LEVERAGING GENOMICS AND TRANSCRIPTOMICS FOR GENE DISCOVERY IN DRY PEA

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Josephine Princy Johnson

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Program:
Genomics, Phenomics and Bioinformatics

April 2024

Fargo, North Dakota

North Dakota State University
Graduate School

LEVERAGING GENOMICS AND TRANSCRIPTOMICS FOR GENE DISCOVERY IN DRY PEA

By

Josephine Princy Johnson

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Nonoy Bandillo
Chair

Dr. Changhui Yan

Dr. Mingao Yuan

Dr. Luis Del Rio Mendoza

Approved:

April 12, 2024
Date

Dr. Changhui Yan
Department Chair

ABSTRACT

Over the past two decades, there has been a significant increase in the utilization of DNA marker-based mapping studies to genetically map and further improve complex quantitative traits. A major caveat of this approach is that genetic mapping of the underlying genes conferring target phenotypes is challenging often due to the extent of long-range linkage disequilibrium (LD) in the genome, particularly in self-pollinated crops. Recent technologies allow us to examine expression-phenotype associations using transcriptome-wide association studies (TWAS) which is independently affected by LD, unlike in the case of genetic markers. This is of greatest utility in species where linkage disequilibrium is extensive such as dry pea, where genes can be prioritized for association with a trait because their expression patterns are independent. The goal of this study is to use gene expression collected from the developing pods of pea and the TWAS approach for mapping and prioritizing likely causal genes underlying seed protein content and yield. As the effective population size (N_e) of the USDA (United States Department of Agriculture) diversity panel provided substantial genetic variation, we utilized 300 USDA pea lines from within the collection and performed a comprehensive single-tissue, multi-environment TWAS across six diverse environments (2 years * 2 locations) in the major pea growing regions in the USA. As we compared the results of TWAS with genome-wide association studies (GWAS), we detected more common and unique set of strongly associated genes. In all TWAS models, the significant genes exhibited clear differentiation, unlike in the case of GWAS. A joint analysis of GWAS and TWAS results using the fisher's combined test (FCT) increased the power of detecting more trait-associated genes including *RGB*. Using GWAS, TWAS and FCT models, we detected 45 genes for protein, 60 genes for yield, and 20 genes that were common to both traits. These results highlight the complex interaction between genetic factors and

environmental influences in shaping the genetic architecture of seed yield and protein. Our study proved that multi-omics strategy increases the gene mapping resolution by surpassing the GWAS and/or TWAS approach, and highlights the potential phenotypic consequences of regulatory variation in dry pea.

ACKNOWLEDGMENTS

I would like to thank my major advisor Dr. Nonoy Bandillo for being a great mentor and for his support through this PhD. As a newcomer in plant science, his guidance proved invaluable in facilitating my transition. I am really honored to be one of his first PhD graduate students. I also want to thank my lab specialist Lisa Piche who taught me a lot of things within the lab and field work. I am really grateful to be a part of this lab where they prioritized the student's education and training.

I extend my sincere gratitude to my esteemed committee members, Dr. Changhui Yan, Dr. Mingao Yuan, and Dr. Luis Del Rio Mendoza, for their unwavering support. I would like to express my heartfelt appreciation to all the professors from my courses for their exceptional dedication to education and their profound brilliance, which has greatly enriched my learning experience.

Finally, I want to thank my family and friends for believing in me.

DEDICATION

To my family and my love who provided endless support in my journey through tough times.

I also want to dedicate this dissertation to myself for not giving up.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BIC	Bayesian Information Criterion
BLUE	Best linear Unbiased Estimates
BLUP	Best Linear Unbiased Prediction
Chr	Chromosome
cM	Centimorgans
DEG	Differentially Expressed Gene
eRD-GWAS	Expression Read Depth – Genome-wide Association Studies
FCT	Fisher's Combined Test
FCT-ME	Fishers' Combined Test – Multi-Environment
FCT-ND	Fisher's Combines Test – North Dakota
FCT-WA	Fisher's Combined Test – Washington
FDR	False Discovery Rates
FFP7	Filament-like protein 7
GWAS	Genome-wide Association Studies
INDEL	Insertion/Deletion
K	Kinship
LD (r^2)	Linkage Disequilibrium
LSE	Least Square Estimates
LTR	Long-Terminal Repeat
MEM	Multi-Environment BLUPs
ME-TWAS	Multi-Environment Transcriptome-wide association Studies
MLMM	Multi-Locus Mixed Model
NCBI	National Center of Biotechnology Information

NDSU	North Dakota State University
ND-TWAS	North Dakota – Transcriptome-wide Association Studies
N_e	Effective Population Size
PC	Principal Components
PCA	Principal Component Analysis
RGB	Cell wall biosynthesis
SNP	Single Nucleotide Polymorphism
TE	Transposable Elements
TIR	Terminal-Inverted Repeat
TWAS	Transcriptome-wide Association Studies
USDA	United States Department of Agriculture
VCF	Variant Calling File
WA-TWAS	Washington – Transcriptome-Wide Association Studies
WEB	Within-Environment BLUPs
ZW6	Zhongwar6

LIST OF APPENDIX FIGURES

CHAPTER 1. LITERATURE REVIEW

Introduction to *Pisum Sativum* (L.)

Pisum sativum (L.) is a diploid, cool-season legume and a member of the Fabaceae/Leguminosae plant family, the third largest flowering plant family with approximately 18,000 species and 800 genera (Lewis 2005). Based on archaeological evidence, the existence of peas goes back to 10,000 BC in Near East (Zohary and Hopf 2000) and Central Asia (Riehl et al., 2013). They have been cultivated throughout the Stone and Bronze Ages in Europe, and since 200 BC in India (Warkentin et al., 2015). Peas were domesticated in the Near East around 9000 BC and are one of the world's oldest domesticated crops, originating in the primary center of origin in the Near and Middle east. The crown Leguminosae divergence is associated with the whole genome duplication event in the pea genome (55 MYA). Since the pea's divergence from other tribes, the pea genome has experienced more nucleotide mutations, gene duplications, and deletions than other sequenced legume genomes (Kreplak et al., 2019). According to homology and synteny computation, there is a synteny relationship identified between pea (*P. sativum*), peanut diploid ancestor (*Arachis duranensis*), lotus (*Lotus japonicus*), barrel medic (*Medicago truncatula*), chickpea (*Cicer arietinum*), pigeon pea (*Cajanus cajan*), soybean (*Glycine max*), common bean (*Phaseolus vulgaris*), mung bean (*Vigna radiata*) and adzuki bean (*Vigna angularis*) (Kreplak et al., 2019). Peas are among the most important pulse crops grown in over 100 countries, with 7,043,605 hectares of dry peas planted worldwide and a total production of 12,403,522 tonnes (FAOSTAT 2021). In the USA alone, the pea production reached one million tonnes in 2019 (USDA 2020). Pea seeds are renowned as a dietary goldmine, containing approximately 32% protein, as well as vitamins, folate, fibers, potassium, and minerals, all of which contribute to human health and aid in the prevention of cardiovascular and certain cancer

diseases (Bari et al., 2021; Tayeh et al., 2015). The genetic diversity of dry peas has been collected from various regions, including Europe, Africa, Asia, America, and Oceania, where they have adapted to diverse environments (Kreplak et al., 2019). Research on the pea genome has lagged behind due to its large genome size compared to other small legume genomes. However, it was the model genome studied by Gregor Mendel in the 18th century as he uncovered the laws of genetics (Ellis et al., 2011). Genes controlling Mendel's seven pea characteristics are known to be located on four chromosomes: chromosome 1 for seed and blossom color, chromosome 4 for height, inflorescence, and pod shape, chromosome 5 for pod color, and chromosome 7 for seed form (Yang et al., 2022; SMÝKAL 2014).

Pisum sativum L. has seven chromosomes (2n=14) with a genome size of 4.45 GB. In recent years, the Caméor pea reference genome has been successfully utilized in genetic mapping studies to detect the variants such as single nucleotide polymorphisms (SNP) and insertion/deletion (Indel) (Kreplak et al., 2019). According to the USDA Agricultural Research Service and Northern Pulse Growers Association databases, 76 genetic maps were generated until 2022. Based on Kreplak et al., (2019), the pea genome has a high occurrence of repetitive sequence and is one of the legume species used as a model for having the most repetitive sequence in its genome. According to the pea reference paper, the annotation step detected 2,225,175 repetitive elements clustered into 2,940 consensuses representing ~83% of the genome. The majority of them are transposable elements (TE), such as the long-terminal repeat (LTR) retrotransposon accounting for 72.7% of the genome. Transposons (class II) make up 5.4% of the genome, with terminal-inverted-repeat (TIR) transposons accounting for 84%. The majority of the gap between the genome and the reference assembly are explained by collapsed sequences of repetitive elements (Kreplak et al., 2019). A total of 57,835 transcripts and 44,756

genes are present in the genome. In chr 1, there are 5,142 genes and 6,599 transcripts; chr 2 (4,365 genes) – 5,775 transcripts; chr 3 (4,751 genes) – 6,115 transcripts; chr 4 (5,276 genes) – 6,924 transcripts; chr 5 (7,098 genes) – 9,208 transcripts; chr 6 (5,593 genes) – 7,586 transcripts and chr 7 (6,087 genes) – 7,943 transcripts. In 2022, a new improved reference genome, CAAS_Psat_ZW6_1.0 was developed from the cultivar Zhongwar6 (ZW6), with a genome size of 3.8 GB and 65,672 genes (Yang et al., 2022a), surpassing the Caméor genome, which has a size of 3.9 GB and 44,756 genes (Kreplak et al., 2019). The latest genome analysis presents a newly generated assembly and annotation of the ZW6 cultivar's genome. The N-50 contig length is 8.98 Mb, representing a 243-fold improvement compared to the Caméor reference genome (Yang et al., 2022a). Since the number of gene counts between the genomes differs, some genes were successfully validated and compared with the *Caméo*r genome, but the functions of the remaining genes are still unknown. Our study will be the first to utilize this new improved genome in the association mapping to determine the trait-associated genes. During the gene validation stages, we will utilize annotations from both genomes, genomic locus, and model information from NCBI (National Center for Biotechnology Information; https://www.ncbi.nlm.nih.gov/gene/?term=pisum+sativum) and the pea genome database developed by the Chinese Academy of Agricultural Sciences (Yang et al., 2022b) to verify the gene's function.

Genetic Diversity and Effective Population Size (N_e)

Genetic diversity is an important factor in a population for gene mapping studies and it also impacts the strength of the population (Frankham 1996). The number of genotypes in the population determines the inbreeding rate and genetic drift, as well as the likelihood of deleterious alleles becoming fixed, leading to a decrease in genetic variation (Hare et al., 2011;

Lonsinger et al., 2018). Effective population size (N_e) is an important parameter in evolution and conservation biology (Waples 2006). N_e refers to the census population size which will be lower than the N number of population (Lonsinger et al., 2018) and represents the strength in the genetic variation. Based on the Wright-Fisher model, this is the number of individuals in an idealized population that would exhibit a comparable genetic response to stochastic processes, similar to those observed in real-world populations (Wang et al., 2016; Wright 1931; Fisher 1930). It is used by the breeders to determine the health of the genotypes and the long-term risk (Frankham, 2005; Hare et al., 2011). The genetic variations will be lost and the population will be at risk, when there is smaller N_e and limited gene flow (Fagan & Holmes 2006; Palstra & Ruzzante 2008). To avoid short-term inbreeding, N_e should be at least greater than or equal to 50, with the population size of minimum 500 (Franklin 1980). N_e can assist in preserving genetic diversity within the breeding population by evaluating the magnitude of genetic variation. This information enables breeders to determine whether adjustments to their strategy, such as introducing new lines, are necessary in order to prevent a reduction in genetic diversity (Morais Júnior et al., 2017).

Commonly used extensions for effective population size theory are variance effective size and inbreeding effective size (Wang et al., 2016). Variance effective size reflects the rate of change in gene frequency variance, while inbreeding effective size corresponds to the observed rate of inbreeding in a population (Crow and Kimura 1970). These measures enable the quantification of the consequences of genetic drift in real populations, based on the characteristics and dynamics of the idealized Wright-Fisher population (Wang et al., 2016).

There are multiple predictive equations available to estimate effective population size (N_e). The equations differ based on the subpopulation division with varying sizes and pedigrees

(Wang et al., 2016; Wang 1997a, b). Wright (1938) derived an initial equation that accounts for how the variance from a parent in a population contributes to the progeny. This equation assumed the population contains equal number of males and females. The predictive equation of animals and X-linked genes were derived by (Wang et al., 2016; Nomura 2002, 2005). We can also estimate N_e using linkage disequilibrium (LD) between the molecular markers, where the equation is derived based on r^2 (LD) and the genetic distance (c) in Morgans. This formula assumes that there is no selective process occurring within the population and that selfing is permitted (Sved 1971). This equation suits well with our populations being studied. We have a genetic dataset which can be used to estimate N_e. From the recent advancements in high-throughput sequencing and the availability of high-density markers such as single nucleotide polymorphisms (SNPs) have increased over the past decade, contributing to the LD-based N_e estimation being acknowledged as more reliable, robust (Novo et al., 2022), cost-effective, and time-effective compared to the temporal approach (Gargiulo et al., 2023).

Peas are one of the highly consumed alternate sources of protein in the recent years, maintaining its variability and stability is the top priority for the breeders in their populations. Since there is no N_e information available, estimation of N_e in peas would be a valuable resource for all the pea breeding programs around the world. Studies such as Juma et al., (2021) estimated the effective population size (N_e) in rice to be 22 using the SNP markers dataset from an elite core panel composed of 72 lines. However, *Ne* may have been underestimated due to limited marker information used in the analysis and some of the markers were specifically designed to *indigo* lines. Similar studies in rice also yielded N_e values within the same range, with calculated values ranging from 23-57 and 40-60; these estimates were based on breeding populations from recurrent selection programs (Grenier et al., 2015) and pedigree data (Morais Júnior et al., 2017).

In soybean, Xavier et al., (2018) estimated N_e for the USDA (United States Department of Agriculture) soybean germplasm collection, composed of 19,652 accessions from Bandillo et al., (2015), reporting it to be 106 individuals. Recent studies have revealed several genetic bottlenecks in soybean (Guo et al., 2010), leading to a reduction in its genetic diversity (Li et al., 2013; Min et al., 2010). Zhao et al., (2013) estimated N_e in wild rice using 11 Chinese Oryza rufipogon populations, including 32 landraces, with reported values between 96 and 158. Other ways such as detecting the ratio of effective population size and the census size (N_e / N) which ranges from 0 to 1 to determine the genetic diversity of the population. For grass germplasm collection species, Johnson et al., (2002) estimated 0.69 for *L. perenne* which was lower than 0.86 *P. spicata* species.

Linkage Disequilibrium (LD)

Linkage Disequilibrium (LD) is a non-random correlation of alleles at various loci (Hill and Robertson 1968). This concept has become popular in estimating effective population size (N_e) (Antao et al., 2011) and in mapping studies. Correlations between alleles are usually generated by genetic drift, which is inversely proportional to N_e (Gargiulo et al., 2023), leading to changes in allele frequencies in a population over time. The advantage of LD over the temporal method (Pollak 1983) lies in the strength of associations between markers, allowing for accurate N_e calculations at any given time (generations) from a single population, without relying on longitudinal data. This makes LD a valuable tool for studying populations where temporal information may be limited or unavailable. Recombination and mutation rates are the main factors that shape the genetic landscape (Ardlie et al., 2002), and by analyzing LD, we can gain a better understanding of their history and apply this knowledge to plant breeding and population genetics (Sved and Hill 2018).

LD findings in previous studies were observed in peas, where both wild and spring peas exhibited a decay distance of approximately 200 kb, whereas wild/landrace peas showed a distance around 100 kb (Siol et al., 2017). In the Beji et al., (2020) study using 365 accessions from a diversity panel, the LD decayed to 0.22 with the distance of 0.9 cM. Based on each chromosome, the decay rate ranges from 0.3 to 1.4 cM. Comparing the LD of peas to other selfing crops such as rice, soybeans, and barley, the physical distances found were more or less similar depending on the type of populations. Huang et al., (2010) estimated LD using *O. indica* and *O. japonica* landraces of rice at 123 and 167 kb, respectively, with r^2 declining to 0.25 and 0.28. Additionally, soybean landraces extended from 90 to 500 kb (Hyten et al., 2007), while improved cultivars hit 133 kb (Zhou et al., 2015). A recent LD analysis from soybean USDA germplasm revealed that the r^2 dropped intragenically within a few kilobases (Xavier et al., 2018), and in barley's landraces, it reached 90 kb (Caldwell et al., 2006). The LD-decay of elite varieties in barley extends up to 212 kb (Caldwell et al., 2006) and in *O. japonica* elite lines at ∼318 kb (Li et al. 2020), but it declined faster in *O. indica* elite lines, around ∼124 kb (Li et al., 2020).

Genome-Wide Association Studies (GWAS)

Genome-wide association studies (GWAS) is one of the popular mapping studies for the past two decades. GWAS detects the correlation among the molecular markers and the phenotypic variations by utilizing historical recombination events (Hirschhorn and Daly, 2005) and genetic diversity (Korte and Farlow, 2013; Gali et al., 2019) in many crops including pulses (Sun et al., 2017). GWAS has higher resolution compared to traditional linkage mapping which uses biparental population to detect the causal trait-associated marker (Korte and Farlow, 2013; Gali et al., 2019). Due to the cost-effective genotyping technologies and developments of

statistical methods, GWAS is widely used to understand the genetic complexity in most species. In most GWAS studies, researchers prefer to use single nucleotide polymorphisms (SNPs), as they are easy to genotype and abundant in the respective genome. As the higher number of markers in the analysis will help tag the causal gene for the respective traits.

In Gali et al., (2019), the authors conducted a GWAS study across multiple trials for different traits. They found highly significant SNPs such as Chr1LG6_57305683 and *Chr1LG6_366513463* for seed yield while *Chr3LG5_194530376* for protein concentration was found commonly in all of their trials. For Aphanomyces root rot in peas, Desgroux et al., (2016) determined 52 significant markers using a 13.2k SNP array. Tafesse et al., (2020) detected 32 markers associated with heat stress responsive traits using 16,877 SNPs. Beji et al., (2020) identified 62 highly significant SNPs associated with frost tolerance in winter pea crop. In the soybean genome, Priyanatha et al., (2022) determined *Glyma.19 g171000* as a significant gene associated with seed oil concentration. Zhang et al., (2023) detected two significant SNPs such as *Gm09_39012959* and *Gm20_24678362* for protein concentration whereas for fat content, they identified four SNPs such as *Gm09_39012959, Gm12_35492373, Gm16_9297124,* and *Gm20_24678362*. Researchers also identified 55 significant genes linked to seven root traits in soybean (Kim et al., 2023). In the study conducted by Zeng et al., (2022) on the maize genome, they identified 59 SNPs that were highly significant for yield-related traits. Through the analysis of the LD rate, they discovered that these SNPs were associated with 58 annotated genes. Zhang et al., (2021) detected 22 significant SNPs associated to yield-related phenotypes in *japonica* rice. Multi-Locus Mixed Model (MLMM) approach evaluated by Segura et al., (2012) also increased the statistical power of GWAS in mapping the genes but the accuracy and gene validation still remains unpredictable.

Tools to Perform GWAS

There are more resources and tools developed to conduct GWAS analysis. Statistical models such as naïve, population structure (Q), kinship (K) and Q+K models were used to perform GWAS (Sharma et al., 2018). Since GWAS is being conducted using germplasms collected all over the world, the population structure and the genetic relatedness should be accounted for in the analysis to prevent false positive SNPs (Yu et al., 2005). Still, the performance of GWAS is not sufficient due to its low statistical power and high false positive signals. More advanced tools have been developed to perform GWAS including GAPIT (Tang et al., 2016), ECMLM (Li et al., 2014), EMMA (Kang et al., 2008), GEMMA (Zhou and Stephens, 2012), FaST-LMM (Lippert et al., 2011), SUPER (Wang et al., 2014) and GenABEL (Svishcheva et al., 2012) among others. The power of GWAS is influenced by factors such as phenotypic variation, population structure, number of genotypes, allele frequency, and LD. Quality control, data preparation, and the use of best linear unbiased predictor (BLUP) and best linear unbiased estimator (BLUE) are necessary steps to adjust for phenotypic variation in GWAS. Additionally, it is crucial to carefully consider the decay of linkage disequilibrium (LD) when conducting GWAS (Alqudah et al., 2020).

Transcriptome-Wide Association Studies (TWAS)

Majority of research focusing on detecting the genetic complexity underlying the phenotypic variation are based on association mapping studies. For the past two decades, researchers have been using molecular markers such as SNPs in GWAS to tag the causal gene associated with the trait-of-interest. But in some species, due to linkage disequilibrium, they might end up retrieving multiple genes linked to the causal variant (highly significant marker). It is impossible to narrow it down which one of those genes are actually our targeted gene of

interest. Long-range of LD decay increases the probability of detecting false positive genes. Having a high LD in a GWAS population means, we need to have a smaller number of SNP markers to reduce false positives. The mapping resolution in GWAS studies is based on the number of markers and LD decay rate (Alqudah et al., 2020). Among previous studies in plants, researchers have tried to incorporate the expression data into the mapping studies and performed transcriptome-wide association studies. Transcriptome-wide association study (TWAS) examines the expression-phenotype associations which are independently affected by LD, unlike in the case of genetic markers (Li et al., 2021). In simpler terms, even if multiple genes are closely connected and cannot be observed separately in different individuals, they can still be given priority for association with a trait because their patterns of expression are independent. Other studies conducted by Lin et al., (2017) and Zheng et al., (2020) have already demonstrated that TWAS can achieve single gene resolution and found partially overlapping gene sets in cross-pollinated species. Additionally, Li et al., (2021) found that TWAS is also effective in self-pollinated species. These studies have shown that TWAS overcomes issues related to Linkage Disequilibrium (LD) and have stated that TWAS is a valuable complement to GWAS.

As discussed before, there are so many options to conduct GWAS analysis but for TWAS, the choices were limited. Different TWAS approaches that were published have different results. For example, results from Hirsch et al., (2014) and Lin et al., (2017) only had one or few genes that passed the false discovery rate threshold (FDR) while in Kremling et al., (2019) and Wu et al., (2022) they used top 0.5-1% hits based on p-values and ended up getting more significant genes which included false positive discoveries as well. Reasons behind these varying results are due to the dataset format used in their respective analysis. Li et al., (2021) converted the expression data to categorical format (0,1,2) to utilize it in the GAPIT R package and

compared it with the GWAS results using the same tool whereas (Kremling et al., 2019; Wu et al., 2022) directly used the normalized gene expression data into the mixed linear model. Since the datasets utilized in GWAS and TWAS analysis were different, the authors used the outlier approach to study the genes to avoid direct comparison of p-values resulted from different datasets rather than using the multiple corrections approach (Kremling et al., 2019; Wu et al., 2022). There is also another method called expression read depth GWAS (eRD-GWAS) was developed by (Lin et al., 2017) in maize which is a TWAS-based Bayesian statistical method. The authors found 13 trait-associated genes and observed that the functions of these genes align with the characteristics of those traits and also one of those genes underwent functional characterization (Lin et al., 2017). In contrast to certain human TWAS methods, which predicted gene expression levels, eRD-GWAS utilized explanatory variables that were explicitly used to measure gene expression levels (Li et al., 2021).

Combining GWAS and TWAS

In 2019, there was a new approach introduced within mapping studies where genomics and transcriptomics were combined to increase the power of gene mapping. Kremling et al., (2019), used Fisher's combined statistical test for the top 10% GWAS highly significant SNPs and TWAS results which led to detect more numbers of known significant genes than running GWAS or TWAS alone in 30 grain carotenoid abundance, 22 agronomic and 20 tocochromanol abundance traits. Following their study, this approach has been successfully applied in sorghum to identify targeted genes correlated with variations of water use efficiency-related traits (Ferguson et al., 2021; Pignon et al., 2021) and in tocochromanol levels in maize grain (Wu et al., 2022). The Fisher's combined test has also been successfully employed to retrieve the candidate genes associated with the maize leaf cuticular conductance (g_c) trait (Lin et al., 2022).

References

Alqudah AM, Sallam A, Baenziger PS, Börner A (2020). GWAS: Fast-forwarding gene identification and characterization in temperate Cereals: lessons from Barley – A review. *Journal of Advanced Research* **22**: 119–135.

Antao T, Pérez-Figueroa A, Luikart G (2011). Early detection of population declines: high power of genetic monitoring using effective population size estimators. *Evol Appl* **4**: 144–154.

