EXPLORING THE GENETICS OF AROMA PRODUCTION IN APPLES

By

Tayab Soomro

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
July 2022

To my parents to whom I am forever indebted for their incredible sacrifices and support for making me who I am today.

Table of Contents

List of Tables

List of Figures

Abstract

Apple aroma is a major determinant of consumer acceptability. In order to better understand the aroma of apples, gas-chromatography mass-spectrometry (GCMS) was used to quantify apple volatiles across 515 apple varieties from Canada's Apple Biodiversity Collection (ABC). Among the volatiles identified, esters and aldehydes were the most abundant classes of compounds with butyl acetate and hexyl acetate present in nearly every variety. Using principal component analysis (PCA), It was determined that the primary axis of variation of the apple volatilome is correlated with harvest date: early-harvested apples tend to express larger numbers and higher amounts of volatiles than late-harvested apples. Through genome-wide association studies (GWAS) with 250,579 single nucleotide polymorphisms (SNPs) I identified a significant association between SNPs at a *NAC18.1* transcription factor and the abundances of two key volatile compounds: 1-hexanol and 1-butanol. Taken together, these results provide a foundation for understanding the genetic basis of apple aroma production.

List of Abbreviations Used

ABC	apple biodiversity collection
GC-MS	gas chromatography mass spectrometry
GWA	genome-wide association
GWAS	genome-wide association study
HRM	high-resolution melting
M. domestica	*Malus domestica*
PC	principal component
PCA	principal component analysis
QQ	quantile-quantile
SNP	single nucleotide polymorphism
VOC	volatile organic compound

Acknowledgements

First and foremost, I am extremely thankful to Dr. Sean Myles. His unconditional support and guidance in academia and my personal life have been instrumental in my ability to complete this work and to be able to present my thesis. It is rare to find a mentor who goes above and beyond to create opportunities for success for their students and I am extremely grateful to have such a mentor. I would also like to extend my sincere thanks to the Myles Lab (Dr. Zoë Migicovisky, Sophie Watts & Thomas Davies) for contributing to the creative and positive lab environment every day.

I would like to thank my parents Wali Soomro and Sabira Soomro. Their incredible support, patience and struggles have shaped me into who I am today. Words fail to express my gratitude for my father Wali Soomro whose resilience and perseverance in the face of adversity have been exemplary lessons and motivation for me throughout my life. I would be remiss in not mentioning my siblings: Usra Soomro and Zaid Soomro. They have provided unending support, love and fun throughout my life and for that, I am extremely thankful.

I would like to thank the funding agencies who supported this work as well as my committee members for their support, time commitment and valuable lessons.

Sir Issac Newton once said, *"If I have seen further, it is by standing on the shoulders of the Giants"*. Therefore, I must thank everyone who has contributed to my learning throughout my life.

CHAPTER 1 — INTRODUCTION

The domesticated apple (*Malus domestica*) represented 43.5% of total Canadian fruit production in 2020, with 390,995 tonnes of apples produced (Canada 2021), making it one of the most consumed fruit crops in Canada. The fruit quality, especially the flavour of the fruit, is an important trait that determines consumer acceptability and by extension, its marketability. However, due to intense selection for traits that enable mass production and worldwide distribution such as storability, firmness and post-harvest shelf-life, little attention has been given to the flavour and aroma of apples, which directly impacts consumer satisfaction (Klee & Tieman, 2018). The lagging of this selection for fruit flavour is attributed to the high costs of breeding and phenotyping (Klee & Tieman, 2018).

Traditional apple breeding is time-consuming and expensive: it takes more than two decades to properly evaluate a new apple variety before commercial release due to its long juvenile period (Igarashi et al., 2016). For example, in a breeding programme spanning 26 years, 52,000 seedlings were originally planted, and only 3 of these were subsequently commercialized (Kole, 2011). Therefore, apple varieties that perform well are clonally propagated, for decades or even centuries, and there is a slow rate of change in variety composition on supermarket shelves despite consumer demand for new apple varieties with novel traits (Klee & Tieman, 2018).

The small number of elite varieties that dominate worldwide markets means that only a fraction of the available genetic diversity in apples is being explored (Migicovsky, Gardner, et al., 2021), and this makes them prone to evolving pests and pathogens (Cornille et al., 2014). To mitigate this, a wide array of agro-chemicals is used to keep apples commercially viable. Indeed, a primary target for most apple breeding programmes worldwide is disease resistance. However, for the long-term sustainability of Canada's apple industry, new apple varieties must not only require less chemical input to grow, but should also contain desirable flavours that result in commercial success (Brown & Maloney, 2003). Marker-assisted selection (MAS) can significantly improve the efficiency of traditional apple breeding by enabling breeders to select offspring using genetic markers associated with desirable traits, and markers that predict disease resistance are commonly used by numerous breeding programmes worldwide (Bus et al., 2011).

Selection for flavour using MAS lags far behind the selection for disease resistance, however. The reason for this is that flavour is more genetically complex than disease resistance: while resistance to a particular disease is frequently controlled by a single gene of large effect, flavour is controlled by numerous interacting genes and environmental variations (Myles, 2013). Flavour is composed of a large collection of volatile organic compounds (VOCs) that are recognized by our olfactory systems. Humans have 350 olfactory receptors which provide the foundation for diversity of flavour experiences (Klee, 2010). Genetic markers that predict apple flavour would therefore be highly desirable, and an opportunity exists to find them and use them to reduce the labour and costs associated with growing trees to maturity for flavour evaluation (Myles, 2013).

Human perception of the flavour of fruit is determined by a complex interaction between taste receptors on the tongue and olfactory receptors located in the nose (Tieman et al., 2006). While taste is largely determined by sugars and acids, apple aroma is a complex trait determined by various VOCs that vary between varieties (Espino-Díaz et al., 2016). A recent study found that 56% of the variance associated with overall consumer liking of blueberry and tomato can be attributed to VOCs (Colantonio et al., 2022). VOCs originate from major pathways of secondary metabolism in plants and can therefore be characterized as secondary metabolites (Abdullah et al., 2015). VOCs can be further classified based on their chemical structure, which includes categories such as alcohols, aldehydes and esters. Among these, esters are the largest group of VOCs that are found in apples and they contribute to the fresh and fruity flavour of apples (Sugimoto et al., 2021). The most comprehensive evaluation of the genetic control of apple VOCs to date discovered a few genetic markers that may be useful for predicting apple aroma, but its scope was limited: it involved the quantification of only 33 VOCs across 162 apple varieties and the use of fewer than 10,000 genetic markers (Farneti et al., 2017). Recent advances in high-throughput analytical chemistry have enabled the untargeted assessment of the apple's entire "volatilome" (Mansurova et al., 2018), which, when paired with next-generation DNA sequencing of hundreds of apple varieties, provides a powerful platform to elucidate the genetic architecture of apple aroma. The pairing of volatilome quantification and next-generation genomics technologies has uncovered the genetic underpinnings of commercially important flavour molecules in a diversity of vegetable and fruit crops such as soybean (Ravi et al., 2019), pear (Qin et al., 2012) and melon (Shi et al., 2020).

The aim of this study is to harness multi-dimensional apple VOC and genomic data to elucidate the mechanism of production of VOCs, as well as to identify genetic markers associated with key apple VOCs.

CHAPTER 2 — LITERATURE REVIEW

2.1 — DOMESTICATION HISTORY AND IMPROVEMENT OF APPLES

The cultivated apple (*Malus domestica*) belongs to the Rosaceae family and was first domesticated from its wild progenitor, *Malus sieversii*, in the Tian Shan Mountains in Central Asia. These early cultivated apples travelled with humans along the Silk Route to the west, and they interbred with other species such as *Malus baccata*, *Malus orientalis*, and *Malus sylvestris* (Cornille et al., 2012), all of which together contributed to the genetic pool of the domesticated apples we enjoy today (Migicovsky, Gardner, et al., 2021; Sun et al., 2020).

The genome of the apple was first sequenced in 2010 using the diploid (2x=34) 'Golden Delicious' variety to produce a draft assembly, which revealed profound insights into the biology of this fruit. One such insight was that the apple went through a gene duplication event with the final number of chromosomes reaching 17 from 9 ancestral chromosomes (Velasco et al., 2010). Another genome assembly was performed in 2017 of the double-haploid Golden Delicious variety by combining the high-quality short-read sequencing and long-read sequencing data. The size of this new genome is 649.7-Mb containing roughly 57,000 genes and now serves as the reference genome (Peace et al., 2019; Zhang et al., 2019).

Ever since the start of apple breeding through the efforts of Thomas Andrew King (1759-1839) with his experiments in apple crossbreeding and selection (Janick, 2012), vegetative propagation has been the default method of mass production of apples. Vegetative production ensures that the fruits produced are uniform in terms of the desired traits such as taste, firmness and shelf-life. However, it leads to a lack of genetic diversity and opens up a whole set of new problems. Pathogens evolve over time via mutation, selection and genetic drift, whereas vegetatively propagated crops do not evolve, leaving them unshielded from evolving pathogens and prone to novel diseases (Myles, 2013). For example, the 'McIntosh' apple has been vegetatively propagated for over 200 years and is thus prone to novel diseases (Migicovsky, Gardner, et al., 2021). To overcome pathogen pressures, heavy pest management practices are undertaken to sustain the apple industry (Myles, 2013).

Additionally, the repeated use of a small number of elite varieties suited for yield, firmness and improved postharvest shelf-life leads to an indirect reduction in flavour and nutrient content (Klee & Tieman, 2013). This is due to the fact that growers are not paid for flavour

quality, and they do not demand it. The customer of the breeder is the grower, and consumers are frequently left out of cultivar development, even though there are major opportunities for market growth with the improvement in fruit quality such as flavour and nutrients (Klee & Tieman, 2018).

One of the most important issues associated with improving the flavour of perennial crops is that the flavour of the fruit can be assessed only after the plant has grown to maturity and is bearing fruit, which takes 5-8 years in apples (Klee, 2010). Marker-assisted selection (MAS), whereby traits can be predicted at the seedling stage, provides the solution for the early selection of promising varieties (Klee & Tieman, 2018).

