ADVANCING THE DISCOVERY OF CAUSAL ALLELES CONTROLLING AGRICULTURAL PHENOTYPES IN APPLE (*MALUS DOMESTICA*)

by

Thomas Davies

Submitted in partial fulfilment of the requirements for
the degree of Doctorate of Philosophy

at

Dalhousie University
Nova Scotia, Canada
August, 2024

Dalhousie University is located in Mi’kma’ki, the
ancestral and unceded territory of the Mi’kmaq.
We are all Treaty people.

Table of Contents

List of Figures

Abstract

The apple is an economically and culturally important fruit crop grown in temperate regions around the world. The apple stands to benefit greatly from modern genomics, including the application of genomics informed breeding and gene editing technologies. However, the discovery of the DNA sequences, or causal alleles, that control apple phenotypes remains as a key barrier to the rapid advancement of genomics informed breeding and gene editing in apple. The objective of this thesis is to advance the current state of knowledge in the areas of apple phenomics and genomics by leveraging the wealth of phenotypic and genetic diversity in Canada's Apple Biodiversity Collection (ABC). The present thesis presents a series of studies aimed at quantifying multiple apple phenotypes and revealing the identity of causal alleles that control them. Specifically, this research aims to provide a detailed comparison of the phenotypic differences between domesticated and wild apples, and to make contributions toward the discovery of causal alleles controlling apple phenotypes. In chapter 1, I compare the domestic apple to its wild progenitor by analysing 10 key agricultural phenotypes. This analysis reveals significant differences in key agricultural phenotypes between the two species. For example, domesticated apples contain 68% less phenolic content than their wild counterparts, on average. This investigation suggests that domesticated apples have significantly diverged from their primary wild progenitor species, and that wild apples may offer important germplasm for breeding in the future. In chapter 2, I use a pool-sequencing genomics approach to scan the genome for regions that control ripening time, phenolic content, and fruit softening. This investigation identifies a number of regions in the genome that are likely to harbour causal alleles. For instance, the analysis in chapter three identifies a region upstream of a well-known transcription factor gene *NAC18.1* as being potentially causal for ripening time. In chapter 4, a genome wide association study is conducted using high depth DNA sequence data from 97 diverse samples from the ABC with the aim of discovering the causal allele for ripening time in apple. Results from chapter 4 delimit a narrow region of the genome on chromosome 3 probable to harbour the causal allele for ripening time, and illustrate some challenges associated with generating gene editing targets from association studies. Finally, a reference panel is generated for future genotype imputation in the ABC population. Chapter 5 summarises the findings of this thesis and provides context and prospective directions for enabling gene editing technologies in apple in the near future. Causal allele discovery in apple will remain challenging moving forward, however the research presented here represents key steps towards identifying causal alleles for a few key phenotypes and provides a foundation for unlocking the full mapping potential of Canada's ABC.

List of Abbreviations Used

ABC	Apple Biodiversity Collection
ACC	1-amino-cyclopropane-1-carboxylic acid
ACO	1-amino-cyclopropane-1-carboxylic acid oxidase
ACS	1-amino-cyclopropane-1-carboxylic acid synthase
CAS	CRISPR associated
CNV	Copy number variant
CRW	Crop wild relative
CRISPR	Clustered regularly interspaced palindromic repeats
DSB	Double stranded break
ERF	Ethylene response factor
gRNA	Guide ribonucleic acid
GWAS	Genome-wide association study
HDR	Homology directed repair
INDEL	Insertion/deletion
LAR	Leucoanthocyanidin reductase
LD	Linkage disequilibrium
MAF	Minor allele frequency
MAS	Marker associated selection
MLM	Mixed linear model
MLMM	Multi-locus mixed-model
NAC	No Apical Meristem, ATAF1/2, and CUC
NHEJ	Non-homologous end joining

NOR	Non-ripening
PAM	Protospacer adjacent motif
PAV	Presence/absence variant
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
pegRNA	prime editing guide RNA
QTL	Quantitative trait loci
RNA	Ribonucleic acid
RNAi	Ribonucleic acid interference
RNP	Ribonucleic protein particle
SAM	Sterile alpha motif
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SV	Structural variant
SSC	Soluble solids content
T-DNA	Transfer-DNA
TALEN	Transcription activator-like effector nuclease
VOC	Volatile organic compounds
WGS	Whole genome sequencing
ZFN	Zinc finger nuclease

Acknowledgements

Thank you to Dr. Sean Myles for the thoughtful mentorship and support during my graduate studies. Through your teachings I learned the meaning of generosity, value, and opportunity. I hope that future graduate students are fortunate enough to learn about science and themselves under the guidance and vision of leaders like you.

Thank you to Dr. Sophie Watts and Tayab Soomro for comradery, friendship, and technical support during my studies. It was a pleasure to be a part of a team together.

Thank you Dr. Vasantha Rupasinghe for your flexibility and support. Thank you Dr. Fraser Clark for your feedback and support. Thank you Dr. Sephanie Colombo for your support. Thank you Dr. Zoë Migicovski for your mentorship and passion.

Thank you to the Natural Sciences and Engineering Research Council of Canada (PGS-D scholarship) and Dalhousie (Faculty of Agriculture Scholarship, L2M Research Scholarship) for providing me with the ability to fund my education.

Thank you to Dr. Nathan Pumplin for your mentorship and guidance. Thank you to Dr. Pat Brown for the opportunity to visit and learn from your research team.

Thank you to my family for the relentless support and the encouragement to pursue the whichever path calls to me.

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

INTRODUCTION

Global food production must increase by at least 55% by 2050 in order to sustain the growing human population, which is projected to reach 9.7 billion (McKenzie & Williams, 2015). In meeting this demand, it will be critical to leverage the world's diversity of crops for sustainable and robust food production. While the majority of human-consumed calories come from a small number of widely produced annual crops (maize, rice, wheat, and soy), perennial crops play an essential role in our food system: they include species that account for approximately one eighth of the world’s food producing surface (Gaut et al., 2015).

The apple (*Malus domestica*) is a perennial crop that is the world's third most produced fruit by weight (FAOSTAT, 2020), and Canada’s second most valuable fruit crop (Agriculture and Agri-Food Canada, 2022). While tremendous progress has been made via traditional breeding in apple, there remains much to be gained in terms of apple nutrition, fruit quality, climate resilience, and disease resistance. However, apple improvement remains a serious challenge and the slow breeding cycle leaves the apple industry slow to adapt to changes in agricultural practices, conditions, regulations, and consumer expectations (Pereira-Lorenzo et al., 2018). In a time of powerful biological and data analytics tools, improvement of apple varieties through an understanding of the plant's molecular biology will be critical for the timely development of improved apple varieties. Given the importance of the apple industry and apple variety improvement, it is critical that all available biological approaches be used to further understand the fundamental biology of apple, and novel strategies be employed to accelerate variety improvement.

Genomics is a field that holds promise to dramatically advance the improvement of agricultural species (Abberton et al., 2016), including apple. Genomics, the study of an organism's DNA sequence, has already had a significant impact on apple improvement and agricultural crops more broadly. The increasing number of genomic approaches and genetic resources, like accessible DNA sequencing and reference genomes, has ushered in an era of unrivalled ability to explore and understand the apple genome (Peace et al., 2019; X. Sun, Jiao, et al., 2020; Y. Sun, Shang, et al., 2022). Understanding the genetic control of traits is a core objective of genetics, and an understanding of the genetic underpinnings of key traits in apple will be essential for rapid improvement of varieties in the future, particularly by leveraging technologies such as gene editing.

To advance apple variety improvement, whether it be through genomics assisted breeding or gene editing, the genetic control of key traits must be sufficiently characterised. Genetic mapping, particularly using diversity panel populations, whole genome sequencing, and statistical association methods such as the genome-wide association study (GWAS), stands as a promising method to discover causal DNA sequences that control apple plant traits. Given the demand for high quality apple varieties, and the persistent challenges in improvement, high resolution genetic mapping experiments hold immense promise for the future of apple improvement.

The goal of this work is to explore and analyse the phenome and the genome of the apple in an attempt to discover DNA sequences that control important agricultural traits in apple. First, phenotype data from the domesticated apple and its wild progenitor species is analysed to better understand the distribution of apple phenotypes and the influence of domestication on key apple phenotypes. Then, phenotype data from a diverse collection of domesticated apple species is used in tandem with next generation sequencing technologies to scan the genome for regions that could be impacting key agricultural traits. Finally, a diverse apple population is leveraged to conduct association mapping for the genetic control of ripening time and to lay the groundwork for future genetic mapping studies.

LITERATURE REVIEW

HISTORIC AND MODERN CONTEXT OF THE APPLE

History and evolution of the apple

The apple (*Malus domestica*) belongs to *Rosaceae*, a large plant family containing multiple fruiting species including peaches, pears, plums, apricots, cherries, strawberries and almonds. The apple is one of the world's most ancient crops, domesticated more than 5,000 years ago in the Tian Shan region of central Asia (Cornille et al., 2014, 2019). The first apples to be cultivated were domesticated from *M. sieversii* (Cornille et al., 2012; Harris et al., 2002). Following domestication, apples were brought West along the Silk Road trading routes where multiple Malus species, including *M. bacatta*, *M. orientalis*, and *M. sylvestris* hybridised (Cornille et al., 2014). A complex history of bidirectional hybridization over thousands of years makes accurately resolving the relationships between these species difficult, but it is clear that each contributed significantly to the genome of the modern domesticated apple (X. Sun, Jiao, et al., 2020). Archeological evidence suggests that apples were established as a cultivated crop in Greek societies as early as 300 BC, and that agriculturalists at the time had already discovered advanced grafting and storage techniques (Juniper & Mabberley, 2006). Over the last 2,000 years, the apple has become one of the globe's most important agricultural fruit species.

Economic, cultural, and agricultural importance of the apple

The apple is the world's third most valuable fruit crop, worth $79B annually (FAOSTAT, 2022). In temperate countries, apple production often represents a significant proportion of fruit production. For example, in Canada the farm gate value of apple was $242M as of 2021, accounting for more than one fifth of the total farm gate value of fruit for the nation (Agriculture and Agri-Food Canada, 2022). In addition to being a major economic driver in the fruit production industry, the apple continues to play a critical role in contemporary culture. The apple appears in modern folklore, art, interpretations of the Bible, and is frequently used to symbolise fruit, health, and nutrition (Juniper & Mabberley, 2006). Apples are important for agricultural systems in temperate regions, as they provide key economic opportunities for food producers, sequester more carbon than annual crops, and offer unique opportunities for connectivity between producers and consumers (Kreitzman et al., 2020; Vallebona et al., 2016).

Wild apple species

There are at least 30 species within the genus *Malus*, and the phylogeny of the genus is likely to be rewritten and updated in the future (Robinson et al., 2001). Apple species in this genus are able to hybridise, making the relationships between the species ambiguous. Domesticated apples have long been cultivated in regions where wild apples grow naturally, facilitating hybridization events spanning hundreds or thousands of years, further complicating the relationships between Malus species (Cornille et al., 2014). Wild apples, which have already contributed significantly to the genome of *M. domestica (X. Sun, Jiao, et al., 2020)*, offer unique traits and genetics that could be of value for modern apple improvement in the future. Fruit traits among wild species are highly variable: *M. baccata* produces small berry-like fruits (Cornille et al., 2012) while *M. sieversii* can produce large fruits with eating quality comparable to modern cultivars (Volk et al., 2013). Wild species are also frequently examined for disease resistance traits, and have been a major source of genetic material for breeding programmes focussed on resistance to apple canker and fire blight (Harshman et al., 2017; Kost et al., 2015; X. Liu et al., 2021; Schlathölter et al., 2023). Importantly, wild apple species are also a source of important phenolic compounds (Volz & McGhie, 2011; N. Wang et al., 2018), that produce the red flesh apple phenotype and play essential roles in human nutrition and health (Huber & Rupasinghe, 2009). While some understanding of wild apple traits has been established, there is still much to be gained through comprehensive phenotypic comparisons between wild and domesticated apples. Wild apple species will be a crucial source of genetic breeding material in the future, and as such it will be important to comprehensively understand the traits of wild apple species, as well as their relationship to the domesticated apple.

APPLE BIOCHEMISTRY AND RIPENING

The nutritional value of apples

The apple plays a key role in the modern human diet as a source of vitamins, minerals and bioactive compounds. Phenolic compounds are a well-studied group of secondary metabolites present in apples (Lu & Foo, 1997) which are more abundant in fruit skin tissues (Takos et al., 2006) and contribute significantly to the nutritional value of apples (Dixon et al., 2005). There are five major groups of polyphenols: flavan-3-ols/procyanidins, dihydrochalcones, flavonols, anthocyanins, and phenolic acids (Huber & Rupasinghe, 2009; Rupasinghe et al., 2017). Nutrition studies have elucidated the impressive human health benefits of consuming phenolic compounds, which include decreased risk of cardiovascular disease, type 2 diabetes, inflammation, metabolic disorders, and some cancers (D. Lin et al., 2016; Shetty & Wahlqvist, 2004). Phenolic compounds are also known to be antimicrobial and are a focus for the basis of plant-derived medicines (Ahmed et al., 2016; Kawabata et al., 2019; Wijesundara et al., 2021). In the US, 22% of the polyphenols in the human diet originate from apples, which makes apples a primary dietary source of these antioxidant compounds (Vinson et al., 2001). Unfortunately, as of 2019, only 10% of North Americans were meeting their daily recommended fruit and vegetable requirements (S. H. Lee et al., 2022), undoubtedly leading to poorer population health outcomes. While most of the population in North America fails to reach adequate fruit consumption, apples play a vital role as the most available fruit for consumption (Weber et al., 2023) due to their excellent storability, affordable cost, and production on both hemispheres of the planet. The health benefits of eating apples has been long recognized, and apples are likely to continue to be a major source of human nutrition in the future.

Climacteric fruit ripening and ethylene

Fruit ripening is an important aspect of fruit biology and can be broadly divided into two classes: climacteric and non-climacteric. Apples are a climacteric fruit, meaning they undergo a ripening process that is characterised by peaks in respiration and ethylene production (McMurchie et al., 1972; Oeller et al., 1991). While there is variation in ripening mechanisms between species, ripening in climacteric fruits is controlled by similar regulatory networks (Adams-Phillips et al., 2004; Giovannoni, 2004; Lü et al., 2018). Many fruits including tomato, banana, pear, peach, apricot, fig, and papaya are considered climacteric.

Ripening is a complex physiological process characterised by changes in fruit colour, texture, firmness, volatile organic compound (VOC) production, sugar metabolism, and anthocyanin production (Dandekari et al., 2004; Defilippi et al., 2005). The ripening process in apple (and other climacteric fruits) is largely mediated by the plant hormone ethylene (Blanpied, 1972; McMurchie et al., 1972), a small gaseous hydrocarbon. Ethylene production can be viewed as a two step process. First, S-adenosyl-l-Met (SAM) is converted into 1-amino-cyclopropane-1-carboxylic acid (ACC) by the enzyme ACC synthase (ACS). Second, ACC is converted into ethylene via ACC oxidase (ACO) (S. F. Yang & Hoffman, 1984). Two systems of ethylene production have been defined in fruit: system 1 and system 2, which are autoinhibitory and autocatalytic, respectively. System 1 primarily functions during normal plant growth, whereas system 2 functions during fruit maturity and ripening (Barry & Giovannoni, 2007; McMurchie et al., 1972). The crucial role of ethylene in fruit ripening is difficult to overstate, as ethylene insensitive plants are unable to produce ripe fruit (Lanahan et al., 1994; Oeller et al., 1991; Yen et al., 1995). Further, evidence strongly suggests that ethylene impacts the regulation of dozens or hundreds of genes crucial to the ripening network (Tadiello et al., 2016). Although ethylene is arguably the most important hormone in the development and ripening of climacteric fruits, it fits into a larger network of genes and hormones that make up a highly complex fruit ripening mechanism.

Numerous genes are involved in the apple ripening process, many of which have poorly understood roles. For example, evidence suggests that *NAC18.1*, the homolog of tomato *non-ripening* (*NOR*), is a high-level regulator of the ripening mechanism in apple (Migicovsky et al., 2021), but the precise details of its modes and degree of interaction with other genes in the ripening network remain unclear. As another example of complex ripening interactions, *ethylene response factor 4* (*ERF4)* mutants have been discovered to have differential influence on apple ripening, including ERF4s ability to bind TOPLESS co-repressor 4 (TPL4) and regulate *ERF3* and ethylene. However, computational analysis suggests that ERF4 may be interacting with thousands of genes, many of which have unknown functions (Y. Hu et al., 2020). Research is ongoing to understand the many interacting genes involved in apple ripening.

It is important to understand ripening because it directly impacts the quality and production of fruit. In the case of apple, ripening time is important for multiple entities in the fruit supply chain, as fruits are often stored for up to a year before reaching supermarket shelves. The timing and effects of ripening are also important for apple producers, who must carefully plan the harvest, storage, and ripening of each year's crop in order to ensure farm profitability (R. Blakey, personal communications, 2022; M. Van Meekran, personal communications, 2021). For example, if an entire orchard is planted with a single variety, an entire years harvest will ripen and require picking within the same 10 day period, creating an unmanageable logistics challenge for growers. The industrial importance of ripening time in apples is well evidenced by the fact that the top 9 cultivars in the USA are homozygous for an allele linked to late harvest date (Migicovsky et al., 2021). A deeper understanding of the mechanisms controlling apple ripening is key for unlocking accelerated breeding for key traits in apple, including and improved fruit storage and firmness.

APPLE BREEDING

Breeding history of apple

The fundamental life-cycle traits of the apple make for a unique model of domestication and breeding that is dramatically different from annual crops. Apples have long generation times (requiring 5-7 years to produce fruit) and are self-incompatible, meaning they display obligate outcrossing. These two biological constraints create long timelines for apple improvement via traditional crossing. Given the evidence suggesting that grafting techniques have been commonplace since 300

BCE (Juniper & Mabberley, 2006; Zohary & Hopf, 2000), one may expect that the clonal propagation of a small number of elite varieties would have narrowed diversity among apples during the thousands of years since domestication of the crop. However, there is little to suggest that a major domestication bottleneck occurred during apple domestication and breeding (Cornille et al., 2012), and high levels of diversity still remain in breeding populations today (S. Kumar et al., 2010; Watts et al., 2021). It has been hypothesised that the explanation for this paradox is accounted for by long-standing traditional breeding methods in multiple regions (Cornille et al., 2014). It is widely accepted that for thousands of years apple growers were independently selecting varieties generated through natural pollinations (also known as “chance seedlings”). It is suspected that the outcrossing nature of apple, the geographical distance between breeding groups, and differences in taste preferences between human communities has preserved high levels of diversity within the crop (Cornille et al., 2014). In addition, the hybridization of wild species in these breeding areas is likely to have worked to further preserve genetic diversity.

Modern breeding techniques

Apple variety improvement remains a difficult challenge for modern breeders due to a number of major barriers. Apple trees are large plants, with a long juvenile phase, a highly heterozygous genome, self-incompatibility, and expensive maintenance, evaluation, and tree care. Together these biological constraints result in a lengthy and costly breeding cycle that has largely inhibited the rapid improvement of apple cultivars. For example, modern breeding programs still require about 25 years to produce just 3 new marketable cultivars (Peil et al., 2008). Despite the industry's high annual value, these major barriers remain and the vast majority of apple breeding today is done via traditional breeding, a process that largely resembles the techniques used by the earliest cultivators of the crop. In recent decades novel tools such as Marker Assisted Selection (MAS) have been developed to help decrease the cost of apple breeding programmes (Edge-Garza et al., 2015; Luby & Shaw, 2001; Migicovsky & Myles, 2017). MAS uses genetic markers that are in linkage with quantitative trait loci (QTL), genetic regions associated with plant traits, providing breeders a method to screen plants for desirable genotypes at early developmental stages. While genetic screening via MAS can significantly reduce the required capital to grow potentially winning varieties to maturity, the overall breeding process is essentially the same as those described previously and still requires decades of investment. In addition, the markers used in MAS are often not strong predictors of the traits of interest, leaving significant guesswork and error in the breeding process (Migicovsky et al., 2021). Apple breeding is further complicated by the fact that many alleles controlling important traits, such as disease resistance, are present exclusively in wild germplasm (Kostick, Teh, & Evans, 2021). Genetic sources of these alleles often have many undesirable fruit quality traits, and introgressing wild alleles into elite backgrounds can take multiple decades (Iezzoni et al., 2020). To truly accelerate variety improvement, precise molecular tools that address multiple challenges faced in the breeding process must be discovered, adapted, and applied in apple.

GENETIC MODIFICATION AND GENE EDITING IN AGRICULTURE

Genetic modification in agricultural plants

Molecular biology techniques that enable direct changes to the genome of an organism have provided tremendous progress in agriculture. In plant species, methods that could create precise and permanent changes to the DNA of the genome date back to about the 1980's, upon the discovery and investigation of *Agrobacterium tumefaciens*, the causative bacterial pathogen of crown gall disease (E. Nester et al., 2005; Zaenen et al., 1974; Zambryski et al., 1980). The *A. tumefaciens* transformation system makes use of a unique suite of *Virulence* (*Vir*) genes to move Transfer DNA (T-DNA), encoded on a plasmid, into the target plant cell where the VIR proteins mediate T-DNA integration into the genome (SB Gelvin, 2003). Since its discovery, *A. tumefaciens* has been engineered to transfer precise T-DNA segments into the plant cell for permanent genome integration to produce plants, such as soy and rice, with novel or improved traits (Do et al., 2019; Pompili et al., 2020; Schlathölter et al., 2023; Ye et al., 2000).

Direct engineering of the plant genome was a major step for plant science and agriculture, and has led to the production of numerous transgenic plants that serve as important inputs of the modern agricultural system (ISAAA, 2023). Addition of genetic material to the plant genome via *Agrobacterium*-mediated transformation remains an area of intense focus today, despite some negative public sentiment towards transgenic organisms. Some examples of important agricultural crop varieties generated with transgenic *Agrobacterium* techniques include Golden Rice (Ye et al., 2000), the Arctic® Apple (Stowe & Dhingra, 2020), the Norfolk Purple Tomato® (Butelli et al., 2008), the

PinkGlow® Pineapple (Fabricant, 2022), and most recently Conscious Greens ® (Karlson et al., 2022).

Although it is an extremely useful genome modification tool, *Agrobacterium* mediated transfer has some significant drawbacks. First, only some plant families are susceptible to infection by *Agrobacterium*, meaning that genome modification is challenging or impossible for some crop species using this method (E. W. Nester, 2014; SB Gelvin, 2003). Second, integration of T-DNA into the genome is random, and T-DNA insertions can often cause undesirable consequences if the integration interrupts functional regions of the genome. Finally, in many countries transgenic plants are regarded as Genetically Modified Organisms (GMOs) and are either highly regulated or banned altogether (Jenkins et al., 2023). Taken together, these aspects of *Agrobacterium*-mediated transfer have pushed plant biologists to look towards other tools, such as gene editing, as a less controversial approach for making meaningful agricultural improvements.

Gene Editing and CRISPR technology

Gene editing is capable of making targeted mutations to the genomes of plant species (Bortesi & Fischer, 2015; Doudna & Charpentier, 2014; Jinek et al., 2012) (as well as bacteria and mammals) and represents a novel approach for the improvement of agricultural crops. Gene editing systems have already been used to generate improved traits in numerous crops, including rice, wheat, and tomato (Nekrasov et al., 2017; H. Zhang et al., 2014; Y. Zhang et al., 2018). As the successor of previous editing systems such as transcription activator-like effector nucleases (TALENs) and zinc-finger nucleases (ZFNs), gene editing using the clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated protein (Cas) system (CRISPR/Cas) has risen to popularity due to its simplicity and modularity (Randhawa & Sengar, 2021).

The CRISPR based system achieves gene editing through two basic components: a ribonucleic acid (RNA) and an active enzyme (Jinek et al., 2012). These components are bound together into a ribonucleoprotein complex (RNP), which is the complete functional gene editing complex. The guide RNA (gRNA) in the CRISPR system provides specific DNA site targeting: the easily-programmable gRNA can be modified to match a sequence of the host DNA to guide the complex to the desired genomic location for action. The enzyme in CRISPR systems is most commonly a nuclease (however other enzymes can be substituted), which cleaves DNA at the targeted region (Doudna & Charpentier, 2014). Cleaved DNA is recognized by the cell's repair machinery as a double stranded break (DSB) and typically repaired through one of two pathways (K. Chen et al., 2019). Non-homologous end joining (NHEJ) is the default repair mechanism and usually results in insertions or deletions of several DNA nucleotides. If the DNA cut is targeted to a gene coding region, this repair pathway is useful in generating knockout, or loss-of-function, mutations. This mechanism has been used to generate commercial varieties of waxy corn and high oleic soy (Voytas, 2019; Waltz, 2016), as well as mustard greens lacking bitter compounds (Karlson et al., 2022). The less-frequent repair pathway is homology-directed repair (HDR), whereby DNA is repaired by the cell based on overlapping template sequences of nearby DNA. By leveraging the template-matching mechanism of HDR, it is possible to add desired nucleotide sequences to a targeted region of a genome by supplying the cell with exogenous DNA with homologous flanking sequences (Van Vu et al., 2020), offering powerful opportunities for introducing precise genetic changes.

Perhaps the simplest genetic variation that can be produced by gene editing is small insertions/deletions (indels). Indels were among the first forms of variation introduced using gene editing (Doudna & Charpentier, 2014), and are typically introduced via a double stranded break to the DNA at a specific location in the genome. The cell's endogenous repair mechanisms then repairs the DNA break, typically through NHEJ, reliably introducing small indels at the break site. Indels are useful in generating knock-outs of target genes to produce novel or improved phenotypes (S. Tian et al., 2017; Y. Wang et al., 2014; Zheng et al., 2020). For example, CRISPR gene editing has been used in *Brassica* species to knock out *myrosinase* genes that produce pungent bitter molecules, resulting in a less bitter nutrient-dense leafy green (Karlson et al., 2022). Indels can also be used to generate variation in regulatory elements that control gene expression. Targeted editing of regulatory elements introduces novel cis-regulatory alleles that give rise to gene expression patterns beyond what is observed in natural populations. This approach, often referred to as "promoter bashing", can provide beneficial quantitative variation in gene regulation and expression (Rodríguez-Leal et al., 2017). Already, the promoter bashing approach has been used to alter known promoter sequences in tomato to produce varieties with different fruit sizes and organ numbers (Rodríguez-Leal et al., 2017). The mutation of regulatory and non-coding regions of the genome has been applied to produce multiple phenotypes, including disease resistance, increased yield, and decreased amylose content in rice (C. Li et al., 2022; J. Li, Chen, et al., 2022; X. Song et al., 2022). The introduction of indels using gene editing has already proved useful in generating novel crop varieties, but only represents one of many types of genetic variation that can be generated through gene editing.

Recent advancements in gene editing have discovered alternative enzymes that can be substituted in the CRISPR gene editing system to produce a myriad of specific DNA mutations. Base editing (Komor et al., 2016), prime editing (Anzalone et al., 2019), and twin prime editing (Anzalone et al., 2021) are novel approaches for making specific non-random mutations to the genome. Base editors consist of a modified CRISPR cassette with an alternate enzyme module capable of making specific point mutations. There are two major classes of base editors, cytosine base editors (Komor et al., 2016) and adenosine base editors (Gaudelli et al., 2017), that can collectively mediate all possible transition mutations (C→T, G→A, A→G, and T→C). These powerful DNA editors enable single nucleotide alleles to be specifically altered to desired genotypes, potentially enabling specific alleles to be changed within the genome. Single point mutations in acetolactate synthase genes via cytosine base editing have been generated to create herbicide resistant tomato and potato varieties (Veillet et al., 2019). Base editors represent the first CRISPR-based gene editing technology to enable specific allele changes, but two more technologies have recently built on this advancement. Prime editing and twin prime editing are adaptations on the classic CRISPR gene editing system, and enable a wide range of specific mutations to be introduced into the genome. Prime editing uses a reverse transcriptase enzyme, a DNA nicase enzyme, and an extended prime editing guide RNA (pegRNA) to produce all twelve possible base-to-base point mutations and sequence-specific insertions and deletions up to 44 bp and 80bp, respectively (Anzalone et al., 2019). Twin prime editing is an advanced application of prime editing, whereby two CRISPR-pegRNA constructs are used in tandem to reverse transcribe complementary single stranded DNA sequences, which then anneal and replace endogenous DNA (Anzalone et al., 2021). Twin prime editing can achieve large insertions (up to 5000 bp) and inversions (up to 40kb) in human cells. While significant ground has been made in optimising prime editing (Zong et al., 2022), twin prime editing has yet to be successfully applied to plants. Genome modification through CRISPR-based approaches is a rapidly evolving field, with advances being reported frequently in the literature (P. J. Chen & Liu, 2022; C. Sun et al., 2023). The full CRISPR gene editing suite of technologies offers the ability to generate a wide array of genetic variation that offer the potential to revolutionise agriculture.

Gene editing technologies are in many cases limited by the presence of a sequence specific motif, known as the protospacer adjacent motif (PAM), that is required for CRISPR/Cas enzymatic action. The PAM is a short nucleotide NGG sequence that must be present at the 3’ end of the gene editing target sequence in order for the Cas9 enzyme to cleave DNA (Jinek et al., 2012). Therefore, if a desired gene editing target does not have an NGG sequence in sufficient proximity, then the edit is often impossible to create (Sander & Joung, 2014). This requirement, until recently, produced a significant constraint on the types of DNA sequences that could be targeted by gene editing. However, new developments have largely alleviated this constraint through the discovery of Cas family proteins with fewer, or in some cases no, PAM requirements (Endo et al., 2019; Q. Ren et al., 2021). Given the progress that has been made in a few short years since the 2012 discovery of gene editing, there is strong potential for the discovery of enzymes that will entirely alleviate the constraints of PAM sequences in the future.

Since its discovery in 2012, CRISPR systems have been applied to more than 100 crops and model plants (Cardi et al., 2023). The field of gene editing is a rapidly evolving field, with modifications continually increasing the specificity, capabilities, and applicability of the CRISPR editing system. In the near future, gene editing will likely be a breeding tool used across almost all agricultural crops, including apples.

Genetic modification and gene editing in apple

Apple is among the increasingly long list of perennial fruit species, including citrus, grape, cacao, and kiwi (Fister et al., 2018; H. Jia et al., 2017; Pompili et al., 2020; C. Ren et al., 2016; Varkonyi-Gasic et al., 2019), in which gene editing systems have, to one degree or another, been enabled. Gene editing systems in apple offer potential solutions to the long breeding cycle and probabilistic nature of traditional apple improvement. Apple improvement via gene editing could, in theory, be used to make precise alterations to the genome of the apple to produce novel improved varieties without the need to cross two samples, which requires random genetic recombination and the growth and maintenance of hundreds or thousands of progeny. Further, altering known causal alleles in apple via gene editing would enable the production of novel varieties with higher certainty of improving a desired trait, or to introduce commercially viable heritage or wild varieties by making precise allele switches. So far, CRISPR knock-out varieties have been generated in apples, successfully producing early flowering, albino, and disease resistant phenotypes in plantlettes (Charrier et al., 2019;

Dalla Costa et al., 2020; Malnoy et al., 2016; Nishitani et al., 2016; Pompili et al., 2020). Indeed, CRISPR genome editing holds tremendous promise for accelerated apple improvement, however this technology is still in its early stages in apple and faces significant challenges in both delivery and target discovery.

A significant challenge for apple gene editing is the successful regeneration of edited apple tissue. Regardless of which transformation technique is used to create gene-edited apple varieties, entire plants must be regenerated from a small number of cells with the desired genetic alteration (Dalla Costa et al., 2020). To achieve this, apple tissues are cultured on media to enable gene editing, and then must be promoted to regenerate into full plants (Malnoy et al., 2016; Nishitani et al., 2016). Unfortunately, there are dramatic differences across apple varieties in the ability to regenerate full plants from edited tissue (Magyar-Tábori et al., 2010). This challenge is evidenced by the fact that nearly all of the successful gene editing experiments in apple have been in the variety 'Gala' (Dalla Costa et al., 2020; Malabarba et al., 2020; Pompili et al., 2020), which appears to be the variety most receptive to tissue regeneration techniques. Unfortunately, the biological explanation for this phenomenon is not understood, and is cited by some as "the biggest bottleneck in plant genome engineering" (Laforest & Nadakuduti, 2022). For apple variety improvement, this poses a significant challenge as the editing of only a few apple varieties offers limited potential for apple improvement as a whole. Further, it is documented that tissue culture protocols frequently result in unintended mutagenesis or changes to the epigenetic landscape (Phillips et al., 1994; D. Zhang et al., 2014), although it is unclear whether this is the case for perennial species. Taken together, there are significant challenges in tissue culture and regeneration that must be overcome in order to unlock the full potential of gene editing in apple.

Gene editing target discovery is another significant barrier to gene editing in apple. The prerequisite to gene editing a crop for a targeted phenotype improvement is the discovery of the causal genetic variant or genomic region controlling that phenotype. To efficiently and strategically enable gene editing, a target allele must first be identified before it can be targeted. This is arguably the most difficult step of the gene editing process, and has frequently been cited as the rate limiting step in crop improvement (Weigel & Nordborg, 2005). Knowledge of the precise identity of the causal allele is required to program the gRNA component of the gene editing system (Jinek et al., 2012), and without this knowledge the technology cannot be effectively applied. Without the discovery of causal genetic variants at near-single nucleotide resolution, gene editing technology cannot be used in a meaningful way to improve apple varieties. It is clear that efforts leveraging whole genome sequencing, statistical mapping approaches, and bioinformatics will be key to discovering gene editing targets necessary for accelerated apple improvement in the future.

GENETIC MAPPING IN APPLE

The importance of genetic mapping and causal allele discovery in apple

Substantial efforts in agricultural genomics are aimed at connecting phenotypes and genotypes. The experimental process of discovering and characterising DNA segments that control a particular trait, also known as causal alleles, is often referred to as genetic mapping. The discovery of causal alleles is crucial for both genomics informed traditional breeding methods (such as MAS) as well as new plant breeding techniques (such as gene editing) (Babu et al., 2004; Jinek et al., 2012). In the case of MAS, causal alleles are by definition the best possible genetic markers for screening progeny. For gene editing purposes, causal allele discovery and characterization is a prerequisite for enabling the technology to be used beyond gene knockouts, which have limited scope for the improvement of plant traits. Therefore, a gap in causal allele discovery leaves only traditional breeding techniques and limited gene editing applications available to those tasked with improving apples. In fact, the "chasm" between genomics and breeders in Rosaceous crops was deemed so important that a multi-institution international project, RosBreed, received $17.7 million in funding from the USDA between 2009 and 2019 with an express objective to characterise causal alleles which were coined as "jewels of the genome" (Iezzoni et al., 2020). Without genetic mapping and causal allele discovery, scientific and industrial progress in apple improvement is likely to remain slow.

Genetic mapping in agriculture

The principle of mapping a quantitative trait locus (QTL), a genetic locus which correlates with variation of a quantitative trait, in plants was first introduced in the 1920's in bean (Sax, 1923). However, QTL mapping was not a major focus in agricultural sciences until the 1980's, after the discovery of polymerase chain reaction (PCR)-based molecular markers, development of appropriate statistical tests, and the advent of accessible computer software (Ahmad & Anjum, 2018; Bernardo, 2008; Paterson et al., 1988). In the decades following, tens of thousands of QTLs have been reported in agricultural plants (Kumawat et al., 2016; Malik et al., 2014; Miura et al., 2011; Peace et al., 2019; Wisser et al., 2006). Today, there are two primary approaches used for genetic mapping in agricultural crops: linkage mapping in parental cross populations and association studies in diversity panels. Both methods have strengths and limitations, and both are used by groups aimed at discovering causal alleles in important agricultural crops.

Linkage Mapping

Parental cross populations are commonly used to map genotype-phenotype relationships via a method called “linkage mapping”. Parental cross (often called bi-parental cross) populations are generated by crossing two parent varieties and then assessing segregating phenotypes and genotypes in the resulting offspring. Linkage mapping is an approach that associates genomic regions with phenotypes, ultimately connecting QTLs with traits. The primary strength of linkage mapping is the ability to detect rare alleles, which are often commercially valuable (Carbonell-Bejerano et al., 2019; Chagné et al., 2007; Conner et al., 1998). Because parental cross mapping populations are generated from only two individual plants, rare alleles carried by those parents will segregate in the progeny and can be detected. However, because these populations often take time to establish and consist of a small number of generations, particularly in perennial species, there are typically few recombination events within the population leading to large blocks of genetic linkage disequilibrium (LD) (Myles et al., 2009). As a consequence, linkage mapping experiments frequently result in the discovery of QTLs that span hundreds of thousands of nucleotides of DNA sequence and contain dozens or even hundreds of genes (S. A. Khan, Chibon, et al., 2012; Kostick, Teh, Norelli, et al., 2021). With large regions of the genome associated with phenotypes, confidently determining the causal allele within the candidate QTL becomes difficult or impossible. From a breeding standpoint, large QTLs derived from linkage mapping can be impractical, as transferring multiple QTLs across genotypes could require the generation of millions of crosses and is likely to result in unwanted genetic drag (Bernardo, 2008). Further, QTLs for a given trait are often specific to the population in which they were derived, which makes them useful in some mapping populations, but renders them ineffective across diverse populations (Sorkheh et al., 2008). An additional limitation of linkage studies is that they are limited to querying only phenotypes and QTLs that segregate within the parents of the population, which restricts the number of phenotypes that can be mapped with a given parental cross population. This is particularly troublesome when working with perennial species, as mapping populations can take decades to establish (Peace et al., 2019). While linkage mapping is frequently sufficient for deriving relationships between phenotypes and large genetic regions, this strategy rarely provides the genomic resolution required for discovering effective breeding markers or gene editing targets within *M. domestica*.

Genome-wide association studies (GWAS) in agriculture

Given the drawbacks of linkage mapping approaches, genome wide association studies (GWAS) are becoming increasingly popular in agricultural genomics. GWAS use statistical models to associate genetic markers from across the genome with a phenotype of interest in populations of unrelated samples (Schaid et al., 2018). GWAS are a powerful approach for identifying causal alleles in crop species, but an understanding of the assumptions and limitations of the GWAS method is important.

To generate statistical mapping power, GWAS must be conducted using populations of largely unrelated samples. The use of such populations captures a large number of genetic recombination events, leading to higher genomic resolution than can be achieved in linkage mapping studies (M. A. Khan & Korban, 2012; Myles et al., 2009). This is because decreased relatedness and increased recombination in a population typically results in decreased LD, the non-random association of alleles at different loci. When two or more alleles are in close physical proximity they are more likely to be inherited together (or "in LD"), and thus associate, which can contribute to suboptimal genomic resolution. However, by capturing historic sexual recombination across unrelated individuals over a large number of generations, large haplotypes can be broken by recombination, and LD decay can be rapid (S. Kumar et al., 2014; Myles et al., 2010; Remington et al., 2001; Tenaillon et al., 2001). GWAS approaches in diverse populations exploit rapid LD decay to produce high mapping resolution. When LD decay in the mapping population is extremely rapid, genotype-phenotype associations from GWAS often lead to associations of variants that are causal or within <100 nucleotides from the causal variants (Liao et al., 2021). However, because GWAS leverages historical recombination and LD decay, the resolution of association studies also relies upon the generation of a high density of genetic markers across the genome, which has historically been a major technological challenge.

There are numerous statistical models that can be used to perform a GWAS. Of the many models, the most common are Mixed Linear Models (MLM) or Multi-Locus Mixed Models (MLMM) (Tibbs Cortes et al., 2021; Zhou & Stephens, 2012). MLM approaches were developed to account for relatedness within mapping populations, as traits that are correlated with population structure often contribute to high false positive rates (Yu et al., 2006). MLMs therefore include a population structure matrix (Q), which accounts for broad level relatedness among samples, and a sample structure matrix (K) that accounts for fine-scale relatedness among samples. MLMMs are similar to MLMs, except that the former is capable of considering multiple markers as covariates in the model, making it the appropriate model for mapping traits with genetic architectures controlled by multiple loci (Tibbs Cortes et al., 2021). There are a number of other genetic mapping techniques akin to the GWAS approach, including pool-seq (Kofler et al., 2011) and k-mer based designs (Voichek & Weigel, 2020), however MLM and MLMMs remain the current dominant GWAS mapping methods.

The primary limitation of GWAS is the inability to detect genotype-phenotype associations when the causal allele is rare or when the phenotype is highly correlated with population structure (Myles et al., 2009; Schaid et al., 2018). Even when appropriate measures are taken to account for relatedness (use of Q and K matrices), traits strongly linked with population structure often cannot be reliably mapped. The collection of highly diverse germplasm is another significant limitation of the GWAS method, as successful sourcing, planting, and maintenance of highly diverse germplasm produces significant logistical and budget challenges. However, the high genetic resolution achieved through GWAS approaches makes it a powerful method for pinpointing causal alleles that can serve as future genetic markers and gene editing targets.

Recent advancements in genetic mapping in apple

The primary focus of most genetic mapping in apple over the last several decades has been fruit quality improvement and disease resistance. Firmness, storability, colour, acidity, sugar content (Brix), along with fire blight and powdery mildew resistance have been traits of great importance to breeders and growers and have therefore been the primary traits of investigation in mapping populations (Chagné et al., 2007, 2019; Chagné, Krieger, et al., 2012; Ding et al., 2022; S. Kumar et al., 2022; McClure et al., 2018, 2019; Watts et al., 2021; B. Wu et al., 2020).

Hundreds of genetic mapping experiments in apple have documented QTLs, and some causal alleles, in the literature. Since causal alleles require more data to discover, the number of published QTLs far outnumber published discoveries of causal alleles in apple. Perhaps the best documented causal allele in apple lies in the *Ma1* gene: a SNP in the coding region of a malate channel protein controls variation in malic acid content in fruit (Bai et al., 2012). While linkage mapping experiments first described QTL that contained the causal allele in the late 1990s (Maliepaard et al., 1998), the mutation was not characterised at the nucleotide level until 2012 (Bai et al., 2012). Another instance of causal allele discovery in apple comes from a biparental cross population between the domestic apple and a wild relative, *M. robusta* (Peil et al., 2007). This population was used to map a QTL containing a fire blight resistance gene, *FIRE BLIGHT RESISTANCE MALUS ROBUSTA 5* (*FBMR5)*, which has since been cloned, transformed into an elite *M. domestica* genetic background (Fahrentrapp et al., 2013), and tested in the field to confirm the alleles function (Schlathölter et al., 2023). The causal allele impacting variation in apple skin colour, a transposable element in the promoter region of a MYB transcription factor, has also been successfully mapped and characterised in recent years (Y. Ban et al., 2007; Espley et al., 2007; Takos et al., 2006; L. Zhang et al., 2019). While numerous groups have identified the genes that contribute to various traits (X. Jia et al., 2022; Migicovsky et al., 2021; Y. Wang et al., 2023; X. Zhang et al., 2022), the precise alleles controlling variation in those traits still have not been discovered and malic acid, skin colour, and fire blight resistance represent some of the only discovered causal alleles in apple to date.

Although less than a dozen causal alleles have been discovered in apple, significant progress has been made in narrowing the genetic location of causal alleles for many important traits. Fruit firmness is an important consumer trait in apple and has historically been a breeding target in apple. Linkage mapping experiments have suggested that variation in *polygalacturonase-1* (*PG1*), a gene related to pectin degradation in the cell wall, is likely responsible for variation in fruit firmness (Chagné et al., 2019; S. Kumar et al., 2013). However, genetic mapping studies in different populations have produced conflicting results and some have proposed ERF and *ACO* genes as putatively causal (Di Guardo et al., 2017; McClure et al., 2018). Currently, the causal allele impacting fruit firmness variation remains to be identified from an approximately 1 Mb region on chromosome 10. Phenolic content is another important trait in apple, as phenolic compounds contribute to both the nutritional value and the bitterness of apple fruits. Linkage mapping studies investigating phenolic content production have suggested a QTL on chromosome 16 as putatively causal (S. A. Khan, Schaart, et al., 2012; McClure et al., 2018). While some research suggests that the *leucoanthocyanidin reductase* (*LAR1*) gene could harbour the causal allele, a number of other genes in close proximity to *LAR1* are also strong candidates and the causal allele has yet to be confidently resolved on chromosome 16 (S. A. Khan, Schaart, et al., 2012; McClure et al., 2019). Significant progress has also made in understanding the alleles that control the self-compatibility locus in apple, which is ultimately determined by alleles in RNA-coding regions (Sheik et al. 2020). Ripening time, an important trait for growers, has also been the focus of multiple genetic mapping studies that have been unable to discover the causal allele. Although multiple research teams have converged on a signal close to the transcription factor *NAC18.1*, the causal allele has yet to be characterised (M. Jung et al., 2020; Larsen et al., 2019; Migicovsky et al., 2021). In each case, genetic markers for these traits have been generated and are used for breeding programmes, despite the fact that the markers are not the causal allele for the given traits (Migicovsky et al., 2021). While markers correlated with causal alleles still hold some breeding value, greater apple improvement can be made by replacing low confidence markers with causal alleles.

Due to the major logistic and technological barriers to establishing genetically diverse mapping populations, genetic mapping in apple has largely been done in biparental cross populations using linkage mapping techniques (Conner et al., 1998; Di Guardo et al., 2017; Peace et al., 2019). Because linkage mapping experiments often yield results that are impractical for apple breeding, apple breeding still lags behind other crops despite the often stated benefits of genomics informed breeding for long lived woody fruit species. In response, multiple groups have mounted efforts to generate genetic diversity panels in the hopes of leveraging genetic recombination for mapping experiments (Johnson et al., 1999; S. Kumar et al., 2010; Watts et al., 2021).

Table 1. List of causal alleles discovered in apple.

Phenotype	Chr	Associated gene (or suspected)	Variant Type	Reference
Malic acid content	16	*Ma1*	SNP	Bai et al. 2012
Fire blight resistance	3	*FBMR5*	Full gene	Peil et al. 2007
Red skin colour intensity	9	*MYB10*	Indel	Espley et al. 2007 Y. Ban et al. 2007
Self-compatibility	17	*S-RNAse*	PAV	Sheik et al. 2020

Challenges in genetic mapping in apple

Genetic mapping in apples faces a number of barriers that make the discovery of causal alleles a persistent challenge. First, establishing a highly diverse germplasm of sufficient size (hundreds or thousands of samples) for powerful association mapping is expensive and laborious. Apple cuttings need to be sourced and transported from around the world, grafted to uniform rootstock, planted with replicates and controls, and maintained over long periods of time in order to conduct genetic mapping. Efforts of this magnitude are rare, but have been achieved by some groups (Gross et al., 2013; M. Jung et al., 2020; S. Kumar et al., 2010; Migicovsky et al., 2022). Second, genetic mapping studies in apple have long been plagued by a limited number of DNA markers. The generation of high density marker datasets has been, and to a large degree remains, prohibitively expensive. Without dense marker datasets, GWAS lack power and resolution and causal genomic regions go undiscovered even if other aspects of the experimental design are strong. Finally, recent studies have demonstrated that plant genomes are extraordinarily diverse, containing thousands of structural variants and large repetitive regions (Saxena et al., 2014), which creates serious genotyping and bioinformatic challenges. The apple genome is no different: it is highly diverse and heterozygous, and has recently undergone a whole genome duplication (Bianco et al., 2016; X. Sun, Jiao, et al., 2020; Velasco, Zharkikh, Affourtit, Dhingra, Cestaro, et al., 2010). These features of the apple genome make accurately capturing and quantifying genetic variation difficult, and doing so remains a persistent challenge in genetic mapping in apple. In the future, it will be crucial to address each of these challenges in apple breeding and genomics.

MODERN GENOMICS AND DIVERSITY PANELS FOR IMPROVED GENETIC MAPPING IN APPLE

Advancements in DNA sequencing technologies

DNA sequencing data is central to the field of genomics. In recent decades, DNA sequencing technologies have experienced significant advancements, revolutionising genomics and enabling researchers to explore the genomes of organisms with unprecedented resolution and throughput (J. Chen et al., 2023; Nurk et al., 2022). The most important improvements in DNA sequencing technology have been the dramatic decrease in sequencing costs, the availability of multiple sequencing types, and the development of multiplexed library preparations (Jain et al., 2016; Wong et al., 2013).

Perhaps the most important development in DNA sequencing technology since the turn of the millennium has been the substantial reduction in sequencing costs. The advent of high-throughput sequencing platforms, commonly known as next-generation sequencing (NGS) technologies, has enabled researchers to sequence DNA at a fraction of the cost compared to traditional Sanger sequencing methods. This cost reduction has largely democratised genomic research, making large-scale sequencing projects more feasible and accessible to a broader scientific community (Mardis, 2017; Shendure et al., 2017).

The field of genomics has also witnessed a rapid expansion in the range of available DNA sequencing approaches, offering researchers various options to generate genetic data. Short-read sequencing, which involves sequencing short DNA fragments in parallel, has been widely adopted due to its high throughput and cost-effectiveness (Metzker, 2010; Shendure et al., 2017). Chemical advancements allowing for paired-end (PE) reads for short-read sequencing have further improved the reliability, quality, and alignment of short-read DNA sequencing. In addition to short-read sequencing, the emergence of long-read sequencing technologies has addressed the limitations associated with short reads, such as difficulties in resolving repetitive regions and structural variation (J. Chen et al., 2023; Shi et al., 2023). Technologies such as Pacific Biosciences' HiFi sequencing and Oxford Nanopore Technologies' nanopore sequencing have enabled the generation of long DNA sequencing reads, offering improved genome assembly, haplotype phasing, gap reduction, and the ability to detect complex genomic rearrangements in plants (Jain et al., 2016; Yue et al., 2023). Recently, long-read technologies have enabled a telomere-to-telomere assembly of the corn genome, in which 5 of 10 chromosomes were covered by a single contig (J. Chen et al., 2023). Long-read technologies provide a significant step forward in sequencing technologies and genome assembly. Further, these single-molecule based approaches do not necessarily require template amplification, meaning they are not prone to bias introduced through copy errors and sequence-dependent bias (Shendure et al., 2017).

To maximise the throughput and cost-effectiveness of DNA sequencing, the development of multiplexed library preparations has played a pivotal role, particularly for experimental designs requiring short-read sequencing. Multiplexing allows multiple DNA samples to be sequenced in a single sequencing run, using unique DNA barcodes (index sequences) to demultiplex the resulting reads downstream (Wong et al., 2013). This approach has significantly increased the efficiency and scalability of DNA sequencing experiments, enabling researchers to simultaneously analyse numerous samples and achieve higher sample throughput with reduced budgets (Kircher & Kelso, 2010).

Improvements in the availability and quality of DNA sequencing data has had a dramatic impact on agriculture, particularly in the area of genetic mapping. Accessible DNA data has enabled advancements in the production of DNA breeding markers, reference genomes, genetic modification, and gene editing, all of which have contributed significantly to accelerated plant breeding and scientific understanding of plant biology over the past 20 years. As of 2023, reference genomes (often multiple) exist for dozens of agricultural crops like rice, corn, soy, wheat, peach, kiwi, grape, sweet basil, garlic, black pepper, and vanilla, (J. Chen et al., 2023; Gonda et al., 2020; Hasing et al., 2020; L. Hu et al., 2019; Y. Liu et al., 2020; Shi et al., 2023; X. Sun, Zhu, et al., 2020; Yue et al., 2023; A. Zhang et al., 2021; H. Zhang et al., 2022). Continuously expanding databases, such as Phytozome (Goodstein et al., 2012) and MaizeGDB (Woodhouse et al., 2021), store publicly available DNA sequences that further enable affordable crop genomics research, and play an important role in genomics research. The ubiquity of DNA data in modern crop research, particularly in non-model or orphan crops (Islam et al., 2022; Mansfeld et al., 2021), is a strong indicator of the value and utility of DNA sequencing.

In terms of causal allele discovery in apple, perhaps the most important impact of DNA sequencing availability is the generation of high density genetic markers. Until about 2015, SNP arrays containing 20k genetic markers were the primary source of genotype data for genetic mapping experiments in apple (Laurens et al., 2018), largely due to their affordability. Although SNP arrays were developed with the goal of gaining sufficient coverage to conduct powerful mapping studies, even the largest apple SNP arrays produced an average marker density of one marker per 1.4 kb (Bianco et al., 2016). This may have been a considerable improvement on previous satellite marker or SNP arrays of the past, but the 480k SNP array still does not provide the resolution required to identify causal alleles, considering the extremely rapid LD decay in apple (S. Kumar et al., 2014; Leforestier et al., 2015; McClure et al., 2018). The use of modern DNA sequencing technologies, however, can nearly exhaustively catalogue variants across entire genomes. For example, a recent study in peach used next-generation DNA sequencing to generate a marker density of one marker every 66 bp (Tan et al., 2021). Affordable whole genome DNA sequencing is key to unlocking the high resolution marker datasets that enable causal allele discovery in apple.