Ardlie KG, Kruglyak L, Seielstad M (2002). Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3**: 299–309.

Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J et al. (2015). A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection. *Plant Genome* **8**: eplantgenome2015.04.0024.

Bari MAA, Zheng P, Viera I, Worral H, Szwiec S, Ma Y, *et al.* (2021). Harnessing Genetic Diversity in the USDA Pea Germplasm Collection Through Genomic Prediction. *Front Genet* **12**: 707754.

Beji S, Fontaine V, Devaux R, Thomas M, Negro SS, Bahrman N, et al. (2020). Genome-wide association study identifies favorable SNP alleles and candidate genes for frost tolerance in pea. *BMC Genomics* **21**: 1–21.

Caldwell KS, Russell J, Langridge P, Powell W (2006). Extreme Population-Dependent Linkage Disequilibrium Detected in an Inbreeding Plant Species, Hordeum vulgare. *Genetics* **172**: 557–567.

Crow JF, Kimura M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row: New York, USA.

Desgroux A, L'Anthoëne V, Roux-Duparque M, Rivière J-P, Aubert G, Tayeh N, *et al.* (2016). Genome-wide association mapping of partial resistance to Aphanomyces euteiches in pea. *BMC Genomics* **17**: 1–21.

Ellis TH, Hofer JM, Timmerman-Vaughan GM, Coyne CJ, Hellens RP (2011). Mendel, 150 years on. *Trends Plant Sci* **16**: 590 - 596.

Fagan, WF & Holmes EE (2006). Quantifying the extinction vortex. *Ecology Letters* **9**: 51–60.

FAOSTAT (2021). Food and Agricultural Organization of the United Nations. Available at: https://www.fao.org/faostat/

Ferguson JN, Fernandes SB, Monier B, Miller ND, Allen D, Dmitrieva A, *et al.* (2021). Machine learning-enabled phenotyping for GWAS and TWAS of WUE traits in 869 field-grown sorghum accessions. *Plant Physiol* **187**: 1481 - 1500.

Fisher RA (1930). *The genetical theory of natural selection*. Oxford University Press. Oxford

Frankham R (1996). Relationship of Genetic Variation to Population Size in WildlifeLa Relación Entre la Variación Genética y el Tamaño Poblacional en Vida Silvestre. *Conserv Biol* **10**: 1500–1508.

Frankham R (2005). Genetics and extinction. Biological Conservation, **126**: 131–140.

Franklin IR (1980). Evolutionary change in small populations In Soulé M. E., & Wilcox B. A. (Eds.), Sunderland, MA: Sinauer Associates. *Conservation biology: An evolutionary-ecological perspective* (pp. 135–149).

Gali KK, Sackville A, Tafesse EG, Lachagari VBR, McPhee K, Smýkal P, *et al.* (2019). Genome-wide association mapping for agronomic and seed quality traits of field pea (*Pisum sativum* L.). *Front Plant Sci* **10**: 466624.

Gargiulo R, Decroocq V, González-Martínez SC, Paz-Vinas I, Aury J-M, Kupin IL, et al. (2023). Estimation of contemporary effective population size in plant populations: limitations of genomic datasets. *bioRxiv*: 2023.07.18.549323.

Grenier C, Cao T-V, Ospina Y, Quintero C, Châtel MH, Tohme J, et al. (2015). Accuracy of Genomic Selection in a Rice Synthetic Population Developed for Recurrent Selection Breeding. *PLoS One* **10**: e0136594.

Guo J, Wang Y, Song C, Zhou J, Qiu L, Huang H, et al. (2010). A single origin and moderate bottleneck during domestication of soybean (Glycine max): implications from microsatellites and nucleotide sequences. *Ann Bot* **106**: 505–514.

Hare MP, Nunney L, Schwartz MK, Ruzzante DE, Burford M, Waples RS, et al. (2011). Understanding and estimating effective population size for practical application in marine species management. *Conserv Biol* **25**: 438–449.

Hill WG, Robertson A (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet* **38**: 226–231.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, *et al.* (2014). Insights into the Maize Pan-Genome and Pan-Transcriptome. *Plant Cell* **26**: 121–135.

Hirschhorn JN, Daly MJ (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**: 95–108.

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* **42**: 961–967.

Hyten DL, Choi I-Y, Song Q, Shoemaker RC, Nelson RL, Costa JM, et al. (2007). Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* **175**: 1937–1944.

Johnson RC, Bradley VL, Evans MA (2002). Effective Population Size during Grass Germplasm Seed Regeneration. *Crop Sci* **42**: 286–290.

Juma RU, Bartholomé J, Thathapalli Prakash P, Hussain W, Platten JD, Lopena V, et al. (2021). Identification of an Elite Core Panel as a Key Breeding Resource to Accelerate the Rate of Genetic Improvement for Irrigated Rice. *Rice* **14**: 92.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, *et al.* (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**: 1709–1723.

Kim S-H, Tayade R, Kang B-H, Hahn B-S, Ha B-K, Kim Y-H (2023). Genome-Wide Association Studies of Seven Root Traits in Soybean (Glycine max L.) Landraces. *Int J Mol Sci* **24**: 873.

Korte A, Farlow A (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9**: 1–9.

Kremling KAG, Diepenbrock CH, Gore MA, Buckler ES, Bandillo NB (2019). Transcriptome-wide association supplements genome-wide association in *Zea mays*. *G3: Genes|Genomes|Genetics* **9**: 3023.

Kreplak J, Madoui M-A, Cápal P, Novák P, Labadie K, Aubert G, *et al.* (2019). A reference genome for pea provides insight into legume genome evolution. *Nat Genet* **51**: 1411–1422.

Lewis GP (2005). *Legumes of the World*. Royal Botanic Gardens Kew.

Li D, Liu Q, Schnable PS (2021). TWAS results are complementary to and less affected by linkage disequilibrium than GWAS. *Plant Physiol* **186**: 1800–1811.

Li M, Liu X, Bradbury P, Yu J, Zhang Y-M, Todhunter RJ, *et al.* (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biol* **12**: 1–10.

Li X, Chen Z, Zhang G, Lu H, Qin P, Qi M, *et al.* (2020). Analysis of genetic architecture and favorable allele usage of agronomic traits in a large collection of Chinese rice accessions. *Sci China Life Sci* **63**: 1688–1702.

Li Y-H, Zhao S-C, Ma J-X, Li D, Yan L, Li J, et al. (2013). Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* **14**: 579.

Lin H-Y, Liu Q, Li X, Yang J, Liu S, Huang Y, *et al.* (2017). Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS. *Genome Biol* **18**: 192.

Lin M, Qiao P, Matschi S, Vasquez M, Ramstein GP, Bourgault R, et al. (2022). Integrating GWAS and TWAS to elucidate the genetic architecture of maize leaf cuticular conductance. *Plant Physiol* **189**: 2144–2158.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011). FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**: 833–835.

Lonsinger RC, Adams JR, Waits LP (2018). Evaluating effective population size and genetic diversity of a declining kit fox population using contemporary and historical specimens. *Ecol Evol* **8**: 12011–12021.

Luikart G, Ryman N, Tallmon DA, Schwartz MK, Allendorf FW (2010). Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conserv Genet* **11**: 355–373.

Min W, Run-zhi L, Wan-ming Y, Wei-jun D (2010). Assessing the genetic diversity of cultivars and wild soybeans using SSR markers. *African Journal of Biotechnology* **9**: 4857–4866.

Morais Júnior OP, Breseghello F, Duarte JB, Morais OP, Rangel PHN, Coelho ASG (2017). Effectiveness of Recurrent Selection in Irrigated Rice Breeding. *Crop Sci* **57**: 3043–3058.

Nomura T (2002). Effective size of populations with unequal sex ratio and variation in mating success. *J Anim Breed Genet* **119**: 297–310.

Nomura T (2005). Effective population size under random mating with a finite number of matings. *Genetics* **171**: 1441–1442.

Novo I, Santiago E, Caballero A (2022). The estimates of effective population size based on linkage disequilibrium are virtually unaffected by natural selection. *PLoS Genet* **18**: e1009764.

Palstra FP & Ruzzante DE (2008). Genetic estimates of contemporary effective population size: What can they tell us about the importance of genetic stochasticity for wild population persistence? *Molecular Ecology* **17**: 3428–3447.

Petr SMÝKAL (2014). Pea (*Pisum sativum* L.) in Biology prior and after Mendel's Discovery. *Czech J. Genet. Plant Breed* **50**: 52 – 64.

Pignon CP, Fernandes SB, Valluru R, Bandillo N, Lozano R, Buckler E, *et al.* (2021). Phenotyping stomatal closure by thermal imaging for GWAS and TWAS of water use efficiency-related genes. *Plant Physiol* **187**: 2544 - 2562.

Pollak E (1983). A new method for estimating the effective population size from allele frequency changes. *Genetics* **104**: 531–548.

Priyanatha C, Torkamaneh D, Rajcan I (2022). Genome-wide association study of soybean germplasm derived from Canadian × Chinese crosses to mine for novel alleles to improve seed yield and seed quality traits. *Front Plant Sci* **13**: 866300.

Riehl S, Zeidi M, Conard NJ (2013). Emergence of Agriculture in the Foothills of the Zagros Mountains of Iran. *Science* **341**: 65 - 67.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, *et al.* (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* **44**: 825–830.

Sharma SK, MacKenzie K, McLean K, Dale F, Daniels S, Bryan GJ (2018). Linkage Disequilibrium and Evaluation of Genome-Wide Association Mapping Models in Tetraploid Potato. *G3: Genes|Genomes|Genetics* **8**: 3185.

Siol M, Jacquin F, Chabert-Martinello M, Smýkal P, Le Paslier M-C, Aubert G, et al. (2017). Patterns of Genetic Structure and Linkage Disequilibrium in a Large Collection of Pea Germplasm. *G3* **7**: 2461–2471.

Sun C, Zhang F, Yan X, Zhang X, Dong Z, Cui D, *et al.* (2017). Genome-wide association study for 13 agronomic traits reveals distribution of superior alleles in bread wheat from the Yellow and Huai Valley of China. *Plant Biotechnol J* **15**: 953–969.

Sved JA (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol* **2**: 125–141.

Sved JA, Hill WG (2018). One Hundred Years of Linkage Disequilibrium. *Genetics* **209**: 629–636.

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012). Rapid variance components–based method for whole-genome association analysis. *Nat Genet* **44**: 1166–1170.

Tafesse EG, Gali KK, Lachagari VBR, Bueckert R, Warkentin TD (2020). Genome-Wide Association Mapping for Heat Stress Responsive Traits in Field Pea. *Int J Mol Sci* **21**: 2043.

Tang Y, Liu X, Wang J, Li M, Wang Q, Tian F, *et al.* (2016). GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *Plant Genome* **9**: lantgenome2015.11.0120.

Tayeh N, Klein A, Le Paslier M-C, Jacquin F, Houtin H, Rond C, *et al.* (2015). Genomic Prediction in Pea: Effect of Marker Density and Training Population Size and Composition on Prediction Accuracy. *Front Plant Sci* **6**: 162438.

USDA (2020). United States Acreage. *National Agricultural Statistics Service*. https://www.nass.usda.gov/Publications/Todays_Reports/reports/acrg0620.pdf. Accessed 15 August 2023

Wang J (1997a). Effective size and F-statistics of subdivided populations. I. Monoecious species with partial selfing. *Genetics* **146**: 1453–1463.

Wang J (1997b). Effective size and F-statistics of subdivided populations. II. Dioecious species. *Genetics* **146**: 1465–1474.

Wang J, Santiago E, Caballero A (2016). Prediction and estimation of effective population size. *Heredity* **117**: 193–206.

Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z (2014). A SUPER Powerful Method for Genome Wide Association Study. *PLoS One* **9**: e107684.

Waples RS (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci*. *Conserv Genet* **7**: 167–184.

Warkentin, T.D., P. Smykal, C.J. Coyne, N. Weeden, C. Domoney, D. Bing, A. Leonforte, Z. Xuxiao, P. Dixit, L. Boros, K.E. McPhee, R.J. McGee, J. Burstin, and T.H.N. Ellis. (2015). Pea. In: De Ron, A.M. (ed). Handbook in Plant Breeding: Grain Legumes. *Springer*, pp 37-73.

Wright S (1931). Evolution in Mendelian Populations. *Genetics* **16**: 97–159.

Wright S (1938). Size of population and breeding structure in relation to evolution. *Science* **87**: 430–431.

Wu D, Li X, Tanaka R, Wood JC, Tibbs-Cortes LE, Magallanes-Lundback M, et al. (2022). Combining GWAS and TWAS to identify candidate causal genes for tocochromanol levels in maize grain. *Genetics* **221**: iyac091.

Xavier A, Thapa R, Muir WM, Rainey KM (2018). Population and quantitative genomic properties of the USDA soybean germplasm collection. *Plant Genet Resour* **16**: 513–523.

Yang T, Liu R, Luo Y, Hu S, Wang D, Wang C, *et al.* (2022a). Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nat Genet* **54**: 1553–1563.

Yang T, Liu R, Luo Y, Hu S, Wang D, Wang C, et al. (2022b). Pea Genome Database developed by Institute of Crop Sciences, Chinese Academy of Agricultural Sciences. Available at: https://www.peagdb.com/index/

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, *et al.* (2005). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 203–208.

Zeng T, Meng Z, Yue R, Lu S, Li W, Li W, *et al.* (2022). Genome wide association analysis for yield related traits in maize. *BMC Plant Biol* **22**: 1–11.

Zhang G, Wang R, Ma J, Gao H, Deng L, Wang N, *et al.* (2021). Genome-wide association studies of yield-related traits in high-latitude japonica rice. *BMC Genomic Data* **22**: 1–12.

Zhang Q, Sun T, Wang J, Fei J, Liu Y, Liu L, et al. (2023). Genome-wide association study and high-quality gene mining related to soybean protein and fat. *BMC Genomics* **24**: 1–12.

Zhao Y, Vrieling K, Liao H, Xiao M, Zhu Y, Rong J, et al. (2013). Are habitat fragmentation, local adaptation and isolation-by-distance driving population divergence in wild rice *Oryza rufipogon*? *Mol Ecol* **22**: 5531–5547.

Zheng Z, Hey S, Jubery T, Liu H, Yang Y, Coffey L, *et al.* (2020). Shared Genetic Control of Root System Architecture between Zea mays and Sorghum bicolor. *Plant Physiol* **182**: 977–991.

Zhou X, Stephens M (2012). Genome-wide Efficient Mixed Model Analysis for Association Studies. *Nat Genet* **44**: 821.

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol* **33**: 408–414.

Zohary D, Hopf M (2000). Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Asia, Europe, and the Nile Valley. Oxford University Press, Oxford.

CHAPTER 2. EFFECTIVE POPULATION SIZE IN DRY PEA[1]

Introduction

Dry pea (*Pisum sativum* L.), is a diploid, cool-season legume and a member of the Leguminosae family (Abbo et al., 2017). Pea is one of the most important pulse crops grown in more than 100 countries, where 7,043,605 hectares of dry pea were planted around the world with a total production of 12,403,522 tons (FAOSTAT 2021). In the USA alone, the pea production reached one million tons in 2019 (USDA 2020). In recent years, pea protein has become more popular in the market for plant-based diets e.g., Beyond® Meat Burger (Bari et al. 2021). Pea seeds have earned a reputation as a dietary goldmine with around 15 – 32% protein content, vitamins, folate, fibers, potassium and minerals, which is good for human health and helps prevent cardiovascular and specific cancer diseases (Bari et al., 2021; Tayeh et al., 2015). The increasing popularity of plant-based proteins in the market has further propelled the demand for peas. Therefore, the study of genetic diversity should expand to accelerate the genetic gain of pea varieties to meet future demands, maintaining the diversity in peas is the top priority for plant breeders (Bari et al., 2021; Gali et al., 2019).

Estimation of effective population size (N_e) determines the rate of inbreeding (Rahimmadar et al., 2021; Tenesa et al., 2007) and genetic changes due to genetic drift (Gargiulo et al., 2023). N_e is an important parameter in population genetics and breeding introduced by Sewall Wright in 1931, which helps breeders to maintain and monitor the level of genetic

[1] This chapter has been published in a preprint server as *Johnson, J.P., Piche, L., Worral, H., Atanda, S.A., Coyne, C. J., McGee, R., McPhee, K., & N. Bandillo. 2024. Effective Population Size in Field Pea. doi: https://doi.org/10.1101/2024.02.19.581041*. It was also previously submitted for publication to an open-access journal and is currently under review. Josephine Princy Johnson developed the pipeline and conducted the formal analysis, prepared the original draft and edited the manuscript.

diversity in their species (Cobb et al., 2019). The estimated N_e is expected to be smaller than the census size (N), as it influences the rate at which genetic diversity decreases within a population (Lonsinger et al., 2018; Hare et al., 2011). Relatively smaller N_e indicates limited population diversity, which, in turn, can restrict genetic advancement within a breeding program (Hayes et al., 2003). Moreover, N_e parameter retrieves the population dynamics of the genes (Nei and Tajima 1981).

The effective size of a population refers to the hypothetical number of individuals in an idealized population that would exhibit a comparable genetic response to stochastic processes, similar to that observed in a real-world population which is based on the Wright-Fisher model (Wang et al., 2016; Wright 1931; Fisher 1930). This model shows genetic drift as the main operating factor, and that changes in allelic and genotypic frequencies over generations are solely influenced by the population size (N) (Wang et al., 2016). In real-world breeding populations, factors such as mutation, migration, natural selection, and non-random mating come into play (Wang et al., 2016) These factors affect the actual rates of inbreeding and changes in gene frequency variance observed in a population (Charlesworth 2009). This will indeed impact N_e and therefore, reduce the genetic variation and diversity. The most commonly used extensions for effective population size theory are variance effective size and inbreeding effective size (Wang et al., 2016). The variance effective size reflects the rate of change in gene frequency variance, while inbreeding effective size corresponds to the rate of inbreeding observed in a population (Crow and Kimura 1970). These measures allow us to quantify the consequences of genetic drift in a real population, based on the characteristics and dynamics of the idealized Wright-Fisher population (Wang et al., 2016).

While N_e of a population can be estimated either from demographic data or genetic markers, the latter is preferred (Gilbert and Whitlock 2015; Luikart et al., 2010; Fernández et al., 2005). Demographic data involves using census size and variance of reproductive success whereas genetic markers reveal changes in allele frequencies over time and are based on linkage disequilibrium (LD). When the pedigree or demographic data is not available, N_e can be estimated using genetic markers (Wang 2005). The most popular and widely-employed genetic approach has been the temporal method, which relies on temporal fluctuations in allele frequencies observed on multiple samples collected from the same population (Nei and Tajima 1981). N_e, however, can also be directly estimated using LD between loci at various distances along the genome (Hayes et al., 2003; Sved 1971). Recent advancements in high-throughput sequencing and the availability of high-density markers such as single nucleotide polymorphisms (SNPs) have increased over the past decade, contributing to the LD-based approach now being acknowledged as more reliable, robust (Novo et al., 2022), cost and time effective than the temporal approach (Gargiulo et al., 2023).

Linkage disequilibrium (represented as r^2) is a phenomenon characterized by the non-random association of alleles at various loci (Hill and Robertson 1968) which became popular in recent years for predicting N_e (Antao et al., 2011). Correlations between alleles are generated by genetic drift when it is inversely proportional to N_e (Gargiulo et al., 2023), which changes the allele frequencies in a population over time. The biggest advantage of LD over the temporal method (Pollak 1983), is the strength of associations between markers that can be used to calculate N_e at any time (generations) from a single population accurately without relying on longitudinal data. This makes LD a valuable tool for studying populations where temporal information may be limited or unavailable. Recombination and mutation rates are fundamental

processes that shape the genetic landscape (Ardlie et al., 2002), and by analyzing LD, we can better understand their history and apply it to plant breeding and population genetics (Sved and Hill 2018).

In this study, we estimated the extent of LD decay in the dry pea genome and utilized the relationship between LD and recombination frequency, as initially described by Sved (1971), to estimate N_e which is convenient as it only requires one sampling time (García-Cortés et al. 2019; Hill 1981). We used two sets of populations: 1) NDSU modern breeding lines, hereafter referred to as NDSU set, and 2) USDA diversity panel, hereafter referred to as USDA set. Our objectives were two-fold: (i) to estimate N_e for these two germplasms set in dry pea and (ii) to compare the genetic variation between these germplasms. To achieve these goals, we developed a comprehensive R package that implements the Sved (1971) formula for N_e prediction. This package not only caters to the specific needs of dry pea research but can also be adapted for use in other crop species. Since there has been no information on N_e for peas, our findings serve as a valuable reference for researchers seeking to determine the minimum number of lines required for designing experiments. Furthermore, comparing the genetic variation between NDSU modern breeding lines and USDA multi-environmental lines provides valuable information about the diversity and potential of these germplasm collections. This knowledge can guide breeding programs and conservation efforts, ensuring the maintenance and enhancement of genetic resources in dry pea cultivation.

Materials and Methods

Plant Materials

In this study, we used plant materials from two distinct germplasms. The first population comes from the NDSU Pulse Breeding Program (NDSU set) where 300 advanced elite lines were

generated from multiple bi-parental populations. These lines were created specifically with a focus on phenotypes including high yield, grain quality, resistance to disease and some other desirable agronomic traits. The breeding lines used in this experiment were carefully chosen and contain both contemporary and past elite germplasm. (Bari et al. 2023; Atanda et al. 2022).

The second population is from a USDA diversity panel (USDA set), and contained 482 accessions, of which 292 samples were from the Pea Single Plant Plus Collection (Pea PSP) (Bari et al., 2021; Holdsworth et al., 2017; Cheng et al., 2015). The USDA set was composed of accessions that represent most of available diversity within the USDA pea germplasm collection based on the knowledge of geography, taxonomy, morphology and genotyping-by-sequencing data generated previously (Holdsworth et al., 2017).

DNA Extraction, Sequencing and Variant Calling

Leaf tissues from the greenhouse were collected at different stages for all NDSU elite lines and USDA accessions. The DNA from the lyophilized tissues were extracted using the DNeasy Plant Mini Kit (Qiagen). Detailed information regarding the tissue collections and extractions are provided in Bari et al., (2023) and Bari et al., (2021). Both NDSU set and USDA set were sequenced using genotyping-by-sequencing (GBS). Using the restriction enzyme *ApeKI*, dual-indexed GBS libraries for both populations were prepared (Elshire et al., 2011). Samples were sequenced using NovaSeq S1 × 100 Illumina sequencing technologies. The NDSU set sequenced libraries were retrieved with a quality score ≥ 30. For the USDA set, FASTQC (Andrews 2010) was utilized to perform a quality check and removed reads with lengths < 50 bases. All reads that passed the quality check were aligned with the reference genome (Kreplak et al., 2019) (https://www.pulsedb.org). Finally, the aligned reads were analyzed using SAMtools (v1.10) and generated the variant files (VCF) using FreeBayes (V1.3.2).

The amount of single nucleotide polymorphisms (SNPs) identified for the NDSU set was 28,832, while 380,527 SNP markers were identified in the USDA set (Bari et al., 2021, 2023; Atanda et al., 2022). For these marker datasets, we filtered minor allele frequency (MAF), since alleles with < 5% could produce bias to the LD and N_e calculations (Toosi et al. 2010, Lee et al., 2014). We also removed markers with more than 20% missing values using Plink v1.9 (Purcell et al., 2007) and heterozygosity > 20% using Tassel v5.0 (Bradbury et al., 2007). The resulting marker sets consisted of 7,157 (NDSU set) and 19,826 (USDA set) SNP markers that were used for downstream analysis.

Calculation of Linkage Disequilibrium (r^2)

LD was calculated using Plink v1.9 (Purcell et al., 2007) with a maximum distance of 750 kb. Using "ggplot2" R package, the genome-wide and chromosome-wide LD-decay (r^2) were visualized against the physical distance (kb) to show the recombination history (see Fig. 2.1 & 2.2).

LD scores were also estimated using Genome-wide Complex Trait Analysis (GCTA) software for a window size of 1000 kb and r^2 cutoff of 0 (Yang et al., 2011). This approach was employed to visualize the distribution of mean LD throughout the genome.