2.2 — MARKER ASSISTED SELECTION (MAS)

MAS is the process by which offspring are selected using genetic markers that are linked to a phenotype of interest. This is referred to as indirect selection because instead of selecting for the traits directly, the markers that are linked to those traits are used for selection (Ribaut & Hoisington, 1998). This technique is particularly useful for perennial crops such as apples where the screening of the traits directly is labour intensive and time-consuming due to the plant's long juvenile period. On top of that, the significant reduction in costs associated with generating enormous genotyping datasets make MAS an attractive choice (Schatz & Langmead, 2013). However, one of the considerations with MAS is that most phenotypes that are of interest to breeders are complex and vary quantitatively, meaning that multiple genetic loci and environmental factors have effects on the observable phenotype. Therefore, in order to find meaningful associations, genotyping a large set of single nucleotide polymorphisms (SNPs) throughout the genome is necessary for increasing the chances of finding markers associated with the causal SNPs (Ribaut & Hoisington, 1998).

MAS has been widely used in plant breeding and can provide significant reductions in the cost and time associated with traditional selection. For example, genetic screening of varieties at the seedling stage in the Washington Apple Breeding program was estimated to reduce the conventional operating costs by 60% (Edge-Garza & Peace, 2010). In another example, marker-assisted selection has been successfully applied in pest and disease management strategies in the New Zealand apple breeding programme (Bus et al., 2000).

Even though there is enormous potential for improvement of apple quality via MAS, it has thus far almost exclusively been used to select for resistance to diseases such as apple scab and powdery mildew (Migicovsky et al., 2016). MAS could be applied to improve flavour, but a major roadblock to performing MAS for flavour is the lack of markers associated with flavour-related phenotypes. Therefore, genetic mapping studies first need to be performed to identify genetic markers (i.e., SNPs) associated with desirable flavours.

2.3 — GENOME-WIDE ASSOCIATION MAPPING

Genome-wide association (GWA) mapping is a technique used for the discovery of associations between genetic variants (i.e., SNPs) and phenotypes using populations of unrelated individuals (Gupta et al., 2019). It differs from the most common genetic mapping technique applied in apples to date called quantitative trait loci (QTL) mapping where the offspring of bi-parental crosses are genotyped and phenotyped. Both QTL and GWA are used for identifying genetic variants that explain phenotypic variation (Gupta et al., 2019). One of the advantages of GWA is that it most often results in higher mapping resolution because it makes use of large numbers of unrelated individuals resulting in the rapid decay of linkage disequilibrium (Rafalski, 2010). However, most genetic mapping studies in apples have made use of simple bi-parental crosses, primarily because these are widely available in breeding programmes: breeders make bi-parental crosses as part of the breeding process and use these populations to search for markers that may be useful for MAS. GWA requires the use of large and diverse germplasm collections, such as the USDA's apple germplasm collection, which has served as a GWA population in the past (Migicovsky et al., 2016). However, GWA is optimally performed in populations planted specifically for the purposes of discovering genotype-phenotype relationships. In apple, such a population was established in Kentville in 2011 and is called Canada's Apple Biodiversity Collection (ABC).

2.4 — GENETIC BASIS OF APPLE VOCs

Apple aroma is determined by the concentrations of VOCs, which are primarily composed of alcohols, ketones, aldehydes and esters. Among these various compounds, esters are the most abundant: they account for 80 to 95% of the total volatiles emitted in apples and are generally perceived as "fruity" and "floral" (Sugimoto et al., 2021). These VOCs bind to the receptors in the human olfactory system which gives the perception of aroma. Each apple variety

emits different compounds at varying concentrations and this variability in the amount and type of VOCs that are emitted give us different flavours (Klee, 2010). Studies of apple aroma aim to quantify VOCs in order to better understand the chemical composition of flavour (Espino-Díaz et al., 2016). Genetic mapping studies then help identify the candidate genes that might be involved in aroma production. For example, in a recent study in blueberries, it was shown that key VOCs controlling important aspects of blueberry flavour are controlled by a small number of genes with large effects on VOC concentrations. This suggests that markers within these genes can be used to predict VOCs in blueberry, which forms the basis of MAS for VOCs in a commercially important fruit crop (Ferrão et al., 2020). In another study, a single region on chromosome 2 was identified as a crucial "hot-spot" for genes involved in ester production, which account for most of the VOCs in apples (Larsen et al., 2019). This suggests that a small number of genetic loci may be responsible for VOC production in apples. More functional studies aim to elucidate the mechanism of a single pathway involved in production of specific types of compounds. For example, a transgenic 'Royal Gala' variety in which the expression of *AAT1* gene was reduced showed reduced levels of most key esters, suggesting that the *AAT1* gene is a critical gene responsible for the biosynthesis of esters contributing to 'ripe apple' flavour in 'Royal Gala' and 'Granny Smith' apple varieties (Souleyre et al., 2014).

Studies investigating apple aroma have thus far focused on targeted compounds with an assumption that the VOCs important for the flavour of the fruits are already known (Farneti et al., 2017; Larsen et al., 2019). This approach is therefore limited in terms of gaining new insights about fruit aroma, and it ignores the potential of rare or low concentration VOCs that may contribute to consumer preference. In contrast, an untargeted metabolic approach enables the quantification of the entire "volatilome" and thus does not suffer from the ascertainment bias of targeted approaches. Therefore, it is desirable to use untargeted approaches for exploring the VOCs in apples.

CHAPTER 3 — MATERIALS AND METHODS

3.1 — SAMPLE PREPARATION

Apple varieties in this study were from Canada's ABC located at the Kentville Research and Development Centre in Kentville, Nova Scotia, Canada. A comprehensive description of the statistical design of the ABC and phenotyping protocols are provided in Watts et al. (2021). Briefly, the harvest date for each variety was determined according to several established ripeness metrics, including seed colour, taste, skin colour, firmness and starch index. For each variety, 10-20 apples were collected from either one or two trees. For volatile quantification, the harvested fruits were stored at 3-3.5˚C for 1 month. After storage, a sample of 5-10 fruits was randomly selected, cored and cut into 8 slices using an 8-piece apple slicer and corer. One or 2 random slices from each apple were selected and frozen in liquid nitrogen. The slices were bagged, labelled and held at -80˚C until analysis. The total mass for each sample ranged from ~300-500g.

3.2 — VOLATILE QUANTIFICATION

A 5 g composite frozen sample (-80°C) was blended with 95 g of a saturated salt solution (NaCl, Fisher Scientific Canada, certified ACS) for 1 min using a Kinematica model MB 800 laboratory mixer (Kinematica AG, Luzern, Switzerland) at setting 4. A 10 g sample of the homogenate was placed in a 20 mL headspace vial, capped and 5 µL of an internal standard (10.0 mg/L Benzaldehyde-d6) was added using a MultiPurpose Sampler (MPS, Gerstel, Linthicum, MD, USA). The VOCs were extracted and analyzed by solid-phase microextraction-gas chromatography x gas chromatography-time of flight–mass spectrometry. Vials were incubated at 30 ºC for 300 s and then the divinylbenzene/Carboxen/polydimethylsiloxane SPME fiber (Supelco Analytical, Bellefonte, PA, USA) was exposed to the headspace for 900s with agitation (on for 60 s; off for 1 s). The fiber was desorbed at 250 ºC for 7 min. The injector was operated at 250 °C in the split mode of 1:20 for 1 min. Helium was used as the carrier gas at a flow rate of 1.4 mL/min. The MPS system was installed on a unit-mass resolution Pegasus 4D TOFMS (LECO, St Joseph, MI, USA). The modulator was mounted in an Agilent 7890 GC gas chromatograph equipped with a secondary oven and a quad-jet dual-stage thermal modulator. Liquid nitrogen was used for cooling the cold jet lines. The first dimension (1D) column was a

polar Stabilwax® (30 m x 0.25 mm x 0.25 µm), and the second dimension (2D) column was a mid-polar Rxi®-5Sil MS (1.09 m x 0.25 mm x 0.25 µm). The optimized 1D GC oven temperature was initially set at 50 °C for 0.20 min, before increasing from 10.3 °C/min to 220 °C. The temperature offset for the secondary oven was 44°C and the modulator temperature offset was +15℃. The modulation period (PM) was 1.2 s, with a hot pulse time of 0.35 s on each jet. The transfer line was held at 250 °C. The TOF-MS was operated in electron ionization (EI) mode at 70 eV, with an acquisition mass range of 35–300 amu, area count calculation applied apex masses, an acquisition rate of 200 Hz, and a detector voltage of 1500V with an optimized voltage offset of 200V. The ion source was heated to 250 °C. Daily mass calibration and tuning were performed using perfluorotributylamine (PFTBA). An acquisition delay of 100s was applied. The chemical identification of the peaks was determined based on the retention index and correspondence of the mass spectra with the 'mainlib' and 'replib' of the 2017 National Institute of Standards and Technology (NIST) Mass Spectral Virtual Library (ChemSW, Fairfield, CA, USA). The VOCs that had NIST similarity scores below 850 were discarded. The retention index for compounds was identified from the retention time using Kovats Retention Index formula (Kováts, 1958).

3.3 — VOLATILE DATA STANDARDIZATION AND CURATION

The concentration of the internal standard, benzaldehyde-d8 was normally distributed across varieties (Shapiro-Wilks test p=2.30x10^{-8}; Figure S1). The peak area value for each VOC within each variety was divided by the peak area value for benzaldehyde-d8 as a standardization procedure and the resulting units for the abundance values were normalized total ion counts (TIC). The compounds were manually categorized into 13 different classes including alcohols, aldehydes, esters, etc. The totals for each compound category were calculated by adding up the standardized peak area values for all compounds within each category. The final table contained 106 compounds across 515 varieties (Table S1). Every compound was present in at least 35 (6.8%) apple varieties and every variety had at least 24 (22.6%) VOCs present.

