Patterns of Genomic and Phenomic Diversity in Apple and Grape

by

Zoë Migicovsky

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

at

Dalhousie University
Halifax, Nova Scotia
June 2017

Dedicated to my Grandma Roz (Roslyn Glickman, 1929-2014)
and Grandma Wylma (Wylma Migicovsky, 1925-2015),
for all the years of love and encouragement
and for inspiring me to always keep learning.

Table of Contents

List of Tables

List of Figures

Abstract

Apples and grapes are two long-lived perennial crops which are economically valuable but whose genomic and phenomic diversity has not yet been fully described and exploited. In this thesis, there were two main objectives: to characterize the genetic basis of several traits in apple and to estimate the degree to which wild relatives have been exploited in modern grape breeding. To achieve these objectives, I first describe variation in apple leaf morphology and demonstrate that comprehensive morphometric analyses of leaf shape can capture hidden, heritable phenotypes. Next, I performed a genome-wide association study using historical data from a diverse apple collection. I identified numerous genotype-phenotype associations, including an amino acid substitution in the transcription factor NAC18.1 that is a strong functional candidate for fruit firmness and harvest date. I also assessed ancestry in some of the most widely grown commercial hybrid grape cultivars. Over one third of hybrids derived approximately half of their ancestry from wild *Vitis* and half from the domesticated grape *Vitis vinifera*, suggesting hybrid grape breeding is in its infancy. Finally, I conclude by describing the potential of using genomics to improve perennial crops through introgression of valuable traits from wild relatives, such as disease resistance. Genetic mapping in wild relatives is difficult since genomic tools are often ill-suited to wild-relatives and phenotyping is an expensive and difficult process. However, there is an urgent need to immediately begin the collection and characterization of wild relatives to enable introgression of these valuable traits using genomics-assisted breeding. Overall, the results of this thesis lead to the following conclusions: comprehensive morphometric techniques capture heritable variation, novel genomic insights can be generated using historical phenotype data from gene banks, hybrid grape breeding is still in its infancy, and wild relatives should be exploited for genomics-assisted breeding of perennial crops.

List of Abbreviations Used

Abbreviation	Description
AIC	Akaike Information Criterion
ALSV	Apple latent spherical virus
ARS	Agricultural Research Service
CI	Confidence interval
CNV	Copy number variation
CWR	Crop wild relative
EFD	Elliptical Fourier descriptor
GBS	Genotyping-by-sequencing
GCTA	Genome-wide complex trait analysis
GEBV	Genomic estimated breeding value
GM	Genetic modification
GMO	Genetically modified organism
GRIN	Germplasm Resources Information Network
GS	Genomic selection
GWAS	Genome-wide association study
HT	High-throughput
IBD	Identity-by-descent
IBS	Identity-by-state
InDels	Insertions and deletions
kNNI	*k*-nearest neighbors imputation
LD	Linkage disequilibrium
MAF	Minor allele frequency
MAGIC	Multi-parent Advanced Generation InterCross
MAS	Marker-assisted selection
NCGR	National Clonal Germplasm Repository
NGS	Next-generation sequencing
OR	Odds-Ratio
PAV	Presence/absence variation
PC	Principal component
PCA	Principal components analysis
PD	Pierce's disease
PH	Persistent homology
PRSV	Papaya ringspot virus
QTL	Quantitative trait locus
REML	REstricted Maximum Likelihood
SNP	Single nucleotide polymorphism
USDA	United States Department of Agriculture
VIGS	Virus-induced gene silencing

Acknowledgments

Thank you to Dr. Sean Myles, for his knowledge, guidance and encouragement. I am incredibly grateful for the opportunity to have done this work, and for his continued mentorship. I would also like to thank Dr. Mark Johnston for his support, as well as my committee members Dr. Rob Beiko and Dr. Bob Latta, for their insightful feedback. Additionally, thank you to the many co-authors and collaborators without whom this work would not have been possible.

Thank you to my fellow lab members: Gavin Douglas, Dr. Kyle Gardner, Dr. Kendra McClure, Dr. Mike McElroy, Dr. Daniel Money and Jason Sawler, as well as honorary lab member Dr. Kate Crosby. I am thankful for your help and advice, both as friends and scientists.

Thank you to the Natural Sciences and Engineering Research Council of Canada (CGS-D3) and Dalhousie University (Killam Predoctoral Fellowship and President's Award) for the financial support which allowed me to pursue this work.

Thank you to my entire family for all the love and encouragement. In particular, thank you to Rose, Jonah, Esty, my dad, and my mom, for listening to me talk about my work, and always being proud of me even if what I was saying didn't always make sense. Your support means everything to me and I could not have done it without you.

Finally, a huge thank you to Mathew, who I can always count on to distract or motivate me, as needed. I am so grateful to have had you with me throughout this experience.

Chapter 1: Introduction

Global food availability must double within the next 25 years to meet demands of a growing population, and yet, perennial species occupy less than 14% of the world's surface area dedicated to food production (McCouch et al., 2013; Food and Agriculture Organization of the United Nations, 2017). Increasing biodiversity should be a major priority for agriculture, and while the historical focus has been on annual crops, over 13% of the world's surface area dedicated to food production grows perennial crops (Food and Agriculture Organization of the United Nations, 2017). Perennial species, which live for 2 years or longer, will be critical to increasing food supply and sustainability. In comparison to annual species, perennials generally have longer growing seasons (Dohleman and Long, 2009), increased root carbon (Glover et al., 2010a), and reduced soil erosion risk (Vallebona et al., 2016). Despite the vital—and increasing—importance of perennial crops, most work towards understanding agricultural species has centered on annual models such as maize (*Zea mays* L.) and rice (*Oryza sativa* L.). It is essential that future work focuses on understanding the availability of unique and desirable traits in perennial species, as well as the genomic architecture underlying these traits, to improve breeding of new and desirable cultivars.

While the benefits of perennial crops are evident, advances in breeding are limited by several major barriers. Many important perennial crops such as apple (*Malus* X. *domestica* Borkh.) and grape (*Vitis vinifera* L.) have a lengthy juvenile phase, which makes the breeding of new cultivars time-consuming. For example, the recent breeding of 3 commercial apple cultivars took 26 years (Peil et al., 2008). Additionally, apple and grape are large plants requiring substantial space and money to grow. Further, when breeding new perennial cultivars, often a small number of elite cultivars are used. For example, work on 439 commercial apple cultivars and breeding selections found 64% were descended from five founding cultivars (Noiton and Alspach, 1996). Limited diversity mean that perennial crops are threatened by continually evolving pathogens, which can easily devastate entire agricultural industries. Breeding new perennial crops capable of surviving a changing environment, including disease pressure and climate change, is critical. However, the time and expense required to breed long-lived perennial crops demands that we use new tools to improve the speed and efficiency of the process, thus also decreasing the cost.

The most critical tool available for decreasing the time and expense of breeding new perennial cultivars is genomics. By using marker-assisted selection (MAS) or genomic selection (GS) breeders can select progeny possessing a trait of interest at the seed or seedling stage. Such tools are particularly valuable for perennials which are expensive and time-consuming to grow to maturity and evaluate. For example, in grape, MAS was estimated to save up to 34% of operational costs over the first 6 to 8 years of a breeding program, while in apple savings of up to 43% were predicted (Edge-Garza et al., 2015). MAS relies on either a small number of genetic markers, or even only one, that are strongly linked to a phenotype. MAS is valuable for traits which are controlled by a small number of large effect loci. In comparison, GS uses genome-wide markers to predict a polygenic phenotype controlled by a large number of small effect loci. Thus, in order to determine which method of genomics-assisted breeding is most valuable, it is necessary to first have an understanding of the genetic architecture of the trait of interest.

Understanding the genetic basis of a trait requires both phenotype and genotype data. Fortunately, the development of next-generation DNA sequencing technologies (NGS), such as genotyping-by-sequencing (GBS), have greatly reduced the costs of acquiring genome-wide genotype data (Elshire et al., 2011). Over the past decade reference genomes have been developed for many perennial crops including apple (Velasco et al., 2010b) and grape (Jaillon et al., 2007), and this number continues to increase. Reference genomes are an essential tool for the alignment of DNA sequence data in order to detect genetic variants—in particular single nucleotide polymorphisms (SNPs)—throughout the genome. The increasing availability and decreasing cost of genomic tools and resources in perennial crops can facilitate a better understanding of the genetic basis of both simple (i.e. Mendelian) and polygenic traits.

While it is reasonable to assume that the cost of acquiring genotype data for the purposes for genetic mapping and genomics-assisted breeding will continue to decrease, the cost of acquiring high quality phenotype data is unlikely to follow a similar trajectory. Thus, there is an increasing interest in developing high-throughput (HT) phenotyping methods that enable the rapid collection of high quality phenotype information. Of particular interest is the use of automated procedures to generate high dimensional comprehensive phenotypic evaluations of plant morphology. This new field of HT phenotyping is often referred to as phenomics (Houle et al., 2010; Furbank and Tester, 2011).

One valuable tool for HT phenotyping is automated image analysis of 2-dimensional shapes, such as leaves. Traditional estimates of leaf shape were restricted to linear measurements such as length and width, but advanced image analysis tools and comprehensive morphometric techniques can now be used to quantify the outline of a shape. Recent work on leaf shape has made use of two comprehensive morphometric techniques in particular. In the first method, Generalized Procrustes analysis uses landmarks representing homologous points, such as vein architecture and lobes, scales the data to eliminate the effect of size, and thereby allows for a comparison between leaf shapes. Alternatively, elliptical Fourier descriptors (EFDs) converts the outline of a leaf into a chain of numbers, indicating step-by-step movements that comprise the outline, after which a Fourier decomposition is applied converting the shape into a harmonic series. Both landmarks and EFDs have been applied to leaves from species such as tomato (Chitwood et al., 2012) and *Passiflora* (Chitwood and Otoni, 2017). In comparison, comprehensive morphometric techniques have not yet been performed on apple leaves, although EFDs have been applied to the fruit (Currie et al., 2000). In addition to landmarks and EFDs, a novel, topology-based morphometric technique, persistent homology, was recently described. Unlike landmarks and EFDs, persistent homology can be applied to 3-dimensional structures, such as branching patterns and root architecture (Li et al., 2017b). Regardless of the morphometric method used for estimating shape, Principal Components Analysis (PCA) can be applied to the resulting numeric dataset. In grape, PCA of leaf shape using comprehensive morphometrics has been used to distinguish between species (Chitwood et al., 2016), cultivars (Chitwood et al., 2014), and even clones of a particular genotype (Klein et al., 2017). However, while comprehensive morphometrics can be used to capture variation in leaf shape, the genetic basis and heritability of the phenotypes estimated from these techniques is unknown.

Decreasing sequencing costs are increasing access to genomic information and facilitating our understanding of complex traits such as leaf shape, but phenotyping traits of interest remains a slow and expensive process, resulting in a "phenotyping bottleneck" (Furbank and Tester, 2011). In addition to the use of HT phenotyping, one highly supported mechanism for addressing the barrier of limited phenotype data—especially in long-lived perennials where the collection of new phenotype data may take decades—is the use of historical phenotype data from gene banks. Despite the difficulties associated with using phenotype data not specifically collected for genetic mapping purposes, historical phenotype data has been successfully used for genetic mapping in barley (*Hordeum vulgare* L.) and potato (*Solanum tuberosum* L.) (Baldwin et al., 2011; Matthies et al., 2014). The first step of genomics-assisted breeding is to link a trait with genomic data for marker discovery, and thus gene banks are a valuable resource for historical phenotype data that can facilitate this process.

In addition to the use of historical phenotype data for genetic mapping, another valuable resource for improvement of perennial crops are crop wild relatives (CWRs). Perennial breeding makes use of a small number of elite cultivars and over 75% of perennials are clonally propagated and thus remain genetically frozen (Miller and Gross, 2011). In comparison, wild relatives continue to evolve in response to natural selection such as disease pressure. As a result, wild relatives often harbor valuable disease resistance genes. Introgression of these traits has been the primary area of interest when breeding with wild relatives thus far: in a review of 19 different crops over 80% of traits from CWRs were involved in disease and pest resistance (Hajjar and Hodgkin, 2007). The use of wild relatives for perennial crop breeding can result in cultivars resistant to evolving pathogens, thus increasing yield and decreasing chemical input.

In perennial crops, wild relatives have been an important source of disease resistance in grape. In the 1860s, the North American louse phylloxera (*Daktulosphaira vitifoliae*) devastated European vineyards that were planted exclusively with grapes from the domesticated species, *V. vinifera*. It was only through making use of wild *Vitis* species for hybrid rootstocks that breeders rescued the wine industry (Alleweldt and Possingham, 1988; Zhang et al., 2009). More recently, Pierce's disease (PD) (*Xylella fastidiosa*) has become a costly threat to the California wine industry (Alston et al., 2013). Fortunately, wild grapes harbor disease-resistance, and breeders can use MAS to track this resistance when breeding new commercial grape cultivars. By repeated backcrossing to *V. vinifera*, grape breeders have managed to generate breeding lines with both PD resistance from the wild and over 97% *V. vinifera* ancestry (Walker et al., 2014). However, the proportion *V. vinifera* ancestry in most other hybrid grapes is much less clear. Even cultivars identified as *V. vinifera* may be incorrect. For example, recent work revealed that 'Koshu', a Japanese wine grape widely believed to be 100% *V. vinifera*, contains 30% wild ancestry (Goto-Yamamoto et al., 2015). Quantifying the level of wild ancestry in commercially grown hybrid grapes is the first step towards facilitating the breeding of new cultivars which possess both desirable traits, such as disease resistance, from wild relatives as well as a high proportion of *V. vinifera* ancestry to allow for commercially acceptable wine.

It is not simply grape breeding that can benefit from traits present in wild relatives. The combination of genomics-assisted breeding and introgression of desirable traits—including diseases resistance, fruit quality and rootstock traits—provides a critical opportunity for improving perennial crops while decreasing the cost and time required for breeding. For example, a recently released commercial apple cultivar, 'WA 5', underwent genetic testing to determine that it carried alleles for scab resistance from a wild relative and desirable alleles for fruit firmness (Evans et al., 2011). Many other perennial crops could also benefit from genomics-assisted breeding using wild relatives and it is time to begin harvesting their potential.

Given the global importance of perennial crops and the need to improve our understanding of them, the main goals of my thesis were to characterize patterns of genomic and phenomic diversity in two important perennial species, apple and grape. In Chapter 2, I begin by using comprehensive morphometrics to describe variation in leaf shape in apple, linking this information with genome-wide SNP data to describe the genetic architecture underlying this trait. In Chapter 3, I use historical phenotype data available from the United States Department of Agriculture to perform genetic mapping in apple for agriculturally important traits like fruit color and harvest date. In Chapter 4, I estimate the wild ancestry of commercially important hybrid grape cultivars from Canada, the United States and Germany. In Chapter 5, I conclude with a discussion of the current limitations and future promise of wild relatives for genomics-assisted breeding in perennial crops. Overall, my thesis describes the genetic basis of immense phenomic diversity in two economically important crops, while emphasizing the potential of both genomics and wild relatives for further improvement of perennial crops.

Chapter 2: Quantifying the genetic basis of leaf shape in apple

Introduction

Apples (*Malus spp.*) are one of the world's most widely grown fruit crops, with the third highest global production quantity of over 84 million tonnes in 2014 (Food and Agriculture Organization of the United Nations, 2017). Apple leaves are generally simple, with an elliptical-to-ovate shape. Previous studies in apple used linear measurements, such as length and width, to quantify leaf shape (Liebhard et al., 2003; Bassett et al., 2011). The length-to-width aspect ratio is a major source of variation in leaf shape. Differing aspect ratios lead to a disproportionate increase or decrease in length relative to width, or allometric variation, in leaves (Gurevitch, 1992; Chitwood et al., 2013). While linear measurements such as leaf length and width are useful, they fail to capture the full extent of leaf shape diversity. Failing to measure leaf shape comprehensively also limits our ability to discern the total underlying genetic contributions.

Elliptical Fourier descriptors (EFDs) are a valuable, well-recognized tool for quantifying the outline of a shape. EFD analysis first converts a contour to a chaincode, a lossless data compression method that encodes shape by a chain of numbers, in which each number indicates step-by-step movements to reconstruct the pixels comprising the shape. A Fourier decomposition is subsequently applied to the chain code, quantifying the shape as a harmonic series. EFDs have been used extensively to quantify leaf shape in diverse species, such as grape (Chitwood et al., 2014), tomato (Chitwood et al., 2012), and *Passiflora* (Chitwood and Otoni, 2017). Previous work used EFDs to assess apple fruit shape (Currie et al., 2000), but this technique has not yet been applied to apple leaves. A newly developed morphometric technique, persistent homology (PH), provides another method for estimating leaf shape. PH, like EFDs, is normalized to differences in size, but it is also orientation invariant. PH treats the pixels of a contour as a 2D point cloud before applying a neighbor density estimator to each pixel. A series of annulus kernels of increasing radii are used to isolate and smooth the contour densities. The number of connected components is recorded as a function of density for each annulus, resulting in a persistence barcode that quantifies shape as topology. The topology-based PH approach can also be applied to serrations and root architecture, allowing the same method to be used across different plant structures (Li et al., 2017a; Li et al., 2017b).

Comprehensively measuring leaf shape, using approaches such as EFDs and PH, is important, as shape features may be associated with agriculturally important traits. Leaves are present during the lengthy juvenile phase in apple but fruits appear only on mature trees and thus, leaf traits can enable early selection without the need for genetic markers. In apple, it generally takes 5 years for significant fruiting to occur and thus, any ability to discard trees not possessing a trait of interest earlier in development is extremely valuable (Kumar et al., 2012a). There are already several cases of unique leaf characteristics providing an early marker for other genetic differences in apple. For example, the gene underlying red fruit flesh color may lead to anthocyanin accumulation in the leaves, causing red foliage (Chagne et al., 2007; Espley et al., 2009) while columnar tree architecture may be accompanied by an increase in leaf number, area, weight per unit area and length-to-width ratio (Talwara et al., 2013). Leaf pH has also been proposed as an early indicator of low acid fruit (Visser and Verhaegh, 1978).

In addition to serving as early markers for other traits, leaf shape and size may influence the amount of light a tree receives, and light exposure is crucial for flowering in apple. Light penetration results in higher levels of flowering, while leaf injury or defoliation can reduce flowering (Dennis, 2003). Thinning apple trees to a particular leaf-to-fruit ratio is a common practice to attain optimal fruit color and size (Fletcher, 1932; Preston, 1954). Contrastingly, trees with fewer fruit may increase vegetative growth and thus leaf area (Wünsche et al., 2000). In previous work, several leaf traits such as area and perimeter were correlated with apple fruit size (Khan et al., 2014). Clearly, there is an important relationship between the leaves and the fruit, and comprehensively quantifying the variation in leaf shape is a crucial component to understanding this relationship in apple.

Leaves are the main photosynthetic organs of apple, but the genetic basis underlying their shape and size remains unknown. In cotton, a single locus controls the major leaf shapes (Andres et al., 2017), but in most instances leaf shape appears to be controlled by numerous small-effect loci (Tian et al., 2011; Chitwood et al., 2013). There are limited examples of genomic analyses of leaf shape in apple, however, a previous bi-parental linkage mapping study found two suggestive quantitative trait loci for leaf size (Liebhard et al., 2003). Previous work also measured several leaf traits such as area, perimeter and circularity, in 158 apple accessions. The study linked these measurements with 901 single nucleotide polymorphisms (SNPs) but found no significant genotype-phenotype relationships (Khan et al., 2014). Thus far, efforts have not been made to estimate the genetic heritability of comprehensive morphometric leaf phenotypes, such as those described using EFDs and PH. It therefore remains unclear to what extent these methods are capturing biologically meaningful, heritable variation.

To fully understand the genetic basis of leaf shape, it is essential to include both linear and morphometric estimates of shape. Decreasing sequencing costs and access to a large and diverse germplasm collection allowed us to analyze approximately 9,000 leaves from over 800 unique accessions which we linked to over 122,000 genome-wide SNPs. We present the first comprehensive analysis of leaf shape in apple, revealing that both accessions and species show allometric variation due to differences in the width of the leaf blade. While the primary axis of variation in apple using EFDs and PH is due to this allometric variation, we find high narrow-sense heritability values even at later principal components, indicating that comprehensive estimates of shape capture heritable variation which would be missed by linear estimates alone.

Materials and Methods

Sample collection

Apple trees in Kentville, Nova Scotia, Canada were budded onto M.9 rootstocks in spring 2012. In the fall, the trees were uprooted and kept in cold storage until spring 2013, when trees were planted in an incomplete block design (see “REstricted Maximum Likelihood

(REML)” below). Leaves from over 900 trees were collected from August 24th to September 16th 2015. Ten leaves were collected from each tree. Leaves were flattened and placed to avoid touching, then scanned using Canon CanoScan (LiDE 220) Colour Image Scanners. Leaves were then dried for 48 hours at 65 °C and weighed to estimate the total dry weight (g) for each tree.

Morphometric analyses

Leaf scans were converted into a separate binary image for each leaf using custom ImageJ macros, which included the ‘make binary’ function (Abràmoff et al., 2004). A new image file was created for each leaf and named after the tree ID. Images were converted to RGB .bmp files and a chain code analysis was performed using SHAPE (Iwata and Ukai, 2002). The chain code was used to calculate normalized elliptical Fourier descriptors (EFDs) in SHAPE. The normalized EFDs were read into Momocs v1.1.5 (Bonhomme et al., 2014) in R (R Core Team, 2016) where harmonics B and C were removed to eliminate asymmetrical variation in leaf shape.

The binary leaf images were also analyzed using persistent homology (PH) (Li et al., 2017b). To numerically estimate the shape of the leaves using PH, we extracted the leaf contour using a 2D point cloud (Figure 2-1a). After centering and normalizing the contour to its centroid size, we used a Gaussian density estimator (Figure 2-1b), which assigns high values (red) to pixels with many neighboring pixels, and low values (blue) to pixels with fewer neighboring pixels. We multiplied the density estimator by an annulus kernel, or ring (Figure 2-1c), which emphasizes the shape in an annulus at the centroid and is thus invariant to orientation (Figure 2-1d). The resulting function can also be visualized from the side view (Figure 2-1e,f). As we moved a plane from top to the bottom, we recorded the number of connected components above the plane, forming a curve. With each new component this value increased, and each time components were merged, it decreased (Figure 2-1g). For each leaf, we computed 16 curves corresponding to 16 expanding rings. For computational purposes, each curve is divided into 500 numbers, ultimately resulting in the shape of each leaf being represented by 8,000 (16*500) values.

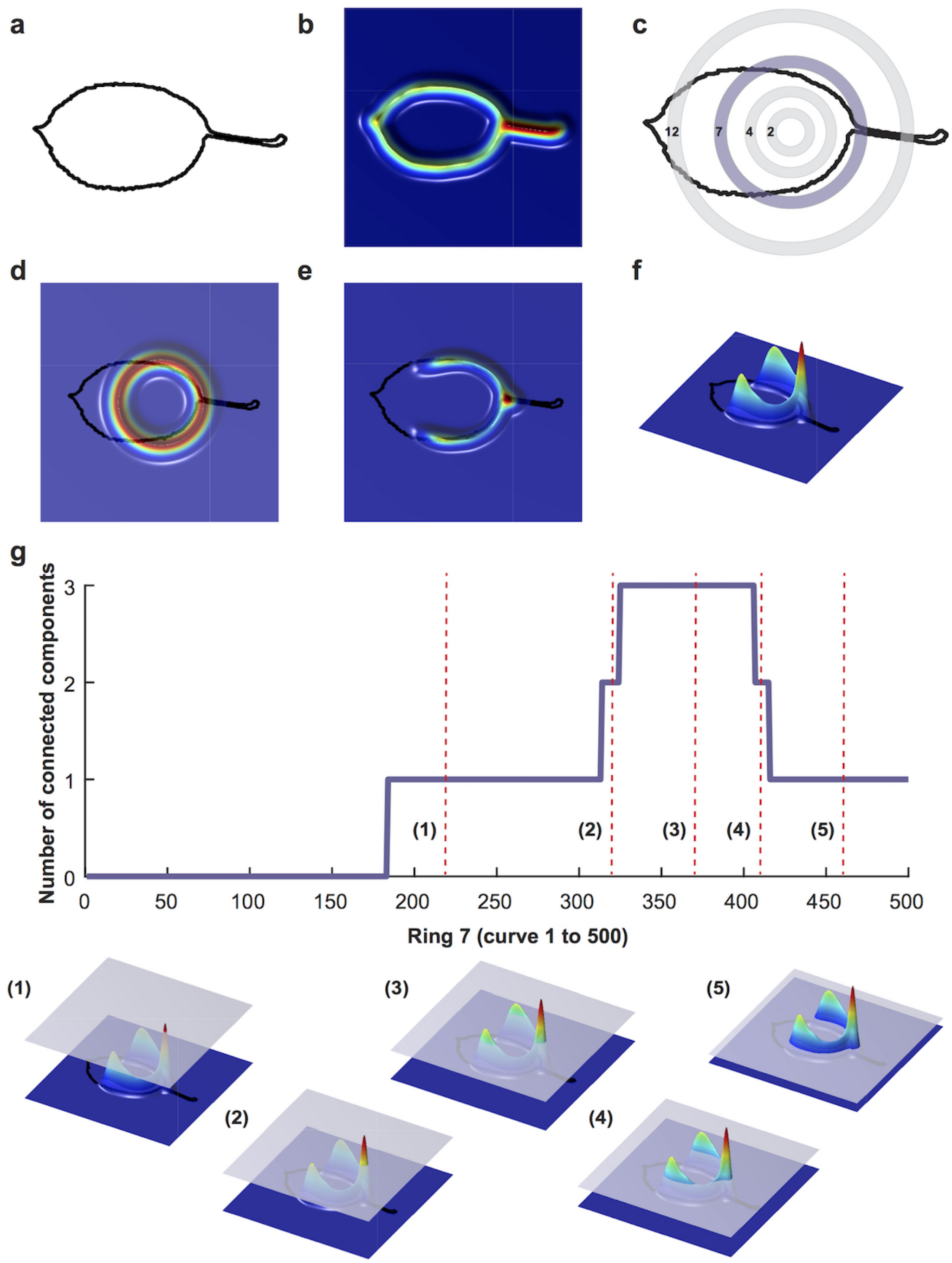


Figure 2-1. Visualization of persistent homology technique for annulus kernel 7. Binary images were converted into a 2D point cloud (a) which was then normalized using a Gaussian density estimator (b). For each leaf, 16 annulus kernels were used. Annulus kernel 7, indicated in purple (c) is used as an example for this visualization. The density estimator is multiplied by annulus kernel 7 (d). The function can also be visualized from the side view (e, f). As a plane moves from top to bottom, the number of connected components is recorded along the curve (g). Below (g) are five visualizations of curves that are represented as red vertical dotted lines in (g).

Only leaves for which both EFDs and PH shape estimations were successfully calculated were included in subsequent analyses. Additionally, only trees with 8-10 leaves were included, as leaves were sometimes removed due to tears, folding, or the absence of a petiole which did not allow for accurate quantification of shape. The final dataset consisted of 915 trees with 8-10 leaves, which included 869 unique accessions and 8,995 leaves.

EFDs and PH values were averaged across leaves from an individual tree. The contribution of EFD harmonics 1 to 15 to the mean leaf shape across all trees was visualized using the 'hcontrib' function in the Momocs R package (Figure 2-2). To allow for discrimination between accessions based on leaf shape, principal component analysis (PCA) was performed using the Momocs 'PCA' function (Bonhomme et al., 2014) for EFDs, and the 'prcomp' function in R for PH values, which center but do not scale the data. The resulting PC values were adjusted using REstricted Maximum Likelihood (see below). Subsequently, we identified the accession with the minimum and maximum value along each of the first 5 PCs.

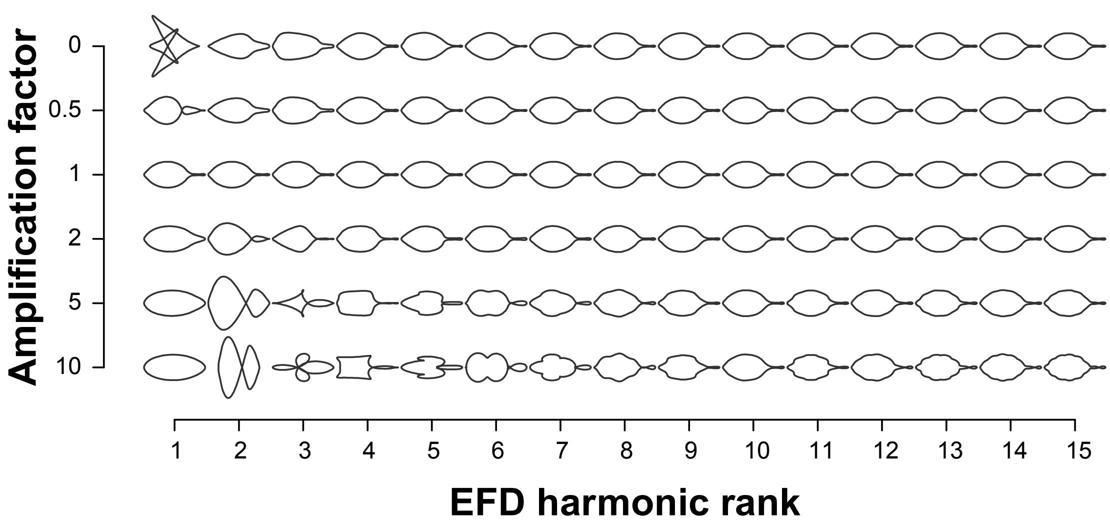


Figure 2-2. Contribution of elliptical Fourier descriptor harmonics to leaf shape. The leaf shapes depicted are the mean leaf shapes based on all 915 trees. Harmonics 1 to 15 are represented on the x-axis and each harmonic is multiplied by the amplification factor on the y-axis to visualize their contribution to mean leaf shape. An amplification factor of 0 indicates the removal of the harmonic; a factor of 1 results in the normal shape; and values above 1 exaggerate effects to better visualize the harmonic's contribution to the final shape.

In addition to estimating the contour of the leaf using EFDs and PH, we used several more metrics to describe the leaves. Using ImageJ, we automated the measurement of leaf surface area (cm^2), length (cm) of the leaf and width (cm) of the leaf as well as major (blade length) and minor (blade width) axes of the best fitting ellipse—which excluded the petiole—through batch processes (Abràmoff et al., 2004). Throughout the manuscript, we use 'major' when referring to the length of the leaf blade, and 'minor' when referencing the width of the leaf blade. We also calculated the aspect ratio of the leaf, by dividing the major axis by the minor axis. Additionally, leaf mass per area was calculated for 780 trees where we possessed surface area data for all 10 leaves, by calculating the ratio of dry weight to surface area (g/cm^2).

While linear phenotypes were calculated as an average value for a particular tree, we also estimated variance within a tree for aspect ratio, length, width, major and minor axis, and surface area. Variance was calculated as the coefficient of variation using the 'cv' function in the raster package (Hijmans, 2016) in R to estimate within-tree variability in leaf size, which is indicated as 'var' throughout this manuscript.

REstricted Maximum Likelihood (REML) adjustment of phenotype data

The orchard sampled in this study is an incomplete block design with 1 of 3 standards per grid. The standards, or "control trees"—'Honeycrisp', 'SweeTango', and 'Ambrosia'—are replicated across the grid. Leaves from these trees were sampled multiple times across the orchard, which allowed us to correct for positional effects. Each phenotype was adjusted using a REstricted Maximum Likelihood (REML) model which resulted in one adjusted value per accession, even when multiple trees were measured. The impact of row grid (rGrid), column grid (cGrid) and rGrid x cGrid effects were adjusted for using the following REML model:

$$phenotype \sim accession + (1 \mid rGrid) + (1 \mid cGrid) + (1 \mid cGrid{:}rGrid)$$

We fit a linear mixed-effects model via REML using the ‘lmer’ function in the lme4 package in R (Bates et al., 2015) and then calculated the least squares means using the ‘lsmeans’ function in the lsmeans R package (Lenth, 2016).

Thus, while the initial phenotype data was collected for 915 trees, following REML adjustment, one value remained per unique accession, resulting in 869 accessions. REML-adjustment was applied directly to all size, weight and variance estimates. For PH and EFDs, we applied the REML following PCA and thus the percent contribution for each PC was calculated using unadjusted values.

Phenomic analyses

The correlation between leaf phenotypes was calculated using Pearson’s correlation and p-values were Bonferroni-corrected for multiple comparisons. The resulting heatmap was visualized using the ‘geom_tile’ function in ggplot2 in R (Wickham, 2009). Next, we examined the leaves for allometry using the ‘SMA’ function in the smartr R package (Warton et al., 2012) to estimate if the slope between the log-transformed minor and major axis differed from 1.

Accessions were labelled as either *Malus* x. *domestica* Borkh. or *Malus sieversii* Lebed. based on information provided by the United States Department of Agriculture (USDA) Germplasm Resources Information Network website (http://www.ars-grin.gov/). We used a Mann-Whitney U test to test if any phenotypes differed between species and Bonferroni-corrected all p-values for multiple comparisons.

Genomic analyses

DNA was extracted using commercial extraction kits. Genotyping-by-sequencing (GBS) libraries were prepared using ApeKI and PstI-EcoT221I restriction enzymes according to Elshire et al. (2011). GBS libraries were sequenced using Illumina Hi-Seq 2000 technology. Reads which failed Illumina's "chastity filter" were removed from raw fastq files. Remaining reads were aligned to the *Malus* x. *domestica* v1.0 pseudo haplotype reference sequence (Velasco et al., 2010a) using the Burrows-Wheeler aligner tool v0.7.12 (Li and Durbin, 2009) and the Tassel version 5 pipeline (Glaubitz et al., 2014). Tassel parameters included a minKmerL of 30, mnQS of 20, mxKmerNum of 50000000 and batchSize of 20. The kmerlength was set to 82 for ApeKI and 89 for PstI-EcoT22I based on the max barcode size. The minMAF for the DiscoverySNPCallerPluginV2 was set to 0.01. All other default parameters were used. Non-biallelic sites and indels were removed using VCFtools v.0.1.14 (Danecek et al., 2011). VCFs for both enzymes were then merged using a custom perl script, preferentially keeping SNPs called by PstI-EcoT22I at overlapping sites, since those sites tended to be at higher coverage.

Missing data was imputed using LinkImputeR v0.9 (Money et al., Submitted, available: http://www.cultivatingdiversity.org/software.html) with global thresholds of 0.01 for minor allele frequency (MAF) and 0.70 for missingness. We examined depths of 3 to 8 and selected a case for imputation with a max position/sample missingness of 0.70, a minimum depth of 5, and an imputation accuracy of 94.9%. The VCF was converted to a genotype table using PLINK v1.07 (Purcell et al., 2007; Purcell, 2009b).

Of the 869 accessions assessed in this study, 816 had genomic data following imputation and filtering and were included in downstream analyses. The resulting genotype table consisted of 816 accessions and 197,565 SNPs. Subsequently, a 0.05 MAF filter was applied using PLINK, after which 128,132 SNPs remained. SNPs with more than 90% heterozygous genotypes were removed. The final genotype table consisted of 816 samples and 122,596 SNPs.

To perform PCA, SNPs were pruned for linkage disequilibrium (LD) using PLINK. We considered a window of 10 SNPs, removing one SNP from a pair if $R^2 > 0.5$, then shifting the window by 3 SNPs and repeating (PLINK command: indep-pairwise 10 3 0.5). This resulted in a set of 75,973 SNPs for 816 accessions. PCA was performed on the

LD-pruned genome-wide SNPs using 'prcomp' in R with data that were centered but not scaled. The first 2 genomic PCs were visualized using ggplot2 in R (Wickham, 2009).

We performed a genome-wide association study (GWAS) using the mixed linear model in Tassel (version 5) for each phenotype, adjusting for relatedness among individuals using a kinship matrix as well as the first 3 PCs for population structure (Bradbury et al., 2007; Zhang et al., 2010). The threshold for significance was calculated using simpleM (Gao et al., 2008; Gao et al., 2010) which estimates the number of PCs needed to explain 0.995 of the variance, or the number of independent SNPs. The inferred Meff used to calculate the significance threshold was 91,667 SNPs.

We searched the regions surrounding any significant GWAS SNPs using the Genome Database for Rosaceae GBrowse tool for *Malus* x. *domestica* v1.0 p genome (Jung et al., 2014). We used a window of +/- 5,000 bp (10 kb) surrounding the significant SNP to check for genes, and when identified, we used the basic local alignment search tool (BLAST) from NCBI to search for the mRNA sequence and reported the result with the max score (Altschul et al., 1990).

Genomic prediction was performed using the 'x.val' function in the R package PopVar (Mohammadi et al., 2015). The rrBLUP model was selected and 5-fold (nFold=5) cross-validation was repeated 3 times (nFold.reps=3) with no further filtering (min.maf=0) from the set of 122,596 SNPs used for GWAS. All other default parameters were used. In addition to performing genomic prediction on the main 24 phenotypes examined in this study, we performed genomic prediction on all 40 PCs for EFDs and on the first 40 PCs for PH values. We also used the 'rnorm' function in R to generate 1,000 random phenotypes with a mean of 0 and a standard deviation of 1, and performed genomic prediction using these random phenotypes to obtain the range of genomic prediction accuracies one can expect at random. Lastly, we used genome-wide complex trait analysis (GCTA) v.1.26.0 which estimates the genetic relationships between individuals based on genome-wide SNPs and uses this information to calculate the variance explained by these SNPs. The ratio of additive genetic variation to phenotypic variance is used to calculate narrow-sense heritability (h^2), or SNP heritability, of a trait (Yang et al., 2011). We used GCTA to estimate heritability for each phenotype, including the first 40 PCs for EFD and PH. We also estimated the correlation between genomic prediction accuracy (r) and narrow-sense heritability (h^2) using a Pearson's correlation.

Results

Variation in apple leaf shape

We examined 24 phenotypes related to apple leaf shape and size including length, width, surface area, dry weight, leaf mass per area, within-tree variance, and overall shape estimated using PCs derived from EFD (elliptical Fourier descriptor) and PH (persistent homology) data (Figure 2-1, Figure 2-2).

To visualize the primary axes of morphometric variation, we chose a representative leaf from accessions with the minimum and maximum values along the first 5 PCs for EFDs and PH (Figure 2-3a). The accessions with extreme values along PC1 for both methods are similar. In fact, 'Binet Rouge' has the lowest value along PC1 for EFD and PH, with the axis clearly representing a decrease in the length-to-width (aspect) ratio. The annulus kernels most strongly contributing to PH PC1 (Figure 2-4) provide further evidence that this PC captures variation in aspect ratio. Variation in leaf shape captured by higher-order PCs is more complex and cryptic, and is thus not captured using linear measurements alone. In addition, while the primary axis of variation (PC1) using EFDs and PH may explain similar aspects of leaf morphology, the morphospaces resulting from the two techniques differ (Figure 2-3b).

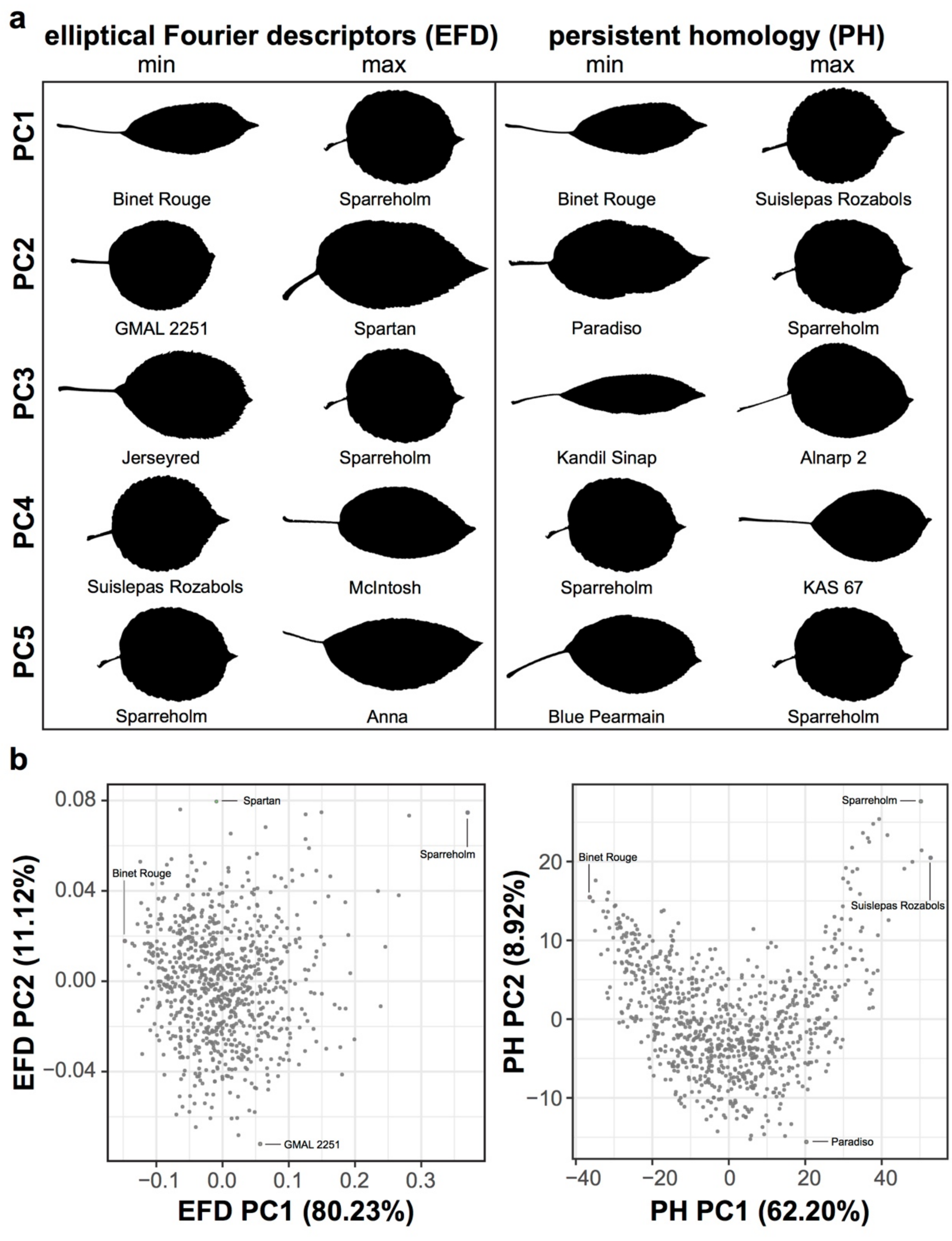


Figure 2-3. Examples of leaf shape across PCs derived from EFDs and PH. Binary images of leaves from accessions with minimum and maximum values along PCs 1 to 5 for EFD and PH estimates. PCs were calculated using values estimated as the average across 8-10 leaves but only a single representative leaf is displayed. PCs were REML-adjusted based on tree position in the orchard. The accession name is also listed (a). Visualization of PC1 vs PC2 for EFD and PH data. Accession with minimum and maximum values along PC1 and PC2 are indicated (b).

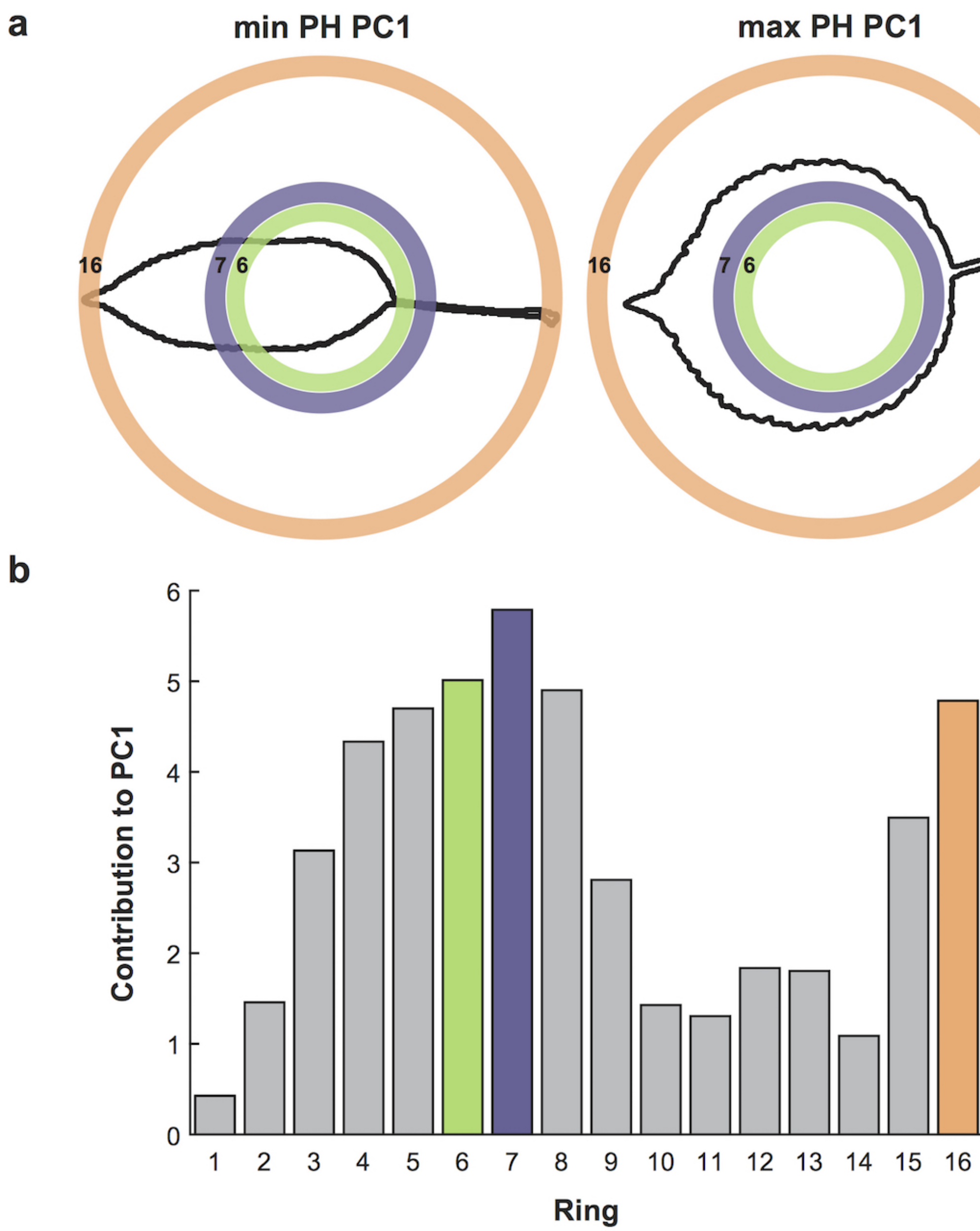


Figure 2-4. Visualization of contributions of each ring to PH PC1. Rings 6, 7 and 16 contribute the most to leaf shape according to PH PC1. The placement of each ring is visualized on a leaf representing the minimum and maximum value along PC1 (a). The contribution to PC1 of each of the 16 rings is also shown (b).

Next, we examined the correlation between all measured traits. By assessing the correlation of PCs resulting from a classical morphometric technique such as EFDs with a novel, topology-based morphometric approach like PH, we reveal how complementary the methods are (Figure 2-5). While there is a highly significant correlation between PC1 for both methods ($R^2 = 0.949$, $p < 1 \times 10^{-15}$), later PCs are often not significantly correlated, with the most notable exception being EFD PC2 and PH PC3 ($R^2 = 0.432$, $p < 1 \times 10^{-15}$), although several other PCs also show weak correlations. Thus, while the primary axis of variation (PC1) is consistent and highly correlated between methods, each method captures distinct aspects of leaf morphology in subsequent PCs.

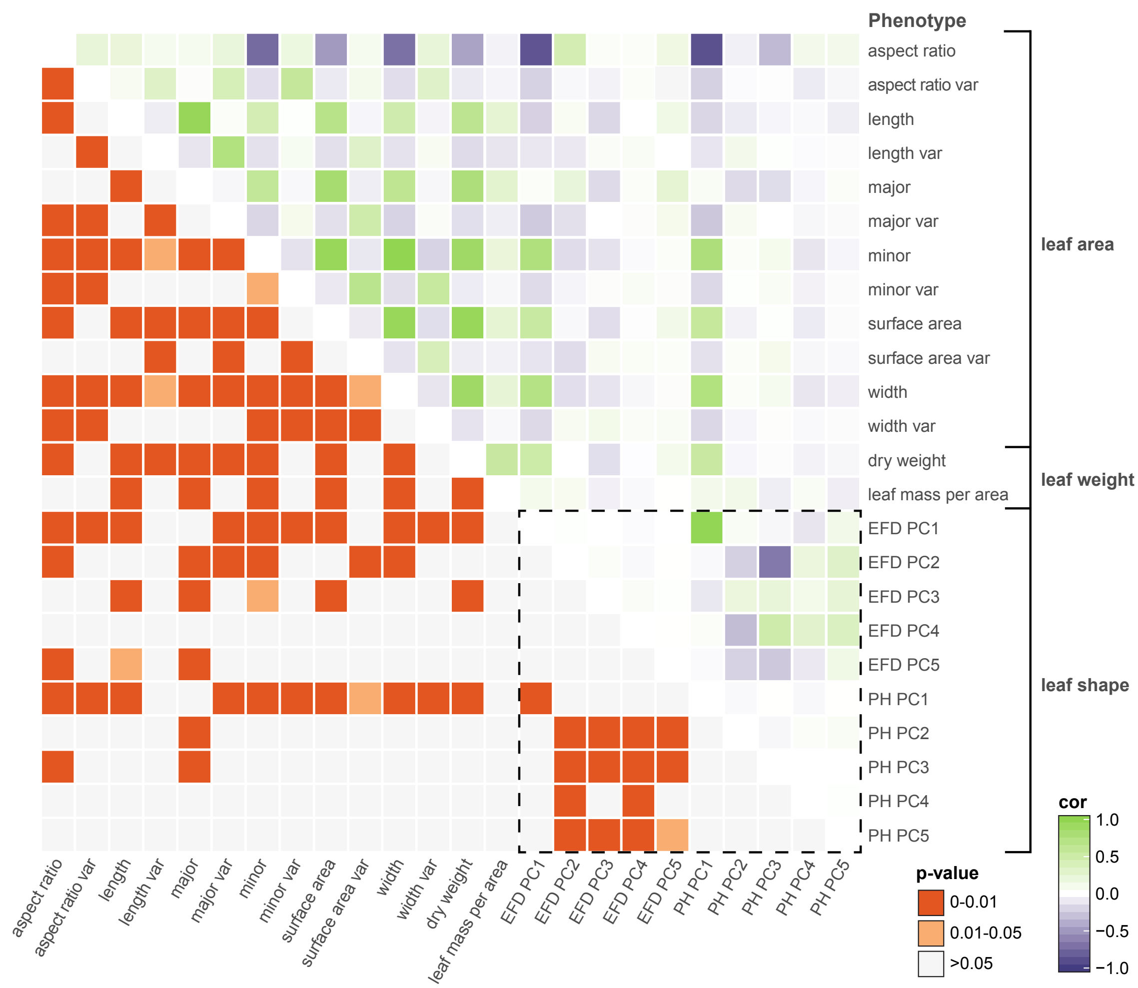


Figure 2-5. Correlations among leaf phenotypes. Values above the diagonal are colored according to the Pearson's correlation coefficient, and those below the diagonal indicate Bonferroni-corrected p-values. The box enclosed by the dotted lines include comparisons only between phenotypes captured by comprehensive morphometric analyses.

Many of the leaf phenotypes show a strong correlation with each other (Figure 2-5). In particular, aspect ratio is highly correlated with PH PC1 ($r = -0.878$, $p < 1 \times 10^{-15}$), EFD PC1 ($r = -0.855$, $p < 1 \times 10^{-15}$) and minor axis (leaf blade width) ($r = -0.734$, $p < 1 \times 10^{-}$

^{15}). The correlation between the minor axis of a leaf and surface area ($r = 0.939$, $p < 1 \times 10^{-15}$) is higher than the correlation between the major axis (blade length) and surface area ($r = 0.810$, $p < 1 \times 10^{-15}$). As expected, leaf surface area is also highly correlated with average leaf dry weight ($r = 0.934$, $p < 1 \times 10^{-15}$), indicating that larger leaves are heavier.

Allometry in apple leaves

The high correlation between aspect ratio and PC1 for both EFD and PH methods indicates that length-to-width ratio is the primary source of variation in apple leaf shape. If there is an allometric relationship between the minor and major axis, and thus, the length and width of a leaf do not increase at equal rates, a slope significantly differing from 1 is expected. We find that the slope between the two measurements is significantly greater than 1 (95% CI = 1.506-1.678, $R^2 = 0.343$, $p < 1 \times 10^{-15}$), indicating that the minor axis increases at a greater rate than the major axis. While there is no significant correlation between the major axis (blade length) and EFD PC1 ($R^2 = 0.001$, $p = 1$) or PH PC1 ($R^2 = 0.002$, $p = 1$), there is a significant correlation for the minor axis (blade width) and EFD PC1 ($R^2 = 0.541$, $p < 1 \times 10^{-15}$) and PH PC1 ($R^2 = 0.573$, $p < 1 \times 10^{-15}$) (Figure 2-6). As PC1 explains 80.23% of the variation in the leaf shape for EFDs, and 62.20% for PH, it is apparent that the width of the leaf blade, and not length, is the major source of leaf shape variation in apple. In fact, the aspect ratio, calculated as the ratio of major axis to minor axis, is even more strongly correlated with EFD and PH PC1, with an R^2 of 0.732 for EFD PC1 ($p < 1 \times 10^{-15}$) and R^2 of 0.771 for PH PC1 ($p < 1 \times 10^{-15}$). Given the significant correlation between EFD PC1 and PH PC1, it is not surprising that aspect ratio is highly correlated with both.