Calculation of Effective Population Size

Effective population size (N_e) for both the NDSU set and the USDA set were estimated based on LD using the Sved (1971) equation 1.1. The recombination rate (cM) was calculated using cM/Mb conversion ratio from a recent pea genetic linkage map (Sawada et. al., 2022) and then transformed to Morgan's (c).

$$N_e = \frac{1}{4c}\left(\frac{1}{E(r^2)} - 1\right) \quad (1.1)$$

Where, N_e = effective population size

c = genetic distance in Morgan's

$E(r^2)$ = expected r^2

The expected r^2 was predicted by linear regression model using least square estimation (LSE),

Prediction of r^2:

$$\hat{\mu} = X\hat{\beta} \tag{1.2}$$

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{1.3}$$

$$\mathbf{X} = \begin{bmatrix} 1 & c_1 \\ 1 & c_2 \\ . & . \\ . & . \\ 1 & c_n \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} \text{mean\_r}_1^2 \\ . \\ . \\ \text{mean\_r}_n^2 \end{bmatrix} \tag{1.4}$$

The mean r^2 from the $\boldsymbol{Y}$ parameter was calculated by LD (r^2) for the genetic distance 'c' using 'group by' mean function in R Environment (R Core Team 2023). Now with the availability of all required parameters, we finally estimated N_e from Eq. 1.1 using LSE.

According to the formula (Eq. 1.1), we assigned the variables as predictor (**X**) and response (**Y**) and calculated the coefficient $\boldsymbol{\beta}_1$ without the intercept term $\boldsymbol{\beta}_0$, following Juma et al., (2021).

$$\boldsymbol{Y} = \left(\frac{1}{\hat{\mu}}\right) - 1, \ \mathbf{X} = 4 \times c \tag{1.5}$$

$$\boldsymbol{X} = \begin{bmatrix} 4c_1 \\ 4c_2 \\ . \\ . \\ . \\ 4c_n \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} \left(\frac{1}{\hat{\mu}_1}\right) - 1 \\ . \\ . \\ \left(\frac{1}{\hat{\mu}_n}\right) - 1 \end{bmatrix} \tag{1.6}$$

Again, we used equation 1.3 to calculate the coefficient $\boldsymbol{\beta}_1$ which represents N_e**.**

Results and Discussion

Linkage Disequilibrium Decay Rate and Scores

The decay of linkage disequilibrium (r^2) was examined in both the NDSU set and USDA set by utilizing 7,157 and 19,826 SNP markers, respectively. This analysis allowed for the identification of the physical distance at which the decay rate occurred. Appendix Figure A.1 depicts the distribution of SNPs within and across chromosomes for both populations, illustrating the marker density. The NDSU set's genome-wide LD-decay plot (Figure 2.1) demonstrates that the r^2 reached its peak value of 0.57 within the initial kilobases and subsequently exhibited a gradual decline. The r^2 showed a decrease from 0.3 to 0.25 when the genomic distance increased from 150 kb to 250 kb. Following that, the LD within each chromosome was observed visually in Figure 2.2 in order to improve comprehension of the decay pattern. Chromosomes 1 and 6 exhibited a rapid decay at approximately 175 kb, while chromosomes 2 and 5 demonstrated a comparatively slower decay rate of around 350 kb. Furthermore, it is worth noting that chromosome 5 had a higher r^2 value of 0.61 compared to other chromosomes. Whereas, the genome-wide LD of USDA set showed that r^2 started at a lower value of 0.34 and dropped rapidly and reached 0.2 and 0.1 at 100 kb and 200 kb (Figure 2.1). From the chromosome-wide LD-decay (Figure 2.2), we observed that chromosome 3 dropped faster around ~150 kb, but the r^2 decreased below 0.1 for chromosomes 4 and 7. Also, chromosomes 1, 5 and 6 decayed slowly (~250 kb) and reached r^2 0.1. We also observed that chromosome 1 exhibited a higher r^2 of 0.37. LD-decay figures show the trend of the r^2 decaying from LD to linkage equilibrium (LE).

Additionally, we performed calculations of LD scores as an alternative metric for inferring LD. The analysis of local LD in the NDSU set indicates a notable rise in the average r^2 of 0.6 across all chromosomes. The average r^2 of chromosomes 5 and 6 was the highest with 0.8.

The genomic interval encompassing the centromeric region of chromosome 2 was missing. In contrast, the USDA set exhibited low average r^2, with chromosome 2 hardly reaching 0.4, and chromosomes 1, 4, and 7 having few sets that reached 0.3. It is worth noting that the LD density of the NDSU set is comparatively lower than the USDA set (Figure 2.3).

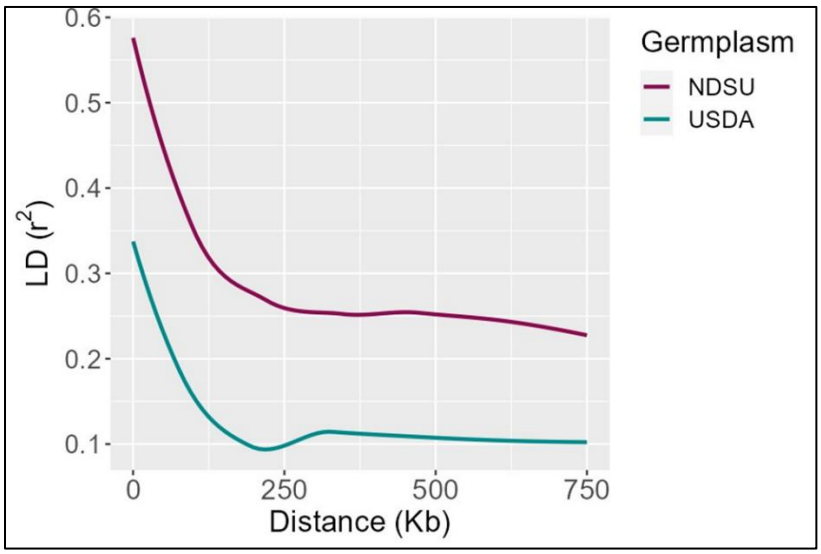

Figure 2.1. Genome-wide linkage disequilibrium - decay of NDSU set and USDA set

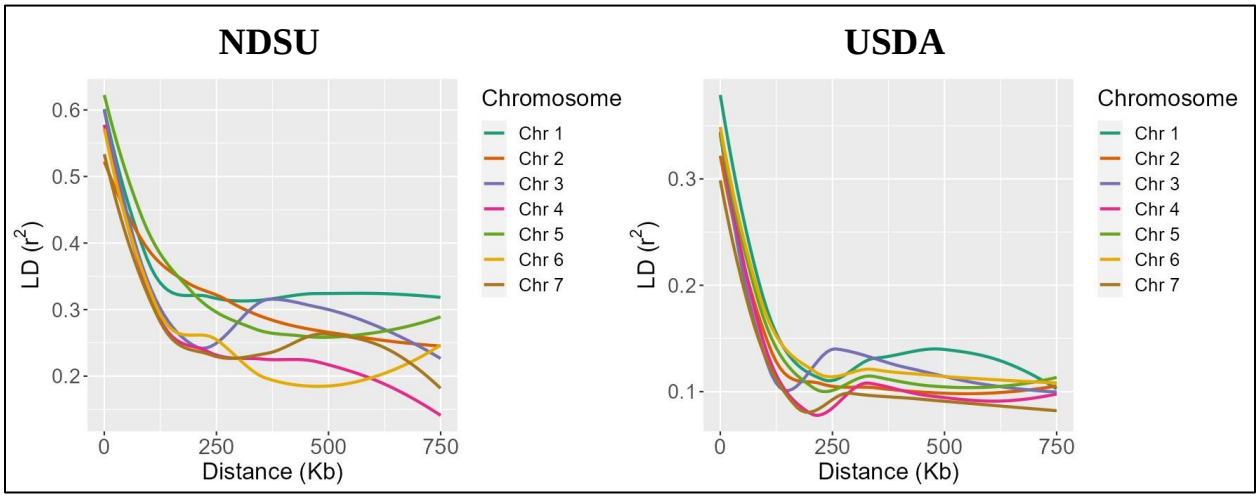

Figure 2.2. Chromosome-wide linkage disequilibrium - decay of NDSU set and USDA set

With respect to recombination rate (centimorgans - cM), the genome-wide r^2 on average decayed from 0.54 to 0.27 at 0.7 cM for the NDSU set, indicating a moderate level of correlation within this specific genetic distance across the genome. In contrast, the USDA set had a lower average r^2 (0.28) which dropped within a shorter genetic distance (0.5 cM). This implies that as the distance between the markers increases to 0.5 cM, they tend to be less correlated with each other (Appendix Figure A.2)

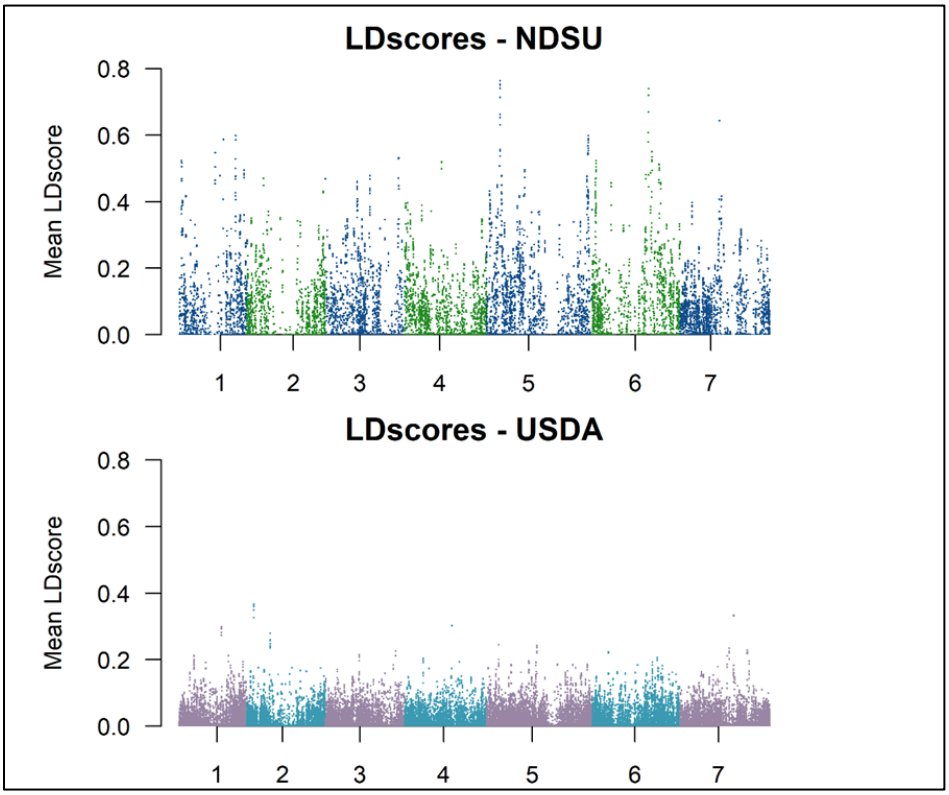

Figure 2.3. The mean LD scores estimated in 1000kb windows. There is a significant increase in LD of NDSU set compared to USDA set

The level of LD exhibited significant variation across distinct genomic regions and populations of dry peas. The impracticality of conducting whole-genome scanning can be attributed to the excessive number of markers required for such studies, particularly in cases where there is a low level of linkage disequilibrium (Kruglyak 1999). The USDA set reported a low LD value, indicating a higher occurrence of recombination events. In contrast, the NDSU set

showed a higher LD score, suggesting a greater frequency of linked markers presumably due to limited recent recombination to date (Siol et al., 2017).

Effective Population Size (N_e)

Based on LD, the estimated effective population size (N_e) for both the populations are shown in Figure 2.4. The smaller N_e and high LD in NDSU set indicates that it has undergone selective pressures leading to reduced diversity and increased correlation between the markers.

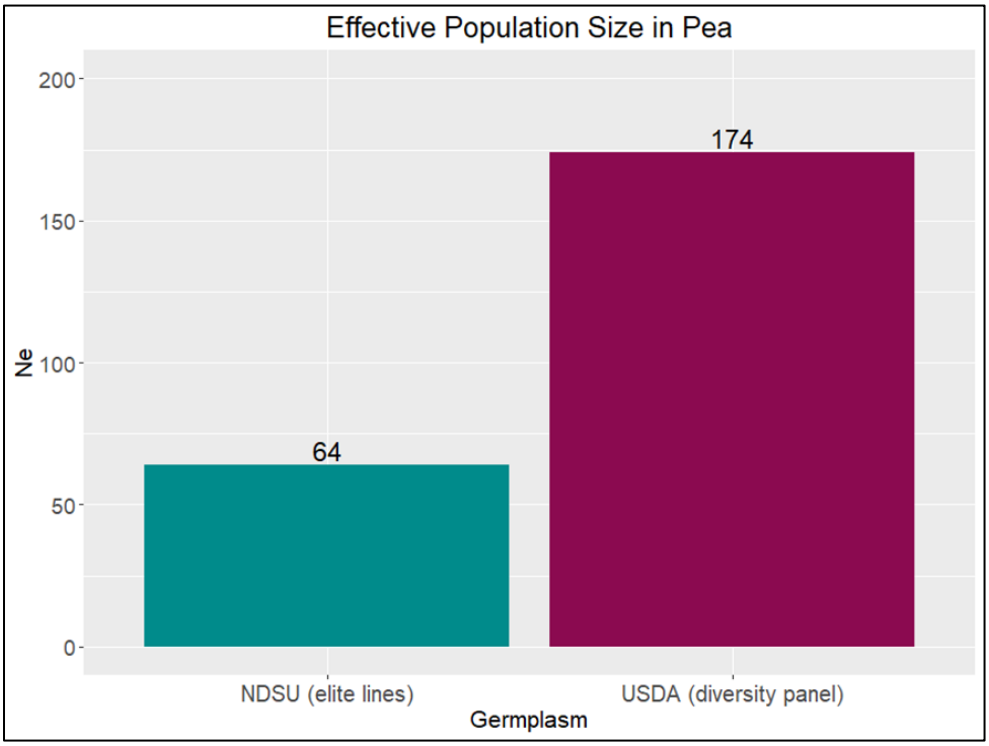

Figure 2.4. Estimated effective population size (N_e) for NDSU set is 64 and USDA set is 174.

Given NDSU set's population history and marker density, it is acceptable to state that despite lower N_e, it holds a reasonable level of diversity that may help maintain its genetic variability which is essential for long-term viability and adaptability. The USDA set resulted in lower LD and higher N_e, meaning it has more diversity and has encountered relatively fewer instances of selective pressures or genetic bottlenecks. It is important to note that the low LD can also be observed in a population with high N_e. Thus, it was expected to see NDSU set with lower N_e

compared to USDA set. These estimates explain how genetic drift and selections have shaped these populations over time.

The N_e estimate for the NDSU set was within the same range as those reported in other self-pollinating crops such as rice (*Oryza sativa*) and soybean (*Glycine max*), with calculated N_e ranging from 20 to 60. Juma et. al. (2021) estimated the N_e in rice to be 22 using an elite core panel comprised of 72 lines, but N_e may have been underestimated due to limited marker information used in the analysis. Similar studies in rice also had the same range of N_e, with calculated values ranging from 23-57 and 40-60; these were estimated based on breeding populations from recurrent selection programs (Grenier et al., 2015) and pedigree data (Morais Júnior et al., 2017). The estimated N_e of USDA set was within the range of N_e values reported in studies conducted on other crops. In soybean, Xavier et al., (2018) estimated N_e for the USDA soybean germplasm collection comprised of 19,652 accessions from Bandillo et al., (2015) and reported it to be 106 individuals. Recent studies have shown that soybean possess several genetic bottlenecks (Guo et al., 2010) and its genetic diversity has been reduced (Li et al., 2013, Min et al., 2010). The N_e estimate of USDA set is relatively higher than soybean, implying greater diversity. Zhao et al., (2013) estimated N_e in wild rice using 11 Chinese *Oryza rufipogon* populations including 32 landraces and reported it between 96-158, which is in a similar range to the USDA set. Thus, the N_e of USDA set offers greater potential for adaptation, maintaining rare alleles, population stability, and reduced risk for inbreeding.

The results of our study also suggest that the use of GBS holds good potential for making inferences of N_e regardless of the germplasm type. Using GBS-based markers, we approximated the LD pattern within and across chromosomes of both germplasms and then used the LD information for estimation of N_e. Genome-wide LD (r^2) of the USDA set decayed from

lower LD at 200 kb, while the NDSU set had the highest LD declined at a longer distance of around 250 kb. These results provided consistency of higher genetic variations of the former over the latter. Similar LD findings have been observed in previous studies conducted on peas, wherein both wild and spring peas exhibited a decay distance of approximately 200 kb, whereas wild/landrace peas were around 100 kb (Siol et al., 2017) which is a bit lower than the USDA set. Comparing the LD of USDA set and the NDSU set to other selfing crops such as rice, soybeans, and barley, the physical distances found were more or less similar depending on the populations. For instance, Huang et al. (2010) estimated LD using *O. indica* and *O. japonica* landraces of rice at 123 and 167 kb, respectively, with r^2 declining to 0.25 and 0.28. Additionally, soybean landraces extended from 90 to 500 kb (Hyten et al., 2007) while improved cultivars hit 133 kb (Zhou et al., 2015) which is similar to the USDA set. Alternatively, a recent LD analysis from soybean USDA germplasm revealed that the r^2 dropped intragenically within a few kilobases (Xavier et al., 2018) and the one in barley's landraces hit 90 kb (Caldwell et al., 2006), both shorter than the USDA set. The LD-decay of the NDSU set was also found to be in a similar range with elite varieties of barley which extended to at least 212 kb (Caldwell et al., 2006) and *O. japonica* elite lines at ~318 kb (Li et al., 2020), but had a higher distance compared to *O. indica* elite lines (~124 kb) (Li et al., 2020). The LD-decay rate of a crop does depend on the genetic background of the populations being studied, and it can be affected due to mutations, genetic drift, non-random mating, and a small N_e (Flint-Garcia et al., 2003).

Since public plant breeding programs are moving toward more quantitative methods, the importance of the dynamic exchange of genetic material and the maintenance of diversity within the population has increased. Effective population size helps breeders preserve and remodel their selection strategies to enhance the stability and variability in their breeding populations (Cobb et

al., 2019). Breeders can also implement marker-based mating experiments known as optimum contribution selection (OCS) (Juma et al., 2021) to maintain diversity in selection candidates for long-term gain. As pulse crop breeders navigate through challenges in their breeding programs, the information from this study provides valuable insights by demonstrating the strength of contemporary populations and possibly contributing to the long-term goal of increasing genetic gain while maintaining diversity in breeding programs.

Conclusion

These research findings shed light on the range of genetic diversity in both the NDSU set and the USDA set. The evaluation of N_e can be a bit more challenging and there is a possibility of potential biases if certain crucial factors including sample size, marker density, population history and LD are not accounted for appropriately (Waples and Yokota 2007; Waples and Do 2010; Gilbert and Whitlock 2015; Marandel et al., 2020). Even though genetic markers have become a more widely utilized approach for estimating N_e in recent years, there are still more obstacles to overcome in its N_e accuracy. Future estimation of N_e could be complemented with gene expression along with DNA markers and demographic history, which would increase the understanding of breeders regarding the population dynamics and potential for adaptation in different environments.

References

Abbo S, Gopher A, Lev-Yadun S (2017). The Domestication of Crop Plants. In: Thomas B, Murray BG, Murphy DJ (eds) Encyclopedia of Applied Plant Sciences (Second Edition), Academic Press: Oxford, pp 50–54.

Andrews S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. http://Www.bioinformatics.babraham.ac.uk/Projects/Fastqc/

Antao T, Pérez-Figueroa A, Luikart G (2011). Early detection of population declines: high power of genetic monitoring using effective population size estimators. *Evol Appl* **4**: 144–154.

Ardlie KG, Kruglyak L, Seielstad M (2002). Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3**: 299–309.

Atanda SA, Steffes J, Lan Y, Al Bari MA, Kim J-H, Morales M, et al. (2022). Multi-trait genomic prediction improves selection accuracy for enhancing seed mineral concentrations in pea. *Plant Genome* **15**: e20260.

Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J et al. (2015). A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection. *Plant Genome* **8**: eplantgenome2015.04.0024.

Bari MAA, Fonseka D, Stenger J, Zitnick-Anderson K, Atanda SA, Morales M et al. (2023). A greenhouse-based high-throughput phenotyping platform for identification and genetic dissection of resistance to Aphanomyces root rot in field pea. *Plant Phenome* **6**: e20063.

Bari MAA, Zheng P, Viera I, Worral H, Szwiec S, Ma Y et al. (2021). Harnessing Genetic Diversity in the USDA Pea Germplasm Collection Through Genomic Prediction. *Front Genet* **12**: 707754.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**: 2633 - 5.

Caldwell KS, Russell J, Langridge P, Powell W (2006). Extreme Population-Dependent Linkage Disequilibrium Detected in an Inbreeding Plant Species, Hordeum vulgare. *Genetics* **172**: 557–567.

Charlesworth B (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**: 195–205.

Cheng P, Holdsworth W, Ma Y, Coyne CJ, Mazourek M, Grusak MA et al. (2015). Association mapping of agronomic and quality traits in USDA pea single-plant collection. *Mol Breed* **35**: 75.

Cobb JN, Juma RU, Biswas PS, Arbelaez JD, Rutkoski J, Atlin G et al. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. *Theor Appl Genet* **132**: 627–645.

Crow JF, Kimura M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row: New York, USA.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**: e19379.

FAOSTAT (2021). Food and Agricultural Organization of the United Nations. Available at: https://www.fao.org/faostat/

Fernández J, Villanueva B, Pong-Wong R, Toro MA (2005). Efficiency of the use of pedigree and molecular marker information in conservation programs. *Genetics* **170**: 1313–1321.

Fisher RA (1930). *The genetical theory of natural selection*. Oxford University Press. Oxford

Flint-Garcia SA, Thornsberry JM, Buckler ES 4th (2003). Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* **54**: 357–374.

Gali KK, Sackville A, Tafesse EG, Lachagari VBR, McPhee K, Hybl M, et al. (2019). Genome-Wide Association Mapping for Agronomic and Seed Quality Traits of Field Pea (Pisum sativum L.). *Front Plant Sci* **10**: 1538.

García-Cortés LA, Austerlitz F, de Cara MAR (2019). An evaluation of the methods to estimate effective population size from measures of linkage disequilibrium. *J Evol Biol* **32**: 267–277.

Gargiulo R, Decroocq V, González-Martínez SC, Paz-Vinas I, Aury J-M, Kupin IL, et al. (2023). Estimation of contemporary effective population size in plant populations: limitations of genomic datasets. *bioRxiv*: 2023.07.18.549323.

Gilbert KJ, Whitlock MC (2015). Evaluating methods for estimating local effective population size with and without migration. *Evolution* **69**: 2154–2166.

Grenier C, Cao T-V, Ospina Y, Quintero C, Châtel MH, Tohme J, et al. (2015). Accuracy of Genomic Selection in a Rice Synthetic Population Developed for Recurrent Selection Breeding. *PLoS One* **10**: e0136594.

Guo J, Wang Y, Song C, Zhou J, Qiu L, Huang H, et al. (2010). A single origin and moderate bottleneck during domestication of soybean (Glycine max): implications from microsatellites and nucleotide sequences. *Ann Bot* **106**: 505–514.

Hare MP, Nunney L, Schwartz MK, Ruzzante DE, Burford M, Waples RS, et al. (2011). Understanding and estimating effective population size for practical application in marine species management. *Conserv Biol* **25**: 438–449.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* **13**: 635–643.

Hill WG (1981). Estimation of effective population size from data on linkage disequilibrium. *Genet Res* **38**: 209–216.

Hill WG, Robertson A (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet* **38**: 226–231.

Holdsworth WL, Gazave E, Cheng P, Myers JR, Gore MA, Coyne CJ, et al. (2017). A community resource for exploring and utilizing genetic diversity in the USDA pea single plant plus collection. *Hortic Res* **4**: 17017.

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* **42**: 961–967.

Hyten DL, Choi I-Y, Song Q, Shoemaker RC, Nelson RL, Costa JM, et al. (2007). Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* **175**: 1937–1944.

Juma RU, Bartholomé J, Thathapalli Prakash P, Hussain W, Platten JD, Lopena V, et al. (2021). Identification of an Elite Core Panel as a Key Breeding Resource to Accelerate the Rate of Genetic Improvement for Irrigated Rice. *Rice* **14**: 92.

Kreplak J, Madoui M-A, Cápal P, Novák P, Labadie K, Aubert G, et al. (2019). A reference genome for pea provides insight into legume genome evolution. *Nat Genet* **51**: 1411–1422.

Kruglyak L (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22**: 139–144.

Lee Y-S, Woo Lee J, Kim H (2014). Estimating effective population size of thoroughbred horses using linkage disequilibrium and theta (4Nμ) value. *Livest Sci* **168**: 32–37.

Li YH, Zhao S-C, Ma J-X, Li D, Yan L, Li J, et al. (2013). Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* **14**: 579.

