri.001G095600*
	ABCC	*Potri.012G033600*
	ABCC	*Potri.016G123000*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCC	*Potri.012G033400*
	ABCC	*Potri.012G033500*
	ABCC	*Potri.015G023800*
	ABCC	*Potri.002G205500*
	ABCC	*Potri.014G130500*
	ABCC	*Potri.003G034700*
	ABCC	*Potri.003G034800*
	ABCC	*Potri.011G091200*
	ABCC	*Potri.011G043100*
	ABCC	*Potri.004G034700*
	ABCC	*Potri.004G034800*
	ABCD	*Potri.010G176400*
	ABCD	*Potri.005G076200*
	ABCD	*Potri.007G092200*
	ABCE	*Potri.004G235900*
	ABCE	*Potri.003G045700*
	ABCF	*Potri.010G254400*
	ABCF	*Potri.001G087300*
	ABCF	*Potri.005G084800*
	ABCF	*Potri.007G080600*
	ABCG	*Potri.004G236500*
	ABCG	*Potri.003G046800*
	ABCG	*Potri.010G131900*
	ABCG	*Potri.001G404000*
	ABCG	*Potri.011G123200*
	ABCG	*Potri.001G370600*
	ABCG	*Potri.011G094700*
	ABCG	*Potri.006G199500*
	ABCG	*Potri.016G065900*
	ABCG	*Potri.013G111900*
	ABCG	*Potri.019G083000*
	ABCG	*Potri.010G225200*
	ABCG	*Potri.006G126100*
	ABCG	*Potri.001G058900*
	ABCG	*Potri.003G169000*
	ABCG	*Potri.001G312200*
	ABCG	*Potri.001G312300*
	ABCG	*Potri.001G311500*
	ABCG	*Potri.001G311300*
	ABCG	*Potri.017G051300*
	ABCG	*Potri.017G051000*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCG	*Potri.005G073100*
	ABCG	*Potri.004G214000*
	ABCG	*Potri.009G008200*
	ABCG	*Potri.017G155300*
	ABCG	*Potri.018G152600*
	ABCG	*Potri.006G049000*
	ABCG	*Potri.016G056600*
	ABCG	*Potri.016G056700*
	ABCG	*Potri.001G255800*
	ABCG	*Potri.009G051200*
	ABCG	*Potri.001G255900*
	ABCG	*Potri.009G051300*
	ABCG	*Potri.001G393300*
	ABCG	*Potri.011G112100*
	ABCG	*Potri.013G117700*
	ABCG	*Potri.005G064300*
	ABCG	*Potri.015G027600*
	ABCG	*Potri.009G070100*
	ABCG	*Potri.012G045100*
	ABCG	*Potri.015G036100*
	ABCG	*Potri.007G004800*
	ABCG	*Potri.002G156900*
	ABCG	*Potri.014G080200*
	ABCG	*Potri.006G215100*
	ABCG	*Potri.008G047900*
	ABCG	*Potri.010G213400*
	ABCG	*Potri.010G115000*
	ABCG	*Potri.001G251100*
	ABCG	*Potri.001G189500*
	ABCG	*Potri.003G049600*
	ABCG	*Potri.008G099000*
	ABCG	*Potri.010G153800*
	ABCG	*Potri.015G006000*
	ABCG	*Potri.002G188900*
	ABCG	*Potri.002G189100*
	ABCG	*Potri.010G153600*
	ABCG	*Potri.010G153700*
	ABCG	*Potri.003G178900*
	ABCG	*Potri.003G179000*
	ABCG	*Potri.001G048700*
	ABCG	*Potri.001G048800*
	ABCG	*Potri.001G049000*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCG	*Potri.001G048900*
	ABCG	*Potri.006G248500*
	ABCG	*Potri.018G032900*
	ABCG	*Potri.001G175700*
	ABCG	*Potri.003G183200*
	ABCG	*Potri.018G074500*
	ABCG	*Potri.009G147100*
	ABCG	*Potri.006G115000*
	ABCG	*Potri.008G024400*
	ABCG	*Potri.003G197100*
	ABCG	*Potri.003G197200*
	ABCI	*Potri.004G214300*
	ABCI	*Potri.009G007800*
	ABCI	*Potri.008G003300*
	ABCI	*Potri.005G062800*
	ABCI	*Potri.002G212700*
	ABCI	*Potri.T149700.1.p*
	ABCI	*Potri.014G137000*
	ABCI	*Potri.002G036300*
	ABCI	*Potri.005G226600*
	ABCI	*Potri.006G086000*
	ABCI	*Potri.003G155800*
	ABCI	*Potri.018G110200*
	ABCI	*Potri.T044000.1.p*
	ABCI	*Potri.008G012500*
	ABCI	*Potri.018G132000*
	ABCI	*Potri.001G144500*
	ABCI	*Potri.001G417700*
	ABCI	*Potri.003G089700*
	ABCI	*Potri.004G016900*
	ABCI	*Potri.004G162000*
	ABCI	*Potri.006G071900*
	ABCI	*Potri.006G074200*
	ABCI	*Potri.006G125500*
	ABCI	*Potri.007G095900*
	ABCI	*Potri.007G131700*
	ABCI	*Potri.010G235800*
	ABCI	*Potri.011G133500*
	ABCI	*Potri.012G088200*
	ABCI	*Potri.014G113400*
	ABCI	*Potri.015G080600*
	ABCI	*Potri.015G136700*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCI	*Potri.016G091600*
	ABCI	*Potri.017G020100*
	ABCI	*Potri.017G026100*
	ABCI	*Potri.017G039100*
	ABCI	*Potri.017G108200*
	ABCI	*Potri.019G089600*
	ABCI	*Potri.017G095400*
	ABCI	*Potri.001G095700*
	ABCA	*GSVIVT01033258001*
	ABCA	*GSVIVT01008042001*
	ABCA	*GSVIVT01025902001*
	ABCA	*GSVIVT01008040001*
	ABCA	*GSVIVT01029736001*
	ABCB	*GSVIVT01005757001*
	ABCB	*GSVIVT01032404001*
	ABCB	*GSVIVT01033860001*
	ABCB	*GSVIVT01010634001*
	ABCB	*GSVIVT01024527001*
Vitis vinifera	ABCB	*GSVIVT01008121001*
	ABCB	*GSVIVT01032898001*
	ABCB	*GSVIVT01038687001*
	ABCB	*GSVIVT01016617001*
	ABCB	*GSVIVT01025040001*
	ABCB	*GSVIVT01036799001*
	ABCB	*GSVIVT01036801001*
	ABCB	*GSVIVT01032577001*
	ABCB	*GSVIVT01007586001*
	ABCB	*GSVIVT01000580001*
	ABCB	*GSVIVT01014625001*
	ABCB	*GSVIVT01017696001*
	ABCB	*GSVIVT01028256001*
	ABCB	*GSVIVT01021365001*
	ABCB	*GSVIVT01021366001*
	ABCB	*GSVIVT01036093001*
	ABCB	*GSVIVT01037027001*
	ABCB	*GSVIVT01032575001*
	ABCB	*GSVIVT01032578001*
	ABCB	*GSVIVT01015306001*
	ABCB	*GSVIVT01016706001*
	ABCB	*GSVIVT01011381001*
	ABCB	*GSVIVT01033645001*
	ABCB	*GSVIVT01009946001*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCB	*GSVIVT01013125001*
	ABCC	*GSVIVT01014629001*
	ABCC	*GSVIVT01014633001*
	ABCC	*GSVIVT01014632001*
	ABCC	*GSVIVT01014628001*
	ABCC	*GSVIVT01037789001*
	ABCC	*GSVIVT01014631001*
	ABCC	*GSVIVT01028458001*
	ABCC	*GSVIVT01028440001*
	ABCC	*GSVIVT01028459001*
	ABCC	*GSVIVT01010326001*
	ABCC	*GSVIVT01018403001*
	ABCC	*GSVIVT01028722001*
	ABCC	*GSVIVT01028744001*
	ABCC	*GSVIVT01019467001*
	ABCC	*GSVIVT01019471001*
	ABCC	*GSVIVT01019473001*
	ABCC	*GSVIVT01019469001*
	ABCC	*GSVIVT01019475001*
	ABCC	*GSVIVT01019476001*
	ABCC	*GSVIVT01016879001*
	ABCC	*GSVIVT01016880001*
	ABCC	*GSVIVT01015960001*
	ABCC	*GSVIVT01021589001*
	ABCC	*GSVIVT01021594001*
	ABCC	*GSVIVT01021595001*
	ABCC	*GSVIVT01027990001*
	ABCD	*GSVIVT01036685001*
	ABCE	*GSVIVT01036876001*
	ABCF	*GSVIVT01003144001*
	ABCF	*GSVIVT01019609001*
	ABCF	*GSVIVT01001694001*
	ABCF	*GSVIVT01034906001*
	ABCF	*GSVIVT01031505001*
	ABCF	*GSVIVT01022235001*
	ABCG	*GSVIVT01002139001*
	ABCG	*GSVIVT01002319001*
	ABCG	*GSVIVT01002949001*
	ABCG	*GSVIVT01015771001*
	ABCG	*GSVIVT01037566001*
	ABCG	*GSVIVT01002136001*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCG	*GSVIVT01003413001*
	ABCG	*GSVIVT01015767001*
	ABCG	*GSVIVT01015768001*
	ABCG	*GSVIVT01031516001*
	ABCG	*GSVIVT01024228001*
	ABCG	*GSVIVT01031528001*
	ABCG	*GSVIVT01031529001*
	ABCG	*GSVIVT01022526001*
	ABCG	*GSVIVT01001660001*
	ABCG	*GSVIVT01014733001*
	ABCG	*GSVIVT01034463001*
	ABCG	*GSVIVT01014402001*
	ABCG	*GSVIVT01011981001*
	ABCG	*GSVIVT01036869001*
	ABCG	*GSVIVT01011781001*
	ABCG	*GSVIVT01028218001*
	ABCG	*GSVIVT01016993001*
	ABCG	*GSVIVT01035785001*
	ABCG	*GSVIVT01014222001*
	ABCG	*GSVIVT01020687001*
	ABCG	*GSVIVT01022346001*
	ABCG	*GSVIVT01037274001*
	ABCG	*GSVIVT01028809001*
	ABCG	*GSVIVT01003847001*
	ABCG	*GSVIVT01006767001*
	ABCG	*GSVIVT01008456001*
	ABCG	*GSVIVT01032625001*
	ABCG	*GSVIVT01025582001*
	ABCG	*GSVIVT01025230001*
	ABCG	*GSVIVT01025712001*
	ABCG	*GSVIVT01016240001*
	ABCG	*GSVIVT01024743001*
	ABCG	*GSVIVT01016994001*
	ABCG	*GSVIVT01016998001*
	ABCG	*GSVIVT01016991001*
	ABCG	*GSVIVT01016999001*
	ABCG	*GSVIVT01017676001*
	ABCG	*GSVIVT01031314001*
	ABCG	*GSVIVT01031377001*
	ABCG	*GSVIVT01031380001*
	ABCG	*GSVIVT01031378001*
	ABCG	*GSVIVT01036184001*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCG	*GSVIVT01034745001*
	ABCG	*GSVIVT01034748001*
	ABCG	*GSVIVT01034741001*
	ABCG	*GSVIVT01034746001*
	ABCG	*GSVIVT01033804001*
	ABCG	*GSVIVT01015456001*
	ABCG	*GSVIVT01035715001*
	ABCG	*GSVIVT01017188001*
	ABCG	*GSVIVT01017198001*
	ABCG	*GSVIVT01017197001*
	ABCG	*GSVIVT01017201001*
	ABCG	*GSVIVT01017202001*
	ABCG	*GSVIVT01017187001*
	ABCG	*GSVIVT01017196001*
	ABCG	*GSVIVT01017184001*
	ABCG	*GSVIVT01017204001*
	ABCG	*GSVIVT01017185001*
	ABCG	*GSVIVT01017336001*
	ABCG	*GSVIVT01015461001*
	ABCG	*GSVIVT01017194001*
	ABCG	*GSVIVT01035780001*
	ABCG	*GSVIVT01035784001*
	ABCG	*GSVIVT01035786001*
	ABCI	*GSVIVT01033294001*
	ABCI	*GSVIVT01028360001*
	ABCI	*GSVIVT01031286001*
	ABCI	*GSVIVT01033060001*
	ABCI	*GSVIVT01026484001*
	ABCI	*GSVIVT01019131001*
	ABCI	*GSVIVT01026677001*
	ABCI	*GSVIVT01037976001*
	ABCI	*GSVIVT01003781001*
	ABCI	*GSVIVT01021593001*
	ABCI	*GSVIVT01029255001*
	ABCI	*GSVIVT01010362001*
	ABCI	*GSVIVT01016794001*
	ABCI	*GSVIVT01036588001*
	ABCI	*GSVIVT01002663001*
	ABCI	*GSVIVT01007689001*
	ABCI	*GSVIVT01010343001*
	ABCI	*GSVIVT01010853001*
	ABCI	*GSVIVT01012742001*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCI	*GSVIVT01013180001*
	ABCI	*GSVIVT01015822001*
	ABCI	*GSVIVT01016992001*
	ABCI	*GSVIVT01017189001*
	ABCI	*GSVIVT01017190001*
	ABCI	*GSVIVT01017191001*
	ABCI	*GSVIVT01017193001*
	ABCI	*GSVIVT01017195001*
	ABCI	*GSVIVT01017199001*
	ABCI	*GSVIVT01017200001*
	ABCI	*GSVIVT01017337001*
	ABCI	*GSVIVT01018890001*
	ABCI	*GSVIVT01021591001*
	ABCI	*GSVIVT01025723001*
	ABCI	*GSVIVT01028207001*
	ABCI	*GSVIVT01028216001*
	ABCI	*GSVIVT01028457001*
	ABCI	*GSVIVT01029545001*
	ABCI	*GSVIVT01031381001*
	ABCI	*GSVIVT01031517001*
	ABCI	*GSVIVT01032557001*
	ABCI	*GSVIVT01036183001*
	ABCA	*Bradi3g35400*
	ABCA	*Bradi3g08020*
	ABCA	*Bradi1g37230*
	ABCA	*Bradi3g35377*
	ABCA	*Bradi3g35390*
	ABCA	*Bradi4g28617*
	ABCB	*Bradi1g63433*
	ABCB	*Bradi4g03310*
	ABCB	*Bradi2g58380*
	ABCB	*Bradi5g09380*
Brachypodium distachyon	ABCB	*Bradi1g08280*
	ABCB	*Bradi3g17010*
	ABCB	*Bradi3g17020*
	ABCB	*Bradi1g28416*
	ABCB	*Bradi1g72517*
	ABCB	*Bradi5g13617*
	ABCB	*Bradi2g48610*
	ABCB	*Bradi5g12307*
	ABCB	*Bradi3g12627*
	ABCB	*Bradi2g62540*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCB	*Bradi1g66290*
	ABCB	*Bradi3g06577*
	ABCB	*Bradi3g52220*
	ABCB	*Bradi5g23600*
	ABCB	*Bradi2g20170*
	ABCB	*Bradi2g20177*
	ABCB	*Bradi2g47400*
	ABCB	*Bradi2g47410*
	ABCB	*Bradi2g11210*
	ABCB	*Bradi2g36897*
	ABCB	*Bradi3g10680*
	ABCB	*Bradi2g17710*
	ABCB	*Bradi2g17720*
	ABCB	*Bradi3g20045*
	ABCB	*Bradi2g47330*
	ABCB	*Bradi2g47360*
	ABCB	*Bradi2g47337*
	ABCB	*Bradi2g47347*
	ABCC	*Bradi1g48697*
	ABCC	*Bradi1g48710*
	ABCC	*Bradi1g48670*
	ABCC	*Bradi3g00540*
	ABCC	*Bradi1g75590*
	ABCC	*Bradi5g19787*
	ABCC	*Bradi2g04577*
	ABCC	*Bradi4g39174*
	ABCC	*Bradi2g00417*
	ABCC	*Bradi2g57900*
	ABCC	*Bradi2g33210*
	ABCC	*Bradi1g37840*
	ABCC	*Bradi1g37826*
	ABCC	*Bradi1g37850*
	ABCC	*Bradi2g12727*
	ABCC	*Bradi5g03460*
	ABCC	*Bradi5g03477*
	ABCC	*Bradi1g47277*
	ABCC	*Bradi5g22077*
	ABCD	*Bradi2g39420*
	ABCD	*Bradi2g61777*
	ABCD	*Bradi2g07076*
	ABCD	*Bradi2g54830*
	ABCE	*Bradi3g10390*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCE	*Bradi3g33470*
	ABCE	*Bradi4g16800*
	ABCF	*Bradi1g15780*
	ABCF	*Bradi2g42170*
	ABCF	*Bradi4g38960*
	ABCF	*Bradi4g14020*
	ABCF	*Bradi3g56780*
	ABCF	*Bradi3g12610*
	ABCG	*Bradi1g66110*
	ABCG	*Bradi1g66120*
	ABCG	*Bradi2g54300*
	ABCG	*Bradi1g62810*
	ABCG	*Bradi2g37630*
	ABCG	*Bradi2g26560*
	ABCG	*Bradi4g29810*
	ABCG	*Bradi2g26800*
	ABCG	*Bradi5g27650*
	ABCG	*Bradi1g40464*
	ABCG	*Bradi2g43570*
	ABCG	*Bradi4g24120*
	ABCG	*Bradi3g16130*
	ABCG	*Bradi1g36410*
	ABCG	*Bradi1g74640*
	ABCG	*Bradi3g29847*
	ABCG	*Bradi5g16400*
	ABCG	*Bradi2g31910*
	ABCG	*Bradi1g00760*
	ABCG	*Bradi4g32850*
	ABCG	*Bradi4g08130*
	ABCG	*Bradi3g57540*
	ABCG	*Bradi5g02010*
	ABCG	*Bradi5g02030*
	ABCG	*Bradi2g01610*
	ABCG	*Bradi1g31130*
	ABCG	*Bradi4g45397*
	ABCG	*Bradi1g38025*
	ABCG	*Bradi2g48620*
	ABCG	*Bradi2g43150*
	ABCG	*Bradi2g10110*
	ABCG	*Bradi3g07850*
	ABCG	*Bradi1g26040*
	ABCG	*Bradi1g00460*

Chapter 7: Appendices

Species	Subfamilies	Gene ID
	ABCG	*Bradi2g04847*
	ABCG	*Bradi3g41987*
	ABCG	*Bradi3g34890*
	ABCG	*Bradi4g28270*
	ABCG	*Bradi1g35680*
	ABCG	*Bradi2g43120*
	ABCG	*Bradi5g02870*
	ABCG	*Bradi3g27647*
	ABCG	*Bradi3g55580*
	ABCG	*Bradi1g29187*
	ABCI	*Bradi5g22580*
	ABCI	*Bradi4g19047*
	ABCI	*Bradi4g38210*
	ABCI	*Bradi1g63380*
	ABCI	*Bradi1g33777*
	ABCI	*Bradi2g51200*
	ABCI	*Bradi1g75570*
	ABCI	*Bradi1g64110*
	ABCI	*Bradi3g54870*
	ABCI	*Bradi1g51940*
	ABCI	*Bradi1g62900*
	ABCI	*Bradi1g74916*
	ABCI	*Bradi1g74922*
	ABCI	*Bradi2g01870*
	ABCI	*Bradi2g53960*
	ABCI	*Bradi3g21310*
	ABCI	*Bradi3g27912*
	ABCI	*Bradi3g43010*
	ABCI	*Bradi4g38187*

Chapter 7: Appendices

Appendix 3.5 Gene duplications of the syntenic gene pairs in flax.

