DEVELOPMENT OF COMPUTATIONAL MODELS, BIOMARKERS, AND TOOLS FOR POSTHARVEST TRAITS IN *MALUS DOMESTICA* FRUIT

By

JOHN ANTHONY HADISH

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
Program in Molecular Plant Sciences

DECEMBER 2023

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of JOHN ANTHONY HADISH find it satisfactory and recommend that it be accepted.

_____

Stephen P. Ficklin, Ph.D., Chair

_____

Loren A. Honaas, Ph.D.

_____

Michael M. Neff, Ph.D

_____

Zhiwu Zhang, Ph.D.

ACKNOWLEDGMENT

First and foremost, I acknowledge my advisor Stephen Ficklin. I thank you for your mentorship and friendship over these past years. I appreciate your patience and guidance in shaping me into the scientist I have become today.

I acknowledge my committee members, Loren Honaas, Michael Neff, and Zhiwu Zhang for their patience and support in helping me develop as a scientist.

I acknowledge my colleagues and friends: Matt McGowan, Huiting Zhang, Itsuhiro Ko, Andrei Smertenko, Václav Svoboda, Rachael DeTar, Skylar Johnson, Yunus Sahin, Kathleen Hickey, Joel Sowders, Matt Marcec, Sharol Marcec, Josh Polito, Alyssa Parish, and Se Eun Jung. Board games, floating, hiking, mushroom hunting, skiing, campfires, and coffee shops with you have made this Ph.D. possible.

DEVELOPMENT OF COMPUTATIONAL MODELS, BIOMARKERS,

AND TOOLS FOR POSTHARVEST TRAITS IN

MALUS DOMESTICA FRUIT

Abstract

by John Anthony Hadish, Ph.D.
Washington State University
December 2023

Chair: Stephen P. Ficklin

Understanding *Malus domestica* (apple) postharvest biology under typical storage conditions is important for ensuring that a high-quality product reaches consumers and for food waste reduction. Despite this, minimal research has been performed investigating the molecular mechanisms at work during storage. The following dissertation uses transcriptomic data and modeling techniques to investigate this biology. Chapter one is a brief literature review on modeling techniques and apple postharvest biology. Chapter two investigates how the core apple hypoxia response differs from other plants and how different postharvest treatments impact apple biology over long-term storage. Chapter three investigates how we can use machine learning models to develop transcriptomic biomarkers for predicting phenotypic traits in apples.

Chapter four investigates currently open questions about data quantity and normalization techniques for modeling transcriptomic traits. Finally, chapter 5 reflects on the lessons learned from this research and on my experiences as a Ph.D. student. This research uncovers potential neo-functionalizations of genes, transcriptomic biomarkers, and a better understanding of modeling using transcriptomic data.

TABLE OF CONTENTS

LIST OF TABLES

Page

CHAPTER FOUR

LIST OF FIGURES

Page

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

APPENDIX ONE

Dedication

I dedicate this dissertation to my family

My mother Ruthann Hadish

My father Gregg Hadish

My brother Paul Hadish

I love you

CHAPTER ONE:

TRANSCRIPTOMIC MODELING AND ITS APPLICATION IN POSTHARVEST POME FRUIT

The Transcriptomics Era

The three-dimensional helix structure of DNA was discovered in 1953 by James Watson and Francis Crick using X-ray crystallography data from Rosland Frankin and Maurice Wilkins. These words are repeated in biology textbooks worldwide, paired alongside the famous “Photo 51,” showing the fuzzy X-ray crystallography image critical for this discovery. This glimpse at the building block of life is considered one of the most significant achievements in biological research and has captured the imagination of budding biologists ever since.

What is often not mentioned in these introductory textbooks is that it would be another decade before researchers could read the code contained within these mysterious molecules. In 1965--after 2½ years of effort--Robert Holley and colleagues released the sequence of the 77 nucleotide alanine tRNA from *Saccharomyces cerevisiae* (Holley et al., 1965). This marked the first sequenced nucleotides, making RNA the first nucleic acid molecule to be sequenced--years before the first DNA molecule (Heather & Chain, 2016; Sanger et al., 1977; Wu, 1970; Xue et al., 2016). This sequencing effort is arguably the beginning of the field we now call transcriptomics.

The term “transcriptomics” first gained popularity in the 1990s to describe the

entire population of coding and non-coding RNA within a sample (Lowe et al., 2017). This RNA population changes in size and content, but typical eukaryotic cells have a ratio of around 1:2 for RNA:DNA with only 1-5% of the RNA being protein-coding (mRNA) (Palazzo & Lee, 2015; Shinohara et al., 2019). Unlike DNA, this pool of RNA is highly dynamic and constantly changing, with new RNA being synthesized and old RNA being recycled. This dance of RNA synthesis and recycling is one of the ways how cells--and by extension organisms--respond to their environment. Cells upregulate different RNA molecules to grow, divide, specialize, and respond to outside pressures.

Modern transcriptomics seeks to identify and quantify these levels of RNA. These molecules are only one step away from DNA, making RNA the most basal phenotype of any organism. Both the genotype and the environment impact the amount and species of RNA within the cell. Understanding this “RNA phenotype” is crucial for gaining a deeper understanding of biology, but observing them has not always been simple.

The decades following the initial sequencing of alanine tRNA resulted in a number of improvements in RNA identification and quantification. Methods were established that could effectively identify and/or quantify individual transcripts. These included techniques such as Expressed Sequence Tags (ESTs) (Parkinson & Blaxter, 2009) Northern Blotting (He & Green, 2013), and RT-qPCR (Adams, 2020). These low throughput methods have been supplemented--and in some cases replaced-- by high-throughput technologies capable of measuring the entire transcriptome. The two dominant high-throughput technologies used today are hybridization-based microarrays developed in the mid-1990s, and sequencing-based RNA-seq developed in the 2000s.

Microarrays continue to be used in niche experimental circumstances and healthcare (Negi et al., 2021) but the decreasing cost of sequencing has made RNA-seq an attractive option and is largely replacing microarray experiments (Lowe et al., 2017). RNA-seq also is advantageous because, unlike microarrays, knowledge of the genome is not required. This makes RNA-seq especially useful in non-model organisms with poor or no genomes available. RNA-seq also allows for the discovery of novel transcripts and the differentiation of transcript isoforms (Lowe et al., 2017).

Transcriptomic Analysis Techniques

Exponential advancement in high-throughput sequencing has made probing the transcriptome a trivial task when compared to two decades ago. We are now in an era where the bottleneck is not gathering data but rather interpreting data. Additional processing of transcriptomic data is required for meaningful interpretation due to the massive size of the data generated. If the letters of all the nucleotides from a single RNA-seq sample were printed on US standard 8.5 by 11-inch paper using the same formatting requirements used by this dissertation, the resulting paper stack would be approximately the height of the Washington Monument*. Various tools and techniques have been established to help process this data into interpretable results, with several novel techniques currently in development. This brief review will concentrate on RNA-seq data interpretation, but many of the techniques described here can be applied to microarray measurements and other high-throughput technologies which produce count data such as proteomics, metabolomics, and lipidomics.

**Assuming a sample with a moderate sequencing depth of 20 million reads and 150 bp reads. Formatting requirements for WSU dissertations result in 1800 characters in 12pt font with 1-inch margins double-spaced. Assuming a ream of paper (500 sheets) is 5.5 cm thick, the resulting paper stack would be 166m tall, just 3m shy of the Washington Monument's 169m aluminum capstone. This back-of-the-napkin calculation only uses nucleotides and does not include sequence quality information. If sequence quality information is included, our precarious stack of paper would more than double in height.*

Pre-processing of Transcriptomic Data

The first step in processing RNA-seq data is to quantify the number of reads in each sample. A "read" is a complete or partial sequence of a single RNA molecule produced by a sequencing machine. A single RNA-seq sample consists of millions of these reads, with the number of the reads in a sample being referred to as its "sequencing depth". These reads are quantified by aligning (or pseudo-aligning (Bray et al., 2016; Patro et al., 2017)) them to a genome sequence or previously assembled transcriptome sequence to identify from which feature (gene or other transcribed portion of the genome) they were transcribed. The number of reads aligned to each feature is then counted to obtain a feature count. Several tools have been developed to perform this read quantification, with popular tools including STAR (Dobin et al., 2013), Hisat2 (Kim et al., 2015), Kallisto (Bray et al., 2016), and Salmon (Patro et al., 2017). Results from these tools can be combined to produce a gene expression matrix (GEM) which is an *n x m* matrix of *n* genes and *m* samples where each value in the matrix represents the

expression of a single feature in a single sample (Hadish et al., 2022).

In addition to the quantification tools listed above, a number of monitoring tools such as FastQC and multiQC (Andrews, 2010; Ewels et al., 2016) and helper analysis tools such as Aspera for data retrieval (Ncbi, 2014), Trimmomatic for quality trimming (Bolger et al., 2014) and Stringtie for read counting (Pertea et al., 2015) have been developed. These are important for ensuring that the RNA-seq data is checked, cleaned, and of high quality to prevent potential errors that may impact downstream analysis. These auxiliary tools can be combined with quantification tools into workflows that ease the computational burden of RNA-seq analysis and make the handling of large datasets manageable. The workflow *GEMmaker* (Hadish et al., 2022) is included as Appendix A1 in this dissertation and represents a high throughput workflow for RNA-seq processing. It integrates popular quantification and analysis tools into an easy-to-use workflow capable of processing thousands of publicly available RNA-seq experiments from NCBI (NCBI Resource Coordinators, 2016). *GEMmaker* is the workflow used to process all RNA-seq data described within the subsequent chapters. This author is the first author of the GEMmaker paper and provided significant contributions to the development of GEMmaker.

After GEM creation, normalization of the quantified RNA-seq data is required for downstream applications due to varying sequencing depths, gene lengths, and library selection methods (Conesa et al., 2016). Unnormalized datasets report the number of reads per feature identified within the sample which is the number of reads the

sequencing machine sequenced for each feature. This unnormalized data is adequate if only considering features within a single sample, but is not useful for comparing samples with different sequencing depths. Transcripts per Kilobase Million (TPM) and Read Per Kilobase Million (RPKM) are very similar and consider the sequencing depth of each sample as well as the length of each transcript/gene to normalize each sample to one million counts per sample. These two techniques are suitable for analysis between samples of the same sample group (i.e. usually a group within the same experiment) (S. Zhao et al., 2020; Y. Zhao et al., 2021). The Median of Ratios Normalization (MRN) (implemented in the DESeq2 package) (Love et al., 2014) and Trimmed Mean of M values (TMM) (implemented in the EdgeR package) (Robinson et al., 2010; Robinson & Oshlack, 2010) are typically used for differential expression and comparison between samples. Optimal normalization methods are well-defined for some transcriptomic analysis techniques (i.e., differential expression) but are still open areas of research for others (i.e., machine learning methods). These pre-processing steps are an often overlooked first step before performing transcriptomic analysis techniques, which can significantly impact the outcome if performed incorrectly (Conesa et al., 2016).

Differential Expression Models

Differential Gene Expression (DEG) is arguably the most common transcriptomic analysis technique. DEG compares a control condition to a treatment condition to identify which genes are up and downregulated. While simple in concept, DE is more

complicated than a Student's T-Test. Comparing thousands of genes using only a few samples means that DEG is a high-dimensionality problem, which can lead to accidental false positives. This high dimensionality issue can be mediated by sharing information across genes. Genes with similar expressions are assumed to have similar dispersion (a similar metric to variance in gene expression), which allows us to effectively increase our sample size without paying for additional replicates (Love et al., 2014). This makes DE a powerful tool that can be used on relatively low replicates (as few as three) (Schurch et al., 2016)). Larger sample sizes (as many as 12 or more) are needed for comparing samples with high gene dispersions (i.e. situations where there is large biological or technical differences in gene expression between samples) (Ching et al., 2014; Schurch et al., 2016).

Variants of simple DEG make it possible to perform more complicated analyses than just "control vs. treatment." Controlling the linear model coefficients of experimental variables via a design matrix (a matrix of 0's and 1's indicating if an experimental variable should be considered), can allow for the testing of multiple comparisons at once, as well as potential interaction terms between variables. The design matrix allows for the analysis of time series experiments, allows genotypic effects to be parsed out from treatment-specific effects, and allows for the identification of treatments that have additive effects (Ritchie et al., 2015).

Popular tools for DEG include edgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014), and limma (Ritchie et al., 2015). Variations on these ideas first fit splines to time series data which allow for more succinct analysis. These spline-based tools

include maSigPro (Conesa et al., 2006; Nueda et al., 2014) and splineTC (Michna et al., 2016; Spies et al., 2017).

Predictive Models

Predictive models have become popular across science and industry domains with the advent of large datasets made possible by the internet and our modern society. These models can become rather complicated--making use of many different methods and datatypes--but the underlying goals behind all of them is the same: to predict a dependent variable using a set of independent variables and to identify which of the independent variables is most important in the prediction. The easiest way to explain these goals of predictive modeling is through a practical example. For this, I will use the famous “Fisher’s Iris” dataset (Fisher, 1936). This is a small dataset that data scientists often use to showcase a new technique or explain a complicated process. It consists of 50 measurements of three species of Iris (*Iris setosa*, *Iris virginica* and *Iris versicolor*). Each Iris has four measurements: the length and width of both the sepals and the petals.

The first goal of predictive modeling is (as its name suggests) to predict some value. This value is referred to as the “dependent” variable. In the Iris dataset the dependent variable we wish to predict is the species of iris. This is therefore a *Classification* problem where 2 or more categorical measurements (i.e. the species) are the dependent variable. This is in contrast to *Regression* problems, where the dependent variable is continuous (for example, if we were trying to predict the height of

the iris) (Brownlee, 2019). The second part of the dataset consists of 4 different measurements, which we refer to as the “independent” variables. The predictive model makes associations between these independent variables to see if they can be used to predict the dependent variable. In our case this means predicting the species of Iris based only on measurements of its floral parts.

This brings us to the second goal of predictive modeling, which is “feature selection”. When a model is using the independent variables to predict the dependent variable, some of the independent variables are more valuable than others. The model can rank these independent variables based on how important they are to predict the dependent variable. We refer to this as “feature selection”. In the case of the iris data, the model may decide that the length of the petal is the most important for distinguishing the species while the sepal width appears to be random. These two goals--prediction and feature selection--are the main principles that connect different types of predictive models.

In real-life predictive modeling applications, datasets tend to be larger than the Iris dataset. Datasets are available for stock performance, housing prices, shopping preference, and disease risk to name a few (Khalilia et al., 2011). The methods used to create models and perform feature selection using this data include methods such as Random Forest (Breiman, 2001), Elastic Net (Zou & Hastie, 2005), gradient boosting(Friedman, 2002), XGBoost (Chen & Guestrin, 2016), Naive Bayes (Rish, 2001), k-nearest neighbors (Fix & Hodges, 1951) and Boruta (Kursa & Rudnicki, 2010).

In a transcriptomic biological context, predictive models can be used to associate large-scale transcriptomic data with a phenotypic variable. This allows researchers to

make predictions about the phenotype on unlabeled datasets and to identify which gene features contribute to the phenotype at the level of the transcriptome. The majority of these types of studies have been performed in human medicine for predicting disease state and cancer type (Feng et al., 2019; Smith et al., 2020; Supplitt et al., 2021) but recent studies have expanded into agricultural settings. Experiments have been performed in *Zea mays* to model flowering time (Azodi et al., 2020), in potato tubers to model tuber quality traits (Acharjee et al., 2016) and in pome fruit (discussed in the following sections).

However, the large-scale adoption of trait association modeling within biological datasets has been slow, mainly due to limited sample size and datasets with many independent variables. Non-biological measurements--such as the price of a house or the cost of a stock--can consist of thousands or even millions of samples. Each of these samples consists of only a few independent variable measurements. This is in comparison to transcriptomic experiments where large experiments are currently only a few hundred samples and the number of measured independent variables (genes) is very large. This type of dataset--few samples and many measurements--is referred to as “wide”. This is compared to “long” datasets which consist of many samples each with only a few measurements. Wide datasets can be an issue since it is difficult for some models to distinguish between important and unimportant variables when few samples are present. Some methods perform better than others with wide datasets, and a discussion and practical application of this is discussed in chapter 3 of this dissertation.

Despite issues associated with wide datasets, transcriptomic trait association models are becoming more prevalent. RNA-seq datasets paired with physiological

measurements allow researchers to create associative models where gene expression levels (independent variable) are used to predict phenotypic traits (dependent variable). This allows for the construction of predictive models (i.e. disease vs no-disease) and for feature selection which allows for the identification of genes associated with these predictions (i.e gene X expression over value y means diseased).

An important model type that is used in the research in the following chapters is random forests (Breiman, 2001; Ho, 1995). Random forests are a type of model that can be used for both regression (continuous dependent variables) and classification (categorical dependent variable) type problems. They are constructed by making a large number of decision trees (Fürnkranz, 2010), each made from a random sub-sample of features and samples of the complete dataset. Once created, data can be sent through this forest of random decision trees, and a prediction will be output. For classification problems, a vote of all of the decision trees determines the predicted label, whereas for regression problems a mean of their decisions determines the predicted value. Feature Importance is calculated in random forest models by taking an average of the feature importance of each decision tree.

Research presented in chapter 3 of this dissertation investigates using modeling to predict important quality traits in genetically identical apples in a postharvest environment. After model creation, feature selection is used to identify the most important genes within the model which are further verified within a separate year via preliminary qPCR analysis. Another large open question within transcriptomic modeling is how many samples are required to get reliable and reproducible results in transcriptomic data modeling experiments. This question remains open largely due to

the lack of massive transcriptomic datasets required for addressing these questions. Chapter 4 of this dissertation addresses this issue by creating a massive Arabidopsis RNA-seq dataset from publicly available data (53260 samples) and using it to model dependent variables such as age and tissue type.

Networks

Networks are one of the core tools systems biologists use to look at the relationships between data. Networks are extremely versatile and consist of two parts: nodes and edges (Barabási, 2016). In a biological network nodes represent a single biological entity--such as genes, metabolites, proteins, organisms, and ecosystems--and edges represent the interactions between these entities. Examples of some biological networks include gene-regulatory interactions (transcription factors and their targets) (Harrington et al., 2020), protein-protein interactions (Schwikowski et al., 2000), and predator-prey interactions (Bruder et al., 2019).

Systems biologists use a number of different methods for constructing networks which can be loosely classified into two types: bottom-up and top-down. Bottom-up network construction seeks to gather and curate current biological knowledge from the literature into meaningful summaries of the current state of knowledge. This method concentrate on the individual components and their local interactions. These interactions are then used to build a complex network that describes the system as a whole (Pezzulo & Levin, 2016). These detailed and curated networks can be used for modeling how an organism, organ, or pathway will respond to a change in environment through flux balance analysis (Orth et al., 2010). In contrast, top-down approaches start

with data measuring the entire system at a given instance and reconstructing the individual interactions using appropriate statistical and association analysis tools (Shahzad & Loor, 2012). The data used to create top-down networks is often omics-level data such as transcriptomic, metabolomic, or proteomic. The data is gathered over a number of different conditions, time points, and/or species/varieties that encompass the desired scope of the experiment.

Top-down transcriptomic networks are a way to identify potentially co-functional or interacting genes under biological conditions. They represent how an organism is responding to its environment via the regulation of its mRNA. This can provide hints to downstream phenotypes such as the metabolome and proteome These networks can provide information about which genes are co-expressed and which genes are regulating others.

The first top-down transcriptomic network construction technique is co-expression. Co-expression networks are based on the principle of "guilt-by-association" which assumes that genes with similar expression across conditions are co-functional (Gillis & Pavlidis, 2012; Wolfe et al., 2005).
A co-expression network is created by gathering transcriptomic data across several treatments and then performing a pairwise correlation between every gene in this dataset. Correlation is usually done using either Pearson's or Spearman's distance correlation (Hou et al., 2022). Correlation generates a complete network where every value is connected to every other value. This is called a "complete network". This complete network must then be thresholded to create a reduced network where only high-confidence edges remain. Thresholding of edges can be done using a hard

threshold (i.e. picking a correlation value of .9 or more) or via network topology properties (Barabási, 2016). After thresholding, modules of highly interconnected genes are identified. These modules consist of genes that are co-expressed and possibly co-functional. There are a number of tools available for researchers to create trancriptomic co-expression networks (Burns et al., 2022; Faith et al., 2007; Ficklin & Feltus, 2013; Langfelder & Horvath, 2008; Liang et al., 2015; Marwah et al., 2018; Meyer et al., 2007; Petereit et al., 2016). The most popular of these tools is “Weighted Gene Correlation Network Analysis” (WGCNA) (Langfelder & Horvath, 2008).

A second way to create a top-down transcriptomic network is to create a regulatory network. In addition to RNA-seq data, regulatory networks integrate additional information during network construction with the goal of identifying how transcription factors regulate other genes. This additional information can include information such as transcription factors, binding motifs, chromatin immunoprecipitation sequencing data (ChIP-seq), and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) data (En Chai et al., 2014; Tu et al., 2020). The goal of adding this additional information is to identify how transcription factors are interacting with the genome to regulate gene expression during a specific condition.

Several techniques have been developed to create these networks, including regression-based (Haury et al., 2012), mutual information (Margolin et al., 2006) bayesian (Sanchez-Castillo et al., 2018) and machine learning-based methods (Huynh-Thu et al., 2010). Of these, random-forest machine-learning style networks first described in the GENIE3 paper (Huynh-Thu et al., 2010; Huynh-Thu & Geurts, 2018) are prevalent due to their high performance and relatively low computational demands

(Marbach et al., 2012). This style of regulatory network construction uses a list of transcription factors and a GEM as input. It then uses random forest regression feature selection (Breiman, 2001) to identify which transcription factors best describe gene expression of each gene. This results in a directed network (which transcription factor regulates which gene) that hypothesizes how genes are regulated in the experimental dataset.

Regulatory networks are used in chapter two of this dissertation to identify putative transcription factors which are predictive of the hypoxia response within detached apple fruits. The response is different than in model organisms, and these regulatory networks provide a method for predicting possible causes which can be addressed in future research.

Pome Fruit Postharvest Modeling

Pome Fruit Basic Biology and Significance

Apples and other pome fruit (Roaceae; tribe Maleae; subtribe Malinae) are part of a lineage that is the result of a genome duplication event around 38 to 42 million years ago which transitioned the 9 ancestral chromosomes to the 17 chromosomes we see today (Li et al., 2019; Velasco et al., 2010). Malinae is the only tribe within Rosaceae with an accessory fruit that develops from multiple fused carpels. (Schulze-Menz, 1964; Sun et al., 2018). This accessory fruit is referred to as a “pome” and can vary from small and hard to large and fleshy. A number of these “pome fruit” bearing species are agriculturally important such as Apple (*Malus domestica*), Pear (*Pyrus communis*), and Quince (*Cydonia oblonga*).

Pome fruit plays a significant role in the economy of Washington State USA. Apples alone add over $2.1 billion dollars to the economy in 2021, representing 21% of the state's agricultural value. Pears add an additional 150 million dollars (USDA, National Agricultural Statistics Service, 2022). The high value and desirability of these fruits make it essential that proper agricultural methods are used to decrease losses and ensure that production can meet demands in an environmentally conscious manner.

Modern agricultural practices have significantly increased the ability of producers and packing houses' ability to store pome fruit--especially apples--for an extended period of time. Significant technologies adopted by the US apple industry include refrigeration in the early 1900s, controlled atmosphere in the 1930s -1950s (DeLong et al., 1999; Sigler, 2011), the scald-preventing antioxidant diphenylamine (DPA) in the 1950s (Dias et al., 2020), the ethylene receptor inhibitor 1-Methylcyclopropene (1-MCP) in the early 2000s (DeEll et al., 2002), and dynamic Controlled Atmosphere in the 2000s (Mditshwa et al., 2018). Despite these numerous modern techniques, postharvest fruit loss continues to be an issue for the industry. Packing houses annually cull millions of tons of apples that have gone bad due to physiological disorders and diseases.

Previous Modeling Experiments and Gene Identification

One way to reduce apple culling is by ensuring that apples are sent to market prior to the development of disease or loss in quality. To meet this goal, apple fruits are monitored before and after harvest so that producers can anticipate and mitigate possible fruit quality issues. This monitoring is currently done through physiological

measurements of the fruits such as starch clearing (Blanpied & Silsby, 1992), fruit firmness (Harker et al., 1996), peel color (Hamza & Chtourou, 2018), and acid and sugar content (Goffings, 1993). These measurements do a reasonable job of predicting future outcomes, but differences in orchards, environmental conditions, and postharvest handling mean that fruit with the same measurements may have additional unaccounted-for variability not captured by physiological measurements. This unaccounted-for variability results in the culling of billions of kilograms of apple fruit from the fresh fruit market annually (USDA, National Agricultural Statistics Service, 2022). Another major issue with measuring physiological traits is that they are slow to develop, lagging behind the event which induced them. Some of these traits--such as soft-scald--do not appear until months after the stress that induced them.

One promising way to account for this variability and get quicker information is to directly measure the apple fruit's transcriptome. The transcriptome directly responds to the environment--sometimes in a matter of seconds--whereas physiological traits can take months to develop. This rapid change to nuance information (such as temperature, sun exposure, soil moisture, and nutrient availability) is valuable for determining physiological traits which will not develop until months in the future. Additionally, Apples are a particularly good subject for investigating transcriptomic biomarkers because they are clonally propagated. Since all individuals are identical, genetic variation in transcriptome response is removed. This means that changes to the transcriptome are due solely to the environmental conditions the fruits are experiencing on the tree and after harvest. These features of the apple transcriptome--rapid response to the

environment, recording nuanced information, and variation only due to the environment--make it an ideal candidate for monitoring fruit quality.

There has already been some investigation into using transcriptomic data to monitor postharvest pome fruit. Studies have investigated differentiating between four harvest time points in ‘Royal Gala’ apples (Favre et al., 2022), soft scald in ‘Honeycrisp’ apple fruits (Leisso et al., 2016), external CO_2 injury in ‘Empire’ apple fruits (Gapper et al., 2013), internal browning in ‘Braeburn’ apples (Hatoum et al., 2016; Mellidou et al., 2014), and superficial scald in ‘Granny Smith’ (Farneti et al., 2015). Additionally, the apple industry has also made use of transcriptomic markers, with the company AgroFresh marketing a test for predicting soft scald and bitter pit risk in ‘Honeycrisp’ apples for the 2019 harvest (Karst, 2019; Prengaman, 2019). It was marketed as a way to determine if lots of organic apples were high risk, which would allow for treatment with conventional chemicals to reduce loss.

Conclusion

Today, most basic molecular plant science work is performed in model organism systems such as *Arabidopsis thaliana, Brachypodium distachyon, Populus trichocarpa, Medicago truncatula* and *Nicotiana benthamiana* (Cesarino et al., 2020) and applied to major annual row crops such as *Zea mays, Oryza* spp., and *Triticum aestivum.* None of these are closely related to the longlived trees and shrubs of the pome fruit, with the most major difference from an economical--and possibly biological--standpoint being the fruit. A well-studied organism with the closest fruit biology is the tomato (*Solanum lycopersicum*) which, like the pome fruits is climacteric (ripening in response to

ethylene). However, this similarity is due to convergent evolution, with several genome duplication events separating the two (Lü et al., 2018).

While it is possible to transfer information from the above model organisms to pome fruit, it is also necessary to directly investigate apple biology to ensure that physiological differences are accounted for. To better understand pome fruit postharvest biology the following chapters use transcriptomics and modern top-down modeling experiments to probe apple biology. Chapter two concentrates on the response to hypoxia in postharvest apples. Apple fruit can be kept alive in hypoxic environments for months, slowing their metabolism and resulting in a number of other changes. Most other plant hypoxia experiments have been performed in either the roots or shoots of Arabidopsis plants, which can only survive for a few hours (Ellis et al., 1999). Apple fruit, therefore, offers an interesting model to investigate long-term hypoxia responses. Chapter three concentrates on modeling apple firmness, a trait important to the industry. Our technique selects genes from transcriptomic experimental data rather than previous knowledge and performs further verification using qPCR. This knowledge-independent gene selection is an essential step in understanding complex traits using top-down transcriptomics for non-model organisms. These techniques provide a way in which further hypotheses can be tested using traditional bottom-up methods. The end of chapter three has a brief section discussing how these same techniques can be applied to modeling apple maturity in postharvest apples in a separate dataset. Chapter four moves away from apple to address issues of transcriptomic modeling using a massive (53260 sample) Arabidopsis dataset created from publicly available data. While modeling using transcriptomic data is becoming more popular, there has not been much

investigation into the size of dataset required for accurate predictions. This massive dataset allows us to test hypotheses about novel techniques and investigate theoretical sample size requirements. This can be directly applied to future research in non-model organisms such as apple for determining necessary experimental design size.

REFERENCES

Acharjee, A., Kloosterman, B., Visser, R. G. F., & Maliepaard, C. (2016). Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics*, *17 Suppl 5*(Suppl 5), 180.

Adams, G. (2020). A beginner's guide to RT-PCR, qPCR and RT-qPCR. *The Biochemist*, *42*(3), 48–53.

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data* (Version 0.11.9) [Computer software].

Azodi, C. B., Pardo, J., VanBuren, R., de Los Campos, G., & Shiu, S.-H. (2020). Transcriptome-Based Prediction of Complex Traits in Maize. *The Plant Cell*, *32*(1), 139–151.

Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.

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

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Brownlee, J. (2019, May 22). *Difference Between Classification and Regression in Machine Learning*. Machine Learning Mastery. https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/

Bruder, A., Frainer, A., Rota, T., & Primicerio, R. (2019). The Importance of Ecological Networks in Multiple-Stressor Research and Management. *Frontiers of*

Environmental Science & Engineering in China, *7*. https://doi.org/10.3389/fenvs.2019.00059

Burns, J. J. R., Shealy, B. T., Greer, M. S., Hadish, J. A., McGowan, M. T., Biggs, T., Smith, M. C., Feltus, F. A., & Ficklin, S. P. (2022). Addressing noise in co-expression network construction. *Briefings in Bioinformatics*, *23*(1). https://doi.org/10.1093/bib/bbab495

Cesarino, I., Dello Ioio, R., Kirschner, G. K., Ogden, M. S., Picard, K. L., Rast-Somssich, M. I., & Somssich, M. (2020). Plant science's next top models. *Annals of Botany*, *126*(1), 1–23.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1603.02754

Ching, T., Huang, S., & Garmire, L. X. (2014). Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* , *20*(11), 1684–1696.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13.

Conesa, A., Nueda, M. J., Ferrer, A., & Talón, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* , *22*(9), 1096–1102.

DeEll, J. R., Murr, D. P., Porteous, M. D., & Vasantha Rupasinghe, H. P. (2002). Influence of temperature and duration of 1-methylcyclopropene (1-MCP) treatment on apple quality. *Postharvest Biology and Technology*, *24*(3), 349–353.

DeLong, J. M., Prange, R. K., Harrison, P. A., Andrew Schofield, R., & DeEll, J. R.

(1999). Using the Streif Index as a Final Harvest Window for Controlled-atmosphere Storage of Apples. In *HortScience* (Vol. 34, Issue 7, pp. 1251–1255). https://doi.org/10.21273/hortsci.34.7.1251

Dias, C., L Amaro, A., C Salvador, Â., Silvestre, A. J. D., Rocha, S. M., Isidoro, N., & Pintado, M. (2020). Strategies to Preserve Postharvest Quality of Horticultural Crops and Superficial Scald Control: From Diphenylamine Antioxidant Usage to More Recent Approaches. *Antioxidants (Basel, Switzerland)*, *9*(4). https://doi.org/10.3390/antiox9040356

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21.

Ellis, M. H., Dennis, E. S., & Peacock, W. J. (1999). Arabidopsis roots and shoots have different mechanisms for hypoxic stress tolerance. *Plant Physiology*, *119*(1), 57–64.

En Chai, L., Kuan Loh, S., Thing Low, S., Saberi Mohamad, M., Deris, S., & Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*, *48*, 55–65.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* , *32*(19), 3047–3048.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., & Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, *5*(1), e8.

Farneti, B., Busatto, N., Khomenko, I., Cappellin, L., Gutierrez, S., Spinelli, F., Velasco, R., Biasioli, F., Costa, G., & Costa, F. (2015). Untargeted metabolomics investigation of volatile compounds involved in the development of apple superficial scald by PTR-ToF–MS. *Metabolomics: Official Journal of the Metabolomic Society*, *11*(2), 341–349.

Favre, L., Hunter, D. A., O'Donoghue, E. M., Erridge, Z. A., Napier, N. J., Somerfield, S. D., Hunt, M., McGhie, T. K., Cooney, J. M., Saei, A., Chen, R. K. Y., McKenzie, M. J., Brewster, D., Martin, H., Punter, M., Carr, B., Tattersall, A., Johnston, J. W., Gibon, Y., … Brummell, D. A. (2022). Integrated multi-omic analysis of fruit maturity identifies biomarkers with drastic abundance shifts spanning the harvest period in "Royal Gala" apple. *Postharvest Biology and Technology*, *193*, 112059.

Feng, X., Zhang, R., Liu, M., Liu, Q., Li, F., Yan, Z., & Zhou, F. (2019). An accurate regression of developmental stages for breast cancer based on transcriptomic biomarkers. *Biomarkers in Medicine*, *13*(1), 5–15.

Ficklin, S. P., & Feltus, F. A. (2013). A Systems-Genetics Approach and Data Mining Tool to Assist in the Discovery of Genes Underlying Complex Traits in Oryza sativa. *PloS One*, *8*(7), e68551.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Fix, E., & Hodges, J. L. (1951). *Discriminatory Analysis Nonparametric Discrimination Consistency Properties*. https://apps.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378.