3.4 — SNP GENOTYPING

The genotyping-by-sequencing (GBS) method (Elshire et al., 2011) was used as previously described in Migicovsky et al. (2022). The initial genotype data consisted of 260,399 SNPs across 1,054 varieties. Two additional markers were genotyped in the same varieties because of their potential role in volatile synthesis and were then combined with the GBS data. First, a Kompetitive Allele Specific PCR (KASP) genotyping assay was used to genotype a functional non-synonymous SNP that results in glutamine to glutamate change at position 387 of the citramalate synthase (CMS) gene that may account for variability in ester synthesis across apple varieties (Sugimoto et al., 2021). Second, high-resolution DNA melting (HRM)-based assay was used to detect the presence of a long terminal repeat (LTR) retrotransposon upstream of the *MYB1* transcription factor that is associated with red skin (Zhang et al., 2019). Primer sequences and reaction details can be found in Appendix C. Out of the 550 varieties for which VOC data were collected, genotype data were available for only 515 varieties, and thus the final SNP genotype matrix was filtered for only those 515 varieties. SNPs with minor allele frequency (MAF) <1% and heterozygosity > 90% were removed. The final genotype matrix contained 250,579 SNPs across 515 varieties.

3.5 — STATISTICAL ANALYSES

All statistical analyses were performed in R version 4.0.2 (R Core Team, 2021). Principal component analysis (PCA) was performed using the *prcomp* function with scale and center parameters from the *stats* package. GWAS was performed using the *mlmm_cof* function from the *mlmm* package (v0.1.1) (Segura et al., 2012). A previous genetic analysis of apple varieties in the ABC found a high degree of relatedness (i.e., siblings and first-degree relationships) among apple varieties found within the ABC, and that the population structure as determined using PCA is strongly correlated with the harvest date (Migicovsky et al., 2021). In order to decrease the impact of the observed population structure and relatedness on the GWA performed here, the first 5 principal components (PCs) and kinship matrix were included in the GWA model as covariates, which is standard practice when performing GWA (e.g., (Myles et al., 2009; Wang et al., 2005)). Thus, a simplified GWA model according to Yu et al. (2006) can be represented as follows:

$$\mathbf{Y} \sim \boldsymbol{\alpha} + \mathbf{Q} + \mathbf{K} + \mathbf{e}$$

where **Y** is a vector of phenotypic observations (i.e., abundance of a particular VOC across varieties); **α** is a vector of SNP effects (i.e., SNP genotypes across varieties); **Q** is a matrix including the values from 5 PCs across varieties that controls for population structure; **K** is the pairwise kinship matrix that controls for close relatedness among varieties; and **e** is the error term (i.e., vector of residual effects).

The kinship matrix accounts for the dependency among SNPs correlated with the phenotypes due to relatedness among apple varieties. The kinship matrix was calculated using a standalone version of Tassel (v5.0) GBS pipeline (v2) (Bradbury et al., 2007). To visualize GWA results, Manhattan plots and quantile-quantile (QQ) plots were generated using ggplot2 (v3.3.5) package in R. The fit of the mixed model to the data can be evaluated by observing the QQ plots: the closer the observed values are to the expected values, the better the model fit. The model fit was quantitatively evaluated using the genomic inflation factor (λ), which expresses the deviation of the distribution of the observed test statistic compared to the distribution of the expected test statistic (Devlin & Roeder, 1999). High genomic inflation factors ($\lambda >> 1$) indicate an excess of false positive genotype-phenotype associations that most frequently result from the model's inability to correct properly for population structure (Reich & Goldstein, 2001). The λ values are shown within each QQ plot to enable an evaluation of model fit.

The pairwise correlations among all pairs of compounds and among pairs of compound classes were calculated as Pearson correlations using the *cor.test* function from the *stats* package (v4.1.2). All computational tasks for this study were carried out on a high-performance computing (HPC) cluster from Compute Canada. Computer code was primarily written in the programming language R, but bash scripts were used for implementing procedures via the command line, especially in cases where parallel computing was required. All code and input files are available via the GitHub repository (2022) at the following URL: https://github.com/MylesLab/apple-aroma.

CHAPTER 4 — RESULTS AND DISCUSSION

4.1 — THE ARCHITECTURE OF APPLE AROMA

To investigate the aroma of apples, 106 volatiles were extracted and analyzed across 515 apple varieties using gas chromatography mass spectroscopy (GC-MS). The names of these compounds were manually curated to resemble the commonly written notation. The abundance of a compound is defined as the area under its chromatogram peak normalized by a benzaldehyde-d8 standard.

After harvesting the apples, they were subjected to one month of cold storage at 3-3.5˚C. Subsequently, these apples were sliced, and two random slices were flash frozen with liquid nitrogen and then stored at -80°C before GC-MS analysis. These pre-conditions that apples are subjected to before the extraction of volatiles have an effect on the food matrix and thus can alter the *in vitro* volatile profile (Dewulf et al., 2002; Farneti et al., 2013). Due to this, the volatile profile that is exhibited by the apple during the chewing process will likely differ from the volatile profile exhibited by the apple as processed using the extraction process employed here. Thus, the degree to which the measurement of the volatilome of an apple employed here fails to accurately represent the volatilome expressing during the chewing process is a caveat of the present study. However, the variation in the relative abundances of compounds and their correlation with genetic variation can nonetheless provide valuable insights about the genetic basis of aroma production, and this is the primary focus of present study.

Each compound was manually categorized into a class, which resulted in 13 different classes of compounds: acids, alcohols, aldehydes, C13-norisoprenoid, esters (straight chain), esters (branched chain), furans, hydrocarbons, ketones, lactones, monoterpenoids, sesquiterpenes, and sulfur/nitrogen compounds (Figure 1B and 1C). In order to assess the data across classes of compounds, I calculated the total volatile abundance of each compound class by summing the abundances of each VOC belonging to that class. I identified that esters, aldehydes, and alcohols are not only the most ubiquitous (Figure 1B) but also the most abundant compound classes (Figure 1C). In fact, nearly the entire apple volatilome as measured in this study is composed of esters, alcohols, and aldehydes: they make up ~ 98% of the total volatile abundance in our dataset (Figure 1C). This observation is in line with previous work showing that esters, aldehydes, and alcohols are the main contributors to fruit aroma (Espino-Díaz et al., 2016).

Esters were not only the most abundant, but there were also a relatively large number of ester compounds in our dataset: 40 (38%) of the compounds I identified were esters. Esters alone are known to account for 80% of the fruit volatiles in apples (De Pooter H Schamp N, 1989), and our results support the notion that esters are likely the largest contributors to the apple volatilome.

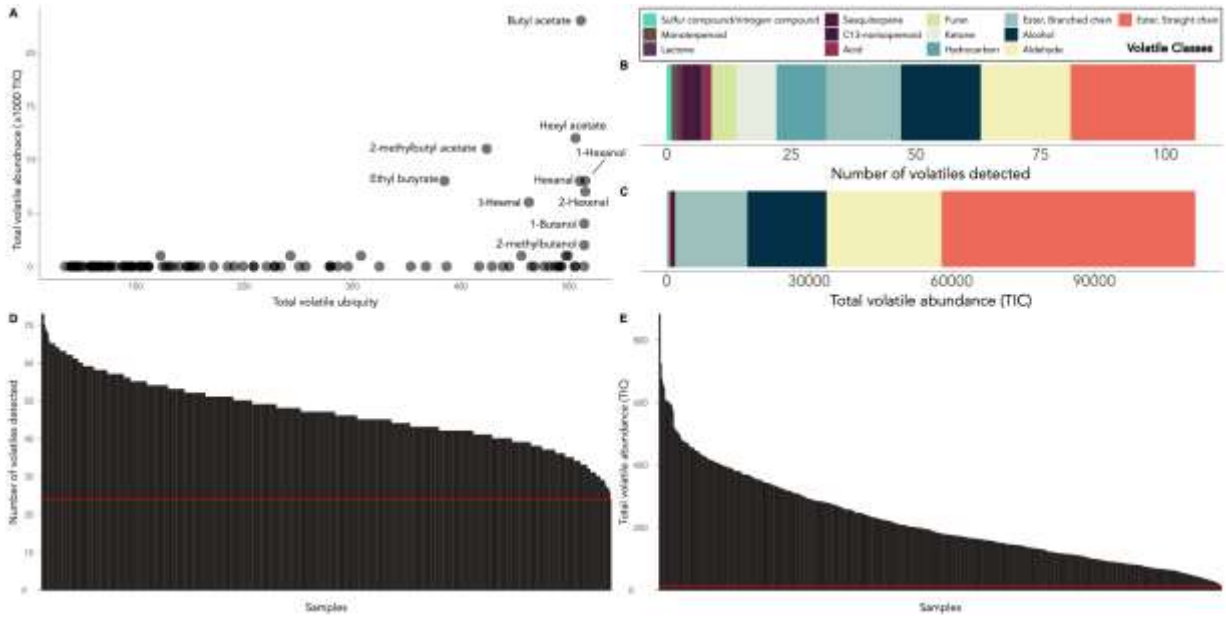

Figure 1 — **Volatile composition across 515 apple varieties.** A) The ubiquity of each volatile (x-axis) is plotted against its abundance (y-axis). B) The number of VOCs detected by compound class. C) The total volatile abundance by compound class. D) The number of volatiles detected (y-axis) sorted in descending order across 515 apple varieties (x-axis) E) The total volatile abundance (y-axis) sorted in descending order across 515 apple varieties. The red horizontal lines represent the minimum values used as thresholds for inclusion in the present study (see Materials and Methods).