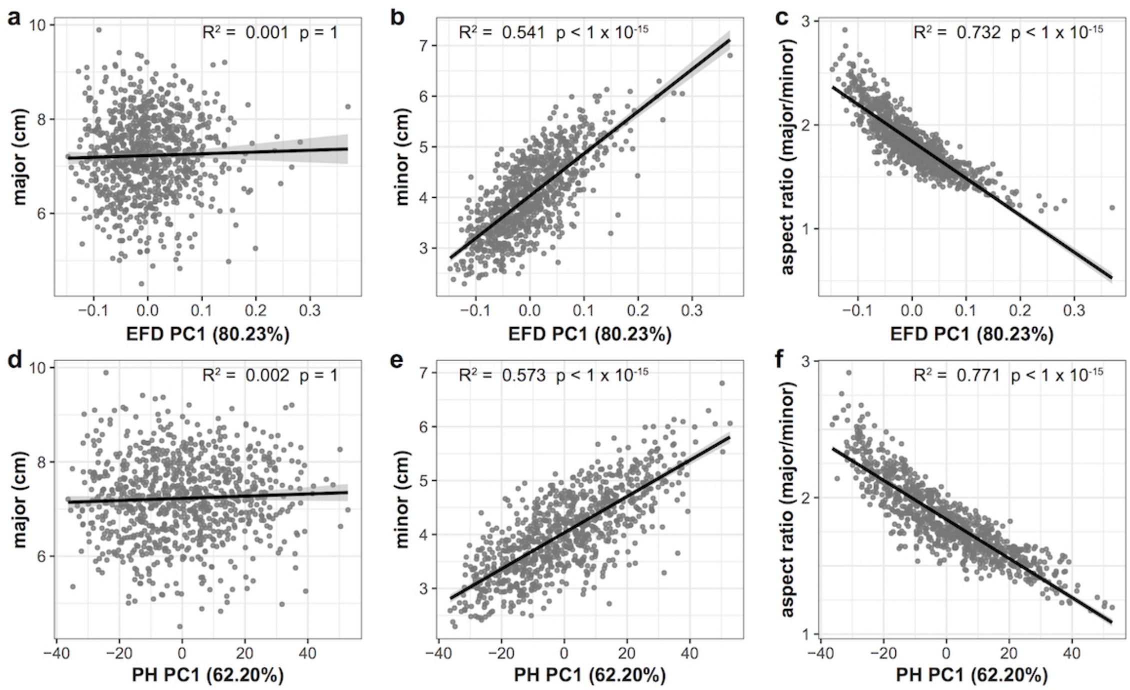


Figure 2-6. Correlation between the primary axis of variation (PC1) captured using EFD and PH values and leaf shape measures. The EFD PC1 is plotted against the major axis (length of leaf blade) (a), minor axis (width of leaf blade) (b) and aspect ratio (ratio of length-to-width of blade) (c). The PH PC1 is plotted against the same measures in panels d-f. The percent variances explained by PC1, prior to REML-adjustment, is shown in parentheses. All p-values are Bonferonni-corrected based on the number of comparisons in Figure 2-5. A regression line from a linear model with a shaded 95% confidence interval is also shown.

In addition to variation between accessions, we investigated differences in leaf shape and size between species by comparing *Malus domestica*, the domesticated apple, with its primary progenitor species, *Malus sieversii* (Appendix I: Table I-I). PCA of the genome-wide SNP data reveals a primary axis of genetic variation that separates *M. domestica* and *M. sieversii,* although separation is incomplete (Figure 2-7a). The major axis (p = 0.975) of the leaves does not differ between species (Figure 2-7b). However, the minor axis ($p = 4 \times 10^{-4}$) of *M. domestica* leaves are significantly larger than *M. sieversii* (Figure 2-6c) and the aspect ratio (p = 0.023) is significantly less (Figure 2-7d). Thus, there is allometric variation both within (Figure 2-6) and between (Figure 2-7) *Malus* species.

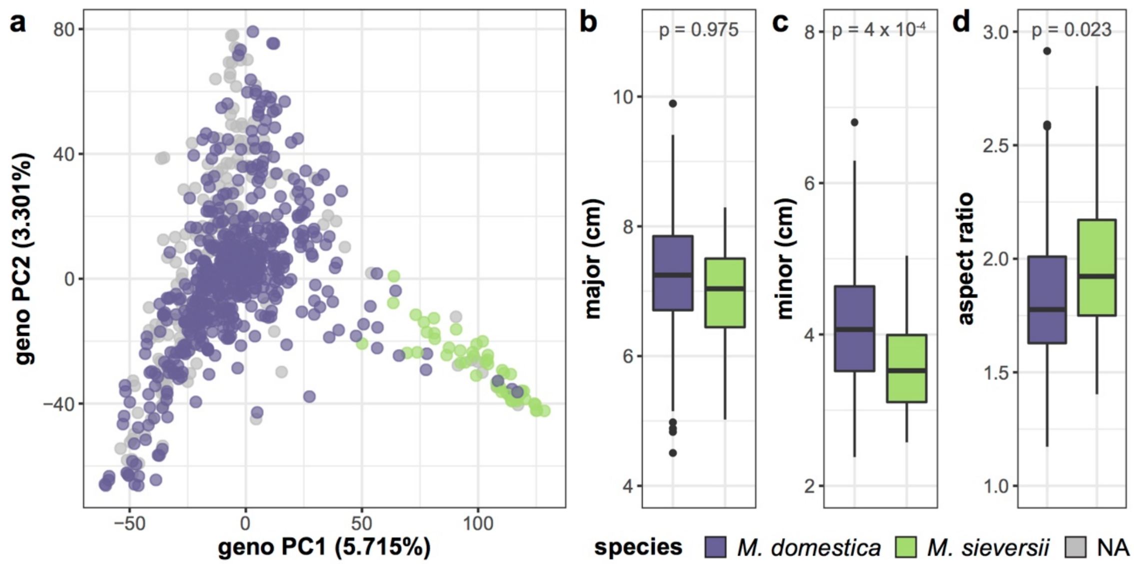


Figure 2-7. Genetic and phenotypic comparison of the domesticated apple and its wild ancestor. PCs 1 and 2 were derived from 75,973 genome-wide SNPs and samples are labeled as *M. domestica* (purple), *M. sieversii* (green) or unknown (gray). *M. domestica* leaves do not differ from *M. sieversii* leaves along the major axis (b), but they have a larger minor axis (c) and aspect ratio (d). P-values reported are Bonferroni-corrected based on multiple comparisons (Appendix I: Table I-I). Species labels are based on USDA classification.

The genetic basis of leaf shape in apple

GWAS of the 24 leaf phenotypes examined in this study yielded few significant results (Appendix I: Figure I-I). We identified 70 significant SNPs representing 5 phenotypes which are reported in Appendix I: Table-II. We examined the regions surrounding significant SNPs for candidate genes using the GBrowse tool (Appendix I: Table-III) (Jung et al., 2014). We searched within a +/- 5,000 bp window, which should capture any linked causal variation given the rapid LD decay observed in a diverse collection of apples that is largely replicated in the germplasm studied here (Migicovsky et al., 2016a). However, no strong candidate genes were identified.

While GWAS examines the genome for single, large-effect loci, genomic prediction estimates our ability to predict a phenotype using genome-wide marker data. We complimented our GWAS with genomic prediction and observed prediction accuracies (r) ranging from -0.10 to 0.52 (Figure 2-8). Aspect ratio is the primary source of variation in leaf shape (Figure 2-5c) and it is also the leaf measurement that had the highest genomic prediction accuracy (0.52). Other phenotypes highly correlated with aspect ratio, such as leaf width (0.51), minor axis (0.49), EFD PC1 (0.48) and PH PC1 (0.47), all had relatively high prediction accuracies. PH PC3 (0.51) was also among the most well-predicted using genetic data.

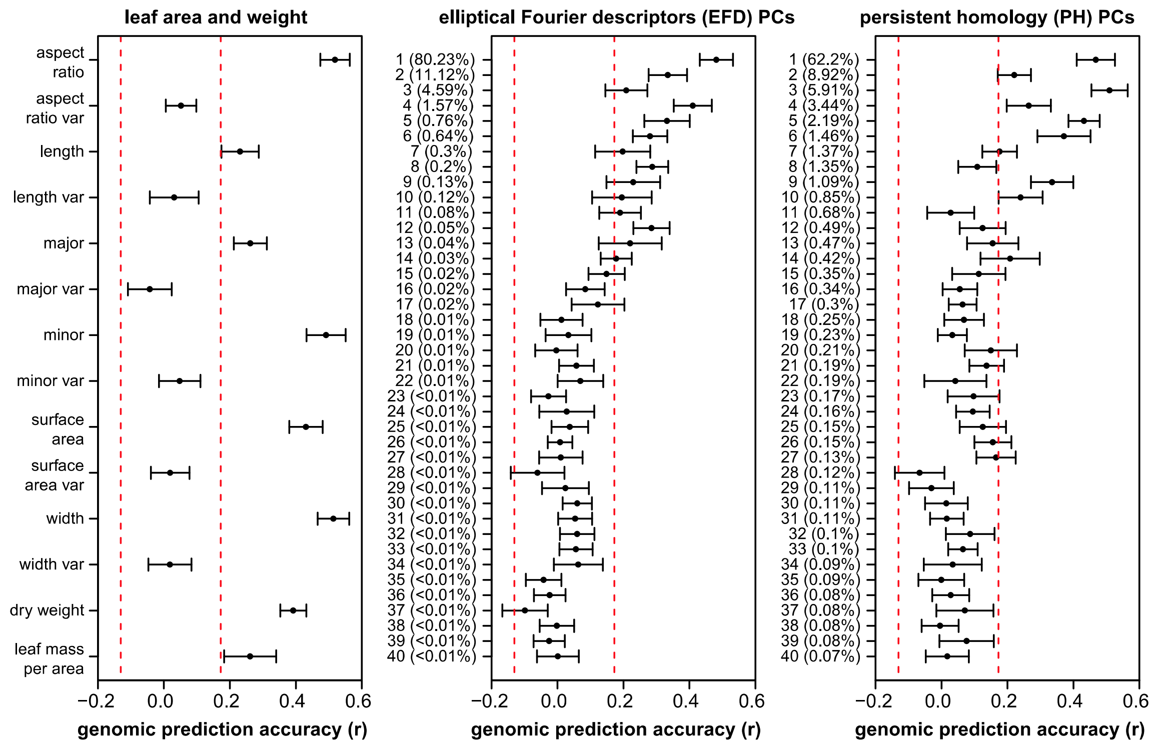


Figure 2-8. Genomic prediction accuracy (*r*). Values represent the average correlation (+/- standard deviation) between observed and predicted phenotype scores, based on 5-fold cross-validation with 3 iterations. Dotted red lines indicate the minimum and maximum prediction average accuracy (*r*) achieved using 1,000 randomly generated phenotypes. The percent variance explained by each PC was calculated prior to REML-adjusted and is indicated in parenthesis.

Similarly, estimates of narrow-sense heritability (h^2) calculated using GCTA (Yang et al., 2011) ranged from 0 to 0.75, with the highest heritability observed for aspect ratio (0.75) followed by leaf width (0.71), EFD PC1 (0.71), minor axis (0.69) and PH PC1 (0.65) (Figure 2-9). Heritability estimates were highly correlated with genomic prediction accuracies (Figure 2-10, $R^2 = 0.936$, $p < 1 \times 10^{-15}$), which is not surprising given that both techniques involve predicting a phenotype from genome-wide SNP data. None of the phenotypes measuring variance within the 8-10 leaves sampled had heritability estimates significantly different from 0.

While the principal component of variation in leaf shape detected by EFDs and PH is aspect ratio, we were also interested in determining if higher-order PCs, which capture variation not readily visible to the eye, are extracting information that is biologically meaningful. Using genomic prediction and heritability estimates, we found evidence of a genetic basis for these "hidden phenotypes", which are unmeasurable using linear techniques. For example, the heritability of phenotypes such as PH PC6 (0.48), PH PC9 (0.35), PH PC10 (0.33) and EFD PC9 (0.33) are similar to traditionally measured phenotypes such as leaf length (0.44) and leaf mass per area (0.40). While higher PCs may have relatively high heritability values, after a certain point the values (+/- standard error) overlap with 0, indicating that they are not heritable. The cutoff for morphometric PCs with a heritable genetic basis is approximately PC17. These results suggest that by making use of morphometric techniques that measure shape comprehensively, we are describing biologically meaningful, heritable phenotypes which would be missed by simple measurements such as leaf length, width and surface area.

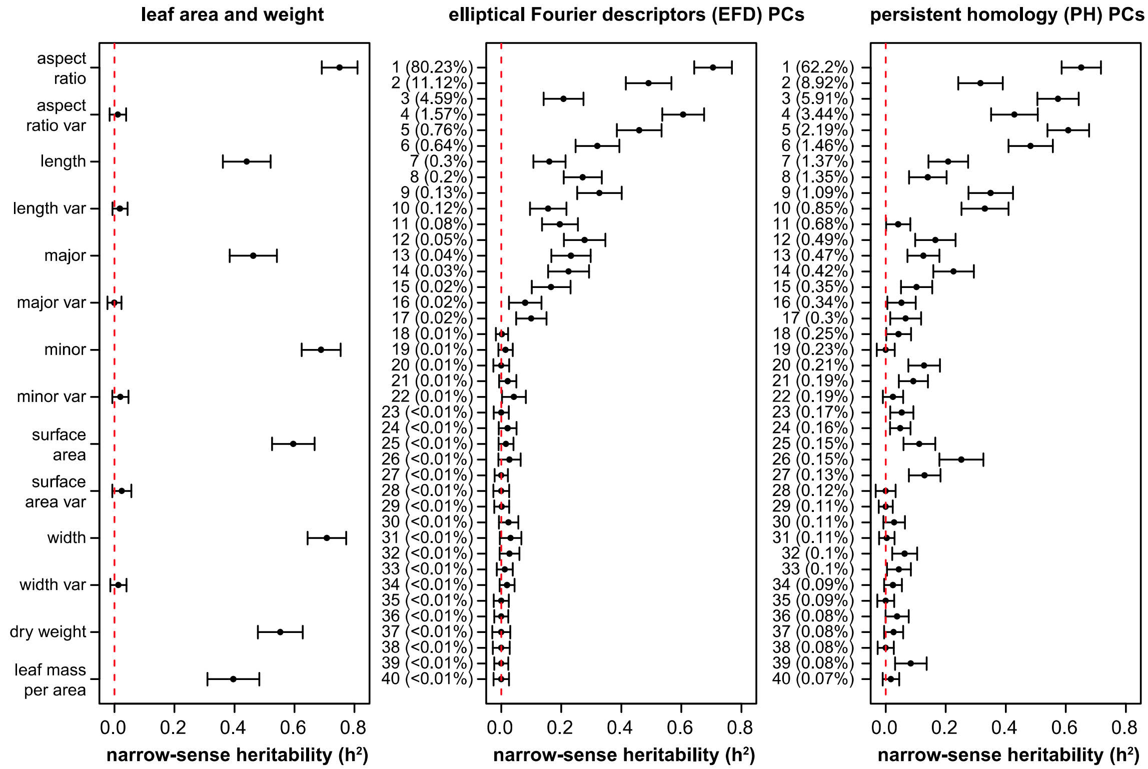


Figure 2-9. Narrow-sense heritability (h^2) for leaf phenotypes. Values represent the additive genetic variance divided by the phenotypic variance (+/- standard error), as calculated using GCTA. Dotted red lines indicate $h^2 = 0$, at which point the phenotypic variation is not heritable. The percent variance explained by each PC was calculated prior to REML-adjusted and is indicated in parenthesis.

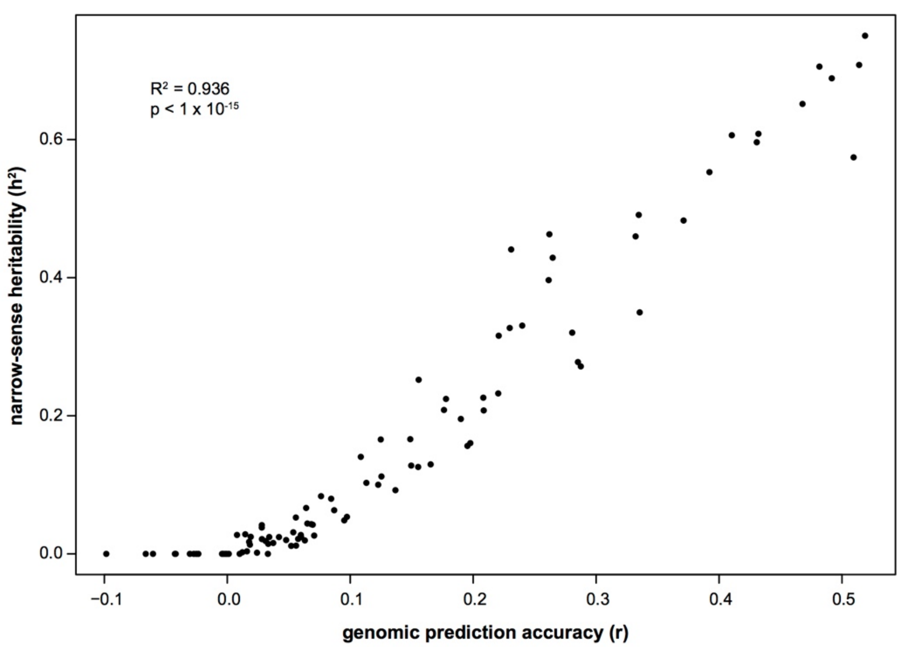


Figure 2-10. Correlation between genomic prediction accuracy (r) and narrow-sense heritability estimates (h^2) for all leaf phenotypes.

Discussion

Leaf shape and size play a crucial role in the growth and development of apple trees, including the fruit. To elucidate the genetic basis of this variation, we quantified leaf shape in apple using traditional linear measurements and comprehensive morphometric techniques. Our work offers the first comparison between the novel topology-based technique, PH, and EFDs, which we find are complementary but distinct methods. For both methods, PC1 was highly correlated with the aspect ratio, thus providing evidence that the primary axis of variation in apple leaf shape can be captured using linear measurements. The minor axis, or width of the leaf blade, was also highly correlated with PC1, while the major axis was not. Thus, variation in the aspect ratio is due to variation in the leaf blade width, not length. Leaf surface area was also more highly correlated with the minor axis than the major axis. Variation in leaf width is therefore essential to both the size and shape of apple leaves, similar to previous work in tomato (Schwarz and Kläring, 2001).

The width of the leaf blade is not only the source of variation between apple accessions, but also between *M. domestica* and *M. sieversii.* The presence of the same allometric relationship within and between species suggests that the genetic loci controlling intra-specific leaf shape variation within *M. domestica* may be the same as those controlling the divergence in leaf shape observed between the domesticated apple and its wild ancestor. For example, in birds, while PC1 and PC2 of bill shape explain the majority of variation across 2,000 species, they are also consistently associated with the variation between higher taxa (possessing >20 species) (Cooney et al., 2017). Our results suggest that the increase in leaf size since domestication has not been an overall increase in leaf size but specifically an increase in blade width leading to larger leaves with a reduced length-to-width ratio.

Our work provides evidence that allometry is the primary source of morphometric variation in apple leaves. These findings are consistent with work reported in other species such as tomato, where the length-to-width ratio was the major source of shape variation (>40%) (Chitwood et al., 2013). Similarly, work in *Passiflora* and *Vitis* species performed using two independent morphometric techniques identified allometric variation as the primary source of variation in PC1, which explained at least 40% of the variation in leaf shape (Chitwood and Otoni, 2017; Klein et al., 2017). Thus, linear measurements—in particular aspect ratio—are an important source of information when describing leaf shape. However, linear measurements are not sufficient for capturing the full spectrum of diversity. In our study, PC1 accounts for 62.20% or 80.23% of the variation, depending on the technique used. By simply quantifying apple leaves using linear measurements, we would miss nearly 40% of the variation in some cases. While PC1 is highly correlated with aspect ratio, later PCs represent orthogonal variation that can likely only be captured through morphometric techniques such as EFDs and PH. To fully quantify variation in leaf shape, comprehensive morphometric techniques are therefore essential.

To discern the genetic contributions to leaf shape, we paired both linear and comprehensive morphometric estimates of shape with genome-wide SNP data. There are examples of a simple genetic basis of leaf shape, such as in *Arabidopsis thaliana*, where the *ANGUSTIFOLIA* and *ROTUNDIFOLIA3* independently control leaf width and length (Tsuge et al., 1996). In barley, transcript levels of *BFL1* limit leaf width, with overexpression resulting in narrower leaves and loss of *BFL1* function resulting in a reduced length-to-width ratio (Jöst et al., 2016). Using GWAS, we found no robust associations with shape phenotypes, observed a low ratio of significant SNPs to the number of phenotypes examined, and found that significant SNPs were sparsely distributed across multiple chromosomes. In addition, the small number of significant SNPs are likely spurious associations due to poor correction for cryptic relatedness, as evidenced by the QQ plots (Appendix I: Figure I-I). These observations suggest that leaf shape is likely polygenic and controlled by a large number of small effect loci, such as in tomato and maize (Tian et al., 2011; Chitwood et al., 2013). In comparison, GWAS on apple fruit phenotypes, such as color and firmness, have revealed strong associations resulting from a small number of large effect loci (Migicovsky et al., 2016a). However, it is possible that large effect loci were missed in the present study, either because of poor reference genome assembly or inadequate marker density. Improvements in genome assembly and increases in marker number will aid to further reveal the genetic architecture of apple leaf shape variation.

Lastly, we investigated the degree to which leaf shape is heritable and can be predicted using genome-wide SNP data. We find that the genomic prediction accuracies of the primary axes of leaf shape variation are similar to previously reported estimates for fruit width (0.48) and length (0.47), indicating that leaf shape is as heritable as fruit shape (Migicovsky et al., 2016a). In combination with few significant GWAS results, high narrow-sense heritability estimates support a polygenic basis for leaf shape. Aspect ratio was identified as the primary source of variation in leaf shape in apple and had the highest genomic prediction and heritability estimates, indicating that there is a genetic, heritable basis for allometric variation in apple. Further, although the first 5 PCs for both EFDs and PH explain the majority of the variation in apple leaf shape, most PCs from 1 to 14 have heritability estimates above 0.20 and may still represent crucial differences in leaf shape from an ecological, evolutionary, or agricultural perspective. Thus, while our ability to detect the primary axes of variation in leaf shape using genome-wide data is expected, our observation that higher level PCs are also heritable confirms that these comprehensive morphometric methods capture biologically meaningful variation that would be missed by linear measurements alone.

Conclusions

It is clear from our work that variation in apple leaf shape and size are under genetic control. Further, high genomic prediction and heritability estimates for higher morphometric PCs indicate that techniques such as EFDs and PH are capturing heritable biological variation that will be missed if researchers restrict leaf shape estimates to linear measurements. Based on these results, it may be possible to perform genomic selection for a phenotype that could only be detected using morphometrics. If a higher order PC was correlated with a trait that was difficult or expensive to measure, assessing leaf shape could potentially be used as proxy for that phenotype, in the same manner that red leaf color can be used to select for red fruit flesh color in apples (Chagne et al., 2007; Espley et al., 2009). Additionally, a better understanding of the variation in leaf shape and size in apple could ultimately have important implications for canopy management, where light exposure is crucial to flowering (Dennis, 2003). Ultimately, through the first in-depth study of leaf shape in apple, we uncover allometry between accessions and species, as well as evidence that complex and heritable phenotypes can be captured using comprehensive morphometric techniques.

Acknowledgments

We would like to acknowledge Gavin Douglas and Sherry Fillmore for their help setting up the statistical design of the orchard and SNP-calling pipeline. We also thank all past and present members of the Myles Lab for their work in maintaining the apple orchard. This article was written, in part, thanks to funding from the Canada Research Chairs program, the National Sciences and Engineering Research Council of Canada and Genome Canada. Z.M. was supported in part by a Killam Predoctoral Scholarship from Dalhousie University.

Chapter 3: Genome to phenome mapping in apple using historical data

Introduction

To meet the needs of a growing global population, food availability must double within 25 years (McCouch et al., 2013). Fortunately, advances in genomics allow breeders to more accurately and quickly improve crops (Lusser et al., 2012). However, continued food improvement relies on increasing our understanding of the genome-phenome relationship (Morrell et al., 2011; Varshney et al., 2014). Association mapping can detect causal genes of interest using phenotyped populations of unrelated individuals, such as those already available in germplasm collections. Once genetic markers linked to important traits are discovered, marker-assisted breeding can enable more efficient selection for plants with these desirable characteristics (Morrell et al., 2011; McCouch et al., 2013).

Apple (*Malus* X. *domestica* Borkh.) had the third highest global gross production value among fruit crops in 2013 and is well-poised to benefit from marker-assisted selection (MAS) that would eliminate undesirable genotypes at the early seedling stage (Myles, 2013; McClure et al., 2014; Food and Agriculture Organization of the United Nations, 2015). Apples have a long juvenile period: significant fruiting generally occurs 5 years after germination, even when using a quickly maturing dwarf rootstock (Kumar et al., 2012a). Two to three additional years may be required to phenotype fruit quality traits before selecting parents for crosses, and a large percentage of offspring are discarded within the first decade of fruit evaluation (Kumar et al., 2013b; Myles, 2013). Apple breeding is also limited by self-incompatibility and a large tree size that requires substantial space and money to breed (Brown and Maloney, 2003; Myles, 2013; McClure et al., 2014). As a result, one recent breeding program required 26 years to generate 3 commercial cultivars from a starting population of 52,000 seedlings (Peil et al., 2008).

Markers linked to numerous phenotypes have been discovered in apple and MAS is already being applied for traits such as disease resistance, postharvest storability, skin color and fruit texture as well as dwarfing and precocity in rootstocks (Fazio et al., 2014; Ru et al., 2015). In one example, using MAS to select for a single marker linked to postharvest storability resulted in an estimated savings of at least 60% of the operating costs associated with first-stage seedling selection (Edge-Garza et al., 2010).

A major barrier to establishing robust genotype-phenotype relationships that can be leveraged for MAS is poor quality phenotype data (Benfey and Mitchell-Olds, 2008; Cobb et al., 2013; Meneses and Orellana, 2013). While technological advances continue to increase the speed and decrease the cost of acquiring genetic data, slow and expensive phenotyping results in a "phenotyping bottleneck" (Houle et al., 2010; Kumar et al., 2012a; Burleigh et al., 2013). It is well known that high quality phenotype data often results in far better powered quantitative trait locus (QTL) analyses (Van Eerdewegh et al., 2002). Fortunately, improvements to phenotyping technology have begun and the scientific community generally recognizes the need for high quality phenotypic measures (Houle et al., 2010; Furbank and Tester, 2011; Meneses and Orellana, 2013; Deans et al., 2015).

There has been great support for the use of historical phenotype data from gene banks for genetic mapping. However, phenotypic evaluation is especially challenging and costly over long time periods, and using data not collected specifically for genetic mapping is often problematic (Myles et al., 2009; Houle et al., 2010; McCouch et al., 2012). Different observers measuring traits over multiple years in varying environments cause phenotyping discrepancies in historical data sets. While DNA sequences are comparable between studies, phenotype data are much more difficult to compare due to missing data, inconsistent replication, and the frequent use of non-quantitative measurements (Peace and Norelli, 2009; Houle et al., 2010; McCouch et al., 2012).

Despite the difficulty of acquiring reliable data and the subsequent need for curation, historical phenotype data has been successfully employed to identify genotype-phenotype relationships in barley (*Hordeum vulgare* L.) and potato (*Solanum tuberosum* L.) (Baldwin et al., 2011; Matthies et al., 2014). Here we examine historical phenotype data available from a large apple gene bank, the United States Department of Agriculture (USDA) apple germplasm collection, and link them to genotypes collected using genotyping-by-sequencing (GBS) (Elshire et al., 2011). We find relationships of interest between phenotypes, identify several genotype-phenotype associations using genome-wide association study (GWAS), describe the very rapid linkage disequilibrium (LD) decay in the domesticated apple and quantify our ability to predict phenotypes using genomic prediction.

Materials and Methods

Phenotype scoring and filtering

Publicly available phenotype data were downloaded from the USDA- Germplasm Resources Information Network (GRIN) website (http://www.ars-grin.gov/cgi-bin/npgs/html/crop.pl?115) on July 18th, 2011. A description of the steps to edit and curate the phenotype data is provided in Figure 3-1. Phenotype data were first trimmed to exclude accessions not labeled as *Malus domestica*, as well as outliers that were clearly mislabeled and did not fall within the *M. domestica* variation observed according to a principal components analysis (PCA) of the genetic data. We manually curated and recoded phenotypes when phenotype scoring was incompatible with downstream applications. In several cases, such as recoding color as a binary trait, data points were removed as a result. Using the genetic data, we determined which accessions exhibited clonal relationships and measurements across clones were averaged (see "Genetic analysis" below). Phenotypes were also combined across years and averaged in cases of replication for a particular accession. Categorical phenotypes were excluded from analysis. We removed phenotypes containing data for fewer than 100 accessions, as well as invariable phenotypes scored entirely as one value. Binary phenotypes with highly uneven distributions of trait scores (i.e. one of the two values was present at a frequency > 95%) were also excluded. The final phenotype data set included binary, ordinal and quantitative phenotypes.

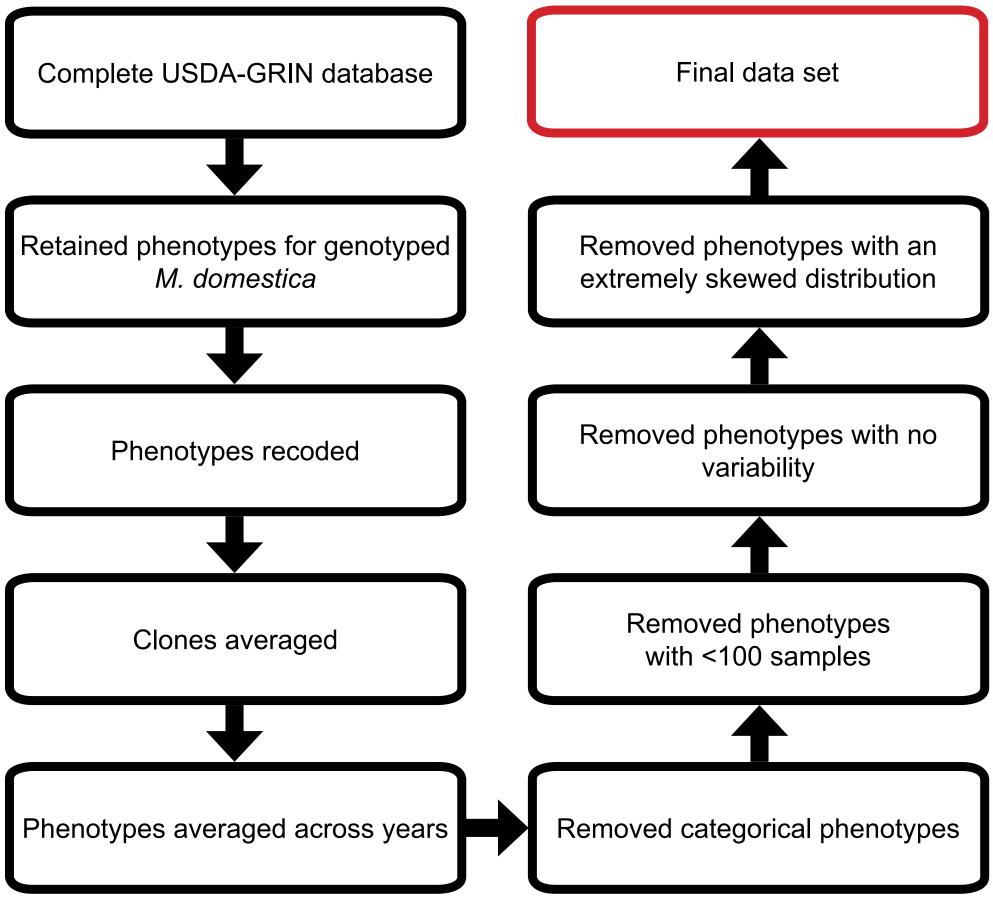


Figure 3-1. Flowchart of processing for phenotype data.

Associations between phenotypes were tested using Pearson's correlation for binary-binary, quantitative-quantitative, and binary-quantitative comparisons. Spearman's rank correlation was used for binary-ordinal and quantitative-ordinal comparisons, while Kendall's rank correlation was used for ordinal-ordinal comparisons. Performing correlations between every possible pair of phenotypes generated a pairwise correlation matrix. To correct for multiple comparisons, a Bonferroni correction was applied by multiplying *p*-values by the number of pairwise comparisons (630).

We divided accessions into several binary categories including harvest season (early and late), color (red and green/yellow), use (cider and eating/cooking) and origin (New World and Old World) using information from the USDA-GRIN database when possible and online sources otherwise. We tested whether phenotypes showed differences according to these categories using a Fisher's Exact Test for binary phenotypes and a Mann-Whitney U test for ordinal and quantitative phenotypes. For Fisher's Exact test we report the Odds-Ratios (OR) and for the Mann-Whitney *U* we report the W test statistic. *P*-values were Bonferroni-corrected for multiple comparisons. All analyses were performed in R (R Core Team, 2015).

Genetic analysis

Genotypes from the *M. domestica* evaluated here were generated using genotyping-by-sequencing (GBS) described in Gardner et al. (unpublished data, 2016). Single nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) <0.01 were excluded. Accessions with >30% missing data were excluded and SNPs with <20% missing data were retained. The resulting genotype matrix contained 8657 SNPs and 929 accessions and 9.3% missing data. Missing genotypes were imputed using LinkImpute (Money et al., 2015). For imputation, we used values of 5 and 20 for parameters *l* and *k*, which resulted in an estimated genotype imputation accuracy of 92%.

To determine if two or more accessions were clonally related, we calculated identity-by-descent (IBD) using PLINK (Purcell et al., 2007; Purcell, 2009a). When two or more accessions had IBD ($\hat{\pi}$)>0.9, one accession from the clonal group was randomly chosen and its genotype data were retained while the genotype data from the other clones were removed. The remaining dataset contained 840 accessions, including 689 with phenotype data.

For principal components analysis (PCA), SNPs were pruned for LD using PLINK by considering a window of 10 SNPs, removing one SNP from a pair if LD was >0.5 then shifting the window by 3 SNPs and repeating the procedure (PLINK command: indep-pairwise 10 3 0.5). After removing SNPs with MAF <0.05, 4395 SNPs and 672 accessions remained for PCA.

We calculated LD decay using PLINK and used only SNPs with MAF >0.05. The apple reference genome contains numerous large gaps of unknown sequence of varying length represented by long series of 'N's (Velasco et al., 2010b). To avoid bias in our LD decay measures, we discarded LD estimates generated from SNP pairs separated by a gap >10,000 'N's. There are 3,590 gaps >10kb in the apple reference genome v1.0 (Genome Database for Rosaceae, GDR available at www.rosaceae.org) (Jung et al., 2014).

A genome-wide association study (GWAS) was performed using EMMAX (Kang et al., 2010). The k matrix was generated in EMMAX (command: emmax-kin –v –h –s –d 10) and we corrected for relatedness using the k matrix without any additional covariates. We used the GBrowse tool (GDR) (Jung et al., 2014) for *Malus* X. *domestica* v1.0p to check for potential genes of interest near GWAS hits that passed the Bonferroni corrected threshold for significance ($p < 0.05$). We examined the distribution of phenotype data for the most significant GWAS SNPs and represented them using the tableplot package in R (Kwan and Friendly, 2012). We also tested which model of inheritance fit best based on the single most significant SNP for overcolor intensity using SNPStats (Sole et al., 2006).

A significant GWAS result for firmness and harvest time was identified in a NAC protein and we aligned NAC proteins from various plant species using ClustalW (Larkin et al., 2007). A phylogenetic tree was built using MEGA6 with the Dayhoff model (Dayhoff et al., 1978) and neighbor joining method. We used a pairwise deletion option for dealing with gaps and a consensus of 1000 bootstrap replicates (Tamura et al., 2013).

Genomic prediction was performed using the x.val function in the R package PopVar (Mohammadi et al., 2015). The rrBLUP model was selected and prediction accuracy was assessed using a 5-fold (nFold=5) cross-validation procedure which masked 20% of the samples' phenotypes and then predicted them using a model generated from the other 80% of the samples' data. All other default parameters were used. Genomic prediction accuracy was calculated as the correlation between the predicted phenotypes and the observed values.

Results and Discussion

Historical data curation

We downloaded 121,950 phenotypic observations for the genus *Malus* from the USDA-GRIN database. These observations came from 105 different phenotypes measured in 4123 accessions spanning 15 years. For the purposes of GWAS and genomic prediction, only accessions with shared segregating polymorphism are useful. We therefore restricted the data to members of the domesticated apple, *M. domestica,* which make up 32% (1339) of the accessions in the database. After filtering (Figure 3-1), the resulting dataset contained 24,778 measurements from 36 phenotypes across 689 different accessions, which represents approximately 20% of the data available for the genus *Malus* in the USDA-GRIN database. It is worth noting that only 36 of the initial 105 phenotypes (34%) were deemed useful for downstream analyses. Most phenotypes were not measured in enough accessions, or were not measured in an appropriate manner, to be useful for genetic mapping and genomic prediction.

Figure 3-2A shows the frequency of phenotypes according to amount of data available. Although promising upon initial inspection, it does not account for the fact that data are often collected across multiple years. In only 26% of cases were there >100 data points for a given phenotype within a particular year. Sample sizes were often small and data collection was highly uneven across years, potentially due to available funding and access to resources (Figure 3-2B). Of the 36 phenotypes included, seven were measured in a single year while 24 were measured across 10 or more years (Figure 3-2C). The bimodal distribution in Figure 3-2C is the result of having a core set of phenotypes which were measured frequently while the remaining phenotypes were measured once. Even when a phenotype was measured across multiple years, the same trees were often not phenotyped each time so the data are highly unbalanced and corrections for year effects could not be implemented.

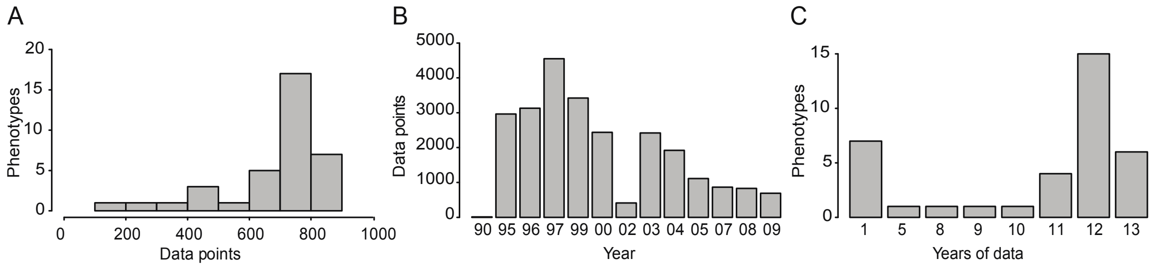


Figure 3-2. Description of phenotype data available from USDA-GRIN database for accessions belonging to the domesticated *M. domestica*. (A) Frequency of phenotypes according to number of data points available. (B) Number of data points by year. Year with no data available are not shown. (C) Number of years of data available for each phenotype; only values that apply to at least one phenotype are shown.

Both inconsistent data collection across years and small sample sizes within years make exploring genotype-phenotype associations using historical data challenging. When an accession was measured across multiple years for a phenotype, we used the mean phenotype score across years in our analyses. Extensive data curation was required to generate data sets that could be successfully linked to genotype data. Figure 3-1 provides a flow chart of how the data were handled from the initial database download to the final data set.

Correlations among phenotypes

Even without genotype data, assessing patterns within the phenotype data help assess data reliability while potentially exposing novel relationships worthy of further inquiry. We therefore investigated correlations between all pairs of the 36 phenotypes remaining after data curation (Figure 3-3). All p-values reported below are Bonferroni-corrected for multiple comparisons. The strongest correlations were between fruit length and width ($r = 0.850$, $p < 1 \times 10^{-15}$), fruit length and weight ($r = 0.517$, $p < 1 \times 10^{-15}$) and fruit weight and width ($r = 0.567$, $p < 1 \times 10^{-15}$). In these cases, an increase in one fruit size measurement is positively correlated with an increase in another, indicating that longer fruit are also heavier and wider.

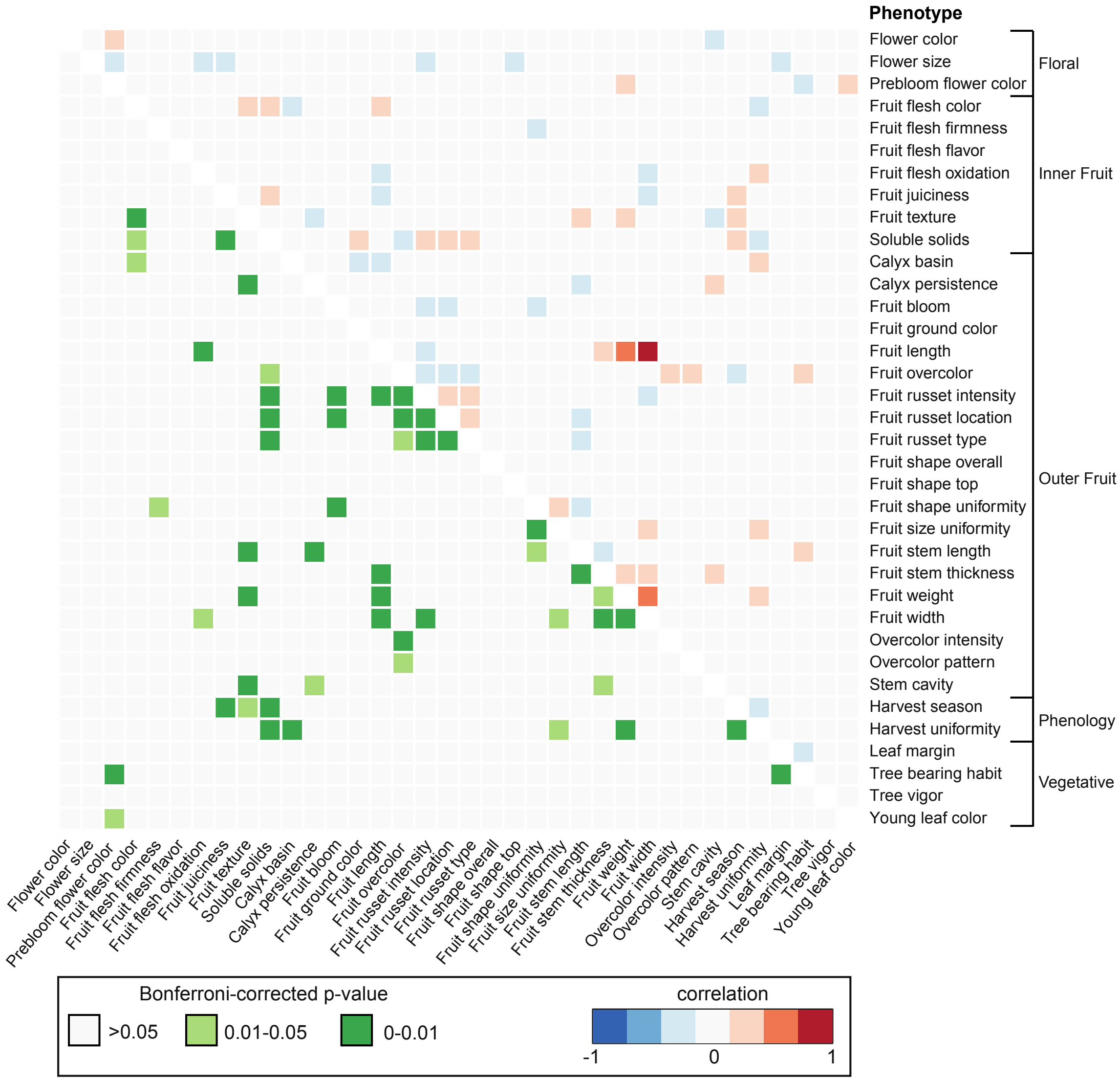


Figure 3-3. Correlations among apple phenotypes. Values above the diagonal are colored to indicate the correlation results (r) and those below the diagonal indicate Bonferroni-corrected p-values.

Three different measurements evaluating the amount and location of russeting (rough, brown skin) were taken: intensity was measured as the percent of fruit surface covered in russet (0-100%), location of russeting indicated which area of the fruit was russeted (either one end, both ends, or entire fruit), and fruit russet type was a binary trait describing the russeting as either extremely fine or medium heavy to cracked. All measurements of fruit russeting including intensity and location ($r = 0.370$, $p = 1.093 \times 10^{-15}$), intensity and type ($r = 0.378$, $p < 1 \times 10^{-15}$), and type and location ($r = 0.3396$, $p = 4.725 \times 10^{-13}$) were positively correlated with each other. Fruit overcolor was negatively correlated with all russeting measurements, indicating that green apples were more russeted than red apples (intensity: $r = -0.196$, $p = 3.792 \times 10^{-3}$; type: $r = -0.181$, $p = 0.0499$; location: $r = -0.22799$, $p = 3.402 \times 10^{-4}$). These correlations are all expected and provide confidence in the reliability of the phenotype data.

Thicker fruit stems were found to be shorter ($r = -0.207$, $p = 5.906 \times 10^{-5}$) and attached to heavier ($r = 0.164$, $p = 0.016$), longer ($r = 0.198$, $p = 2.1496 \times 10^{-4}$), wider fruits ($r = 0.249$, $p = 9.992 \times 10^{-8}$), potentially enabling larger, heavier apples to better remain attached to the tree.

Fruit uniformity in shape and size was scored as either uniform or variable by visually comparing 10 apples from the same cultivar. Fruit that were more uniform in shape were also more uniform in size ($r = 0.361$, $p < 1 \times 10^{-15}$). Uniformity facilitates the processing of commercial cultivars, thus the correlation between uniform size and shape may be due to selective breeding for this desirable trait (Brown and Maloney, 2003).

In agreement with previous work (Jan et al., 2012), varieties harvested late in the season tended to be juicier ($r = 0.210$, $p = 0.007$) and have higher soluble solids ($r = 0.359$, $p = 2.142 \times 10^{-18}$). Harvest time was also negatively correlated with harvest uniformity ($r = 0.348$, $p = 3.090 \times 10^{-10}$), indicating that apples harvested earlier in the season required fewer visits to the tree during harvest. Staff being occupied by other activities early in the growing season may partially account for this observation. Accessions with heavier apples required more visits during harvest ($r = 0.382$, $p = 9.792 \times 10^{-13}$) and tended to have a coarser fruit texture ($r = 0.192$, $p = 4.981 \times 10^{-4}$).

Cultivars with more russeting locations ($r = -0.205$, $p = 1.183 \times 10^{-3}$) or a higher intensity of russeting ($r = -0.202$, $p = 4.887 \times 10^{-4}$) tended to have lower natural bloom, or wax, on the fruit at maturity. In our study, wax was scored as simply present or absent and did not distinguish further based on cuticle properties. However, wax impacts the cuticle of the fruit, and as a result could play a role in the susceptibility to russeting, although this may be a complicated relationship (Khanal et al., 2013). In previous work, genes involved in cutin and wax synthesis were downregulated in russeted apple skin

(Legay et al., 2015). Increased russeting type ($r = 0.260$, $p = 7.505 \times 10^{-7}$), location ($r = 0.276$, $p = 6.556 \times 10^{-8}$) and intensity ($r = 0.259$, $p = 1.246 \times 10^{-7}$) as well as later harvest season ($r = 0.359$, $p = 2.142 \times 10^{-18}$) were also correlated with higher soluble solids.

Differences between apple types

Cultivars were divided into several binary categories according to information from GRIN and/or an online search. Wherever possible, an accession was categorized as either from the Old World or the New World; as a primarily red or other (green/yellow) apple; as primarily used for cider or other (eating/cooking) purposes; and as a late (October/November) or early (August/September) variety. All p-values reported below are Bonferroni-corrected for multiple comparisons.

Similar to the results in Figure 3-3, we found that apples we scored as "late" tended to be juicier (OR = 3.714, $p = 0.006$) and have a higher soluble solids concentration ($W = 5590.5$, $p = 6.11 \times 10^{-7}$) than apples we scored as "early", providing support for the accuracy of our scoring (Figure 3-4). In agreement with previous work, firmness was also higher for later varieties (OR = 4.461, $p = 1.3 \times 10^{6}$) (Watkinsa et al., 2000; Oraguzie et al., 2004; Nybom et al., 2012). We also found that red apples had a higher level of fruit bloom or wax on the fruit compared to green/yellow apples (OR = 0.473, $p = 0.007$).

New World apples were generally larger in size including length ($W = 30705$, $p = 9.651 \times 10^{-10}$), width ($W = 31220.5$, $p = 4.284 \times 10^{-7}$) and weight (OR = 1.958, $p = 0.010$), while Old World apples had a higher russet intensity ($W = 47328.5$, $p = 1.937 \times 10^{-7}$) and were less likely to be red (OR = 2.626, $p = 1.791 \times 10^{-3}$). These observations suggest there may have been stronger selection in New World breeding programs for large red apples with less russeting. As expected based on previous work, apples primarily used for cider were smaller in fruit weight (OR = 5.791, $p = 1.966 \times 10^{-5}$), length ($W = 3034.5$, $p = 1.618 \times 10^{-10}$), and width ($W = 3158.5$, $p = 6.136 \times 10^{-9}$) compared to other (eating and cooking) apples (Miles and King, 2014). Cider apples also oxidized more than eating/cooking apples ($W = 10860$, $p = 2.592 \times 10^{-6}$).

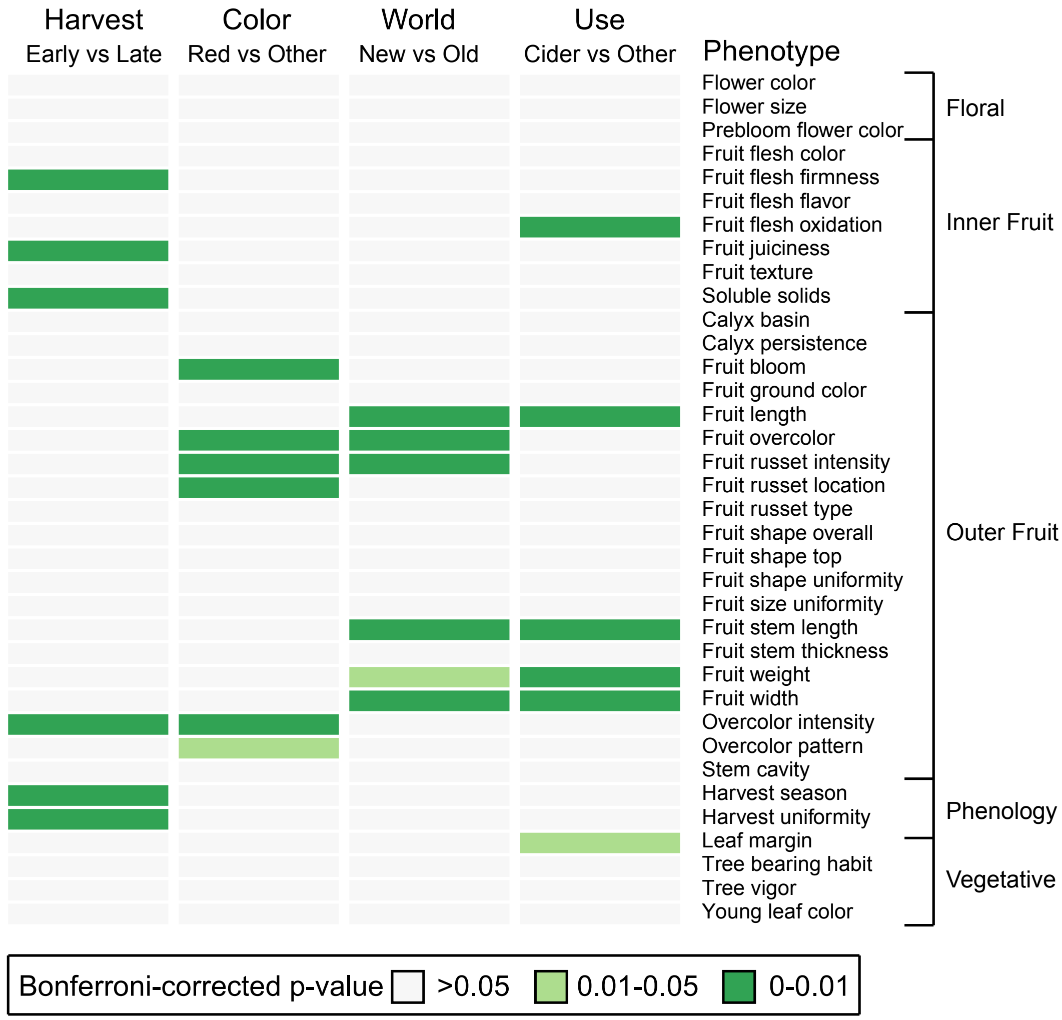


Figure 3-4. The relationship between apple categories and phenotypes. Each phenotype was divided into two groups according to various categories (harvest time, color, geography and used) and compared. *p*-values are Bonferroni-corrected.

Population structure

The genetic structure of the accessions from the USDA collection was investigated using PCA. Principal components (PCs) were calculated from the genome-wide SNP data. Accessions were plotted along the first two PCs and labeled by harvest time and geography (Figure 3-5A). The primary axis of genetic structure (PC1) in apple distinguishes early ripening from late ripening accessions (Figure 3-5B) (W = 5045.5, p =

2.441 x 10^{-25}) while Old World and New World apples differ along PC2 (Figure 3-5C) (W = 117404.5, p = 1.317 x 10^{-44}) indicating that population structure within the domesticated apple is at least partially due to differences in origin and harvest time.