Fürnkranz, J. (2010). Decision Tree. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 263–267). Springer US.

Gapper, N. E., Rudell, D. R., Giovannoni, J. J., & Watkins, C. B. (2013). Biomarker development for external CO2 injury prediction in apples through exploration of both transcriptome and DNA methylation changes. *AoB Plants*, *5*, 1–9.

Gillis, J., & Pavlidis, P. (2012). “Guilt by Association” Is the Exception Rather Than the Rule in Gene Networks. *PLoS Computational Biology*, *8*(3), e1002444.

Hadish, J. A., Biggs, T. D., Shealy, B. T., Bender, M. R., McKnight, C. B., Wytko, C., Smith, M. C., Feltus, F. A., Honaas, L., & Ficklin, S. P. (2022). GEMmaker: process massive RNA-seq datasets on heterogeneous computational infrastructure. *BMC Bioinformatics*, *23*(1), 1–11.

Harrington, S. A., Backhaus, A. E., Singh, A., Hassani-Pak, K., & Uauy, C. (2020). The Wheat GENIE3 Network Provides Biologically-Relevant Information in Polyploid Wheat. *G3* , *10*(10), 3675–3686.

Hatoum, D., Hertog, M. L. A. T. M., Geeraerd, A. H., & Nicolai, B. M. (2016). Effect of browning related pre- and postharvest factors on the “Braeburn” apple metabolome during CA storage. *Postharvest Biology and Technology*, *111*, 106–116.

Haury, A.-C., Mordelet, F., Vera-Licona, P., & Vert, J.-P. (2012). TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Systems Biology*, *6*, 145.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8.

He, S. L., & Green, R. (2013). Northern blotting. *Methods in Enzymology*, *530*, 75–87.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., & Zamir, A. (1965). Structure of a Ribonucleic Acid. *American Association for the Advancement of Science*. https://www.jstor.org/stable/1715055

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282 vol.1.

Hou, J., Ye, X., Feng, W., Zhang, Q., Han, Y., Liu, Y., Li, Y., & Wei, Y. (2022). Distance correlation application to gene co-expression network analysis. *BMC Bioinformatics*, *23*(1), 81.

Huynh-Thu, V. A., & Geurts, P. (2018). dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. *Scientific Reports*, *8*(1), 3384.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS One*, *5*(9). https://doi.org/10.1371/journal.pone.0012776

Karst, T. (2019, October 7). *AgroFresh Solutions helps foresee bitter pit in Honeycrisp*. The Packer. https://www.thepacker.com/news/packer-tech/agrofresh-solutions-helps-foresee-bitter-pit-honeycrisp

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, *11*, 51.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package.

Journal of Statistical Software, *36*, 1–13.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559.

Leisso, R. S., Gapper, N. E., Mattheis, J. P., Sullivan, N. L., Watkins, C. B., Giovannoni, J. J., Schaffer, R. J., Johnston, J. W., Hanrahan, I., Hertog, M. L. A. T. M., Nicolaï, B. M., & Rudell, D. R. (2016). Gene expression and metabolism preceding soft scald, a chilling injury of “Honeycrisp” apple fruit. *BMC Genomics*, *17*(1), 1–23.

Liang, M., Zhang, F., Jin, G., & Zhu, J. (2015). FastGCN: a GPU accelerated tool for fast gene co-expression networks. *PloS One*, *10*(1), e0116776.

Li, H., Huang, C.-H., & Ma, H. (2019). Whole-Genome Duplications in Pear and Apple. In S. S. Korban (Ed.), *The Pear Genome* (pp. 279–299). Springer International Publishing.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, *13*(5), e1005457.

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

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., DREAM5 Consortium, Kellis, M., Collins, J. J., & Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, *9*(8),

796–804.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., & Califano, A. (2006). ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, *7*(Suppl 1), S7.

Marwah, V. S., Kinaret, P. A. S., Serra, A., Scala, G., Lauerma, A., Fortino, V., & Greco, D. (2018). INfORM: Inference of NetwOrk Response Modules. *Bioinformatics* , *34*(12), 2136–2138.

Mditshwa, A., Fawole, O. A., & Opara, U. L. (2018). Recent developments on dynamic controlled atmosphere storage of apples—A review. *Food Packaging and Shelf Life*, *16*, 59–68.

Mellidou, I., Buts, K., Hatoum, D., Ho, Q. T., Johnston, J. W., Watkins, C. B., Schaffer, R. J., Gapper, N. E., Giovannoni, J. J., Rudell, D. R., Hertog, M. L. A. T. M., & Nicolai, B. M. (2014). Transcriptomic events associated with internal browning of apple during postharvest storage. *BMC Plant Biology*, *14*(1). https://doi.org/10.1186/s12870-014-0328-x

Meyer, P. E., Kontos, K., Lafitte, F., & Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP Journal on Bioinformatics & Systems Biology*, *2007*(1), 79879.

Michna, A., Braselmann, H., Selmansberger, M., Dietz, A., Hess, J., Gomolka, M., Hornhardt, S., Bluethgen, N., Zitzelsberger, H., & Unger, K. (2016). Natural cubic spline regression modeling followed by dynamic network reconstruction for the identification of radiation-sensitivity gene association networks from time-course

transcriptome data. *PloS One*, *11*(8), e0160791.

Ncbi. (2014). *SRA Handbook [Internet] - Aspera Transfer Guide*. https://www.ncbi.nlm.nih.gov/books/NBK242625/

NCBI Resource Coordinators. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *44*(D1), D7–D19.

Negi, A., Shukla, A., Jaiswar, A., Shrinet, J., & Jasrotia, R. S. (2021). Applications and challenges of microarray and RNA-sequencing. In D. B. Singh & R. K. Pathak (Eds.), *Bioinformatics : Methods and Applications*. Elsevier Science & Technology.

Nueda, M. J., Tarazona, S., & Conesa, A. (2014). Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics* , *30*(18), 2598–2602.

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–248.

Palazzo, A. F., & Lee, E. S. (2015). Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics*, *6*, 2.

Parkinson, J., & Blaxter, M. (2009). Expressed Sequence Tags: An Overview. In J. Parkinson (Ed.), *Expressed Sequence Tags (ESTs)* (pp. 1–13). Humana Press.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295.

Petereit, J., Smith, S., Harris, F. C., Jr, & Schlauch, K. A. (2016). petal: Co-expression

network modelling in R. *BMC Systems Biology*, *10 Suppl 2*(Suppl 2), 51.

Pezzulo, G., & Levin, M. (2016). Top-down models in biology: explanation and control of complex living systems above the molecular level. *Journal of the Royal Society, Interface / the Royal Society*, *13*(124). https://doi.org/10.1098/rsif.2016.0555

Prengaman, K. (2019). *Revealing risks with RNA*. Good Fruit Grower. https://www.goodfruit.com/revealing-risks-with-rna/

Rish, I. (2001). *An empirical study of the naive Bayes classifier*. https://www.cc.gatech.edu/home/isbell/classes/reading/papers/Rish.pdf

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , *26*(1), 139–140.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25.

Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I. M., Carrion, M. C., & Huang, Y. (2018). A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics* , *34*(6), 964–970.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467.

Schulze-Menz, G. K. (1964). Rosaceae. In H. Melchior (Ed.), *A. Engler's Syllabus der*

Pflanzenfamilien (pp. 209–218).

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* , *22*(6), 839–851.

Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature Biotechnology*, *18*(12), 1257–1261.

Shahzad, K., & Loor, J. J. (2012). Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism. *Current Genomics*, *13*(5), 379–394.

Shinohara, K., Toné, S., Ejima, T., Ohigashi, T., & Ito, A. (2019). Quantitative Distribution of DNA, RNA, Histone and Proteins Other than Histone in Mammalian Cells, Nuclei and a Chromosome at High Resolution Observed by Scanning Transmission Soft X-Ray Microscopy (STXM). *Cells* , *8*(2). https://doi.org/10.3390/cells8020164

Sigler, D. (2011, July 28). CA storage has become staple of the fruit industry. *Fruit Growers News*. https://fruitgrowersnews.com/article/ca-storage-has-become-staple-of-the-fruit-industry/

Smith, A. M., Walsh, J. R., Long, J., Davis, C. B., Henstock, P., Hodge, M. R., Maciejewski, M., Mu, X. J., Ra, S., Zhao, S., Ziemek, D., & Fisher, C. K. (2020). Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics*, *21*(1), 119.

Spies, D., Renz, P. F., Beyer, T. A., & Ciaudo, C. (2017). Comparative analysis of

differential gene expression tools for RNA sequencing time course data. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbx115

Sun, J., Shi, S., Li, J., Yu, J., Wang, L., Yang, X., Guo, L., & Zhou, S. (2018). Phylogeny of Maleae (Rosaceae) Based on Multiple Chloroplast Regions: Implications to Genera Circumscription. *BioMed Research International*, *2018*, 7627191.

Supplitt, S., Karpinski, P., Sasiadek, M., & Laczmanska, I. (2021). Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. *International Journal of Molecular Sciences*, *22*(3). https://doi.org/10.3390/ijms22031422

Tu, X., Mejía-Guerra, M. K., Valdes Franco, J. A., Tzeng, D., Chu, P.-Y., Shen, W., Wei, Y., Dai, X., Li, P., Buckler, E. S., & Zhong, S. (2020). Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nature Communications*, *11*(1), 5089.

USDA, National Agricultural Statistics Service. (2022). *Noncitrus Fruits and Nuts 2021 Summary*. https://downloads.usda.library.cornell.edu/usda-esmis/files/zs25x846c/4q77gv96p/t722jd76c/ncit0522.pdf

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.

Wolfe, C. J., Kohane, I. S., & Butte, A. J. (2005). Systematic survey reveals general

applicability of “guilt-by-association”; within gene coexpression networks. *BMC Bioinformatics*, *6*(1), 227.

Wu, R. (1970). Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of bacteriophage lambda and 186 DNA. *Journal of Molecular Biology*, *51*(3), 501–521.

Xue, Y., Wang, Y., & Shen, H. (2016). Ray Wu, fifth business or father of DNA sequencing? *Protein & Cell*, *7*(7), 467–470.

Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* , *26*(8), 903–909.

Zhao, Y., Li, M.-C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H., & McShane, L. M. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine*, *19*(1), 269.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *67*(2), 301–320.

CHAPTER TWO:

CHARACTERIZATION OF THE TRANSCRIPTOMIC RESPONSE OF LONG-TERM HYPOXIA IN *MALUS DOMESTICA FRUIT*

Authors:

John A. Hadish[1,2], Heidi L. Hargarten[3], Huiting Zhang[2,3], Stephen P. Ficklin[1,2], Loren A. Honaas[3*]

[1] Molecular Plant Science Department, Washington State University, Pullman WA, 99164

[2] Department of Horticulture, Washington State University, Pullman WA, 99163

[3] USDA Agricultural Research Service Physiology and Pathology of Tree Fruits Research: Wenatchee, WA, 98801

*Corresponding Author. Email: loren.honaas@usda.gov

Notification: This Chapter is currently prepared for submission

Attributions

JAH worked on experimental design, performed the analyses and wrote the manuscript. **HLH** gathered samples and assisted with writing. **HZ** assisted with phylogenetic analysis. **SPF** assisted with editing and experimental design. **LAH** worked on experimental design, provided funding guidance.

Abstract

Research on how plants respond to hypoxia has concentrated on model organisms where tissues can only survive hypoxic conditions for a few hours to a few days. In contrast, hypoxic conditions are used commercially as a method to prolong the life of *Malus Domestica* (apple) fruit for up to a year of storage without tissue death. This ability of apples to remain alive in hypoxic conditions is an interesting adaptation that has had limited molecular investigation despite its economic importance. Here we investigate the long-term apple hypoxia response using a time-course RNA-seq dataset collected for several postharvest storage conditions. We use comparative phylogenies, differential expression, and regulatory networks to identify genes that regulate and are regulated by the hypoxia response. We identify potential neo-functionalizations of core-hypoxia response genes in apples, including novel regulation of sub-group VII Ethylene Response Factor (ERFVII) and plant cysteine oxidases (PCO) family members.

1 Introduction

Groundbreaking work on the identification and revelation of the molecular mechanisms of sub-group VII Ethylene Response Factor (ERFVII) Transcription Factors (TFs) laid the foundation for our current understanding of plant responses to (and their recovery from) low oxygen environments (Gibbs et al., 2011; Hattori et al., 2009; Hinz et al., 2010; Licausi et al., 2010, 2011; Mustroph et al., 2010). Shortly after, the identification of plant cysteine oxidases (PCOs) as critical enzymes which (in Arabidopsis) act in combination with constitutively expressed ERFVII proteins to

regulate a ‘tunable oxygen-sensing system’ (Gasch et al., 2016; Weits et al., 2014). To briefly summarise this system, Arabidopsis has five ERFVII TFs. Three of these ERFVII TFs (RAP2.2, RAP2.3, RAP2.12: RAP type ERFVII) are constitutively expressed (Bui et al., 2015) whereas the remaining two (HRE1 and HRE2: HRE type ERFVII) are upregulated during hypoxia (Licausi et al., 2010). The activity of these ERFVII genes is regulated at the protein level, where they are degraded via the n-degron pathway in the presence of oxygen and are functionally active during hypoxic conditions. The n-degron pathway is able to quickly respond to hypoxia by oxygen sensing through the PCO gene family. The PCO family needs oxygen to oxidize the N-terminal cysteine residue present on all ERFVII-type proteins after the removal of Met by MAP (Giuntoli & Perata, 2018). This cysteine oxidation event flags the ERFVII genes for degradation. However, in the absence of oxygen, PCO is unable to perform this cysteine oxidation, which allows the ERFVII TFs to rapidly activate the hypoxia response by binding to hypoxia-responsive promoter element (HRPE) motifs of hypoxia-responsive genes (HRGs) (Gasch et al., 2016)

Studies using model organisms are critical for developing our understanding of molecular pathways. In model organisms, these core hypoxia response pathways have been studied using molecular biology techniques, such as knock-out mutants, transformations, CRISPR, etc., often observing gene expression patterns in tissues like leaves, shoots, and roots (Mustroph et al., 2009). Non-model organisms, such as *Malus domestica*, present excellent opportunities to explore how core hypoxia stress response mechanisms are adapted across plants.

Malus domestica (apple) fruit is an intriguing plant organ for studying hypoxia due to its unique adaptations. Plants such as Arabidopsis and *Oryza sativa* (rice) are useful for studying transient responses to hypoxia during stress response and recovery over a matter of hours or days (Fu & Xu, 2023; León et al., 2021; Papdi et al., 2015).In contrast, with apples, hypoxia is used as a tool in commercial storehouses to inhibit fruit ripening, where fruit are kept in low or ultra-low oxygen conditions (often below 1% oxygen) for months and sometimes as long as a year (Brizzolara et al., 2020). In apples (and other large storage organs) this prolonged exposure to low oxygen is not necessarily unnatural, as such storage organs have been shown to have constitutive hypoxic conditions internally [aka hypoxic niches (Cukrov, 2018; Geigenberger et al., 2000; Licausi et al., 2011; Loreti & Perata, 2020; Rolletschek et al., 2002)]. The nature of the success of low oxygen storage, coupled with the natural physiology of fruits, provides an opportunity to explore how fruits (and other equivalent species) have evolved the ability to withstand such prolonged hypoxia.

Another important trait of apple biology is that they are climacteric: ripening in response to ethylene. This is intriguing because recent research has shown cross-talk between hypoxia and ethylene pathways. In tomatoes, ERFVII orthologs were observed to be involved in regulating fruit ripening (Liu et al., 2016), including acting as negative regulators of carotenoid accumulation (Lee et al., 2012), indicating an important relationship between hypoxia signaling, ethylene response factors, and fruit ripening (Cukrov 2018). In persimmon, hypoxia-responsive ERFs were demonstrated to have dual roles in both regulation of low oxygen metabolism genes and deastringency associated with ripening (Min et al., 2012, 2014; Wang et al., 2017; Zhu et al., 2018). In

grapes, the development of hypoxic conditions was observed in berry flesh during ripening (Xiao et al., 2018). Furthermore, pre-climacteric fruit (immature fruits) were observed as more tolerant to low oxygen stress compared to postclimacteric fruits (mature fruits) (Ke et al., 1994), and fruit of more advanced maturity have a higher potential to ferment in low oxygen conditions (Both et al., 2016).

In addition to the treatment of low-oxygen of apples by the industry, referred to as Controlled Atmosphere (CA), there are two additional strategies for postharvest management of apple fruit in storage: refrigeration and the application of ethylene inhibitors, such as 1-Methylcyclopropene (1-MCP). These strategies are used to slow metabolic activity (chilling) and interfere with the molecular mechanisms of the ripening process by inhibiting ethylene sensing and signaling (CA and MCP (Cukrov, 2018)). By blocking ethylene perception, 1-MCP generally inhibits respiration rates, reduces the rate of softening, prevents loss of greenness, inhibits greasiness, and has a mixed effect on titratable acidity, volatile content, soluble solid concentration, and physiological postharvest disorders (Watkins, 2006), and references therein, (Lv et al., 2020)). 1-MCP treatment has also been demonstrated to reduce the accumulation of reactive oxygen species in fruit exposed to chilling stress (Sabban-Amin et al., 2011). As an inhibitor of ethylene perception, 1-MCP application may affect apple fruit's adaptive responses to low oxygen environments, either through enhancement (Mattheis et al., 2005; Rupasinghe et al., 2000; Watkins, 2006; Watkins et al., 2000) or impairment, as evidenced by certain cultivars (such as 'Honeycrisp') becoming more susceptible to low oxygen injury when treated with 1-MCP (Chiu et al., 2015).

Exploration of the impacts of 1-MCP in concert with CA on apple fruit has received interest with regard to fruit quality outcomes (J. DeEll et al., 2016; J. R. DeEll & Lum, 2017; Poirier et al., 2020; Watkins et al., 2015; Zanella, 2003), metabolomics (Bekele et al., 2014; Hatoum et al., 2016), but rarely through the lens of transcriptomics (Johnson & Zhu, 2015). The experiment There has been significant molecular work performed on how model organisms respond to short-term hypoxic conditions. However, it is likely that there are novel adaptations in gene function in apple fruit which allow them to survive for up to a year in hypoxic conditions. In this paper, we use transcriptomics to 1) identify core genes associated with responses to long-term hypoxic conditions in apple fruit; 2) gain insight into the role ethylene and chilling temperatures play in these long-term responses; 3) suggest apple homolog-specific adaptations to the current hypoxia molecular model (Giuntoli & Perata, 2018). We identified potential neo-functionalizations of core-hypoxia response genes in apples, including novel transcriptomic regulation of sub-group VII Ethylene Response Factor (ERFVII) and plant cysteine oxidases (PCO) family members. We also showed that apple fruit transcriptomic response to long-term hypoxic storage can be loosely divided into two responses, where the first is rapid to respond and potentially controlled by the n-degron pathway whereas the latter does not respond until months after storage and is likely controlled by ethylene.

2 Materials and Methods

2.1 Plant material, experimental design, and fruit texture

‘Gala’ apples were collected and sorted as detailed in Hadish et al. (2023, in submission, Chapter 2 of this dissertation). The current analysis uses the RNA-seq data described in this paper. The fruit was received from a commercial facility in Quincy, WA on August 21st, 2018. Upon arrival at the USDA-ARS Tree Fruit Research Laboratory in Wenatchee, WA, apples were randomly sorted by hand and stored in air at 1°C for 7 days. Four RNA-seq samples were taken during this ripening period. After 7 days of conditioning, apples were randomly assigned into treatment and storage conditions. These fruits were divided into six treatment condition categories (A1, A10, A20, MCP, CA, MCPCA). Fruit in the MCP and MCPCA treatments were treated with SmartFresh™ (AgroFresh Solutions, Inc., Philadelphia, PA USA), also known as 1-Methylcyclopropene (1-MCP), overnight and then stored at 1°C in either air (MCP) or controlled atmosphere (MCPCA, 2% O_2, 1% CO_2). 1-MCP was applied at 1°C and in accordance with SmartFresh™ product recommendations. The fruit not treated with 1-MCP were stored at 1°C in a controlled atmosphere (CA, 2% O_2, 1% CO_2), at 1°C in air conditions (A1), at 10°C in air conditions (A10), or at 20°C in air conditions (A20). Postharvest sampling was done at condition-relevant time intervals, as untreated fruit stored in air (A1 treatment) was expected to lose firmness faster than long-term fruit. Please see the methods section of (Hadish et al. 2023, in submission, Chapter 2 of this dissertation)**,** for a more detailed description of experimental conditions, treatments, and sampling time points.

2.2 Tissue collection, RNA extraction, and Quality Control

Tissue collection and RNA extraction and quality control were performed as described in (Hadish et al. 2023, in submission, Chapter 2 of this dissertation). The fruit was kept at respective temperatures until the moment of tissue harvest. Fruit stored in CA was removed prior to tissue harvest (in air for ~30 minutes from removal to completion of tissue collection). Three slices of cortex from six 'Gala' apples each were pooled to create a biological replicate, and three biological replicates were collected (18 apples total) for each treatment. Slices were coarsely diced and immediately flash-frozen in liquid nitrogen and stored at -80°C.

RNA was extracted using a CTAB/Chloroform protocol modified for use on pome fruit tissue in the postharvest period (Honaas & Kahn, 2017). Extracted RNA was analyzed for purity using the NanodropOne (Thermo Fisher Scientific, Waltham, MA USA), for integrity on the Agilent Bioanalzyer (Agilent, Santa Clara, CA USA, Agilent-RNA Pico Kit, cat#: 5067-1513), and quantity using the Invitrogen™ Qubit™3 (Thermo Fisher Scientific, Waltham, MA USA, Qubit™ RNA HS Assay Kit, cat#: Q32852). Only RNA that met the following standards was used for downstream analysis: A260/A280 ≈ 2.0, RNA Integrity Number (RIN) ≥ 8.0.

2.3 Transcriptome Sequencing, Quality Control, and Reference Genome Selection

This analysis is described in Hadish et al. (2023, in submission, Chapter 2 of this dissertation) and summarized here. Briefly, libraries using Lexogen's QuantSeq 3' mRNA-Seq Library Prep Kit FWD (Cat# 015; www.lexogen.com) were prepared at the Penn State Genomics Core Facility (University Park, PA, United States) per (Honaas et

al., Jan 12-16 2019). Libraries were sequenced on a 150 bp single-end protocol to a target volume of ~20 million reads per biological replicate on Illumina's HiSeq 2,500 in Rapid Mode. Raw read data are publicly available at the NCBI Sequencing Read Archive (SRA - BioProject PRJNA938164).

RNA-seq reads were preprocessed with Trimmomatic (Bolger et al., 2014) prior to genome alignment, per (Lexogen, 2020) recommendations. These reads were then processed using the GEMmaker Workflow (Hadish et al., 2022) running Hisat2 (Kim et al., 2015) using default settings to create a Gene Expression Matrix (GEM). The 'Golden Delicious' doubled-haploid genome (GDDH13) (Daccord et al., 2017) was downloaded from the Genome Database for Rosaceae (GDR) (Jung et al., 2019) and used for alignment. The GEM was normalized using Deseq2's (Love et al., 2014) median of ratio normalization (Anders & Huber, 2010). Samples with low alignment (4 samples appeared to have significant issues as they had less than 20% alignment) and genes with zero RNA-Seq reads across the sample set were removed prior to downstream analyses. Both the GEM and MultiQC reports can be found in Hadish et al. (2023, in submission, Chapter 2 of this dissertation).

2.4 Differential Expression

The DESeq2 package (Love et al., 2014) was used for differential gene expression (DEG) analysis. Three different DEG analyses were performed. For the first two analyses, samples were grouped by treatment (Number of Samples for each: PreTreat: 12, A1: 33, A10: 20, A20: 16, MCP: 20, MCPCA: 21, CA: 21). The first analysis sought to identify all genes which were upregulated during hypoxia. To be classified as

upregulated during hypoxia, gene needed to be at least 1 log2 fold change up-regulated with an adjusted p-value (padj) of < 0.05 in MCPCA or in CA compared (as oxygen is at low levels in these treatments) to each of the other conditions (PreTreat, A1, A10, A20, MCP). A total of 606 genes were identified as upregulated in this analysis.

In the second DEG analysis the upregulated MCPCA and CA (hypoxia sample groups) genes needed to be differentially expressed from each other. This was done to identify genes that are hypoxia responsive as well as ethylene responsive. This consisted of two subsets: subset A where CA was upregulated compared to all other treatments (PreTreat, A1, A10, A20, MCP, and MCPCA), and subset B where MCPCA was upregulated compared to all other treatments. A gene was classified as a DEG if it had a least a 1 log2 fold change up-regulated and a padj of < 0.05). A total of 52 and 20 genes were identified for CA and MCPCA upregulation respectively.

A third DEG analysis focused on long-term fruit (CA, MCP, and MCPCA fruit) over the course of the time period (2,4,5,6,7,8, and 9 months postharvest). At each timepoint differential expression was performed for CA versus MCP and MCPCA, and for MCPCA versus MCP. This resulted in two gene sets--“ethylene” and “n-degron”--which are visualized in **Figure 4**. The “ethylene” gene set is where genes in CA fruit show upregulation when compared to MCP and MCPCA fruit, which has the goal of identifying genes upregulated by ethylene. The “n-degron” gene set is where MCPCA and CA are upregulated when compared to MCP, which has the goal of identifying genes upregulated in response to hypoxia.

2.5 Phylogenetic Analysis

To identify homologs of DEGs from apples, other closely related species (i.e. Rosaceae species), and plant model organisms (for example *Arabidopsis thaliana* (Arabidopsis) and *Oryza sativa* (rice)), and to investigate the evolutionary history of gene families of interest, a phylogenetic approach was taken. First, DEGs were classified into orthogroups pre-computed with the 26Gv2.0 scaffold using the both BLAST and HMM option implemented in the GeneFamilyClassifier tool from PlantTribes2 (Wafula et al., 2022). The list of DEGs and their corresponding orthogroups are listed in **Supplementary Tables 5 and 6**. Next, all genes classified into the same orthogroup were identified from 16 Rosaceae genomes [the same 15 from (Zhang et al., 2022) plus Malus baccata (W. Chen et al., 2019)] and were merged with sequences from the 26Gv2.0 scaffolding species following methods from (Zhang et al., 2022). These resulting files which contain homologs of DEGs all across land plants are available in **Supplementary File 1,** and were used as input for gene family alignment and phylogeny. Some of the DEGs belong to large orthogroups (e.g. OG1 contains 8275 sequences from all the investigated genomes) and the number of sequences in these orthogroups exceeded the input sequence limit of the alignment software, MAFFT, thus a subset of sequences were used - sequences from 7 genomes were used in OG1-10 (*Malus domestica* GDDH13 and Honeycrisp; *Pyrus betulifolia*; *Fragaria vesca* v4.0a2; *Arabidopsis thaliana* TAIR10; *Vitis vinifera* v2.1; *Oryza sativa* v7.0); sequences from 13 genomes were used for OG11-30 (the 7 mentioned above plus *Rosa chinensis* v2; *Populus trichocarpa* v3.0; *Theobroma cacao* v1.1; *Solanum lycopersicum* v2.4; *Nelumbo nucifera* v1.0; *Amborella trichopoda* v1.0). Orthogroup multiple sequence

alignment, phylogenetic tree estimation, homology inference, and gene model evaluation were performed following methods from (Zhang et al., 2022). Phylogenetic trees were visualized using Dendroscope (version 3.8.8) (Huson & Scornavacca, 2012).

Members of the ERF gene families were classified into several orthogroups, thus, to construct a gene family tree containing all the ERFs, the SuperOrthogroup classification from the PlantTribes2 was investigated. First, orthogroups belonging to the same SuperOrthogroup (under MCL stringency 3.0 from GeneFamilyClassifier output, **Supplemental Table 1**) as OG7, which contains most of the known Arabidopsis ERF genes, were extracted. This resulted in 6 orthogroups - OG7, OG2171, OG7665, OG14248, OG16955, and OG17668. Because the PlantTribes2 functional annotations of OG7665 and OG14248 indicate that proteins in these 2 orthogroups are involved in abscisic acid signal transduction pathway and dehydration response, respectively, these 2 orthogroups were removed from the list. For the rest of the 4 orthogroups, sequences from the 7 genomes mentioned above (genomes used for OG1-10) plus *Malus domestica* golden delicious v1.0 were used for alignment construction and phylogeny inference. The same method as described above was used for multiple sequence alignment, but the phylogenetic tree was inferred using IQ-TREE version 2.0.3 (Nguyen et al., 2015)) with 2000 ultrafast bootstrap replicates (Hoang et al., 2018) and -bnni for bootstrap optimization.

2.6 GO Enrichment Analyses and Gene Annotation

Gene Ontology functional enrichment analysis was performed on all gene sets using the AgriGO v2 database “Go Analysis Tool” (http://bioinformatics.cau.edu.cn/AppleMDO/

accessed May 2023). The “GDDH13 V1.1 homology with Arabidopsis reference” was used (Du et al., 2010; Tian et al., 2017). **Supplemental Figure 7** was generated using the “Graphical Result” tool available after the AgriGO v2 GO analysis tool.

For *cis*-motif identification, the 1000 nucleotide genomic sequence of all genes was extracted from the Golden Delicious Double Haploid (GDDH13) genome file using the provided annotations (Daccord et al., 2017) (accessed using the Genome Database for Rosaceae (GDR, https://www.rosaceae.org/) (Jung et al., 2019)) and the tool seqkit (version 2.1.0)(Shen et al., 2016). Annotated 5’-untranslated regions were included in this 100 bp sequence. The presence of the *cis*-motif HRPE (GCCVCYGGTTTY) (Gasch et al., 2016) was detected in these 1000 nucleotide genomic sequences using the FIMO package (Grant et al., 2011) of Meme-suite v 5.5.2 (Bailey et al., 2015) (https://meme-suite.org/meme/tools/fimo accessed May 2023).

A mappingof gene names from the Arabidopsis genome to the GDDH13 genome was provided by an orthologue table available from GDR (Jung et al., 2019) (https://www.rosaceae.org/species/malus/malus_x_domestica/genome_GDDH13_v1.1 accessed January 2023). Mapping of gene names from the Golden Delicious version 1.0 genome (Velasco et al., 2010) to the GDDH13 genome was done using OrthoFinder (version 2.5.5) (Emms & Kelly, 2019). Gene descriptions of putative Arabidopsis orthologues were retrieved from The Arabidopsis Information Resource (TAIR) (Berardini et al., 2015) (https://www.arabidopsis.org/tools/bulk/genes/index.jsp accessed June 2023).

Heatmap visualization and clustering of DEG was done using the package ‘pretty heatmaps’ (pheatmap) version 1.0.12 (Kolde, 2018) in R version 4.1.3 (R Core Team,

2022). Visualization of DEG counts was done using ggplot2 (Wickham, 2009) and dplyr (Wickham et al., 2023) of the tidyverse package (Wickham et al., 2019) as well as ggrepel for intelligent labeling (Slowikowski, 2023).

2.7. GENIE3 Network Construction and Analysis

A GENIE3 (Huynh-Thu et al., 2010) style network was constructed using the sklearn python implementation (Pedregosa et al., 2011). Pre-processing was performed to remove genes that did not have at least 10 counts in 3 genes, resulting in a matrix with 143 samples and 23813 genes.

A list of apple transcription factors identified using iTAK (Zheng et al., 2016) was retrieved from the AppleMDO database http://bioinformatics.cau.edu.cn/AppleMDO/gene_family/ (Da et al., 2019). This list consisted of 2965 putative transcriptive factors in the GDDH13 Apple Genome. 1557 of these had significant gene expression (at least 3 samples with a count of 10) in the dataset and were used in the analysis. GENIE3 (Huynh-Thu et al., 2010) was used to predict the putative targets of the transcription factors in our dataset. Settings for GENIE3 were set at *max_features* = sqrt(number of transcription factors) (which equated to 39) and *n_estimators* = 1000. The completed regulatory network, with relationships thresholded at 5 potential TF per gene, is available as **Supplemental Table 2**. An additional step was taken to record r^2 and m_rmse metrics for how well transcription factors were able to predict the value of each gene **Supplemental Table 3**. A histogram of the r^2 values is visualized as **Supplemental Figure 1**.

After regulatory network construction, genes identified as differentially expressed in hypoxia and their putative transcription factors were selected from the network to form subgraphs and graphed using cytoscape version 3.9.1 (Shannon et al., 2003). Thresholding of the network was performed by selecting the top 5 transcription factors for each gene in the 606 hypoxia gene set. This resulting subgraph was then reduced by excluding transcription factor which did not regulate at least 10 of the hypoxia genes (ad hoc threshold) and is available as **Supplemental Table 4**. Network analysis within cytoscape was used to generate node degree. Coloring of transcription factors corresponds to iTAK putative function, with a key available as **Supplemental Figure 2**. TFs were classified as targeting ethylene-responsive hypoxia genes if the majority of the genes they targeted belonged to the DEG group of 72 genes, otherwise, they were classified as hypoxia targeting if the majority of the genes they targeted were in the remaining 606 hypoxia genes. The “majority” was normalized to account for a number of genes in each category so that the majority for “ethylene responsive hypoxia targeting” was defined by anything over the line $y = 24/143 * x$ where as “hypoxia targeting” were classified as anything below this line. The line was calculated based on the maximum number of genes targeted by TF in either category. See **Figure 4** for the line plot and information on TFs. “Ethylene responsive hypoxia targeting” genes were subclassified as “CA only upregulated” or “MCPCA only upregulated” based on the set of genes they primarily targeted **Supplemental Figure 3 A**.