In addition to our observation that a small number of compound classes dominate the apple volatilome, I observed that a relatively small number of individual VOCs account for a large proportion of the overall volatile abundance. These compounds are both ubiquitous and abundant: they were detected in nearly every variety and were present in high amounts. The most abundant compound in our dataset was butyl acetate (a volatile ester), which was detected in 511 of the 515 varieties and whose abundance across all varieties represented 20.8% of the overall volatile abundance in our data set. Butyl acetate is commonly used as a flavouring agent in various foods (*JECFA Evaluations*, 2022). The apple that has the highest amount of butyl acetate

is ‘Dukat’, a variety from Kazakhstan primarily eaten fresh or dried (*Dukat Apple*, 2022). Dukat is a cross between Golden Delicious (female parent) and Cox’s Orange (male parent). One of Dukat’s parents, Golden Delicious, is known to have high concentrations of butyl acetate, which is designated as an ‘impact compound’ for its ability to have a decisive impact on the sensory quality of the fruit (Song & Bangerth, 1996). Further, there generally are higher concentrations of volatile esters in Cox’s Orange and Golden Delicious apples, the two parents of Dukat (Dixon & Hewett, 2000). While butyl acetate was present in 511 varieties, it remained completely undetected in 4 varieties. The ubiquity and abundance in butyl acetate observed in the present study supports the notion that this compound likely plays a key role in the diversity of sensory quality across apple varieties.

The compound which is present in all of the 515 varieties (*i.e.,* the most ubiquitous) is 1-hexanol and its total abundance across all varieties represented 7.8% of the overall volatile abundance in our dataset. It has been previously shown that the exogenous application of 1-hexanol onto apple fruit induces soft scald, a common post-harvest disorder that appears in response to cold storage after about 2-8 weeks (Wills, 1972). The variety ‘Honeycrisp’ is known to be highly susceptible to soft scald (Xu et al., 2017), but its level of 1-hexanol is only slightly above the median value of the 515 varieties evaluated here. It is therefore unclear whether endogenous production of 1-hexanol mediates soft scold susceptibility and, by extension, whether selection against 1-hexanol production by apple breeders, either phenotypically or using genetic markers, may be an effective strategy for selecting for resistance to soft scald.

The apple with the highest number of VOCs is Red Cinnamon, which expressed 73 out of the 106 VOCs. The apple with the highest cumulative VOCs abundance was Krapchatoe. Descriptions of these two varieties online failed to reveal why they may lie at the extremes of our data distributions. Conversely, the apple that has the lowest volatile abundance is ‘Black Ben Davis’. This apple is known for its ruggedness, but generally has a poor flavour. It was famous in the 19th and 20th centuries because it was easier to ship. However, as the shipping and packing improved, this variety fell out of favour (Beach et al., 1905). This is a prime example of how flavour is an essential consideration for consumers but is often overlooked because of a focus on production-related traits.

Figure 2 — **Pairwise correlation heatmap among the volatiles.** The correlation represents the coefficient (r) values ranging from -1 to 1. Colours range from blue to red, representing negative to positive correlations, respectively.

In order to understand the relationships among VOCs, I created a heatmap displaying correlations among all pairwise combinations of VOCs (Figure 2). I observed a statistically significant excess of positive correlations (N = 704) compared to negative correlations (N = 22; Pearson's Chi-squared test, $p = 7.39 \times 10^{-91}$). This suggests that an increase in the abundance of one compound generally causes the abundance of another compound to increase rather than decrease. Furthermore, significant positive correlations were more often observed between pairs of compounds within the same class, and less often between compounds of different classes. For example, one of the strongest correlations observed is between (E)-2-octenal and 2-heptenal (r=0.97; p=1.45x10^{-317}), two aldehydes that appear to be extremely tightly co-expressed (Figure 3). Both of these compounds are known autoxidation products of linoleic acid and may contribute to the off-flavour of apple juice in the presence of light (Hashizume et al., 2007). A full investigation of each pair of positively and negatively correlated compounds is beyond the

scope of the present study. However, these data may be further analyzed to elucidate groups of compounds involved in similar metabolic pathways.

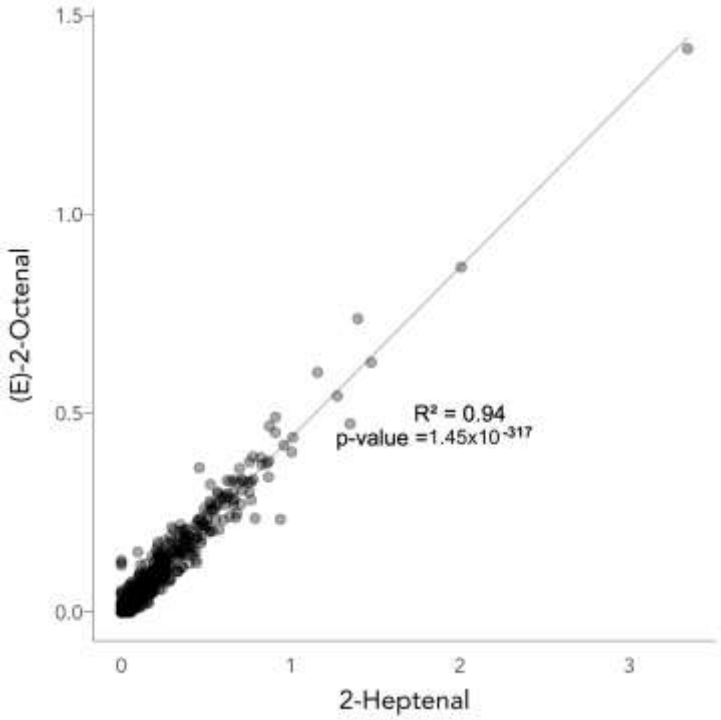

Figure 3 — **Correlation of the abundance of (E)-2-octenal and 2-heptenal.** The coefficient of correlation (R^2) and significance value is calculated by fitting a linear model.

4.2 — HARVEST DATE SHAPES THE APPLE VOLATILOME

In order to assess the relationships among varieties based on their volatilomes, I performed PCA on the entire VOC data set. The first two PCs explained 17.8% of the total variance and PC1 generally separated varieties based on their harvest date (Figures 4A and 4B).

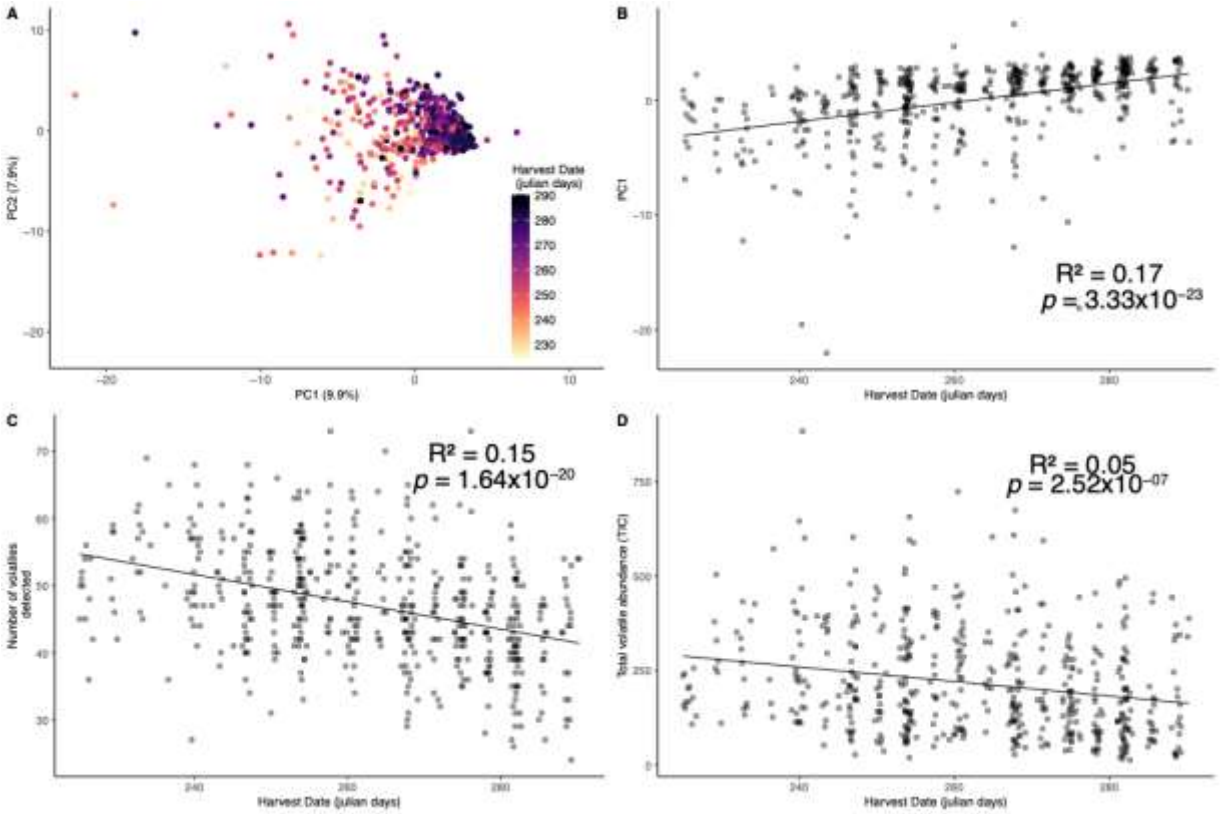

Figure 4 — **Variation in volatilome is correlated with harvest date.** A) PCA bi-plot of PC1 and PC2 derived from a matrix of 106 VOCs across 515 apple varieties. Each variety is coloured according to its harvest date. B) Scatter plot of PC1 values and harvest date across 515 varieties. C) Scatter plot of ubiquity (i.e. the number of VOCs detected) and harvest date across 515 varieties D) Scatter plot of total volatile abundance and harvest date across 515 varieties. Lines of best fit, R^2 and P values result from Pearson correlations between variables.