Gene 1	Gene 2	*Ka*	*Ks*	*Ka/Ks*	Selection Type	Duplication time (MYA)
LuABCA1	*LuABCA2*	0.01	0.09	0.12	Purifying	2.97
LuABCB43	*LuABCB45*	0.10	0.85	0.11	Purifying	28.27
LuABCB38	*LuABCB42*	0.20	0.53	0.37	Purifying	17.70
LuABCB14	*LuABCB13*	0.00	0.11	0.03	Purifying	3.60
LuABCB33	*LuABCB32*	0.02	0.20	0.08	Purifying	6.63
LuABCB16	*LuABCB15*	0.01	0.12	0.10	Purifying	3.83
LuABCB11	*LuABCB12*	0.01	0.18	0.08	Purifying	5.87
LuABCB28	*LuABCB29*	0.03	0.09	0.38	Purifying	2.93
LuABCB35	*LuABCB34*	0.03	0.05	0.72	Purifying	1.53
LuABCB19	*LuABCB20*	0.01	0.09	0.06	Purifying	3.10
LuABCB27	*LuABCB26*	0.00	0.08	0.04	Purifying	2.50
LuABCB22	*LuABCB23*	0.02	0.10	0.16	Purifying	3.23
LuABCB25	*LuABCB24*	0.06	0.14	0.44	Purifying	4.70
LuABCC15	*LuABCC16*	0.00	0.05	0.06	Purifying	1.67
LuABCC13	*LuABCC12*	0.02	0.18	0.10	Purifying	5.93
LuABCC17	*LuABCC18*	0.03	0.13	0.25	Purifying	4.23
LuABCC1	*LuABCC2*	0.01	0.14	0.10	Purifying	4.50
LuABCC6	*LuABCC5*	0.01	0.16	0.09	Purifying	5.37
LuABCC10	*LuABCC9*	0.00	0.08	0.05	Purifying	2.80
LuABCE2	*LuABCE1*	0.00	0.11	0.01	Purifying	3.50
LuABCF8	*LuABCF7*	0.01	0.17	0.03	Purifying	5.80
LuABCF2	*LuABCF1*	0.01	0.06	0.13	Purifying	1.87
LuABCF5	*LuABCF6*	0.00	0.20	0.02	Purifying	6.50
LuABCG71	*LuABCG72*	0.21	0.10	2.18	Positive	3.23
LuABCG61	*LuABCG64*	0.45	0.36	1.25	Positive	11.90
LuABCG67	*LuABCG66*	0.03	0.38	0.09	Purifying	12.73
LuABCG42	*LuABCG68*	0.02	0.16	0.14	Purifying	5.20
LuABCG75	*LuABCG74*	0.07	0.13	0.50	Purifying	4.43
LuABCG80	*LuABCG69*	0.54	0.53	1.02	Positive	17.77
LuABCG81	*LuABCG78*	0.08	0.66	0.13	Purifying	21.93
LuABCG7	*LuABCG6*	0.00	0.08	0.02	Purifying	2.73
LuABCG14	*LuABCG12*	0.19	0.84	0.23	Purifying	28.10
LuABCG37	*LuABCG40*	0.06	0.59	0.10	Purifying	19.67
LuABCG38	*LuABCG41*	0.05	0.62	0.08	Purifying	20.67
LuABCG43	*LuABCG28*	0.02	0.27	0.06	Purifying	8.83
LuABCG29	*LuABCG30*	0.06	0.18	0.34	Purifying	6.10
LuABCG47	*LuABCG46*	0.03	0.30	0.08	Purifying	9.83
LuABCG44	*LuABCG45*	0.03	0.19	0.18	Purifying	6.17
LuABCG48	*LuABCG49*	0.03	0.14	0.21	Purifying	4.80
LuABCG27	*LuABCG26*	0.01	0.05	0.20	Purifying	1.53

Chapter 7: Appendices

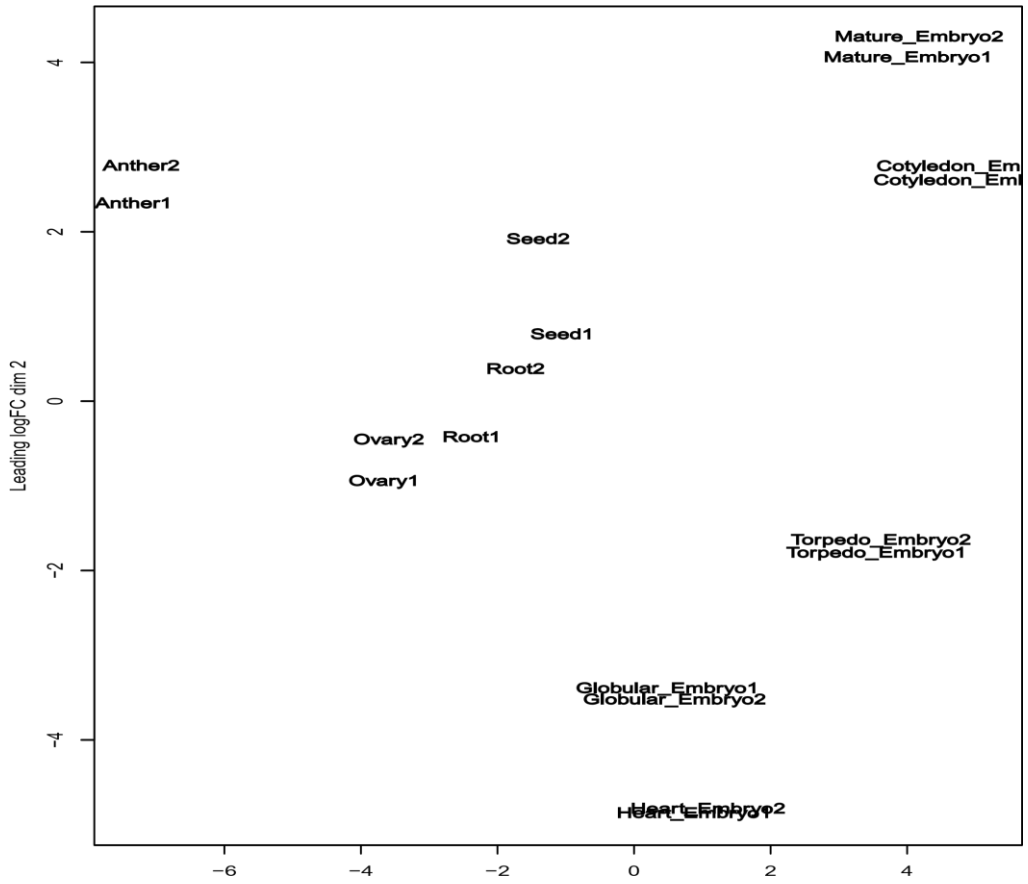
Gene 1	Gene 2	*Ka*	*Ks*	*Ka/Ks*	Selection Type	Duplication time (MYA)
LuABCG24	*LuABCG25*	0.01	0.11	0.09	Purifying	3.57
LuABCG20	*LuABCG19*	0.02	0.09	0.23	Purifying	2.90
LuABCG50	*LuABCG54*	0.06	0.49	0.12	Purifying	16.30
LuABCG4	*LuABCG3*	0.23	0.21	1.12	Positive	6.93
LuABCH21	*LuABCH20*	0.00	0.06	0.05	Purifying	2.03
LuABCG16	*LuABCG11*	0.32	0.83	0.39	Purifying	27.73
LuABCD2	*LuABCD3*	0.55	0.57	0.97	Purifying	19.03
LuHMA1	*LuHMA2*	0.01	0.16	0.07	Purifying	5.40
LuHMA6	*LuHMA8*	0.71	0.67	1.06	Positive	22.20

Appendix 3.6 Gene ontology (GO) of ATP binding cassette (ABC) transporter and heavy metal associated (HMA) genes in flax.

Gene	GO ID	GO function	Function category	Roles of gene
ABC transporter	GO:0005524	ATP binding	molecular function	Interacting selectively and non-covalently with ATP, adenosine 5'-triphosphate, a universally important coenzyme and enzyme regulator.
	GO:0006810	transport	biological process	The directed movement of substances (such as macromolecules, small molecules, ions) into, out of or within a cell, or between cells, or within a multicellular organism by means of some agent such as a transporter or pore.
	GO:0016021	integral component of membrane	cellular component	The component of a membrane consisting of the gene products and protein complexes having at least some part of their peptide sequence embedded in the hydrophobic region of the membrane.
	GO:0016887	ATPase activity	molecular function	Catalysis of the reaction: ATP + H_2O = ADP + phosphate + 2 H+. May or may not be coupled to another reaction.
	GO:0042626	ATPase activity, coupled to transmembrane movement of substances	molecular function	Catalysis of the reaction: ATP + H_2O = ADP + phosphate, to directly drive the active transport of a substance across a membrane.
	GO:0055085	transmembrane transport	biological process	The process in which a solute is transported across a lipid bilayer, from one side of a membrane to the other
	GO:0005215	transporter activity	molecular function	Enables the directed movement of substances (such as macromolecules, small molecules, ions) into, out of or within a cell, or between cells.
	GO:0003824	catalytic activity	molecular function	Catalysis of a biochemical reaction at physiological temperatures. In biologically catalyzed reactions, the reactants are known as substrates, and the catalysts are naturally occurring macromolecular substances known as enzymes. Enzymes possess specific binding sites for substrates, and are usually composed wholly or largely of protein, but RNA that has catalytic activity (ribozyme) is often also regarded as enzymatic.
	GO:0016301	kinase activity	molecular function	Catalysis of the transfer of a phosphate group, usually from ATP, to a substrate molecule.

Chapter 7: Appendices

Gene	GO ID	GO function	Function category	Roles of gene
	GO:0005525	GTP binding	molecular function	Interacting selectively and non-covalently with GTP, guanosine triphosphate.
	GO:0005622	intracellular	cellular component	The living contents of a cell; the matter contained within (but not including) the plasma membrane, usually taken to exclude large vacuoles and masses of secretory or ingested material. In eukaryotes it includes the nucleus and cytoplasm.
	GO:0007165	signal transduction	biological process	The cellular process in which a signal is conveyed to trigger a change in the activity or state of a cell. Signal transduction begins with reception of a signal (e.g. a ligand binding to a receptor or receptor activation by a stimulus such as light), or for signal transduction in the absence of ligand, signal-withdrawal or the activity of a constitutively active receptor. Signal transduction ends with regulation of a downstream cellular process, e.g. regulation of transcription or regulation of a metabolic process. Signal transduction covers signaling from receptors located on the surface of the cell and signaling via molecules located within the cell. For signaling between cells, signal transduction is restricted to events at and within the receiving cell.
	GO:0007264	small GTPase mediated signal transduction	biological process	Any series of molecular signals in which a small monomeric GTPase relays one or more of the signals.
	GO:0016020	membrane	cellular component	A lipid bilayer along with all the proteins and protein complexes embedded in it an attached to it.
HMA	GO:0000166	nucleotide binding	molecular function	Interacting selectively and non-covalently with a nucleotide, any compound consisting of a nucleoside that is esterified with (ortho)phosphate or an oligophosphate at any hydroxyl group on the ribose or deoxyribose.
	GO:0006812	cation transport	biological process	The directed movement of cations, atoms or small molecules with a net positive charge, into, out of or within a cell, or between cells, by means of some agent such as a transporter or pore.
	GO:0016021	integral component of membrane	cellular component	The component of a membrane consisting of the gene products and protein complexes having at least some part of their peptide sequence embedded in the hydrophobic region of the membrane.
	GO:0019829	cation-transporting ATPase activity	molecular function	Catalysis of the transfer of a solute or solutes from one side of a membrane to the other according to the reaction: ATP + H_2O + cation(out) = ADP + phosphate + cation(in).
	GO:0046872	metal ion binding	molecular function	Interacting selectively and non-covalently with any metal ion.
	GO:0030001	metal ion transport	biological process	The directed movement of metal ions, any metal ion with an electric charge, into, out of or within a cell, or between cells, by means of some agent such as a transporter or pore.
	GO:0005507	copper ion binding	molecular function	Interacting selectively and non-covalently with copper (Cu) ions.



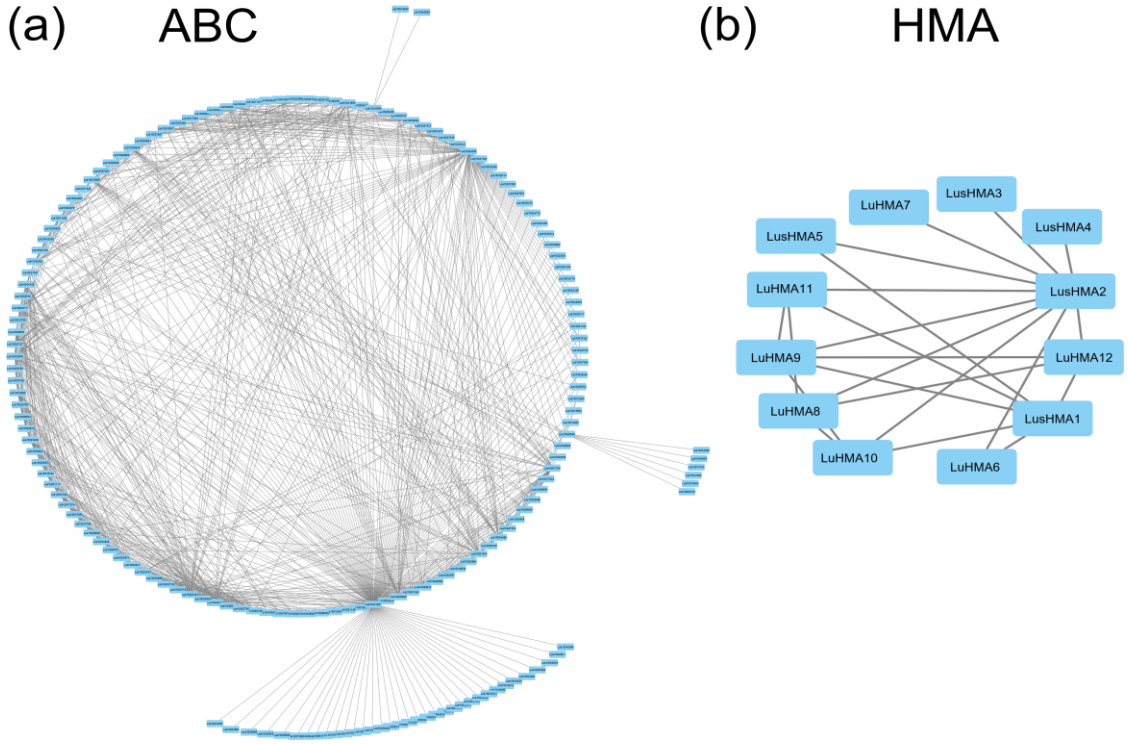
Appendix 3.7 Multidimensional scaling (MDS) plot of nine different organs displaying the relative similarities between the biological replicates based on the log fold change values.

Appendix 3.8 Pearson correlation coefficients of the flax syntenic pairs and their functional evolution.