Li X, Chen Z, Zhang G, Lu H, Qin P, Qi M, et al. (2020). Analysis of genetic architecture and favorable allele usage of agronomic traits in a large collection of Chinese rice accessions. *Sci China Life Sci* **63**: 1688–1702.

Lonsinger RC, Adams JR, Waits LP (2018). Evaluating effective population size and genetic diversity of a declining kit fox population using contemporary and historical specimens. *Ecol Evol* **8**: 12011–12021.

Luikart G, Ryman N, Tallmon DA, Schwartz MK, Allendorf FW (2010). Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conserv Genet* **11**: 355–373.

Marandel F, Charrier G, Lamy J-B, Le Cam S, Lorance P, Trenkel VM (2020). Estimating effective population size using RADseq: Effects of SNP selection and sample size. *Ecol Evol* **10**: 1929–1937.

Min W, Run-zhi L, Wan-ming Y, Wei-jun D (2010). Assessing the genetic diversity of cultivars and wild soybeans using SSR markers. *African Journal of Biotechnology* **9**: 4857–4866.

Morais Júnior OP, Breseghello F, Duarte JB, Morais OP, Rangel PHN, Coelho ASG (2017). Effectiveness of recurrent selection in irrigated rice breeding. *Crop Sci* **57**: 3043–3058.

Nei M, Tajima F (1981). Genetic drift and estimation of effective population size. *Genetics* **98**: 625–640.

Novo I, Santiago E, Caballero A (2022). The estimates of effective population size based on linkage disequilibrium are virtually unaffected by natural selection. *PLoS Genet* **18**: e1009764.

Pollak E (1983). A new method for estimating the effective population size from allele frequency changes. *Genetics* **104**: 531–548.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**: 559.

R Core Team. (2023). R: A Language and Environment for Statistical Computing. https://www.r-project.org/

Rahimmadar S, Ghaffari M, Mokhber M, Williams JL (2021). Linkage Disequilibrium and Effective Population Size of Buffalo Populations of Iran, Turkey, Pakistan, and Egypt Using a Medium Density SNP Array. *Front Genet* **12**: 608186.

Sawada C, Moreau C, Robinson GHJ, Steuernagel B, Wingen LU, Cheema J, et al. (2022). An Integrated Linkage Map of Three Recombinant Inbred Populations of Pea (Pisum sativum L.). *Genes* **13**: 196

Siol M, Jacquin F, Chabert-Martinello M, Smýkal P, Le Paslier M-C, Aubert G, et al. (2017). Patterns of Genetic Structure and Linkage Disequilibrium in a Large Collection of Pea Germplasm. *G3* **7**: 2461–2471.

Sved JA (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol* **2**: 125–141.

Sved JA, Hill WG (2018). One Hundred Years of Linkage Disequilibrium. *Genetics* **209**: 629–636.

Tayeh N, Klein A, Le Paslier M-C, Jacquin F, Houtin H, Rond C, et al. (2015). Genomic Prediction in Pea: Effect of Marker Density and Training Population Size and Composition on Prediction Accuracy. *Front Plant Sci* **6**: 941.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, et al. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Res* **17**: 520–526.

Toosi A, Fernando RL, Dekkers JCM (2010). Genomic selection in admixed and crossbred populations. *J Anim Sci* **88**: 32–46.

USDA (2020). United States Acreage. *National Agricultural Statistics Service*. https://www.nass.usda.gov/Publications/Todays_Reports/reports/acrg0620.pdf. Accessed 15 August 2023

Wang J (2005). Estimation of effective population sizes from data on genetic markers. *Philos Trans R Soc Lond B Biol Sci* **360**: 1395–1409.

Wang J, Santiago E, Caballero A (2016). Prediction and estimation of effective population size. *Heredity* **117**: 193–206.

Waples RS, Do C (2010). Linkage disequilibrium estimates of contemporary *Ne* using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evol Appl* **3**: 244–262.

Waples RS, Yokota M (2007). Temporal estimates of effective population size in species with overlapping generations. *Genetics* **175**: 219–233.

Wright S (1931). Evolution in Mendelian Populations. *Genetics* **16**: 97–159.

Xavier A, Thapa R, Muir WM, Rainey KM (2018). Population and quantitative genomic properties of the USDA soybean germplasm collection. *Plant Genet Resour* **16**: 513–523.

Yang J, Lee SH, Goddard ME, Visscher PM (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**: 76–82.

Zhao Y, Vrieling K, Liao H, Xiao M, Zhu Y, Rong J, et al. (2013). Are habitat fragmentation, local adaptation and isolation-by-distance driving population divergence in wild rice *Oryza rufipogon*? *Mol Ecol* **22**: 5531–5547.

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol* **33**: 408–414.

CHAPTER 3. REGULATORY LANDSCAPE OF DEVELOPING PODS IN DRY PEA

Introduction

The aim of plant breeders is to better understand and improve the phenotypes in their species. To achieve this, researchers prioritize detecting genes underlying the phenotypic variations and eventually leading to crop improvement (Li et al., 2021). Recent advancements in high-throughput sequencing made it easier for us to access the molecular markers across the genome, which is a key dataset required to map the genes. The most popular method used over the past two decades to genetically map and identify trait-associated genes is genome-wide association studies (GWAS). GWAS identifies the association between genetic variants and phenotypic variation (red arrow, Figure 3.1) by utilizing historical recombination events (Hirschhorn and Daly 2005) and genetic diversity (Korte and Farlow 2013; Gali et al., 2019). GWAS has a higher resolution compared to traditional linkage mapping which uses a bi-parental population to detect the causal trait-associated marker (Korte and Farlow, 2013; Gali et al., 2019). Due to the cost-effective genotyping technologies and statistical methods, GWAS is widely used to understand the genetic complexity in most species. In most GWAS studies, researchers prefer to use single nucleotide polymorphisms (SNPs), as they are easy to genotype and abundant in the respective genome. SNPs can distinguish between closely related individuals and have been used in many genetic mapping studies in pea (Korte and Farlow, 2013; Gali et al., 2019; Sindhu et al., 2014; Tayeh et al., 2015), soybean (Priyanatha et al., 2022), maize (Zeng et al., 2022), and rice (Ashfaq et al., 2023); allowing the plant research community to detect a plethora of causal markers associated with hundreds of phenotypes (Tian et al., 2020).

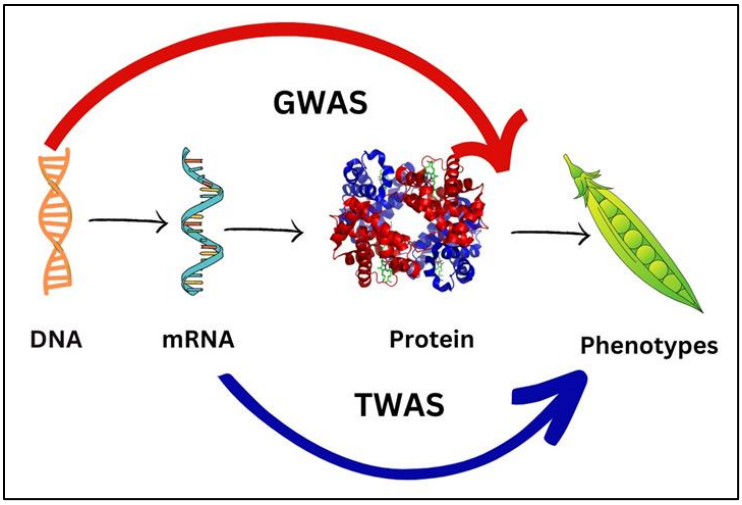

Figure 3.1. The ultimate central dogma of life which links DNA markers to the phenotypes for detecting the trait-associated genes through GWAS, whereas mRNA is linked directly to phenotype through methods like TWAS for identifying the genes.

The major caveat of GWAS is that, due to long-range linkage disequilibrium (LD), it is often not possible to verify which of those multiple genes linked to the genetic marker is, in fact, the causal gene, particularly in self-pollinated crops (Wallace et al., 2014; Li et al., 2021). A transcriptome-wide association study (TWAS) examines the expression-phenotype associations (blue arrow, Fig. 3.1), which are independently affected by LD, unlike in the case of genetic markers. In simpler terms, even if multiple genes are closely connected and cannot be observed separately in different individuals, they can still be given priority for association with a trait because their patterns of expression are independent. This is particularly useful in species where there is a significant amount of linkage disequilibrium or in cases where it is not possible to create high-resolution mapping populations (Kremling et al., 2019). TWAS has been proven to be an assuring development for gene mapping studies in plants and is also equivalent to GWAS. This gene-level approach has been compared against GWAS in a qualitative trait of the soybean genome, which has slow rates of LD-decay, whereas the authors were able to prove that TWAS

can provide single-gene resolution for candidate genes and overcome the limitations of LD (Li et al., 2021).

There are plenty of approaches to conduct GWAS analysis but for TWAS, the choices were limited. A TWAS-based Bayesian statistical method called expression read depth GWAS (eRD-GWAS) was developed by Lin et al., (2017) in maize. The authors found 13 trait-associated genes and observed that the functions of those genes aligned with the characteristics of the traits and furthermore, one of those genes underwent functional characterization (Lin et al., 2017). In contrast to certain human TWAS methods, which predicted gene expression levels, eRD-GWAS utilized explanatory variables that explicitly measured gene expression levels (Li et al., 2021). Kremling et al., (2019) used a different approach employing multiple regression-based TWAS with seven tissues and overlapping results from GWAS and TWAS to find the causal genes. Both these Bayesian-based and regression-based TWAS studies resulted in hundreds of genes but have not yet been evaluated quantitatively after publication. In Li et al., (2021), the authors performed TWAS using a simple approach by transforming the normalized expression matrix to a numerical range (0 to 2) for use in GAPIT (R package). Different TWAS approaches that were published gave differing results. For example, results from Hirsch et al., (2014) and Lin et al., (2017) only had one or few genes that passed the false discovery rate threshold (FDR), while in Kremling et al., (2019) and Wu et al., (2022) they had used the top 0.5-1% hits based on p-values and ended up getting more significant genes which also included false positive discoveries as well. Reasons behind these varying results are due to the dataset format used in their respective analyses. Li et al., (2021) converted the expression data to a categorical format (0,1,2) to utilize in the GAPIT R package and compared it with the GWAS results using the same tool, whereas Kremling et al., (2019) directly used the normalized gene expression data in

the mixed linear model. Since the datasets utilized in GWAS and TWAS analysis were different, the authors used the outlier approach to study the genes in order to avoid direct comparison of p-values resulting from different datasets rather than using the multiple corrections approach (Kremling et al., 2019).

TIn this study, a gene expression resource was created specifically for pea pod development in two environments (North Dakota (ND) and Washington (WA)) for evaluating their expression levels and the complex interactions between the genotype and environment using TWAS. In addition, we have also identified SNP-based molecular markers to perform GWAS and conducted a comparative study with TWAS analysis. The efficacy of the TWAS approach was evaluated in this self-pollinated crop for quantitative traits such as protein content and yield. Given the recent rise in popularity of peas as an alternative protein source, it is crucial to prioritize the identification of causal genes correlated to protein content as well as yield. Our objectives were threefold: (i) identify genetic markers and develop a gene expression resource targeting pod development, (ii) quantify regulatory variations in gene expression profiles and assess the abundance of identified genetic markers and (iii) map trait-associated genes through genome-wide and transcriptome-wide association studies. Our results established that even in quantitative traits of a self-pollinating crop, TWAS can be an additional resource to validate GWAS results and is less affected by LD-decay rate.

Materials and Methods

Plant Materials

This study utilized 300 diverse Pisum sativum accessions which represent most of the available diversity in the USDA (United States Department of Agriculture) pea germplasm collection. The population size was determined using the available germplasm, statistical power

for making inferences, and the phenotyping capability of researchers involved in North Dakota (ND) and Washington (WA) environment testing sites (2 locations * 2 years). Through the utilization of effective population size (Ne) estimation (see Chapter 2), it has been demonstrated that USDA accessions exhibit substantial genetic variations, which are essential for mapping studies. This germplasm offered ample genetic diversity and a significant number of historical recombination events, leading to successful downstream analysis. The total 300 accessions consisted of 72 lines of early-maturity (< 55 days to 50% flowering), 120 lines of mid-maturity (56-63 days to 50% flowering) and 108 late-maturity (> 64 days to 50% flowering).

DNA Extraction, Whole Genome Sequencing, SNP Calling

After two rounds of single seed descent performed in the greenhouse, young leaves were collected from each accession to extract DNA using the DNeasy Plant Mini Kit (Qiagen). The extracted DNA was sent to HudsonAlpha following their guidelines for whole-genome resequencing (10x depth) using Illumina sequencing technology. A total of 5.9 terabytes of WGS raw data was produced, consisting of 103 billion paired-end reads. The quality of these reads was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed by Trimmomatic (version 0.36; Bolger et al., 2014). The trimmed reads were aligned to the Pisum sativum Chinese reference genome (Yang et al., 2022) by bwa-mem2 (Vasimuddin et al., 2019). PCR duplicates were marked by Picard's function 'MarkDuplicates' (http://broadinstitute.github.io/picard/) and removed using Samtools-1.10 (Li et al., 2009). The variant (SNP) file was called using Bcftools and filtered by vcftools (Danecek et al., 2011). The filtering parameters were min-meanDP 5, min-alleles 2 --max-alleles 2 (for bi-allelic), max-missing 0.05 and maf 0.05. The initial number of SNPs retrieved was 6,720,968. We further narrowed down the SNPs by removing any with more than 20% missingness using Plink v1.9

(Purcell et al., 2007), heterozygosity > 15% and imputed the missing SNPs using k-nearest neighbor genotype imputation method (Money et al., 2017) in Tassel v5.0 (Bradbury et al., 2007). The final number of SNPs used for downstream analysis was 137,725.

Field Experimental Design for RNA Expression Analysis

All accessions were planted following diagonal check augmented incomplete block design with two replications. Pseudo blocks were assigned to approximate maturity groups and randomization of accessions was made within each maturity groups such as early, mid and late. We also included four check varieties per location: ND used AC Agassiz, Arcadia, DS Admiral, and Hampton while WA used ND Dawn, Columbian, DS Admiral, and Hampton. The expression analysis was conducted at the NDSU Prosper, ND, site and at Pullman, WA, during the 2022 growing season.

Determination of Optimal Tissue Sampling Stage

To capture the highest amount of differentially expressed genes (DEG) within the developing pods across all three maturity groupings represented within the USDA population, a pilot field experiment was conducted to optimize the tissue sampling timepoint. The timepoints were defined by utilizing the reproductive growth stage 3 (R3) corresponding to the early presence of a flat pod at one or more nodes immediately after flowering (timepoint T_0), with each successive timepoint having an additional 6 days added, establishing the specific timeframe for tagging and collecting samples (Table 3.1).

RNA was isolated from a single seed within the pod across all timepoints with expression analysis conducted to determine the timepoint having the highest number of expressed genes. Figure 3.3 displays the number of genes expressed at each timepoint and across the maturity groupings, unambiguously demonstrating that T_1 exhibits the greatest number of genes expressed

Table 3.1. Detailed description of the sampling timepoints.

Timepoints	Description
T_0	Equivalent to R3 growth stage (at first sight of pod after flower)
T_1	T_0 + 6 days
T_2	T_0 + 12 days
T_3	T_0 + 18 days

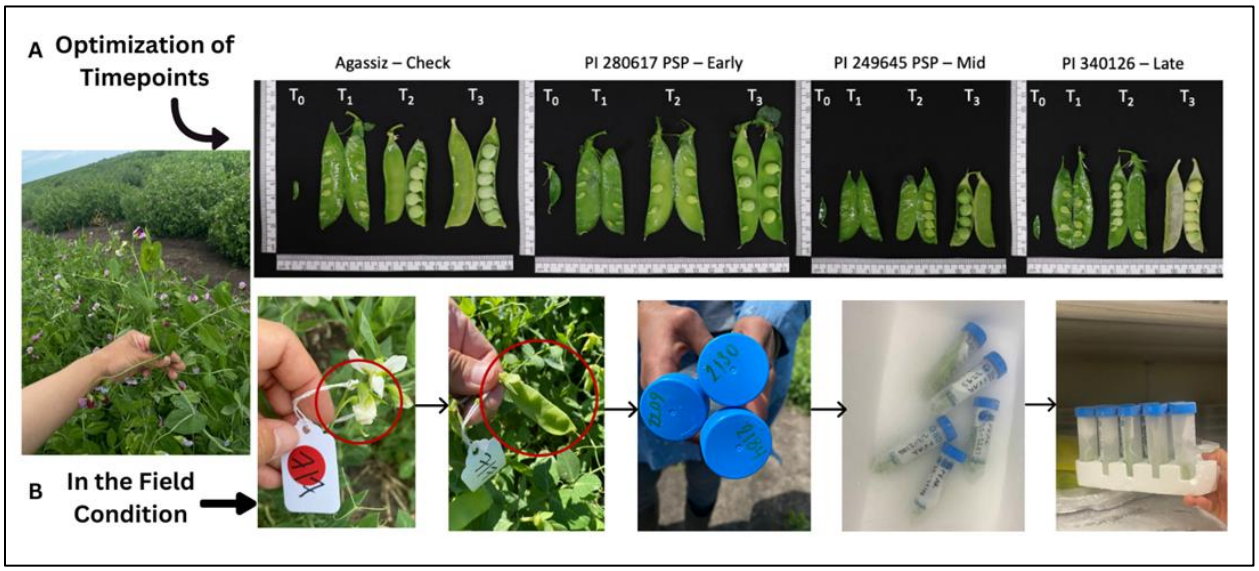

Figure 3.2. A) Optimization of the sampling timepoints displaying the growth stages of the developing pods and seeds for representative lines of each maturity group (the check variety along with early, mid, and late maturity) and B) Shows tagging, collection of desired pods from T_1 and storing the sample tubes in the -80° freezer prior to RNA extraction.

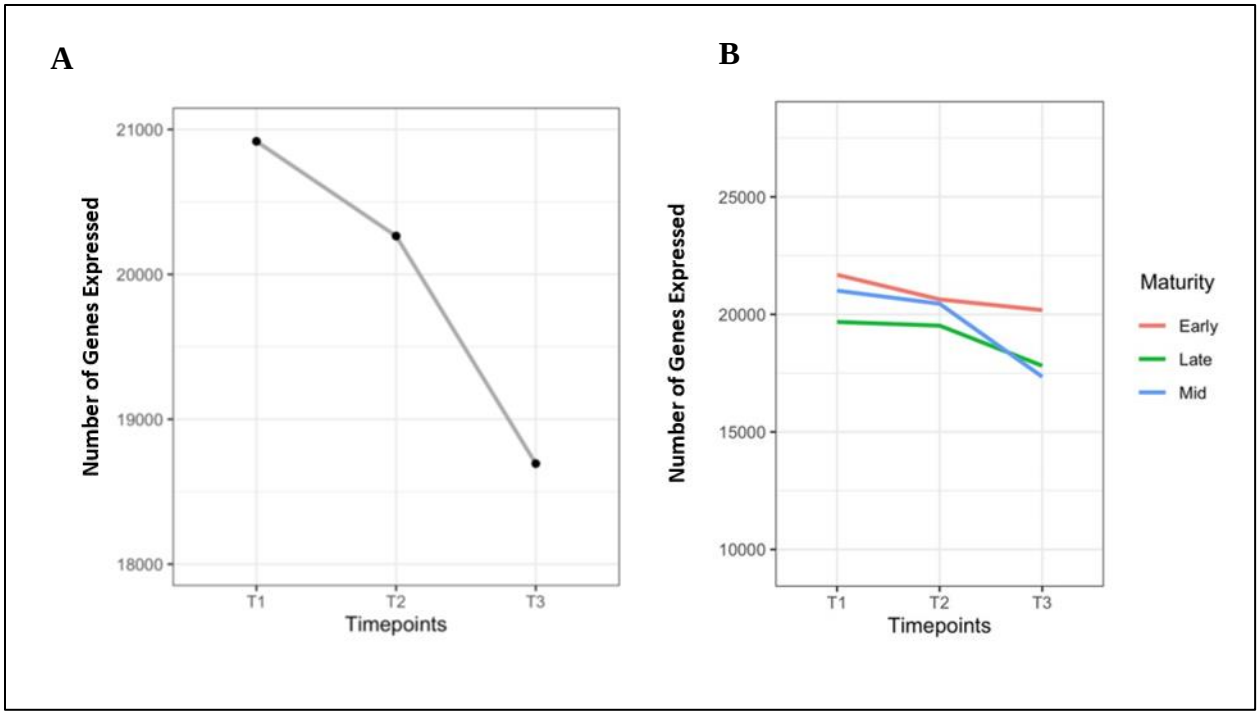

Figure 3.3. Graphs depicting the number of genes expressed at each timepoint (A) and across the maturity lines (B). T_1 clearly displays the optimal timepoint having the highest number of gene expressed.

across all maturity groups. With that evidence, we used T_1 as the optimal time point for tissue sampling.

During the 2022 field experiment that contained all 300 accessions, pods from three plants within the plot were tagged at the T_0 stage. Tissue samples were then collected six days later (1 pod from 3 plants) representing the T_1 stage, placed in a 50ml tube and immediately stored on dry ice until they were transported back to the lab where they could be kept in a -80°C freezer until RNA could be extracted.

RNA Extraction, 3' RNA-Seq Library Preparation and Sequencing and Quantitative Expression Analysis

The developing pods at timepoint T_1 were sampled at each environment: WA = replication 1 and ND = replication 2. RNA was isolated from a single seed within each pods at T_1 using the *Quick*-RNA^{TM} Plant Miniprep (ZYMO Research, Orange, CA, USA) according to

the manufacturer's protocol including proteinase K treatment, with elution performed using 100µl. RNA sample concentrations were quantified and quality assessed using the Qubit RNA BR and RNA IQ Assay kits with the Qubit 4.0 fluorometer (Life Technologies Corporation, Eugene, OR, USA), and standardized to a final concentration of 100 ng/µl. 3' RNA-Seq libraries were prepared from the total RNA per sample using the Lexogen QuantSeq 3' mRNA-Seq Library Prep Kit FWD for Illumina (https://www.lexogen.com/quantseq-3mrna-sequencing/). The libraries were quantified on a Molecular Devices Spectra Max M2 plate reader and pooled them to achieve optimal uniformity. Later, the pool was quantified by digital PCR and sequenced on 1 lane of Illumina NextSeq500 sequencer, single-end 1x86bp, and then six base i7 indices were utilized to de-multiplex the samples using Illumina bcl2fastq software (version 2.20; Illumina, Inc., San Diego, CA) which was performed at the Cornell Institute of Biotechnology. In quality control stage, the first 12 bases and adapters were removed from the sequences for samples with minimum 200K reads using Trimmomatic (version 0.36; Bolger et al., 2014). Poly-A tails and poly-G stretches of at least 10 bases in length were also removed using the BBDuk program from the package BBMap (https://sourceforge.net/projects/bbmap/; version 37.50), keeping reads at least 18 bases in length after trimming. Poly-G stretches result from sequencing past the ends of short fragments (G = no signal). The trimmed reads were aligned to the Pea ZW6 genome assembly (Yang et al., 2022a, b) using the STAR aligner (version 2.7.10b; Dobin et al., 2012). For the STAR indexing step, the gff3 annotation file (https://www.peagdb.com/static/data/download/genes.gff3.zip) was converted to gtf format with the gffread program (version 0.10.4) from cufflinks (Trapnell et al., 2012). Key parameters used in the STAR indexing step (--runMode genomeGenerate) were: --genomeChrBinNbits 18 and --sjdbOverhang 73. The output SAM files were converted to BAM using SAMtools (version

1.15.1; Li et al., 2009), and the number of reads overlapping each gene in the gff3 file on the forward strand were counted using HTSeq-count (version 0.6.1; Anders et al., 2014).

The R package DESeq2 (version 1.36.0; Love et al., 2014) was used to obtain both normalized and variance stabilized counts, to conduct a principal components analysis of the 500 most variably expressed genes after count normalization and variance stabilizing transformation. For the variance stabilized counts, genes with fewer than two counts per sample on average were excluded. Tests to detect genes differentially expressed between the WA versus ND samples were performed both with and without Bayesian shrinkage of the log fold change (LFC) estimates via the "apeglm" method (Zhu et al., 2018).