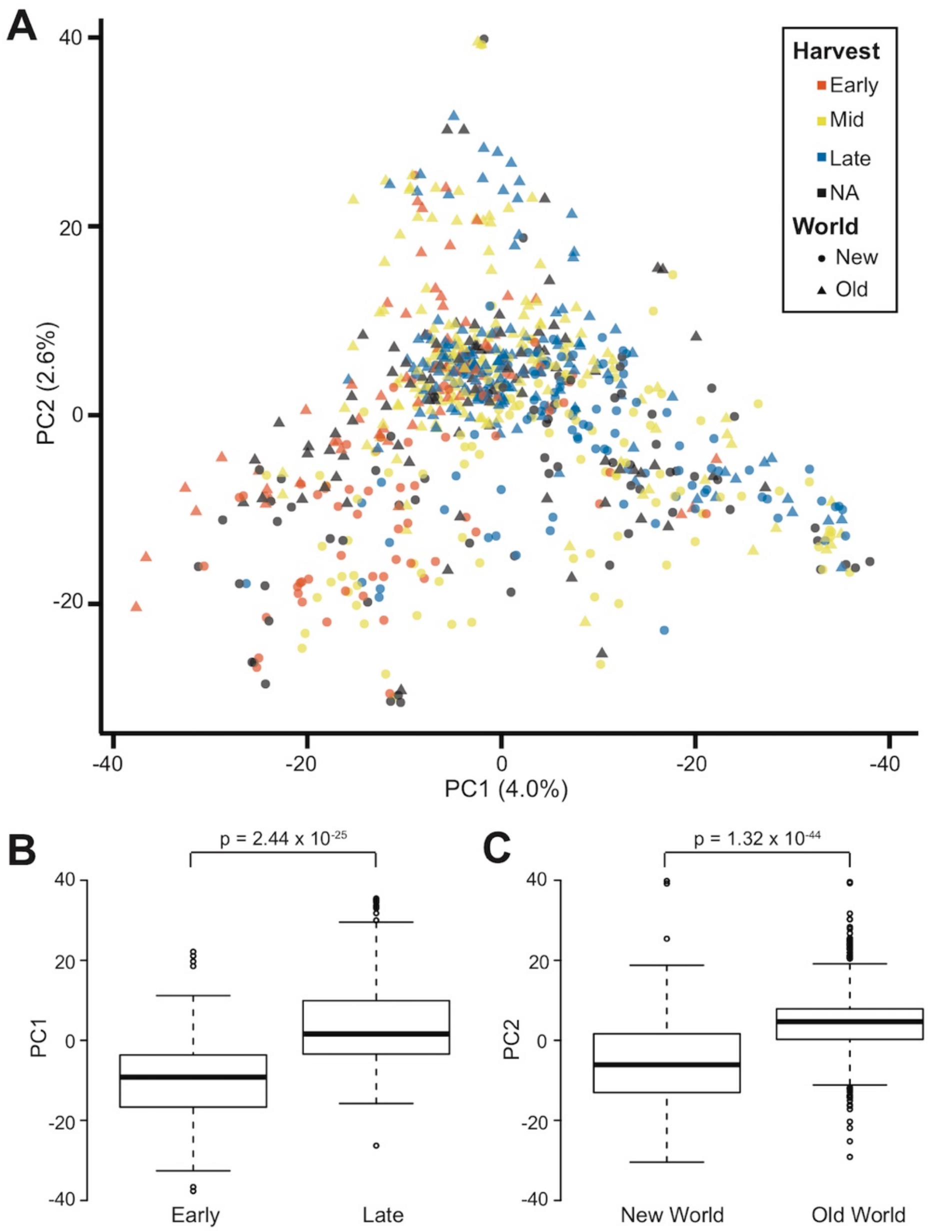


Figure 3-5. Genetic relatedness based on harvest time and geographic origin. (A) PCA was performed using genome-wide SNP data. All samples with known geography information were retained and labeled based on geography with point shape as well as harvest time with point color when possible. Unknown harvest times are marked as NA. The percentage of variance explained by each PC is indicated in parenthesis along each axis. (B) Boxplot of PC1 values for early vs. late harvest varieties of apples. (C) Boxplot of PC2 values for New and Old World varieties of apple. Results are reported from a Mann-Whitney U test.

According to a PCA of genome-wide SNPs, cider and other (eating and cooking) varieties differ significantly along PC1 ($W = 8796$, $p = 7.082 \times 10^{-4}$) and PC2 ($W = 19264$, $p = 8.048 \times 10^{-17}$) (Figure 3-6). In contrast, two previous studies found weak genetic differentiation between cider and dessert cultivars (Cornille et al., 2012; Leforestier et al., 2015).

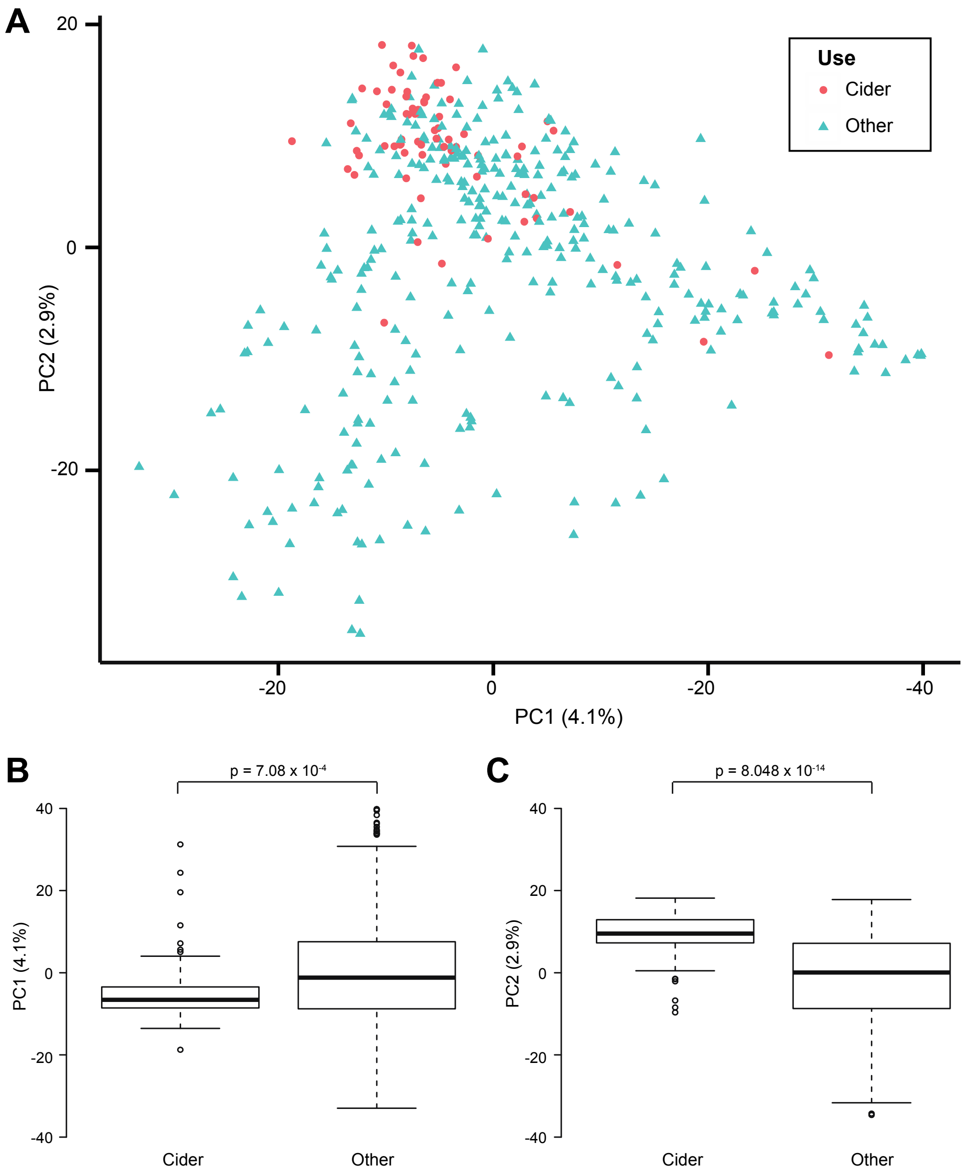


Figure 3-6. Cider was compared to Other (Eating and Cooking) varieties using PCA of genome-wide SNPs. (A) PC1 vs. PC2 for apples based on primary use. Percentage indicates amount of total variance explained by a particular PC. (B) Boxplot of PC1 values for cider and other varieties of apple. (C) Boxplot of PC2 values for cider and other varieties of apple. Values were compared using a Mann–Whitney U test.

We also investigated the degree to which phenotypes were correlated with population structure. We determined the proportion of phenotypic variance (R^2) explained by the first ten genetic PCs for each phenotype and values ranged from 0.07% to 28% (Figure 3-7). In agreement with Figure 3-5, harvest season was most strongly correlated with population structure: the first 10 PCs explained 28% of the variance in harvest season, with PC1 explaining 16% of the variance. The harvest season of an accession is a proxy for the amount of time it requires for that accession to mature. Like many other phenological traits (e.g. Hall and Willis, 2006; Grillo et al., 2013), this measure of ripening time has likely evolved in response to local climates. Thus, the harvest season of an apple may reflect, to some degree, its geographic ancestry. Geography is often a strong predictor of genetic relatedness and models of isolation-by-distance are commonly supported by population genomic data (e.g. Cao et al., 2011). Our observation of harvest season as the trait most strongly correlated with population structure in apples is consistent with a model of isolation-by-distance, with the notion that the time required for an accession to mature has been shaped by the geographic origin of its ancestors.

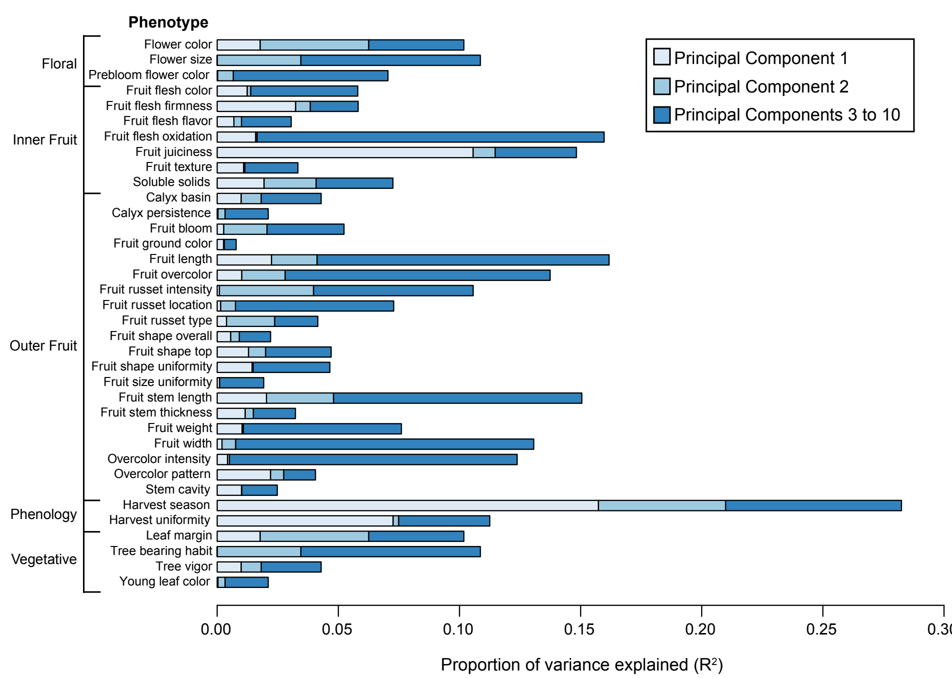


Figure 3-7. Proportion of variance explained for different phenotypes using PCs 1 to 10. PCs were calculated using genome-wide SNPs.

Phenotypic measures are often used as proxies of genetic relatedness: it is assumed that apples that are more closely related should also be more phenotypically similar. We investigated the extent to which a summary of all phenotype measurements captured the genetic relatedness among samples. To accomplish this, we calculated a phenotypic Euclidean distance using the dist() function in R among all pairs of cultivars. We used varying numbers of phenotypes and compared this pairwise distance matrix to a pairwise kinship matrix generated from the genome-wide SNP data using a Mantel test. In our study, phenotypic measurements captured at most 24% of the variance of the kinship matrix. Interestingly, increasing the number of phenotypes measured may decrease the ability to explain genetic relatedness among cultivars, potentially due to the inclusion of low quality phenotypes (Figure 3-8). Thus, assessments of relatedness based on phenotype data alone are unlikely to accurately capture the genetic relatedness of germplasm collections such as this one. Inferences about relatedness in germplasm collections should make use of genotype data and not be based on phenotype data alone whenever possible (Jansky et al., 2015).

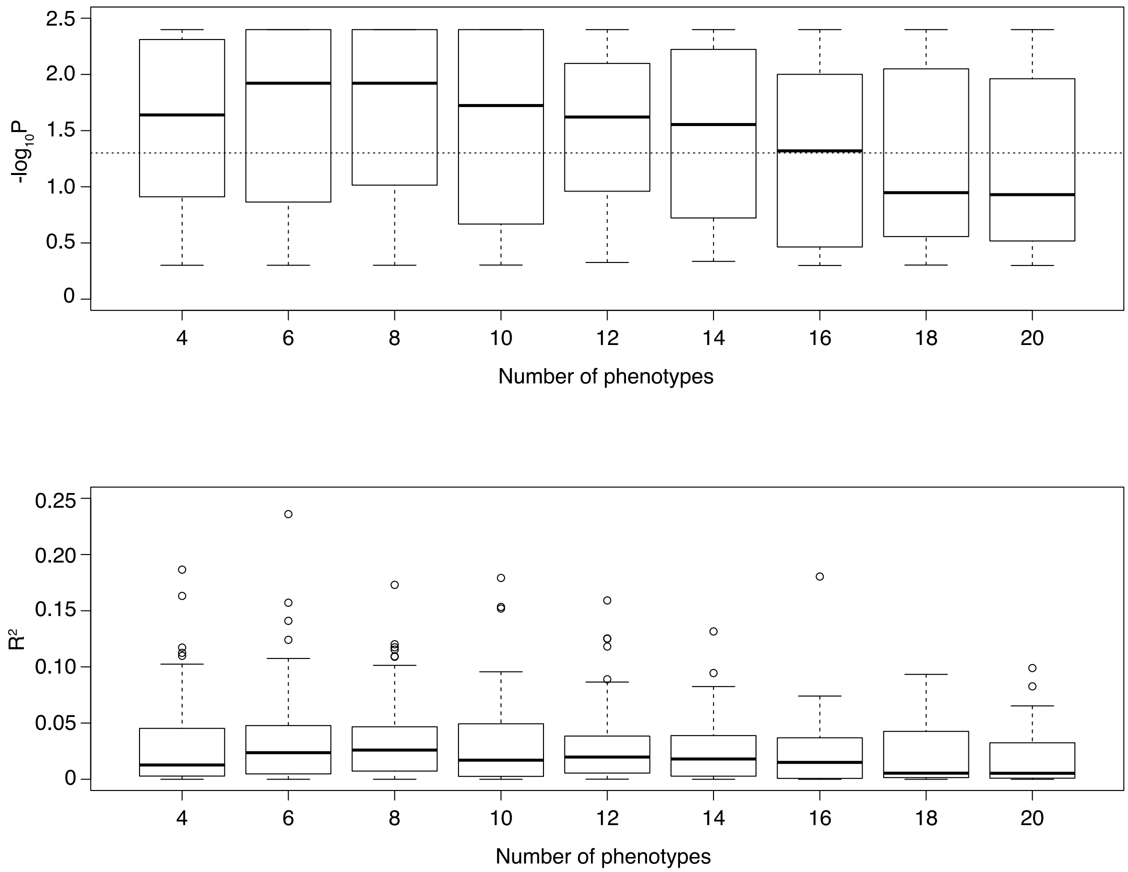


Figure 3-8. Comparison of distance among samples calculated from phenotype and genotype data. (A) -$\log_{10}$ transformed p-values and (B) R^2 were calculated by comparing a phenotypic distance matrix to a kinship matrix generated using genome-wide SNP data. The x axis indicates the number of phenotypes used to generate a phenotype distance matrix. For each sample size, a random set of phenotypes was sampled 100 times and the resulting phenotypic distance matrices were compared to the genetic kinship matrix using a Mantel test.

LD and GWAS

The power of GWAS is dictated in part by the degree to which genotyped SNPs are in LD with causal alleles. We examined the extent of LD decay in the apple genome based on 4096 SNPs, which were 124 kb apart on average. LD in apple is generally low, even at close distances, and decays rapidly (Figure 3-9A). We had 1900 SNP pairs which were <500 bp apart, and this allowed us to assess LD decay even at short distances (Figure 3-9B). At distances of <100 bp, there is a bimodal distribution of LD, with many SNPs in high LD (r^2 > 0.8) and many in low LD (r^2 < 0.2). Very few SNPs have high LD when the inter-SNP distance is >100 bp.

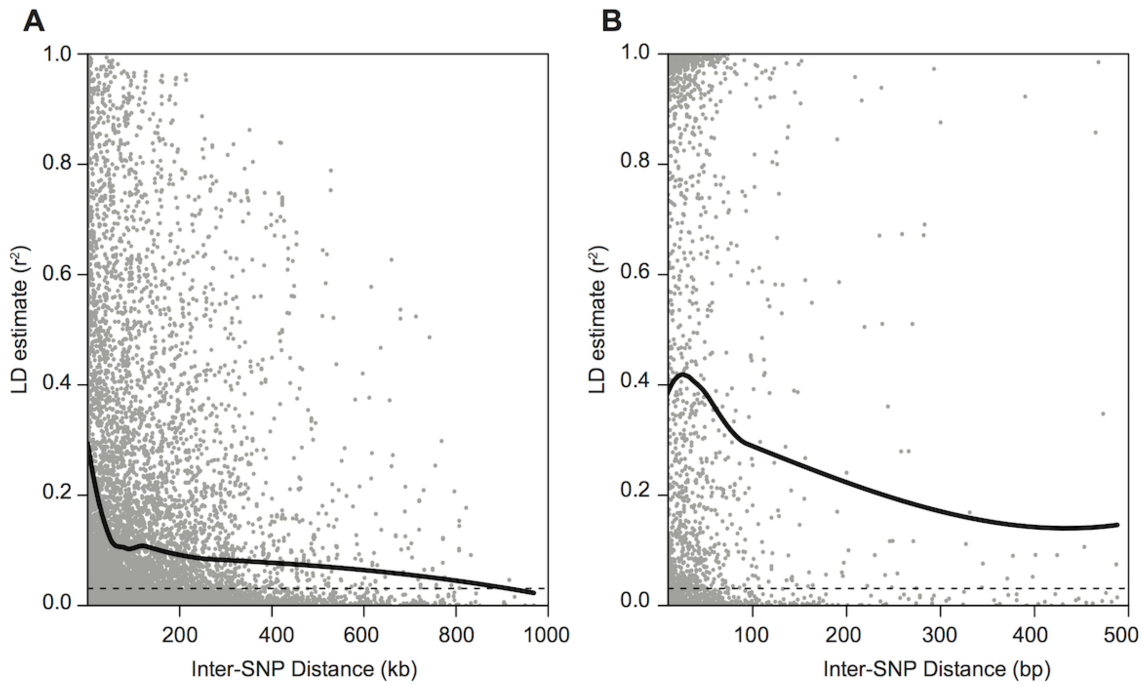


Figure 3-9. Linkage disequilibrium (LD) decay curve in apple. (A) LD decay using comparisons of inter-SNP distances up to 1 Mb. (B) LD decay comparisons of inter-SNP distances of 10 to 500 bp. Smoothed fitted lines were calculated using the LOESS method. The horizontal dotted lines represent background LD: it is the upper 95% confidence interval from 10,000 LD measures generated from comparisons between SNP pairs from different chromosomes.

The rapid LD decay we observe in apple is likely due to the lack of a true founder population resulting in high species diversity, as occurs in many fruit trees such as the domesticated grape (*Vitis vinifera* L.) (Myles et al., 2010; Khan and Korban, 2012). Given that LD decays to r^2 < 0.2 at approximately 100 bp, we reason that, on average, a SNP is needed every 100 bp in order to perform well-powered GWAS in a diverse collection of domesticated apples. The apple has a genome size of ~750 Mb (Velasco et al., 2010b), and we therefore estimate that millions of SNPs are needed for robust GWAS in diverse collections of apples. It also implies that GWAS in apple collections such as the one studied here will achieve extraordinarily high mapping resolution and it will likely be possible to localize the precise nucleotide positions of many causal variants without the need to perform additional fine mapping, which holds great promise for advancing both MAS and genome editing.

While rapid LD decay, relatively low SNP density and the use of historical phenotype data provided potential barriers to successful GWAS, we found several results of note (Figure 3-10) including hits significant associations with fruit firmness (Figure 3-10A), harvest time (Figure 3-10B) and fruit color (Figure 3-10C, Figure 3-10D). All GWAS results are visualized in Appendix II: Figure II-I. The alleles, p-value, MAF, and effect for all 31 SNPs significantly associated with a phenotype after Bonferonni correction are listed in Appendix II: Table II-I.

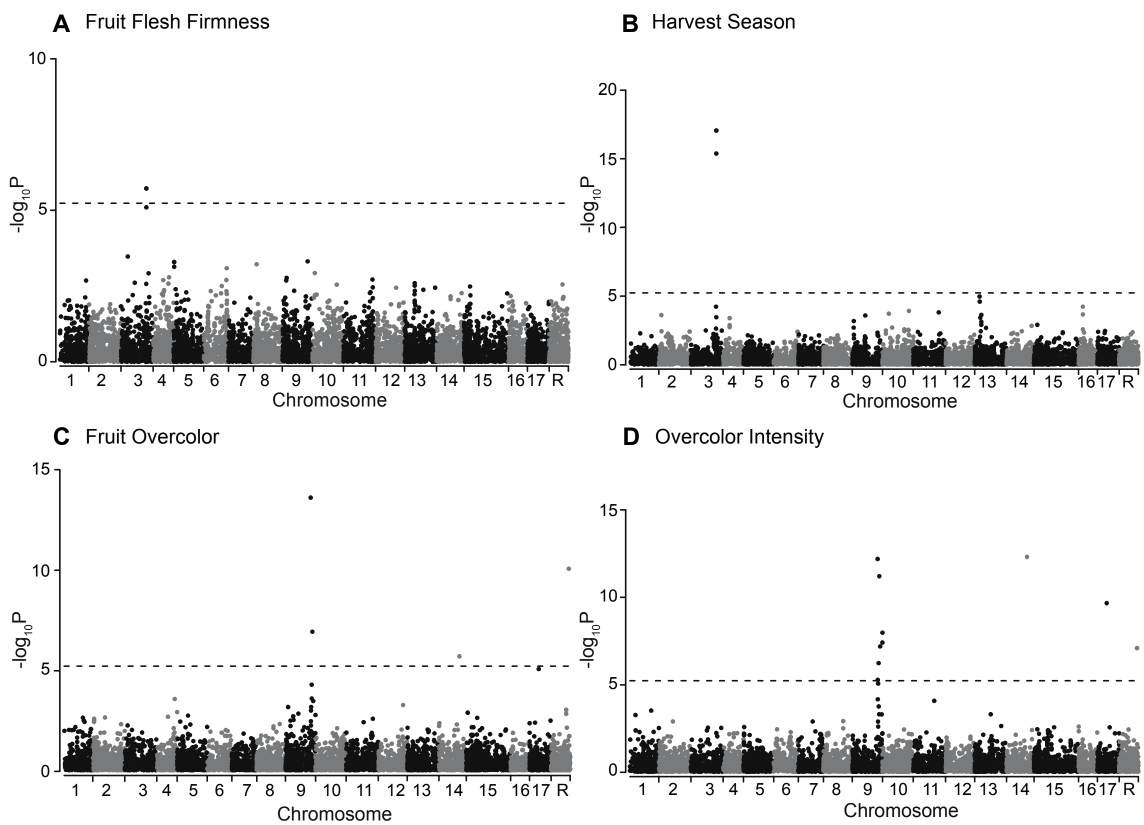


Figure 3-10. Manhattan plot of GWAS results for traits of interest, including (A) fruit flesh firmness, (B) harvest season, (C) fruit overcolor, and (D) overcolor intensity. p-values are log-transformed, and the threshold for significance is Bonferroni-corrected and indicated by the horizontal dotted lines. Chromosome R indicates SNPs found on contigs that remain unanchored to the reference genome.

The apple is a climacteric fruit in which ethylene production during fruit ripening drives a loss of fruit firmness. Excessive softening is undesirable and leads to lower consumer acceptability, so there is a strong interest in breeding cultivars that retain their firmness during extended storage (Johnston et al., 2002). Cultivars harvested later in the season also tend to have slower softening rate, potentially due to having smaller cells and smaller intercellular spaces that may result in stronger tissue (Johnston et al., 2002; Nybom et al., 2012).

Alleles at two genes in ethylene's biosynthetic pathway, *Md-ACO1* on chromosome 10 and *Md-ACS1* on chromosome 15, have been repeatedly associated with fruit firmness (Oraguzie et al., 2004; Costa et al., 2005; Costa et al., 2010) and markers at these loci are used in marker-assisted breeding (Ru et al., 2015). We did not find significant associations with fruit firmness on chromosome 10 or 15, likely due to low SNP density surrounding the causal loci. While we were unable to determine the exact position of *Md-ACS1* in the reference genome, the SNPs closest to *Md-ACO1* were >100 kb away. Given the rapid extent of LD decay, it is not surprising we were unable to identify a significant peak for these two loci.

Fruit firmness is a complicated physiological process and in addition to *Md-ACO1* and *Md-ACS1*, other genes are also involved (Atkinson et al., 2012; Costa et al., 2014). Here we report a single GWAS hit on chromosome 3 for firmness (Figure 3-10A) that overlaps with the hit for harvest time (Figure 3-10B). The overlap of GWAS hits for firmness and harvest time is likely due to the physiological relationship between fruit maturity and firmness. Previous work, as well as our own (Figure 3-4), have reported positive correlations between later harvest and firmer fruit (Watkinsa et al., 2000; Oraguzie et al., 2004; Nybom et al., 2012).

Previous linkage mapping studies in biparental apple populations identified QTLs for both firmness and harvest date on chromosome 3 (Liebhard et al., 2003; Kenis et al., 2008). A recent GWAS also identified significant SNPs for fruit firmness on chromosome 3 (Kumar et al., 2013a). Our GWAS hit on chromosome 3 is consistent with these previously identified QTL for firmness and harvest date. Our high resolution

GWAS enabled a refinement of the position of this QTL on the distal end of chromosome 3 and we found that it falls within the coding region of NAC18.1 (GenBank ID: NM_001294055.1; chr3:31407982..31409374). NAC18.1 is a transcription factor that belongs to the *NAC* gene family, which is one of the largest families of plant-specific transcription factors (Olsen et al., 2005). NAC proteins are known to be involved in fruitlet abscission in apple (Botton et al., 2011) and ripening in peach (*Prunus persica* (L.) Batsch) (Pirona et al., 2013) and banana (*Musa acuminate* Colla) (Shan et al., 2012).

In our study, the most significant GWAS hit results in a nonsynonymous substitution from aspartic acid (D) to tyrosine (Y) at the fifth amino acid of NAC18.1, which we refer to as D5Y. Although D5Y does not appear to fall within a functional domain (de Castro et al., 2006), a D to Y amino acid substitution results in a score of -3 according to the BLOSUM62 matrix (Henikoff and Henikoff, 1992). Most notably, D5Y lies within a motif that we refer to as the TDSS motif. Using phylogenetic analysis, we demonstrate that NAC proteins possessing a TDSS motif cluster together with a bootstrap value of 88. The NAC proteins showing the highest homology to NAC18.1 have the TDSS motif, and the D residue is conserved in 22 of the 25 proteins analyzed (Figure 3-11, Appendix II: Figure II-II). We hypothesize that the possession of this motif indicates shared evolutionary function. For example, the tomato (*Solanum lycopersicum* L.) NOR protein is required for ripening in tomato (Karlova et al., 2014). Likewise, in *Arabidopsis thaliana*, NAC2 is involved in ethylene-mediated senescence (Qiu et al., 2015). In kiwifruit (*Actinidia arguta* var. *arguta*), one member of the NAC protein family, NAC2, which encodes a variant of the TDSS motif (PDSS), activates the promoter of terpene synthase 1 (*TPS1*) more strongly than NAC1 and NAC3, which carry the conserved TDSS motif. *TPS1* expression results in aromatic terpene production in ripe fruit (Nieuwenhuizen et al., 2015). The D to Y substitution we describe here correlates with softer apples that ripen earlier. If D5Y is indeed causal, it may act as a gain of function by activating the expression of downstream proteins and accelerating ripening in apple. Further functional studies are required to reveal the possible function of D5Y.

Species	NAC Sequence
Apple derived (*Malus domestica*)	MECTYSSAGS
Apple ancestral (*Malus domestica*)	MECTDSSAGS
Pear (*Pyrus bretschneideri*)	MESTDSSAGS
Grape (*Vitis vinifera*)	MESTDSSSGS
Arabidopsis (*Arabidopsis thaliana*)	MESTDSSGGP
Poplar (*Populus euphratic*)	MEGTISSSGS
Kiwifruit (*Actinidia arguta*) NAC1	MESTDSSTGS
Kiwifruit (*Actinidia arguta*) NAC2	MESPDSSVGL
Rice (*Oryza sativa* Japonica Group)	MESPDSSSGS
Wheat (*Aegilops longissima*)	MGSPDSSSGS

Figure 3-11. Multiple species alignment of NAC proteins. Proteins that include the TDSS motif were chosen from pear (NAC 18 GenBank ID: XP_009334622.1), grape (NAC25 GenBank ID: CBI20351.3), *Arabidopsis thaliana* (NAC2 GenBank ID: AEE75684.1), poplar (NAC25 GenBank ID: XP_011027905.1), kiwifruit (NAC1 GenBank ID: AID55348.1 and NAC2 GenBank ID: AID55349.1), rice (Os07 g0566500, GenBank ID: NP_001060017.1) and wheat (NAM-1, GenBank ID: AFD54040.1). The D5Y substitution is highlighted.

In addition to the D5Y substitution identified, harvest season had 2 additional significant SNPs near the NAC18.1 coding region (chr3:31409376 and chr3:31409480, Appendix II: Table II-I), both in high LD ($r^2 > 0.94$) with D5Y. Ultimately, SNPs in and around NAC18.1 are potential markers for MAS, which would help determine harvest time and firmness in seedlings from new crosses with unknown maturity dates and ensure cultivars were grown in the appropriate climate to ripen before winter.

Identification of markers for fruit skin color in apple using GWAS may be useful for marker-assisted breeding by allowing for skin color selection during the juvenile phase, prior to fruit production (Zhang et al., 2014). Anthocyanins are responsible for the red coloration in apples and the transcription factor MdMYB1 regulates anthocyanin genes, forming the genetic basis for apple skin color (Takos et al., 2006; Ban et al., 2007). In the present study, we sought associations with fruit overcolor (Figure 3-10C), which is a binary assessment (i.e., red vs green), and with overcolor intensity (Figure 3-10D), which is a quantitative measurement of the percentage of overcolor (generally red) on a fruit.

We confirmed the association between MYB1 and fruit color: the most convincing peaks for fruit overcolor and overcolor intensity occur near position 32.8 Mb on chromosome 9 where the MYB1 is located (Zhu et al., 2010; Gardner et al., 2014).

To investigate the ability of single GWAS hits (Figure 3-10) to explain phenotypic variance, we examined the phenotype data for each of the three possible genotypes at each of the most significant GWAS SNPs (Figure 3-12). We found that an accession with at least one minor allele (A) at chr3:31409362 was 'soft' in 67% of cases (Figure 3-12A). Accessions with this same allele had an 'early' harvest time in 63% of cases (Figure 3-12B). When using a binary assessment for fruit color, accessions with at least one minor allele (A) at position chr9:31448296 were 'red' 87% of the time (Figure 3-12C). Further studies are required to evaluate the utility of these markers for predicting phenotypes in other populations.

We also used quantitative measurements of fruit overcolor intensity to estimate the mode of inheritance using SNPStats (Figure 3-12D) (Sole et al., 2006). Based on a single SNP, a codominant model fit best in which a single A allele resulted in a 22% increase in fruit overcolor intensity, while two A alleles resulted in a 28% increase compared to the mean overcolor value of TT (Akaike Information Criterion (AIC) = 5684). This codominant model fits only slightly better than a dominant model (AIC = 5692) in which a single A allele results in the full phenotypic effect. Based on our dataset, these SNPs all have potential for MAS and the ability to help select for desirable characteristics such as firmness and color using a single SNP.

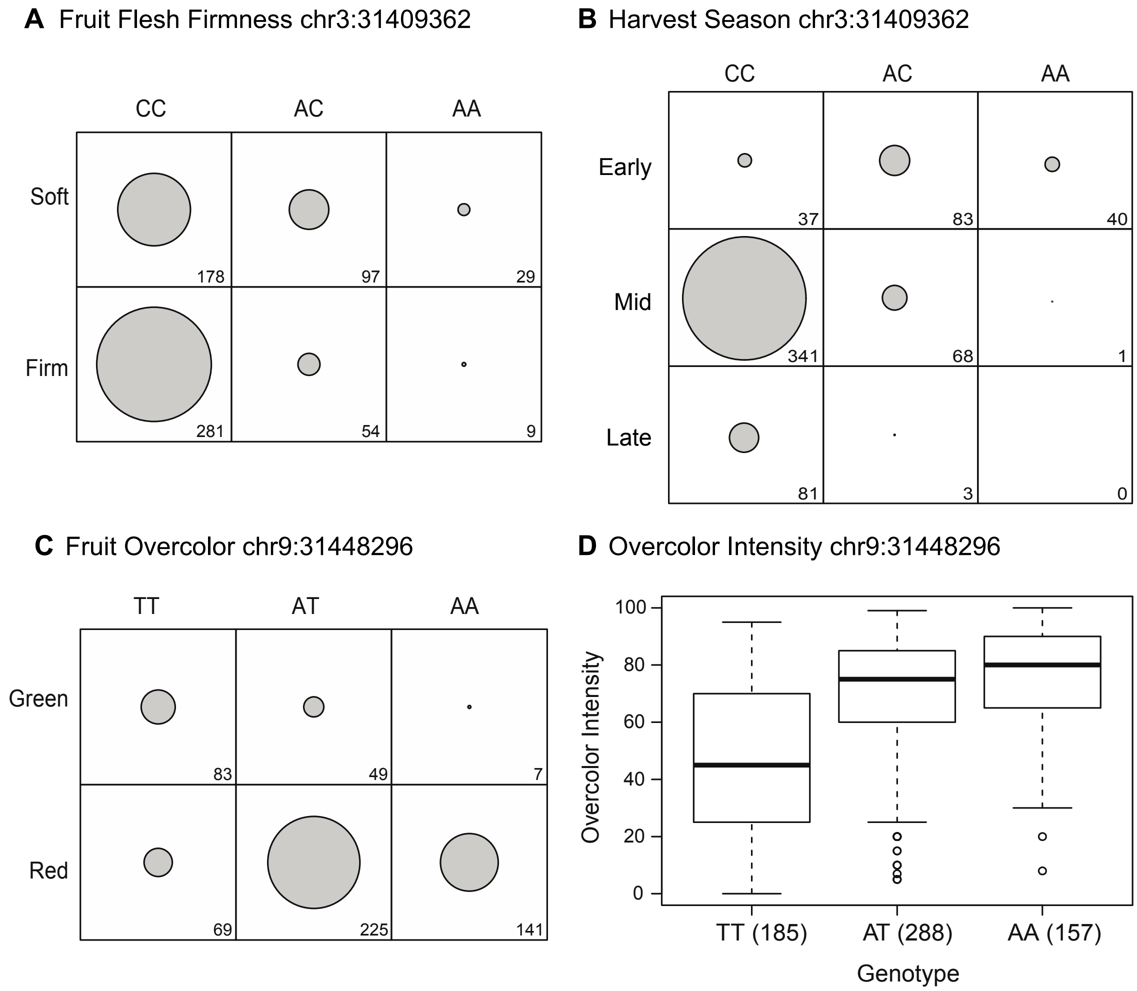


Figure 3-12. Distribution of phenotype scores stratified by genotype at the most significant GWAS SNPs for (A) firmness, (B) harvest season, (C) fruit overcolor, and (D) overcolor intensity. The sequence for each potential genotype is indicated. The number of observations within a particular genotype or phenotype category is listed. Circled areas are proportional to the number of observations.

Significant GWAS SNPs that are relatively isolated and not near any other significant SNPs may still be true positives given that we observed such rapid LD decay. However, some of the genotype-phenotype associations may be either false positives or simply mismapped SNPs in the reference genome. For example, LD between the lone hit on chromosome 17 at 11.6 Mb for overcolor intensity is as strong as the LD between the top chromosome 9 hit and many physically adjacent SNPs. Therefore, the chromosome 17 SNP may have been assigned an incorrect physical position in the reference genome sequence used in this study. In two previous linkage mapping studies, 13.7% and 18.3%

of SNPs were assigned to linkage groups that conflicted with the predicted chromosomal locations according to the reference genome (Antanaviciute et al., 2012; Gardner et al., 2014). Caution is therefore warranted in drawing strong conclusions about the number of QTL or the genetic architecture of a trait when dealing with incomplete or poor quality reference genome sequences. The associations we report in Appendix II: Table II-I should be interpreted taking this into consideration.

Phenotype prediction accuracy using historical data in apple

Although the rapid LD decay we observe here presents challenges for GWAS in apple using relatively low-density genotyping platforms such as GBS, we still observed suggestive hits including several that overlap with known QTL. However, the strict thresholds of GWAS only enable the detection of loci of relatively large effect. Most traits of interest to breeders are arguably highly polygenic and thus controlled by numerous small effect loci. As genome-wide marker data become available for an increasing number of organisms, it is becoming common to evaluate the extent to which phenotypes can be predicted using genomic prediction methods (Heffner et al., 2011). Genomic prediction is a particularly useful tool for researchers interested in predicting complex, polygenic phenotypes as it uses all markers simultaneously to predict phenotypes without identifying QTL (Gibson, 2010; Endelman, 2011).

To investigate the degree to which we can predict the phenotypes explored in this study, we performed genomic prediction analysis and determined that the highest prediction accuracy (0.57) was found for harvest season followed by fruit width (0.48) and length (0.47) (Figure 3-13). Similar prediction accuracies ranging from 0.31 for grain yield to 0.63 for flowering time have been found in rice (Spindel et al., 2015). Although many phenotypes are not well-predicted using genome-wide SNP data, or were collected in a way that does not enable genetic mapping, this was not always the case. Several phenotypes appeared to be heritable, predictable, and even controlled by loci that were detected using GWAS. For example, color had fairly high prediction accuracies of 0.41 for overcolor intensity and 0.32 for fruit overcolor.

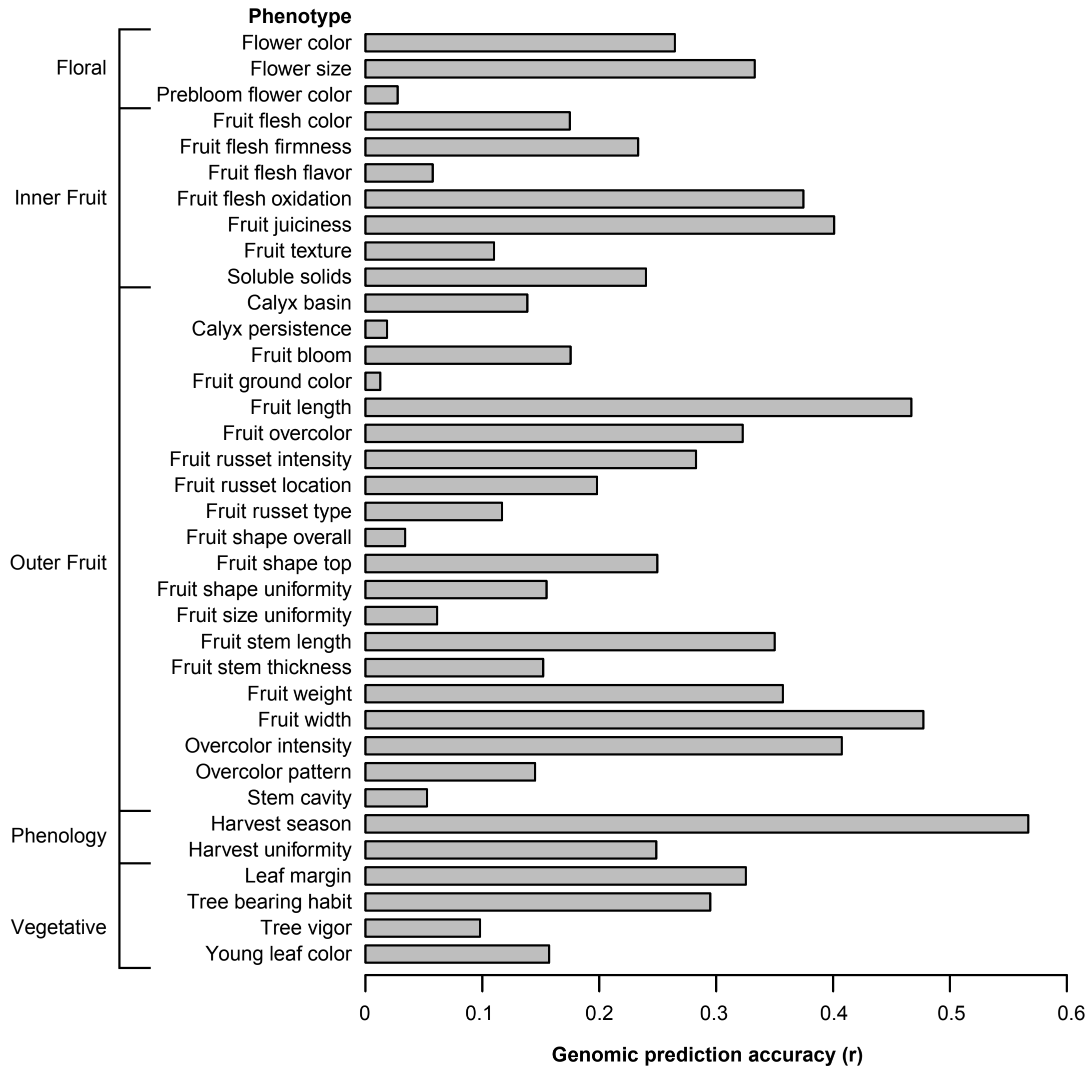


Figure 3-13. Genomic prediction accuracy: *r* values represent the correlation between observed and prediction phenotype scores from genomic prediction using a five-fold cross-validation procedure.

Phenotype prediction accuracy was highly correlated with the proportion of phenotypic variance explained by genetic PCs 1 to 10 ($r = 0.898$, $p = 1.141 \times 10^{-13}$) (Figure 3-14). While genetic PCs capture the principal axes of population genetic structure, phenotype prediction relies on a genetic relationship matrix among all samples derived from all SNPs. The observed positive relationship between these two methods of explaining phenotypic variation is expected given that both methods capture genetic relatedness among accessions in different manners.

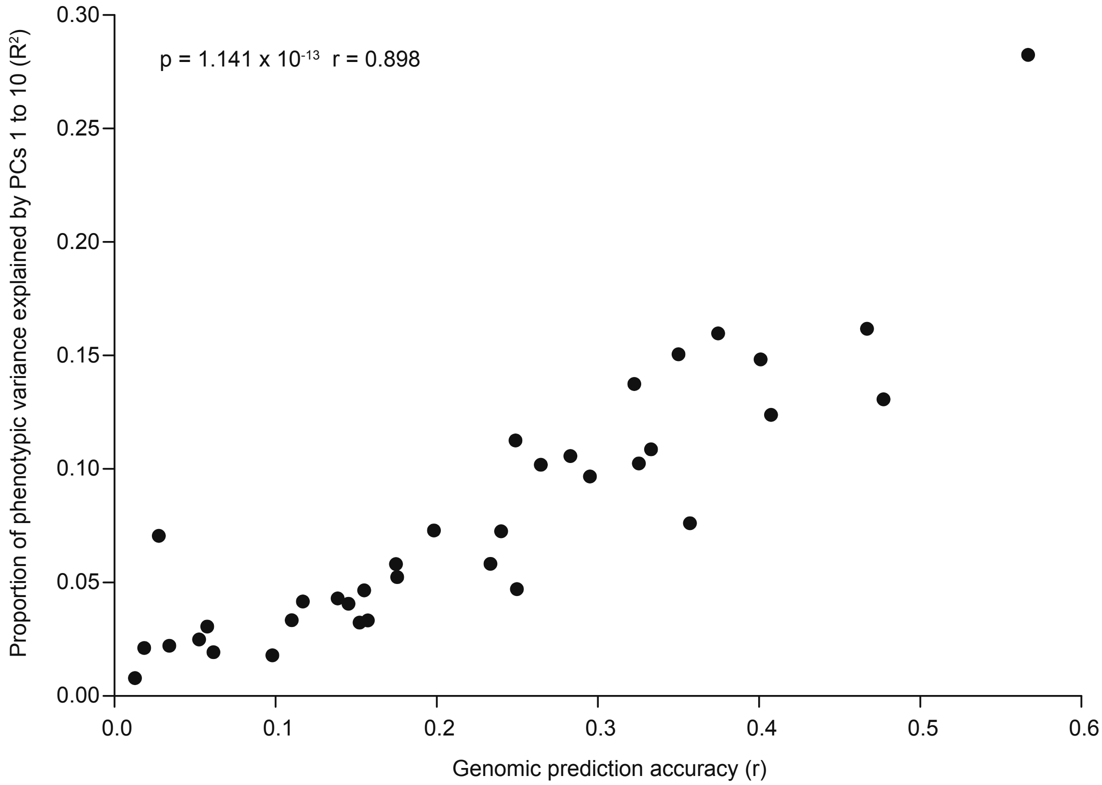


Figure 3-14. Correlation between genomic prediction accuracy and proportion of phenotypic variance explained by the first 10 genetic PCs.

Conclusions

There are many difficulties associated with the use of historical phenotype data for association mapping and genomic prediction, including small samples sizes, use of subjective measurements, and inconsistent data collection across years. However, by recoding phenotype data and combining across years we were able to observe relationships between phenotypes as well as provide evidence of several GWAS hits including color, fruit firmness and harvest time. In particular, we report a novel nonsynonymous SNP in transcription factor NAC18.1, which is correlated with softer apples that ripen earlier and warrants further functional investigation. The continued advent of high-throughput phenotyping technologies and improvement in phenotype measurement collection will increase our ability to understand the genome-phenome relationship in apple. Due to rapid LD decay, whole genome sequencing may be required to enable well-powered association mapping in diverse collections like the one studied here. However, using sufficiently dense genotype data, we expect loci to be detected at nearly nucleotide resolution, which will allow for more widespread adoption of marker-assisted selection.

Acknowledgments

This project was enabled through the Cooperative Agreement (58-1910-02- 029FN) entitled "Characterization of genomic diversity and evolutionary origins of the USDA-ARS *Malus* germplasm collection" between the USDA-ARS and Dalhousie University. We thank the people who manage and maintain the USDA apple germplasm collection in Geneva, NY, and also those who made the phenotype data from the apple collection available via the USDA-GRIN database. We would like to acknowledge Angela Baldo, Chris Richards, and Gayle Volk for their contributions to initial discussions of data processing. We would also like to acknowledge Daryl Somers for useful input. This article was written, in part, thanks to funding from the Canada Research Chairs program, the National Sciences and Engineering Research Council of Canada, and Genome Canada. Z.M. was supported in part by a Killam Predoctoral Scholarship from Dalhousie University

Chapter 4: Genomic ancestry estimation quantifies use of wild species in grape breeding

Introduction

Grapes are one of the world's most valuable crops and although grown primarily for wine, they are also used fresh, dried and in juice (Reisch et al., 2012). In 2013, grapes had the 2nd highest global gross production value among fruit crops, exceeded only by tomato (Food and Agriculture Organization of the United Nations, 2015). Grapes belong to the genus *Vitis*, which includes over 60 inter-fertile species spread broadly across the northern hemisphere (This et al., 2006). However, based on total global area in 2010, over 98% of wine grapes belong to a single species, *Vitis vinifera* (Anderson and Aryal, 2013). Almost all grape cultivars grown commercially are either *V. vinifera* or hybrids that include *V. vinifera* parentage (Reisch et al., 2012).

In addition to the use of one *Vitis* species for almost all grape growing, grapes are predominately grown using vegetative propagation, which has resulted in extensive clonal relationships and limited diversity. The wine industry's preference for traditional varieties makes the acceptance of even new *V. vinifera* cultivars difficult (Alleweldt and Possingham, 1988; Bisson et al., 2002). A study by Myles et al. (2011) found 58% of the 950 grape cultivars examined had at least one clonal relationship. Among the unique cultivars, 74.8% had a first-degree relationship with at least one other cultivar. This extensive inter-relatedness and lack of diversity have left grape cultivars susceptible to many continually evolving pathogens (Myles et al., 2011; Myles, 2013). For example, Pierce's disease currently costs the California wine industry approximately $92 million annually (Alston et al., 2013). The future of the wine industry relies on the exploration of new genetic diversity through breeding.

Crop wild relatives (CWRs) provide a useful source of genetic variation for crop improvement (Tanksley and McCouch, 1997; Hajjar and Hodgkin, 2007; McCouch et al., 2013). An overview of 19 different crops found that more than 80% of beneficial traits from CWR genes were involved in pest and disease resistance (Hajjar and Hodgkin, 2007). By 1997, genomic crop improvements made due to CWRs had an estimated global benefit of $115 billion annually (Pimentel et al., 1997). Due to disease susceptibility of *V. vinifera* cultivars, settlers to North America had great difficulty growing the vine. These early settlers grew native, wild vines such as *V. labrusca* and *V. aestivalis* and hybridized them with *V. vinifera* (Alleweldt, 1997). Significant exploitation of CWRs began in the 1850s when the phylloxera louse devastated European vineyards. Breeders used American wild *Vitis* species to develop rootstocks resistant to phylloxera, rescuing the wine industry. Commercial *V. vinifera* wine cultivars are still grafted onto these phylloxera-resistant rootstocks (Alleweldt and Possingham, 1988; Zhang et al., 2009).

Largely in response to the phylloxera crisis, wild *Vitis* were also used in scion breeding. However, the initial hybrids were generally considered undesirable for wine production due to unfavorable aromas and tastes inherited from the wild *Vitis* parents (Liang et al., 2008; Sun et al., 2011; Liang et al., 2012; Narduzzi et al., 2015). Sustained breeding enabled the development of hybrid cultivars with improved disease resistance and without the undesirable flavor compounds, including German varieties such as 'Phoenix' and 'Orion' (Alleweldt and Possingham, 1988). Early French breeders, including Eugene Kuhlmann and Pierre Castel, also created well-known hybrids such as 'Marechal Foch' and 'Castel'. However, despite the promise of novel hybrid grape cultivars, their use was met with strong resistance. France introduced several wine "quality laws" prohibiting the use of many French-American hybrids (Reisch et al., 2012; Meloni and Swinnen, 2014). French regulations influenced the perception of hybrid grape cultivars, as well as the European Union wine classification, which outlawed hybrids from the highest quality level (Meloni and Swinnen, 2014).

Although it is widely believed that nearly all commercial grape varieties derive their entire ancestry from *V. vinifera*, there is increasing evidence that wild *Vitis* species may have been incorporated more often than previously assumed. Estimates of *V. vinifera* ancestry frequently rely on historical pedigrees from breeders, but these records may be flawed. Genomics provides a powerful tool for detecting pedigree errors and wild *Vitis* ancestry. For example, a recent study used nuclear microsatellite markers to determine that 33% of the 381 breeder pedigrees examined were incorrect. In most cases, the paternal parent was incorrectly identified, likely due to pollen contamination (Lacombe et al., 2013). Most recently, a genomic analysis uncovered that the most important Japanese wine cultivar, 'Koshu', contained 30% wild ancestry despite being commonly classified as entirely *V. vinifera* (Goto-Yamamoto et al., 2015).

In addition to illuminating the contribution of wild *Vitis* to commercial grapes, genomics can help breeders introgress desirable traits from wild relatives into new grape cultivars. Marker-assisted selection (MAS) uses genetic markers either responsible for a phenotype or strongly linked to it. MAS is especially helpful in long-lived perennial crops, like grapes, where selection can be made at the seed or seedling stage, eliminating the time and money required for the plant to fully mature (McClure et al., 2014). Moreover, combining markers linked to key traits with genomic ancestry estimates can enable breeders to select the progeny with the highest *V. vinifera* content as well as the desirable trait from the wild relative.

To enable genomics-assisted ancestry estimation in grapes, Sawler et al. (2013) estimated *V. vinifera* ancestry in interspecific *Vitis* hybrids using single nucleotide polymorphism (SNP) array data from 127 accessions in the grape germplasm collection of the United States Department of Agriculture (USDA). However, the USDA collection contains relatively few commonly grown commercial hybrid cultivars. To gain insight into the ancestry across the most common commercial hybrids, we generated genotyping-by-sequencing (GBS) data and quantified *V. vinifera* ancestry from 64 of the most widely grown commercial hybrids from North America and Europe. We find that *V. vinifera* ancestry ranged from 11% to 76% across our sample of hybrid varieties. The distribution of ancestry across hybrids suggests the unusual practice of breeders backcrossing more frequently to wild *Vitis* species than to *V. vinifera* during hybrid grape breeding.

Methods

Sample Collection and Genotype Calling

Leaf tissue was collected from 63 commercial grape varieties from Canada (Nova Scotia), Germany and the United States. We also used samples from 11 *V. vinifera*, 6 hybrids, and 15 wild accessions from the USDA grape germplasm collection which were previously genotyped in Sawler et al. (Sawler et al., 2013). DNA was extracted using commercial extraction kits. A list of all samples is available in Appendix III: Table III-I.

A single GBS library from 96 samples was generated according to Elshire et al. (2011) using two different pairs of restriction enzymes (HindIII-HF/BfaI, HindIII-HF/MseI) and was sequenced using Illumina Hi-Seq 2000 technology. Reads were aligned to the 12X grape reference genome from GENOSCOPE (http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/) using the Tassel/BWA version 5 pipeline with minimum quality score (mnQS) of 20 and minimum kmer count (c) of 3 to generate a genotype matrix with 830,822 sites (Li and Durbin, 2009; Glaubitz et al., 2014). All other default parameters were used.

Data Curation

VCFtools v0.1.12b (Danecek et al., 2011) was used to filter for biallelic sites as well as a minimum number of reads (minDP) of 8. The file was converted into PLINK format and SNPs with <20% missing data were retained using PLINK v1.07 (Purcell et al., 2007; Purcell, 2009a). Accessions with >20% missing data were removed, followed by SNPs with a minor allele frequency (MAF) <0.05. SNPs with excess heterozygosity (i.e. failed a Hardy-Weinberg equilibrium test with a p-value < 0.001) were also removed, resulting in 80 accessions and 6664 sites remaining. An identity-by-state (IBS) similarity matrix was calculated using PLINK for hybrid samples. Missing genotypes were imputed using LinkImpute (Money et al., 2015) with optimized values of 6 and 17 for parameters l and k, respectively, which resulted in an estimated genotype imputation accuracy of 91%.

In order to perform a PCA-based admixture analysis, equal ancestral sample sizes are required (McVean, 2009). We removed two random *V. vinifera* samples, and the resulting dataset contained 78 samples, which included ancestral populations of 7 wild *Vitis* samples and 7 *V. vinifera*. Only SNPs with MAF >0.1 in the ancestral populations were retained across all samples.