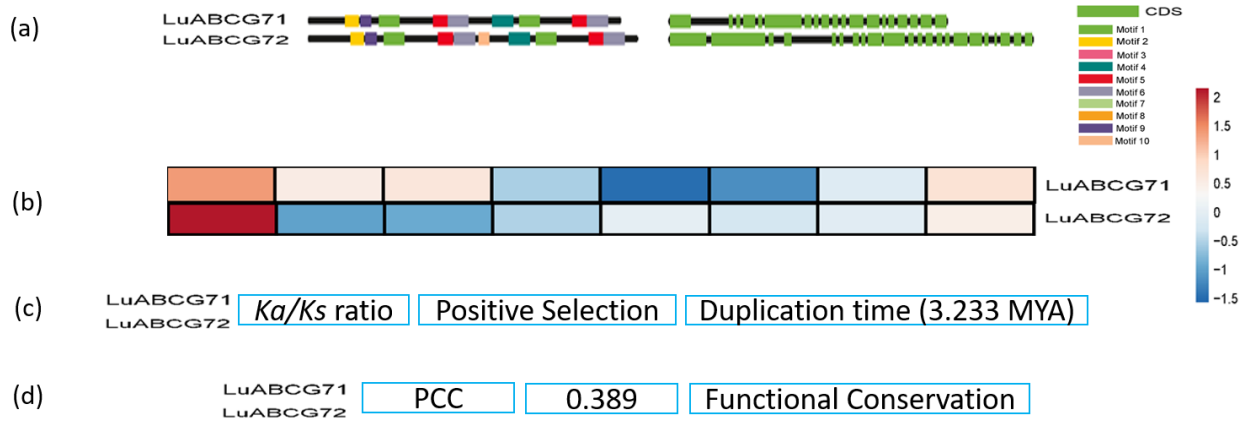
Gene 1	Gene 2	Gene duplications	Pearson correlation coefficients	Functional evolution
ABCA1	*ABCA2*	WGD or Segmental	0.814	Functional Conservation
ABCB43	*ABCB45*	WGD or Segmental	-0.522	Neofunctionalization
ABCB38	*ABCB42*	WGD or Segmental	0.704	Functional Conservation
ABCB14	*ABCB13*	WGD or Segmental	0.875	Functional Conservation
ABCB33	*ABCB32*	WGD or Segmental	0.915	Functional Conservation
ABCB16	*ABCB15*	WGD or Segmental	0.906	Functional Conservation
ABCB11	*ABCB12*	WGD or Segmental	0.835	Functional Conservation
ABCB28	*ABCB29*	WGD or Segmental	0.944	Functional Conservation
ABCB35	*ABCB34*	WGD or Segmental	0.707	Functional Conservation
ABCB19	*ABCB20*	WGD or Segmental	0.068	Neofunctionalization
ABCB27	*ABCB26*	WGD or Segmental	0.893	Functional Conservation
ABCB22	*ABCB23*	WGD or Segmental	0.871	Functional Conservation
ABCB25	*ABCB24*	WGD or Segmental	-0.072	Neofunctionalization

Chapter 7: Appendices

Gene 1	Gene 2	Gene duplications	Pearson correlation coefficients	Functional evolution
ABCC15	*ABCC16*	WGD or Segmental	0.970	Functional Conservation
ABCC13	*ABCC12*	WGD or Segmental	0.712	Functional Conservation
ABCC17	*ABCC18*	WGD or Segmental	0.501	Functional Conservation
ABCC1	*ABCC2*	WGD or Segmental	0.924	Functional Conservation
ABCC6	*ABCC5*	WGD or Segmental	0.958	Functional Conservation
ABCC10	*ABCC9*	WGD or Segmental	0.900	Functional Conservation
ABCE2	*ABCE1*	WGD or Segmental	0.423	Functional Conservation
ABCF8	*ABCF7*	WGD or Segmental	0.897	Functional Conservation
ABCF2	*ABCF1*	WGD or Segmental	0.917	Functional Conservation
ABCF5	*ABCF6*	WGD or Segmental	0.924	Functional Conservation
ABCG71	*ABCG72*	WGD or Segmental	0.389	Functional Conservation
ABCG61	*ABCG64*	WGD or Segmental	0.831	Functional Conservation
ABCG67	*ABCG66*	WGD or Segmental	0.620	Functional Conservation
ABCG42	*ABCG68*	WGD or Segmental	0.521	Functional Conservation
ABCG81	*ABCG78*	WGD or Segmental	0.655	Functional Conservation
ABCG7	*ABCG6*	WGD or Segmental	0.876	Functional Conservation
ABCG14	*ABCG12*	WGD or Segmental	0.483	Functional Conservation
ABCG4	*ABCG2*	WGD or Segmental	-0.658	Neofunctionalization
ABCG37	*ABCG40*	WGD or Segmental	0.864	Functional Conservation
ABCG38	*ABCG41*	WGD or Segmental	0.052	Neofunctionalization
ABCG43	*ABCG28*	WGD or Segmental	0.558	Functional Conservation
ABCG29	*ABCG30*	WGD or Segmental	0.431	Functional Conservation
ABCG47	*ABCG46*	WGD or Segmental	0.699	Functional Conservation
ABCG44	*ABCG45*	WGD or Segmental	0.737	Functional Conservation
ABCG48	*ABCG49*	WGD or Segmental	-0.681	Neofunctionalization
ABCG27	*ABCG26*	WGD or Segmental	0.537	Functional Conservation
ABCG21	*ABCG23*	WGD or Segmental	-0.203	Neofunctionalization
ABCG24	*ABCG25*	WGD or Segmental	0.608	Functional Conservation
ABCG20	*ABCG19*	WGD or Segmental	0.750	Functional Conservation
ABCG50	*ABCG54*	WGD or Segmental	0.703	Functional Conservation
ABCG4	*ABCG3*	WGD or Segmental	0.527	Functional Conservation
ABCG51	*ABCG56*	WGD or Segmental	-0.264	Neofunctionalization
LuABCH21	*ABCH20*	WGD or Segmental	0.981	Functional Conservation
ABCG16	*ABCG11*	Tandem	0.858	Functional Conservation
LuABCD2	*ABCD3*	Tandem	0.972	Functional Conservation
LuHMA1	*LuHMA2*	WGD or Segmental	0.409	Functional Conservation
LuHMA6	*LuHMA8*	WGD or Segmental	-0.464	Neofunctionalization
LuHMA9	*LuHMA10*	WGD or Segmental	0.300	Functional Conservation



Appendix 3.9 Interaction network of ATP binding cassette (ABC) transporter genes (a) and heavy metal associated (HMA) genes (b).



Appendix 3.10 Schematic representations of the two most stable genes among nine cadmium (Cd) candidate genes based on their conserved gene structure (a), gene expression (b), non-synonymous/synonymous substitution rates (c), and Pearson correlation coefficient (PCC) (d).

Chapter 7: Appendices

7.2 Supplementary Materials for Chapter 4

Appendix 4.1 List of 168 flax accessions used in the study.

Accessions	Alternative name
Andro	
CDCArras	
CDCBuryu	
CDCDorado	
CDCGlas	
CDCGold	
CDCMelyn	
CDCMons	
CDCNeela	
CDCNormandy	
CDCPlava	
CDCRowland	
CDCSanctuary	
CDCSorrel	
CDCValour	
CN100795	Tammes Pale Blue
CN100848	Ottawa 2152
CN100864	Bjelo Katjacs
CN100883	Beta 201
CN100952	VIR-1270
CN101052	L-93-2
CN101055	L-140-16
CN101094	Torzhokskij 4
CN101115	AR-2
CN101116	G-1781-4-18
CN101127	G-1847-4-1
CN101132	No 3860 (from Krasmodar)
CN101154	Belochka
CN101265	Amason
CN101286	Dakota Line 8
CN101289	A-93
CN101296	L. 270-68
CN101308	VNIIL-180
CN101310	VNIIL-571
CN101325	VNIIL-1104
CN101329	VNIIL-519
CN101332	VNIIL-1046

Chapter 7: Appendices

Accessions	Alternative name
CN101338	VNIIL-655
CN101364	VNIIL-776
CN101366	VNIIL-725
CN101373	VNIIL-868
CN101378	VNIIL-409
CN101385	TR 35141
CN101392	Tajga
CN101395	Line 629-01
CN101401	G 2063-5-10
CN101402	VNIIL-5631
CN101406	L-41
CN101416	China 1
CN101419	China 4
CN101421	China 6
CN101469	Sel. of CIli-1484 (C4)
CN101471	Sel CIli -1490 (C4)
CN101472	Sel. of CIli-1493 (C4)
CN101486	Sel CIli -1761 (short)
CN101493	Sel CIli -1819 (C4)
CN101535	Sel CIli -2085 (C2)
CN101536	Sel CIli -2085 (C4)
CN101554	Sel CIli -2225 (C4)
CN101560	Sel. of CIli-2289 (C2)
CN101565	Sel CIli -2410 (C6)
CN101580	Sel. Of CIli-2611 (C4)
CN101594	Sel. Of CIli-2699 (C4)
CN101595	Sel CIli -2703 (C6/C4)
CN101596	Sel CIli -2719 (C4)
CN101598	Sel. Of CIli-2734 (C4)
CN101600	Sel CIli -2748 (C4)
CN101610	Sel. of VIR-2404 (Sterility)
CN113300	
CN113302	
CN113304	
CN113310	
CN113312	
CN113314	
CN113316	
CN113324	
CN113328	
CN113330	
CN113332	

Chapter 7: Appendices

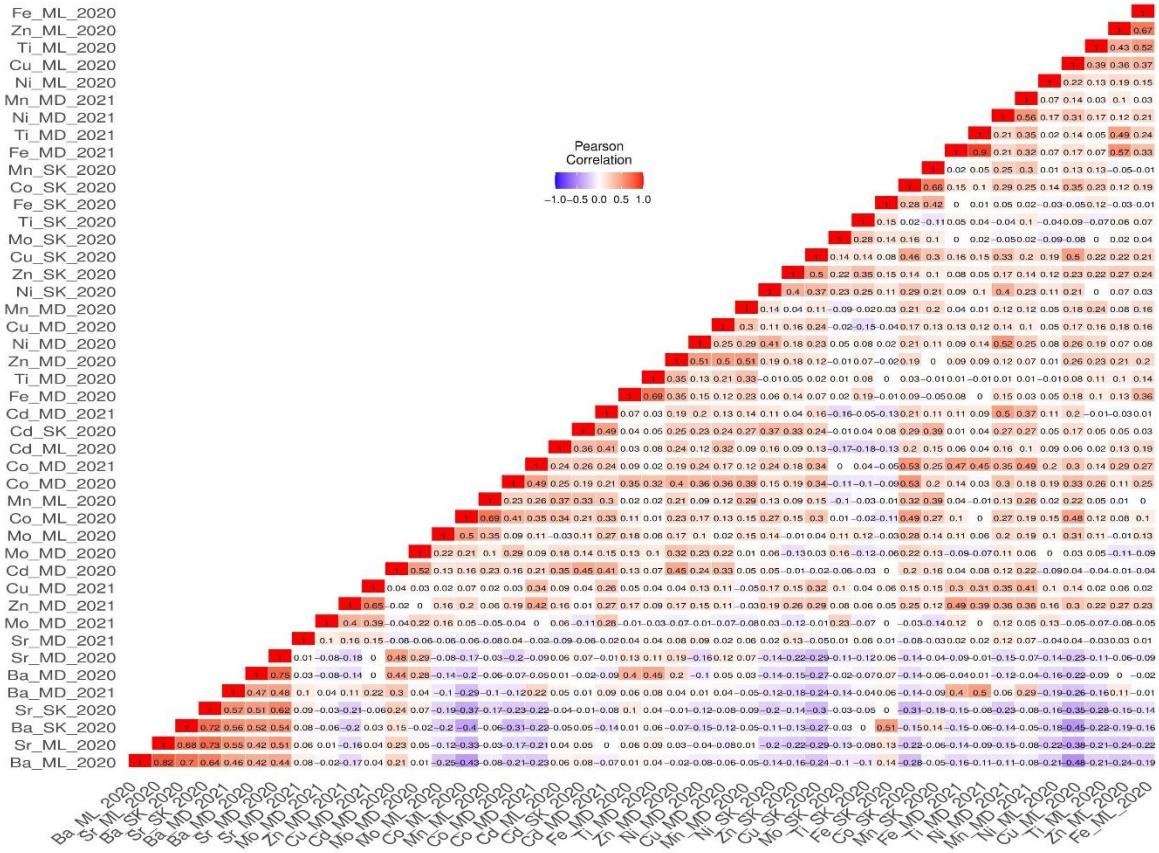
Accessions	Alternative name
CN113340	
CN113342	
CN113344	
CN113348	
CN113352	
CN113354	
CN113358	
CN113360	
CN113368	
CN113372	
CN114041	Linola 989
CN114057	Prairie Thunder
CN114067	Shape
CN114693	Prairie Grande
CN116251	CDC Bethune-14
CN116281	Flanders
CN116282	Vimy
CN116283	Somme
CN116286	AAC Bravo
CN18983	Laura
CN18986	Hermes
CN18988	Ariane
CN18989	Atalante
CN18998	Escalina
CN32542	Vniil-17
CN32546	Korostenskij 3
CN33386	Noralta
CN33399	Bison
CN33992	Culbert
CN37286	McGregor
CN52732	Norlin
CN72583	Hanley
CN72584	Macbeth
CN96992	CIli-1503
CN97050	CIli-1924
CN97056	CIli-1930
CN97083	TMP 2724-10
CN97092	CIli-1991
CN97129	CIli-2028
CN97139	CIli-2038
CN97180	Sorth Behbehan
CN97214	CIli-2295

Chapter 7: Appendices

Accessions	Alternative name
CN97306	N.P. (R.R.) 9
CN97341	TMP 8152-12
CN97351	de metcha 1-3-6 Vilm
CN97392	Novelty
CN97397	Sel. C.I. 19-47 Pale Blue
CN97404	Buda Sel.
CN97430B	TMP 2998-9
CN97520	CIli-576
CN97533	TMP 7619-2
CN97571	Cyprus
CN97616	Tammes Type 12
CN97649	Sel. 176 x (19 x 112)
CN97718	C.I. 342 x 644, F8 Sel. 6
CN97740	Redson
CN97881	Biwing x C.I. 980 (II-40-35)
CN97886	Lusatia
CN97980	10410/46
CN98037	10479/46
CN98037B	TMP 2467-8
CN98150	Z 11637
CN98165	1546-S
CN98193	L.G. 0189B
CN98263B	TMP 2534-5
CN98279	Karnobat 5
CN98363	N.P. 30
CN98397	N.P. 65
CN98398	N.P. 66
CN98475	Flachskopf
CN98566	Neelum (3/2)
CN98566C	TMP 2945-4
CN98569	R.R. 9 (Ind. Agr. Res. Inst.), PI 3
CN98610	Brawley R0001 (yel.sd. sel.)
CN98634	Toba
CN98639	W5618RO-41
CN98644	TMP 8168-3
CN98689	TMP 8216-5
CN98712	Safi 1.4-2-1
CN98742	TMP 8274-8
CN98752	Lina grosses graines Vilmorin No1
CN98794	Lino de Cabiro
CN98829	Dolgunetz
CN98961	248918

Chapter 7: Appendices

Accessions	Alternative name
Norland	Norland
Redwood65	Redwood-65
Royal	Royal
SP2047	SP 2047
Winona	Winona



Appendix 4.2 Pearson correlation coefficients (*r*) of heavy metal (HM) contents among all environments (year/location combinations). ML represents Melita, SK; Saskatoon, and MD; Morden

Appendix 4.3 Quantitative trait nucleotides (QTNs) identified for ten heavy metal (HM) contents.

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu1_10413121	1	10413121	Co	pLARmEB	3.70E-01	*	5.20
Lu1_10857504	1	10857504	Co	pLARmEB	6.70E-01	ns	0.19
Lu1_11473639	1	11473639	Ti	pLARmEB	3.50E-01	ns	2.09

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu1_11482977	1	11482977	Ti	pLARmEB	5.90E-01	ns	0.35
Lu1_11862929	1	11862929	Zn	pLARmEB	4.00E-01	ns	0.03
Lu1_11862929	1	11862929	Zn	mrMLM	1.60E-03	ns	0.03
Lu1_11906143	1	11906143	Sr	GLM	4.00E-01	*	5.16
Lu1_11906143	1	11906143	Sr	MLM	4.00E-01	*	5.16
Lu1_12260693	1	12260693	Ti	pLARmEB	6.20E-01	ns	0.42
Lu1_12668673	1	12668673	Fe	pLARmEB	1.30E-01	ns	0.53
Lu1_12668673	1	12668673	Sr	pLARmEB	2.70E-01	*	5.31
Lu1_13413202	1	13413202	Sr	FASTmrEMMA	8.70E-03	**	3.43
Lu1_13530165	1	13530165	Ti	pLARmEB	2.70E-01	ns	1.12
Lu1_13700450	1	13700450	Sr	GLM	3.10E-03	**	3.96
Lu1_13700450	1	13700450	Sr	MLM	3.10E-03	**	3.96
Lu1_15267435	1	15267435	Co	pLARmEB	8.10E-01	ns	0.36
Lu1_15906860	1	15906860	Ni	FASTmrMLM	3.00E-01	*	3.37
Lu1_15906860	1	15906860	Ni	FASTmrEMMA	3.00E-01	*	3.37
Lu1_15906860	1	15906860	Ni	ISIS EM-BLASSO	3.00E-01	*	3.37
Lu1_15906860	1	15906860	Ni	pLARmEB	3.00E-01	*	3.37
Lu1_15906860	1	15906860	Ni	mrMLM	3.00E-01	*	3.37
Lu1_15912187	1	15912187	Ti	pLARmEB	5.90E-01	ns	0.45
Lu1_1673460	1	1673460	Cd	FASTmrMLM	9.70E-02	*	5.24
Lu1_16858148	1	16858148	Co	pLARmEB	9.00E-01	ns	2.71
Lu1_17036384	1	17036384	Ni	pLARmEB	0.79	ns	0.35
Lu1_17083119	1	17083119	Co	pLARmEB	4.30E-01	ns	0.37
Lu1_17878494	1	17878494	Ni	FASTmrMLM	4.70E-01	ns	1.78
Lu1_17878494	1	17878494	Ni	ISIS EM-BLASSO	4.70E-01	ns	1.78
Lu1_17907764	1	17907764	Ti	pLARmEB	8.10E-01	ns	0.39
Lu1_194465	1	194465	Sr	pLARmEB	8.90E-03	**	3.92
Lu1_19847074	1	19847074	Ti	pLARmEB	7.30E-01	ns	0.64
Lu1_20365416	1	20365416	Co	pLARmEB	8.60E-01	ns	0.21
Lu1_20416547	1	20416547	Ti	pLARmEB	6.20E-01	ns	0.18
Lu1_20416547	1	20416547	Zn	pLARmEB	2.40E-01	ns	1.39
Lu1_20723678	1	20723678	Ti	pLARmEB	7.00E-01	ns	0.53
Lu1_21214983	1	21214983	Ba	FASTmrMLM	1.70E-01	ns	0.47
Lu1_25318220	1	25318220	Fe	FarmCPU	8.00E-01	ns	0.77
Lu1_25318220	1	25318220	Fe	GLM	8.00E-01	ns	0.77
Lu1_25318220	1	25318220	Fe	MLM	8.00E-01	ns	0.77