I observed a significant correlation between harvest date and PC1 (R^2=0.17; p=3.33x10^{-23}): the varieties that ripen quickly and are harvested early tend to have lower PC1 values while the varieties that ripen slowly and are harvested late tend to have higher PC1 values (Figure 4B). This relationship is driven by our observation that harvest date is negatively correlated with the ubiquity and abundance of VOCs (Figures 4C and 4D). This suggests that early harvested apple varieties tend to have a larger number of VOCs and more aroma overall than late-harvested apples. One of the possible mechanisms for this phenomenon could be that the ripening process is simply accelerated in the early-ripening varieties and is slower in late-ripening varieties. In other words, the overall metabolic activity in early-harvested varieties may be higher than in late ripening varieties. This acceleration among early-ripening varieties may result in more volatiles

being produced at any given time since the breakdown and build-up of volatile compounds is higher in these early-harvested varieties in comparison to the late-harvested varieties. Taken together, these results suggest that the time it takes an apple to ripen on the tree is intimately linked to its volatilome.

4.3 — GENETIC CONTROL OF VOCs

I performed GWAS to identify the loci in the genome that are responsible for the production of various VOCs. The interpretation of the resulting Manhattan and QQ plots was often challenging due to many insignificant genomic associations to the compounds. This suggested that numerous compounds in our dataset were not adequately abundant or ubiquitous to provide the statistical power for GWA. These compounds sometimes generated statistically significant SNP-phenotype associations, but an assessment of the QQ plots and the genomic inflation factors (λ) suggested an excess of false positives: the distribution of the values across varieties frequently resulted in a poor fit for the mixed model GWA algorithm that I employed. Thus, I instead focus on the most reliable genotype-phenotype associations by only considering those that form a clear and reliable peak in a Manhattan plot, a signal that suggests the identification of a single genetic locus of large effect on the concentration of a VOC. The Manhattan plots, along with their QQ plots, are shown as supplementary figures (Figures S2-S7). While a further exploration of more complex GWA statistical models suitable for challenging data distributions could be explored in the future, only the most promising associations identified using a single GWA model are explored in detail below.

1-butanol and 1-hexanol are among the most ubiquitous and abundant VOCs detected and both their concentrations appear to be mediated by genetic variation at a single locus on chromosome 3. The most significant SNP associated with these two compounds lies within the *NAC18.1* gene, a member of the NAC family of transcription factors. Functional genomics studies across diverse species have demonstrated that NAC transcription factors are implicated in ripening phenotypes across diverse agricultural crops, including tomato (Kumar et al., 2018), melon (Ríos et al., 2017), banana (Shan et al., 2012), peach (Pirona et al., 2013), and apricot (García-Gómez et al., 2019). Notably, the homolog of *NAC18.1* in tomatoes (*Solanum lycopersicum*) is the *NON-RIPENING (NOR)* gene, a well-studied gene that, when knocked out in tomatoes, produces the *nor* mutant tomato that does not ripen (Tigchelaar, 1973). In apple,

numerous recent GWASs have repeatedly identified associations between *NAC18.1* and harvest date (Jung et al., 2020; Larsen et al., 2019; McClure et al., 2018; Migicovsky et al., 2016; Urrestarazu et al., 2017). A recent study demonstrated that introducing the apple *NAC18.1* transgene into a *nor* mutant tomato recovers ripening, providing strong evidence that genetic variation within *NAC18.1* mediates the apple ripening process (Migicovsky et al., 2021). Our observation that concentrations of 1-butanol and 1-hexanol were associated with genetic variation in *NAC18.1* is consistent with our observation that harvest date is negatively correlated with both 1-butanol ($R = -0.46$, $p = 2.86 \times 10^{28}$) and 1-hexanol ($R = -0.44$, $p = 9.75 \times 10^{-26}$). Both of these key aromatic compounds are more abundant in early ripening varieties than in late-ripening varieties, and our results suggest that the reduction in expression of these compounds over the harvest season is mediated by genetic variation in or near the *NAC18.1* gene. It is noteworthy that firmness is also strongly associated with harvest date, whereby early harvested varieties tend to be softer than late-harvested varieties (Johnston et al., 2002; Migicovsky et al., 2016; Nybom et al., 2013; Oraguzie et al., 2004). We, therefore, propose that the *NAC 18.1* locus is a master regulator of apple ripening, and alleles at this locus modulate numerous ripening-associated phenotypes including harvest date, firmness, and the expression of VOCs.

I found strong associations at a single locus (chromosome 2 at 1,120,527 bp) in the genome for various ester compounds: 2-methyl-butyl acetate, isobutyl acetate, and hexyl acetate. This region in the genome encodes genes such as ribosomal protein S11-beta (MD02G1015400) and HXXXD-type acyltransferase (*AAT1*) (MD02G1013900). The *AAT1* gene was involved in the production of esters by using the alcohols as substrates in an oxygen-dependent reaction where an acyl group is transferred from acyl-CoA to the oxygen and an alcohol's hydroxyl group., forming an ester (Espino-Díaz et al., 2016). At the aforementioned locus on chromosome 2, I also found a strong association for hexyl acetate. This compound has been linked to soft scald in apples (Wills, 1972). Exploring the genes around this locus may hold the key to addressing this disease in apples. Finally the results also showed a significant association for 2-4-hexadienal in the middle of 2-oxoglutarate (2OG) Fe(II) dependent oxygenase superfamily (MD07G1005600). It was not clear what impact this gene or the markers nearby this locus have on the production of this VOC. This lack of clarity on potential functions of the loci around the strongly associated compounds is due to relatively poor annotation of the apple genome. Non-model eukaryotic genomes generally are poorly annotated due to their large size and intron-

containing genes make them difficult substrates for annotation (Liolios, 2006). Future studies with high-density markers with improved reference genome annotations will provide more details in terms of the potential causal alleles.

Table 1 — The VOCs strongly associated with the genomic variation tabulated according to their chromosome positions. The p-value represents the significance of association after Bonferroni correction for multiple comparisons. The SNP positions are based on GDDH13 v1.1 reference genome (Zhang et al., 2019).

Compound	**Chr**	**Position**	**Allele**	***p*-value**	**Nearby Candidate Genes**
2-4-hexadienal	7	576999	G	5.28×10^{-08}	2-oxoglutarate (2OG) Fe(II) dependent oxygenase superfamily (MD07G1005600) ; Rad21/Rec8-like family protein (MD07G1005000) ; C2 calcium/lipid-binding plant phosphoribosyltransferase family protein (MD07G1005400) ; Ribosomal protein L6 family protein (MD07G1005300) ; tetratricopeptide repeat (TPR)-containing protein (MD07G1005500), and AAA-type ATPase family protein (MD07G1006000)
2-methylbutyl acetate	2	1120527	A	3.64×10^{-18}	ribosomal protein S11-beta (MD02G1015400) ; RNA-binding KH domain-containing protein (MD02G1015700) ; HXXXD-type acyl-transferase family protein (AAT1) (MD02G1013900)
hexyl acetate	2	1120527	A	1.29×10^{-16}	
isobutyl acetate	2	1120527	A	3.74×10^{-20}	
butyl acetate	2	1164704	A	5.16×10^{-19}	N-acetyltransferase (MDP0000162960)
n-propyl-acetate	2	1331428	T	7.15×10^{-18}	translation initiation factor activity molecular function (MDP0000160803)
pentyl acetate	2	1331428	T	1.28×10^{-20}	hydrolase activity molecular function (MDP0000199630)
1-butanol	3	30698039	A	1.05×10^{-11}	*NAC18.1*
1-hexanol	3	30698039	A	3.42×10^{-11}	*NAC18.1*

CHAPTER 5 — CONCLUSIONS

Apples are one of the world's most important fruit crops. However, the few elite varieties that are on market have been bred for production-related traits such as firmness and storability. Apple flavour, which is arguably the most important trait for consumers, has been largely ignored in commercial breeding programmes. Flavour is a complex trait determined in part by the aroma of apples. In this study, I have laid the groundwork for understanding the aroma-producing compounds (i.e., VOCs) in apples. I quantified 106 VOCs across 515 apple varieties and coupled these data with over 250,000 genetic markers from across the genome with the aim of identifying genomic regions that control apple aroma production.

By analyzing this multi-dimensional VOC dataset, I found that esters were the largest group of VOCs identified. Knowing from the literature that these are the most important compounds involved in aroma production, my work makes the case that improving apple aroma will likely involve altering the abundance of esters in some shape or form. Further, I find a single locus in the genome on chromosome 2 (1120527 bp) which is responsible for the variation in the abundance of most important esters such as hexyl acetate and 2-methylbutyl acetate. This suggests that there is a single regulator at the top of the pathway involved in ester production. Taken together, these insights about potential ester regulation provide novel avenues for evaluating the changes in ester on the fruit flavour.

Previous works in the literature have shown the association of *NAC18.1* gene with the variation in harvest date, firmness and other ripening processes in apples. My research expands on this understanding by providing additional layer of information that the variation in 1-hexanol and 1-butanol (two of the most important alcohols) is also associated with *NAC18.1* gene. We, therefore, propose that the *NAC 18.1* locus is a master regulator of apple ripening, and alleles at this locus modulate numerous ripening-associated phenotypes including harvest date, firmness, and the expression of VOCs.

This exploratory study confirms previous knowledge of VOCs (such as esters being the largest group of VOCs in apples), as well as brings novel insights such as the involvement of *NAC18.1* in alcohol production. Most importantly, this study provides potential gene targets for altering and evaluating the aroma of apples.

REFERENCES

Abdullah, A. A., Altaf-Ul-Amin, M., Ono, N., Sato, T., Sugiura, T., Morita, A. H., Katsuragi, T., Muto, A., Nishioka, T., & Kanaya, S. (2015). Development and mining of a volatile organic compound database. *BioMed Research International*, *2015*, 139254.

Beach, S. A., Booth, N. O., & Taylor, O. M. (1905). *The Apples of New York* (pp. 68–71). J.B. Lyon.

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

Brown, S. K., & Maloney, K. E. (2003). Genetic improvement of apple: breeding, markers, mapping and biotechnology. In *Apples: botany, production and uses* (pp. 31–59). https://doi.org/10.1079/9780851995922.0031

Bus, V. G. M., Rikkerink, E. H. A., Caffier, V., Durel, C.-E., & Plummer, K. M. (2011). Revision of the nomenclature of the differential host-pathogen interactions of Venturia inaequalis and Malus. *Annual Review of Phytopathology*, *49*, 391–413.