Phenotyping

The 300 USDA accessions were cultivated and had complete agronomic datasets gathered over a span of two years at the NDSU Minot (2021)/Prosper (2022) sites and at the WSU Pullman site for both 2021 and 2022. The designated planting dates for the Minot and Prosper site were May 1, 20221 and May 28, 2022, while the Pullman site was on May 4th and May 11th for each respective year. The following traits included in the dataset are: protein concentration from the harvested seeds which was analyzed using near infrared (NIR) spectrometry and the seed yield (lb/A) calculated based on the following formula,

$$\text{Yield} = \frac{\text{Harvest wt (lbs)}}{\text{Plot size (acres)}} \text{ x } \frac{0.865}{1-\frac{\text{moisture (\%)}}{100}} \text{ x } \frac{1}{60} \quad (2.1)$$

Phenotypic Data Analysis

The best linear unbiased prediction (BLUPs) was used to perform a mixed linear model to analyze the unbalanced dataset and retrieve the genetic merit of each genotype using ASReml-R package (Version 4.1; Butler et al., 2022). We conducted the phenotypic analysis for two

environments x years such as 2021_North Dakota, 2021_Washington, 2022_North Dakota and 2022_Washington. Each environment was assessed individually to determine the heritability in order to measure the precision of single field trials. Since our dataset is unbalanced with spatial trends and complex experimental factors, we calculated heritability using the method proposed by Cullis *et al.,* (2006) based on BLUP (Eq. 2.2) rather than using the basic formula to calculate the entry mean based heritability (Hussain et al., 2022; https://github.com/whussain2/Analysis-pipeline).

$$H_C = 1 - \frac{\bar{v}_{BLUP}}{2\sigma_g^2} \quad (2.2)$$

Where $\bar{v}_{BLUP}$ is a mean variance difference between two genotypes based on BLUPs and the σ_g^2 is the genotype variance. In order to calculate BLUPs for multi-environment (MEB) to conduct GWAS, we used stage-wise analysis based on the pipeline generated by Hussain et al., (2022) comprised of two stages. In the first stage, we calculated adjusted means as Best Linear Unbiased Estimators (BLUEs) and residuals for each environment by treating genotypes as a fixed effect (Eq. 2.3). We adjusted the means of genotypes for their blocks and replications.

$$Y_{nmi} = \mu + G_n + R_m + B_{nmi} + \epsilon_{nmi} \quad (2.3)$$

Y_{nmi} = is the effect of n^{th} genotype in m^{th} replications and i^{th} block within m replication

μ = overall mean

G_n = random effect of the n^{th} genotype

R_m = fixed effect of the m replication

B_{nmi}= random effect of i^{th} block nested with m replication

ϵ_{nmi} = residual error

In the second stage, a mixed linear model was fitted across all environments using the BLUEs obtained from the previous step (Eq. 2.4). Here, the assumption of this model is the error

was obtained from the first stage. The weights fitted in this model were estimated by the reciprocal of the squared standard error of BLUEs. These weighted BLUEs were utilized to address the varying error variance.

$$Y_{nm} = \mu + G_n + E_m + (G \times E)_{nm} + \epsilon_{nm} \quad (2.4)$$

Y_{nm} = n is the BLUE observation in m environment

μ = overall mean

G_n = random effect of the nth genotype

E_m = random effect of the m environment

$(G \times E)_{nm}$ = genotype by environment interaction term

ϵ_{nm} = residual error

Finally, the BLUPs were extracted from this model for GWAS (MEB). For TWAS within-environment (WEB), we used a similar method to extract BLUPs by using data from North Dakota (2021 & 2022) and Washington (2021 & 2022) separately in the analysis and combined the results to perform the multi-environment TWAS (Hussain et al., 2022).

Genome-Wide and Transcriptome-Wide Association Studies (GWAS & TWAS)

The GWAS and TWAS was conducted using the linear mixed model (LMM) in R (R Core Team, 2023). Covariates and kinship were calculated using SNP and gene expression datasets to address population structure and relatedness among the genotypes. The principle components (PCs) and kinship (Ks) were derived using prcomp() and VanRaden() (Van Raden, 2008) functions in R. The number of PCs and Ks required in each of these models was detected by Bayesian Information Criterion (BIC) (Schwarz 1978). SNPs or genes from the top 1% hits based on the p-value were counted and considered as the gene-of-interest. We used the LD-

window size ~250kb that was estimated in Chapter 2 to identify the genes in the GWAS study from the top 1% SNPs.

$$\text{Phenotypes} = \text{SNPs or Gene expression} + \text{PCs} + \text{Ks} + \text{Error} \quad (2.5)$$

Statistical Analyses

The difference between the average expression levels of the KIW84_010029 and KIW84_065350 genes in the two protein groups (high vs low) was analyzed using the Welch two-sample t-test (Welch 1947).

Results and Discussion

The gene expression of developing pods in peas was evaluated based on the optimized timepoint of T_1 (see Materials and Methods). Using the ND and WA field trials, the differences in expression patterns in pods between these environments and the genes associated with the quantitative traits such as seed protein and yield were evaluated. In this study, we performed single-tissue within and multi-environment TWAS and a comparative analysis with GWAS.

Phenotypic Analysis

The quantitative variations of seed protein and yield were assessed in each environment to evaluate the effectiveness and differences of the trials. In the 2021 trials, the heritability of ND's protein and yield were lower than that of WA. However, in the 2022 trials, ND's heritability increased and surpassed that of WA. This difference suggests that the genetic factors improved its contribution to the seed protein and yield between the years 2021 and 2022 in ND (Figure 3.4). This could be due to the changing weather conditions, genetic adaptations or changes in the trial practices. Multi-environment and within-environment BLUPs (MEB & WEB) (Materials and Methods) were calculated to perform the regression-based association

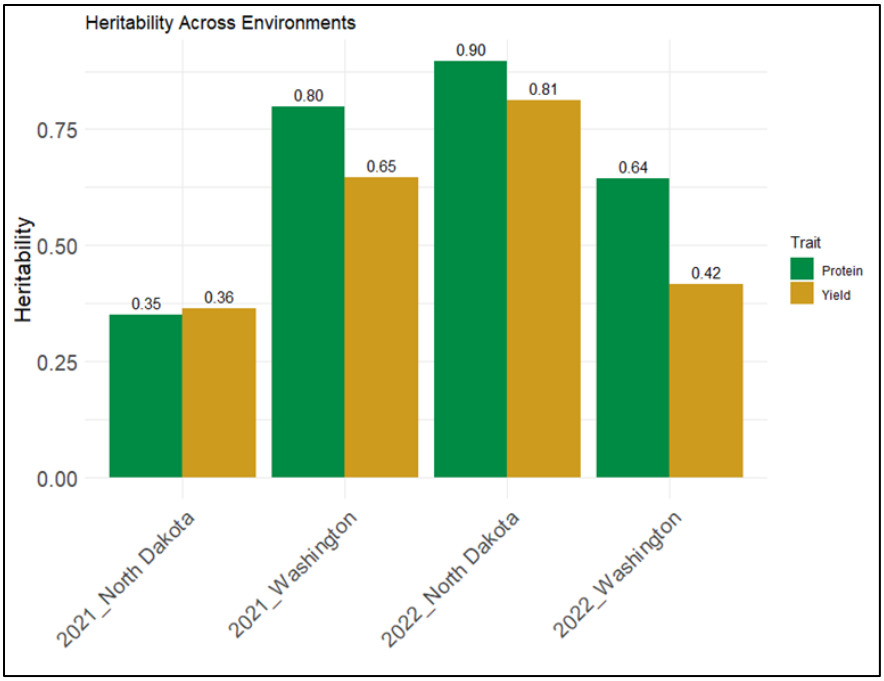

Figure 3.4. Heritability estimates of two environments x two years for protein and yield traits. The 2022_North Dakota environment yielded much higher as compared to the 2021_North Dakota environment.

studies. The protein range in MEB was around 27.38 to 32.44 whereas for WEB-ND the range was 27.48 to 36.07 (Figure 3.6A) and for WEB-WA it was 25.44 to 31.95 (Figure 3.7A).

Expression Analysis

Including the check varieties, 700 tissues were collected from both environments but only 620 samples surpassed the RNA quality for sequencing. After further quality control, we retrieved 505 samples and 15,358 genes from both environments. ND had 247, and WA had 258 samples with same set of 15,358 genes. The principal component analysis (PCA) of the top 500 genes and the sample distribution revealed similarities due to the common genetic backgrounds and some differences among the genotypes and locations explaining the complex interaction of genotype by environment (GXE) (Figure 3.5A and B). We used shrinkage of effect sizes (log fold change) to visualize the differences in gene expression between ND and WA. There are 18% upregulated (1,337), 32% downregulated (3,871) and 50% non-significant (10,190) genes found

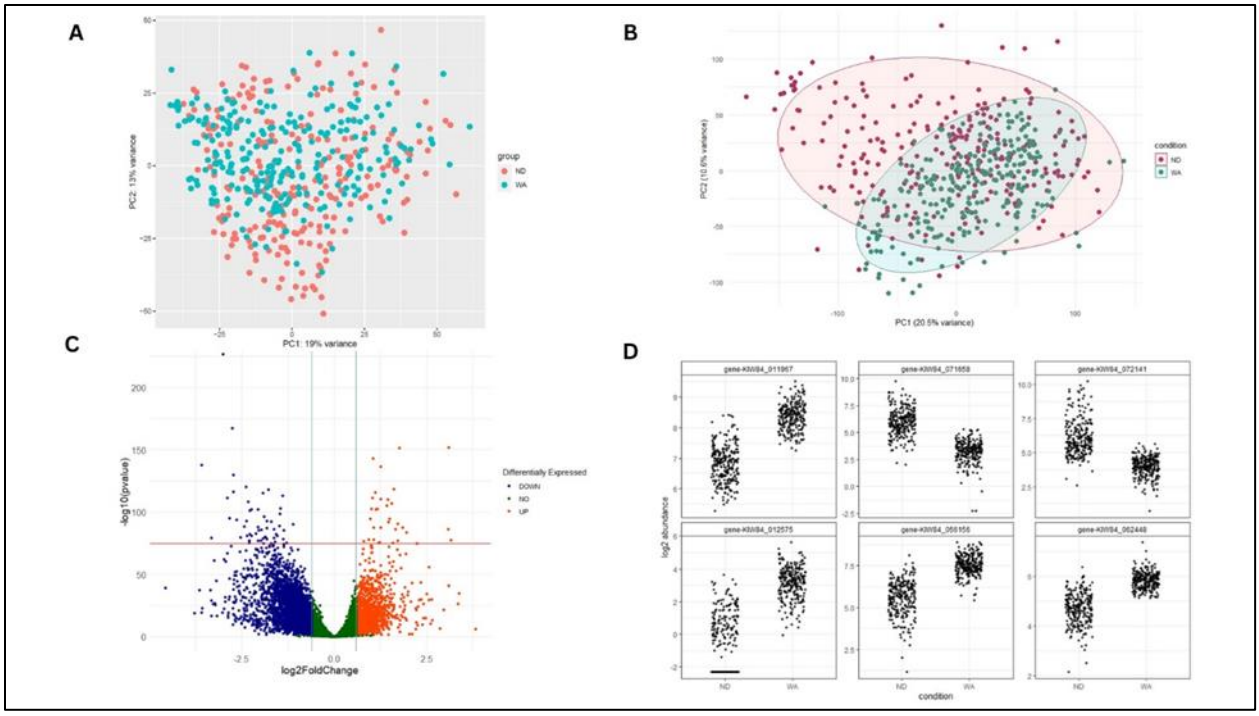

Figure 3.5. A) PCA of the top 500 highly expressed genes revealed similarity of genes being expressed between ND and WA, B) PCA shows the relationship between the samples across two environments where they formed two groups of clusters with overlapping expression patterns, C) Volcano plot representing differential gene expression between the two environmental conditions (ND & WA), and D) the top six genes expressed amongst both environments.

in this condition (Figure 3.4C). We also found one of the most significant upregulated (gene-KIW84_071658) and downregulated (gene-KIW84_011967) genes with -log10(p-value) >200 (Figure 3.5C). Such extreme p-values supporting the differential expression of these genes indicates a substantial change in expression levels compared to other genes.

We also took a detour to delve into the differential gene expression between high and low protein conditions in ND and WA. Based on the histogram of protein phenotype distribution of ND and WA, we extracted the 10% genotypes of lower and upper intervals (Figure 3.6A and 3.7A). Overall, we had 48 genotypes for ND (24 lower and 24 upper - Figure 3.6A) and 46 genotypes for WA (23 lower and 23 upper – Figure 3.7A). The DESEQ2 was analyzed for these samples with low vs high protein conditions separately for each environment. The number of

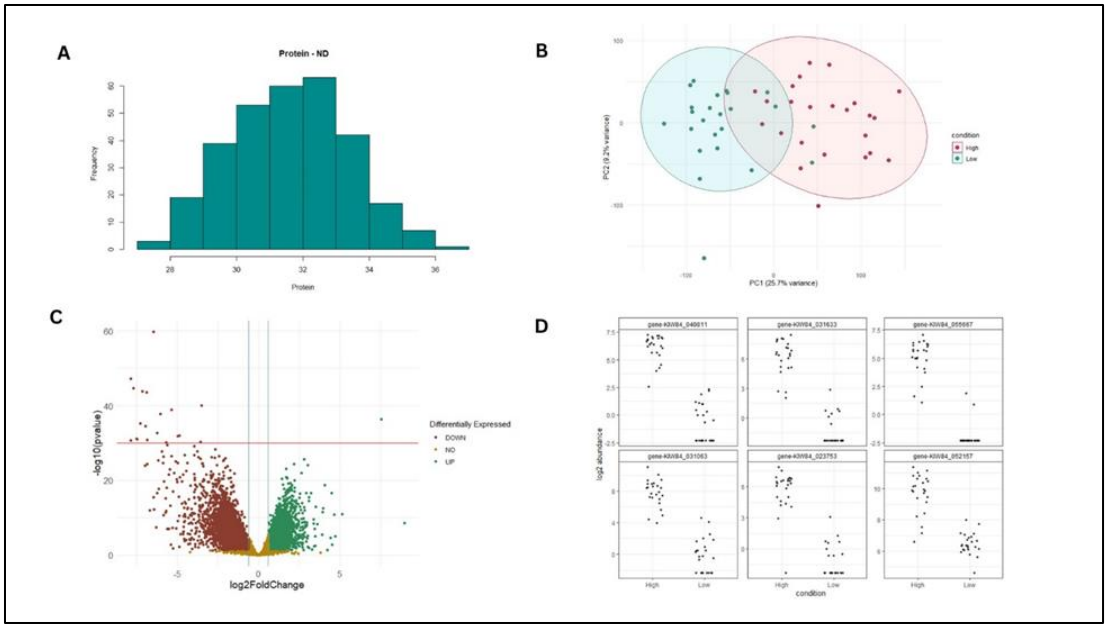

Figure 3.6. Analysis of the ND environment dataset showing A) the distribution of protein, B) PCA showing the relationship between the samples across two protein groups (high 10% & low 10%), C) Volcano plot representing differential gene expression between two protein conditions, and D) the top six genes expressed in two protein groups.

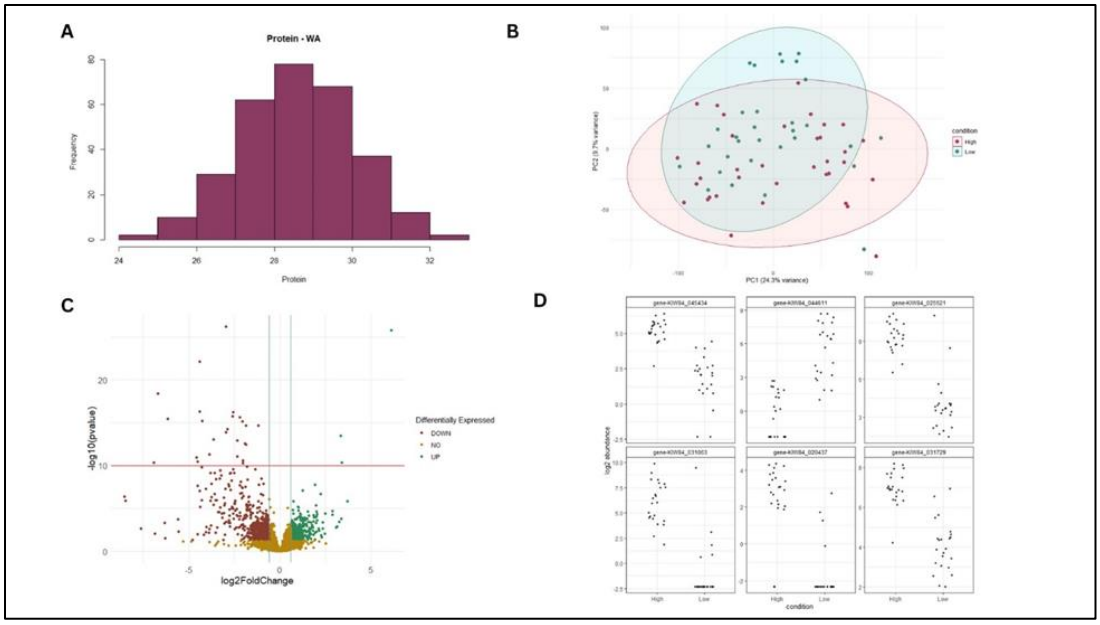

Figure 3.7. Analysis of the WA environment dataset showing A) distribution of protein, B) PCA showing the relationship between the samples across two protein groups (high 10% & low 10%), C) Volcano plot representing differential gene expression between two protein conditions, and D) the top six genes expressed in two proteins.

genes used for further analysis after quality controls was 15,083 (ND) and 14,832 (WA). A portion of differentially expressed genes was retrieved in both environments for high vs low protein, as shown in Figure 3.6C and Figure 3.7C. In ND, we detected 4,338 upregulated and 5,040 downregulated genes, whereas in WA, we only had 132 upregulated and 333 downregulated genes.

We found a common and highly significant upregulated gene tag "KIW84_031063" between ND and WA protein conditions. This gene encodes the WW domain-containing protein (Sudheesh et al., 2015). As one of the upregulated gene, it was expected to increase in expression levels as the protein content increases. Consequently, the correlation with the ND protein (R = 0.51 - Appendix Figure B.1(A)) was higher than with the WA protein (R = 0.29 –Appendix Figure B.1(B)); both are positively correlated. We also performed combined correlation analysis with ND and WA protein to understand whether this could increase the correlation between the gene and the protein, unfortunately it did not increase the correlation (R = 0.46 – Appendix Figure B.1(C)). This WW domain-containing protein may play a role in the biological pathways responsible during changes in the environment.

TWAS and GWAS

Transcriptome-wide association studies were performed using the filtered gene expression data (15,398 genes) for within and multi-environments (ND ~275 lines and WA ~248 lines) and the Genome-wide association studies were conducted using a filtered SNP dataset (~137,725). Using the four mixed linear models, we detected the top 1% genes based on their p-values (Kremling et al., 2019). Since we are using continuous TWAS and discrete GWAS datasets, this method of detecting the genes will avoid direct comparisons of p-values. We identified a unique and common set of genes in all of these models for protein and yield traits.

Considering protein trait, we found four similar genes such as KIW84_060149, KIW84_046360, KIW84_051242, and KIW84_063439 between ND-TWAS and WA-TWAS. One of those genes KIW84_046360 located in chromosome 4 was also present in the ME-TWAS model. KIW84_046360 was the only gene similar between WA-TWAS and ME-TWAS while seven additional genes were found to be similar between ND-TWAS and ME-TWAS. When we compared GWAS with the ND-TWAS, we detected another 10 significant genes. Additionally, 8 genes were similar between WA-TWAS and GWAS. One of the genes KIW84_063439 in chromosome 6 was found in all GWAS, ND-TWAS, and WA-TWAS models. Moreover, another gene KIW84_010029 in chromosome 1 was detected similarly between GWAS, ND-TWAS, and ME-TWAS. Apart from that, 10 more genes were similar between GWAS and ME-TWAS. For yield trait, we found two genes KIW84_032148, and KIW84_021648 similar between ND-TWAS and WA-TWAS. With ME-TWAS, we found 8 genes similar to ND-TWAS and 6 genes with WA-TWAS. When compared with GWAS, we detected 10 genes similar to ND-TWAS and 14 genes with WA-TWAS. Finally, when comparing the GWAS with the multi-environment model, we also identified 14 similar genes. Gene KIW84_023481 on chromosome 2 was found to be similar in all GWAS, ND-TWAS and ME-TWAS models (See Table 3.2).

Furthermore, similar genes were identified between the protein and yield traits. Common sets of genes were also observed among different models. For example, KIW84_073247 (See Figure 3.8A & 3.9A) was also similarly identified in both GWAS and WA-TWAS. This gene encodes the filament-like protein 7 (FFP7); belonging to the long coiled-coil plant protein family (Gindullis et al., 2002). We retrieved three more genes that were present in both protein and yield trait GWAS models which were KIW84_065802, KIW84_055084, and KIW84_055103. One of the genes KIW84_065802 encodes the LOB domain-containing protein 38 which

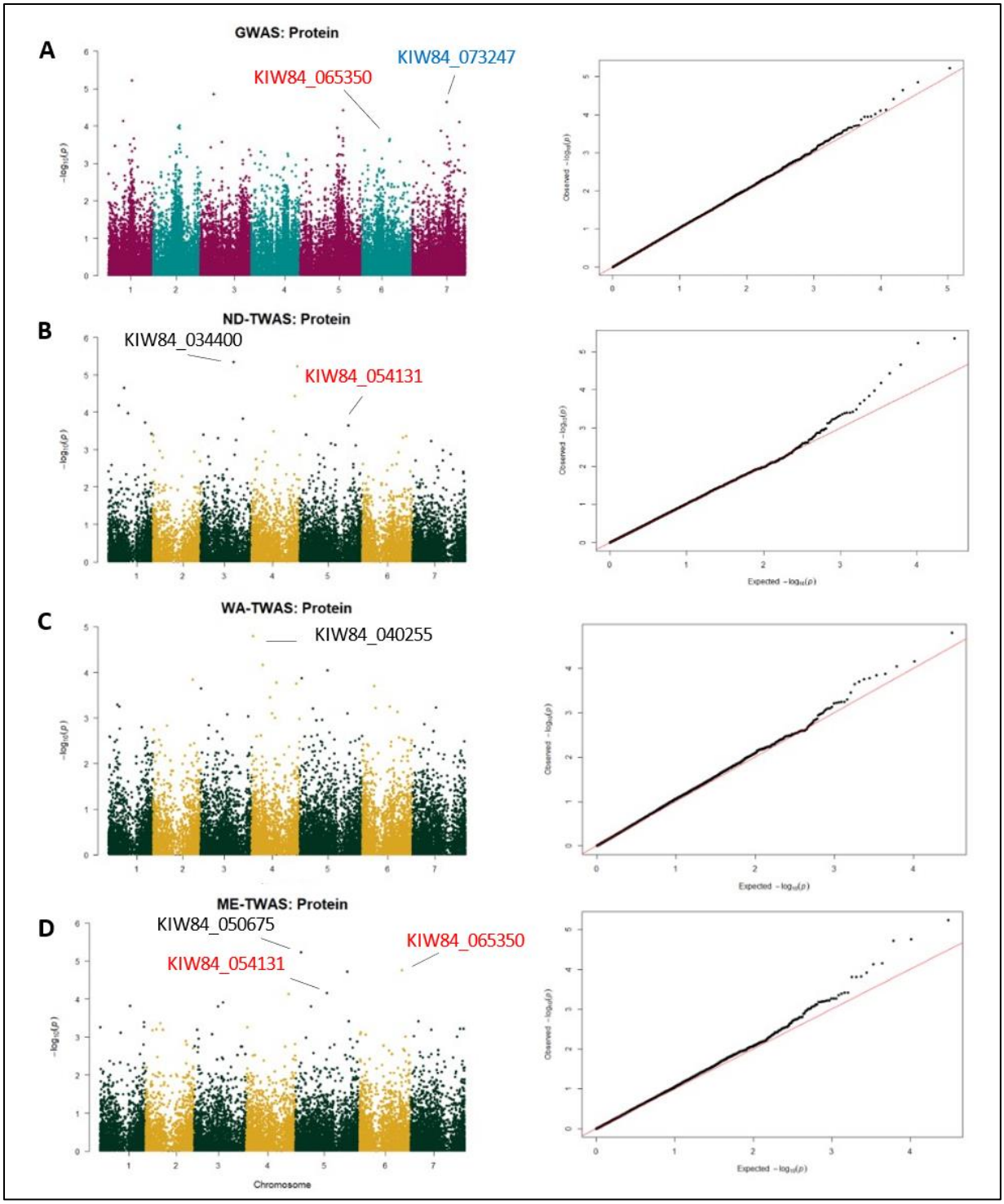

Figure 3.8. Mixed linear model - Manhattan plots of protein (A) GWAS (B) ND-TWAS (C) WA-TWAS, and (D) ME-TWAS. The genes highlighted in red falls within the top 10 genes for that model and also are common across other models. The blue highlighted gene from (A) is common in the yield trait from Figure 3.9A.