A second subgraph was created containing the ERVII genes and their putative targets. An ERVII gene was classified as a putative regulator if it was in the top 5 TFs

regulating a gene. The number of putative regulators of each of the 1557 genes is graphed as a histogram in **Supplemental Figure 4**.

3 Results

3.1 Apple PCO Genes Classification

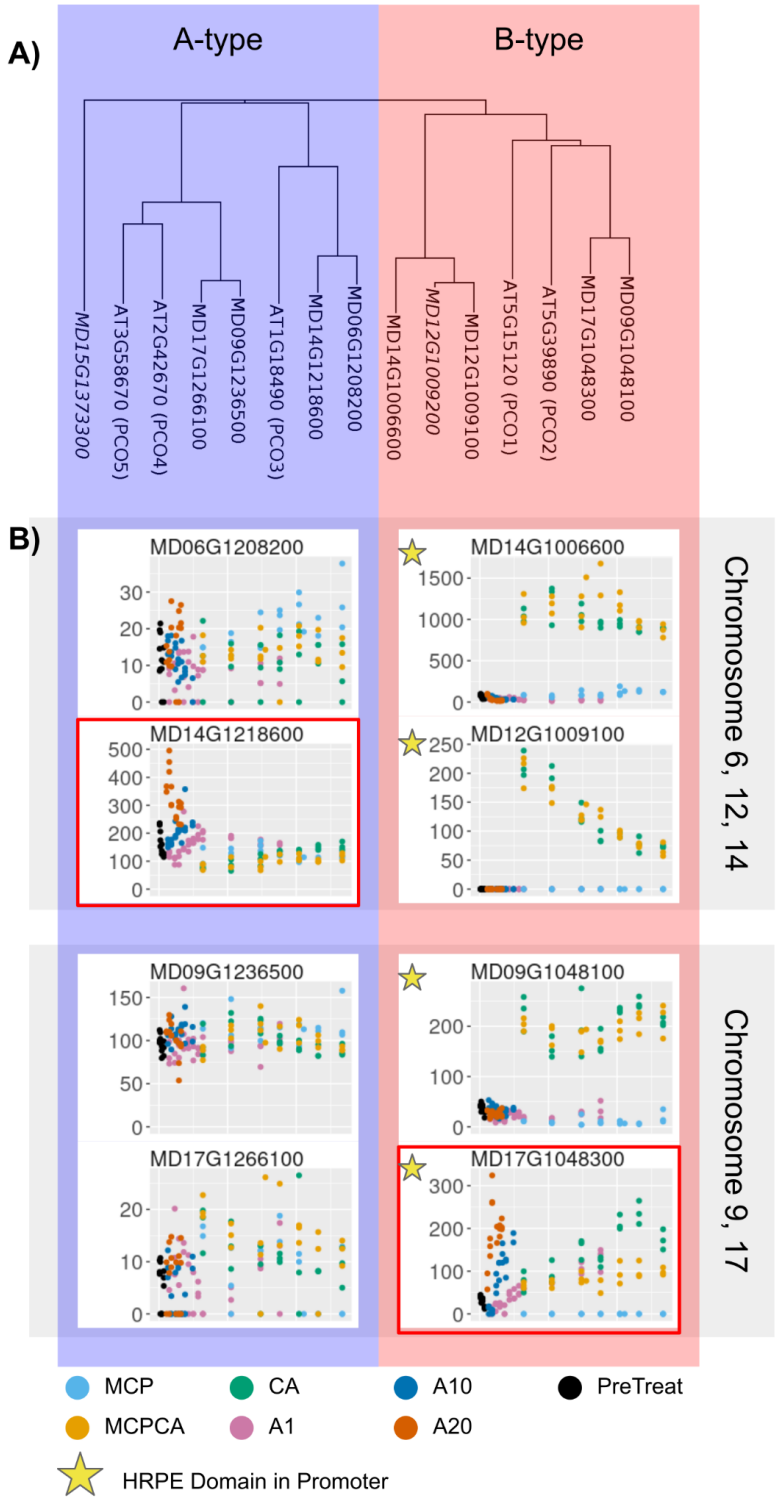


Figure 1: Plant Cysteine Oxidase (PCO) family in Apple, grouped into A-type (stable, non-hypoxia inducible) and B-type (hypoxia-responsive) categories proposed by (Weits et al., 2023). **A) Phylogeny** shows a nucleotide-based grouping of *Arabidopsis thaliana* PCO genes and Apple GDDH13 PCO genes. An expanded phylogeny is included in **Supplemental Figure 5** and includes PCO genes from 41 species. Apple Genes in italics (MD15G1373300 and MD12G1009200) showed no expression in our dataset, and are excluded from the expression plots of this figure. **B) Gene Expression Plots** show Apple GDDH13 PCO genes with expression in our dataset. The x-axis represents time in storage, and the y-axis represents gene count normalized using the DESeq2 package (Love et al., 2014). Genes are surrounded by gray boxes according to recent genome duplication events (Velasco et al., 2010). Duplication events were verified to be in *Pyrus communis* (Pear) as well (**Supplemental Figure 5**), making this consistent with current knowledge of the subtribe Malinae duplication event (Li et al., 2019). Genes with unexpected transcriptomic levels based on PCO A/B classification (MD14G1218600 and MD17G1048300) are highlighted using a red outline.

PCO genes have been shown to be important in the Arabidopsis hypoxia response due to their ability to detect oxygen concentration and their role in the n-degron pathway (Giuntoli & Perata, 2018). Ten apple homologs of Arabidopsis PCOs were identified in the GDDH13 genome (**Supplemental Figure 5**). Of these, two showed no expression in our dataset (MD15G1373300 and MD12G1009200). The remaining eight PCOs were categorized into type A and type B (Weits et al., 2023) based on phylogeny and HRPE elements (**Figure 1 A and Supplemental Figure 5**), with 4 being characterized as type A (MD06G1208200, MD14G1218600, MD09G1236500, and MD17G1266100) and 4 being characterized as type B (MD14G1006600, MD12G1009100, MD09G1048100, and MD17G1048300). All type B Apple PCO genes contained a *cis* HRPE motif in the 1000bp upstream of their translational start site whereas none of the type A apple PCO genes contained this motif. These type A and B were further categorized based on the recent Maleae (apple tribe) genome duplication event (Velasco et al., 2010) with 4 likely paralogous pairs: MD06G1208200 with MD14G1218600, MD09G1236500 with

MD17G1266100, MD14G1006600 with MD12G1009100, and MD09G1048100 with MD17G1048300 **Figure 1 B**.

Gene expression of the 143 samples was visualized across time to assess similarities and differences in expression (**Figure 1 B**). For type A PCO genes, 3 had relatively constant expression across all experimental conditions where as MD14G1218600 showed upregulation in short-term fruits in the A10 and A20 treatments. For type B PCO genes 3 (MD14G1006600, MD12G1009100 and MD09G1048100) showed upregulation under hypoxic treatments (CA and MCPCA) when compared to all other treatments whereas the remaining PCO gene MD17G1048300 had upregulation over the course of the short-term treatments (A1, A10, A20). Additionally, MD17G1048300 showed a difference in expression between the two hypoxia treatments (CA and MCPCA) and complete elimination of expression under MCP treatment which indicates partial regulation via ethylene-based mechanisms.

3.2 Apple ERFVII Gene Family

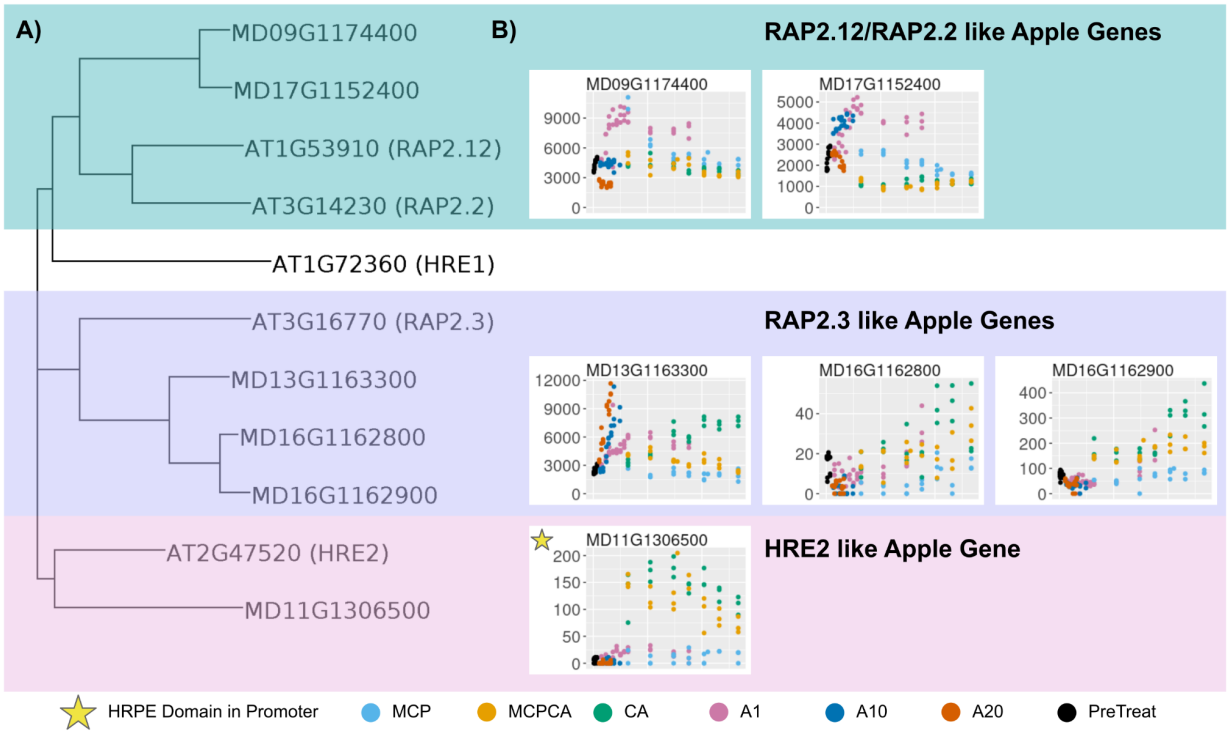


Figure 2: Ethylene Response Factor group VII family in Apple GDDH13 genome. These are the only genes in the GDDH13 genome with the N-terminal degron (N-degron) motif MCGGAI/V. **A**) Phylogeny of apple genes compared to their Arabidopsis homologs. Colored boxes break the phylogeny into three categories based on similarity to Arabidopsis homologs, expression patterns, and presence of *cis* HRPE motif. The sea-green box contains RAP2.12/RAP2.2-like genes, the purple contains RAP2.3-like genes, and the pink containsHRE2 like genes. **B)** Expression profiles of the 6 ERFVII genes with the x-axis representing time in storage, and the y-axis representing gene counts normalized using the DESeq2 package (Love et al., 2014).

Six apple ERFVII genes were identified from the GDDH13 genome using phylogenetic analysis (**Figure 2 A** and **Supplemental Figure 6**). These were categorized into three groups (RAP2.12/RAP2.2-like, RAP2.3-like, and HRE2-like) based on their similarity to Arabidopsis homologs, expression patterns (constitutive expression or hypoxia responsive), and presence of cis HRPE motif. Two were

classified as RAP2.12/RAP2.2 like (MD09G1174400 and MD17G1152400) three as RAP2.3 like (MD13G1163300, MD16G1162800, MD16G1162900) and the remaining gene was classified as HRE2 like (MD11G1306500). These were the only six genes in the apple GDDH13 genome that contained the N-terminal degron (N-degron) motif MCGGAI/V, which is conserved across kingdoms for ERFVII transcription factors and is used for degradation via the “type I” PRT6 cys/arg branch of the n-degron pathway (Dissmeyer, 2019; Gibbs et al., 2011; Licausi et al., 2011).

Expression of these six genes is visualized across time to assess similarities and differences (**Figure 2 B)**. Genes that were the closest homologs to Arabidopsis RAP2.12 and RAP2.2 (MD09G1174400, MD17G1152400, top of **Figure 2 B**) showed dramatic changes in short-term fruit (A1, A10, and A20), with low temperatures (A1) causing higher transcriptomic upregulation than in hypoxic conditions (MCPCA and CA) which were also at the same temperature. Genes whose closest homolog was Arabidopsis RAP2.3 (MD13G1163300, MD16G1162800, and MD16G1162900, middle of **Figure 2 B**) did not have large transcriptomic responses to temperature. MD16G1162800 and MD16G1162900 showed a modest increase during hypoxic conditions. MD13G1163300 showed an opposing regulatory pattern to the cold-induced RAP2.12/RAP2.2 type TF, with increased regulation during warmer conditions, indicating it may be induced by ethylene production. The remaining gene (MD11G1306500, bottom of **Figure 2 B**), which was the closest homolog to Arabidopsis HRE2 was upregulated during hypoxia when compared to other treatments.

3.3 Hypoxia Responses

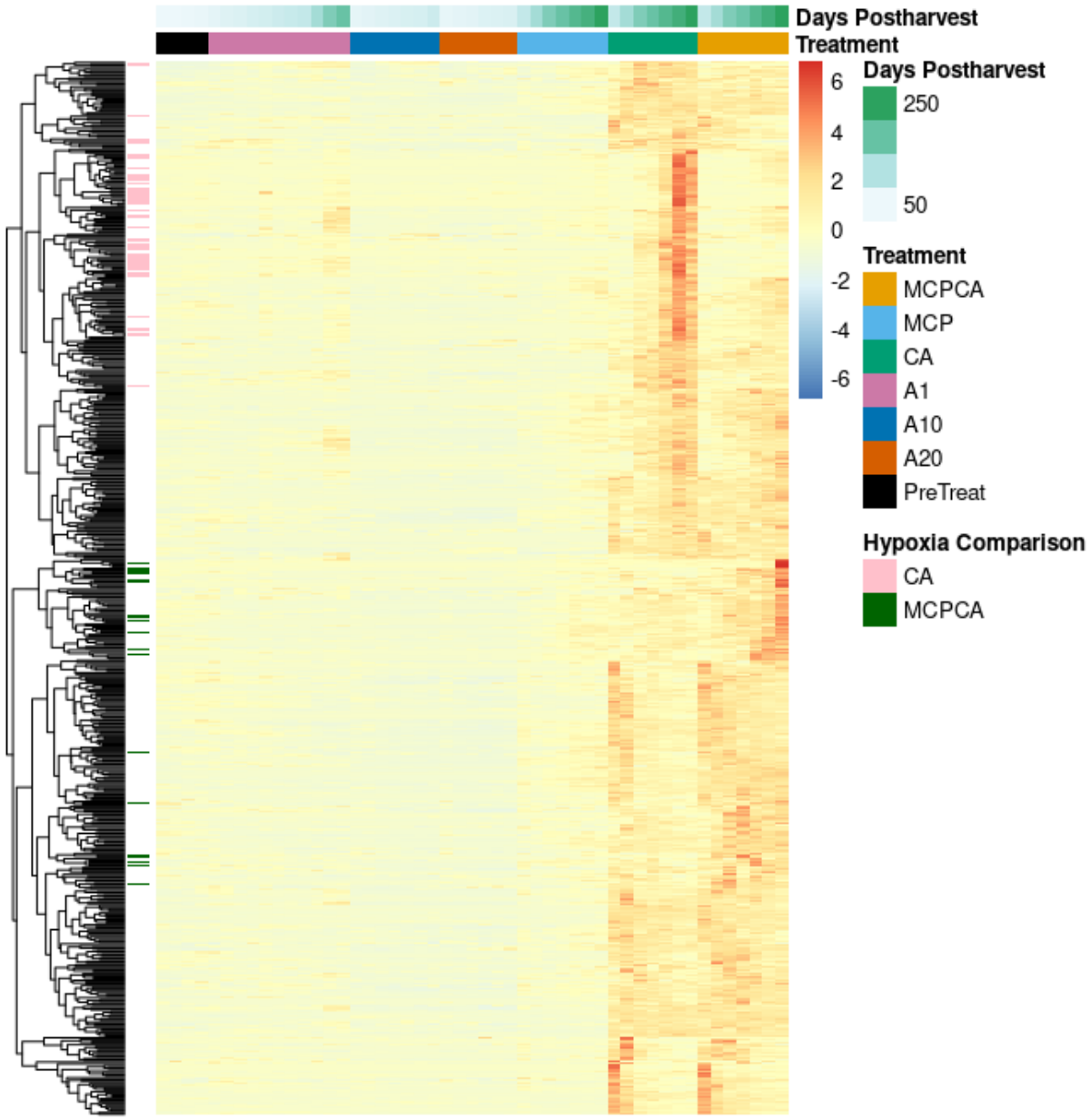


Figure 3: Heatmap of the 606 genes upregulated by Hypoxia over each treatment and condition tested. Rows (genes) were clustered based on expression similarity. The y-axis annotation between the dendrogram and heatmap, titled “Hypoxia Comparison”, highlights genes that are differentially upregulated under one of the hypoxia treatments (MCPCA and CA) and not the other. The x-axis annotations above the heatmap indicate samples grouped by treatment and ordered by days postharvest (the earliest timepoint sample is first within each treatment). Each cell represents the average of 3 replicates.

A total of 606 genes were identified as differentially upregulated during the hypoxia response when performing pairwise comparisons of all non-hypoxia treatment groups (PreTreat, A1, A10, A20, MCP) to each of the hypoxia groups (CA and MCPCA) (heatmap of genes **Figure 3, Supplemental Table 5**). Within this group of 606, 52 were identified as differentially upregulated only during the CA hypoxia response, and 20 were only upregulated during the MCPCA hypoxia response (**Figure 3** y-axis annotation "Hypoxia Comparison", **Supplemental Table 6**).

Functional Enrichment of the 606 hypoxia genes revealed terms related to oxygen sensing and hypoxia (GO:0036293, GO:0070482, GO:0001666, GO:1901700), energy management (GO:0015979, GO:0019684, GO:0009765, GO:0009055, and GO:0004022) and terms related to and various sensing molecules (GO:0010310, GO:0042743). All Go terms reported here had p-values below 0.00005 with exact p-values and a complete list of enriched GO terms available as **Supplemental Table 7** and a hierarchical tree view of these terms is available as **Supplemental Figure 7.** GO term enrichment analysis for the smaller 52 and 20 gene sets did not produce any significantly enriched terms. Genes within these groups (52 and 20) had Arabidopsis homologs which are known to be upregulated in response to ethylene and hypoxia such as members of the ACC OXIDASE (ACO1), and ACC SYNTHASE (ACS10) (Cukrov et al., 2016; Ireland et al., 2014), ERF1 and ERF2 (Hartman et al., 2019), and HYPOXIA RESPONSE UNKNOWN PROTEIN 26 (HUP26) (Huh, 2021). A list of Arabidopsis homologs for these genesets is available as part of **Supplemental Table 6**.

Of the 49 genes reported in Arabidopsis as core-induced genes during hypoxia (Mustroph et al., 2009) 27 were seen as differentially expressed in the list of 606

hypoxia genes (out of 59 GDDH13 genome orthologs whose value was over 3 for at least 10 samples using the DESeq2 normalized dataset). These genes are visualized as a heatmap as **Supplemental Figure 8**.

3.4 Upregulation of Hypoxia vs Ethylene genes in long-term fruit time-series data

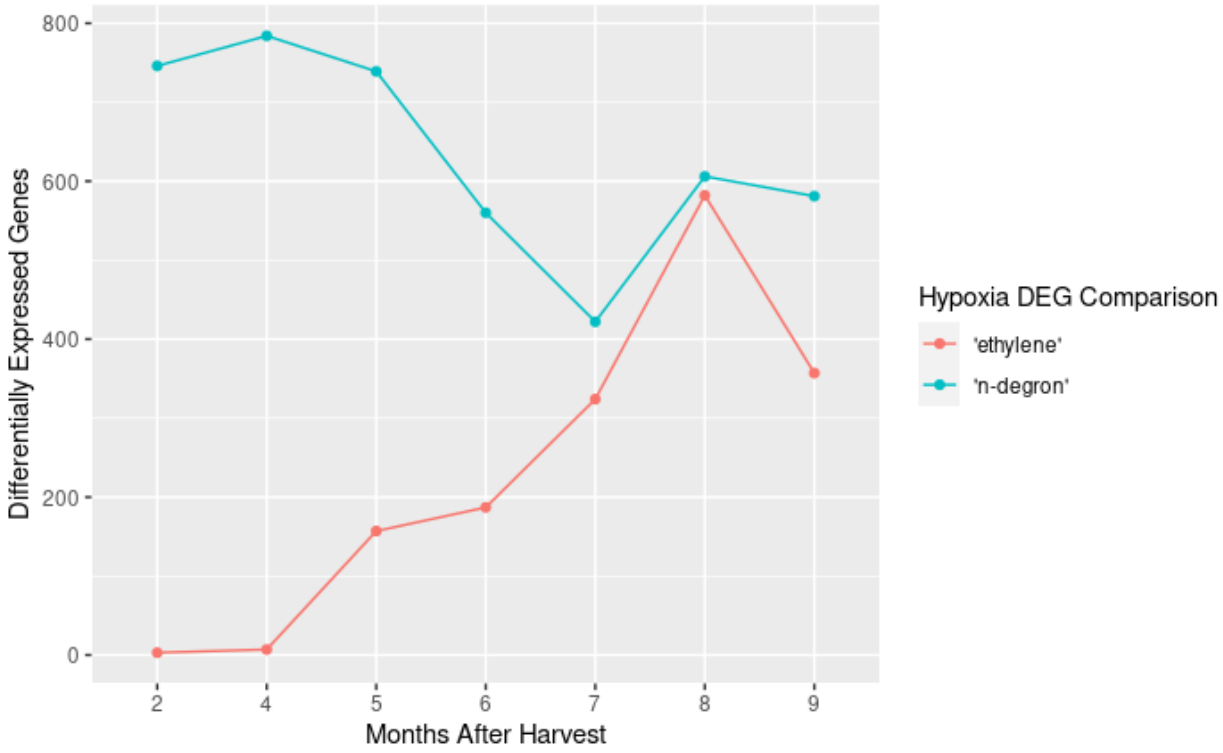


Figure 4: Number of DEGs in long-term fruit (CA, MCPCA, MCP) by months postharvest. The cyan-colored line represents genes that were upregulated in CA and MCPCA when compared to MCP (genes putatively involved in the n-degron pathway), while the salmon-colored line represents genes that were upregulated in CA when compared to MCPCA and MCP (genes putatively involved in ethylene response).

Pairwise differential expression analysis was performed between the long-term fruit (MCP, CA, MCPCA) at each month to assess trends in gene upregulation. Genes that were upregulated in CA and MCPCA compared to MCP were considered putative

hypoxia ‘n-degron’ related genes since they are upregulated in response to hypoxia irrespective of ethylene. Genes upregulated in CA compared to MCPCA and MCP were considered ‘ethylene’ related genes because they are only upregulated when ethylene is present (CA condition only). The number of DEGs at each timepoint is visualized in **Figure 4.** The number of DEG ‘n-degron’ genes was at least 400 genes at each timepoint over the 7 months that the long-term fruit was measured (**Supplemental Table 8**). In contrast, the number of DEGs classified as ‘ethylene’ was only three genes at the 2-month mark, but increased dramatically over the next 7- months (**Supplemental Table 9**). Heatmaps showing the expression patterns of these genes are available as **Supplemental Figure 9** and **Supplemental Figure 10**.

GO Functional Enrichment Terms are available as **Supplemental Table 10**. To summarise, the ‘n-degron’ genes top terms (p-value < 0.00005, see **Supplemental Table 10** for precise values), at each time point, were those related to decreased oxygen and hypoxia response (GO:0036293, GO:0070482,GO:0015979, GO:0001666, GO:1901700). Other terms included those related to energy metabolism (GO:0006091, GO:0019684, GO:0009767, GO:0009773, GO:0009765) and stress (GO:0080135, GO:0006970, GO:0009651). There were no terms enriched for the first two time points (two and four months), but later time points were enriched for terms related to ethylene (GO:0009873, GO:0071369), low oxygen response (GO:1901700) and a variety of other signaling pathways (**Supplemental Table 10**).

3.5 Predicted Transcriptomic Regulation of Hypoxia Up-regulated Genes

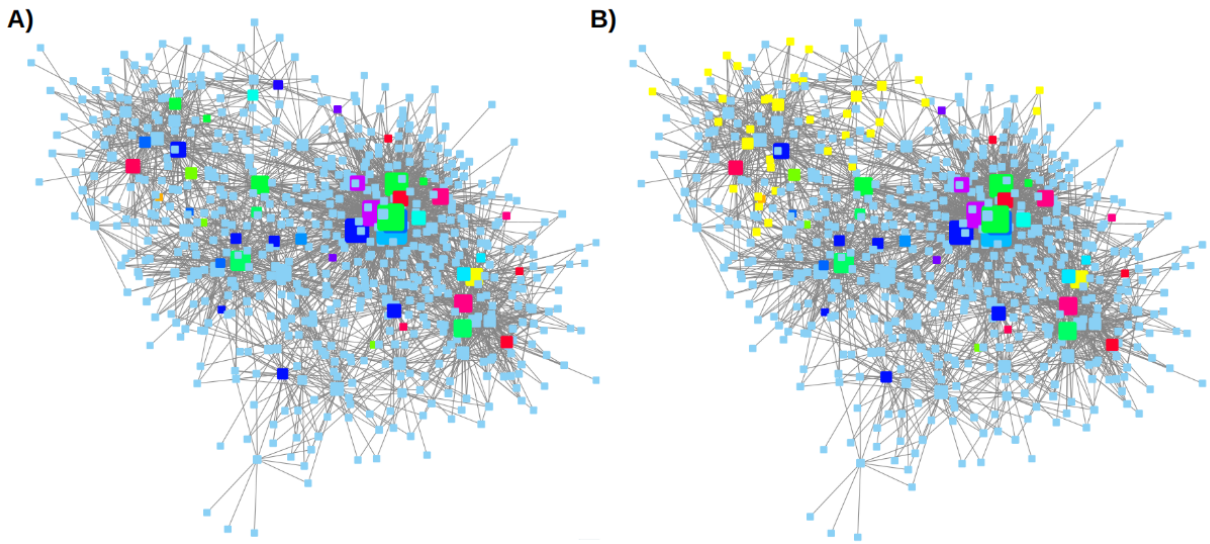


Figure 5: Regulatory network of 606 genes upregulated by hypoxic conditions. **A)** Network with transcription factors colored based on class (see **Supplemental Figure 2** for classes). Light blue nodes are genes identified as DEG during Hypoxia. The size of the node is relative to degree **B)** The same network as **A** with genes up-regulated only in CA conditions (not MCPCA) are visualized in yellow.

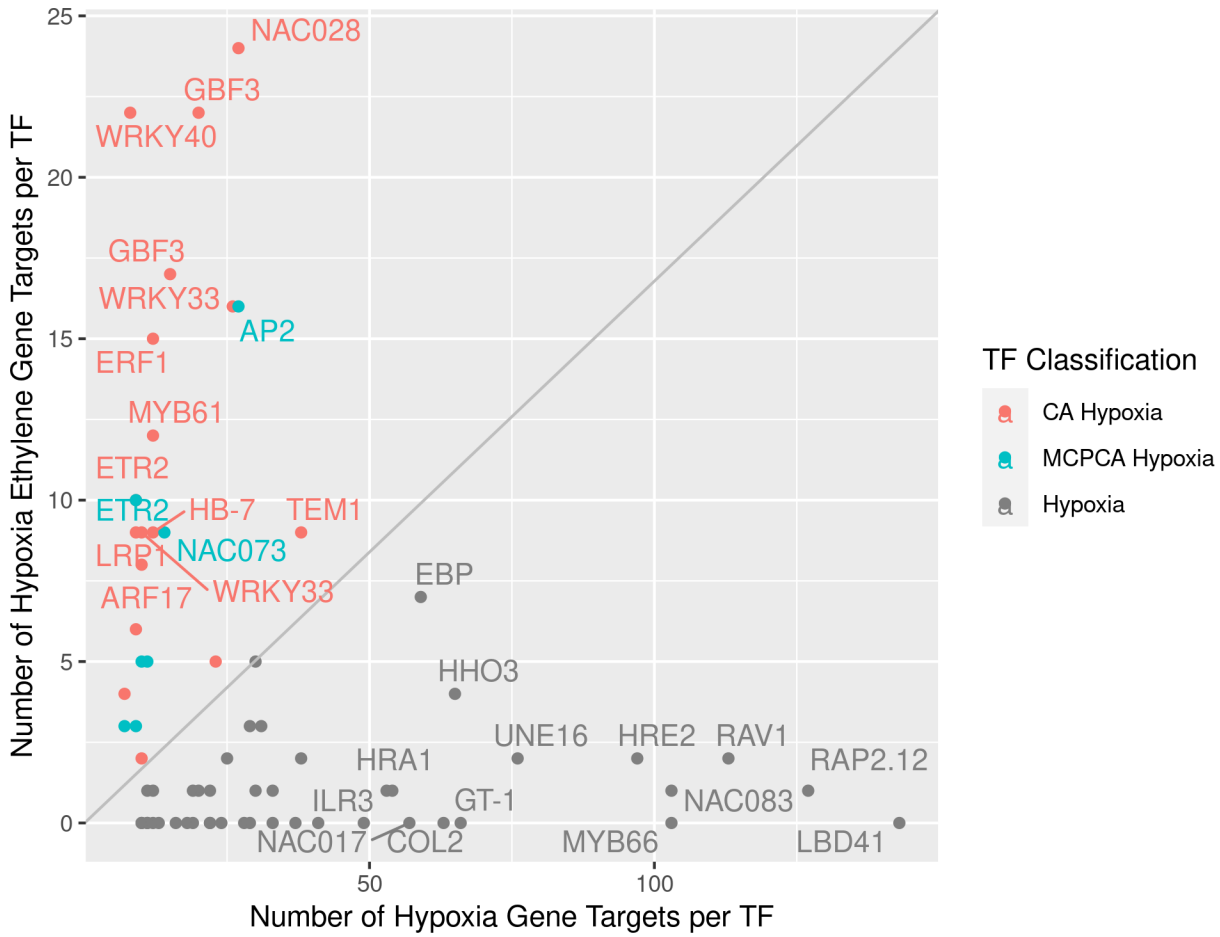


Figure 6: Transcription factor gene regulation of hypoxia and ethylene-dependent hypoxia genes. Each point represents a single transcription factor, with the adjacent label the closest Arabidopsis homolog (note that in some instances multiple apple homologs are present (i.e. ETR2, WRKY33, and GBF3). Arabidopsis homolog names are present for TFs regulating more than 7 Hypoxia Ethylene Dependent genes or at least 50 Hypoxia genes. The x-axis is the number of DEG genes related to hypoxia response that the transcription factor is connected to (potentially regulating) in the regulatory network thresholded at 10 TF for each gene. The y-axis is the same as the x-axis but for DEGs related to the hypoxia ethylene response (72 genes). See **Supplemental Table 11** for a description of each gene.

A regulatory network was created using GENIE3 (Huynh-Thu et al., 2010) to investigate transcription factors (TFs) with potential influence on hypoxia-related genes in apples. Of the 1557 apple TFs with expression in our dataset, 67 were present within the

regulatory network (**Figure 5 A, Supplemental Table 4**). The regulatory network was created without knowledge of which genes were members of the 606 hypoxia gene set and the subset of 52 CA. However, genes that were members of the 52 CA grouped together in a module providing more evidence for their co-regulation **Figure 5 B, Supplemental Figure 3 B.** The divide between these groupings is further illustrated in **Figure 6**, which shows a scatterplot of the transcription factors positioned according to the number of genes they putatively regulate, according to the regulatory network. The x-axis indicates the number of “Hypoxia” DEGs minus the “Ethylene Hypoxia” DEGs (606 - 72), and the y-axis represents the number of DEGs in the “Ethylene Hypoxia” group (72 genes). In the figure “Ethylene Hypoxia” is further broken down into genes that were DEG upregulated during CA only and those which were DEG upregulated in MCPCA only (salmon and colored points respectively). Additional information about the TFs visualized in **Figure 6** can be found in **Supplemental Table 11**.

TFs regulating the DEG gene set (**Figure 5**) included known hypoxia-related TF such as the apple homologs of LBD41 (MD09G1088700), RAV1 (MD13G1046100), and HRA1 (Giuntoli et al., 2017) (MD14G1094300). In addition, three of the ERFVII TFs previously mentioned were also HRE2 (MD11G1306500), RAP2.12 (MD17G1152400), and RAP2.3 (MD16G1162900) (Giuntoli & Perata, 2018). TFs that were connected to the Hypoxia ethylene genes included TFs known to be involved in cross-talk between ethylene and stress-related pathways such as apple homologs of ERF1 (MD10G1184800) and ETR2 (MD13G1209700) (Hartman et al., 2019; Zhao et al., 2012), TEM1 (MD16G1047700) and AP2 (MD15G1286400) (Licausi et al., 2013), as well as WRKY33 (MD04G1167700 and MD12G1181000) (Tang et al., 2021). Most of the

genes identified in the regulatory network are already implicated in the hypoxia response in Arabidopsis, providing supporting evidence for the validity of the network. Additionally, several novel putative regulators have been identified, which include members of the NAC family proteins (NAC017, NAC028, NAC083) known to be involved in plant stress (Bian et al., 2020), members of the MYB family (MYB66, MYB61) known to be involved in cell morphology, primary and secondary metabolism (Cao et al., 2020) and other members of the WRKY family (WRKY40) which are known to be involved in stress response and developmental processes (F. Chen et al., 2017). These provide hypotheses of potential additional mediators of the hypoxia response in apple fruit which are potentially working in conjunction with currently identified mechanisms.