Bus, V., Ranatunga, C., Gardiner, S., Bssett, H., & Rikkerink, E. (2000). MARKER ASSISTED SELECTION FOR PEST AND DISEASE RESISTANCE IN THE NEW ZEALAND APPLE BREEDING PROGRAMME. In *Acta Horticulturae* (Issue 538, pp. 541–547). https://doi.org/10.17660/actahortic.2000.538.95

Colantonio, V., Ferrão, L. F. V., Tieman, D., Bliznyuk, N., Sims, C., Klee, H., Munoz, P. R., & Resende, M. F. R. (2022). *Metabolomic Selection for Enhanced Fruit Flavor*. https://doi.org/10.1101/2020.09.17.302802

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

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

De Pooter H Schamp N, D. P. (1989). Flavor chemistry: Trends and development. In Teranishi R Buttery R Shahidi (Ed.), *Flavor chemistry: Trends and development* (pp. 24–34). American Chemical Society.

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997–1004.

Dewulf, J., Van Langenhove, H., & Wittmann, G. (2002). Analysis of volatile organic compounds using gas chromatography. *Trends in Analytical Chemistry: TRAC*, *21*(9–10), 637–646.

Dixon, J., & Hewett, E. W. (2000). Factors affecting apple aroma/flavour volatile concentration: A Review. *New Zealand Journal of Crop and Horticultural Science*, *28*(3), 155–173.

Dukat Apple. (2022, Summer 7). Speciality Produce. https://specialtyproduce.com/produce/Dukat_Apple_18566.php

Edge-Garza, & Peace. (2010). Enabling marker-assisted seedling selection in the Washington apple breeding program. *On Molecular Markers in Horticulture 859*. https://www.actahort.org/books/859/859_44.htm

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One*, *6*(5), e19379.

Espino-Díaz, M., Sepúlveda, D. R., González-Aguilar, G., & Olivas, G. I. (2016). Biochemistry of Apple Aroma: A Review. *Food Technology and Biotechnology*, *54*(4), 375–397.

Farneti, B., Alarcón, A. A., Cristescu, S. M., Costa, G., Harren, F. J. M., Holthuysen, N. T. E., & Woltering, E. J. (2013). Aroma volatile release kinetics of tomato genotypes measured by PTR-MS following artificial chewing. *Food Research International (Ottawa, Ont.)*, *54*(2), 1579–1588.

Farneti, B., Di Guardo, M., Khomenko, I., Cappellin, L., Biasioli, F., Velasco, R., & Costa, F. (2017). Genome-wide association study unravels the genetic control of the apple volatilome and its interplay with fruit texture. *Journal of Experimental Botany*, *68*(7), 1467–1478.

Ferrão, L. F. V., Johnson, T. S., Benevenuto, J., Edger, P. P., Colquhoun, T. A., & Munoz, P. R. (2020). Genome-wide association of volatiles reveals candidate loci for blueberry flavor. *The New Phytologist*, *226*(6), 1725–1737.

García-Gómez, B. E., Salazar, J. A., Dondini, L., Martínez-Gómez, P., & Ruiz, D. (2019). Identification of QTLs linked to fruit quality traits in apricot (Prunus armeniaca L.) and biological validation through gene expression analysis using qPCR. *Molecular Breeding: New Strategies in Plant Improvement*, *39*(2). https://doi.org/10.1007/s11032-018-0926-7

GitHub Repository. (2022, August 3). GitHub. https://github.com

Gupta, P. K., Kulwal, P. L., & Jaiswal, V. (2019). Association mapping in plants in the post-GWAS genomics era. *Advances in Genetics*, *104*, 75–154.

Hashizume, M., Gordon, M. H., & Mottram, D. S. (2007). Light-induced off-flavor development in cloudy apple juice. *Journal of Agricultural and Food Chemistry*, *55*(22), 9177–9182.

Igarashi, M., Hatsuyama, Y., Harada, T., & Fukasawa-Akada, T. (2016). Biotechnology and apple breeding in Japan. *Breeding Science*, *66*(1), 18–33.

Janick. (2012). Fruit breeding: past, present and future. *XXII Congresso Brasilerio de. Fruiticultura, Bento*. https://www.hort.purdue.edu/newcrop/pdfs/fruit-breeding-brazil.pdf

JECFA Evaluations. (2022, Winter 7). IPCS InChem. https://inchem.org/documents/jecfa/jeceval/jec_252.htm

Johnston, J. W., Hewett, E. W., Hertog, M. L. A. T. M., & Harker, F. R. (2002). Harvest date and fruit size affect postharvest softening of apple fruit. *The Journal of Horticultural Science & Biotechnology*, *77*(3), 355–360.

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

Klee, H. J. (2010). Improving the flavor of fresh fruits: genomics, biochemistry, and biotechnology. *The New Phytologist*, *187*(1), 44–56.

Klee, H. J., & Tieman, D. M. (2013). Genetic challenges of flavor improvement in tomato. *Trends in Genetics: TIG*, *29*(4), 257–262.

Klee, H. J., & Tieman, D. M. (2018). The genetics of fruit flavour preferences. *Nature Reviews. Genetics*, *19*(6), 347–356.

Kole, C. (Ed.). (2011). *Wild Crop Relatives: Genomic and Breeding Resources: Temperate Fruits*. Springer, Berlin, Heidelberg.

Kováts, E. (1958). Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helvetica Chimica Acta*, *41*(7), 1915–1932.

Kumar, R., Tamboli, V., Sharma, R., & Sreelakshmi, Y. (2018). NAC-NOR mutations in tomato Penjar accessions attenuate multiple metabolic processes and prolong the fruit shelf life. *Food Chemistry*, *259*, 234–244.

Larsen, B., Migicovsky, Z., Jeppesen, A. A., Gardner, K. M., Toldam-Andersen, T. B., Myles, S., Ørgaard, M., Petersen, M. A., & Pedersen, C. (2019). Genome-Wide Association Studies in Apple Reveal Loci for Aroma Volatiles, Sugar Composition, and Harvest Date. In *The Plant Genome* (Vol. 12, Issue 2, p. 180104). https://doi.org/10.3835/plantgenome2018.12.0104

Liolios, K. (2006). The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. In *Nucleic Acids Research* (Vol. 34, Issue 90001, pp. D332–D334). https://doi.org/10.1093/nar/gkj145

Mansurova, M., Ebert, B. E., Blank, L. M., & Ibáñez, A. J. (2018). A breath of information: the volatilome. *Current Genetics*, *64*(4), 959–964.

McClure, K. A., Gardner, K. M., Douglas, G. M., Song, J., Forney, C. F., DeLong, J., Fan, L., Du, L., Toivonen, P. M. A., Somers, D. J., Rajcan, I., & Myles, S. (2018). A Genome-Wide Association Study of Apple Quality and Scab Resistance. *The Plant Genome*, *11*(1). https://doi.org/10.3835/plantgenome2017.08.0075

Migicovsky, Z., Gardner, K. M., Money, D., Sawler, J., Bloom, J. S., Moffett, P., Chao, C. T., Schwaninger, H., Fazio, G., Zhong, G.-Y., & Myles, S. (2016). Genome to Phenome Mapping in Apple Using Historical Data. *The Plant Genome*, *9*(2). https://doi.org/10.3835/plantgenome2015.11.0113

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

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

Migocovsky, Z., Douglas, G. M., & Myles, S. (2022). Genotyping-by-Sequencing of Canada's Apple Biodiversity Collection. *Frontiers in Genetics*.

Myles, S. (2013). Improving fruit and wine: what does genomics have to offer? *Trends in Genetics: TIG*, *29*(4), 190–196.

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

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

Oraguzie, N. C., Iwanami, H., Soejima, J., Harada, T., & Hall, A. (2004). Inheritance of the Md-ACS1 gene and its relationship to fruit softening in apple (Malus × domestica Borkh.). In *Theoretical and Applied Genetics* (Vol. 108, Issue 8, pp. 1526–1533). https://doi.org/10.1007/s00122-003-1574-8

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

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

Qin, G., Tao, S., Cao, Y., Wu, J., Zhang, H., Huang, W., & Zhang, S. (2012). Evaluation of the volatile profile of 33 Pyrus ussuriensis cultivars by HS-SPME with GC–MS. *Food Chemistry*, *134*(4), 2367–2382.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rafalski, J. A. (2010). Association genetics in crop improvement. *Current Opinion in Plant Biology*, *13*(2), 174–180.

Ravi, R., Taheri, A., Khandekar, D., & Millas, R. (2019). Rapid Profiling of Soybean Aromatic Compounds Using Electronic Nose. *Biosensors*, *9*(2). https://doi.org/10.3390/bios9020066

Reich, D. E., & Goldstein, D. B. (2001). Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology*, *20*(1), 4–16.

Ribaut, J.-M., & Hoisington, D. (1998). Marker-assisted selection: new tools and strategies. *Trends in Plant Science*, *3*(6), 236–239.

Ríos, P., Argyris, J., Vegas, J., Leida, C., Kenigswald, M., Tzuri, G., Troadec, C., Bendahmane, A., Katzir, N., Picó, B., Monforte, A. J., & Garcia-Mas, J. (2017). ETHQV6.3 is involved in melon climacteric fruit ripening and is encoded by a NAC domain transcription factor. *The Plant Journal: For Cell and Molecular Biology*, *91*(4), 671–683.

Schatz, M. C., & Langmead, B. (2013). The DNA data deluge. *IEEE Spectrum*, *50*(7), 28–33.

Section, H. (2021). Statistical Overview of the Canadian Fruit Industry 2020. *Horticulture Section - Crops and Horticulture Division Agriculture and Agri-Food Canada*. https://publications.gc.ca/collections/collection_2021/aac-aafc/A71-33-2020-eng.pdf

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, *44*(7), 825–830.