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu1_25323035	1	25323035	Co	pLARmEB	9.40E-01	**	1.46
Lu1_25324909	1	25324909	Co	pLARmEB	2.30E-01	ns	0.76
Lu1_25325248	1	25325248	Sr	MLM	4.00E-02	*	0.23
Lu1_25326421	1	25326421	Sr	pLARmEB	3.00E-02	*	0.25
Lu1_25326874	1	25326874	Co	pLARmEB	7.40E-01	ns	4.54
Lu1_25331081	1	25331081	Ni	pLARmEB	0.86	ns	0.24
Lu1_25331081	1	25331081	Sr	pLARmEB	3.20E-01	ns	0.71
Lu1_26221446	1	26221446	Co	pLARmEB	7.80E-01	**	2.52
Lu1_26222995	1	26222995	Co	mrMLM	5.90E-01	ns	2.52
Lu1_26235391	1	26235391	Ti	pLARmEB	5.90E-01	ns	0.34
Lu1_26236507	1	26236507	Fe	FASTmrMLM	5.90E-01	ns	0.11
Lu1_26236696	1	26236696	Sr	MLM	1.80E-02	*	0.48
Lu1_26236696	1	26236696	Sr	pLARmEB	1.80E-02	*	0.48
Lu1_26239219	1	26239219	Ti	pLARmEB	1.10E-01	ns	0.25
Lu1_26244820	1	26244820	Fe	pLARmEB	9.70E-01	*	10.65
Lu1_26263359	1	26263359	Fe	pLARmEB	4.10E-01	ns	3.06
Lu1_26263736	1	26263736	Co	pLARmEB	6.60E-01	ns	0.60
Lu1_26265593	1	26265593	Co	pLARmEB	5.20E-01	ns	0.22
Lu1_26265593	1	26265593	Ti	pLARmEB	4.60E-01	ns	0.38
Lu1_26268239	1	26268239	Zn	pLARmEB	2.00E-01	ns	0.41
Lu1_26268239	1	26268239	Co	pLARmEB	1.70E-01	ns	2.06
Lu1_26270719	1	26270719	Sr	pLARmEB	1.60E-02	*	3.22
Lu1_27186534	1	27186534	Sr	pLARmEB	0.013	*	0.56
Lu1_27752577	1	27752577	Co	pLARmEB	8.90E-01	ns	1.17
Lu1_27758116	1	27758116	Ni	FASTmrMLM	4.80E-01	ns	0.34
Lu1_27758116	1	27758116	Ni	ISIS EM-BLASSO	4.80E-01	ns	0.34
Lu1_27758116	1	27758116	Ni	mrMLM	4.80E-01	ns	0.34
Lu1_28480337	1	28480337	Co	pLARmEB	8.30E-01	ns	0.17
Lu1_28480516	1	28480516	Co	pLARmEB	5.00E-01	ns	0.50
Lu1_29338975	1	29338975	Co	pLARmEB	5.30E-01	ns	0.30
Lu1_7310590	1	7310590	Ti	pLARmEB	6.20E-01	ns	0.21
Lu1_8791977	1	8791977	Ti	pLARmEB	2.80E-01	ns	0.47
Lu1_9376954	1	9376954	Sr	FASTmrEMMA	2.40E-01	ns	0.05
Lu1_9376954	1	9376954	Sr	mrMLM	2.40E-01	ns	0.05
Lu2_10037668	2	10037668	Mn	mrMLM	1.40E-01	ns	2.08
Lu2_11254639	2	11254639	Ti	pLARmEB	1.80E-01	ns	1.10

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu2_11355622	2	11355622	Cu	mrMLM	2.00E-01	ns	1.34
Lu2_11355622	2	11355622	Cu	FASTmrMLM	2.00E-01	ns	1.34
Lu2_11567234	2	11567234	Fe	GLM	4.70E-01	**	3.32
Lu2_11567234	2	11567234	Fe	ISIS EM-BLASSO	4.70E-01	**	3.32
Lu2_11567234	2	11567234	Fe	mrMLM	4.70E-01	**	3.32
Lu2_11805774	2	11805774	Co	FASTmrMLM	5.00E-01	*	5.10
Lu2_11917060	2	11917060	Ti	pLARmEB	0.17	ns	1.35
Lu2_11923349	2	11923349	Ti	pLARmEB	4.50E-01	ns	0.85
Lu2_11976307	2	11976307	Fe	mrMLM	1.10E-01	ns	1.60
Lu2_11979691	2	11979691	Sr	mrMLM	6.40E-03	**	0.53
Lu2_11979691	2	11979691	Sr	GLM	6.40E-03	**	0.53
Lu2_11983094	2	11983094	Sr	GLM	7.90E-03	**	2.78
Lu2_12547562	2	12547562	Co	pLARmEB	6.10E-01	ns	0.29
Lu2_12565807	2	12565807	Ti	pLARmEB	9.40E-01	ns	2.76
Lu2_12567563	2	12567563	Ti	pLARmEB	5.40E-01	ns	0.12
Lu2_12609212	2	12609212	Co	pLARmEB	7.40E-01	ns	0.18
Lu2_14511171	2	14511171	Ti	pLARmEB	5.20E-01	ns	0.48
Lu2_17240363	2	17240363	Co	pLARmEB	6.70E-01	**	1.20
Lu2_17523187	2	17523187	Sr	pLARmEB	2.80E-01	ns	0.74
Lu2_17531152	2	17531152	Ti	pLARmEB	3.50E-01	ns	0.52
Lu2_17744333	2	17744333	Co	pLARmEB	7.30E-01	ns	0.64
Lu2_19721217	2	19721217	Ni	FASTmrMLM	1.60E-01	ns	1.13
Lu2_19721217	2	19721217	Ni	ISIS EM-BLASSO	1.60E-01	ns	1.13
Lu2_21365734	2	21365734	Ni	ISIS EM-BLASSO	0.033	*	4.64
Lu2_3830339	2	3830339	Ba	mrMLM	8.50E-01	ns	0.36
Lu2_5517288	2	5517288	Cd	ISIS EM-BLASSO	4.40E-01	ns	0.68
Lu2_6186376	2	6186376	Ti	pLARmEB	9.30E-01	ns	0.24
Lu2_653721	2	653721	Sr	GLM	6.40E-01	ns	1.88
Lu2_653721	2	653721	Sr	MLM	6.40E-01	ns	1.88
Lu2_6634979	2	6634979	Sr	FASTmrMLM	4.00E-01	ns	0.53
Lu2_6634979	2	6634979	Sr	mrMLM	4.00E-01	ns	0.53
Lu2_8804882	2	8804882	Co	pLARmEB	6.60E-01	ns	0.47
Lu2_9701680	2	9701680	Sr	GLM	9.40E-01	ns	0.21
Lu2_9701680	2	9701680	Sr	MLM	9.40E-01	ns	0.21
Lu3_10318345	3	10318345	Ti	pLARmEB	2.00E-01	ns	0.81
Lu3_12076843	3	12076843	Sr	pLARmEB	1.90E-01	ns	0.21

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu3_12260701	3	12260701	Fe	GLM	7.00E-01	ns	0.27
Lu3_12483044	3	12483044	Co	pLARmEB	2.10E-01	ns	1.14
Lu3_13386515	3	13386515	Co	pLARmEB	9.90E-01	ns	0.73
Lu3_14467769	3	14467769	Sr	GLM	7.70E-01	ns	0.91
Lu3_14467769	3	14467769	Sr	MLM	7.70E-01	ns	0.91
Lu3_14468211	3	14468211	Sr	GLM	6.20E-01	ns	1.32
Lu3_14468211	3	14468211	Sr	MLM	6.20E-01	ns	1.32
Lu3_16473464	3	16473464	Co	pLARmEB	8.60E-01	ns	2.64
Lu3_17173441	3	17173441	Ti	pLARmEB	5.90E-01	*	5.66
Lu3_17971708	3	17971708	Ti	ISIS EM-BLASSO	2.40E-01	ns	0.42
Lu3_17971708	3	17971708	Fe	FASTmrMLM	3.10E-01	ns	1.21
Lu3_19238415	3	19238415	Sr	FASTmrMLM	5.00E-02	**	2.50
Lu3_19258260	3	19258260	Sr	GLM	2.20E-01	ns	0.35
Lu3_19258260	3	19258260	Sr	MLM	2.20E-01	ns	0.35
Lu3_19423268	3	19423268	Sr	GLM	1.40E-02	*	0.45
Lu3_19423268	3	19423268	Sr	MLM	1.40E-02	*	0.45
Lu3_19445132	3	19445132	Sr	GLM	7.40E-02	ns	0.43
Lu3_19445178	3	19445178	Sr	GLM	2.90E-02	*	0.28
Lu3_22029670	3	22029670	Sr	GLM	2.10E-01	**	0.94
Lu3_22029670	3	22029670	Sr	MLM	2.10E-01	**	0.94
Lu3_22029673	3	22029673	Sr	GLM	2.30E-01	ns	0.60
Lu3_22029673	3	22029673	Sr	MLM	2.30E-01	ns	0.60
Lu3_22029675	3	22029675	Sr	GLM	5.90E-02	ns	1.52
Lu3_22029675	3	22029675	Sr	MLM	5.90E-02	ns	1.52
Lu3_26404939	3	26404939	Cd	ISIS EM-BLASSO	5.70E-01	ns	0.35
Lu3_26420475	3	26420475	Sr	GLM	2.10E-02	*	0.48
Lu3_26422028	3	26422028	Sr	GLM	2.80E-01	ns	0.84
Lu3_26450929	3	26450929	Ti	pLARmEB	8.90E-01	ns	0.09
Lu3_26523437	3	26523437	Co	pLARmEB	3.20E-01	ns	0.58
Lu3_26578377	3	26578377	Sr	GLM	8.20E-02	ns	1.02
Lu3_26578377	3	26578377	Sr	MLM	8.20E-02	ns	1.02
Lu3_26578392	3	26578392	Sr	GLM	5.90E-01	ns	0.62
Lu3_26578392	3	26578392	Sr	MLM	5.90E-01	ns	0.62
Lu3_26578461	3	26578461	Sr	MLM	1.60E-03	**	0.91
Lu3_4183590	3	4183590	Fe	FarmCPU	7.10E-01	ns	3.32
Lu3_4183590	3	4183590	Fe	GLM	7.10E-01	ns	3.32

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu3_493530	3	493530	Sr	GLM	6.70E-02	ns	0.41
Lu3_5479935	3	5479935	Sr	GLM	7.00E-03	**	0.68
Lu3_5479935	3	5479935	Sr	MLM	7.00E-03	**	0.68
Lu3_5508154	3	5508154	Sr	GLM	8.80E-03	**	0.67
Lu3_7487534	3	7487534	Cd	ISIS EM-BLASSO	6.20E-01	*	7.15
Lu3_7487534	3	7487534	Cd	mrMLM	6.20E-01	*	7.15
Lu3_7487534	3	7487534	Cd	FASTmrMLM	6.20E-01	*	7.15
Lu3_943534	3	943534	Sr	GLM	4.80E-01	ns	0.66
Lu3_943534	3	943534	Sr	MLM	4.80E-01	ns	0.66
Lu4_11162201	4	11162201	Co	pLARmEB	3.30E-01	*	4.66
Lu4_11270490	4	11270490	Sr	GLM	2.50E-02	*	0.42
Lu4_11270490	4	11270490	Sr	MLM	2.50E-02	*	0.42
Lu4_127929	4	127929	Zn	mrMLM	5.90E-01	ns	0.11
Lu4_15433671	4	15433671	Sr	GLM	6.50E-01	ns	0.21
Lu4_15433671	4	15433671	Sr	MLM	6.50E-01	ns	0.21
Lu4_163752	4	163752	Sr	GLM	2.80E-02	*	2.37
Lu4_163752	4	163752	Sr	MLM	2.80E-02	*	2.37
Lu4_17131337	4	17131337	Sr	GLM	6.30E-01	ns	0.54
Lu4_17131337	4	17131337	Sr	MLM	6.30E-01	ns	0.54
Lu4_18692951	4	18692951	Ti	pLARmEB	4.10E-01	ns	0.61
Lu4_19757820	4	19757820	Sr	GLM	5.20E-02	ns	0.21
Lu4_19757820	4	19757820	Sr	MLM	5.20E-02	ns	0.21
Lu4_19757947	4	19757947	Sr	GLM	8.30E-03	**	0.87
Lu4_19757947	4	19757947	Sr	MLM	8.30E-03	**	0.87
Lu4_19767345	4	19767345	Sr	GLM	5.40E-02	ns	0.13
Lu4_19773831	4	19773831	Sr	FASTmrEMMA	6.60E-01	ns	0.38
Lu4_19848927	4	19848927	Fe	GLM	6.60E-01	**	1.40
Lu4_19848927	4	19848927	Fe	MLM	6.60E-01	**	1.40
Lu4_19848927	4	19848927	Fe	ISIS EM-BLASSO	6.60E-01	**	1.40
Lu4_19880270	4	19880270	Sr	GLM	9.10E-03	**	0.56
Lu4_19905774	4	19905774	Sr	GLM	3.80E-02	*	0.47
Lu4_19905774	4	19905774	Sr	MLM	3.80E-02	*	0.47
Lu4_2192402	4	2192402	Co	pLARmEB	7.10E-01	ns	0.13
Lu4_3785453	4	3785453	Mn	FarmCPU	8.90E-01	ns	0.12
Lu4_3785453	4	3785453	Mn	GLM	8.90E-01	ns	0.12
Lu4_3785453	4	3785453	Mn	mrMLM	8.90E-01	ns	0.12

Chapter 7: Appendices

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu4_729742	4	729742	Sr	GLM	1.30E-01	ns	1.67
Lu4_729742	4	729742	Sr	MLM	1.30E-01	ns	1.67
Lu4_729751	4	729751	Sr	GLM	1.70E-02	*	0.48
Lu4_729751	4	729751	Sr	MLM	1.70E-02	*	0.48
Lu4_7409724	4	7409724	Fe	GLM	7.30E-01	ns	0.34
Lu4_7859768	4	7859768	Mn	FASTmrMLM	9.60E-01	ns	0.01
Lu4_7859768	4	7859768	Mn	ISIS EM-BLASSO	9.60E-01	ns	0.01
Lu5_1542156	5	1542156	Sr	GLM	6.00E-02	ns	0.35
Lu5_3768287	5	3768287	Mn	mrMLM	1.20E-01	ns	0.99
Lu5_6117987	5	6117987	Ti	pLARmEB	7.10E-01	ns	0.16
Lu5_6198348	5	6198348	Fe	GLM	7.40E-02	ns	0.86
Lu5_6198348	5	6198348	Fe	FASTmrMLM	7.40E-02	ns	0.86
Lu5_6198348	5	6198348	Fe	mrMLM	7.40E-02	ns	0.86
Lu5_6790709	5	6790709	Ti	pLARmEB	7.90E-01	ns	0.38
Lu5_9349523	5	9349523	Sr	mrMLM	4.00E-01	ns	0.16
Lu6_10310871	6	10310871	Fe	FASTmrMLM	9.80E-01	ns	0.04
Lu6_11424106	6	11424106	Co	pLARmEB	9.90E-01	ns	0.10
Lu6_11471289	6	11471289	Sr	FarmCPU	1.00E-01	ns	0.23
Lu6_11506867	6	11506867	Co	pLARmEB	2.30E-01	ns	1.05
Lu6_11564892	6	11564892	Co	pLARmEB	7.50E-01	ns	1.36
Lu6_12894439	6	12894439	Ti	pLARmEB	2.80E-01	ns	1.73
Lu6_14232082	6	14232082	Ti	pLARmEB	2.40E-01	ns	1.25
Lu6_149354	6	149354	Fe	GLM	3.60E-01	ns	0.96
Lu6_149501	6	149501	Sr	GLM	3.50E-01	ns	0.82
Lu6_15818175	6	15818175	Sr	GLM	8.80E-03	**	0.62
Lu6_15818175	6	15818175	Co	pLARmEB	6.80E-01	ns	1.36
Lu6_158363	6	158363	Fe	GLM	6.30E-01	ns	1.06
Lu6_16239024	6	16239024	Co	pLARmEB	6.00E-01	ns	0.18
Lu6_2121884	6	2121884	Sr	FarmCPU	8.80E-03	**	0.75
Lu6_3523454	6	3523454	Co	pLARmEB	8.50E-02	ns	2.26
Lu6_3660509	6	3660509	Zn	FASTmrMLM	0.0419	ns	0.11
Lu6_3660509	6	3660509	Zn	mrMLM	7.30E-01	ns	0.11
Lu6_3844939	6	3844939	Sr	GLM	2.10E-02	*	0.20
Lu6_445918	6	445918	Ti	ISIS EM-BLASSO	8.30E-01	*	5.62
Lu6_8487801	6	8487801	Co	pLARmEB	3.90E-01	ns	1.54
Lu6_9739632	6	9739632	Mn	ISIS EM-BLASSO	7.10E-01	ns	0.21

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu7_10374118	7	10374118	Fe	mrMLM	5.40E-01	**	9.48
Lu7_12701094	7	12701094	Ti	pLARmEB	6.00E-01	ns	0.28
Lu7_17019553	7	17019553	Co	pLARmEB	4.50E-01	ns	0.62
Lu7_17086138	7	17086138	Ba	mrMLM	9.80E-01	ns	1.30
Lu7_17086138	7	17086138	Ba	FASTmrMLM	9.80E-01	ns	1.30
Lu7_17089067	7	17089067	Co	pLARmEB	5.30E-01	ns	1.80
Lu7_17091064	7	17091064	Ti	pLARmEB	5.30E-01	ns	0.44
Lu7_17098332	7	17098332	Sr	GLM	2.30E-02	*	0.47
Lu7_17098332	7	17098332	Sr	MLM	2.30E-02	*	0.47
Lu7_17099882	7	17099882	Ti	ISIS EM-BLASSO	9.10E-01	**	5.34
Lu7_17107562	7	17107562	Sr	GLM	5.50E-01	**	1.17
Lu7_17107562	7	17107562	Sr	MLM	5.50E-01	**	1.17
Lu7_18294757	7	18294757	Sr	GLM	2.80E-01	ns	0.44
Lu7_214233	7	214233	Co	pLARmEB	5.70E-01	ns	0.24
Lu7_221619	7	221619	Ti	pLARmEB	4.70E-01	ns	0.33
Lu7_2513788	7	2513788	Cu	ISIS EM-BLASSO	6.30E-01	ns	0.15
Lu7_4744454	7	4744454	Co	pLARmEB	8.80E-01	ns	1.39
Lu7_6891825	7	6891825	Sr	pLARmEB	3.70E-01	ns	0.64
Lu7_7795225	7	7795225	Fe	GLM	9.20E-01	*	7.76
Lu7_7795225	7	7795225	Fe	MLM	9.20E-01	*	7.76
Lu7_7841206	7	7841206	Fe	GLM	3.80E-01	*	7.97
Lu7_7841206	7	7841206	Fe	MLM	3.80E-01	*	7.97
Lu7_7841299	7	7841299	Fe	GLM	6.60E-01	**	7.58
Lu7_7841299	7	7841299	Fe	MLM	6.60E-01	**	7.58
Lu7_7841299	7	7841299	Fe	mrMLM	6.60E-01	**	7.58
Lu7_7841299	7	7841299	Fe	pLARmEB	6.60E-01	**	7.58
Lu7_7841300	7	7841300	Fe	GLM	4.60E-01	*	7.58
Lu7_7841300	7	7841300	Fe	MLM	4.60E-01	*	7.58
Lu7_849864	7	849864	Co	pLARmEB	1.40E-01	ns	0.34
Lu7_9640759	7	9640759	Ni	FASTmrMLM	7.50E-01	*	1.48
Lu7_9640759	7	9640759	Ni	ISIS EM-BLASSO	7.50E-01	*	1.48
Lu7_9640759	7	9640759	Ni	mrMLM	7.50E-01	*	1.48
Lu8_10346733	8	10346733	Sr	FarmCPU	6.40E-02	ns	2.40
Lu8_12047724	8	12047724	Co	pLARmEB	1.30E-01	**	1.57
Lu8_12304779	8	12304779	Sr	GLM	4.80E-01	ns	0.39
Lu8_12304779	8	12304779	Sr	MLM	4.80E-01	ns	0.39