3.6 Predicted Transcriptomic Regulation by ERFVII Family Genes

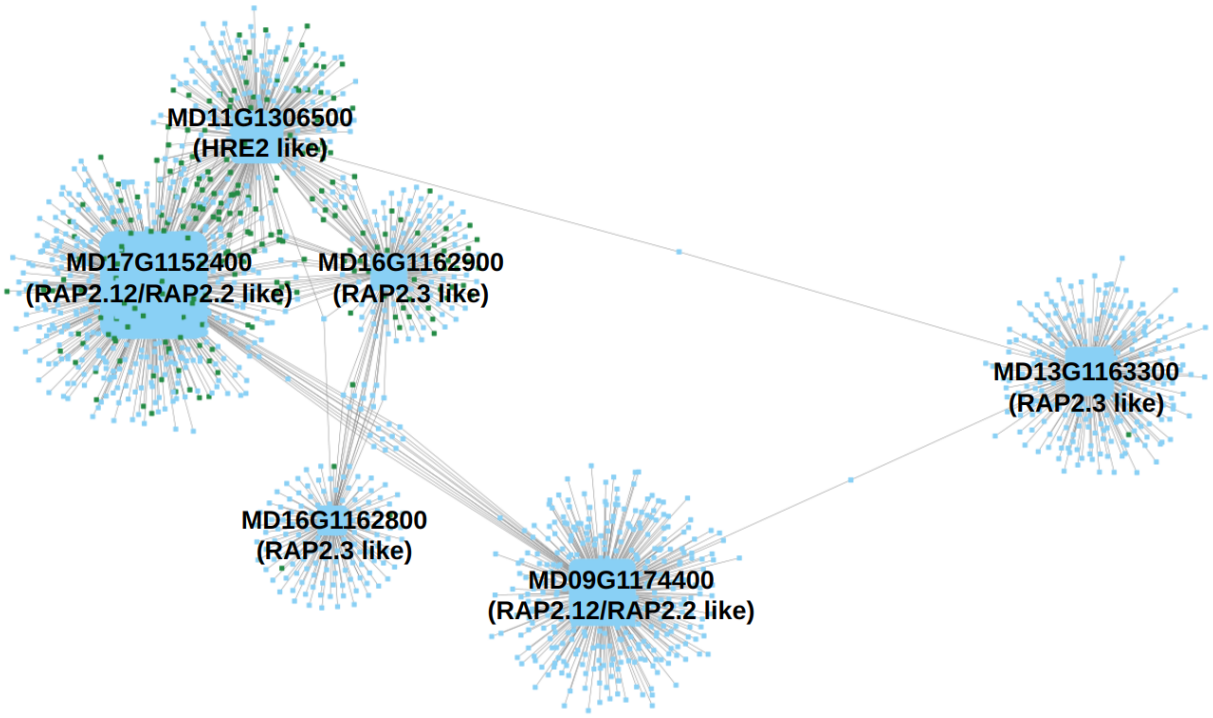


Figure 7: ERFVII transcription factor family and putatively regulated genes identified from the regulatory network. ERFVII genes are sized by total degree and labeled. Green-colored nodes are those in the DEG 606 hypoxia set. This subgraph highlights two major results: there is little overlap between genes putatively regulated by each ERFVII transcription factor, and three of these TFs are predicted to control genes from the 606 hypoxia set while the others are largely not.

The large differences in ERFVII expression patterns seen in **Figure 2** suggest they regulate different gene sets in apples. To investigate which genes these TFs could putatively be regulating, a reduced network was extracted from the regulatory network where predicted targets of each of the ERFVIIs were investigated **Figure 7, Supplemental Table 12**. The six ERFVII TFs had different numbers of genes which they were predicted to regulate, with MD09G1174400 regulating 356, MD11G1306500 273, MD13G1163300 253, MD16G1162800 143, MD16G1162900 184, and MD17G1152400 588. All of these were predicted to regulate more genes than the remaining TFs in the network, which on average regulated 77 genes each. **Supplemental Figure 4** shows a distribution of the number of genes regulated by each transcription factor present in the network.

GO functional enrichment analysis revealed some overlap between these gene sets, but also terms unique to each (**Supplemental Table 13)**. The two RAP2.2/RAP2.12 genes were enriched for different functional terms, with MD09G1174400 enriched for terms related to cellular components, while MD17G1152400 was enriched for oxygen-related terms and stress responses. The RAP2.3 TFs had few enriched terms, with MD13G1163300 enriched for one term which was organ morphogenesis (GO:0009887); MD16G1162800 enriched for “carbohydrate derivative biosynthetic process” (GO:1901137); and MD16G1162900 having no

enriched terms. The HRE2-related TF MD11G1306500 was enriched for three low oxygen-related terms (GO:0070482, GO:0036293, and GO:0001666). The network in **Figure 7** demonstrates that there is little overlap between the genes potentially being regulated by each ERFVII transcription factor.

4 Discussion

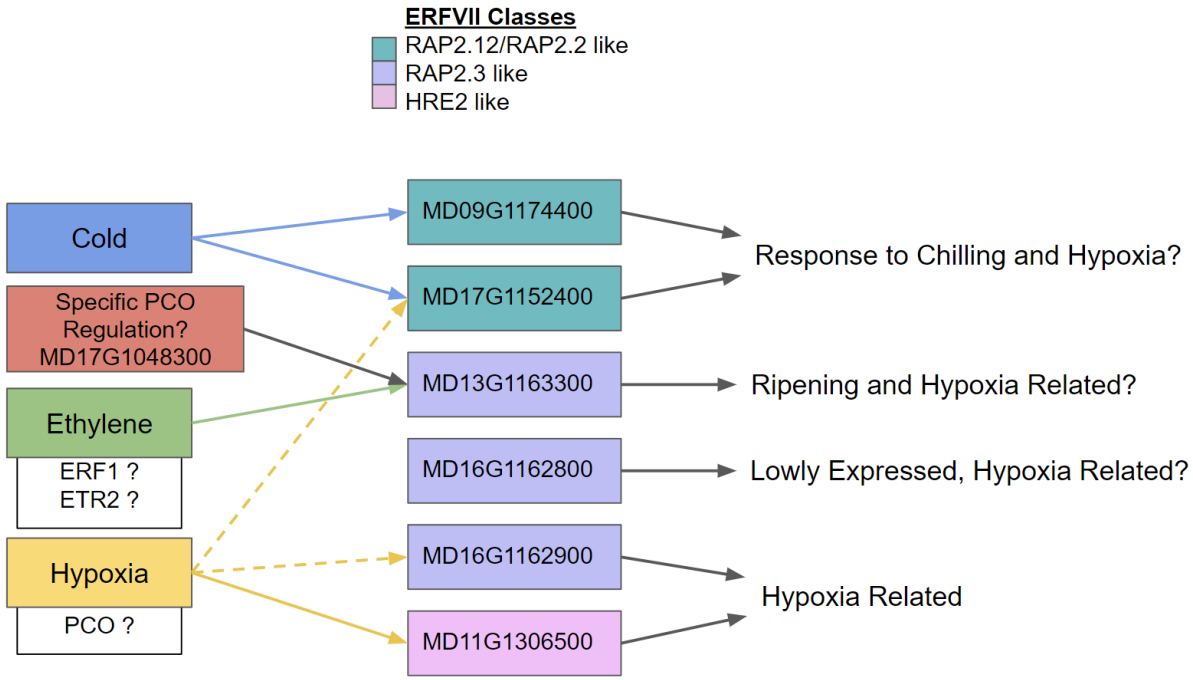


Figure 8: Putative diagram of transcriptomic regulation of ERFVII transcription factors. Regulators of transcription are Cold, PCO, Ethylene, and Hypoxia. While PCO acts post-translationally on ERFVII, upregulation of PCO's at the transcription level by ERFVII results in a transcriptional feedback loop.

Postharvest treatments such as chilling, CA, and 1-MCP are important tools used for storing apples for up to a year. However, the molecular mechanisms for why this is effective are not well understood in apples. This study investigated the time-series

transcriptomic response of apples under six different postharvest storage regimes. We sought to identify the core hypoxic response in apples, gain insight into how ethylene may be impacting the hypoxia response, and suggest homolog-specific adaptations to the current hypoxia molecular model.

4.1 Neo-functionalization of core hypoxia pathway genes in Apple fruit

We observed apparent neo-functionalization of the core n-degron pathway genes--specifically ERFVII and PCO--at the transcriptomic level during postharvest storage in apples. We chose to focus on the n-degron pathway genes due to their importance in regulating downstream hypoxia genes (ERFVII family) and the direct sensing of oxygen (PCO family). Additionally, other members of the n-degron pathway (e.g. homologs of ATE, PRT6, ACBP, MAP, SINAT1/2) did not show differential expression in our data; consistent with previous experiments indicating these genes tend to be constitutively expressed. (Giuntoli & Perata, 2018).

The apple PCO family consists of eight active homologs of the five Arabidopsis PCOs. These eight genes consist of four paralogous pairs that are likely the result of the Malinae duplication event (Velasco et al., 2010), with two pairs belonging to the type A PCOs and two pairs belonging to the type B PCOs. Previous work has shown that transcriptomic levels of type B PCO genes are directly induced by ERFVII proteins binding the *cis* HRPE motif during hypoxic conditions whereas type A PCO genes are constitutively expressed at low levels (Gasch et al., 2016; Weits et al., 2023). The induction of the type B PCO on exposure to hypoxic conditions “primes” a negative feedback loop where the increased levels of type B PCO proteins can rapidly reduce the

levels of ERFVII genes once the plant returns to normoxic conditions. This rapid attenuation of the hypoxia response is important to ensure that the expensive anaerobic metabolism is substituted for the more efficient aerobic metabolism (van Dongen & Licausi, 2015). Our data verified this response for 3 of the 4 active type B apple PCO genes (MD14G1006600, MD12G1009100 and MD09G1048100), but the response of the fourth gene (MD17G1048300) was unexpected (**Figure 1**). MD17G1048300 showed a response to hypoxia, but is also responsive to ethylene which suggests a neo-functionalization, which supported by phylogenetic analysis, likely arose after the Malinae duplication event (**Figure 1 B**). Previous studies of Arabidopsis PCOs have shown preferential selectivity of different ERFVII TFs and differences in expression patterns indicating the groups are not completely homogenous in function (White et al., 2018). For PCO MD17G1048300 the presence of the *cis* HRPE binding motif, similarities in expression patterns to one of the ERFVII RAP2.3 orthologues (MD13G1163300), and putative regulatory control support the assumption that these two genes play a role in responding to hypoxia during ethylene ripening (**Figure 8**).

MD14G1218600, a type A PCO gene, also showed unexpected expression patterns compared to previously described type A PCOs (Weits et al., 2023). It is the highest expressed of the type A PCO genes in our dataset and is up-regulated during warmer conditions (A10 and A20) which indicates that it may function to control ERFVII gene activity during ripening. Like the other apple PCO genes in this paper, it appears to be a result of the Malinae genome duplication event, sharing homology with MD06G1208200. This duplication would allow for neo-functionalization without influencing the core hypoxia response.

Our results show that the apple ERFVII gene family has a wide variety of responses to postharvest treatments at the transcriptomic level (**Figure 2**). Upregulation occurs during chilling for RAP2.12 and RAP2.2-like apple genes (MD09G1174400 and MD17G1152400) whereas one RAP2.3-like gene (MD13G1163300) shows upregulation during ripening and hypoxia conditions. The remaining RAP2.3-like genes (MD16G1162800, MD16G1162900) and the HRE-like gene (MD11G1306500) showed a more typical hypoxia upregulation response, albeit at varying levels. The behavior of the first three of these genes has not been recorded in apples. However, transcriptomic control of ERFVII genes during stress and developmental conditions has been recorded and verified in Arabidopsis models (Giuntoli & Perata, 2018). Confirmed Arabidopsis responses include RAP2.2 upregulated by ethylene and partner recorded and verified in Arabidopsis models (Giuntoli and Perata 2018). Confirming with Med25 to induce Botrytis resistance (Zhao et al., 2012), and RAP2.3 upregulated by ethylene and downregulated by DELLA to mediate apical hook development (Marín-de la Rosa et al., 2014). In both of these instances, the ERFVII gene acts as a way to integrate signals from multiple inputs. The ERFVII is transcriptionally upregulated by some (induced by the EIN2-EIN3/EIL signaling cascade or repressed by DELLA), or acting in concert with (MED25) to induce different downstream responses.

In our dataset, these different downstream responses induced by each ERFVII gene are predicted in **Figure 7**. Each of the ERFVII TFs are predicted to regulate dramatically different sets of genes with little overlap. Three of the genes (MD11G1306500, MD17G1152400, and MD16G1162900) are expected to regulate the

majority of the hypoxia response genes, whereas the other three are poorly enriched for hypoxia-related genes and are predicted to regulate genes with other functions.

While this transcriptomic data is suggestive of novel new roles of ERFVII and PCO apple genes, further investigation is needed to verify that they are functionally active at the protein level. Recent experiments have shown that PCO activity is repressed by cold temperature, which would suggest that the ERFVII MD09G1174400 and MD17G1152400 may play an active role during chilling (Gibbs et al., 2018).

4.2 Expansion of the downstream Hypoxia response

A large number of transcripts were identified as upregulated in response to hypoxic conditions (**Figure 3**). These covered a variety of GO terms related to oxygen sensing, rerouting of metabolism, and energy management (**Supplemental Table 7**). This is consistent with previous studies (Cukrov et al., 2016; Mustroph et al., 2009), with our DEG set covering half of the genes identified as “core hypoxia” in Arabidopsis from Mustroph et al. 2009 (Mustroph et al., 2009) (**Supplemental Figure 8)**. In addition, MCPCA-treated fruit allowed us to look at the set of hypoxia-related DEGs upregulated only in the presence of ethylene. This list was much smaller (72 in total) but consisted of genes that are responsive to ethylene in other experiments such as ACC OXIDASE (ACO1), ACC SYNTHASE (ACS10) (Cukrov et al., 2016; Ireland et al., 2014), ERF1 and ERF2 (Hartman et al., 2019), and HYPOXIA RESPONSE UNKNOWN PROTEIN 26 (HUP26) **Supplemental Table 6**.

The regulatory network was thresholded using the set of 606 differentially expressed hypoxia genes to identify potential TFs, with a total of 67 TF being identified (**Figure 5).**

Several of these TFs are known regulators of aspects of the plant hypoxia response. This included ERFVII genes which directly induce a hypoxia response, and genes that co-regulate with or are induced by the ERFVII genes, to initiate downstream hypoxia responses. These included LBD41, an anaerobic metabolism regulator (Mustroph et al., 2010), HRA1 a mediator of the hypoxia response downstream of ERFVII proteins (Giuntoli et al., 2014, 2017), and RAV1 a regulator of plant growth during stress conditions (Sengupta et al., 2020). These regulatory network results are exciting because hypoxia response TFs were implicated as potential regulators of the hypoxia DEGs without the inclusion of previous knowledge of their function. Replicating this biological knowledge in apples shows the power of transcriptomic network approaches for gene identification. It also provides hypotheses for additional TFs involved in the apple hypoxia response. Newly identified TFs are listed in the results as well as **Supplemental Table 11.**

DEGs identified as being ethylene dependent (gene set of 72) had predicted regulators which are induced by ethylene, such as ERF1, ETR2, and AP2 (Dolgikh et al., 2019). These genes have been shown to regulate ERFVII genes during other stress responses in arabidopsis (Marín-de la Rosa et al., 2014; Zhao et al., 2012) which suggests that our predicted TFs could explain the interesting transcriptomic levels we saw in the ERFVII genes **Figure 2**.

4.3 Differences between Controlled Atmosphere and 1-Methylcyclopropene fruit are most dramatic after over 7 months in storage

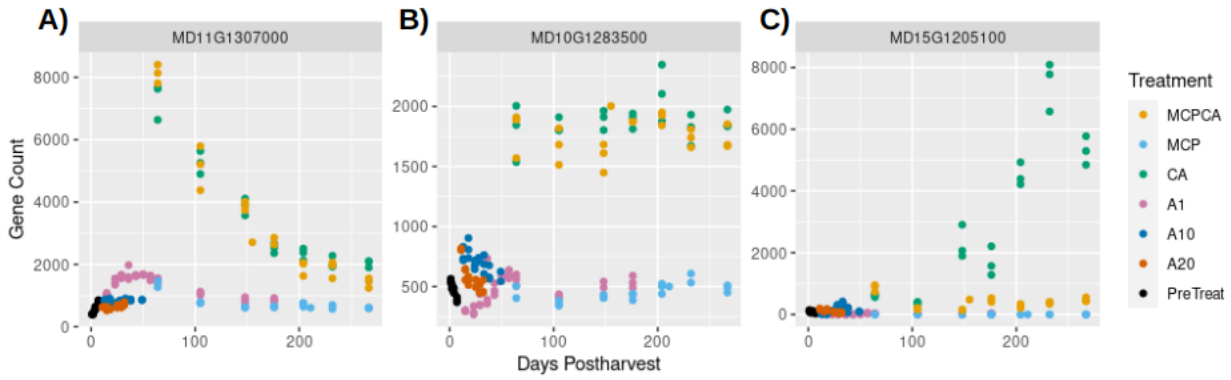


Figure 9: Representative model of three hypoxia response genes. **A** and **B** are included in the “n-degron” response group because they are induced in hypoxia (MCPCA or CA) regardless of ethylene (CA has some ethylene), whereas the **C** is referred to as “ethylene” as it is induced in CA but not in MCP treatments. **A** is rapidly induced then decreases with time, **B** is constitutively expressed, and **C** is upregulated late in the treatment, but only in CA fruit. Example genes shown in this figure are MD11G1307000 an ortholog of AT4G02280 (SUS3), a sucrose synthase; MD10G1283500, an ortholog of AT4G33070 (PDC1), a pyruvate decarboxylase that is the first step in ethanolic fermentation; and MD15G1205100 and ortholog of AT2G19590 (ACO1), a member of the yang cycle.

Hypoxia experiments performed in model organisms such as Arabidopsis and rice have focused on short-term low-oxygen environments from a few hours to a few days (Ellis et al., 1999; Klecker et al., 2014; Mustroph et al., 2009). This is because longer exposure results in tissue death. Apple fruit, in contrast, can be stored for periods of up to a year in low oxygen conditions (Gapper et al., 2022). This long-term storage ability likely has novel adaptations compared to short-term responses. Our results indicate that during long-term, low-oxygen storage, apple fruit responses can be split into at least two categories--a rapid response which is putatively controlled by the n-degron pathway, and a long-term response which is putatively controlled by ethylene. The rapid response can additionally be split into two categories, those whose expression are sustained over the course of the hypoxia experiment, or decrease after initial induction. An example of

these responses shown across time is visualized in **Figure 9**, and divisions based on these principles can also be seen in the clustering of **Figure 3.**

To further illustrate these differences in response, **Figure 4** shows how “n-degron” genes are upregulated over the course of the long-term hypoxia experiment whereas “ethylene” dependent genes are increasingly upregulated as the treatment progresses. In fact, DEGs at the 2 and 4-month timepoints did not show differential expression for the ethylene-dependent category, whereas the 7, 8, and 9-month timepoints showed hundreds (**Supplemental Table 8** and **Supplemental Table 9)**.

Our results show that the metabolism of an apple continues to change as it remains in a controlled atmosphere and that transcriptomic differences between MCPCA and CA fruit are most dramatic very late in their storage regime. This has implications for scientific studies of controlled atmosphere fruit, as previous studies are often conducted over a time scale of one or two months (Cukrov et al., 2016; Sanhueza et al., 2015)--ending before we observed differences in long-term treated fruit. This also has important implications for industry, as differences in gene transcript levels point to potential candidate genes that may explain differences seen in fruit quality between long-term treated fruit (Lu et al., 2018).

5 Conclusion

Long-term storage is important for supplying the demand for fresh apple fruit year-round. Little is known about how genes respond at the transcript level to common storage treatments such as 1-MCP, CA, and refrigeration. Such an understanding is important because pathways such as the n-degron and ethylene pathways are currently informed from Arabidopsis studies, yet apples have unique biology that supports

long-term storage and need greater understanding. In our efforts to uncover the unique differences in these pathways, major results from our study include:

i) There are neo-functionalizations in the transcriptomic response of core n-degron-related genes that are members of the PCO and ERFVII families. These functions appear to be in response to chilling stress and ethylene production.

ii) The downstream hypoxia response in apples consists of a large gene set (perhaps at least 606 from our DEG results) which are predicted to regulate similar functions as other plant hypoxia responses. Homologues of Arabidopsis hypoxia response TFs and several new putative TFs have been identified to play a role in regulating these genes during hypoxia. Ethylene appears to play a role in regulating a portion of this response.

iii) The apple transcriptome during long-term storage consists of genes that are rapidly induced and those whose expression slowly increases with time. The rapid response is consistent with previous characterizations of the n-degron pathway, whereas the long-term response appears to be primarily controlled by ethylene. This has implications for fruit management when considering the application of 1-MCP.

The observations of novel transcriptomic levels of PCO and ERFVII genes in postharvest apples is intriguing, but additional verification of their activity at the protein level is necessary. Future direction can focus on characterizing the interactors which cause these novel transcriptomic patterns and on the downstream targets of their

activity. TF identified as potential regulators in our networks are promising candidates for these genes. A better understanding of the hypoxia pathway at the molecular level has the potential for both breeding and postharvest storage management.

REFERENCES

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106.

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Research*, *43*(W1), W39–W49.

Bekele, E. A., Annaratone, C. E. P., Hertog, M. L. A. T. M., Nicolai, B. M., & Geeraerd, A. H. (2014). Multi-response optimization of the extraction and derivatization protocol of selected polar metabolites from apple fruit tissue for GC-MS analysis. *Analytica Chimica Acta*, *824*, 42–56.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The Arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *Genesis* , *53*(8), 474–485.

Bian, Z., Gao, H., & Wang, C. (2020). NAC Transcription Factors as Positive or Negative Regulators during Ongoing Battle between Pathogens and Our Food Crops. *International Journal of Molecular Sciences*, *22*(1). https://doi.org/10.3390/ijms22010081

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

Both, V., Thewes, F. R., Brackmann, A., Ferreira, D. de F., Pavanello, E. P., & Wagner, R. (2016). Effect of low oxygen conditioning and ultralow oxygen storage on the volatile profile, ethylene production and respiration rate of “Royal Gala” apples. *Scientia Horticulturae*, *209*, 156–164.

Brizzolara, S., Manganaris, G. A., Fotopoulos, V., Watkins, C. B., & Tonutti, P. (2020).

Primary Metabolism in Fresh Fruits During Storage. *Frontiers in Plant Science*, *11*, 80.

Bui, L. T., Giuntoli, B., Kosmacz, M., Parlanti, S., & Licausi, F. (2015). Constitutively expressed ERF-VII transcription factors redundantly activate the core anaerobic response in Arabidopsis thaliana. *Plant Science: An International Journal of Experimental Plant Biology*, *236*, 37–43.

Cao, Y., Li, K., Li, Y., Zhao, X., & Wang, L. (2020). MYB Transcription Factors as Regulators of Secondary Metabolism in Plants. *Biology*, *9*(3). https://doi.org/10.3390/biology9030061

Chen, F., Hu, Y., Vannozzi, A., Wu, K., Cai, H., Qin, Y., Mullis, A., Lin, Z., & Zhang, L. (2017). The WRKY Transcription Factor Family in Model Plants and Crops. *Critical Reviews in Plant Sciences*, *36*(5-6), 311–335.

Chen, W., Zhang, M., Zhang, G., Li, P., & Ma, F. (2019). Differential Regulation of Anthocyanin Synthesis in Apple Peel under Different Sunlight Intensities. *International Journal of Molecular Sciences*, *20*(23). https://doi.org/10.3390/ijms20236060

Chiu, G. Z., Shelp, B. J., Bowley, S. R., DeEll, J. R., & Bozzo, G. G. (2015). Controlled atmosphere-related injury in “Honeycrisp” apples is associated with γ-aminobutyrate accumulation. *Canadian Journal of Plant Science. Revue Canadienne de Phytotechnie*, *95*(5), 879–886.

Cukrov, D. (2018). Progress toward Understanding the Molecular Basis of Fruit Response to Hypoxia. *Plants*, *7*(4). https://doi.org/10.3390/plants7040078

Cukrov, D., Zermiani, M., Brizzolara, S., Cestaro, A., Licausi, F., Luchinat, C., Santucci,

C., Tenori, L., Van Veen, H., Zuccolo, A., Ruperti, B., & Tonutti, P. (2016). Extreme Hypoxic Conditions Induce Selective Molecular Responses and Metabolic Reset in Detached Apple Fruit. *Frontiers in Plant Science*, *7*, 146.

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

Da, L., Liu, Y., Yang, J., Tian, T., She, J., Ma, X., Xu, W., & Su, Z. (2019). AppleMDO: A Multi-Dimensional Omics Database for Apple Co-Expression Networks and Chromatin States. *Frontiers in Plant Science*, *10*, 1333.

DeEll, J., Lum, G., & Behrouz, E.-M. (2016). Effects of delayed controlled atmosphere storage on disorder development in “Honeycrisp” apples. *Canadian Journal of Plant Science. Revue Canadienne de Phytotechnie*. https://doi.org/10.1139/cjps-2016-0031

DeEll, J. R., & Lum, G. B. (2017). Effects of Low Oxygen and 1-Methylcyclopropene on Storage Disorders of “Empire” Apples. *HortScience: A Publication of the American Society for Horticultural Science*, *52*(9), 1265–1270.

Dissmeyer, N. (2019). Conditional Protein Function via N-Degron Pathway–Mediated Proteostasis in Stress Physiology. *Annual Review of Plant Biology*, *70*(1), 83–117.

Dolgikh, V. A., Pukhovaya, E. M., & Zemlyanskaya, E. V. (2019). Shaping Ethylene Response: The Role of EIN3/EIL1 Transcription Factors. *Frontiers in Plant Science*,

10, 1030.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research*, *38*(Web Server issue), W64–W70.

Ellis, M. H., Dennis, E. S., & Peacock, W. J. (1999). Arabidopsis roots and shoots have different mechanisms for hypoxic stress tolerance. *Plant Physiology*, *119*(1), 57–64.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238.

Fu, X., & Xu, Y. (2023). Dynamic Metabolic Changes in Arabidopsis Seedlings under Hypoxia Stress and Subsequent Reoxygenation Recovery. *Stresses*, *3*(1), 86–101.

Gapper, N. E., Bowen, J. K., & Brummell, D. A. (2022). Biotechnological approaches for predicting and controlling apple storage disorders. *Current Opinion in Biotechnology*, *79*, 102851.

Gasch, P., Fundinger, M., Müller, J. T., Lee, T., Bailey-Serres, J., & Mustroph, A. (2016). Redundant ERF-VII Transcription Factors Bind to an Evolutionarily Conserved cis-Motif to Regulate Hypoxia-Responsive Gene Expression in Arabidopsis. *The Plant Cell*, *28*(1), 160–180.

Geigenberger, P., Fernie, A. R., Gibon, Y., Christ, M., & Stitt, M. (2000). Metabolic activity decreases as an adaptive response to low internal oxygen in growing potato tubers. *Biological Chemistry*, *381*(8), 723–740.

Gibbs, D. J., Lee, S. C., Isa, N. M., Gramuglia, S., Fukao, T., Bassel, G. W., Correia, C. S., Corbineau, F., Theodoulou, F. L., Bailey-Serres, J., & Holdsworth, M. J. (2011). Homeostatic response to hypoxia is regulated by the N-end rule pathway in plants.

Nature, *479*(7373), 415–418.

Gibbs, D. J., Tedds, H. M., Labandera, A.-M., Bailey, M., White, M. D., Hartman, S., Sprigg, C., Mogg, S. L., Osborne, R., Dambire, C., Boeckx, T., Paling, Z., Voesenek, L. A. C. J., Flashman, E., & Holdsworth, M. J. (2018). Oxygen-dependent proteolysis regulates the stability of angiosperm polycomb repressive complex 2 subunit VERNALIZATION 2. *Nature Communications*, *9*(1), 5438.

Giuntoli, B., Lee, S. C., Licausi, F., Kosmacz, M., Oosumi, T., van Dongen, J. T., Bailey-Serres, J., & Perata, P. (2014). A trihelix DNA binding protein counterbalances hypoxia-responsive transcriptional activation in Arabidopsis. *PLoS Biology*, *12*(9), e1001950.

Giuntoli, B., Licausi, F., van Veen, H., & Perata, P. (2017). Functional Balancing of the Hypoxia Regulators RAP2.12 and HRA1 Takes Place in vivo in Arabidopsis thaliana Plants. *Frontiers in Plant Science*, *8*, 591.

Giuntoli, B., & Perata, P. (2018). Group VII Ethylene Response Factors in Arabidopsis: Regulation and Physiological Roles. *Plant Physiology*, *176*(2), 1143–1155.

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* , *27*(7), 1017–1018.

Hadish, J. A., Biggs, T. D., Shealy, B. T., Bender, M. R., McKnight, C. B., Wytko, C., Smith, M. C., Feltus, F. A., Honaas, L., & Ficklin, S. P. (2022). GEMmaker: process massive RNA-seq datasets on heterogeneous computational infrastructure. *BMC Bioinformatics*, *23*(1), 1–11.

Hartman, S., Liu, Z., van Veen, H., Vicente, J., Reinen, E., Martopawiro, S., Zhang, H.,

van Dongen, N., Bosman, F., Bassel, G. W., Visser, E. J. W., Bailey-Serres, J., Theodoulou, F. L., Hebelstrup, K. H., Gibbs, D. J., Holdsworth, M. J., Sasidharan, R., & Voesenek, L. A. C. J. (2019). Ethylene-mediated nitric oxide depletion pre-adapts plants to hypoxia stress. *Nature Communications*, *10*(1), 4020.

Hatoum, D., Hertog, M. L. A. T. M., Geeraerd, A. H., & Nicolai, B. M. (2016). Effect of browning related pre- and postharvest factors on the “Braeburn” apple metabolome during CA storage. *Postharvest Biology and Technology*, *111*, 106–116.

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

Hinz, M., Wilson, I. W., Yang, J., Buerstenbinder, K., Llewellyn, D., Dennis, E. S., Sauter, M., & Dolferus, R. (2010). Arabidopsis RAP2.2: an ethylene response transcription factor that is important for hypoxia survival. *Plant Physiology*, *153*(2), 757–772.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522.

Honaas, L. A., Hargarten, H., Ficklin, S. P., Hadish, J., Wafula, E. K., dePamphilis, C. W., Mattheis, J., & Rudell, D. (Jan 12-16 2019). *Co-Expression networks provide insight into postharvest fruit physiology*. International Plant and Animal Genome Conference, San Diego, CA.

Honaas, L. A., & Kahn, E. (2017). A practical examination of RNA isolation methods for

European pear (Pyrus communis). *BMC Research Notes*, *10*(1), 237.

Huh, S. U. (2021). New function of Hypoxia-responsive unknown protein in enhanced resistance to biotic stress. *Plant Signaling & Behavior*, *16*(3), 1868131.

Huson, D. H., & Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic Biology*, *61*(6), 1061–1067.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS One*, *5*(9). https://doi.org/10.1371/journal.pone.0012776

Ireland, H. S., Gunaseelan, K., Muddumage, R., Tacken, E. J., Putterill, J., Johnston, J. W., & Schaffer, R. J. (2014). Ethylene regulates apple (Malus × domestica) fruit softening through a dose × time-dependent mechanism and through differential sensitivities and dependencies of cell wall-modifying genes. *Plant & Cell Physiology*, *55*(5), 1005–1016.

Johnson, F. T., & Zhu, Y. (2015). Transcriptome changes in apple peel tissues during CO2 injury symptom development under controlled atmosphere storage regimens. *Horticulture Research*, *2*. https://doi.org/10.1038/hortres.2015.61

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

Ke, D., Yahia, E., Mateos, M., & Kader, A. A. (1994). Ethanolic Fermentation of `Bartlett' Pears as Influenced by Ripening Stage and Atmospheric Composition. *Journal of*

the American Society for Horticultural Science. American Society for Horticultural Science, *119*(5), 976–982.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360.

Klecker, M., Gasch, P., Peisker, H., Dörmann, P., Schlicke, H., Grimm, B., & Mustroph, A. (2014). A Shoot-Specific Hypoxic Response of Arabidopsis Sheds Light on the Role of the Phosphate-Responsive Transcription Factor PHOSPHATE STARVATION RESPONSE1. *Plant Physiology*, *165*(2), 774–790.

Kolde, R. (2018). *Package "pheatmap"* (Version 1.0.12) [Computer software]. https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf

Lee, J. M., Joung, J.-G., McQuinn, R., Chung, M.-Y., Fei, Z., Tieman, D., Klee, H., & Giovannoni, J. (2012). Combined transcriptome, genetic diversity and metabolite profiling in tomato fruit reveals that the ethylene response factor SlERF6 plays an important role in ripening and carotenoid accumulation. *The Plant Journal: For Cell and Molecular Biology*, *70*(2), 191–204.

León, J., Castillo, M. C., & Gayubas, B. (2021). The hypoxia-reoxygenation stress in plants. *Journal of Experimental Botany*, *72*(16), 5841–5856.

Lexogen. (2020). *Quant Seq 3' mRNA-Seq Library Prep Kit User Guide* (Issue 015UG009V0223).

Licausi, F., Kosmacz, M., Weits, D. A., Giuntoli, B., Giorgi, F. M., Voesenek, L. A. C. J., Perata, P., & van Dongen, J. T. (2011). Oxygen sensing in plants is mediated by an N-end rule pathway for protein destabilization. *Nature*, *479*(7373), 419–422.