Advancements in genetic resources for apple

As discussed previously, DNA sequencing data holds tremendous value for understanding and improving crops, and this fact remains true for apple. However, to maximise the value and insight of DNA sequencing data, it is important to have access to other genomic resources for a target organism. Fortunately, important genomic resources have been publicly published for apple in recent years, including multiple reference genomes and a large -omics database.

Arguably the most important genetic resource for genomics studies is the reference genome, which is often the first milestone in enabling genomics-focussed research in a crop. The first apple genome, that of the ‘Golden Delicious’ apple variety, was published in 2010 (Velasco, Zharkikh, Affourtit, Dhingra, Viola, et al., 2010) and since then multiple reference genome builds have been constructed, including the ‘Honeycrisp’ genome and an apple pan-genome (A. Khan et al., 2022; Peace et al., 2019; Sun et al., 2020). Publication of the reference genome led to advancements in apple genomics including single nucleotide polymorphism (SNP) chip development (Chagné, Crowhurst, et al., 2012; Chagné et al., 2019), genotyping-by-sequencing datasets (GBS) (Larsen et al., 2018; Migicovsky et al., 2022), and ultimately enabled higher-resolution genome wide association studies (GWAS) (Di Guardo et al., 2017; Larsen et al., 2018; Noh et al., 2020). The combined application of these genomic resources has also led to the implementation of MAS in apple breeding programmes (Baumgartner et al., 2016; Chagné et al., 2019; Jänsch et al., 2015; S. A. Khan, Chibon, et al., 2012).

The Genome Database for Rosacea (GDR) is another key genetic resource for apple (S. Jung et al., 2019). The GDR provides access to reference genomes, raw sequence data, genetic markers, genome annotation files, various Basic Local Alignment Search Tools (BLAST), and genome browsing interfaces. The GDR not only organises and stores freely available apple data, but acts as a connective hub for various tools that aim to maximise the value and impact of biological data collected by individual groups. Databases such as GDR are important for advancing research in apple and are an excellent example of the benefits of a collaborative spirit of scientific inquiry.

Canada’s Apple Biodiversity Collection (ABC)

Canada’s Apple Biodiversity Collection (ABC) is among the world's most diverse apple germplasm collections, and is an invaluable source of genetic information for genetic mapping and variety improvement in apple. Located at the Agriculture and Agri Food Canada (AAFC) Kentville Research Station, in Nova Scotia, Canada, (45.071767, −64.480466), the orchard contains 1,119 apple varieties, which includes 78 unique accessions from the wild progenitor species, *M. sieversii*. The ABC contains apple varieties sourced from the United States Department of Agriculture (USDA) Plant Genetic Resources Unit apple germplasm collection (Geneva, New York, USA), the Nova Scotia Fruit Growers’ Association Cultivar Evaluation Trial, and breeding material from the AAFC Kentville breeding program. The orchard is planted in a randomised block design, with each accession grafted onto M.9 rootstock and planted in duplicate.

The immense genetic diversity combined with the design of the orchard makes the ABC among the most powerful genetic mapping populations ever established in apple.

The ABC has already been the focus of phenomics and genomics research in recent years. Historically, the collection of high quality phenotype data from large populations has been a major challenge for perennial crop species, including apple, resulting in what has been cited as the “phenotyping bottleneck” (Furbank & Tester, 2011), which limits the ability to conduct high powered experiments (Burleigh et al., 2013). Importantly, a large phenomics effort was recently undertaken within the ABC to address the phenotyping bottleneck (Watts et al., 2021). Tremendous amounts of human labour, aided by advances in phenotyping equipment such as penetrometers and cold storage facilities, have resulted in a comprehensive phenomic characterization of the ABC (Watts et al., 2021). This wealth of data is not only essential for genetic mapping, but also for advancing our understanding of phenotypic variation of apple, historical improvement of apple, and relationships between wild and domesticated apples.

Given the unique challenges in genetically mapping traits in apple (outlined above), research leveraging the ABC holds great potential to generate impactful insights into apple biology, domestication and accelerated variety improvement. The ABC is the source germplasm for the research contained within this thesis, and is used to advance apple improvement, agricultural genomics, and plant biology more broadly.

CONCLUDING REMARKS

At the time of this writing, gene editing technology appears to offer the greatest potential for rapid apple variety improvement and numerous advancements in the field of gene editing have made its application in apple a real possibility in the near future. Therefore, the overall goal of this thesis is to move the current state of knowledge closer towards enabling gene editing in apple. Of the major challenges that make gene editing difficult in apple, developing a deeper understanding of apple phenotypes and discovering the causal allele(s) that control them are arguably the most obstructive. The ABC is uniquely suited to address these challenges. The present thesis presents a series of studies aimed at quantifying multiple apple phenotypes and revealing the identity of causal alleles that control them. Causal alleles discovered in this thesis will serve as gene editing targets once other important discoveries are made in plant science. Other research groups will make progress in areas of apple tissue culture, plant regeneration, and gene editing protocols, each of which represents an independent challenge that must be addressed to enable effective gene editing in apple. In the near future, causal alleles from this work will serve as gene editing targets that can be integrated with discoveries from other groups to provide breeders with gene editing capabilities that hold potential to dramatically accelerate the speed at which apple varieties can be improved and improve apple agriculture as a whole.

RESEARCH OBJECTIVES AND HYPOTHESES

Project overview

A deeper understanding of apple phenotypes and the genetics controlling those phenotypes is crucial for timely and meaningful variety improvement in the near future. In chapter 2, I conduct an experiment to understand how the apple has changed since domestication, and where the apple phenome is likely to progress in the future. Using comprehensive phenomic data, I analyse the trait differences between the domesticated apple and its primary wild progenitor, and observe recent trends in apple phenotypes over a 200 year period of breeding. In chapter 3, I leverage whole genome sequencing (WGS) across pooled DNA samples to scan the genome for signals of genetic differentiation between groups of apples that show dramatically different ripening times, phenolic content production, and softening during storage. Following chapter 3, it becomes clear that whole genome sequencing across many samples in the ABC is likely necessary to determine the causal allele(s) controlling plant traits. Therefore, in chapter 4 I conduct a genetic mapping experiment for ripening time by sequencing the genomes of 97 diverse apple varieties to high depth and conducting a GWAS, while setting the stage for future mapping experiments that can leverage advancements in low pass sequencing and imputation. In chapter 5, I conclude with a discussion of the overall results from this thesis, my scientific opinions and learnings during the project, and outline future directions for genetic mapping and improvement in apple. Overall, my thesis describes phenotypic diversity and change in apple, improves our understanding of the apple genome, lays the groundwork for large-scale mapping experiments in the future, and moves the scientific state of knowledge closer to enabling gene editing for apple improvement.

Thesis objective

The present thesis aims to improve the current understanding of apple phenotypes and the genetic mechanisms that control them.

Research objectives

Chapter 2

The objective of chapter 2 is to quantify the differences in ten phenotypes between the domesticated apple (*M. domestica*) and its primary wild progenitor (*M. sieversii*), and evaluate how the apple has changed during recent apple improvement.

Hypotheses:

- Phenology traits will differ between domesticated and wild progenitor apple species.
- Domesticated apples will be heavier, less acidic, and less phenolic.
- Phenotypic effects of apple domestication over the last 200 years will be detectable: apple cultivars released in more recent years will be less phenolic, have higher soluble solid content, and retain firmness in storage better than older cultivars.

Chapter 3

The objective of chapter 3 is to discover the causal allele(s) responsible for ripening, softening, and phenolic content in apple (*M. domestica*) using a pool-sequencing approach.

Hypotheses:

- The causal allele for ripening will be within the coding region of NAC18.1 on chromosome 3.
- The causal allele for softening (percent firmness lost during storage) will be within the coding region of ERF (MDP0000855671) located on chromosome 10.
- The causal allele for phenolic content will be within the coding region of LAR1 on chromosome 16.

Chapter 4

The objective of chapter 4 is to discover and characterise the causal allele(s) responsible for ripening time in apple by leveraging whole genome sequencing from a diverse reference panel.

Hypothesis:

- The causal allele for ripening time in apple will be in the region immediately upstream of *NAC18.1* on chromosome 3.

Chapter 2: Phenotypic divergence between the wild and cultivated apple

The contents of this chapter are published as:

Davies, T., Watts, S., McClure, K., Migicovsky, Z., & Myles, S. (2022). Phenotypic divergence between the cultivated apple (*Malus domestica*) and its primary wild progenitor (*Malus sieversii*). PLoS One, 17(3), e0250751. https://doi.org/10.1371/journal.pone.0250751

Rationale

If the domesticated apple is to experience meaningful phenotype improvement in the future, it is important to quantify and understand apple phenotypic change over time. Here, I quantify the phenotypic relationship between wild and domesticated apple species and the phenotype changes that have occurred in the domesticated apple over time. This analysis will demonstrate where the phenome of the apple has been historically and what trends have likely shaped the apple phenotype in the recent past. Further, it provides insight into the potential of wild apple germplasm as a source of valuable alleles for agricultural improvement. Perhaps most importantly, a deeper understanding of apple phenotypes helps inform researchers and breeders about which phenotypes could potentially be altered in the future via breeding or gene editing. Overall, the first chapter of this thesis aimed to accurately quantify the phenotypic changes that have occurred since apple domestication and during recent centuries of breeding.

Abstract

An understanding of the relationship between the cultivated apple (*Malus domestica*) and its primary wild progenitor species (*M. sieversii*) not only provides an understanding of how apples have been improved in the past, but may be useful for apple improvement in the future. We measured 10 phenotypes in over 1000 unique apple accessions belonging to *M.* domestica and *M. sieversii* from Canada's Apple Biodiversity Collection. Using principal components analysis (PCA), we determined that *M. domestica* and *M. sieversii* differ significantly in phenotypic space and are nearly completely distinguishable as two separate groups. We found that *M. domestica* had a shorter juvenile phase than *M. sieversii* and that cultivated trees produced flowers and ripe fruit later than their wild progenitors. Cultivated apples were also 3.6 times heavier, 43% less acidic, and had 68% less phenolic content than wild apples. Using historical records, we found that apple breeding over the past 200 years has resulted in a trend towards apples that have higher soluble solids, are less bitter, and soften less during storage. Our results quantify the significant changes in phenotype that have taken place since apple domestication, and provide evidence that apple breeding has led to continued phenotypic divergence of the cultivated apple from its wild progenitor species.

Introduction

A detailed understanding of crop wild relatives (CRWs) is crucial for the future of agricultural production and sustainability. CRWs represent unique genetic pools that could provide important traits such as disease and pest resistance and drought tolerance to agriculturally important crop species (Cowan et al., 2020; Seiler et al., 2017). As the world's population continues to grow and the effects of climate change become more pronounced, crops will face new and more severe challenges (Lesk et al., 2016; Luck et al., 2011). Wild relatives of crops offer a rich source of genetic diversity that can be used to develop new varieties that are better adapted to changing environmental conditions (Brozynska et al., 2016). By conserving and utilizing the genetic resources found among CRWs, food production systems can be built to be more resilient, sustainable, and able to meet the demands of a growing global population.

CRWs of apple hold significant value for the future of apple agriculture. Wild relatives of the apple possess a wealth of genetic diversity, including traits such as disease resistance, nutritional content, and tolerance to environmental stress (Kost et al., 2015; Smanalieva et al., 2020; Volk et al., 2015). By incorporating the genes from these wild relatives into apple varieties, breeders can develop apples that are better adapted to changing climates and more resistant to pests and diseases (Migicovsky & Myles, 2017). Additionally, wild apple relatives can be used to conserve the genetic diversity of apples, which is essential for maintaining the long-term health and sustainability of the apple industry. Further, a thorough understanding of apple wild relatives could provide insights into the evolution of the domesticated apple and the patterns of artificial selection over the last few millennia. Finally, a comprehensive comparison of the domesticated apple and its wild ancestors could reveal valuable information about specific phenotype structure and trait heredity, which would be informative for downstream genetic mapping efforts. By characterizing the wild relative species of the apple, a more accurate vision of both what apples have been historically and what apples could be in the future can be formed.

The domesticated apple (*Malus domestica*) belongs to the genus *Malus*, which consists of 30-55 interfertile species that grow primarily in temperate climates. Archaeological evidence suggests that apples have been cultivated for at least 3,000 years (Zohary & Hopf, 2000) and that they have had immense cultural, religious, culinary and economic importance for centuries (Cornille et al., 2014; Ferree & Warrington, 2003; Juniper & Mabberley, 2006). Genomic evidence suggests that as apples were transported west into Europe along the Silk Road from Central Asia, hybridization and introgression from multiple *Malus* species created the modern cultivated apple (*M. domestica*) (Cornille et al., 2014; Duan et al., 2017). While there has been introgression from multiple species, including *Malus sylvestris* and *Malus baccata*, to the *M. domestica* genome, *Malus sieversii* of Kazakhstan is widely recognized as the primary ancestor of the cultivated apple (Duan et al., 2017; X. Sun, Jiao, et al., 2020; Velasco, Zharkikh, Affourtit, Dhingra, Viola, et al., 2010).

Today, the cultivated apple is the 3rd most produced fruit crop in the world (FAOSTAT, 2020). Accordingly, apple fruit quality and phenology traits have been a major focus for breeding programs around the world (M. Jung et al., 2020; McClure et al., 2018; Urrestarazu et al., 2017), and both wild and domesticated germplasm are routinely evaluated for their potential use by apple breeders (Gottschalk & van Nocker, 2013; M.

A. Khan et al., 2014). Traits such as precocity, harvest date and flowering date have practical implications for apple producers, as these traits influence investment timelines, crop quality and fruit damage risk. Weight, firmness, sugar content, acidity and phenolic content are important considerations for processors and consumers, who have specific preferences for these quality attributes when choosing to purchase apples (Cliff et al., 2016). Many of these fruit quality traits have been targets for improvement in breeding programs around the world, and current genetic mapping efforts remain focused on these phenotypes (Iezzoni et al., 2020; McClure et al., 2019; B. Wu et al., 2020). Cost-effective trait improvement in apples is critical since the investment costs of growing apple trees are high. Apple trees are large plants with a long juvenile phase: new trees often only start bearing fruit 5 years into the life cycle, requiring growers to invest heavily before generating revenue. Thus, producers typically grow only thoroughly evaluated and historically successful apple varieties. As a result, a small number of well-established varieties dominate the cultivated population. For example, in 2019 over 50% of all commercially produced apples in the US consisted of only 4 apple cultivars (WAPA, 2018). The global population of apples is dominated by a small number of elite varieties, despite an immense source of genetic and phenotypic diversity available for apple improvement (Migicovsky, Gardner, et al., 2021). Decreased diversity in apples, and agricultural crops more broadly, has resulted in an increased interest in the use of crop wild relatives (CWRs) for agricultural improvement. CWRs offer genetic and phenotypic diversity that can be leveraged in the breeding of novel cultivars with desirable traits such as disease resistance or flesh colour (McCouch et al., 2013). By 1997 the world economy had gained approximately $115 billion in benefits from the use of CWRs as sources of resistance to environmental change and disease (Pimentel et al., 1997). An understanding of how fruit quality and phenology vary within the cultivated apple's wild relatives is essential to future apple improvement.

Phenotyping large and diverse populations of plants is labour intensive and frequently results in a "phenotyping bottleneck" (Furbank & Tester, 2011), leaving crop researchers without powerful fruit quality data for analysis. Recently, comprehensive phenotyping of Canada's Apple Biodiversity Collection (ABC) generated measurements for fruit phenotypes in a collection of more than 1000 wild and cultivated apple accessions (Watts et al., 2021). In the present work, we explored ten phenotypes from the ABC and determined the degree to which the cultivated apple differed from its primary wild progenitor, *M. sieversii,* and how cultivated apples have changed over the past 200 years of breeding and improvement.

Materials and methods

Phenotype data

The phenotype data analysed here were collected from Canada's Apple Biodiversity Collection (ABC) and were part of previously published work (Watts et al., 2021). Briefly, the ABC is an apple germplasm collection located at the Agriculture and Agri-Food Canada (AAFC) Kentville Research Station in Nova Scotia, Canada (45.071767, -64.480466). The ABC contains 1119 unique accessions of apples planted in duplicate on M.9 rootstock in an incomplete randomised block design. The apple accessions in the ABC consist of accessions from the United States Department of Agriculture

(USDA) Plant Genetic Resources Unit apple germplasm collection in Geneva, NY, USA; commercial cultivars from the Nova Scotia Fruit Growers' Association Cultivar Evaluation Trial; and diverse breeding material from AAFC Kentville. The orchard consists largely of *M. domestica* accessions, but also contains 78 *M. sieversii* accessions.

Phenotype data from the ABC were collected in 2016 and 2017 (Watts et al., 2021). Here we focus on 10 phenotypes most relevant for assessing how apples have changed during domestication, breeding and improvement. Precocity was measured as a score of 1-4, indicating year of bloom; 1 (2014), 2 (2015), 3 (2016) and score 4 indicated that the tree had not yet bloomed as of 2016. Flowering date was measured in 2016 as the date in Julian days when the youngest wood displayed >80% of flowers at king bloom stage. Since it often took more than one day to harvest the entire orchard, harvest date was recorded in Julian days as the Monday of the week of harvest. Firmness was measured as the average firmness in kg/cm^2 at harvest of five apples measured using a penetrometer. Weight was measured as the average weight in grams of five apples at harvest. Acidity was measured as the malic acid content in mg/mL of combined juice from five apples measured using titration. Soluble solids were measured as °Brix of the juice of five apples using a refractometer. Phenolic content was measured as µmol GAE/g of fresh weight. Percent acidity change was measured by subtracting the acidity at harvest from the acidity after 90 days storage and then dividing by the acidity at harvest. Percent firmness change was measured by subtracting the firmness at harvest from the firmness after 90 days storage and then dividing by the firmness at harvest. Sample sizes for each phenotype are listed in Table 1.

Table 2-1. Sample sizes by phenotype.

Phenotype	***M. domestica***	***M. sieversii***
Precocity	797	76
Flowering Date	768	74
Harvest Date	647	59
Firmness	644	59
Weight	644	58
Acidity	626	56
Soluble Solids	644	56
Phenolic Content	399	9
% Change in acidity during storage	449	19
% Change in firmness during storage	409	27

Data analysis

Principal components analysis (PCA) was conducted using a scaled and centred matrix of the 10 phenotypes listed in Table 1 using the prcomp() function in R 4.0.2 (R Core

Team, 2020). A Wilcoxon signed-rank test was used to determine whether the phenotypes and PC values differed significantly between wild and cultivated apples.

A Pearson correlation was used to assess relationships between phenotypes and the release year of cultivated apples. Where appropriate, the significance threshold was Bonferroni-corrected to account for 10 comparisons. Data visualisation was performed using the ggplot2 R package (Wickham, 2016).

Results

PCA of the 10 phenotypes revealed modest overlap between cultivated and wild apples in phenotypic space (Fig. 2-1A, 2-1B). Wild and cultivated apples were significantly different along PC1 (W = 53893, p = 3.56 x 10^{-26}), PC2 (W = 13066, p = 2.07 x 10^{-17}) and PC3 (W = 39203, p = 0.0002; Fig. 2-1C).

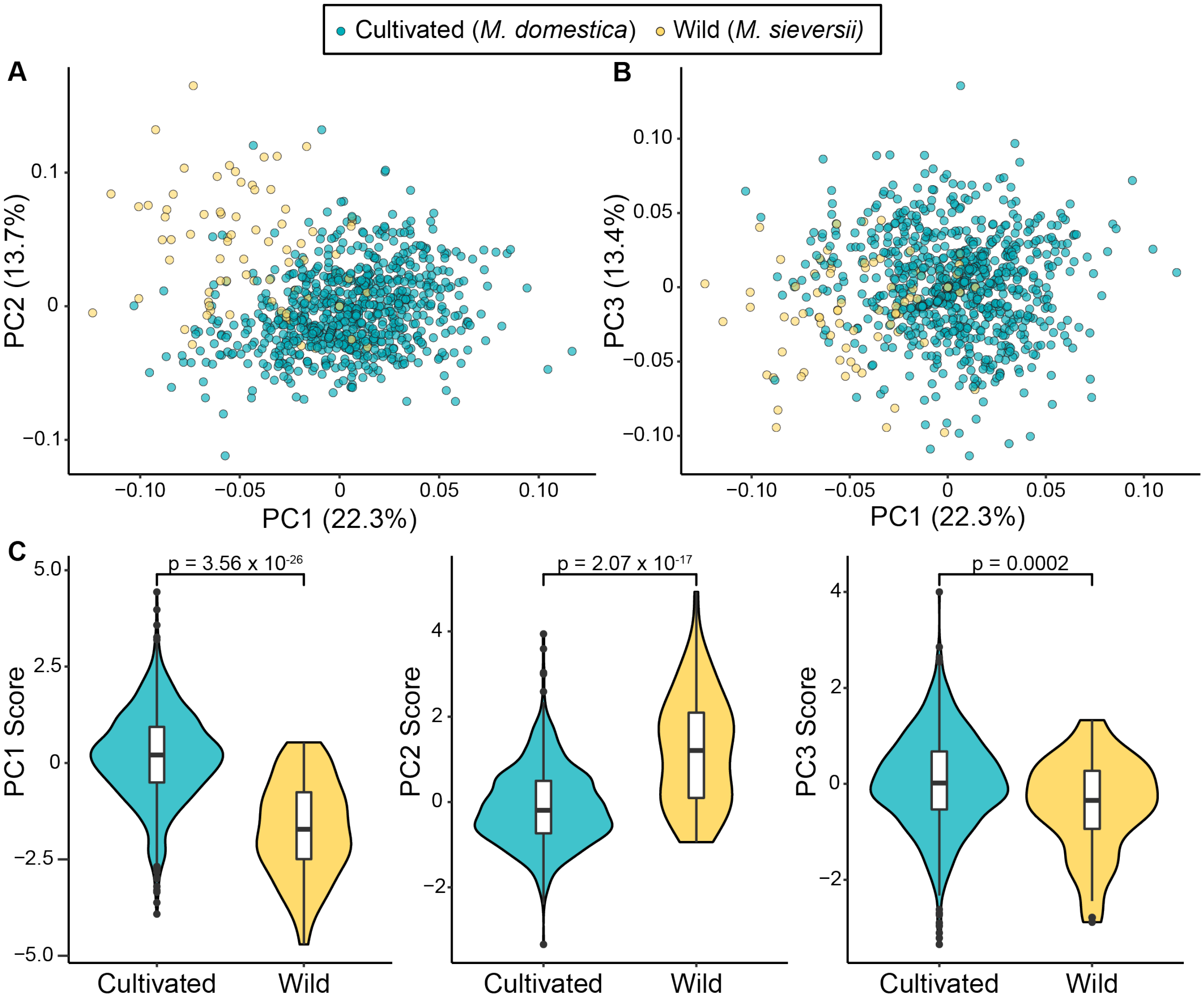


Fig 2-1. PCA of ten phenotypes in wild (N = 79) and cultivated apples (N = 801). A) PC1 vs PC2. B) PC1 vs PC3. The proportion of the variance explained by each PC is shown in parentheses on each axis. C) The difference between wild and cultivated apples for PCs 1, 2 and 3 are shown as violin plots. P values from a Wilcoxon test comparing PC values between cultivated and wild apples are shown for each of the first three PCs.

To visualise and assess the difference between cultivated and wild apples for each individual phenotype, we produced density plots to visualise each species' distribution for each phenotype and tested whether phenotypes differed between the two species (Fig. 2-2).

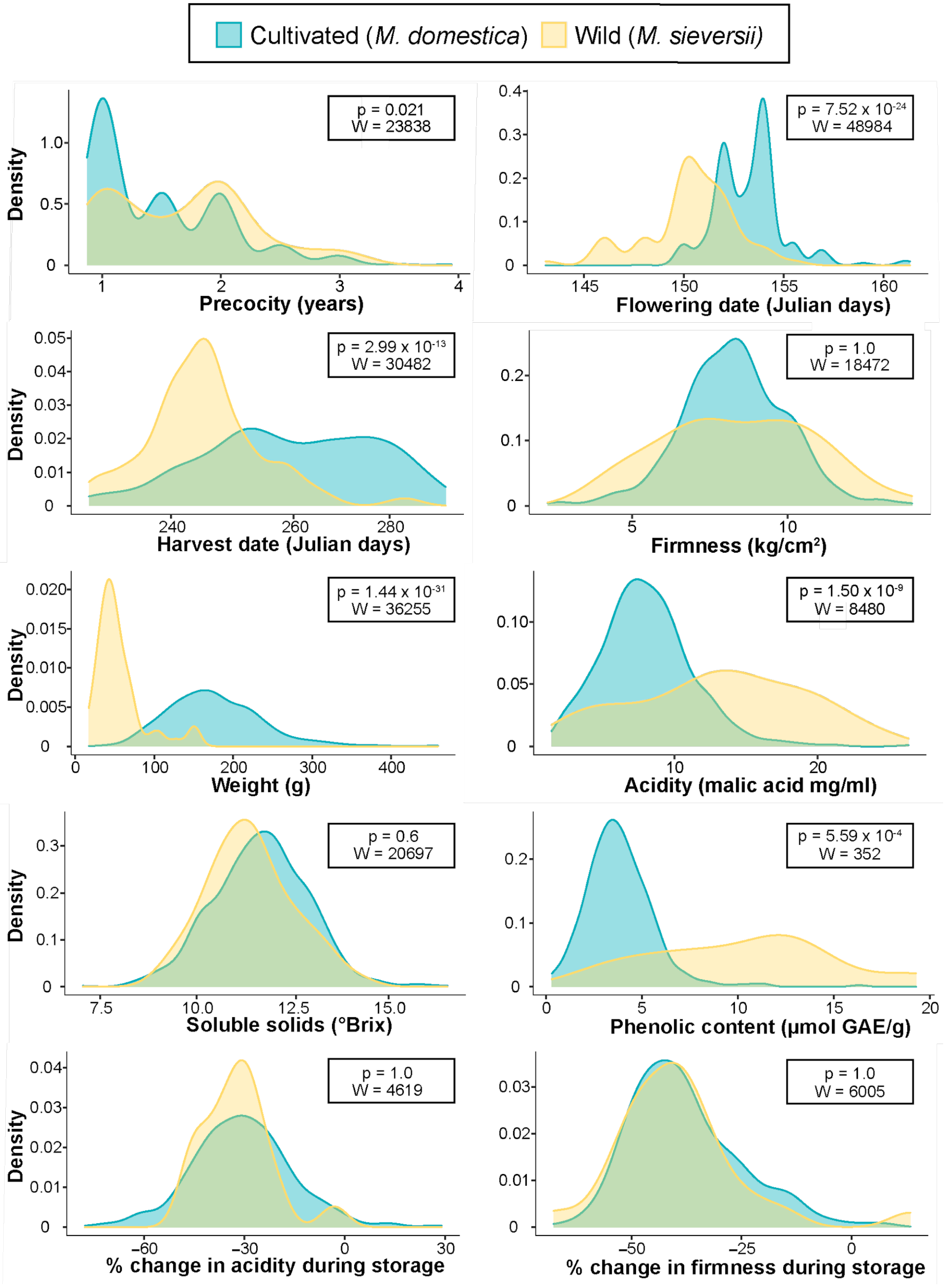

Cultivated (*M. domestica*)
Wild (*M. sieversii*)
Density
p = 0.021
W = 23838
Precocity (years)
0
0.5
1.0
1
2
3
4
p = 7.52 x 10^{-24}
W = 48984
Flowering date (Julian days)
0
0.1
0.2
0.3
0.4
145
150
155
160
Density
p = 2.99 x 10^{-13}
W = 30482
Harvest date (Julian days)
0
0.01
0.02
0.03
0.04
0.05
240
260
280
p = 1.0
W = 18472
Firmness (kg/cm^2)
0
0.1
0.2
5
10
Density
p = 1.44 x 10^{-31}
W = 36255
Weight (g)
0
0.005
0.010
0.015
0.020
0
100
200
300
400
p = 1.50 x 10^{-9}
W = 8480
Acidity (malic acid mg/ml)
0
0.05
0.10
10
20
Density
p = 0.6
W = 20697
Soluble solids (°Brix)
0
0.1
0.2
0.3
7.5
10.0
12.5
15.0
p = 5.59 x 10^{-4}
W = 352
Phenolic content (μmol GAE/g)
0
0.1
0.2
0
5
10
15
20
Density
p = 1.0
W = 4619
% change in acidity during storage
0
0.01
0.02
0.03
0.04
–60
–30
0
30
p = 1.0
W = 6005
% change in firmness during storage
0
0.01
0.02
0.03
–50
–25
0

Fig 2-2. Overlapping density plots of 10 phenotypes comparing values from wild and cultivated apples. The phenotype associated with each plot is shown along the X axis. The W and Bonferroni-corrected P values report the results of performing a Wilcoxon rank sum test of the difference between the phenotypic distributions of wild and cultivated apples.

Wild and cultivated apples differed significantly for 6 of the 10 phenotypes tested, including precocity (W = 23838, p = 0.021), flowering date (W = 48984, p = 7.52x10^{-24}), harvest date (W = 30482, p = 2.99x10^{-13}), weight (W = 36255, p = 1.44x10^{-31}), acidity (W = 8480, p = 5.1x10^{-9}), and phenolic content (W = 352, p = 5.59x10^{-5}). We found that, on average, cultivated apples produce flowers for the first time 21% (0.38 years) earlier than wild apples. Within a growing season, cultivated apples flower 3 days later, and are harvested 15 days later than wild apples. Cultivated apples are also 3.6 times heavier, 43% less acidic, and 68% lower in phenolic content than their wild progenitors. In comparison, wild and cultivated apples did not differ significantly for firmness, soluble solids, or changes in acidity or firmness during storage.

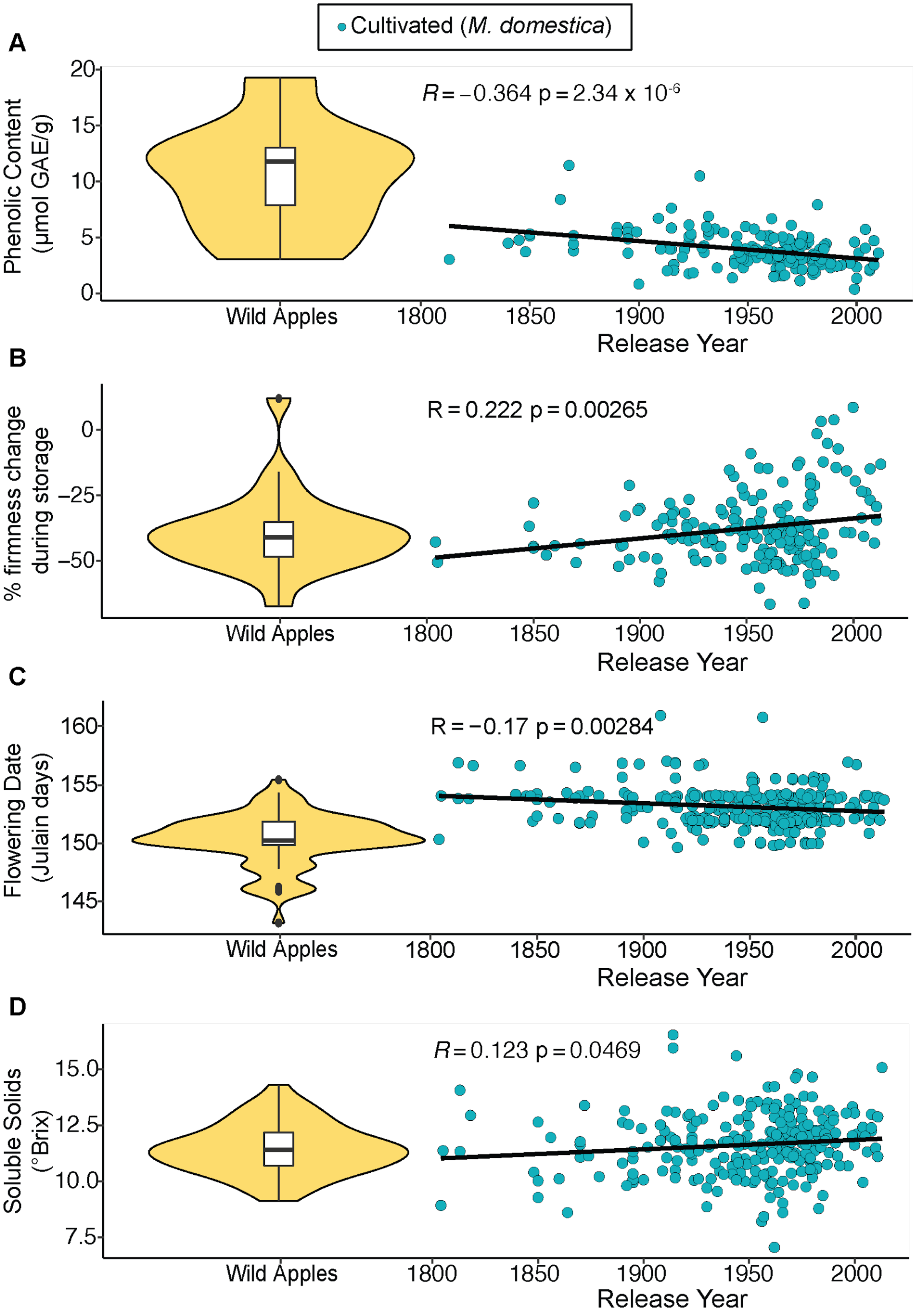


Fig 2-3. Phenotype values of cultivated apples as a function of their release year with a comparison to values in their wild ancestor, *M. sieversii*. Phenotypes include phenolic

content (A), firmness change during storage (B), flowering date (C), and soluble solids (D). Values for cultivated apples are blue, and the values observed for *M. sieversii* are represented in yellow as a violin plot on the left side of each plot. The R and P values from a Pearson correlation between phenotypic values and release year are shown within each scatter plot.

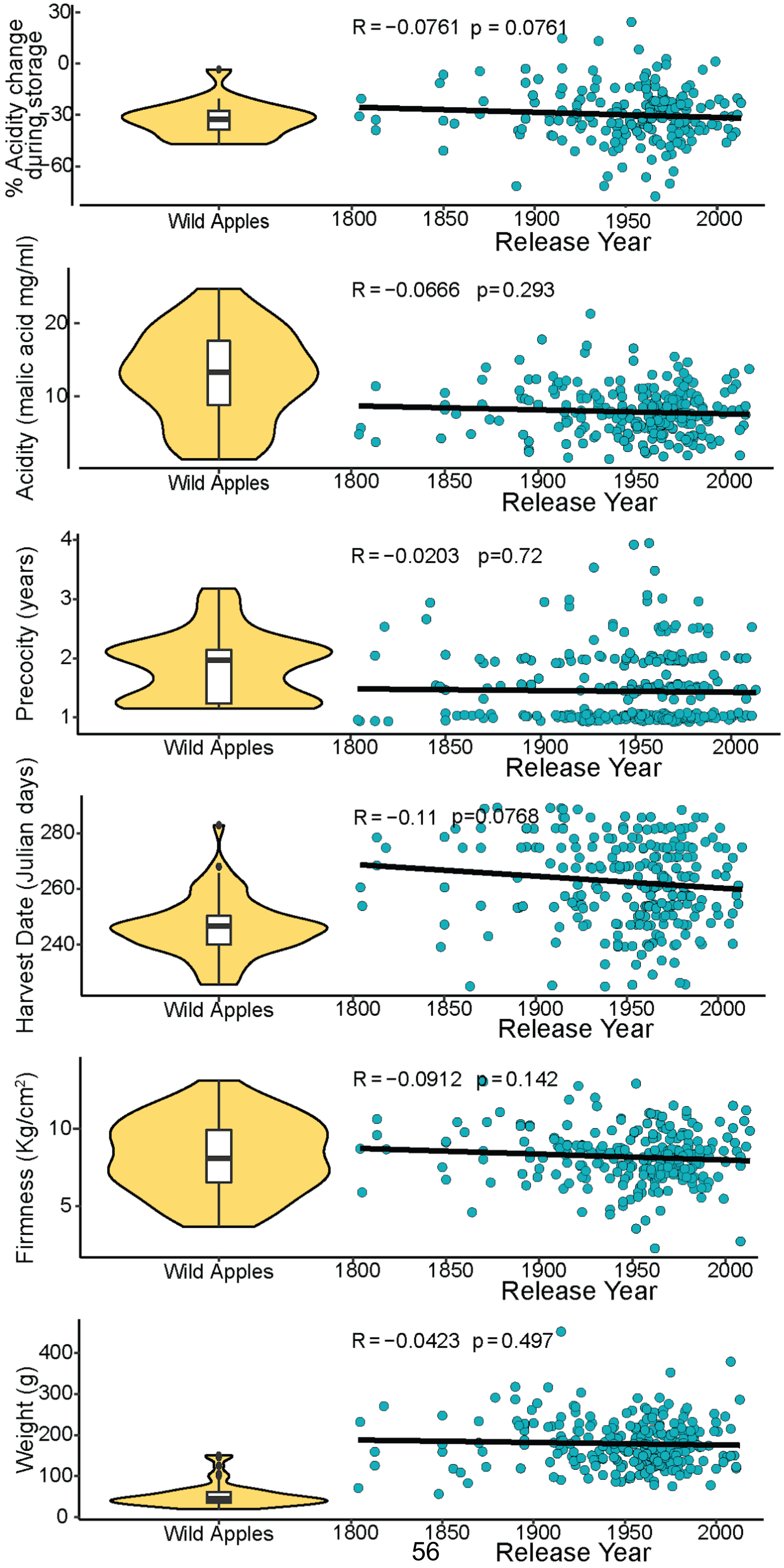

% Acidity change during storage
30
0
-30
-60
R = −0.0761 p = 0.0761
Wild Apples
1800
1850
1900
1950
2000
Release Year
Acidity (malic acid mg/ml)
20
10
R = −0.0666 p=0.293
Wild Apples
1800
1850
1900
1950
2000
Release Year
Precocity (years)
4
3
2
1
R = −0.0203 p=0.72
Wild Apples
1800
1850
1900
1950
2000
Release Year
Harvest Date (Julian days)
280
260
240
R = −0.11 p=0.0768
Wild Apples
1800
1850
1900
1950
2000
Release Year
Firmness (Kg/cm²)
10
5
R = −0.0912 p = 0.142
Wild Apples
1800
1850
1900
1950
2000
Release Year
Weight (g)
400
300
200
100
0
R = −0.0423 p = 0.497
Wild Apples
1800
1850
1900
1950
2000
Release Year

Fig 2-4. Phenotypes of cultivated apples as a function of their release year with a comparison to the ancestral state. Phenotypes include acidity change during storage, acidity, precocity, harvest date, firmness, and weight. Cultivated apple scores for each phenotype are shown in blue, and the ancestral state of each phenotype is represented in yellow as a density distribution of values from *M. sieversii*. The R and P values from a Pearson correlation between phenotypic values and release year are shown within each scatter plot.

To visualize phenotypic change within cultivated apples over time, apples' phenotypes are displayed as a function of their release year (Fig. 2-3 & Appendix I-I). We found significant correlations with release year for phenolic content (R = -0.364, p = 2.34x10^{-6}), change in firmness during storage (R = 0.222, p = 0.00265), flowering date (R = -0.172, p = 0.00247), and soluble solids (R = 0.123 , p = 0.0469) and determined that cultivated apples have shifted closer to the mean of wild apples for flowering date and firmness change, but further from the mean of wild apples for phenolic content and soluble solids.

Discussion

Apples have been cultivated for over 3000 years, but because vegetative propagation has been practised for 2000 years, it has been suggested that only about 100 generations have elapsed since apple domestication (Spengler, 2019). Despite this relatively short window for apple improvement, we found that cultivated apples are nearly entirely phenotypically distinct from their primary wild progenitor, *M. sieversii* (Fig.

2-1). Phenotypic differences are frequently used as an approximate measure of relatedness, and the separation in principal component space observed here is in agreement with genomic studies that have shown significant differentiation between the genomes of *M. domestica* and *M. sieversii* (Duan et al., 2017; Migicovsky, Gardner, et al., 2021). It is worth acknowledging that we observed some overlap between wild and cultivated apples in phenotypic space. The PCA performed here made use of only 10 phenotypes, and it is possible that more differentiation would be observed with more measures of the apple phenome. Further, each variable in PCA should ideally capture an independent biological feature of apples. However, some phenotypes analysed here are correlated, such as harvest date and firmness (Watts et al., 2021), and their variation may be driven by the same biological feature (Migicovsky et al., 2021). Therefore, interpreting our PCA as a quantification of the degree of phenotypic differentiation between cultivated and wild apples should take these caveats into consideration.

We found significant differences between wild and cultivated apples for several phenology traits including precocity, flowering date, and harvest date (Fig. 2-2). Cultivated apple trees flower and bear fruit at a younger age. Due to the long juvenile phase of apple trees, plants with the ability to bear fruit earlier in their life cycle are desirable for growers because revenue is generated earlier. It is therefore possible that precocity has been selected for during apple improvement.

Flowering date was 17% (3 days) later in cultivated apples than wild apples. Frost during blossoming can cause loss, damage or reduced marketability of fruits (Eccel et al., 2009), making flowering time an important consideration for growers when planting orchards. Additionally, apples with later flowering dates tend to be firmer (Nybom et al., 2013; Watts et al., 2021), and firmer apples are preferred by consumers (Harker et al., 2008). The later flowering date in cultivated apples could therefore be a by-product of selection for firm apples. Similarly, selection for firm apples may explain why cultivated apples were harvested 15 days later than wild apples, since harvest date and firmness are strongly correlated (Nybom et al., 2013; Watts et al., 2021). It is well established that harvest date is a reliable predictor of fruit firmness, and these two phenotypes may be regulated by a common molecular pathway (Migicovsky et al., 2021). Thus, preference for firm fruit could be directly impacting the selection for apples with later harvest dates.

We found significant differences between cultivated and wild apples across multiple fruit traits including weight, acidity, and phenolic content (Fig. 2-2). Cultivated apples are 3.6x heavier than wild apples, in agreement with previous comparisons between these two species (S. Kumar et al., 2014). Consumers prefer large, visually appealing fruit (Carew & Smith, 2004; Skreli & Imami, 2012), so selection for large fruit size may explain our observation. We also found that cultivated apples are 43% less acidic than wild counterparts. Acidity contributes to the sour taste of apples, and apple preference is heavily influenced by acid/sugar ratios (Hampson et al., 2000). Given this relationship, it is not surprising that cultivated apples, which are primarily consumed as fresh fruit (Lutes, 2019), have lower acid than wild apples but do not differ in soluble solid content. Finally, cultivated apples have, on average, 68% less phenolic content than wild apples. Phenolic compounds, which offer nutritional benefits (D. Lin et al., 2016), are partially responsible for the enzymatic browning that occurs when apple flesh is exposed to oxygen (Holderbaum et al., 2010). Browned flesh is visually unappealing and typically results in negative effects on flavour, making apples that resist browning more appealing to producers and consumers (Holderbaum et al., 2010). In fact, the only genetically modified apple variety on the market today, Arctic^{TM} Apples, was designed to silence genes related to enzymatic browning and was advertised as “the original nonbrowning apple” (Stowe & Dhingra, 2020). The human aversion to apple browning has likely contributed to the decline in phenolic content in cultivated apples, despite the nutritional benefits of such compounds. In addition, some evidence suggests that fruit size impacts polyphenol accumulation in apples (Busatto et al., 2019), which could help explain why we observe lower phenolic content in cultivated apples.

According to the present analysis, many phenotypes of cultivated apples have dramatically changed since divergence from the primary progenitor species, *M. sieversii*. These differences represent phenotypic separation that could be leveraged in the improvement of cultivated apples, and emphasizes the potentially functional diversity provided by CWRs. While wild apples from this investigation may not offer improved fruit quality phenotypes that are currently attractive to consumers, they hold phenotypic variation that could be important for apple improvement in the future. For example, breeders could exploit the high phenolic content of wild apples to improve the nutritional quality of cultivated apples. Further, traits from wild apple varieties could potentially benefit the cider industry, which values high acidity and phenolic content (Mattila et al., 2006).

Analysis of cultivated apple phenotypes as a function of release year revealed changes over the past 200 years in phenolic content, change in firmness during storage, flowering date, and soluble solids (Fig. 2-3). In particular, as shown previously (Watts et al., 2021), phenolic content has decreased over time. Phenolic content is associated with bitter taste (Soares et al., 2013), and modern varieties therefore likely taste less bitter on average than older varieties. Although selection for decreased bitterness could explain our observation, the relationship between low phenolic content and decreased flesh browning could also explain why modern cultivated apples tend to have less phenolics (Toivonen, 2006). In comparison, wild apples tend to have higher phenolic content, indicating that cultivated varieties are diverging from the ancestral state. Similarly, more recently released apple cultivars soften less during storage than older cultivars, diverging from the ancestral state. The extended storage and long-distance shipment of apples has become increasingly routine over the past several decades, and selection for reduced softening during storage may explain why firmness retention has improved over time. Storage and transport have also been key targets in tomato breeding (Kramer & Redenbaugh, 1994), and the demand for fruit that performs well during extended storage and transport is unlikely to subside.

Flowering date is an important trait for apple production, and varies widely across the genus *Malus* (Gottschalk & van Nocker, 2013). Later flowering apple trees are less likely to be impacted by frost damage (Eccel et al., 2009) and more likely to be firm (Watts et al., 2021), which is preferred by consumers. Despite the understood benefits of growing apples with later flowering dates, we found that more recently released varieties had earlier flowering dates. The trend towards earlier flowering varieties could indicate that selection for other traits has indirectly impacted flowering date. Alternatively, growers could be preferring earlier flowering varieties in an attempt to manage fruit ripening times during the harvest season. Cultivated varieties are trending towards the ancestral state of earlier flowering dates, which suggests that wild apples could offer valuable genetic material for breeding earlier harvested varieties.

Finally, we found that more modern cultivated apples are only slightly higher in soluble solid content. Previous investigations have reported that firm apples tend to have higher sugar content (McClure et al., 2018; Migicovsky et al., 2016; Nybom et al., 2013), so our observation that modern apple varieties tend to have higher soluble solids content (SSC) may be at least partially be driven by recent selection for increased firmness. Further, a number of studies have suggested that the sugar content of apples is a key factor affecting consumer preference (Cliff et al., 2016; Harker et al., 2008). Although SSC is only a modest predictor of perceived sweetness (Aprea et al., 2017), consumer’s preference for sweet apples could underlie the upward trend in soluble solid content seen in modern cultivated apples.

Several caveats of the present analysis are worth noting. First, we only considered one of the multiple progenitor species of *M. domestica* here (X. Sun, Jiao, et al., 2020). Therefore, only a fraction of the ancestry of the cultivated apple is captured by *M. sieversii*, and a more inclusive pool of ancestral species would yield a more comprehensive comparison of wild and cultivated apples. Second, it is unknown how representative the current sample of wild apples is of the broader *M. sieversii* population. It is possible that the wild apple varieties within the ABC represent only an unrepresentative subset of *M. sieversii*, and thus do not accurately capture the diversity of the species. Further, there has been evidence of gene flow between cultivated and wild apples (Cornille et al., 2013), which could mean that the wild species from the current investigation have experienced gene introgression from cultivated trees, and thus do not accurately represent the wild progenitor. Finally, the relatively small sample size in several comparisons limited the power of some of our analyses (Table 3-1).

Our work demonstrates that cultivated and wild apples have diverged phenotypically, and that hundreds of years of apple improvement have shaped the variation in fruit and phenology we observe among cultivated apples today. Wild apples offer potentially valuable pools of genetic material that may be helpful for apple improvement. Future comprehensive phenomic evaluations, including metabolomic and transcriptomic analyses, across diverse wild apple species will help further assess the degree to which the apple's wild relatives may contribute to improving apple cultivar development.

Conclusions

There are many difficulties associated with finding suitable populations for the comparison of cultivated and wild relative species, particularly in long lived woody perennial crops such as apple. However, by leveraging the vast diversity and large size of the ABC, the present study provides a meaningful comparison of the cultivated apple and its primary wild progenitor. According to the present analysis, the domesticated apple and its primary wild progenitor are distinguishable across multiple phenotypes. This analysis revealed that cultivated apples are significantly heavier, less acidic, and have lower total phenolic contents than their wild ancestor. Further, cultivated apples flower later, have later harvest dates, and bear fruit within fewer growing seasons than wild apples. Historical analysis revealed that more modern cultivated apple varieties are higher in soluble solids, have lower total phenolic content, flower earlier, and retain more firmness during storage in comparison to older varieties. Overall, this analysis revealed that the phenome of these two species of apple have significantly diverged, and that historical breeding trends are detectable in this population.

Importantly, this analysis detected recent trends in apple phenotypes, particularly increases in firmness retention, decreases in total phenolic content, earlier flowering dates, and increases in soluble solids. Patterns observed in these phenotypes indicate there may have been significant selection for these phenotypes in the last 200 years. This conclusion is supported by the fact that many contemporary apple breeding programs continue to target improvements in these phenotypes today (Rob Blakey, personal communications, 2021, Kevin Brandt, personal communications, 2021, Erin

Wallich, personal communications, 2021). The discovery of the alleles controlling these traits could therefore offer a path to rapid phenotype improvement and provide value to the apple industry. In the future, it may be wise to put efforts towards discovering the causal DNA sequences controlling these traits, as such discoveries could aid apple breeders and benefit our agricultural system.

Data availability

All data presented are freely available to the public via Watts et al. (Watts et al., 2021). Statistical analyses presented here can be found on GitHub at https://github.com/MylesLab/Wild_vs_cultivated.

Author contributions

Thomas Davies[1]

ROLES: Conceptualization, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing

Sophie Watts[1]

ROLES: Data curation, Methodology, Writing – review & editing

Kendra McClure[1]

ROLES: Data curation, Methodology, Writing – review & editing

Zoë Migicovsky[1]

ROLES: Conceptualization, Data curation, Supervision, Writing – review & editing

Sean Myles[1]

ROLES: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – review & editing

*E-mail: sean.myles@dal.ca

[1]Department of Plant, Food, and Environmental Sciences, Faculty of Agriculture, Dalhousie University, Truro, NS, Canada

Acknowledgements

The authors thank the Nova Scotia Fruit Growers' Association and the Farm Services team at AAFC-Kentville for their work in establishing and maintaining the trees studied here. This work was supported by the National Science and Engineering Research Council of Canada. Z. Migicovski was supported by the National Science Foundation Plant Genome Research Program 1546869.

Chapter 3: Pool-seq of diverse apple germplasm reveals candidate loci underlying ripening time, phenolic content, and softening.

The contents of this chapter are published as:

Davies, T., & Myles, S. (2023). Pool-seq of diverse apple germplasm reveals candidate loci underlying ripening time, phenolic content, and softening. Fruit Research, 3(1). https://doi.org/10.48130/FruRes-2023-0011

Rationale

Previous genetic mapping efforts from our research group and others indicate that GBS data does not offer the marker density and genomic resolution required to detect the causal allele(s) controlling apple phenotypes (Larsen et al., 2018; McClure et al., 2018). Because the GBS data in the ABC population is relatively sparse (approximately 1 variant per 2kb), it is unlikely that causal alleles are to be found among the GBS markers. Rather, GBS markers are likely to be in linkage with causal alleles but offer little additional information as to the specific identity or location of the allele(s). In all but the most fortunate of circumstances, in which a GBS marker happens to be the causal variant, our GBS marker dataset will not enable the confident detection of causal alleles at the nucleotide level in the ABC. Therefore, increased genetic resolution in the ABC population needs to be generated to move closer to discovering causal alleles. The aforementioned mapping attempts have made it clear that a precise mapping of causal alleles for key traits will remain difficult without the resolution provided by whole genome sequencing (WGS) data. Pooled sequencing approaches are cost-effective and provide high-density marker data, essentially saturating the genome with markers. For these reasons, the pool-sequencing approach was selected as the method of choice in this chapter.