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu8_13063010	8	13063010	Mn	FASTmrMLM	3.70E-01	ns	0.21
Lu8_13847239	8	13847239	Sr	GLM	2.40E-02	*	0.35
Lu8_13847239	8	13847239	Sr	MLM	2.40E-02	*	0.35
Lu8_14073080	8	14073080	Ti	pLARmEB	5.90E-01	ns	0.24
Lu8_1436156	8	1436156	Ni	ISIS EM-BLASSO	0.86	ns	0.42
Lu8_21509453	8	21509453	Fe	mrMLM	8.00E-01	ns	1.26
Lu8_21580044	8	21580044	Fe	GLM	1.50E-01	ns	0.48
Lu8_21594201	8	21594201	Ti	pLARmEB	5.00E-01	ns	0.69
Lu8_21611937	8	21611937	Sr	GLM	1.20E-01	ns	0.10
Lu8_21611937	8	21611937	Sr	MLM	1.20E-01	ns	0.10
Lu8_21617485	8	21617485	Sr	GLM	6.80E-01	*	0.94
Lu8_21617485	8	21617485	Sr	FASTmrEMMA	6.80E-01	*	0.94
Lu8_21618742	8	21618742	Sr	GLM	4.50E-01	ns	0.19
Lu8_21618742	8	21618742	Sr	MLM	4.50E-01	ns	0.19
Lu8_21633476	8	21633476	Ni	FASTmrMLM	9.50E-01	ns	0.58
Lu8_21680245	8	21680245	Sr	GLM	4.40E-01	ns	0.05
Lu8_21680245	8	21680245	Sr	MLM	4.40E-01	ns	0.05
Lu8_23274330	8	23274330	Ti	pLARmEB	7.60E-01	ns	0.63
Lu8_49140	8	49140	Sr	GLM	2.10E-01	ns	1.24
Lu8_585308	8	585308	Sr	GLM	1.10E-01	ns	0.17
Lu8_585308	8	585308	Sr	MLM	1.10E-01	ns	0.17
Lu8_587393	8	587393	Sr	GLM	2.10E-01	ns	1.04
Lu8_587393	8	587393	Sr	MLM	2.10E-01	ns	1.04
Lu8_60629	8	60629	Ti	pLARmEB	0.018	*	0.04
Lu8_60629	8	60629	Sr	GLM	1.80E-02	*	3.16
Lu8_60658	8	60658	Sr	GLM	3.40E-02	*	0.79
Lu8_60658	8	60658	Sr	MLM	3.40E-02	*	0.79
Lu8_60659	8	60659	Sr	GLM	7.00E-02	ns	1.06
Lu8_60659	8	60659	Sr	MLM	7.00E-02	ns	1.06
Lu8_6318953	8	6318953	Co	pLARmEB	3.90E-01	ns	0.33
Lu8_6715730	8	6715730	Ti	pLARmEB	1.60E-01	*	1.21
Lu8_75724	8	75724	Sr	GLM	5.20E-01	ns	0.74
Lu8_75724	8	75724	Sr	MLM	5.20E-01	ns	0.74
Lu8_82211	8	82211	Sr	GLM	6.50E-01	ns	0.55
Lu8_82211	8	82211	Sr	MLM	6.50E-01	ns	0.55
Lu8_8839678	8	8839678	Sr	FASTmrEMMA	5.90E-01	ns	0.48

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu8_98342	8	98342	Sr	GLM	3.50E-01	ns	0.95
Lu8_98342	8	98342	Sr	MLM	3.50E-01	ns	0.95
Lu8_9873727	8	9873727	Sr	GLM	4.30E-02	*	2.41
Lu9_10808332	9	10808332	Sr	MLM	1.00E-02	*	0.61
Lu9_12465613	9	12465613	Zn	FASTmrMLM	3.20E-01	ns	1.54
Lu9_13338325	9	13338325	Cd	mrMLM	3.00E-01	ns	0.74
Lu9_13338325	9	13338325	Cd	ISIS EM-BLASSO	3.00E-01	ns	0.74
Lu9_13338325	9	13338325	Cd	FASTmrEMMA	3.00E-01	ns	0.74
Lu9_14102433	9	14102433	Sr	GLM	7.00E-01	ns	0.11
Lu9_14102433	9	14102433	Sr	MLM	7.00E-01	ns	0.11
Lu9_14152623	9	14152623	Ti	pLARmEB	7.60E-01	ns	0.10
Lu9_14239878	9	14239878	Sr	GLM	1.00E-01	ns	1.24
Lu9_14239878	9	14239878	Sr	MLM	1.00E-01	ns	1.24
Lu9_14239881	9	14239881	Sr	MLM	1.00E-02	*	0.72
Lu9_15862043	9	15862043	Sr	GLM	2.80E-02	*	0.43
Lu9_15862043	9	15862043	Sr	MLM	2.80E-02	*	0.43
Lu9_16778244	9	16778244	Cu	ISIS EM-BLASSO	1.40E-01	ns	1.78
Lu9_16843165	9	16843165	Cd	mrMLM	3.90E-01	ns	0.25
Lu9_16843165	9	16843165	Cd	FASTmrMLM	3.90E-01	ns	0.25
Lu9_16878254	9	16878254	Co	pLARmEB	9.20E-01	ns	0.46
Lu9_16881904	9	16881904	Co	pLARmEB	8.30E-01	ns	1.17
Lu9_16882661	9	16882661	Co	pLARmEB	6.50E-01	ns	1.09
Lu9_16907775	9	16907775	Ti	pLARmEB	1.70E-01	ns	0.18
Lu9_16917634	9	16917634	Ti	pLARmEB	3.20E-01	ns	0.62
Lu9_16966095	9	16966095	Co	pLARmEB	7.70E-01	ns	0.31
Lu9_17017285	9	17017285	Co	pLARmEB	6.90E-01	ns	1.21
Lu9_17056635	9	17056635	Ti	pLARmEB	9.10E-01	ns	3.83
Lu9_17073447	9	17073447	Co	pLARmEB	8.30E-01	ns	0.56
Lu9_17105028	9	17105028	Sr	GLM	8.80E-02	ns	0.53
Lu9_17105028	9	17105028	Sr	MLM	8.80E-02	ns	0.53
Lu9_17131228	9	17131228	Cd	FASTmrMLM	3.70E-01	ns	0.74
Lu9_17131228	9	17131228	Cd	mrMLM	3.70E-01	ns	0.74
Lu9_17131228	9	17131228	Cd	ISIS EM-BLASSO	3.70E-01	ns	0.74
Lu9_17192342	9	17192342	Sr	GLM	3.50E-02	*	0.87
Lu9_17192342	9	17192342	Sr	MLM	3.50E-02	*	0.87
Lu9_17200027	9	17200027	Co	pLARmEB	9.30E-01	ns	0.44

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu9_17201722	9	17201722	Fe	mrMLM	7.80E-01	**	10.57
Lu9_17243307	9	17243307	Ti	pLARmEB	8.30E-01	ns	0.58
Lu9_17246640	9	17246640	Co	pLARmEB	7.30E-01	ns	0.57
Lu9_17246640	9	17246640	Ti	pLARmEB	1.80E-01	ns	0.67
Lu9_17261987	9	17261987	Co	pLARmEB	5.70E-01	ns	0.39
Lu9_17280992	9	17280992	Sr	MLM	5.90E-01	ns	0.55
Lu9_1730550	9	1730550	Cu	FASTmrMLM	3.70E-01	ns	1.60
Lu9_1730550	9	1730550	Cu	mrMLM	3.70E-01	ns	1.60
Lu9_17341167	9	17341167	Ni	pLARmEB	0.03	ns	1.13
Lu9_17392318	9	17392318	Fe	GLM	7.20E-01	ns	0.06
Lu9_17415003	9	17415003	Ba	ISIS EM-BLASSO	7.30E-01	ns	1.34
Lu9_17453937	9	17453937	Co	pLARmEB	6.90E-01	ns	0.24
Lu9_17454206	9	17454206	Co	pLARmEB	6.10E-01	ns	3.12
Lu9_17542072	9	17542072	Ti	pLARmEB	2.40E-01	ns	0.65
Lu9_17591047	9	17591047	Co	pLARmEB	3.00E-01	ns	0.52
Lu9_21546287	9	21546287	Ti	pLARmEB	9.20E-01	ns	1.68
Lu9_26183	9	26183	Sr	pLARmEB	2.30E-02	*	0.40
Lu9_47515	9	47515	Sr	GLM	1.00E-01	ns	0.45
Lu9_47515	9	47515	Sr	MLM	1.00E-01	ns	0.45
Lu9_53969	9	53969	Co	pLARmEB	9.40E-01	ns	0.21
Lu9_62855	9	62855	Co	pLARmEB	6.50E-01	ns	1.38
Lu9_63995	9	63995	Fe	GLM	6.80E-01	ns	0.48
Lu9_6843813	9	6843813	Ba	mrMLM	6.60E-01	ns	0.23
Lu9_6843813	9	6843813	Ba	FASTmrMLM	6.60E-01	ns	0.23
Lu9_7775009	9	7775009	Cd	ISIS EM-BLASSO	8.00E-01	ns	0.16
Lu9_79814	9	79814	Co	pLARmEB	9.30E-01	ns	0.72
Lu9_8387539	9	8387539	Co	ISIS EM-BLASSO	6.30E-01	*	2.52
Lu9_8387539	9	8387539	Co	mrMLM	6.30E-01	*	2.52
Lu9_9585748	9	9585748	Cu	mrMLM	5.30E-01	*	6.21
Lu9_9585748	9	9585748	Cu	FASTmrMLM	5.30E-01	*	6.21
Lu10_10245781	10	10245781	Sr	MLM	4.10E-02	*	0.26
Lu10_15338938	10	15338938	Sr	GLM	6.80E-01	ns	0.26
Lu10_15338938	10	15338938	Sr	MLM	6.80E-01	ns	0.26
Lu10_15338941	10	15338941	Sr	GLM	2.90E-03	**	0.97
Lu10_15338941	10	15338941	Sr	MLM	2.90E-03	**	0.97
Lu10_4074991	10	4074991	Co	pLARmEB	7.40E-01	ns	0.20

Chapter 7: Appendices

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu10_4074991	10	4074991	Ti	pLARmEB	2.80E-01	ns	0.76
Lu10_4508210	10	4508210	Ti	pLARmEB	8.70E-01	*	1.59
Lu10_7238569	10	7238569	Cd	FASTmrMLM	6.10E-01	ns	0.39
Lu10_7238569	10	7238569	Cd	mrMLM	6.10E-01	ns	0.39
Lu10_7704389	10	7704389	Ti	pLARmEB	8.50E-01	ns	0.47
Lu10_7998746	10	7998746	Sr	FASTmrEMMA	0.013	*	2.52
Lu11_10265112	11	10265112	Fe	GLM	9.10E-01	ns	0.52
Lu11_11079129	11	11079129	Fe	FASTmrMLM	3.50E-01	ns	1.23
Lu11_11079129	11	11079129	Fe	mrMLM	3.50E-01	ns	1.23
Lu11_11079132	11	11079132	Mn	ISIS EM-BLASSO	7.30E-01	ns	0.38
Lu11_12010795	11	12010795	Ni	FASTmrMLM	4.10E-02	*	4.24
Lu11_14610978	11	14610978	Ni	FASTmrMLM	7.10E-01	ns	0.28
Lu11_14610978	11	14610978	Ni	ISIS EM-BLASSO	7.10E-01	ns	0.28
Lu11_14610978	11	14610978	Ni	mrMLM	7.10E-01	ns	0.28
Lu11_19131603	11	19131603	Sr	FASTmrMLM	1.40E-02	*	5.43
Lu11_3620428	11	3620428	Co	pLARmEB	8.60E-01	ns	0.10
Lu11_4017458	11	4017458	Sr	GLM	4.40E-02	*	0.15
Lu11_4017458	11	4017458	Sr	MLM	4.40E-02	*	0.15
Lu11_4586460	11	4586460	Fe	FarmCPU	8.30E-01	*	10.36
Lu11_5999195	11	5999195	Ba	mrMLM	1.20E-01	ns	2.40
Lu12_14237214	12	14237214	Co	ISIS EM-BLASSO	9.70E-01	ns	0.30
Lu12_14507321	12	14507321	Mn	mrMLM	8.50E-01	ns	0.29
Lu12_15916883	12	15916883	Ba	ISIS EM-BLASSO	1.70E-02	*	1.69
Lu12_16265369	12	16265369	Ni	ISIS EM-BLASSO	4.90E-01	ns	1.35
Lu12_16528816	12	16528816	Ba	FASTmrMLM	2.80E-03	**	3.22
Lu12_16568252	12	16568252	Sr	GLM	7.30E-03	**	0.65
Lu12_16568252	12	16568252	Sr	MLM	7.30E-03	**	0.65
Lu12_16603042	12	16603042	Co	ISIS EM-BLASSO	1.80E-01	ns	1.21
Lu12_16615248	12	16615248	Ba	mrMLM	2.00E-01	ns	1.54
Lu12_16615248	12	16615248	Ba	FASTmrMLM	2.00E-01	ns	1.54
Lu12_16627926	12	16627926	Sr	GLM	1.10E-02	*	0.50
Lu12_16627926	12	16627926	Sr	MLM	1.10E-02	*	0.50
Lu12_18693435	12	18693435	Sr	FarmCPU	2.00E-01	*	6.96
Lu12_1885304	12	1885304	Co	ISIS EM-BLASSO	6.60E-01	ns	2.21
Lu12_1946222	12	1946222	Sr	GLM	9.00E-04	**	1.16
Lu12_3817638	12	3817638	Sr	GLM	4.30E-02	*	0.24

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu12_3866751	12	3866751	Cd	FASTmrMLM	6.20E-01	ns	0.95
Lu12_3884097	12	3884097	Sr	FarmCPU	8.90E-02	ns	2.06
Lu12_3898188	12	3898188	Fe	FASTmrMLM	5.00E-02	ns	1.96
Lu12_5815671	12	5815671	Sr	GLM	5.60E-03	**	0.74
Lu12_5815732	12	5815732	Sr	GLM	3.70E-01	ns	0.66
Lu12_5815732	12	5815732	Sr	MLM	3.70E-01	ns	0.66
Lu12_6900096	12	6900096	Cd	mrMLM	5.60E-01	ns	0.53
Lu12_6900096	12	6900096	Cd	FASTmrMLM	5.60E-01	ns	0.53
Lu12_6900096	12	6900096	Cd	ISIS EM-BLASSO	5.60E-01	ns	0.53
Lu12_7583472	12	7583472	Sr	GLM	1.20E-04	**	6.06
Lu12_7583472	12	7583472	Sr	MLM	1.20E-04	**	6.06
Lu13_10600206	13	10600206	Sr	GLM	3.30E-01	ns	1.34
Lu13_10600206	13	10600206	Sr	MLM	3.30E-01	ns	1.34
Lu13_10600226	13	10600226	Sr	MLM	5.20E-01	ns	1.08
Lu13_10606843	13	10606843	Sr	GLM	3.00E-02	*	0.31
Lu13_10606843	13	10606843	Sr	MLM	3.00E-02	*	0.31
Lu13_10636361	13	10636361	Sr	GLM	3.90E-03	**	0.74
Lu13_10636361	13	10636361	Sr	MLM	3.90E-03	**	0.74
Lu13_12416850	13	12416850	Ba	ISIS EM-BLASSO	1.60E-01	ns	1.64
Lu13_14078477	13	14078477	Sr	GLM	8.20E-02	ns	0.58
Lu13_14078477	13	14078477	Sr	MLM	8.20E-02	ns	0.58
Lu13_17334670	13	17334670	Sr	GLM	3.70E-02	*	5.40
Lu13_17334670	13	17334670	Sr	MLM	3.70E-02	*	5.40
Lu13_18200591	13	18200591	Ba	FASTmrMLM	8.50E-01	ns	1.07
Lu13_18200591	13	18200591	Sr	FarmCPU	2.60E-01	ns	2.09
Lu13_1825772	13	1825772	Fe	pLARmEB	7.70E-01	ns	0.27
Lu13_20132825	13	20132825	Sr	GLM	2.20E-02	*	0.43
Lu13_20132825	13	20132825	Sr	MLM	2.20E-02	*	0.43
Lu13_20270139	13	20270139	Co	pLARmEB	7.60E-01	ns	0.82
Lu13_3701997	13	3701997	Mn	FASTmrMLM	4.60E-01	ns	0.33
Lu13_3701997	13	3701997	Mn	mrMLM	4.60E-01	ns	0.33
Lu13_4114623	13	4114623	Sr	GLM	5.50E-03	**	0.70
Lu13_4114623	13	4114623	Sr	MLM	5.50E-03	**	0.70
Lu13_4884368	13	4884368	Ti	pLARmEB	5.20E-02	*	6.67
Lu13_4967495	13	4967495	Sr	FASTmrEMMA	2.70E-01	ns	0.11
Lu13_6588638	13	6588638	Sr	MLM	1.90E-03	**	1.03