Licausi, F., Ohme-Takagi, M., & Perata, P. (2013). APETALA2/Ethylene Responsive

Factor (AP2/ERF) transcription factors: mediators of stress responses and developmental programs. *The New Phytologist*, *199*(3), 639–649.

Licausi, F., van Dongen, J. T., Giuntoli, B., Novi, G., Santaniello, A., Geigenberger, P., & Perata, P. (2010). HRE1 and HRE2, two hypoxia-inducible ethylene response factors, affect anaerobic responses in Arabidopsis thaliana. *The Plant Journal: For Cell and Molecular Biology*, *62*(2), 302–315.

Li, H., Huang, C.-H., & Ma, H. (2019). Whole-Genome Duplications in Pear and Apple. In S. S. Korban (Ed.), *The Pear Genome* (pp. 279–299). Springer International Publishing.

Liu, M., Gomes, B. L., Mila, I., Purgatto, E., Peres, L. E. P., Frasse, P., Maza, E., Zouine, M., Roustan, J.-P., Bouzayen, M., & Pirrello, J. (2016). Comprehensive Profiling of Ethylene Response Factor Expression Identifies Ripening-Associated ERF Genes and Their Link to Key Regulators of Fruit Ripening in Tomato. *Plant Physiology*, *170*(3), 1732–1744.

Loreti, E., & Perata, P. (2020). The Many Facets of Hypoxia in Plants. *Plants*, *9*(6). https://doi.org/10.3390/plants9060745

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Lu, X., Meng, G., Jin, W., & Gao, H. (2018). Effects of 1-MCP in combination with Ca application on aroma volatiles production and softening of “Fuji” apple fruit. *Scientia Horticulturae*, *229*, 91–98.

Lv, J., Zhang, M., Bai, L., Han, X., Ge, Y., Wang, W., & Li, J. (2020). Effects of 1-methylcyclopropene (1-MCP) on the expression of genes involved in the

chlorophyll degradation pathway of apple fruit during storage. *Food Chemistry*, *308*, 125707.

Marín-de la Rosa, N., Sotillo, B., Miskolczi, P., Gibbs, D. J., Vicente, J., Carbonero, P., Oñate-Sánchez, L., Holdsworth, M. J., Bhalerao, R., Alabadí, D., & Blázquez, M. A. (2014). Large-scale identification of gibberellin-related transcription factors defines group VII ETHYLENE RESPONSE FACTORS as functional DELLA partners. *Plant Physiology*, *166*(2), 1022–1032.

Mattheis, J. P., Fan, X., & Argenta, L. C. (2005). Interactive responses of gala apple fruit volatile production to controlled atmosphere storage and chemical inhibition of ethylene action. *Journal of Agricultural and Food Chemistry*, *53*(11), 4510–4516.

Min, T., Fang, F., Ge, H., Shi, Y.-N., Luo, Z.-R., Yao, Y.-C., Grierson, D., Yin, X.-R., & Chen, K.-S. (2014). Two novel anoxia-induced ethylene response factors that interact with promoters of deastringency-related genes from persimmon. *PloS One*, *9*(5), e97043.

Min, T., Yin, X.-R., Shi, Y.-N., Luo, Z.-R., Yao, Y.-C., Grierson, D., Ferguson, I. B., & Chen, K.-S. (2012). Ethylene-responsive transcription factors interact with promoters of ADH and PDC involved in persimmon (Diospyros kaki) fruit de-astringency. *Journal of Experimental Botany*, *63*(18), 6393–6405.

Mustroph, A., Lee, S. C., Oosumi, T., Zanetti, M. E., Yang, H., Ma, K., Yaghoubi-Masihi, A., Fukao, T., & Bailey-Serres, J. (2010). Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses. *Plant Physiology*, *152*(3), 1484–1500.

Mustroph, A., Zanetti, M. E., Jang, C. J. H., Holtan, H. E., Repetti, P. P., Galbraith, D.

W., Girke, T., & Bailey-Serres, J. (2009). Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(44), 18843–18848.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274.

Papdi, C., Pérez-Salamó, I., Joseph, M. P., Giuntoli, B., Bögre, L., Koncz, C., & Szabados, L. (2015). The low oxygen, oxidative and osmotic stress responses synergistically act through the ethylene response factor VII genes RAP2.12, RAP2.2 and RAP2.3. *The Plant Journal: For Cell and Molecular Biology*, *82*(5), 772–784.

Pedregosa, F., Varoquaux, G., & Gramfort, A. (2011). Scikit-learn: Machine learning in Python. *Of Machine Learning …*. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com

Poirier, B. C., Mattheis, J. P., & Rudell, D. R. (2020). Extending “Granny Smith” apple superficial scald control following long-term ultra-low oxygen controlled atmosphere storage. *Postharvest Biology and Technology*, *161*, 111062.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rolletschek, H., Borisjuk, L., Koschorreck, M., Wobus, U., & Weber, H. (2002). Legume embryos develop in a hypoxic environment. *Journal of Experimental Botany*,

53(371), 1099–1107.

Rupasinghe, H. P. V., Murr, D. P., Paliyath, G., & Skog, L. (2000). Inhibitory effect of 1-MCP on ripening and superficial scald development in “McIntosh” and “Delicious” apples. *The Journal of Horticultural Science & Biotechnology*, *75*(3), 271–276.

Sabban-Amin, R., Feygenberg, O., Belausov, E., & Pesis, E. (2011). Low oxygen and 1-MCP pretreatments delay superficial scald development by reducing reactive oxygen species (ROS) accumulation in stored “Granny Smith” apples. *Postharvest Biology and Technology*, *62*(3), 295–304.

Sanhueza, D., Vizoso, P., Balic, I., Campos-Vargas, R., & Meneses, C. (2015). Transcriptomic analysis of fruit stored under cold conditions using controlled atmosphere in Prunus persica cv. “Red Pearl.” *Frontiers in Plant Science*, *6*, 788.

Sengupta, S., Ray, A., Mandal, D., & Nag Chaudhuri, R. (2020). ABI3 mediated repression of RAV1 gene expression promotes efficient dehydration stress response in Arabidopsis thaliana. *Biochimica et Biophysica Acta, Gene Regulatory Mechanisms*, *1863*(9), 194582.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504.

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PloS One*, *11*(10), e0163962.

Slowikowski, K. (2023). *ggrepel: Automatically Position Non-Overlapping Text Labels with “ggplot2.”* https://github.com/slowkow/ggrepel

Tang, H., Bi, H., Liu, B., Lou, S., Song, Y., Tong, S., Chen, N., Jiang, Y., Liu, J., & Liu, H. (2021). WRKY33 interacts with WRKY12 protein to up-regulate RAP2.2 during submergence induced hypoxia response in Arabidopsis thaliana. *The New Phytologist*, *229*(1), 106–125.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., & Su, Z. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, *45*(W1), W122–W129.

van Dongen, J. T., & Licausi, F. (2015). Oxygen sensing and signaling. *Annual Review of Plant Biology*, *66*, 345–367.

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.

Wafula, E. K., Zhang, H., Von Kuster, G., Leebens-Mack, J. H., Honaas, L. A., & dePamphilis, C. W. (2022). PlantTribes2: tools for comparative gene family analysis in plant genomics. In *bioRxiv* (p. 2022.11.17.516924). https://doi.org/10.1101/2022.11.17.516924

Wang, M.-M., Zhu, Q.-G., Deng, C.-L., Luo, Z.-R., Sun, N.-J., Grierson, D., Yin, X.-R., & Chen, K.-S. (2017). Hypoxia-responsive ERFs involved in postdeastringency softening of persimmon fruit. *Plant Biotechnology Journal*, *15*(11), 1409–1419.

Watkins, C. B. (2006). The use of 1-methylcyclopropene (1-MCP) on fruits and vegetables. *Biotechnology Advances*, *24*(4), 389–409.

Watkins, C. B., Gapper, N. E., Nock, J. F., Rudell, D. A., Leisso, R., Lee, J., Buchanan, D., Mattheis, J., Johnston, J., Schaffer, R., Giovannoni, J. J., Hertog, M. L. A. T. M., & Nicolaï, B. M. (2015). Interactions between 1-MCP and controlled atmospheres on quality and storage disorders of fruits and vegetables. *Acta Horticulturae*, *1071*(February), 45–58.

Watkins, C. B., Nock, J. F., & Whitaker, 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*(1), 17–32.

Weits, D. A., Giuntoli, B., Kosmacz, M., Parlanti, S., Hubberten, H.-M., Riegler, H., Hoefgen, R., Perata, P., van Dongen, J. T., & Licausi, F. (2014). Plant cysteine oxidases control the oxygen-dependent branch of the N-end-rule pathway. *Nature Communications*, *5*, 3425.

Weits, D. A., Zhou, L., Giuntoli, B., Carbonare, L. D., Iacopino, S., Piccinini, L., Lombardi, L., Shukla, V., Bui, L. T., Novi, G., van Dongen, J. T., & Licausi, F. (2023). Acquisition of hypoxia inducibility by oxygen sensing N-terminal cysteine oxidase in spermatophytes. *Plant, Cell & Environment*, *46*(1), 322–338.

White, M. D., Kamps, J. J. A. G., East, S., Taylor Kearney, L. J., & Flashman, E. (2018). The plant cysteine oxidases from Arabidopsis thaliana are kinetically tailored to act as oxygen sensors. *The Journal of Biological Chemistry*, *293*(30), 11786–11795.

Wickham, H. (2009). *ggplot2*. Springer New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., Mc Gowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller,

E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., & Spinu, V. (2019). Welcome to the Tidyverse. *The Journal of Open Source Software*. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*.

Xiao, Z., Rogiers, S. Y., Sadras, V. O., & Tyerman, S. D. (2018). Hypoxia in grape berries: the role of seed respiration and lenticels on the berry pedicel and the possible link to cell death. *Journal of Experimental Botany*, *69*(8), 2071–2083.

Zanella, A. (2003). Control of apple superficial scald and ripening—a comparison between 1-methylcyclopropene and diphenylamine postharvest treatments, initial low oxygen stress and ultra low oxygen storage. *Postharvest Biology and Technology*, *27*(1), 69–78.

Zhang, H., Wafula, E. K., Eilers, J., Harkess, A. E., Ralph, P. E., Timilsena, P. R., dePamphilis, C. W., Waite, J. M., & Honaas, L. A. (2022). Building a foundation for gene family analysis in Rosaceae genomes with a novel workflow: A case study in Pyrus architecture genes. *Frontiers in Plant Science*, *13*, 975942.

Zhao, Y., Wei, T., Yin, K.-Q., Chen, Z., Gu, H., Qu, L.-J., & Qin, G. (2012). Arabidopsis RAP2.2 plays an important role in plant resistance to Botrytis cinerea and ethylene responses. *The New Phytologist*, *195*(2), 450–460.

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., Banf, M., Dai, X., Martin, G. B., Giovannoni, J. J., Zhao, P. X., Rhee, S. Y., & Fei, Z. (2016). iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. *Molecular Plant*, *9*(12),

1667–1670.

Zhu, Q.-G., Gong, Z.-Y., Wang, M.-M., Li, X., Grierson, D., Yin, X.-R., & Chen, K.-S. (2018). A transcription factor network responsive to high CO2/hypoxia is involved in deastringency in persimmon fruit. *Journal of Experimental Botany*, *69*(8), 2061–2070.

Supplemental Materials

Supplemental Tables

Supplemental Table 1: ERF SuperOrthogorup Plant Tribes.
Included as a separate file.

Supplemental Table 2: Entire Apple regulatory network, thresholded at 5 putative transcription factors regulating each gene.
Included as a separate file.

Supplemental Table 3: Metrics describing the support of each gene in the hypoxia regulatory network. Metrics include m_rmse and r^2 values. Supplemental Figure 1 shows this data plotted to show the distribution of r^2 value.
Included as a separate file.

Supplemental Table 4: Thresholded regulatory network based on the Hypoxia upregulated genes. Referred to as the hypoxia regulatory network.
Included as a separate file.

Supplemental Table 5: results of the 606 Genes. Some genes are repeated multiple times (see methods section).
Included as a separate file.

Supplemental Table 6: results of the 72 Genes DEG for Ethylene dependent Hypoxia genes.
Included as a separate file.

Supplemental Table 7: GO terms for hypoxia genes (606).
Included as a separate file.

Supplemental Table 8: DEG Genes Time Series at each time point. “n-degron”.
Included as a separate file.

Supplemental Table 9: DEG Genes Time Series at each time point. “Ethylene”.
Included as a separate file.

Supplemental Table 10: Go Enrichment at each time point for ‘n-degron’ and ‘ethylene’ multiple comparisons. T5 corresponds to 64 days postharvest (dph), T6 = 105 DPH. T7 = 148DPH, T8 = 176 DPH*, T9 = 204 DPH T10 = 232 DPH *, T11 = 267 DPH. * Of the 9

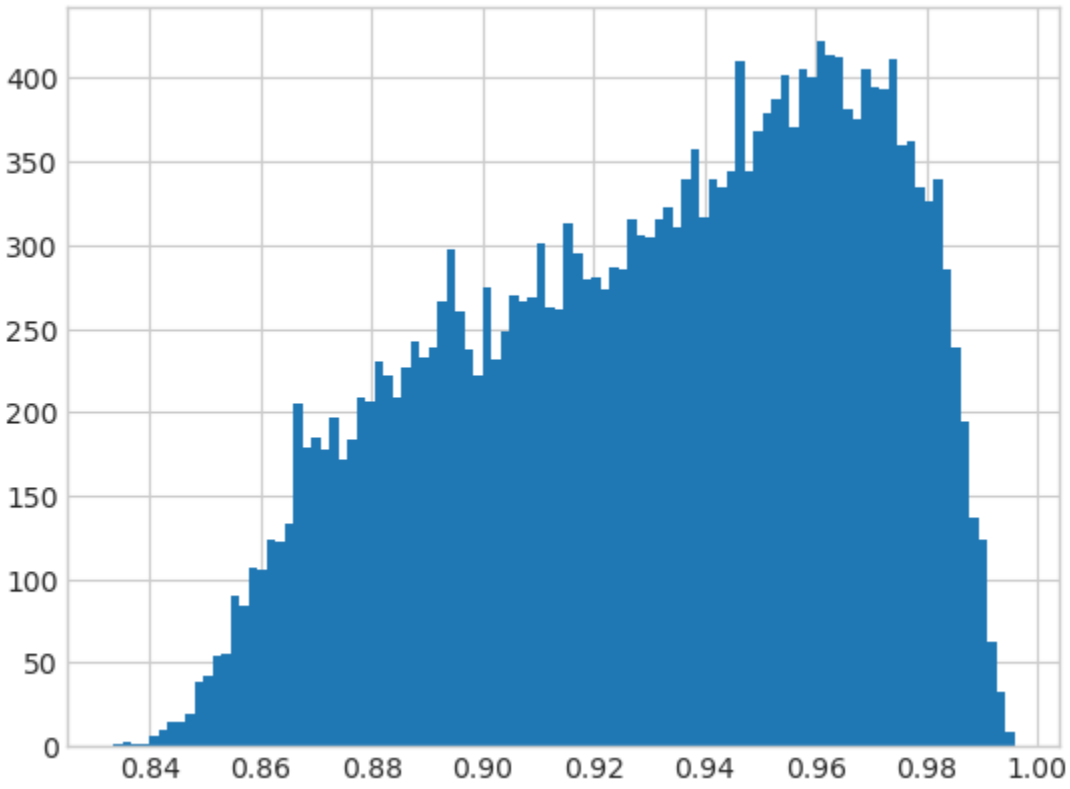
samples in T8, one was actually sampled at day 155 (136C_Clean_S40_R1_001), and of the 9 samples in T10, one was actually sampled on day 211 (100_Clean_S4_R1_001).
Included as a separate file.

Supplemental Table 11: Table of Transcription Factors which were identified to regulate hypoxia genes from the GENIE3 network. Columns: gene_GDDH13: apple gene identifier from the GDDH13 genome **gene_ath:** Arabidopsis gene identifier, identified by using the ath to GDDH13 conversion file from GDR. **n_hypoxia:** number of genes from the hypoxia condition this TF was seen regulating **h_hypoxia_eth:** number of genes from the hypoxia ethylene group this TF was seen regulating. **description:** description of Arabidopsis homolog which is closest related to apple gene **percent_identity, align_length, mismatch, gap, e_value, score**: information about arabidopsis to apple TF similarity **gene_symbol_long, gene_symbol_short** long and short Arabidopsis gene symbols **all_gene_symbols**: other symbols this gene has been identified as. *Included as a separate file.*

Supplemental Table 12: ERFVII Regulatory Network.
Included as a separate file.

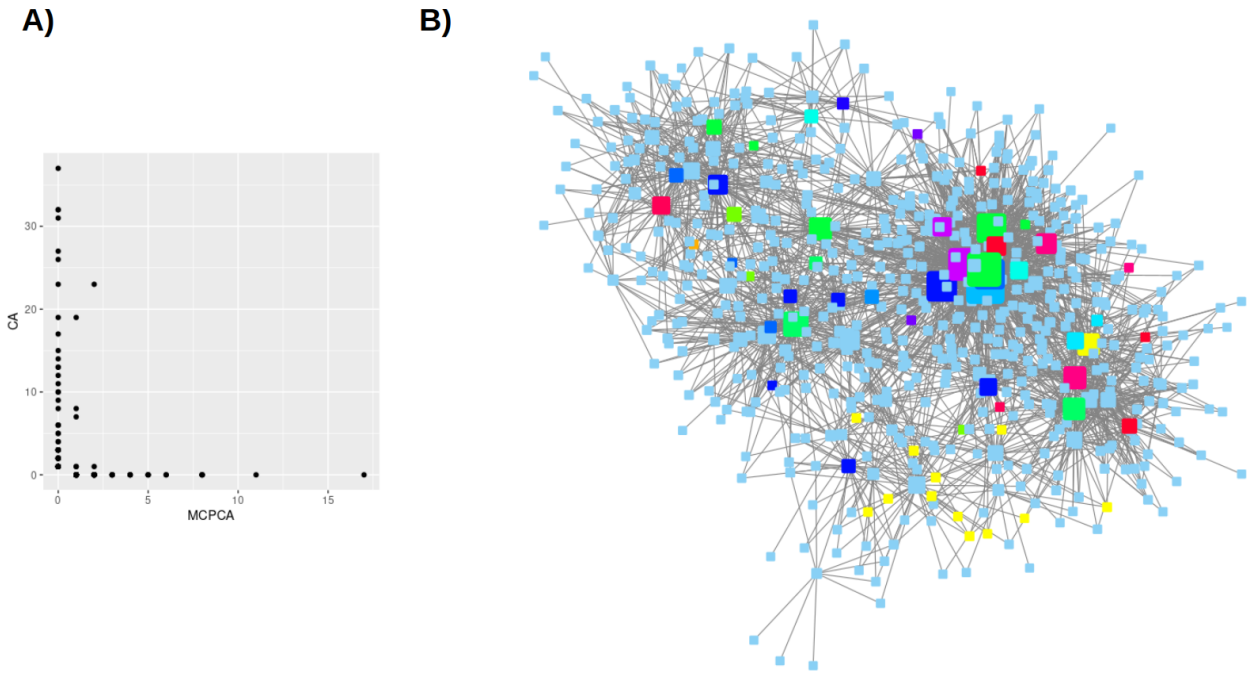
Supplemental Table 13: GO terms for each ERFVII gene putatively regulated.
Included as a separate file.

Supplemental Figures

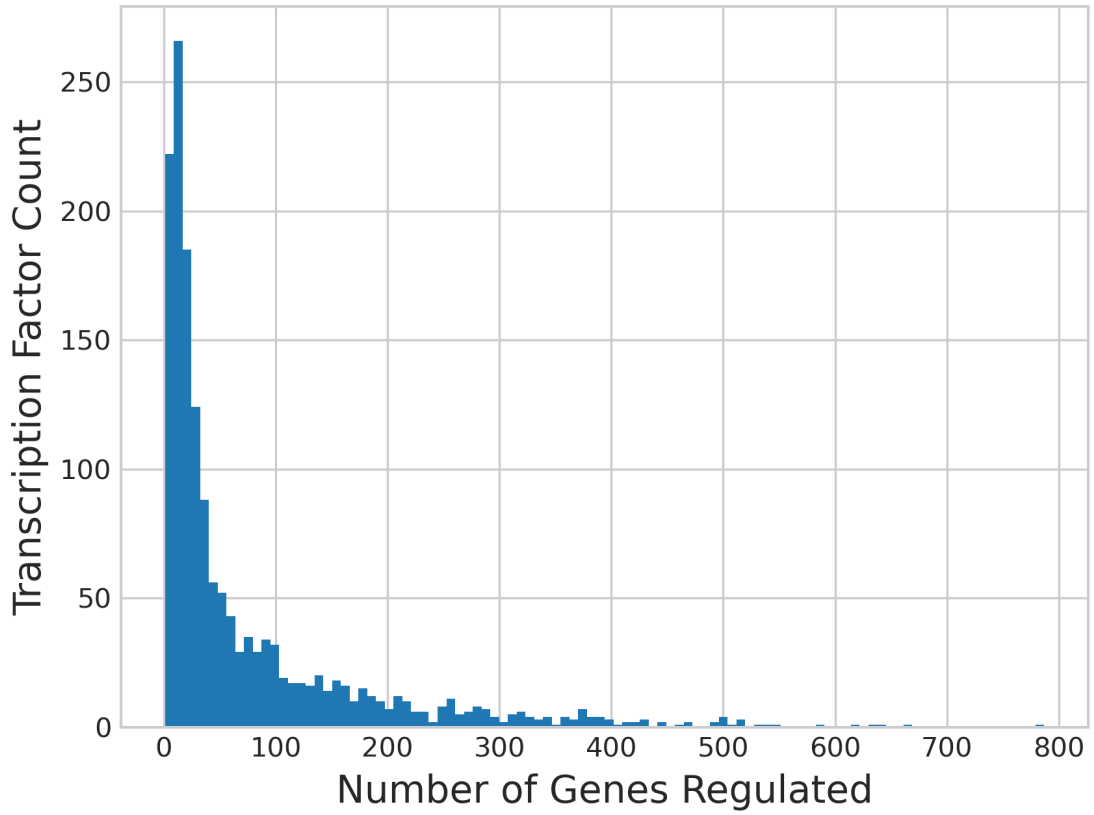


Supplemental Figure 1: r2 histogram distribution of genes in the regulatory network. x-axis is the r2 value and y axis is the count.

Supplemental Figure 2: Supplemental Legend for **Figure 5.**



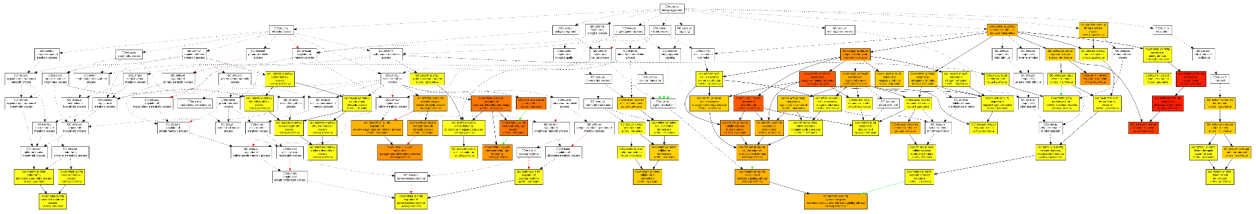
Supplemental Figure 3: Hypoxia Network Supplemental. **A)** Transcription factor regulation of either MCPCA hypoxia ethylene or CA hypoxia ethylene DEG. Notice there is very little overlap between the TF's predicted to regulate either category. **B)** MCPCA hypoxia ethylene upregulated module highlighted as yellow nodes in bottom portion of the graph.



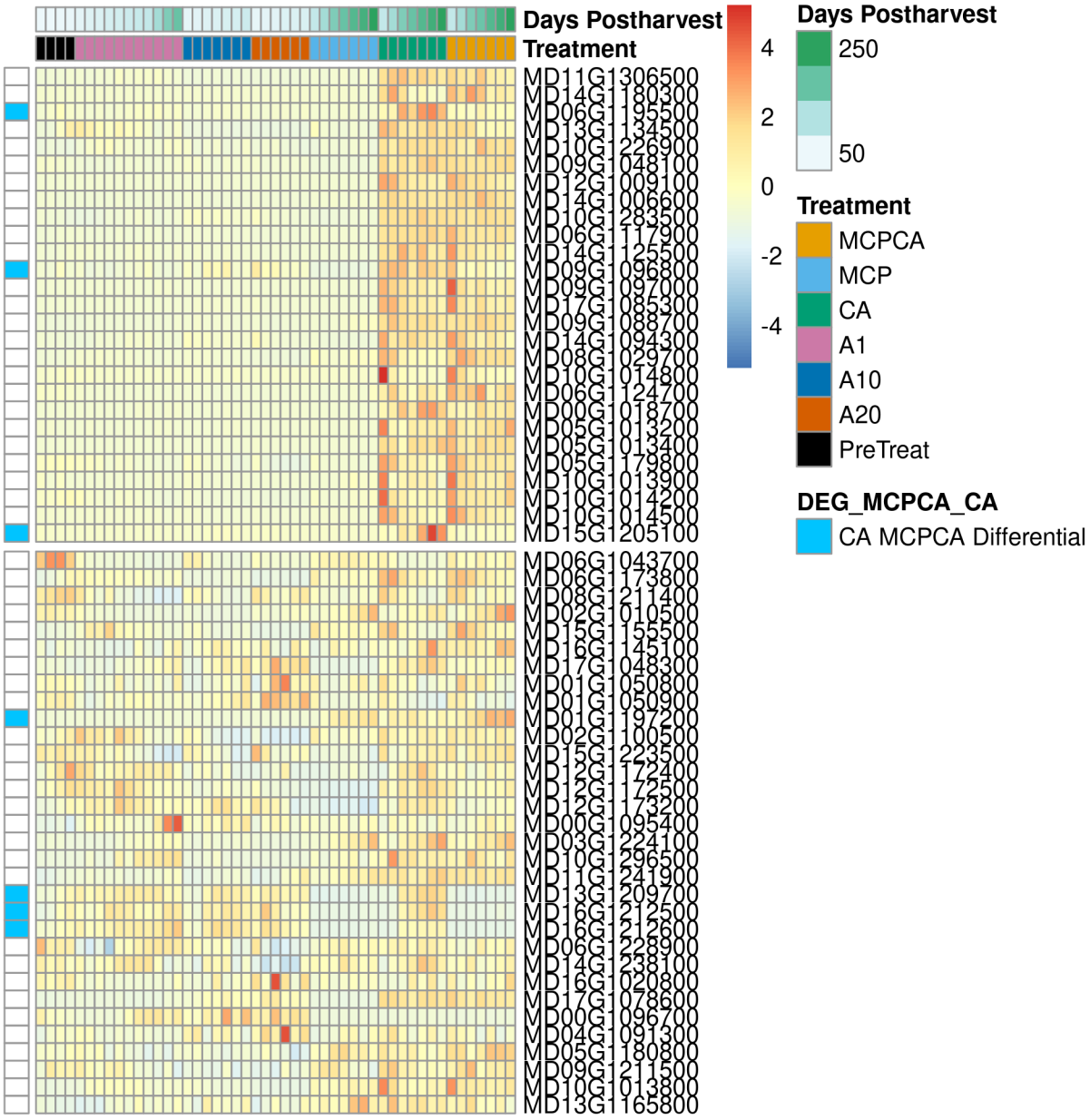
Supplemental Figure 4: Number of genes each Transcription Factor is regulating in the complete regulatory network thresholded at 5 potential regulators per gene.

Supplemental Figure 5: PCO phylogeny.
Included as a separate file.

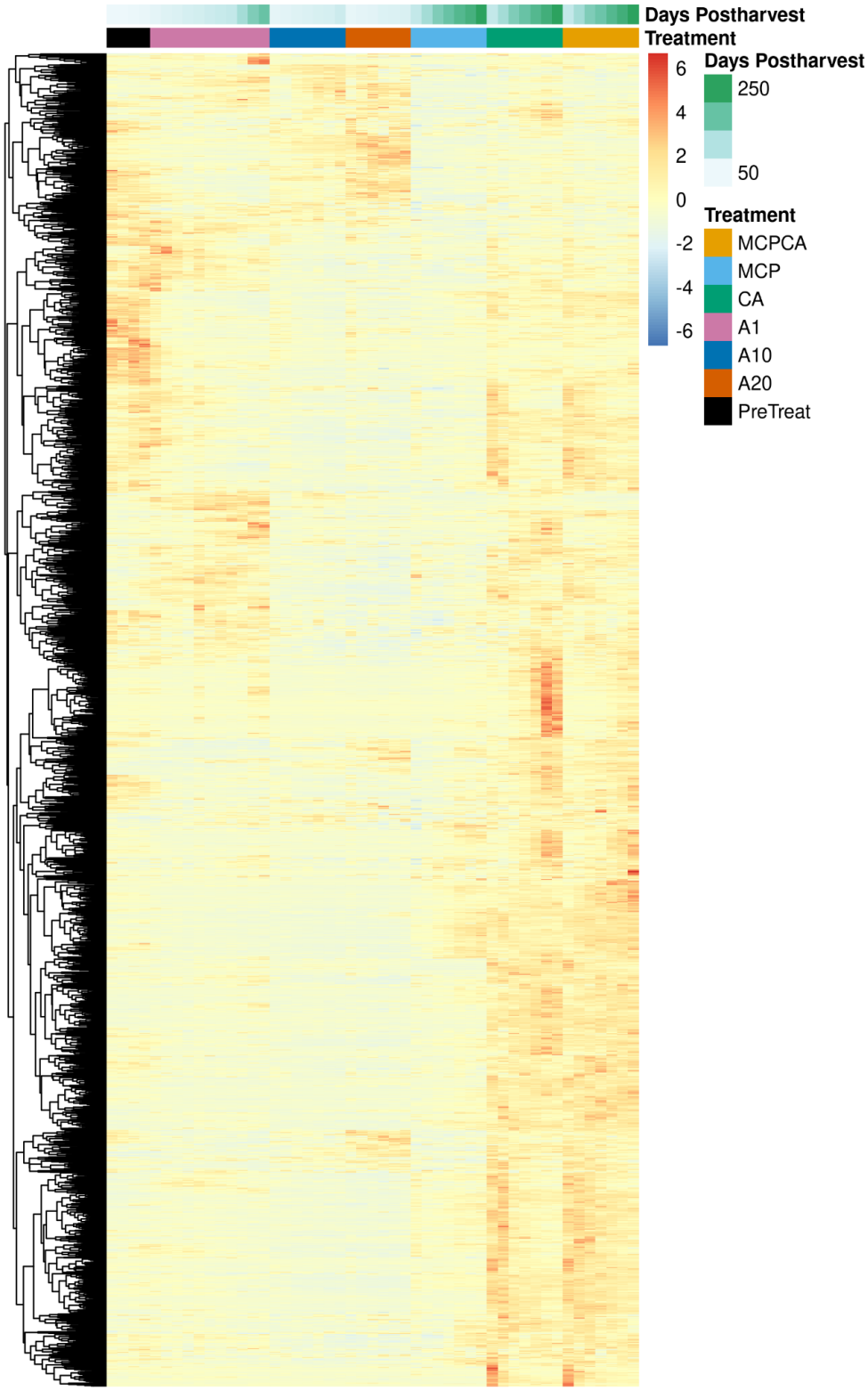
Supplemental Figure 6: ERF family phylogeny made using Plant Tribes.
Included as a separate file.



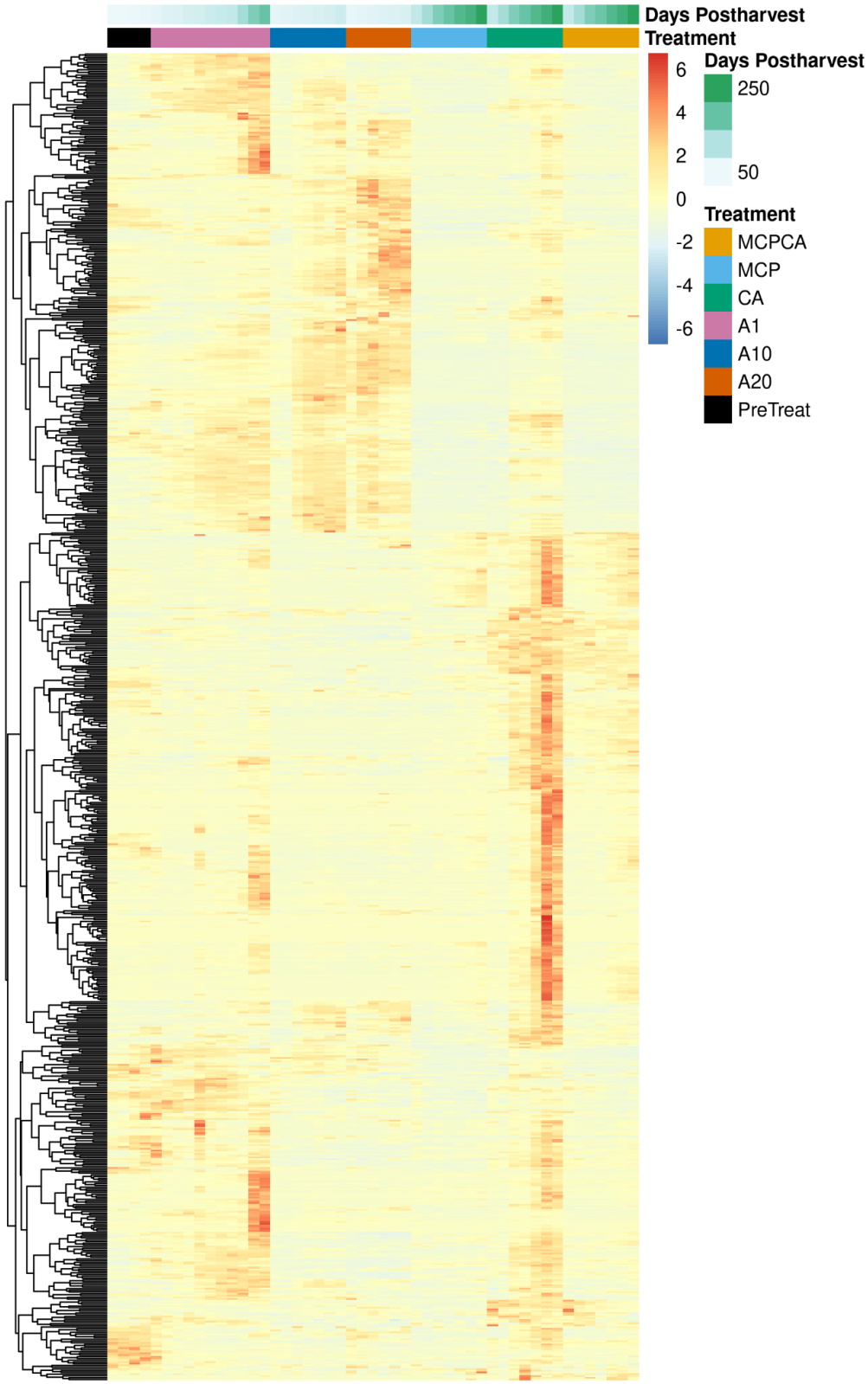
Supplemental Figure 7: GO term enrichment hierarchical Pclustering for the 606 DEG genes.
The full version is Included as a separate file.