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., & 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*(14), 5171–5187.

Shi, J., Wu, H., Xiong, M., Chen, Y., Chen, J., Zhou, B., Wang, H., Li, L., Fu, X., Bie, Z., & Huang, Y. (2020). Comparative analysis of volatile compounds in thirty nine melon cultivars by headspace solid-phase microextraction and gas chromatography-mass spectrometry. In *Food Chemistry* (Vol. 316, p. 126342). https://doi.org/10.1016/j.foodchem.2020.126342

Song, J., & Bangerth, F. (1996). The effect of harvest date on aroma compound production from 'Golden Delicious' apple fruit and relationship to respiration and ethylene production. In *Postharvest Biology and Technology* (Vol. 8, Issue 4, pp. 259–269). https://doi.org/10.1016/0925-5214(96)00020-8

Souleyre, E. J. F., Chagné, D., Chen, X., Tomes, S., Turner, R. M., Wang, M. Y., Maddumage, R., Hunt, M. B., Winz, R. A., Wiedow, C., Hamiaux, C., Gardiner, S. E., Rowan, D. D., & Atkinson, R. G. (2014). The AAT1 locus is critical for the biosynthesis of esters contributing to "ripe apple" flavour in "Royal Gala" and "Granny Smith" apples. *The Plant Journal: For Cell and Molecular Biology*, *78*(6), 903–915.

Sugimoto, N., Engelgau, P., Jones, A. D., Song, J., & Beaudry, R. (2021). Citramalate synthase yields a biosynthetic pathway for isoleucine and straight- and branched-chain ester formation in ripening apple fruit. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(3). https://doi.org/10.1073/pnas.2009988118

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

Tieman, D. M., Zeigler, M., Schmelz, E. A., Taylor, M. G., Bliss, P., Kirst, M., & Klee, H. J. (2006). Identification of loci affecting flavour volatile emissions in tomato fruits. *Journal of Experimental Botany*, *57*(4), 887–896.

Tigchelaar, E. C. (1973). A new ripening mutant, non-ripening (nor). *Report of the Tomato Genetics Cooperative. Tomato Genetics Cooperative*, *35*, 20.

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

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

Wang, W. Y. S., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews. Genetics*, *6*(2), 109–118.

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

Wills, R. B. H. (1972). Effect of hexyl compounds on soft scald of apples. *Phytochemistry*, *11*(6), 1945–1946.

Xu, Y., Ma, Y., Howard, N. P., Chen, C., Tong, C. B. S., Celio, G., DeEll, J. R., & Moran, R. E. (2017). Microstructure of Soft Scald in 'Honeycrisp' Apples. In *Journal of the American Society for Horticultural Science* (Vol. 142, Issue 6, pp. 464–469). https://doi.org/10.21273/jashs04250-17

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

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

APPENDIX A — STUDENT CONTRIBUTION TO THE MANUSCRIPT

Soomro T, Jordan M, Watts S, Migicovsky Z, Forney C, Song J, Myles S. (2022). **Harvest date shapes apple volatilome**. Horticulture Research. (in preparation).

Conceived, designed and performed the biological experiments: Charles Forney, Michael Jordan and Sean Myles

Conceived the research questions: **Tayab Soomro**, Sean Myles

Curated and analyzed the data: **Tayab Soomro,** Sean Myles

Phenotype data collection: Sophie Watts, Zoe Migicovsky

Contributed in code-reviews, ideation, and other moral support: Tommy Davies, Sophie Watts, Zoe Migicovsky

Wrote the manuscript: **Tayab Soomro**

APPENDIX B — ASSAY METHODS FOR DETECTING INDIVIDUAL MARKERS

B.1 — Kompetitive Allele Specific PCR (KASP) genotyping assay for citrimalate synthease SNP

Methods: KASP assays were run on genomic DNA samples normalized in 384-well plates. Primer sequences for the assay can be found in **Table B1**.

Primers used in the KASP assay.

Primer Name	Primer Sequence (5' – 3')
MD_CMS_KASP_G2	GAAGGTGACCAAGTTCATGCTatgccagtggaattcacg
MD_CMS_KASP_C2	GAAGGTCGGAGTCAACGGATTatgccagtggaattcacc
MD_CMS_387_R3	AACTGCAAAATAAAAGTTAATATGGAAA

Each assay (5ul reaction) was run using the following reagents:

0.07 µl of 100 µM primers
2.5 µl KASP mastermix
2.5 µl of 3 ng/µl DNA

We used the following program on a CFX cycler:

Stage 1: 94 °C for 15 minutes (Hot-start activation)

Stage 2 (Touchdown):
94 °C for 20 seconds
61 °C for 1 minute
(decrease by 0.6°C per cycle)
[REPEAT 10x]
Stage 3 (amplification):
94 °C for 20 seconds
55 °C for 1 minute
[REPEAT 26x]

The plates were then read in BioRad CFX for 1 minute at 37 °C.

B.2 — High-resolution DNA melting (HRM)-based assay method to detect redTE

Methods:

High-Resolution DNA melting (HRM)-based genotyping assays were run on genomic DNA samples normalized in 384-well plates. Primer sequences for the assay were taken from Zhang et al. (2019) and can be found in **Table 1**.

Table B1. Primers used in the HRM assay.

Primer Name	Primer Sequence (5' – 3')
redTE_F	GGTCACCCAACCCACACTGGGCCTTG
redTE_R	CGGCCGCAATCGCAAGACGCAGA

Each assay was run using the following reagents:

0.1ul of 100uM primers
2ul of 5ng/ul DNA
2.025 ul sterile water
0.5ul of 10x PCR buffer
0.05ul of 5U/ul Taq
0.125ul 10x Evagreen Dye
0.2 ul of 10mM dNTPs

We used the following program on a CFX cycler:

95°C for 3:30 minutes
95°C for 30 seconds
60°C for 30 seconds
72°C for 30 seconds
[REPEAT 34x]
72°C for 5 minutes

The plates were denatured for 3-5 minutes before they were read using a Lightscanner HR384 (Biofire) and LightScanner Software with Call-IT 2.5 in ‘amplicon genotyping’ mode.

APPENDIX C — SUPPORTING INFORMATION

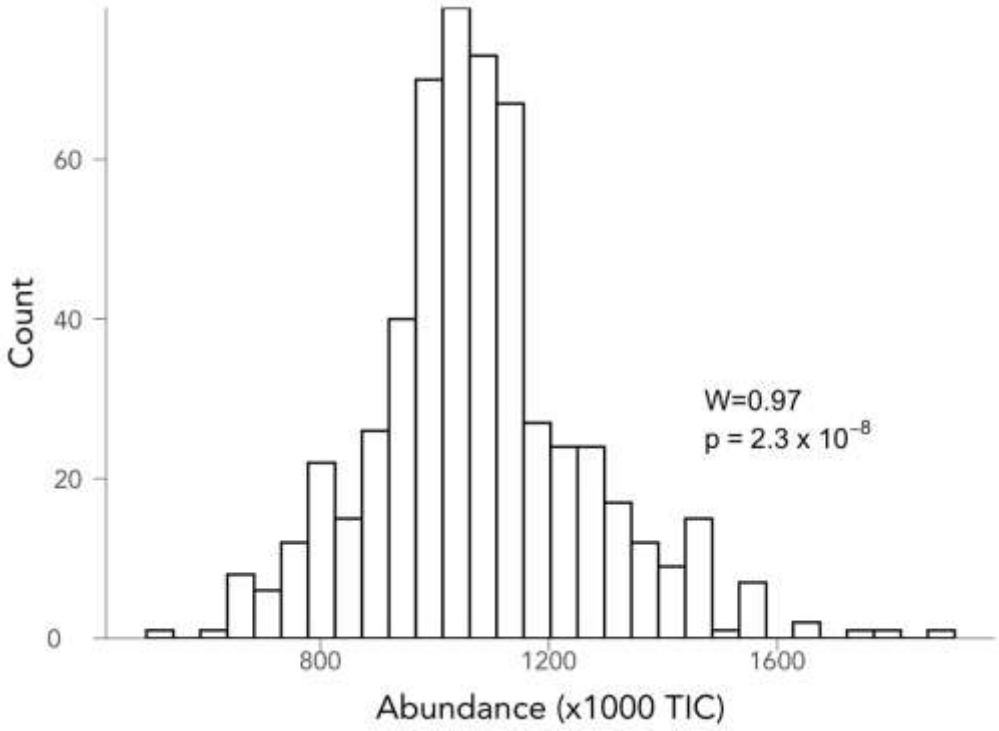

Figure S 1 — **Distribution of abundance value for benzaldehyde-d8 standard across the dataset.** The test statistic and significance value where calculated using a Shapiro-Wilks test for normality.