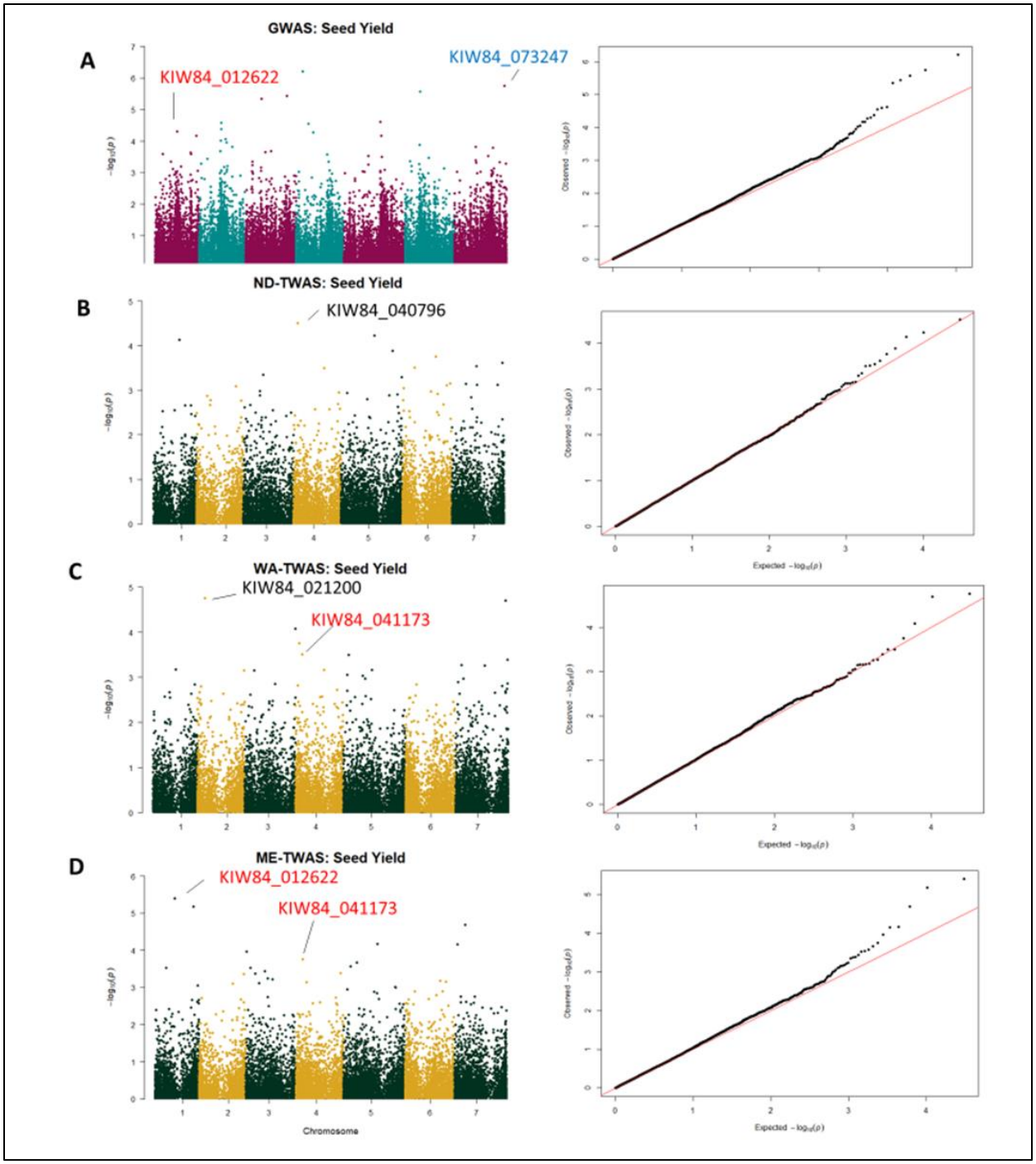

Figure 3.9. Mixed linear model - Manhattan plots of yield (A) GWAS (B) ND-TWAS (C) WA-TWAS, and (D) ME-TWAS. The genes highlighted in red falls within the top 10 genes for that model and also are common across other models. The blue highlighted gene from (A) is common in the protein trait from Figure 3.8A.

regulates gene expression and was also found in the ND-TWAS (yield) and ME-TWAS (protein) models. Additionally, we also detected one more gene in the ND-TWAS, three more in the WA-TWAS, and another three in the ME-TWAS models that are common between the traits (See Table 3.2). Another interesting gene was the KIW84_063439 which was commonly found in the ND-TWAS models for both traits, and was also detected in the GWAS and WA-TWAS protein models. KIW84_063439 is a 60S ribosomal protein-L4 which enables RNA binding and structural constituent of ribosome (Zhukov et al., 2015).

Unfortunately, many significant genes from TWAS models were not identified by GWAS and there were certain differences among the candidate genes. How does TWAS identify these significant genes that GWAS was unable to detect? One possible explanation could be the variation in the phenotypic dataset utilized in the GWAS model. This dataset consisted of 300 genotypes and incorporated the multi-environment factor over a span of 2 years. In contrast, the TWAS models (ND & WA) were based on specific environments over the same 2-year period, but with varying sample sizes (See Results). Considering that we are dealing with quantitative traits, these sample variations could lead to minor differentiation in the results. Also, if there is no notable SNP in close proximity to the causal gene, GWAS will be unable to establish a correlation between that gene and the variation in the trait (Li et al., 2021). Likewise, TWAS also did not detect significant genes that were present in the GWAS model. Some of the results may be false negatives due to the possibility that certain genes were not expressed or their expression levels were too low to be included in the subsequent analysis. According to the GWAS model, some of them could be false-positives as well or the variations in the GWAS panel are not associated with the variations in the expression data (Li et al., 2021).

Table 3.2. Top significant genes in all GWAS and TWAS models.

Gene Id	Transcript Id	Chromosome	Start	End	GWAS[a]	ND-TWAS[b]	WA-TWAS[c]	ME-TWAS[d]
KIW84_045675	*Psat04G0567500*	4	472722881	472729994	P[e]	P		
KIW84_065197	*Psat06G0519700*	6	420537202	420538798	P	P		
KIW84_010029	*Psat01G0002900*	1	665002	669503	P	P		P
KIW84_058009	*Psat05G0800900*	5	628385604	628392041	P	P		
KIW84_044335	*Psat04G0433500*	4	360303549	360323974	P	P		
KIW84_040625	*Psat04G0062500*	4	36145992	36146783	P	P		
KIW84_053407	*Psat05G0340700*	5	258658213	258660736	P	P		
KIW84_041639	*Psat04G0163900*	4	107103524	107105614	P	P		
KIW84_062665	*Psat06G0266500*	6	196634039	196635126	P	P		
KIW84_025557	*Psat02G0555700*	2	491529795	491531150	P	P		
KIW84_073610	*Psat07G0361000*	7	243998139	243998555	P		P	
KIW84_042868	*Psat04G0286800*	4	213255385	213261441	P		P	
KIW84_073247	*Psat07G0324700*	7	217451536	217454987	P, Y[f]		P, Y	
KIW84_076861	*Psat07G0686100*	7	540090158	540095165	P		P	
KIW84_071645	*Psat07G0164500*	7	110041233	110049169	P		P	
KIW84_035877	*Psat03G0587700*	3	509316691	509318930	P		P	
KIW84_035910	*Psat03G0591000*	3	510480858	510483200	P		P	
KIW84_063439	*Psat06G0343900*	6	286280948	286283037	P	P, Y	P	
KIW84_065350	*Psat06G0535000*	6	430008547	430010313	P			P
KIW84_065802	*Psat06G0580200*	6	465659895	465660653	P, Y	Y		P
KIW84_025299	*Psat02G0529900*	2	483016380	483022295	P			P
KIW84_065566	*Psat06G0556600*	6	445620844	445626359	P			P
KIW84_074115	*Psat07G0411500*	7	278789677	278791363	P			P
KIW84_074782	*Psat07G0478200*	7	342132938	342136434	P			P
KIW84_070780	*Psat07G0078000*	7	49461182	49461736	P			P
KIW84_062493	*Psat06G0249300*	6	171263956	171264525	P			P
KIW84_056247	*Psat05G0624700*	5	513018686	513020033	Y	Y		
KIW84_041746	*Psat04G0174600*	4	113824200	113830669	Y	Y		
KIW84_052148	*Psat05G0214800*	5	167464755	167468313	Y	Y		

Table 3.2. Top significant genes in all GWAS and TWAS models (continued).

Gene Id	Transcript Id	Chromosome	Start	End	GWAS[a]	ND-TWAS[b]	WA-TWAS[c]	ME-TWAS[d]
KIW84_070011	*Psat07G0001100*	7	622750	625474	Y[f]	Y		
KIW84_073282	*Psat07G0328200*	7	219989262	219990096	Y	Y		
KIW84_075976	*Psat07G0597600*	7	487030458	487031029	Y	Y		
KIW84_073633	*Psat07G0363300*	7	245009736	245010842	Y	Y		
KIW84_076941	*Psat07G0694100*	7	542446963	542448948	Y	Y		
KIW84_036286	*Psat03G0628600*	3	535390308	535393372	Y		Y	
KIW84_076937	*Psat07G0693700*	7	542368800	542371571	Y		Y	
KIW84_012817	*Psat01G0281700*	1	235782657	235788553	Y		Y	
KIW84_075987	*Psat07G0598700*	7	487435525	487435971	Y		Y	
KIW84_076804	*Psat07G0680400*	7	537619904	537621427	Y		Y	
KIW84_051500	*Psat05G0150000*	5	119085200	119087383	Y		Y	
KIW84_043428	*Psat04G0342800*	4	255938249	255946186	Y		Y	
KIW84_073994	*Psat07G0399400*	7	270698256	270709076	Y		Y	
KIW84_057896	*Psat05G0789600*	5	622745648	622749471	Y		Y	
KIW84_060282	*Psat06G0028200*	6	8574465	8577797	Y		Y	
KIW84_020560	*Psat02G0056000*	2	28761566	28766103	Y		Y	
KIW84_040445	*Psat04G0044500*	4	27320070	27323682	Y		Y	
KIW84_073140	*Psat07G0314000*	7	209779835	209794432	Y		Y	
KIW84_030145	*Psat03G0014500*	3	11597847	11603506	Y			Y
KIW84_064113	*Psat06G0411300*	6	329085938	329091987	Y			Y
KIW84_020590	*Psat02G0059000*	2	29367473	29370856	Y			Y
KIW84_014158	*Psat01G0415800*	1	378245551	378247143	Y			Y
KIW84_075051	*Psat07G0505100*	7	386498734	386500951	Y			Y
KIW84_040318	*Psat04G0031800*	4	18683999	18685489	Y			Y
KIW84_052812	*Psat05G0281200*	5	217561742	217563766	Y			Y
KIW84_055565	*Psat05G0556500*	5	458610976	458615751	Y			Y
KIW84_073249	*Psat07G0324900*	7	217877072	217883174	Y			Y
KIW84_033789	*Psat03G0378900*	3	298852358	298859092	Y			Y
KIW84_073996	*Psat07G0399600*	7	270817500	270819862	Y			Y

Table 3.2. Top significant genes in all GWAS and TWAS models (continued).

Gene Id	Transcript Id	Chromosome	Start	End	GWAS[a]	ND-TWAS[b]	WA-TWAS[c]	ME-TWAS[d]
KIW84_023481	*Psat02G0348100*	2	352575604	352577167	Y[f]	Y		Y
KIW84_055084	*Psat05G0508400*	5	389847498	389848500	P[e], Y			
KIW84_055103	*Psat05G0510300*	5	390105090	390106151	P, Y			
KIW84_064531	*Psat06G0453100*	6	364504406	364517071		Y	P	
KIW84_010174	*Psat01G0017400*	1	10755069	10761421			P, Y	
KIW84_071645	*Psat07G0164500*	7	110041233	110049169			P, Y	
KIW84_045090	*Psat04G0509000*	4	428480281	428483899				P, Y
KIW84_033002	*Psat03G0300200*	3	246495016	246504433				P, Y
KIW84_030624	*Psat03G0062400*	3	47930025	47933975				P, Y
KIW84_060149	*Psat06G0014900*	6	4302634	4305223		P	P	
KIW84_046360	*Psat04G0636000*	4	496792717	496794721		P	P	P
KIW84_051242	*Psat05G0124200*	5	98455060	98462047		P	P	
KIW84_032148	*Psat03G0214800*	3	179367126	179372251		Y	Y	
KIW84_021648	*Psat02G0164800*	2	128545485	128856972		Y	Y	
KIW84_012599	*Psat01G0259900*	1	202408817	202417410		P		P
KIW84_035274	*Psat03G0527400*	3	444081166	444085328		P		P
KIW84_056167	*Psat05G0616700*	5	507558968	507562592		P		P
KIW84_074531	*Psat07G0453100*	7	312436144	312444093		P		P
KIW84_054131	*Psat05G0413100*	5	320376797	320383213		P		P
KIW84_072119	*Psat07G0211900*	7	146487227	146492881		Y		Y
KIW84_064620	*Psat06G0462000*	6	373577608	373583949		Y		Y
KIW84_025099	*Psat02G0509900*	2	475245850	475248918		Y		Y
KIW84_073207	*Psat07G0320700*	7	214386369	214389703		Y		Y
KIW84_011503	*Psat01G0150300*	1	104701196	104713429		Y		Y
KIW84_056801	*Psat05G0680100*	5	553140383	553143580		Y		Y
KIW84_075241	*Psat07G0524100*	7	413876881	413892168		Y		Y
KIW84_040712	*Psat04G0071200*	4	41406319	41408291			Y	Y
KIW84_041173	*Psat04G0117300*	4	70339146	70350168			Y	Y
KIW84_050524	*Psat05G0052400*	5	42068139	42071428			Y	Y

Table 3.2. Top significant genes in all GWAS and TWAS models (continued).

Gene Id	Transcript Id	Chromosome	Start	End	GWAS[a]	ND-TWAS[b]	WA-TWAS[c]	ME-TWAS[d]
KIW84_050942	*Psat05G0094200*	5	75224349	75228493			Y[f]	Y
KIW84_060965	*Psat06G0096500*	6	35096931	35099670			Y	Y
KIW84_060904	*Psat06G0090400*	6	33083559	33085744			Y	Y

[a] Genome-wide association study
[b] North Dakota - Transcriptome-wide association study
[c] Washington - Transcriptome-wide association study
[d] Multi-Environment Transcriptome-wide association study
[e] Protein
[f] Yield

In the GWAS model, LD could also have resulted in some false-positive genes (Li et al., 2021), especially in self-pollinating species like pea which has a slow rate of LD-decay of ~250kb. When we used 250kb upstream and downstream from the significant markers, overall, we detected all the genes within the 500kb window size. In that case, we retrieved a greater number of genes from the top 1% SNPs than with any of the TWAS models. For example, the commonly detected gene KIW84_010029 was found in the GWAS protein model from the top 1% significant SNPs in chromosome 1 along with 52 annotated genes due to LD but in ND-TWAS and ME-TWAS models, we detected this gene directly within the top 1% genes. To further validate its significance, we conducted a Welch two sample t-test for KIW84_010029 differential gene expression between high and low protein groups using ND lines which showed that it is statistically significant with p-value < 2.2e-16 (Appendix Figure B.2(A)). Despite the high LD-decay rate, TWAS managed to detect the KIW84_010029 gene associated with protein. In a similar way, Li et al., (2021) mapped a known gene associated with a qualitative trait directly in a high-LD soybean genome using TWAS but in the GWAS, the authors detected 80 genes along with the known gene. One of the highly significant downregulated genes, KIW84_065350, found in chromosome 6 from the differential expression analysis between high vs low protein conditions was also detected in the GWAS and ME-TWAS models. This gene was further studied and found to exhibited a significance having a p-value 2.861e-12 (Welch two sample t-test) using the differential gene expression between high and low protein across multi-environment lines (Appendix Figure B.2(B)).

As previously discussed, the LD decay rate might affect the detection power of GWAS using SNP data. At the same time, if the expression levels of neighboring genes are highly correlated, it can also affect the TWAS resolving power which was noted in some of the human

TWAS studies (Wainberg et al., 2019; Mancuso et al., 2019). More false-positive signals could have been detected due to the high correlation (Zheng et al., 2020), but this was not observed in our study. Since we are dealing with quantitative traits, it is also difficult to narrow down the top significant genes as easily as it is in qualitative traits even with low correlation. In the case of soybean (Li et al., 2021), the authors knew the exact gene that they were trying to find in their studies. Since we are using the new reference genome in our study, some of the genes have not been validated and their functions are unknown. Most of the significant genes from Table 3.2 are not validated but with the model information from NCBI (https://www.ncbi.nlm.nih.gov/gene/?term=pisum+sativum) and the pea genome database developed by the Chinese Academy of Agricultural Sciences (Yang et al., 2022b), we were able to shortlist the important genes. Based on the results generated from this study, it is vital to indicate that TWAS is less affected by LD than GWAS as proved by Li et al., (2017) and Li et al., (2021) however it could supplement GWAS analysis.

Conclusion

We extended the TWAS study from qualitative traits to quantitative traits in a self-pollinating crop and noted that TWAS was also an additional resource for GWAS and they are less affected by LD (Li et al., 2017; Li et al., 2021). By utilizing the developing pod tissue that was associated with the traits, we were able to effectively integrate their expression data with the phenotypes and readily identify the genes. We noticed more similarities between the genes expressed in different environments. By analyzing protein and yield traits in individual environments, TWAS was able to detect the most significant genes even with high LD which makes this tool more helpful. We also were able to detect a greater number of significant genes that were commonly found in GWAS with multi-environment TWAS than within-environment

TWAS and the differentiation between the expression levels of the two environments explains the complex interaction of genotype by environment. Finally, it has been proven that TWAS can be used to detect the trait-associated genes in quantitative traits as well and serves as a complementary resource for GWAS.

References

Anders S, Pyl PT, Huber W (2014). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.

Ashfaq M, Rasheed A, Zhu R, Ali M, Javed MA, Anwar A, *et al.* (2023). Genome-wide association mapping for yield and yield-related traits in rice (*Oryza Sativa* L.) Using SNPs Markers. *Genes* **14**: 1089.

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

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**: 2633–2635.

Butler DG, Cullis BR, Gilmour AR, Gogel BJ, Thompson R (2022). ASReml-R reference manual version 4.1. *VSN International Ltd, Hemel Hempstead, HP1 1ES, UK*. Available at: http://www.vsni.co.uk/.

Cullis BR, Smith AB, Coombes NE (2006). On the design of early generation variety trials with correlated data. *J Agric Biol Environ Stat* **11**: 381–393.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al.* (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Gali KK, Sackville A, Tafesse EG, Lachagari VBR, McPhee K, Smýkal P, *et al.* (2019). Genome-wide association mapping for agronomic and seed quality traits of field pea (*Pisum sativum* L.). *Front Plant Sci* **10**: 466624.

Gindullis F, Rose A, Patel S, Meier I (2002). Four signature motifs define the first class of structurally related large coiled-coil proteins in plants. *BMC Genomics* **3**: 9–9.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, *et al.* (2014). Insights into the Maize Pan-Genome and Pan-Transcriptome. *Plant Cell* **26**: 121–135.

Hirschhorn JN, Daly MJ (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**: 95–108.

Hussain W, Anumalla M, Catolos M, Khanna A, Sta. Cruz MT, Ramos J, *et al.* (2022). Open-source analytical pipeline for robust data analysis, visualizations and sharing in crop breeding. *Plant Methods* **18**: 1–12.

Korte A, Farlow A (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9**: 1–9.

Kremling KAG, Diepenbrock CH, Gore MA, Buckler ES, Bandillo NB (2019). Transcriptome-wide association supplements genome-wide association in *Zea mays*. *G3: Genes|Genomes|Genetics* **9**: 3023.

Li D, Liu Q, Schnable PS (2021). TWAS results are complementary to and less affected by linkage disequilibrium than GWAS. *Plant Physiol* **186**: 1800–1811.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078.

Lin H-Y, Liu Q, Li X, Yang J, Liu S, Huang Y, *et al.* (2017). Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS. *Genome Biol* **18**: 192.

Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 1–21.

Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, *et al.* (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nat Genet* **51**: 675–682.

Money D, Migicovsky Z, Gardner K, Myles S (2017). LinkImputeR: user-guided genotype calling and imputation for non-model organisms. *BMC Genomics* **18**: 1–12.

Priyanatha C, Torkamaneh D, Rajcan I (2022). Genome-wide association study of soybean germplasm derived from Canadian × Chinese crosses to mine for novel alleles to improve seed yield and seed quality traits. *Front Plant Sci* **13**: 866300.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559–575.

R Core Team. (2023). R: A language and environment for statistical computing. https://www.r-project.org/.

Schwarz G (1978). Estimating the dimension of a model. *Ann Stat* **6**: 461–464.

Sindhu A, Ramsay L, Sanderson L-A, Stonehouse R, Li R, Condie J, *et al.* (2014). Gene-based SNP discovery and genetic mapping in pea. *Theor Appl Genet* **127**: 2225–2241.

Sudheesh S, Sawbridge TI, Cogan NO, Kennedy P, Forster JW, Kaur S (2015). De novo assembly and characterization of the field pea transcriptome using RNA-Seq. *BMC Genomics* **16**: 611.

Tayeh N, Aluome C, Falque M, Jacquin F, Klein A, Chauveau A, *et al.* (2015). Development of two major resources for pea genomics: the GenoPea 13.2K SNP Array and a high-density, high-resolution consensus genetic map. *Plant J* **84**: 1257–1273.

Tian D, Wang P, Tang B, Teng X, Li C, Liu X, *et al.* (2020). GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals. *Nucleic Acids Res* **48**: D927–D932.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, *et al.* (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**: 562–578.

Van Raden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**(11): 4414–4423.

Vasimuddin M, Misra S, Li H, Aluru S (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, IEEE pp 314–324.

Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, *et al.* (2019). Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* **51**: 592–599.

Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES (2014). Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet* **10**: e1004845.

Wu D, Li X, Tanaka R, Wood JC, Tibbs-Cortes LE, Magallanes-Lundback M, *et al.* (2022). Combining GWAS and TWAS to identify candidate causal genes for tocochromanol levels in maize grain. *Genetics* **221**: iyac091.

Yang T, Liu R, Luo Y, Hu S, Wang D, Wang C, *et al.* (2022a). Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nat Genet* **54**: 1553–1563.

Yang T, Liu R, Luo Y, Hu S, Wang D, Wang C, et al. (2022b). Pea Genome Database developed by Institute of Crop Sciences, Chinese Academy of Agricultural Sciences. Available at: https://www.peagdb.com/index/

Zeng T, Meng Z, Yue R, Lu S, Li W, Li W, *et al.* (2022). Genome wide association analysis for yield related traits in maize. *BMC Plant Biol* **22**: 1–11.

Zheng Z, Hey S, Jubery T, Liu H, Yang Y, Coffey L, *et al.* (2020). Shared Genetic Control of Root System Architecture between Zea mays and Sorghum bicolor. *Plant Physiol* **182**: 977–991.

Zhu A, Ibrahim JG, Love MI (2018). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35**: 2084–2092.

Zhukov VA, Zhernakov AI, Kulaeva OA, Ershov NI, Borisov AY, Tikhonovich IA (2015). De Novo Assembly of the Pea (Pisum sativum L.) Nodule Transcriptome. *Int J Genomics Proteomics* **2015**: 695947.