Supplemental Figure 8: The 49 Mustroph et al. 2009 gene Apple homologues (out of 59 GDDH13 genome orthologs whose value was over 3 for at least 10 samples using the DESeq2 normalized dataset) shown as a heatmap. Split based on if the gene was shown to be DEG during hypoxic conditions in the apple dataset (top 27 rows) or if it was not (bottom 32 rows). Samples are columns and are the average of replicates for that column. They are organized by condition and time. Row annotation indicates if gene is DEG between the CA fruit dataset and the MCPCA fruit dataset (abs(log2foldchange > 1 and padj < 0.05)). From top to bottom, “CA MCPCA Differential” genes closest arabidopsis homologs are ACHT5 - AT5G61440, WIP4 - AT4G10270, ACO1 - AT2G19590, PP2-A13 - AT2G19590 and three copies of ETR2 - AT3G23150.



Supplemental Figure 9: Heatmap of ‘n-degron’ DEGs identified as upregulated in at least one time point.



Supplemental Figure 10: Heatmap of 'ethylene' DEGs identified as upregulated in at least one-time point.

CHAPTER THREE:

TOWARDS IDENTIFICATION OF POSTHARVEST FRUIT QUALITY TRANSCRIPTOMIC MARKERS IN *MALUS DOMESTICA*

Authors:

John A. Hadish[1,2], Heidi L. Hargarten[3], Huiting Zhang[2], Loren A. Honaas[3] , Stephen P. Ficklin[1,2,*]

[1] Molecular Plant Science Department, Washington State University, Pullman WA, 99164

[2] Department of Horticulture, Washington State University, Pullman WA, 99163

[3] USDA Agricultural Research Service Physiology and Pathology of Tree Fruits Research: Wenatchee, WA, 98801

*Corresponding Author. Email: stephen.ficklin@wsu.edu

Notification: This Chapter is currently in submission

Attributions

JAH: Writing - Original Draft, Writing - Review & Editing, Formal analysis, Methodology, Conceptualization **HLH:** Writing - Original Draft, Writing - Review & Editing, Investigation, Formal analysis, Conceptualization **HZ:** Formal analysis, Writing - Review & Editing **LAH:** Writing - Review & Editing, Supervision, Project administration, Funding

acquisition, Resources, Conceptualization **SPF:** Writing - Original Draft, Writing - Review & Editing, Supervision, Funding acquisition, Resources.

Abstract

Prognostic Transcriptomic Biomarkers (PTBs) are gene expression profiles that are representative of the underlying molecular biology of an organism. Gene expression profiles are highly impacted by the environment, including past events that affect developmental processes. As a result, it is expected that one or more genes could serve as a signal of a current or future phenotypic state to which their expression profile can be associated using statistical or machine learning approaches. Such an association results in a set of genes whose expression patterns become a PTB for populations that are genetically similar or for traits that are highly affected by the environment (low heritability). This makes PTBs suitable for monitoring traits of pome fruit such as Malus domestica (apple) because all individuals of a cultivar are clones, and differences in fruit quality--both at harvest and during the postharvest period--are largely due to the environment. The prediction of future fruit quality could enhance supply chain efficiency, reduce crop loss, and provide higher and more consistent quality for consumers. However, several questions must be addressed to determine if PTBs are viable, such as if common modeling approaches are robust; the minimum number of genes needed for a model to be predictive; and if technologies such as qPCR can be used as an inexpensive substitute for application of PTBs in a commercial setting. To address these questions we conducted a pilot study that sought to explore the potential for use of PTBs for fruit texture in the ‘Gala’ variety of apples, across several postharvest storage regiments. Fruit texture in ‘Gala’ apples is highly

controllable by postharvest treatments and thus it becomes a good candidate to explore broader use of PTBs. Results show that PTBs show promise for postharvest traits and that further research is justified to explore more complex, less controllable traits.

1 Introduction

Apples (*Malus domestica*) are one of the world's most consumed fruits, with over 4.8 billion kilograms produced annually in the United States (Gerlach, 2022). Advancements in breeding, orchard management, and postharvest storage technologies have made it possible to store apples for up to a year after harvest (Gapper et al., 2022). Despite this, producers and packing houses cull millions of kilograms of apples annually due to losses of fruit quality (e.g. loss of firmness and acid), including physiological disorders such as internal browning, bitter pit, superficial scald, and watercore (Gapper et al., 2022; L. A. Honaas et al., 2019; Nicolaï et al., 2006; Shewa et al., 2022). The propensity for losses in quality may not be apparent at harvest and instead develop after apples have been in storage for several months.

Improved predictions about the risk for losses in apple fruit quality could enhance efficiency throughout the supply chain, from the field to the consumer table. For example, apples at relatively high risk for losses in quality could be marketed first, reducing storage costs, and increasing pack-out among fruit lots. The current apple management toolkit is largely composed of physiological indices including starch clearing (Blanpied & Silsby, 1992), fruit firmness (Harker et al., 1996), peel color (Hamza & Chtourou, 2018), and acid and sugar content (Goffings, 1993). However, these methods are often insufficient to estimate risk for losses in quality; indeed these

limitations certainly play a role in the diversion of billions of kilograms of apple fruit from the fresh fruit market (USDA, National Agricultural Statistics Service, 2022).

Potential alternatives to physiological measurements exist, so-called prognostic biomarkers, that consist of biomolecules that are relatable to future fruit quality. Apple cultivars are clones of one another which means that differences in traits at harvest and after postharvest storage are due to environmental effects at harvest and during development. The apple fruit transcriptome can respond early and rapidly to changes in production and postharvest environments and could therefore be indicative of future fruit quality. Previously, the term prognostic transcriptomic biomarker (PTB) has been used as describing one or more genes whose expression profile is associated with the occurrence of a phenotypic trait (Pedrotty et al., 2012). In the present case, a PTB would be associated with a fruit quality metric, and changes thereof, during the postharvest period. Researchers can identify useful PTBs from RNA-seq data using statistical and machine-learning methods that identify expression profiles associated with complex phenotypic traits (Acharjee et al., 2020). These machine-learning methods do not use prior knowledge of molecular pathways and instead rely only on RNA-seq observations. This means that the identified PTBs are not biased by existing molecular knowledge, which can be incomplete. This is particularly valuable in non-model organisms where direct evidence of gene function is oftentimes lacking. Indeed, the majority of research using PTBs has been in the context of human medicine where they have been used to predict complex diseases such as cancer (Feng et al., 2019; Supplitt et al., 2021), heart disease (Deng, 2018), and Alzheimer's (Hadar & Gurwitz, 2018). However, there have been recent investigations into developing PTBs for predicting

commercially relevant apple disorders; biomarker tests were launched commercially in 2019 for risk assessment of bitter pit and soft scald in ‘Honeycrisp’ apples (Conklin, 2019; Karst, 2019; Prengaman, 2019). Although these particular markers have since been discontinued, interest in this area remains (Gapper et al., 2022). For example, a recent study sought to develop PTBs to distinguish among harvest times in ‘Royal Gala’ fruit as a method for determining an optimal harvest date (Favre et al., 2022). Despite these efforts, multiple questions remain, such as if common modeling methods are robust; the minimum number of genes needed for a model to be predictive; and if technologies such as qPCR can be used as an inexpensive substitute for application of PTBs in a commercial setting. The answers to the latter two questions are important to know if high-throughput assays can be cost-effective.

In this study, we seek to answer these questions by developing preliminary PTBs for predicting firmness in ‘Gala’ apples, a variety particularly susceptible to loss of firmness during storage (Volz et al., 2003). We chose firmness as a proof of concept for PTB identification for three reasons: (1) there already exist several candidate genes within the literature related to firmness identified by genomic means which we can use for model assessment (McClure et al., 2018), (2) firmness can be easily and accurately measured using a penetrometer (Harker et al., 1996), and (3) firmness is strongly impacted by postharvest treatments (Ganai et al., 2018; Kolniak-Ostek et al., 2014) making it controllable within our project's experimental design. Our experiment was designed to track changes in fruit texture across commercially relevant storage regimes. These included refrigeration, controlled atmosphere (DeLong et al., 1999), and the

ethylene perception inhibitor 1-Methylcyclopropene (1-MCP) (DeEll et al., 2002) identification of PTBs that are robust across different postharvest conditions.

We explore two common methods for association analysis, Elastic Net (EN) (Zou & Hastie, 2005) and Random Forest (RF) (Breiman, 2001) feature selection, for the identification of potential PTBs using a large RNA-seq and fruit quality dataset. We further verify our PTBs using Boruta feature selection and demonstrate that a relatively small set (15 PTBs) is sufficient for accurate predictions of loss of firmness within our dataset. Finally, we used qPCR to test the robustness of potential PTBs in a different lot of fruit not used for model development to explore the potential for qPCR being used as a lower-cost assay for models built with a large, global-scale gene activity data set.

2 Materials and Methods

2.1 Fruit harvest, sorting, and storage

'Gala' fruit was harvested from two different locations in two different years. Fruit harvested in Year 1 (2018) was used for RNA-Seq and model development for selecting PTB candidates, while fruit harvested in Year 2 (2019) was used for qPCR validation of the selected PTB candidates. A different orchard was selected for subsequent qPCR validation to determine if the candidate PTBs selected were robust to variations in different growing conditions experienced by the fruit.

Year 1 fruit was received from a commercial facility in Quincy, WA on August 21st, 2018. Upon arrival at the USDA-ARS Tree Fruit Research Laboratory in Wenatchee, WA, apples were randomly sorted by hand onto pressed fiber fruit trays holding 18 apples each. Trays were placed in cardboard boxes and stored in air at 1 °C for seven

days. This 7d conditioning period is a standard commercial practice (Lum et al., 2016) used to mitigate negative storage outcomes that can be associated with early application of controlled atmosphere and 1-MCP.

After seven days of conditioning, apples were randomly assigned into one of 6 possible treatment conditions, three “short-term” storage conditions and three “long-term” storage conditions.” The fruit designated for short-term storage were placed in normal air at either 1 °C (A1, n = 396), 10 °C (A10, n = 252) or 20 °C (A20, n = 216). Fruit designated for long-term storage were divided into three treatment categories (MCP, CA, MCPCA). Two-thirds of these fruits were treated with SmartFresh™ (AgroFresh Solutions, Inc., Philadelphia, PA USA), also known as 1-Methylcyclopropene (1-MCP), overnight. Post treatment, fruit were stored at 1 °C in either air (MCP, n = 252) or controlled atmosphere (MCPCA, n = 252; 2 % O_2, 1 % CO_2). The remaining one-third of the fruit not treated with 1-MCP were stored in a controlled atmosphere at 1 °C (CA, n = 252; 2 % O_2, 1 % CO_2). 1-MCP was applied at 1 °C and in accordance with SmartFresh™ product recommendations. Sampling points and experimental layout are illustrated in Figure 1 and Supplemental Table 1.

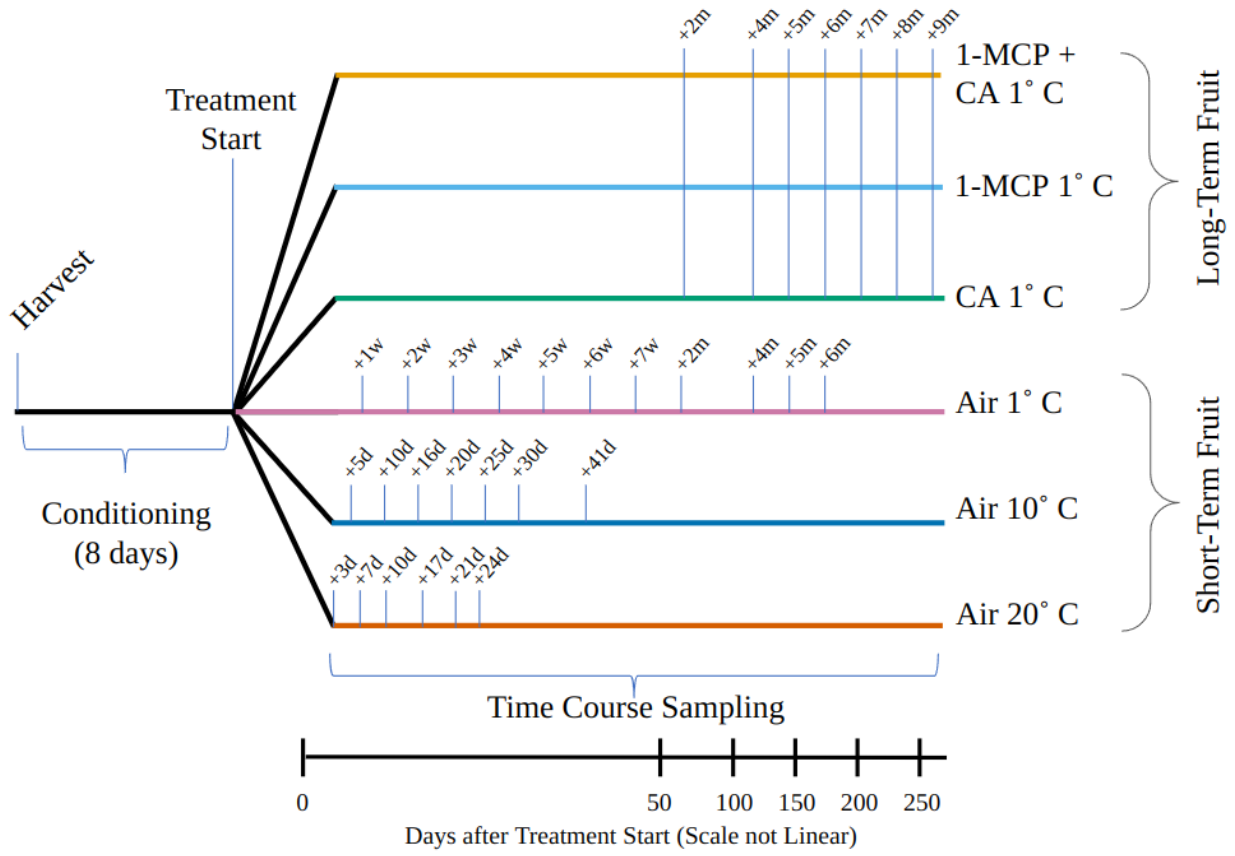


Figure 1: Sampling time points and treatments for 2018 RNA-seq data. Each time point includes the following samples: 3 biological replicates (6 apples per replicate) used for RNA-Seq, a texture analysis of 9 apple fruit at pullout, and a texture analysis of 9 apple fruit after a 7d ripening period (stored in air at 20 °C).

The following year, a second batch of fruit was harvested for PTB evaluation. These Year 2 ‘Gala’ apples were harvested from the Washington State University Tree Fruit Research and Extension Center’s Sunrise Research Orchard near Rock Island, WA on August 27th, 2019. These fruits were noted to be larger (at harvest diameter, Welch two sample t-test $p = 0.0077$, Supplemental Figure 1 A) and more mature (at harvest creep, Welch two sample t-test $p = 0.084$, Supplemental Figure 1 B) at harvest compared to fruit from Year 1. Year 2 fruit were treated and stored in a similar manner as Year 1, with the A20 experimental group omitted. For postharvest sampling timelines,

condition-relevant time intervals were chosen, as short-term fruit was expected to lose firmness faster than long-term fruit. See Supplemental Table 1 for a detailed description of experimental conditions, treatments, and sampling time points for Years 1 and 2.

2.2 Fruit quality, firmness, and tissue collection

For both experiment years, fruit cortex from 18 ‘Gala’ apples was harvested for RNA extraction for each treatment in three biological replicates of six apples each. Fruit was kept at its storage condition temperature (A20 = 20 °C; A10 = 10 °C; A1, MCP, MCPCA and CA = 1 °C) until the time of tissue harvest. CA fruit was removed from CA prior to tissue harvest (in air for ~ 30 min from removal to completion of tissue collection). Cortex tissue was harvested by first using a vegetable peeler to remove the peel of the apple around the apple’s equator (~ 2 cm wide) to expose cortex tissue. Then, using a knife, an ~ 0.5 cm wide disc was cut from the center of the apple. From this disc, three equal slices were taken around the core. Slices were coarsely diced and immediately flash-frozen in liquid nitrogen and stored at -80 °C. Frozen tissue was ground using a SPEX® Freezer/Mill® 6875 (SPEX®SamplePrep, Metuchen, NJ USA). Ground, frozen tissue was stored at -80 °C.

At each sampling time point, two additional sets of fruit were removed for texture analysis via the Mohr Digi-Test MDT-2 Penetrometer (MOHR Test and Measurement LLC, Richland, WA USA) using the standard 11 mm probe for apple fruit. The first set of fruit was measured at pull out (Year 1 n = 9; Year 2 n = 18), and the second set was placed in a dark room at 20 °C for 7 d to simulate time in a supply chain and assessed texture after this ripening period (Year 1 n = 9; Year 2 n = 18).

2.3 RNA extraction, quality control, and transcriptome sequencing

RNA was extracted from the ground frozen tissue using a CTAB/Chloroform protocol modified for use on pome fruit tissue in the postharvest period (L. A. Honaas & Kahn, 2017). Extracted RNA was analyzed for purity using the NanodropOne (Thermo Fisher Scientific, Waltham, MA USA) for integrity on the Agilent Bioanalyzer (Agilent, Santa Clara, CA USA, Agilent-RNA Pico Kit, Cat #: 5067-1513), and quantity using the Invitrogen™ Qubit™3 (Thermo Fisher Scientific, Waltham, MA USA, Qubit™ RNA HS Assay Kit, Cat #: Q32852). Only RNA that met the following standards was used for downstream RNA-Seq and qPCR: A260/A280 ≈ 2.0, RNA Integrity Number (RIN) ≥ 8.0.

To generate RNA-Seq data, libraries using Lexogen's QuantSeq 3' mRNA-Seq Library Prep Kit FWD (Cat # 015; www.lexogen.com) were prepared at the Penn State Genomics Core Facility (University Park, PA, United States) as described in (L. A. Honaas et al., 2019). Libraries were sequenced on a 150 base pair single-end protocol to a target volume of ~ 8-10 million reads per biological replicate on Illumina's HiSeq 2500 in Rapid Mode. Read data are publicly available at the Sequencing Read Archive (BioProject PRJNA938164).

2.4 Processing of RNA-seq data for analysis

Raw RNA-seq data was preprocessed with Trimmomatic (Bolger et al., 2014) to remove the leading 12 nucleotides. This trimming is recommended for QuantSeq 3' FWD sequencing prior to genome alignment (Lexogen, 2020). Transcripts were then aligned and counted using the GEMmaker workflow (Hadish et al., 2022) using the

Hisat2 (Kim et al., 2015). The ‘Golden Delicious' doubled-haploid genome (GDDH13) (Daccord et al., 2017) was downloaded from the Genome Database for Rosaceae (GDR) (Jung et al., 2019) and used for alignment. The Hisat2 option in GEMmaker automatically runs the following bioinformatic tools: Fastqc (Andrews, 2010), Trimmomatic (Bolger et al., 2014), Hisat2 (Kim et al., 2015), Samtools (Li et al., 2009), Stringtie (Pertea et al., 2015), and MultiQC (Ewels et al., 2016). The output of this workflow is a Gene Expression Matrix (GEM) with counts reported in Transcripts per Million (TPM). The average read alignment was 71 %, with an average of 53 % of reads assigned unambiguously. The multiQC (Ewels et al., 2016) report of the alignment is included in Supplemental Table 2. Seven samples were removed due to low alignment (Supplemental Table 2). A total of 46559 genes are present in the GDDH13 genome, but we did not have alignment to all of these with our dataset. Any gene which had zero RNA-seq reads was removed from this analysis. The final GEM consisted of 128 samples and 32303 genes (Supplemental File 1).

2.5 Random Forest modeling of RNA-seq samples

The GEM and the Overall Average Hardness post-simulated supply chain (OAH post) measurements were used for modeling firmness using RandomForestRegressor from the sklearn package (version 1.1.3) (Pedregosa et al., 2011). OAH post was used as the dependent (target) variable, and the expression values of all genes in the GEM were used as the independent (explanatory) variables. While the industry uses M1 and other metrics (including aggregated and/or proprietary ones), we use OAH because it is a convenient proxy for fruit texture and offered more contrast in our experiment than

other metrics reported by the MORH texture analyzer. It is therefore an appropriate metric to use for exploration of biomarkers that may be viable PTBs.

The RF full model (RF-fm) used all 32303 genes and was bootstrapped 100 times. Parameters for the RandomForestRegressor were: n_estimators = 1000, max_features = 0.5, and min_samples_leaf = 5. The feature importance of each gene was recorded after each run, and the total feature importance was calculated as the sum of feature importance over 100 runs. The top 15 genes based on summed feature importance were selected as a reduced sample set to be used for both stability measurements and for the RF reduced model (RF-rm).

The RF-rm was created using the same parameters as the RF-fm with the exception that only the top 15 genes from the RF-fm were used. This was intended to use the best genes from the previous model while reducing the chance of overfitting due to too many features. The reduced feature set also represents a more realistic number of genes that could be sampled in a commercial setting.

2.6 Elastic Net modeling of RNA-seq samples

Elastic Net (EN) feature selection was performed using the sklearn packages ElasticNetCV (version 1.1.3) (Pedregosa et al., 2011). As in RF, OAH post was used as the dependent variable, and TPM expression values were used as independent variables. For the full model, all 100 bootstraps were performed using randomized train test splits for each run. l1 ratios searched were 0.1, 0.5, 0.7, 0.9, 0.95, 0.99 and 1.0. Other parameters included cross-validation of 5 (cv = 5) and a maximum of 1000

iterations (max_iter = 1000). Like the RF-fm, the EN full model (EN-fm) used all 32303 genes during feature selection.

The coefficients of the best model were recorded after each iteration. The absolute value of these coefficients was taken and then normalized to a total of 1 to make coefficients comparable to RF's feature importance. The top 15 genes were those with the highest total normalized coefficients. The top 15 genes were then used to create the EN reduced model (EN-rm). Besides the number of genes, the EN-rm was created using the same parameters as the EN-fm.

2.7 Stability measurements

The stability of both RF and EN models was assessed using techniques outlined in (Harrell, 2022). In short, genes were numerically ranked in each run according to either their feature importance (RF) or coefficients (EN). For each of the 100 runs of EN-fm and RF-fm, the top 15 genes ranks were plotted. The rank of an individual gene is expected to remain relatively the same if a model is stable and to change drastically if a model is unstable.

2.8 Boruta feature selection of samples

Boruta Random Forest (BRF) feature selection was performed on the mean OAH post measurements with the GEM. BRF does not attempt to make predictive models, but instead uses the entire dataset to see which features perform better than a randomized "shadow feature" (Kursa & Rudnicki, 2010). We use it here to verify feature predictions selected from previously discussed models. BorutaPy (*BorutaPy*, n.d.; Kursa

& Rudnicki, 2010) was used with the following parameters: max_iter = 200, perc = 90. BRF was run 100 times on bootstrapped resampling of the data set.

2.9 Gene of interest selection for qPCR validation

2.9.1 Criteria for gene selection

The top 45 genes identified as being predictive of texture loss from the RF regression model were selected and classified into orthogroups pre-computed with the 26Gv2.0 scaffold using PlantTribes2 (Wafula et al., 2022). Orthogroup multiple sequence alignment, phylogenetic tree estimation, homology inference, and gene model evaluation were performed using genes from 16 Rosaceae genomes [the same 15 from (Zhang et al., 2022) plus *Malus baccata* (Chen et al., 2019)] plus the scaffolding species following methods from (Zhang et al., 2022). These top 45 genes were further filtered for primer development following criteria from Honaas et al. (2021), giving priority to genes in (in no particular order): 1) small orthogroups (ideally < 15 members in *Prunus persica*), 2) high expression, 3) low variance between biological replicates, and 4) those with linear expression profiles. From these criteria, a set of 15 genes were selected for primer development. To guide the selection of regions for targeted primer development, orthogroup multiple sequence alignments produced by PlantTribes2 were visualized and manually examined in Geneious R9 (Kearse et al., 2012). Only gene regions with highly homologous sequences across apple cultivars were selected for primer design. The PlantTribes2 Orthogroup Classification and Annotations for these 15 genes in Supplemental Table 3.

2.9.2 Primer development and qPCR

Primers were developed in Geneious using the Primer3 plug-in [v2.3.4 (Untergasser et al., 2012)], following parameters detailed in Supplemental Table 4. For candidate genes with highly similar homologous sequences, primer development was targeted to specific regions in alignment with the highest dissimilarity. Candidate genes and their primer characteristics can be found in Supplemental Table 5 (CDS and primer alignments in Supplemental File 2). Reference genes previously used in (Hargarten et al., 2018) and (L. A. Honaas et al., 2019) were selected from the literature: MDP0000274900 (Perini et al., 2014), MDP0000173025 (Bowen et al., 2014), and MDP0000223691 (Storch et al., 2015). The GDDH13 homologous sequences for these reference genes were identified, using PlantTribes2, as MD09G1190100, MD16G1209000, and MD15G1211100 respectively (Zhang et al., 2022). Primers were synthesized by Integrated DNA Technologies (IDT, Coralville, IA), dissolved in qPCR-grade water (catalog no. W4502; Sigma-Aldrich, St. Louis, MO) to produce 100 µm solutions, and stored at –20 °C. qPCR was performed as described in (Hargarten et al., 2018) for ‘Granny Smith’ on a subset of 33 Year 2 samples (corresponding to similar early and late postharvest storage time points from each experimental treatment and condition in Year 1 - Supplemental Table 6) with a slight modification to the protocol: the qPCR reaction volume was increased to 15 µL [to accommodate automated liquid handling by an epMotion 5073 (Eppendorf, Hamburg, Germany)] by increasing the volume of SYBR per reaction while maintaining the template mass per reaction (10 pg cDNA).

2.9.3 qPCR post-processing for normalized expression

Amplification and melt curves for each triplicated group of technical replicates were manually inspected for Ct variance and melt curve anomalies. Individual technical replicates were removed from downstream analyses if Ct variance was > 1.5. Next, reaction efficiency was calculated based on raw amplification data using the R v4.1.2(R Core Team, 2021) package 'qpcR' v1.4-1 (Ritz & Spiess, 2008). First, statistical model selection was performed for sigmoidal fit testing of the raw real-time PCR data for the reference gene PCR runs using the *mselect*() function with comparing nonlinear sigmoidal models (l4, l5, b4, b5) and exponential models (expGrowth, expSDM, linexp), a bilinear model (lin2), and a mechanistic model (cm3) (Spiess et al., 2015). The *mselect*() function had the following parameters set: fctList = list(l5, l4, b5, b4, cm3, lin2, linexp, expGrowth, expSDM), crit = "weights". The best overall models (lin2 and linexp) were selected based on model goodness of fit Akaike Information Criterion (AIC) and r^2. Using the best model, the *modlist*() function was rerun, with the following parameters set: remove = "none", smooth = "spline". Threshold cycles were determined using the 'Cy0' method (Guescini et al., 2008) to calculate efficiency using the *pcrbatch*() function. In the *pcrbatch*() function output, the efficiency was calculated using the best overall model (lin2 or linexp) on a gene-by-gene basis. Computed reaction efficiencies were entered into their corresponding PTBs in BioRad's CFX Maestro 1.0 software (4.0.2325.0418) and used to calculate relative normalized expression values with the Pfaffl method (Pfaffle 2001) and three reference genes. Reference gene stability was assessed using the CFX Maestro software. One reference gene, MD15G1211100, had

moderate variation across the samples indicating less than ideal stability. This reference gene was removed from analysis prior to normalized expression computation.

2.10 Literature genes random forest

Genes were selected from previous literature concentrating on loss of firmness and texture in apple fruits (Chang & Tong, 2020; Hu et al., 2020; Migicovsky et al., 2021; Wu et al., 2021). If necessary, literature gene names were converted from other apple genomes nomenclature [i.e. (Velasco et al., 2010)] to GDDH13 (Daccord et al., 2017) using homology through OrthoFinder (Emms & Kelly, 2019). In total, 98 genes from the literature were identified (Supplemental Table 7). 85 of these genes had at least one read aligned to them in the GEM. and were used to create RF models in the same manner as previously discussed. 100 bootstrap replicates were performed using the 85 gene set, and the total feature importance of each gene set was determined. The genes with the top 15 summed feature importance were then used in a new model, with model performance and feature importance being calculated.

3 Results and Discussion

3.1 Firmness loss

Firmness declined more rapidly in the short-term fruit than in the long-term fruit. Fruit stored at room temperature of 20 °C (A20) lost firmness most quickly, followed by fruit stored at 10 °C (A10) and 1 °C fruit (A1). Long-term fruit maintained firmness throughout the experiment (Figure 2). These firmness trends are consistent with previous knowledge of apple ripening (Bai et al., 2005).

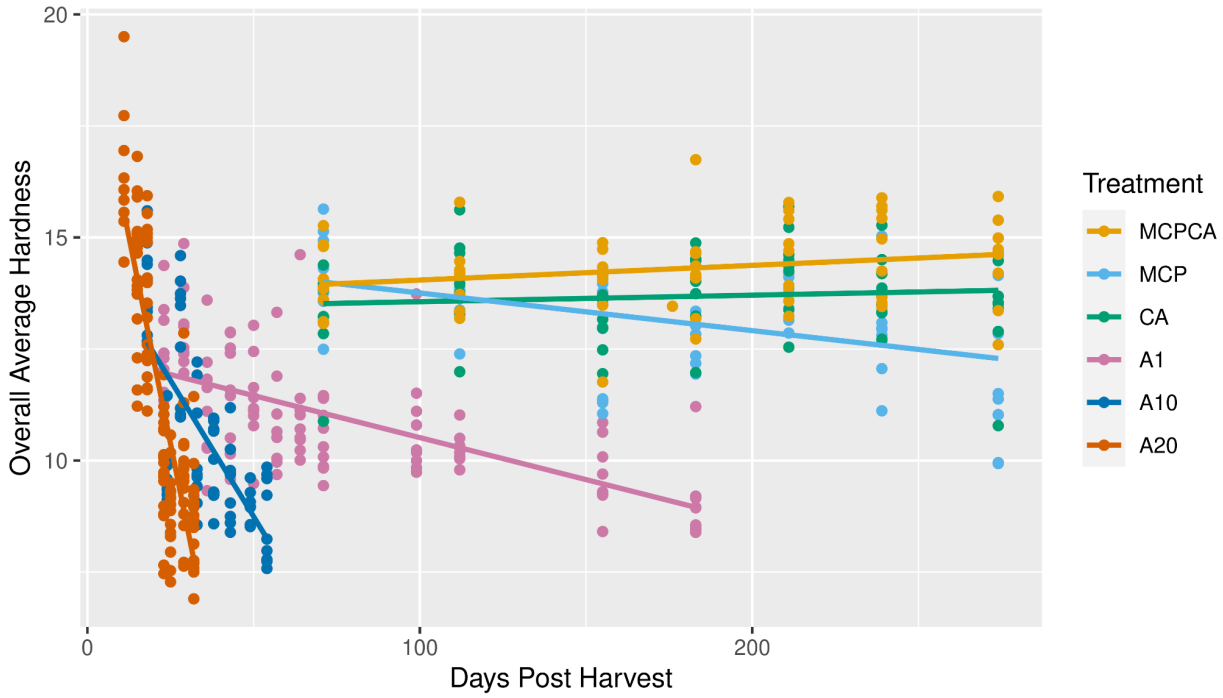


Figure 2: Overall average hardness after a 7d simulated supply chain (in air at 20 °C).

3.2 Transcriptomic data pre-processing

A principle component analysis (PCA) of the TPM (transcript per million) read counts shows that the expression data is separated primarily based on storage temperature (1 °C, 10 °C, 20 °C) along PC1, and by storage condition (1-MCP and CA) along PC2 (Figure 3A). This shows that we have variance within the transcriptome that describes both experimental conditions and days after harvest. Coloring with Days Postharvest also shows considerable variation over the PCA plots (Figure 3B).