We pruned for linkage disequilibrium using PLINK by considering a window of 10 SNPs, removing one SNP from a pair if r^2 was >0.5 then shifting the window by 3 SNPs and repeating the procedure (PLINK command: indep-pairwise 10 3 0.5). 2538 sites, which included 56 indels and 2482 SNPs, remained for principal component analysis (PCA). The median distance between SNPs remaining after filtering was 1086 bp and the inter-SNP distribution can be seen in Figure 4-1.

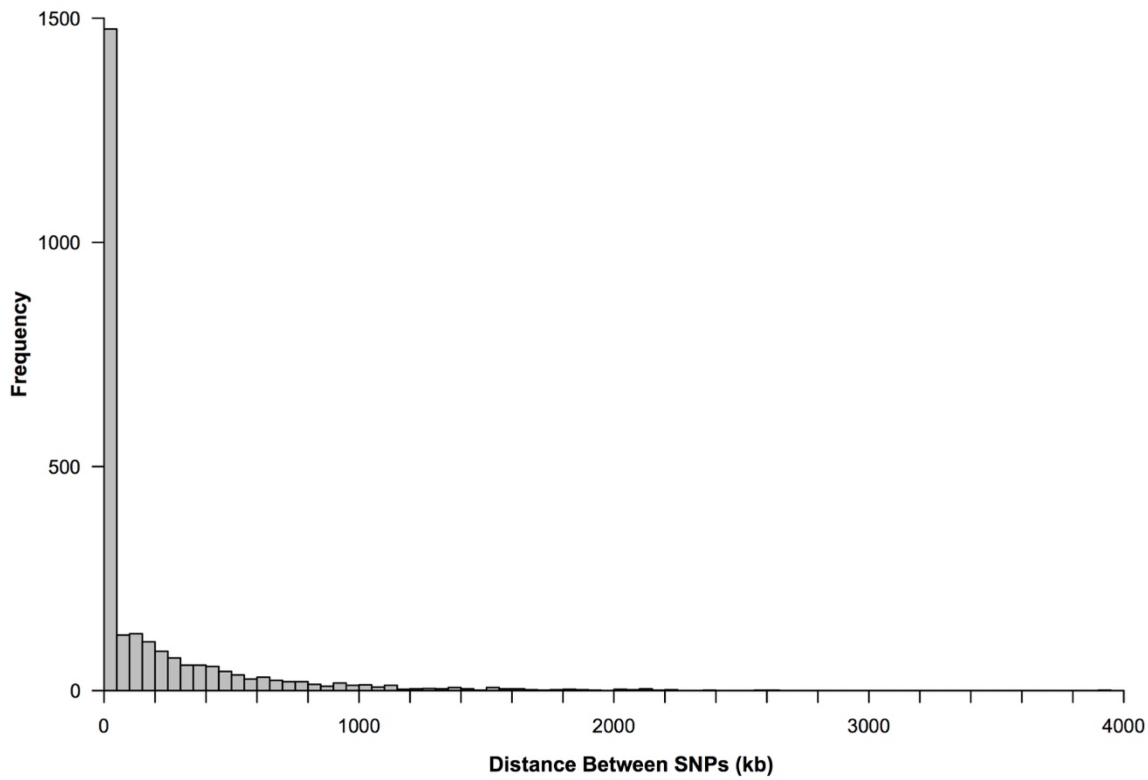


Figure 4-1. Distance (kb) between filtered SNPs used for ancestry estimation.

Ancestry estimation

We calculated principal component (PC) axes using the ancestral *V. vinifera* and wild *Vitis* samples and then projected hybrid cultivars onto these axes using smartpca from the EIGENSOFT v.6.0.4 software package (Figure 4-2A) (Patterson et al., 2006; Price et al., 2006). Based on PCA projection, ancestry coefficients for each hybrid were estimated using a similar approach described in (Bryc et al., 2010; Sawler et al., 2013). The Euclidean distance between a particular hybrid cultivar and the mean value for *V. vinifera* (a) and wild *Vitis* (b) populations along PC1 was calculated and the percentage of *V. vinifera* was determined using the formula '% *V. vinifera* = b/(a+b)*100' (Figure 4-2B).

Simulations of Admixture

In order to determine the accuracy of the PCA-based ancestry estimates, we generated simulated offspring using data from the genotyped samples as described in Sawler et al. (2013). We estimated the proportion *V. vinifera* ancestry from simulated F1 hybrids, F1 x *V. vinifera* backcrosses and F1 x wild *Vitis* backcrosses, which are expected to have 50%, 75% and 25% *V. vinifera* ancestry, respectively. For the F1 hybrids, a parent was randomly selected from each two ancestral populations, and parental genotypes were combined by randomly sampling one allele at each site. Linkage disequilibrium between sites was ignored and the process was repeated 10,000 times in order to generate 10,000 F1 offspring. The procedure was repeated with a randomly chosen simulated F1 as one parent and a randomly chosen wild *Vitis* (n = 10,000) or *V. vinifera* (n = 10,000) as the other, in order to simulate backcrossing to the ancestral populations. The percentage *V. vinifera* ancestry and 95% confidence interval were calculated for all simulated populations.

Results and Discussion

Method verification

Wild *Vitis* species can be used in grape breeding programs to introgress disease and abiotic stress resistance into susceptible germplasm belonging to the domesticated grape, *V. vinifera*. Commercial cultivars with wild *Vitis* ancestry are often referred to as "hybrids". An evaluation of ancestry across commercial hybrids can provide insight into the history of hybrid grape breeding and a foundation for future efforts to select for ancestry based on marker data. Previous work provided accurate ancestry estimates of interspecific grape cultivars using Vitis9KSNP array data for cultivars belonging to the USDA germplasm collection (Sawler et al., 2013). We applied the same PCA-based method to evaluate the ancestry of some of the most widely grown hybrid cultivars sampled from North America and Europe using GBS data.

PCA provides a clear separation of wild *Vitis* and *V. vinifera* samples along PC1, with commercial hybrids found between the two ancestral groups (Figure 4-2A). The projected position of a hybrid along PC1 was used to calculate its percentage *V. vinifera* ancestry (Figure 4-2B).

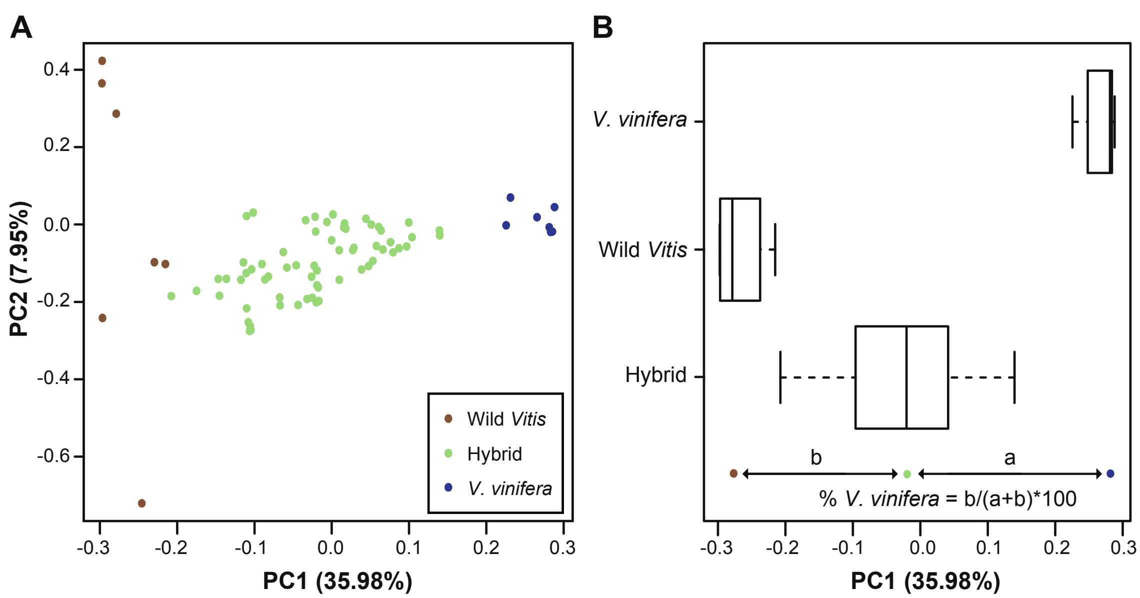


Figure 4-2. PCA-based ancestry estimation using 2482 SNPs and 56 indels for 7 wild *Vitis*, 7 *V. vinifera*, and 64 hybrid samples. (a) PCs were generated using wild *Vitis* and *V. vinifera* samples. The proportion of the variance explained by each PC is shown in parentheses along each axis. Hybrids were projected onto the axes. (b) Boxplots of PC1 values for wild *Vitis*, *V. vinifera*, and hybrid cultivars as well as a visual description of the calculation used for ancestry estimation. Further details are found in the Methods.

In order to evaluate the accuracy of our ancestry estimates, we performed *in silico* crosses between wild *Vitis* and *V. vinifera* populations using our genome-wide SNP data to simulate F1 hybrids as well as hybrids generated from F1 simulated hybrids backcrossed to *V. vinifera* or wild *Vitis*. The simulated progeny were projected onto PC axes determined using the ancestral populations and the resulting PCA plot is shown in Figure 4-3A.

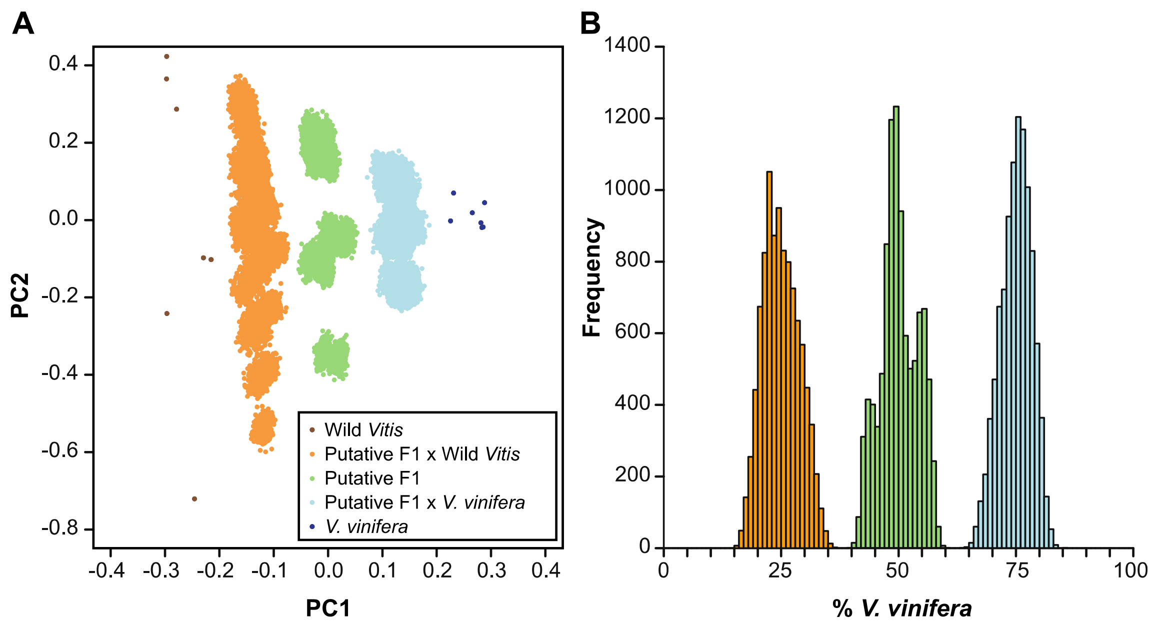


Figure 4-3. Simulation of hybrids (10,000 of each). (a) Simulated hybrids including F1 hybrids, F1 backcrossed to *V. vinifera* and F1 backcrossed to wild *Vitis* were projected onto axes generated using wild *Vitis* and *V. vinifera* samples (b) Distribution of ancestry estimates for simulated populations

The expected *V. vinifera* content in an F1 offspring with one *V. vinifera* and one wild *Vitis* parent is 50%, and the mean estimated content in the simulated F1 population described here was 50.1%, with a 95% confidence interval (CI) ranging from 42.7% to 57.2%. In progeny produced by an F1 hybrid backcrossed to wild *Vitis*, the expected *V. vinifera* content is 25%, which was the mean estimate of our simulated data, with a 95% CI of 18.4% to 32.6%. Finally, the mean *V. vinifera* content in simulated F1 hybrids backcrossed to *V. vinifera* is expected to be 75%, and our results have a mean value of

75.1%, with a 95% CI of 68.5% to 80.9%. The proximity of our simulated values to expected values provides support for the accuracy of our method, but it is worth noting that our 95% confidence intervals indicate that estimates may deviate by as much as 7-8% from the expected value. Moreover, the accuracy of our estimates may decrease in cases where crosses are generated from parents whose ancestry differs significantly from the samples used as ancestral populations in the present study. Ancestry estimates for simulated progeny are shown in Figure 4-3B.

Commercial Grape Ancestry Estimation

The distribution of *V. vinifera* content estimated for the hybrid grape cultivars examined in this work is found in Figure 4-4A, and the ancestry estimates for each cultivar are listed in Figure 4-4B.

Hybrids previously genotyped in Sawler et al. (2013) and replicated in this study using GBS include ‘Bertille-seyve 5563’ (DVIT 169), ‘Van Buren’ (DVIT 1129), ‘Rofar Vidor’ (DVIT 2258), DVIT 2180, ‘Jackson Sel. #3’ (DVIT 2916), and ‘Marechal Foch’ (California) (DVIT 214). The ancestry estimates for these samples differed by 2-5% from those previously estimated, with the exception of DVIT 2180 where our estimate of *V. vinifera* ancestry was 19% higher than in the previous work. DVIT 2180 is an unnamed accession simply identified as a *Vitis* species by the USDA. Given that the tissue for both studies was collected separately, the large difference in our estimates may be due to mislabelling or sample mix-up. Regardless of this discrepancy, the position of this sample in PC space confirms that it is indeed a hybrid sample (Figure 4-2A).

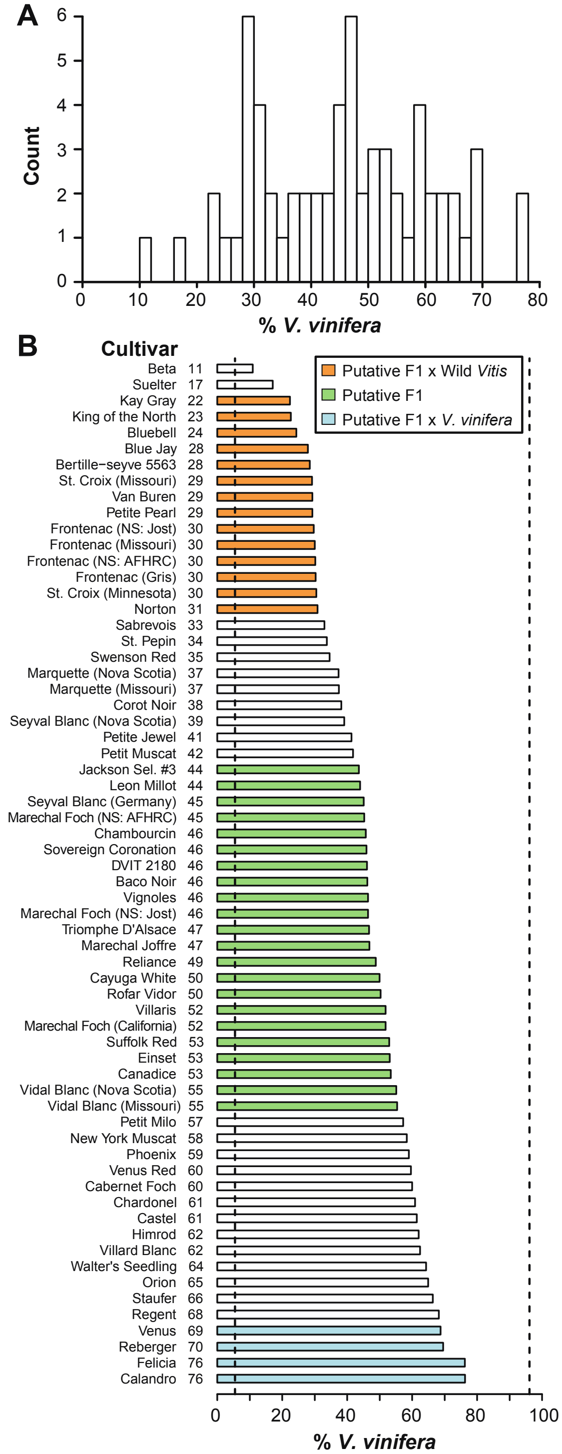
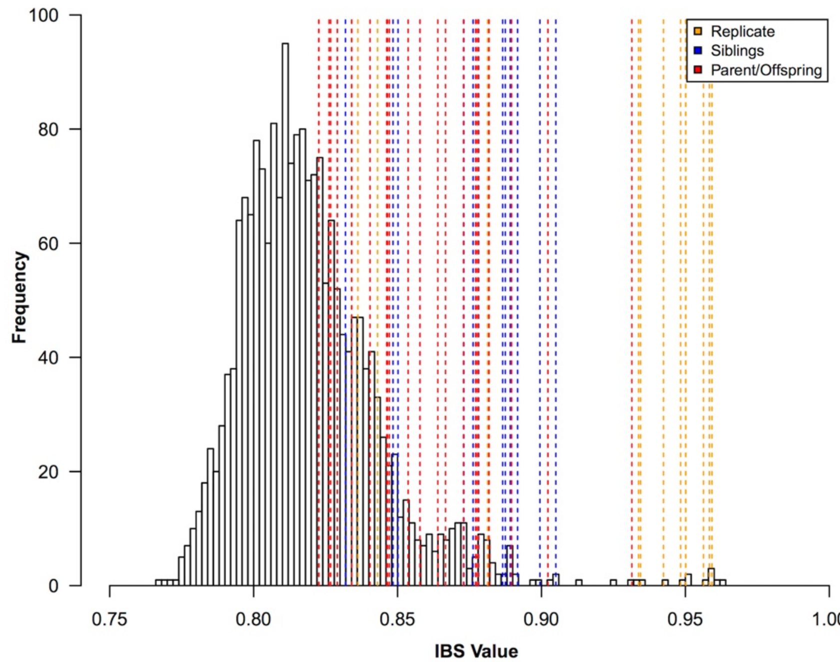


Figure 4-4. Estimated *V. vinifera* content in 64 commercial grape hybrids. Estimates are based on 2538 sites. (a) Distribution of *V. vinifera* ancestry estimates in hybrids (b) *V. vinifera* ancestry estimates for each cultivar. Bars are colored if a hybrid cultivar's ancestry estimate falls within the 95 % confidence interval of a F1, F1 x wild *Vitis*, or F1 x *V. vinifera* cross, based on simulated values. Dotted lines indicate mean values for the wild *Vitis* and *V. vinifera* samples

In order to further confirm the accuracy of our ancestry estimates, we compared *V. vinifera* ancestries inferred from well-known pedigrees to our genomics-based ancestry estimates. For example, 'Beta' is a cross between *Vitis riparia* and 'Concord', a *Vitis labrusca* cross thought to possess some *V. vinifera* ancestry due in part to its hermaphroditic flowers (Swenson, 1985; Cahoon, 1986). Sawler et al. (2013) estimated the *V. vinifera* content of 'Concord' as 31%. Based on these values, the percentage *V. vinifera* found in 'Beta' is expected to be approximately 16%, and it was estimated as 11% here (Figure 4-4A). 'Baco Noir' is a known F1 hybrid between 'Folle Blanc' (*V. vinifera*) and *V. riparia*, and therefore it is expected to be 50% *V. vinifera.* Our estimate is 46%, which falls within the 95% confidence interval of the *V. vinifera* ancestry estimates from our simulated F1 hybrid offspring. In these two cases, our genomics-based ancestry estimates are consistent with pedigree-based estimates.

Our study also included several cultivars collected from multiple locations, and the ancestry estimates were generally similar or equivalent for these replicates from different geographic regions. For example, 'Frontenac' sampled from two locations in Nova Scotia, Missouri, as well as a Gris sport, were all estimated to be 30% *V. vinifera.* 'Marquette' samples from both Nova Scotia and Missouri were estimated to contain 37% *V. vinifera*. However, the ancestry estimate (52%) for a 'Marechal Foch' accession retrieved from the USDA germplasm collection was 6% and 7% higher than the samples collected from two different locations in Nova Scotia. IBS values indicate that this sample is likely not the same cultivar as the 'Marechal Foch' grown in Nova Scotia (Figure 4-5). Still, all ancestry estimates of 'Marechal Foch' fall within the putative F1 range, which is expected given 'Marechal Foch' is the offspring of '101-14 Mgt.' (*V. riparia* x *V. rupestris*) x 'Goldriesling' (*V. vinifera*). 'Leon Millot' (44%) and 'Marechal Joffre' (47%) are siblings of 'Marechal Foch', and their ancestry estimates also fall within the range expected from an F1 hybrid (Figure 4-4B) (Pollefeys and Bousquet, 2003).

A



B

Cultivar 1	Cultivar 2	Relationship	IBS Value
Felicia	Vidal Blanc (Missouri)	Parent/Offspring	0.82
Cayuga	Seyval Blanc (Germany)	Parent/Offspring	0.83
Felicia	Vidal Blanc (Nova Scotia)	Parent/Offspring	0.83
Chardonel	Seyval Blanc (Germany)	Parent/Offspring	0.83
Villaris	Felicia	Sibling	0.83
Seyval Blanc (Nova Scotia)	Seyval Blanc (Germany)	Replicate	0.83
Cabernet Foch	Marechal Foch (California)	Parent/Offspring	0.83
Marechal Foch (Nova Scotia: Jost)	Marechal Foch (California)	Replicate	0.84
Phoenix	Villard Blanc	Parent/Offspring	0.84
Marechal Foch (Nova Scotia: AFHRC)	Marechal Foch (California)	Replicate	0.84
Beta	Bluebell	Parent/Offspring	0.85
Orion	Villard Blanc	Parent/Offspring	0.85
Staufer	Villard Blanc	Parent/Offspring	0.85
Leon Millot	Marechal Foch (California)	Sibling	0.85
Marechal Foch (California)	Marechal Joffre	Sibling	0.85
Beta	Blue Jay	Parent/Offspring	0.85
Beta	Suelter	Sibling	0.86
Cabernet Foch	Marechal Foch (Nova Scotia: Jost)	Parent/Offspring	0.86
Cabernet Foch	Marechal Foch (Nova Scotia: AFHRC)	Parent/Offspring	0.86
Reliance	Suffolk Grape	Parent/Offspring	0.87
Regent	Chambourcin	Parent/Offspring	0.87
Sabrevois	St. Croix (Minnesota)	Sibling	0.87
Phoenix	Orion	Sibling	0.88
Vidal Blanc (Missouri)	Villaris	Parent/Offspring	0.88
Vidal Blanc (Nova Scotia)	Villaris	Parent/Offspring	0.88
Petite Jewel	Canadice	Parent/Offspring	0.88
Cayuga	Seyval Blanc (Nova Scotia)	Parent/Offspring	0.88
Phoenix	Staufer	Sibling	0.88
Sabrevois	St. Croix (Missouri)	Sibling	0.88
Marechal Foch (Nova Scotia: AFHRC)	Marechal Foch (Nova Scotia: Jost)	Replicate	0.88
Himrod	Canadice	Parent/Offspring	0.88
Chardonel	Seyval Blanc (Nova Scotia)	Parent/Offspring	0.88
Orion	Staufer	Sibling	0.89
Leon Millot	Marechal Foch (Nova Scotia: Jost)	Sibling	0.89
Leon Millot	Marechal Foch (Nova Scotia: AFHRC)	Sibling	0.89
Sovereign Coronation	Himrod	Parent/Offspring	0.89
Marechal Foch (Nova Scotia: Jost)	Marechal Joffre	Sibling	0.89
Marechal Foch (Nova Scotia: AFHRC)	Marechal Joffre	Sibling	0.90
Leon Millot	Marechal Joffre	Sibling	0.90
Regent	Calandro	Parent/Offspring	0.91
Regent	Reberger	Parent/Offspring	0.93
St. Croix (Missouri)	St. Croix (Minnesota)	Replicate	0.93
Marquette (Missouri)	Marquette (Nova Scotia)	Replicate	0.93
Vidal Blanc (Nova Scotia)	Vidal Blanc (Missouri)	Replicate	0.94
Frontenac (Gris)	Frontenac (Missouri)	Replicate	0.95
Frontenac (Gris)	Frontenac (Nova Scotia: Jost)	Replicate	0.95
Frontenac (Nova Scotia: AFHRC)	Frontenac (Gris)	Replicate	0.95
Frontenac (Nova Scotia: AFHRC)	Frontenac (Missouri)	Replicate	0.96
Frontenac (Missouri)	Frontenac (Nova Scotia: Jost)	Replicate	0.96
Frontenac (Nova Scoia: AFHRC)	Frontenac (Nova Scotia: Jost)	Replicate	0.96

Figure 4-5. Distribution of IBS values for expected replicates (orange), siblings (blue) and parent/offspring (red). (a) Histogram of IBS values calculated in hybrid samples only. Dotted lines are drawn at values for expected first degree relationships as well as replicates. (b) Expected relationships between cultivars with their associated IBS values.

Within cultivar differences in ancestry estimates may be due partially to genotyping error. Curation error also leads to the mislabeling of samples and misidentification of cultivars. Previous work on *V. vinifera* cultivars from the USDA collection revealed widespread curation error (Myles et al., 2011), and recent work on the same collection found that the species names assigned to samples were incorrect in approximately 4% of cases (Sawler et al., 2013). In another example, three different Italian varieties all referred to as 'Bonarda' had no direct genetic relationship with each other (Martínez et al., 2008). Thus, curation error represents a likely source for the discrepancies we observe between samples with identical names.

While our data do not allow us to resolve first-degree relationships, we did examine the distribution of IBS values based on expected relationships derived from pedigree data (Figure 4-5). We found that, while many cultivars do share alleles in a manner that supports their expected relationship, several pairs of samples that are supposed to be either geographic replicates or first-degree relatives did not have IBS values consistent with their pedigrees. For example, the IBS value for 'Villaris' and 'Felicia' (0.83) was at least 0.02 lower than all other sibling pairs examined. Additionally, the 'Seyval Blanc' sampled from Germany does not resemble the 'Seyval Blanc' from Nova Scotia to the degree we expect. In both cases, the *V. vinifera* ancestry estimates also differed. Furthermore, 'Orion', 'Staufer' and 'Phoenix' are all progeny of crosses between 'Villard Blanc' (62%) and *V. vinifera* varieties, which has been confirmed by simple sequence repeat genotyping (Rudolf Eibach, personal communication). However, the expected ancestry for these progeny based on pedigree information should be higher (~81%) than what we observe (59%-65%). Further work is required in order to confirm potential sample mislabeling, cross-contamination, or genotyping error.

Wild Species Introgression

Often the best source for improvement of a crop plant is its wild relatives (Tanksley and McCouch, 1997). One crop that has benefited greatly from the use of wild relatives in breeding is tomato. Disease resistance in most commercial tomato cultivars is the result of genes introgressed from wild species (Foolad, 2007; Menda et al., 2014). However, recurrent backcrossing to elite varieties is performed for several generations in order to remove undesirable genes introduced from the wild relative (Menda et al., 2014). In tomato, it is customary to continue backcrossing to elite germplasm for 4 to 6 generations before the resulting hybrid is tested commercially (Bai and Lindhout, 2007).

In comparison to tomato, grape breeding appears to still be in its infancy. Approximately one third (22/64) of the hybrids analyzed in this study have *V. vinifera* content consistent with F1 hybridization (Figure 4-4B). Our results suggest that grape breeders have not extensively backcrossed with *V. vinifera* in order to introgress wild genes of interest. The distribution of *V. vinifera* ancestry across hybrids actually implies that backcrosses to wild *Vitis* species have been more frequent than backcrosses to *V. vinifera* during hybrid grape breeding. Breeders may have generated hybrids with high wild content when aiming to introgress numerous beneficial traits from wild relatives over a small number of generations. Further local ancestry estimates would be required in order to determine the number of generations of crossing.

The high number of hybrids consistent with F1 hybridization suggests that, overall, recent hybrid grape breeding has not followed standard breeding practices that aim to introgress desirable traits from wild species by repeatedly backcrossing to elite germplasm. Alternatively, because breeders often target numerous traits for introgression from the wild, the optimal *V. vinifera* content may be lower than the desired elite content in other crops. Ultimately, the crucial factor will be which desirable parts of each ancestral genome are captured, rather than the final *V. vinifera* percentage.

One instance where repeated backcrossing to V. vinifera has been exploited is in the development of Pierce's disease (PD) resistant wine grapes by tracking PD resistance alleles from the wild species *V. arizonica* through MAS (Riaz et al., 2009). Seedlings resistant to PD were repeatedly backcrossed to *V. vinifera*, resulting in progeny with 97% *V. vinifera* ancestry in the fifth generation, a value much higher than any estimates of commercial cultivars examined in this study (Walker et al., 2014). There are many more opportunities for desirable trais, such as cold hardiness, to be introgressed from wild *Vitis* species into novel elite cultivars (Zhang et al., 2015).

The use of molecular markers can also allow breeders to introgress multiple resistance genes into a single variety, a process called pyramiding (Joshi and Nayak, 2010). ‘Regent’ is a cross between ‘Diana’, a *V. vinifera* variety, and the hybrid grape ‘Chambourcin’, which has 46% *V. vinifera* ancestry according to our work. Based on these values, the expected *V. vinifera* ancestry of ‘Regent’ is approximately 73%, and our estimate is 68%. The complex pedigree of ‘Regent’ enabled the introgression of mildews and botrytis disease resistance from several *Vitis* species as well as high frost tolerance and early maturity (Eibach and Töpher, 2002). In 2013, ‘Regent’ ranked 12th in Germany according to total acreage (Ruehl et al., 2015). Recently, Regent was also crossed with VHR 3082-1-42 (*Muscadinia rotundifolia* x *V. vinifera*, then backcrossed four times with *V. vinifera)* to successfully combine powdery and downy mildew resistance genes into a single variety whose ancestry likely exceeds 80% *V. vinifera* (Eibach et al., 2007).

The Institute for Grapevine Breeding Geilweilerhof, which developed ‘Regent’, bred 6 of the 7 cultivars with the highest *V. vinifera* content in our study (Figure 4-4B). Thus, some breeders have produced hybrids with a high percentage of *V. vinifera* ancestry while retaining desirable characteristics from wild species. However, the overall lack of evidence for repeated backcrossing to *V. vinifera* in hybrid grape breeding indicates that grape breeders have yet to fully exploit the potential of combining key traits from wild species into novel cultivars with high *V. vinifera* content.

Conclusions

By examining the ancestry of 64 commercially grown grape hybrids using PCA-based ancestry estimation we found that approximately one third of hybrids have ancestry consistent with F1 hybridization: they derive half of their ancestry from wild *Vitis* and half from *V. vinifera*, suggesting that hybrid grape breeding is in its infancy. If backcrossing to *V. vinifera* was more widely adopted, we anticipate increased acceptance of hybrid grape varieties. Improved hybrid cultivars with higher *V. vinifera* ancestry could eventually lead to the relaxation of regulations against planting hybrid grapes, and ultimately a proliferation of grape cultivars with increased abiotic and disease resistance as well as favored wine qualities.

We anticipate our method can be extended to facilitate marker-assisted selection by allowing for offspring with the highest *V. vinifera* content to be selected at the seedling stage. In combination with MAS, ancestry estimates, such as those described here, can enable the continued improvement of grape by exploiting the diversity of wild *Vitis* species while maintaining desirable *V. vinifera* characteristics.

Acknowledgments

We would like to acknowledge Bruce Reisch for useful input. This article was written, in part, thanks to funding from the Canada Research Chairs program, the National Sciences and Engineering Research Council of Canada and Genome Canada. Z.M. was supported in part by a Killam Predoctoral Scholarship from Dalhousie University.

Chapter 5: Exploiting wild relatives for genomics-assisted breeding of perennial crops

Introduction

Perennials, or species that live for more than 2 years, include herbaceous plants, woody shrubs, and trees (Miller and Gross, 2011). Although most agriculturally important crops are annuals, perennials occupy over 13% of the world's surface area dedicated to food production (Table 5-1) (Food and Agriculture Organization of the United Nations, 2017). Not only are perennial crops a vital contributor to global food production and nutrition, but many offer advantages over annual crops. For example, perennial species generally have longer growing seasons (Dohleman and Long, 2009), increased root carbon (Glover et al., 2010a), and reduced soil erosion risk (Vallebona et al., 2016) when compared to annuals. As a result, there is increasing interest in perennializing annual grains (Glover et al., 2010b; Kane et al., 2016). While there are many benefits to growing perennials, breeding new cultivars is expensive and time-consuming due to the large size and lengthy juvenile phase of many species. For example, an avocado tree (*Persea americana* Mill.) may take up to 15 years to mature before flowering (Berg and Lahav, 1996). The recent breeding of 3 commercial apple (*Malus* X. *domestica* Borkh.) cultivars took 26 years (Peil et al., 2008), and thus, it is common for a limited number of elite cultivars to be propagated widely for long periods of time. For example, the 'McIntosh' apple is over 200 years old, while the 'Pinot Noir' grape (*Vitis vinifera* L.) has been grown for a millennium. Propagation of the same cultivars for decades—if not centuries—results in increasing susceptibility to disease, since these crops remain genetically frozen while pathogens continue to evolve (Myles, 2013). Over 75% of perennial crops are vegetatively propagated and the extensive use of a small number of elite cultivars fails to exploit the immense phenotypic and genetic diversity available (Miller and Gross, 2011). Expanding the breeding pool to include wild relatives can provide a crucial new source of desirable traits for introgression into perennial crops.

Table 5-1. The top 20 perennial crops based on total global area. Total global area, in million hectares, is listed as well as the proportion of total area (annuals and perennials) and proportion of perennial area for each of these crops. The total global area for all crops is estimated at 1335.37 million hectares, while the global area for perennial crops is estimated at 177.90 million hectares. Values are calculated based on the most recent available year of data from the Food and Agriculture Organization of the United Nations (2014) website (http://ww.fao.org/faostat). Crops we were unable to categorize due to ambiguous names which included both perennial and annual species were excluded.

Crop	Global Area (Million Hectares)	Global Contribution (%)	Perennial Contribution (%)
Sugar cane (*Saccharum* spp.)	27.1	2.03	15.2
Palm fruit (*Elaeis* spp.)	18.7	1.4	10.5
Coconuts (*Cocos nucifera* L.)	11.9	0.894	6.71
Rubber (*Hevea brasiliensis* (Willd. ex A.Juss.) Müll.Arg.)	11.1	0.831	6.24
Coffee (*Coffea arabica* L.)	10.5	0.785	5.89
Cocoa (*Theobroma cacao* L.)	10.4	0.781	5.87
Olives (*Olea europaea* L.)	10.3	0.769	5.77
Grapes (*Vitis vinifera* L.)	7.12	0.534	4
Pigeon peas (*Cajanus cajan* (L.) Millsp.)	7.03	0.527	3.95
Cashew (*Anacardium occidentale* L.)	6.04	0.452	3.39
Mangoes (*Mangifera indica* L.), mangosteens *(Garcinia* x *mangostana* L.), guavas (*Psidium guajva* L.)	5.64	0.423	3.17
Bananas (*Musa* spp.)	5.39	0.404	3.03
Apples (*Malus* X. *domestica* Borkh.)	5.05	0.378	2.84
Plantains (*Musa* spp.)	4.5	0.337	2.53
Oranges (*Citrus* spp.)	3.89	0.291	2.18
Tea (*Camellia* spp.)	3.8	0.285	2.14
Plums (*Prunus domestica* L.), sloes (*Prunus spinosa* L.)	2.52	0.189	1.42
Tangerines, mandarins, clementines, satsumas (*Citrus* spp.)	2.28	0.171	1.28
Almonds (*Prunus dulcis* (Mill.) D.A. Webb)	1.73	0.13	0.974
Pears (*Pyrus communis* L.)	1.57	0.118	0.885

Crop wild relatives (CWRs) provide an invaluable resource for improving perennial crops through disease resistance, fruit quality, and rootstocks. By 1997, improvements to crops due to CWRs had an estimated global benefit of $115 billion annually (Pimentel et al., 1997). However, the definition of what constitutes a ‘crop wild relative’ can be unclear, especially in perennial species where only a few generations of breeding may have occurred since domestication. For example, in kiwifruit (*Actinidia* spp.) almost all cultivars were either taken directly from the wild or are the result of only two-to-three generations of breeding. Commercial kiwifruit cultivars, including ‘Hayward’ (*Actinidia chinensis* var. *deliciosa* (A.Chev.) A.Chev.), the most widely grown cultivar, are very similar to wild plants (Ferguson, 2007; Ferguson and Huang, 2007). Likewise, cranberry (*Vaccinium macrocarpon* Ait.) cultivars are generally either wild selections or only a few generations removed (Fajardo et al., 2012). Finally, most banana cultivars (*Musa* spp.) are vegetatively propagated wild individuals collected by farmers due to the presence of parthenocarpic fruit which develop without seeds, pollination, or fertilization (Heslop-Harrison and Schwarzacher, 2007). When many elite perennial cultivars are in fact simply wild plants selected for cultivation with minimal improvement or domestication, the concept of CWRs becomes blurred.

While many cultivated perennial crops are essentially wild, even crops that have been bred for millennia are often not genetically distinct from their wild ancestors. To demonstrate this, we used genome-wide single nucleotide polymorphism (SNP) data to compare the primary progenitor and cultivated species of grape (*Vitis*) and apple (*Malus*) using principal component analysis (PCA) (Figure 5-1) (Myles et al., 2011, Gardner et al., submitted). Figure 5-1 suggests no clear differentiation between the domesticated *Vitis vinifera* and the wild progenitor, *Vitis sylvestris* and the same is true of the domesticated *Malus domestica* and its primary progenitor species, *Malus sieversii* (Ledeb.) M. Roem. This is consistent with previous analyses, which found evidence of gene flow between wild and cultivated grapes in Western Europe, as well as between wild and domesticated apples (Myles et al., 2011; Cornille et al., 2012). Thus, it is worth noting that the distinction between cultivated crop and CWR, or progenitor species, in perennial crops is often blurred, as there may be shared segregating polymorphism and ongoing gene flow after domestication. Nevertheless, the notion of introgressing wild traits into elite germplasm is applicable across a diverse range of perennial crops, even those without a clear distinction between wild and cultivated species.

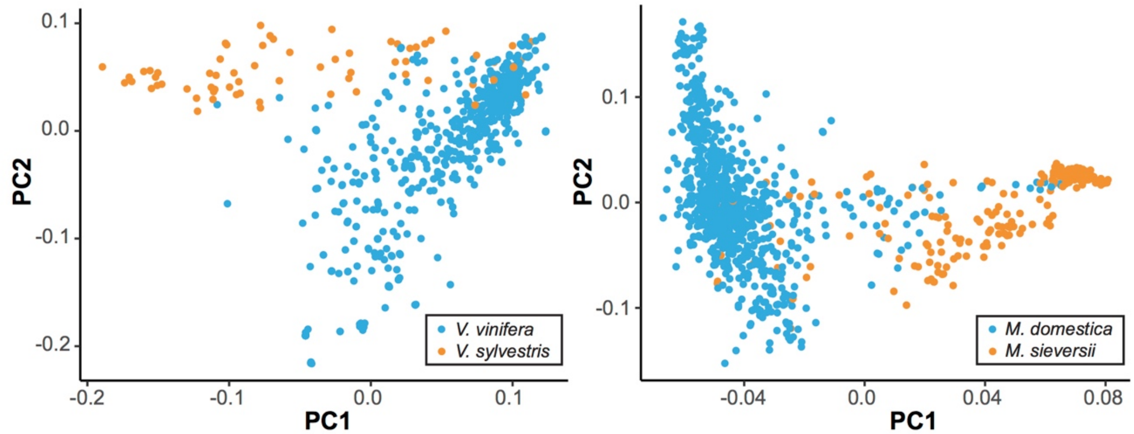


Figure 5-1. Lack of differentiation between wild and domesticated perennial species. By plotting the two major axes of variation against each other (i.e. PC1 vs PC2) we gain an overview of the genetic relatedness among samples. The primary wild ancestors and domesticated species cannot be clearly separated. PCA was performed using SNP data to compare primary progenitor species and cultivated accessions of grape and apple. Cultivated accessions, as labelled by the USDA, are indicated in blue, while the primary progenitor species are indicated in orange. Equal sample sizes were used for both species and additional samples were projected onto the PCA axes

Marker-assisted selection (MAS) can increase the efficiency of incorporating desirable traits present in wild germplasm into domesticated, or elite, cultivars. MAS relies on genetic markers that are either causal for, or strongly linked to, a phenotype. The primary benefit of MAS is the ability to select individuals possessing a trait of interest at the seed or seedling stage using genetic markers. MAS allows the breeder to eliminate plants that do not possess the desired trait and may otherwise require a decade of cultivation to assess phenotypically. Instead, resources and space can be dedicated only to individuals with the desired characteristic. Plants with the desired trait can then be backcrossed to elite germplasm to maintain the wild trait of interest, while preserving important commercial traits (Figure 5-2).

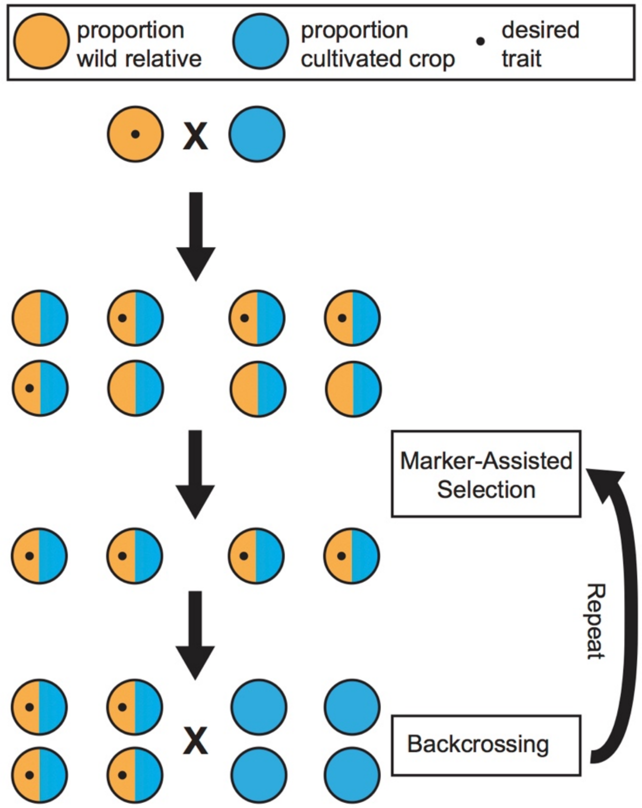


Figure 5-2. Schematic of breeding using MAS. Wild relatives containing a trait of interest are crossed with a cultivated crop. In this example, the wild parent is heterozygous for a dominant Mendelian trait. With a marker associated with this trait, offspring can be screened for the trait and eliminated at the seedling stage. MAS ensures that the trait of interest is present in the progeny through several generations of backcrossing. Not shown here is that, with each generation, there is an increase in the proportion ancestry derived from the cultivated compartment while maintaining the desirable wild trait.

Backcrossing to elite germplasm is crucial to ensuring traits of agricultural importance are maintained when breeding with wild relatives: the goal is to retain all desirable characteristics of the elite cultivars while introducing only the small number of desirable loci from the wild. However, a genomic assessment of wild ancestry in over 60 commercially grown hybrid grape cultivars found that one third had ancestry consistent with F1 hybridization. In fact, the study demonstrated that backcrosses to wild *Vitis* were more frequent than backcrosses to *V. vinifera*, indicating that repeated backcrossing to elite germplasm is not yet widely practiced (Migicovsky et al., 2016b). Breeding of perennial crops using wild relatives is still in its infancy. Through use of MAS and repeated backcrossing we can anticipate new, superior cultivars possessing useful traits from wild relatives while still maintaining the desirable characteristics of elite cultivars.

In addition to saving time, MAS can decrease the cost of perennial breeding using wild relatives. When compared to traditional fruit breeding, MAS was estimated to save up to 43% of operational costs over the first 6 to 8 years of an apple breeding program (Edge-Garza et al., 2015). MAS eliminates the need to phenotype and therefore offers the greatest cost and time savings for traits that may be difficult or expensive to measure, such as disease resistance, as well as traits expressed late in development, such as fruit quality (Töpfer et al., 2011). This review addresses the current use, future potential, and limitations of using wild germplasm for genomics-assisted breeding in perennial crops.

Benefits: disease resistance

The majority of perennial crops are vegetatively propagated for decades, or even centuries, and are increasingly susceptible to evolving pathogens (Miller and Gross, 2011). In contrast, wild relatives have undergone natural selection in response to disease pressure and often harbor crucial resistance genes, which can be exploited through breeding. The monogenic nature of many resistance genes means MAS is especially feasible for introgression of disease resistance loci. Indeed, in a review of 19 different crops, over 80% of the traits incorporated from CWRs were involved in disease and pest resistance (Hajjar and Hodgkin, 2007). Similarly, another review of 104 MAS studies from 1995 to 2012 found 74% focused on disease and pest resistance (Brumlop et al., 2013). This demonstrates the widely-acknowledged potential of improving crops through introgression of disease resistance traits from wild relatives.

The introgression of disease resistance from wild germplasm is perhaps best exemplified by modern grape breeders. While the genus *Vitis* contains over 60 inter-fertile species, approximately 99% of the world's vineyards are planted with a single species, *V. vinifera* (This et al., 2006; Anderson and Aryal, 2013). While not commonly grown for commercial purposes, wild *Vitis* relatives possess many desirable traits not found within

V. vinifera. For example, the effects of Pierce's disease (PD) (*Xylella fastidiosa*) cost the California wine industry approximately $92 million annually (Alston et al., 2013). *Vitis arizonica* Engelm., a wild grape, is resistant to PD and has been used to develop PD-resistant wine grapes. MAS allows breeders to track PD resistance while backcrossing offspring repeatedly to *V. vinifera*. Breeding lines now possess PD resistance as well as 97% *V. vinifera* ancestry (Walker et al., 2014). Thus, MAS can facilitate the introgression of disease resistance from wild relatives while allowing for progeny that maintain desirable quality traits due to a high proportion of domesticated ancestry.

In addition to facilitating introgression of a single source of disease resistance, MAS is a valuable tool for introgression of several sources of resistance to the same disease, or even resistance to multiple diseases, through a process called pyramiding. For example, a *Muscadinia rotundifolia* (Michx.) Small x *V. vinifera* cross was backcrossed 4 times to *V. vinifera*, resulting in the progeny 'VHR 3082-1-42' (Pauquet et al., 2001). 'VHR 3082-1-42' was then crossed with 'Regent', a hybrid grape variety that is approximately 68% *V. vinifera*. The resulting progeny possess both powdery and downy mildew resistance genes from wild relatives as well as *V. vinifera* ancestry that likely exceeds 80% (Eibach et al., 2007; Migicovsky et al., 2016b). Wild grape species are also resistant to diseases such as black rot, crown gall, and others, all of which provide the opportunity for further improvement of commercial grape cultivars (Owens, 2008). The use of MAS to pyramid either several sources of resistance to a single disease, or resistance to multiple diseases, into a single cultivar is in its infancy. However, pyramiding of disease resistance markers promises to eventually result in grapes which require less chemical input to grow but still possess other commercially desirable traits.

There is also great potential for MAS in improving disease resistance in apple breeding programs. Apple scab (*Venturia inaequalis*) is one of the most destructive diseases in apple (*M. domestica*) and may require 20-30 fungicide treatments per season in commercial orchards. The wild relative *Malus floribunda* Siebold ex Van Houtte is widely used as a source of apple scab resistance. However, the resistance offered from *M. floribunda* is ineffective against certain strains of apple scab and a broader base of resistance is needed (Parisi et al., 1993; Soriano et al., 2009). Fortunately, resistance genes from several other wild relatives, including *Malus baccata jackii* (Gygax et al., 2004) and *Malus micromalus* Makino. (Patocchi et al., 2005), have been identified. Recent work used MAS to pyramid three scab resistance genes as well as genes for powdery mildew (*Podosphaera lecotricha*) resistance and enhanced fire blight (*Erwinia amylovora*) resistance into a single apple tree (Baumgartner et al., 2015). Many other desirable traits, including both abiotic and biotic stress resistance, are also found in wild *Malus* species (Volk et al., 2015). Thus, cultivars are being developed that contain ancestry from several wild relatives, each contributing desirable alleles to achieve the breeder's target. However, the achievements of breeding programs that have successfully exploited numerous wild perennial species are not yet widespread. The use of wild diversity will only increase in importance as pathogens continue to evolve.

In many instances, there is a great urgency to identify and exploit sources of disease resistance. In the case of banana, 'Cavendish' cultivars, were first grown due to their resistance to *Fusarium* wilt (*Fusarium oxysporum* f. sp. *cubense*) (Heslop-Harrison and Schwarzacher, 2007). Over 40% of bananas produced worldwide are 'Cavendish' cultivars and there are now reports of an evolved form of the pathogen to which it is susceptible (Hwang and Ko, 2004; Ploetz et al., 2007). Once infected with *Fusarium* wilt, the disease cannot be controlled and banana plants must be replaced with a new, resistant cultivar (Daly and Walduck, 2006). Resistant wild banana populations which co-evolved with the pathogen have been found and offer a valuable source of resistance to the newly evolved and highly pathogenic forms of *Fusarium* wilt (Javed et al., 2004). Recently, a marker for *Fusarium* wilt susceptibility with a discriminatory power of 93% was developed (Cunha et al., 2015). MAS is likely to facilitate the development of new resistant cultivars that will eventually replace the 'Cavendish' banana. It is possible for a single virulent strain to devastate an entire industry, and efforts to exploit wild relatives will become critical if the evolvability of pathogens is ignored.

The intense pressure to rapidly develop new, disease-resistant cultivars is not exclusive to the banana industry. Cacao (*Theobroma cacao* L.), used in the production of chocolate, is a perennial tree native to South America. As a result of disease outbreak in South and Central America over the past 200 years, 70% of the world production now occurs in Africa, 10% in Asia and only 20% in South America (Brown et al., 2005). Brazil, once the third largest producer of cacao, became a net importer of the crop following the arrival of *Moniliophthora perniciosa*, which causes Witches' broom disease (Meinhardt et al., 2008). The use of a small number of cacao cultivars has left the crop vulnerable to disease and requires the continued expansion of production regions. However, pathogens continue to move to new cacao plantations. The only viable longterm solution in cacao, like banana, is the development of new, disease-resistant cultivars. Fortunately, wild populations of cacao still exist and have evolved in the presence of these pathogens. These wild relatives can be easily crossed with cultivated varieties, using molecular markers to accelerate the breeding process (Brown et al., 2005; Meinhardt et al., 2008; Zhang and Motilal, 2016). The recent evaluation of 520 wild cacao trees for important traits such as disease resistance, bean quality and flavor will provide a valuable resource for future breeding (Zhang and Motilal, 2016). The cacao industry's renewed focus on wild diversity serves as a warning to others who have yet to face the challenges that arise from evolving pathogen pressures. Only by establishing, maintaining, and evaluating diverse germplasm collections will the sources of pathogen resistance required in the future be readily available to breeders.

Benefits: fruit quality

While most crop wild relatives don't taste very good, they may still possess unique fruit quality traits that can be incorporated into domesticated germplasm to create novel cultivars. Prior to the use of genomics, the fruity and aromatic "foxy" flavor found in the wild grape, *Vitis labrusca* L., was introgressed into the domesticated grape, *V. vinifera*, for use in table grapes (Reisch et al., 2012). North Americans now commonly associate foxiness with "grape flavor", especially in confectionary products. Although wild relatives are exploited primarily for their disease resistance, in some cases unique fruit characteristics possessed by wild relatives, but absent from cultivated germplasm, are targeted by breeders as well.

The appearance of a fruit, including color, is a critical breeding target in many fruit species. Most kiwifruits (*A. chinensis*) have green or yellow flesh, but red flesh is highly valued by consumers (Harker et al., 2007). The first red-fleshed commercial cultivar in the Chinese market, ‘Hongyang’, required 20 years of breeding and selection to produce (Wang et al., 2002). Only a few red-fleshed kiwifruits have been collected for use in breeding. Wild kiwifruit with red flesh, including both *A. chinensis* and other *Actinida* species, remain largely unexploited (Sui et al., 2013). Genomic work has begun in an effort to develop markers to easily identify red-fleshed kiwifruit. The identification of genetic markers for red flesh from wild relatives would allow breeders to select for this trait in kiwifruit, while minimizing the influence of any negative wild characteristics through repeated backcrossing (Wang et al., 2012). Fruit characteristics, such as color, are only visible in perennial crops after a juvenile phase and provide an excellent example of the potential for MAS to reduce the cost of breeding by allowing breeders to eliminate plants which do not possess the trait at an early stage. Reducing the cost of breeding through genomics can facilitate the development of more cultivars possessing unique fruit characteristics from wild relatives.

In addition to fruit appearance, improving nutritional qualities such as antioxidant capacity is an area of major interest, especially in raspberry (*Rubus idaeus* L.) and blackberry (*Rubus* spp.) breeding. A comparison between wild and cultivated raspberries found the highest antioxidant capacity in *Rubus caucasicus* Focke, indicating the potential of increasing antioxidants in commercial cultivars through use of this species in breeding (Deighton et al., 2000). Similarly, work on blackberries found that wild genotypes had much higher levels of a key antioxidant than a commercial cultivar. Therefore, wild blackberries may be of use to breeding programs aiming to increase antioxidant content (Cuevas-Rodriguez et al., 2010). Raspberry and blackberry are just two examples of perennial crops which could benefit from breeding with wild relatives for desirable nutritional qualities, and it will be interesting to see how quickly—if at all—genomics-assisted approaches are adopted in these cases.

As evidenced by the examples provided in this review, MAS is incredibly useful for tracking traits that a breeder aims to introgress from wild relatives. However, in most cases, MAS will be used to maintain desirable traits from cultivated ancestors, rather than to introduce desirable quality traits from the wild. For example, in apple MAS is already in use for traits such as postharvest storability, firmness, acidity and skin color (Ru et al., 2015). When introgressing disease resistance from a wild relative, a breeder wants to retain only progeny with the desired fruit quality traits from the elite parent. Markers can be used to simultaneously track these desirable traits from the elite parent and the disease resistance from the wild parent. Thus, genomics is a valuable tool that enables breeders to efficiently select for the benefits offered by both wild and cultivated germplasm.