CHAPTER 4. INCREASING THE POWER OF GENETIC MAPPING BY COMBINING GWAS AND TWAS IN DRY PEA

Introduction

Dry peas (*Pisum sativum)* are one of the widely consumed pulse crops around the world, and Mendel's favorite species. The 21st century has mostly become a plant-based food century where the food markets, are always seeking for the optimal alternate source of protein, among which peas hold a strong position. This is largely attributed to their high protein content, around 32%, as well as their vitamins, fibers and minerals (Bari et al., 2021). Peas are also known to improve gut health and prevent certain cancers (Bari et al., 2021; Mudryj et al., 2014). In recent years, pea cultivation has increased due to market demands, with a greater emphasis on research in the field. Researchers in pulse crops are consistently evaluating and performing various field trials to maximize protein content. To pursue this, understanding the biological pathway underlying protein variations is crucial for the researchers. Methods such as genome-wide association studies (GWAS) and linkage mapping have been used for the past two decades to identify the genes associated with their trait-of-interest. The accuracy of linkage mapping in identifying the causal variant is constrained by the fewer recombination events from the bi-parental population (Beji et al., 2020). Following linkage mapping, GWAS emerged as a more popular method for conducting gene mapping, as it utilizes genetic diversity and ancestral recombination from diversity panels (Gupta et al., 2014). Both linkage mapping and GWAS are based on linkage disequilibrium (LD) between markers and targeted loci. However, the difference lies in the fact that LD from linkage mapping is created by biparental mating, whereas in GWAS, LD comes from the diversity panels (Bangarwa et al., 2020). LD is solely influenced by recombination in linkage mapping bi-parental populations; however, in GWAS, LD is also

influenced by other factors such as genetic drift, selection, mutation, population structure, relatedness and mating (Flint-Garcia et al., 2003). Besides linkage mapping, GWAS surpasses pedigree-based QTL mapping as it utilizes LD and recombination rates from much broader germplasm collections (Bangarwa et al., 2020). GWAS has been employed in plants, animal and human genome studies to dissect genes associated with phenotypes (Gangurde et al., 2022). GWAS is also a powerful tool in plant breeding for conducting marker-assisted selection (MAS) experiments, where it detects molecular markers linked to targeted genomic regions (Gupta et al., 2014). Over the past few years, this study has detected many significant markers associated with the trait-of-interest in peas (Desgroux et al., 2016; Gali et al., 2019; Tafesse et al., 2020; Beji et al., 2020; Martins et al., 2022).

The efficacy of gene mapping continues to advance with the development of additional novel models and methodologies. Statistical models such as naïve, population structure (Q), kinship (K) and Q+K models were employed to conduct GWAS (Sharma et al., 2018). However, the performance of GWAS alone is not sufficient due to its low statistical power and high false positive signals. To address this, more advanced tools have been developed to perform GWAS, including GAPIT (Tang et al., 2016), ECMLM (Li et al., 2014), EMMA (Kang et al., 2008), GEMMA (Zhou and Stephens, 2012), FaST-LMM (Lippert et al., 2011), SUPER (Wang et al., 2014) and GenABEL (Svishcheva et al., 2012). Despite these advancements, there is a limitation in resolution due to LD, as it cannot provide single-gene resolution. A promising emerging methodology in genetic mapping, known as transcriptome-wide association studies (TWAS), utilizes expression levels directly to pinpoint genes associated with phenotypic variations. This approach has shown success in qualitative traits and in achieving single gene resolution (Li et al., 2021). We have extended this methodology to quantitative traits such as protein and yield in

peas, demonstrating that TWAS can effectively detect highly significant genes in both within- and multi-environment analysis (see Chapter 3).

Kremling et al., (2019) enhanced gene mapping efficacy by integrating GWAS and TWAS results, thereby identifying more highly significant known genes in maize using Fisher's combined test. Following their study, this approach has been successfully applied in sorghum to identify targeted genes correlated with variations of water use efficiency-related traits (Ferguson et al., 2021; Pignon et al., 2021) and in tocochromanol levels in maize grain (Wu et al., 2022). In this study, we aim to employ this approach to enhance genetic mapping in peas. The objective of this study is to integrate GWAS and TWAS results (from Chapter 3) for seed protein and yield traits to identify highly significant genes.

Materials and Methods

Fisher's Combined Test for GWAS and TWAS

We retrieved the top 10% of GWAS (Q+K model) results with the lowest p-value SNPs (~137,725 - from Chapter 3) and used them to perform Fisher's combined test (FCT) with TWAS results based on Kremling et al., (2019). The nearest gene to each of the top 10% SNPs was extracted using the gene annotation file (CAAS_Psat_ZW6_1.0) and assigned to its respective GWAS p-values. We prioritized only the top 10% SNPs to reduce the computational burden, given the large number of 137,925 SNPs and their nearest genes. We also adjusted the TWAS results, as some genes identified in the top 10% of GWAS may not be available in TWAS; for these cases, their p-values were set to 1. Subsequently, we combined both TWAS and GWAS results based on their p-values. Finally, we conducted Fisher's combined test for each gene using the '*sumlog()'* function from the "metap" R package (version 1.1; Dewey 2019), focusing on protein and yield traits.

Statistical Analysis

We utilized the differential gene expression data of the genes KIW84_023874, KIW84_063439 and KIW84_073247 in two protein groups with multi-environment lines (high vs low) from Chapter 3 to analyze the significance of these genes using the Welch two-sample t-test (Welch 1947; Li et al., 2021). Additionally, we extracted high and low yield multi-environment lines phenotypic data, along with their respective gene expression levels, to conduct t-tests for the genes KIW84_012622 and KIW84_030145.

Results and Discussion

FCT of Protein and Seed Yield

The Fisher's combined test (FCT) was conducted by integrating GWAS (top 10%) and TWAS results retrieved from Chapter 3. Employing three FCT models such as Fisher's combined test – North Dakota (FCT-ND), Fisher's combined test – Washington (FCT-WA) and Fisher's combined test – Multi-Environment (FCT-ME) for each trait, we identified the top 1% significant genes, following Kremling et al., (2019) and Wu et al., (2022). We detected a greater number of unique and common gene sets in these models associated with protein and seed yield phenotypes. For protein traits, we narrowed down three highly significant genes from the top 10 genes: KIW84_031658, KIW84_031667 on chromosome 3 and KIW84_012863 on chromosome 1 (Figure 4.1A, B & C) which were consistent across all FCT-ND, FCT-WA and FCT-ME models. Additionally, four more genes were commonly found between FCT-ND and FCT-WA, with seven genes common in FCT-ME, and three genes between FCT-WA and FCT-ME (Table 4.1). Considering all the top 1% genes, we detected 76 genes common between FCT-ND and FCT-ME, and 431 with FCT-ME. Fifty-two genes were similar among FCT-WA and FCT-ME. One highly significant gene, KIW84_065350 (Figure 4.1C), found in FCT-ME, was also

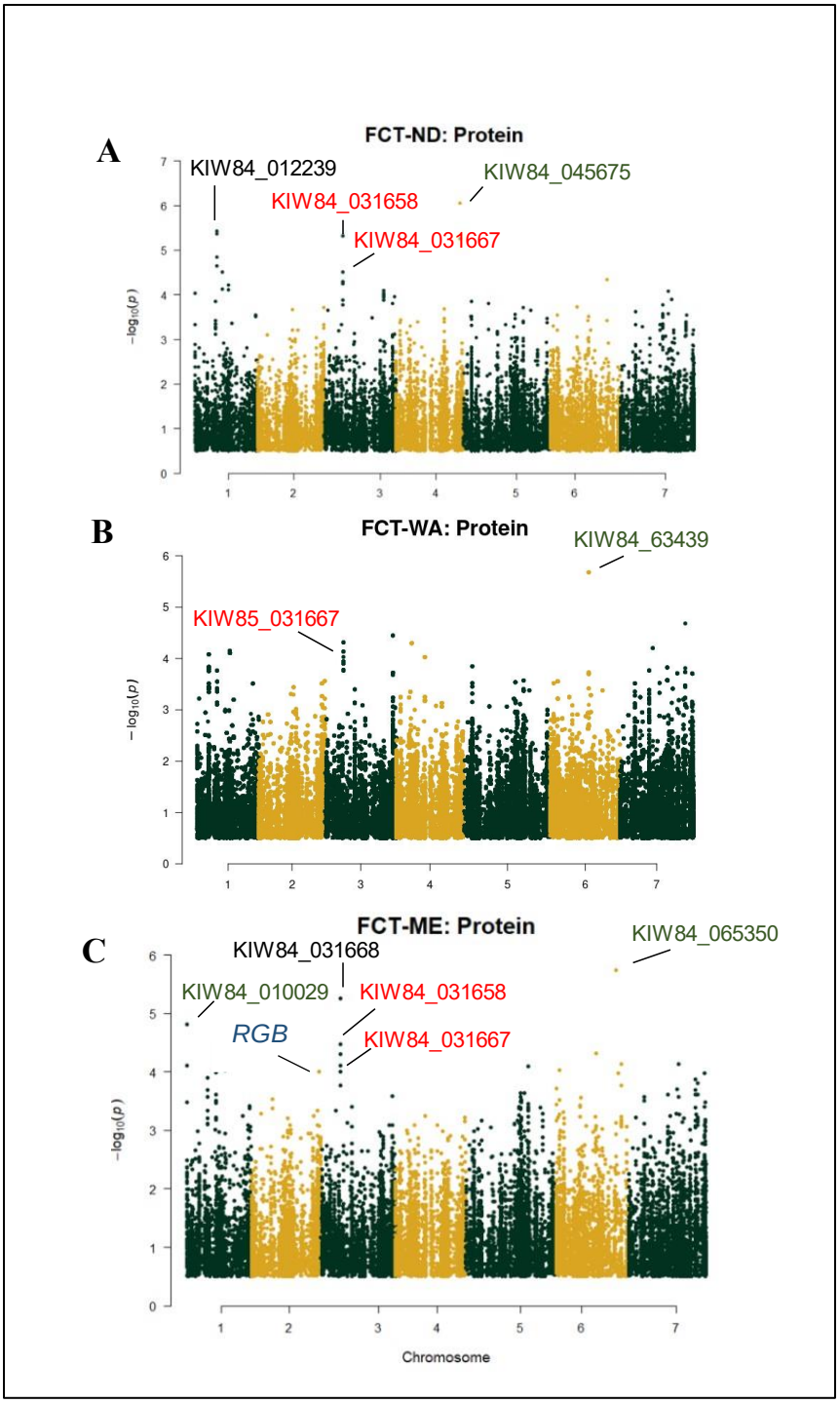

Figure 4.1. Fisher's Combined Test – Manhattan plots for protein (A) FCT-ND, (B) FCT-WA, and (C) FCT-ME. The genes highlighted in red fall within the top 10 genes for that model and also are common across other models. The green highlighted genes are similar to models from Chapter 3 (Figure 3.7A, D & Table 3.2). The *RGB* gene is highlighted in blue (C).

identified in GWAS and ME-TWAS, demonstrating its significance among high vs. low protein lines (see Chapter 3). Additionally, we identified the gene KIW84_063439 (Figure 4.1B) in WA-TWAS, GWAS, and in both traits for ND-TWAS. To assess the significance of this gene in high and low protein groups from ME, we conducted the Welch two sample t-test, confirming its high significance with a p-value of 1.766e-10 (Appendix Figure C.1(A)).

Regarding yield traits, the top 10 significant genes that were similar across all models include KIW84_041159 and KIW84_041156 on chromosome 4, KIW84_076507 on chromosome 7, and KIW84_055077 on chromosome 5. Additionally, six genes were common between FCT-ND and FCT-ME, while three genes were common between FCT-WA and FCT-ME. Two highly significant genes, such as KIW84_012622 (Chromosome 1) and KIW84_030145 (Chromosome 3) (Figure 4.2C), were also detected in GWAS and ME-TWAS models (Chapter 3). Subsequently, we performed t-test again, revealing their high significance among high vs low yield ME lines, with p-values < 2.22-16 and 4.56e-07, respectively. (Appendix Figure C.2(A) & (B))

When comparing genes associated with protein and yield traits, we found the same gene i.e., KIW84_055077 that were common in all models for both traits. Despite being detected in all TWAS models, GWAS was unable to map this gene from the significant SNPs (Chapter 3). Furthermore, we identified three additional genes in FCT-ND models for both protein and yield, two in FCT-WA models, and six more in FCT-ME models. One of these genes, KIW84_073247 from FCT-WA models, was also detected in GWAS and TWAS-WA models for both protein and yield, which is a filament-like protein 7 (Chapter 3). To further validate its significance in protein, we conducted a t-test for KIW84_073247 differential gene expression between high and

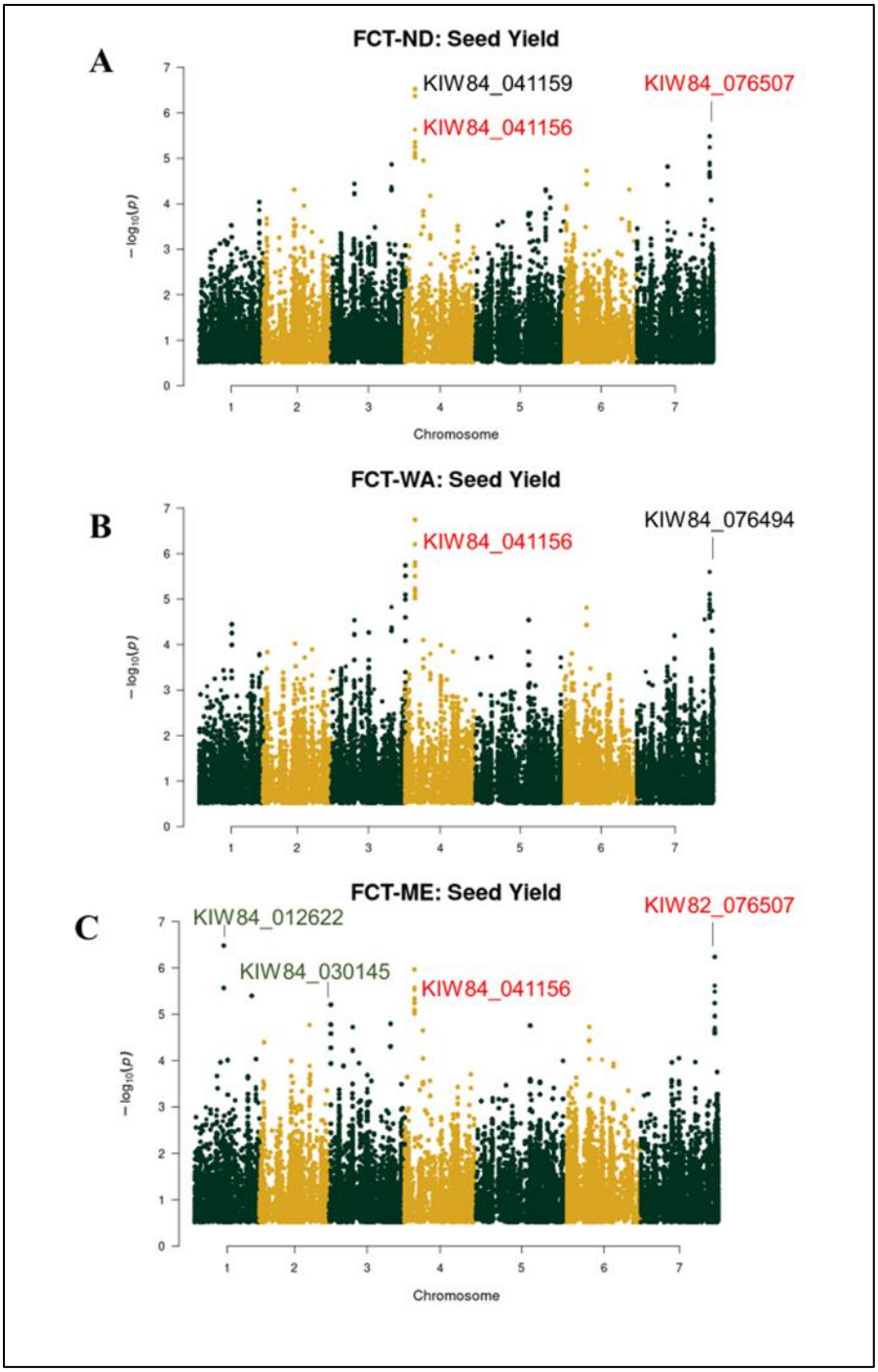

Figure 4.2. Fisher's Combined Test – Manhattan plots for seed yield (A) FCT-ND, (B) FCT-WA, and (C) FCT-ME. The genes highlighted in red fall within the top 10 genes for that model and also are common across other models. The green highlighted genes are similar to models from Chapter 3 (Figure 3.8A & Table 3.2).

Table 4.1. Top significant genes in all Fisher's combined test models.

Gene Id	Transcript Id	Chromosome	Start	End	FCT-ND[a]	FCT-WA[b]	FCT-ME[c]
KIW84_031658	*Psat03G0165800*	3	138948792	138949790	P[d]	P	P
KIW84_031667	*Psat03G0166700*	3	139552600	139556455	P	P	P
KIW84_012863	*Psat01G0286300*	1	244550420	244553547	P	P	
KIW84_036137	*Psat03G0613700*	3	525424245	525427282	P	P	
KIW84_074991	*Psat07G0499100*	7	377973060	377973738	P	P	
KIW84_050838	*Psat05G0083800*	5	66767705	66770453	P	P	
KIW84_012599	*Psat01G0259900*	1	202408817	202414484	P		P
KIW84_012863	*Psat01G0286300*	1	244550420	244553547	P	P	P
KIW84_035274	*Psat03G0527400*	3	444081166	444085328	P		P
KIW84_012047	*Psat01G0204700*	1	149571491	149572814	P		P
KIW84_035875	*Psat03G0587500*	3	509283027	509306740	P		P
KIW84_074531	*Psat07G0453100*	7	312431644	312444093	P		P
KIW84_022726	*Psat02G0272600*	2	257754793	257755365	P		P
KIW84_055501	*Psat05G0550100*	5	450734801	450743844	P		P
KIW84_063439	*Psat06G0343900*	6	286280948	286283037		P	P
KIW84_075984	*Psat07G0598400*	7	487257166	487257612		P	P
KIW84_035877	*Psat03G0587700*	3	509316691	509318930		P	P
KIW84_041159	*Psat04G0115900*	4	69617914	69620069	Y[e]	Y	Y
KIW84_041156	*Psat04G0115600*	4	69499199	69499777	Y	Y	Y
KIW84_076507	*Psat07G0650700*	7	624834489	524836686	Y	Y	Y
KIW84_062392	*Psat06G0239200*	6	156672382	156671540	Y		Y
KIW84_073282	*Psat07G0328200*	7	219989262	219990096	Y		Y
KIW84_035193	*Psat03G0519300*	3	434937960	434938322	Y		Y
KIW84_056247	*Psat05G0624700*	5	513018686	513020033	Y		Y
KIW84_022531	*Psat02G0253100*	2	222374551	222378465	Y		Y
KIW84_076494	*Psat07G0649400*	7	524197599	524199524		Y	Y
KIW84_035191	*Psat03G0519100*	3	434784481	434784826		Y	Y
KIW84_055077	*Psat05G0507700*	5	389683369	389683475	P, Y	P, Y	Y
KIW84_031661	*Psat03G0166100*	3	139247130	139247642	P, Y		

Table 4.1. Top significant genes in all Fisher's combined test models (continued).

Gene Id	Transcript Id	Chromosome	Start	End	FCT-ND[a]	FCT-WA[b]	FCT-ME[c]
KIW84_071643	*Psat07G0164300*	7	109973449	109974365	P[d], Y[e]		
KIW84_044335	*Psat04G0433500*	4	360303549	360323974	P, Y		
KIW84_073247	*Psat07G0324700*	7	217451536	217454987		P, Y	
KIW84_071645	*Psat07G0164500*	7	110041233	110041233		P, Y	
KIW84_065802	*Psat06G0580200*	6	465659895	465659895			P, Y
KIW84_076804	*Psat07G0680400*	7	537619904	537621427			P, Y
KIW84_055206	*Psat05G0520600*	5	403803138	403805391			P, Y
KIW84_032656	*Psat03G0265600*	3	223525052	223547303			P, Y
KIW84_075899	*Psat07G0589900*	7	478543931	478546088			P, Y
KIW84_056876	*Psat05G0687600*	5	557297578	557302697			P, Y

[a] Fisher's combined test – North Dakota
[b] Fisher's combined test – Washington
[c] Fisher's combined test – multi-Environment
[d] Protein
[e] Yield

low protein groups using ME lines, establishing its significance with a p-value < 2.2e-16 (Appendix Figure C.1(B)).

Some genes detected in FCT models were not identified in either GWAS or TWAS models. For instance, we identified the validated gene KIW84_023874 (*RGB* – Figure 4.1C) from the top 1% FCT-ME protein, which was absent in both GWAS and TWAS models. This gene is known to regulate cell wall biosynthesis (Daba et al., 2022; Dhugga et al., 1997), and was previously reported as a positional candidate gene by Burstin et al., (2007). Its significance was further evaluated with a t-test on high vs low protein group ME lines, confirming its significance with a p-value of 2.20e-10. Given that this study was based on the new reference genome (CAAS_Psat_ZW6_1.0- Yang *et al.*, 2022a), we were unable to confirm the functions of some genes as they are yet to be validated. However, leveraging available gene information from NCBI (https://www.ncbi.nlm.nih.gov/gene/?term=pisum+sativum) and the pea genome database developed by the Chinese Academy of Agricultural Sciences (Yang et al., 2022b), we were able to shortlist genes and identify an increased number of validated genes using Fisher's method compared to running GWAS or TWAS alone (Kremling et al., 2019). Other studies in maize grain for tocochromanol levels have also demonstrated the detection of a greater number of known genes with the FCT model compared to GWAS or TWAS alone (Wu et al., 2022), as well in maize kernel traits (Kremling et al., 2019). The Fisher's combined test has been successfully employed to retrieve candidate genes associated with maize leaf cuticular conductance (g_c) trait (Lin et al., 2022). As demonstrated in these studies, combining GWAS and TWAS increases the statistical power to detect candidate genes for the trait-of-interest (Lin et al., 2022).

Conclusion

The research findings from this study represent the first combined GWAS and TWAS models in peas and underscore their significance in genetic mapping. Utilizing GWAS, TWAS, and FCT models, we detected 45 genes for protein, 60 genes for yield, and 20 genes that were common to both traits (See Table 3.2 and 4.1), based on similarities observed across the models. Considering the top 1% genes following the methodology outlined by Kremling et al., (2019), we retrieved a greater number of significant genes associated with each phenotype. Notably, we detected a positional candidate gene, RGB, for protein using the FCT-ME model, which remained undetermined in both GWAS and TWAS models. GWAS can identify genes even in regulatory elements, while TWAS validates genes through their expression levels; hence, combining both approaches capitalize on the strengths of each study. Recent researches have increasingly proved that integrating genomics and transcriptomics yields improved results. The FCT statistical approach holds promising power for mapping candidate genes in other species as well. Future studies in pea genetic mapping should consider incorporating GWAS and TWAS into investigating other agronomic traits.

References

Bangarwa EK, Kumawat R, Barupal HL, Kumar A, Yadav MK (2020). Association Mapping in Crops Plants. *IntJCurrMicrobiolAppSci* **9**: 1313–1325.

Bari MAA, Zheng P, Viera I, Szwiec S, Ma Y, Main D, *et al.* (2021). Harnessing Genetic Diversity in the USDA Pea Germplasm Collection Through Genomic Prediction. *Front Genet* **12**: 707754.

Beji S, Fontaine V, Devaux R, Thomas M, Negro SS, Bahrman N, *et al.* (2020). Genome-wide association study identifies favorable SNP alleles and candidate genes for frost tolerance in pea. *BMC Genomics* **21**: 1–21.

Burstin J, Marget P, Huart M, Moessner A, Mangin B, Duchene C, *et al.* (2007). Developmental Genes Have Pleiotropic Effects on Plant Morphology and Source Capacity, Eventually Impacting on Seed Protein Content and Productivity in Pea. *Plant Physiol* **144**: 768–781.

Daba SD, Morris CF (2022). Pea proteins: Variation, composition, genetics, and functional properties. *Cereal Chem* **99**: 8–20.

Desgroux A, L'Anthoëne V, Roux-Duparque M, Rivière J-P, Aubert G, Tayeh N, *et al.* (2016). Genome-wide association mapping of partial resistance to *Aphanomyces euteiches* in pea. *BMC Genomics* **17**: 1–21.

Dewey M. (2019). metap: Meta-Analysis of Significance Values. R package version 1.1. https://cran.r-project.org/package¼metap

Dhugga KS, Tiwari SC, Ray PM (1997). A reversibly glycosylated polypeptide (RGP1) possibly involved in plant cell wall synthesis: purification, gene cloning, and trans-Golgi localization. *Proc Natl Acad Sci U S A* **94**(14): 7679-84.

Ferguson JN, Fernandes SB, Monier B, Miller ND, Allen D, Dmitrieva A, *et al.* (2021). Machine learning-enabled phenotyping for GWAS and TWAS of WUE traits in 869 field-grown sorghum accessions. *Plant Physiol* **187**(3): 1481-1500.

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003). Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* **54**: 357-74.

Gali KK, Sackville A, Tafesse EG, Lachagari VBR, McPhee K, Hybl M, *et al.* (2019). Genome-Wide Association Mapping for Agronomic and Seed Quality Traits of Field Pea (Pisum sativum L.). *Front Plant Sci* **10**: 466624.