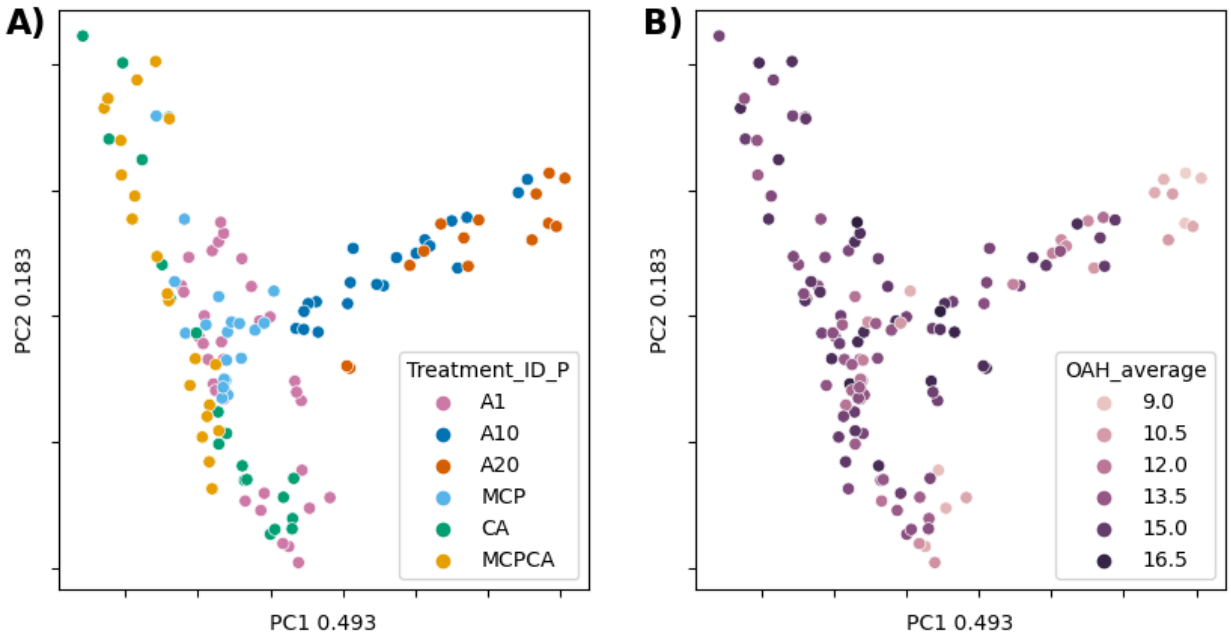


Figure 3: PCA 1 and 2 of TPM transcriptomic data counts colored by A) treatment and B) days postharvest. PC1 primarily separates fruit based on temperature and days postharvest of the warmer fruit whereas PC2 separates based on days postharvest of fruit treated with 1-MCP and CA

3.3 Model performance - random forest vs. elastic net

The first step was to create models using the full feature set of genes which we refer to as full models. Model performance for the Random Forest full model (RF-fm) and the Elastic Net full model (EN-fm) were comparable (Table 1), with testing r^2 for EN-fm (r^2 = 0.767 ± 0.099 SD) performing better than RF-fm (r^2 0.687 ± 0.124 SD). A visualization of a single run of both EN-fm and RF-fm is visualized in Figure 4 which is split into training (Figure 4 **A** for RF-fm and Figure 4 **C** for EN-fm) and testing (Figure 4 **B** for RF-fm and Figure 4 **D** for EN-fm) sets.

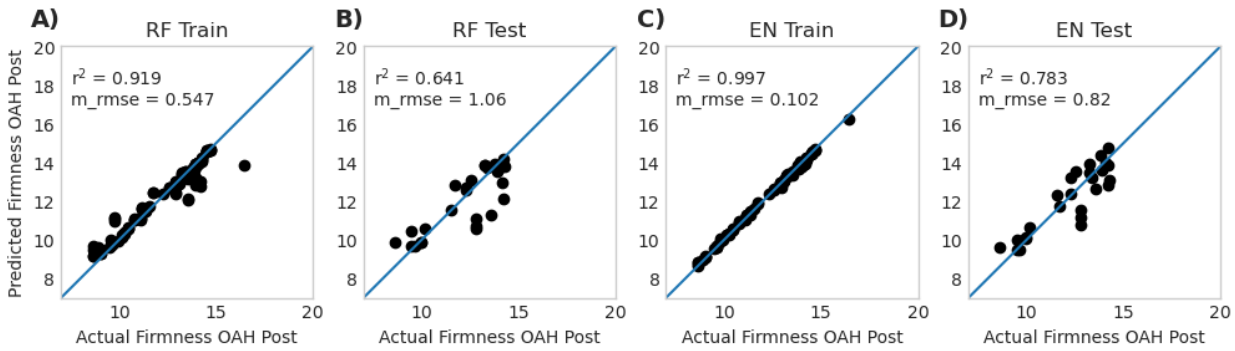


Figure 4: Model performance for a single run of Random Forest full model (RF-fm) **(A** and **B)** and Elastic Net full model (EN-fm) **(C** and **D)** using all genes. The data used in each model is split between train (RF-fm **A** and EN-fm **C**) and test data (RF-fm **B** and EN-fm **D**). Data for replicated 100 runs of these models is presented in Table 1. Reported r^2 and m_rmse values in this figure represent a single run of a representative model, whereas data reported in Table 1 represents the average of 100 replicates.

Table 1: r^2 and m_rmse values with standard deviation of 100 bootstrap runs for all models. For Random Forest and Elastic Net 'All Genes' summary, the 'Full Models' use all genes from data (32303) whereas the 'Reduced Models' use only the top 15 genes from the full model. The Random Forest Literature full model was done with 85 genes referenced in Supplemental Table 7, and the reduced model was run with only the top 15 genes from this subset. The models run for qPCR used only the 15 genes selected for qPCR, therefore a reduced model was not applicable. Performance metrics are reported with ± standard deviation.

Random Forest All Genes				
	r^2 Train	m_rmse Train	r^2 Test	m_rmse Test
Full Model	0.928 ± 0.007	0.506 ± 0.025	0.687 ± 0.124	1.02 ± 0.216
Reduced Model	0.887 ± 0.012	0.633 ± 0.039	0.727 ± 0.099	0.954 ± 0.19
Random Forest Literature Genes Only				
	r^2 Train	m_rmse Train	r^2 Test	m_rmse Test
Full Model (85 Genes)	0.897 ± 0.008	0.606 ± 0.027	0.68 ± 0.134	1.02 ± 0.215
Reduced Model	0.882 ± 0.009	0.647 ± 0.026	0.711 ± 0.106	0.992 ± 0.209
Random Forest qPCR Genes Only				
	r^2 Train	m_rmse Train	r^2 Test	m_rmse Test
15 Gene Model	0.897 ± 0.011	0.604 ± 0.03	0.748 ± 0.077	0.925 ± 0.15
Elastic Net All Genes				
	r^2 Train	m_rmse Train	r^2 Test	m_rmse Test
Full Model	0.949 ± 0.057	0.311 ± 0.295	0.767 ± 0.099	0.889 ± 0.184
Reduced Model	0.85 ± 0.011	0.731 ± 0.025	0.784 ± 0.061	0.845 ± 0.098

In addition to the RF-fm and EN-fm, Boruta Random Forest (BRF) was also performed. BRF selects “all relevant” genes through the use of randomized shadow features. If a gene does not perform better than a shadow feature for predictions, it is eliminated (Kursa & Rudnicki, 2010). BRF does not attempt to predict firmness but instead concentrates on identifying all relevant features which are able to predict firmness better than a shadow feature. Out of 100 bootstrap runs of BRF on randomized train test sets, 51 genes were selected as performing better than shadow features (Supplemental Table 8). Twelve of the top 15 genes from the RF model (based on feature importance rank) were present in the BRF 51 genes set. This overlap indicates that for this data the feature selection for RF prediction models is consistent with BRF feature selection models (Speiser et al., 2019). A strong overlap indicates that the features selected by RF are sufficient for predicting actual firmness in this dataset rather than just fitting on noise. None of the top 15 genes from EN (based on normalized coefficients) were represented within the BRF gene set.

3.4 Model stability

While RF-fm and EN-fm performed similarly in terms of r^2 and m_rmse, stability differed. The stability of each model was determined by bootstrap re-sampling the data and re-running the model 100 times (Harrell, 2022). After each run, features were ranked by importance, and the variance of the ranks of the top 15 RF-fm and EN-fm genes was explored. The RF-fm was more stable than the EN-fm when compared across 100 runs. The feature importance ranks these runs are visualized in Figure 5A

for the top 15 genes of the RF-fm and Figure 5B for the EN-fm. Each point in the figure represents the rank of the gene from a single bootstrap run. Genes with high importance for firmness will have a lower rank and appear as a point on the left-hand side of the plot and the variance of rank can be visualized by the spread of points along the x-axis. The left-hand y-axis lists the top 15 genes in the model. In some of the bootstrap runs, the top 15 genes were not ranked within the top 500 genes. The frequency that each gene appeared in the top 500 is indicated on the right-hand y-axis. RF-fm genes on average appeared in 82.07 (+/- 8.83 SD) out of 100 bootstrap runs, whereas EN-fm genes appeared in 45.87 out of 100 (std 29.78) which was significantly less than the RF-fm (Unpaired t-test $p < 0.0001$). The minimum number of times one of the top 15 genes was selected in the RF-fm was 71, compared to 13 in the EN-fm. Figure 5 also shows that top genes in the EN-fm had a larger spread of rank when compared to RF-fm which displayed lower variability for all genes. Neither method selected all top genes in respective models every time.

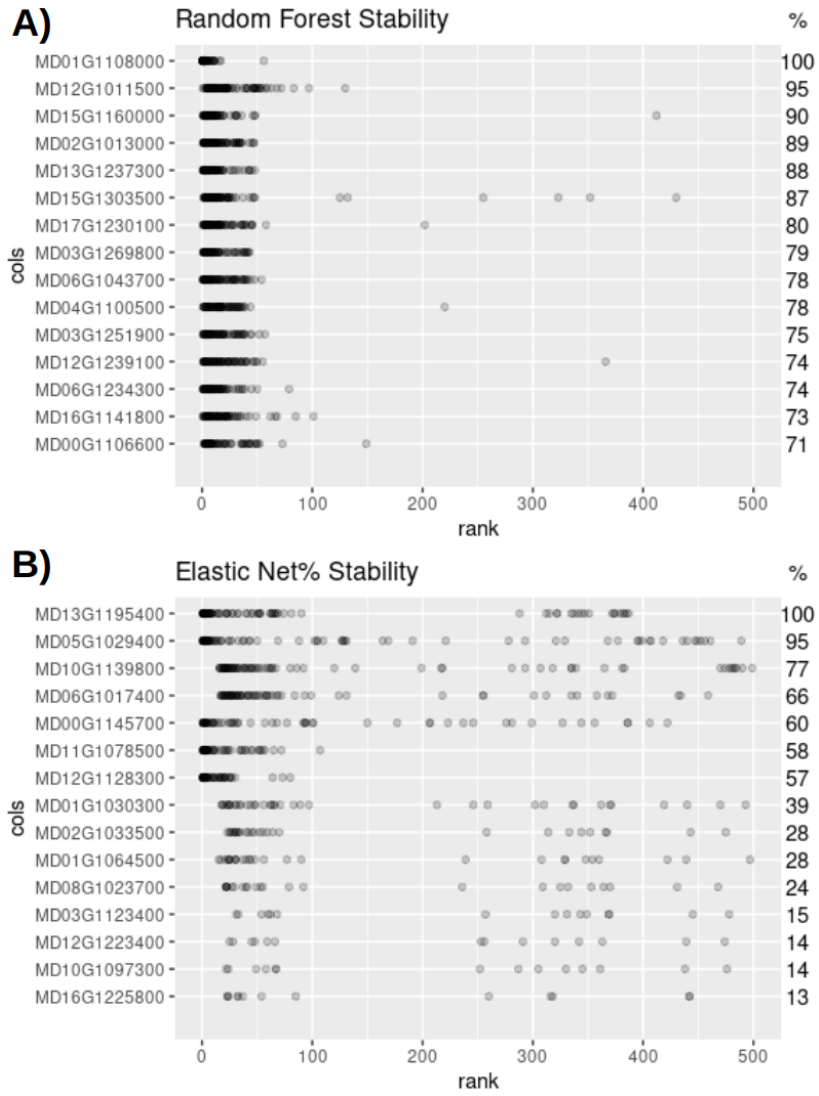


Figure 5: Model stability of A) Random Forest (RF) Full Model (FM) and B) Elastic Net (EN) FM. The top 15 genes of each model are shown along the y-axis. Each point represents a single bootstrap re-run of the model and its position along the x-axis is the gene's rank in importance from the re-run model. A rank of 1 is given to a gene if it is the most important in the respective model. A point is present for a re-run model only if the gene occurred in the top 500 important genes. Numbers on the right side y-axis of the graph indicate how many times the gene was selected by the model.

3.5 Model performance of top genes

The top 15 genes from each model were used to create new models, referred to as reduced models, to explore if a model using a small subset of genes could perform

as well as the full model (32303 genes). This was done for two reasons, first, to simulate a realistic number of genes that could be sampled economically in a commercial setting, and second to help improve model performance. A smaller feature set can cause a model to be more generalizable and reduce the chance of overfitting (Menze et al., 2009). Generalizability is desirable as this can make models more robust to inherent variability in novel testing data due to, for example, environmental variation among years and orchards.

Both EN-rm and RF-rm had increased performance within the testing data when compare with EN-fm and RF-fm (Table 1), with RF-rm achieving r^2 of 0.727 ± 0.099 in the testing set and EN-rm achieving r^2 of 0.784 ± 0.061. The reduced number of genes allowed the models to be more generalizable than full models and illustrate the importance of feature reduction when dealing with datasets that have many features. The performance of a single run that is representative of most runs of both RF-rm and EN-rm is visualized in Figure 6.

Both reduced models performed similarly, but there were differences in the genes selected for each. The genes which were selected for RF-rm and EN-rm do not overlap with each other and exhibit different expression patterns. Several of the genes selected by EN have low expression levels. The TPM expression levels of these top 15 genes are visualized in Supplemental Figure 2 for RF and Supplemental Figure 3 for EN.

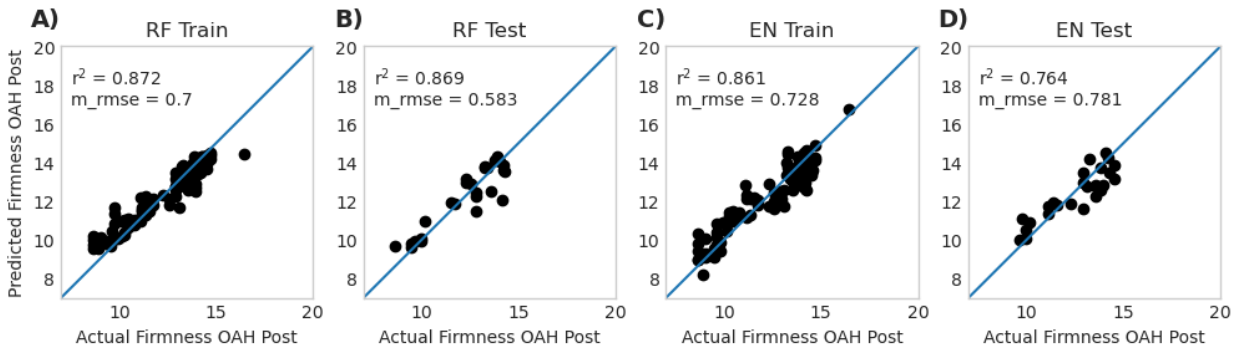


Figure 6: Model performance of a single random forest reduced model (RF-rm) and elastic net reduced model (EN-rm) training **A** and testing **B.** training **C** and testing **D**. Data for replicated 100 runs of these model is presented in Table 1. Reported r^2 and m_rmse values in this figure represent a single run of a representative model, whereas data reported in Table 1 represents the average of 100 replicates.

3.6 Literature genes random forest

Random Forest model performance was also assessed for a set of 85 genes related to firmness and other texture traits identified from the literature. There were no genes identified from the literature that also appeared in the top 15 genes identified by any of the previous models described thus far. Literature genes were assessed using two models. The full model used the 85 genes referenced in Supplemental Table 7 and the reduced model used the top 15 genes from the initial 85-gene model. Both models were comparable, with statistics on r^2 and m_rmse present in Table 1. The performance of a single run of the literature-reduced model is visualized in Figure 7. Both models had slightly lower performance than models created using the entire dataset (Table 1). Gene expression patterns and gene names from literature for the top 15 literature genes used in the reduced model can be seen in Supplemental Figure 4.

We suspect that the slightly lower performance of the models using genes from the literature may be due to some of the genes being transcription factors that are only

turned on for short periods of time and that this gene set is expressed prior to firmness loss (Chang & Tong, 2020; Hu et al., 2020; Migicovsky et al., 2021; Wu et al., 2021). However, the fact that these genes were able to perform comparably to our models indicates the depth of information contained within RNA-seq datasets and the power of RF models.

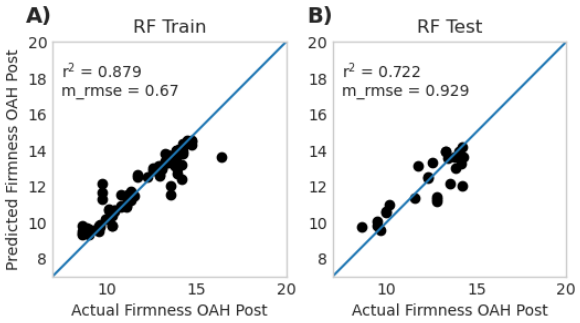


Figure 7: Random Forest model for the top 15 literature genes selected from the literature gene full model (85 genes). **A** is training data and **B** is testing data. The literature models had reduced performance compared to models using all genes. Data for replicated 100 runs of this model is presented in Table 1. Reported r^2 and m_rmse values in this figure represent a single run of a representative model, whereas data reported in Table 1 represents the average of 100 replicates.

3.7 Exploration of qPCR for model evaluation

3.7.1 qPCR Validation Data

A commercially viable PTB must consist of a tractable number of targets, in this case mRNA transcripts, in order for the test to be feasible. Here, we performed a targeted evaluation using qPCR of 15 genes (selected from the RF-fm) in a fully independent replication of the initial experiment (i.e. Year 2 samples). A positive, significant correlation between RNA-seq and qPCR measurements would indicate that genes selected by our models are more likely to be robust across different sample sets and could be investigated further for future PTB development. Eight of the 15 selected

genes had Pearson correlation coefficients over 0.67, with three of these over 0.90. However, the other seven genes evaluated showed low, non-significant correlations ($p > 0.05$). These trends comparing the qPCR data from 2019 and the transcriptome data from 2018 are visualized in Supplemental Figure 5.

To explore why some genes may have a higher agreement than others, we considered expression patterns between RNA-Seq and qPCR on a treatment-by-treatment basis (Supplementary File 3). Generally, when considering all PTBs and the overall expression patterns, the concordance is low, as indicated by global correlation analyses above. However, as we focus in on specific treatment comparisons, there are apparent patterns. Fruit stored in air and not treated with 1-MCP or stored in CA had generally had higher agreement between RNA-Seq and qPCR measurements, with the A10 treatment having the most similar expression patterns overall. When CA and 1-MCP treatments were considered, the agreement between expression patterns of these two sample sets was more tenuous, with the MCPCA treatment generally having the lowest concordance. Breaking expression pattern comparisons down by time revealed indications that the length of storage and the timing of expression assessment may influence the predictive ability of the assessed biomarkers. Notably, fruit stored in air and assessed early in the storage period (Harvest to Early Postharvest) had higher agreement between sample sets than fruit stored in CA and treated with 1-MCP. When comparisons of longer time intervals were considered (Early to Late Postharvest or Harvest to Late Postharvest), the agreement between technologies improved, especially for fruit stored in CA and/or treated with 1-MCP.

Taken all together, observed discrepancies between the expected (model) and observed (qPCR) performance of GOIs assessed here indicate there are additional variables that need to be considered and incorporated into the models for enhanced performance. PBT efficacy may be influenced by a variety of factors such as maturity at harvest, annual weather patterns, other physiological indices, etc., that warrant further investigation. Furthermore, it is possible that certain GOIs identified in the models are better suited for certain conditions or applications than others, given the dynamic nature of gene expression in pome fruits during postharvest storage in modified storage conditions (Busatto et al., 2019; Gapper et al., 2013; Hargarten et al., 2018; L. Honaas et al., 2021).

3.7.2 Model evaluation of genes selected for qPCR

For qPCR, we selected 15 genes out of the top 30 genes from the RF-fm, about half of which were within the top 15 most important genes. In order to assess if model performance would drop further if 'less' important genes were included, we ran the Random Forest model using the 15 genes selected for qPCR (Random Forest qPCR Genes from Table 1). The performance for this gene set was 0.897 ± 0.011 SD for the training data and 0.748 ± 0.077 for the testing data, which was similar to the performance of the top 15 overall gene model (RF-rm, Table 1). The performance of a single run of the qPCR model is visualized in Figure 8. Gene expression patterns and gene names from the qPCR model can be seen in Supplemental Figure 6. The genes for this This provides another layer of evidence to suggest that the model is not overfitting and that despite variation in years, correlation holds. Second, because the

set of genes used for qPCR had a wider range of importance scores (selected from the top 30 genes from the RF-fm) and these genes still performed well in the RF model, it implies that there may not be a single set of genes that are predictive of the phenotype being investigated. This is important for future PTB development because it allows for flexibility in terms of gene selection; other pragmatic criteria could be considered, such as signal-to-noise ratios, variability, and ease of testing.

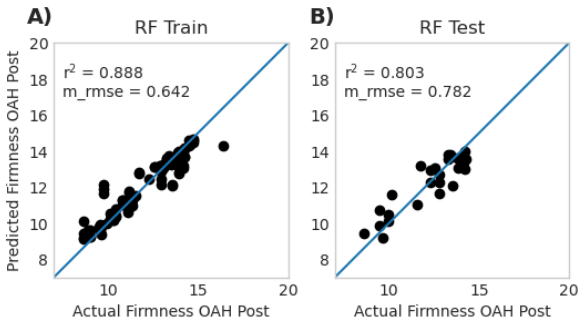


Figure 8: Single RF Model performance of genes selected for qPCR using the RNA-seq 2018 dataset. Data for replicated 100 runs of this model is presented in Table 1. Reported r^2 and m_rmse values in this figure represent a single run of a representative model, whereas data reported in Table 1 represents the average of 100 replicates.

4 Conclusions

Our results provide answers to several key questions about PTBs. First, these results provide more evidence supporting that gene expression profiles can be used in models that predict outcome. We identified a putative set of prognostic transcriptomic biomarkers (PTBs) capable of predicting postharvest fruit texture in ‘Gala’ apples within our experiment that included a range of commercially relevant postharvest treatments. Second, we explored two popular association methods and show that PTBs were

identified from a Random Forest regression-based feature selection model that outperformed Elastic Net Regression. Importantly, feature set stability varied across different train-test splits indicating a propensity for error in the Elastic Net models. This work shows that Random Forest regression in apples can be robust, especially as the gene set identified using Random Forest modeling outperformed putative firmness genes identified in the literature as associated with fruit texture. This illustrates the value of using an *ab initio* approach for PTB rather than relying solely on current knowledge. Some of the selected genes may underly or control expression of the trait as well as providing a signal for it. Third, we show that as few as 15 genes can be used to predict outcome and that qPCR has potential, although requires more exploration, for application of a PTB. While these results show that transcriptomics can be used for predictive biomarkers of traits that are highly impacted by the environment, it should be noted that this study is not exhaustive. Fruit texture in apples is well understood and highly controllable, which aided our work, but the models from this study may not ever be used in practice. However, it provides evidence to justify further studies and more data collection for traits that are more complex. More research is needed to understand other limitations (i.e., sample sizes), and how other factors (such as gene expression normalization) affect model robustness. Because the number of environmental variables is large, more data would be needed for more complex traits before a PTB can be used in an applied setting. However, this study provides evidence that such an undertaking is worthwhile. Further work in this area may yield information that can be used to expand the postharvest toolkit for managing apple fruit quality, leading to increased supply chain efficiency and less waste. Moreover, because gene expression is ubiquitous across all

living organisms, PTBs show promise as a tool for any species with traits that are highly impacted by the environment.

Acknowledgments

The authors would like to thank Bruno Torres and Sophia Reed for their excellent technical assistance.

REFERENCES

Acharjee, A., Larkman, J., Xu, Y., Cardoso, V. R., & Gkoutos, G. V. (2020). A random forest based biomarker discovery and power analysis framework for diagnostics research. *BMC Medical Genomics*, *13*(1), 178.

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data* (Version 0.11.9) [Computer software].

Bai, J., Baldwin, E. A., Goodner, K. L., Mattheis, J. P., & Brecht, J. K. (2005). Response of four apple cultivars to 1-methylcyclopropene treatment and controlled atmosphere storage. *HortScience: A Publication of the American Society for Horticultural Science*, *40*(5), 1534–1538.

Blanpied, G. D., & Silsby, K. J. (1992). Predicting Harvest Date Window for Apples. *Cornell Cooperative Extension*.

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

BorutaPy. (n.d.). https://github.com/scikit-learn-contrib/boruta_py

Bowen, J., Ireland, H. S., Crowhurst, R., Luo, Z., Watson, A. E., Foster, T., Gapper, N., Giovanonni, J. J., Mattheis, J. P., Watkins, C., Rudell, D., Johnston, J. W., & Schaffer, R. J. (2014). Selection of low-variance expressed Malus x domestica (apple) genes for use as quantitative PCR reference genes (housekeepers). *Tree Genetics & Genomes*. https://doi.org/10.1007/s11295-014-0720-6

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

Chang, H.-Y., & Tong, C. B. S. (2020). Identification of Candidate Genes Involved in Fruit Ripening and Crispness Retention Through Transcriptome Analyses of a “Honeycrisp” Population. *Plants*, *9*(10). https://doi.org/10.3390/plants9101335

Chen, X., Li, S., Zhang, D., Han, M., Jin, X., Zhao, C., Wang, S., Xing, L., Ma, J., Ji, J., & An, N. (2019). Sequencing of a Wild Apple (Malus baccata) Genome Unravels the Differences Between Cultivated and Wild Apple Species Regarding Disease Resistance and Cold Tolerance. *G3: Genes, Genomes, Genetics*, *9*(7), g3.400245.2019.

Conklin, C. C. (2019, March 26). *Genomics reveals secrets of optimal harvest, storage*. Fruit Growers News. https://fruitgrowersnews.com/article/genomics-reveals-secrets-of-optimal-harvest-storage/

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

DeEll, J. R., Murr, D. P., Porteous, M. D., & Vasantha Rupasinghe, H. P. (2002). Influence of temperature and duration of 1-methylcyclopropene (1-MCP) treatment on apple quality. *Postharvest Biology and Technology*, *24*(3), 349–353.

DeLong, J. M., Prange, R. K., Harrison, P. A., Andrew Schofield, R., & DeEll, J. R. (1999). Using the Streif Index as a Final Harvest Window for Controlled-atmosphere

Storage of Apples. In *HortScience* (Vol. 34, Issue 7, pp. 1251–1255). https://doi.org/10.21273/hortsci.34.7.1251

Deng, M. C. (2018). A peripheral blood transcriptome biomarker test to diagnose functional recovery potential in advanced heart failure. *Biomarkers in Medicine*, *12*(6), 619–635.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* , *32*(19), 3047–3048.

Favre, L., Hunter, D. A., O'Donoghue, E. M., Erridge, Z. A., Napier, N. J., Somerfield, S. D., Hunt, M., McGhie, T. K., Cooney, J. M., Saei, A., Chen, R. K. Y., McKenzie, M. J., Brewster, D., Martin, H., Punter, M., Carr, B., Tattersall, A., Johnston, J. W., Gibon, Y., … Brummell, D. A. (2022). Integrated multi-omic analysis of fruit maturity identifies biomarkers with drastic abundance shifts spanning the harvest period in "Royal Gala" apple. *Postharvest Biology and Technology*, *193*, 112059.

Feng, X., Zhang, R., Liu, M., Liu, Q., Li, F., Yan, Z., & Zhou, F. (2019). An accurate regression of developmental stages for breast cancer based on transcriptomic biomarkers. *Biomarkers in Medicine*, *13*(1), 5–15.

Ganai, S. A., Ahsan, H., Tak, A., Mir, M. A., Rather, A. H., & Wani, S. M. (2018). Effect of maturity stages and postharvest treatments on physical properties of apple during storage. *Journal of the Saudi Society of Agricultural Sciences*, *17*(3), 310–316.

Gapper, N. E., Bowen, J. K., & Brummell, D. A. (2022). Biotechnological approaches for

predicting and controlling apple storage disorders. *Current Opinion in Biotechnology*, *79*, 102851.

Gapper, N. E., McQuinn, R. P., & Giovannoni, J. J. (2013). Molecular and genetic regulation of fruit ripening. *Plant Molecular Biology*, *82*(6), 575–591.

Gerlach, C. (2022). USApple Industry Outlook 2022. *U.S. Apple Association*. https://usapple.org/wp-content/uploads/2022/08/USAPPLE-INDUSTRYOUTLOOK-2022.pdf

Goffings, M. H. (1993). Variability in Maturity, Quality and Storage Ability of Jonagold Apples on a Tree. *Acta Horticulturae*, *326*, 59–64.

Hadar, A., & Gurwitz, D. (2018). Peripheral transcriptomic biomarkers for early detection of sporadic Alzheimer disease? *Dialogues in Clinical Neuroscience*, *20*(4), 293–300.

Hadish, J. A., Biggs, T. D., Shealy, B. T., Bender, M. R., McKnight, C. B., Wytko, C., Smith, M. C., Feltus, F. A., Honaas, L., & Ficklin, S. P. (2022). GEMmaker: process massive RNA-seq datasets on heterogeneous computational infrastructure. *BMC Bioinformatics*, *23*(1), 1–11.

Hamza, R., & Chtourou, M. (2018). Apple Ripeness Estimation Using Artificial Neural Network. *2018 International Conference on High Performance Computing & Simulation (HPCS)*, 229–234.

Hargarten, H., Waliullah, S., Kalcsits, L., & Honaas, L. A. (2018). Leveraging Transcriptome Data for Enhanced Gene Expression Analysis in Apple. *Journal of the American Society for Horticultural Science. American Society for Horticultural Science*, *143*(5), 333–346.

Harker, F. R., Maindonald, J. H., & Jackson, P. J. (1996). Penetrometer measurement of

apple and kiwifruit firmness: Operator and instrument differences. *Journal of the American Society for Horticultural Science. American Society for Horticultural Science*, *121*(5), 927–936.

Harrell, F. (2022, October 6). *Statistical Thinking - How to Do Bad Biomarker Research*. https://www.fharrell.com/post/badb/

Honaas, L. A., Hargarten, H. L., Ficklin, S. P., Hadish, J. A., Wafula, E., dePamphilis, C. W., Mattheis, J. P., & Rudell, D. R. (2019). Co-expression networks provide insights into molecular mechanisms of postharvest temperature modulation of apple fruit to reduce superficial scald. *Postharvest Biology and Technology*, *149*, 27–41.

Honaas, L. A., & Kahn, E. (2017). A practical examination of RNA isolation methods for European pear (Pyrus communis). *BMC Research Notes*, *10*(1), 237.

Honaas, L., Hargarten, H., Hadish, J., Ficklin, S. P., Serra, S., Musacchi, S., Wafula, E., Mattheis, J., dePamphilis, C. W., & Rudell, D. (2021). Transcriptomics of Differential Ripening in “d”Anjou’ Pear (Pyrus communis L.). *Frontiers in Plant Science*, *12*, 609684.

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

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

Karst, T. (2019, October 7). *AgroFresh Solutions helps foresee bitter pit in Honeycrisp*. The Packer. https://www.thepacker.com/news/packer-tech/agrofresh-solutions-helps-foresee-bitter-pit-honeycrisp

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360.

Kolniak-Ostek, J., Wojdyło, A., Markowski, J., & Siucińska, K. (2014). 1-Methylcyclopropene postharvest treatment and their effect on apple quality during long-term storage time. *European Food Research and Technology = Zeitschrift Fur Lebensmittel-Untersuchung Und -Forschung. A*, *239*(4), 603–612.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*, 1–13.

Lexogen. (2020). *Quant Seq 3' mRNA-Seq Library Prep Kit User Guide* (Issue 015UG009V0223).

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , *25*(16), 2078–2079.

Lum, G. B., Brikis, C. J., Deyman, K. L., Subedi, S., DeEll, J. R., Shelp, B. J., & Bozzo, G. G. (2016). Pre-storage conditioning ameliorates the negative impact of 1-methylcyclopropene on physiological injury and modifies the response of antioxidants and γ-aminobutyrate in "Honeycrisp" apples exposed to controlled-atmosphere conditions. *Postharvest Biology and Technology*, *116*,

115–128.

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

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, *10*, 213.

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.

Nicolaï, B. M., Lötze, E., Peirs, A., Scheerlinck, N., & Theron, K. I. (2006). Non-destructive measurement of bitter pit in apple fruit using NIR hyperspectral imaging. *Postharvest Biology and Technology*, *40*(1), 1–6.

Pedregosa, F., Varoquaux, G., & Gramfort, A. (2011). Scikit-learn: Machine learning in Python. *Of Machine Learning …*. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com

Pedrotty, D. M., Morley, M. P., & Cappola, T. P. (2012). Transcriptomic biomarkers of cardiovascular disease. *Progress in Cardiovascular Diseases*, *55*(1), 64–69.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S.