Benefits: rootstocks

A primary use of wild relatives in perennial breeding thus far has been for the development of rootstock varieties. Vegetatively propagated woody perennial crops are often shoots, or scions, grafted onto wild or hybrid rootstocks. Rootstocks can be used to improve perennial crops both above and below ground. Above ground, rootstocks can confer unique traits to the scion, such as precocity, or the reduction of time until a tree bears fruit, as well as the dwarfing of large trees. Below ground, targeted rootstock traits include drought tolerance, salt tolerance and disease resistance (Warschefsky et al., 2016). While use of MAS in rootstock breeding has been limited to date, genomics can further improve rootstocks by facilitating the use of wild germplasm.

Given that most perennial crops are clonally propagated from a small number of elite cultivars, increased ease of travel and evolving pathogens pose a dangerous threat both to the scion as well as the portion of the plant found below ground. In the 1860s, the North American phylloxera aphid (*Phylloxera vastatrix*) devastated European vineyards. By attacking the roots of the plant, phylloxera kills *V. vinifera* vines within one to two years. Breeders used American wild *Vitis* species to develop resistant rootstocks, rescuing the wine industry, and *V. vinifera* wine cultivars are still grafted onto these rootstocks today (Alleweldt and Possingham, 1988; Zhang et al., 2009). Currently, bacterial canker

(*Pseudomonas syringae* pv. *actinidiae*) poses a serious threat to kiwifruit worldwide. Fortunately, resistance to bacterial canker has been found in wild Chinese kiwifruit germplasm. A recent interspecies cross between wild *Actinidia eriantha* Benth. and cultivated *A. deliciosa* resulted in a rootstock resistant to bacterial canker. The same work discovered a genomic marker potentially useful for identifying bacterial-canker resistant hybrid rootstocks (Lei et al., 2014). As pathogens continue to evolve and spread, wild germplasm will be indispensable for use in rootstock breeding. Discovery of disease-resistant rootstocks using MAS can allow for the continued use of commercially successful scions while protecting the plant from diseases below ground.

Of the 25 most-produced fruit and nut crops, 20 may be grafted onto rootstocks, including grape and walnut (*Juglans regia* L.). The 5 crops not grafted are all monocots where grafting is not possible (Warschefsky et al., 2016). Given the global value of grafted perennial crops, the breeding of superior rootstocks is an area of great importance. While several generations of backcrossing may be necessary when crossing wild relatives with commercial scions to maintain fruit quality, wild trait introgression in rootstocks can be accomplished in fewer generations because the fruit quality of a rootstock cultivar is irrelevant. Use of wild relatives is further facilitated by graft compatibility between more distant relatives. For example, many stone fruits can be budded onto rootstocks developed for other *Prunus* species (Beckman and Lang, 2002). Peach (*Prunus persica* (L.) Batsch) and almond (*Prunus amydalus* (Mill.) D.A. Webb) x peach hybrid rootstocks with resistance to root-knot nematodes (*Meloidoyne* spp.) as well as adaption to calcareous soil have been released. Both peaches and almonds, as well as some plum and apricot cultivars, can be grafted onto these rootstocks (Felipe, 2009). However, the most widely used rootstocks in almonds are still susceptible to lesion and ring nematodes, crown gall, and bacterial canker. The National Clonal Germplasm Repository (NCGR) of the United States Department of Agriculture Agricultural Research Service (USDA-ARS) in Davis, California is using almond relatives such as peach, wild almond species, and plums as potential donors for disease resistance and drought tolerance (Aradhya et al., 2015). The phenotypic evaluation of wild relatives, in combination with genomic data, enables the identification of markers linked to these desirable traits, allowing for donors to be efficiently selected. The development of disease resistant rootstocks through MAS using wild relatives is a topic of intense research interest in several perennial crops, and it is anticipated that this pursuit will result in substantial reductions in chemical input.

In addition to almond, the USDA-ARS is performing research on the use of wild relatives as potential sources of disease resistance for rootstocks in walnut. The primary walnut rootstock is ‘Paradox’, a California black walnut *Juglans hindsii* x cultivated walnut *J. regia* hybrid which is tolerant of wet soil conditions, but susceptible to crown gall (*Agrobacterium tumefaciens*) (Hasey et al., 2013; Aradhya et al., 2015). Promising sources of disease resistance to crown gall and *Phytophthora* rots have been identified in wild species such as the North American black walnut (*Juglans hindsii*, *Juglans major*, and *Juglans microcarpa*) and Asian butternut species (*Juglans cathayensis* Dode and *Juglans ailantifolia* Carr.). Mapping populations are currently being developed to identify disease resistance markers for MAS in walnut rootstocks (Aradhya et al., 2015). While molecular markers have not yet been used extensively in rootstock breeding, MAS can be used to screen hybrid progeny at a reduced cost and without the need to expose plants to pathogens in order to determine resistance status. Additionally, many of the traits important for rootstock breeding, such as disease resistance, precocity, and dwarfing of the scion are targeted and defined. In comparison, in scion breeding far more complex traits, such as overall fruit quality, may be targeted. In turn, desirable rootstock traits are more likely to be controlled by a small number of genetic loci with large effects. Thus, the simple genetic architecture of most rootstock traits makes them amenable to genetic mapping as well as MAS. While wild relatives have long been viewed as a valuable tool for rootstock breeding, combining such benefits with genomics-assisted approaches is the crucial next step.

Genomic resources and limitations: mapping and breeding

Despite the promise of wild relatives for improvement through MAS in perennial crops, there are several challenges to consider. In order to make use of wild relatives for MAS, the first step is to discover markers for traits of interest. Genome-wide association studies (GWAS) and linkage mapping are two methods used to establish genotype-phenotype relationships. GWAS relies on differences within a population of diverse, unrelated individuals in order to discover correlations between markers and traits. In comparison, linkage mapping exploits bi-parental crosses to map traits in the resulting progeny. One of the main advantages of GWAS over traditional linkage mapping is its superior mapping resolution. GWAS markers correlated with a phenotype are likely to be very close to the causal locus. In some cases, the likely causal genetic variant itself can be identified through GWAS (Migicovsky et al., 2016a). In linkage mapping, large genomic intervals, often spanning millions of nucleotides, are identified while the causal genetic variant is unlikely to be pinpointed. GWAS is particularly promising in perennials because of the time and cost required to generate bi-parental crosses. An additional benefit is that GWAS can be applied to germplasm collections that are already in the ground and waiting to be exploited (Chitwood et al., 2014). The discrepancy in mapping resolution between the two methods is a function of the number of recombination events captured by each method. In GWAS, a large number of unrelated individuals means that a large number of recombination events have occurred in the history of the genetic material being assessed. In linkage mapping, only the recombination events captured through the generation of the bi-parental cross can be exploited, resulting in relatively large chunks of DNA that share co-ancestry among individuals.

The high mapping resolution offered by GWAS is amplified in many perennials because of the relatively rapid linkage disequilibrium (LD) decay in high-diversity perennial crops. For example, LD decays within 200 bp in grape (Lijavetzky et al., 2007) and within 100 bp in apple (Migicovsky et al., 2016a) and Norway spruce (*Picea abies* (L.) H.Karst.) (Heuertz et al., 2006). This level of LD decay is far more rapid than in diverse populations of most well-studied annuals like rice (*Oryza sativa* L., ~75 to >500 kb; Mather et al., 2007), maize (*Zea mays* L., 1 to 10 kb; Yan et al., 2009), and soybean (*Glycine max* L. Merr., 336 to 574kb; Hyten et al., 2007). The correlation between a marker and a causal variant is related to the level of LD between the two: the higher the LD, the more likely the marker will serve as an indicator for the presence of the causal variant. While rapid LD decay results in high mapping resolution, it also means that a very high density of markers is required for effective GWAS because the correlation among markers surrounding the causal variant decays so quickly. In some cases, generating sufficient coverage for GWAS by saturating the genome with markers may be prohibitively expensive due to rapid LD decay. However, the cost of marker discovery and genotyping is likely to continue to decrease, and it will therefore surely be feasible in the future for researchers to acquire the genotype data required for effective GWAS.

While GWAS in perennials is an attractive option, it is not always viable. Traits targeted by breeders are often present only within a wild relative species, and are completely absent within cultivated germplasm. Attempts to map such a trait in a population composed of the wild relative and the cultivated germplasm using GWAS would be futile because the trait co-segregates perfectly with ancestry. The marker you aim to uncover will be present in the wild relative but absent in the cultivated germplasm, but that is also the case for millions of other markers across the genome (Figure 5-3). When the phenotypes are perfectly segregated, GWAS is of no help and a bi-parental cross between the wild and cultivated populations must be made to genetically map the trait. Linkage mapping in the resulting bi-parental population allows for such co-segregating traits to be genetically mapped, because the confounding effects of population structure are broken through crossing. Thus, when mapping traits of interest found only in wild relatives, linkage mapping studies may be necessary due to co-segregation. However, it is sometimes the case that wild and domesticated germplasm share segregating polymorphism and are not significantly genetically differentiated, as is the case with apples and grapes (Figure 5-1). In such instances, the confounding effects of co-ancestry may not be too severe and GWAS may be the genetic mapping option of choice. Additionally, when a phenotype is not perfectly co-segregated with ancestry, but rather differentially expressed in the two populations, it may be possible to perform GWAS using wild and domesticated plants. In this scenario, including both population structure and the SNP-by-population interaction in the GWAS model would help avoid false positives and ensure that SNPs are consistently associated with the trait across wild and domesticated populations (Biscarini et al., 2010). For each crop and phenotype of interest, the optimal genetic mapping approach, and the desired genetic composition of the population, will vary.

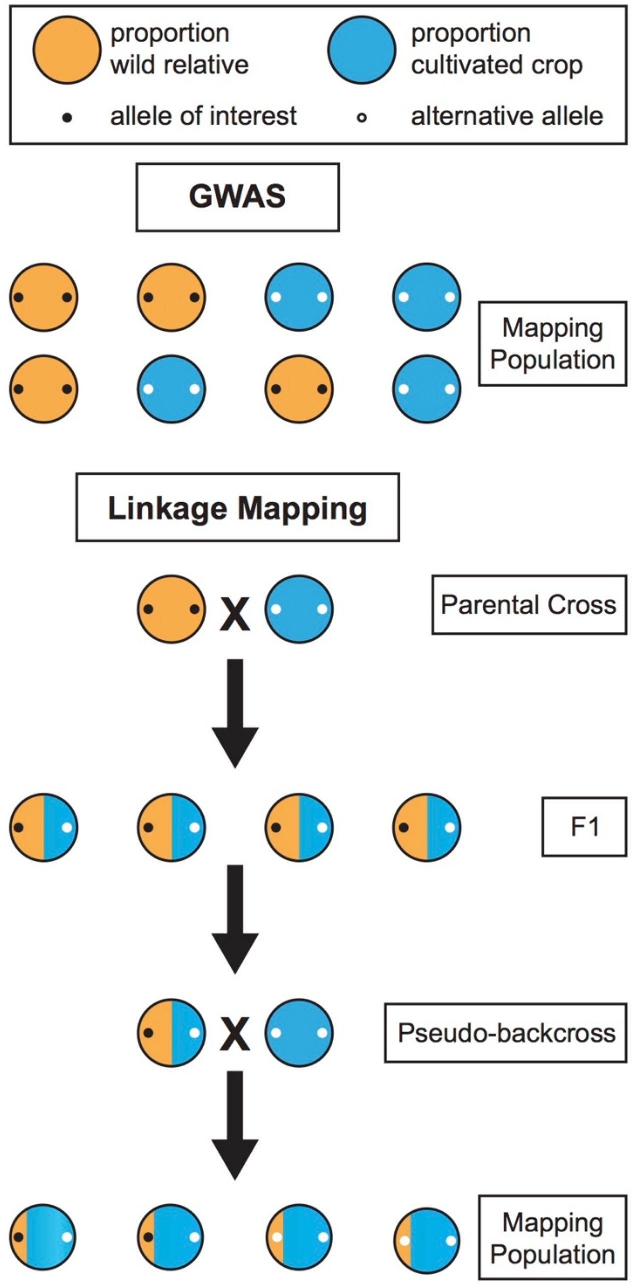


Figure 5-3. Comparison of the effectiveness of GWAS and linkage mapping for mapping alleles of interest in wild relatives. When an allele of interest is found only in wild germplasm it co-segregates with population structure and cannot be mapped using GWAS. Linkage mapping provides a viable alternative for mapping traits in wild relatives. However, in the F1 generation, alleles homozygous for alternative states in the wild and cultivated parent will not segregate. Thus, a backcross, or pseudo-backcross, is required to map most alleles of interest.

Linkage mapping provides a viable alternative to GWAS for co-segregating traits. In annual crops, it is typically performed crops through a cross of highly homozygous parents, often as a result of selfing. In perennials, the severe inbreeding depression and high level of heterozygosity requires a mapping design in which parents are not selfed. As an alternative, the two-way pseudo-testcross design, in which two highly heterozygous parents are crossed, has been successfully applied in many perennials, beginning in 1994 with an analysis of an interspecific *Eucalyptus grandis* W. Hill x *Eucalyptus urophylla* S. T. Blake cross (Grattapaglia and Sederoff, 1994). However, the progeny resulting from a two-way pseudo-testcross will not segregate for markers homozygous for alternative alleles in the parental plants (Figure 5-3). Given that many wild traits of interest will likely fall into this category, mapping will require at least one generation of backcrossing before linkage mapping can be applied. However, many perennials also have high levels of inbreeding depression, so close relatives cannot be used when performing backcrosses. Instead, a cultivar that is not one of the parents from the initial cross should be used to perform pseudo-backcrossing. The combination of a two-way pseudo-testcross design and pseudo-backcrossing can enable the detection of markers for valuable traits in wild perennial relatives.

When introgressing regions of the genome associated with a phenotype, or quantitative trait loci (QTL), from wild germplasm, linkage drag may lead to undesirable phenotypes in the resulting progeny. Linkage drag is the result of unfavorable genes linked to a desirable QTL also being incorporated into the domesticated germplasm (Varshney et al., 2014). Additional generations of pseudo-backcrossing can reduce the effects of linkage drag. If undesirable loci are tightly linked to the locus of interest, it may be difficult to eliminate the impact of linkage drag through conventional breeding. Fine-mapping of a QTL can allow for the selection of individuals with specific recombination events that minimize linkage drag. Unfortunately, fine mapping requires generating a large number of crosses for sufficient recombination (Khan and Korban, 2012). Reduced recombination frequencies have also been reported surrounding loci introgressed for resistance from a related species, such a 25-fold reduction in poplar (*Populus* spp.), providing further evidence that large populations will likely be needed for fine mapping (Stirling et al., 2001). As a result, the fine-mapping process is both expensive and time-consuming (Khan and Korban, 2012). Once a recombinant individual is identified, they can be used as a donor in breeding and backcrossing can continue for several generations using MAS.

In addition to eliminating linkage drag, fine-mapping may lead to the identification of causal alleles which can be subsequently incorporated into the genomes of domesticated crops through genetic modification (GM) or genome editing techniques. These techniques can be applied directly to the cultivar of interest, immediately incorporating the trait, and does not require multiple generations of backcrossing to eliminate linkage drag. This is especially valuable in perennial crops with a lengthy juvenile phase or infertile hybrid progeny. Previous work successfully generated transgenic bananas with resistance to *Fusarium* wilt, the major pathogen threatening banana production (Paul et al., 2011). Simiarly, transgenic plantains (*Musa* spp.) with resistance to nematode pests *Radopholus similis* and *Helicotylenchus multicinctus* have been developed (Tripathi et al., 2015). In papaya (*Carica papaya* L.), the limiting production factor is the papaya ringspot virus (PRSV). While there have been attempts to transfer PRSV resistance from related wild *Vasconcellea* species to *C. papaya*, initially only F_1 hybrids were possible as the resulting offspring were often infertile, preventing further backcrossing to *C. papaya* (Gonsalves et al., 2006). The first successfully backcrossed PRSV-resistant papaya was only reported in 2011, after 50 years of attempts (Siar et al., 2011). Instead, for almost two decades, papaya with transgenic resistance to PRSV have been cultivated in Hawaii (Gonsalves et al., 1998; Suzuki et al., 2007). Thus, GM is a valuable tool that can expedite the breeding of disease-resistant cultivars.

Currently, the most promising genome editing technique is CRISPR/Cas9, which is simple, flexible and efficient. CRISPR/Cas9 has been successfully employed in perennial species including apple (Nishitani et al., 2016) and sweet orange (*Citrus sinensis* (L.) Osbeck) (Jia and Wang, 2014). Clearly, incorporation of desirable traits from wild relatives into perennial crops is not limited to MAS, but can also be achieved through GM. However, the social and regulatory acceptance of GM crops, including papaya outside of Hawaii, is often limited (Davidson, 2008). Acceptance of GM perennials is especially difficult since many are fruit crops that are consumed fresh. However, CRISPR/Cas9 may result in modified crops acceptable to those opposed to traditional GM techniques. For example, in 2015, Sweden confirmed that some plants edited using CRISPR/Cas9 were not considered GMOs under the European definition (Wolter and Puchta, 2017). One method for achieving further acceptance of crops modified using CRISPR/Cas9 is to avoid the use of foreign DNA, as recently achieved in maize (Svitashev et al., 2016) and bread wheat (*Triticum aestivum* L.) (Liang et al., 2017). Until global acceptance of CRISPR/Cas9 occurs, MAS continues to be a useful genomic tool for the introgression of desirable traits. Additionally, unlike genome editing, MAS remains useful when precise detection of causal loci is not possible and only markers highly correlated with the trait of interest are available.

The simple distinction between GWAS and linkage mapping is useful, but experimental designs that blur this distinction, and exploit the benefits of both methods, are uncovering numerous genotype-phenotype associations. For example, a Multi-parent Advanced Generation InterCross (MAGIC) population is created by intercrossing multiple parental lines rather than a single bi-parental cross. The increased level of recombination in the progeny allows for improved precision of mapping using inbred offspring (Cavanagh et al., 2008). In perennials, where the creation of inbred lines is often not possible, other designs have been implemented. For example, work in apple made use of a factorial mating design consisting of 4 female parents and 2 pollen parents (Kumar et al., 2012b). This family-based design allowed for the discovery of markers for traits such as fruit firmness, internal browning and titratable acidity, which could be implemented in MAS (Kumar et al., 2013a). Therefore, alternative mating designs are a promising tool for increased mapping resolution when performing linkage mapping between wild and domesticated crops.

The limited diversity—often a single bi-parental cross—exploited in traditional linkage mapping results in a mapping population where many QTL will not segregate and therefore not be detected. Further, due to a potentially small population size, small-effect

QTL may not exceed the significance threshold. Significant markers identified are often only relevant to populations that share significant co-ancestry with the parents of the bi-parental mapping population. Thus, in comparison to GWAS, markers discovered using linkage mapping may not be predictive in diverse collections of germplasm (Owens, 2011). However, when identifying a marker for a trait from a wild relative, it is only necessary that the marker functions within that population, as a single source can be used as a donor for MAS. For example, while several sources of PD resistance have been used in grape breeding, the most important donor has been from a single *V. arizonica* accession, b43-17, which likely hybridized with *Vitis candicans* and is homozygous for monogenic resistance (Walker et al., 2014). Given that a single wild individual possessing a desirable trait is often sufficient for introgression into elite cultivars through MAS, transferability is of limited concern when exploiting alleles derived from a single wild relative.

A form of genomics-assisted breeding that is increasingly being used for complex traits is genomic selection (GS). GS is particularly useful when the breeder aims to predict a complex trait controlled by numerous QTL. In these cases, a small number of markers will not be sufficient for phenotype prediction. Many economically important traits, such as fruit quality, are polygenic and therefore controlled by a large number of loci. MAS uses specific molecular markers discovered through linkage mapping or GWAS. In comparison, GS uses all marker data as well as phenotype data from a population to predict a genomic estimated breeding value (GEBV) for an individual. Once a model has been validated, GEBVs can be calculated using only genotype information. However, while particular markers for MAS can be used to track a trait of interest across multiple generations, as breeding populations evolve, GS requires additional rounds of phenotyping in order to maintain an accurate prediction model (Varshney et al., 2014). Additionally, in contrast to MAS, GS requires genotyping a large number of markers, which may still be cost-prohibitive in many breeding programs. A combination of MAS and GS has been proposed in apple, in which monogenic traits are screened using MAS, followed by GS for complex traits. Such a strategy may benefit many perennial crops when introgressing multiple traits from wild relatives, especially to allow for durable disease resistance (Kumar et al., 2012a).

There are many tools and designs for genetic mapping and implementation of genomics-assisted breeding. The decision of which strategy to employ will vary depending on the genetic architecture of the trait as well as the genetic structure of the mapping and breeding populations. Similarly, the specific tool for introgression of markers is a complex decision that will require weighing factors such as the urgency of developing a new cultivar, the extent of linkage drag, and the acceptance of GM technology. While the optimal combination of genomic tools will differ by crop, the adoption of genomics-assisted breeding will ultimately enable breeders to more efficiently and cost effectively incorporate desirable wild traits that would otherwise remain locked away in wild germplasm.

Genomic resources and limitations: sequencing

Despite the immense potential of wild relatives for improving perennial crops, the first step to exploiting this resource through genomics-assisted breeding is discovering markers linked to useful phenotypes. While the genetic divergence between cultivated germplasm and wild relatives is precisely why wild relatives offer such unique and diverse traits, it may also cause difficulties for marker discovery and breeding. For example, when relatives differ in ploidy levels or total chromosome number, it may be difficult to produce fertile interspecific hybrids. The domesticated grape, *V. vinifera,* has 19 chromosomes while its relative, the American wild grape *Muscadinia rotundifolia*, has 20. However, progeny from *V. vinifera* x *M. rotundifolia* have been generated and used for backcrossing to *V. vinifera.* Despite occasional sterility, successful pseudo-backcrossing occurred for 6 subsequent generations, allowing for the introgression of the *M. rotundifolia* gene for powdery mildew resistance, *Run1*, while maintaining a high proportion of *V. vinifera* (Bouquet et al., 2000). In cases of differing ploidy, one solution is the use of protoplast fusion, which has allowed for the creation of somatic hybrids in *Citrus* with ploidy differences as well as pollen/ovule sterility and abnormal chromosome pairing (Guo and Deng, 2001; Rauf et al., 2013). When fertile hybrids are still not possible and a causal locus has been identified, genome editing provides a viable alternative for introgression of valuable traits from wild germplasm.

In addition to the difficulties potentially associated with crossing more distant relatives, wild germplasm may have higher levels of diversity, and as such, DNA sequencing and genotyping tools designed for domesticated species may not function as successfully. For example, SNP arrays are widely used in humans, but do not function as well on organisms with greater genetic diversity because they are designed based on a reference genome. Insertion/deletion polymorphisms (InDels), copy number variants (CNVs) and presence-absence variants (PAVs) all reduce hybridization of a sample's DNA to the probes on an array. Recent work in grape using the Vitis9KSNP array found 33-44% of genotype calls were discarded due to poor quality. In this case, hybridization intensities were more useful than genotype calls for genetic mapping precisely because of the probe-sequence hybridization issues caused by high levels of genetic divergence across grape species (Myles et al., 2015). Thus, when mapping in high diversity perennial crops with SNP arrays such as grape (Myles et al., 2010), peach (Verde et al., 2012) and apple (Bianco et al., 2016), use of hybridization intensities rather than genotype calls is a viable option to overcome the inevitably poor genotype quality.

As an alternative to a genotyping microarray, next-generation DNA sequencing technologies (NGS) such as restriction site associated DNA (RAD) sequencing (Baird et al., 2008) and genotyping-by-sequencing (GBS) (Elshire et al., 2011) do not require markers to be discovered prior to genotyping. The simultaneous discovery and genotyping of markers eliminates the need for DNA to hybridize to previously designed probes and makes NGS well-suited to high diversity species as well as wild relatives. However, in many cases, a reference genome is still used to map DNA sequence reads resulting from NGS and identify SNPs for association mapping or genomic selection. Despite the proliferation of reference genome sequences, there is a lack of reference genomes for wild relatives. More than 100 plant genomes were sequenced between 2000 and 2014, but only 15 were wild relatives and over half of those were soybean (Michael and VanBuren, 2015). Thus, there is a clear need for reference genomes in wild relatives in order to map sequence reads allowing for the detection of SNPs for downstream analyses, ultimately allowing for genomics-assisted breeding.

While more genomic resources are still needed for wild species, the number of reference genomes available has continued to increase. Resequencing of several *Citrus* species including oranges, pummelos and mandarins enabled researchers to determine the contributions of various wild progenitor species to cultivated citrus (Wu et al., 2014). Currently, a dozen wild *Prunus* species useful in hybrid breeding for rootstocks are undergoing genome resequencing by the USDA-ARS (Aradhya et al., 2015). Yet, in many cases, resequencing may not be sufficient for the detection of crucial genomic differences between wild and cultivated crops. Resequencing can detect SNPs as well as InDels when aligned to a reference genome. However, structural differences such as CNVs and PAVs are more difficult to detect. Within species, a large portion of the genome is present in only a subset of individuals. For example, transcriptome sequencing in maize was used to determine that only 16.4% of representative transcript assemblies were expressed in all 503 inbred lines examined (Hirsch et al., 2014). The divergence between wild relatives and cultivated plants is likely much greater. As a result, the genomic region of interest in a wild relative may be a sequence not present in the domesticated crop. DNA sequences present only in wild relatives require *de novo* assembly rather than resequencing to be mapped. The improvement of genomic resources, such as *de novo* assembly of wild relative reference genomes, can enable the discovery of markers for MAS and GS.

Finally, most sequencing results in some degree of missing data in the final table of genotypes. Missing sequence data can be filled in using imputation. However, imputation generally requires that genomic data be aligned to a reference genome. Popular imputation software, including Beagle (Browning and Browning, 2007) and fastPhase (Scheet and Stephens, 2006), rely on the input of SNPs ordered according to a reference genome, which is not possible for many wild relatives with limited genomic resources. Several methods such as Random Forest and *k*-nearest neighbors imputation (kNNI) can be used when a reference genome is not available (Nazzicari et al., 2016). LinkImpute is an imputation software based on kNNI, which updates the method to use linkage between markers rather than distance between samples when calculating neighbors. When compared to existing imputation methods, LinkImpute had a similar run time and accuracy to Beagle, despite not requiring positional information for markers (Money et al., 2015). As the ability to impute missing data without a reference genome improves, reduced representation sequencing techniques with high missing data, such as GBS, will continue to facilitate the discovery of new markers for genomics-assisted breeding in wild relatives.

While there is opportunity for great improvement to elite perennial crops through genomics-assisted introgression of traits from wild relatives, many barriers remain. Genomic tools designed for domesticated species are either not well-suited to more diverse wild relatives, or may be lacking completely. The same genetic divergence that has resulted in wild relatives harboring unique and desirable traits for breeding also results in difficulties in developing markers to introgress these traits into elite germplasm. However, given that DNA sequencing costs are likely to continue decreasing, it is essential that researchers begin planning for a future where the collection and analysis of DNA sequence data will not be the bottleneck to successful genetic mapping. Especially for perennial breeders used to working on timescales of decades, the focus should be on the collection of high-quality phenotype data that can always be paired later with genotype data as it becomes available. Now is the time to establish GWAS and linkage mapping populations that will enable powerful genetic mapping in a future where genotyping costs are negligible and the available genomic analysis tools are far superior to those available today.

Further limitations

Although the primary focus of this review is the use of genomics, it is worth noting that there are several difficulties unrelated to genomics that may limit the use of improvement using wild relatives. First, in order to make use of wild relatives for breeding, new germplasm must be collected. While some wild relative collections are well-characterized and actively in use, such as those described in this review, there are likely many benefits of wild germplasm that remain undiscovered. A focus on the collection and characterization of wild germplasm is the first step towards discovering which relatives and traits will be useful for breeding, and thus be exploitable through genomics.

Among the major barriers to improved characterization of wild germplasm are the locations where such germplasm may be found. Often, wild relatives must be collected from locations that are difficult to access, and thus collecting new wild germplasm can be an expensive and time-consuming process. For example, wild cacao is found in the tropical rainforests of South America (Lachenaud et al., 2007), while fruits and nuts may be expensive and difficult to retrieve from tall trees, and even vegetative samples may be bulky to transport (Aradhya et al., 2015). There are also compulsory quarantine requirements when transferring material between political boundaries. Several decades may pass between the collection of wild germplasm and their use by growers (Lachenaud et al., 2007). Finally, it is important to consider the cultural and financial ramifications of collecting wild relatives. In the past, germplasm has been collected from farmers and communities without compensation or recognition. In such a scenario, seeds may be taken from one country and used to benefit the private sector in another country. While there is ample opportunity for commercial crops to benefit from wild relatives, it is necessary that farmers and communities which have preserved wild relatives receive adequate credit and compensation for use of such resources (Montenegro, 2016).

The introgression of valuable wild traits into domesticated crops can only occur when breeders have access to these relatives through gene banks. The collection of new samples for marker discovery poses a major limitation to establishing such collections. Wild relatives are very under-represented in gene bank collections. A recent overview of over 1,000 taxa in 81 crops found that no CWR germplasm existed in gene banks for 29% of taxa, while 24% had fewer than 10 accessions. Over 95% of taxa had insufficient wild relative representation in gene banks, clearly supporting the need for better collection of wild germplasm, in order to make use of it in breeding (Castañeda-Álvarez et al., 2016). Future collection of germplasm is also threatened due to habitat destruction and climate change (Maxted et al., 2012). As the power of genomic tools increases, genomics will become increasingly effective for introgression of wild traits into perennial crops. However, the ability to exploit wild relatives for breeding requires that this diversity be protected for future use through gene banks and habitat conservation. Preservation of wild relatives will require a complex approach across many environments on a local, national and international scale (Montenegro, 2016). It is crucial to begin exhaustive sampling and extensive evaluation of wild germplasm for all major perennial crops, an enormously expensive and time-consuming undertaking. However, such projects are essential to ensuring a safe and secure future food supply as clonally propagated cultivars continue to be threatened by a constantly-evolving environment.

Future directions

An essential step towards the adoption of genomic markers from wild relatives will be methods that accelerate the juvenile period in order to increase the efficiency of backcrossing progeny to domesticated germplasm. While the use of genomics-assisted breeding can increase the efficiency of selecting for traits of interest and decrease the number of plants that must be propagated, the long juvenile period of many perennials still poses a constraint on the rate of crop improvement.

A solution to the problem of long juvenile periods has been found in grapes. In grapes, 'microvines' possessing a *Vvgai1* mutant allele display dwarfism, a short generation time and continuous flowering. In comparison to the 2-5 years of juvenility generally required for grapes, the *Vvgai1* mutant produces fruit 2 months after germination. In addition to allowing for the rapid cycling of generations, microvines take up less space and could be a valuable tool for genomics studies and MAS (Chaïb et al., 2010). For example, recent work used microvines to aide in QTL identification for traits such as berry acidity (Houel et al., 2015). In apple, an early flowering transgenic line containing the *BpMADS4* gene from silver birch (*Betula pendula*) was combined with MAS to pyramid resistance to apple scab, powdery mildew, and fire blight (Flachowsky et al., 2011). However, while transgenic lines are incredibly helpful for decreasing the generation time while breeding, it is often desirable to have a final cultivar for release that does not contain the transgene and is not considered a genetically modified organism (GMO). This scenario is facilitated by a transgene that is dominant and heterozygous, resulting in only 50% of offspring possessing the gene in each generation. Thus, once the rapid cycling of generations is completed, a non-GMO tree possessing desirable traits from wild relatives—but not the transgene—can easily be selected (Flachowsky et al., 2011). The creation of similar mutants in other species, which reduce the juvenile phase in long-lived perennials, will be essential to the efficient application of MAS.

As an alternative to transgenics, virus-induced gene silencing (VIGS) can also be used to shorten the juvenile phase in perennials. VIGS uses a viral vector to infect a plant with a particular gene, resulting in an RNA-mediated defense which silences expression of the gene within the plant (Lu et al., 2003). The *apple latent spherical virus* (ALSV) does not induce disease symptoms in the infected plant and can be used as a vector for VIGS (Igarashi et al., 2009). When ALSV is used to express *Arabidopsis thaliana* florigen while silencing expression of *MdTFL1-1* in apple or *PcTFL1-1* in pear, flowering time can be reduced to 3 months or less. As genes involved in flowering are identified in other perennials, VIGS could be used to silence these genes and thus shorten the juvenile period (Yamagishi et al., 2011; Yamagishi et al., 2016). ALSV has several other valuable characteristics which make it attractive for use in breeding. The virus was not detected in neighboring trees in an orchard where it had been present since 1984, suggesting there was no vector for transmission present in the sampled orchard and horizontal transmission via pollen did not occur (Nakamura et al., 2011). Additionally, approximately 99% of seedlings from ALSV-infected trees can be considered virus-free (Kishigami et al., 2014). Finally, ALSV can be eliminated from an infected tree using high temperature, allowing for vegetative propagation of that tree and resulting in fruit exempt from restrictions on GMOs (Yamagishi et al., 2016). Therefore, VIGS is a promising method for reducing the juvenile phase in perennials, allowing for a shorter generation time and thus facilitating backcrossing when breeding with wild relatives.

The ability to genotype plants using MAS at the earliest stage of development will allow for the least amount of time and resources to be spent propagating plants which do not carry the marker of interest. While extraction of DNA from seeds is possible for several plants, in perennials it is generally required that plants germinate in order to collect DNA from leaf tissue. Many tree fruits and nuts require a seed dormancy period of up to 12 weeks at low temperatures prior to germination. The development and improvement of methods which overcome seed dormancy could decrease the time prior to genotyping and the generation time between crosses. Several techniques for overcoming seed dormancy include the dissection of embryos and application of bioactive gibberellins or nitric oxide (van Nocker and Gardiner, 2014). Work describing the nondestructive ability to extract DNA from seeds, although recently published in soybean, has been limited so far (Al-Amery et al., 2016). In such a scenario, only the seeds with the desired trait would be germinated, greatly improving the efficiency and decreasing the cost of each breeding cycle.

To facilitate DNA sequence mapping and marker discovery for wild relatives, improvement of genomic resources is needed. As such, there is an urgent need for reference genomes in wild species, or the development of pan-genome sequences that include sequence from both wild and domesticated relatives. To characterize the pan-genome of poplar, recent work performed genome-wide analysis of structural variation in three intercrossable poplar species (Pinosio et al., 2016). Similar efforts are required in most other perennial species. Resequencing of wild germplasm in combination with *de novo* assembly will not only improve our understanding of the domestication history of perennial crops, but also enable the genetic mapping of important traits that can be used for genomics-assisted breeding.

While this review focuses on the potential of genomics-assisted breeding, and in particular MAS, it is worth noting that these tools will always be used in combination with traditional evaluation of cultivars when selecting new varieties. Breeders will always grow and evaluate plants prior to commercial release, but genomics can speed up reaching that final evaluation. Moreover, there are certainly cases where MAS may not even be desirable. For example, when selecting for red fruit flesh in apple, the same anthocyanin-regulating transcription factor often leads to red foliage and therefore trees with this trait can be easily identified before fruit production (Chagne et al., 2007; Espley et al., 2009). However, there is also a paralogous gene for red fruit flesh color where red foliage does not occur and MAS could be valuable in those instances (Chagne et al., 2013). Due to the cost and labour expense of MAS, previous work selecting for downy and powdery mildew resistance in grape included both phenotypic and marker-assisted selection. The initial population of interest consisted of 119 plants inoculated with downy mildew. Seedlings resistant to downy mildew were then screened for powdery mildew resistance. Finally, the 20 seedlings resistant to both diseases were tested using MAS, resulting in a final reduction to only 4 seedlings (Eibach et al., 2007). In this case, while phenotype selection was effective, MAS allowed for an improved reduction in the number of seedlings. When applying MAS to perennial crops, the greatest cost-savings will occur if testing occurs at the seed or seedling stage. MAS is particularly useful for traits that are difficult, expensive, or time-consuming to phenotype, such as fruit traits and disease resistance. To be of use, the markers must be economical to discover as well as test. Lastly, MAS requires a robust marker-trait association which improves the breeder's ability to select for individuals possessing a particular trait. Thus, while low cost MAS can facilitate the introgression of specific traits of interest from wild relatives, it will ultimately only be useful when the cost of phenotyping is higher than the cost of discovering markers and genotyping (Luby and Shaw, 2001).

Lastly, a major barrier to more widespread adoption of MAS is often not the lack of genomic resources for wild relatives or the cost of genotyping, but the 'phenotyping bottleneck' present when characterizing germplasm. While the cost and speed of collecting genomic data has continued to decrease, phenotyping remains slow and expensive (Burleigh et al., 2013). Given that high-quality phenotype data is required for well-powered QTL analyses, the improvement of phenotyping technologies is a major area of current research interest. The development of new, high-throughput (HT) phenotyping technologies has begun, including advances in image analysis and robotics (Furbank and Tester, 2011). Improvement to phenotyping technologies will aid in the characterization of wild germplasm, a task which is particularly challenging due to the high level of diversity present. Thus far, HT phenotyping technologies have focused on annual crops such as rice (Tanger et al., 2017) and cotton (*Gossypium barbadense* L.) (Andrade-Sanchez et al., 2014), neglecting diverse perennial crops. One example of a technology useful for wild relatives is Field Book, an open-source application for collecting field data that eliminates the need to transcribe handwritten notes (Rife and Poland, 2014). As phenotyping technology for perennial crops and wild relatives improves, so will the ability to detect markers which can be exploited for genomics-assisted breeding. Thus, phenotyping of wild relatives, while expensive, is a necessary task. Additionally, good quality phenotype data will continue to have value in the future. Phenotype data can be collected now but analyzed in the future when, for example, the cost of whole genome sequencing is no longer prohibitive.

Ultimately, although genomics-assisted breeding has been used to introgress traits from wild relatives into perennial crops in the past, there are still many areas in which future work is required to improve this process. The use of genomic tools such as those which reduce the generation time for long-lived perennial crops and allow for DNA extraction from seeds—and the continued development of such tools—are two crucial steps in facilitating the use of MAS in perennials. To make use of markers in breeding, they must first be discovered, and as such improvement to genetic mapping techniques and resources will be necessary. Finally, MAS is especially valuable for the introgression of multiple traits as well as those that are difficult or expensive to phenotype. However, the usefulness of MAS relies on the ability to discover and genotype markers for less than the cost of phenotyping all progeny. As technology improves and the cost of marker discovery decreases, it will become increasingly feasible to introgress useful traits from wild relatives into elite perennial cultivars, resulting in the much-needed improvement of crops that may have been clonally propagated for centuries.

Conclusions

There are clearly many traits such as disease resistance, fruit quality and rootstock characteristics which would benefit domesticated perennials but are locked in undesirable, wild germplasm. Use of MAS can enable breeders to unlock the potential of wild germplasm by facilitating selection at an early stage of development—or even as a seed—allowing for less time and money to be spent growing plants which will inevitably be discarded. However, when crossing wild relatives and elite cultivars there are certain limitations and difficulties. Often many generations of backcrossing are required to decrease linkage drag and other wild characteristics. Use of GM technology can help reduce the amount of time required for breeding, but decades may still be required for consumer and regulatory acceptance. The development of CRISPR/Cas9 is a promising alternative to traditional methods. Both MAS and genome editing require the initial discovery of markers, which is complicated by the fact that alleles for traits of interest often co-segregate with millions of other alleles in wild germplasm. Yet, the potential benefit of accessing unique and desirable traits in wild germplasm could revolutionize perennial crop improvement. Unfortunately, the discovery of useful markers using GWAS and linkage mapping may still require decades to yield results. Thus, it is essential the collection and characterization of wild relatives begin immediately, while genomic and phenomic tools suited to diverse germplasm continue to improve. The continued vegetative propagation of domesticated perennial cultivars affords pathogens the opportunity to become increasingly effective while robbing both growers and consumers of the unique and desirable traits present in wild germplasm. After decades, or even millennia, of growing the same perennial cultivars frozen in genetic time, the decreasing costs of sequencing can finally allow us to harvest the potential of wild relatives through genomics-assisted breeding. We have only begun to enjoy the benefits of wild relatives in perennial crop improvement, and continued technological advances will surely result in the more efficient development of tastier food that requires less chemical input to grow.

Acknowledgements

We would like to acknowledge Mark O. Johnston, Christophe M. Herbinger, Robert G. Beiko, Robert G. Latta and Michel S. McElroy for helpful discussion. This article was written, in part, thanks to funding from the Canada Research Chairs program, the National Sciences and Engineering Research Council of Canada and Genome Canada. Z.M. was supported in part by a Killam Predoctoral Scholarship from Dalhousie University.

Chapter 6: : Conclusion

Summary of findings

There were two main objectives to this thesis: to characterize the genetic basis of several simple and complex traits in apple and to estimate the current use of wild relatives in grape breeding while describing the potential of genomics and wild relatives for further improvement of perennial crops.

In Chapter 2, we examined 9,000 leaves from 869 unique apple accessions using linear measurements and comprehensive morphometric techniques. We identified allometric variation in the length-to-width aspect ratio between accessions and species of apple. The allometric variation was due to variation in the width of the leaf blade, not length. Aspect ratio was highly correlated with the primary axis of morphometric variation (PC1) quantified using elliptical Fourier descriptors (EFDs) and persistent homology (PH). While the primary source of variation was aspect ratio, subsequent PCs corresponded to complex shape variation not captured by linear measurements. After linking the morphometric information with over 122,000 genome-wide SNPs, we found high narrow-sense heritability values even at later PCs, indicating that comprehensive morphometrics can capture hidden, heritable phenotypes. Thus, techniques such as EFDs and PH are capturing heritable biological variation that would be missed using linear measurements alone, and which could potentially be used to select for a hidden phenotype only detectable using comprehensive morphometrics. Ultimately, a better understanding of the genetic basis of quantitative traits is important for genomics-assisted breeding.

In Chapter 3, we mined historical phenotype data from the USDA GRIN database and linked this information with genome-wide SNP data for 689 apple accessions. We identified genetic structure based on the geographic origin of the accession as well as the time required for ripening. Performing GWAS, we confirmed a known association between fruit color and MYB1. We also identified an amino acid substitution in the transcription factor NAC18.1 that is a strong functional candidate for fruit firmness and harvest date. Finally, we demonstrated that traits such as harvest time and fruit size can be predicted with relatively high accuracy using genomic prediction. This work indicates the potential of using gene banks for genetic mapping, which holds great potential for continued improvement of apples through MAS.

In Chapter 4, we used genome-wide SNP data and a PCA-based ancestry estimation procedure to assess ancestry in some of the most widely grown commercial hybrid grape cultivars. We verified the method with both empirical and simulated data and found the ancestry of commercial hybrid grapes ranged from 11% to 76% *V. vinifera*. Approximately one third of hybrids had ancestry consistent with F1 hybridization: they derived half of their ancestry from wild *Vitis* and half from *V. vinifera*. Our results suggest that hybrid grape breeding is in its infancy. If backcrossing to *V. vinifera* were more widely adopted in grape, acceptance of hybrid grape cultivars could improve. The method described in this chapter could be combined with MAS to facilitate the breeding of hybrid grapes which require less chemical input to grow and produce high quality wine.

In Chapter 5, we described the potential of using genomics to improve perennial crops through introgression of valuable traits from wild relatives. Given that perennial crops are expensive and time-consuming to breed, genomics provides a valuable tool for improving the efficiency of breeding by allowing for selection of progeny possessing a trait of interest at an early stage. Genomics will always be used in combination with traditional breeding techniques, but it is a valuable tool for accelerating the speed and decreasing the cost of breeding. Wild relatives are a largely untapped source of desirable traits such as disease resistance, fruit quality, and rootstock characteristics. We described examples from wild relatives of perennial crops possessing these traits, as well as current efforts to incorporate traits from wild relatives using genomics-assisted breeding. Genetic mapping in wild relatives is made difficult by genomic tools that are often ill-suited to the diversity of wild-relatives. Additionally, phenotyping wild relatives is an expensive and difficult process. However, there is an urgent need to immediately begin the collection and characterization of wild relatives of perennial species in order to discover which relatives and traits can be used for genomics-assisted breeding and improvement of perennial crops.

Overall, the results of this thesis lead to the following conclusions: comprehensive morphometric techniques capture heritable variation, novel genomic insights can be generated using historical phenotype data from gene banks, hybrid grape breeding is still in its infancy, and wild relatives should be exploited for genomics-assisted breeding of perennial crops.

Future directions

While the cost of DNA sequencing continues to decrease, the ability to phenotype has instead become the limitation—or "phenotyping bottleneck"—to genetic mapping (Furbank and Tester, 2011). High-throughput (HT) phenotyping using comprehensive morphometrics enabled us to capture heritable variation in apple leaf shape that traditional phenotyping such as linear measurements would have missed. Further work on the apple fruit shape would complement this analysis and potentially provide evidence of whether leaf shape can be used as an early indicator of fruit shape or other agricultural important characteristics. We also exploited historical phenotype data from the USDA and demonstrated that this information can be linked with new genetic data for genetic mapping. Indeed, we presented a novel functional candidate for variation in fruit firmness and harvest date in apple that should be explored further. However, image-based analysis and historical phenotype data, though valuable, are not sufficient. HT phenotyping technology better-suited to diverse perennial species and wild relatives is critical to the success of genetic mapping and thus genomics-assisted breeding. High quality phenotype data collected now will continue to have value in the future, as the cost of genotyping becomes negligible to the cost of phenotyping.

Once genetic mapping has been performed for important traits, this information can be exploited in combination with ancestry estimates—such as those we describe in hybrid grapes—to select for both desirable traits and domesticated ancestry when breeding. Our estimates of domesticated ancestry in commercial hybrid grapes indicate that backcrossing to wild *Vitis* has been more frequent than backcrossing to *V. vinifera*, and as such, ancestry estimation could be a valuable tool for increasing *V. vinifera* content. The same technique could be applied to other hybrid crops. Wild relatives harbour many valuable traits that can help improve perennial crops, but remain locked inside germplasm with many other undesirable characteristics. Thus, ancestry estimation is a valuable tool for allowing the breeder to limit hybrid offspring to those with the highest level of domesticated ancestry and therefore the highest proportion of desirable characteristics. In combination with MAS or genomic selection, these ancestry estimates can be used to ensure both a high level of domesticated ancestry as well as the presence of desirable traits—both those that originate from wild relatives and domesticated parents—of interest in the progeny.

Finally, the results of this thesis focus on characterizing phenomic diversity in apple and grape through genetic mapping, but future work is required to implement the methods and markers described. It is essential to focus not simply on the discovery of markers, but also ensuring the information is accessible and genetic tests are easily available so that breeders can make use of these markers. Publications describing genetic markers are not sufficient (Peace, 2017). If appropriate genetic tests are designed for breeders, it will improve efficiency of breeding and allow breeders to make use of more diverse germplasm. While genetic mapping and ancestry estimation, such as that performed in this thesis, are the first steps, it is ultimately only through applied genetic testing that breeders will be able to save time and money, while selecting for desirable traits and ancestry in perennial cultivars, ultimately improving crops for a safe and secure food supply.

References

Abràmoff, M.D., Magalhães, P.J., and Ram, S.J. (2004). Image processing with ImageJ. *Biophotonics international* 11**,** 36-42.

Al-Amery, M., Fukushige, H., Serson, W., and Hildebrand, D. (2016). Nondestructive DNA Extraction Techniques for Soybean (Glycine Max) Seeds. *Journal of Crop Improvement* 30**,** 165-175.

Alleweldt, G. (1997). "Genetics of grapevine breeding," in *Progress in Botany*. Springer), 441-454.

Alleweldt, G., and Possingham, J.V. (1988). Progress in grapevine breeding. *Theoretical and Applied Genetics* 75**,** 669-673.

Alston, J.M., Fuller, K.B., Kaplan, J.D., and Tumber, K.P. (2013). The economic consequences of Pierce's disease and related policy in the California winegrape industry. *Journal of Agricultural and Resource Economics* 38**,** 269-297.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology* 215**,** 403-410.

Anderson, K., and Aryal, N.R. (2013). "A guide to where in the world various winegrape varieties are grown," in *Which Winegrape Varieties are Grown Where?: A Global Empirical Picture,* eds. K. Anderson & N.R. Aryal. University of Adelaide Press), 1-11.

Andrade-Sanchez, P., Gore, M.A., Heun, J.T., Thorp, K.R., Carmo-Silva, A.E., French, A.N., Salvucci, M.E., and White, J.W. (2014). Development and evaluation of a field-based high-throughput phenotyping platform. *Functional Plant Biology* 41**,** 68-79.

Andres, R.J., Coneva, V., Frank, M.H., Tuttle, J.R., Samayoa, L.F., Han, S.W., Kaur, B., Zhu, L., Fang, H., Bowman, D.T., Rojas-Pierce, M., Haigler, C.H., Jones, D.C., Holland, J.B., Chitwood, D.H., and Kuraparthy, V. (2017). Modifications to a LATE MERISTEM IDENTITY1 gene are responsible for the major leaf shapes of Upland cotton (Gossypium hirsutum L.). *Proc Natl Acad Sci U S A* 114**,** E57-E66.

Antanaviciute, L., Fernández-Fernández, F., Jansen, J., Banchi, E., Evans, K.M., Viola, R., Velasco, R., Dunwell, J.M., Troggio, M., and Sargent, D.J. (2012). Development of a dense SNP-based linkage map of an apple rootstock progeny using the Malus Infinium whole genome genotyping array. *BMC Genomics* 13**,** 203-203.

Aradhya, M.K., Preece, J., and Kluepfel, D.A. (Year). "Genetic Conservation, Characterization and Utilization of Wild Relatives of Fruit and Nut Crop at the USDA Germplasm Repository in Davis, California", in: *II International Symposium on Wild Relatives of Subtropical and Temperate Fruit and Nut Crops 1074*: International Society for Horticultural Science (ISHS), Leuven, Belgium), 95-104.

Atkinson, R.G., Sutherland, P.W., Johnston, S.L., Gunaseelan, K., Hallett, I.C., Mitra, D., Brummell, D.A., Schröder, R., Johnston, J.W., and Schaffer, R.J. (2012). Down-regulation of POLYGALACTURONASE1 alters firmness, tensile strength and water loss in apple (Malus x domestica) fruit. *BMC Plant Biology* 12.

Bai, Y., and Lindhout, P. (2007). Domestication and breeding of tomatoes: what have we gained and what can we gain in the future? *Ann Bot* 100**,** 1085-1094.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one* 3**,** e3376.

Baldwin, S.J., Dodds, K.G., Auvray, B., Genet, R.A., Macknight, R.C., and Jacobs, J.M.E. (2011). Association mapping of cold-induced sweetening in potato using historical phenotypic data. *Annals of Applied Biology* 158**,** 248-256.

Ban, Y., Honda, C., Hatsuyama, Y., Igarashi, M., Bessho, H., and 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 Physiol* 48**,** 958-970.

Bassett, C.L., Glenn, D.M., Forsline, P.L., Wisniewski, M.E., and Farrell, R.E. (2011). Characterizing Water Use Efficiency and Water Deficit Responses in Apple (Malus× domestica Borkh. and Malus sieversii Ledeb.) M. Roem. *HortScience* 46**,** 1079-1084.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *2015* 67**,** 48.

Baumgartner, I.O., Patocchi, A., Frey, J.E., Peil, A., and Kellerhals, M. (2015). Breeding Elite Lines of Apple Carrying Pyramided Homozygous Resistance Genes Against Apple Scab and Resistance Against Powdery Mildew and Fire Blight. *Plant Molecular Biology Reporter* 33**,** 1573-1583.

Beckman, T., and Lang, G. (Year). "Rootstock breeding for stone fruits", in: *XXVI International Horticultural Congress: Genetics and Breeding of Tree Fruits and Nuts 622*), 531-551.

Benfey, P.N., and Mitchell-Olds, T. (2008). From Genotype to Phenotype: Systems Biology Meets Natural Variation. *Science* 320**,** 495-497.

Berg, B.O., and Lahav, E. (1996). "Avocados," in *Fruit breeding, tree and tropical fruits,* eds. J. Janick & J.N. Moore. John Wiley & Sons), 113-166.

Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denance, C., Theron, 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., and Troggio, M. (2016). Development and validation of the Axiom Apple480K SNP genotyping array. *Plant J.*

Biscarini, F., Bovenhuis, H., Van Arendonk, J., Parmentier, H., Jungerius, A., and Van Der Poel, J. (2010). Across-line SNP association study of innate and adaptive immune response in laying hens. *Animal genetics* 41**,** 26-38.

Bisson, L.F., Waterhouse, A.L., Ebeler, S.E., Walker, M.A., and Lapsley, J.T. (2002). The present and future of the international wine industry. *Nature* 418**,** 696-699.

Bonhomme, V., Picq, S., Gaucherel, C., and Claude, J. (2014). Momocs: outline analysis using R. *Journal of Statistical Software* 56**,** 1-24.

Botton, A., Eccher, G., Forcato, C., Ferrarini, A., Begheldo, M., Zermiani, M., Moscatello, S., Battistelli, A., Velasco, R., Ruperti, B., and Ramina, A. (2011). Signaling pathways mediating the induction of apple fruitlet abscission. *Plant Physiol* 155**,** 185-208.

Bouquet, A., Pauquet, J., Adam-Blondon, A., Torregrosa, L., Merdinoglu, D., and Wiedemann-Merdinoglu, S. (2000). Towards obtaining grapevine varieties resistant to powdery and downy mildews by conventional breeding and biotechnology. *Bulletin de l'OIV* 73**,** 445-452.

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

Brown, J.S., Schnell, R., Motamayor, J., Lopes, U., Kuhn, D.N., and Borrone, J.W. (2005). Resistance gene mapping for witches' broom disease in Theobroma cacao L. in an F2 population using SSR markers and candidate genes. *Journal of the American Society for Horticultural Science* 130**,** 366-373.

Brown, S.K., and Maloney, K.E. (2003). "Genetic Improvement of Apple: Breeding, Markers, Mapping and Biotechnology," in *Apples: Botany, Production and Uses,* eds. D.C. Ferree & I.J. Warrington. CABI Publishing), 31-60.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81**,** 1084-1097.

Brumlop, S., Reichenbecher, W., Tappeser, B., and Finckh, M.R. (2013). What is the SMARTest way to breed plants and increase agrobiodiversity? *Euphytica* 194**,** 53-66.

Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.M., Wambebe, C., Tishkoff, S.A., and Bustamante, C.D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci USA* 107**,** 786-791.

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., Simmons, N.B., Stevenson, D.W., Thacker, R.W., Theriot, E.C., Todorovic, S., Velazco, P.M., Walls, R.L., Wolfe, J.M., and Yu, M. (2013). Next-generation phenomics for the Tree of Life. *PLoS Curr* 5.

Cahoon, C. (1986). The Concord Grapes. *Fruit Varieties Journal* 40**,** 106-107.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Muller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K.J., and Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet* 43**,** 956-963.

Castañeda-Álvarez, N.P., Khoury, C.K., Achicanoy, H.A., Bernau, V., Dempewolf, H., Eastwood, R.J., Guarino, L., Harker, R.H., Jarvis, A., Maxted, N., Müller, J.V., Ramirez-Villegas, J., Sosa, C.C., Struik, P.C., Vincent, H., and Toll, J. (2016). Global conservation priorities for crop wild relatives. *Nature Plants***,** 16022.

Cavanagh, C., Morell, M., Mackay, I., and Powell, W. (2008). From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Curr Opin Plant Biol* 11**,** 215-221.

Chagne, 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., and Gardiner, S.E. (2007). Mapping a candidate gene (MdMYB10) for red flesh and foliage colour in apple. *BMC Genomics* 8**,** 212.

Chagne, D., Lin-Wang, K., Espley, R.V., Volz, R.K., How, N.M., Rouse, S., Brendolise, C., Carlisle, C.M., Kumar, S., De Silva, N., Micheletti, D., Mcghie, T., Crowhurst, R.N., Storey, R.D., Velasco, R., Hellens, R.P., Gardiner, S.E., and Allan, A.C. (2013). An ancient duplication of apple MYB transcription factors is responsible for novel red fruit-flesh phenotypes. *Plant Physiol* 161**,** 225-239.

Chaïb, J., Torregrosa, L., Mackenzie, D., Corena, P., Bouquet, A., and Thomas, M.R. (2010). The grape microvine–a model system for rapid forward and reverse genetics of grapevines. *The Plant Journal* 62**,** 1083-1092.

Chitwood, D.H., Headland, L.R., Kumar, R., Peng, J., Maloof, J.N., and Sinha, N.R. (2012). The developmental trajectory of leaflet morphology in wild tomato species. *Plant Physiol* 158**,** 1230-1240.

Chitwood, D.H., Klein, L.L., O'hanlon, R., Chacko, S., Greg, M., Kitchen, C., Miller, A.J., and Londo, J.P. (2016). Latent developmental and evolutionary shapes embedded within the grapevine leaf. *New Phytol* 210**,** 343-355.

Chitwood, D.H., Kumar, R., Headland, L.R., Ranjan, A., Covington, M.F., Ichihashi, Y., Fulop, D., Jimenez-Gomez, J.M., Peng, J., Maloof, J.N., and Sinha, N.R. (2013). A quantitative genetic basis for leaf morphology in a set of precisely defined tomato introgression lines. *Plant Cell* 25**,** 2465-2481.

Chitwood, D.H., and Otoni, W.C. (2017). Morphometric analysis of Passiflora leaves: the relationship between landmarks of the vasculature and elliptical Fourier descriptors of the blade. *GigaScience* 6**,** 1-13.

Chitwood, D.H., Ranjan, A., Martinez, C.C., Headland, L.R., Thiem, T., Kumar, R., Covington, M.F., Hatcher, T., Naylor, D.T., Zimmerman, S., Downs, N., Raymundo, N., Buckler, E.S., Maloof, J.N., Aradhya, M., Prins, B., Li, L., Myles, S., and Sinha, N.R. (2014). A modern ampelography: a genetic basis for leaf shape and venation patterning in grape. *Plant Physiol* 164**,** 259-272.

Cobb, J.N., Declerck, G., Greenberg, A., Clark, R., and Mccouch, S. (2013). Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. *Theor Appl Genet* 126**,** 867-887.

Cooney, C.R., Bright, J.A., Capp, E.J., Chira, A.M., Hughes, E.C., Moody, C.J., Nouri, L.O., Varley, Z.K., and Thomas, G.H. (2017). Mega-evolutionary dynamics of the adaptive radiation of birds. *Nature* 542**,** 344-347.

Cornille, A., Gladieux, P., Smulders, M.J., Roldan-Ruiz, I., Laurens, F., Le Cam, B., Nersesyan, A., Clavel, J., Olonova, M., Feugey, L., Gabrielyan, I., Zhang, X.G., Tenaillon, M.I., and 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 Genet* 8**,** e1002703.

Costa, F., Cappellin, L., Farneti, B., Tadiello, A., Romano, A., Soukoulis, C., Sansavini, S., Velasco, R., and Biasioli, F. (2014). Advances in QTL mapping for ethylene production in apple (Malus×domestica Borkh.). *Postharvest Biology and Technology* 87**,** 126-132.

Costa, F., Peace, C.P., Stella, S., Serra, S., Musacchi, S., Bazzani, M., Sansavini, S., and Van De Weg, W.E. (2010). QTL dynamics for fruit firmness and softening around an ethylene-dependent polygalacturonase gene in apple (Malus x domestica Borkh.). *J Exp Bot* 61**,** 3029-3039.

Costa, F., Stella, S., Van De Weg, W.E., Guerra, W., Cecchinel, M., Dallavia, J., Koller, B., and Sansavini, S. (2005). Role of the genes Md-ACO1 and Md-ACS1 in ethylene production and shelf life of apple (Malus domestica Borkh). *Euphytica* 141**,** 181-190.

Cuevas-Rodriguez, E.O., Yousef, G.G., Garcia-Saucedo, P.A., Lopez-Medina, J., Paredes-Lopez, O., and Lila, M.A. (2010). Characterization of anthocyanins and proanthocyanidins in wild and domesticated Mexican blackberries (Rubus spp.). *J Agric Food Chem* 58**,** 7458-7464.

Cunha, C.M.S., Hinz, R.H., Pereira, A., Tcacenco, F.A., Paulino, E.C., and Stadnik, M.J. (2015). A SCAR marker for identifying susceptibility to Fusarium oxysporum f. sp. cubense in banana. *Scientia Horticulturae* 191**,** 108-112.

Currie, A.J., Ganeshanandam, S., Noiton, D.A., Garrick, D., Shelbourne, C.J.A., and Oraguzie, N. (2000). Quantitative evaluation of apple (Malus × domestica Borkh.) fruit shape by principal component analysis of Fourier descriptors. *Euphytica* 111**,** 219–227.

Daly, A., and Walduck, G. (2006). Fusarium wilt of bananas (Panama disease). *Northern Territory Government. USA*.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., Depristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., Mcvean, G., Durbin, R., and Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics* 27**,** 2156-2158.

Davidson, S.N. (2008). Forbidden fruit: transgenic papaya in Thailand. *Plant Physiol* 147**,** 487-493.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978). A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure***,** 345–352.

De Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., and Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* 34**,** W362-365.

Deans, A.R., Lewis, S.E., Huala, E., Anzaldo, S.S., Ashburner, M., Balhoff, J.P., Blackburn, D.C., Blake, J.A., Burleigh, J.G., Chanet, B., Cooper, L.D., Courtot, M., Csösz, S., Cui, H., Dahdul, W., Das, S., Dececchi, T.A., Dettai, A., Diogo, R., Druzinsky, R.E., Dumontier, M., Franz, N.M., Friedrich, F., Gkoutos, G.V., Haendel, M., Harmon, L.J., Hayamizu, T.F., He, Y., Hines, H.M., Ibrahim, N., Jackson, L.M., Jaiswal, P., James-Zorn, C., Köhler, S., Lecointre, G., Lapp, H., Lawrence, C.J., Le Novère, N., Lundberg, J.G., Macklin, J., Mast, A.R., Midford, P.E., Mikó, I., Mungall, C.J., Oellrich, A., Osumi-Sutherland, D., Parkinson, H., Ramírez, M.J., Richter, S., Robinson, P.N., Ruttenberg, A., Schulz, K.S., Segerdell, E., Seltmann, K.C., Sharkey, M.J., Smith, A.D., Smith, B., Specht, C.D., Squires, R.B., Thacker, R.W., Thessen, A., Fernandez-Triana, J., Vihinen, M., Vize, P.D., Vogt, L., Wall, C.E., Walls, R.L., Westerfeld, M., Wharton, R.A., Wirkner, C.S., Woolley, J.B., Yoder, M.J., Zorn, A.M., and Mabee, P. (2015). Finding Our Way through Phenotypes. *PLoS Biol* 13**,** e1002033.

Deighton, N., Brennan, R., Finn, C., and Davies, H.V. (2000). Antioxidant properties of domesticated and wild Rubus species. *Journal of the Science of Food and Agriculture* 80**,** 1307-1313.

Dennis, F., Jr. (2003). "Flowering, Pollination and Fruit Set and Development," in *Apples: Botany, Production and Uses,* eds. D.C. Ferree & I.J. Warrington. CABI Publishing), 153-166.

Dohleman, F.G., and Long, S.P. (2009). More productive than maize in the Midwest: How does Miscanthus do it? *Plant Physiol* 150**,** 2104-2115.

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

Edge-Garza, D.A., Peace, C.P., and Zhu, Y. (2010). Enabling marker-assisted seedling selection in the Washington apple breeding program. *Acta Horticult***,** 369–373.

Eibach, R., and Töpher, R. (Year). "Success in resistance breeding: "Regent" and its steps into the market", in: *VIII International Conference on Grape Genetics and Breeding 603*), 687-691.

Eibach, R., Zyprian, E., Welter, L., and Töpfer, R. (2007). The use of molecular markers for pyramiding resistance genes in grapevine breeding. *VITIS-Journal of Grapevine Research* 46**,** 120-124.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6**,** e19379.

Endelman, J.B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal* 4**,** 250.

Espley, R.V., Brendolise, C., Chagne, D., Kutty-Amma, S., Green, S., Volz, R., Putterill, J., Schouten, H.J., Gardiner, S.E., Hellens, R.P., and Allan, A.C. (2009). Multiple repeats of a promoter segment causes transcription factor autoregulation in red apples. *Plant Cell* 21**,** 168-183.

Evans, K.M., Barritt, B.H., Konishi, B.S., Dilley, M.A., Brutcher, L.J., and Peace, C.P. (2011). 'WA 5' Apple. *HortScience* 46**,** 958-960.

Fajardo, D., Morales, J., Zhu, H., Steffan, S., Harbut, R., Bassil, N., Hummer, K., Polashock, J., Vorsa, N., and Zalapa, J. (2012). Discrimination of American Cranberry Cultivars and Assessment of Clonal Heterogeneity Using Microsatellite Markers. *Plant Molecular Biology Reporter* 31**,** 264-271.

Fazio, G., Wan, Y., Kviklys, D., Romero, L., Adams, R., Strickland, D., and Robinson, T. (2014). Dw2, a new dwarfing locus in apple rootstocks and its relationship to induction of early bearing in apple scions. *Journal of the American Society for Horticultural Science* 139**,** 87-98.

Felipe, A.J. (2009). 'Felinem','Garnem', and 'Monegro'almond× peach hybrid rootstocks. *HortScience* 44**,** 196-197.

Ferguson, A.R. (2007). The need for characterisation and evaluation of germplasm: kiwifruit as an example. *Euphytica* 154**,** 371-382.

Ferguson, A.R., and Huang, H. (2007). Genetic resources of kiwifruit: domestication and breeding. *Horticultural reviews* 33**,** 1-121.

Flachowsky, H., Le Roux, P.M., Peil, A., Patocchi, A., Richter, K., and Hanke, M.V. (2011). Application of a high-speed breeding technology to apple (Malus x domestica) based on transgenic early flowering plants and marker-assisted selection. *New Phytol* 192**,** 364-377.

Fletcher, L. (1932). Effect of thinning on size and color of apples. *Proc. Amor. Soc. Hort. Sci.(29)***,** 51-56.

Food and Agriculture Organization of the United Nations (2015). *FAOSTAT* [Online]. Available: http://faostat3.fao.org/browse/rankings/commodities_by_regions/E [Accessed October 19 2015].

Food and Agriculture Organization of the United Nations (2017). *FAOSTAT* [Online]. Available: http://www.fao.org/faostat/en/ - data/QC [Accessed March 9 2017].

Foolad, M.R. (2007). Genome mapping and molecular breeding of tomato. *Int J Plant Genomics* 2007**,** 64358.

Furbank, R.T., and Tester, M. (2011). Phenomics--technologies to relieve the phenotyping bottleneck. *Trends Plant Sci* 16**,** 635-644.

Gao, X., Becker, L.C., Becker, D.M., Starmer, J.D., and Province, M.A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic epidemiology* 34**,** 100-105.

Gao, X., Starmer, J., and Martin, E.R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic epidemiology* 32**,** 361-369.

Gardner, K.M., Brown, P., Cooke, T.F., Cann, S., Costa, F., Bustamante, C., Velasco, R., Troggio, M., and Myles, S. (2014). Fast and cost-effective genetic mapping in apple using next-generation sequencing. *G3 (Bethesda)* 4**,** 1681-1687.

Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nat Genet* 42**,** 558-560.

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., and Buckler, E.S. (2014). TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLoS ONE* 9**,** e90346.

Glover, J.D., Culman, S.W., Dupont, S.T., Broussard, W., Young, L., Mangan, M.E., Mai, J.G., Crews, T.E., Dehaan, L.R., and Buckley, D.H. (2010a). Harvested perennial grasslands provide ecological benchmarks for agricultural sustainability. *Agriculture, Ecosystems & Environment* 137**,** 3-12.

Glover, J.D., Reganold, J., Bell, L., Borevitz, J., Brummer, E., Buckler, E., Cox, C., Cox, T.S., Crews, T., and Culman, S. (2010b). Increased food and ecosystem security via perennial grains. *Science* 328**,** 1638-1639.

Gonsalves, C., Cai, W., Tennant, P., and Gonsalves, D. (1998). Effective development of Papaya ringspot virus resistant papaya with untranslatable coat protein gene using a modified microprojective transformation method. *Acta Horticult***,** 311-314.

Gonsalves, D., Vegas, A., Prasartsee, V., Drew, R., Suzuki, J., and Tripathi, S. (2006). Developing papaya to control papaya ringspot virus by transgenic resistance, intergeneric hybridization, and tolerance breeding. *Plant Breeding Reviews* 26**,** 35-73.

Goto-Yamamoto, N., Sawler, J., and Myles, S. (2015). Genetic Analysis of East Asian Grape Cultivars Suggests Hybridization with Wild *Vitis*. *Plos One* 10**,** e0140841.

Grattapaglia, D., and Sederoff, R. (1994). Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* 137**,** 1121-1137.

Grillo, M.A., Li, C., Hammond, M., Wang, L., and Schemske, D.W. (2013). Genetic architecture of flowering time differentiation between locally adapted populations of Arabidopsis thaliana. *New Phytol* 197**,** 1321-1331.

Guo, W., and Deng, X. (2001). Wide somatic hybrids of Citrus with its related genera and their potential in genetic improvement. *Euphytica* 118**,** 175-183.

Gurevitch, J. (1992). Sources of Variation in Leaf Shape among Two Populations of Achillea Lanulosa. *Genetics* 130**,** 385-394.

Gygax, M., Gianfranceschi, L., Liebhard, R., Kellerhals, M., Gessler, C., and Patocchi, A. (2004). Molecular markers linked to the apple scab resistance gene Vbj derived from Malus baccata jackii. *Theor Appl Genet* 109**,** 1702-1709.

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

Hall, M.C., and Willis, J.H. (2006). Divergent Selection on Flowering Time Contributes to Local Adaptation in Mimulus guttatus Populations. *Evolution* 60**,** 2466-2477.

Harker, F., Jaeger, S., Lau, K., and Rossiter, K. (Year). "Consumer perceptions and preferences for kiwifruit: a review", in: *VI International Symposium on Kiwifruit 753*), 81-88.

Hasey, J., Kluepfel, D., and Anderson, K. (Year). "Crown gall incidence and severity: seedling walnut rootstock versus clonally propagated rootstock", in: *VII International Walnut Symposium 1050*), 305-308.

Heffner, E.L., Jannink, J.-L., and Sorrells, M.E. (2011). Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. *The Plant Genome* 4**,** 65-75.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* 89**,** 10915-10919.

Heslop-Harrison, J.S., and Schwarzacher, T. (2007). Domestication, genomics and the future for banana. *Ann Bot* 100**,** 1073-1084.

Heuertz, M., De Paoli, E., Kallman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M., and Gyllenstrand, N. (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [Picea abies (L.) Karst]. *Genetics* 174**,** 2095-2105.

Hijmans, R.J. (2016). "raster: Geographic Data Analysis and Modeling", in: *R package version 2.5-8.*).

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

Houel, C., Chatbanyong, R., Doligez, A., Rienth, M., Foria, S., Luchaire, N., Roux, C., Adivèze, A., Lopez, G., Farnos, M., Pellegrino, A., This, P., Romieu, C., and Torregrosa, L. (2015). Identification of stable QTLs for vegetative and reproductive traits in the microvine (Vitis vinifera L.) using the 18 K Infinium chip. *BMC Plant Biology* 15.

Houle, D., Govindaraju, D.R., and Omholt, S. (2010). Phenomics: the next challenge. *Nat Rev Genet* 11**,** 855-866.

Hwang, S.-C., and Ko, W.-H. (2004). Cavendish banana cultivars resistant to Fusarium wilt acquired through somaclonal variation in Taiwan. *Plant disease* 88**,** 580-588.

Hyten, D.L., Choi, I.-Y., Song, Q., Shoemaker, R.C., Nelson, R.L., Costa, J.M., Specht, J.E., and Cregan, P.B. (2007). Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* 175**,** 1937-1944.

Igarashi, A., Yamagata, K., Sugai, T., Takahashi, Y., Sugawara, E., Tamura, A., Yaegashi, H., Yamagishi, N., Takahashi, T., Isogai, M., Takahashi, H., and Yoshikawa, N. (2009). Apple latent spherical virus vectors for reliable and effective virus-induced gene silencing among a broad range of plants including tobacco, tomato, Arabidopsis thaliana, cucurbits, and legumes. *Virology* 386**,** 407-416.

Iwata, H., and Ukai, Y. (2002). SHAPE: a computer program package for quantitative evaluation of biological shapes based on elliptic Fourier descriptors. *Journal of Heredity* 93**,** 384-385.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M.E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.F., Weissenbach, J., Quetier, F., Wincker, P., and French-Italian Public Consortium for Grapevine Genome, C.

(2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449**,** 463-467.

Jan, I., Rab, A., and Sajid, M. (2012). Storage Performance of Apple Cultivars Harvest at Different Stages of Maturity. *The Journal of Animal & Plant Sciences* 22**,** 438-447.

Jansky, S.H., Dawson, J., and Spooner, D.M. (2015). How do we address the disconnect between genetic and morphological diversity in germplasm collections? *Am J Bot* 102**,** 1213-1215.

Javed, M., Chai, M., and Othman, R. (2004). Study of resistance of Musa acuminata to Fusarium oxysporum using RAPD markers. *Biologia Plantarum* 48**,** 93-99.

Jia, H., and Wang, N. (2014). Targeted genome editing of sweet orange using Cas9/sgRNA. *PLoS One* 9**,** e93806.

Johnston, J.W., Hewett, E.W., and Hertog, M.L.a.T.M. (2002). Postharvest softening of apple (Malus domestica) fruit: A review. *New Zealand Journal of Crop and Horticultural Science* 30**,** 145-160.

Joshi, R.K., and Nayak, S. (2010). Gene pyramiding-A broad spectrum technique for developing durable stress resistance in crops. *Biotechnology and Molecular Biology Review* 5**,** 51-60.

Jöst, M., Hensel, G., Kappel, C., Druka, A., Sicard, A., Hohmann, U., Beier, S., Himmelbach, A., Waugh, R., Kumlehn, J., Stein, N., and Lenhard, M. (2016). The INDETERMINATE DOMAIN Protein BROAD LEAF1 Limits Barley Leaf Width by Restricting Lateral Proliferation. *Current Biology* 26**,** 903-909.

Jung, S., Ficklin, S.P., Lee, T., Cheng, C.H., Blenda, A., Zheng, P., Yu, J., Bombarely, A., Cho, I., Ru, S., Evans, K., Peace, C., Abbott, A.G., Mueller, L.A., Olmstead, M.A., and Main, D. (2014). The Genome Database for Rosaceae (GDR): year 10 update. *Nucleic Acids Res* 42**,** D1237-1244.

Kane, D.A., Roge, P., and Snapp, S.S. (2016). A Systematic Review of Perennial Staple Crops Literature Using Topic Modeling and Bibliometric Analysis. *PLoS One* 11**,** e0155788.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42**,** 348-354.

Karlova, R., Chapman, N., David, K., Angenent, G.C., Seymour, G.B., and De Maagd, R.A. (2014). Transcriptional control of fleshy fruit development and ripening. *J Exp Bot* 65**,** 4527-4541.

Kenis, K., Keulemans, J., and Davey, M.W. (2008). Identification and stability of QTLs for fruit quality traits in apple. *Tree Genetics & Genomes* 4**,** 647-661.

Khan, M.A., and Korban, S.S. (2012). Association mapping in forest trees and fruit crops. *Journal of experimental botany* 63**,** 4045-4060.

Khan, M.A., Olsen, K.M., Sovero, V., Kushad, M.M., and Korban, S.S. (2014). Fruit Quality Traits Have Played Critical Roles in Domestication of the Apple. *The Plant Genome* 7**,** 0.

Khanal, B.P., Shrestha, R., Hückstädt, L., and Knoche, M. (2013). Russeting in apple seems unrelated to the mechanical properties of the cuticle at maturity. *HortScience* 48**,** 1135-1138.

Kishigami, R., Yamagishi, N., Ito, T., and Yoshikawa, N. (2014). Detection of apple latent spherical virus in seeds and seedlings from infected apple trees by reverse transcription quantitative PCR and deep sequencing: evidence for lack of transmission of the virus to most progeny seedlings. *Journal of General Plant Pathology* 80**,** 490-498.

Klein, L.L., Caito, M., Chapnick, C., Kitchen, C., O'hanlon, R., Chitwood, D.H., and Miller, A.J. (2017). Digital Morphometrics of Two North American Grapevines (Vitis: Vitaceae) Quantifies Leaf Variation between Species, within Species, and among Individuals. *Frontiers in Plant Science* 8.

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

Kumar, S., Chagne, D., Bink, M.C., Volz, R.K., Whitworth, C., and Carlisle, C. (2012b). Genomic selection for fruit quality traits in apple (Malus x domestica Borkh.). *PLoS One* 7**,** e36674.

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

Kumar, S., Volz, R.K., Chagne, D., and Gardiner, S. (2013b). "Breeding for Apple (Malus x domestica Borkh.) Fruit Quality Traits in the Genomics Era," in *Genomics of Plant Genetic Resources: Volume 2. Crop productivity, food security and nutritional quality,* eds. R. Tuberosa, A. Graner & E. Frison. Springer Science & Business Media), 387-416.

Kwan, E., and Friendly, M. (2012). tableplot: Represents tables as semi-graphic displays. *R package version 0.3-5*.

Lachenaud, P., Paulin, D., Ducamp, M., and Thevenin, J.M. (2007). Twenty years of agronomic evaluation of wild cocoa trees (Theobroma cacao L.) from French Guiana. *Scientia Horticulturae* 113**,** 313-321.

Lacombe, T., Boursiquot, J.M., Laucou, V., Di Vecchi-Staraz, M., Peros, J.P., and This, P. (2013). Large-scale parentage analysis in an extended set of grapevine cultivars (*Vitis vinifera* L.). *Theor Appl Genet* 126**,** 401-414.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P.A., Mcwilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23**,** 2947-2948.

Leforestier, D., Ravon, E., Muranty, H., Cornille, A., Lemaire, C., Giraud, T., Durel, C.-E., and Branca, A. (2015). Genomic basis of the differences between cider and dessert apple varieties. *Evolutionary Applications***,** n/a-n/a.

Legay, S., Guerriero, G., Deleruelle, A., Lateur, M., Evers, D., Andre, C.M., and Hausman, J.F. (2015). Apple russeting as seen through the RNA-seq lens: strong alterations in the exocarp cell wall. *Plant Mol Biol* 88**,** 21-40.

Lei, Y., Jing, Z., and Li, L. (Year). "Selection and Evaluation of a New Kiwifruit Rootstock Hybrid for Bacterial Canker Resistance", in: *VIII International Symposium on Kiwifruit 1096*), 413-420.

Lenth, R.V. (2016). Least-squares means: the R package lsmeans. *J Stat Softw* 69**,** 1-33.

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

Li, M., Duncan, K., Topp, C.N., and Chitwood, D.H. (2017a). Persistent homology and the branching topologies of plants. *Am J Bot* 104**,** 349-353.

Li, M., Frank, M.H., Coneva, V., Mio, W., Topp, C.N., and Chitwood, D.H. (2017b). Persistent homology: a tool to universally measure plant morphologies across organs and scales. *bioRxiv*.

Liang, Z., Chen, K., Li, T., Zhang, Y., Wang, Y., Zhao, Q., Liu, J., Zhang, H., Liu, C., Ran, Y., and Gao, C. (2017). Efficient DNA-free genome editing of bread wheat using CRISPR/Cas9 ribonucleoprotein complexes. *Nat Commun* 8**,** 14261.

Liang, Z., Wu, B., Fan, P., Yang, C., Duan, W., Zheng, X., Liu, C., and Li, S. (2008). Anthocyanin composition and content in grape berry skin in *Vitis* germplasm. *Food Chemistry* 111**,** 837-844.

Liang, Z., Yang, Y., Cheng, L., and Zhong, G.-Y. (2012). Polyphenolic composition and content in the ripe berries of wild *Vitis* species. *Food Chemistry* 132**,** 730-738.

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

Lijavetzky, D., Cabezas, J.A., Ibanez, A., Rodriguez, V., and Martinez-Zapater, J.M. (2007). High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* 8**,** 424.

Lu, R., Martin-Hernandez, A.M., Peart, J.R., Malcuit, I., and Baulcombe, D.C. (2003). Virus-induced gene silencing in plants. *Methods* 30**,** 296-303.

Luby, J.J., and Shaw, D.V. (2001). Does marker-assisted selection make dollars and sense in a fruit breeding program? *HortScience* 36**,** 872-879.

Lusser, M., Parisi, C., Plan, D., and Rodriguez-Cerezo, E. (2012). Deployment of new biotechnologies in plant breeding. *Nat Biotech* 30**,** 231-239.

Martínez, L., Cavagnaro, P., Boursiquot, J.-M., and Agüero, C. (2008). Molecular characterization of Bonarda-type grapevine (*Vitis vinifera* L.) cultivars from Argentina, Italy, and France. *American journal of enology and viticulture* 59**,** 287-291.

Mather, K.A., Caicedo, A.L., Polato, N.R., Olsen, K.M., Mccouch, S., and Purugganan, M.D. (2007). The extent of linkage disequilibrium in rice (Oryza sativa L.). *Genetics* 177**,** 2223-2232.

Matthies, I.E., Malosetti, M., Roder, M.S., and Van Eeuwijk, F. (2014). Genome-wide association mapping for kernel and malting quality traits using historical European barley records. *PLoS One* 9**,** e110046.

Maxted, N., Kell, S., Ford-Lloyd, B., Dulloo, E., and Toledo, Á. (2012). Toward the Systematic Conservation of Global Crop Wild Relative Diversity. *Crop Science* 52**,** 774.

Mcclure, K.A., Sawler, J., Gardner, K.M., Money, D., and Myles, S. (2014). Genomics: a potential panacea for the perennial problem. *Am J Bot* 101**,** 1780-1790.

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., Lawton-Rauh, A., Lijua, Q., Lusty, C., Michael, T., Myles, S., Naito, K., Nelson, R.L., Pontarollo, R., Richards, C.M., Rieseberg, L., Ross-Ibarra, J., Rounsley, S., Hamilton, R.S., Schurr, U., Stein, N., Tomooka, N., Van Der Knaap, E., Van Tassel, D., Toll, J., Valls, J., Varshney, R.K., Ward, J., Waugh, R., Wenzl, P., and Zamir, D. (2013). Agriculture: Feeding the future. *Nature* 499**,** 23-24.

Mccouch, S.R., Mcnally, K.L., Wang, W., and Sackville Hamilton, R. (2012). Genomics of gene banks: A case study in rice. *Am J Bot* 99**,** 407-423.

Mcvean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genet* 5**,** e1000686.

Meinhardt, L.W., Rincones, J., Bailey, B.A., Aime, M.C., Griffith, G.W., Zhang, D., and Pereira, G.A. (2008). Moniliophthora perniciosa, the causal agent of witches' broom disease of cacao: what's new from this old foe? *Mol Plant Pathol* 9**,** 577-588.

Meloni, G., and Swinnen, J. (2014). The Political Economy of European Wine Regulations. *Journal of Wine Economics* 8**,** 244-284.

Menda, N., Strickler, S.R., Edwards, J.D., Bombarely, A., Dunham, D.M., Martin, G.B., Mejia, L., Hutton, S.F., Havey, M.J., and Maxwell, D.P. (2014). Analysis of wild-species introgressions in tomato inbreds uncovers ancestral origins. *BMC Plant Biology* 14**,** 287.

Meneses, C., and Orellana, A. (2013). Using genomics to improve fruit quality. *Biological Research* 46**,** 347-352.

Michael, T.P., and Vanburen, R. (2015). Progress, challenges and the future of crop genomes. *Curr Opin Plant Biol* 24**,** 71-81.

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

Migicovsky, Z., Sawler, J., Money, D., Eibach, R., Miller, A.J., Luby, J.J., Jamieson, A.R., Velasco, D., Von Kintzel, S., Warner, J., Wuhrer, W., Brown, P.J., and Myles, S. (2016b). Genomic ancestry estimation quantifies use of wild species in grape breeding. *BMC Genomics* 17**,** 478.

Miles, C.A., and King, J. (2014). Yield, Labor, and Fruit and Juice Quality Characteristics of Machine and Hand-harvested ‘Brown Snout’ Specialty Cider Apple. *HortTechnology* 24**,** 519-526.

Miller, A.J., and Gross, B.L. (2011). From forest to field: perennial fruit crop domestication. *Am J Bot* 98**,** 1389-1414.

Mohammadi, M., Tiede, T., and Smith, K.P. (2015). PopVar: A Genome-Wide Procedure for Predicting Genetic Variance and Correlated Response in Biparental Breeding Populations. *Crop Science* 55**,** 2068.

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.Y., and Myles, S. (2015). LinkImpute: Fast and Accurate Genotype Imputation for Non-Model Organisms. *G3* 5**,** 23383-22390.

Montenegro, M. (2016). Banking on Wild Relatives to Feed the World. *Gastronomica: The Journal of Critical Food Studies* 16**,** 1-8.

Morrell, P.L., Buckler, E.S., and Ross-Ibarra, J. (2011). Crop genomics: advances and applications. *Nat Rev Genet* 13**,** 85-96.

Myles, S. (2013). Improving fruit and wine: what does genomics have to offer? *Trends Genet* 29**,** 190-196.

Myles, S., Boyko, A.R., Owens, C.L., Brown, P.J., Grassi, F., Aradhya, M.K., Prins, B., Reynolds, A., Chia, J.-M., Ware, D., Bustamante, C.D., and Buckler, E.S. (2011). Genetic structure and domestication history of the grape. *PNAS* 108**,** 3530-3535.

Myles, S., Chia, J.-M., Hurwitz, B., Simon, C., Zhong, G.Y., Buckler, E., and Ware, D. (2010). Rapid Genomic Characterization of the Genus Vitis. *PLoS ONE* 5**,** e8219.

Myles, S., Mahanil, S., Harriman, J., Gardner, K.M., Franklin, J.L., Reisch, B.I., Ramming, D.W., Owens, C.L., Li, L., Buckler, E.S., and Cadle-Davidson, L. (2015). Genetic mapping in grapevine using SNP microarray intensity values. *Molecular Breeding* 35.

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

Nakamura, K., Yamagishi, N., Isogai, M., Komori, S., Ito, T., and Yoshikawa, N. (2011). Seed and pollen transmission of Apple latent spherical virus in apple. *Journal of General Plant Pathology* 77**,** 48-53.

Narduzzi, L., Stanstrup, J., and Mattivi, F. (2015). Comparing Wild American Grapes with *Vitis vinifera*: A Metabolomics Study of Grape Composition. *J Agric Food Chem* 63**,** 6823-6834.

Nazzicari, N., Biscarini, F., Cozzi, P., Brummer, E.C., and Annicchiarico, P. (2016). Marker imputation efficiency for genotyping-by-sequencing data in rice (Oryza sativa) and alfalfa (Medicago sativa). *Molecular Breeding* 36**,** 1-16.

Nieuwenhuizen, N.J., Chen, X., Wang, M.Y., Matich, A.J., Perez, R.L., Allan, A.C., Green, S.A., and Atkinson, R.G. (2015). Natural variation in monoterpene synthesis in kiwifruit: transcriptional regulation of terpene synthases by NAC and ETHYLENE-INSENSITIVE3-like transcription factors. *Plant Physiol* 167**,** 1243-1258.

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

Noiton, D.A., and Alspach, P.A. (1996). Founding clones, inbreeding, coancestry, and status number of modern apple cultivars. *Journal of the American Society for Horticultural Science* 121**,** 773-782.

Nybom, H., Ahmadi-Afzadi, M., Sehic, J., and Hertog, M. (2012). DNA marker-assisted evaluation of fruit firmness at harvest and post-harvest fruit softening in a diverse apple germplasm. *Tree Genetics & Genomes* 9**,** 279-290.

Olsen, A.N., Ernst, H.A., Leggio, L.L., and Skriver, K. (2005). NAC transcription factors: structurally distinct, functionally diverse. *Trends Plant Sci* 10**,** 79-87.

Oraguzie, N.C., Iwanami, H., Soejima, J., Harada, T., and Hall, A. (2004). Inheritance of the Md-ACS1 gene and its relationship to fruit softening in apple (Malus x domestica Borkh.). *Theor Appl Genet* 108**,** 1526-1533.

Owens, C.L. (2008). "Grapes," in *Temperate Fruit Crop Breeding*. Springer), 197-233.

Owens, C.L. (2011). "Linkage disequilibrium and prospects for association mapping in *Vitis*," in *Genetics, genomics and breeding of grapes,* eds. A.-F. Adam-Blondon, J.M. Martínez-Zapater & C. Kole. CRC Press), 93-110.

Parisi, L., Lespinasse, Y., Guillaumes, J., and Krüger, J. (1993). A new race of Venturia inaequalis virulent to apples with resistance due to the Vf gene. *Phytopathology* 83**,** 533-537.

Patocchi, A., Walser, M., Tartarini, S., Broggini, G.A., Gennari, F., Sansavini, S., and Gessler, C. (2005). Identification by genome scanning approach (GSA) of a microsatellite tightly associated with the apple scab resistance gene Vm. *Genome* 48**,** 630-636.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet* 2**,** e190.

Paul, J.Y., Becker, D.K., Dickman, M.B., Harding, R.M., Khanna, H.K., and Dale, J.L. (2011). Apoptosis-related genes confer resistance to Fusarium wilt in transgenic 'Lady Finger' bananas. *Plant Biotechnol J* 9**,** 1141-1148.

Pauquet, J., Bouquet, A., This, P., and Adam-Blondon, A.-F. (2001). Establishment of a local map of AFLP markers around the powdery mildew resistance gene Run1 in grapevine and assessment of their usefulness for marker assisted selection. *Theoretical and Applied Genetics* 103**,** 1201-1210.

Peace, C., and Norelli, J. (2009). "Genomics Approaches to Crop Improvement in the Rosaceae," in *Genetics and Genomics of Rosaceae,* eds. K. Folta & S. Gardiner. Springer New York), 19-53.

Peace, C.P. (2017). DNA-informed breeding of rosaceous crops: promises, progress and prospects. *Horticulture Research* 4**,** 17006.

Peil, A., Dunemann, F., Richter, K., Hoefer, M., Király, I., Flachowsky, H., and Hanke, M.-V. (Year). "Resistance breeding in apple at Dresden-Pillnitz", in: *Ecofruit-13th International Conference on Cultivation Technique and Phytopathological Problems in Organic Fruit-Growing: Proceedings to the Conference from 18thFebruary to 20th February 2008 at Weinsberg/Germany*), 220-225.

Pimentel, D., Wilson, C., Mccullum, C., Huang, R., Dwen, P., Flack, J., Tran, Q., Saltman, T., and Cliff, B. (1997). Economic and environmental benefits of biodiversity. *BioScience***,** 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., and Morgante, M. (2016). Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. *Mol Biol Evol* 33**,** 2706-2719.

Pirona, R., Eduardo, I., Pacheco, I., Linge, C.D.S., Miculan, M., Verde, I., Tartarini, S., Dondini, L., Pea, G., and Bassi, D. (2013). Fine mapping and identification of a candidate gene for a major locus controlling maturity date in peach. *BMC plant biology* 13**,** 166.

Ploetz, R.C., Kepler, A.K., Daniells, J., and Nelson, S.C. (2007). Banana and plantain—an overview with emphasis on Pacific island cultivars. *Species profiles for Pacific Island agroforestry***,** 21-32.

Pollefeys, P., and Bousquet, J. (2003). Molecular genetic diversity of the French-American grapevine hybrids cultivated in North America. *Genome* 46**,** 1037-1048.

Preston, A. (1954). Effects of Fruit Thinning by the Leaf Count Method on Yield, Size and Biennial Bearing of the Apple Duchess: Favourite. *Journal of horticultural science* 29**,** 269-277.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38**,** 904-909.

Purcell, S. 2009a. PLINK v1.07. Available: http://pngu.mgh.harvard.edu/purcell/plink/.

Purcell, S. (2009b). *PLINK v.1.07* [Online]. Available: http://pngu.mgh.harvard.edu/purcell/plink/ [Accessed].

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.a.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* 81**,** 559-575.

Qiu, K., Li, Z., Yang, Z., Chen, J., Wu, S., Zhu, X., Gao, S., Gao, J., Ren, G., Kuai, B., and Zhou, X. (2015). EIN3 and ORE1 Accelerate Degreening during Ethylene-Mediated Leaf Senescence by Directly Activating Chlorophyll Catabolic Genes in Arabidopsis. *PLoS Genet* 11**,** e1005399.

R Core Team (2015). " R: A language and environment for statistical computing". (Vienna, Austria: R Foundation for Statistical Computing).

R Core Team (2016). "R: A Language and Environment for Statistical Computing". (Vienna, Austria: R Foundation for Statistical Computing).

Rauf, S., Iqbal, Z., and Shahzad, M. (2013). Genetic improvement of Citrus for disease resistance. *Archives of Phytopathology and Plant Protection* 46**,** 2051-2061.

Reisch, B.I., Owens, C.L., and Cousins, P.S. (2012). "Grape," in *Fruit Breeding,* eds. M.L. Badenes & D.H. Byrne. Springer US), 225-262.

Riaz, S., Tenscher, A.C., Graziani, R., Krivanek, A.F., Ramming, D.W., and Walker, M.A. (2009). Using marker-assisted selection to breed Pierce's disease-resistant grapes. *American journal of enology and viticulture* 60**,** 199-207.

Rife, T.W., and Poland, J.A. (2014). Field Book: An Open-Source Application for Field Data Collection on Android. *Crop Science* 54**,** 1624.

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

Ruehl, E., Schmid, J., Eibach, R., and Töpfer, R. (2015). "Grapevine breeding programmes in Germany," in *Grapevine Breeding Programs for the Wine Industry,* ed. A. Reynolds. (Oxford: Woodhead Publishing), 77-101.

Sawler, J., Reisch, B., Aradhya, M.K., Prins, B., Zhong, G.Y., Schwaninger, H., Simon, C., Buckler, E., and Myles, S. (2013). Genomics assisted ancestry deconvolution in grape. *PLoS One* 8**,** e80791.

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78**,** 629-644.

Schwarz, D., and Kläring, H.-P. (2001). Allometry to estimate leaf area of tomato. *Journal of Plant Nutrition* 24**,** 1291-1309.

Shan, W., Kuang, J.-F., Chen, L., Xie, H., Peng, H.-H., Xiao, Y.-Y., Li, X.-P., Chen, W.-X., He, Q.-G., Chen, J.-Y., and Lu, W.-J. (2012). Molecular characterization of banana NAC transcription factors and their interactions with ethylene signalling component EIL during fruit ripening. *Journal of Experimental Botany* 63**,** 5171-5187.

Siar, S.V., Beligan, G.A., Sajise, A.J.C., Villegas, V.N., and Drew, R.A. (2011). Papaya ringspot virus resistance in Carica papaya via introgression from Vasconcellea quercifolia. *Euphytica* 181**,** 159-168.

Sole, X., Guino, E., Valls, J., Iniesta, R., and Moreno, V. (2006). SNPStats: a web tool for the analysis of association studies. *Bioinformatics* 22**,** 1928-1929.

Soriano, J.M., Joshi, S.G., Van Kaauwen, M., Noordijk, Y., Groenwold, R., Henken, B., Van De Weg, W.E., and Schouten, H.J. (2009). Identification and mapping of the novel apple scab resistance gene Vd3. *Tree Genetics & Genomes* 5**,** 475-482.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., Atlin, G., Jannink, J.L., and Mccouch, S.R. (2015). Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. *PLoS Genet* 11**,** e1004982.

Stirling, B., Newcombe, G., Vrebalov, J., Bosdet, I., and Bradshaw Jr., H.D. (2001). Suppressed recombination around the MXC3 locus, a major gene for resistance to poplar leaf rust. *Theoretical and Applied Genetics* 103**,** 1129-1137.

Sui, L., Liu, Y., Zhong, C., and Huang, H. (2013). Geographical distribution and morphological diversity of red-fleshed kiwifruit germplasm (Actinidia chinensis Planchon) in China. *Genetic Resources and Crop Evolution* 60**,** 1873-1883.

Sun, Q., Gates, M.J., Lavin, E.H., Acree, T.E., and Sacks, G.L. (2011). Comparison of Odor-Active Compounds in Grapes and Wines from *Vitis vinifera* and Non-Foxy American Grape Species. *J Agric Food Chem* 59**,** 10657-10664.

Suzuki, J., Tripathi, S., and Gonsalves, D. (2007). Virus-resistant transgenic papaya: commercial development and regulatory and environmental issues. *Biotechnology and plant disease managment. CAB International, Wallingford*, 436-461.

Svitashev, S., Schwartz, C., Lenderts, B., Young, J.K., and Mark Cigan, A. (2016). Genome editing in maize directed by CRISPR-Cas9 ribonucleoprotein complexes. *Nat Commun* 7**,** 13274.

Swenson, E.P. (1985). Wild *Vitis riparia* from northern US and Canada--breeding source for winter hardiness in cultivated grapes--a background of the Swenson hybrids. *Fruit Varieties Journal* 39**,** 28-31.

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

Talwara, S., Grout, B.W.W., and Toldam-Andersen, T.B. (2013). Modification of leaf morphology and anatomy as a consequence of columnar architecture in domestic apple (Malus×domestica Borkh.) trees. *Scientia Horticulturae* 164**,** 310-315.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 30**,** 2725-2729.

Tanger, P., Klassen, S., Mojica, J.P., Lovell, J.T., Moyers, B.T., Baraoidan, M., Naredo, M.E., Mcnally, K.L., Poland, J., Bush, D.R., Leung, H., Leach, J.E., and Mckay, J.K. (2017). Field-based high throughput phenotyping rapidly identifies genomic regions controlling yield components in rice. *Sci Rep* 7**,** 42839.

Tanksley, S.D., and Mccouch, S.R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277**,** 1063-1066.

This, P., Lacombe, T., and Thomas, M.R. (2006). Historical origins and genetic diversity of wine grapes. *Trends Genet* 22**,** 511-519.

Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., Mcmullen, M.D., Holland, J.B., and Buckler, E.S. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43**,** 159-162.

Töpfer, R., Hausmann, L., and Eibach, R. (2011). "Molecular Breeding," in *Genetics, genomics and breeding of grapes,* eds. A.-F. Adam-Blondon, J.M. Martínez-Zapater & C. Kole. CRC Press), 160-185.

Tripathi, L., Babirye, A., Roderick, H., Tripathi, J.N., Changa, C., Urwin, P.E., Tushemereirwe, W.K., Coyne, D., and Atkinson, H.J. (2015). Field resistance of transgenic plantain to nematodes has potential for future African food security. *Sci Rep* 5**,** 8127.

Tsuge, T., Tsukaya, H., and Uchimiya, H. (1996). Two independent and polarized processes of cell elongation regulate leaf blade expansion in Arabidopsis thaliana (L.) Heynh. *Development* 122**,** 1589-1600.

Vallebona, C., Mantino, A., and 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 Eerdewegh, P., Little, R.D., Dupuis, J., Del Mastro, R.G., Falls, K., Simon, J., Torrey, D., Pandit, S., Mckenny, J., Braunschweiger, K., Walsh, A., Liu, Z., Hayward, B., Folz, C., Manning, S.P., Bawa, A., Saracino, L., Thackston, M., Benchekroun, Y., Capparell, N., Wang, M., Adair, R., Feng, Y., Dubois, J., Fitzgerald, M.G., Huang, H., Gibson, R., Allen, K.M., Pedan, A., Danzig, M.R., Umland, S.P., Egan, R.W., Cuss, F.M., Rorke, S., Clough, J.B., Holloway, J.W., Holgate, S.T., and Keith, T.P. (2002). Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* 418**,** 426-430.

Van Nocker, S., and Gardiner, S.E. (2014). Breeding better cultivars, faster: applications of new technologies for the rapid deployment of superior horticultural tree crops. *Hortic Res* 1**,** 14022.

Varshney, R.K., Terauchi, R., and Mccouch, S.R. (2014). Harvesting the Promising Fruits of Genomics: Applying Genome Sequencing Technologies to Crop Breeding. *PLoS Biol* 12**,** e1001883.

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., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L.M., Gutin, N., Lanchbury, J., Macalma, T., Mitchell, J.T., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V.T., King, S.T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M.M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A.C., Bus, V., Chagne, D., Crowhurst, R.N., Gleave, A.P., Lavezzo, E., Fawcett, J.A., Proost, S., Rouze, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R.P., Durel, C.-E., Gutin, A., Bumgarner, R.E., Gardiner, S.E., Skolnick, M., Egholm, M., Van De Peer, Y., Salamini, F., and Viola, R. (2010a). The genome of the domesticated apple (Malus [times] domestica Borkh.). *Nat Genet* 42**,** 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., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L.M., Gutin, N., Lanchbury, J., Macalma, T., Mitchell, J.T., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V.T., King, S.T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M.M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A.C., Bus, V., Chagne, D., Crowhurst, R.N., Gleave, A.P., Lavezzo, E., Fawcett, J.A., Proost, S., Rouze, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R.P., Durel, C.E., Gutin, A., Bumgarner, R.E., Gardiner, S.E., Skolnick, M., Egholm, M., Van De Peer, Y., Salamini, F., and Viola, R. (2010b). The genome of the domesticated apple (Malus x domestica Borkh.). *Nat Genet* 42**,** 833-839.

Verde, I., Bassil, N., Scalabrin, S., Gilmore, B., Lawley, C.T., Gasic, K., Micheletti, D., Rosyara, U.R., Cattonaro, F., Vendramin, E., Main, D., Aramini, V., Blas, A.L., Mockler, T.C., Bryant, D.W., Wilhelm, L., Troggio, M., Sosinski, B., Aranzana, M.J., Arus, P., Iezzoni, A., Morgante, M., and Peace, C. (2012). Development and evaluation of a 9K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. *PLoS One* 7**,** e35668.

Visser, T., and Verhaegh, J. (1978). Inheritance and selection of some fruit characters of apple. II. The relation between leaf and fruit pH as a basis for preselection. *Euphytica* 27**,** 761-765.

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

Walker, M.A., Riaz, S., and Tenscher, A. (Year). "Optimizing the Breeding of Pierce's Disease Resistant Winegrapes with Marker-Assisted Selection", in: *International Society for Horticultural Science (ISHS)*: Acta Hortic), 139-143.

Wang, M., Li, M., and Meng, A. (Year). "Selection of a new red-fleshed kiwifruit cultivar 'Hongyang'", in: *V International Symposium on Kiwifruit 610*), 115-117.

Wang, Y.-C., Zhang, L., Man, Y.-P., Li, Z.-Z., and Qin, R. (2012). Phenotypic characterization and simple sequence repeat identification of red-fleshed kiwifruit germplasm accessions. *HortScience* 47**,** 992-999.

Warschefsky, E.J., Klein, L.L., Frank, M.H., Chitwood, D.H., Londo, J.P., Von Wettberg, E.J., and Miller, A.J. (2016). Rootstocks: Diversity, Domestication, and Impacts on Shoot Phenotypes. *Trends Plant Sci* 21**,** 418-437.

Warton, D.I., Duursma, R.A., Falster, D.S., and Taskinen, S. (2012). smatr 3- an R package for estimation and inference about allometric lines. *Methods in Ecology and Evolution* 3**,** 257-259.

Watkinsa, C.B., Nocka, J.F., and Whitakerb, B.D. (2000). Responses of early, mid and late season apple cultivars to postharvest application of 1-methylcyclopropene (1-MCP) under air and controlled atmosphere storage conditions. *Postharvest Biology and Technology* 19**,** 17–32.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wolter, F., and Puchta, H. (2017). Knocking out consumer concerns and regulator's rules: efficient use of CRISPR/Cas ribonucleoprotein complexes for genome editing in cereals. *Genome Biol* 18**,** 43.

Wu, G.A., Prochnik, S., Jenkins, J., Salse, J., Hellsten, U., Murat, F., Perrier, X., Ruiz, M., Scalabrin, S., Terol, J., Takita, M.A., Labadie, K., Poulain, J., Couloux, A., Jabbari, K., Cattonaro, F., Del Fabbro, C., Pinosio, S., Zuccolo, A., Chapman, J., Grimwood, J., Tadeo, F.R., Estornell, L.H., Munoz-Sanz, J.V., Ibanez, V., Herrero-Ortega, A., Aleza, P., Perez-Perez, J., Ramon, D., Brunel, D., Luro, F., Chen, C., Farmerie, W.G., Desany, B., Kodira, C., Mohiuddin, M., Harkins, T., Fredrikson, K., Burns, P., Lomsadze, A., Borodovsky, M., Reforgiato, G., Freitas-Astua, J., Quetier, F., Navarro, L., Roose, M., Wincker, P., Schmutz, J., Morgante, M., Machado, M.A., Talon, M., Jaillon, O., Ollitrault, P., Gmitter, F., and Rokhsar, D. (2014). Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat Biotechnol* 32**,** 656-662.

Wünsche, J.N., Palmer, J.W., and Greer, D.H. (2000). Effects of Crop Load on Fruiting and Gas-exchange Characteristics of 'Braeburn'/M. 26 Apple Trees at Full Canopy. *Journal of the American Society for Horticultural Science* 125**,** 93-99.

Yamagishi, N., Li, C., and Yoshikawa, N. (2016). Promotion of Flowering by Apple Latent Spherical Virus Vector and Virus Elimination at High Temperature Allow Accelerated Breeding of Apple and Pear. *Front Plant Sci* 7**,** 171.

Yamagishi, N., Sasaki, S., Yamagata, K., Komori, S., Nagase, M., Wada, M., Yamamoto, T., and Yoshikawa, N. (2011). Promotion of flowering and reduction of a generation time in apple seedlings by ectopical expression of the Arabidopsis thaliana FT gene using the Apple latent spherical virus vector. *Plant Mol Biol* 75**,** 193-204.

Yan, J., Shah, T., Warburton, M.L., Buckler, E.S., Mcmullen, M.D., and Crouch, J. (2009). Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS One* 4**,** e8451.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88**,** 76-82.

Zhang, D., and Motilal, L. (2016). "Origin, dispersal, and current global distribution of cacao genetic diversity," in *Cacao Diseases*. Springer), 3-31.

Zhang, J., Hausmann, L., Eibach, R., Welter, L.J., Topfer, R., and Zyprian, E.M. (2009). A framework map from grapevine V3125 (*Vitis vinifera* 'Schiava grossa' x 'Riesling') x rootstock cultivar 'Börner' (*Vitis riparia* x *Vitis cinerea*) to localize genetic determinants of phylloxera root resistance. *Theor Appl Genet* 119**,** 1039-1051.

Zhang, J., Wu, X., Niu, R., Liu, Y., Liu, N., Xu, W., and Wang, Y. (2015). Cold-resistance evaluation in 25 wild grape species. *VITIS-Journal of Grapevine Research* 51**,** 153.

Zhang, X.J., Wang, L.X., Chen, X.X., Liu, Y.L., Meng, R., Wang, Y.J., and Zhao, Z.Y. (2014). A and MdMYB1 allele-specific markers controlling apple (Malus x domestica Borkh.) skin color and suitability for marker-assisted selection. *Genet Mol Res* 13**,** 9103-9114.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., and Ordovas, J.M. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics* 42**,** 355-360.

Zhu, Y., Evans, K., and Peace, C. (2010). Utility testing of an apple skin color MdMYB1 marker in two progenies. *Molecular Breeding* 27**,** 525-532.

Appendix I: Quantifying the genetic basis of leaf shape (Chapter 2)

Table I-I. Comparison of leaf phenotypes between accessions based on metadata. Bonferroni-adjusted *p*-values resulting from a Mann-Whitney U test estimating the difference between accessions based on species (*Malus domestica*/*Malus sieversii*) for the leaf phenotypes examined.