This chapter aims to scan the genome for causal alleles controlling three agriculturally important fruit phenotypes: ripening time, softening, and total phenolic content. Of the three phenotypes, ripening time has been the focal point of a substantial body of work (M. Jung et al., 2020; Larsen et al., 2019; McClure et al., 2018; Urrestarazu et al., 2017). In previous mapping experiments it was hypothesised that a coding variant in the first exon of NAC18.1, a transcription factor, controlled harvest date (Migicovsky et al., 2021; Watts et al., 2023). However, functional experiments by our group determined that not only were there multiple coding sequence variants impacting the protein structure of NAC18.1, but that two coding region haplotypes at *NAC18.1* produced no significant differences in transgenic tomato fruit ripening (Migicovsky et al., 2021). Therefore, the question addressed in the following chapter is: Is the causal variant controlling ripening time in the coding region of NAC18.1 or in a nearby non-coding region? To address this question, WGS data is leveraged to examine the genome in high resolution. In this chapter, I employ a WGS pool-sequencing approach with the goal of discovering the causal alleles for ripening time, fruit softening, and total phenolic content.

Abstract

Ripening time, softening, and phenolic content are phenotypes of considerable commercial importance in apples. Identifying causal genetic variants controlling these traits not only advances marker-assisted breeding, but it is also an essential step for the application of gene editing technologies in apples. To advance the discovery of genetic variants associated with these phenotypes, we examined allele frequency differences between groups of phenotypically extreme samples from Canada's Apple Biodiversity Collection using pooled whole genome sequencing (pool-seq). We sequenced pooled DNA samples to an average read depth of 150x and scanned the genome for allelic differentiation between pools. For each phenotype, we identified >20 million genetic variants and identified numerous candidate genes. We identified loci on chromosomes 3 and 4 associated with ripening time, the former suggesting that regulatory variants upstream of a previously identified transcription factor *NAC18.1* may be causal. Our analysis identified candidate regions on chromosomes 4, 8, and 16 associated with phenolic content, and suggested a cluster of *UDP-Glycosyltransferase family* genes as candidates for polyphenol production. Further, we identified regions on chromosomes 17 and 10 associated with softening and suggest a *Long-chain fatty alcohol dehydrogenase family* gene as putatively causal.

Introduction

Apples (*Malus x domestica Borhk*) are an ancient crop species, with evidence of domestication dating to at least 3,000 years ago (Zohary & Hopf, 2000). Today, apples are the world's third most valuable fruit crop worth $77 billion annually (FAOSTAT, 2021), and they are widely recognized as an important source of sustenance and nutrition for the human population. Continuous improvement of apple varieties is important for the sustainability and success of the industry, but breeding improved apple varieties remains a difficult challenge. Apple trees are highly heterozygous and require expensive maintenance, resulting in a costly breeding process. Further, when breeding for fruit quality traits, new varieties cannot be assessed until trees have matured through the juvenile phase, which can take 4-7 years. These biological characteristics make apples an excellent candidate for the use of molecular breeding tools that can accelerate breeding cycles and reduce the costs of bringing new apple varieties to market.

Molecular breeding tools offer valuable strategies for breeders to reduce breeding costs and more efficiently improve crops. For complex traits controlled by numerous small effect loci, the use of genome-wide genetic markers is now widely used in a genomic selection (GS) framework (Heffner et al., 2009). For traits controlled by a small number of large effect loci, however, marker assisted selection (MAS) using a small number of markers, can significantly decrease costs during apple variety improvement (Edge-Garza et al., 2015; Luby & Shaw, 2001). Ideally, genetic markers used for MAS are causal alleles that control traits targeted for improvement. However, many genetic markers used for apple breeding are only linked to desirable traits based on genetic mapping studies but have not been shown to be causal (Migicovsky et al., 2021; Nybom et al., 2013). Thus, there remains uncertainty about the degree to which markers used for MAS in diverse apple germplasm accurately predict phenotypes and are effective in reducing breeding costs.

In recent years, molecular techniques such as gene editing have become valuable tools for crop improvement, allowing researchers to make targeted changes to DNA sequences in elite germplasm in numerous crops (H. Jia et al., 2017; Svitashev et al., 2015; F. Wang et al., 2016). While a number of barriers must be overcome before genome editing can be effectively applied for apple cultivar improvement, the approach holds tremendous promise for apple cultivar improvement, possibly through gene knock-outs (Charrier et al., 2019) or targeted allele swaps mediated via the application of base editors (Malabarba et al., 2020). In most cases, gene editing will require the identification of causal genetic variants for commercially important traits, however few have been previously identified. To date, genetic mapping studies in apple have generally lacked the sample size, diversity and marker density required to identify causal genetic variants at nucleotide resolution. The discovery of causal genetic variants underpinning important agricultural traits thus continues to be a challenge in apple, and ultimately limits the ability of breeders to make improvements in key agricultural traits via genome editing technologies.

To advance apple improvement via gene editing, it is critical to identify causal alleles controlling important agricultural traits. Ripening time, phenolic content, and softening are three important fruit traits in apple as they impact labor management, fruit nutrition, and fruit storage, respectively. Numerous genetic mapping studies have investigated these traits in the past (Bink et al., 2014; Chagné, Krieger, et al., 2012;

McClure et al., 2019; Migicovsky et al., 2021; Nybom et al., 2013) and they are likely to remain target traits for apple improvement in the future. Therefore, an understanding of the genetic architecture and causal genetic variants underlying these traits is important for future apple variety improvement.

Numerous attempts have been made to map the causal alleles underpinning ripening time, phenolic content, and softening in apple. Multiple genome wide association studies (GWAS)(M. Jung et al., 2020; Larsen et al., 2019; Migicovsky et al., 2016; Urrestarazu et al., 2017) and functional genomics evidence (Migicovsky et al., 2021) suggest that *NAC18.1* (MD03G1222600), a transcription factor on chromosome 3, is a key gene involved in ripening time variation in apple. However, the causal allele(s) in or around *NAC18.1* responsible for ripening time variation remain unknown. Similarly, the causal allele(s) for phenolic content in apple remain elusive, despite a number of investigations proposing *leucoanthocyanidin reductase* (*LAR1)*, on chromosome 16 (Chagné, Krieger, et al., 2012; S. A. Khan, Chibon, et al., 2012; McClure et al., 2019) as a candidate gene for phenolic content production. While QTLs associated with fruit softening have been identified on multiple chromosomes, and the genes *PG1* and *ERF* have been either functionally validated or proposed as putatively causal in determining the storability of apple fruits (Di Guardo et al., 2017; McClure et al., 2018; B. Wu et al., 2021), causal alleles for softening also remain unknown. Despite numerous attempts through both linkage mapping and GWAS, the precise locations of causal genetic variants underlying these three traits remain unknown.

The discovery of causal alleles for key traits in apple has remained challenging in large part due to the costs of gathering high quality phenotype and genotype data across sufficiently diverse populations. With the rapid expansion of high-throughput DNA sequencing in recent years, whole genome sequencing of pooled DNA samples has become a powerful cost-effective approach to identify allele frequency differences between populations that differ in phenotype. By pooling DNA samples from extremes of a phenotype distribution, genomic regions with extreme allele frequency differences between pools are identified as loci that potentially harbour causal genetic variants for the phenotype of interest. This method has been successfully used to identify causal loci in non-model organisms such as geese, watermelon, and cannabis (Dong et al., 2018; S. Ren et al., 2021; Welling et al., 2020). In apple, pool-seq approaches have been used to investigate the genetic basis of acidity, weeping, and internal browning traits (S. Ban & Xu, 2020; Dougherty et al., 2018; S. Kumar et al., 2022). Here, we use a pool-sequencing approach (Kofler et al., 2011) to evaluate allele frequency differences between sub-populations of apples from a diverse population that vary markedly for ripening time, polyphenol production, and softening. Allele frequency differences and a modified chi-squared test (Spitzer et al., 2019) were used here to scan the genome for regions with the largest allele frequency differences between groups, and genes in these regions were curated and discussed.

Materials and methods

Pool selection and DNA sequencing

DNA was extracted from leaf tissue collected from Canada's Apple Biodiversity Collection (ABC) in Kentville, Nova Scotia, Canada, as described in Migicovsky et al (Migicovsky et al., 2021). For each phenotype examined here (ripening time, phenolic content, and softening), 50 *M. domestica* accessions from the ABC with the most extreme phenotypic values were selected from each tail of the phenotype distribution (Fig. 3-1), forming two groups of 50 accessions (except in cases where DNA extraction failed) for each phenotype. DNA from accessions within each of the selected groups was combined into a pool, with DNA from each sample represented in equimolar concentration. DNA extraction and pooling was performed by Platform Genetics Inc. A total of six equimolar DNA pools were formed: late harvested (N= 50), early harvested (N= 49), high phenolic content (N= 50), low phenolic content (N= 49), low softening (N= 50), high softening (N= 50). For phenotypic selection, apple phenotype measurements from 2017 measured by Watts et al.(Watts, Migicovsky, McClure, et al., 2021) were used. Ripening time was measured as the Julian day of the year in which the fruit were deemed ripe and ready for harvest. Phenolic content was measured as micromolar of gallic acid equivalents per gram of fresh weight (µmolGAE/g) via a Folin–Ciocalteu assay. Softening was measured as the percent change in firmness between harvest and 3 months post-storage, as measured by a penetrometer. Details of the germplasm used, the experimental design of the orchard, and the phenotyping protocols are provided in Watts et al (Watts, Migicovsky, McClure, et al., 2021).

Pooled libraries were prepared and whole genome sequencing was performed by the McGill Genome Centre. DNA libraries were prepared using a Lucigen PCR-free NxSeq kit. Each pool was sequenced on a single lane of Illumina NovaSeq6000 S4 v1.5 PE150 in high output mode.

Sequence data pre-processing, mapping and variant calling

FastQ files from each pool were aligned to the Golden Delicious double haploid reference genome (Daccord et al., 2017) using the MiniMap2 alignment tool (H. Li, 2018). Binary Alignment Map (BAM) files were produced following the GATK Best Practices guidelines (https://gatk.broadinstitute.org). Mapped BAM files were coordinate sorted and indexed with Samtools sort and index functions(H. Li & Durbin, 2009). Sequencing duplicates were marked with Picard MarkDuplicates (http://broadinstitute.github.io/picard/). Samtools was used to produce 3 mpileup files, one for each of the three phenotypes. The Popoolation2 pipeline (Kofler et al., 2011) was used to produce three sync files from each mpileup file (Appendix II-I). To reduce the number of false positive variants, only variants supported by a read depth of 50-500x in each pool with a combined alternate allele count of at least 10 were considered for downstream analyses (Ries et al., 2016; Welling et al., 2020).

Allele frequency estimation, candidate region identification, and gene model curation

Allele frequency estimates (AFe) for each pool were generated for each site in the genome using the snp-frequency-diff.pl script within Popoolation2 (Kofler et al., 2011). Delta-AFe values were calculated as the absolute difference of AFe at each variant between pools. AFe and delta-AFe values were calculated and analyzed with the poolSeq package in R (Taus et al., 2017). Allele counts at each position were used to conduct a modified chi-squared test (CST) in R using the adapted.chi.squared function within the ACER package (Spitzer et al., 2019). Delta-AFe values and CST p-values for each variant site were visualised using the qqman package in R (D. Turner, 2018). Candidate regions for each phenotype were defined as regions of the genome within 20 kb of the top and bottom 0.001% of delta-AFe and CST p-values, respectively. Gene annotations were produced by Daccord et al (Daccord et al., 2017).

Gene Ontology (GO) Enrichment analysis

To identify candidate genes involved in each phenotype, protein coding genes within 20 kb of variants within both the 0.001% lowest CST p-values and 0.001% highest delta-AFe values for each phenotype were curated and reduced to a unique set (MD IDs) (Table S2,S3,S4). This resulted in 21, 385, and 321 candidate genes for ripening time, phenolics, and softening, respectfully. Genome wide annotations as well as annotations for genes associated with top hits from each phenotype were imported using the topGO package in R (Alexa & Rahnenfuhrer, 2020). Gene enrichment in biological process ontology was tested using the topGO package with algorithm parameters ‘weight01’, to account for GO hierarchy, and ‘fishers’ as the test statistic.

Results

Phenotype distributions

Ripening time, phenolic content, and softening trait values were each roughly normally distributed in the ABC population, with the phenolic content distribution showing an extended tail containing apples with high phenolic content (Fig. 1). Ripening time in the population ranged from 225-290 Julian days, with a mean ripening time of 261 Julian days. The early and late pools ranged from 225-236 (mean 229) and 282-290 (mean 289) Julian days, respectively. The mean value for total phenolic content in the ABC population was 4.34 µmolGAE/g. The low and high phenolic content pools had phenolic content values that ranged from 0.3-2.2 (mean 1.4) and 6.1-27.9 (mean 10.0) µmolGAE/g, respectively. On average, apples lost about a third of their firmness during storage: change in firmness within the population ranged from -67.7% to 13.4%, with a mean change in firmness of -37.8%. The high and low softening pools had percent change in firmness values that ranged from -67.7 to -51.1% (mean -56.2%) and -19.7-13.4% (mean -10.9%), respectively.

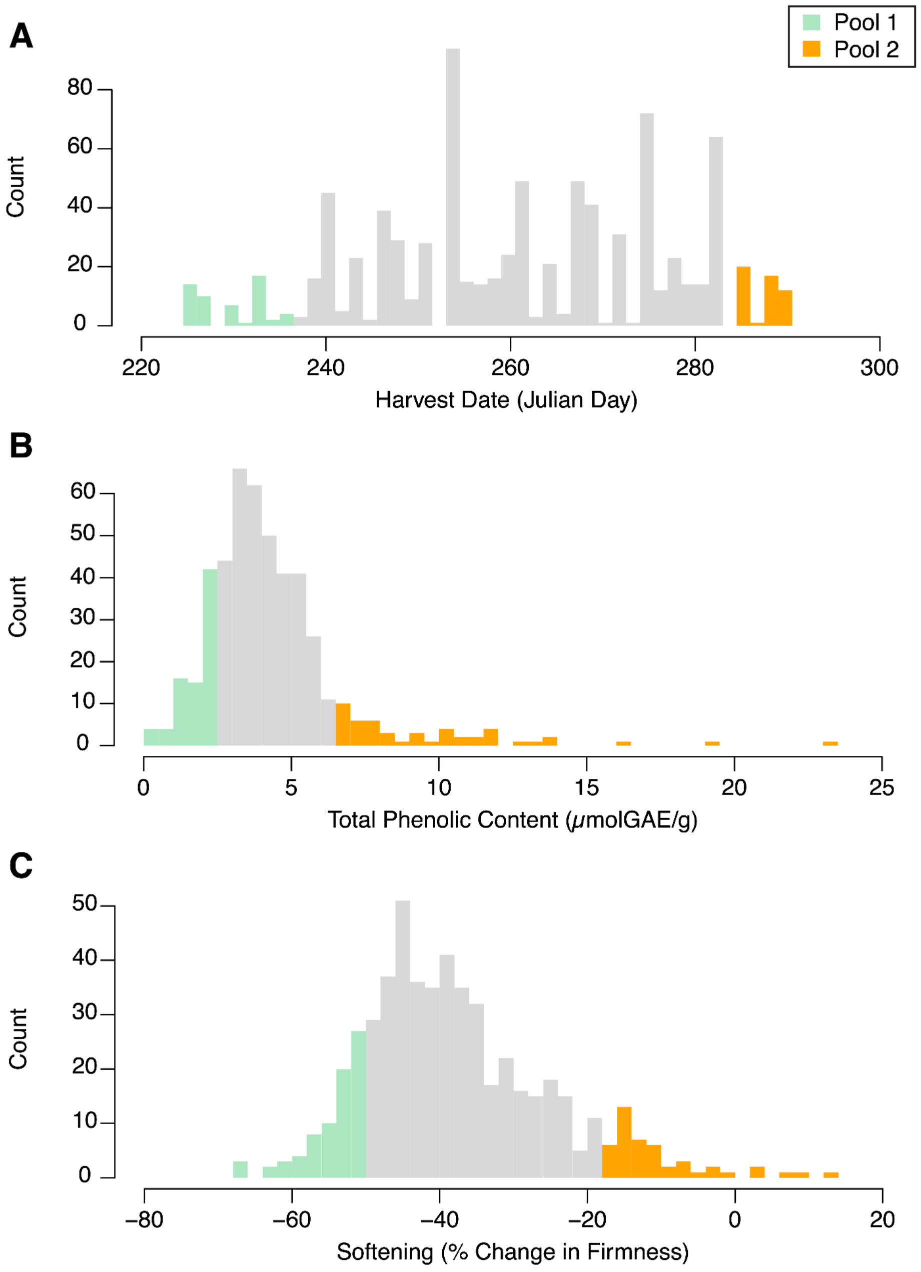


Fig. 3-1. Phenotype distributions for ripening time, phenolic content, and softening.

Green and orange bars represent accessions selected for pooled sequencing.

Genome sequencing and variant calling

DNA sequencing produced a combined 2.8 billion reads comprising more than 864 billion base pairs of DNA sequence. Mapping rates for the libraries varied from 95.96 to 96.54%. Read depth for pools ranged from 128.4-184.6x (Table S1). Average read depth across all six pools was 150.4x (Appendix II-II). After filtering for positions with read depths within the acceptable read depth range (50-500x), we obtained 81%, 82% and 81% coverage of the apple reference genome for ripening time, phenolic content, and softening, respectively. The mean number of variants called for each phenotype was 25,506,587 (Table S1).

Candidate region identification

The highest observed delta-AFe value for ripening time was 0.923 found on chromosome 4. Chromosomes 3, 4, 7, and 16 harbored variants with delta-AFe values greater than 0.8 (Fig. 3-2a). Two notable peaks, on chromosomes 3 and 4, were identified by delta-AFe and CST analysis (Fig. 3-2a,b). The signal on chromosome 3 consists of a 76.7 kb region, from 30,656,169 to 30,732,938 bp (Fig. 3-2c). Within this window, 259 variants had delta-AFe values > 0.8. The variant with the highest local delta-AFe (0.907) was an A/G SNP at bp 30,702,958, approximately 4.7 kb upstream of a NAC transcription factor previously associated with ripening time(Larsen et al., 2019; Migicovsky et al., 2016; Migicovsky et al., 2021). The same variant scored the lowest local CST p-value (1.81×10^{-126}). The second peak on chromosome 4 spanned approximately 10 kb (Fig. 3-2d). This signal contains a C/A SNP located at 1,482,075 bp, with the single highest delta-AFe (0.923) and lowest CST p-value (9.39×10^{-127}) for ripening time. This variant window contained 12 variants with delta-AFe values > 0.8. None of the variants from the peak on chromosome 4 were within annotated gene-coding regions, however the peak is within 15 kb of the coding region of a *histidine kinase* gene (MD04G1013100) and a *methionine tRNA ligase* gene (MD04G1013000). Twenty-one unique genes, 9 of which had associated GO terms, were within candidate regions for ripening time. We report significantly enriched GO terms (Table S2) for genes in these regions, which included metabolic processes and phosphatidylinositol phosphate biosynthetic processes.

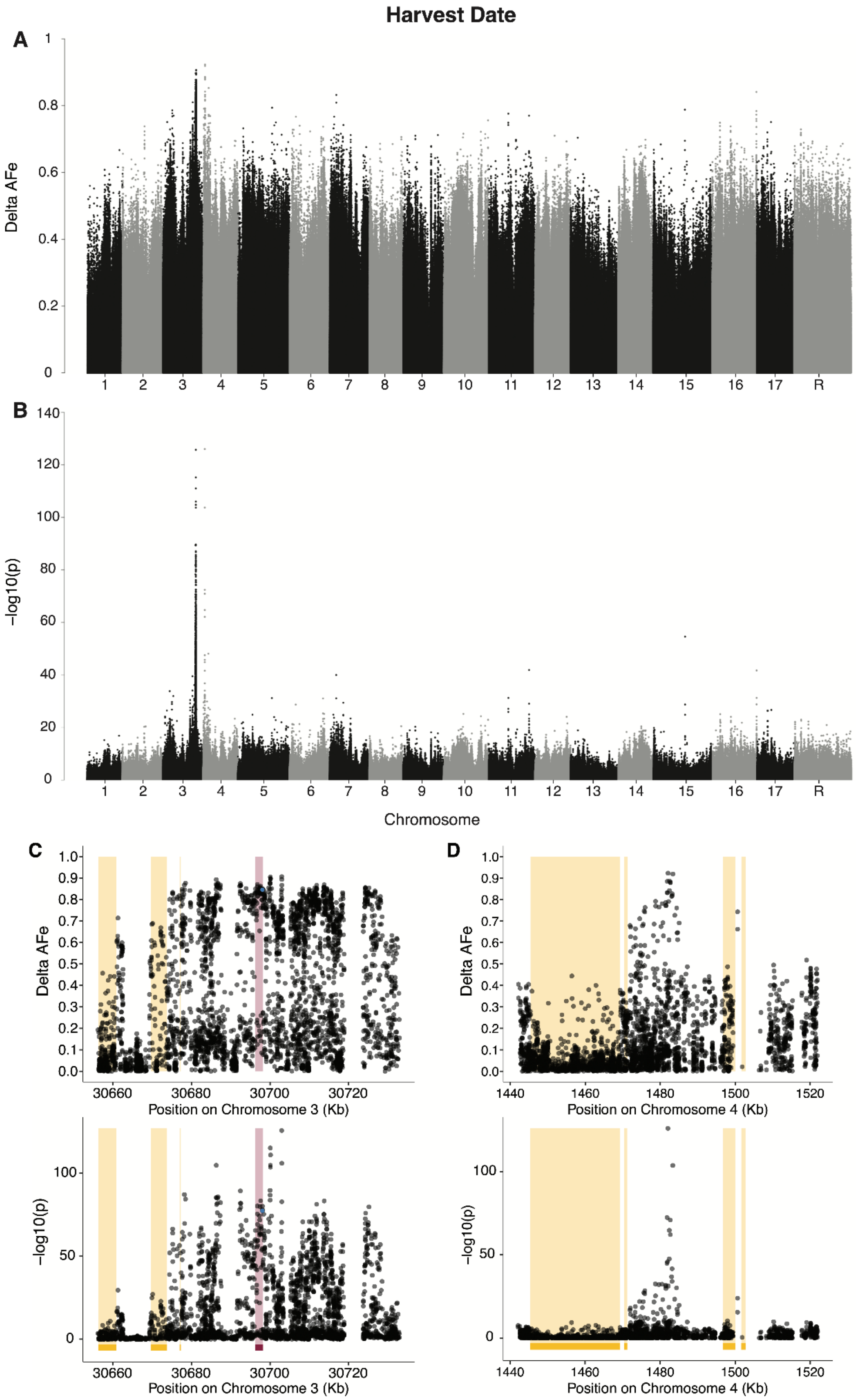

Harvest Date
A
Delta AFe
0
0.2
0.4
0.6
0.8
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
R
B
−log10(p)
0
20
40
60
80
100
120
140
Chromosome
C
Delta AFe
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
30660
30680
30700
30720
Position on Chromosome 3 (Kb)
−log10(p)
0
50
100
D
1440
1460
1480
1500
1520
Position on Chromosome 4 (Kb)

Fig. 3-2 Manhattan plots for genome wide delta-AFe and chi-squared test p-values for ripening time**. a** Delta-AFe values and **b** chi-squared test p-values from variants detected across the genome. **c-d** Zoom-in plots for signals on chromosome 3 and chromosome 4. Yellow bars indicate gene coding regions. Red bar outlines the *NAC18.1* coding region. The red dot is the D5Y SNP, a putatively causal non-synonymous mutation previously identified in the *NAC18.1* gene (Migicovsky et al., 2021) . “R” on the X-axis of the genome-wide plots indicates the “random” chromosome containing contigs that remain unanchored to the reference genome.

For phenolic content, four candidate regions were identified: chromosomes 4, 7, 8 and 16 harboured variants with delta-AFe values greater than 0.7 (Fig. 3-3a). The variant with the single highest delta-AFe between pools (0.784) was a C/T SNP at 3,857,519 bp on chromosome 4 (Fig. 3-3c), within the 3’-UTR region of a *Tetratricopeptide repeat (TPR)-like superfamily protein* gene (MD04G1034700). The signal on chromosome 4 is also within 11.5 kb of two *Transcriptional factor B3 family protein* (MD04G1034500, MD04G1034600) genes and a *glutathione peroxidase 2* (MD04G1034400) gene. A signal on chromosome 8 was identified (Fig. 3-3d) and contained a T/C SNP at 28,726,105 bp with the smallest p-value ($9.8x10^{-30}$) for phenolic content and a delta-AFe of 0.766. There were 7 variants with delta-AFe values above 0.7 in this region on chromosome 8. Of these variants, none were within coding regions of genes, and the nearest gene was *Ubiquinol-cytochrome C reductase hinge protein* gene (MD08G1223400) approximately 8.9 kb downstream. Another signal on chromosome 8 was identified containing 5 variants with delta-AFe values above 0.7, the highest of which is 0.77 at 11,325,518 bp. This group of variants does not fall within any annotated gene coding regions, but is within 5 kb of suspected coding regions of two genes of unknown function (MD08G1122700 and MD08G1122800). Two candidate regions on were detected on chromosome 16 (Fig. 3-3a,b): a single variant with the highest delta-AFe (0.77) on chromosome 16, and a group of variants forming a 49 kb window (3,839,333-3,889,319 bp) approximately 1.1 MB downstream of the aforementioned variant. The single variant was a T/C SNP at 2,727,461 bp, 46 bp upstream of an unannotated gene (MD16G1038200). Within the large window of variants on chromosome 16 (Fig. 3-3e), a C/A SNP at 3,864,330 bp had the smallest p-value (4.9x10^{-29}) and had the highest local delta-AFe (0.75). This variant was the only variant in the region with a delta-AFe greater than 0.7, while 7 other variants had delta-AFe > 0.6. While the SNP with the strongest signal in this region was not within the coding region of any gene, it was 668 bp upstream of a *UDP-Glycosyltransferase superfamily protein* (*UGT*) gene (MD16G1054500). Additionally, multiple variants within the candidate region on chromosome 16 were within coding sequences of *Tetratricopeptide repeat (TPR)-like superfamily protein* (MD16G1054700) and a *UGT protein* (MD16G1054400). Additionally, another 4 *UGT* genes (MD16G1054300, MD16G1054400, MD16G1054500, MD16G1054600) are within 7.1 kb of the variant with the highest delta-AFe at this locus. 358 unique genes, 188 of which had associated GO terms, were within candidate regions for phenolic content. We report the top 10 GO enrichment terms (Table S3), which included menaquinone biosynthetic processes, heme A biosynthetic processes, and polyamine metabolic processes.

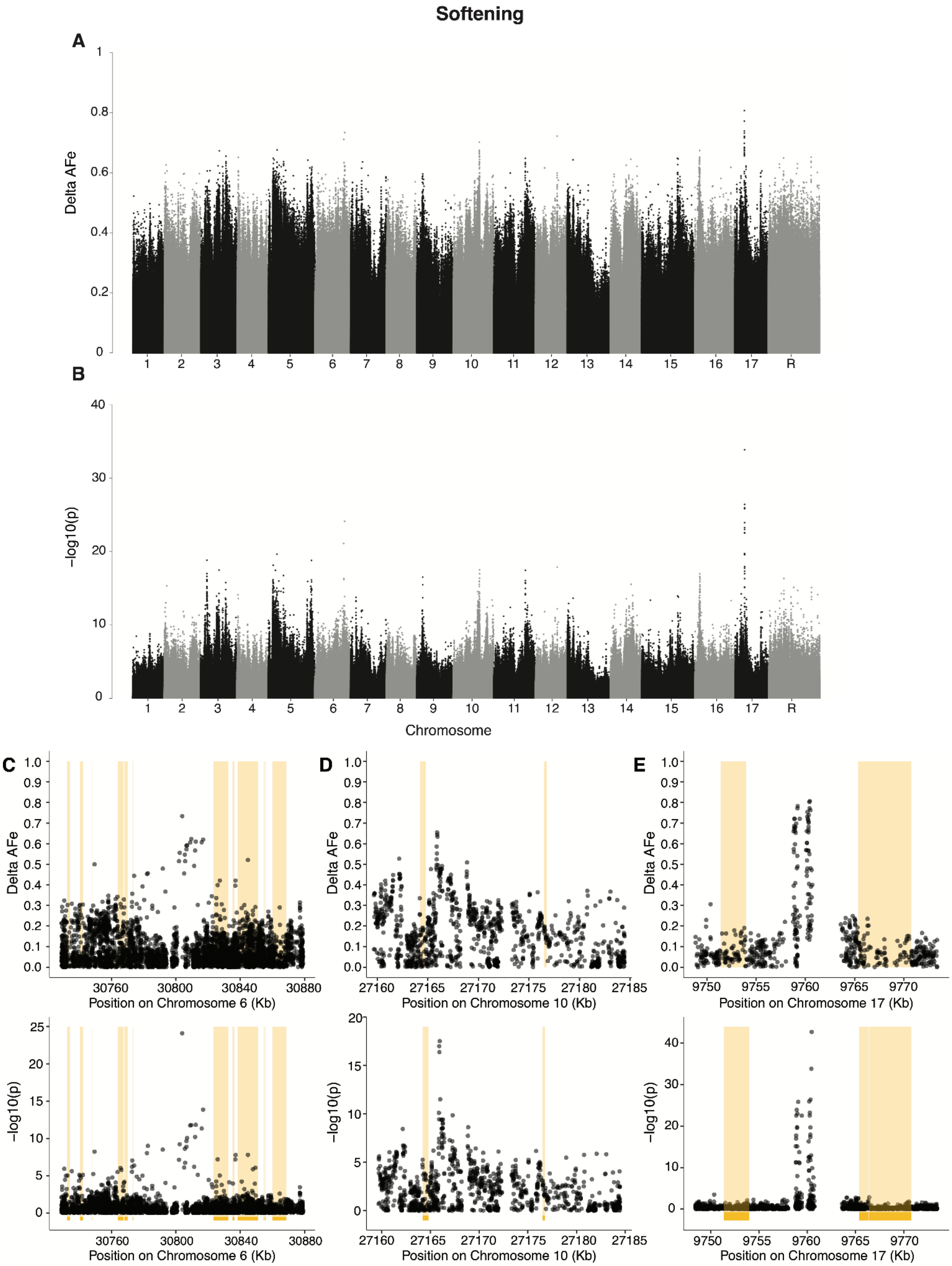

Softening
A
Delta AFe
0
0.2
0.4
0.6
0.8
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 R
B
−log10(p)
0
10
20
30
40
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 R
Chromosome
C
Delta AFe
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
30760 30800 30840 30880
Position on Chromosome 6 (Kb)
−log10(p)
0 5 10 15 20 25
30760 30800 30840 30880
Position on Chromosome 6 (Kb)
D
Delta AFe
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
27160 27165 27170 27175 27180 27185
Position on Chromosome 10 (Kb)
−log10(p)
0 5 10 15 20
27160 27165 27170 27175 27180 27185
Position on Chromosome 10 (Kb)
E
Delta AFe
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
9750 9755 9760 9765 9770
Position on Chromosome 17 (Kb)
−log10(p)
0 10 20 30 40
9750 9755 9760 9765 9770
Position on Chromosome 17 (Kb)

Fig. 3-3 Manhattan plots of delta-AFe and chi-squared test p-values for phenolic content. **a** Delta-AFe values and **b** chi-squared test p-values from variants detected across the genome. **c-d** Zoom-in plots for signals on chromosome 4, chromosome 8, and 16. Yellow bars indicate protein coding regions.

Candidate regions for apple softening during storage were identified on chromosomes 6, 10, 12, and 17 (Fig. 3-4a,b). The strongest signal for softening was on chromosome 17 (Fig. 3-4d), and the variant with both the lowest p-value ($1.95\text{x}10^{-43}$) and highest delta-AFe (0.807) for softening was a G/A SNP at position 9,760,456 bp on chromosome 17. This signal spanned an approximately 1.6 kb region, from 9,758,808-9,760,456 bp. Eight other variants in this region had delta-AFe values > 0.7. This signal overlaps with a gap (approximately 5.5 kb) of variants (Fig. 3-4d) as reads from the high softening pool, on average, failed to satisfy the minimum read depth cut off (average depth 38x) in this region, while reads from the low softening pool aligned to this region with sufficient depth (average depth 52x). The coding regions of a *Sterile alpha motif (SAM) domain-containing protein* (MD17G1113700), *vacuolar protein sorting 11* (MD17G1113900), as well as two other unannotated genes (Table S4) were within 10 kb of the signal on chromosome 17. The variant on chromosome 6 most strongly associated with softening was a C/T SNP at 30,803,965 bp, had a delta-AFe of 0.734 and a p-value of $7.8\text{x}10^{-25}$. The signal in this region spans roughly 12.9 kb (30,803,965-30,816,936 bp) (Fig. 3-4c). The nearest gene to this signal is approximately 18 kb downstream and encodes a 5S RNA (MD06G1167800). 321 unique genes, 158 of which had associated GO terms, were candidate regions for softening. We report the top 10 GO terms for enrichment for these genes (Table S4), which included mitochondrial fission, regulation of DNA-templated transcription, leaf senescence, and cold acclimation.

Total Phenolic Content

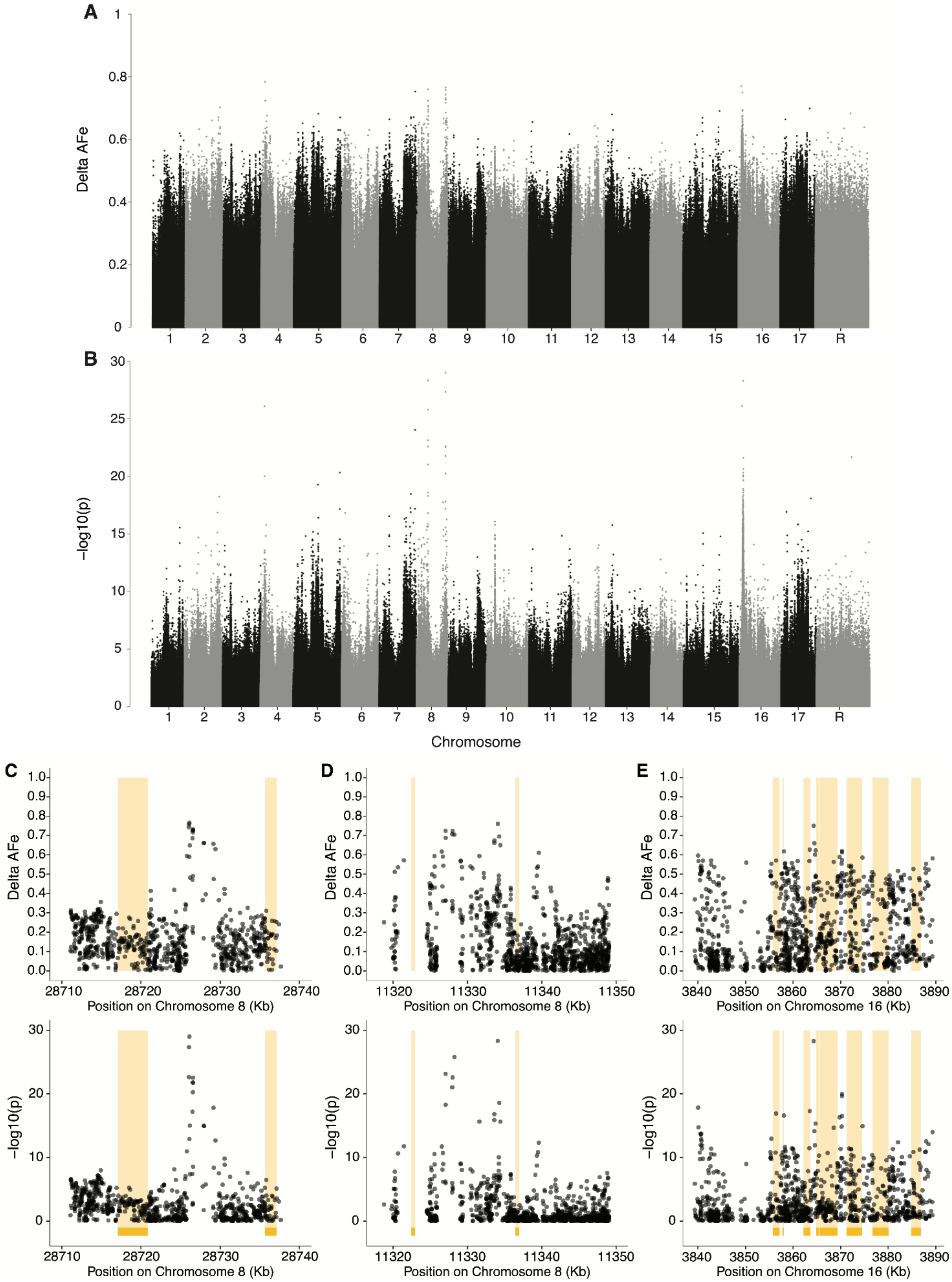


Fig. 3-4 Manhattan plots for genome wide delta-AFe and chi-squared test p-values for apple softening. **a** Delta-AFe values and **b** chi-squared test p-values from variants detected across the genome. **c-d** Zoom-in plots for signals on chromosome 6 and chromosome 17. Yellow bars indicate protein coding regions.

Discussion

We aimed to identify candidate genes and putatively causal variants underpinning three economically important apple phenotypes: ripening time, phenolic content, and softening. Here, we used a pool-seq approach (Kofler et al., 2011), a cost-effective WGS method that has been successfully employed to identify putatively causal alleles for phenotypes in other plant species (Dong et al., 2018; Ranavat et al., 2021; Welling et al., 2020), to scan the genome for regions of genetic differentiation between groups of samples with extreme phenotypes. Candidate regions discussed below were defined as regions of the genome within 20 kb of the strongest signals from our genome-wide scan for each trait.

Further evidence that *NAC18.1* impacts ripening time in apple and a novel signal on chromosome 4

There is strong evidence that ripening time in apple is controlled by a transcription factor on chromosome 3, *NAC18.1* (M. Jung et al., 2020; Larsen et al., 2019; Migicovsky et al., 2021), the homolog of the well-studied *NOR* ripening gene in tomato. Numerous variants in the coding region of this gene have been discovered in apple (Migicovsky et al., 2021), however, no strong evidence of causal variant(s) underlying ripening time has been revealed to date. Our pool-seq approach successfully identified a candidate locus encompassing the *NAC18.1* region on chromosome 3. The candidate region for ripening time is a roughly 80 kb window of variants showing high delta-AFe values (Fig. 3-2c). While hundreds of high delta-AFe variants exist within the coding region of nearby genes, including the previously identified nonsynonymous SNP D5Y within *NAC18.1* (Fig. 3-2c)(Migicovsky et al., 2021), the most extreme delta-AFe and CST p-values for ripening time did not lie within the coding region of NAC18.1. The strongest signal within the chromosome 3 window was 4.6 kb upstream of the gene *NAC18.1*, which suggests that the causal variants for ripening time may be regulatory variants impacting the expression of *NAC18.1*. Thus, our results suggest that ripening time in apple is likely impacted by genetic changes in regulatory elements that affect the expression of *NAC18.1* rather than non-synonymous changes to its coding region. Our findings here are similar to those in peach, in which genomic variation approximately 10 kb upstream of a NAC transcription factor has been found to influence the ripening period of peach fruit through modulated gene expression (Tan et al., 2021).

Additionally, a 4 kb gap in variant detection appears approximately 20 kb upstream of *NAC18.1* (Fig. 3-2c). Read depths from the early ripening time pool fell below the depth threshold in this region (see Methods), resulting in a segment within the chromosome 3 signal in which no variants were called (Fig. 3-2c). This gap in variant calling was caused by low sequence coverage in the early harvest pool, but not in the late harvest pool. This observation suggests that a deletion of sequence upstream of the *NAC18.1* locus may result in earlier harvested apples. This is consistent with observations in peach in which a tandem repeat variant is associated with elevated NAC expression in early-ripening accessions (Tan et al., 2021). Similar gaps in delta-AFe values were identified in a pool-seq approach examining cannabinoid synthesis in cannabis and suggest that presence/absence variants may be involved in that phenotype (Welling et al., 2020). A recent study demonstrated that the presence/absence of TEs can impact the regulation of transcription factors, ultimately influencing plant traits like flower colour (Y. Tian et al., 2022). Taken together, our results suggest genetic variation in the regulatory region of *NAC18.1* is likely playing a key role in ripening time in apple.

We also detected a candidate region on chromosome 4 for ripening time, which represents a novel locus for this phenotype. The strongest signal in this region does not include variants within coding regions of any nearby genes, but could indicate variants impacting gene regulation. The closest gene to the top hit on chromosome 4 is a *histidine kinase* (MD04G1013100), belonging to a family of multi-functional proteins that often play a role in signal transduction and cellular reception in plants (Hoang et al., 2021). Given that apple ripening is regulated in large part by cell signalling and plant hormones (Busatto et al., 2017; Seymour et al., 2013), it follows that variation in or near genes related to signal transduction and reception may lead to variation in ripening time. A recent RNA-seq study determined that genes on chromosome 4 likely impact ripening period in apple (Nawaz et al., 2021). The top hit on chromosome 4 in the present work is approximately 0.5 Mb downstream of a *Homeodomain-like superfamily* gene (MD04G1008300) that has been previously been linked to the early ripening phenotype in a mutant Hanfu apple variety (Nawaz et al., 2021). Because this *Homeodomain-like superfamily* gene is over 0.5 Mb downstream from our strongest signal on chromosome 4, it remains unclear if this gene and the signal detected in the present work are linked.

The signal on chromosome 4 was unexpected given that numerous previous genetic mapping studies of apple ripening time (Larsen et al., 2019; McClure et al., 2018, 2019) identified only a single peak near *NAC18.1* on chromosome 3 but never yielded a signal on chromosome 4 for ripening time. The signal on chromosome 4 was likely detected in the present study due to the higher marker density obtained here compared to previous studies that relied on relatively low-density genotyping-by-sequencing (GBS) data. The variants that make up the signal on chromosome 4 fall in a region of the genome that lacked markers completely in previous mapping studies (Appendix II-III). While this signal could be an artefact of erroneous read mapping or reference genome misconstruction, we suggest that this novel candidate region for apple ripening time on chromosome 4 is worthy of future investigation.

Signals for phenolic content production detected across the genome

Multiple candidate genomic regions for total phenolic content were detected, including signals on chromosomes 4, 7, 8 and 16 (Fig. 3-3a,b). This suggests a complex genetic architecture underlying total phenolic content involving numerous loci, consistent with both the way in which the phenotype was measured and the complexity of phenolic content production in apple fruit. Total phenolic content captures the total reductive potential of apple tissue and therefore measures the collective concentration of many phenolic compounds. Given that the measure of total phenolic content captures the cumulative reductive capacity of multiple secondary metabolites, it is unsurprising that we detect numerous candidate regions for this phenotype across the genome.

The candidate region containing the variant with the largest delta-AFe value for total phenolic content was detected on chromosome 4. This signal is a single SNP (Fig. 3-3c) in the 3'-UTR region of a *(TPR)-like superfamily protein* gene (MD04G1034700). TPR motifs facilitate protein-protein interactions and TPR-containing proteins have long been implicated in complex plant processes and plant hormone signalling networks including cytokinin and gibberellin responses as well as ethylene biosynthesis (Greenboim-Wainberg et al., 2005; Z. Lin et al., 2009; Schapire et al., 2006). Because the production of phenolic compounds is often linked to stress and various environmental cues, it is possible that this *(TPR)-like superfamily protein* plays a role in hormone signalling networks that influence polyphenol production. 3'-UTR regions are untranslated regulatory regions of mRNA, and 3'-UTR sequences can impact polyadenylation, translation efficiency, and stability of mRNAs (Srivastava et al., 2018). Therefore, while the exact role of *TPR-like superfamily protein* remains unclear, the variant detected here may be influencing translational regulation of the *(TPR)-like superfamily protein* and downstream total phenolic content production.

We also detected two candidate regions for total phenolic content on chromosome 8 (Fig. 3-3a,b). The first region, located at approximately 11.3 Mb, consisted of multiple variants with high delta-AFe values. However, none of these variants fell within protein coding regions and the nearest coding regions are unannotated. It is possible that one or both of the unannotated genes in the region are involved in phenolic content production, but without proper annotation, their involvement in phenolic content production remains uncertain. The second region on chromosome 8, located at approximately 28.7 Mb, consists of a peak of variants centered around a T/C SNP at 28,726,105 bp, which had the smallest CST p-value for the phenolic content phenotype. The nearest gene to this signal is a *Ubiquinol-cytochrome C reductase hinge protein* gene (MD08G1223400), approximately 8.9 kb downstream of the top SNP in the region. By measuring total phenolic content with the Folin–Ciocalteu assay, it is assumed that redox potential from substrates other than polyphenols is approximately constant across cultivars. However, if there is variation in reducing substrates other than polyphenols, then signals in the genome contributing to variance in non-polyphenolic substrates may be detected instead. Given that *Ubiquinol-cytochrome C reductase* encodes a key enzyme in the oxidative phosphorylation process within the mitochondria, the signal we detected at this locus may be picking up on genetic variation contributing to the amount of Ubiquinol-cytochrome C reductase produced in the cell rather than genetic variation contributing to phenolic content production. To our knowledge, while many other attempts to map phenolic content production in apple have been made (Chagné, Krieger, et al., 2012; S. A. Khan, Schaart, et al., 2012; S.

Kumar et al., 2022; McClure et al., 2019), only one has provided evidence for the involvement of chromosome 8 (S. A. Khan, Chibon, et al., 2012), suggesting that at least one of the signals found on chromosome 8 could be an artefact of measuring other reducing compounds in apple. Further investigations in discovering genes underlying phenolic content in apple would be wise to use phenotyping methods such as liquid chromatography–mass spectrometry or high-performance liquid chromatography, which can accurately quantify specific polyphenols.

Two candidate regions were also detected on chromosome 16 (Fig. 3-3a,b). The first region consisted of a cluster of variants around 3.8 Mb (Fig. 3-3e) and the second of a single variant at 2.7 Mb. The former cluster, a 50 kb window of variants with high delta-AFe and CST p-values (Fig. 3-3e), is roughly centred around a C/A SNP at 3,864,330 bp. Notably, there are 4 annotated *UGT* gene coding regions within 7.1 kb of this SNP. *UGT*s belong to a large gene family that produce glycosides by catalysing the transfer of sugar subunits between molecules (Lairson et al., 2008). Some UGTs are understood to catalyse the final steps in producing phenolic compounds in apple including phloridzin, quercetin glycosides, cyanidin pentoside, and kaempterol glycosides (Jugdé et al., 2008; S. A. Khan, Schaart, et al., 2012). One of the *UGT*s in this region is *UDP-glycosyltransferase 89B1*, also known as *flavonol 3-O-glucosyltransferase*, which catalyses the glucosylation of flavonols (Lim et al., 2004) and contributes to the production of diverse phenolic compounds (Holton & Cornish, 1995). Further, *flavonol 3-O-glucosyltransferase* has been previously implicated in the production of anthocyanin in strawberries and apple (Given et al., 1988; Ju et al., 1995).

Moreover, decreased expression of *flavonoid 3-glucosyltransferase* was found to be associated with lower anthocyanin production in sweet cherry (*Prunus avium*), a closely related species (Qi et al., 2022). This is consistent with previous linkage mapping studies in apple that have suggested *UGT*s as candidate genes for phenolic content production in apple (S. A. Khan, Chibon, et al., 2012). Taken together, the strong signal detected in this cluster of *UGT* genes suggest that *UGT*s on chromosome 16 may play a role in phenolic compound production in apple fruit. Further, variation impacting one or more *UGT*s on chromosome 16 could explain the QTL for kaempferol glycosides and phloridzin observed by Khan et al. (2012). We propose that *UGT*s on chromosome 16 represent strong candidate genes for polyphenol production in apples.

The variant with the single highest delta-AFe value on chromosome 16 was a T/C SNP at 2,727,461 bp. This SNP is approximately 1.1 Mb upstream of the *UGT* cluster discussed above, but only 189 kb upstream of *LAR1*, a gene identified by multiple previous studies as a strong candidate gene for phenolic content production in apple (Chagné, Krieger, et al., 2012; S. A. Khan, Chibon, et al., 2012; McClure et al., 2019). In other plant species, *LAR1* is directly involved in the production of catechin, a precursor component of procyanidins (Tanner et al., 2003). McClure et al. (McClure et al., 2019) suggested that *LAR1* may be involved in the production of many apple polyphenols after detecting signals near *LAR1* for multiple individually measured phenolic compounds. Linkage mapping experiments have also implicated a QTL hotspot on chromosome 16 for phenolic content that includes *LAR1* (Chagné, Krieger, et al., 2012; S. A. Khan, Chibon, et al., 2012). Khan et al. (S. A. Khan, Schaart, et al.,

2012) provided evidence that differences in *LAR1* expression, rather than coding region variation, was responsible for differences in polyphenol production among apple accessions. We did not detect a strong signal in the *LAR1* region in this study, however this is not the result of low sequence read depth in the *LAR1* region. It is possible that the SNP detected here is impacting a regulatory element and influencing *LAR1* expression, but given the distance between this variant and *LAR1* (189kb), we view this explanation as improbable. Instead, it seems more likely that this variant is picking up a signal related to another gene in the region, perhaps a transcription factor, that acts upstream of *LAR1*, as postulated by Khan et al (S. A. Khan, Schaart, et al., 2012). Despite the relatively high marker density employed in this experiment, the precise location of variants on chromosome 16 affecting phenolic content in apple remain unclear.

Multiple loci implicated in apple fruit softening

We found signals of allelic differentiation between softening pools on chromosomes 5, 6, 10, 16, and 17 (Fig. 3-4a,b). Evidence of loci on multiple chromosomes affecting softening is consistent with the hypothesis that fruit firmness is multigenic(Bink et al., 2014). Of the signals detected in the present study, those on chromosomes 6 and 17 were the strongest (Fig. 3-4a,b). The candidate region on chromosome 6 spans roughly 13 kb, with the nearest protein coding sequence 18 kb downstream. As none of the variants with the highest AFe values from this region were within the coding sequences of nearby genes, this signal may be detecting genetic variation in regulatory elements. While there are numerous genes within 20 kb in either direction of this signal (Table S4), a group of three *Tetratricopeptide repeat (TPR)-like superfamily proteins* (MD06G1168400, MD06G1168500, MD06G1168800) about 20 kb downstream are noteworthy. Proteins with TPR domains are common in plant hormone signalling (Kou et al., 2021; R. Kumar et al., 2014; Moya-León et al., 2019), and since fruit softening is largely driven through hormone-mediated ripening (Schapire et al., 2006), it could be that these *(TPR)-like superfamily proteins* are impacting softening related pathways in apple. Linkage experiments have identified QTLs for fruit firmness on chromosome 6 in the past (Bink et al., 2014; Liebhard et al., 2003), but could not identify putatively causal genes. Our results are in agreement with these linkage studies, and suggest that a locus on chromosome 6 plays a significant role in fruit softening.

We detected a candidate region for softening on chromosome 17 made up of two narrow regions of variants with high allelic differentiation between pools (Fig. 3-4d). The signal in this region is approximately 6 kb downstream of coding sequences for both a *Sterile alpha motif (SAM) domain-containing* (MD17G1113700) gene as well as a *vacuolar protein sorting 11* (*vps11*) (MD17G1113900) gene. The former is from a family of plant proteins that is still not fully understood, but known to function in a vast number of cellular processes in plants, from DNA protection to stomatal light response(Denay et al., 2017). The latter, *vps11*, belongs to a large family of proteins involved in diverse cellular processes from virus resistance to leaf growth and senescence in plants (Agaoua et al., 2022; Yamazaki et al., 2008). Interestingly, the narrow regions that make up the signal on chromosome 17 are formed of variants in a region of low variant detection (Fig. 3-4d) due to low read depth in the high softening pool. As seen in other pool-seq studies in plants (Welling et al., 2020), large differences in read depths between pools may indicate a region containing structural variation. Here, such a difference could indicate that the signal on chromosome 17 is driven by presence/absence variation or a complex genomic rearrangement responsible for variation in softening among accessions. This signal, and the discrepancy in read depth between pools, could represent a transposable or repetitive element that is largely absent in the high softening group, and present in the low softening group, for example. Further mapping studies using diverse germplasm with high-density marker data is required to understand the structure of this genomic region and its relationship to fruit softening.