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu13_8325128	13	8325128	Sr	FASTmrEMMA	6.90E-01	ns	0.56
Lu13_8325128	13	8325128	Sr	mrMLM	6.90E-01	ns	0.56
Lu14_10213635	14	10213635	Sr	GLM	2.70E-03	**	0.85
Lu14_10213635	14	10213635	Sr	MLM	2.70E-03	**	0.85
Lu14_10213659	14	10213659	Sr	GLM	2.90E-01	ns	0.54
Lu14_10213659	14	10213659	Sr	MLM	2.90E-01	ns	0.54
Lu14_10213660	14	10213660	Sr	GLM	6.80E-01	ns	0.55
Lu14_10213660	14	10213660	Sr	MLM	6.80E-01	ns	0.55
Lu14_1224788	14	1224788	Fe	FASTmrMLM	6.50E-01	ns	3.40
Lu14_13333449	14	13333449	Sr	GLM	3.00E-02	*	0.35
Lu14_13333449	14	13333449	Sr	MLM	3.00E-02	*	0.35
Lu14_13793995	14	13793995	Ti	pLARmEB	2.10E-01	ns	2.44
Lu14_13862354	14	13862354	Ni	FASTmrMLM	5.00E-01	**	1.55
Lu14_13862354	14	13862354	Ni	ISIS EM-BLASSO	5.00E-01	**	1.55
Lu14_13881752	14	13881752	Sr	GLM	2.50E-01	ns	0.26
Lu14_13881752	14	13881752	Sr	MLM	2.50E-01	ns	0.26
Lu14_13881755	14	13881755	Sr	GLM	1.70E-01	ns	0.26
Lu14_13881755	14	13881755	Sr	MLM	1.70E-01	ns	0.26
Lu14_14410121	14	14410121	Sr	GLM	1.30E-02	*	0.53
Lu14_14410121	14	14410121	Sr	MLM	1.30E-02	*	0.53
Lu14_15938571	14	15938571	Sr	MLM	2.00E-02	*	0.32
Lu14_15938571	14	15938571	Sr	GLM	2.00E-02	*	0.32
Lu14_2151593	14	2151593	Co	pLARmEB	8.20E-01	**	1.18
Lu14_383230	14	383230	Co	pLARmEB	8.60E-01	**	6.47
Lu14_399600	14	399600	Cd	ISIS EM-BLASSO	8.50E-01	ns	0.78
Lu14_399600	14	399600	Cd	FASTmrMLM	8.50E-01	ns	0.78
Lu14_399600	14	399600	Cd	mrMLM	8.50E-01	ns	0.78
Lu14_4030777	14	4030777	Sr	GLM	9.90E-03	**	3.11
Lu14_4030777	14	4030777	Sr	MLM	9.90E-03	**	3.11
Lu14_7209674	14	7209674	Co	pLARmEB	8.70E-01	ns	1.15
Lu14_8719542	14	8719542	Cd	FASTmrMLM	8.50E-01	ns	0.14
Lu14_9145099	14	9145099	Co	pLARmEB	5.90E-01	ns	2.78
Lu15_12721874	15	12721874	Co	pLARmEB	7.60E-01	ns	0.32
Lu15_14276734	15	14276734	Co	mrMLM	6.00E-01	ns	3.30
Lu15_14844853	15	14844853	Sr	GLM	8.30E-01	ns	0.21
Lu15_14844853	15	14844853	Sr	MLM	8.30E-01	ns	0.21

QTN	Chr[1]	Position	Elements	Models	P-value	KW	R^2 (%)
Lu15_15419144	15	15419144	Sr	GLM	5.90E-01	ns	0.20
Lu15_15419144	15	15419144	Sr	MLM	5.90E-01	ns	0.20
Lu15_15423357	15	15423357	Cd	FASTmrMLM	5.00E-01	**	2.49
Lu15_15558804	15	15558804	Sr	GLM	7.70E-03	**	0.67
Lu15_15558804	15	15558804	Sr	MLM	7.70E-03	**	0.67
Lu15_1635409	15	1635409	Sr	GLM	1.40E-02	*	0.41
Lu15_1635409	15	1635409	Sr	MLM	1.40E-02	*	0.41
Lu15_1635590	15	1635590	Fe	GLM	8.70E-02	ns	0.60
Lu15_4676709	15	4676709	Sr	GLM	1.50E-02	*	3.09
Lu15_4676709	15	4676709	Sr	MLM	1.50E-02	*	3.09
Lu15_5337666	15	5337666	Co	pLARmEB	9.90E-01	ns	3.73
Lu15_5366242	15	5366242	Co	pLARmEB	9.60E-01	ns	0.32
Lu15_5856575	15	5856575	Sr	FASTmrEMMA	2.50E-02	*	0.32
Lu15_5861941	15	5861941	Zn	mrMLM	0.007	ns	0.16
Lu15_8475393	15	8475393	Sr	GLM	1.40E-02	*	0.42
Lu15_8475393	15	8475393	Sr	MLM	1.40E-02	*	0.42
Lu15_950215	15	950215	Sr	GLM	9.90E-03	**	2.01
Lu15_950215	15	950215	Sr	MLM	9.90E-03	**	2.01
Lu15_950215	15	950215	Sr	FASTmrMLM	9.90E-03	**	2.01
Lu15_950215	15	950215	Sr	mrMLM	9.90E-03	**	2.01

[1]Chr: chromosome; Ba: barium; Cd: cadmium; Co: cobalt; Cu: copper; Fe: iron; Mn: manganese; Ni: nickel; Sr: strontium; Ti: titanium; Zn: zinc; KW, Kruskal-Wallis test significance level where "ns," "*," "**," and correspond to non-significant and P-value $\leqslant$ 0.05 and 0.01 respectively.

Appendix 4.4 List of the quantitative trait nucleotides (QTNs) identified by eight genome-wide association studies (GWAS) and the statistical models used for the analysis of ten heavy metal (HM) contents.

QTN	Elements	No. of models	Models
Lu1_10413121	Co	1	pLARmEB
Lu1_10857504	Co	1	pLARmEB
Lu1_11473639	Ti	1	pLARmEB
Lu1_11482977	Ti	1	pLARmEB
Lu1_12260693	Ti	1	pLARmEB
Lu1_12668673	Fe	1	pLARmEB
Lu1_12668673	Sr	1	pLARmEB
Lu1_13413202	Sr	1	FASTmrEMMA
Lu1_13530165	Ti	1	pLARmEB

Chapter 7: Appendices

QTN	Elements	No. of models	Models
Lu1_15267435	Co	1	pLARmEB
Lu1_15912187	Ti	1	pLARmEB
Lu1_1673460	Cd	1	FASTmrMLM
Lu1_16858148	Co	1	pLARmEB
Lu1_17036384	Ni	1	pLARmEB
Lu1_17083119	Co	1	pLARmEB
Lu1_17907764	Ti	1	pLARmEB
Lu1_194465	Sr	1	pLARmEB
Lu1_19847074	Ti	1	pLARmEB
Lu1_20365416	Co	1	pLARmEB
Lu1_20416547	Ti	1	pLARmEB
Lu1_20416547	Zn	1	pLARmEB
Lu1_20723678	Ti	1	pLARmEB
Lu1_21214983	Ba	1	FASTmrMLM
Lu1_25323035	Co	1	pLARmEB
Lu1_25324909	Co	1	pLARmEB
Lu1_25325248	Sr	1	MLM
Lu1_25326421	Sr	1	pLARmEB
Lu1_25326874	Co	1	pLARmEB
Lu1_25331081	Ni	1	pLARmEB
Lu1_25331081	Sr	1	pLARmEB
Lu1_26221446	Co	1	pLARmEB
Lu1_26222995	Co	1	mrMLM
Lu1_26235391	Ti	1	pLARmEB
Lu1_26236507	Fe	1	FASTmrMLM
Lu1_26239219	Ti	1	pLARmEB
Lu1_26244820	Fe	1	pLARmEB
Lu1_26263359	Fe	1	pLARmEB
Lu1_26263736	Co	1	pLARmEB
Lu1_26265593	Co	1	pLARmEB
Lu1_26265593	Ti	1	pLARmEB
Lu1_26268239	Zn	1	pLARmEB
Lu1_26268239	Co	1	pLARmEB
Lu1_26270719	Sr	1	pLARmEB
Lu1_27186534	Sr	1	pLARmEB
Lu1_27752577	Co	1	pLARmEB
Lu1_28480337	Co	1	pLARmEB
Lu1_28480516	Co	1	pLARmEB
Lu1_29338975	Co	1	pLARmEB
Lu1_7310590	Ti	1	pLARmEB
Lu1_8791977	Ti	1	pLARmEB
Lu10_10245781	Sr	1	MLM

QTN	Elements	No. of models	Models
Lu10_4074991	Co	1	pLARmEB
Lu10_4074991	Ti	1	pLARmEB
Lu10_4508210	Ti	1	pLARmEB
Lu10_7704389	Ti	1	pLARmEB
Lu10_7998746	Sr	1	FASTmrEMMA
Lu11_10265112	Fe	1	GLM
Lu11_11079132	Mn	1	ISIS EM-BLASSO
Lu11_12010795	Ni	1	FASTmrMLM
Lu11_19131603	Sr	1	FASTmrMLM
Lu11_3620428	Co	1	pLARmEB
Lu11_4586460	Fe	1	FarmCPU
Lu11_5999195	Ba	1	mrMLM
Lu12_14237214	Co	1	ISIS EM-BLASSO
Lu12_14507321	Mn	1	mrMLM
Lu12_15916883	Ba	1	ISIS EM-BLASSO
Lu12_16265369	Ni	1	ISIS EM-BLASSO
Lu12_16528816	Ba	1	FASTmrMLM
Lu12_16603042	Co	1	ISIS EM-BLASSO
Lu12_18693435	Sr	1	FarmCPU
Lu12_1885304	Co	1	ISIS EM-BLASSO
Lu12_1946222	Sr	1	GLM
Lu12_3817638	Sr	1	GLM
Lu12_3866751	Cd	1	FASTmrMLM
Lu12_3884097	Sr	1	FarmCPU
Lu12_3898188	Fe	1	FASTmrMLM
Lu12_5815671	Sr	1	GLM
Lu13_10600226	Sr	1	MLM
Lu13_12416850	Ba	1	ISIS EM-BLASSO
Lu13_18200591	Ba	1	FASTmrMLM
Lu13_18200591	Sr	1	FarmCPU
Lu13_1825772	Fe	1	pLARmEB
Lu13_20270139	Co	1	pLARmEB
Lu13_4884368	Ti	1	pLARmEB
Lu13_4967495	Sr	1	FASTmrEMMA
Lu13_6588638	Sr	1	MLM
Lu14_1224788	Fe	1	FASTmrMLM
Lu14_13793995	Ti	1	pLARmEB
Lu14_2151593	Co	1	pLARmEB
Lu14_383230	Co	1	pLARmEB
Lu14_7209674	Co	1	pLARmEB
Lu14_8719542	Cd	1	FASTmrMLM
Lu14_9145099	Co	1	pLARmEB

QTN	Elements	No. of models	Models
Lu15_12721874	Co	1	pLARmEB
Lu15_14276734	Co	1	mrMLM
Lu15_15423357	Cd	1	FASTmrMLM
Lu15_1635590	Fe	1	GLM
Lu15_5337666	Co	1	pLARmEB
Lu15_5366242	Co	1	pLARmEB
Lu15_5856575	Sr	1	FASTmrEMMA
Lu15_5861941	Zn	1	mrMLM
Lu2_10037668	Mn	1	mrMLM
Lu2_11254639	Ti	1	pLARmEB
Lu2_11805774	Co	1	FASTmrMLM
Lu2_11917060	Ti	1	pLARmEB
Lu2_11923349	Ti	1	pLARmEB
Lu2_11976307	Fe	1	mrMLM
Lu2_11983094	Sr	1	GLM
Lu2_12547562	Co	1	pLARmEB
Lu2_12565807	Ti	1	pLARmEB
Lu2_12567563	Ti	1	pLARmEB
Lu2_12609212	Co	1	pLARmEB
Lu2_14511171	Ti	1	pLARmEB
Lu2_17240363	Co	1	pLARmEB
Lu2_17523187	Sr	1	pLARmEB
Lu2_17531152	Ti	1	pLARmEB
Lu2_17744333	Co	1	pLARmEB
Lu2_21365734	Ni	1	ISIS EM-BLASSO
Lu2_3830339	Ba	1	mrMLM
Lu2_5517288	Cd	1	ISIS EM-BLASSO
Lu2_6186376	Ti	1	pLARmEB
Lu2_8804882	Co	1	pLARmEB
Lu3_10318345	Ti	1	pLARmEB
Lu3_12076843	Sr	1	pLARmEB
Lu3_12260701	Fe	1	GLM
Lu3_12483044	Co	1	pLARmEB
Lu3_13386515	Co	1	pLARmEB
Lu3_16473464	Co	1	pLARmEB
Lu3_17173441	Ti	1	pLARmEB
Lu3_17971708	Ti	1	ISIS EM-BLASSO
Lu3_17971708	Fe	1	FASTmrMLM
Lu3_19238415	Sr	1	FASTmrMLM
Lu3_19445132	Sr	1	GLM
Lu3_19445178	Sr	1	GLM
Lu3_26404939	Cd	1	ISIS EM-BLASSO

QTN	Elements	No. of models	Models
Lu3_26420475	Sr	1	GLM
Lu3_26422028	Sr	1	GLM
Lu3_26450929	Ti	1	pLARmEB
Lu3_26523437	Co	1	pLARmEB
Lu3_26578461	Sr	1	MLM
Lu3_493530	Sr	1	GLM
Lu3_5508154	Sr	1	GLM
Lu4_11162201	Co	1	pLARmEB
Lu4_127929	Zn	1	mrMLM
Lu4_18692951	Ti	1	pLARmEB
Lu4_19767345	Sr	1	GLM
Lu4_19773831	Sr	1	FASTmrEMMA
Lu4_19880270	Sr	1	GLM
Lu4_2192402	Co	1	pLARmEB
Lu4_7409724	Fe	1	GLM
Lu5_1542156	Sr	1	GLM
Lu5_3768287	Mn	1	mrMLM
Lu5_6117987	Ti	1	pLARmEB
Lu5_6790709	Ti	1	pLARmEB
Lu5_9349523	Sr	1	mrMLM
Lu6_10310871	Fe	1	FASTmrMLM
Lu6_11424106	Co	1	pLARmEB
Lu6_11471289	Sr	1	FarmCPU
Lu6_11506867	Co	1	pLARmEB
Lu6_11564892	Co	1	pLARmEB
Lu6_12894439	Ti	1	pLARmEB
Lu6_14232082	Ti	1	pLARmEB
Lu6_149354	Fe	1	GLM
Lu6_149501	Sr	1	GLM
Lu6_15818175	Sr	1	GLM
Lu6_15818175	Co	1	pLARmEB
Lu6_158363	Fe	1	GLM
Lu6_16239024	Co	1	pLARmEB
Lu6_2121884	Sr	1	FarmCPU
Lu6_3523454	Co	1	pLARmEB
Lu6_3844939	Sr	1	GLM
Lu6_445918	Ti	1	ISIS EM-BLASSO
Lu6_8487801	Co	1	pLARmEB
Lu6_9739632	Mn	1	ISIS EM-BLASSO
Lu7_10374118	Fe	1	mrMLM
Lu7_12701094	Ti	1	pLARmEB
Lu7_17019553	Co	1	pLARmEB

QTN	Elements	No. of models	Models
Lu7_17089067	Co	1	pLARmEB
Lu7_17091064	Ti	1	pLARmEB
Lu7_17099882	Ti	1	ISIS EM-BLASSO
Lu7_18294757	Sr	1	GLM
Lu7_214233	Co	1	pLARmEB
Lu7_221619	Ti	1	pLARmEB
Lu7_2513788	Cu	1	ISIS EM-BLASSO
Lu7_4744454	Co	1	pLARmEB
Lu7_6891825	Sr	1	pLARmEB
Lu7_849864	Co	1	pLARmEB
Lu8_10346733	Sr	1	FarmCPU
Lu8_12047724	Co	1	pLARmEB
Lu8_13063010	Mn	1	FASTmrMLM
Lu8_14073080	Ti	1	pLARmEB
Lu8_1436156	Ni	1	ISIS EM-BLASSO
Lu8_21509453	Fe	1	mrMLM
Lu8_21580044	Fe	1	GLM
Lu8_21594201	Ti	1	pLARmEB
Lu8_21633476	Ni	1	FASTmrMLM
Lu8_23274330	Ti	1	pLARmEB
Lu8_49140	Sr	1	GLM
Lu8_60629	Ti	1	pLARmEB
Lu8_60629	Sr	1	GLM
Lu8_6318953	Co	1	pLARmEB
Lu8_6715730	Ti	1	pLARmEB
Lu8_8839678	Sr	1	FASTmrEMMA
Lu8_9873727	Sr	1	GLM
Lu9_10808332	Sr	1	MLM
Lu9_12465613	Zn	1	FASTmrMLM
Lu9_14152623	Ti	1	pLARmEB
Lu9_14239881	Sr	1	MLM
Lu9_16778244	Cu	1	ISIS EM-BLASSO
Lu9_16878254	Co	1	pLARmEB
Lu9_16881904	Co	1	pLARmEB
Lu9_16882661	Co	1	pLARmEB
Lu9_16907775	Ti	1	pLARmEB
Lu9_16917634	Ti	1	pLARmEB
Lu9_16966095	Co	1	pLARmEB
Lu9_17017285	Co	1	pLARmEB
Lu9_17056635	Ti	1	pLARmEB
Lu9_17073447	Co	1	pLARmEB
Lu9_17200027	Co	1	pLARmEB