Phenotype	***M. domestica* / *M. sieversii***
EFD PC1	1
EFD PC2	0.004048534
EFD PC3	1
EFD PC4	1
EFD PC5	0.003316918
PH PC1	0.683551321
PH PC2	1
PH PC3	1.06E-05
PH PC4	1
PH PC5	2.99E-05
dry weight	0.004296286
leaf mass per area	0.318844203
surface area	0.002547523
surface area var	1
length	0.21019243
length var	1
width	0.000206938
width var	1
minor	0.000413503
minor var	1
major	0.974530433
major var	1
aspect ratio	0.023214743
aspect ratio var	1

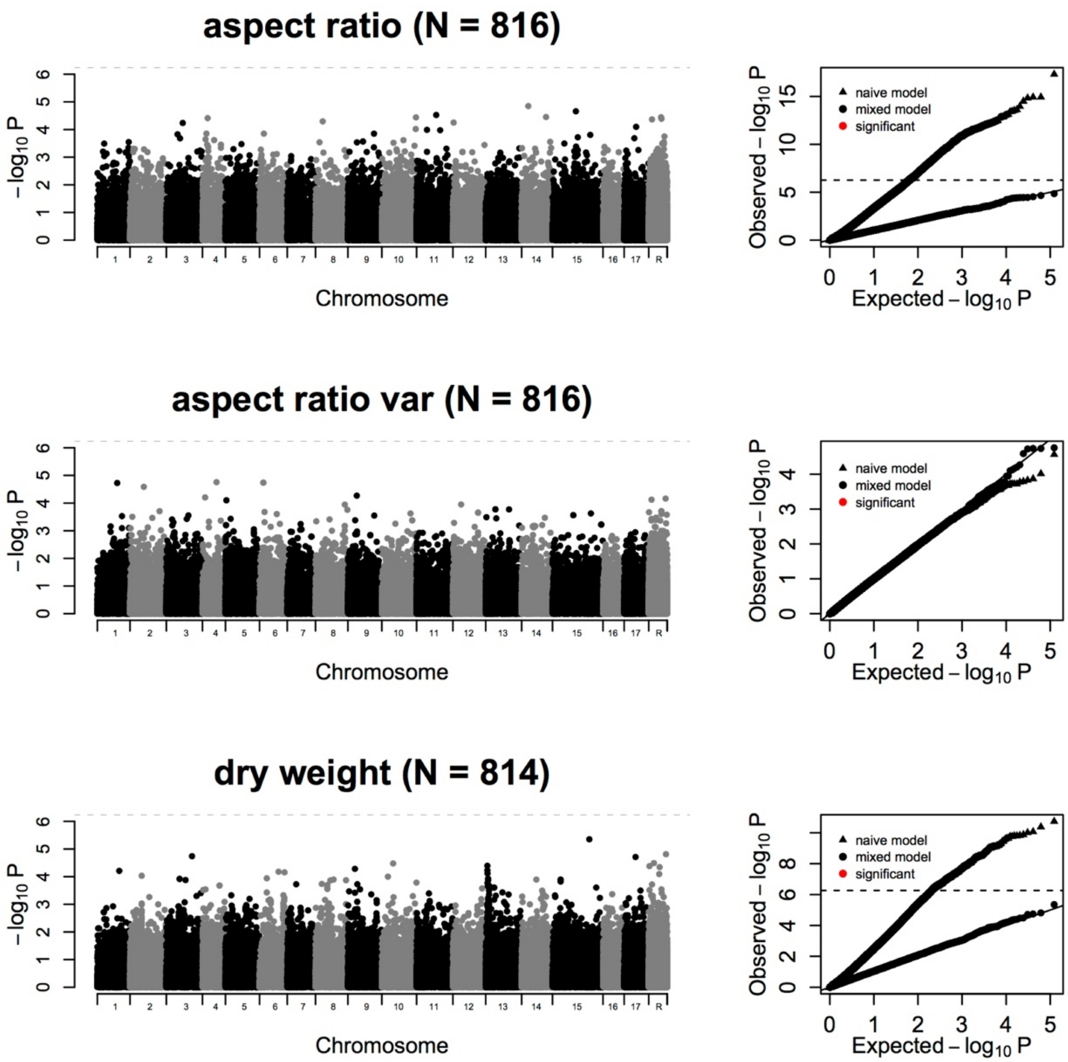


Figure I-I. GWAS results for leaf phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. P-values are log-transformed and the threshold for significance is simpleM-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

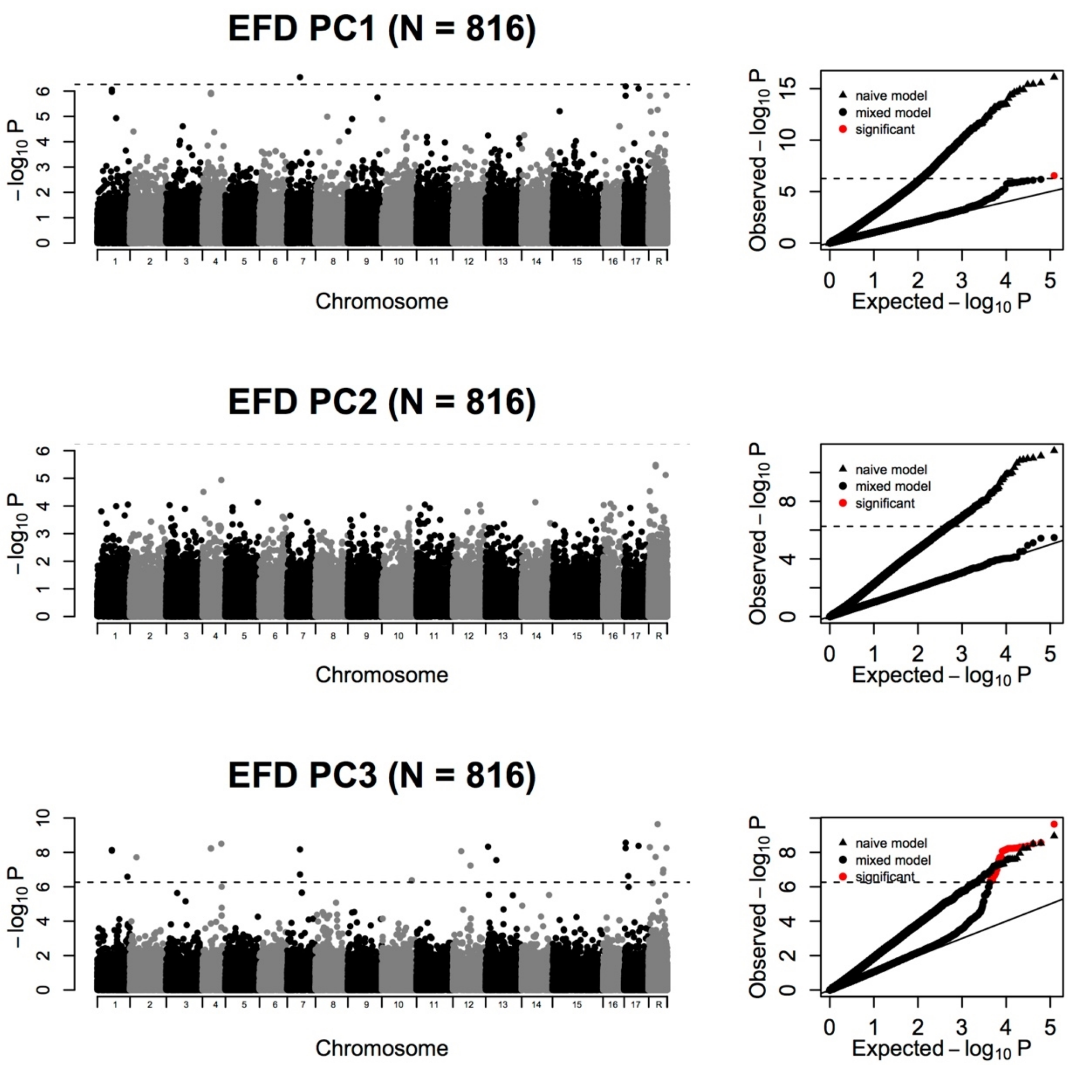


Figure I-I. GWAS results for leaf phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. P-values are log-transformed and the threshold for significance is simpleM-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

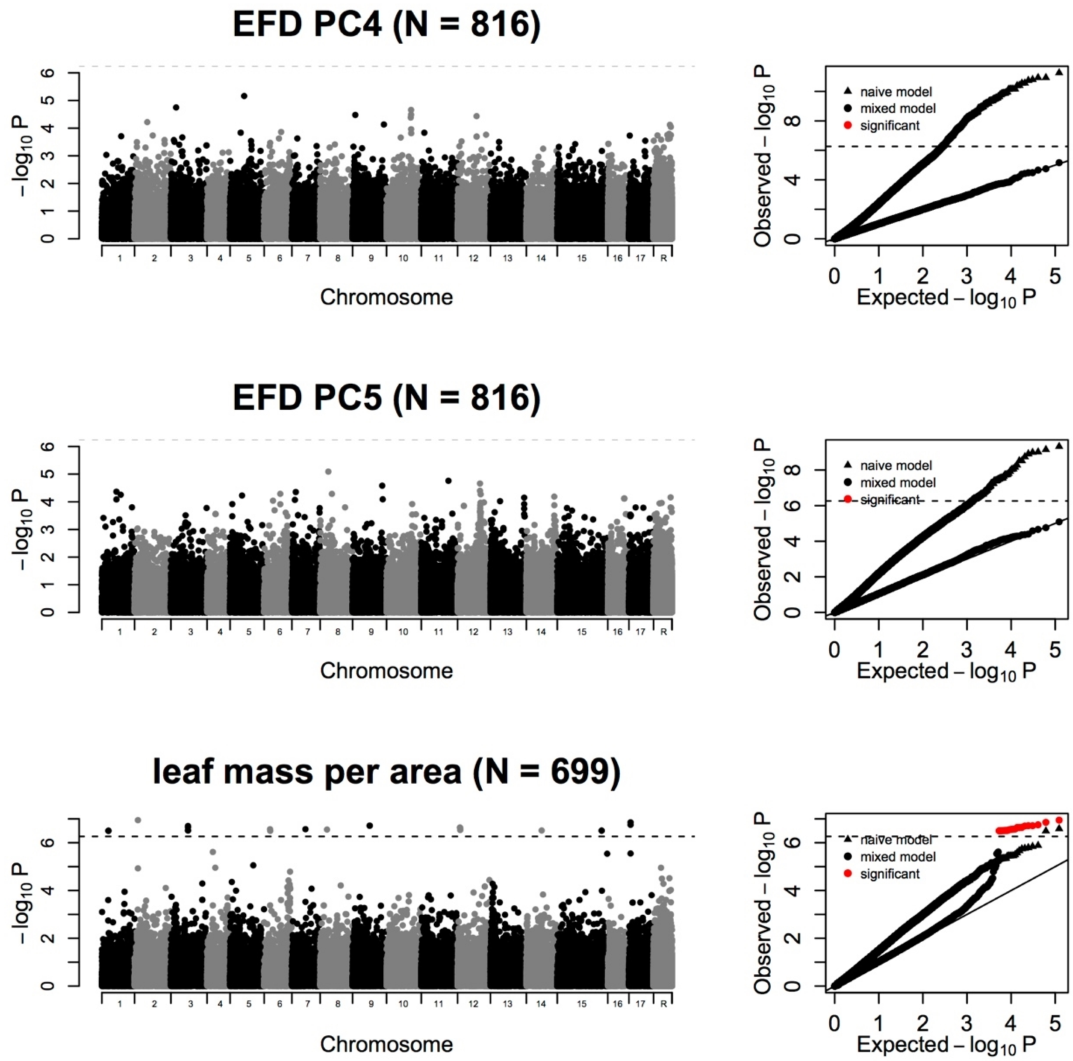


Figure I-I. GWAS results for leaf phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is simpleM-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

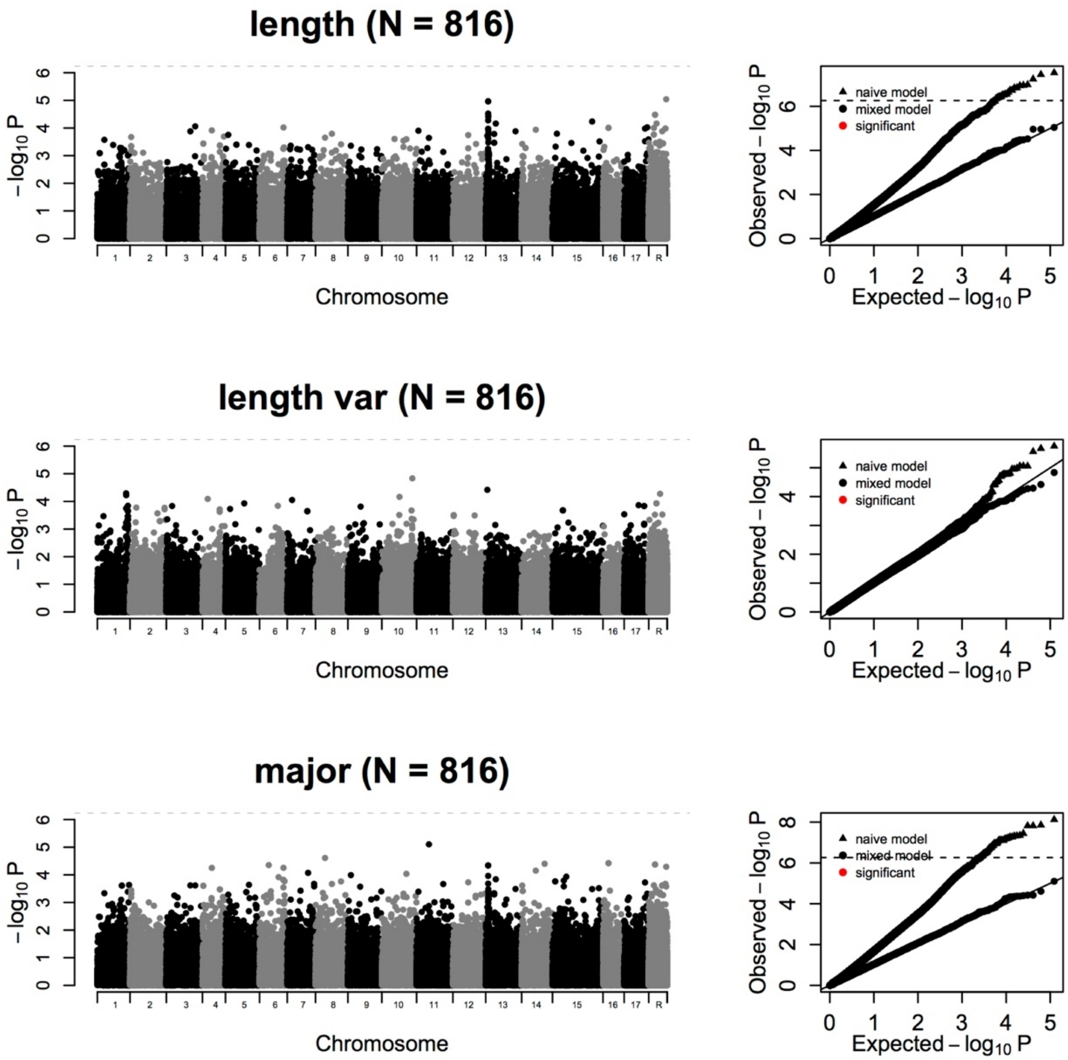


Figure I-I. GWAS results for leaf phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is simpleM-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

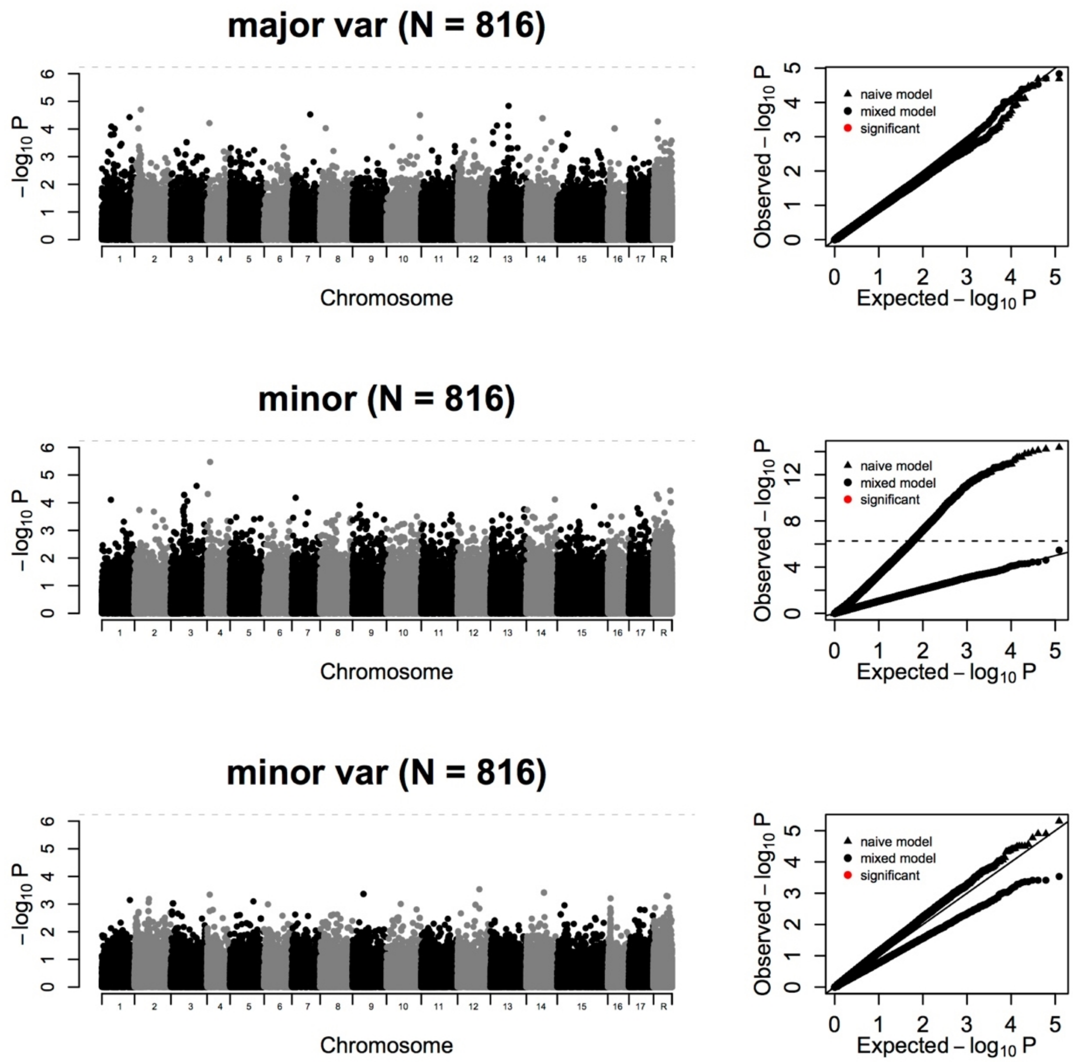


Figure I-I. GWAS results for leaf phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is simpleM-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

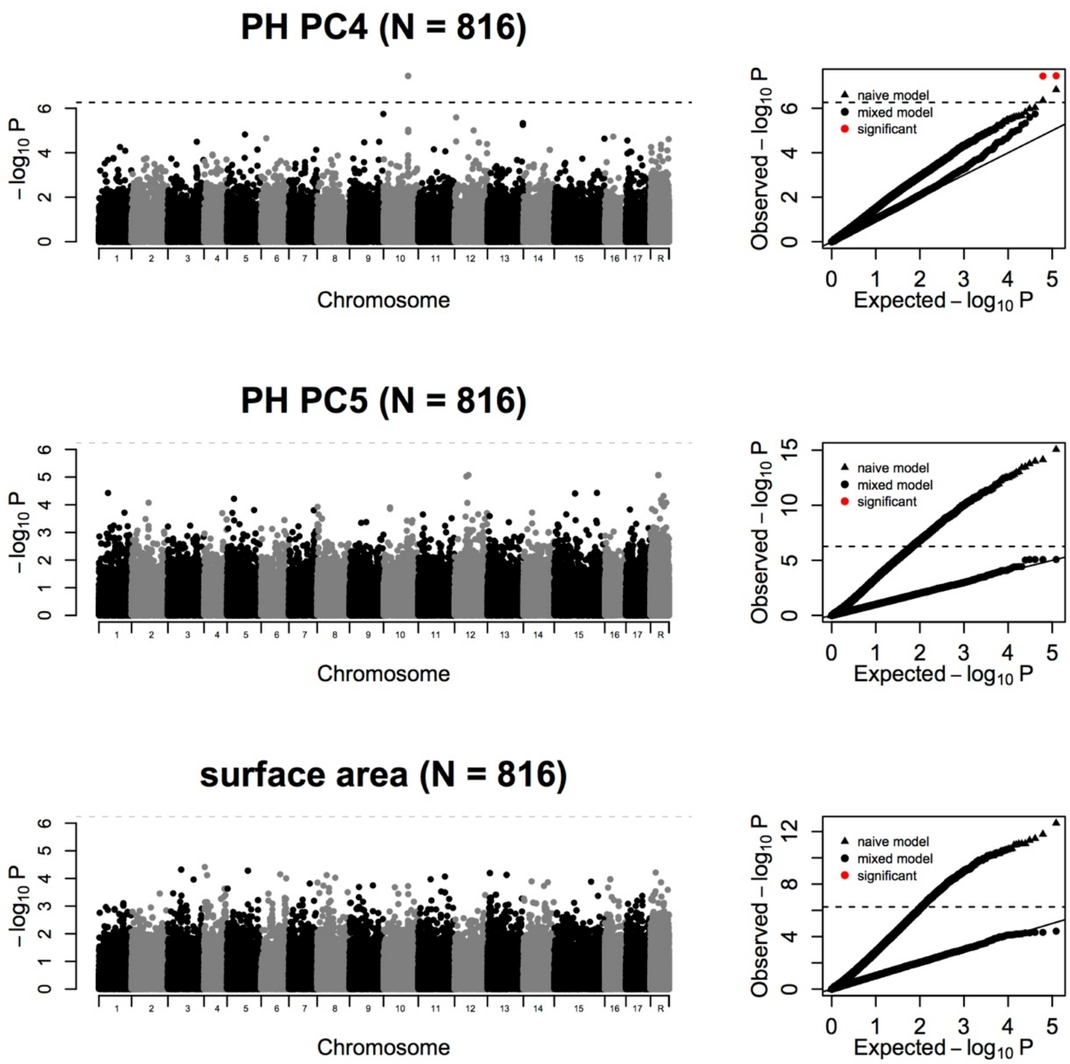


Figure I-I. GWAS results for leaf phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is simpleM-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

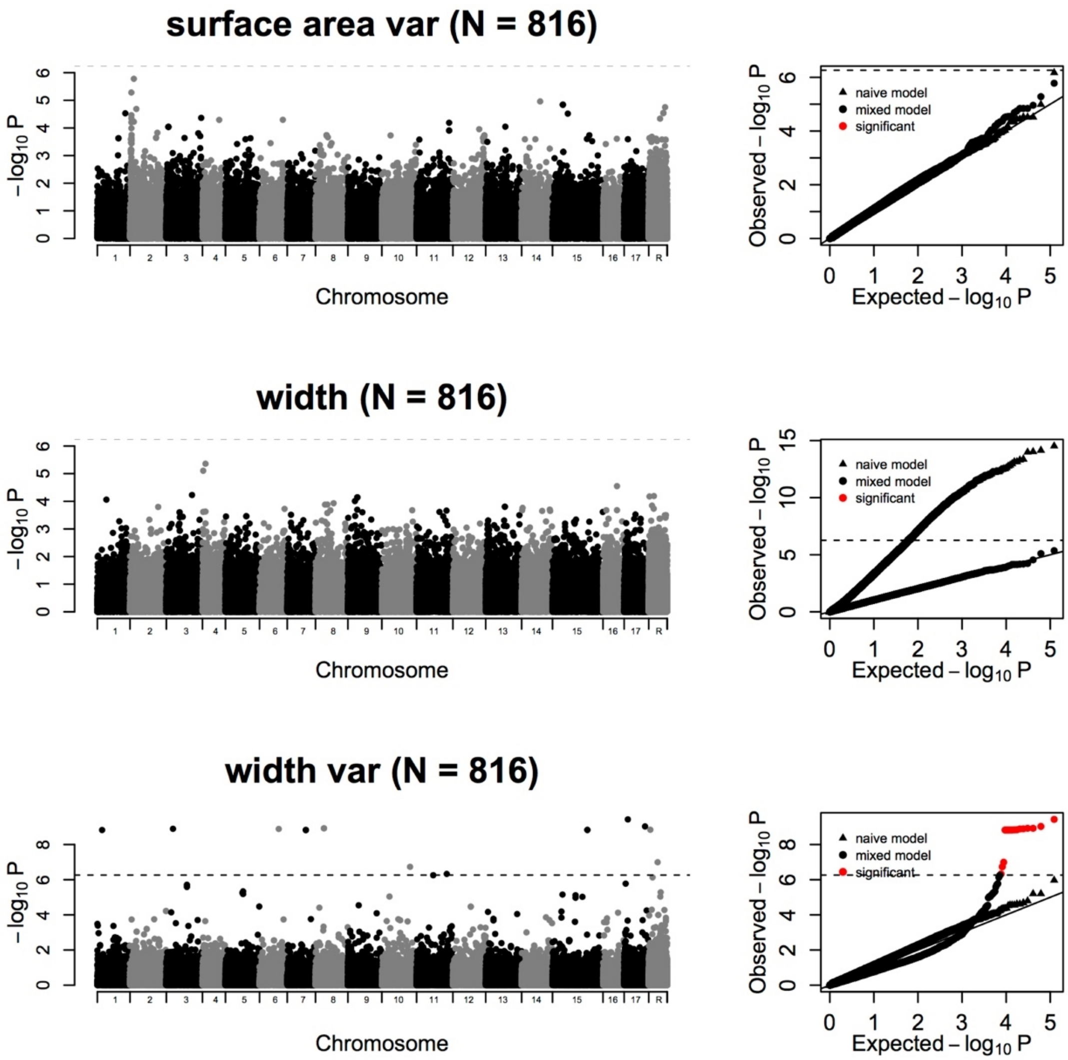


Figure I-I. GWAS results for leaf phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is simpleM-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

Table I-II. Positional information for significant leaf GWAS results. *p*-value, minor allele, minor effect, major allele, major effect and MAF are included.

Trait	Chr	Position	*p*-value	Minor Allele	Minor Effect	Major Allele	Major Effect	MAF
EFD PC1	7	14057829	2.83E-07	A	0.230	T	-0.006	0.275
EFD PC3	1	16040767	7.20E-09	C	0.100	A	-0.001	0.112
EFD PC3	1	16040782	7.95E-09	C	0.100	A	0.000	0.117
EFD PC3	1	33244662	2.59E-07	A	0.075	G	-0.003	0.265
EFD PC3	2	7167203	1.92E-08	A	0.046	G	0.000	0.073
EFD PC3	4	9336777	5.95E-09	A	0.100	G	0.001	0.346
EFD PC3	4	9336793	5.87E-09	A	0.100	C	0.001	0.345
EFD PC3	4	9336818	5.87E-09	G	0.100	A	0.001	0.345
EFD PC3	4	20837364	3.12E-09	G	0.043	A	-0.002	0.079
EFD PC3	7	14057813	6.70E-09	A	0.054	T	0.002	0.312
EFD PC3	7	14057829	1.90E-07	A	0.057	T	0.003	0.275
EFD PC3	10	33021752	4.24E-07	G	0.060	T	0.000	0.095
EFD PC3	10	33021772	4.24E-07	C	0.060	T	0.000	0.095
EFD PC3	10	33021786	4.24E-07	C	0.060	T	0.000	0.095
EFD PC3	10	33021810	4.24E-07	C	0.060	A	0.000	0.095
EFD PC3	12	9162531	8.49E-09	A	0.047	G	-0.002	0.164
EFD PC3	12	19393888	5.80E-08	T	0.067	C	0.000	0.064
EFD PC3	13	2561476	4.65E-09	A	0.068	T	0.004	0.092
EFD PC3	13	11744890	2.80E-08	A	0.065	G	0.002	0.085
EFD PC3	17	1595328	5.66E-09	C	0.099	T	-0.002	0.054
EFD PC3	17	1678938	2.74E-09	G	0.097	A	-0.003	0.051
EFD PC3	17	4569327	2.36E-07	T	0.029	C	0.002	0.108
EFD PC3	17	15848936	4.17E-09	G	0.100	A	-0.001	0.270
EFD PC3	18	6723804	4.92E-09	C	0.100	T	0.002	0.430
EFD PC3	18	39265418	1.84E-08	T	0.066	C	-0.001	0.243
EFD PC3	18	51548255	2.27E-10	T	0.075	C	0.001	0.384
EFD PC3	18	85482786	1.53E-07	T	0.062	G	0.000	0.225
EFD PC3	18	87574849	1.01E-07	C	0.051	G	0.000	0.134
EFD PC3	18	110856726	5.52E-09	T	0.101	A	0.001	0.146
leaf mass per area	1	7234679	3.15E-07	G	0.007	A	0.000	0.101
leaf mass per area	1	7234686	3.13E-07	T	0.007	A	0.000	0.104

Trait	Chr	Position	*p*-value	Minor Allele	Minor Effect	Major Allele	Major Effect	MAF
leaf mass per area	1	7234687	3.15E-07	A	0.007	G	0.000	0.101
leaf mass per area	2	3700705	1.13E-07	C	0.007	G	0.000	0.057
leaf mass per area	3	18865039	3.05E-07	C	0.007	T	0.000	0.071
leaf mass per area	3	18865048	3.05E-07	A	0.007	G	0.000	0.071
leaf mass per area	3	18865054	2.27E-07	A	0.007	G	0.000	0.197
leaf mass per area	3	18865092	1.99E-07	A	0.007	T	0.000	0.195
leaf mass per area	6	6548794	2.71E-07	C	0.007	T	0.000	0.050
leaf mass per area	6	6765216	3.16E-07	T	0.007	C	0.000	0.052
leaf mass per area	7	14985838	2.70E-07	A	0.007	G	0.000	0.061
leaf mass per area	8	7591799	2.80E-07	G	0.007	A	0.000	0.098
leaf mass per area	9	18897545	1.93E-07	A	0.007	G	0.000	0.050
leaf mass per area	9	18897546	1.93E-07	T	0.007	G	0.000	0.050
leaf mass per area	12	2599754	2.31E-07	G	0.007	A	0.000	0.072
leaf mass per area	12	2599755	2.31E-07	A	0.007	G	0.000	0.072
leaf mass per area	12	2875079	2.66E-07	T	0.007	C	0.000	0.134
leaf mass per area	12	2875081	2.92E-07	A	0.007	G	0.000	0.064
leaf mass per area	14	16895798	3.06E-07	T	0.007	C	0.000	0.070
leaf mass per area	15	48974300	3.11E-07	G	0.007	C	0.000	0.075
leaf mass per area	15	48974306	3.11E-07	C	0.007	G	0.000	0.075
leaf mass per area	17	1752929	1.39E-07	G	0.007	C	0.000	0.137
leaf mass per area	17	1752972	1.78E-07	T	0.007	C	0.000	0.335
PH PC4	10	27087002	3.43E-08	C	-11.747	G	-0.117	0.074
PH PC4	10	27106339	3.53E-08	G	-11.710	T	0.002	0.072
width var	1	5231105	1.49E-09	A	25.484	C	0.046	0.317

Trait	**Chr**	**Position**	***p*-value**	**Minor Allele**	**Minor Effect**	**Major Allele**	**Major Effect**	**MAF**
width var	3	7321723	1.27E-09	T	25.432	C	-0.170	0.263
width var	6	21230247	1.30E-09	C	25.284	T	-0.258	0.055
width var	7	20529860	1.49E-09	A	25.512	C	0.045	0.148
width var	7	20529908	1.49E-09	A	25.512	G	0.045	0.148
width var	7	20529924	1.49E-09	T	25.512	C	0.045	0.148
width var	8	9370617	1.19E-09	C	25.681	T	0.211	0.305
width var	8	9370647	1.19E-09	C	25.681	T	0.211	0.305
width var	10	31172600	1.85E-07	T	15.534	C	-0.104	0.127
width var	11	33287797	4.71E-07	G	15.100	A	0.319	0.306
width var	15	38493217	1.46E-09	A	25.544	G	0.120	0.061
width var	15	38493221	1.49E-09	A	25.519	G	0.073	0.060
width var	17	4084876	3.71E-10	A	24.931	T	-0.608	0.097
width var	17	23055349	9.33E-10	G	25.785	T	0.336	0.142
width var	18	9197912	1.44E-09	C	25.563	G	0.101	0.135
width var	18	51270468	1.02E-07	T	16.007	C	0.163	0.133

Table I-III. Genes found within ±5 kb of SNPs exceeding significance threshold. Results are listed according to the Genome Database for Rosaceae GBrowse (accessed January 27 2017). Trait, positional information, overlapping mRNA, Go Term Name and InterPro description are listed. Only SNPs with overlapping mRNA are reported. Go Term Names and InterPro Description is only listed once for SNPs overlapping the same mRNA.

Trait	Chr	Pos	mRNA Name	GO Term Name	InterPro Description
EFD PC3	1	16040767	MDP0000 168790	O-methyltransferase activity	O-methyltransferase, family 2
				methyltransferase activity	Winged helix-turn-helix transcription repressor DNA-binding
				protein dimerization activity	Plant methyltransferase dimerisation
					O-methyltransferase, COMT, eukaryota
EFD PC3	1	16040782	MDP0000 168790		
EFD PC3	1	33244662	MDP0000 266452	protein binding	Homeodomain-like
					Myb/SANT-like domain
			MDP0000 266451	protein kinase activity	Protein kinase, catalytic domain
				ATP binding	Legume lectin, alpha chain, conserved site
				binding	Legume lectin domain
				protein serine/threonine kinase activity	Serine/threonine- / dual-specificity protein kinase, catalytic domain
				protein tyrosine kinase activity	Serine/threonine-protein kinase, active site
				protein phosphorylation	Concanavalin A-like lectin/glucanase
					Protein kinase-like domain
					Concanavalin A-like lectin/glucanase, subgroup
					Protein kinase, ATP binding site
					Legume lectin, beta chain, Mn/Ca-binding site
					Tyrosine-protein kinase, catalytic domain

Trait	Chr	Pos	mRNA Name	GO Term Name	InterPro Description
EFD PC3	2	7167203	MDP0000 246957		
			MDP0000 134259	protein binding	von Willebrand factor, type A
			MDP0000 134260		Organic solute transporter Ost-alpha
EFD PC3	4	20837364	MDP0000 304608	intracellular protein transport	Zinc finger, Sec23/Sec24-type
				ER to Golgi vesicle-mediated transpor	Sec23/Sec24, trunk domain
				zinc ion binding	Sec23/Sec24, helical domain
				COPII vesicle coat	Sec23/Sec24 beta-sandwich
EFD PC3	4	9336777	MDP0000 260328	nucleotide binding	Nucleotide-binding, alpha-beta plait
			MDP0000 316458	hydrolase activity	Nucleoside phosphatase GDA1/CD39
EFD PC3	4	9336793	MDP0000 260328		
			MDP0000 316458		
EFD PC3	4	9336818	MDP0000 260328		
			MDP0000 316458		
EFD PC3	10	33021752	MDP0000 555589	catechol oxidase activity	Polyphenol oxidase, C-terminal
				oxidation-reduction process	
EFD PC3	10	33021772	MDP0000 555589		
EFD PC3	10	33021786	MDP0000 555589		
EFD PC3	10	33021810	MDP0000 555589		
EFD PC3	12	19393888	MDP0000 168714	cell death	Mlo-related protein
				integral to membrane	
EFD PC3	12	9162531	MDP0000 666968	transport	Nucleoporin interacting component Nup93/Nic96
				nuclear pore	

Trait	Chr	Pos	mRNA Name	GO Term Name	InterPro Description
EFD PC3	13	11744890	MDP0000 836932		
			MDP0000 193683	catalytic activity	Pyruvate carboxyltransferase
					Aldolase-type TIM barrel
EFD PC3	13	2561476	MDP0000 743350	sequence-specific DNA binding transcription factor activity	AP2/ERF domain
				DNA binding	DNA-binding, integrase-type
EFD PC3	17	1595328	MDP0000 312625	binding	Pentatricopeptide repeat
			MDP0000 208843		Uncharacterised protein family UPF0118
			MDP0000 694848	protein binding	Armadillo
				binding	Armadillo-like helical
					Armadillo-type fold
EFD PC3	17	1678938	MDP0000 261713		
leaf mass per area	1	7234679	MDP0000 520923	extracellular space	Allergen Ole e 1, conserved site
					Pollen Ole e 1 allergen/extensin
leaf mass per area	1	7234686	MDP0000 520923		
leaf mass per area	1	7234687	MDP0000 520923		
leaf mass per area	2	3700705	MDP0000 273540		
			MDP0000 873812	serine-type endopeptidase activity	Peptidase S8/S53, subtilisin/kexin/sedolisi n
				identical protein binding	Proteinase inhibitor I9
				proteolysis	Peptidase S8, subtilisin-related
				negative regulation of catalytic activity	Peptidase S8, subtilisin, Asp-active site

Trait	Chr	Pos	mRNA Name	GO Term Name	InterPro Description
leaf mass per area	6	6548794	MDP0000 266305	protein kinase CK2 complex	Casein kinase II, regulatory subunit
				protein kinase regulator activity	MORN motif
					Casein kinase II, regulatory subunit, alpha-helical
					Casein kinase II, regulatory subunit, beta-sheet
leaf mass per area	6	6765216	MDP0000 166312	amino acid transmembrane transporter activity	Amino acid/polyamine transporter I
				membrane	Cationic amino acid transporter
				amino acid transmembrane transport	
leaf mass per area	7	14985838	MDP0000 134641		
leaf mass per area	8	7591799	MDP0000 142597	NA	
leaf mass per area	9	18897545	MDP0000 241908		Auxin responsive SAUR protein
leaf mass per area	9	18897546	MDP0000 241908		
leaf mass per area	15	48974300	MDP0000 281060	phosphoric ester hydrolase activity	Synaptojanin, N-terminal
leaf mass per area	15	48974306	MDP0000 281060		
PH PC4	10	27087002	MDP0000 222705		Protein of unknown function DUF1639
PH PC4	10	27106339	MDP0000 930936	protein kinase activity	Serine-threonine/tyrosine-protein kinase catalytic domain
				protein phosphorylation	Protein kinase-like domain
			MDP0000 930939	nucleic acid binding	DNA methylase, N-6 adenine-specific, conserved site
				methyltransferase activity	Putative methylase
				methylation	Methyltransferase small
			MDP0000 259036	protein binding	Bromodomain
					DREPP plasma membrane polypeptide

Trait	Chr	Pos	mRNA Name	GO Term Name	InterPro Description
width var	1	5231105	MDP0000259316	3-deoxy-7-phosphoheptulonate synthase activity	DAHP synthetase, class II
				aromatic amino acid family biosynthetic process	
width var	3	7321723	MDP0000145529	integral to membrane	Major facilitator superfamily
				transmembrane transport	Major facilitator superfamily domain, general substrate transporter
					Major facilitator superfamily domain
width var	6	21230247	MDP0000136052	protein binding	Armadillo
				binding	Armadillo-like helical
					Armadillo-type fold
width var	7	20529860	MDP0000318495	DNA binding	Histone H3
				protein binding	Ubiquitin
				nucleosome	Histone core
				nucleosome assembly	Histone-fold
					Ubiquitin supergroup
					Ubiquitin subgroup
width var	7	20529908	MDP0000318495		
width var	7	20529924	MDP0000318495		
width var	8	9370617	MDP0000265371	protein serine/threonine phosphatase activity	Protein phosphatase 2C, manganese/magnesium aspartate binding site
				protein kinase activity	Protein kinase, catalytic domain
				ATP binding	AGC-kinase, C-terminal
				protein serine/threonine kinase activity	Protein phosphatase 2C-like
				catalytic activity	Serine/threonine- / dual-specificity protein kinase, catalytic domain
				protein tyrosine kinase activity	Serine/threonine-protein kinase, active site

Trait	Chr	Pos	mRNA Name	GO Term Name	InterPro Description
				protein dephosphorylation	Protein kinase-like domain
				protein phosphorylation	Protein kinase, ATP binding site
				protein serine/threonine phosphatase complex	Tyrosine-protein kinase, catalytic domain
			MDP0000 265372	metallopeptidase activity	Peptidase M1, alanine aminopeptidase/leukotri ene A4 hydrolase
				zinc ion binding	Peptidase M1, membrane alanine aminopeptidase, N-terminal
				proteolysis	Domain of unknown function DUF3358
width var	8	9370647	MDP0000 265371		
			MDP0000 265372		
width var	10	31172600	MDP0000 328070		
width var	11	33287797	MDP0000 948862	protein kinase activity	Protein kinase, catalytic domain
				ATP binding	Legume lectin domain
				binding	Serine/threonine- / dual-specificity protein kinase, catalytic domain
				protein serine/threonine kinase activity	Serine/threonine-protein kinase, active site
				protein tyrosine kinase activity	Concanavalin A-like lectin/glucanase
				protein phosphorylation	Protein kinase-like domain
					Concanavalin A-like lectin/glucanase, subgroup
					Protein kinase, ATP binding site
					Tyrosine-protein kinase, catalytic domain
width var	15	38493217	MDP0000 406249		Protein of unknown function DUF1264
width var	15	38493221	MDP0000 406249		

Trait	Chr	Pos	mRNA Name	GO Term Name	InterPro Description
width var	17	23055349	MDP0000278380	protein kinase activity	Protein kinase, catalytic domain
				ATP binding	Serine-threonine/tyrosine-protein kinase catalytic domain
				protein serine/threonine kinase activity	Thaumatin, pathogenesis-related
				protein tyrosine kinase activity	Serine/threonine- / dual-specificity protein kinase, catalytic domain
				protein phosphorylation	Serine/threonine-protein kinase, active site
					Protein kinase-like domain
					Thaumatin, conserved site
					Tyrosine-protein kinase, catalytic domain
width var	17	4084876	MDP0000544455	metabolic process	UDP-glucuronosyl/UDP-glucosyltransferase
			MDP0000253523	metabolic process	UDP-glucuronosyl/UDP-glucosyltransferase

Appendix II: Genome to phenome mapping in apple using historical data (Chapter 3)