Chromosome 10 has been suggested to harbour alleles responsible for apple fruit softening by multiple groups (Costa et al., 2010; S. Kumar et al., 2013; Longhi et al., 2012; McClure et al., 2018; X. Yang et al., 2022), and a signal on chromosome 10 is detected in the present study (Fig. 3-4a,b). Previous attempts to map fruit softening in the diverse apple population used here have detected SNPs associated with softening near an *ethylene response factor* (ERF)(MD10G1184800)(McClure et al., 2018). However, the strongest signal on chromosome 10 in the present study is 972 kb upstream of *ERF*, and closer to *PG1* (Appendix II-IV), a well-studied fruit firmness gene (Costa, 2015), which has been suggested by many groups as a promising candidate gene for apple softening (Costa et al., 2010; S. Kumar et al., 2013; Longhi et al., 2012).

In fact, a variant in *PG1* is considered by many as a “functional SNP” and is frequently used to predict firmness in apple germplasm (Baumgartner et al., 2016). Further, Di Guardo et al. (2017) have provided evidence that expression of *PG1* is correlated with apple fruit softening. Interestingly, the variant on chromosome 10 most strongly associated with softening in the present study is 451 kb upstream of *PG1*. This suggests that the signal we detect may be caused by a long-range regulatory element impacting *PG1* expression, consistent with the relationship proposed by Di Guardo et al (Di Guardo et al., 2017). However, given the density of genetic variants in the present work, the rapid LD decay in our population (Migicovsky et al., 2022), and the questionable utility of *PG1* variants to predict fruit firmness (Chagné et al., 2019; McClure et al., 2018; Migicovsky et al., 2021), it is also possible that this signal is detecting another gene that influences fruit softening nearby. The strongest signal detected on chromosome 10 in the present study is immediately upstream of a *Long-chain fatty alcohol dehydrogenase family protein* (MD10G1176100). This family of genes is known to be involved in the production of fatty alcohols, which contribute to forming plant cuticular waxes(Bernard & Joubès, 2013; Samuels et al., 2008). Plant waxes are important for preventing non-stomatal water loss (Riederer & Schreiber, 2001), and have been implicated in contributing to the storability of blueberry fruits (Chu et al., 2018). While there is some evidence that wax composition impacts apple softening (Chai et al., 2020), the link between the production of waxes on the peel of apple and fruit storability remains unclear. Nonetheless, we suggest that *Long-chain* *fatty alcohol dehydrogenase family protein* should be considered as a candidate gene for apple fruit softening.

The discovery of many regions of the genome associated with softening is in agreement with previous studies suggesting that this trait is multigenic. QTLs for softening have been mapped to chromosomes 5, 6, 10, and 16 (Amyotte et al., 2017; Bink et al., 2014; Costa, 2015; Di Guardo et al., 2017; Liebhard et al., 2003; McClure et al., 2018; B. Wu et al., 2021), all of which are detected in the present work. Given the complexity of fruit softening during storage and the number of loci discovered, our work is in agreement with previous suggestions that the genetic architecture of apple softening is multigenic.

It is worth noting that in the present study, DNA sequencing reads from each of the pools covered roughly 80% of the reference genome, meaning that nearly 20% of the positions in the reference genome were not considered in the present analysis. This leaves a considerable portion of the genome unexamined. Future works should aim to examine as much of the genome as possible, perhaps through the use of pangenomes or alternative DNA sequencing methods.

Conclusions

To date, there have been few causal alleles discovered in apple. With the rise of gene editing technologies and the continued need for improved fruit varieties, the discovery of causal alleles is key for accelerated fruit improvement. In this study, we scanned the apple genome for genetic differentiation between groups of phenotypically divergent samples with the aim of finding regions of the genome responsible for variation in ripening time, phenolic content production, and fruit softening. Our study provides further evidence that *NAC18.1* is involved in controlling ripening time, suggests that genetic variation impacting the expression of *NAC18.1* may be causal, and implicates a novel locus for ripening time on chromosome 4. Further, this investigation identified multiple loci across the genome related to phenolic content production, and suggests that a cluster *UGT* genes on chromosome 16, among others, may be responsible for variation in phenolic content production. Finally, the strong signals detected on multiple chromosomes in the present work suggest a complex genetic architecture for softening, and implicates many candidate genes, including a gene related to fruit wax production on chromosome 10. Across all phenotypes, there is a recurring theme: many signal peaks appear in non-coding regions of the genome in relatively close proximity to genes. This indicates that regulatory variants are likely to play a larger role in the control of plant phenotypes than I had anticipated or predicted. Together, the genomic resolution provided by the data in this work sheds light on the genomic control of important phenotypes and will support future efforts to enable genomics-assisted improvement of apples.

The pool-seq experiment in this chapter was successful in generating numerous novel insights, and provided strong evidence that the causal variant for ripening time is in the regulatory region upstream of *NAC18.1*. Further, the pool-seq approach emphasises the power of WGS for causal allele discovery: it is clear that the ability to examine the genome at the nucleotide level is essential for identifying causal alleles. This is well evidenced by the relatively narrow signal peaks in the present study, which allow for delimiting of small regions of the genome likely to harbour causal alleles. However, while the pool-seq experiment from the present chapter does provide a number of important findings, causal alleles cannot be identified with confidence in part due to key drawbacks inherent to the pool-seq approach. First, pool-seq data cannot link specific DNA sequences to the samples in which they are derived. This means that variants and haplotypes cannot be discerned from the genetic information on a sample specific basis. Second, the pool-seq approach does not accurately consider insertions and deletions, and fails to detect some genetic variation due to extreme read depths, thereby disregarding key variant types known to impact phenotypes. In the future, methodologies that make use of WGS from many samples will be key for generating detailed genomic data that is high resolution and sample specific. Under ideal circumstances, a large diverse apple population is sequenced to high depth and queried for variants. Then, association mapping methods could leverage the high resolution WGS data and a large population size to effectively map causal alleles with precision. However, such projects require substantial budgets and smaller, more manageable projects are likely to be the most pragmatic path forward.

Data availability

The datasets generated during and/or analysed during the current study are available in the NCBI SRA repository (https://www.ncbi.nlm.nih.gov/sra/PRJNA929465).

Author contributions

Thomas Davies[1]

ROLES: Conceptualization, Data collection, Data curation, Formal analysis,

Visualization, Writing – original draft, Writing – review & editing

Sean Myles[1]

ROLES**:** Conceptualization, Writing – review & editing

[1]Department of Plant, Food, and Environmental Sciences, Faculty of Agriculture, Dalhousie University, Truro, NS, Canada

Chapter 4: Mapping of the causal allele(s) controlling ripening time in apple

Rationale

After analysing the results and considering the strengths and weaknesses of the genetic mapping approaches used thus far (see Chapter 3: Conclusions), it became clear that whole genome sequencing across many samples from the ABC was going to be essential for effective causal allele discovery. Therefore, sequencing a highly diverse group of samples from the ABC to high depth is a logical and reasonable step towards mapping causal alleles. The generation of WGS from a diverse subset of samples from the ABC provides the opportunity to reach two primary outcomes. First, it immediately provides enough genetic information to conduct a high resolution genome wide association study for ripening time. Second, the WGS data can serve as a “reference panel” for genotype imputation across the rest of the ABC in the future. A reference panel is a subset of samples from the ABC sequenced to high depth that can be used to fill in missing or low confidence genotype calls for the remainder of the ABC population through a process called imputation. Imputation using a high quality reference panel enables high accuracy genotype information to be generated for an entire population at significantly lower cost. The generation of a reference panel represents a significant investment in the mapping potential of the ABC in the future, and enables association mapping of numerous traits measured in the population at a later time.

With regards to ripening time, multiple studies have indicated that this phenotype is likely controlled by a single large effect locus on chromosome 3 (Davies & Myles, 2023; M. Jung et al., 2020; Larsen et al., 2019; Migicovsky et al., 2021; Watts et al., 2023), making ripening time a strong candidate for a WGS GWAS using reference panel samples. A more detailed association study aimed at ripening time would provide evidence for either a coding or regulatory variant as the causal allele, helping to resolve ambiguous lines of evidence from the previous reports (Davies & Myles, 2023; Migicovsky et al., 2021; Watts et al., 2023). Ideally, an association study leveraging high depth sequencing and a diverse sample set provides enough resolution to uncover the causal allele(s), although complex genetic variation may remain hard to query. In this chapter, I select a diverse group of samples from the ABC and generate a high quality reference panel from high depth WGS data. Then, I conduct a genome wide association study with the aim of discovering the causal allele for ripening time.

Abstract

Elucidating the causal allele(s) controlling ripening time in apple is important not only for fully understanding the genetic control of the trait, but also for enabling the use of precise breeding techniques and gene editing technologies. Genetic data was collected from 97 of the most genetically diverse samples from Canada’s Apple Biodiversity Collection and sequenced at high depth using short read whole genome sequencing. Genetic polymorphisms were called from DNA sequencing data and resulted in 49M genetic variants across the apple genome. A genome wide association study was conducted and identified a 45.7kb signal on chromosome 3 associated with ripening

time. Within this signal region, transcription factor *NAC18.1* is the only annotated gene. Genetic variants 11.3kb upstream of *NAC18.1* produced strongest associations with ripening time, and suggest that genetic variants in the upstream regulatory region control ripening time in apple. A read-depth analysis across samples suggests that another region, also in the *NAC18.1* upstream region, could be influencing ripening time. Together the results here suggest that genetic variation in the promoter or upstream regulatory region of *NAC18.1* likely harbours the causal allele(s) controlling ripening time. Finally, the region associated with ripening time in apple was analysed to illustrate the potential viability and challenges of employing gene editing technologies based solely on association study results. Importantly, the genetic variants produced in this chapter represent a high quality reference panel that will be well suited for the imputation of variants across the remainder of Canada's Apple Biodiversity Collection in the future, laying the groundwork for some of the largest and most detailed association mapping in apple in the near future.

Introduction

The breeding of improved fruit cultivars is the primary objective of fruit breeding science. Consumer desires, supply chain preferences, environmental conditions, agricultural practices, and governmental policies are a few of the many factors considered by modern breeders and experimentalists attempting to produce novel improved fruit cultivars (Dalla Costa et al., 2017; Finger & Möhring, 2024; Iezzoni et al., 2020; Ru et al., 2015). In recent decades, apple fruit breeding science has largely transitioned from traditional trial-and-error breeding methods to genomics-informed breeding approaches that glean valuable information from plant DNA sequences (S. Kumar et al., 2012; Peace et al., 2019; Troggio et al., 2012). Today, the discovery of causal alleles — the DNA nucleotide sequences that control traits — is an important step for accelerated apple cultivar improvement (Iezzoni et al., 2020). A detailed understanding of causal alleles controlling key traits not only offers key information for marker assisted selection (MAS) and genomic selection (GS), but it is necessary to enable novel gene editing technologies, such as Clustered Regularly Interspaced Palindromic Repeats (CRISPR) (Jinek et al., 2012). Despite the clear value of discovering causal alleles in apple, few are known to date, rendering the effective application of the newest gene editing technologies challenging in apple. If apple trait improvement efforts are to experience significant acceleration in the near future, causal alleles controlling key traits must be confidently mapped and characterised.

Locating the precise genomic identity of causal alleles remains a significant challenge for biologists. In recent decades the use of statistical models, particularly genome wide association studies (GWAS), that detect associations between genetic markers and phenotypes have risen in popularity (Abdellaoui et al., 2023). Association studies in plants have a number of advantages over traditional mapping techniques (Korte & Farlow, 2013). First, because association studies do not require plant crosses or detailed pedigree information, large datasets are often possible to collect. Second, since large groups of unrelated samples are typically used in association studies, generations of historical recombination can be leveraged to generate high mapping power and genomic resolution. Well designed GWA studies can be powerful enough to discover causal alleles across the tree of life, as seen in bears, cannabis, and humans (Kenny et al., 2012; Leckie et al., 2023; Puckett et al., 2023).

The ability of GWAS approaches to detect causal alleles directly from genotype and phenotype data means that, in theory, useful breeding markers and gene editing targets can be used by breeders without the need for expensive functional investigations. The key to efficiently identifying causal alleles via GWAS-based genetic mapping in diverse populations is high variant density: only the query of genetic variants that cover the genome at high density will enable the discovery of causal alleles from genotype and phenotype data. Therefore, to generate the genotype data necessary for such experiments, genetic variants of high quality and density must be collected from the target population. Collection and generation of genomic data that can support high density variant calling typically requires the generation of whole genome sequence (WGS) data at considerable depth (>10x), which can be prohibitively expensive. Further, effective use of GWAS models requires a diverse population with enough historical recombination that linkage disequilibrium (LD) decays rapidly at short genomic intervals. Establishing, maintaining, and accessing such a population is both expensive and laborious, and has historically been a major barrier for conducting a high powered GWAS in long-lived woody perennials like apple (Iezzoni et al., 2020). Together, the cost of high coverage genotype data and the access to a diverse population have been the primary barriers to high resolution and high power GWAS-based mapping attempts in apple.

Fortunately, recent advances in bioinformatics, specifically in the area of genotype imputation, have made it possible for genotype data from large populations to be accurately imputed from low-pass DNA sequencing at significantly reduced cost (Buckley et al., 2022; J. H. Li et al., 2021, 2023; Martin et al., 2021; Snelling et al., 2020). Low-pass imputation uses bioinformatic algorithms to make informed inferences about the genotype identities of samples sequenced at low depth (<1x). However, this method is only viable provided a reference panel of high sequencing depth (>10x) and quality is established first, typically from a diverse subset of the larger population. Low-pass imputation methods require a reference panel as the basis for imputation, and use the high quality variant information from the reference panel to fill in genotypes at variant sites among samples sequenced at low depths. Through this method, a reference panel from a small number of samples (<100) can be leveraged to generate accurate genotype information across large populations (Buckley et al., 2022; Snelling et al., 2020). Therefore, if cost effective GWAS mapping approaches are to be used in the future, it is important for high quality variants to be called from diverse reference panels, particularly in organisms where DNA sequencing costs remain a primary barrier to genetic mapping. As such, a reference panel for Canada’s Apple Biodiversity Collection (ABC) would set the stage for accurate imputation across thousands of samples, and open the door for arguably the most powerful genetic mapping experiments in apple to date.

Association mapping studies have discovered causal alleles in apple, including malic acid production (Bai et al., 2012) and budbreak (Watson et al., 2024) . Discovery of causal alleles controlling key agricultural traits holds value for apple breeders as the fruit industry continues to put efforts towards producing new and improved apple varieties. Of the traits most deeply investigated in apple, the genomic control of ripening time has garnered attention for its impact on harvesting logistics and regional growability. A substantial body of scientific work focuses on the genetic control of ripening time in apple (Larsen et al., 2019; Migicovsky et al., 2021; Watts et al., 2023; G. Zhang et al., 2018). As climates in temperate regions of the world change and experience increased severe weather events, alterations to the ripening time of apple trees will be key for breeding cultivars that are climate-change resistant and well suited for existing apple production regions (Pfleiderer et al., 2019). Despite decades of previous investigations into the molecular basis of ripening time, including QTL mapping (Chagné et al., 2014; Liebhard et al., 2003), GWAS (Larsen et al., 2019; Watts et al., 2023), and transgenics (Migicovsky et al., 2021), the causal alleles for ripening time in apple remain unknown. It is clear that *NAC18.1*, a transcription factor on chromosome 3, plays a key role in the ripening time of apple cultivars. However, the location and nature of the different alleles at the *NAC18.1* gene that control ripening time remain unidentified. While mutations to the coding sequence of the gene have been suggested as causal (Migicovsky et al., 2021; Watts et al., 2023), there is also evidence suggesting a genetic mutation in the promoter region of the gene influences variation in ripening time (Davies & Myles, 2023).

The purpose of this chapter is twofold: to generate a high quality reference panel for future imputation efforts at Canada’s Apple Biodiversity Collection (ABC), and to conduct a genetic mapping experiment to elucidate the causal allele(s) controlling ripening time in apple. To achieve this, I generated WGS data from a diverse subset of samples from Canada’s ABC to 20x depth. I then implemented a bioinformatics pipeline to generate high-quality and high-density genotype data from across the genome. The result is, to my knowledge, the highest resolution genetic variant data produced in apples to date. Then, I used a GWAS-based mapping approach using this diverse group of apple cultivars from Canada’s ABC in an attempt to map the causal allele for ripening time.

Materials and methods

Sample selection and GWAS power analysis

To determine if a GWAS using a small number of samples (109) from the ABC yielded enough power to detect large effect loci, a GWAS for ripening time was performed using 250k SNPs and 109 diverse samples. To select a group of 109 samples that maximise genetic diversity for a GWAS, SVcollecter (Ranallo-Benavidez et al., 2021) was employed. Genome-wide SNP data from previous work (Migicovsky et al., 2022) was randomly downsampled to 125k genome wide SNPs across the 109 samples. Phenotype data was retrieved from previously published literature (Watts et al., 2021). A sample structure matrix (k-matrix) was generated from the bed, bim, and fam PLINK files via PLINK. A population stratification matrix (q-matrix) was generated in R using the princomp function. The PLINK files, k-matix, q-matrix and phenotype data were used to perform a GWAS using a linear mixed model (LMM) via GEMMA (Zhou & Stephens, 2012) with the flag -gk and a minor allele frequency (MAF) threshold of 0.05.

Manhattan plots and QQ plot for the GWAS were visualized via the qqman package in R.

DNA collection

Fresh bud tissue from selected samples was collected from the ABC in the spring of 2022 and sent to Platform Genetics (Vineland, Ontario) for DNA extraction and library preparation. DNA libraries were sent to Gencove (New York, New York) for Illumina NovaSeq paired end (PE) short read (150bp) sequencing to an average depth of 20x per sample. Out of the 109 samples sent to Gencove for sequencing, 12 failed quality tests, primarily due to low DNA yield from extractions, and data from 97 samples were thus used for downstream analyses.

DNA sequencing pre-processing, read alignment, and variant calling

DNA data was processed using a custom bioinformatics pipeline generated for this project (Figure 4-1). Raw DNA reads in the form of fastq files were retrieved from the Gencove platform and the quality of PE reads for each sample were accessed using the quality control software FastQC (Andrews, 2010). Library adapter sequences were removed with Trimmomatic (Bolger et al., 2014) using the ILLUMINACLIP function and sequences with low quality read scores were removed with the SLIDINGWINDOW function using a window size of 5 and a minimum average base phred score of 20. Sequencing read pairs and read depth distributions are shown in Appendix III-II. PE reads were aligned to the Golden Delicious reference genome GDDH13 version 1.1 (Velasco, Zharkikh, Affourtit, Dhingra, Viola, et al., 2010) using the Burrows-Wheeler Aligner (BWA) (H. Li & Durbin, 2009) using the ‘mem’ algorithm and the -M flag. Sequence Alignment Map (SAM) files for each sample were coordinate sorted using the Picard SortSam tool (http://broadinstitute.github.io/picard/). SAM files for each sample were scanned for duplicate reads using Picard MarkDuplicates and duplicated reads were assigned appropriate flags. SAM files were converted to Binary Alignment Mapping (BAM) files to optimise storage and efficiency in downstream processing. Variants were called from sample-specific BAM files using GATK HaplotypeCaller in GVCF mode, using a local re-alignment variant calling algorithm (Van der Auwera & O’Connor, 2020), producing GVCF files for each sample. GVCF files from each sample were combined by chromosome to optimise computational efficiency (i.e. variants from each of the 18 chromosomes across samples were combined separately) using GATK GenomicsDBImport, creating a Genomics Data Base (GDB) for each of the 18 chromosomes. GATK GenotyopeGVCF was used in GDB mode to perform joint genotyping for each chromosome, resulting in a Variant Calling Format (VCF) file for each chromosome containing all variants across all samples. From each chromosome VCF file, SNPs and indels were extracted using GATK SelectVariants to include biallelic SNPs as well as indels up to 60bp in length. Picard MergeVCFs was used to combine SNPs and indels, respectively, from each chromosome into separate VCF files. MergeVCFs was used to combine SNPs and indels from across all samples to produce a final dataset of raw variants in a single VCF file. All bioinformatic steps were performed using custom bash scripts to reduce computational load and runtimes on the ComputeCanada Beluga cluster.

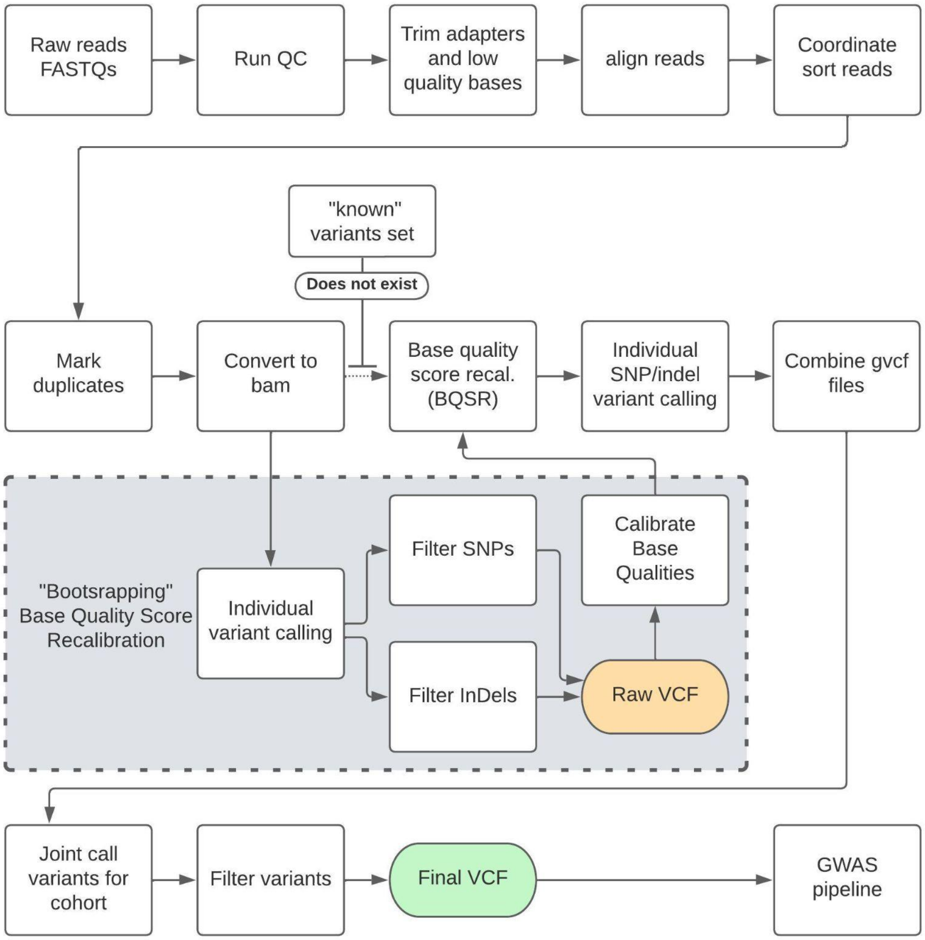


Fig 4-1. Bioinformatics workflow. Yellow box indicates the raw (unfiltered) variant dataset. The green box indicates the final filtered variant dataset.

Genotype quality control

Genotype quality filtering was performed according to GATK best practices (Van der Auwera & O'Connor, 2020), with some adjustments based on the structure of the raw variant data. Thresholds for Base Quality Score Recalibration (BQSR) and bootstrapping of SNPs were QD < 1.0, FS > 60.0, MQ < 40, MQRankSum < 12.5, and ReadPosRankSum < 8.5. Thresholds for BQSR and bootstrapping of indels were QD < 1.0, FS > 200.0, and SOR < 10.0. Base recalibration tables were generated for each sample using GATK BaseRecalibrator and the filtered SNPs and indels generated previously. BAMs from each sample were then recalibrated using GATK ApplyBQSR with the -bqsr flag, producing recalibrated BAMs for each sample.

Following BAM recalibration, variants were re-called from BAM files using the same steps as outlined in the previous subsection, resulting in a final dataset of high quality variants from across the cohort contained in a single VCF file. BCFtools-vstats software package was used to generate quality metrics across variants (H. Li, 2011). The final variant dataset was filtered to only include biallelic SNPs and indels. Genotype concordance was calculated between the final WGS VCF file and previous GBS data from the population (Migicovsky et al., 2022) using BCFtools stats package (H. Li, 2011).

Genome-Wide Association Study

Phenotype data was retrieved from previous work (Watts et al., 2021), and formatted in R to conform to the standards required by GEMMA. 76 of 97 samples had ripening time data and were suitable for GWAS. A sample structure matrix (k-matrix) was generated from the VCF file via PLINK and GEMMA with the flag -gk. A population stratification matrix (q-matrix) was generated in R using the princomp PCA function. The VCF, k-matix, q-matrix and phenotype data were used to perform a GWAS using a linear mixed model via GEMMA using a minor allele frequency (MAF) threshold of 0.02. Manhattan plots and QQ plots for the GWAS were visualised via the qqman package in R. Zoom-in plots were generated using the ggplot2 package (Wickham, 2016) and p values were multiple test corrected with a Bonferroni correction. Genic regions shown in manhattan plots were taken from the Golden Delicious reference genome annotation (*Genome Database for Rosaecea*, 2023). The peach structural variant homologous region was located by retrieving the causal SV from (Tan et al., 2021) and aligning the variant region to the apple reference genome using NCBI blastx algorithm (Altschul et al., 1990).

Sliding window WGS read depth analysis

Read depth at each genomic position in the genome was extracted from each sample using Samtools depth using the -a flag. A sliding window algorithm was applied to the entire genome to calculate the mean read depth for non overlapping 100bp windows across the genome for each sample. Window read depths were then normalised by subtracting the mean and dividing by the standard deviation of all windows across the genome. A Pearson correlation test was conducted between mean window read depths and the ripening time for each sample, using the cor.test function in R. Results of the correlation test were visualised using the plot function and p values were corrected for multiple comparisons using the Bonferroni correction.

Pooled read depth analysis

Read depths at all positions in the genome were extracted from pooled sequencing data from the harvest date pools generated in Chapter 3. For each genomic position, read depths were standardised by subtracting the genome-wide mean read depth and dividing by genome-wide standard deviation using the ‘scale’ function in R. The absolute value of the difference between standardised read depths of pools was calculated and plotted.

Linkage disequilibrium comparisons

Linkage disequilibrium (LD) across the 76 WGS samples was calculated for all variants within 1kb of one another using PLINK. The WGS VCF file was subsampled down to contain only the 17 SNPs in the *NAC18.1* region captured by GBS data (Migicovsky et al., 2022) using VCFtools --snps function. VCFs containing only those 17 SNPs across all available samples (76 for WGS, 1116 for GBS) were converted into hapmap file format using Tassel (Bradbury et al., 2007), and pairwise LD was calculated between all 17 SNPs using the LDHeatmap package in R (Shin et al., 2006). Visualisations were produced with ggplot2 (Wickham, 2016) and LDHeatmap packages in R. The p values generated by the MLMM GWAS for GBS data were taken directly from Watts et al. (2023).

CRISPR editing sites

The reference genome sequence from the 50kb signal region on chromosome 3 (30680020-30720000bp) was analysed for CRISPR/cas9 target sequences using the CRISPRscan software (Moreno-Mateos et al., 2015). CRISPR target sequences were counted by CRISPRscan quality score, defined as <40 = low, 40-60 = medium, >60 = high quality.

Code availability

All custom scripts generated for the materials and methods of this chapter are available at https://github.com/MylesLab/Ripening_time_WGS.

Results

Sample selection

To examine the overall genetic diversity of the ABC sample pool and ensure that the samples selected for the experiment were not biassed, I selected 109 diverse samples from the ABC that capture the maximum genetic diversity from the ABC. First, the genetic landscape of the ABC population was visualised via principal components analysis (PCA) (Fig 4-2). Samples from the ABC population were then chosen using an algorithm designed to select subpopulations that capture the maximum genetic diversity from large populations, and plotted in PC space to ensure they captured the majority of the genetic landscape. To ensure that the subset of 109 samples capture most of the variation in ripening time across the ABC, a comparison was made between ripening times of the subset of 109 samples and the ripening times of 837 samples from the ABC for which ripening time data were available. The result (Appendix III-I) suggests that the method employed to select the subset of 109 samples was relatively unbiased in terms of ripening time as the comparison between ripening time distributions described above (Appendix III-I) revealed no significant difference (p = 0.06).

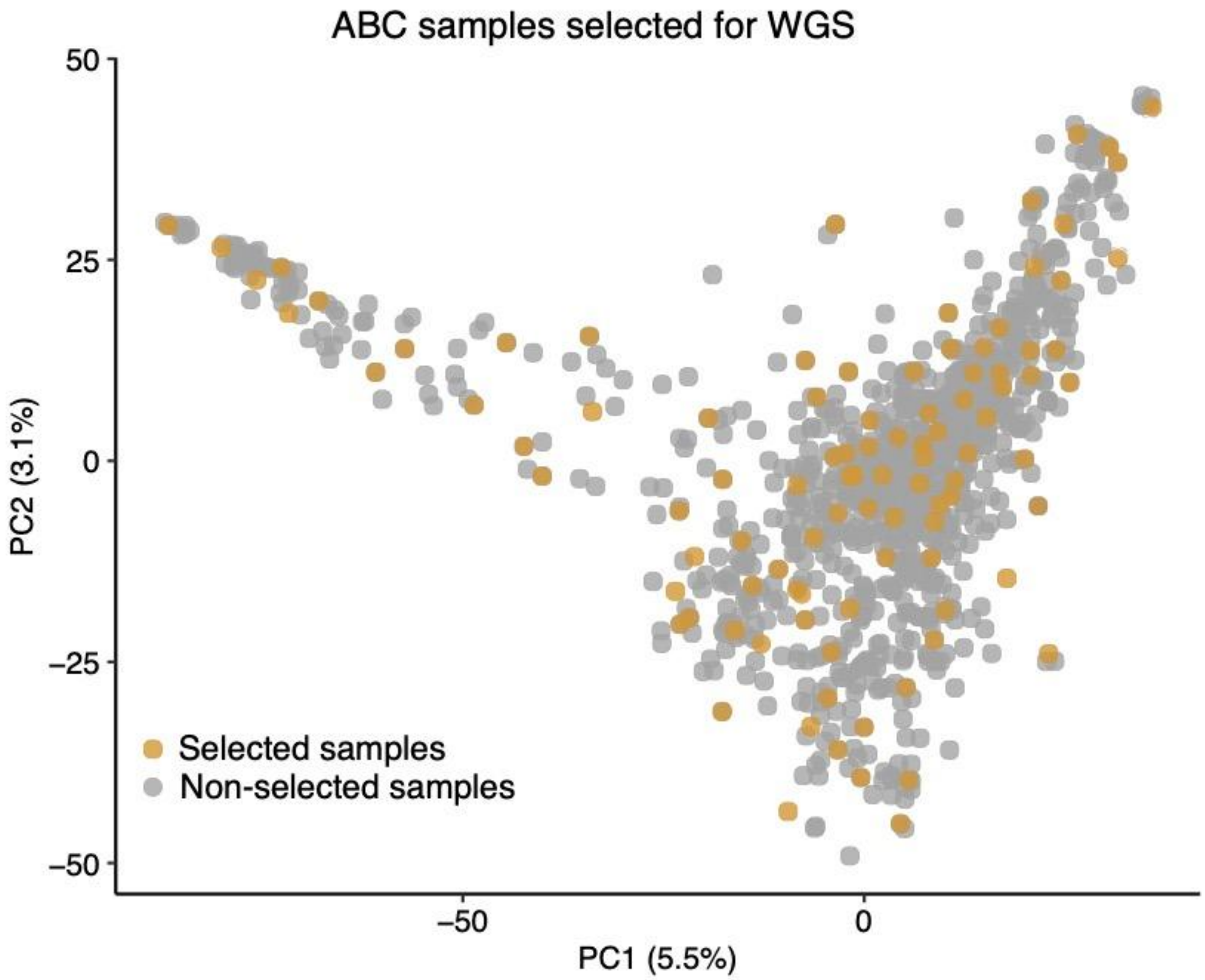


Fig 4-2. Genomic PCA plot for ABC population on the first two principal components using genotype data from previous work (Migicovsky et al., 2022). Each dot represents a sample from the ABC population in genomic principal component space. X and Y axes represent the first and second PCs, respectively, and the variance explained by each PC is shown in parentheses. Samples highlighted in orange were selected for WGS downstream.

GWAS power analysis

To investigate the degree to which whole-genome sequencing paired with GWAS can lead to the discovery of large effect loci even when using small sample sizes, I selected 109 samples from the full set of 1116 samples from the ABC (described above) and ran GWAS for ripening time, which has a known large effect locus on chromosome 3. To determine if GWAS signals for large effect loci may be detectable even in relatively small sample sizes from the population studied here, a Linear Mixed Model (LMM) using 125k genome-wide GBS SNPs from 109 selected samples was conducted to ensure the proposed experiment would yield enough power before proceeding with WGS sequencing. The GWAS proved sufficient to detect a strong signal on chromosome 3 previously associated with ripening time (Fig 4-3). The top SNP associated with ripening time was the D5Y SNP, as previously discovered by Migicovsky et al. (2021).

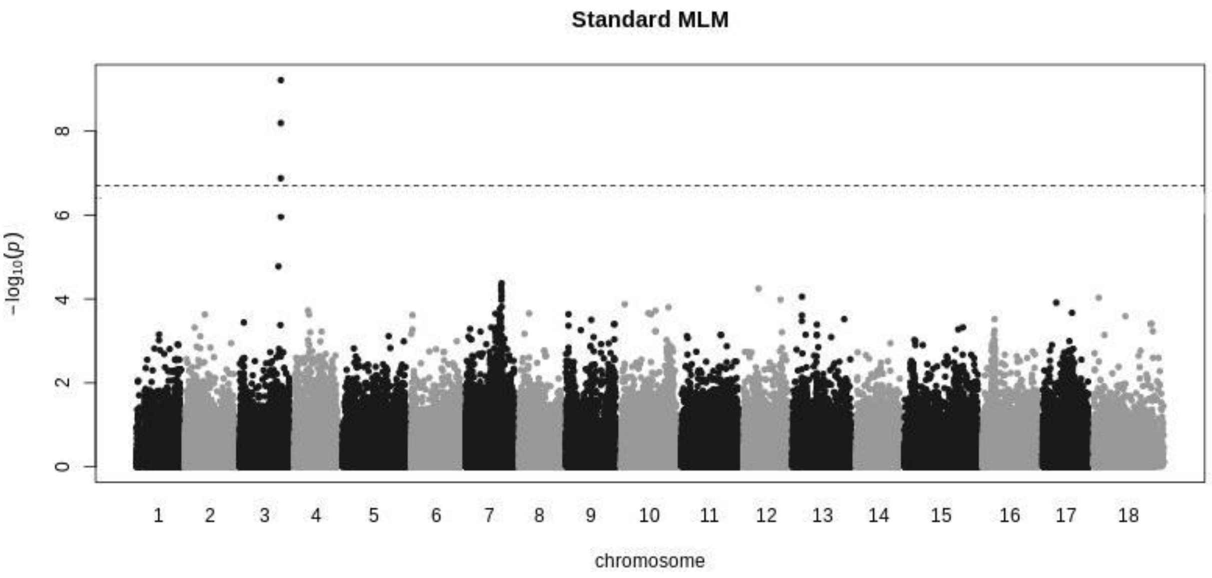


Fig 4-3. Genome wide manhattan plot from a standard MLM using 125k GBS SNPs and ripening time data from 109 diverse samples from the ABC. Each dot represents a SNP.

The horizontal dotted line is the Bonferroni-corrected significance threshold. The SNP most significantly associated with ripening time is the same SNP that was discovered previously in other studies (Larsen et al., 2019; Watts et al., 2023).

WGS and the final variants

The mean read depth across the 97 WGS samples was 17.5x (Appendix III-II). The final dataset of high quality variants from these samples in VCF format contained 43,506,958 SNPs and 5,785,765 indels, for a total of 49,292,723 biallelic variants across the genome. A distribution of indel lengths called from the variant calling pipeline is shown in Appendix III-III. SNP abundance per chromosome varied between 1.8M and 3.1M, and indel abundance per chromosome varied between 252k and 422k (Appendix III, Table I). Mean variant frequency across the genome was one variant/14.4bp or 69.4 variants/kb (Appendix III, Table I). The mean minor allele frequency (MAF)(Fig 4-4) at variant sites was 0.07. Between-sample genotype non-reference discordance (NRD) between WGS samples and GBS sequenced samples was 39.6%. NRD reported here was calculated as the ratio of mismatching sites and the total number of sites, excluding matching homozygous reference allele sites (Appendix III-V).

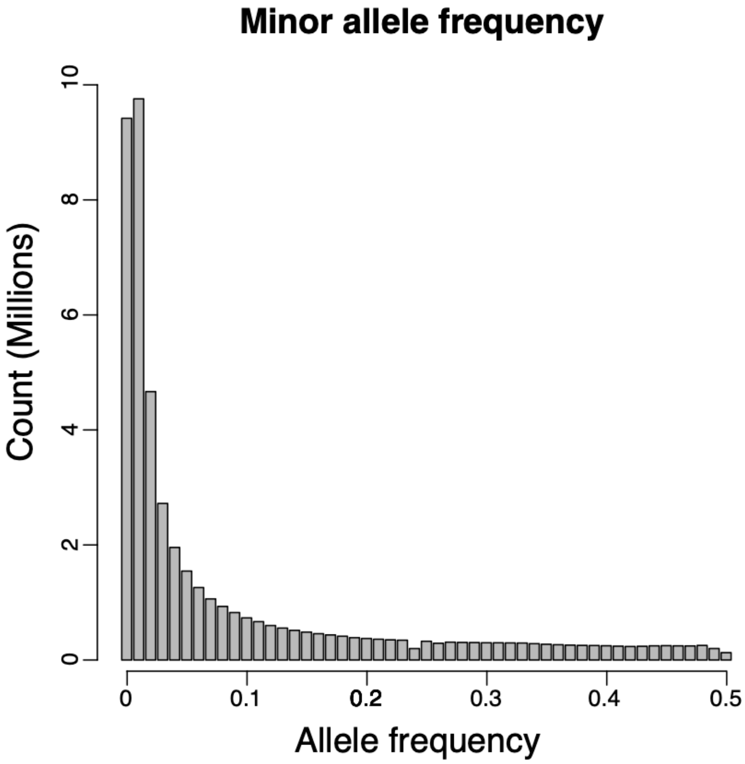


Fig 4-4. Minor allele frequency distribution at variant sites across samples, which totals 43,506,958 SNPs and 5,785,765 indels.

GWAS

To conduct the GWAS on samples with WGS data, LMM was run using the GEMMA software. After filtering variants (MAF = 0.02, missingness = 0.05), a total of 25,769,077 variants were used to perform a GWAS with 76 samples that had ripening time phenotype data. The Linear Mixed Model revealed a single significant peak on chromosome 3 (Fig 4-5A), which consists of a 45.7 kb region (30681.7-30727.5kb) containing 2,832 variants (Fig 4-5B), 343 of which were significantly associated with ripening time after multiple test correction. Within this peak three variants shared the lowest p-value ($3.0x10^{-11}$), 30709586 (T/C SNP), 30709593 (C/T SNP), and 30709888 (T/A SNP) (Fig 4-6). The minor allele frequencies of each of these SNPs was also the same (0.32). For these three variants, there are only two haplotypes present in the samples (haplotype A: T/C/T, haplotype B: C/T/A)(Fig 4-6). These variants are approximately 11.3kb upstream of *NAC18.1*, which is the nearest gene coding region. Many significantly associated variants within this peak formed a roughly 45.7kb block of variants that constitutes the signal on chromosome 3. The QQ plot from the model indicated a mild skew towards low p values (Fig 4-5C). The genomic inflation factor (lambda value) for this model was 1.045.

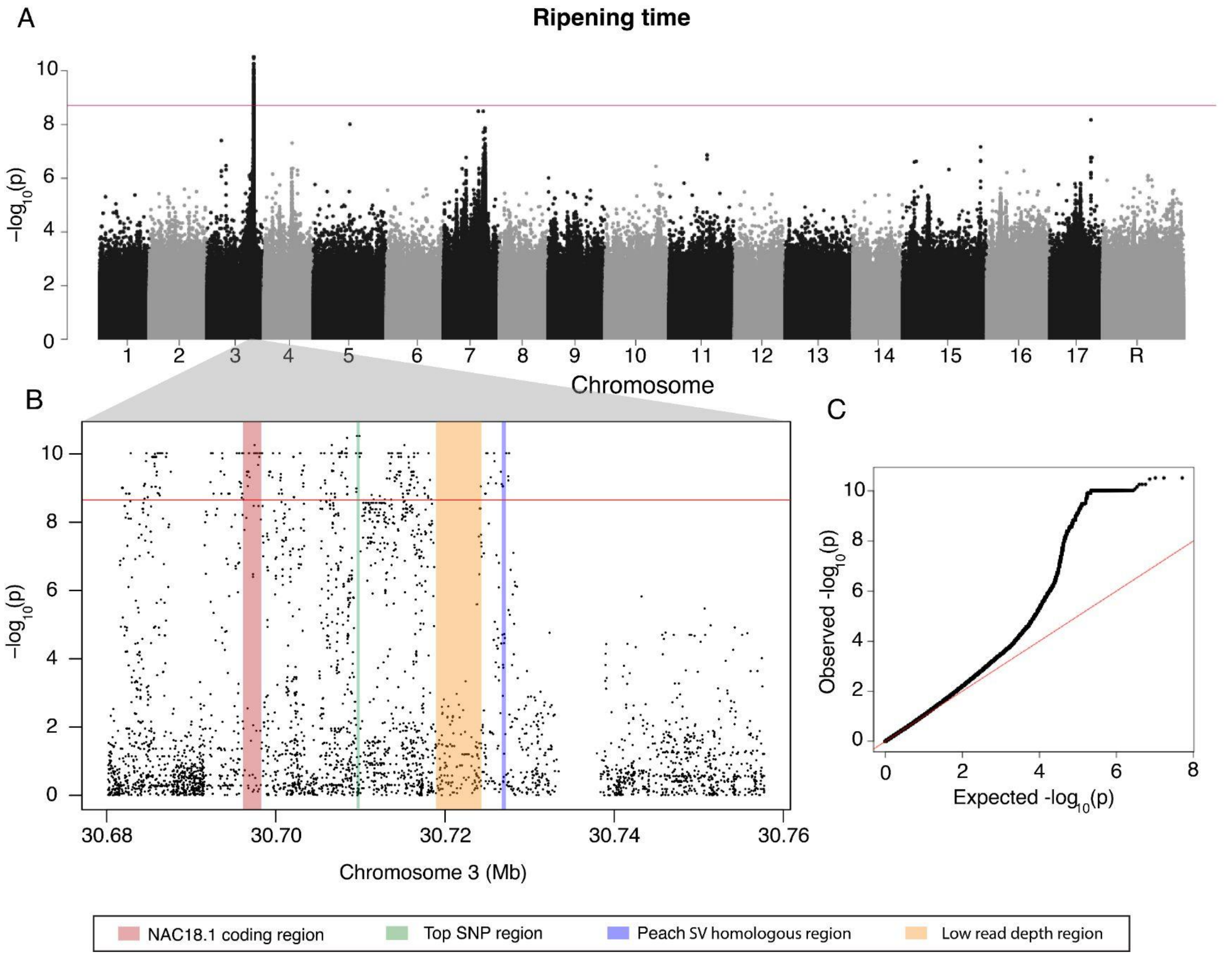


Fig 4-5. Manhattan and QQ plots for the GWAS of ripening time using 76 samples and 25,769,077 variants derived from WGS data. (A) Genome-wide manhattan plot for ripening time showing a single strong peak on chromosome 3. (B) A zoom-in plot of the significant peak on chromosome 3, the only region of the genome showing a significant GWAS signal. The red vertical bar indicates the NAC18.1 coding region. The green vertical bar represents the region with top GWAS SNPs. The yellow bar represents a region of low read depth. The blue vertical bar represents the homologous genomic region in apple of the causal mutation for ripening time discovered in peach (Tan et al., 2021). (C) A QQ plot from MLM model showing a skew towards an excess of low observed P values from the GWAS.

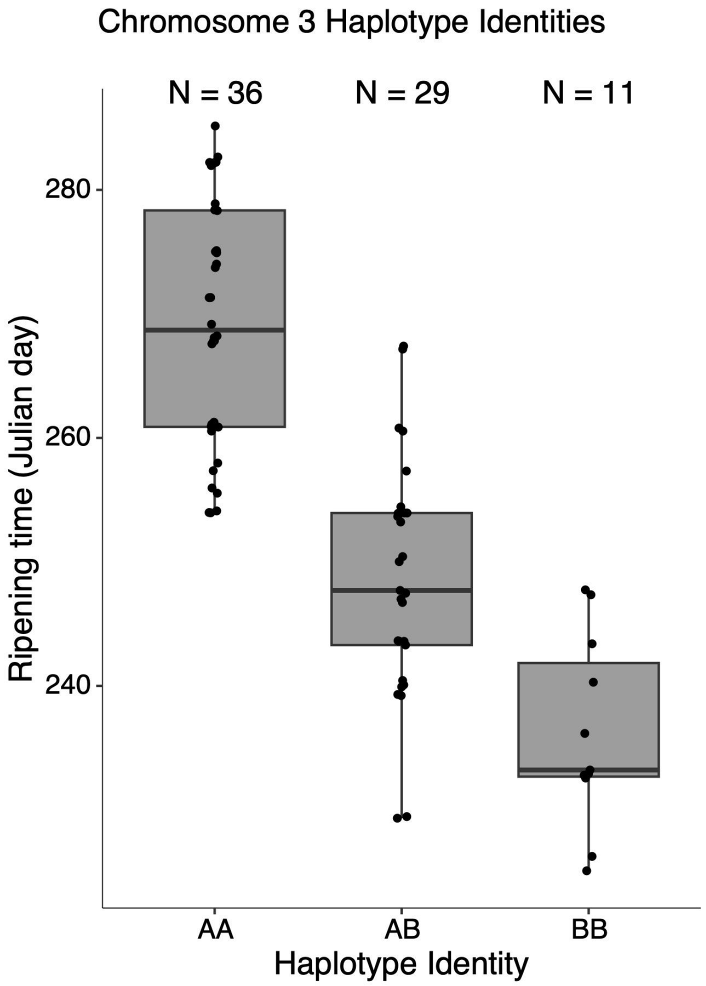


Figure 4-6. Ripening times for haplotype identities for the interval capturing the top 3 strongest associated SNPs (30709586-30709888bp) from the 76 sample GWAS for ripening time. The distribution of ripening time values is shown for each haplotype identity: each dot represents a single sample from the GWAS.

Sliding window WGS read depth analysis

To investigate the possible impact of structural variants or copy number variants on ripening time, a correlational analysis between normalised read depths in 100pb non overlapping windows and ripening time across the 76 selected samples was undertaken. This analysis revealed a strong association between read depth windows on chromosome 3 and ripening time (Appendix III-V)(Fig 4-7B). In addition, the R values from the correlation test indicated a single strong signal on chromosome 3 (Appendix III-VI). The test identified a positive correlation between read depths and ripening time in the chromosome 3 peak region: later ripening varieties had a larger number of sequencing reads at this genomic region. This signal consists of an approximately 45.7 kb region spanning 30680-30725kb on chromosome 3. The single strongest correlation, and smallest p value, from the test was found in the window spanning 30721701-30721800bp on chromosome 3. For this specific window, the normalised mean read depth was significantly correlated with ripening time after correcting for multiple comparisons ($r = 0.75$, $p = 1.16\text{x}10^{-14}$) (Appendix III-VII). When compared to results from the GWAS, the strongest read depth associations were found in the same 45.7 kb region as the most strongly associated variants from the GWAS model. However, although the 100bp windows most strongly correlated with ripening time were in the same 45.7kb region (30.68 - 30.72 Mb, Fig 4-7B) as the top hits identified by the GWAS (Fig 4-7A), the strongest signals from the two tests did not directly overlap. More precisely, the normalised read depth analysis identified the strongest signal at 30718-30724kb (yellow bar indicated on Fig 4-7B), and GWAS identified strong associations spanning from 30689-30718kb and 30724-30728kb (Fig 4-7A)(on either side of the normalised read depth signal). However, the GWAS did not detect significant associations within the region identified by the normalised read depth analysis (Fig 4-7B), only on either side of that signal. To summarise, the read depth analysis and the GWAS identify a signal in the same 45.7kb region on chromosome 3, however the strongest associations from both tests do not overlap within that window.

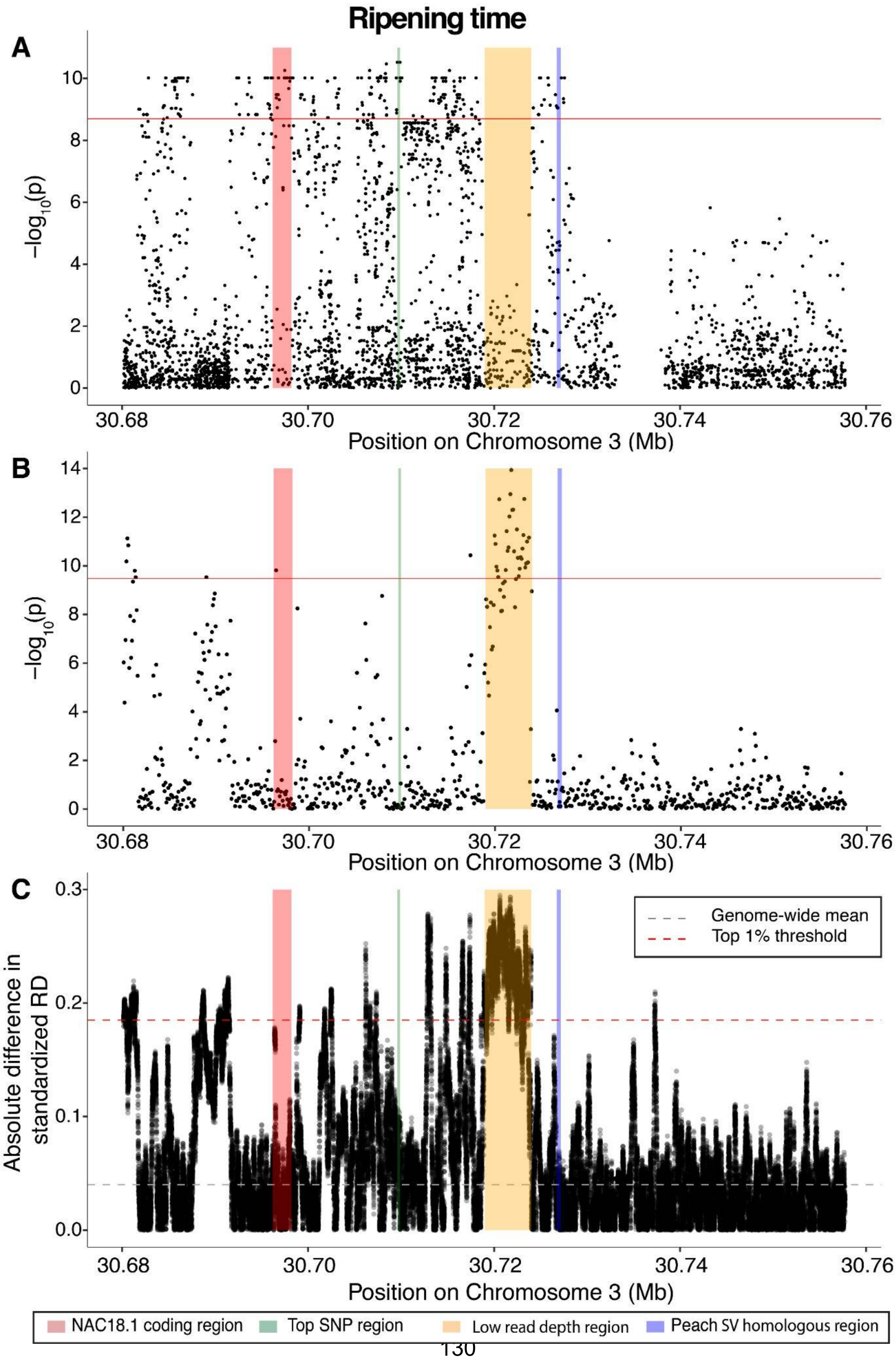

Ripening time
A
B
C
−log10(p)
Position on Chromosome 3 (Mb)
30.68
30.70
30.72
30.74
30.76
0
2
4
6
8
10
12
14
Absolute difference in standardized RD
0.0
0.1
0.2
0.3
Genome-wide mean
Top 1% threshold
NAC18.1 coding region
Top SNP region
Low read depth region
Peach SV homologous region

Fig 4-7. Evidence of association with ripening time for genetic markers and read depth on chromosome 3. (A) A Manhattan plot from the GWAS showing 80kb of chromosome 3, including the 45.7 kb region significantly associated with ripening time. The horizontal line is the Bonferroni-corrected significance threshold. Markers above the dotted line are considered significantly associated with ripening time. (B) A zoom-in plot for P values from normalised 100 bp mean read depth correlations (C) A zoom-in plot for standardised read depth analysis from chapter 3. The red bar indicates the NAC18.1 coding region. The green bar represents the region with top GWAS SNPs.The yellow bar represents a low read depth region. The blue bar represents the homologous region of the causal mutation for ripening time in peach.