QTN	Elements	No. of models	Models
Lu9_17201722	Fe	1	mrMLM
Lu9_17243307	Ti	1	pLARmEB
Lu9_17246640	Co	1	pLARmEB
Lu9_17246640	Ti	1	pLARmEB
Lu9_17261987	Co	1	pLARmEB
Lu9_17280992	Sr	1	MLM
Lu9_17341167	Ni	1	pLARmEB
Lu9_17392318	Fe	1	GLM
Lu9_17415003	Ba	1	ISIS EM-BLASSO
Lu9_17453937	Co	1	pLARmEB
Lu9_17454206	Co	1	pLARmEB
Lu9_17542072	Ti	1	pLARmEB
Lu9_17591047	Co	1	pLARmEB
Lu9_21546287	Ti	1	pLARmEB
Lu9_26183	Sr	1	pLARmEB
Lu9_53969	Co	1	pLARmEB
Lu9_62855	Co	1	pLARmEB
Lu9_63995	Fe	1	GLM
Lu9_7775009	Cd	1	ISIS EM-BLASSO
Lu9_79814	Co	1	pLARmEB
Lu1_11862929	Zn	2	mrMLM,pLARmEB
Lu1_11906143	Sr	2	MLM,GLM
Lu1_13700450	Sr	2	MLM,GLM
Lu1_17878494	Ni	2	ISIS EM-BLASSO,FASTmrMLM
Lu1_26236696	Sr	2	pLARmEB,MLM
Lu1_9376954	Sr	2	mrMLM,FASTmrEMMA
Lu10_15338938	Sr	2	MLM,GLM
Lu10_15338941	Sr	2	MLM,GLM
Lu10_7238569	Cd	2	mrMLM,FASTmrMLM
Lu11_11079129	Fe	2	mrMLM,FASTmrMLM
Lu11_4017458	Sr	2	MLM,GLM
Lu12_16568252	Sr	2	MLM,GLM
Lu12_16615248	Ba	2	FASTmrMLM,mrMLM
Lu12_16627926	Sr	2	MLM,GLM
Lu12_5815732	Sr	2	MLM,GLM
Lu12_7583472	Sr	2	MLM,GLM
Lu13_10600206	Sr	2	MLM,GLM
Lu13_10606843	Sr	2	MLM,GLM
Lu13_10636361	Sr	2	MLM,GLM
Lu13_14078477	Sr	2	MLM,GLM
Lu13_17334670	Sr	2	MLM,GLM
Lu13_20132825	Sr	2	MLM,GLM

QTN	Elements	No. of models	Models
Lu13_3701997	Mn	2	mrMLM,FASTmrMLM
Lu13_4114623	Sr	2	MLM,GLM
Lu13_8325128	Sr	2	mrMLM,FASTmrEMMA
Lu14_10213635	Sr	2	MLM,GLM
Lu14_10213659	Sr	2	MLM,GLM
Lu14_10213660	Sr	2	MLM,GLM
Lu14_13333449	Sr	2	MLM,GLM
Lu14_13862354	Ni	2	ISIS EM-BLASSO,FASTmrMLM
Lu14_13881752	Sr	2	MLM,GLM
Lu14_13881755	Sr	2	MLM,GLM
Lu14_14410121	Sr	2	MLM,GLM
Lu14_15938571	Sr	2	GLM,MLM
Lu14_4030777	Sr	2	MLM,GLM
Lu15_14844853	Sr	2	MLM,GLM
Lu15_15419144	Sr	2	MLM,GLM
Lu15_15558804	Sr	2	MLM,GLM
Lu15_1635409	Sr	2	MLM,GLM
Lu15_4676709	Sr	2	MLM,GLM
Lu15_8475393	Sr	2	MLM,GLM
Lu2_11355622	Cu	2	FASTmrMLM,mrMLM
Lu2_11979691	Sr	2	GLM,mrMLM
Lu2_19721217	Ni	2	ISIS EM-BLASSO,FASTmrMLM
Lu2_653721	Sr	2	MLM,GLM
Lu2_6634979	Sr	2	mrMLM,FASTmrMLM
Lu2_9701680	Sr	2	MLM,GLM
Lu3_14467769	Sr	2	MLM,GLM
Lu3_14468211	Sr	2	MLM,GLM
Lu3_19258260	Sr	2	MLM,GLM
Lu3_19423268	Sr	2	MLM,GLM
Lu3_22029670	Sr	2	MLM,GLM
Lu3_22029673	Sr	2	MLM,GLM
Lu3_22029675	Sr	2	MLM,GLM
Lu3_26578377	Sr	2	MLM,GLM
Lu3_26578392	Sr	2	MLM,GLM
Lu3_4183590	Fe	2	GLM,FarmCPU
Lu3_5479935	Sr	2	MLM,GLM
Lu3_943534	Sr	2	MLM,GLM
Lu4_11270490	Sr	2	MLM,GLM
Lu4_15433671	Sr	2	MLM,GLM
Lu4_163752	Sr	2	MLM,GLM
Lu4_17131337	Sr	2	MLM,GLM
Lu4_19757820	Sr	2	MLM,GLM

Chapter 7: Appendices

QTN	Elements	No. of models	Models
Lu4_19757947	Sr	2	MLM,GLM
Lu4_19905774	Sr	2	MLM,GLM
Lu4_729742	Sr	2	MLM,GLM
Lu4_729751	Sr	2	MLM,GLM
Lu4_7859768	Mn	2	ISIS EM-BLASSO,FASTmrMLM
Lu6_3660509	Zn	2	mrMLM,FASTmrMLM
Lu7_17086138	Ba	2	FASTmrMLM,mrMLM
Lu7_17098332	Sr	2	MLM,GLM
Lu7_17107562	Sr	2	MLM,GLM
Lu7_7795225	Fe	2	MLM,GLM
Lu7_7841206	Fe	2	MLM,GLM
Lu7_7841300	Fe	2	MLM,GLM
Lu8_12304779	Sr	2	MLM,GLM
Lu8_13847239	Sr	2	MLM,GLM
Lu8_21611937	Sr	2	MLM,GLM
Lu8_21617485	Sr	2	FASTmrEMMA,GLM
Lu8_21618742	Sr	2	MLM,GLM
Lu8_21680245	Sr	2	MLM,GLM
Lu8_585308	Sr	2	MLM,GLM
Lu8_587393	Sr	2	MLM,GLM
Lu8_60658	Sr	2	MLM,GLM
Lu8_60659	Sr	2	MLM,GLM
Lu8_75724	Sr	2	MLM,GLM
Lu8_82211	Sr	2	MLM,GLM
Lu8_98342	Sr	2	MLM,GLM
Lu9_14102433	Sr	2	MLM,GLM
Lu9_14239878	Sr	2	MLM,GLM
Lu9_15862043	Sr	2	MLM,GLM
Lu9_16843165	Cd	2	FASTmrMLM,mrMLM
Lu9_17105028	Sr	2	MLM,GLM
Lu9_17192342	Sr	2	MLM,GLM
Lu9_1730550	Cu	2	mrMLM,FASTmrMLM
Lu9_47515	Sr	2	MLM,GLM
Lu9_6843813	Ba	2	FASTmrMLM,mrMLM
Lu9_8387539	Co	2	mrMLM,ISIS EM-BLASSO
Lu9_9585748	Cu	2	FASTmrMLM,mrMLM
Lu1_25318220	Fe	3	MLM,GLM,FarmCPU,
Lu1_27758116	Ni	3	mrMLM,ISIS EM-BLASSO,FASTmrMLM
Lu11_14610978	Ni	3	mrMLM,ISIS EM-BLASSO,FASTmrMLM
Lu12_6900096	Cd	3	ISIS EM-BLASSO,FASTmrMLM,mrMLM
Lu14_399600	Cd	3	mrMLM,FASTmrMLM,ISIS EM-BLASSO
Lu2_11567234	Fe	3	mrMLM,ISIS EM-BLASSO,GLM

QTN	Elements	No. of models	Models
Lu3_7487534	Cd	3	FASTmrMLM,mrMLM,ISIS EM-BLASSO
Lu4_19848927	Fe	3	ISIS EM-BLASSO,MLM,GLM
Lu4_3785453	Mn	3	mrMLM,GLM,FarmCPU
Lu5_6198348	Fe	3	mrMLM,FASTmrMLM,GLM
Lu7_9640759	Ni	3	mrMLM,ISIS EM-BLASSO,FASTmrMLM
Lu9_13338325	Cd	3	FASTmrEMMA,ISIS EM-BLASSO,mrMLM
Lu9_17131228	Cd	3	ISIS EM-BLASSO,mrMLM,FASTmrMLM
Lu15_950215	Sr	4	mrMLM,FASTmrMLM,MLM,GLM
Lu7_7841299	Fe	4	pLARmEB,mrMLM,MLM,GLM
Lu1_15906860	Ni	5	mrMLM,pLARmEB,ISIS EM-BLASSO,FASTmrEMMA,FASTmrMLM

Ba: barium; Cd: cadmium; Co: cobalt; Cu: copper; Fe: iron; Mn: manganese; Ni: nickel; Sr: strontium; Ti: titanium; Zn: zinc

Appendix 4.5 Average proportion of variance of quantitative trait nucleotides (QTNs) identified using eight genome-wide association studies (GWAS) statistical models for the analysis ten heavy metal (HM) contents.

Model types	Models	No. of QTNs	Average R^2 (%) of QTN
SLSM	GLM	115	1.19 ± 1.65
	MLM	89	1.24 ± 1.78
MLSM	FarmCPU	10	2.91 ± 3.30
	FASTmrEMMA	11	1.17 ± 1.29
	FASTmrMLM	45	1.71 ± 1.79
	ISIS EM-BLASSO	35	1.63 ± 1.73
	mrMLM	45	1.85 ± 2.51
	pLARmEB	141	1.24 ± 1.64

SLSM; single-locus statistical model, MLSM; multi-locus statistical model

Chapter 7: Appendices

Appendix 4.6 Quantitative trait nucleotides (QTNs) identified from 42,959 single nucleotide polymorphisms (SNPs) for ten heavy metal (HM) contents using single- and multi-locus genome-wide association studies (GWAS) models.

Elements	SLSMs	MLSMs	SLSMs+MLSMs	Non-redundant	Major QTN	Major QTN effect (R^2, %)	Minor QTN effect (R^2, %)	All QTN effect (R^2, %)
Ba	0	14	14	11	0	NA	1.38 ± 0.9	1.38 ± 0.90
Cd	0	26	26	14	2	6.20 ± 1.35	0.68 ± 0.63	1.47 ± 2.12
Co	0	75	75	74	3	5.59 ± 0.76	1.11 ± 1.08	1.30 ± 1.40
Cu	0	8	8	5	1	6.21 ± NA	1.22 ± 0.73	2.22 ± 2.32
Fe	24	25	49	33	8	8.99 ± 1.41	1.16 ± 1.08	3.12 ± 3.63
Mn	1	12	13	9	0	NA	0.51 ± 0.65	0.51 ± 0.65
Ni	0	28	28	15	0	NA	1.53 ± 1.44	1.53 ± 1.44
Sr	179	36	215	133	6	5.80 ± 0.73	0.84 ± 0.82	1.03 ± 1.26
Ti	0	54	54	54	4	5.82 ± 0.58	0.77 ± 0.77	1.17 ± 1.57
Zn	0	9	9	7	0	NA	0.39 ± 0.64	0.39 ± 0.58
Total	204	287	491	355	24	6.89 ± 1.56	0.95 ± 0.92	1.39 ± 1.83

SLSM; single-locus statistical model; MLSM: multi-locus statistical model. Major QTN are defined as $R^2 \geq 5\%$, while minor QTN as $R^2 < 5\%$. Ba: barium; Cd: cadmium; Co: cobalt; Cu: copper; Fe: iron; Mn: manganese; Ni: nickel; Sr: strontium; Ti: titanium; Zn: zinc

Chapter 7: Appendices

Appendix 4.7 Analysis of variance (ANOVA) for ten heavy metal (HM) contents and eight genome-wide association studies (GWAS) models.

Source	DF	Sum of Square	Mean Square	F value	*Pr* (>F)
Traits	9	240.40	26.71	9.90	6.82e-14 ***
GWAS Models	7	24.10	3.45	1.28	0.259764
Traits × Models	63	157.90	6.08	2.25	0.000492 ***
Residuals	448	1208.80	2.70		

Appendix 4.8 Quantitative trait nucleotides (QTNs) with significant allelic effect based on the non-parametric Kruskal-Wallis test for each heavy metal (HM) contents.

Elements	SLSMs	MLSMs	SLSMs+MLSMs	Non-redundant QTN	No. major QTN
Ba	0	2	2	2	0
Cd	0	5	5	3	2
Co	0	11	11	10	3
Cu	0	2	2	1	1
Fe	11	9	20	10	8
Mn	0	0	0	0	0
Ni	0	12	12	5	0
Sr	93	18	111	70	6
Ti	0	7	7	7	4
Zn	0	0	0	0	0
Total	104	66	170	108	24

Chapter 7: Appendices

Appendix 4.9 List of common ATP binding cassette (ABC) genes putatively associated with heavy metals (HMs) based on their orthology to functionally-characterized *Arabidopsis* genes (Khan et al. 2020).

Flax Gene ID	SNP	R^2 (%)	Functional Annotation	*Arabidopsis* ortholog	Models	Elements
Lus10025824	*Lu1_1673460*	5.24	ABC-2 and Plant PDR ABC-type transporter family protein	*AT1G59870.1*	FASTmrMLM	Cd
Lus10005249	*Lu13_17334670*	5.40	ATP binding cassette subfamily B19	*AT3G28860.1*	MLM,GLM	Sr
Lus10033470	*Lu3_17173441*	5.66	ATP binding cassette subfamily B19	*AT3G28860.1*	pLARmEB	Ti
Lus10004583	*Lu3_7487534*	7.15	P-glycoprotein 9	*AT4G18050.1*	ISIS EM-BLASSO,mrMLM,FASTmrMLM	Cd
Lus10004896	*Lu9_9585748*	6.21	Multidrug resistance-associated protein 6	*AT3G21250.2*	mrMLM,FASTmrMLM	Cu

Cd: cadmium; Cu: copper; Sr: strontium; Ti: titanium

Chapter 7: Appendices

Appendix 4.10 Analysis of variance (ANOVA) for genomic predictive ability (*r*) of genomic selection (GS) models constructed for ten heavy metal (HM) contents, ten statistical models, and two marker sets.

Source	DF	Sum of Square	Mean Square	F value	*Pr* (>F)
Traits	9	0.48	0.05	129.40	< 2e-16 ***
Models	9	0.04	0.01	11.41	2.44e-12 ***
Marker sets	1	20.25	20.25	46999.87	< 2e-16 ***
Traits×Models	81	0.06	0.00	1.59	0.0196 *
Models×Marker sets	9	0.02	0.00	4.69	5.07e-05 ***
Traits×Marker sets	9	0.48	0.05	123.40	< 2e-16 ***
Residuals	81	0.04	0.00		

Appendix 4.11 Predictive ability ($r \pm s$) of ten heavy metal (HM) contents using the trait-defined quantitative trait nucleotides (QTNs) for each metal content as markers along with genome-wide 43K single nucleotide polymorphisms (SNPs). Ten genomic selection (GS) models were used to estimate *r* values.

Elements	GS models	Marker sets	No. of markers	$r \pm s$
Ba	RR-BLUP	QTNs of Ba	11	0.84 ± 0.13a
		43K Markers	42,959	0.22 ± 0.15a
	GBLUP	QTNs of Ba	11	0.82 ± 0.17a
		43K Markers	42,959	0.19 ± 0.13abc
	BRR	QTNs of Ba	11	0.70 ± 0.30b
		43K Markers	42,959	0.19 ± 0.13abc
	BL	QTNs of Ba	11	0.71 ± 0.29b
		43K Markers	42,959	0.16 ± 0.20bc
	BayesA	QTNs of Ba	11	0.71 ± 0.29b
		43K Markers	42,959	0.20 ± 0.14ab
	BayesB	QTNs of Ba	11	0.71 ± 0.30b
		43K Markers	42,959	0.20 ± 0.14ab
	BayesC	QTNs of Ba	11	0.70 ± 0.30b
		43K Markers	42,959	0.19 ± 0.13ab
	RFR	QTNs of Ba	11	0.69 ± 0.25b
		43K Markers	42,959	0.15 ± 0.11c
	SVR	QTNs of Ba	11	0.69 ± 0.30b
		43K Markers	42,959	0.16 ± 0.17bc
	RKHS	QTNs of Ba	11	0.67 ± 0.30b
		43K Markers	42,959	0.17 ± 0.16bc
Cd	RR-BLUP	QTNs of Cd	14	0.81 ± 0.12a
		43K Markers	42,959	0.20 ± 0.17a
	GBLUP	QTNs of Cd	14	0.80 ± 0.18a

Chapter 7: Appendices

Elements	GS models	Marker sets	No. of markers	*r* ± *s*
	BRR	43K Markers	42,959	0.18 ± 0.21abc
		QTNs of Cd	14	0.79 ± 0.22ab
	BL	43K Markers	42,959	0.18 ± 0.21abc
		QTNs of Cd	14	0.79 ± 0.21ab
	BayesA	43K Markers	42,959	0.17 ± 0.22abc
		QTNs of Cd	14	0.81 ± 0.12a
	BayesB	43K Markers	42,959	0.16 ± 0.23abc
		QTNs of Cd	14	0.74 ± 0.13b
	BayesC	43K Markers	42,959	0.15 ± 0.23abc
		QTNs of Cd	14	0.77 ± 0.13ab
	RFR	43K Markers	42,959	0.19 ± 0.21ab
		QTNs of Cd	14	0.76 ± 0.25ab
	SVR	43K Markers	42,959	0.12 ± 0.21c
		QTNs of Cd	14	0.78 ± 0.23ab
	RKHS	43K Markers	42,959	0.15 ± 0.21abc
		QTNs of Cd	14	0.79 ± 0.20ab
		43K Markers	42,959	0.12 ± 0.24bc
Co	RR-BLUP	QTNs of Co	93	0.67 ± 0.14a
		43K Markers	42,959	0.14 ± 0.10a
	GBLUP	QTNs of Co	93	0.66 ± 0.15a
		43K Markers	42,959	0.11 ± 0.09b
	BRR	QTNs of Co	93	0.64 ± 0.19a
		43K Markers	42,959	0.08 ± 0.11cd
	BL	QTNs of Co	93	0.63 ± 0.19a
		43K Markers	42,959	0.12 ± 0.09ab
	BayesA	QTNs of Co	93	0.65 ± 0.18a
		43K Markers	42,959	0.10 ± 0.09bc
	BayesB	QTNs of Co	93	0.64 ± 0.15a
		43K Markers	42,959	0.10 ± 0.09bc
	BayesC	QTNs of Co	93	0.62 ± 0.23a
		43K Markers	42,959	0.11 ± 0.08bc
	RFR	QTNs of Co	93	0.62 ± 0.24a
		43K Markers	42,959	0.10 ± 0.08bc
	SVR	QTNs of Co	93	0.62 ± 0.24a
		43K Markers	42,959	0.10 ± 0.09bc
	RKHS	QTNs of Co	93	0.62 ± 0.21a
		43K Markers	42,959	0.06 ± 0.12d
Cu	RR-BLUP	QTNs of Cu	5	0.83 ± 0.06a
		43K Markers	42,959	0.03 ± 0.21abcd
	GBLUP	QTNs of Cu	5	0.79 ± 0.23ab