Gangurde SS, Xavier A, Naik YD, Jha UC, Rangari SK, Kumar R, *et al.* (2022). Two decades of association mapping: Insights on disease resistance in major crops. *Front Plant Sci* **13**: 1064059.

Gupta PK, Kulwal PL, Jaiswal V (2014). Association mapping in crop plants: opportunities and challenges. *Adv Genet* **85**: 109–147.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, *et al.* (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**: 1709–1723.

Kremling KAG, Diepenbrock CH, Gore MA, Buckler ES, Bandillo NB (2019). Transcriptome-wide association supplements genome-wide association in *Zea mays*. *G3: Genes|Genomes|Genetics* **9**: 3023.

Li D, Liu Q, Schnable PS (2021). TWAS results are complementary to and less affected by linkage disequilibrium than GWAS. *Plant Physiol* **186**(4): 1800-1811.

Li M, Liu X, Bradbury P, Yu J, Zhang Y-M, Todhunter RJ, *et al.* (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biol* **12**: 1–10.

Lin M, Qiao P, Matschi S, Vasquez M, Ramstein GP, Bourgault R, et al. (2022). Integrating GWAS and TWAS to elucidate the genetic architecture of maize leaf cuticular conductance. *Plant Physiol* **189**: 2144–2158.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011). FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**: 833–835.

Martins LB, Balint-Kurti P, Reberg-Horton SC (2022). Genome-wide association study for morphological traits and resistance to Peryonella pinodes in the USDA pea single plant plus collection. *G3 Genes|Genomes|Genetics* **12**: jkac168.

Mudryj AN, YuNancy, AukemaHarold M (2014). Nutritional and health benefits of pulses. *Appl Physiol Nutr Metab* **39**(11): 1197-204.

Pignon CP, Fernandes SB, Valluru R, Bandillo N, Lozano R, Buckler E, *et al.* (2021). Phenotyping stomatal closure by thermal imaging for GWAS and TWAS of water use efficiency-related genes. *Plant Physiol* **187** (4): 2544-2562.

Sharma SK, MacKenzie K, McLean K, Dale F, Daniels S, Bryan GJ (2018). Linkage Disequilibrium and Evaluation of Genome-Wide Association Mapping Models in Tetraploid Potato. *G3: Genes|Genomes|Genetics* **8**: 3185.

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012). Rapid variance components–based method for whole-genome association analysis. *Nat Genet* **44**: 1166–1170.

Tafesse EG, Gali KK, Lachagari VBR, Bueckert R, Warkentin TD (2020). Genome-Wide Association Mapping for Heat Stress Responsive Traits in Field Pea. *Int J Mol Sci* **21**(6): 2043.

Tang Y, Liu X, Wang J, Li M, Wang Q, Tian F, *et al.* (2016). GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *Plant Genome* **9**: lantgenome2015.11.0120.

Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z (2014). A SUPER Powerful Method for Genome Wide Association Study. *PLoS One* **9**: e107684.

Welch BL (1947). The generalization of ‘student’s’ problem when several different population varlances are involved. *Biometrika* **34**: 28–35.

Wu D, Li X, Tanaka R, Wood JC, Tibbs-Cortes LE, Magallanes-Lundback M, *et al.* (2022). Combining GWAS and TWAS to identify candidate causal genes for tocochromanol levels in maize grain. *Genetics* **221**: iyac091.

Yang T, Liu R, Luo Y, Hu S, Wang D, Wang C, et al. (2022a). Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nat Genet* **54**: 1553–1563.

Yang T, Liu R, Luo Y, Hu S, Wang D, Wang C, et al. (2022b). Pea Genome Database developed by Institute of Crop Sciences, Chinese Academy of Agricultural Sciences. Available at: https://www.peagdb.com/index/

Zhou X, Stephens M (2012). Genome-wide Efficient Mixed Model Analysis for Association Studies. *Nat Genet* **44**: 821.

APPENDIX A

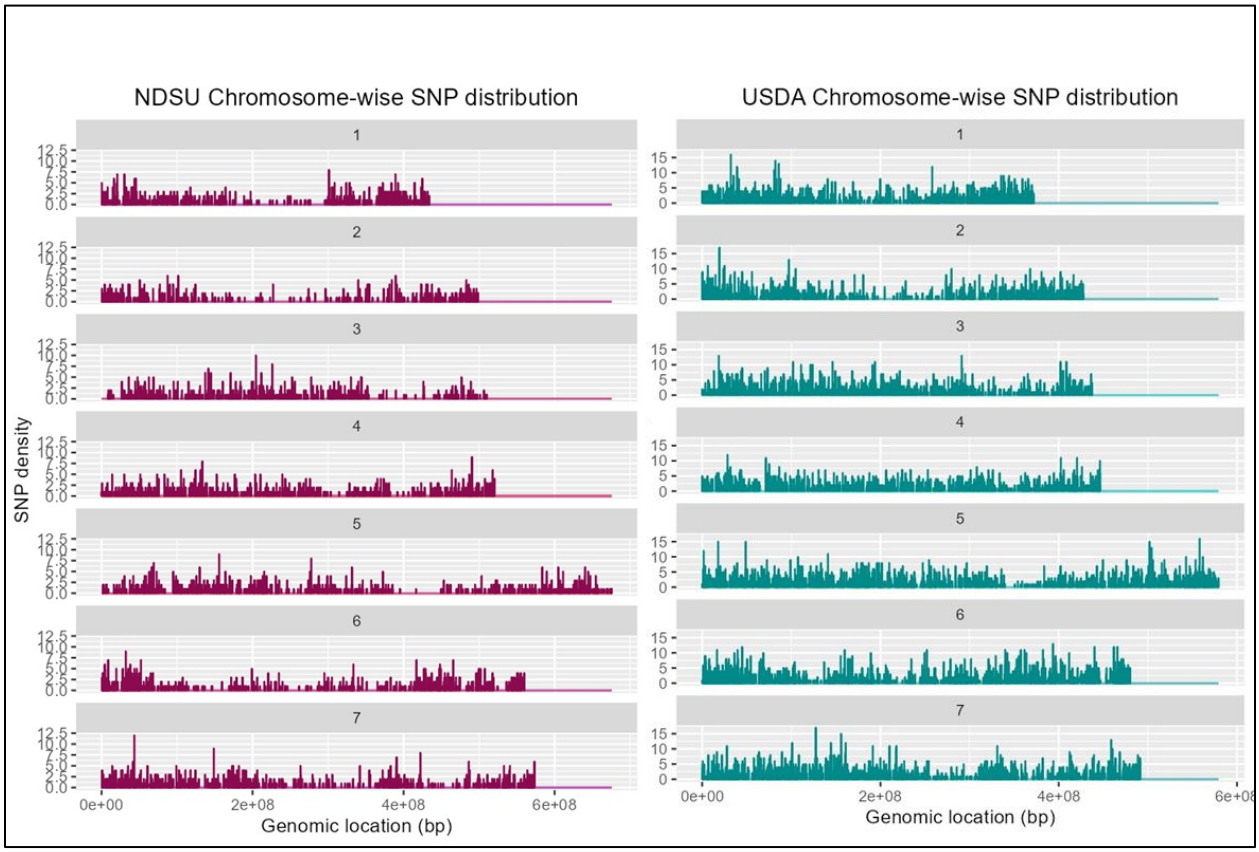

Figure A.1. SNP density of NDSU and USDA set, x-axis is the genomic location (bp) and y-axis is the density.

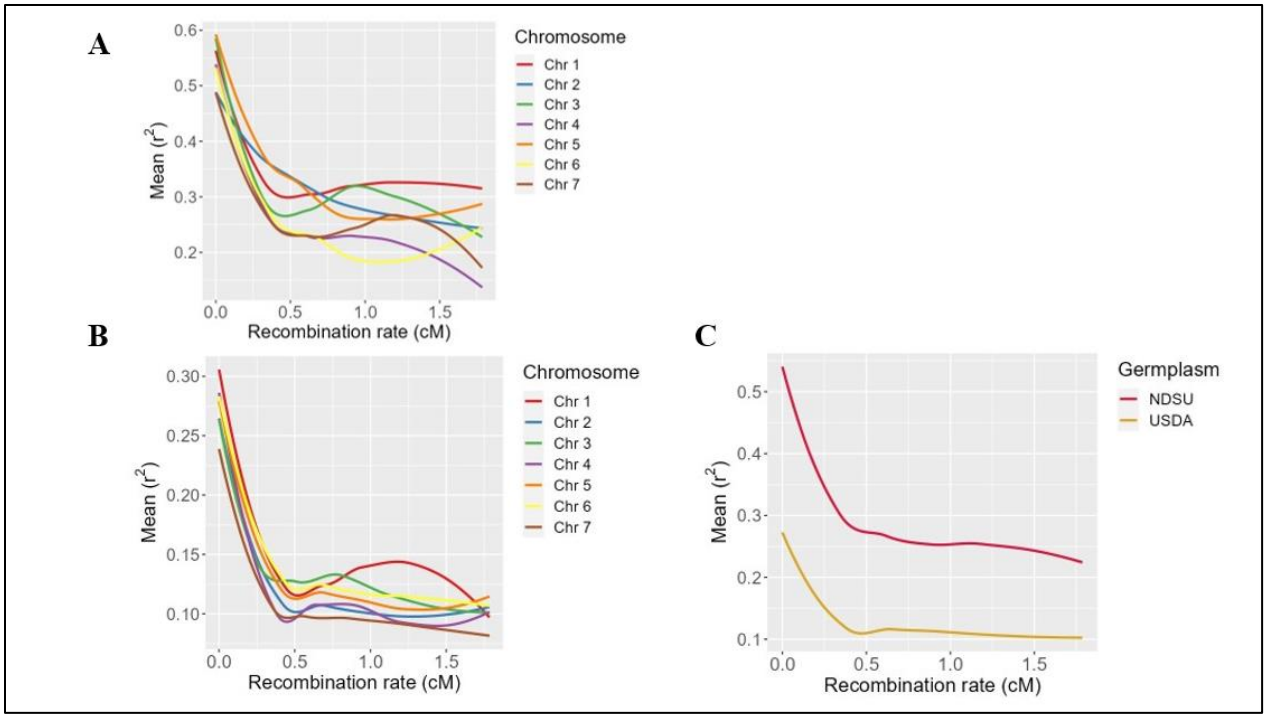

Figure A.2. Genome-wide (C) and Chromosome-wide linkage disequilibrium decay in the NDSU (A) and USDA (B) with mean of r^2 (y-axis) and recombination rate (cM) (x-axis)

APPENDIX B

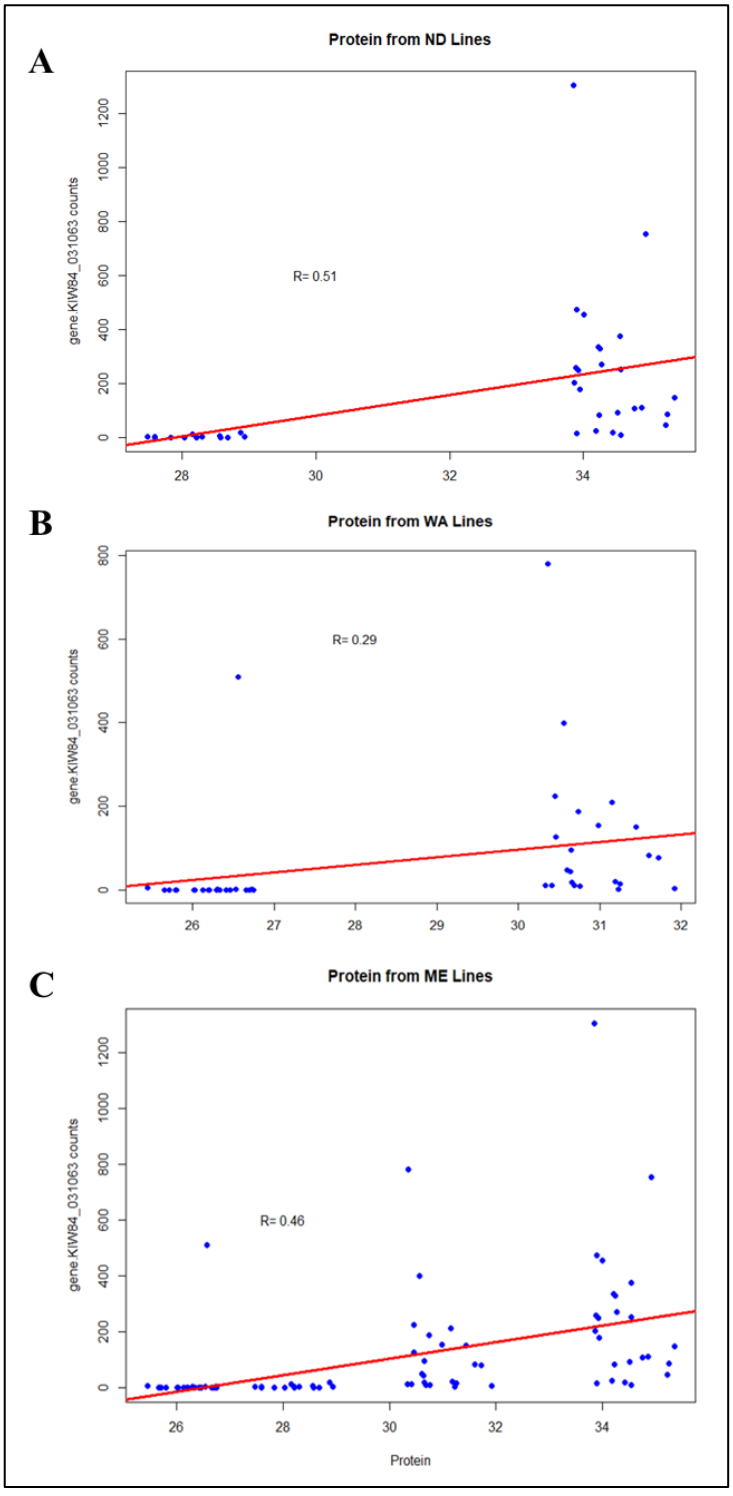

Figure B.1. Relationship between gene KIW84_031063 to protein (high and low genotypes) showing positive correlation A) Correlation of the gene to ND lines with R=0.51 B) Correlation of the gene to WA lines with R=0.29, and C) Correlation of the gene to ME lines with R=0.46, which is higher than the individual environments relationship with the gene.

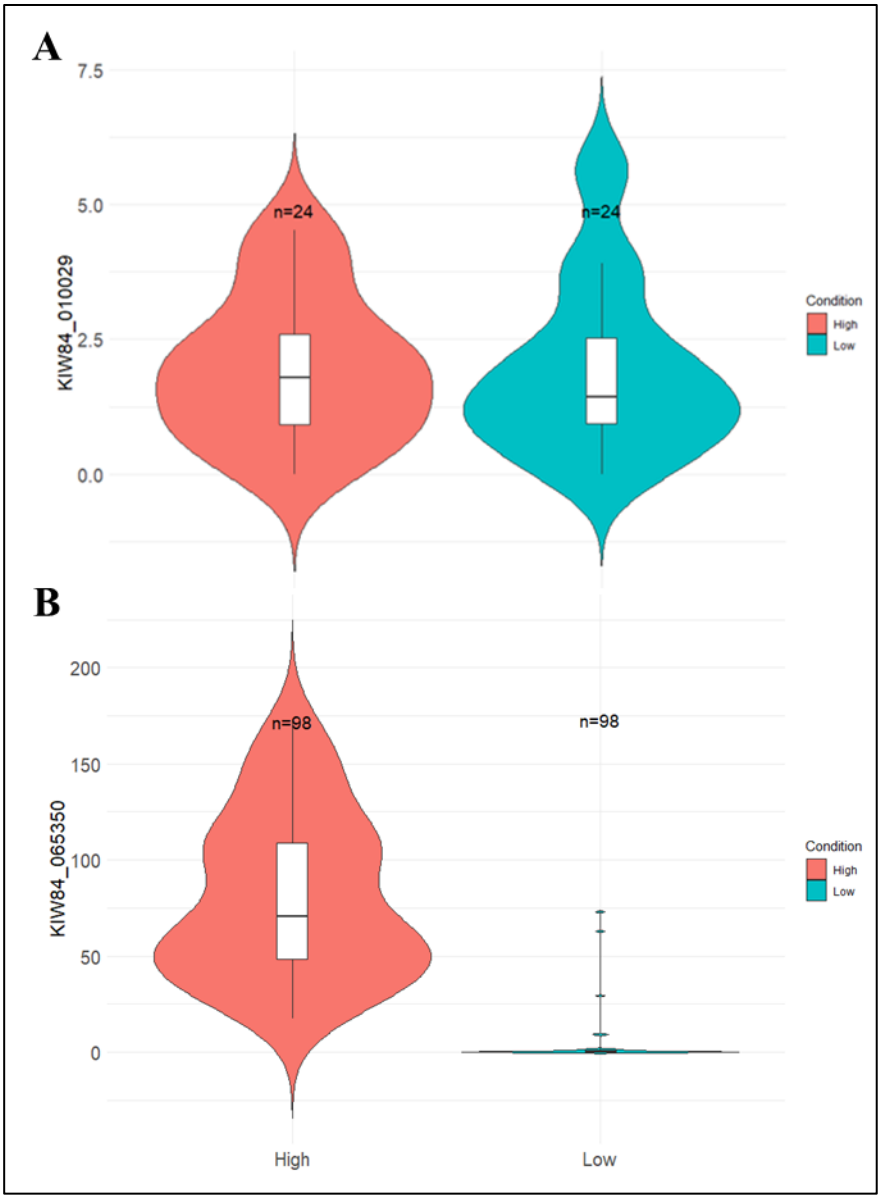

Figure B.2. A) Violin plots of gene KIW84_010029 in ND lines with high and low protein, showing the probability density curves of the KIW84_010029 gene in the two protein groups, B) Violin plots of gene KIW84_065350 in ME lines with high and low protein, showing the probability density curves of the KIW84_065350 gene in the two protein groups.

APPENDIX C

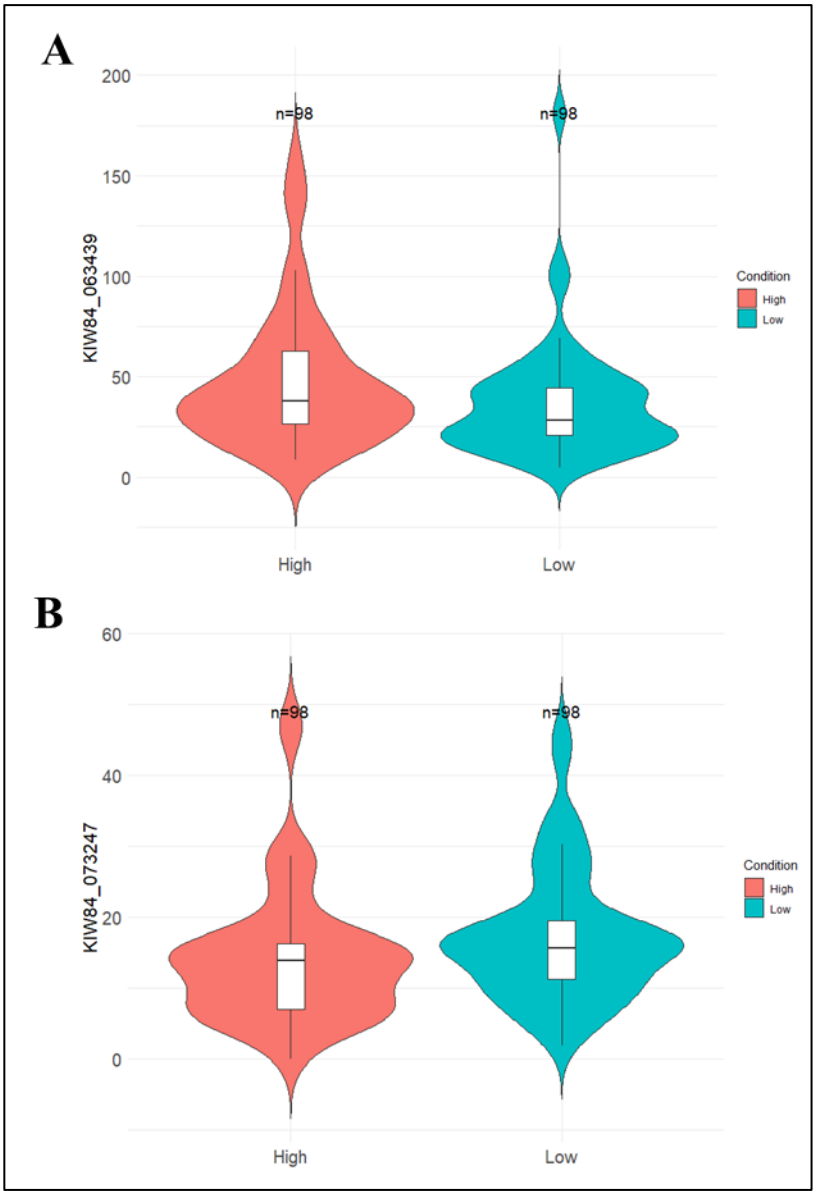

Figure C.1. (A) Violin plots of gene KIW84_063439 in ME lines with high and low protein, showing the probability density curves of the KIW84_063439 gene in the two protein groups, B) Violin plots of gene KIW84_073247 in ME lines with high and low protein, showing the probability density curves of the KIW84_073247 gene in the two protein groups.

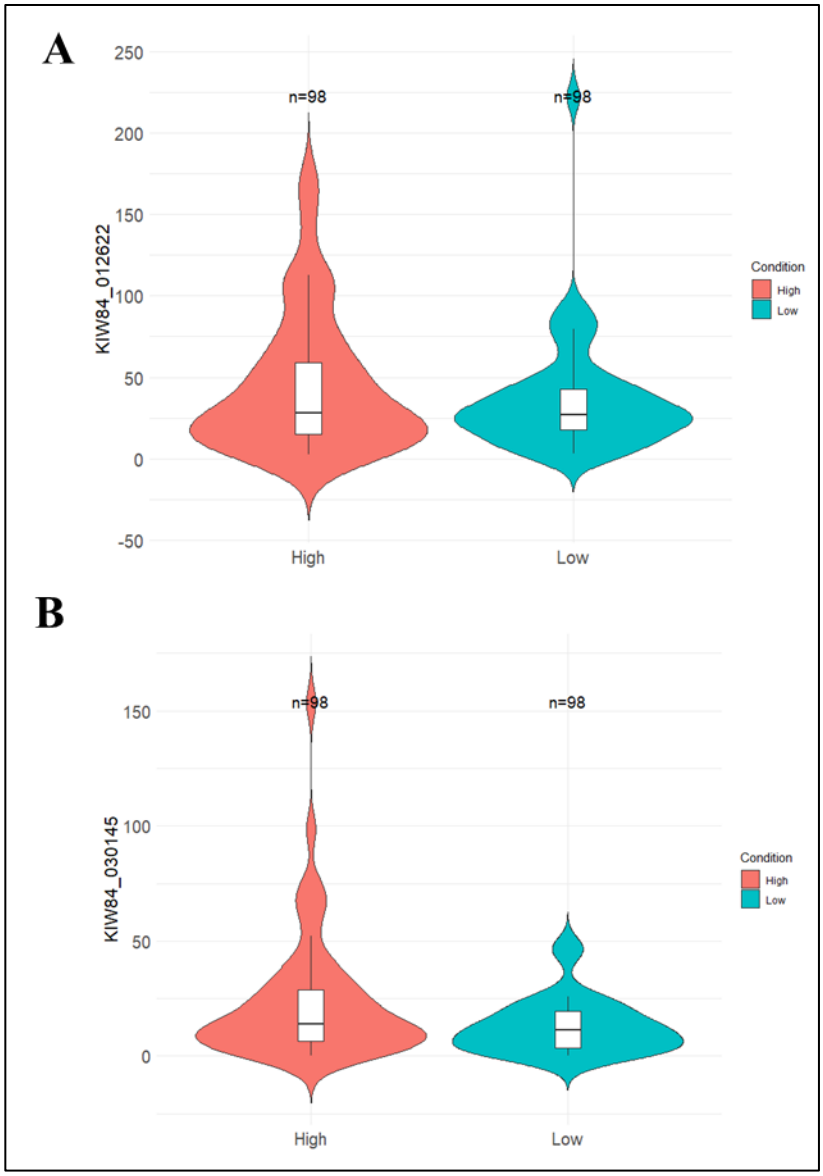

Figure C.2. (A) Violin plots of gene KIW84_012622 in ME lines with high and low yield, showing the probability density curves of the KIW84_012622 gene in the two yield groups, B) Violin plots of gene KIW84_030145 in ME lines with high and low yield, showing the probability density curves of the KIW84_030145 gene in the two yield groups.