Pooled read depth analysis

To further investigate the possible impact of copy number variation or structural variants on ripening time, data from a previous experiment (Chapter 3) in which groups of samples with extreme (early and late) ripening times were sequenced via WGS, was analysed to examine differences in read depths between ripening time groups. The absolute difference between normalised read depths of different ripening time groups from Chapter 3 (Davies & Myles, 2023) at each nucleotide in the genome were compared (Fig 4-7C). This analysis revealed a strong peak in the *NAC18.1* region. The largest absolute difference between normalised read depths was 0.29 at 30720635bp, in the same low read depth region (yellow bar indicated on Fig 4-7C) identified by the sliding window read depth analysis. Further, the largest normalised read depth difference from the pooled analysis (0.29 at 30720635bp) was 1.1kb from the strongest correlation identified by the sliding window read depth analysis.

Linkage disequilibrium comparisons

Mean and median LD (r^2 values) for the 76 WGS samples were 0.25 and 0.09, respectively over 1kb windows (Appendix III-VIII). Pairwise LD among 17 SNPs from the *NAC18.1* region revealed higher mean and median levels of LD in the WGS samples (mean R^2 = 0.42, median R^2 = 0.21) than in the GBS samples (mean R^2 = 0.38, median R^2 = 0.08) (Fig 4-8). In both cases, the SNPs most strongly associated with ripening time, located 20.5kb apart, were in strong linkage (Fig 4-8).

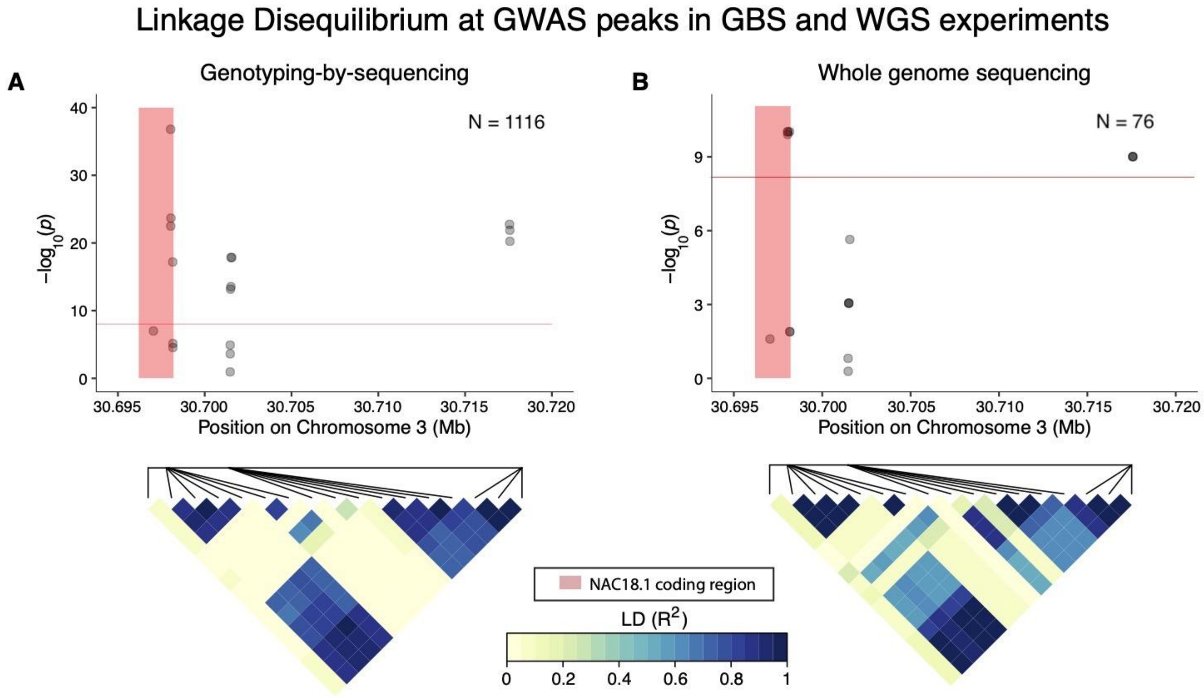


Fig 4-8. LD plots of GWAS results from the GBS samples (A) (Migicovski et al., 2021) and WGS samples (A). Only SNPs found in the *NAC18.1* region from Migicovski et al., 2021 were used to compare LD in the region. Manhattan plots represent results from the GBS experiment and the WGS experiment, respectfully.

CRISPR editing sites

CRISPRScan analysis of the 50kb genomic region containing the GWAS signal for ripening time revealed a total of 4054 possible CRISPR gene editing sites, distributed approximately evenly across the region (Fig 4-9A). Of the CRISPR editing sites, 1601 were of medium quality, and 407 were of high quality (Fig 4-9B). Within the coding sequence of NAC18.1, there were 231 total CRISPR editing sites, 44 of which were medium quality, and 48 of which were high quality.

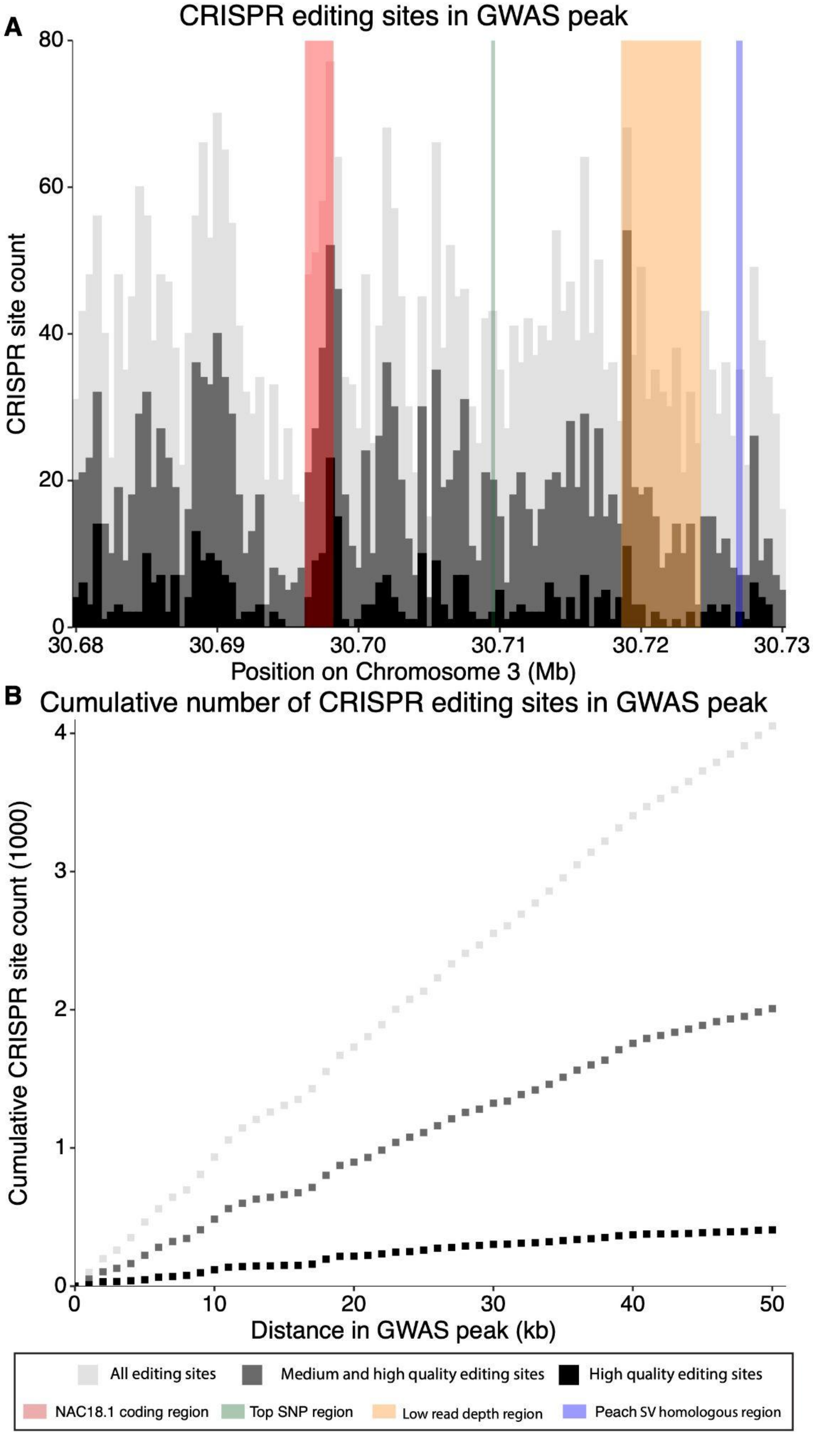

A
CRISPR editing sites in GWAS peak
CRISPR site count
0
20
40
60
80
30.68
30.69
30.70
30.71
30.72
30.73
Position on Chromosome 3 (Mb)
B
Cumulative number of CRISPR editing sites in GWAS peak
Cumulative CRISPR site count (1000)
0
1
2
3
4
0
10
20
30
40
50
Distance in GWAS peak (kb)
All editing sites
Medium and high quality editing sites
High quality editing sites
NAC18.1 coding region
Top SNP region
Low read depth region
Peach SV homologous region

Fig 4-9. Cumulative counts of CRISPR gene editing sites across the 50kb signal region on Chromosome 3 (30.68-30.73Mb). (A) CRISPR editing site counts by quality. (B) The cumulative CRISPR editing sites across the region. High, medium and high, and all quality CRISPR editing site counts are shown in black, dark gray, and light gray, respectively.

Discussion

Sample selection and preliminary GWAS power analysis

While the primary aim of this chapter is to conduct GWAS to discover the causal allele controlling ripening time in apple, it is important to note that the genomic variant dataset (N=97) generated in this chapter lays the foundation for future imputation of genotypes across the ABC. Genotype imputation is the process of filling in missing genotypes by using data from closely related samples and other genotypes in close linkage disequilibrium to the missing genotype. Imputation methods are effective approaches for inferring the genotypes of large numbers of samples in a population without the need to generate DNA sequence data at high depths, which is costly (Guan & Stephens, 2008; Marchini et al., 2007; Servin & Stephens, 2007). With the increasing availability of low-pass sequencing (sequencing at <1x depth), imputation has become critical for generating reliable genotype data across large, diverse mapping populations (J. H. Li et al., 2021, 2023; Martin et al., 2021), and is a viable approach for imputation of genotypes across the ABC. However, reliable imputation requires the generation of a high quality variant dataset from a diverse and representative subset of the larger population, often termed a “reference panel” (Guan & Stephens, 2008; Marchini et al., 2007; Servin & Stephens, 2007). The variants from the reference panel must be of high quality and confidence, and typically require high sequencing depths (>10x) to ensure that downstream imputation of the rest of the population is reliable (Martin et al., 2021). The dataset of variants generated from the 97 samples in this chapter represent a high-quality genotype reference panel for the ABC. In future work, this reference panel will be used to impute genotype data across more than 1000 samples in the population, unlocking genetic mapping approaches with high power and resolution.

When selecting samples to be included in a reference panel for imputation, the aim is to include samples that capture as much of the genetic diversity of the entire population as possible. To select samples from the ABC that were diverse, I employed a selection algorithm that is specifically designed to select subpopulations that contain maximum genetic variation and diversity (Ranallo-Benavidez et al., 2021). After examining the algorithm's sample recommendation, I selected (N=109) samples spanning the entire genetic PC space of the population (Fig 4-2), such that the samples chosen for deep sequencing in this chapter are likely to capture maximum genetic diversity and variation observed across the larger ABC population. The selection of diverse samples at this stage is important for future genotype imputation, as imputed genotypes across the rest of the population will come directly from, and are therefore directly limited by, the variants called from the reference panel.

While the reference panel needed to be diverse, it also needed to provide sufficient power to detect causal alleles via a GWAS. During the experimental design phase of the current chapter, the degree to which sample size would be a limiting factor for the statistical mapping power of the GWAS experiment was unclear. Before sequencing the selected samples (N=109) at high depth, it was important to ensure that a GWAS with this sample size would yield enough power to detect a strong effect locus. Multiple studies in apple have discovered strong hits for ripening time at the *NAC18.1* locus on chromosome 3 (Davies & Myles, 2023; Larsen et al., 2019; Migicovsky et al., 2016; Migicovsky 2021; Watts et al., 2023). As it is clear that a single strong effect locus controls ripening time in apple and related fruit species (Bin et al., 2022; Dirlewanger et al., 2012; Tan et al., 2021), a sufficiently powerful GWAS model should detect a single, strong signal for this trait. Specifically, a sufficiently powerful GWAS model should detect a single peak towards the end of chromosome 3. Following this logic, I conducted a preliminary GWAS by sub-sampling data from Migicovski et al. (2016) to ensure an association study using the reference panel would be viable. The preliminary GWAS (Fig 3) detected a single strong peak on chromosome 3 for ripening time. This was a strong indication that a GWAS using WGS data from 109 samples from the ABC was a viable approach to mapping ripening time in apple.

WGS of selected samples

DNA extractions failed to produce sufficient DNA for 12 samples, leaving 97 samples with WGS data. Low yields from DNA extractions are frequently caused by the presence of high plant fibre, resins, polysaccharides, tannins, and polyphenol content naturally abundant in the leaf tissues which interfere with DNA extraction chemistry (Fang et al., 1992; Murray & Thompson, 1980; Webb & Knapp, 1990). In the future, it would be wise to account for such failures by preparing approximately 15% more samples for library preparation than anticipated to ensure that failures in library preparation and sequencing are mitigated. Although 12 failed samples represented a noteworthy decrease in sample size from the preliminary GWAS, I was confident that the power of the GWAS would still be sufficient to map the ripening time locus. From the 97 samples successfully sequenced, the mean read depth was 17.5x (max: 36.2x, min: 7.2x)(Appendix III-II), sufficiently deep for high quality variant calling, and no samples needed to be excluded on the basis of low read depth.

Reference panel genotype data

The final reference panel genotype data is high resolution: it contains 49.3M variants across the genome, an average of one variant per 15bp. This high density is equal or more high-density than similar variant calling efforts in highly heterozygous crop species (Holušová et al., 2023; Z. Liang et al., 2019; M.-Y. Zhang et al., 2021), and to my knowledge, is the most high-density genome-wide genotype data generated in apple to date. Woody perennials, and plants more generally, are well known to have complex genomes with high levels of genetic variation (Bayer et al., 2020; Z. Liang et al., 2019; Saxena et al., 2014) and the high density of variants seen in this chapter should be expected from the selected samples as they were specifically selected to capture high genetic diversity from the ABC. The MAF distribution observed in the dataset was as expected (Fig 4-4), with most variants in the ABC being rare (Migicovsky et al., 2022). Overall, the high density and high quality genomic data is essential for GWAS mapping and will be important for future imputation based experiments, as a high density genotype data is crucial for leveraging LD decay for high resolution GWAS. This reference panel will likely be the basis of the largest mapping experiments (in terms of sample size and marker density) in apple in the near future.

GWAS

The GWAS using the WGS genotype data detected a single strong signal on chromosome 3 in the *NAC18.1* region (Fig 4-5A&B), as hypothesised. The signal spanned a 45.7kb region in which *NAC18.1* is the only annotated gene. The three variants with the strongest association with ripening time were within 302bp of one another (green bar, Fig 4-5), 11.3kb upstream of *NAC18.1* (coded on the antisense strand) (red bar, Fig 4-5). This suggests that the causal variant impacting ripening time may be a regulatory variant, impacting the upstream regulatory regions controlling transcription of *NAC18.1*. While this evidence does agree with the signal detected by Migicovski et al. (2016), it is not consistent with the hypothesis from that publication that the causal variant for ripening time is the D5Y SNP in the coding region of the *NAC18.1* gene. Although the D5Y SNP was a reasonable putatively causal allele based on previous low-density genetic mapping studies (Larsen et al., 2019; Migicovsky et al., 2021), functional work using transgenic tomatoes showed no significant difference in ripening time between plants bearing alternate alleles at D5Y (Migicovsky et al., 2021). This suggests that coding region SNPs in *NAC18.1* previously thought to impact ripening time are not causal, and that variants upstream of *NAC18.1* within non-coding regions more likely control ripening time. The findings here are also consistent with the results from Chapter 3 (Davies & Myles, 2023), which suggested that a regulatory variant approximately 5kb kilobases upstream of the *NAC18.1* coding region controls ripening time. However, the present experiment indicates that the causal variant is further upstream of *NAC18.1* than previously suggested (Davies & Myles, 2023). Further, in contrast to the data from Chapter 3, the present experiment detected no signal on Chromosome 4 for ripening time, which calls the novel signal on chromosome 4 reported by Davies et al. 2023 into question.

The traditional approach for mapping causal genomic regions in perennial plants, particularly in apple, uses bi-parental crosses with highly related breeding material and low density SNP markers (Bianco et al., 2016; Chagné, Crowhurst, et al., 2012; Chagné et al., 2007, 2019; Kostick, Teh, Norelli, et al., 2021). However, mapping via bi-parental crosses from existing breeding material routinely results in large genomic intervals, often spanning megabases or hundreds of kilobases (S. A. Khan, Chibon, et al., 2012; Kostick, Teh, Norelli, et al., 2021; Myles et al., 2009) making the identification of potentially causal alleles nearly impossible. Here, using a small number of samples, I delimit a region that is 302bp in length using a GWAS in a diverse population that likely contains the causal variant for ripening time in apple. This result is evidence that mapping experiments using largely unrelated samples, high resolution markers, and

GWAS is more effective at identifying potentially causal alleles than traditional genetic mapping approaches that rely on bi-parental crosses. With increases in sample size, the current approach could potentially capture more genetic diversity and resolution and lead to the discovery of causal alleles or narrow regions containing causal alleles controlling numerous apple traits.

Top GWAS hits form a haplotype

The three SNPs most strongly associated with ripening time fall within a 302bp region of the genome in which only two haplotypes are present in the samples (haplotype A: T/C/T, haplotype B: C/T/A)(Fig 4-6). These three SNPs are in complete linkage and have the same allele frequencies and thus the same p values from the GWAS. Samples from the experiment that are homozygous for haplotype A have the latest ripening time and samples homozygous for haplotype B have the earliest ripening times, with the heterozygote samples showing intermediate ripening times (Fig 4-6). This is evidence of incomplete dominance in ripening time, which is consistent with the ripening time literature in fruits (Tan et al., 2021; R. Wang et al., 2020; Watts et al., 2023). Due to the sample size of the present experiment, there is an insufficient amount of recombination to break down the haplotype any further, as all three SNPs are in perfect linkage disequilibrium. Therefore, this 302bp region containing two haplotypes is the narrowest genomic region that the present experiment can delimit for the control of ripening time. Genotype and phenotype data from more samples is required to capture recombination that breaks this region into smaller haplotypes.

While it is likely that the causal allele for ripening time in apple lies within the 302bp region identified above, it remains possible that the causal variant lies outside this region. Complex genetic mechanisms such as allelic heterogeneity, which frequently control plant traits, can create spurious associations between SNPs and phenotypes even when the impacts of relatedness and population structure have been statistically controlled as was done in the present experiment (Korte & Farlow, 2013; Vilhjálmsson & Nordborg, 2013). For example, flowering time in Arabidopsis is perhaps the most well studied plant phenotype, and GWAS has been used repeatedly in the past to fine-map potentially causal alleles for flowering time (Atwell et al., 2010; Brachi et al., 2010; Zan & Carlborg, 2019). Although a well characterised variant in the *AOP2* gene was long thought to be a major contributor to flowering time in Arabidopsis based on GWAS results, recent re-analysis of GWAS data have provided evidence that this association may be caused by fitting an association model for a single causal locus when in fact there are two causative loci in close proximity (Sasaki et al., 2021). Therefore, in the present study it is possible that a complex mechanism, such as allelic heterogeneity, that is not captured by the GWAS model used here could be controlling ripening time. It is clear that statements of causality cannot be confidently made from GWAS associations alone and that further functional work, such as gene expression analysis and transgenic experiments, must be performed to confirm the precise location and effect of causal alleles.

Within the peak region on chromosome 3, I identified a region of low read depth (discussed below) as well as a homologous region of high sequence similarity to that of the causal region for ripening time in peach (Tan et al., 2021). In peach, a complex SV has been documented to control ripening time by impacting the transcription of a peach NAC transcription factor, a homolog of *MdNAC18.1*. The 2kb causal region from the peach reference genome showed a high level of homology (E value = 10^{-41}, identities = 75%) with the apple reference genome in the region upstream of *NAC18.1*. Although the genetic mechanism controlling ripening time is not fully understood in peach, it is possible that the causal variant in apple is of similar identity or acts via a similar mechanism. Although the most strongly associated SNPs from the present GWAS are not directly in the peach homologous region, it is possible that the causal allele in both species is impacting a conserved regulatory structure, and that the same locus controls ripening time in both species. The idea that a conserved locus controls ripening time across species in Rosacea is the logical extension of previous hypotheses (Dirlewanger et al., 2012), supported by evidence whereby ripening time is associated with the same locus in peach, apricot, wild strawberry, and sweet cherry (Dirlewanger et al., 2012; X. Li et al., 2023). To more accurately locate the causal allele in apple at this locus, a higher level of LD decay within the genomic region of interest is required (discussed below). To achieve this, a GWAS with more samples may suffice, although long read sequencing of the region may be required to confidently capture and resolve haplotypes or complex structural variations.

Sliding window WGS read depth analysis

While calling variants from short read data is a robust method to produce high quality genotype data in the form of SNPs and short indels, this approach has limited potential for directly calling structural variants (SVs), such as copy number variants (CNVs) and presence-absence variants (PAVs) (Francia et al., 2015). Instead, SVs are best discovered via long read sequencing (Ho et al., 2020; Sedlazeck et al., 2018). However, in lieu of long read sequencing, read depth information from short read WGS can be used as a proxy to examine patterns of DNA abundance in samples. In fact, read depth information is key for a number of computational models designed to discover SVs (Francia et al., 2015; Ho et al., 2020). SV detection methods leveraging read depth information are often based on the principle that SVs typically change the abundance of a DNA sequence (greater or fewer copies of a given sequence) in relation to the reference genome. If samples are sequenced to a known depth, it is often reasonable to assume that read depths will be relatively uniform across the genome. Following this logic, it is reasonable to compare the read depths for genomic positions to trait data, correlating each read depth position from all samples to phenotype data. For example, a correlation test could be conducted between the read depths at a given genomic position across a group of samples and the ripening time for those samples. This correlation could reasonably be made for each position in the genome, with strong correlations indicating a relationship between DNA sequence copies and the phenotype, providing evidence of structural or copy number genomic variation contributing to the phenotype. This basic approach is the basis of a number of SV detection methods, and sliding windows are often used for read depth calculations in such cases (Francia et al., 2015). Given that the causal allele for ripening time in peach has been characterised as a structural variant and an increasingly large body of literature suggests that SVs often control plant phenotypes (Cardoso et al., 2014; Hattori et al., 2009; Qiao et al., 2023; Saxena et al., 2014; Y. Sun, Wang, et al., 2022; K. Xu et al., 2006), a sliding window read depth algorithm using 100bp windows was applied to the WGS dataset.

The sliding window read depth analysis produced strong and significant correlations between normalised read depth and ripening time (Fig 4-7B), with the only strong signal in the genome on chromosome 3 (Appendix III-V). While the analysis detected the same 45.7kb region on chromosome 3 as the GWAS analysis, the strongest correlations from the sliding window read depth analysis were within an approximately 5kb subregion of the 45.7kb signal (Fig 4-7, yellow bar). This region was termed the “low read depth region” (Fig 4-7, yellow bar) earlier in the GWAS analysis due to the unusually low read depths detected across samples, which was suspected to have decreased the number of variants called in this region. The unusually low read depth in this region across most samples suggests that the reference genome contains DNA sequence, perhaps a SV, that is not common among the samples we sequenced. Interestingly, correlations between read depth and ripening time in this low read depth region were the strongest in the genome, with read depth being strongly and significantly correlated with late ripening time (r = 0.75, Appendix III-VII). Strong correlations between read depth and ripening time suggests that a SV in this region may have generated a duplication or complex insertion/deletion of DNA sequence that is impacting the ripening time phenotype. Given that the region detected in this analysis is approximately 22kb upstream of the NAC18.1 coding region, these results are further evidence that a regulatory variant could be impacting ripening time in apple, as previously hypothesised (Davies & Myles, 2023). This finding is strikingly similar to ripening time experiments in peach, in which samples with a SV upstream of a NAC transcription factor have earlier ripening times (Tan et al., 2021). In peach, the causal SV controlling ripening time is a deletion relative to the reference genome (peach reference cultivar is Lovell, late ripening) and has been shown to decrease expression of NAC in early ripening varieties. In fact, deletion of the 26kb region directly upstream of the NAC gene in peach abolishes fruit ripening altogether (Nuñez-Lillo et al., 2015). Similarly to peach, the apple reference genome, Golden Delicious, is a late ripening variety. These similarities suggest, in both peach and apple, DNA sequences upstream of the NAC gene are essential for modulating NAC expression and the presence of reference sequence DNA is likely responsible for producing the late ripening phenotype. Together, the apple and peach data suggest that the deletion of DNA in the region upstream of NAC likely results in a decrease in gene expression of NAC, leading to an earlier ripening phenotype. Although experiments explicitly measuring gene expression are required to test this hypothesis in apple, it agrees with the assertion made by Watts et al. (2023) that *NAC18.1* is acting as a ‘throttle’ controlling the rate of ripening in apple. Moreover, recent studies in strawberry have identified the differential expression of *FvRIF*, the homolog of *NAC18.1*, as key for controlling fruit ripening (X. Li et al., 2023). These three independent lines of evidence (peach, apple and strawberry) are a strong signal that genetic variation impacting transcription factor expression is controlling fruit ripening in Rosaceous fruit. While the read depth analysis done here was not conclusive, future

GWAS-based mapping experiments should consider the use of long read sequencing to investigate the presence and potential impact of SVs in this region.

Pool-seq read depth analysis

Because the sliding window read depth analysis (above) suggested that a SV or CNV may be impacting ripening time, I hypothesised that read depth data from pooled DNA from early and late ripening samples would detect a signal at the same locus. To test this, pooled WGS data from chapter 3 was used to compare normalised read depths between pools of early and late ripening samples (Davies & Myles, 2023). Interestingly, the largest difference in normalised read depths between early and late ripening pools was within the low read depth region (Fig 4-7C, yellow bar) identified previously. This result further supports the notion that a SV upstream of *NAC18.1* may be controlling ripening time. Together, the sliding window analysis and the pool-seq analysis provide support for the hypothesis that a DNA segment that is present in the late-ripening ‘Golden Delicious’ reference genome but is altered or deleted in early ripening samples is controlling ripening time in apple. Again, to confidently confirm this result and produce more detailed information about the sequence identity in this region, long read sequencing data across a substantial number of samples will be required.

Linkage disequilibrium (LD) decay and comparison

Levels of LD in the GWAS population in the present chapter were similar to levels observed in the ABC previously (Migicovsky et al., 2016) and in other high diversity perennial fruit crops such as peach, sweet cherry, and pear (Appendix III-VIII)(Donkpegan et al., 2023; S. Kumar et al., 2017; Micheletti et al., 2015). Due to the high diversity of the WGS samples, I hypothesised that LD would be relatively low in the WGS genomic data, but reasoned that the LD would be higher in the WGS data than that of recent GBS studies which used more samples from the ABC. A comparison of LD rates between the present study and previous mapping experiments was made by calculating LD for WGS samples and GBS samples from the ABC (Migicovsky et al., 2022; Watts et al., 2023) at the *NAC18.1* signal region. As hypothesised, the present study showed low LD overall (mean r^2 = 0.25 over 1kb), but when comparing SNPs common to both experiments at the *NAC18.1* locus, the WGS genomic data LD is nearly 3x higher than that of the GBS study (Fig 4-8). This difference in LD can be accounted for by the large difference in sample size between WGS (N=76) and GBS (N=1116) experiments, as it is well understood that larger sample sizes capture more historical recombination events and drive down LD (Nordborg & Tavaré, 2002; Single & Thomson, 2016). Because of the key influence of LD on GWAS resolution, this highlights the key trade off that was made in the current experiment: by sequencing a small number of samples at high depth, GWAS resolution remained lower than it would have if more samples were sequenced at lower depth. Therefore, the current GWAS had high variant density, but higher than optimal LD. In theory, given a large number of samples from the ABC, LD could be driven low enough that GWAS could detect causal alleles with near-single base pair resolution. To this end, the large difference in LD decay between the WGS and GBS experiments seen here is a strong indicator that future mapping studies that combine the variant density of WGS (as in the current study) and vast historical recombination of the ABC (as observed in the GBS samples) are likely to enable fine-mapping of causal alleles to narrow genomic regions. This bodes well for future studies of the ABC that leverage WGS and imputation for elucidating causal alleles for breeding and gene editing.

CRISPR examination

To illustrate the possibility of generating gene editing targets from GWAS results, I explored the *NAC18.1* signal region associated with ripening time for potential CRISPR gene editing sites. The 50kb region associated with ripening time in the present chapter was scanned for possible CRISPR gene editing sites with the required protospacer adjacent motif (PAM) sequences. The observation of more than 4000 valid CRISPR gene editing sites (Fig 4-9) within this 50kb region suggests that efficiently transitioning from mapping experiments to functional validation, even when mapping experiments have confidently identified a causal region on the order of kilobases, will likely remain challenging. Even when considering only the highest quality gene editing sites, there are over 400 unique loci that could serve as gene editing targets at this locus. Functional testing of 400 unique gene edits across multiple apple lines would require the regeneration and propagation of thousands of plants, and tree care for several years to appropriately evaluate ripening time. A functional experiment of this magnitude would require prohibitive levels of capital, labour, and time resources. While limited functional experiments have been accomplished in apple in the recent past (Jiang et al., 2022; Kost et al., 2015; Schlathölter et al., 2023), to my knowledge, there are no examples of evaluating gene editing in a single perennial crop on the scale of hundreds of unique gene editing sites at this time. Therefore, the 50kb region associated with ripening time found in this chapter is too large to efficiently inform functional work, as thousands of successful gene edits must be made to exhaustively evaluate potentially causal alleles controlling ripening time. In addition, newly emerging gene editing technologies that are not limited by traditional PAM sequences (Endo et al., 2019; Q. Ren et al., 2021) continue to increase the number of potential gene editing sites available for validation, making fine-mapping of short genomic regions ever more important. The ripening time example provided here emphasises the importance of high resolution fine-mapping, as narrow regions leave fewer possible gene editing sites for functional validation, and decrease the burdon of downstream functional experimentation. In the future, the discovery of extremely narrow genomic regions (<1kb) via association studies similar to the one conducted here will likely be key for enabling the efficient and successful gene editing in apple.

Conclusions

The discovery of causal alleles is arguably the primary barrier to the application of novel gene editing technologies in apple. At present, WGS sequencing of large populations remains prohibitively expensive and the use of reference panel-based imputation approaches is becoming increasingly popular in agriculture (Buckley et al., 2022; J. Li, Wang, et al., 2022; Nosková et al., 2021). Moving forward, imputation based methods are likely to be an effective way to provide the genotype data required for causal allele discovery in apple. Several studies in recent years have imputed genotypes across large populations with comparable reference panel sample sizes to the one developed here (Buckley et al., 2022; Nosková et al., 2021). In this chapter, I generated a high quality reference panel with over 49M variants from a diverse subset of samples from the ABC. This reference panel was then used to conduct GWAS-based mapping experiments for ripening time. Mapping analyses done in this chapter discovered a signal spanning 47.2kb on chromosome 3 associated with ripening time and delimited a 302bp region upstream of a transcription factor, *NAC18.1*, that likely contains the causal allele(s) controlling ripening time in apple. This finding strongly suggests that the causal allele(s) controlling ripening time is/are regulatory variant(s). However, the causal allele(s) for ripening time cannot be delimited with high confidence using the methods in this chapter, primarily due to the high linkage disequilibrium attributable to the small sample size of the reference panel and the inability of the reference panel to effectively query complex genetic variation. This study provides evidence that GWAS using a small number of diverse samples can be effective for detecting relatively narrow genomic regions that control plant traits. However, the analyses done here indicate that signal regions on the order of kilobases are likely too large to inform effective gene editing in apple. Given the large number of gene editing sites in the signal region detected in this chapter, gene editing for ripening time in apple will be challenging without further refining the signal. To more confidently pinpoint the causal allele(s) for ripening time, a mapping study with more samples is required, and would be best followed by multiple sequencing methods across samples, detailed *NAC18.1* expression studies, and transgenic functional research focussed on the *NAC18.1* promoter region. Importantly, the high quality reference panel generated in this chapter can be leveraged in the near future to accurately impute the genotypes of over 1000 samples from Canada's ABC. Imputation on this scale will enable high powered and high resolution genetic mapping in apple, and will almost certainly provide the power to detect narrow regions of the genome controlling dozens apple phenotypes. In some cases, mapping studies using the imputed data may lead to the direct discovery of causal alleles underpinning key agricultural traits in apple. This chapter represents a substantial investment in the mapping potential of Canada's ABC and the completion of an essential foundation for future association mapping experiments in apple.

Chapter 5: Summary, final conclusions, and future work

Summary of findings

The objective of this thesis is to advance the current state of knowledge in the areas of apple phenomics and genomics by leveraging the wealth of phenotypic and genetic diversity in Canada's Apple Biodiversity Collection (ABC). Specifically, this research aims to provide a detailed comparison of the phenotypic differences between domesticated and wild apples, and to make contributions toward the discovery of causal alleles controlling apple traits. The approach employed in the present research is to first comprehensively examine numerous phenotypes of the domesticated apple, and then compare and contrast these phenotypes to its primary wild progenitor species. Following this exploration of phenome evolution, a pool-sequencing genomics approach is applied in an attempt to map the genomic control of valuable apple phenotypes in the domesticated apple. Finally, high depth and high resolution DNA sequence data from 97 diverse samples from the ABC is generated and analysed with the aim of discovering the causal allele for ripening time in apple.

In chapter 2, I compare the phenomes of two important apple species: *Malus domestica* and *Malus sieversii*, the latter being the primary wild progenitor of the former. This statistical comparison analyses 10 plant phenotypes across over 1000 samples from the ABC. The phenomes of these two species are significantly different overall. I found that domesticated apple trees have shorter juvenile phases and produce ripe fruits later than their wild counterparts. Further, on average, fruits from *M. domestica* are 3.6 times heavier, 43% less acidic, and have 68% less phenolic content than wild apples. The historical analysis suggests that breeding practices over the past 200 years have led to apples that are higher in soluble solids, are less bitter, and soften less during storage. This research sheds light on the impacts of domestication on the modern apple, and highlights the value of crop wild relatives as breeding material for cultivar improvement in the future.

In chapter 3, I employ a pool-sequencing approach with the aim of discovering the causal DNA sequences controlling ripening time, softening, and phenolic content production in apples. I use whole genome sequencing data from phenotypically extreme samples from the ABC to scan the apple genome for signals of differentiation between groups of samples for each phenotype. This investigation provides further evidence of the involvement of the transcription factor *NAC18.1*, and suggests that the promoter region upstream of this gene could harbour the causal allele(s). This is a significant step forward in understanding ripening time, in which the nature of the causal allele(s) (coding or regulatory sequence) has been a source of uncertainty in recent years. Further, this study detects multiple loci associated with phenolic content production and implicates a family of *UDP-Glycosyltransferase superfamily proteins* as potentially responsible for variation in this phenotype. Finally, this study suggests a complex genetic architecture underlying softening in apple, and provides evidence that a gene related to skin wax production could be involved in fruit softening during storage. This chapter provides evidence that pooled sequencing approaches are suitable methods for genetic mapping in diploid perennial crops like apple, and reveals a number of novel loci potentially controlling important plant traits for future investigation.

In chapter 4, I take critical steps towards unlocking the full mapping potential of Canada's ABC and make progress toward the discovery of the causal allele for ripening time. Using DNA from 97 diverse samples from the ABC, I generate a high resolution reference panel from high depth WGS data via a custom bioinformatics pipeline. I then conduct a GWAS for ripening time using the reference panel. This mapping experiment detected a strong signal for ripening time on chromosome 3 and delimits a 302bp region upstream of *NAC18.1* that likely harbours the causal allele(s) controlling ripening time in apple. Analyses from this chapter illustrate that the frequency of CRISPR gene editing sites in the genome will likely render the application of gene editing technologies in apple challenging without the identification of extremely narrow causal regions (<1kb) or the definitive identification of the causal allele(s) for a phenotype. While the causal allele(s) for ripening time could not be confidently identified in this experiment, the generation of a high quality reference panel represents the completion of an essential step for enabling accurate imputation of genotypes for thousands of samples from the ABC in the near future. With a reference panel complete, the stage is set for future high-power and high-resolution GWAS using imputed genotypes in this population, which hold the potential to discover numerous valuable causal alleles in apple.

Altogether, the analyses of the apple phenome and genome in this thesis further the scientific community's basic understanding of phenotype variation, the effects of domestication and breeding, the genomic control of agricultural phenotypes, and the path towards causal allele discovery in high-diversity perennial crops. Further, it provides a compelling and pragmatic example of the challenges associated with using GWAS to discover causal alleles that could be used as gene editing targets. This research contributes to a growing body of work that improves our understanding of the genomic control of important apple phenotypes, and also lays important groundwork for future high-powered genetic mapping in a highly diverse apple mapping population.

Final conclusions

The most important conclusion from this work was the realisation that the identification of causal alleles is far more complicated and challenging than I initially anticipated at the beginning of the project. The path from genetic mapping to improved apple varieties using gene editing is not composed of a set of well-defined steps, but is better understood as a series of probabilistic inferences: there is statistical noise in each step of the process that ultimately results in far more uncertainty in the conclusions than I initially anticipated. Even with access to a large and highly diverse mapping population, it is far more challenging to precisely discover the causal DNA sequences that control a phenotype in apple than I predicted. Upon starting this project, I anticipated that multiple single nucleotide variants (SNVs) controlling key phenotypes could be identified given the appropriate sequencing data and experimental design. However, my search for causal SNVs was driven by an oversimplified view of genome organisation and control. Recent work has shown that SNVs likely represent only a small fraction of the genetic changes that cause phenotypic variation. Instead, it has become clear that structural genetic variation such as CNVs, PAVs, indels, and translocations are often the genetic variation that controls plant phenotypes. The pronounced role of structural variation in the control of plant phenotypes adds a significant layer of difficulty to detecting causal alleles because structural variation is challenging to query using short read sequencing. Effective detection of structural genetic variation requires multiple forms of DNA sequencing and the use of multiple genomes or pan-genomes. It has become evident over the course of this project that the detection of complex variation will be a key challenge for causal allele discovery.

Further, it is clear that the genetic control of each phenotype is unique and each phenotype requires a unique analysis. For example, if more evidence were to support that a coding variant within *NAC18.1* was controlling ripening time, a transgenic experiment in which the coding sequence was altered or interrupted (gene mutagenesis or knockouts, e.g., (X. Li et al., 2023)) or transgenic complementation (using multiple *NAC18.1* haplotypes, e.g., Migicovsky et al. 2021) may be the logical follow up experiment. However, in the case presented here, the causal allele is likely in the promoter sequence, making expression (RNAi, virus-induced gene silencing, e.g., Jiang et al., 2022) or promoter analysis experiments more appropriate follow up studies. Therefore, it is likely that after mapping causal alleles to narrow regions of the genome, each phenotype will require unique types of functional experiments depending on GWAS results. I now fully recognize that even with detailed genotype and phenotype information from the ABC, there will be no one-size-fits-all approach for identifying causal alleles across phenotypes. Association mapping may delimit narrow regions of the genome harbouring causal alleles in the future, but having full confidence in the identity of a causal allele at the nucleotide level will likely require orthogonal evidence in the form of functional experiments whose design will depend on themode of genetic control and type of polymorphism involved in each phenotype. In some cases, such as phenolic content production (Chapter 3), phenotypes appear to be controlled by multiple loci of small effect, which is not surprising, but adds another layer of difficulty to understanding the alleles impacting the variance in that phenotype. Given the unique genetic architecture of each trait and the detail with which the causal allele must be delimited before gene editing or genetic modification approaches can be used, each phenotype likely requires individualised lines of research. Individual lines of experimentation for each phenotype necessarily means that discovering causal alleles for each phenotype will require significant resources, and phenotypes that offer greatest value to our agricultural system should therefore be prioritised. It is clear that causal allele discovery will remain challenging and resource intensive, but still offers value both to fruit science and the apple industry.

Although wrought with challenges, association mapping still represents a powerful tool for mapping causal alleles. There are many examples of such approaches yielding causal alleles (Bai et al., 2012; Kenny et al., 2012; Puckett et al., 2023; Qiao et al., 2023; Tan et al., 2021), however I have come to realise the true extent of the investment and time achieve success in this endeavour. Already, genetic mapping in Canada’s ABC has required more than a million dollars and a decade of intense design, logistics, and research, and it seems likely that another 10 years is required to map causal alleles for many traits at our team's current pace. In addition, association mapping is likely to only represent one aspect of effective causal sequence discovery. As seen in related Rosaceous crops like strawberry (X. Li et al., 2023; Martín-Pizarro et al., 2021; Sánchez-Sevilla et al., 2017), gene expression analysis is a powerful tool to discover the genes controlling certain traits, and can be leveraged to determine which DNA sequences impact plant phenotypes. Given the immense challenge of genetic mapping in plants, the use of association mapping and expression analysis strategies in tandem is likely the fastest path towards understanding plant traits.

Importantly, the optimism expressed about gene editing in the introduction of this document is likely belied by the true challenges of enabling gene editing technology in apple. While gene editing does offer novel and powerful strategies for advancing plant science and crop improvement, a number of significant challenges remain over and above causal allele discovery. For apple, tissue culture and regeneration may represent the single greatest barrier to gene editing (Atkins & Voytas, 2020; Venezia & Creasey Krainer, 2021). The successful culturing, editing, and regeneration of apple tissues is currently seen with increasing pessimism by those directly focussed on the problem (Sophie Watts, personal communications, 2024, Shai Lawit, personal communications, 2023, Brian Crawford, personal communications, 2023, Franklin Lewis, personal communications, 2023). There is a significant bottleneck in discovering effective protocols for editing apple genomes and regenerating edited tissue into functional plants. As of this writing, only a single apple cultivar, ‘Gala’, has been edited and regenerated successfully . This means that even with the successful discovery of causal alleles and application of gene editing in apple, only a single apple cultivar is currently available for modification by gene editing. At present, the biological mechanisms underpinning tissue culture and regeneration in apple (and perennial trees in general) remain highly speculative and supposed advancements in this area are frequently guarded by trade secrets (Okanagan Specialty Fruits, Verinomics Inc., Caribou Biosciences). Therefore, without significant progress in tissue culture and regeneration, gene editing only offers benefits to a handful of cultivars, and cannot deliver on the promise of improvements across elite, heirloom, or wild cultivars. This significantly reduces the proposed impact of gene editing on cultivar improvement, novel variety creation, and the promise of crop re-domestication (Hanak et al., 2022; Lyzenga et al., 2021; J. Xu et al., 2019).

Even in crops where gene editing has been successfully applied to generate novel phenotypes and new consumer products, it is unclear if such products represent industrial improvements or commercial successes. For example, the Pairwise Plants Conscious Greens® (Karlson et al., 2022), generated using CRISPR gene editing to knockout the production of bitter tasting metabolites, is understood in industry circles to be a scientific success but a complete market failure that is unlikely to appear on grocery store shelves much longer (Eric Ward, personal communications, 2023). Even with regulatory changes around the world distinguishing between gene edited and GMO food products in recent years (Health Canada, 2022; Turnbull et al., 2021) and reduced consumer scepticism towards gene edited foods (Funk, 2020), it remains unclear how trends in governmental regulation and consumer thought will impact the market success of gene edited food products. These challenges, both scientific and industrial, have dampened my optimism about the promise of gene editing in apple, and have instead strengthened my optimism in high-throughput traditional breeding methods and GMO techniques.

Future directions

Genotype imputation

The studies in chapters 3 and 4, although insightful, are both limited by sample size due to the constraints of DNA sequencing costs. Under ideal circumstances, all 1116 samples from the ABC would be sequenced to high depth, enabling mapping experiments that maximise statistical power and fully leverage the vast historical recombination captured by the ABC population. High-depth DNA sequencing of the entire population would allow for high-confidence genotype calling and thus causal allele identification. However, in the future, it is still unlikely that thousands of samples from the ABC will be sequenced to high depth due to the high costs associated with whole genome sequencing thousands of samples at high depth. Instead, genotype imputation offers an opportunity to generate accurate genotype information for thousands of samples at a fraction of the cost of deep WGS, and offers comparable accuracy to sequencing at high depths (J. H. Li et al., 2021, 2023; Snelling et al., 2020). The reference panel generated in chapter 4 sets the stage for genotype imputation of the entire ABC, and will be used to generate accurate genotype information for all samples in the ABC following low coverage sequencing of all non-reference samples.

The logical next step from the work presented in this thesis is the collection of low-depth sequencing from the remainder of the ABC and the subsequent imputation of genotypes across the entire ABC using the reference panel produced in Chapter 4. Imputation is becoming an increasingly popular approach for agricultural genomics, and has been applied successfully in numerous organisms including pigs, dogs, cattle, and laying chickens (Buckley et al., 2022; J. Li, Wang, et al., 2022; Nosková et al., 2021; Snelling et al., 2020). Indeed, imputation work at the ABC is already underway and, at the time of this writing, DNA from 820 samples from the ABC have been sequenced at low depth (1x). These data will be used to impute the genotypes of 820 non-reference samples, bringing the number of samples with accurate genome-wide genotype information to 917. A population of this size with high density genetic markers will provide high resolution for future association studies, particularly since LD decay has been shown to be rapid in large apple collections, including the ABC (Larsen et al., 2019; McClure et al., 2018; Migicovsky et al., 2016). Future association studies using these data will be able to leverage both the high genomic variant saturation and the high LD decay from the ABC population to delimit causal alleles or narrow regions of the genome harbouring causal alleles. In the latter case, targeted sequencing approaches can then be applied to narrow regions of the genome associated with phenotypes to elucidate causal alleles. Recently, there have been advancements in imputation such that long read sequencing can enable imputation of structural variants (Noyvert et al., 2023). The ability to detect structural variation is likely to be key for discovering causal alleles in the future, and imputation of structural variation could be a reasonable next step following imputation of SNPs and short indels using short read sequencing. In summary, imputation offers a cost effective and accurate method of mapping the genomic control of traits in large populations, and stands as the most reasonable path forward for investigations at the ABC.

The use of multiple sequencing technologies

In the future, it is highly probable that multiple sequencing approaches will be necessary to uncover causal alleles from the ABC. As exemplified in multiple recent studies (Cirilli et al., 2022; J.-M. Song et al., 2020; Tan et al., 2021), the use of both long and short read sequencing data is an effective approach for discovering complex genetic variation that impacts plant phenotypes. For example, two recent publications in peach have confidently delimited complex variants that control ripening time and a double flowering trait using a combination of short and long read sequencing (Cirilli et al., 2022; Tan et al., 2021). These experiments clearly demonstrate that, with sufficient access to diverse samples and multiple sequencing technologies, causal alleles can often be confidently identified using association studies. As discussed earlier, the increasing number of studies finding causal variants that are structural, such as copy number variation, presence absence variation, and complex rearrangements, indicates that structural variants likely underpin a significant proportion of plant phenotypic variation (Alonge et al., 2020; Gabur et al., 2019; Saxena et al., 2014). With this in mind, it will be increasingly important to use long read sequencing such as Oxford Nanopore or PacBio to accurately discover structural variation. Further, given that regulatory sequences are also likely to play a role in the control of numerous plant phenotypes (see Chapters 3 and 4), sequencing methods such as Hi-C may also be reasonable for untangling associations and interactions between genes and regulatory regions that ultimately control plant phenotypes (Song et al. 2020; Zhang et al. 2019; Belton et al. 2012; Eagen 2018; Kim et al. 2022). In the future, multiple sequencing methods will likely need to be applied on a case-by-case basis depending on the genetic structure of the trait to discover causal alleles.

Pan-genomes

Pan-genomes are an emerging resource in plant genetics and are likely to be the standard reference genome format in the future (Bayer et al., 2020). Pan-genomes are reference sequences generated from multiple samples from a species and therefore capture more genetic diversity than can be captured in a single sample's linear reference genome, which is the current standard. To date, dozens of crop pan-genomes exist, including builds for rice, maize, wheat, soy, poplar, pepper, sesame, and walnut (Hirsch et al., 2014; Y.-H. Li et al., 2014; Montenegro et al., 2017; Ou et al., 2018; Pinosio et al., 2016; Schatz et al., 2014; Trouern-Trend et al., 2020; Yu et al., 2019). This recent expansion in the number of pan-genomes is largely due to declines in long-read sequencing and computing costs, both of which are essential resources for the generation of pan-genomes (Bayer et al., 2020).

Pan-genomes offer a number of significant advantages over traditional linear reference genomes. First, pan-genomes capture more structural variation than linear genomes

(Munir et al., 2020). This is important because structural variation has been demonstrated to be key for the control of numerous plant phenotypes (Jiao & Schneeberger, 2020; Nsabiyera et al., 2019; Schatz et al., 2014; J.-M. Song et al., 2020), and pan-genomes enable such variation to be more accurately identified. Second, it has been demonstrated that alignment of short read sequence data to a reference pan-genome significantly improves sequence mapping accuracy and downstream quality of variant calls and can also lead to more accurate gene expression measurements (Golicz et al., 2016; R. Li et al., 2019; X. Tian et al., 2020). Although the accuracy of read mapping and variant calls were relatively high in this work, the use of a pan-genome could have improved the mapping accuracy and variant quality of calls made in Chapter 4 of this thesis. Finally, pan-genomes capture more of the "dispensable genome" — the genes not present in all samples that are captured by a pan-genome (Medini et al., 2005). This is significant because the dispensable genome is now understood to contain genes and causal alleles that provide important and agriculturally relevant phenotypes such as improved flavour in tomato (L. Gao et al., 2019) and seed weight in pea (Zhao et al., 2020). Overall, pan-genomes represent a significant step forward in more completely querying the genomic variation of a crop and avoiding the well known single sample bias introduced by using a linear reference genome.

It is worth noting that a pan-genome has been constructed for apple; however, it is built from the sequences of only three samples from across the genus *Malus* (X. Sun, Jiao, et al., 2020). In the future, apple research is likely to benefit from the addition of dozens of samples to the pan-genome, as seen in other crops in which pan-genome builds incorporate upwards of 3000 samples (W. Wang et al., 2018). Looking forward, it is almost certain that the current gold standard reference genome in apple, derived from the Golden Delicious variety (Daccord et al., 2017), will be replaced soon with an apple pan-genome.