Chapter 7: Appendices

Elements	GS models	Marker sets	No. of markers	*r* ± *s*
	BRR	43K Markers	42,959	0.08 ± 0.17ab
		QTNs of Cu	5	0.79 ± 0.24ab
	BL	43K Markers	42,959	0.08 ± 0.17a
		QTNs of Cu	5	0.79 ± 0.24ab
	BayesA	43K Markers	42,959	0.05 ± 0.17abcd
		QTNs of Cu	5	0.75 ± 0.09b
	BayesB	43K Markers	42,959	0.07 ± 0.17abc
		QTNs of Cu	5	0.78 ± 0.07ab
	BayesC	43K Markers	42,959	0.07 ± 0.17abc
		QTNs of Cu	5	0.81 ± 0.07a
	RFR	43K Markers	42,959	0.08 ± 0.17ab
		QTNs of Cu	5	0.78 ± 0.24ab
	SVR	43K Markers	42,959	0.01 ± 0.16d
		QTNs of Cu	5	0.78 ± 0.23ab
	RKHS	43K Markers	42,959	0.03 ± 0.19bcd
		QTNs of Cu	5	0.79 ± 0.23ab
	RR-BLUP	43K Markers	42,959	0.02 ± 0.16cd
		QTNs of Fe	37	0.78 ± 0.13a
	GBLUP	43K Markers	42,959	0.17 ± 0.09ab
		QTNs of Fe	37	0.77 ± 0.15ab
	BRR	43K Markers	42,959	0.16 ± 0.1abc
		QTNs of Fe	37	0.75 ± 0.17ab
	BL	43K Markers	42,959	0.16 ± 0.1abc
		QTNs of Fe	37	0.76 ± 0.18ab
	BayesA	43K Markers	42,959	0.18 ± 0.10a
		QTNs of Fe	37	0.73 ± 0.27abc
Fe	BayesB	43K Markers	42,959	0.17 ± 0.10abc
		QTNs of Fe	37	0.78 ± 0.14a
	BayesC	43K Markers	42,959	0.17 ± 0.10abc
		QTNs of Fe	37	0.72 ± 0.28abc
	RFR	43K Markers	42,959	0.16 ± 0.10abc
		QTNs of Fe	37	0.67 ± 0.33c
	SVR	43K Markers	42,959	0.14 ± 0.09c
		QTNs of Fe	37	0.72 ± 0.17bc
	RKHS	43K Markers	42,959	0.15 ± 0.10bc
		QTNs of Fe	37	0.75 ± 0.15ab
	RR-BLUP	43K Markers	42,959	0.16 ± 0.09abc
		QTNs of Mn	9	0.78 ± 0.11ab
Mn	GBLUP	43K Markers	42,959	-0.12 ± 0.13c
		QTNs of Mn	9	0.80 ± 0.11a

Chapter 7: Appendices

Elements	GS models	Marker sets	No. of markers	*r* ± *s*
	BRR	43K Markers	42,959	-0.07 ± 0.13b
		QTNs of Mn	9	0.77 ± 0.14ab
	BL	43K Markers	42,959	-0.07 ± 0.13b
		QTNs of Mn	9	0.76 ± 0.15abc
	BayesA	43K Markers	42,959	-0.09 ± 0.14bc
		QTNs of Mn	9	0.78 ± 0.12ab
	BayesB	43K Markers	42,959	-0.09 ± 0.14b
		QTNs of Mn	9	0.77 ± 0.14abc
	BayesC	43K Markers	42,959	-0.08 ± 0.14b
		QTNs of Mn	9	0.77 ± 0.14ab
	RFR	43K Markers	42,959	-0.07 ± 0.13b
		QTNs of Mn	9	0.73 ± 0.12c
	SVR	43K Markers	42,959	-0.03 ± 0.14a
		QTNs of Mn	9	0.76 ± 0.13bc
	RKHS	43K Markers	42,959	-0.07 ± 0.13b
		QTNs of Mn	9	0.77 ± 0.14abc
		43K Markers	42,959	-0.07 ± 0.14b
Ni	RR-BLUP	QTNs of Ni	16	0.81 ± 0.06a
		43K Markers	42,959	0.10 ± 0.16a
	GBLUP	QTNs of Ni	16	0.79 ± 0.14ab
		43K Markers	42,959	0.09 ± 0.15a
	BRR	QTNs of Ni	16	0.76 ± 0.26abc
		43K Markers	42,959	0.09 ± 0.15a
	BL	QTNs of Ni	16	0.78 ± 0.17ab
		43K Markers	42,959	0.10 ± 0.16a
	BayesA	QTNs of Ni	16	0.75 ± 0.28abc
		43K Markers	42,959	0.09 ± 0.16ab
	BayesB	QTNs of Ni	16	0.74 ± 0.28abc
		43K Markers	42,959	0.09 ± 0.16a
	BayesC	QTNs of Ni	16	0.76 ± 0.26abc
		43K Markers	42,959	0.09 ± 0.15a
	RFR	QTNs of Ni	16	0.72 ± 0.19c
		43K Markers	42,959	0.01 ± 0.16c
	SVR	QTNs of Ni	16	0.75 ± 0.16abc
		43K Markers	42,959	0.05 ± 0.15bc
	RKHS	QTNs of Ni	16	0.73 ± 0.31bc
		43K Markers	42,959	0.07 ± 0.15ab
Sr	RR-BLUP	QTNs of Ni	148	0.59 ± 0.33a
		43K Markers	42,959	-0.03 ± 0.16c
	GBLUP	QTNs of Ni	148	0.62 ± 0.28a

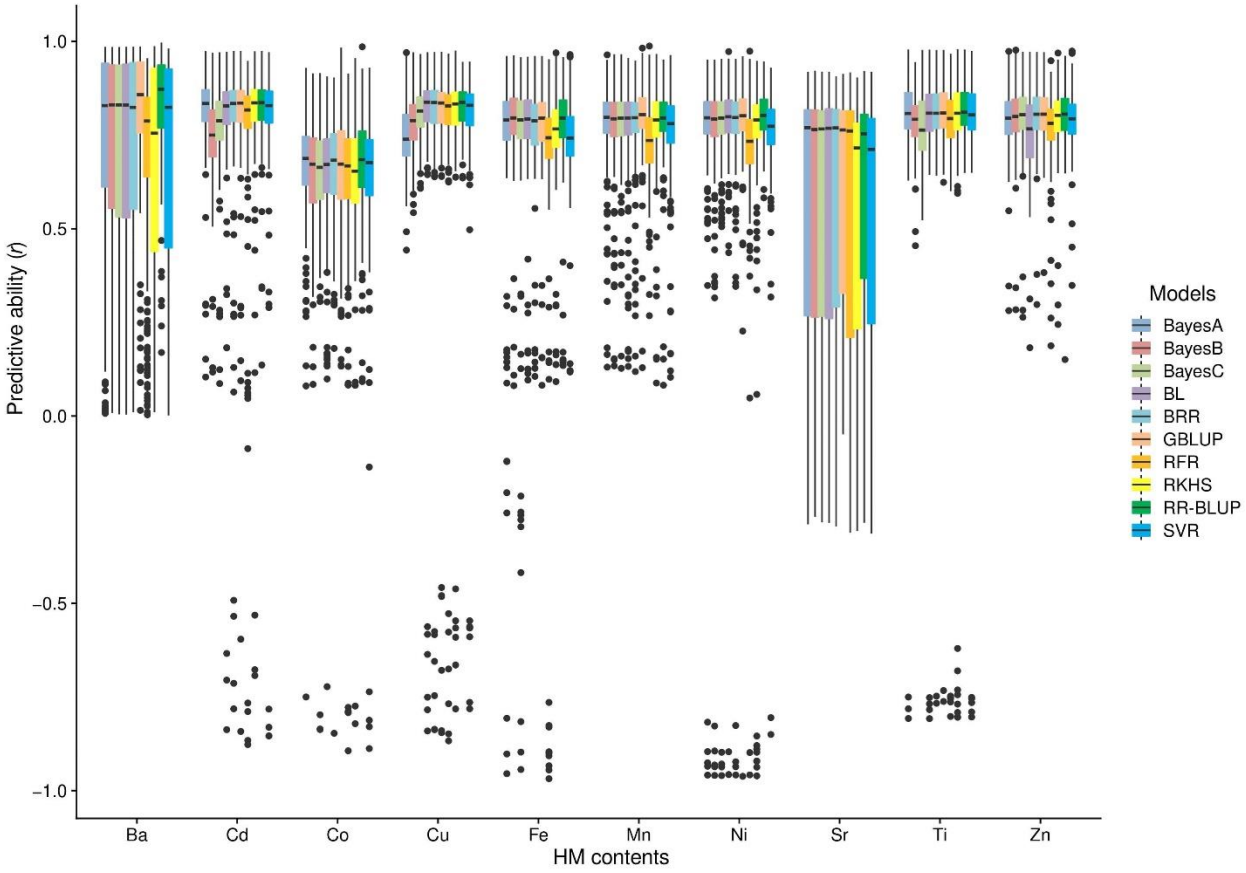
Chapter 7: Appendices

Elements	GS models	Marker sets	No. of markers	*r* ± *s*
	BRR	43K Markers	42,959	0.10 ± 0.15a
		QTNs of Ni	148	0.61 ± 0.31a
	BL	43K Markers	42,959	0.10 ± 0.15a
		QTNs of Ni	148	0.59 ± 0.34a
	BayesA	43K Markers	42,959	0.06 ± 0.15ab
		QTNs of Ni	148	0.61 ± 0.31a
	BayesB	43K Markers	42,959	0.07 ± 0.14ab
		QTNs of Ni	148	0.60 ± 0.31a
	BayesC	43K Markers	42,959	0.07 ± 0.15ab
		QTNs of Ni	148	0.60 ± 0.31a
	RFR	43K Markers	42,959	0.10 ± 0.14a
		QTNs of Ni	148	0.58 ± 0.34a
	SVR	43K Markers	42,959	0.08 ± 0.16ab
		QTNs of Ni	148	0.57 ± 0.31a
	RKHS	43K Markers	42,959	0.08 ± 0.18ab
		QTNs of Ni	148	0.56 ± 0.32a
		43K Markers	42,959	0.06 ± 0.16b
Ti	RR-BLUP	QTNs of Ti	51	0.82 ± 0.07a
		43K Markers	42,959	0.11 ± 0.20c
	GBLUP	QTNs of Ti	51	0.78 ± 0.26ab
		43K Markers	42,959	0.22 ± 0.15a
	BRR	QTNs of Ti	51	0.77 ± 0.3ab
		43K Markers	42,959	0.22 ± 0.15a
	BL	QTNs of Ti	51	0.74 ± 0.36b
		43K Markers	42,959	0.17 ± 0.15ab
	BayesA	QTNs of Ti	51	0.77 ± 0.30ab
		43K Markers	42,959	0.21 ± 0.15a
	BayesB	QTNs of Ti	51	0.78 ± 0.07ab
		43K Markers	42,959	0.20 ± 0.15a
	BayesC	QTNs of Ti	51	0.77 ± 0.10ab
		43K Markers	42,959	0.21 ± 0.15a
	RFR	QTNs of Ti	51	0.72 ± 0.35b
		43K Markers	42,959	0.09 ± 0.18c
	SVR	QTNs of Ti	51	0.74 ± 0.36b
		43K Markers	42,959	0.11 ± 0.25c
	RKHS	QTNs of Ti	51	0.75 ± 0.34ab
		43K Markers	42,959	0.13 ± 0.19bc
Zn	RR-BLUP	QTNs of Zn	6	0.81 ± 0.07a
		43K Markers	42,959	0.05 ± 0.18b
	GBLUP	QTNs of Zn	6	0.80 ± 0.07a

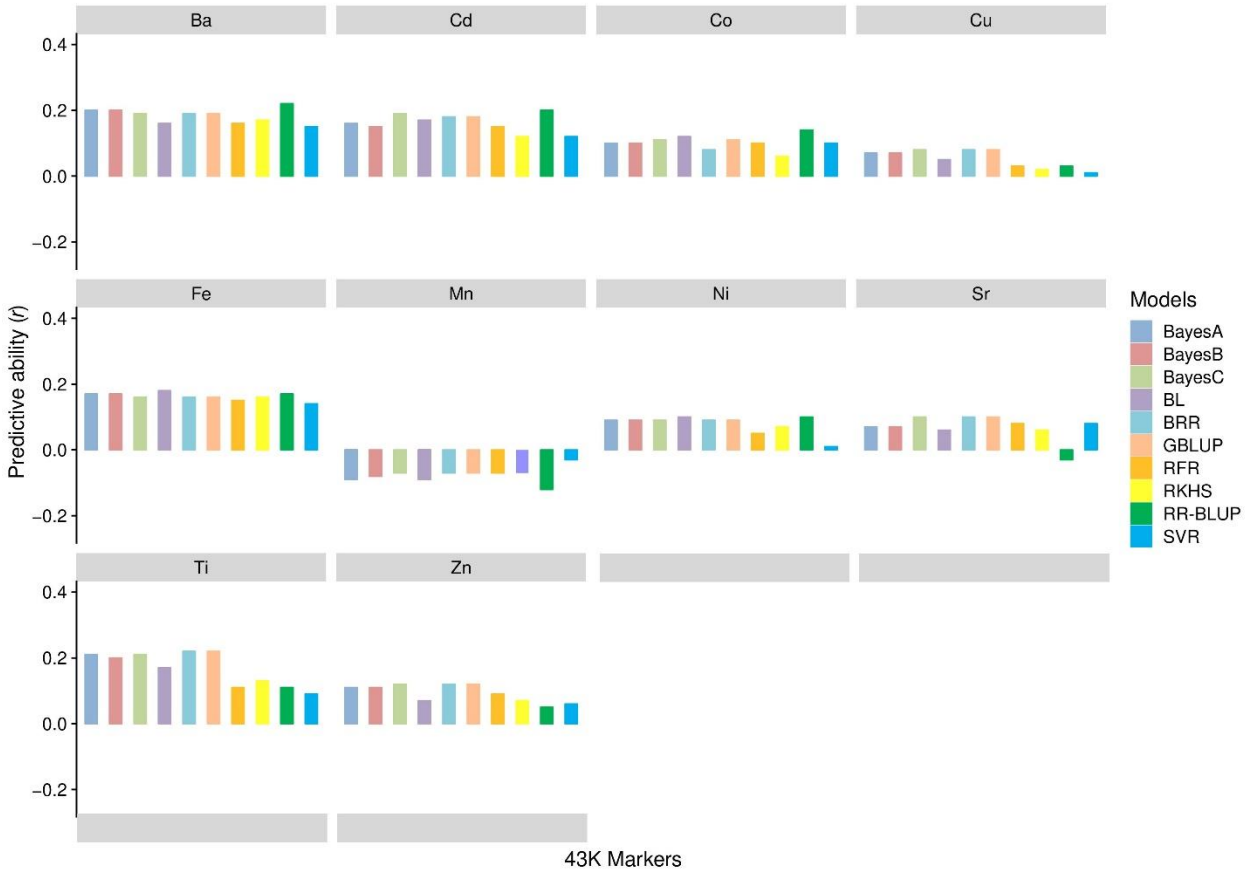
Chapter 7: Appendices

Elements	GS models	Marker sets	No. of markers	*r* ± *s*
	BRR	43K Markers	42,959	0.12 ± 0.15a
		QTNs of Zn	6	0.80 ± 0.08a
	BL	43K Markers	42,959	0.12 ± 0.15a
		QTNs of Zn	6	0.76 ± 0.10c
	BayesA	43K Markers	42,959	0.07 ± 0.16ab
		QTNs of Zn	6	0.79 ± 0.08ab
	BayesB	43K Markers	42,959	0.11 ± 0.15a
		QTNs of Zn	6	0.80 ± 0.08a
	BayesC	43K Markers	42,959	0.11 ± 0.15a
		QTNs of Zn	6	0.80 ± 0.08a
	RFR	43K Markers	42,959	0.12 ± 0.15a
		QTNs of Zn	6	0.77 ± 0.09bc
	SVR	43K Markers	42,959	0.06 ± 0.15b
		QTNs of Zn	6	0.79 ± 0.07ab
	RKHS	43K Markers	42,959	0.09 ± 0.17ab
		QTNs of Zn	6	0.80 ± 0.08a
		43K Markers	42,959	0.07 ± 0.14b

Ba: barium; Cd: cadmium; Co: cobalt; Cu: copper; Fe: iron; Mn: manganese; Ni: nickel; Sr: strontium; Ti: titanium; Zn: zinc



Appendix 4.12 Comparisons of genomic predictive ability (*r*) of ten genomic selection (GS) models for ten heavy metal (HM) contents *per se* using the trait-defined quantitative trait nucleotides (QTNs) for each metal content: 11 (Ba), 14 (Cd), 93 (Co), 5 (Cu), 37 (Fe), 9 (Mn), 16 (Ni), 148 (Sr), 51 (Ti) and 6 (Zn). Each panel displays the *r* values obtained from the ten GS models using the trait-defined QTNs. The number of QTNs used in the models are on the x-axis. Five-fold cross-validation was used to estimate *r* values.



Appendix 4.13 Comparisons of genomic predictive ability (*r*) of ten genomic selection (GS) models for ten heavy metal (HM) contents *per se*. Each panel displays the *r* values obtained from the models using the overall set of 43K single nucleotide polymorphisms (SNPs) detected for all ten metal contents. Five-fold cross-validation was used to estimate *r* values.