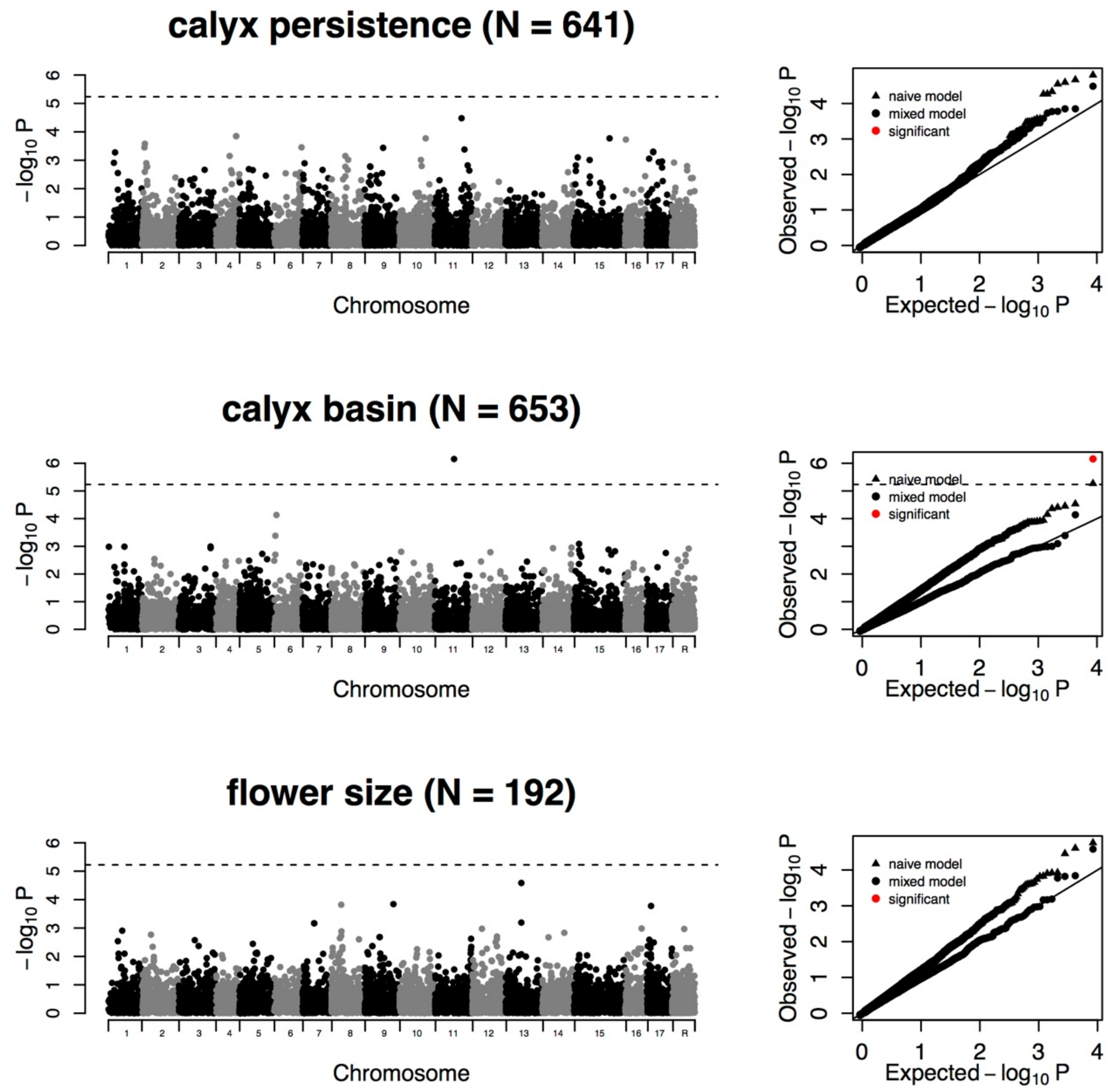


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

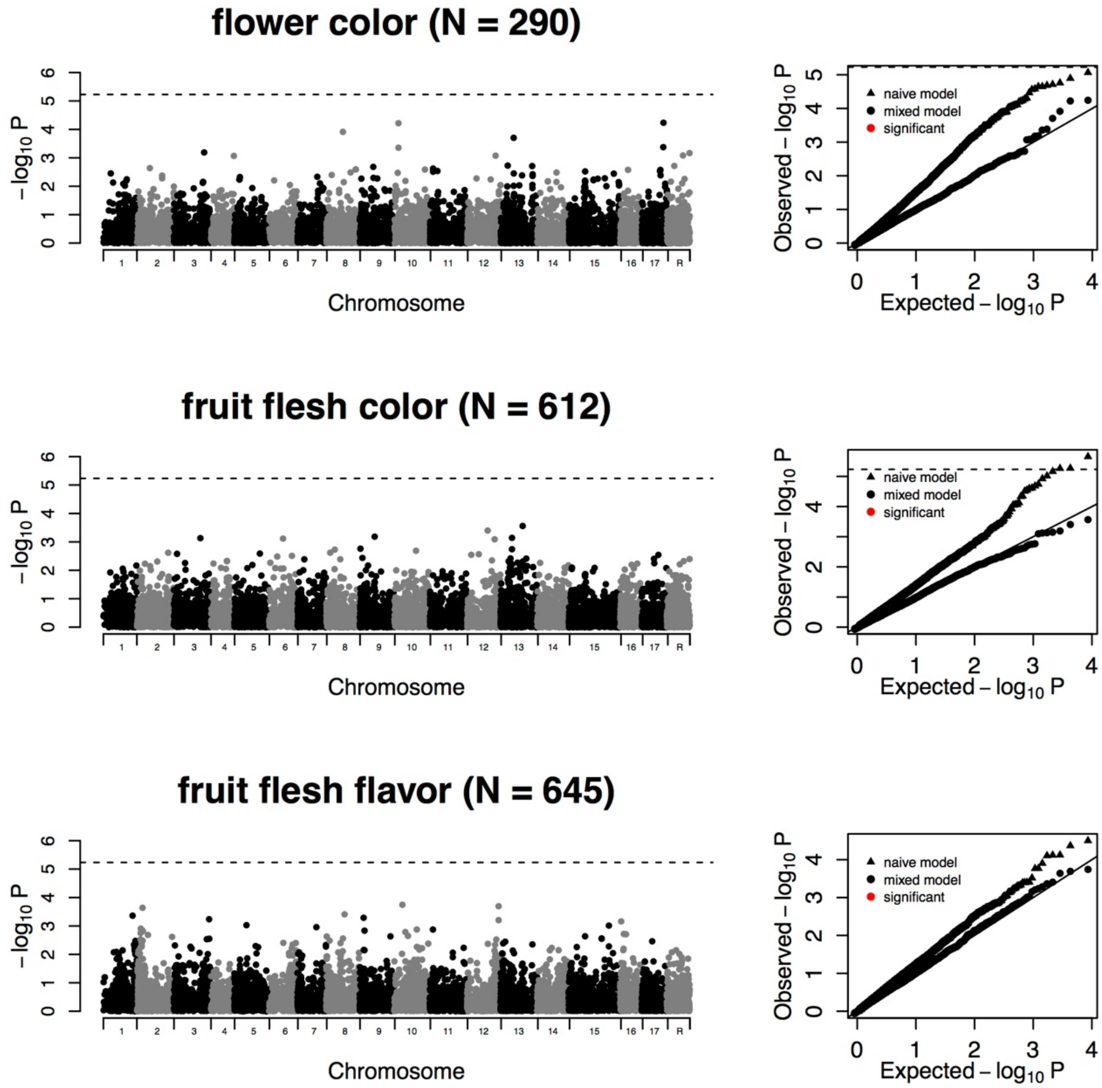


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

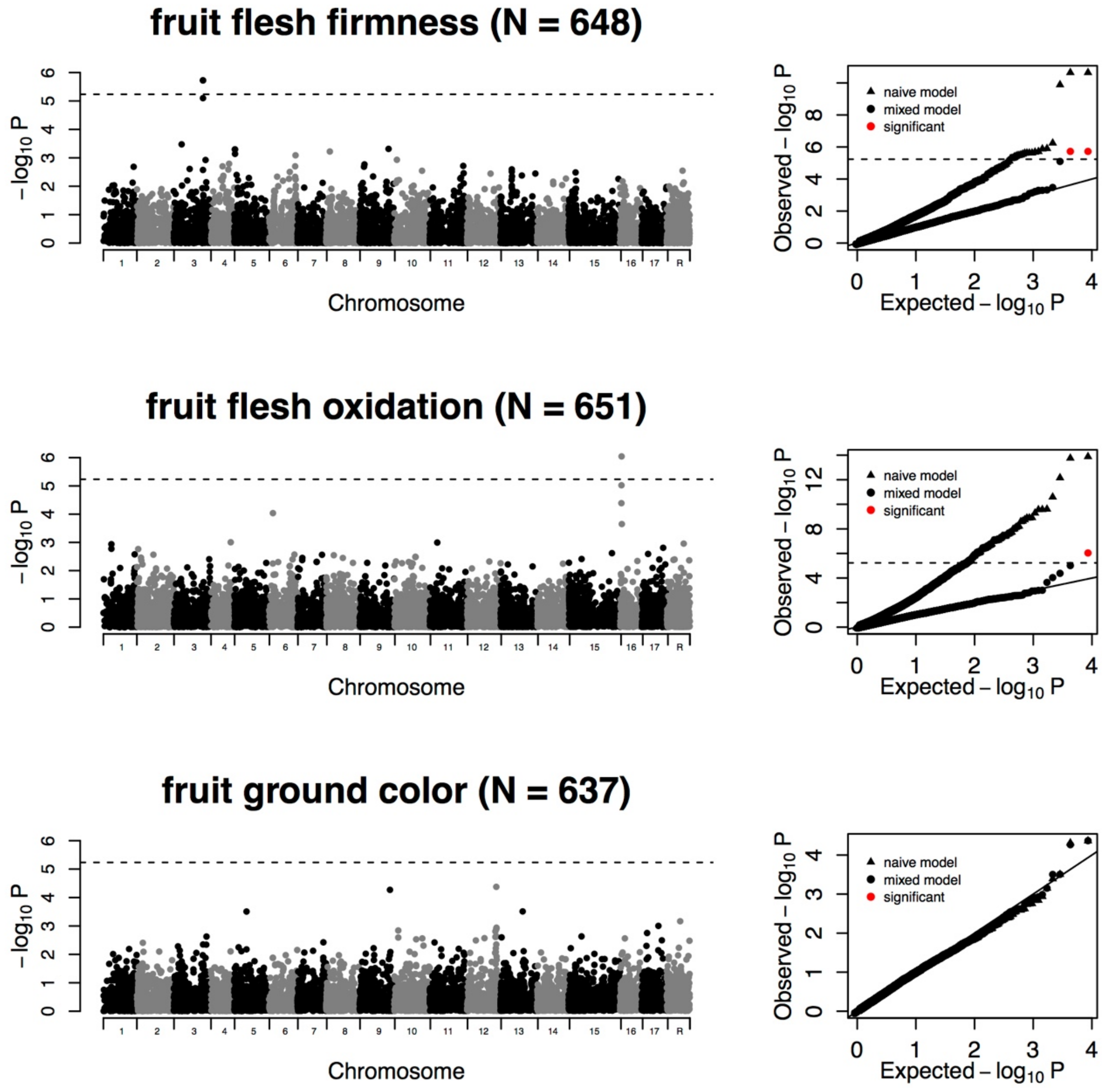


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

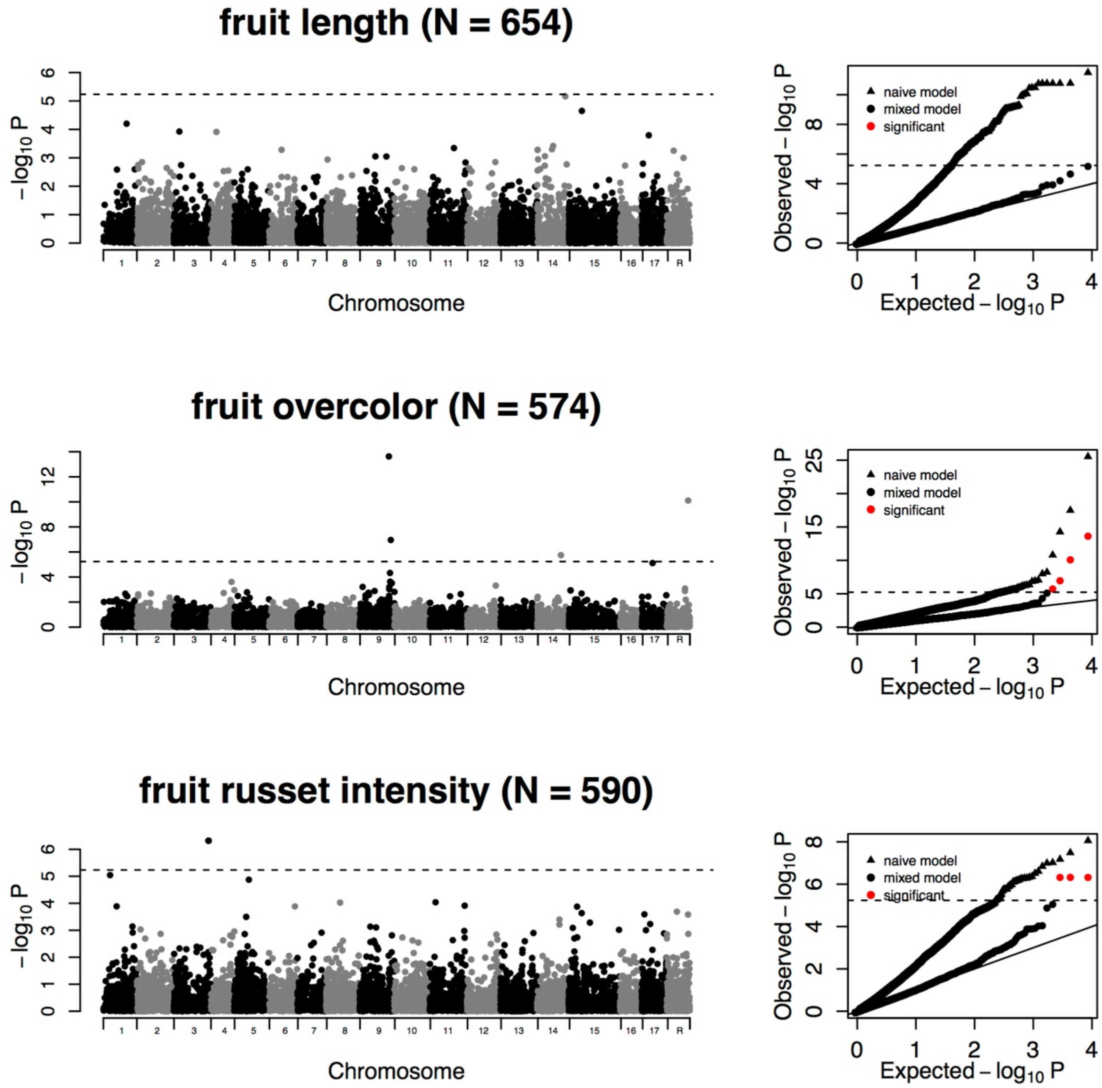


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

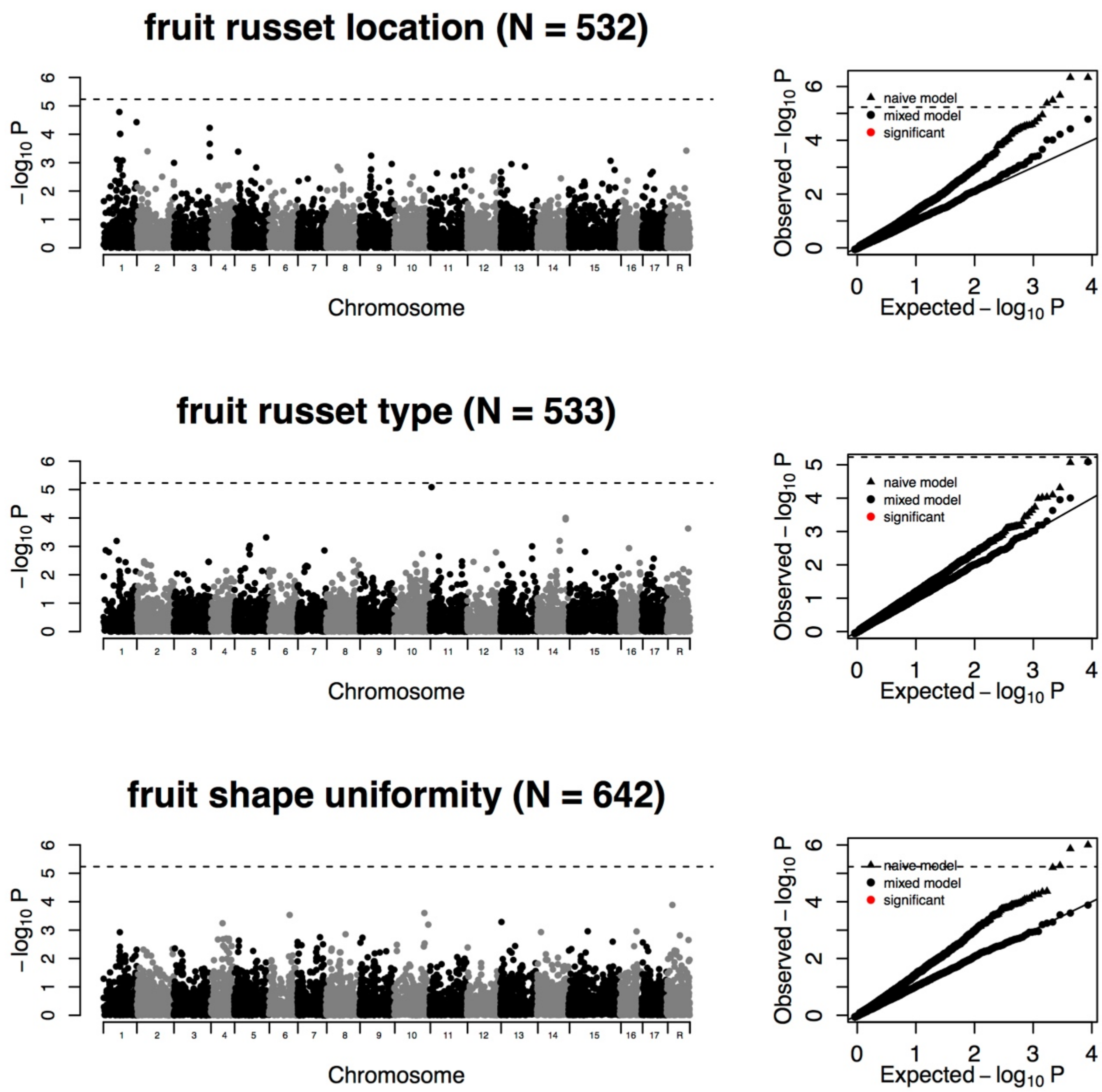


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

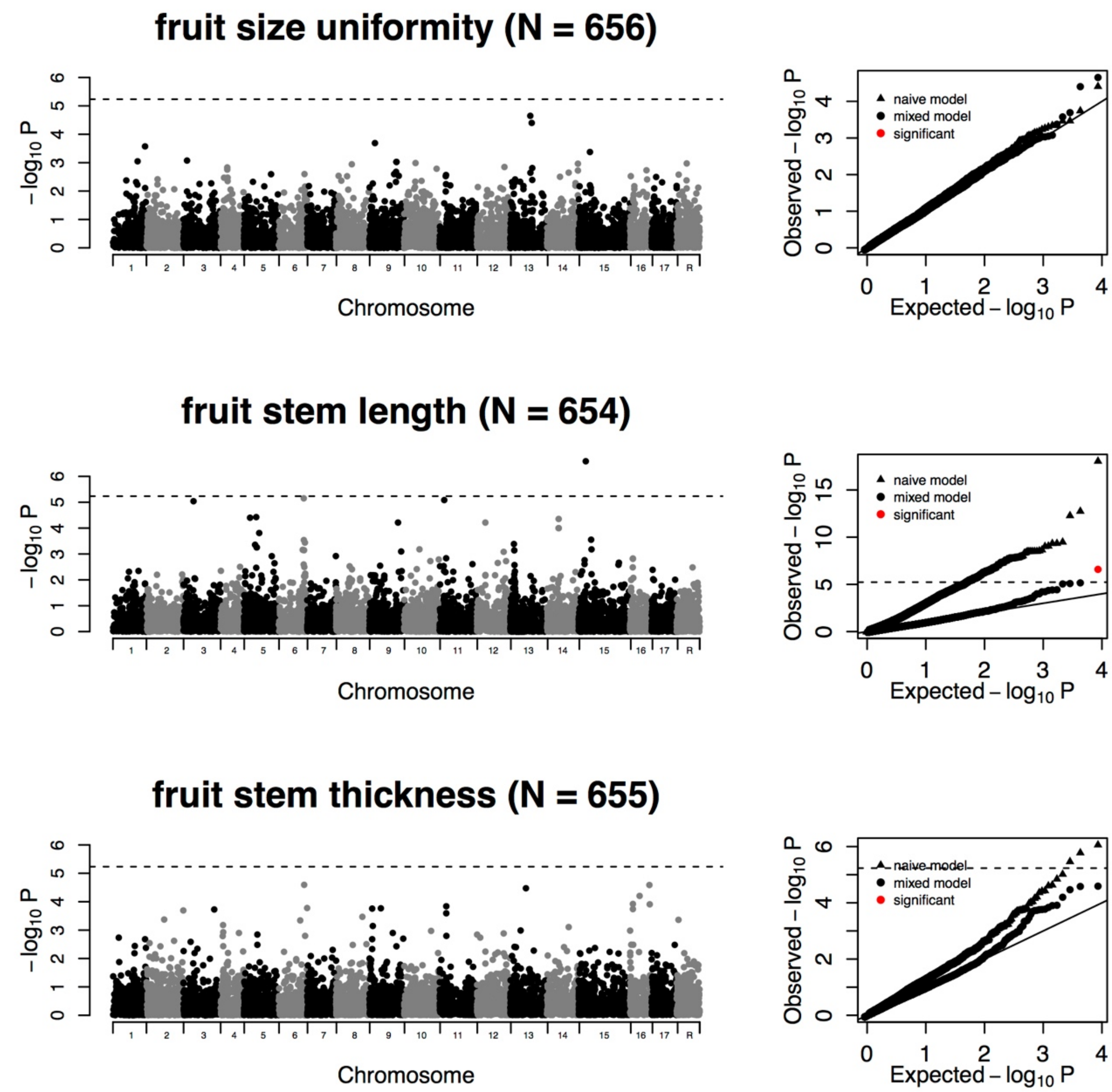


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

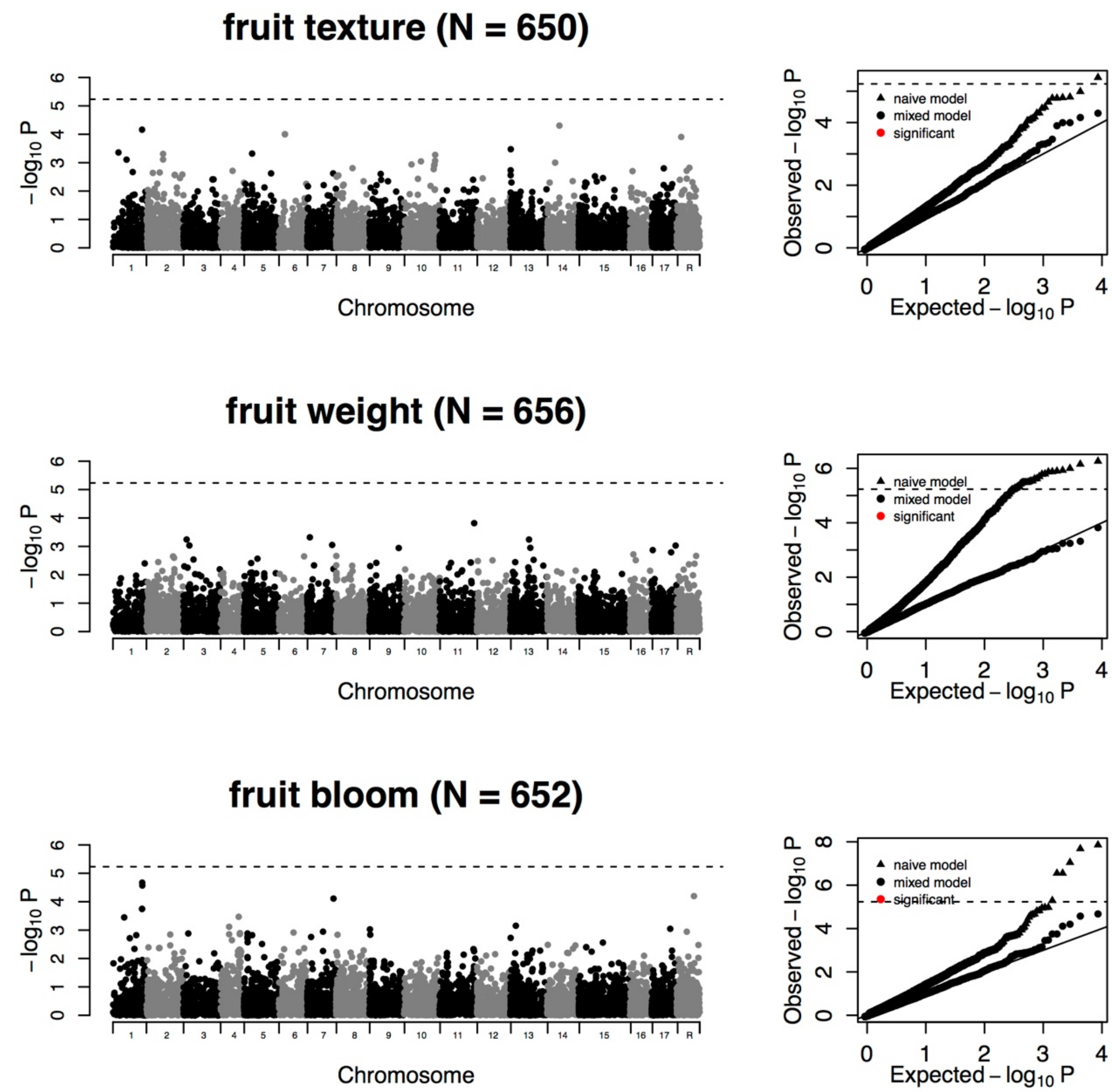


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

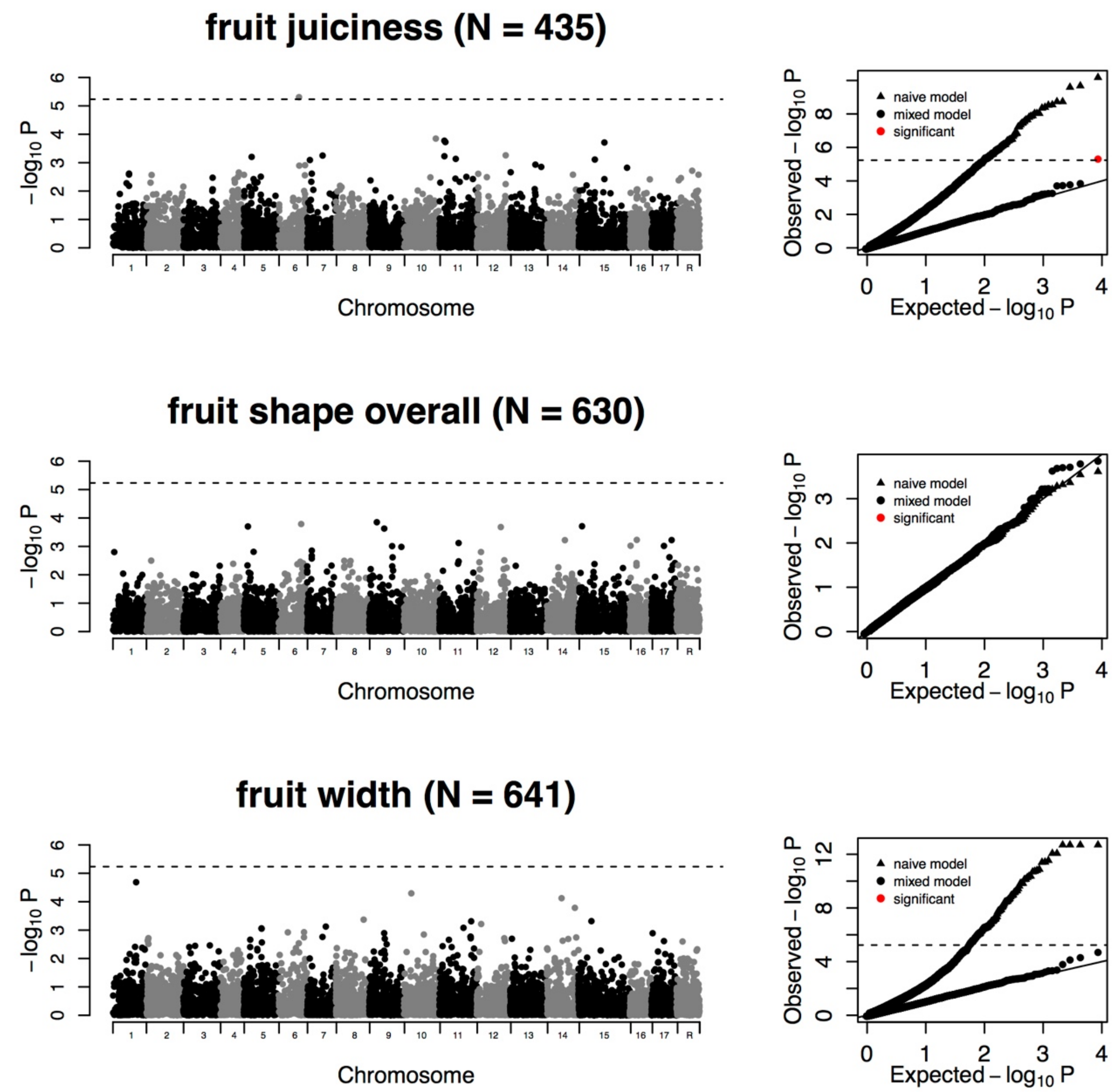


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

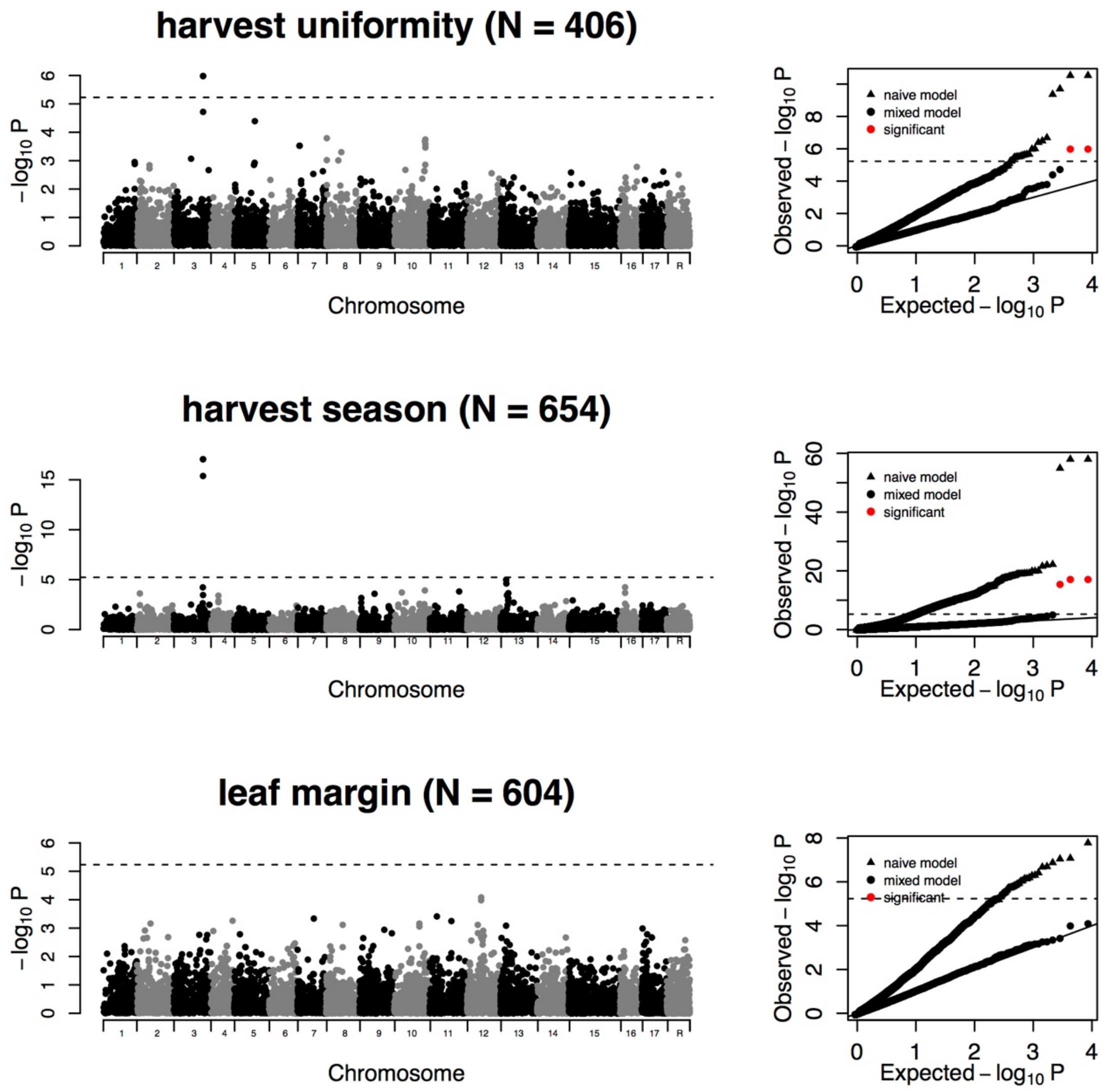


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

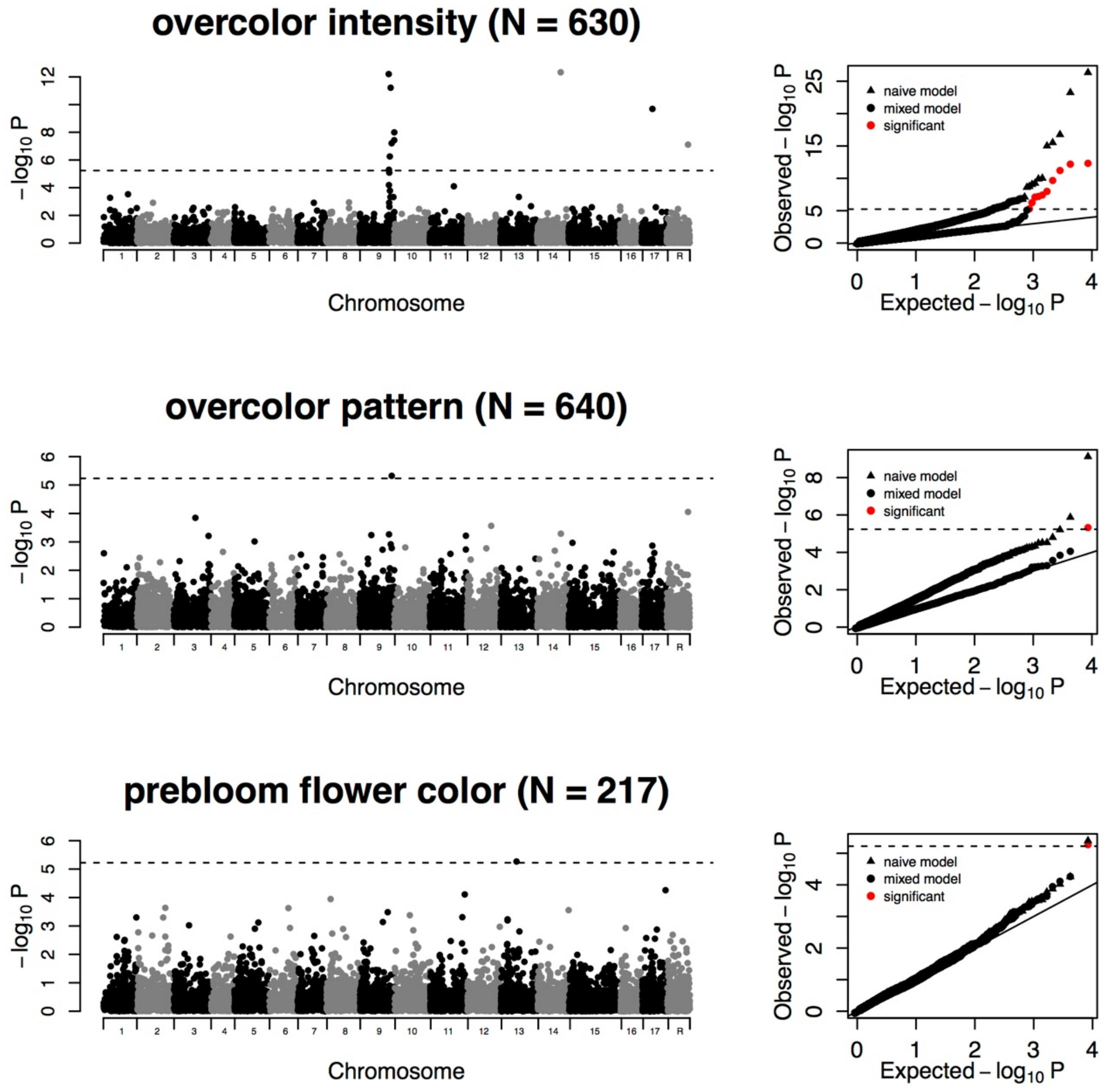


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

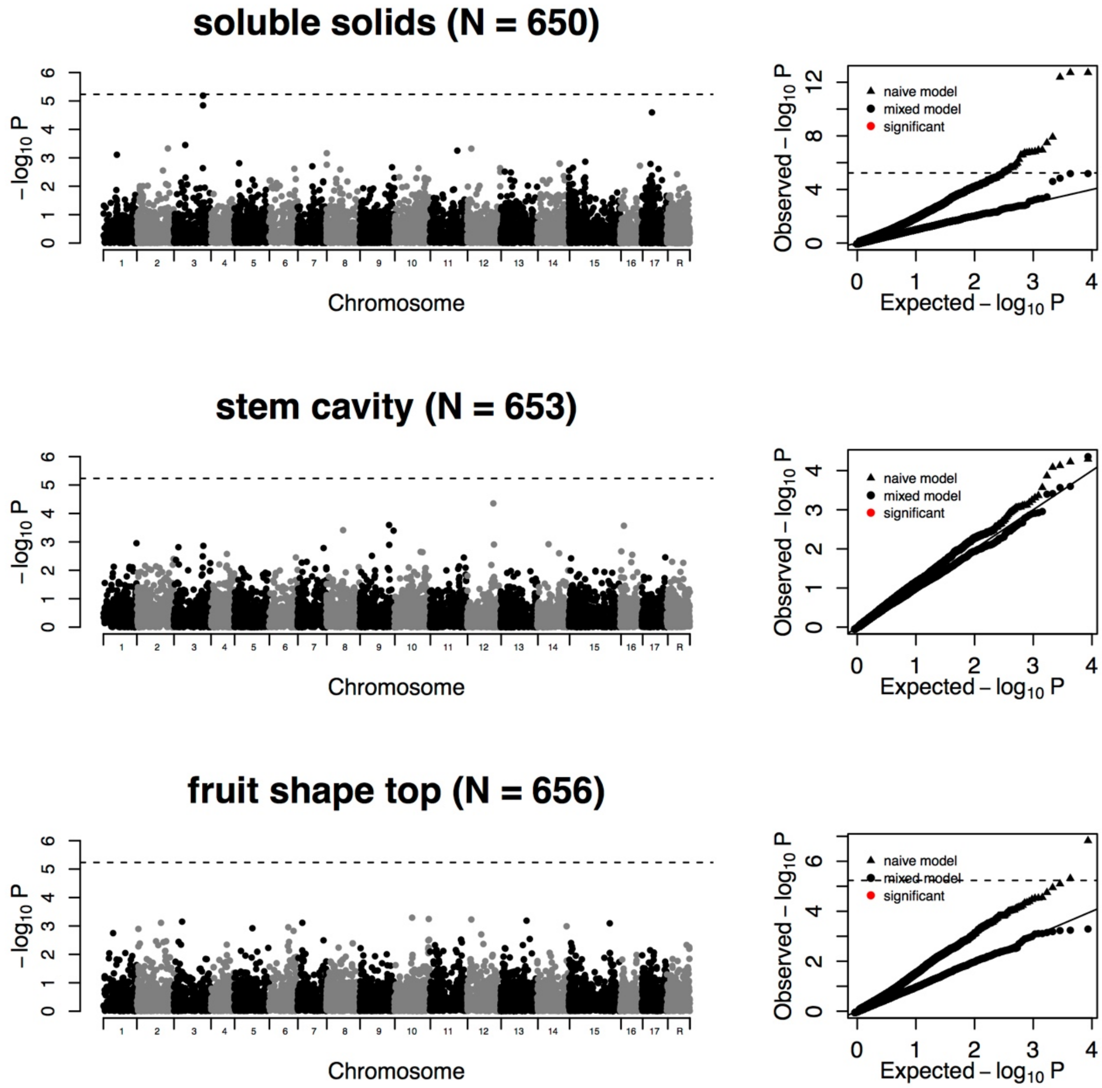


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

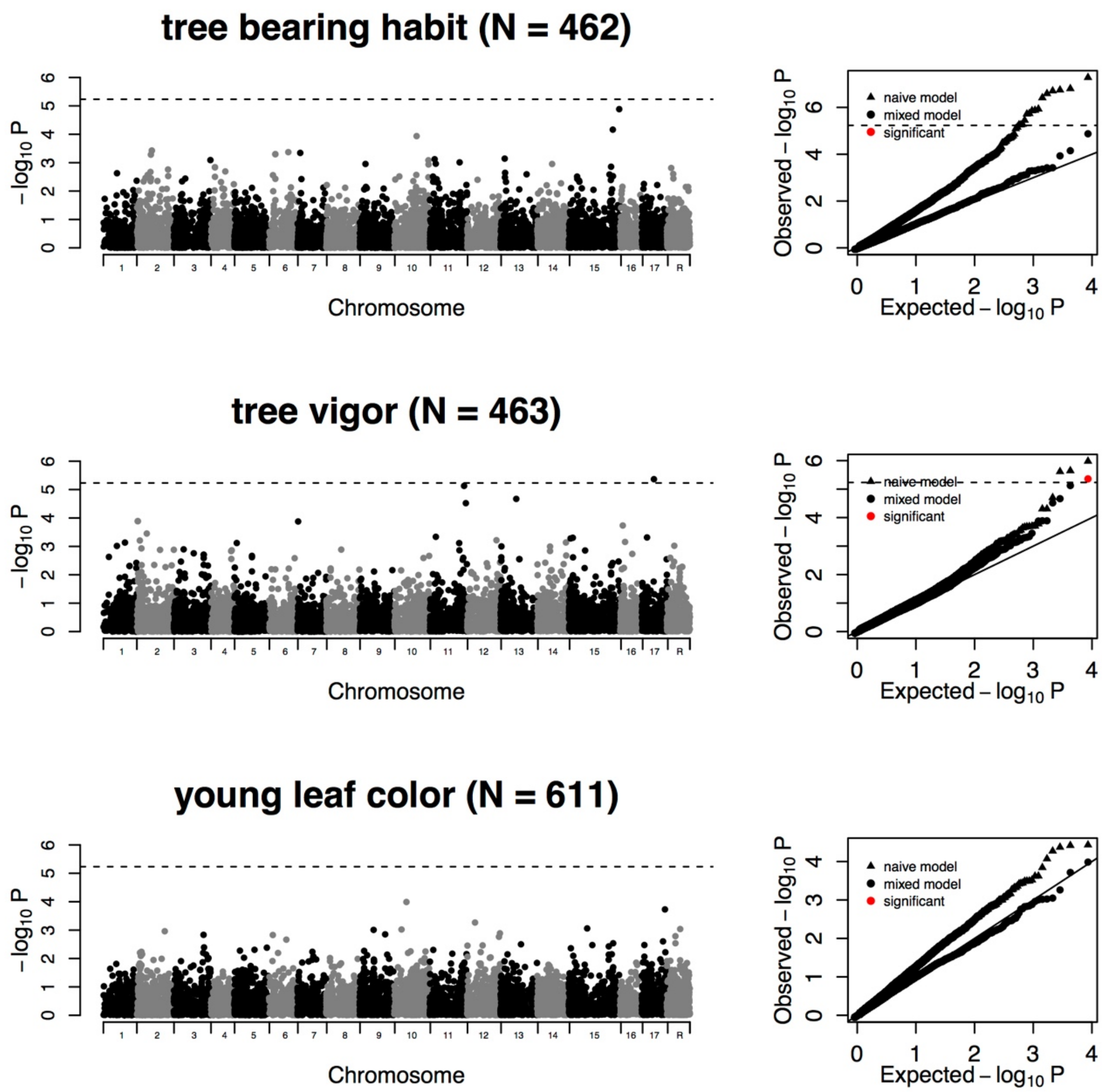


Figure II-I. GWAS results for apple phenotypes. Manhattan and QQ plots are included as well as the naive (Pearson correlation) and mixed model results. *P*-values are log-transformed and the threshold for significance is Bonferroni-corrected and indicated by a dotted line. Chromosome R indicates SNPs found on contigs unanchored to the reference genome.

Table II-I. Position information for significant GWAS hits as well as *p*-value and MAF. Chromosome R refers to concatenated unanchored contigs. The effect indicated is the direction of change in a particular phenotype score when the minor allele is present.

Phenotype	Chr	Position	*P*-value	MAF	Minor	Major	Effect
calyx basin	11	20670942	6.96E-07	0.03139	C	T	-
fruit flesh firmness	3	31409362	1.89E-06	0.1752	A	C	-
fruit flesh firmness	3	31409376	1.89E-06	0.1752	T	C	-
fruit flesh oxidation	16	1426905	9.19E-07	0.4339	A	G	-
fruit juiciness	6	21792902	5.00E-06	0.05287	T	C	-
fruit overcolor	9	31448296	2.45E-14	0.4965	A	T	+
fruit overcolor	9	33551878	1.14E-07	0.2683	A	T	-
fruit overcolor	14	25461478	1.91E-06	0.4756	T	C	+
fruit overcolor	R	88549507	8.34E-11	0.4007	C	A	+
fruit russet intensity	3	37428298	4.79E-07	0.01525	C	A	+
fruit russet intensity	3	37428308	4.79E-07	0.01525	C	T	+
fruit russet intensity	3	37428318	4.79E-07	0.01525	A	G	+
fruit stem length	15	7672279	2.59E-07	0.04587	T	C	+
harvest season	3	31409362	8.72E-18	0.1804	A	C	-
harvest season	3	31409376	8.72E-18	0.1804	T	C	-
harvest season	3	31409480	4.15E-16	0.1774	C	T	-
harvest uniformity	3	31409362	1.05E-06	0.1823	A	C	+
harvest uniformity	3	31409376	1.05E-06	0.1823	T	C	+
overcolor intensity	9	31448296	6.44E-13	0.4778	A	T	+
overcolor intensity	9	31690894	5.23E-06	0.2944	G	C	-
overcolor intensity	9	32681423	5.72E-07	0.2183	G	A	-
overcolor intensity	9	33551878	6.23E-12	0.2817	A	T	-
overcolor intensity	9	34628208	6.40E-08	0.304	T	A	+
overcolor intensity	9	37371284	3.84E-08	0.4302	A	G	+
overcolor intensity	9	37460989	1.04E-08	0.4079	T	C	+
overcolor intensity	14	25461478	4.86E-13	0.4937	T	C	+
overcolor intensity	17	11584126	2.15E-10	0.3841	C	T	+
overcolor intensity	R	88549507	7.99E-08	0.3825	C	A	+
overcolor pattern	9	34628208	4.72E-06	0.3039	T	A	+
prebloom flower color	13	16933147	5.35E-06	0.01843	A	T	+
tree vigor	17	13039979	4.36E-06	0.02592	C	T	-

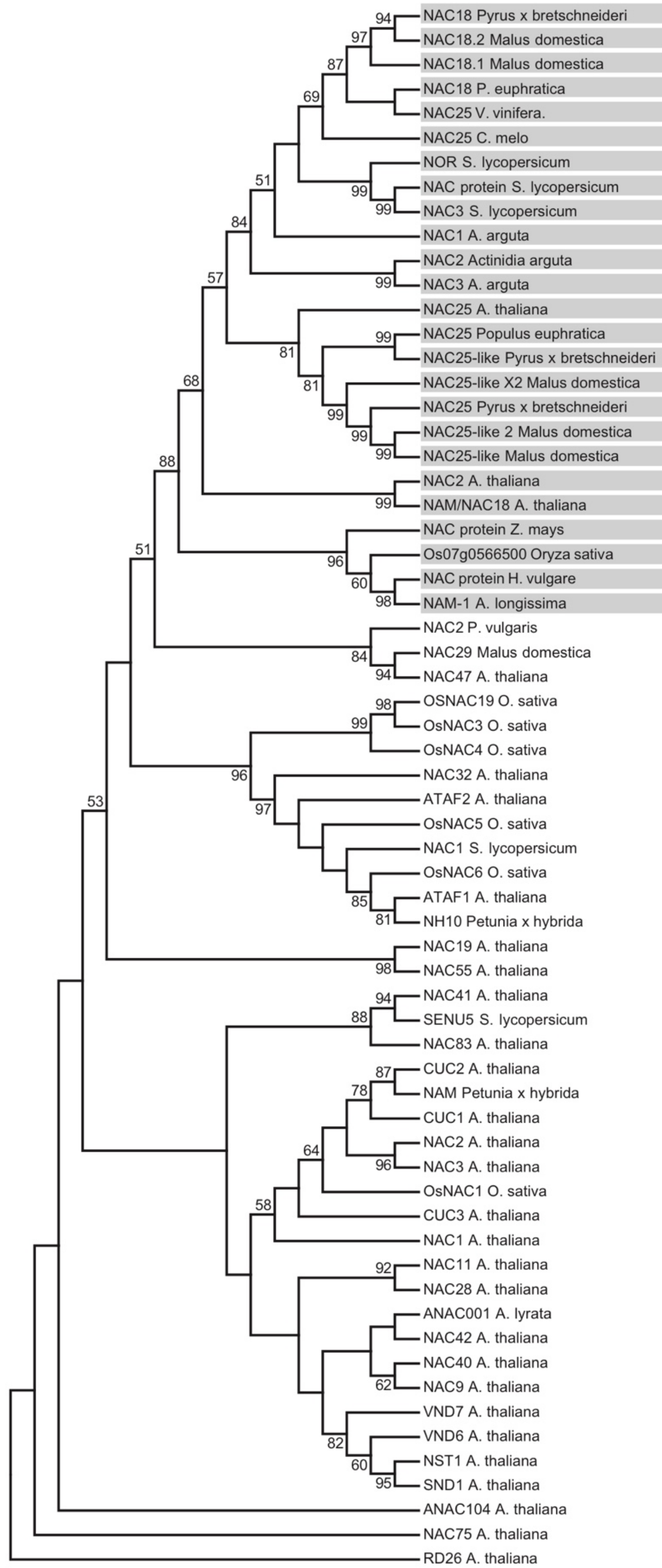

NAC18 Pyrus x bretschneideri
NAC18.2 Malus domestica
NAC18.1 Malus domestica
NAC18 P. euphratica
NAC25 V. vinifera.
NAC25 C. melo
NOR S. lycopersicum
NAC protein S. lycopersicum
NAC3 S. lycopersicum
NAC1 A. arguta
NAC2 Actinidia arguta
NAC3 A. arguta
NAC25 A. thaliana
NAC25 Populus euphratica
NAC25-like Pyrus x bretschneideri
NAC25-like X2 Malus domestica
NAC25 Pyrus x bretschneideri
NAC25-like 2 Malus domestica
NAC25-like Malus domestica
NAC2 A. thaliana
NAM/NAC18 A. thaliana
NAC protein Z. mays
Os07g0566500 Oryza sativa
NAC protein H. vulgare
NAM-1 A. longissima
NAC2 P. vulgaris
NAC29 Malus domestica
NAC47 A. thaliana
OSNAC19 O. sativa
OsNAC3 O. sativa
OsNAC4 O. sativa
NAC32 A. thaliana
ATAF2 A. thaliana
OsNAC5 O. sativa
NAC1 S. lycopersicum
OsNAC6 O. sativa
ATAF1 A. thaliana
NH10 Petunia x hybrida
NAC19 A. thaliana
NAC55 A. thaliana
NAC41 A. thaliana
SENU5 S. lycopersicum
NAC83 A. thaliana
CUC2 A. thaliana
NAM Petunia x hybrida
CUC1 A. thaliana
NAC2 A. thaliana
NAC3 A. thaliana
OsNAC1 O. sativa
CUC3 A. thaliana
NAC1 A. thaliana
NAC11 A. thaliana
NAC28 A. thaliana
ANAC001 A. lyrata
NAC42 A. thaliana
NAC40 A. thaliana
NAC9 A. thaliana
VND7 A. thaliana
VND6 A. thaliana
NST1 A. thaliana
SND1 A. thaliana
ANAC104 A. thaliana
NAC75 A. thaliana
RD26 A. thaliana

Figure II-II. Phylogenetic analysis of NAC protein family members. NAC proteins possessing a TDSS motif are highlighted. Protein sequences were aligned using ClustalW. A phylogenetic tree was built using MEGA6 with the Dayhoff model and neighbor joining method. The pairwise deletion option was used for dealing with gaps and branches were based on a consensus of 1000 bootstrap replicates. Bootstrap percentage values above 50 are shown at branch nodes. Accession numbers are as follows: *Malus domestica* NAC18.1 (NP_001280984.1), *M. domestica* NAC18.2 (XP_008386130.1), *M. domestica* NAC29 (NP_001280963.1), *Pyrus x bretschneideri* NAC18 (XP_009334622.1), *Pyrus x bretschneideri* NAC25 (XP_009378497.1), *Pyrus x bretschneideri* NAC25-like (XP_009379434.1), *Populus euphratica* NAC18 (XP_011029435.1), *P. euphratica* NAC18 (XP_011027905.1), *Cucumis melo* NAC25 (XP_008452274.1), *Arabidopsis thaliana* NAC18 (AT1G52880), *A. thaliana* NAC2 (AT3G15510), *A. thaliana* NAC25 (AT1G61110), *Vitis vinifera* NAC25 (CBI20351.3), *Actinidia arguta* NAC1 (AID55348.1), *A. arguta* NAC2 (AID55349.1), *A. arguta* NAC3 (AID55350.1), *M. domestica* NAC25-like X2 (XP_008383789.1), *Solanum lycopersicum* NOR (SGN-U317381), *S. lycopersicum* NAC3 (SGN-U568609), *S. lycopersicum* NAC protein (NP_001266277.2), *S. lycopersicum* SENU5 (CAA99760), *S. lycopersicum* NAC1 (AAR88435), *Zea mays* NAC protein (AKO90072.1), *Hordeum vulgare* NAC protein (BAK04712.1), *Oryza sativa* Os07 g0566500 (NP_001060017.1), *Arabidopsis lyrata* ANAC001 (XP_002892089), *A. thaliana* NAC47 (AT3G04070), *A. thaliana* NAC19 (AT1G52890), *A. thaliana* NAC55 (AT3G15500), *A. thaliana* ATAF1(X74755), *A. thaliana* ATAF2 (AK118910), *A. thaliana* NAC32 (AT1G77450), *A. thaliana* NAC41 (AT2G33480), *A. thaliana* NAC83 (AT5G13180), *A. thaliana* NAC104 (At5 g64530), *A. thaliana* NAC75 (AT4G29230), *A. thaliana* NST1 (AT2G46770), *A. thaliana* SND1 (AT1G32770), *A. thaliana* VND6 (AT5G62380), *A. thaliana* VND7 (AT1G71930), *A. thaliana* NAC40 (AT2G27300), *A. thaliana* NAC9 (AT4G35580), *A. thaliana* NAC28 (AT1G65910), *A. thaliana* CUC3 (AAP82630), *A. thaliana* CUC1 (BAB20598), *A. thaliana* CUC2 (BAA19529), *Petunia* x hybrida NAM Protein (CAA63101.1), *Petunia* x hybrida NH10 (AF509873), OsNAC1 (AB028180), OsNAC3 (BAA89797), OsNAC4 (AB028183), OsNAC5 (AB028184), OsNAC6 (BAA89800), OsNAC19 (AY596808), *Aegilops longissima* (AFD54040.1) *Phaseolus vulgaris* NAC2 (XP_007158644), *A. thaliana* NAC2 (AAO41710), *A. thaliana* NAC3 (AT3G29035), *A. thaliana* NAC11 (AT1G32510), *A. thaliana* NAC1 (AAF21437), *A. thaliana* NAC42 (AT2G43000), *A. thaliana* RD26 (AT4G27410).

Appendix III: Genomic ancestry estimation quantifies use of wild species in grape breeding (Chapter 4)

Table III-I. A list of the 78 cultivars examined as well as species, location and institute.

Cultivar	Species	Location	Institute
Baco Noir	Hybrid	Nova Scotia, Canada	Jost Vineyards
Beta	Hybrid	Minnesota, USA	University of Minnesota
Bluebell	Hybrid	Minnesota, USA	University of Minnesota
Blue Jay	Hybrid	Minnesota, USA	University of Minnesota
Borner	Vitis riparia x Vitis cinerea	Siebeldingen, Germany	Geilweilerhof Institute for Grape Breeding
Cabernet Foch	Hybrid	Nova Scotia, Canada	Jost Vineyards
Calandro	Hybrid	Siebeldingen, Germany	Geilweilerhof Institute for Grape Breeding
Canadice	Hybrid	Nova Scotia, Canada	Wührer Vineyards
Castel	Hybrid	Nova Scotia, Canada	Warner Vineyards
Cayuga	Hybrid	Nova Scotia, Canada	Wührer Vineyards
Chambourcin	Hybrid	Missouri, USA	Missouri State University Grape Foundation Vineyard
Chardonelle	Hybrid	Missouri, USA	Missouri State University Grape Foundation Vineyard
Corot Noir	Hybrid	Missouri, USA	Missouri State University Grape Foundation Vineyard
Van Buren (DVIT 1129)	Hybrid	California, USA	USDA
DVIT 1588	Vitis x champinii	California, USA	USDA
DVIT 1613	Vitis cinerea (Engelm.) Engelm. ex Millardet	California, USA	USDA
DVIT 1641	Vitis palmata	California, USA	USDA
Bertille-seyve 5563 (DVIT 169)	Hybrid	California, USA	USDA
DVIT 1703	Vitis aestivalis	California, USA	USDA
034-55 (DVIT 1807)	Vitis vinifera	California, USA	USDA
Marechal Foch (California) (DVIT 214)	Hybrid	California, USA	USDA
Loose Perlette (DVIT 2177)	Vitis vinifera	California, USA	USDA

Cultivar	Species	Location	Institute
DVIT 2180	Hybrid	California, USA	USDA
DVIT 2217	Vitis cinerea (Engelm.) Engelm. ex Millardet	California, USA	USDA
DVIT 2224	Vitis cinerea var. helleri (L. H. Bailey) M. O. Moore	California, USA	USDA
Rofar Vidor (DVIT 2258)	Hybrid	California, USA	USDA
Kecskemet (DVIT 2639)	Vitis vinifera	California, USA	USDA
L'Arvine (DVIT 2640)	Vitis vinifera	California, USA	USDA
Dan Ben Hanna (DVIT 2669)	Vitis vinifera	California, USA	USDA
Jackson Sel. #3 (DVIT 2916)	Hybrid	California, USA	USDA
Kandhar (DVIT 2918)	Vitis vinifera	California, USA	USDA
Peagudo (DVIT 887)	Vitis vinifera	California, USA	USDA
Einset	Hybrid	Nova Scotia, Canada	Wührer Vineyards
Felicia	Hybrid	Siebeldingen, Germany	Geilweilerhof Institute for Grape Breeding
Frontenac (Missouri)	Hybrid	Missouri, USA	Missouri State University Grape Foundation Vineyard
Frontenac (Gris)	Hybrid	Nova Scotia, Canada	Jost Vineyards
Frontenac (Nova Scotia: Jost)	Hybrid	Nova Scotia, Canada	Jost Vineyards
Frontenac (Nova Scotia: AFHRC)	Hybrid	Nova Scotia, Canada	Atlantic Food and Horticulture Research Centre
Himrod	Hybrid	Nova Scotia, Canada	Wührer Vineyards
Kay Gray	Hybrid	Minnesota, USA	University of Minnesota
King of the North	Hybrid	Minnesota, USA	University of Minnesota
Leon Millot	Hybrid	Nova Scotia, Canada	Atlantic Food and Horticulture Research Centre
Marechal Foch (Nova Scotia: AFHRC)	Hybrid	Nova Scotia, Canada	Atlantic Food and Horticulture Research Centre
Marechal Foch (Nova Scotia: Jost)	Hybrid	Nova Scotia, Canada	Jost Vineyards
Marechal Joffre	Hybrid	Nova Scotia, Canada	Jost Vineyards

Cultivar	Species	Location	Institute
Marquette (Nova Scotia)	Hybrid	Nova Scotia, Canada	Wührer Vineyards
Marquette (Missouri)	Hybrid	Missouri, USA	Missouri State University Grape Foundation Vineyard
New York Muscat	Hybrid	Nova Scotia, Canada	Wührer Vineyards
Norton	Hybrid	Missouri, USA	Missouri State University Grape Foundation Vineyard
Orion	Hybrid	Siebeldingen, Germany	Geilweilerhof Institute for Grape Breeding
Petit Milo	Hybrid	Nova Scotia, Canada	Jost Vineyards
Petit Muscat	Hybrid	Nova Scotia, Canada	Wührer Vineyards
Petite Jewel	Hybrid	Minnesota, USA	University of Minnesota
Petite Pearl	Hybrid	Missouri, USA	Missouri State University Grape Foundation Vineyard
Phoenix	Hybrid	Siebeldingen, Germany	Geilweilerhof Institute for Grape Breeding
Reberger	Hybrid	Siebeldingen, Germany	Geilweilerhof Institute for Grape Breeding
Regent	Hybrid	Siebeldingen, Germany	Geilweilerhof Institute for Grape Breeding
Reliance	Hybrid	Nova Scotia, Canada	Wührer Vineyards
Sabrevois	Hybrid	Minnesota, USA	University of Minnesota
Seyval Blanc (Germany)	Hybrid	Siebeldingen, Germany	Geilweilerhof Institute for Grape Breeding
Seyval Blanc (Nova Scotia)	Hybrid	Nova Scotia, Canada	Warner Vineyards
Suelter	Hybrid	Minnesota, USA	University of Minnesota
Sovereign Coronation	Hybrid	Nova Scotia, Canada	Wührer Vineyards
St. Croix (Minnesota)	Hybrid	Minnesota, USA	University of Minnesota
St. Croix (Missouri)	Hybrid	Missouri, USA	Missouri State University Grape Foundation Vineyard
St. Pepin	Hybrid	Minnesota, USA	University of Minnesota
Staufer	Hybrid	Siebeldingen, Germany	Geilweilerhof Institute for Grape Breeding
Suffolk Grape	Hybrid	Nova Scotia, Canada	Wührer Vineyards
Swenson Red	Hybrid	Nova Scotia, Canada	Atlantic Food and Horticulture Research Centre
Triomphe D'Alsace	Hybrid	Nova Scotia, Canada	Wührer Vineyards
Venus	Hybrid	Nova Scotia, Canada	Wührer Vineyards

Cultivar	Species	Location	Institute
Venus (Red)	Hybrid	Nova Scotia, Canada	Atlantic Food and Horticulture Research Centre
Vidal Blanc (Nova Scotia)	Hybrid	Nova Scotia, Canada	Warner Vineyards
Vidal Blanc (Missouri)	Hybrid	Missouri, USA	Missouri State University Grape Foundation Vineyard
Vignoles	Hybrid	Missouri, USA	Missouri State University Grape Foundation Vineyard
Villard Blanc	Hybrid	Siebeldingen, Germany	Geilweilerhof Institute for Grape Breeding
Villaris	Hybrid	Siebeldingen, Germany	Geilweilerhof Institute for Grape Breeding
Walter's Seedling	Hybrid	Nova Scotia, Canada	Wührer Vineyards

Appendix IV: Co-authorship and copyright release

Chapter 2: Quantifying the genetic basis of leaf shape in apple

Although this work was led by me, it could not have been completed without the help of my co-authors. All co-authors contributed analysis or feedback on the final manuscript. This chapter has not yet been submitted for publication but a pre-print is available:

Migicovsky, Z., Li, M., Chitwood, D.H., and Myles, S. (2017). "Morphometrics reveals complex and heritable apple leaf shapes". bioRxiv. doi: https://doi.org/10.1101/139303

Chapter 3: Genome to phenome mapping in apple using historical data

Although this work was led by me, it could not have been completed without the help of my co-authors. All co-authors provided data, analysis, or feedback on the final manuscript, which was originally published as:

Migicovsky, Z., Gardner, K.M., Money, D., Sawler, J., Bloom, J.S., Moffett, P., Chao, C.T., Schwaninger, H., Fazio, G., Zhong, G.-Y., and Myles, S. (2016). "Genome to phenome mapping in apple using historical data". The Plant Genome, 9:2. doi: 10.3835/plantgenome2015.11.0113

Permission for reproduction for use in this dissertation was granted on February 21 2017th. Copyright permission from the content published Crop Science Society of America was granted with the license number 4115691356357. The version of the manuscript which appears in this dissertation has been modified from the published version.

Chapter 4: Genomic ancestry estimation quantifies use of wild species in grape breeding

Although this work was led by me, it could not have been completed without the help of my co-authors. All co-authors provided data, analysis, or feedback on the final manuscript, which was originally published as:

Migicovsky, Z., Sawler, J., Money, D., Eibach, R., Miller, A.J., Luby, J.J., Jamieson, A.R., Velasco, D., von Kintzel, S., Warner, J., Wührer, W., Brown, P.J., and Myles, S. (2016). "Genomic ancestry estimation quantifies use of wild species in grape breeding". BMC Genomics, 17:478. doi: 10.1186/s12864-016-2834-8

The open access articles published in BioMed Central's journals are made available under the Creative Commons Attribution (CC-BY) license, which means they are accessible online without any restrictions and can be re-used in any way, subject only to proper attribution (which, in an academic context, usually means citation).

The re-use rights enshrined in our license agreement (http://www.biomedcentral.com/about/policies/license-agreement) include the right for anyone to produce printed copies themselves, without formal permission or payment of permission fees.

The version of the manuscript which appears in this dissertation has been modified from the published version.

Chapter 5: Exploiting wild relatives for genomics-assisted breeding of perennial crops

Although this work was led by me, it could not have been completed without the help of my co-author. The final manuscript was originally published as:

Migicovsky, Z., and Myles, S. (2017). "Exploiting Wild Relatives for Genomics-assisted Breeding of Perennial Crops." Front. Plant Sci. 8:460. doi: 10.3389/fpls.2017.00460

The version of the manuscript which appears in this dissertation has been modified from the published version.