K-mer based association mapping

Methods that allow one to capture a wide spectrum of genetic variation without the cost-intensive genotyping of SVs at scale will be crucial for effective association mapping in the future. One such approach, which observes unique k-mers from short read sequencing data to detect genetic variation (Voichek & Weigel, 2020), has proved to be a significant improvement over traditional GWAS methods that make use only of SNPs or short indels (Lemay et al., 2023). Using this method, k-mers are derived from short sequence reads and are used as a proxy for genetic variants in the genome. The presence/absence of k-mers in a population can be used to conduct association studies to detect polymorphisms associated with a phenotype. The primary advantage of the k-mer approach is that k-mers are able to detect any genetic variation so long as that variation produces a unique k-mer in short read sequencing data. This has been shown to effectively capture the vast majority of SVs in plant genomes (Voichek & Weigel, 2020). Therefore, the k-mer method is not limited to subsets of genetic variation like traditional GWAS datasets (e.g., SNPs and indels only), and provides a method of more comprehensively surveying the genome for variation. Another important advantage of the k-mer based approach is that it does not make use of a reference genome prior to the association test, avoiding much of the bias and error introduced when aligning sequence reads to a reference genome. After the association test, k-mers associated with a phenotype can be traced back to the sequencing reads from which they originated and then mapped to one or more reference genomes, effectively locating causal alleles. A recent comparison in soybean demonstrated that this approach is significantly better at locating causal alleles via association studies than traditional SV or SNP based GWAS (Lemay et al., 2023). However, the k-mer method has been reported to produce higher numbers of false positives and requires more intensive and careful interpretation of results. At present, a primary barrier for the application of the k-mer approach in apple is that it relies on the presence/absence of k-mers, which may not be suitable for heterozygous crops, in which k-mers for both alleles would be observed at heterozygous sites (Lemay et al., 2023; Voichek & Weigel, 2020). Thus, to date, it has only been applied to inbred crops that are homozygous (Colque-Little et al., 2021; Tripodi et al., 2021). Despite this, I applied this methodology to the WGS generated in chapter 4 and did detect a strong signal associated with ripening time at the *NAC18.1* locus (data not shown). That being said, the downstream interpretation of the results was burdensome both computationally and conceptually, and further work on that project is required to be confident in the quality and validity of the results. The detection of a preliminary signal in the *NAC18.1* region is encouraging and suggests potential for k-mer applications in heterozygous crops like apple in the future. Fortunately, a number of tools have been developed to aid in analysis of k-mer results (Lemay et al., 2023), and new methods using k-mer counts (Cheng He, et al., 2021) may be more suitable for heterozygous crops.

Association mapping to gene editing

In the future, targeted mutagenesis via gene editing using a small number of candidate causal alleles from association studies may be the fastest path towards improved apple varieties. As seen with recent studies in peach (Cirilli et al., 2022; Tan et al., 2021), association studies followed by next generation high depth sequencing in large populations can discover the location of causal alleles with a high degree of confidence. Therefore, it is conceivable that targeted mutagenesis of potentially causal alleles could be undertaken immediately following GWAS and high depth sequencing, rather than investing time and resources in functional genomics studies to build confidence in the identity of candidate causal alleles. Although functional experiments often provide unique insights into the mechanistic control of plant phenotypes, the approximate location of the causal allele is not frequently refined using these approaches (X. Li et al., 2023; Migicovsky et al., 2021). Therefore, functional genomics often provides a deeper understanding of the control of a given phenotype than is necessary for producing an improved plant, and may be an ineffective use of resources in some cases. Rather than investing in functional genomics experiments that often produce modest increases in confidence in the identity of causal alleles, it is conceivable to gene edit numerous apple plants using a list of lower-confidence causal alleles from association studies for evaluation. Although most edits would provide no changes to the targeted phenotype, gene editing of potentially causal regions could provide improved phenotypes by screening large numbers of edited plants even if the genetic variation introduced is novel or poorly understood. This approach was discussed earlier in this document (See

Chapter 1) as “promoter bashing”, and could be used to introduce variation to regions of the genome that are highly likely to impact phenotypes. It leverages the sequence specificity of gene editing and could be conceptualised as a more precise form of classic mutagenesis. Studies have demonstrated the efficacy of this approach across multiple agricultural species including tobacco, soybean, and rice (Bao et al., 2019; J. Gao et al., 2015; C. Li et al., 2022; X. Song et al., 2022). By editing numerous lower confidence causal alleles across diverse apple plants at scale, it is plausible that targeted mutagenesis could be the fastest way to improve apple varieties. I am hopeful that with breakthroughs in other areas of gene editing (described above) that targeted mutagenesis approaches can be applied to apple in the future.

Advancements in Gene Editing

Although I described a sense of dampened optimism towards gene editing technologies (see Final Conclusions), it is worth briefly discussing recent advancements in the area and the potential the technology may offer apple improvement in the near future. Novel discoveries in CRISPR-cas protein complexes have opened a number of paths towards genomic improvement in apple, and crops more broadly. PrimeRoots editing (C. Sun et al., 2023) is a recent discovery that produces sequence specific insertions and deletions up to 11kb in plant genomes. In combination with the gene editing capabilities previously discussed, this effectively opens the door for nearly all allele swaps and precision insertions and deletions, which could theoretically allow for the precise addition, removal, or conversion of many causal alleles found in plants. Genome modification through CRISPR-based approaches is a rapidly evolving field, with advances happening frequently (Kweon et al., 2024; R. Liang et al., 2024). Multiple groups are making striking progress on improving the efficiency, specificity and versatility of CRISPR systems though the careful engineering of Cas proteins and CRISPR RNAs (D. Y. Kim et al., 2022; H. Lee et al., 2019; Yan et al., 2017). It is safe to assume that the capabilities of CRISPR based systems are likely to expand in the near future. The full CRISPR gene editing suite of technologies in their current and future forms offer the ability to generate a wide array of genetic variation and could play a vital role in apple improvement in the future if significant organism specific barriers (see Final Conclusions) are addressed.

Ribozyme-based genome editing technologies

In 2024, hydrolytic endonucleolytic ribozymes (HYERs) were reported to have DNA cleaving capabilities, making the HYER system a candidate as a novel gene editing technology (Z.-X. Liu et al., 2024). Early reports of the technology suggest HYERs could work across the tree of life as they have been shown to be active in both *E. coli* and mammalian genomes (Z.-X. Liu et al., 2024). HYERs are a small complex of single stranded RNAs that exhibit sequence-specific DNA targeting. In addition, the HYER system is far smaller (0.6kb) than previously discovered CRISPR (2-3kb) systems, which is attractive in situations where CRISPR construct sizes are prohibitively large. Already, careful engineering of native HYER systems has extended the recognition sequence of HYER to 20bp, making it equally as specific as CRISPR systems (Z.-X. Liu et al., 2024). Because this is an emerging discovery with few formal demonstrations, it remains unclear what role this technology could play in agriculture, or if it will be a suitable method for gene editing. However, the discovery of a class of biological molecules that can be programmed to alter the genome outside of the CRISPR system suggests that numerous similar systems may exist in the natural world, and that we are likely at the beginning of a genome engineering tool discovery era.

Genetic Modification Technology

Genetic modification (GM) refers to the suite of methods that take DNA sequences from another species and integrate those sequences into a host genome (Riva et al., 1998; SB Gelvin, 2003). These methods offer the ability to introduce novel traits or make targeted gene knockouts in an organism, the former being arguably the most powerful crop improvement approach to date. The ability to move sequences across the tree of life has revolutionised agriculture in many crops (Bullock & Nitsi, 2001; Pray et al., 2002). For apple, fire blight resistance sequences have been successfully moved from a wild apple species, *M. robusta*, to domesticated apple to produce a fire blight resistant Gala variety that is now in the 5th year of field trials and showing great promise (Kost et al., 2015; Schlathölter et al., 2023). Given that the ABC contains dozens of wild apple cultivars, the ABC could contribute to apple improvement through this approach if causal sequences are discovered in the wild population. Since disease related traits arguably offer the most impact to the agricultural industry, and because disease resistance often comes from wild relatives or distantly related species (Hajjar & Hodgkin, 2007; Love, 1999; Wilson et al., 2000), this technology offers tremendous potential for apple improvement. GM approaches have already shown market traction in apple, with Okanagan Specialty Fruits' non-browning Arctic Apple® varieties being a commercial success (Neal Carter, personal communications, 2022). Of course, a primary barrier for GM technology is public perception and governmental regulation. While consumer acceptance of GMOs seems to be increasing (Marette et al., 2021; Wunderlich & Gatto, 2015), as demonstrated by the growing movement around the Purple Tomato® (Nathan Pumplin, personal communications, 2024), governmental regulation across countries remains volatile and uncertain. Further, the regulatory process to get GMO plants to market in many countries is burdensome and arguably discourages innovative investment in this area. It is clear that any plant varieties with GMO status will face significant challenges in some areas of the world, and will be subject to some level of disapproval from a portion of the consumer base for the foreseeable future. However, I remain optimistic that GM technologies can enable the generation of significantly improved apple varieties through gene knockouts and the movement of cis- and trans-genes in the future.

Artificial intelligence (AI) in genetic mapping

In the last few years, the use of artificial intelligence (AI) has been touted as the solution to numerous modern challenges from city traffic to online dating (G. Chen & Zhang, 2022; Y. Wu & Kelly, 2021). In biology, numerous groups are working to apply cutting edge AI methods in the hopes of accelerating scientific discovery (Buchelt et al., 2024; Hassoun et al., 2022; Nagarajan et al., 2019). For example Google’s AlphaFold 3 AI model, which has primarily been used to predict protein folding structure, has recently been introduced to the field of science, and AlphaFold 3 founders claim that the model “predicts the structure and interactions of all of life’s molecules”. While no doubt the technology is promising, as evidenced by over 20,000 citations and pioneering of a novel research field (Jumper et al., 2021), I am cautious of the confidence some scientists have in such technologies. Further, other teams which use proprietary AI breeding techniques in the pursuit of advancing agriculture, have yet to realise success in the form of a field-proven plant or market successful product (Mariano Alvarez, personal communications, 2023; Nathan Pumplin, personal communications, 2024). I have no doubt that AI will play a role in the advancement of crop improvement in the future, however it is not clear how this technology will aid in addressing major challenges such as the phenotyping bottleneck (Furbank & Tester, 2011), the collection and curation of diverse crop varieties (Iezzoni et al., 2020), and tissue culture and regeneration (Atkins & Voytas, 2020; Venezia & Creasey Krainer, 2021).

In conclusion, apple breeding and improvement remains challenging and simultaneous advancements in multiple disciplines will be key for marked acceleration of apple variety improvement. In terms of causal allele discovery, association mapping and the use of multiple forms of WGS will be key in the future and the ability to detect complex genetic variation will be essential. Further, it will be important to use multiple mapping methodologies and pangenomes to not only identify efficacious breeding targets for selective breeding, but also to identify causal alleles and genes from other species that can offer benefits to apple via gene editing and genetic modification technologies.

REFERENCES

Abberton, M., Batley, J., Bentley, A., Bryant, J., Cai, H., Cockram, J., de Oliveira, A. C., Cseke, L. J., Dempewolf, H., De Pace, C., Edwards, D., Gepts, P., Greenland, A., Hall, A. E., Henry, R., Hori, K., Howe, G. T., Hughes, S., Humphreys, M., Yano, M. (2016). Global agricultural intensification during climate change: a role for genomics. *Plant Biotechnology Journal*, *14*(4), 1095–1098.

Abdellaoui, A., Yengo, L., Verweij, K. J. H., & Visscher, P. M. (2023). 15 years of GWAS discovery: Realizing the promise. *American Journal of Human Genetics*, *110*(2), 179–194.

Adams-Phillips, L., Barry, C., & Giovannoni, J. (2004). Signal transduction systems regulating fruit ripening. *Trends in Plant Science*, *9*(7), 331–338.

Agaoua, A., Rittener, V., Troadec, C., Desbiez, C., Bendahmane, A., Moquet, F., & Dogimont, C. (2022). A single substitution in Vacuolar protein sorting 4 is responsible for resistance to Watermelon mosaic virus in melon. *Journal of Experimental Botany*, *73*(12), 4008–4021.

Agriculture and Agri-Food Canada. (2021). *Statistical Overview of the Canadian Fruit Industry 2021*. Government of Canada. Retrieved June 29, 2023, from https://agriculture.canada.ca/en/sector/horticulture/reports/statistical-overview-canadian-fruit-industry-2021

Ahmad, R., & Anjum, M. A. (2018). Applications of molecular markers to assess genetic diversity in vegetable and ornamental crops-a review. *Journal of Horticultural Science and Technology*, *1*, 1–7.

Ahmed, S. I., Hayat, M. Q., Tahir, M., Mansoor, Q., Ismail, M., Keck, K., & Bates, R. B. (2016). Pharmacologically active flavonoids from the anticancer, antioxidant and antimicrobial extracts of Cassia angustifolia Vahl. *BMC Complementary and Alternative Medicine*, *16*(1), 460.

Alexa, A., & Rahnenfuhrer, J. (2020, October 30). *topGO: Enrichment Analysis for Gene Ontology*. https://rdrr.io/bioc/topGO/

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., Levy, Y., Harel, T. H., Shalev-Schlosser, G., Amsellem, Z., Razifard, H., Caicedo, A. L., Tieman, D. M., Klee, H., Kirsche, M., Lippman, Z. B. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, *182*(1), 145–161.e23.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Amyotte, B., Bowen, A. J., Banks, T., Rajcan, I., & Somers, D. J. (2017). Mapping the sensory perception of apple using descriptive sensory evaluation in a genome wide association study. *PloS One*, *12*(2), e0171710.

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Anzalone, A. V., Gao, X. D., Podracky, C. J., Nelson, A. T., Koblan, L. W., Raguram, A., Levy, J. M., Mercer, J. A. M., & Liu, D. R. (2021). Programmable deletion, replacement, integration and inversion of large DNA sequences with twin prime editing. *Nature Biotechnology*.

Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., Raguram, A., & Liu, D. R. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, *576*(7785), 149–157.

Aprea, E., Charles, M., Endrizzi, I., Laura Corollaro, M., Betta, E., Biasioli, F., & Gasperi, F. (2017). Sweet taste in apple: the role of sorbitol, individual sugars, organic acids and volatile compounds. *Scientific Reports*, *7*, 44950.

Atkins, P. A., & Voytas, D. F. (2020). Overcoming bottlenecks in plant gene editing. *Current Opinion in Plant Biology*, *54*, 79–84.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, *465*(7298), 627–631.

Babu, R., Nair, S., Prasanna, B., & Gupta, H. (2004). Integrating marker-assisted selection in crop breeding: Prospects and challenges. *Current Science*, *87*, 607–619.

Bai, Y., Dougherty, L., Li, M., Fazio, G., Cheng, L., & Xu, K. (2012). A natural mutation-led truncation in one of the two aluminum-activated malate transporter-like genes at the Ma locus is associated with low fruit acidity in apple. *Molecular Genetics and Genomics: MGG*, *287*(8), 663–678.

Ban, S., & Xu, K. (2020). Identification of two QTLs associated with high fruit acidity in apple using pooled genome sequencing analysis. *Horticulture Research*, *7*, 171.

Ban, Y., Honda, C., Hatsuyama, Y., Igarashi, M., Bessho, H., & Moriguchi, T. (2007). Isolation and functional analysis of a MYB transcription factor gene that is a key regulator for the development of red coloration in apple skin. *Plant & Cell Physiology*, *48*(7), 958–970.

Bao, A., Chen, H., Chen, L., Chen, S., Hao, Q., Guo, W., Qiu, D., Shan, Z., Yang, Z., Yuan, S., Zhang, C., Zhang, X., Liu, B., Kong, F., Li, X., Zhou, X., Tran, L.-S. P., & Cao, D. (2019). CRISPR/Cas9-mediated targeted mutagenesis of GmSPL9 genes alters plant architecture in soybean. *BMC Plant Biology*, *19*(1), 131.

Barry, C. S., & Giovannoni, J. J. (2007). Ethylene and Fruit Ripening. *Journal of Plant Growth Regulation*, *26*(2), 143.

Baumgartner, I. O., Kellerhals, M., Costa, F., Dondini, L., Pagliarani, G., Gregori, R., Tartarini, S., Leumann, L., Laurens, F., & Patocchi, A. (2016). Development of SNP-based assays for disease resistance and fruit quality traits in apple (Malus × domestica Borkh.) and validation in breeding pilot studies. *Tree Genetics & Genomes*, *12*(3).

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, *6*(8), 914–920.

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* , *58*(3), 268–276.

Bernard, A., & Joubès, J. (2013). Arabidopsis cuticular waxes: advances in synthesis, export and regulation. *Progress in Lipid Research*, *52*(1), 110–129.

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Science*, *48*(5), 1649–1664.

Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denancé, C., Théron, A., Poncet, C., Micheletti, D., Kerschbamer, E., Di Pierro, E. A., Larger, S., Pindo, M., Van de Weg, E., Davassi, A., Laurens, F., Velasco, R., Durel, C.-E., & Troggio, M. (2016). Development and validation of the Axiom(®) Apple480K SNP genotyping array. *The Plant Journal: For Cell and Molecular Biology*, *86*(1), 62–74.

Bink, M. C. A. M., Jansen, J., Madduri, M., Voorrips, R. E., Durel, C.-E., Kouassi, A. B., Laurens, F., Mathis, F., Gessler, C., Gobbin, D., Rezzonico, F., Patocchi, A., Kellerhals, M., Boudichevskaia, A., Dunemann, F., Peil, A., Nowicka, A., Lata, B., Stankiewicz-Kosyl, M., van de Weg, W. E. (2014). Bayesian QTL analyses using pedigreed families of an outcrossing species, with application to fruit firmness in apple. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *127*(5), 1073–1090.

Bin, L., Domingo, M. S., Mayobre, C., Martín-Hernández, A. M., Pujol, M., & Garcia-Mas, J. (2022). Knock-out of CmNAC-NOR affects melon climacteric fruit ripening. In *bioRxiv* (p. 2022.02.02.478821). https://doi.org/10.1101/2022.02.02.478821

Blanpied, G. D. (1972). A study of ethylene in apple, red raspberry, and cherry. *Plant Physiology*, *49*(4), 627–630.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* , *30*(15), 2114–2120.

Bortesi, L., & Fischer, R. (2015). The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnology Advances*, *33*(1), 41–52.

Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., Bergelson, J., Cuguen, J., & Roux, F. (2010). Linkage and association mapping of Arabidopsis thaliana flowering time in nature. *PLoS Genetics*, *6*(5), e1000940.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* , *23*(19), 2633–2635.

Brozynska, M., Furtado, A., & Henry, R. J. (2016). Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnology Journal*, *14*(4), 1070–1085.

Buchelt, A., Adrowitzer, A., Kieseberg, P., Gollob, C., Nothdurft, A., Eresheim, S., Tschiatschek, S., Stampfer, K., & Holzinger, A. (2024). Exploring artificial intelligence for applications of drones in forest ecology and management. *Forest Ecology and Management*, *551*, 121530.

Buckley, R. M., Harris, A. C., Wang, G.-D., Whitaker, D. T., Zhang, Y.-P., & Ostrander, E. A. (2022). Best practices for analyzing imputed genotypes from low-pass sequencing in dogs. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, *33*(1), 213–229.

Bullock, D., & Nitsi, E. I. (2001). 4 - GMO Adoption and Private Cost Savings: GR Soybeans and Bt Corn. In G. C. Nelson (Ed.), *Genetically Modified Organisms in Agriculture* (pp. 21–38). Academic Press.

Burleigh, J. G., Alphonse, K., Alverson, A. J., Bik, H. M., Blank, C., Cirranello, A. L., Cui, H., Daly, M., Dietterich, T. G., Gasparich, G., Irvine, J., Julius, M., Kaufman, S., Law, E., Liu, J., Moore, L., O'Leary, M. A., Passarotti, M., Ranade, S., … Yu, M. (2013). Next-generation phenomics for the Tree of Life. *PLoS Currents*, *5*.

Busatto, N., Matsumoto, D., Tadiello, A., Vrhovsek, U., & Costa, F. (2019). Multifaceted analyses disclose the role of fruit size and skin-russeting in the accumulation pattern of phenolic compounds in apple. *PloS One*, *14*(7), e0219354.

Busatto, N., Tadiello, A., Trainotti, L., & Costa, F. (2017). Climacteric ripening of apple fruit is regulated by transcriptional circuits stimulated by cross-talks between ethylene and auxin. *Plant Signaling & Behavior*, *12*(1), e1268312.

Butelli, E., Titta, L., Giorgio, M., Mock, H.-P., Matros, A., Peterek, S., Schijlen, E. G. W. M., Hall, R. D., Bovy, A. G., Luo, J., & Martin, C. (2008). Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors. *Nature Biotechnology*, *26*(11), 1301–1308.

Carbonell-Bejerano, P., Royo, C., Mauri, N., Ibáñez, J., & Miguel Martínez Zapater, J. (2019). Somatic Variation and Cultivar Innovation in Grapevine. In A. Morata & I. Loira (Eds.), *Advances in Grape and Wine Biotechnology*. IntechOpen.

Cardi, T., Murovec, J., Bakhsh, A., Boniecka, J., Bruegmann, T., Bull, S. E., Eeckhaut, T., Fladung, M., Galovic, V., Linkiewicz, A., Lukan, T., Mafra, I., Michalski, K., Kavas, M., Nicolia, A., Nowakowska, J., Sági, L., Sarmiento, C., Yıldırım, K., … Van Laere, K. (2023). CRISPR/Cas-mediated plant genome editing: outstanding challenges a decade after implementation. *Trends in Plant Science*, 4(2).

Cardoso, C., Zhang, Y., Jamil, M., Hepworth, J., Charnikhova, T., Dimkpa, S. O. N., Meharg, C., Wright, M. H., Liu, J., Meng, X., Wang, Y., Li, J., McCouch, S. R., Leyser, O., Price, A. H., Bouwmeester, H. J., & Ruyter-Spira, C. (2014). Natural variation of rice strigolactone biosynthesis is associated with the deletion of two MAX1 orthologs. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(6), 2379–2384.

Carew, R., & Smith, E. G. (2004). The value of apple characteristics to wholesalers in western Canada: A hedonic approach. *Canadian Journal of Plant Science*, *84*(3), 829–835.

Chagné, D., Carlisle, C. M., Blond, C., Volz, R. K., Whitworth, C. J., Oraguzie, N. C., Crowhurst, R. N., Allan, A. C., Espley, R. V., Hellens, R. P., & Gardiner, S. E. (2007). Mapping a candidate gene (MdMYB10) for red flesh and foliage colour in apple. *BMC Genomics*, *8*, 212.

Chagné, D., Crowhurst, R. N., Troggio, M., Davey, M. W., Gilmore, B., Lawley, C., Vanderzande, S., Hellens, R. P., Kumar, S., Cestaro, A., Velasco, R., Main, D., Rees, J. D., Iezzoni, A., Mockler, T., Wilhelm, L., Van de Weg, E., Gardiner, S. E., Bassil, N., & Peace, C. (2012). Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PloS One*, *7*(2), e31745.

Chagné, D., Dayatilake, D., Diack, R., Oliver, M., Ireland, H., Watson, A., Gardiner, S. E., Johnston, J. W., Schaffer, R. J., & Tustin, S. (2014). Genetic and environmental control of fruit maturation, dry matter and firmness in apple (Malus × domestica Borkh.). *Horticulture Research*, *1*, 14046.

Chagné, D., Krieger, C., Rassam, M., Sullivan, M., Fraser, J., André, C., Pindo, M., Troggio, M., Gardiner, S. E., Henry, R. A., Allan, A. C., McGhie, T. K., & Laing, W. A. (2012). QTL and candidate gene mapping for polyphenolic composition in apple fruit. *BMC Plant Biology*, *12*, 12.

Chagné, D., Vanderzande, S., Kirk, C., Profitt, N., Weskett, R., Gardiner, S. E., Peace, C. P., Volz, R. K., & Bassil, N. V. (2019). Validation of SNP markers for fruit quality and disease resistance loci in apple (Malus × domestica Borkh.) using the OpenArray® platform. *Horticulture Research*, *6*, 30.

Chai, Y., Li, A., Chit Wai, S., Song, C., Zhao, Y., Duan, Y., Zhang, B., & Lin, Q. (2020). Cuticular wax composition changes of 10 apple cultivars during postharvest storage. *Food Chemistry*, *324*, 126903.

Charrier, A., Vergne, E., Dousset, N., Richer, A., Petiteau, A., & Chevreau, E. (2019). Efficient Targeted Mutagenesis in Apple and First Time Edition of Pear Using the CRISPR-Cas9 System. *Frontiers in Plant Science*, *10*, 40.

Cheng He, et al. (2021). Trait Association and Prediction Through Integrative K-mer Analysis. *BioRxiv*. https://www.biorxiv.org/content/10.1101/2021.11.17.468725v1.full

Chen, G., & Zhang, J. (2022). Applying Artificial Intelligence and Deep Belief Network to predict traffic congestion evacuation performance in smart cities. *Applied Soft Computing*, *121*, 108692.

Chen, J., Wang, Z., Tan, K., Huang, W., Shi, J., Li, T., Hu, J., Wang, K., Wang, C., Xin, B., Zhao, H., Song, W., Hufford, M. B., Schnable, J. C., Jin, W., & Lai, J. (2023). A complete telomere-to-telomere assembly of the maize genome. *Nature Genetics*, 1–11.

Chen, K., Wang, Y., Zhang, R., Zhang, H., & Gao, C. (2019). CRISPR/Cas Genome Editing and Precision Plant Breeding in Agriculture. *Annual Review of Plant Biology*, *70*, 667–697.

Chen, P. J., & Liu, D. R. (2022). Prime editing for precise and highly versatile genome manipulation. *Nature Reviews. Genetics*, 1–17.

Chu, W., Gao, H., Chen, H., Fang, X., & Zheng, Y. (2018). Effects of cuticular wax on the postharvest quality of blueberry fruit. *Food Chemistry*, *239*, 68–74.

Cirilli, M., Rossini, L., Chiozzotto, R., Baccichet, I., Florio, F. E., Mazzaglia, A., Turco, S., Bassi, D., & Gattolin, S. (2022). Less is more: natural variation disrupting a miR172 gene at the di locus underlies the recessive double-flower trait in peach (*P. persica L. Batsch*). *BMC Plant Biology*, *22*(1), 318.

Cliff, M. A., Stanich, K., Lu, R., & Hampson, C. R. (2016). Use of descriptive analysis and preference mapping for early-stage assessment of new and established apples. *Journal of the Science of Food and Agriculture*, *96*(6), 2170–2183.

Colque-Little, C., Abondano, M. C., Lund, O. S., Amby, D. B., Piepho, H.-P., Andreasen, C., Schmöckel, S., & Schmid, K. (2021). Genetic variation for tolerance to the downy mildew pathogen Peronospora variabilis in genetic resources of quinoa (*Chenopodium quinoa*). *BMC Plant Biology*, *21*(1), 41.

Conner, P. J., Brown, S. K., & Weeden, N. F. (1998). Molecular-marker analysis of quantitative traits for growth and development in juvenile apple trees. *TAG. Theoretical and Applied Genetics*, *96*(8), 1027–1035.

Cornille, A., Antolín, F., Garcia, E., Vernesi, C., Fietta, A., Brinkkemper, O., Kirleis, W., Schlumbaum, A., & Roldán-Ruiz, I. (2019). A Multifaceted Overview of Apple Tree Domestication. *Trends in Plant Science*, *24*(8), 770–782.

Cornille, A., Giraud, T., Smulders, M. J. M., Roldán-Ruiz, I., & Gladieux, P. (2014). The domestication and evolutionary ecology of apples. *Trends in Genetics: TIG*, *30*(2), 57–65.

Cornille, A., Gladieux, P., & Giraud, T. (2013). Crop-to-wild gene flow and spatial genetic structure in the closest wild relatives of the cultivated apple. *Evolutionary Applications*, *6*(5), 737–748.

Cornille, A., Gladieux, P., Smulders, M. J. M., Roldán-Ruiz, I., Laurens, F., Le Cam, B., Nersesyan, A., Clavel, J., Olonova, M., Feugey, L., Gabrielyan, I., Zhang, X.-G., Tenaillon, M. I., & Giraud, T. (2012). New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genetics*, *8*(5), e1002703.

Costa, F. (2015). MetaQTL analysis provides a compendium of genomic loci controlling fruit quality traits in apple. *Tree Genetics & Genomes 11*(1).

Costa, F., Peace, C. P., Stella, S., Serra, S., Musacchi, S., Bazzani, M., Sansavini, S., & Van de Weg, W. E. (2010). QTL dynamics for fruit firmness and softening around an ethylene-dependent polygalacturonase gene in apple (*Malus×domestica Borkh.*). *Journal of Experimental Botany*, *61*(11), 3029–3039.

Cowan, M. F., Blomstedt, C. K., Norton, S. L., Henry, R. J., Møller, B. L., & Gleadow, R. (2020). Crop wild relatives as a genetic resource for generating low-cyanide, drought-tolerant Sorghum. *Environmental and Experimental Botany*, *169*, 103884.

Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., Velasco, R., Di Pierro, E. A., Gouzy, J., Rees, D. J. G., Guérif, P., Muranty, H., Durel, C.-E., Laurens, F., Lespinasse, Y., Gaillard, S., Bucher, E. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genetics*, *49*(7), 1099–1106.

Dalla Costa, L., Malnoy, M., & Gribaudo, I. (2017). Breeding next generation tree fruits: technical and legal challenges. *Horticulture Research*, *4*, 17067.

Dalla Costa, L., Piazza, S., Pompili, V., Salvagnin, U., Cestaro, A., Moffa, L., Vittani, L., Moser, C., & Malnoy, M. (2020). Strategies to produce T-DNA free CRISPRed fruit trees via Agrobacterium tumefaciens stable gene transfer. *Scientific Reports*, *10*(1), 20155.

Dandekari, A. M., Teo, G., Defilippi, B. G., Uratsu, S. L., Passey, A. J., Kader, A. A., Stow, J. R., Colgan, R. J., & James, D. J. (2004). Effect of down-regulation of ethylene biosynthesis on fruit flavor complex in apple fruit. *Transgenic Research*, *13*(4), 373–384.

Davies, T., & Myles, S. (2023). Pool-seq of diverse apple germplasm reveals candidate loci underlying ripening time, phenolic content, and softening. *Fruit Research*, *3*(1).

Defilippi, B. G., Dandekar, A. M., & Kader, A. A. (2005). Relationship of ethylene biosynthesis to volatile production, related enzymes, and precursor availability in apple peel and flesh tissues. *Journal of Agricultural and Food Chemistry*, *53*(8), 3133–3141.

Denay, G., Vachon, G., Dumas, R., Zubieta, C., & Parcy, F. (2017). Plant SAM-Domain Proteins Start to Reveal Their Roles. *Trends in Plant Science*, *22*(8), 718–725.

Di Guardo, M., Bink, M. C. A. M., Guerra, W., Letschka, T., Lozano, L., Busatto, N., Poles, L., Tadiello, A., Bianco, L., Visser, R. G. F., van de Weg, E., & Costa, F. (2017). Deciphering the genetic control of fruit texture in apple by multiple family-based analysis and genome-wide association. *Journal of Experimental Botany*, *68*(7), 1451–1466.

Ding, T., Tomes, S., Gleave, A. P., Zhang, H., Dare, A. P., Plunkett, B., Espley, R. V., Luo, Z., Zhang, R., Allan, A. C., Zhou, Z., Wang, H., Wu, M., Dong, H., Liu, C., Liu, J., Yan, Z., & Yao, J.-L. (2022). microRNA172 targets APETALA2 to regulate flavonoid biosynthesis in apple (Malus domestica). *Horticulture Research*.

Dirlewanger, E., Quero-García, J., Le Dantec, L., Lambert, P., Ruiz, D., Dondini, L., Illa, E., Quilot-Turion, B., Audergon, J.-M., Tartarini, S., Letourmy, P., & Arús, P. (2012). Comparison of the genetic determinism of two key phenological traits, flowering and maturity dates, in three Prunus species: peach, apricot and sweet cherry. *Heredity*, *109*(5), 280–292.

Dixon, R. A., Xie, D.-Y., & Sharma, S. B. (2005). Proanthocyanidins--a final frontier in flavonoid research? *The New Phytologist*, *165*(1), 9–28.

Dong, W., Wu, D., Li, G., Wu, D., & Wang, Z. (2018). Next-generation sequencing from bulked segregant analysis identifies a dwarfism gene in watermelon. *Scientific Reports*, *8*(1), 2908.

Donkpegan, A. S. L., Bernard, A., Barreneche, T., Quero-García, J., Bonnet, H., Fouché, M., Le Dantec, L., Wenden, B., & Dirlewanger, E. (2023). Genome-wide association mapping in a sweet cherry germplasm collection (*Prunus avium L.*) reveals candidate genes for fruit quality traits. *Horticulture Research*, *10*(10).

Do, P. T., Nguyen, C. X., Bui, H. T., Tran, L. T. N., Stacey, G., Gillman, J. D., Zhang, Z. J., & Stacey, M. G. (2019). Demonstration of highly efficient dual gRNA CRISPR/Cas9 editing of the homeologous GmFAD2-1A and GmFAD2-1B genes to yield a high oleic, low linoleic and α-linolenic acid phenotype in soybean. *BMC Plant Biology*, *19*(1), 311.

Doudna, J. A., & Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*(6213), 1258096.

Dougherty, L., Singh, R., Brown, S., Dardick, C., & Xu, K. (2018). Exploring DNA variant segregation types in pooled genome sequencing enables effective mapping of weeping trait in Malus. *Journal of Experimental Botany*, *69*(7), 1499–1516.

D. Turner, S. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software*, *3*(25), 731.

Duan, N., Bai, Y., Sun, H., Wang, N., Ma, Y., Li, M., Wang, X., Jiao, C., Legall, N., Mao, L., Wan, S., Wang, K., He, T., Feng, S., Zhang, Z., Mao, Z., Shen, X., Chen, X., Jiang, Y., Chen, X. (2017). Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nature Communications*, *8*(1), 249.

Eagen, K. P. (2018). Principles of Chromosome Architecture Revealed by Hi-C. *Trends in Biochemical Sciences*, *43*(6), 469–478.

Eccel, E., Rea, R., Caffarra, A., & Crisci, A. (2009). Risk of spring frost to apple production under future climate scenarios: the role of phenological acclimation. *International Journal of Biometeorology*, *53*(3), 273–286.

Edge-Garza, D. A., Luby, J. J., & Peace, C. (2015). Decision support for cost-efficient and logistically feasible marker-assisted seedling selection in fruit breeding. *Molecular Breeding: New Strategies in Plant Improvement*, *35*(12), 223.

Endo, M., Mikami, M., Endo, A., Kaya, H., Itoh, T., Nishimasu, H., Nureki, O., & Toki, S. (2019). Genome editing in plants by engineered CRISPR-Cas9 recognizing NG PAM. *Nature Plants*, *5*(1), 14–17.

Espley, R. V., Hellens, R. P., Putterill, J., Stevenson, D. E., Kutty-Amma, S., & Allan, A. C. (2007). Red colouration in apple fruit is due to the activity of the MYB transcription factor, MdMYB10. *The Plant Journal: For Cell and Molecular Biology*, *49*(3), 414–427.

Fabricant, F. (2022, February 14). Del Monte Pineapples Go Pink. *The New York Times*. https://www.nytimes.com/2022/02/14/dining/del-monte-pinkglow-pineapples.html

Fahrentrapp, J., Broggini, G. A. L., Kellerhals, M., Peil, A., Richter, K., Zini, E., & Gessler, C. (2013). A candidate gene for fire blight resistance in *Malus ×robusta* 5 is coding for a CC–NBS–LRR. *Tree Genetics & Genomes*, *9*(1), 237–251.

Fang, G., Hammar, S., & Grumet, R. (1992). A quick and inexpensive method for removing polysaccharides from plant genomic DNA. *BioTechniques*, *13*(1), 52–54, 56.

FAOSTAT. (2020, December 22). *FAOSTAT*. Food and Agriculture Association of the United States. http://www.fao.org/faostat/en/

FAOSTAT 2021. (2021). Retrieved August 24, 2022, from https://www.fao.org/faostat/en/#data/QV

Ferree, D. C., & Warrington, I. J. (2003). *Apples: Botany, Production, and Uses*. CABI.

Finger, R., & Möhring, N. (2024). The emergence of pesticide-free crop production systems in Europe. *Nature Plants*, *10*(3), 360–366.

Francia, E., Pecchioni, N., Policriti, A., & Scalabrin, S. (2015). CNV and Structural Variation in Plants: Prospects of NGS Approaches. In G. Sablok, S. Kumar, S. Ueno, J. Kuo, & C. Varotto (Eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches* (pp. 211–232). Springer International Publishing.

Funk, C. (2020, March 18). *About half of U.S. adults are wary of health effects of genetically modified foods, but many also see advantages*. Pew Research Center. https://www.pewresearch.org/fact-tank/2020/03/18/about-half-of-u-s-adults-are-wary-of-health-effects-of-genetically-modified-foods-but-many-also-see-advantages/

Furbank, R. T., & Tester, M. (2011). Phenomics--technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, *16*(12), 635–644.

Gabur, I., Chawla, H. S., Snowdon, R. J., & Parkin, I. A. P. (2019). Connecting genome structural variation with complex traits in crop plants. *TAG. Theoretical and Applied Genetics. 132*(3), 733–750.

Gao, J., Wang, G., Ma, S., Xie, X., Wu, X., Zhang, X., Wu, Y., Zhao, P., & Xia, Q. (2015). CRISPR/Cas9-mediated targeted mutagenesis in Nicotiana tabacum. *Plant Molecular Biology*, *87*(1-2), 99–110.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y., van der Knaap, E., Huang, S., Klee, H. J., Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, *51*(6), 1044–1051.

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I., & Liu, D. R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature*, *551*(7681), 464–471.

Gaut, B. S., Díez, C. M., & Morrell, P. L. (2015). Genomics and the Contrasting Dynamics of Annual and Perennial Domestication. *Trends in Genetics: TIG*, *31*(12), 709–719.

Genome Database for Rosaecea. (2023). Retrieved December 8, 2023, from https://www.rosaceae.org/species/malus/malus_x_domestica/genome_GDDH13_v1.1

Giovannoni, J. J. (2004). Genetic regulation of fruit development and ripening. *The Plant Cell*, *16* S170–S180.

Given, N. K., Venis, M. A., & Grierson, D. (1988). Phenylalanine Ammonia-Lyase Activity and Anthocyanin Synthesis in Ripening Strawberry Fruit. *Journal of Plant Physiology*, *133*(1), 25–30.

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., & Edwards, D. (2016). The pangenome of an agronomically important crop plant Brassica oleracea. *Nature Communications*, *7*, 13390.

Gonda, I., Faigenboim, A., Adler, C., Milavski, R., Karp, M.-J., Shachter, A., Ronen, G., Baruch, K., Chaimovitsh, D., & Dudai, N. (2020). The genome sequence of tetraploid sweet basil, *Ocimum basilicum L.*, provides tools for advanced genome editing and molecular breeding. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *27*(5).

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, *40*, D1178–D1186.

Gottschalk, C., & van Nocker, S. (2013). Diversity in seasonal bloom time and floral development among apple species and hybrids. *Journal of the American Society for Horticultural Science, 138*(5), 367–374.

Greenboim-Wainberg, Y., Maymon, I., Borochov, R., Alvarez, J., Olszewski, N., Ori, N., Eshed, Y., & Weiss, D. (2005). Cross talk between gibberellin and cytokinin: the Arabidopsis GA response inhibitor SPINDLY plays a positive role in cytokinin signaling. *The Plant Cell*, *17*(1), 92–102.

Gross, B. L., Volk, G. M., Richards, C. M., Reeves, P. A., Henk, A. D., Forsline, P. L., Szewc-McFadden, A., Fazio, G., & Thomas Chao, C. (2013). Diversity Captured in the USDA-ARS National Plant Germplasm System Apple Core Collection. *Journal of the American Society for Horticultural Science*, *138*(5), 375–381.

Guan, Y., & Stephens, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genetics*, *4*(12), e1000279.

Hajjar, R., & Hodgkin, T. (2007). The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica*, *156*(1-2), 1–13.

Hampson, C. R., Quamme, H. A., Hall, J. W., MacDonald, R. A., King, M. C., & Cliff, M. A. (2000). Sensory evaluation as a selection tool in apple breeding. *Euphytica*, *111*(2), 79–90.

Hanak, T., Madsen, C. K., & Brinch-Pedersen, H. (2022). Genome Editing-accelerated Re-Domestication (GEaReD) - A new major direction in plant breeding. *Biotechnology Journal*, *17*(7), e2100545.

Harker, F. R., Kupferman, E. M., Marin, A. B., Gunson, F. A., & Triggs, C. M. (2008). Eating quality standards for apples based on consumer preferences. *Postharvest Biology and Technology*, *50*(1), 70–78.

Harris, S. A., Robinson, J. P., & Juniper, B. E. (2002). Genetic clues to the origin of the apple. *Trends in Genetics: TIG*, *18*(8), 426–430.

Harshman, J. M., Evans, K. M., Allen, H., Potts, R., Flamenco, J., Aldwinckle, H. S., Wisniewski, M. E., & Norelli, J. L. (2017). Fire Blight Resistance in Wild Accessions of Malus sieversii. *Plant Disease*, *101*(10), 1738–1745.

Hasing, T., Tang, H., Brym, M., Khazi, F., Huang, T., & Chambers, A. H. (2020). A phased Vanilla planifolia genome enables genetic improvement of flavour and production. *Nature Food*, *1*(12), 811–819.

Hassoun, S., Jefferson, F., Shi, X., Stucky, B., Wang, J., & Rosa, E. (2022). Artificial Intelligence for Biology. *Integrative and Comparative Biology*, *61*(6), 2267–2275.

Hattori, Y., Nagai, K., Furukawa, S., Song, X.-J., Kawano, R., Sakakibara, H., Wu, J., Matsumoto, T., Yoshimura, A., Kitano, H., Matsuoka, M., Mori, H., & Ashikari, M. (2009). The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water. *Nature*, *460*(7258), 1026–1030.

Health Canada. (2022). *Guidelines for the Safety Assessment of Novel Foods*. Canada.ca. Retrieved November 10, 2022, from https://www.canada.ca/en/health-canada/services/food-nutrition/legislation-guidelines/guidance-documents/guidelines-safety-assessment-novel-foods-derived-plants-microorganisms/guidelines-safety-assessment-novel-foods-2006.html#a5

Heffner, E. L., Sorrells, M. E., & Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Science*, *49*(1), 1–12.

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., & Buell, C. R. (2014). Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell*, *26*(1), 121–135.

Hoang, X. L. T., Prerostova, S., Thu, N. B. A., Thao, N. P., Vankova, R., & Tran, L.-S. P. (2021). Histidine Kinases: Diverse Functions in Plant Development and Responses to Environmental Conditions. *Annual Review of Plant Biology*, *72*, 297–323.

Holderbaum, D. F., Kon, T., Kudo, T., & Guerra, M. P. (2010). Enzymatic Browning, Polyphenol Oxidase Activity, and Polyphenols in Four Apple Cultivars: Dynamics during Fruit Development. *HortScience*, *45*(8), 1150–1154.

Holton, T. A., & Cornish, E. C. (1995). Genetics and Biochemistry of Anthocyanin Biosynthesis. *The Plant Cell*, *7*(7), 1071–1083.

Holušová, K., Čmejlová, J., Suran, P., Čmejla, R., Sedlák, J., Zelený, L., & Bartoš, J. (2023). High-resolution genome-wide association study of a large Czech collection of sweet cherry (*Prunus avium L.*) on fruit maturity and quality traits. *Horticulture Research.*

Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. *Nature Reviews. Genetics*, *21*(3), 171–189.

Huber, G. M., & Rupasinghe, H. P. V. (2009). Phenolic profiles and antioxidant properties of apple skin extracts. *Journal of Food Science*, *74*(9), C693–C700.

Hu, L., Xu, Z., Wang, M., Fan, R., Yuan, D., Wu, B., Wu, H., Qin, X., Yan, L., Tan, L., Sim, S., Li, W., Saski, C. A., Daniell, H., Wendel, J. F., Lindsey, K., Zhang, X., Hao, C., & Jin, S. (2019). The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nature Communications*, *10*(1), 4702.

Hu, Y., Han, Z., Sun, Y., Wang, S., Wang, T., Wang, Y., Xu, K., Zhang, X., Xu, X., Han, Z., & Wu, T. (2020). ERF4 affects fruit firmness through TPL4 by reducing ethylene production. *The Plant Journal: For Cell and Molecular Biology*, *103*(3), 937–950.

Iezzoni, A. F., McFerson, J., Luby, J., Gasic, K., Whitaker, V., Bassil, N., Yue, C., Gallardo, K., McCracken, V., Coe, M., Hardner, C., Zurn, J. D., Hokanson, S., van de Weg, E., Jung, S., Main, D., da Silva Linge, C., Vanderzande, S., Davis, T. M., Peace, C. (2020). RosBREED: bridging the chasm between discovery and application to enable DNA-informed breeding in rosaceous crops. *Horticulture Research*, *7*(1), 177.

ISAAA. (2023). *GM Crops List*. https://www.isaaa.org/gmapprovaldatabase/cropslist/default.asp

Islam, T., Afroz, N., Koh, C., Hoque, M. N., Rahman, M. J., Gupta, D. R., Mahmud, N. U., Nahid, A. A., Islam, R., Bhowmik, P. K., & Sharpe, A. G. (2022). Whole-genome sequencing of a year-round fruiting jackfruit (Artocarpus heterophyllus Lam.) reveals high levels of single nucleotide variation. *Frontiers in Plant Science*, *13*, 1044420.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239.

Jänsch, M., Broggini, G. A. L., Weger, J., Bus, V. G. M., Gardiner, S. E., Bassett, H., & Patocchi, A. (2015). Identification of SNPs linked to eight apple disease resistance loci. *Molecular Breeding: New Strategies in Plant Improvement*, *35*(1).

Jenkins, D., Juba, N., Crawford, B., Worthington, M., & Hummel, A. (2023). Regulation of plants developed through new breeding techniques must ensure societal benefits. *Nature Plants*, *9*(5), 679–684.

Jia, H., Zhang, Y., Orbović, V., Xu, J., White, F. F., Jones, J. B., & Wang, N. (2017). Genome editing of the disease susceptibility gene CsLOB1 in citrus confers resistance to citrus canker. *Plant Biotechnology Journal*, *15*(7), 817–823.

Jiang, L., Shen, W., Liu, C., Tahir, M. M., Li, X., Zhou, S., Ma, F., & Guan, Q. (2022). Engineering drought-tolerant apple by knocking down six GH3 genes and potential application of transgenic apple as a rootstock. *Horticulture Research*, *9*, uhac122.

Jiao, W.-B., & Schneeberger, K. (2020). Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nature Communications*, *11*(1), 1–10.

Jia, X., Wang, Q., Ye, Y., Li, T., Sun, X., Huo, L., Wang, P., Gong, X., & Ma, F. (2022). MdATG5a positively regulates nitrogen uptake under low nitrogen conditions by enhancing the accumulation of flavonoids and auxin in apple roots. *Environmental and Experimental Botany*, *197*, 104840.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816–821.

Johnson, W. C., Forsline, P. L., Aldwinckle, H. S., Todd Holleran, H., Robinson, T. L., & Norelli, J. J. (1999). 051 The USDA-ARS/Cornell University Apple Rootstock Breeding and Evaluation Program. *HortScience*, *34*(3), 450A – 450.

Jugdé, H., Nguy, D., Moller, I., Cooney, J. M., & Atkinson, R. G. (2008). Isolation and characterization of a novel glycosyltransferase that converts phloretin to phlorizin, a potent antioxidant in apple. *The FEBS Journal*, *275*(15), 3804–3814.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Jung, M., Roth, M., Aranzana, M. J., Auwerkerken, A., Bink, M., Denancé, C., Dujak, C., Durel, C.-E., Font I Forcada, C., Cantin, C. M., Guerra, W., Howard, N. P., Keller, B., Lewandowski, M., Ordidge, M., Rymenants, M., Sanin, N., Studer, B., Zurawicz, E., Muranty, H. (2020). The apple REFPOP-a reference population for genomics-assisted breeding in apple. *Horticulture Research*, *7*(1), 189.

Jung, S., Lee, T., Cheng, C.-H., Buble, K., Zheng, P., Yu, J., Humann, J., Ficklin, S. P., Gasic, K., Scott, K., Frank, M., Ru, S., Hough, H., Evans, K., Peace, C., Olmstead, M., DeVetter, L. W., McFerson, J., Coe, M., Main, D. (2019). 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Research*, *47*(D1), D1137–D1145.

Juniper, B. E., & Mabberley, D. J. (2006). *The story of the apple*. Timber Press (OR).

Ju, Z., Liu, C., & Yuan, Y. (1995). Activities of chalcone synthase and UDPGal: flavonoid-3-o-glycosyltransferase in relation to anthocyanin synthesis in apple. *Scientia Horticulturae*, *63*(3), 175–185.

Karlson, D., Mojica, J. P., Poorten, T. J., Lawit, S. J., Jali, S., Chauhan, R. D., Pham, G. M., Marri, P., Guffy, S. L., Fear, J. M., Ochsenfeld, C. A., (Lincoln) Chapman, T. A., Casamali, B., Venegas, J. P., Kim, H. J., Call, A., Sublett, W. L., Mathew, L. G., Shariff, A., Rapp, R. (2022). Targeted Mutagenesis of the Multicopy Myrosinase Gene Family in Allotetraploid Brassica juncea Reduces Pungency in Fresh Leaves across Environments. *Plants*, *11*(19), 2494.

Kawabata, K., Yoshioka, Y., & Terao, J. (2019). Role of Intestinal Microbiota in the Bioavailability and Physiological Functions of Dietary Polyphenols. *Molecules* , *24*(2).

Kenny, E. E., Timpson, N. J., Sikora, M., Yee, M.-C., Moreno-Estrada, A., Eng, C., Huntsman, S., Burchard, E. G., Stoneking, M., Bustamante, C. D., & Myles, S. (2012). Melanesian blond hair is caused by an amino acid change in TYRP1. *Science*, *336*(6081), 554.

Khan, M. A., & Korban, S. S. (2012). Association mapping in forest trees and fruit crops. In *Journal of Experimental Botany* (Vol. 63, Issue 11, pp. 4045–4060).

Khan, M. A., Olsen, K. M., Sovero, V., Kushad, M. M., & Korban, S. S. (2014). Fruit quality traits have played critical roles in domestication of the apple. *The Plant Genome*, *7*(3), 1.

Khan, S. A., Chibon, P.-Y., de Vos, R. C. H., Schipper, B. A., Walraven, E., Beekwilder, J., van Dijk, T., Finkers, R., Visser, R. G. F., van de Weg, E. W., Bovy, A., Cestaro, A., Velasco, R., Jacobsen, E., & Schouten, H. J. (2012). Genetic analysis of metabolites in apple fruits indicates an mQTL hotspot for phenolic compounds on linkage group 16. *Journal of Experimental Botany*, *63*(8), 2895–2908.

Khan, S. A., Schaart, J. G., Beekwilder, J., Allan, A. C., Tikunov, Y. M., Jacobsen, E., & Schouten, H. J. (2012). The mQTL hotspot on linkage group 16 for phenolic compounds in apple fruits is probably the result of a leucoanthocyanidin reductase gene at that locus. *BMC Research Notes*, *5*, 618.

Kim, D. Y., Lee, J. M., Moon, S. B., Chin, H. J., Park, S., Lim, Y., Kim, D., Koo, T., Ko, J.-H., & Kim, Y.-S. (2022). Efficient CRISPR editing with a hypercompact Cas12f1 and engineered guide RNAs delivered by adeno-associated virus. *Nature Biotechnology*, *40*(1), 94–102.

Kim, K., Kim, M., Kim, Y., Lee, D., & Jung, I. (2022). Hi-C as a molecular rangefinder to examine genomic rearrangements. *Seminars in Cell & Developmental Biology*, *121*, 161–170.

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *32*(6), 524–536.

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* , *27*(24), 3435–3436.

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., & Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, *533*(7603), 420–424.

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, *9*, 29.

Kostick, S. A., Teh, S. L., & Evans, K. M. (2021). Contributions of Reduced Susceptibility Alleles in Breeding Apple Cultivars with Durable Resistance to Fire Blight. *Plants*, *10*(2).

Kostick, S. A., Teh, S. L., Norelli, J. L., Vanderzande, S., Peace, C., & Evans, K. M. (2021). Fire blight QTL analysis in a multi-family apple population identifies a reduced-susceptibility allele in “Honeycrisp.” *Horticulture Research*, *8*(1), 28.

Kost, T. D., Gessler, C., Jänsch, M., Flachowsky, H., Patocchi, A., & Broggini, G. A. L. (2015). Development of the First Cisgenic Apple with Increased Resistance to Fire Blight. *PloS One*, *10*(12), e0143980.

Kou, X., Feng, Y., Yuan, S., Zhao, X., Wu, C., Wang, C., & Xue, Z. (2021). Different regulatory mechanisms of plant hormones in the ripening of climacteric and non-climacteric fruits: a review. *Plant Molecular Biology*, *107*(6), 477–497.

Kramer, M. G., & Redenbaugh, K. (1994). Commercialization of a tomato with an antisense polygalacturonase gene: The FLAVR SAVR^{TM} tomato story. *Euphytica*, *79*(3), 293–297.

Kreitzman, M., Toensmeier, E., Chan, K. M. A., Smukler, S., & Ramankutty, N. (2020). Perennial Staple Crops: Yields, Distribution, and Nutrition in the Global Food System. *Frontiers in Sustainable Food Systems*, *4*.