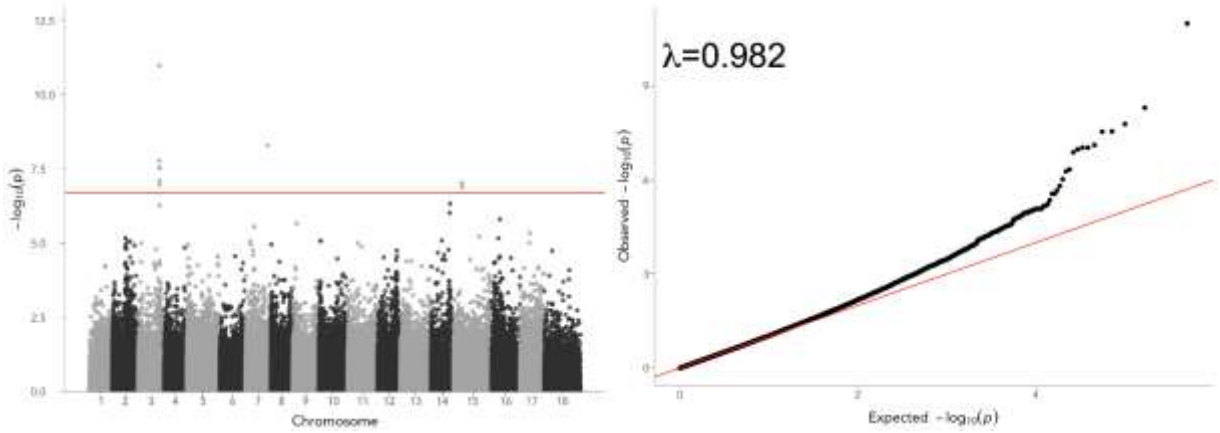

Figure S 2 — **Manhattan plot for 1-butanol along with QQ-plot.**

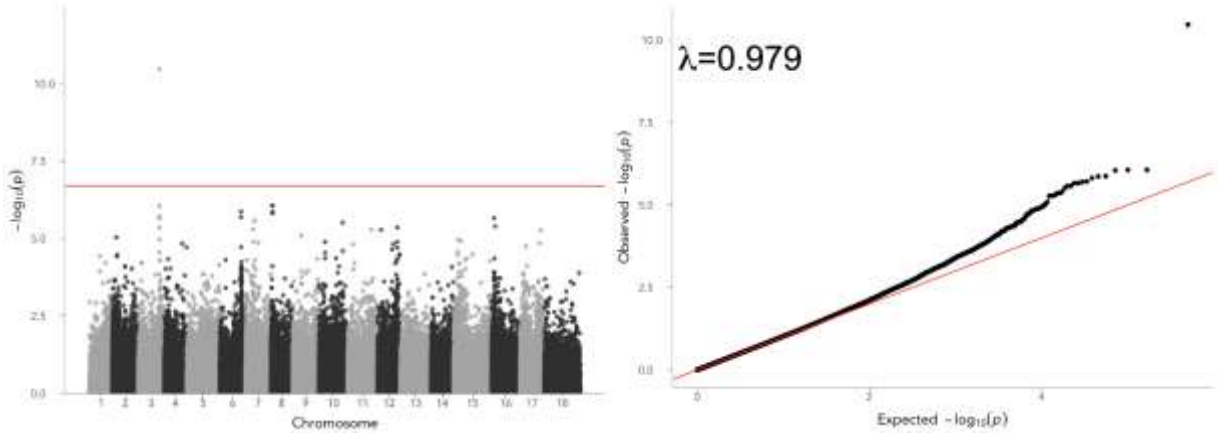

Figure S 3 — **Manhattan plot for 1-hexanol along with QQ-plot.**

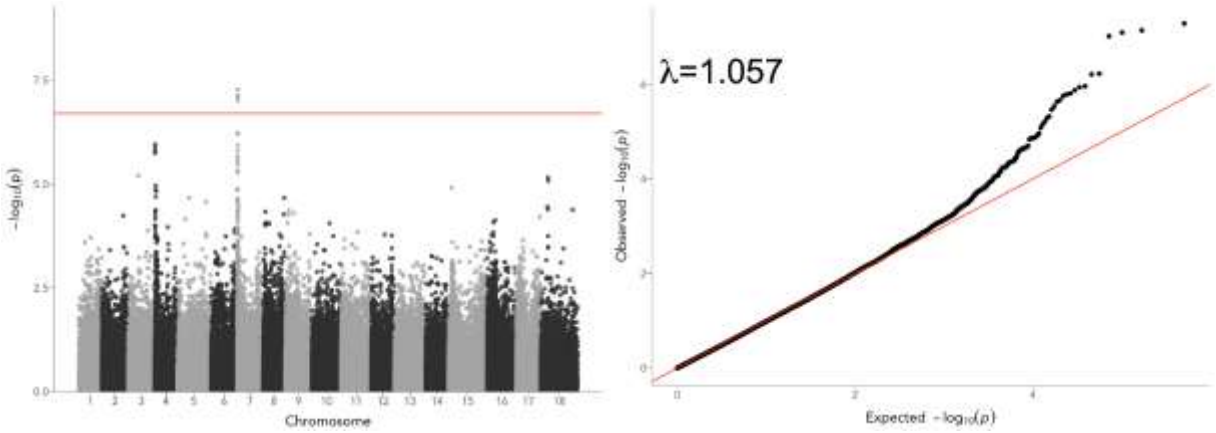

Figure S 4 — **Manhattan plot for 2-4-hexadienal along with QQ-plot**

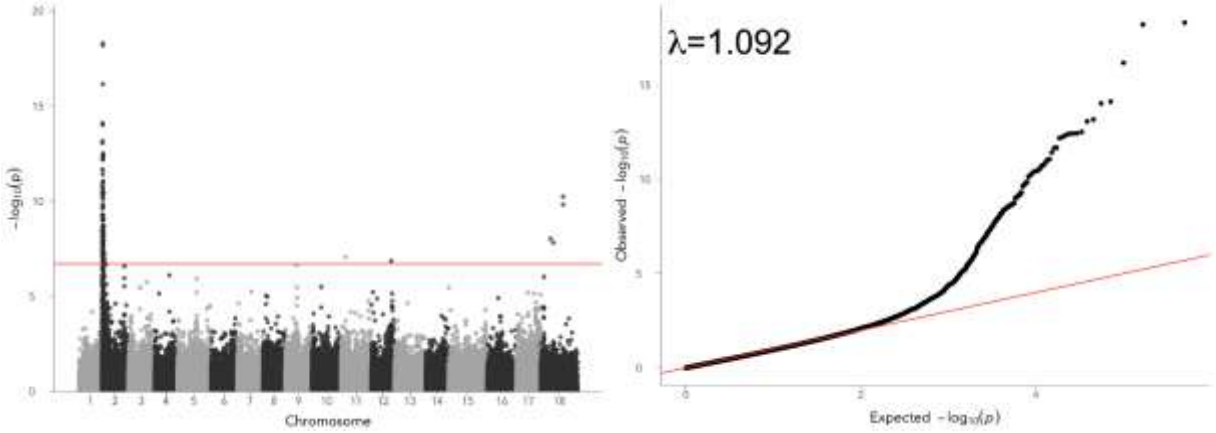

Figure S 5 — **Manhattan plot for butyl acetate along with QQ-plot.**

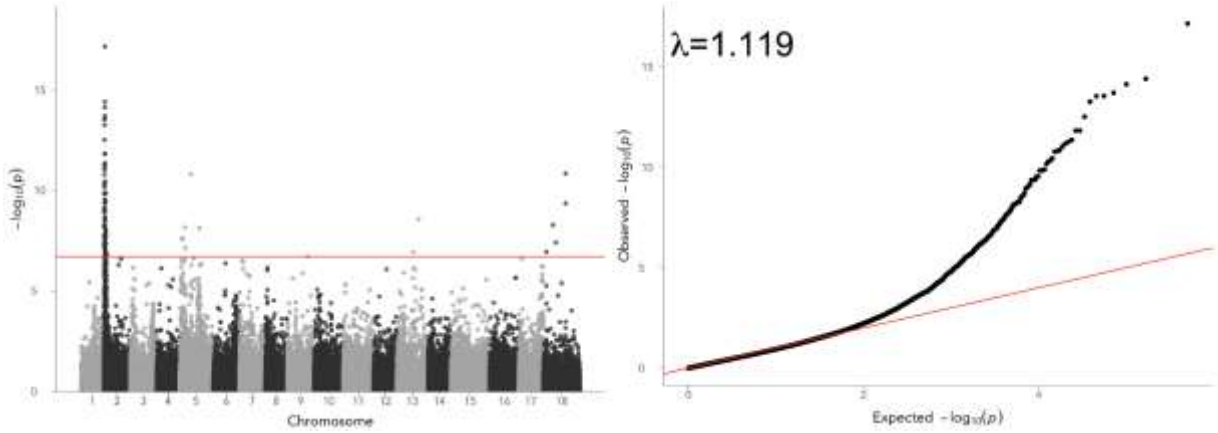

Figure S 6 — **Manhattan plot for n-propyl acetate along with QQ-plot.**

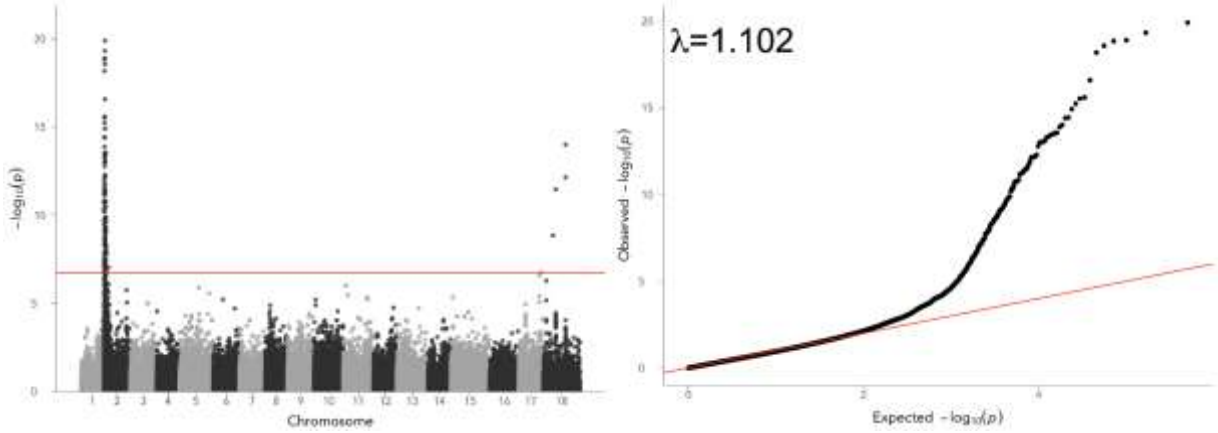

Figure S 7 — **Manhattan plot for pentyl acetate along with QQ-plot.**

Table S 1 — A matrix of volatile compound abundance data for each apple variety. The apple id is a unique identifier of apple variety and is connected with the apple trait data from Watts et al. (2021).