Kumar, R., Khurana, A., & Sharma, A. K. (2014). Role of plant hormones and their interplay in development and ripening of fleshy fruits. *Journal of Experimental Botany*, *65*(16), 4561–4575.

Kumar, S., Bink, M. C. A. M., Volz, R. K., Bus, V. G. M., & Chagné, D. (2012). Towards genomic selection in apple (*Malus ×domestica Borkh.*) breeding programmes: Prospects, challenges and strategies. *Tree Genetics & Genomes*, *8*(1), 1–14.

Kumar, S., Deng, C. H., Molloy, C., Kirk, C., Plunkett, B., Lin-Wang, K., Allan, A., & Espley, R. (2022). Extreme-phenotype GWAS unravels a complex nexus between apple (*Malus domestica*) red-flesh colour and internal flesh browning. *Fruit Research*, *2*(1), 1–14.

Kumar, S., Garrick, D. J., Bink, M. C., Whitworth, C., Chagné, D., & Volz, R. K. (2013). Novel genomic approaches unravel genetic architecture of complex traits in apple. *BMC Genomics*, *14*, 393.

Kumar, S., Kirk, C., Deng, C., Wiedow, C., Knaebel, M., & Brewer, L. (2017). Genotyping-by-sequencing of pear (Pyrus spp.) accessions unravels novel patterns of genetic diversity and selection footprints. *Horticulture Research*, *4*, 17015.

Kumar, S., Raulier, P., Chagné, D., & Whitworth, C. (2014). Molecular-level and trait-level differentiation between the cultivated apple (Malus× domestica Borkh.) and its main progenitor *Malus sieversii*. *Plant Genetic Resources; Cambridge*, *12*(3), 330–340.

Kumar, S., Volz, R. K., Alspach, P. A., & Bus, V. G. M. (2010). Development of a recurrent apple breeding programme in New Zealand: a synthesis of results, and a proposed revised breeding strategy. *Euphytica*, *173*(2), 207–222.

Kumawat, G., Gupta, S., Ratnaparkhe, M. B., Maranna, S., & Satpute, G. K. (2016). QTLomics in Soybean: A Way Forward for Translational Genomics and Breeding. *Frontiers in Plant Science*, *7*, 1852.

Kweon, J., Park, S., Jeon, M. Y., Lim, K., Jang, G., Jang, A.-H., Lee, M., Seok, C., Lee, C., Park, S., Ahn, J., Jang, J., Sung, Y. H., Kim, D., & Kim, Y. (2024). Efficient DNA base editing via an optimized DYW-like deaminase. In *bioRxiv* (p. 2024.05.15.594452). https://doi.org/10.1101/2024.05.15.594452

Laforest, L. C., & Nadakuduti, S. S. (2022). Advances in Delivery Mechanisms of CRISPR Gene-Editing Reagents in Plants. *Frontiers in Genome Editing*, *4*, 830178.

Lairson, L. L., Henrissat, B., Davies, G. J., & Withers, S. G. (2008). Glycosyltransferases: structures, functions, and mechanisms. *Annual Review of Biochemistry*, *77*, 521–555.

Lanahan, M. B., Yen, H. C., Giovannoni, J. J., & Klee, H. J. (1994). The never ripe mutation blocks ethylene perception in tomato. *The Plant Cell*, *6*(4), 521–530.

Larsen, B., Gardner, K., Pedersen, C., Ørgaard, M., Migicovsky, Z., Myles, S., & Toldam-Andersen, T. B. (2018). Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing. *PloS One*, *13*(8), e0201889.

Larsen, B., Migicovsky, Z., Jeppesen, A. A., Gardner, K. M., Toldam-Andersen, T. B., Myles, S., Ørgaard, M., Petersen, M. A., & Pedersen, C. (2019). Genome-Wide Association Studies in Apple Reveal Loci for Aroma Volatiles, Sugar Composition, and Harvest Date. *The Plant Genome*, *12*.

Laurens, F., Aranzana, M. J., Arus, P., Bassi, D., Bink, M., Bonany, J., Caprera, A., Corelli-Grappadelli, L., Costes, E., Durel, C.-E., Mauroux, J.-B., Muranty, H., Nazzicari, N., Pascal, T., Patocchi, A., Peil, A., Quilot-Turion, B., Rossini, L., Stella, A., van de Weg, E. (2018). An integrated approach for increasing breeding efficiency in apple and peach in Europe. *Horticulture Research*, *5*, 11.

Leckie, K. M., Sawler, J., Kapos, P., Mackenzie, J. O., Giles, I., Baynes, K., Lo, J., Celedon, J. M., & Baute, G. J. (2023). Loss of daylength sensitivity by splice site mutation in Cannabis. In *bioRxiv* (p. 2023.03.10.532103). https://doi.org/10.1101/2023.03.10.532103

Lee, H., Dhingra, Y., & Sashital, D. G. (2019). The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *eLife*, *8.*

Lee, S. H., Moore, L. V., Park, S., Harris, D. M., & Blanck, H. M. (2022). Adults Meeting Fruit and Vegetable Intake Recommendations - United States, 2019. *MMWR. Morbidity and Mortality Weekly Report*, *71*(1), 1–9.

Leforestier, D., Ravon, E., Muranty, H., Cornille, A., Lemaire, C., Giraud, T., Durel, C.-E., & Branca, A. (2015). Genomic basis of the differences between cider and dessert apple varieties. *Evolutionary Applications*, *8*(7), 650–661.

Lemay, M.-A., de Ronne, M., Bélanger, R., & Belzile, F. (2023). k-mer-based GWAS enhances the discovery of causal variants and candidate genes in soybean. *The Plant Genome*, e20374.

Lesk, C., Rowhani, P., & Ramankutty, N. (2016). Influence of extreme weather disasters on global crop production. *Nature*, *529*(7584), 84–87.

Liang, R., He, Z., Zhao, K. T., Zhu, H., Hu, J., Liu, G., Gao, Q., Liu, M., Zhang, R., Qiu, J.-L., & Gao, C. (2024). Author Correction: Prime editing using CRISPR-Cas12a and circular RNAs in human cells. *Nature Biotechnology*.

Liang, Z., Duan, S., Sheng, J., Zhu, S., Ni, X., Shao, J., Liu, C., Nick, P., Du, F., Fan, P., Mao, R., Zhu, Y., Deng, W., Yang, M., Huang, H., Liu, Y., Ding, Y., Liu, X., Jiang, J., Dong, Y. (2019). Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses. *Nature Communications*, *10*(1), 1190.

Liao, L., Zhang, W., Zhang, B., Fang, T., Wang, X.-F., Cai, Y., Ogutu, C., Gao, L., Chen, G., Nie, X., Xu, J., Zhang, Q., Ren, Y., Yu, J., Wang, C., Deng, C. H., Ma, B., Zheng, B., You, C.-X., Han, Y. (2021). Unraveling a genetic roadmap for improved taste in the domesticated apple. *Molecular Plant*. *18*(5).

Li, C., Zhou, L., Wu, B., Li, S., Zha, W., Li, W., Zhou, Z., Yang, L., Shi, L., Lin, Y., & You, A. (2022). Improvement of Bacterial Blight Resistance in Two Conventionally Cultivated Rice Varieties by Editing the Noncoding Region. *Cells* , *11*(16).

Liebhard, R., Kellerhals, M., Pfammatter, W., Jertmini, M., & Gessler, C. (2003). Mapping quantitative physiological traits in apple (Malus x domestica Borkh.). *Plant Molecular Biology*, *52*(3), 511–526.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* , *27*(21), 2987–2993.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* , *34*(18), 3094–3100.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* , *25*(14), 1754–1760.

Li, J., Chen, L., Liang, J., Xu, R., Jiang, Y., Li, Y., Ding, J., Li, M., Qin, R., & Wei, P. (2022). Development of a highly efficient prime editor 2 system in plants. *Genome Biology*, *23*(1), 161.

Li, J. H., Findley, K., Pickrell, J. K., Blease, K., Zhao, J., & Kruglyak, S. (2023). Low-pass sequencing plus imputation using avidity sequencing displays comparable imputation accuracy to sequencing by synthesis while reducing duplicates. *G3* *86*(6).

Li, J. H., Mazur, C. A., Berisa, T., & Pickrell, J. K. (2021). Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Research*, *31*(4), 529–537.

Li, J., Wang, Z., Lubritz, D., Arango, J., Fulton, J., Settar, P., Rowland, K., Cheng, H., & Wolc, A. (2022). Genome-wide association studies for egg quality traits in White Leghorn layers using low-pass sequencing and SNP chip data. *Journal of Animal Breeding and Genetics*, *139*(4), 380–397.

Lim, E.-K., Ashford, D. A., Hou, B., Jackson, R. G., & Bowles, D. J. (2004). Arabidopsis glycosyltransferases as biocatalysts in fermentation for regioselective synthesis of diverse quercetin glucosides. *Biotechnology and Bioengineering*, *87*(5), 623–631.

Lin, D., Xiao, M., Zhao, J., Li, Z., Xing, B., Li, X., Kong, M., Li, L., Zhang, Q., Liu, Y., Chen, H., Qin, W., Wu, H., & Chen, S. (2016). An Overview of Plant Phenolic Compounds and Their Importance in Human Nutrition and Management of Type 2 Diabetes. *Molecules* , *21*(10), 1374.

Lin, Z., Ho, C.-W., & Grierson, D. (2009). AtTRP1 encodes a novel TPR protein that interacts with the ethylene receptor ERS1 and modulates development in Arabidopsis. *Journal of Experimental Botany*, *60*(13), 3697–3714.

Li, R., Fu, W., Su, R., Tian, X., Du, D., Zhao, Y., Zheng, Z., Chen, Q., Gao, S., Cai, Y., Wang, X., Li, J., & Jiang, Y. (2019). Towards the Complete Goat Pan-Genome by Recovering Missing Genomic Segments From the Reference Genome. *Frontiers in Genetics*, *10*, 1169.

Liu, X., Li, X., Wen, X., Zhang, Y., Ding, Y., Zhang, Y., Gao, B., & Zhang, D. (2021). PacBio full-length transcriptome of wild apple (*Malus sieversii*) provides insights into canker disease dynamic response. *BMC Genomics*, *22*(1), 52.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M., Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C., & Tian, Z. (2020). Pan-Genome of Wild and Cultivated Soybeans. *Cell*, *182*(1), 162–176.e13.

Liu, Z.-X., Zhang, S., Zhu, H.-Z., Chen, Z.-H., Yang, Y., Li, L.-Q., Lei, Y., Liu, Y., Li, D.-Y., Sun, A., Li, C.-P., Tan, S.-Q., Wang, G.-L., Shen, J.-Y., Jin, S., Gao, C., & Liu, J.-J. G. (2024). Hydrolytic endonucleolytic ribozyme (HYER) is programmable for sequence-specific DNA cleavage. *Science*, *383*(6682), eadh4859.

Li, X., Martín-Pizarro, C., Zhou, L., Hou, B., Wang, Y., Shen, Y., Li, B., Posé, D., & Qin, G. (2023). Deciphering the regulatory network of the NAC transcription factor FvRIF, a key regulator of strawberry (*Fragaria vesca*) fruit ripening. *The Plant Cell*, *35*(11), 4020–4045.

Li, Y.-H., Zhou, G., Ma, J., Jiang, W., Jin, L.-G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., Zhang, S.-S., Zuo, Q., Shi, X.-H., Li, Y.-F., Zhang, W.-K., Hu, Y., Kong, G., Hong, H.-L., Tan, B., Qiu, L.-J. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, *32*(10), 1045–1052.

Longhi, S., Moretto, M., Viola, R., Velasco, R., & Costa, F. (2012). Comprehensive QTL mapping survey dissects the complex fruit texture physiology in apple (*Malus x domestica Borkh*.). *Journal of Experimental Botany*, *63*(3), 1107–1121.

Love, S. L. (1999). Founding clones, major contributing ancestors, and exotic progenitors of prominent North American potato cultivars. *American Journal of Potato Research, 76*(5), 263–272.

Luby, J. J., & Shaw, D. V. (2001). Does Marker-assisted Selection Make Dollars and Sense in a Fruit Breeding Program? *HortScience*, *36*(5), 872–879.

Luck, J., Spackman, M., Freeman, A., Tre˛bicki, P., Griffiths, W., Finlay, K., & Chakraborty, S. (2011). Climate change and diseases of food crops. *Plant Pathology*, *60*(1), 113–121.

Lü, P., Yu, S., Zhu, N., Chen, Y.-R., Zhou, B., Pan, Y., Tzeng, D., Fabi, J. P., Argyris, J., Garcia-Mas, J., Ye, N., Zhang, J., Grierson, D., Xiang, J., Fei, Z., Giovannoni, J., & Zhong, S. (2018). Genome encode analyses reveal the basis of convergent evolution of fleshy fruit ripening. *Nature Plants*, *4*(10), 784–791.

Lutes, H. (2019, July 23). *The Facts on Conventional and Non-Browning Apples - CropLife.ca*. CropLife Canada. https://croplife.ca/facts-figures/apples/

Lu, Y., & Foo, L. Y. (1997). Identification and quantification of major polyphenols in apple pomace. *Food Chemistry*, *59*(2), 187–194.

Lyzenga, W. J., Pozniak, C. J., & Kagale, S. (2021). Advanced domestication: harnessing the precision of gene editing in crop breeding. *Plant Biotechnology Journal*, *19*(4), 660–670.

Magyar-Tábori, K., Dobránszki, J., Teixeira da Silva, J. A., Bulley, S. M., & Hudák, I. (2010). The role of cytokinins in shoot organogenesis in apple. *Plant Cell, Tissue and Organ Culture*, *101*(3), 251–267.

Malabarba, J., Chevreau, E., Dousset, N., Veillet, F., Moizan, J., & Vergne, E. (2020). New Strategies to Overcome Present CRISPR/Cas9 Limitations in Apple and Pear: Efficient Dechimerization and Base Editing. *International Journal of Molecular Sciences*, *22*(1).

Maliepaard, C., Alston, F. H., van Arkel, G., Brown, L. M., Chevreau, E., Dunemann, F., Evans, K. M., Gardiner, S., Guilford, P., van Heusden, A. W., Janse, J., Laurens, F., Lynn, J. R., Manganaris, A. G., den Nijs, A. P. M., Periam, N., Rikkerink, E., Roche, P., Ryder, C., King, G. J. (1998). Aligning male and female linkage maps of apple (*Malus pumila Mill.*) using multi-allelic markers. *TAG. Theoretical and Applied Genetics*, *97*(1), 60–73.

Malik, W., Ashraf, J., Iqbal, M. Z., Khan, A. A., Qayyum, A., Ali Abid, M., Noor, E., Ahmad, M. Q., & Abbasi, G. H. (2014). Molecular markers and cotton genetic improvement: current status and future prospects. *TheScientificWorldJournal*, *2014*, 607091.

Malnoy, M., Viola, R., Jung, M.-H., Koo, O.-J., Kim, S., Kim, J.-S., Velasco, R., & Nagamangala Kanchiswamy, C. (2016). DNA-Free Genetically Edited Grapevine and Apple Protoplast Using CRISPR/Cas9 Ribonucleoproteins. *Frontiers in Plant Science*, *7*, 1904.

Mansfeld, B. N., Boyher, A., Berry, J. C., Wilson, M., Ou, S., Polydore, S., Michael, T. P., Fahlgren, N., & Bart, R. S. (2021). Large structural variations in the haplotype-resolved African cassava genome. *The Plant Journal: For Cell and Molecular Biology*.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, *39*(7), 906–913.

Mardis, E. R. (2017). DNA sequencing technologies: 2006-2016. *Nature Protocols*, *12*(2), 213–218.

Marette, S., Disdier, A.-C., & Beghin, J. C. (2021). A comparison of EU and US consumers' willingness to pay for gene-edited food: Evidence from apples. *Appetite*, *159*, 105064.

Martin, A. R., Atkinson, E. G., Chapman, S. B., Stevenson, A., Stroud, R. E., Abebe, T., Akena, D., Alemayehu, M., Ashaba, F. K., Atwoli, L., Bowers, T., Chibnik, L. B., Daly, M. J., DeSmet, T., Dodge, S., Fekadu, A., Ferriera, S., Gelaye, B., Gichuru, S., NeuroGAP-Psychosis Study Team. (2021). Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *American Journal of Human Genetics*, *108*(4), 656–668.

Martín-Pizarro, C., Vallarino, J. G., Osorio, S., Meco, V., Urrutia, M., Pillet, J., Casañal, A., Merchante, C., Amaya, I., Willmitzer, L., Fernie, A. R., Giovannoni, J. J., Botella, M. A., Valpuesta, V., & Posé, D. (2021). The NAC transcription factor FaRIF controls fruit ripening in strawberry. *The Plant Cell.* 71*(1)*

Mattila, P., Hellström, J., & Törrönen, R. (2006). Phenolic acids in berries, fruits, and beverages. *Journal of Agricultural and Food Chemistry*, *54*(19), 7193–7199.

McClure, K. A., Gardner, K. M., Douglas, G. M., Song, J., Forney, C. F., DeLong, J., Fan, L., Du, L., Toivonen, P. M. A., Somers, D. J., Rajcan, I., & Myles, S. (2018). A genome-wide association study of apple quality and scab resistance. *The Plant Genome*, *11*(1), 1–14.

McClure, K. A., Gong, Y., Song, J., Vinqvist-Tymchuk, M., Campbell Palmer, L., Fan, L., Burgher-MacLellan, K., Zhang, Z., Celton, J.-M., Forney, C. F., Migicovsky, Z., & Myles, S. (2019). Genome-wide association studies in apple reveal loci of large effect controlling apple polyphenols. *Horticulture Research*, *6*, 107.

McCouch, S., Baute, G. J., Bradeen, J., Bramel, P., Bretting, P. K., Buckler, E., Burke, J. M., Charest, D., Cloutier, S., Cole, G., Dempewolf, H., Dingkuhn, M., Feuillet, C., Gepts, P., Grattapaglia, D., Guarino, L., Jackson, S., Knapp, S., Langridge, P., Zamir, D. (2013). Agriculture: Feeding the future. *Nature*, *499*(7456), 23–24.

McKenzie, F. C., & Williams, J. (2015). Sustainable food production: constraints, challenges and choices by 2050. *Food Security*, *7*(2), 221–233.

McMurchie, E. J., McGlasson, W. B., & Eaks, I. L. (1972). Treatment of fruit with propylene gives information about the biogenesis of ethylene. *Nature*, *237*(5352), 235–236.

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, *15*(6), 589–594.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, *11*(1), 31–46.

Micheletti, D., Dettori, M. T., Micali, S., Aramini, V., Pacheco, I., Da Silva Linge, C., Foschi, S., Banchi, E., Barreneche, T., Quilot-Turion, B., Lambert, P., Pascal, T., Iglesias, I., Carbó, J., Wang, L.-R., Ma, R.-J., Li, X.-W., Gao, Z.-S., Nazzicari, N., Aranzana, M. J. (2015). Whole-Genome Analysis of Diversity and SNP-Major Gene Association in Peach Germplasm. *PloS One*, *10*(9), e0136803.

Migicovsky, Z., Douglas, G. M., & Myles, S. (2022). Genotyping-by-sequencing of Canada's apple biodiversity collection. *Frontiers in Genetics*, *13*, 934712.

Migicovsky, Z., Gardner, K. M., Money, D., Sawler, J., Bloom, J. S., Moffett, P., Chao, C. T., Schwaninger, H., Fazio, G., Zhong, G.-Y., & Myles, S. (2016). Genome to Phenome Mapping in Apple Using Historical Data. *The Plant Genome*, *9*(2).

Migicovsky, Z., Gardner, K. M., Richards, C., Thomas Chao, C., Schwaninger, H. R., Fazio, G., Zhong, G.-Y., & Myles, S. (2021). Genomic consequences of apple improvement. *Horticulture Research*, *8*(1), 9.

Migicovsky, Z., & Myles, S. (2017). Exploiting Wild Relatives for Genomics-assisted Breeding of Perennial Crops. *Frontiers in Plant Science*, *8*, 460.

Migicovsky, Z., Yeats, T. H., Watts, S., Song, J., Forney, C. F., Burgher-MacLellan, K., Somers, D. J., Gong, Y., Zhang, Z., Vrebalov, J., van Velzen, R., Giovannoni, J. G., Rose, J. K. C., & Myles, S. (2021). Apple Ripening Is Controlled by a NAC Transcription Factor. *Frontiers in Genetics*, *12*, 671300.

Miura, K., Ashikari, M., & Matsuoka, M. (2011). The role of QTLs in the breeding of high-yielding rice. *Trends in Plant Science*, *16*(6), 319–326.

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C.-K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., & Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, *90*(5), 1007–1013.

Moreno-Mateos, M. A., Vejnar, C. E., Beaudoin, J.-D., Fernandez, J. P., Mis, E. K., Khokha, M. K., & Giraldez, A. J. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nature Methods*, *12*(10), 982–988.

Moya-León, M. A., Mattus-Araya, E., & Herrera, R. (2019). Molecular Events Occurring During Softening of Strawberry Fruit. *Frontiers in Plant Science*, *10*, 615.

Munir, F., Saba, N. U., Arveen, M., Siddiqa, A., Ahmad, J., & Amir, R. (2020). Chapter 14 - Pan-genomics of plants and its applications. In D. Barh, S. Soares, S. Tiwari, & V. Azevedo (Eds.), *Pan-genomics: Applications, Challenges, and Future Prospects* (pp. 285–306). Academic Press.

Murray, M. G., & Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research*, *8*(19), 4321–4325.

Myles, S., Chia, J.-M., Hurwitz, B., Simon, C., Zhong, G. Y., Buckler, E., & Ware, D. (2010). Rapid genomic characterization of the genus vitis. *PloS One*, *5*(1), e8219.

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., & Buckler, E. S. (2009). Association mapping: critical considerations shift from genotyping to experimental design. *The Plant Cell*, *21*(8), 2194–2202.

Nagarajan, N., Yapp, E. K. Y., Le, N. Q. K., Kamaraj, B., Al-Subaie, A. M., & Yeh, H.-Y. (2019). Application of Computational Biology and Artificial Intelligence Technologies in Cancer Precision Drug Discovery. *BioMed Research International*, *2019*, 8427042.

Nawaz, I., Tariq, R., Nazir, T., Khan, I., Basit, A., Gul, H., Anwar, T., Awan, S. A., Bacha, S. A. S., Zhang, L., Zhang, C., & Cong, P. (2021). RNA-Seq profiling reveals the plant hormones and molecular mechanisms stimulating the early ripening in apple. *Genomics*, *113*(1 Pt 2), 493–502.

Nekrasov, V., Wang, C., Win, J., Lanz, C., Weigel, D., & Kamoun, S. (2017). Rapid generation of a transgene-free powdery mildew resistant tomato by genome deletion. *Scientific Reports*, *7*(1), 482.

Nester, E., Gordon, M. P., & Kerr, A. (2005). *Agrobacterium tumefaciens: from plant pathology to biotechnology*. https://www.cabdirect.org/cabdirect/abstract/20053170197

Nester, E. W. (2014). Agrobacterium: nature's genetic engineer. *Frontiers in Plant Science*, *5*, 730.

Nishitani, C., Hirai, N., Komori, S., Wada, M., Okada, K., Osakabe, K., Yamamoto, T., & Osakabe, Y. (2016). Efficient Genome Editing in Apple Using a CRISPR/Cas9 system. *Scientific Reports*, *6*, 31481.

Noh, J., Do, Y. S., Kim, G. H., & Choi, C. (2020). A genome-wide association study for the detection of genes related to apple Marssonina Blotch disease resistance in apples. *Scientia Horticulturae*, *262*, 108986.

Nordborg, M., & Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics: TIG*, *18*(2), 83–90.

Nosková, A., Bhati, M., Kadri, N. K., Crysnanto, D., Neuenschwander, S., Hofer, A., & Pausch, H. (2021). Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in Swiss Large White pigs. *BMC Genomics*, *22*(1), 290.

Noyvert, B., Erzurumluoglu, A. M., Drichel, D., Omland, S., Andlauer, T. F. M., Mueller, S., Sennels, L., Becker, C., Kantorovich, A., Bartholdy, B. A., Brænne, I., Bolivar-Lopez, J. C., Mistrellides, C., Belbin, G. M., Li, J. H., Pickrell, J. K., de Jong, J., Arora, J., Hu, Y., Boehringer Ingelheim - Global Computational Biology and Digital Sciences. (2023). Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations. In *bioRxiv*. https://doi.org/10.1101/2023.12.20.23300308

Nsabiyera, V., Baranwal, D., Qureshi, N., Kay, P., Forrest, K., Valárik, M., Doležel, J., Hayden, M. J., Bariana, H. S., & Bansal, U. K. (2019). Fine Mapping of Lr49 Using 90K SNP Chip Array and Flow-Sorted Chromosome Sequencing in Wheat. *Frontiers in Plant Science*, *10*, 1787.

Nuñez-Lillo, G., Cifuentes-Esquivel, A., Troggio, M., Micheletti, D., Infante, R., Campos-Vargas, R., Orellana, A., Blanco-Herrera, F., & Meneses, C. (2015). Identification of candidate genes associated with mealiness and maturity date in peach [*Prunus persica (L.) Batsch*] using QTL analysis and deep sequencing. *Tree Genetics & Genomes*, *11*(4), 86.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53.

Nybom, H., Ahmadi-Afzadi, M., Sehic, J., & Hertog, M. (2013). DNA marker-assisted evaluation of fruit firmness at harvest and post-harvest fruit softening in a diverse apple germplasm. *Tree Genetics & Genomes*, *9*(1), 279–290.

Oeller, P. W., Lu, M. W., Taylor, L. P., Pike, D. A., & Theologis, A. (1991). Reversible inhibition of tomato fruit senescence by antisense RNA. *Science*, *254*(5030), 437–439.

Ou, L., Li, D., Lv, J., Chen, W., Zhang, Z., Li, X., Yang, B., Zhou, S., Yang, S., Li, W., Gao, H., Zeng, Q., Yu, H., Ouyang, B., Li, F., Liu, F., Zheng, J., Liu, Y., Wang, J., … Zou, X. (2018). Pan-genome of cultivated pepper (Capsicum) and its use in gene presence-absence variation analyses. *The New Phytologist*, *220*(2), 360–363.

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., & Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature*, *335*(6192), 721–726.

Peace, C. P., Bianco, L., Troggio, M., van de Weg, E., Howard, N. P., Cornille, A., Durel, C.-E., Myles, S., Migicovsky, Z., Schaffer, R. J., Costes, E., Fazio, G., Yamane, H., van Nocker, S., Gottschalk, C., Costa, F., Chagné, D., Zhang, X., Patocchi, A., … Vanderzande, S. (2019). Apple whole genome sequences: recent advances and new prospects. *Horticulture Research*, *6*, 59.

Peil, A., Dunemann, F., Richter, K., Hoefer, M., Király, I., Flachowsky, H., & Hanke, M.-V. (2008). *Resistance Breeding in Apple at Dresden-Pillnitz* (M. Boos (Ed.); pp. 220–225).

Peil, A., Garcia-Libreros, T., Richter, K., Trognitz, F. C., Trognitz, B., Hanke, M.-V., & Flachowsky, H. (2007). Strong evidence for a fire blight resistance gene of Malus robusta located on linkage group 3. *Plant Breeding*, *126*(5), 470–475.

Pereira-Lorenzo, S., Fischer, M., Ramos-Cabrer, A. M., & Castro, I. (2018). Apple (Malus spp.) Breeding: Present and Future. In J. M. Al-Khayri, S. M. Jain, & D. V. Johnson (Eds.), *Advances in Plant Breeding Strategies: Fruits: Volume 3* (pp. 3–29). Springer International Publishing.

Pfleiderer, P., Menke, I., & Schleussner, C.-F. (2019). Increasing risks of apple tree frost damage under climate change. *Climatic Change*, *157*(3), 515–525.

Phillips, R. L., Kaeppler, S. M., & Olhoft, P. (1994). Genetic instability of plant tissue cultures: breakdown of normal controls. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(12), 5222–5226.

Pimentel, D., Wilson, C., McCullum, C., Huang, R., Dwen, P., Flack, J., Tran, Q., Saltman, T., & Cliff, B. (1997). Economic and Environmental Benefits of Biodiversity. *Bioscience*, *47*(11), 747–757.

Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M. C., Zaina, G., Bastien, C., Cattonaro, F., Marroni, F., & Morgante, M. (2016). Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. *Molecular Biology and Evolution*, *33*(10), 2706–2719.

Pompili, V., Dalla Costa, L., Piazza, S., Pindo, M., & Malnoy, M. (2020). Reduced fire blight susceptibility in apple cultivars using a high-efficiency CRISPR/Cas9-FLP/FRT-based gene editing system. *Plant Biotechnology Journal*, *18*(3), 845–858.

Pray, C. E., Huang, J., Hu, R., & Rozelle, S. (2002). Five years of Bt cotton in China - the benefits continue. *The Plant Journal: 31*(4), 423–430.

Puckett, E. E., Davis, I. S., Harper, D. C., Wakamatsu, K., Battu, G., Belant, J. L., Beyer, D. E., Jr, Carpenter, C., Crupi, A. P., Davidson, M., DePerno, C. S., Forman, N., Fowler, N. L., Garshelis, D. L., Gould, N., Gunther, K., Haroldson, M., Ito, S., Kocka, D., Barsh, G. S. (2023). Genetic architecture and evolution of color variation in American black bears. *Current Biology: CB*, *33*(1), 86–97.e10.

Qiao, L., Yang, Y., Zhou, Y., Cui, H., Zhou, Y., Liu, C., Zhou, Y., Liu, H., Cheng, Z., & Pan, Y. (2023). Fine genetic mapping of the Mottled Rind Color (Morc) locus reveals a 4895-bp presence-absence variation contributing to the mottled or unmottled fruit rind color in cucumber. *Scientia Horticulturae*, *321*, 112303.

Qi, X., Dong, Y., Liu, C., Song, L., Chen, L., & Li, M. (2022). The *PavNAC56* transcription factor positively regulates fruit ripening and softening in sweet cherry (Prunus avium). *Physiologia Plantarum*, e13834.

Ranallo-Benavidez, T. R., Lemmon, Z., Soyk, S., Aganezov, S., Salerno, W. J., McCoy, R. C., Lippman, Z. B., Schatz, M. C., & Sedlazeck, F. J. (2021). Optimized sample selection for cost-efficient long-read population sequencing. *Genome Research*, *31*(5), 910–918.

Ranavat, S., Becher, H., Newman, M. F., Gowda, V., & Twyford, A. D. (2021). A Draft Genome of the Ginger Species Alpinia nigra and New Insights into the Genetic Basis of Flexistyly. *Genes*, *12*(9).

Randhawa, S., & Sengar, S. (2021). Chapter One - The evolution and history of gene editing technologies. In D. Ghosh (Ed.), *Progress in Molecular Biology and Translational Science* (Vol. 178, pp. 1–62). Academic Press.

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.2) [Darwin17.0].

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., Kresovich, S., Goodman, M. M., & Buckler, E. S., 4th. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11479–11484.

Ren, Q., Sretenovic, S., Liu, S., Tang, X., Huang, L., He, Y., Liu, L., Guo, Y., Zhong, Z., Liu, G., Cheng, Y., Zheng, X., Pan, C., Yin, D., Zhang, Y., Li, W., Qi, L., Li, C., Qi, Y., & Zhang, Y. (2021). PAM-less plant genome editing using a CRISPR–SpRY toolbox. *Nature Plants*, *7*(1), 25–33.

Ren, S., Lyu, G., Irwin, D. M., Liu, X., Feng, C., Luo, R., Zhang, J., Sun, Y., Shang, S., Zhang, S., & Wang, Z. (2021). Pooled Sequencing Analysis of Geese (Anser cygnoides) Reveals Genomic Variations Associated With Feather Color. *Frontiers in Genetics*, *12*, 650013.

Riederer, M., & Schreiber, L. (2001). Protecting against water loss: analysis of the barrier properties of plant cuticles. *Journal of Experimental Botany*, *52*(363), 2023–2032.

Ries, D., Holtgräwe, D., Viehöver, P., & Weisshaar, B. (2016). Rapid gene identification in sugar beet using deep sequencing of DNA from phenotypic pools selected from breeding panels. *BMC Genomics*, *17*, 236.

Riva, G., González-Cabrera, J., Vazquez-Padrón, R., & Ayra-Pardo, C. (1998). Agrobacterium tumefaciens: a natural tool for plant transformation. *Electronic Journal of Biotechnology*, *1*, 24–25.

Robinson, J. P., Harris, S. A., & Juniper, B. E. (2001). Taxonomy of the genus Malus Mill. (Rosaceae) with emphasis on the cultivated apple, Malus domestica Borkh. *Plant Systematics and Evolution*, *226*(1), 35–58.

Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E., & Lippman, Z. B. (2017). Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. *Cell*, *171*(2), 470–480.e8.

Rupasinghe, H. P. V., Balasuriya, N., & Wang, Y. (2017). Prevention of Type 2 Diabetes by Polyphenols of Fruits. In K. H. Al-Gubory & I. Laher (Eds.), *Nutritional Antioxidant Therapies: Treatments and Perspectives* (pp. 447–466). Springer International Publishing.

Ru, S., Main, D., Evans, K., & Peace, C. (2015). Current applications, challenges, and perspectives of marker-assisted seedling selection in Rosaceae tree fruit breeding. *Tree Genetics & Genomes*, *11*(1), 8.

Samuels, L., Kunst, L., & Jetter, R. (2008). Sealing plant surfaces: cuticular wax formation by epidermal cells. *Annual Review of Plant Biology*, *59*, 683–707.

Sánchez-Sevilla, J. F., Vallarino, J. G., Osorio, S., Bombarely, A., Posé, D., Merchante, C., Botella, M. A., Amaya, I., & Valpuesta, V. (2017). Gene expression atlas of fruit ripening and transcriptome assembly from RNA-seq data in octoploid strawberry (Fragaria × ananassa). *Scientific Reports*, *7*(1), 13737.

Sander, J. D., & Joung, J. K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology*, *32*(4), 347–355.

Sasaki, E., Köcher, T., Filiault, D. L., & Nordborg, M. (2021). Revisiting a GWAS peak in Arabidopsis thaliana reveals possible confounding by genetic heterogeneity. *Heredity*, *127*(3), 245–252.

Saxena, R. K., Edwards, D., & Varshney, R. K. (2014). Structural variations in plant genomes. *Briefings in Functional Genomics*, *13*(4), 296–307.

Sax, K. (1923). The Association of Size Differences with Seed-Coat Pattern and Pigmentation in *Phaseolus vulgaris*. *Genetics*, *8*(6), 552–560.

SB Gelvin. (2003). Agrobacterium-Mediated Plant Transformation: the Biology behind the “Gene-Jockeying” Tool. *Microbiology and Molecular Biology Reviews: MMBR*, *67*(1), 16–37.

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews. Genetics*, *19*(8), 491–504.

Schapire, A. L., Valpuesta, V., & Botella, M. A. (2006). TPR Proteins in Plant Hormone Signaling. *Plant Signaling & Behavior*, *1*(5), 229–230.

Schatz, M. C., Maron, L. G., Stein, J. C., Hernandez Wences, A., Gurtowski, J., Biggers, E., Lee, H., Kramer, M., Antoniou, E., Ghiban, E., Wright, M. H., Chia, J.-M., Ware, D., McCouch, S. R., & McCombie, W. R. (2014). Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. *Genome Biology*, *15*(11), 506.

Schlathölter, I., Broggini, G. A. L., Streb, S., Studer, B., & Patocchi, A. (2023). Field study of the fire-blight-resistant cisgenic apple line C44.4.146. *The Plant Journal: For Cell and Molecular Biology*, *113*(6), 1160–1175.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468.

Seiler, G. J., Qi, L. L., & Marek, L. F. (2017). Utilization of sunflower crop wild relatives for cultivated sunflower improvement. *Crop Science*, *57*(3), 1083–1101.

Servin, B., & Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics*, *3*(7), e114.

Seymour, G. B., Østergaard, L., Chapman, N. H., Knapp, S., & Martin, C. (2013). Fruit development and ripening. *Annual Review of Plant Biology*, *64*, 219–241.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, *550*(7676), 345–353.

Shetty, K., & Wahlqvist, M. (2004). A model for the role of the proline-linked pentose-phosphate pathway in phenolic phytochemical bio-synthesis and mechanism of action for human health and environmental applications. *Asia Pacific Journal of Clinical Nutrition*, *13*(1).

Shin, J.-H., Blay, S., McNeney, B., & Graham, J. (2006). LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. *Journal of Statistical Software*, *16*, 1–9.

Shi, X., Cao, S., Wang, X., Huang, S., Wang, Y., Liu, Z., Liu, W., Leng, X., Peng, Y., Wang, N., Wang, Y., Ma, Z., Xu, X., Zhang, F., Xue, H., Zhong, H., Wang, Y., Zhang, K., Velt, A., Zhou, Y. (2023). The complete reference genome for grapevine (Vitis vinifera L.) genetics and breeding. *Horticulture Research*, *10*(5), uhad061.

Single, R. M., & Thomson, G. (2016). Linkage Disequilibrium: Population Genetics of Multiple Loci. In R. M. Kliman (Ed.), *Encyclopedia of Evolutionary Biology* (pp. 400–404). Academic Press.

Skreli, E., & Imami, D. (2012). Analyzing consumers' preferences for apple attributes in Tirana, Albania. *International Food and Agribusiness Management Review*, *15*(1030-2016-82810), 137–157.

Smanalieva, J., Iskakova, J., Oskonbaeva, Z., Wichern, F., & Darr, D. (2020). Investigation of nutritional characteristics and free radical scavenging activity of wild apple, pear, rosehip, and barberry from the walnut-fruit forests of Kyrgyzstan. *European Food Research and Technology. 246*(5), 1095–1104.

Snelling, W. M., Hoff, J. L., Li, J. H., Kuehn, L. A., Keel, B. N., Lindholm-Perry, A. K., & Pickrell, J. K. (2020). Assessment of Imputation from Low-Pass Sequencing to Predict Merit of Beef Steers. *Genes*, *11*(11).

Soares, S., Kohl, S., Thalmann, S., Mateus, N., Meyerhof, W., & De Freitas, V. (2013). Different phenolic compounds activate distinct human bitter taste receptors. *Journal of Agricultural and Food Chemistry*, *61*(7), 1525–1533.

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S., Zhou, R., Xie, W.-Z., Cheng, Y., Zhang, Y., Liu, K., Yang, Q.-Y., Chen, L.-L., & Guo, L. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants*, *6*(1), 34–45.

Song, X., Meng, X., Guo, H., Cheng, Q., Jing, Y., Chen, M., Liu, G., Wang, B., Wang, Y., Li, J., & Yu, H. (2022). Targeting a gene regulatory element enhances rice grain yield by decoupling panicle number and size. *Nature Biotechnology*, *40*(9), 1403–1411.

Sorkheh, K., Malysheva-Otto, L. V., Wirthensohn, M. G., Tarkesh-Esfahani, S., & Martínez-Gómez, P. (2008). Linkage disequilibrium, genetic association mapping and gene localization in crop plants. *Genetics and Molecular Biology*, *31*(4), 805–814.

Spengler, R. N. (2019). Origins of the Apple: The Role of Megafaunal Mutualism in the Domestication of Malus and Rosaceous Trees. *Frontiers in Plant Science*, *10*, 617.

Spitzer, K., Pelizzola, M., & Futschik, A. (2019). Modifying the Chi-square and the CMH test for population genetic inference: adapting to over-dispersion. In *arXiv* http://arxiv.org/abs/1902.08127

Srivastava, A. K., Lu, Y., Zinta, G., Lang, Z., & Zhu, J.-K. (2018). UTR-Dependent Control of Gene Expression in Plants. *Trends in Plant Science*, *23*(3), 248–259.

Stowe, E., & Dhingra, A. (2020). Development of the Arctic® Apple. *Plant Breeding Reviews*, *44*, 273–296.

Sun, C., Lei, Y., Li, B., Gao, Q., Li, Y., Cao, W., Yang, C., Li, H., Wang, Z., Li, Y., Wang, Y., Liu, J., Zhao, K. T., & Gao, C. (2023). Precise integration of large DNA sequences in plant genomes using PrimeRoot editors. *Nature Biotechnology* *18*(7).

Sun, X., Jiao, C., Schwaninger, H., Chao, C. T., Ma, Y., Duan, N., Khan, A., Ban, S., Xu, K., Cheng, L., Zhong, G.-Y., & Fei, Z. (2020). Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nature Genetics*, *52*(12), 1423–1432.

Sun, X., Zhu, S., Li, N., Cheng, Y., Zhao, J., Qiao, X., Lu, L., Liu, S., Wang, Y., Liu, C., Li, B., Guo, W., Gao, S., Yang, Z., Li, F., Zeng, Z., Tang, Q., Pan, Y., Guan, M., Liu, T. (2020). A Chromosome-Level Genome Assembly of Garlic (*Allium sativum*) Provides Insights into Genome Evolution and Allicin Biosynthesis. *Molecular Plant*, *13*(9), 1328–1339.

Sun, Y., Shang, L., Zhu, Q.-H., Fan, L., & Guo, L. (2022). Twenty years of plant genome sequencing: achievements and challenges. *Trends in Plant Science*, *27*(4), 391–401.

Sun, Y., Wang, J., Li, Y., Jiang, B., Wang, X., Xu, W.-H., Wang, Y.-Q., Zhang, P.-T., Zhang, Y.-J., & Kong, X.-D. (2022). Pan-Genome Analysis Reveals the Abundant Gene Presence/Absence Variations Among Different Varieties of Melon and Their Influence on Traits. *Frontiers in Plant Science*, *13*, 835496.

Svitashev, S., Young, J. K., Schwartz, C., Gao, H., Falco, S. C., & Cigan, A. M. (2015). Targeted Mutagenesis, Precise Gene Editing, and Site-Specific Gene Insertion in Maize Using Cas9 and Guide RNA. *Plant Physiology*, *169*(2), 931–945.

Tadiello, A., Longhi, S., Moretto, M., Ferrarini, A., Tononi, P., Farneti, B., Busatto, N., Vrhovsek, U., Molin, A. D., Avanzato, C., Biasioli, F., Cappellin, L., Scholz, M., Velasco, R., Trainotti, L., Delledonne, M., & Costa, F. (2016). Interference with ethylene perception at receptor level sheds light on auxin and transcriptional circuits associated with the climacteric ripening of apple fruit (*Malus x domestica Borkh.*). *The Plant Journal: For Cell and Molecular Biology*, *88*(6), 963–975.

Takos, A. M., Jaffé, F. W., Jacob, S. R., Bogs, J., Robinson, S. P., & Walker, A. R. (2006). Light-induced expression of a MYB gene regulates anthocyanin biosynthesis in red apples. *Plant Physiology*, *142*(3), 1216–1232.

Tanner, G. J., Francki, K. T., Abrahams, S., Watson, J. M., Larkin, P. J., & Ashton, A. R. (2003). Proanthocyanidin biosynthesis in plants. Purification of legume leucoanthocyanidin reductase and molecular cloning of its cDNA. *The Journal of Biological Chemistry*, *278*(34), 31647–31656.

Tan, Q., Li, S., Zhang, Y., Chen, M., Wen, B., Jiang, S., Chen, X., Fu, X., Li, D., Wu, H., Wang, Y., Xiao, W., & Li, L. (2021). Chromosome-level genome assemblies of five Prunus species and genome-wide association studies for key agronomic traits in peach. *Horticulture Research*, *8*(1), 213.

Taus, T., Futschik, A., & Schlötterer, C. (2017). Quantifying Selection with Pool-Seq Time Series Data. *Molecular Biology and Evolution*, *34*(11), 3023–3034.

Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., & Gaut, B. S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). *Proceedings of the National Academy of Sciences of the United States of America*, *98*(16), 9161–9166.

Tian, S., Jiang, L., Gao, Q., Zhang, J., Zong, M., Zhang, H., Ren, Y., Guo, S., Gong, G., Liu, F., & Xu, Y. (2017). Efficient CRISPR/Cas9-based gene knockout in watermelon. *Plant Cell Reports*, *36*(3), 399–406.

Tian, X., Li, R., Fu, W., Li, Y., Wang, X., Li, M., Du, D., Tang, Q., Cai, Y., Long, Y., Zhao, Y., Li, M., & Jiang, Y. (2020). Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. *Science China. Life Sciences*, *63*(5), 750–763.

Tian, Y., Thrimawithana, A., Ding, T., Guo, J., Gleave, A., Chagné, D., Ampomah-Dwamena, C., Ireland, H. S., Schaffer, R. J., Luo, Z., Wang, M., An, X., Wang, D., Gao, Y., Wang, K., Zhang, H., Zhang, R., Zhou, Z., Yan, Z., Yao, J.-L. (2022). Transposon insertions regulate genome-wide allele-specific expression and underpin flower colour variations in apple (Malus spp.). *Plant Biotechnology Journal*. *81*(1).

Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *The Plant Genome*, *14*(1), e20077.

Toivonen, P. M. A. (2006). Fresh-cut apples: Challenges and opportunities for multi-disciplinary research. *Canadian Journal of Plant Science. 86* (Special Issue), 1361–1368.

Tripodi, P., Rabanus-Wallace, M. T., Barchi, L., Kale, S., Esposito, S., Acquadro, A., Schafleitner, R., van Zonneveld, M., Prohens, J., Diez, M. J., Börner, A., Salinier, J., Caromel, B., Bovy, A., Boyaci, F., Pasev, G., Brandt, R., Himmelbach, A., Portis, E., Stein, N. (2021). Global range expansion history of pepper (Capsicum spp.) revealed by over 10,000 genebank accessions. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(34).

Troggio, M., Gleave, A., Salvi, S., Chagné, D., Cestaro, A., Kumar, S., Crowhurst, R. N., & Gardiner, S. E. (2012). Apple, from genome to breeding. *Tree Genetics & Genomes*, *8*(3), 509–529.

Trouern-Trend, A. J., Falk, T., Zaman, S., Caballero, M., Neale, D. B., Langley, C. H., Dandekar, A. M., Stevens, K. A., & Wegrzyn, J. L. (2020). Comparative genomics of six Juglans species reveals disease-associated gene family contractions. *The Plant Journal: For Cell and Molecular Biology*, *102*(2), 410–423.

Turnbull, C., Lillemo, M., & Hvoslef-Eide, T. A. K. (2021). Global regulation of genetically modified crops amid the gene edited crop boom - A review. *Frontiers in Plant Science*, *12*, 630396.

Urrestarazu, J., Muranty, H., Denancé, C., Leforestier, D., Ravon, E., Guyader, A., Guisnel, R., Feugey, L., Aubourg, S., Celton, J.-M., Daccord, N., Dondini, L., Gregori, R., Lateur, M., Houben, P., Ordidge, M., Paprstein, F., Sedlak, J., Nybom, H., Durel, C.-E. (2017). Genome-Wide Association Mapping of Flowering and Ripening Periods in Apple. *Frontiers in Plant Science*, *8*, 1923.

Vallebona, C., Mantino, A., & Bonari, E. (2016). Exploring the potential of perennial crops in reducing soil erosion: A GIS-based scenario analysis in southern Tuscany, Italy. *Applied Geography* , *66*, 119–131.

Van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. "O'Reilly Media, Inc."

Van Vu, T., Sivankalyani, V., Kim, E.-J., Doan, D. T. H., Tran, M. T., Kim, J., Sung, Y. W., Park, M., Kang, Y. J., & Kim, J.-Y. (2020). Highly efficient homology-directed repair using CRISPR/Cpf1-geminiviral replicon in tomato. *Plant Biotechnology Journal*. *43*(2).

Veillet, F., Perrot, L., Chauvin, L., Kermarrec, M.-P., Guyon-Debast, A., Chauvin, J.-E., Nogué, F., & Mazier, M. (2019). Transgene-Free Genome Editing in Tomato and Potato Plants Using Agrobacterium-Mediated Delivery of a CRISPR/Cas9 Cytidine Base Editor. *International Journal of Molecular Sciences*, *20*(2).

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S. K., Troggio, M., Pruss, D., & Others. (2010). The genome of the domesticated apple (Malus× domestica Borkh.). *Nature Genetics*, *42*(10), 833–839.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S. K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., Viola, R. (2010). The genome of the domesticated apple (*Malus × domestica Borkh.*). *Nature Genetics*, *42*(10), 833–839.

Venezia, M., & Creasey Krainer, K. M. (2021). Current Advancements and Limitations of Gene Editing in Orphan Crops. *Frontiers in Plant Science*, *12*, 742932.

Vilhjálmsson, B. J., & Nordborg, M. (2013). The nature of confounding in genome-wide association studies. *Nature Reviews. Genetics*, *14*(1), 1–2.

Vinson, J. A., Su, X., Zubik, L., & Bose, P. (2001). Phenol antioxidant quantity and quality in foods: fruits. *Journal of Agricultural and Food Chemistry*, *49*(11), 5315–5321.

Voichek, Y., & Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nature Genetics*, *52*(5), 534–540.

Volk, G. M., Chao, C. T., Norelli, J., Brown, S. K., Fazio, G., Peace, C., McFerson, J., Zhong, G.-Y., & Bretting, P. (2015). The vulnerability of US apple (Malus) genetic resources. *Genetic Resources and Crop Evolution*, *62*(5), 765–794.

Volk, G. M., Henk, A. D., Richards, C. M., Forsline, P. L., & Thomas Chao, C. (2013). Malus sieversii: A Diverse Central Asian Apple Species in the USDA-ARS National Plant Germplasm System. *HortScience: A Publication of the American Society for Horticultural Science*, *48*(12), 1440–1444.

Volz, R. K., & McGhie, T. K. (2011). Genetic variability in apple fruit polyphenol composition in Malus × domestica and Malus sieversii germplasm grown in New Zealand. *Journal of Agricultural and Food Chemistry*, *59*(21), 11509–11521.

Voytas, D. (2019). Gene editing offers dietary benefits. *Gene*, *33*(4).

Waltz, E. (2016). CRISPR-edited crops free to enter market, skip regulation. *Nature Biotechnology*, *34*(6), 582.

Wang, F., Wang, C., Liu, P., Lei, C., Hao, W., Gao, Y., Liu, Y.-G., & Zhao, K. (2016). Enhanced Rice Blast Resistance by CRISPR/Cas9-Targeted Mutagenesis of the ERF Transcription Factor Gene OsERF922. *PloS One*, *11*(4), e0154027.

Wang, N., Jiang, S., Zhang, Z., Fang, H., Xu, H., Wang, Y., & Chen, X. (2018). Malus sieversii: the origin, flavonoid synthesis mechanism, and breeding of red-skinned and red-fleshed apples. *Horticulture Research*, *5*, 70.

Wang, R., Lammers, M., Tikunov, Y., Bovy, A. G., Angenent, G. C., & de Maagd, R. A. (2020). The rin, nor and Cnr spontaneous mutations inhibit tomato fruit ripening in additive and epistatic manners. *Plant Science: An International Journal of Experimental Plant Biology*, *294*, 110436.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R. R., Zhang, F., Mansueto, L., Copetti, D., Sanciangco, M., Palis, K. C., Xu, J., Sun, C., Fu, B., Zhang, H., Gao, Y., Leung, H. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, *557*(7703), 43–49.

Wang, Y., Cheng, X., Shan, Q., Zhang, Y., Liu, J., Gao, C., & Qiu, J.-L. (2014). Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nature Biotechnology*, *32*(9), 947–951.

Wang, Y., Zhu, Y., Jiang, H., Mao, Z., Zhang, J., Fang, H., Liu, W., Zhang, Z., Chen, X., & Wang, N. (2023). The regulatory module MdBZR1-MdCOL6 mediates brassinosteroid- and light-regulated anthocyanin synthesis in apple. *The New Phytologist*, *238*(4), 1516–1533.

WAPA. (2018, December). *U.S. Apple and Pear Forecast*. WAPA - The World Apple and Pear Association. http://www.wapa-association.org/asp/page_1.asp?doc_id=447

Watson, A. E., Guitton, B., Soriano, A., Rivallan, R., Vignes, H., Farrera, I., Huettel, B., Arnaiz, C., Falavigna, V. da S., Coupel-Ledru, A., Segura, V., Sarah, G., Dufayard, J.-F., Sidibe-Bocs, S., Costes, E., & Andrés, F. (2024). Target enrichment sequencing coupled with GWAS identifies MdPRX10 as a candidate gene in the control of budbreak in apple. *Frontiers in Plant Science*, *15*, 1352757.

Watts, S., Migicovsky, Z., McClure, K. A., Yu, C. H. J., Amyotte, B., Baker, T., Bowlby, D., Burgher-MacLellan, K., Butler, L., Donald, R., Fan, L., Fillmore, S., Flewelling, J., Gardner, K., Hodges, M., Hughes, T., Jagadeesan, V., Lewis, N., MacDonell, E., Myles, S. (2021). Quantifying apple diversity: A phenomic characterization of Canada’s Apple Biodiversity Collection. *PLANTS, PEOPLE, PLANET*, *3*(6), 747–760.

Watts, S., Migicovsky, Z., & Myles, S. (2023). Large-scale apple GWAS reveals *NAC18.1* as a master regulator of ripening traits. *Fruit Research*, *3*(1), 0–0.

Webb, D. M., & Knapp, S. J. (1990). DNA extraction from a previously recalcitrant plant genus. *Plant Molecular Biology Reporter / ISPMB*, *8*(3), 180–185.

Weber, C., Simnitt, S., Lucier, G., & Davis, W. V. (2023, March 30). *Fruit and Tree Nuts*. https://www.ers.usda.gov/webdocs/outlooks/106240/fts-376.pdf?v=2488.5

Weigel, D., & Nordborg, M. (2005). Natural variation in Arabidopsis. How do we find the causal genes? *Plant Physiology*, *138*(2), 567–568.

Welling, M. T., Liu, L., Kretzschmar, T., Mauleon, R., Ansari, O., & King, G. J. (2020). An extreme-phenotype genome-wide association study identifies candidate cannabinoid pathway genes in Cannabis. *Scientific Reports*, *10*(1), 18643.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*.

Wijesundara, N. M., Lee, S. F., Cheng, Z., Davidson, R., & Rupasinghe, H. P. V. (2021). Carvacrol exhibits rapid bactericidal activity against Streptococcus pyogenes through cell membrane damage. *Scientific Reports*, *11*(1), 1487.

Wilson, J. P., Hess, D. E., & Hanna, W. W. (2000). Resistance to Striga hermonthica in Wild Accessions of the Primary Gene Pool of Pennisetum glaucum. *Phytopathology*, *90*(10), 1169–1172.

Wisser, R. J., Balint-Kurti, P. J., & Nelson, R. J. (2006). The genetic architecture of disease resistance in maize: a synthesis of published studies. *Phytopathology*, *96*(2), 120–129.

Wong, K. H., Jin, Y., & Moqtaderi, Z. (2013). Multiplex Illumina sequencing using DNA barcoding. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, *Chapter 7*, Unit 7.11.

Woodhouse, M. R., Cannon, E. K., Portwood, J. L., 2nd, Harper, L. C., Gardiner, J. M., Schaeffer, M. L., & Andorf, C. M. (2021). A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biology*, *21*(1), 385.

Wu, B., Shen, F., Chen, C. J., Liu, L., Wang, X., Zheng, W. Y., Deng, Y., Wang, T., Huang, Z. Y., Xiao, C., Zhou, Q., Wang, Y., Wu, T., Xu, X. F., Han, Z. H., & Zhang, X. Z. (2021). Natural variations in a pectin acetylesterase gene, MdPAE10, contribute to prolonged apple fruit shelf life. *The Plant Genome*, *14*(1), e20084.

Wu, B., Shen, F., Wang, X., Zheng, W. Y., Xiao, C., Deng, Y., Wang, T., Yu Huang, Z., Zhou, Q., Wang, Y., Wu, T., Feng Xu, X., Hai Han, Z., & Zhong Zhang, X. (2020). Role of MdERF3 and MdERF118 natural variations in apple flesh firmness/crispness retainability and development of QTL-based genomics-assisted prediction. *Plant Biotechnology Journal*. *52*(7).

Wunderlich, S., & Gatto, K. A. (2015). Consumer perception of genetically modified organisms and sources of information. *Advances in Nutrition* , *6*(6), 842–851.

Wu, Y., & Kelly, R. M. (2021). Online Dating Meets Artificial Intelligence: How the Perception of Algorithmically Generated Profile Text Impacts Attractiveness and Trust. *Proceedings of the 32nd Australian Conference on Human-Computer Interaction*, 444–453.

Xu, J., Hua, K., & Lang, Z. (2019). Genome editing for horticultural crop improvement. *Horticulture Research*, *6*, 113.

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., Heuer, S., Ismail, A. M., Bailey-Serres, J., Ronald, P. C., & Mackill, D. J. (2006). Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature*, *442*(7103), 705–708.

Yamazaki, M., Shimada, T., Takahashi, H., Tamura, K., Kondo, M., Nishimura, M., & Hara-Nishimura, I. (2008). Arabidopsis VPS35, a retromer component, is required for vacuolar protein sorting and involved in plant growth and leaf senescence. *Plant & Cell Physiology*, *49*(2), 142–156.

Yang, S. F., & Hoffman, N. E. (1984). Ethylene Biosynthesis and its Regulation in Higher Plants. *Annual Review of Plant Physiology*, *35*(1), 155–189.

Yang, X., Wu, B., Liu, J., Zhang, Z., Wang, X., Zhang, H., Ren, X., Zhang, X., Wang, Y., Wu, T., Xu, X., Han, Z., & Zhang, X. (2022). A single QTL harboring multiple genetic variations leads to complicated phenotypic segregation in apple flesh firmness and crispness. *Plant Cell Reports*. *18*(4).

Yan, M.-Y., Yan, H.-Q., Ren, G.-X., Zhao, J.-P., Guo, X.-P., & Sun, Y.-C. (2017). CRISPR-Cas12a-Assisted Recombineering in Bacteria. *Applied and Environmental Microbiology*, *83*(17).

Yen, H. C., Lee, S., Tanksley, S. D., Lanahan, M. B., Klee, H. J., & Giovannoni, J. J. (1995). The tomato Never-ripe locus regulates ethylene-inducible gene expression and is linked to a homolog of the Arabidopsis ETR1 gene. *Plant Physiology*, *107*(4), 1343–1353.

Ye, X., Al-Babili, S., Klöti, A., Zhang, J., Lucca, P., Beyer, P., & Potrykus, I. (2000). Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science*, *287*(5451), 303–305.

Yue, J., Chen, Q., Wang, Y., Zhang, L., Ye, C., Wang, X., Cao, S., Lin, Y., Huang, W., Xian, H., Qin, H., Wang, Y., Zhang, S., Wu, Y., Wang, S., Yue, Y., & Liu, Y. (2023). Telomere-to-telomere and gap-free reference genome assembly of the kiwifruit Actinidia chinensis. *Horticulture Research*, *10*(2), uhac264.

Yu, J., Golicz, A. A., Lu, K., Dossa, K., Zhang, Y., Chen, J., Wang, L., You, J., Fan, D., Edwards, D., & Zhang, X. (2019). Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnology Journal*, *17*(5), 881–892.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208.

Zaenen, I., Van Larebeke, N., Van Montagu, M., & Schell, J. (1974). Supercoiled circular DNA in crown-gall inducing Agrobacterium strains. *Journal of Molecular Biology*, *86*(1), 109–127.

Zambryski, P., Holsters, M., Kruger, K., Depicker, A., Schell, J., Van Montagu, M., & Goodman, H. M. (1980). Tumor DNA structure in plant cells transformed by A. tumefaciens. *Science*, *209*(4463), 1385–1391.

Zan, Y., & Carlborg, Ö. (2019). A Polygenic Genetic Architecture of Flowering Time in the Worldwide Arabidopsis thaliana Population. *Molecular Biology and Evolution*, *36*(1), 141–154.

Zhang, A., Zhou, H., Jiang, X., Han, Y., & Zhang, X. (2021). The Draft Genome of a Flat Peach (Prunus persica L. cv. “124 Pan”) Provides Insights into Its Good Fruit Flavor Traits. *Plants*, *10*(3).

Zhang, D., Wang, Z., Wang, N., Gao, Y., Liu, Y., Wu, Y., Bai, Y., Zhang, Z., Lin, X., Dong, Y., Ou, X., Xu, C., & Liu, B. (2014). Tissue culture-induced heritable genomic variation in rice, and their phenotypic implications. *PloS One*, *9*(5), e96879.'

Zhang, G., Li, T., Zhang, L., Dong, W., & Wang, A. (2018). Expression analysis of NAC genes during the growth and ripening of apples. *Zahradnictvi (Prague, Czech Republic: 1992)*, *45*(1), 1–10.

Zhang, H., Wang, Y., Deng, C., Zhao, S., Zhang, P., Feng, J., Huang, W., Kang, S., Qian, Q., Xiong, G., & Chang, Y. (2022). High-quality genome assembly of Huazhan and Tianfeng, the parents of an elite rice hybrid Tian-you-hua-zhan. *Science China. Life Sciences*, *65*(2), 398–411.

Zhang, H., Zhang, J., Wei, P., Zhang, B., Gou, F., Feng, Z., Mao, Y., Yang, L., Zhang, H., Xu, N., & Zhu, J.-K. (2014). The CRISPR/Cas9 system produces specific and homozygous targeted gene editing in rice in one generation. *Plant Biotechnology Journal*, *12*(6), 797–807.

Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C. M., Zhang, C., Tian, Y., Liu, G., Gul, H., Wang, D., Tian, Y., Yang, C., Meng, M., Yuan, G., Kang, G., Wu, Y., Wang, K., Zhang, H., Cong, P. (2019). A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nature Communications*, *10*(1), 1494.

Zhang, M.-Y., Xue, C., Hu, H., Li, J., Xue, Y., Wang, R., Fan, J., Zou, C., Tao, S., Qin, M., Bai, B., Li, X., Gu, C., Wu, S., Chen, X., Yang, G., Liu, Y., Sun, M., Fei, Z., Wu, J. (2021). Genome-wide association studies provide insights into the genetic determination of fruit traits of pear. *Nature Communications*, *12*(1), 1144.

Zhang, X., Gong, X., Cheng, S., Yu, H., Li, D., Su, X., Lei, Z., Li, M., & Ma, F. (2022). Proline-rich protein MdPRP6 alters low nitrogen stress tolerance by regulating lateral root formation and anthocyanin accumulation in transgenic apple (Malus domestica). *Environmental and Experimental Botany*, *197*, 104841.

Zhang, Y., Li, D., Zhang, D., Zhao, X., Cao, X., Dong, L., Liu, J., Chen, K., Zhang, H., Gao, C., & Wang, D. (2018). Analysis of the functions of TaGW2 homoeologs in wheat grain weight and protein content traits. *The Plant Journal: For Cell and Molecular Biology*, *94*(5), 857–866.

Zhao, J., Bayer, P. E., Ruperao, P., Saxena, R. K., Khan, A. W., Golicz, A. A., Nguyen, H. T., Batley, J., Edwards, D., & Varshney, R. K. (2020). Trait associations in the pangenome of pigeon pea (Cajanus cajan). *Plant Biotechnology Journal*, *18*(9), 1946–1954.

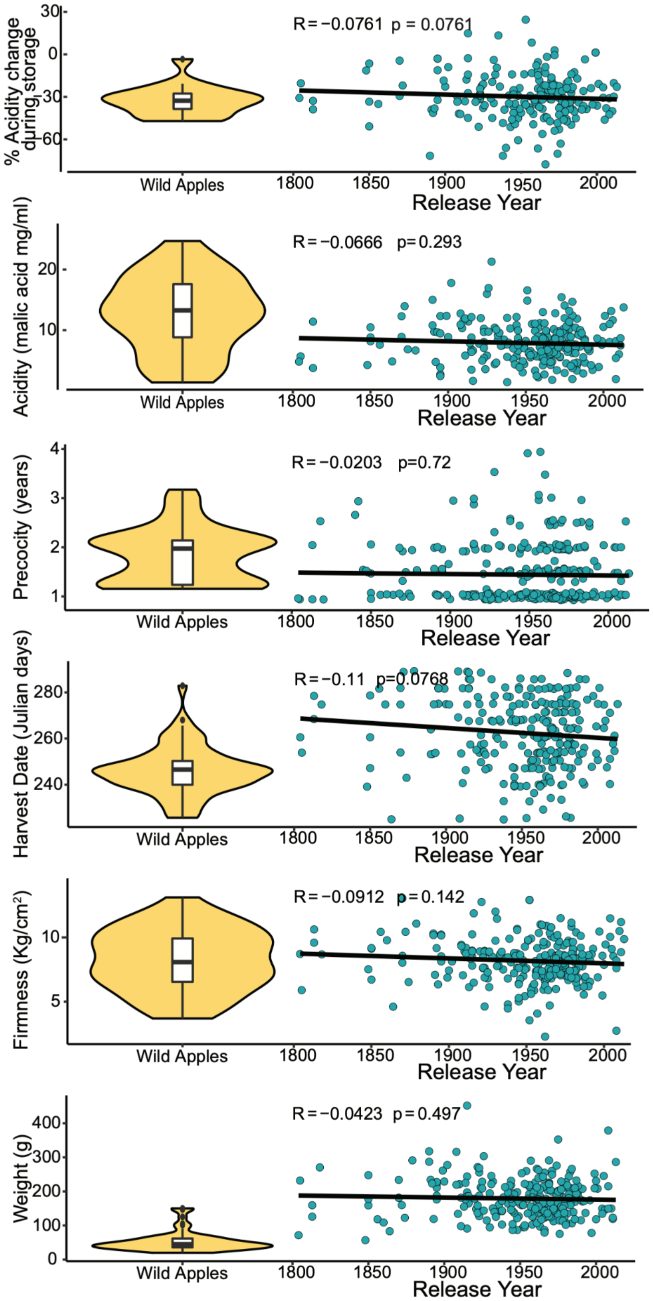
Zheng, M., Zhang, L., Tang, M., Liu, J., Liu, H., Yang, H., Fan, S., Terzaghi, W., Wang, H., & Hua, W. (2020). Knockout of two BnaMAX1 homologs by CRISPR/Cas9-targeted mutagenesis improves plant architecture and increases yield in rapeseed (Brassica napus L.). *Plant Biotechnology Journal*, *18*(3), 644–654.

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824.

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.

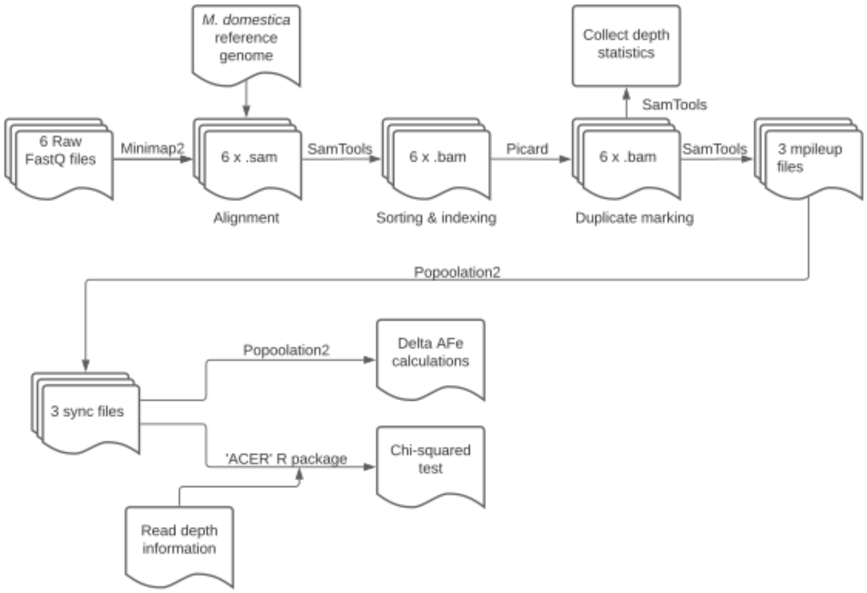
Zong, Y., Liu, Y., Xue, C., Li, B., Li, X., Wang, Y., Li, J., Liu, G., Huang, X., Cao, X., & Gao, C. (2022). An engineered prime editor with enhanced editing efficiency in plants. *Nature Biotechnology*, *40*(9), 1394–1402.

Appendix I: Phenotypic divergence between the wild and cultivated apple (Chapter 2)

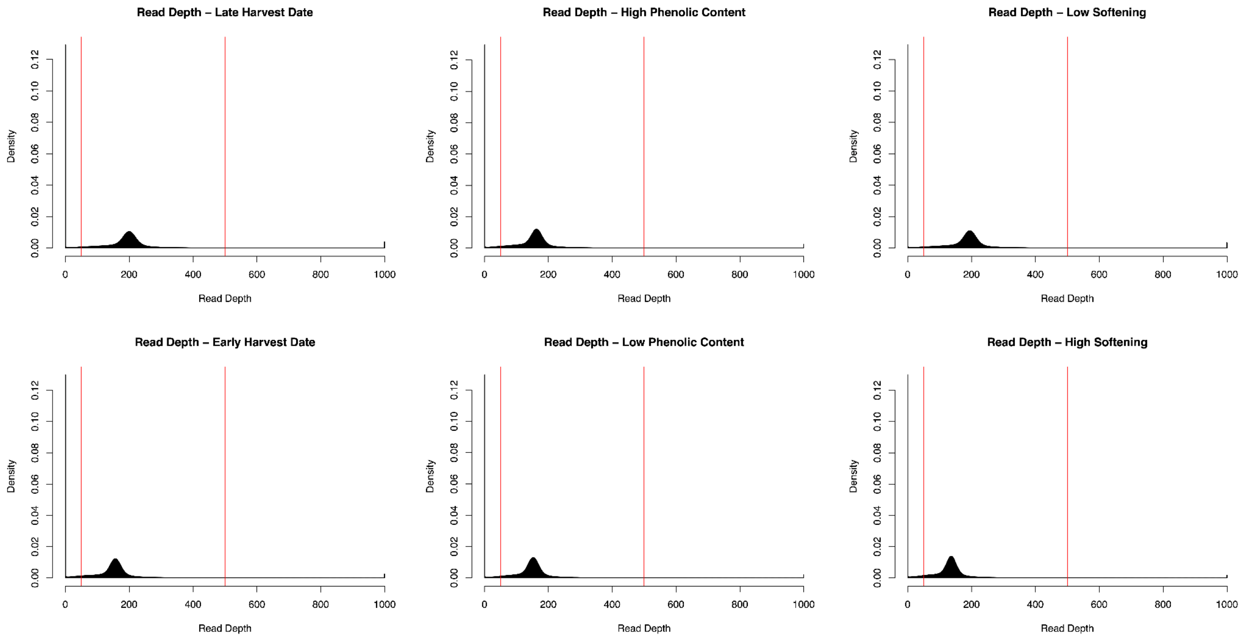


Appendix I-I. Phenotypes of cultivated apples as a function of their release year with a comparison to the ancestral state. Phenotypes include acidity change during storage, acidity, precocity, harvest date, firmness, and weight. Cultivated apple scores for each phenotype are shown in blue, and the ancestral state of each phenotype is represented in yellow as a density distribution of values from *M. sieversii*. The R and p values from a Pearson correlation between phenotypic values and release year are shown within each scatter plot.

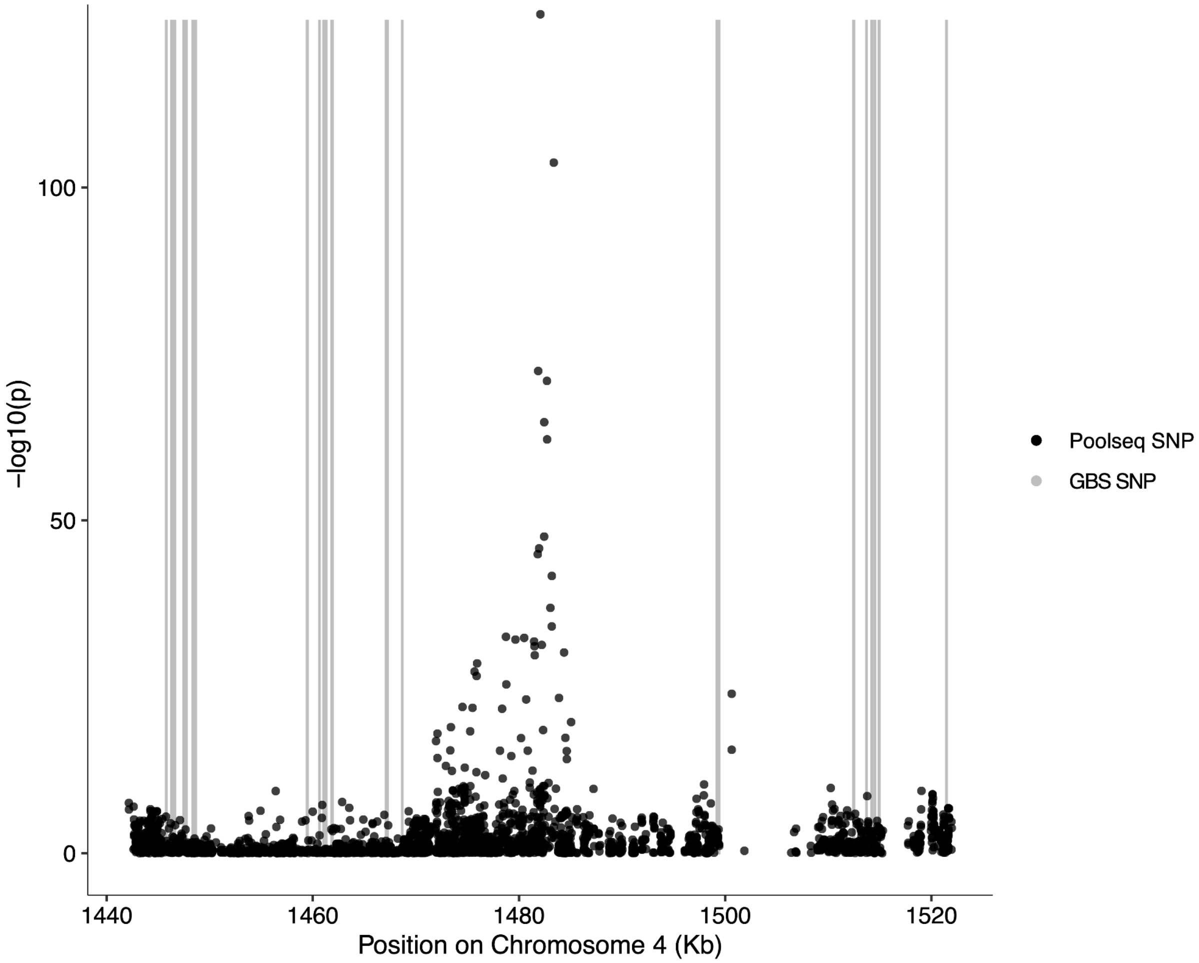
Appendix II: Pool-seq of diverse apple germplasm reveals candidate loci underlying ripening time, phenolic content, and softening (Chapter 3).



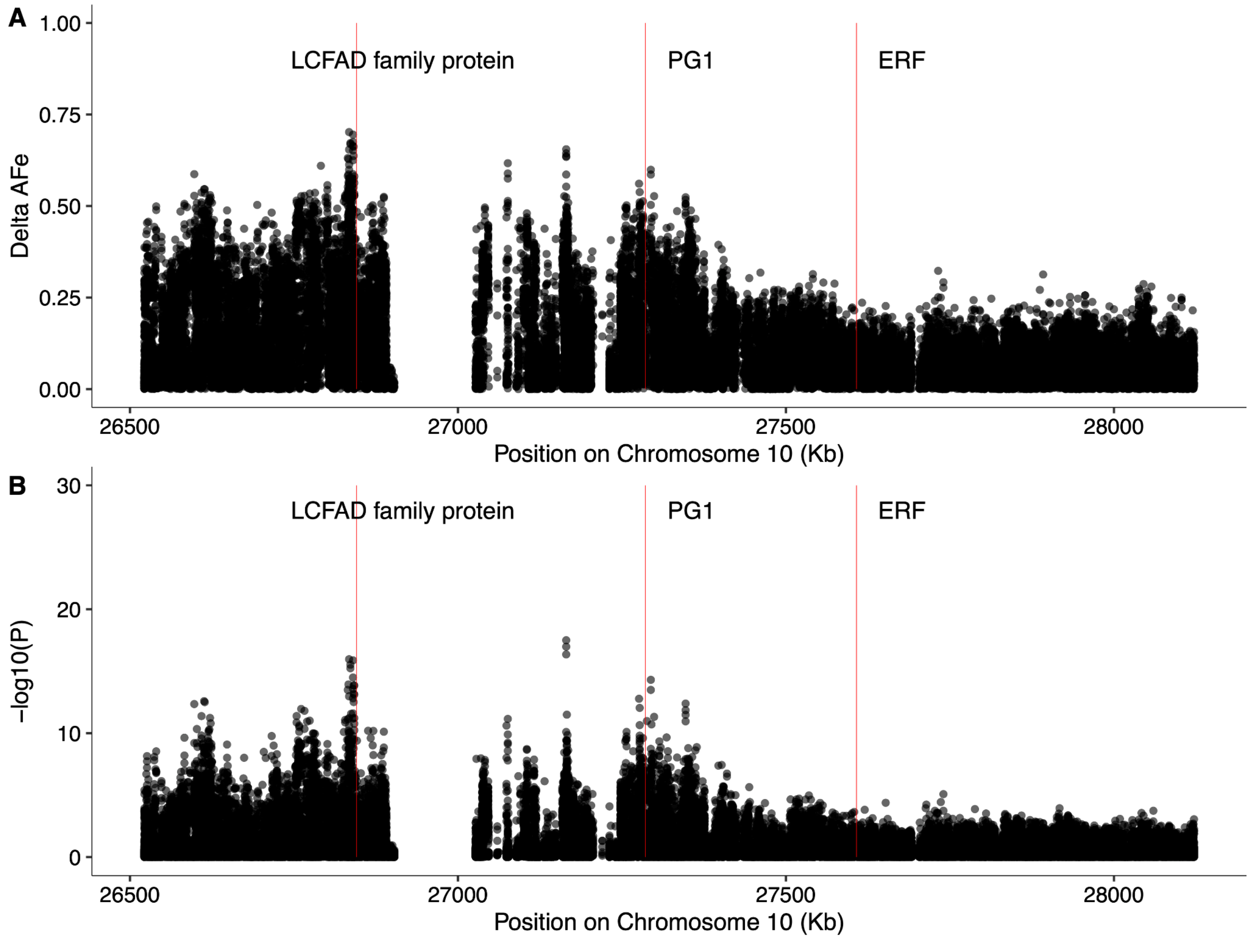
Appendix II-I. Bioinformatic workflow of the pool-seq GWAS.



Appendix II-II. Read depth histograms for each phenotype pool. Red bars indicate read depth cutoff limits (50x and 500x).



Appendix II-III. Overlap of variants from the present study and previous mapping experiments using GBS studies (Migicovsky et al. 2022).



Appendix II-IV. Manhattan plots for softening signal on chromosome 10. Delta-AFe (A) and CST p-values (B) are represented by black dots, red bars indicate coding regions of Long-chain fatty alcohol dehydrogenase family protein (LCFAD) (MD10G1176100), PG1, and ERF (MD10G1184800), respectively.

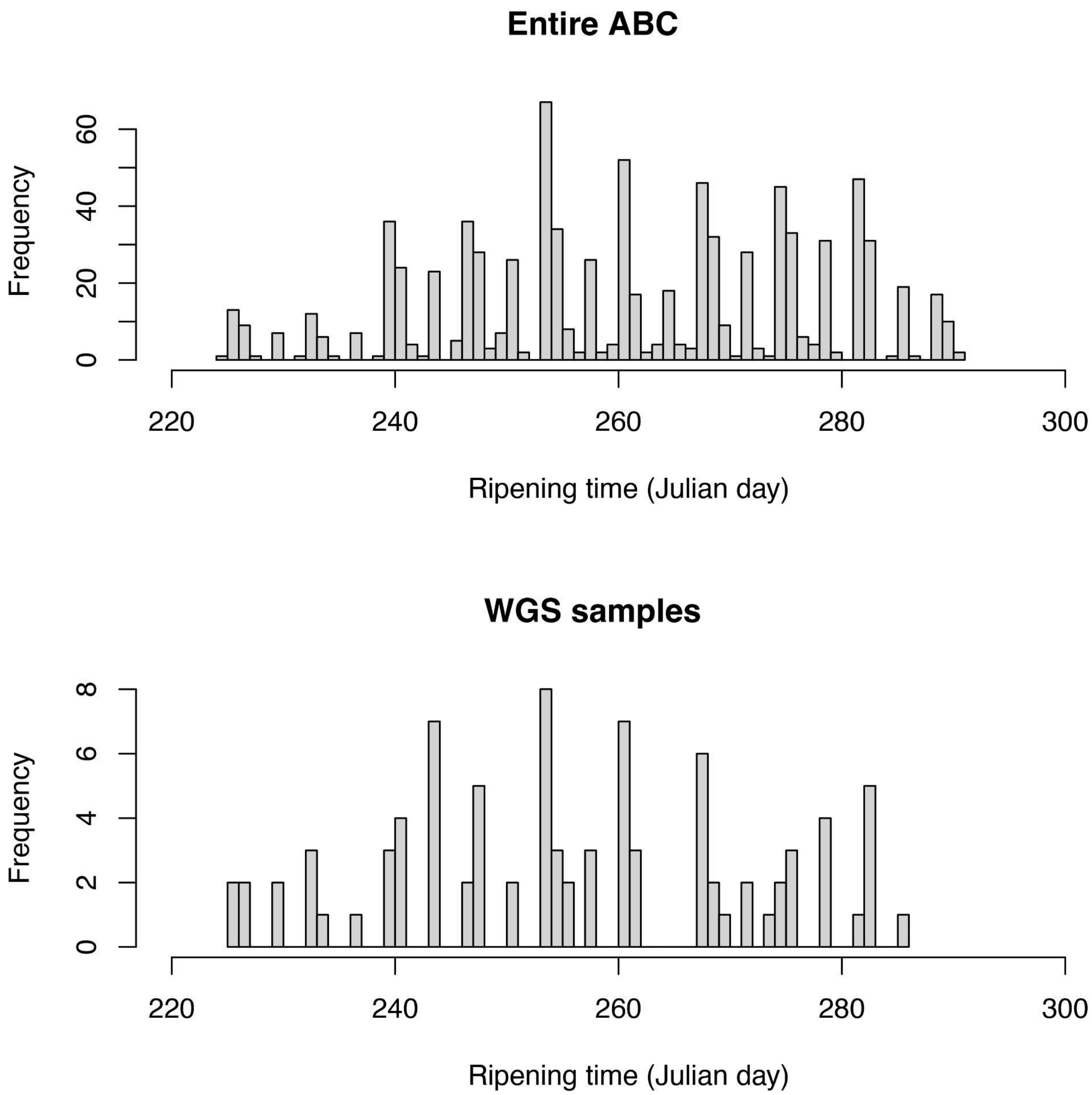
Table II-I. Genome coverage table. Various position read mapping data including total genome coverage by position, total genome coverage by percent, and average read depth. [ELECTRONIC SUPPLEMENT]

Table II-II. Ripening time extended results. Ripening time top variant hits, candidate genes, and top GO enrichment terms. [ELECTRONIC SUPPLEMENT]

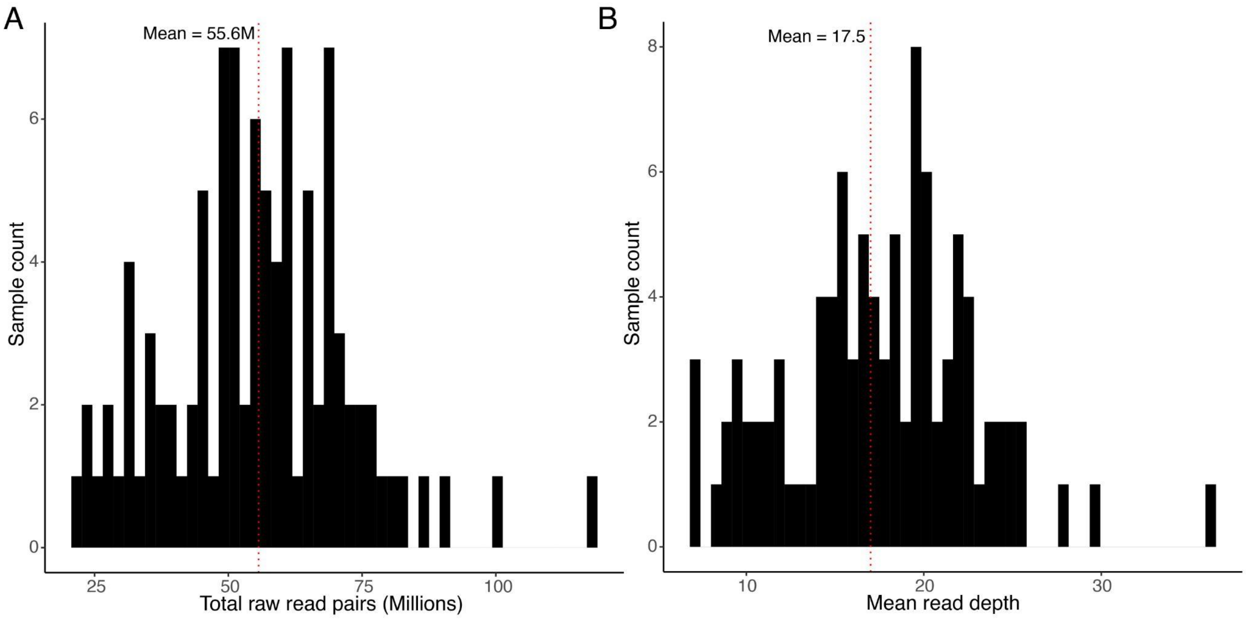
Table II-III. Total phenolic content extended results. Total phenolic content top variant hits, candidate genes, and top GO enrichment terms. [ELECTRONIC SUPPLEMENT]

Table II-IV. Softening extended results. Softening top variant hits, candidate genes, and top GO enrichment terms. [ELECTRONIC SUPPLEMENT]

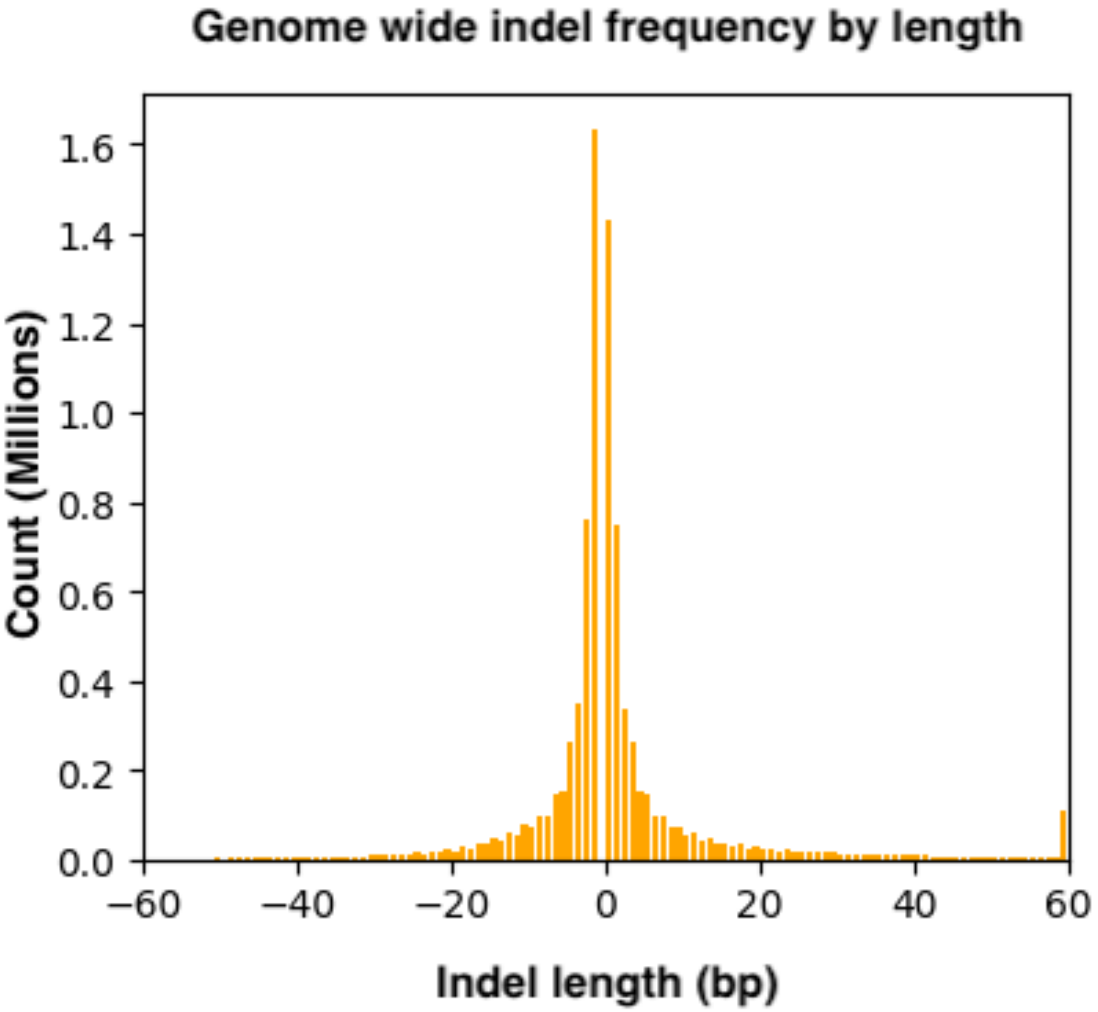
Appendix III: Fine mapping of the causal allele controlling ripening time in apple (Chapter 4)



Appendix III-I. Ripening time distributions of the entire ABC population (N = 837) and the samples selected for WGS (N = 107). The subset for the 107 sample selected for WGS clearly span the entire range of ripening times observed in the entire ABC population.



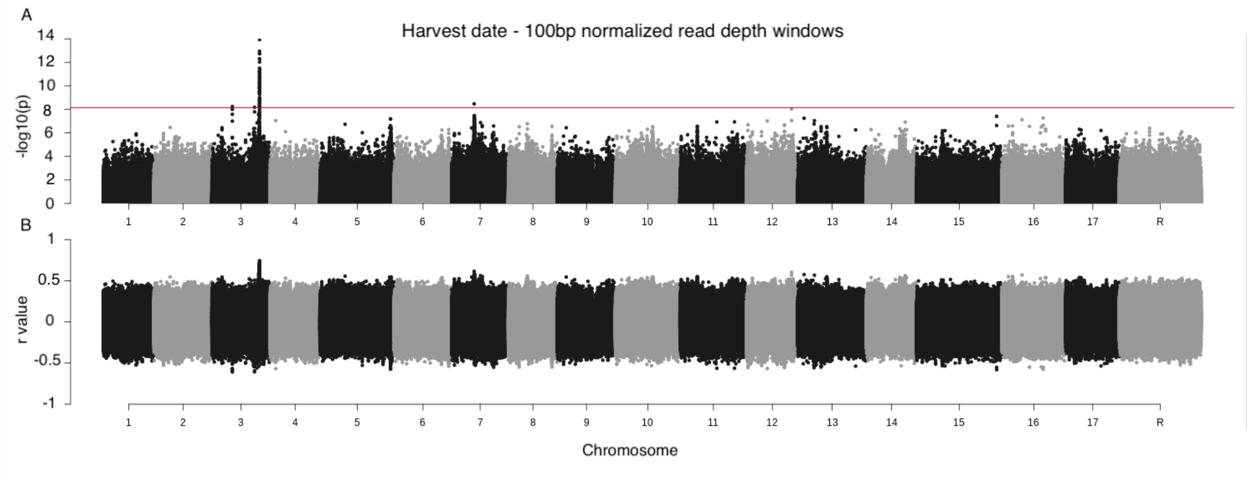
Appendix III-II. DNA sequencing read pairs (A) and mean read depth across samples (B). The mean number of raw read pairs generated for each sample was 55.6 million. The mean and median proportions of reads that passed quality score trimming across all samples was 95.79% and 99.94%, respectively. The average read depth across samples, after filtering and quality control, was 17.5x. The median insert size from PE reads was 292 bp, and the mean insert size was 312 bp.



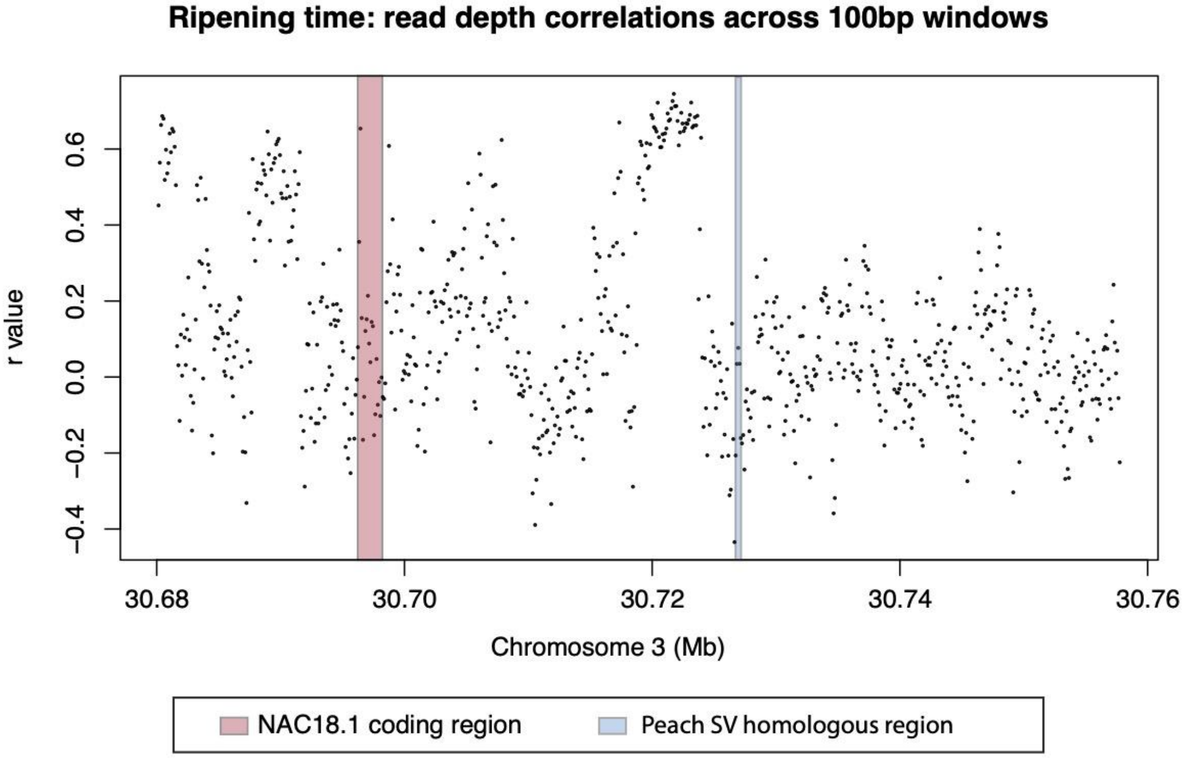
Appendix III-III. Indel length distribution for the WGS final variant dataset. Most indels detected in the data were small (< 10 bp), however indels up to 60 bp in size were detected.

$$NRD = 100 * (xRR + xRA + xAA) / (xRR + xRA + xAA + mRA + mAA)$$

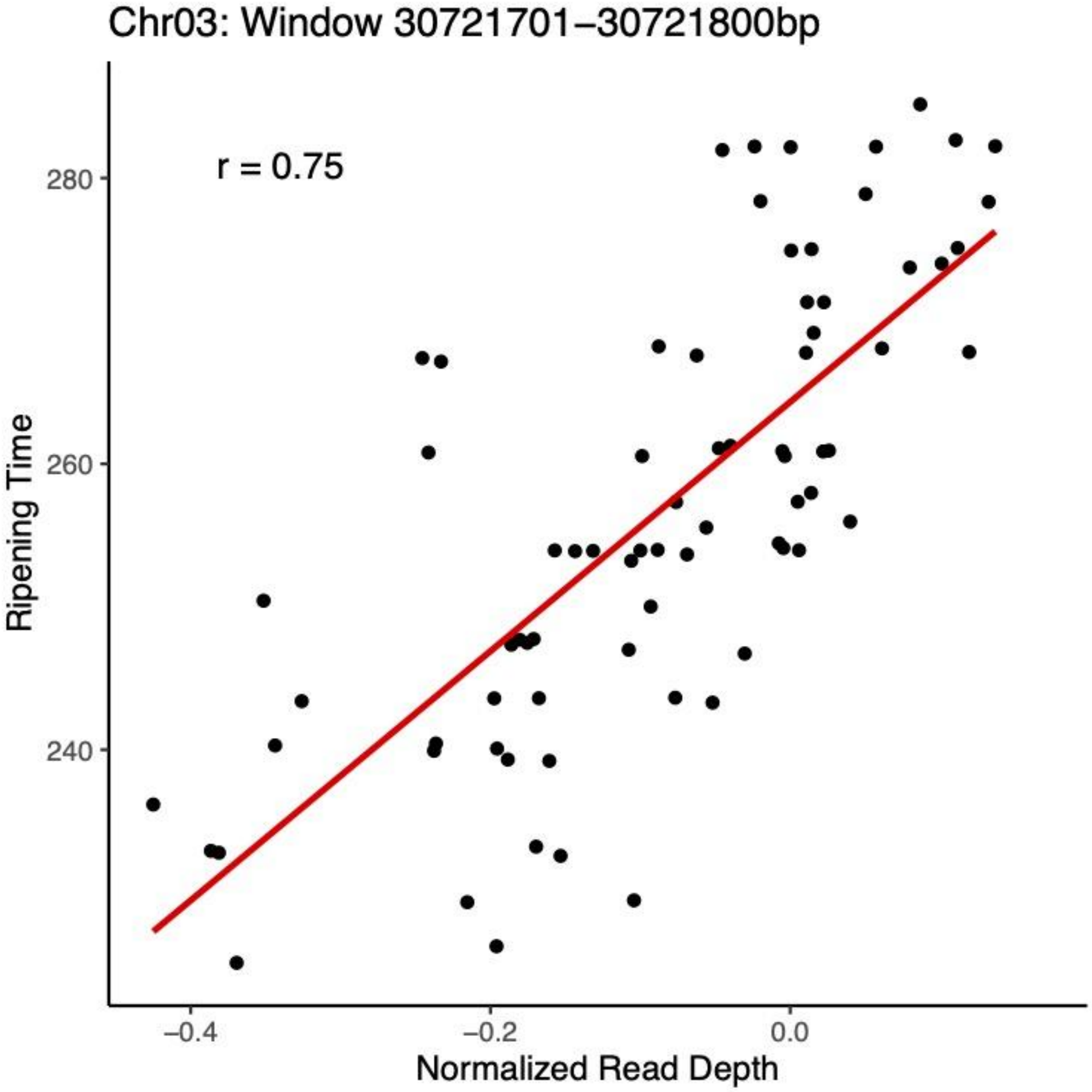
Appendix III-IV. Equation used to calculate non-reference discordance (NRD) for each sample between GBS and WGS VCF files. X = mismatches, m = matches, RR = reference/reference, RA = reference/alternate, AA = alternate/alternate.



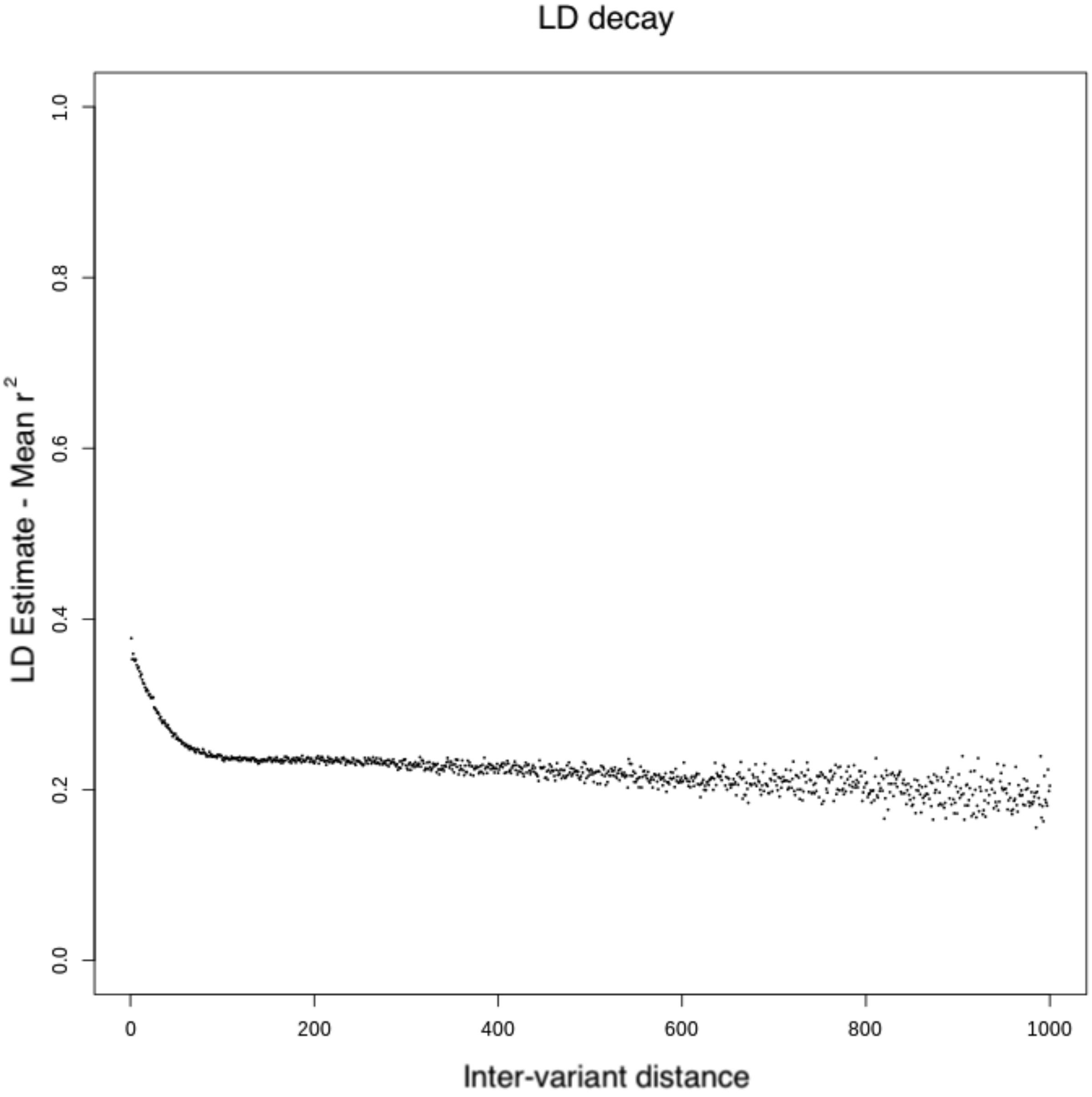
Appendix III-V. Manhattan plots for normalized read depth correlations across 100bp windows of the genome. A) manhattan plot showing P values for the pearson correlation between each window (dot) and ripening time. The red line indicates Bonferroni corrected threshold. B) manhattan plot showing r values for the pearson correlation between each window (dot) and ripening time.



Appendix III-VI. Read depth correlations across 100bp windows on within the strongest signal region for ripening time on Chromosome 3.



Appendix III-VII. Correlation of normalized read depth across the 100bp window spanning 30721701-30721800bp and ripening time for each WGS sample.



Appendix III-VIII. Linkage disequilibrium (LD) decay curve from inter-variant distances of 10 to 1000 bp in 76 WGS samples.

Table I-I. Variant frequency across the genome for WGS variants. [ELECTRONIC SUPPLEMENT]

Appendix IV: Copyright release

Chapter 2: Phenotypic divergence between the wild and cultivated apple

Davies, T., Watts, S., McClure, K., Migicovsky, Z., & Myles, S. (2022). Phenotypic divergence between the cultivated apple (Malus domestica) and its primary wild progenitor (Malus sieversii). PloS One, 17(3), e0250751. https://doi.org/10.1371/journal.pone.0250751

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license. The version of the manuscript which appears in this dissertation has been modified from the published version.

Chapter 3: Pool-seq of diverse apple germplasm reveals candidate loci underlying ripening time, phenolic content, and softening.

The contents of this chapter are published as:

Davies, T., & Myles, S. (2023). Pool-seq of diverse apple germplasm reveals candidate loci underlying ripening time, phenolic content, and softening. Fruit Research, 3(1). https://doi.org/10.48130/FruRes-2023-0011

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license. The version of the manuscript which appears in this dissertation has been modified from the published version.