L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295.

Prengaman, K. (2019). *Revealing risks with RNA*. Good Fruit Grower. https://www.goodfruit.com/revealing-risks-with-rna/

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

Ritz, C., & Spiess, A.-N. (2008). qpcR: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. *Bioinformatics* , *24*(13), 1549–1551.

Shewa, A. G., Gobena, D. A., & Ali, M. K. (2022). Review on postharvest quality and handling of apple. *International Journal of Agricultural Science and Food Technology*, *8*(1), 028–032.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. *Expert Systems with Applications*, *134*, 93–101.

Storch, T. T., Pegoraro, C., Finatto, T., Quecini, V., Rombaldi, C. V., & Girardi, C. L. (2015). Identification of a Novel Reference Gene for Apple Transcriptional Profiling under Postharvest Conditions. *Plos One*. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120599

Supplitt, S., Karpinski, P., Sasiadek, M., & Laczmanska, I. (2021). Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. *International Journal of Molecular Sciences*, *22*(3). https://doi.org/10.3390/ijms22031422

USDA, National Agricultural Statistics Service. (2022). *Noncitrus Fruits and Nuts 2021 Summary*. https://downloads.usda.library.cornell.edu/usda-esmis/files/zs25x846c/4q77gv96p/t722jd76c/ncit0522.pdf

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.

Volz, R. K., Harker, F. R., & Lang, S. (2003). Firmness Decline in `Gala' Apple during Fruit Development. *Journal of the American Society for Horticultural Science. American Society for Horticultural Science*, *128*(6), 797–802.

Wafula, E. K., Zhang, H., Von Kuster, G., Leebens-Mack, J. H., Honaas, L. A., & dePamphilis, C. W. (2022). PlantTribes2: tools for comparative gene family analysis in plant genomics. In *bioRxiv* (p. 2022.11.17.516924). https://doi.org/10.1101/2022.11.17.516924

Wu, B., Shen, F., Wang, X., Zheng, W. Y., Xiao, C., Deng, Y., Wang, T., Yu Huang, Z., Zhou, Q., Wang, Y., Wu, T., Feng Xu, X., Hai Han, Z., & Zhong Zhang, X. (2021). Role of MdERF3 and MdERF118 natural variations in apple flesh firmness/crispness retainability and development of QTL-based genomics-assisted prediction. *Plant Biotechnology Journal*, *19*(5), 1022–1037.

Zhang, H., Wafula, E. K., Eilers, J., Harkess, A. E., Ralph, P. E., Timilsena, P. R., dePamphilis, C. W., Waite, J. M., & Honaas, L. A. (2022). Building a foundation for

gene family analysis in Rosaceae genomes with a novel workflow: A case study in Pyrus architecture genes. *Frontiers in Plant Science*, *13*, 975942.

Busatto, N., Farneti, B., Tadiello, A., Oberkofler, V., Cellini, A., Biasioli, F., Delledonne, M., Cestaro, A., Noutsos, C., & Costa, F. (2019). Wide transcriptional investigation unravel novel insights of the on-tree maturation and postharvest ripening of “Abate Fetel” pear fruit. *Horticulture Research*, *6*, 32.

Supplemental Materials

Supplemental Tables

Supplemental Table 1: Apple tissue collection schedule.
Included as a separate file.

Supplemental Table 2: MultiQC report for RNA-seq alignment.
Included as a separate file.

Supplemental Table 3: PlantTribes2 orthogroup classification top 15 genes.
Included as a separate file.

Supplemental Table 4: qPCR primer design parameters.
Included as a separate file.

Supplemental Table 5: qPCR primer sequences.
Included as a separate file.

Supplemental Table 6: Comparison of year one and year two sample design.
Included as a separate file.

Supplemental Table 7: Firmness genes identified from the literature.
Included as a separate file.

Supplemental Table 8: Genes identified by Boruta Random Forest.
Included as a separate file.

Supplemental Figures

Supplemental Figure 1: Comparison of physiological measurements across years. **A** is for fruit diameter and **B** is for at harvest creep.
Included as a separate file.

Supplemental Figure 2: Top 15 genes of random forest model expression. Counts are in transcripts per million.
Included as a separate file.

Supplemental Figure 3: Top 15 genes of elastic net model expression. Counts are in transcripts per million.
Included as a separate file.

Supplemental Figure 4: Top 15 genes of random forest literature genes expression. Counts are in transcripts per million.
Included as a separate file.

Supplemental Figure 5: Comparison of qPCR and RNA-seq data.
Included as a separate file.

Supplemental Figure 6: The random forest qPCR genes expression. Counts are in transcripts per million.
Included as a separate file.

CHAPTER FOUR:

INVESTIGATING REQUIREMENTS OF TRANSCRIPTOMIC DATASETS FOR PREDICTIVE MODELING USING LARGE *ARABIDOPSIS THALIANA* RNA-SEQ DATASET

Abstract

Transcriptomic data can be combined with phenotypic data to create predictive models which identify transcriptomic biomarkers. These biomarkers are useful for monitoring and prediction of difficult-to-measure phenotypic traits and are becoming increasingly used in high-value agricultural crops. Despite this, little research has been done on how many samples are required for these models to be accurate, and which normalization should be used. Here we create a massive RNA-seq dataset from publicly available *Arabidopsis thaliana* data with corresponding measurements for age and tissue type. We use this dataset to create random forest regression and classification models to determine how many samples are needed for accurate prediction and which normalization method is required. We find that Median Ratios Normalization significantly increases performance when predicting age. We also find that in the case of our dataset, only a few hundred samples are required to predict tissue types, whereas a few thousand samples are necessary to accurately predict age. These are important findings to consider when designing experiments to identify transcriptomic biomarkers.

Introduction

Predictive modeling of phenotypic traits using transcriptomic data is a method that is increasing in popularity as larger datasets become available. This type of modeling uses count data derived from RNA-seq experiments to identify biomarkers that are predictive of a phenotypic trait. Whereas genetic markers remain constant for the life of an organism, transcriptomic biomarkers change readily in response to their environment and to the internal status of an organism. A transcriptomic biomarker is one or more genes whose transcriptomic level is indicative of a current or future phenotypic outcome. This means that they can be used for monitoring and prediction of difficult-to-measure phenotypic traits. Transcriptomic biomarkers are used in medical research with an emphasis on predicting cancer type and stage (Bostanci et al., 2023; Feng et al., 2019; Smith et al., 2020; Supplitt et al., 2021). However, recent research has branched into high-value agricultural crops, where researchers are interested in predicting traits such as flowering time (Azodi et al., 2020), flesh quality traits in apples, pears, and potatoes (Acharjee et al., 2016; Gapper et al., 2013; Hatoum et al., 2016; Leisso et al., 2016), and apple maturity (Favre et al., 2022).

Despite the increasing interest in transcriptomic modeling of traits, there has not been an investigation of how many samples are required to perform these models. The majority of datasets used for this type of modeling often incorporate only a few dozen to a few hundred samples. This contrasts with predictive models in non-biological research areas which can sometimes have thousands to millions of samples (Herman & Schumacher, 2018; Rokach, 2016). Additionally, RNA-seq datasets have a much higher dimensionality in terms of the number of features (genes) they have when compared to

other datasets. An RNA-seq dataset can have measurements for thousands of genes, whereas datasets in other domains typically only have a few hundred (Li & Li, 2018). These RNA-seq datasets are referred to as “wide”, containing many features (genes) and relatively few samples. This can present issues for model methods that were created with the intent of only a few features (Li & Li, 2018; Van Der Maaten et al., 2009).

Despite increased interest in biomarker discovery from gene expression, more information is needed to address the question of the number of samples that might be needed to accurately find biomarkers from an underdetermined system using gene expression data. Here we report a study to, first, explore how many samples may be required for transcriptomic modeling of both categorical and continuous phenotypic traits, and second, to identify the effect that RNA-seq count normalization methods have on large disparate RNA-seq datasets. Normalization is a potential acute problem for both large data sets collected over several years or by multiple collaborating groups as well as for large conglomerate datasets with potentially hundreds of different experiments.

In this paper, we create a large RNA-seq dataset from *Arabidopsis thaliana* (Arabidopsis) retrieved from the National Center for Biotechnology Information’s (NCBI) Sequence Read Archive (SRA) database (NCBI Resource Coordinators, 2016). We also create a large annotation dataset from companion data available on NCBI BioProjects database (Barrett et al., 2012; Federhen et al., 2014) which we manually curate. We chose to use Arabidopsis, as it is a model organism in plant science (Somssich, 2019) with well-defined physiological stages (Boyes et al., 2001) and has a

large amount of RNA-seq data available for it. This makes Arabidopsis an ideal candidate for investigating the size of the dataset required for creating accurate models. The diverse annotations available for this Arabidopsis data allow us to investigate models for both classification-based variables (tissue type) and continuous variables (age).

We use random forest models (Breiman, 2001) for investigating these datasets, as it is robust to outliers and is capable of dealing with a large number of features (in our case genes) (Couronné et al., 2018), and has been shown to be superior to deep learning methods for count data (Smith et al., 2020). We also make use of Boruta (Kursa et al., 2010) for feature reduction, and Synthetic Minority Over-sampling TEchnique SMOTE (Chawla et al., 2002) for over-sampling (supplementing) sparse data. Additionally, we investigate several normalization methods (Trimmed Mean of M values (TMM) (Robinson et al., 2010), Median Ratios Normalization (MRN) (Anders & Huber, 2010; Love et al., 2014), Transcripts Per kilobase Million (TPM), and No Normalization (NoNo) to determine which is best for dealing with large conglomerate datasets.

The following research seeks to address how many samples are required for transcriptomic modeling of both categorical and continuous phenotypic traits, identify the best normalization method for dealing with large conglomerate RNA-seq datasets, and demonstrate how large conglomerate datasets can be mined for additional information beyond their initial intent. This has relevance for experimental design interested in identifying transcriptomic biomarkers for both categorical and continuous phenotypic traits. We found that MRN normalization performed better than other forms

of normalization when predicting age, whereas normalization had no impact when predicting tissue type. We also found that a few hundred samples are sufficient for predicting our categorical variable of tissue type, whereas a few thousand samples are required for modeling the continuous variable of age.

Method

RNA-seq Data Pre-Processing

Arabidopsis RNA-seq data was retrieved from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (NCBI Resource Coordinators, 2016) using the following search parameters:

txid3702[Organism:noexp] AND ("biomol rna"[Properties] AND "platform illumina"[Properties] NOT "strategy wxs"[Properties] NOT "strategy targeted capture"[Properties] NOT ("strategy other"[Properties] NOT ("library selection pcr"[Properties] NOT "library selection padlock probes capture method"[Properties] NOT "library selection hybrid selection"[Properties] NOT "library selection other"[Properties]))) NOT ("strategy chip"[Properties] NOT "strategy mre seq"[Properties] NOT "strategy atac seq"[Properties] NOT "strategy faire seq"[Properties] NOT "strategy mnase seq"[Properties] NOT "strategy dnase hypersensitivity"[Properties] NOT "strategy medip seq"[Properties] NOT "strategy mbd seq"[Properties] NOT "strategy bisulfite seq"[Properties]) NOT "filetype bam"[Properties] AND ("biomol rna"[Properties] AND "platform illumina"[Properties] AND "filetype fastq"[Properties])

This search string retrieves all *Arabidopsis thaliana* (Arabidopsis)(NCBI txid3702) RNA-seq data in the fastq format created using an Illumina sequencing machine (Illumina, San Diego, CA, US). A total of 74833 SRR (RNA-seq runs) were identified with these parameters corresponding to 59428 SRX (RNA-seq experiments) (**Supplemental Table 1**). An SRX experiment can consist of multiple RNA-seq run files.

The list of 74833 SRR accessions was used with the GEMmaker workflow (Hadish et al., 2021)automatically retrieves from NCBI and processes the SRR files. For these data, GEMmaker was run using Kallisto (Bray et al., 2016) for expression quantification. For genome alignment, GEMmaker was given the Arabidopsis genome (TAIR 10 assembly, Araport 11 annotations) retrieved from The Arabidopsis Information Resource (TAIR) (Berardini et al., 2015). Data was split into batches of ~5000 SRR numbers for processing. While splitting the data is not necessary for GEMmaker, doing so allowed for execution on multipole queus on Washington State University's high-performance computing cluster "Kamiak" and took approximately 3 months to complete using available nodes at the time. During execution, 6168 SRX experiments were removed due to improper SRA file formatting, SRA file corruption, or empty SRA files (10.38% removed). The resulting Gene Expression Matrix (GEM) consisted of 53260 samples and 48359 genes representing 2605 NCBI BioProjects. After the creation of the GEM, custom Python code was used to remove samples that did not have at least 1 million reads and did not have at least 70% of reads aligning with the Arabidopsis genome. A total of 288 sample pairs were noticed to be identical, and on closer inspection, it was determined that the same RNA-seq sample had been uploaded to NCBI multiple times (at least twice, as many as five times) under different names

(and often different annotations). These were removed from consideration. Within the remaining samples, genes with expression in less than 1000 samples over a count of 10 (in the NoNo dataset) were removed. This was an ad hoc filter based on “filterByExpr” function of the edgeR package (Robinson et al., 2010), with a number of different posible combinations of “samples” and “count” variables assessed to see how this would influence gene count (**Supplemental Figure 1**). This filtering resulted in a final GEM consisting of 32044 samples and 43224 genes.

Four separate GEMs were created to test model performance for different normalization methods: Trimmed Mean of M values (TMM), Median Ratios Normalization (MRN), Transcripts Per kilobase Million (TPM), and No Normalization (NoNo). TMM normalization (Robinson et al., 2010; Robinson & Oshlack, 2010) and MRN normalization (Anders & Huber, 2010; Love et al., 2014) were performed using the Python “conorm” package 1.2.0 (Meshcheryakov, 2021). TPM and NoNo normalization values were an output of Kallisto (Bray et al., 2016). How these normalizations impacted sample count is visualized as **Supplemental Figure 2**.

Sample Annotations Pre-Processing

Sample annotations were retrieved from the NCBI BioProject database (Barrett et al., 2012; Federhen et al., 2014) using BioSampleParser which was slightly modified to check for successful data retrieval (Limeta, 2020). Annotations were retrieved for 48696 NCBI BioSamples, representing data from 2643 BioProjects. A total of 668 different annotation classes (e.g. “tissue”, “age”, “organism”, “title”, etc.) were retrieved (**Supplemental Table 2** all BioSample Info). A majority of these annotation classes

were present for one or a few BioProjects, and therefore were sparse, only containing a few samples (**Supplemental Figure 3**). These sparse classes were ignored. Annotation classes assigned to over 5000 RNA-seq samples are visualized in **Figure 1 A**. For this experiment, we used annotation classes “tissue” and “days” because of their large number of associated RNA-seq samples and biological relevance. Additionally, “Tissue” is categorical and can therefore be modeled as a classification problem (i.e., which tissue did a sample come from), whereas “days” is continuous and can be modeled as a regression problem (i.e., how old is this sample).

The “tissue” annotations from NCBI included a total of 1186 unique annotation terms (**Figure 1 B**) for the RNA-seq samples. Due to inconsistent use of terms such as misspellings and ambiguity, these terms required manual curation. In summary, we formed six tissue categories with the following terms: ”leaf”, “seedling”, “shoot”, “seed”, “root”, and “flower” (referred to as the “tissue-6” dataset). In addition, a second dataset consisting only of ”leaf”, “seed”, “root”, and “flower” was created to test on precise labels, and is referred to as “tissue-4”. As part of our curation process, we made several changes. First, misspellings were corrected. Generic terms (e.g. “the plant”, “whole” “col-0”), tissue types created for specific laboratory applications which are unlike their donor tissue (e.g., “protoplasts”, “in vitro cotyledon”, “tissue culture callus”), and unknown or inappropriate values (e.g., “usa”, “p100”, “liver”) were excluded. A notable conglomeration of samples combined “inflorescence” (e.g. “immature inflorescence”, “plant inflorescence”, “inflorescence containing stage 8 and younger flowe”) terms with “flower” related terms (e.g. “mature flower”, “immature flower bud cluster”, “young_flower_control”). Our “seedling” term was defined as plants younger than 6 days

post germination (stage 1, before first primary leaves, plate-based) (Boyes et al., 2001), but it should be noted that many samples labeled with the term “seedling” had no information about the day they were collected and were still included. Terms related to “rosette” (e.g. “complete rosette”, “aerial rosette tissue”, “entire vegetative rosette”) were manually changed to “leaf” unless other information was included (e.g., “rosette and inflorescence”). The “shoot” term was assigned to a large group of samples (1142 samples) but is ambiguous in its meaning, as the flat, rosette nature of adult Arabidopsis plants means that this tissue type is difficult to define in plants over a few days old. After manual correction, the tissue labels were combined with the filtered GEMs, resulting in a data frame of 16271 RNA-Seq samples from 1128 BioProjects across these 6 tissue types (**Table 1**).

The age annotation consisted of 857 unique values across all samples. Like tissue, age was reported by researchers in multiple ways. Age was reported from the time of “seeding”/”germination”, after an event such as “flowering” or “inoculation”, or as a raw number without any additional information. Age was also reported with different terms such as “day”, “week” and “month”. Some age values were improper (“Austria: Innsbruck”, “Nitrogen, plus Cycloheximide, Dexamethasone”, “environmental-water”, etc.), were not day specific (“just prior to or at bolting”, “Adult”, etc.) or implausible (“6month”, “67 years”, etc.). Such values were excluded. A summary of the valid annotation values is available in **Table 2** and visualized in **Supplemental Figure 4.** For our experiment, we used age values reported with the term “Days” in the training and testing of models and used other age values to verify the models. Those RNA-seq samples with age in "Days" values combined with the cleaned annotations are hereafter

referred to as “Days” Age annotation Labeled (DAL) dataset. DAL is a subset of the age dataset. Additionally, we used the time period of 0 to 30 days and excluded later days because of sparsity after 30 days **Figure 1 C**– preliminary models tended to perform poorly when they were included. RNA-seq samples with ages between 0 and 30 were combined with the filtered GEMs, which resulted in a dataset with 6136 samples from 485 BioProjects.

After combining the GEMs with the six tissue categories or age in days, respectively, the resulting data frames were split into training and testing datasets. These splits considered BioProject as a batch effect and required that samples from a single BioProject are all either in the training dataset or the testing dataset. This requirement was meant to prevent the overfitting of models due to latent covariates resulting from sample preparation and sequencing (sequencing depth and sample preparation type) within BioProjects. For the tissue dataset, a total of 12749 samples from 902 BioProjects were in the training dataset and 3522 samples from 226 BioProjects were in the testing dataset (the testing set has 28% of the samples and 25% of the BioProject). **Supplemental Table 3** shows a breakdown of these splits based on the 6 conditions. For the DAL dataset, a total of 5062 samples from 388 BioProjects were in the training dataset and 1074 samples from 97 BioProjects in the testing dataset (the testing set has 21% of the samples and 25% of the BioProjects). These splits were used for all models except if otherwise stated.

Table 1: Distribution of the different tissue categories. BioProjects with multiple tissue categories are included in all counts.

Tissue Category	Number of Samples	Number of BioProjects
flower	723	79
leaf	5321	364
root	2436	202
seed	689	59
seedling	5960	440
shoot	1142	79

Table 2: Distribution of the different age categories.

Age Category	Number of Samples	Number of BioProjects
Reported as Time		
Days (DAL)	6457	508
Weeks	2263	195
Months	38	5
Number Only	1177	67
After Event		
Days After Sowing (DAS)	514	19
Days After Germination (DAG)	445	45
Days After Pollination	112	7
Days After Flowering	68	3

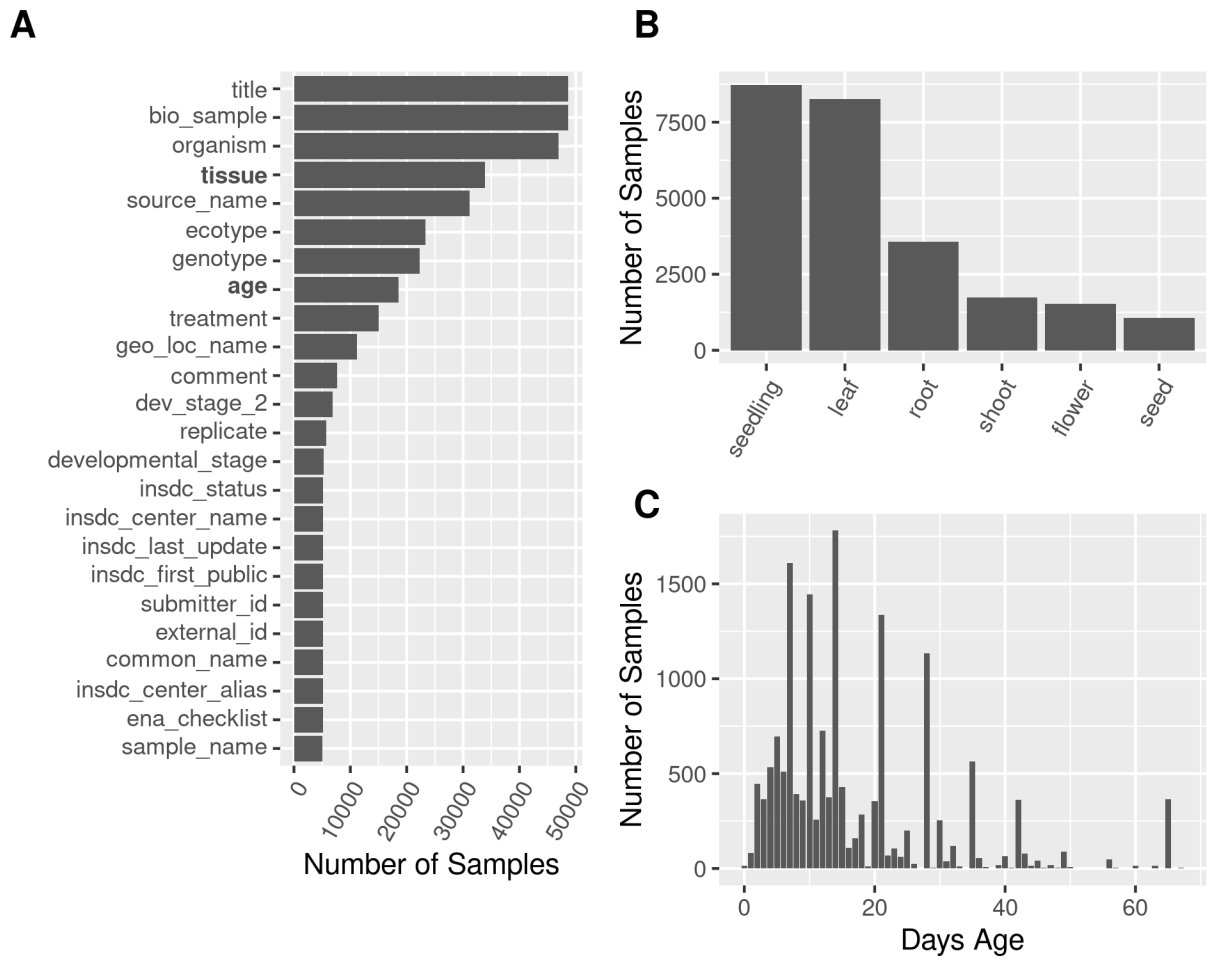


Figure 1: Annotations retrieved from NCBI for the Arabidopsis dataset. **A)** Available annotation columns. The y-axis represents the name of the annotation column, and the x-axis represents the number of samples that have that annotation column. Annotation columns can be both biological related and technical related. Annotation columns with at least 5000 samples are shown. For this experiment, we decided to concentrate on tissue and age. **B)** Tissue annotations. The x-axis represents the tissue annotation type, and the y-axis represents how many samples have that annotation. **C)** Age annotations. The x-axis represents how old the sample annotation is (rounded to the nearest day), and the y-axis represents how many samples are that age. Age annotation was reported on NCBI in many ways, with a breakdown figure reported as (**Supplemental Figure 4**)

Model Parameter Optimization

Random forest model parameter optimization was carried out using the

RandomizedSearchCV class from the Python package Scikit-learn (sklearn) (Pedregosa

et al., 2011). Random Forest models were created using the RandomForestClassifier class for classification of tissues and the RandomForestRegression class was used for age prediction. In both cases, the parameter search space iterated over the following grid: 'bootstrap': [True, False], 'n_estimators': [100, 300, 500, 1000, 1500, 2000], 'n_estimators': [3, 5, 10, 20], 'max_depth': 3 to 100 at an interval of 3), 'min_samples_split': [1,2,4], 'min_samples_leaf': [3, 5, 10, 20, 30], 'max_features': ['sqrt', 'log2']. These parameters were sampled 100 times for each of the four filtered GEMs (NoNo, and three GEMs normalized by TMM, MRN, and TPM respectively). Five-fold cross-validation was performed in each model instance, with folds accounting for BioProject batch effect in a manner similar to the input dataset. Evaluation of each instance was performed using F1 and Accuracy for the tissue classification problem and r^2 for the age regression problem to determine the optimal model parameters and compare the effect of normalization methods of the GEMs.

Assessing Annotation Accuracy

The datasets used in this project are conglomerates of many BioProjects, with samples and labels being collected and classified in many different ways by different researchers. This creates the possibility that our models are fitting on the variation between projects, and not on the actual biological traits of interest (i.e. overfitting). To test that our RNA-seq samples contained information that was reflected in their assigned labels, we conducted a randomization strategy, where a percentage of the dependent variable (tissue or age) from 0 - 100% was randomized in the training dataset. If the model is fitting on true biological information, then model accuracy should

decrease over increased randomization. Model creation was done using the optimized parameters found for each respective dataset, with respective accuracy assessments (F1 and Accuracy for tissue-6 and tissue-4, r^2 for DAL).

Model Performance Metrics

Models using all available data for the tissue dataset (six categories) and DAL dataset (from 0-30 days) were assessed and visualized using respective optimized parameters. Visualization for the categorical tissue dataset uses a confusion matrix and visualization of the quantitative DAL dataset compared to predicted and actual data as a scatterplot.

To determine how many samples are required to accurately predict “tissue” and “age” in our datasets, modeling was performed using different size sample sets. This was performed in an iterative manner, starting with a small training dataset and gradually adding samples. We used two approaches. The first approach added samples from the training dataset by BioProject: each iteration added all the samples from 10 random BioProjects. The average number of samples added at each iteration was 92.9 (std 87.3) for the tissue dataset and 128.2 (std 42.3) for the DAL dataset. For the second approach, the entire training datasets were randomized (irrespective of BioProject) and samples were randomly added in batches (for age: batch size of 20 for the first 500 samples added, batch size of 40 for the next 500 samples added, batch size of 100 for the next 1000 samples, and batch size of 200 for the remaining samples (up to 5062). For tissue-4: batch size of 10 for the first 500, batch size of 30 for the next 500, batch size of 60 for the next 1000, and batch size of 200 for the remaining. For tissue-6: batch size of 20 for the first 1000, batch size of 100 for the remainder, see

Supplemental Figure 5 x-axis) In both cases, the same testing dataset was used every time (independent BioProjects). The first approach is intended to evaluate the required number of independent BioProjects for good model performance. The second approach is intended to simulate a single homogenous--albeit high variance--dataset. This is to simulate how many samples a researcher would need to gather if they were interested in replicating this independently. Additionally, both of these methods were performed on a reduced tissue dataset which excluded the categories “shoots” and “seedlings”. This reduced tissue dataset with only 4 categories (”leaf”, “seed”, “root”, and “flower”) is referred to as the tissue 4 dataset.

Models Using Germination and Sowing Dates

Within the annotations for age were samples labeled with terms related to “Days After Germination” (DAG) and “Days After Sowing” (DAS). These two terms are more specific than data in the DAL dataset, but there were fewer BioProjects using these labels. The DAG dataset had 530 samples from 52 BioProjects and the DAS dataset had 873 samples from 19 BioProjects. Two new random forest regression models were created using just the data from these respective labels. Testing was done in the same manner as described above with the DAL dataset. Additionally, both the DAG and DAS models were used to predict day within the DAL dataset to determine if they were more accurate at predicting DAL due to their more specific nature.

Synthetic Data and Balancing

Data for the DAL dataset were supplemented using Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla et al., 2002). This was done to increase the amount of data available for days with few samples to see if this increased model performance. Prior to SMOTE, dates with half days (e.g. 5.5 days) were rounded up to the nearest day. Days with less than 10 samples were removed, as SMOTE is inaccurate with small sample sizes (days 0,1,19,27, 29 were removed). SMOTE oversampling was performed using the imbalanced-learn package (version 0.11.0) with default parameters (Lemaitre et al., 2017). This resulted in a new data frame that contained 16112 samples, over three times the amount of the original training dataset size of 5062. The distribution of samples before and after SMOTE is visualized in **Supplemental Figure 6**. Synthetic Data was not created for the tissue dataset due to already high model performance.

Feature Selection and Evaluation

Feature selection was performed on the DAL, tissue-6, and tissue-4 datasets to determine which genes were most important for predicting tissue and days of age. Boruta feature selection was performed on both the DAL and the tissue dataset using their respective optimized parameters. Borutapy v 0.3 (Homola & Beanico, 2019; Kursa et al., 2010) parameters were set to *n_estimators*='auto' (defaults to 1000), *max_iter*=200, *perc* = 100 for both the tissue and DAL datasets. The results of these Boruta runs are sets of genes that are capable of predicting the dependent variable (i.e., tissue or days of age) better than a randomized version of themselves. Genes selected by Boruta were used to create new datasets: a new DAL Boruta Dataset which had

15024 genes remaining, a new tissue-6 Boruta Dataset which had 20017 genes remaining, and a new tissue-4 Boruta dataset which had 7837 genes remaining. These reduced Boruta datasets were then used to create new random forest models. While boruta is able to determine if a gene is better than a random version of itself, it does not rank the genes on their importance to the model. To generate feature importance scores, new random forest models were created.

Results

Optimizing Input Parameters and Assessing Normalization Method

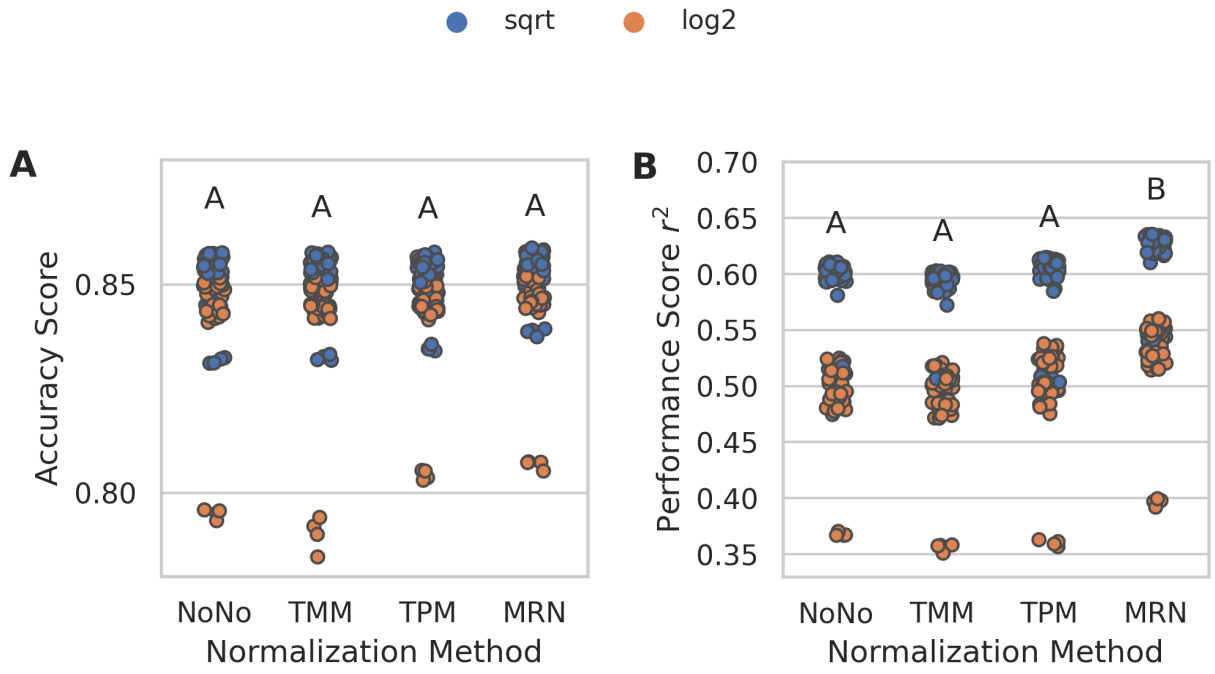


Figure 2: Assessing parameter optimization and normalization methods **A)** Tissue model. There were no significant differences between normalization methods (one-way ANOVA, p = 0.475) **B)** Age model results. A and B represent significant differences based on Tukey's honestly significant difference (HSD) test **Supplemental Table 4**. Of the 4 normalization methods, MRN had the best performance on average. We also saw separation in the chart, which was most impacted by the parameter “max_features” (colored points).Additional “max_features” values were evaluated and can be seen in **Supplemental Figure 7**. Points that are separated from their respective colored

clusters (log2 around 0.36 and sqrt around 0.53) were found to be models where “max_depth” was too low and set to only 3 **Supplemental Figure 8**.

Normalization methods dramatically impacted GEM values and therefore total reads per sample. A distribution of the number of reads per sample for all 4 normalization methods is visualized as **Supplemental Figure 2**. These normalized GEMs were used to create models for both tissue classification and age prediction. For the tissue classification, no significant difference in model performance was seen between the 4 normalization methods (one-way ANOVA, p = 0.485) (**Figure 2 A**). In contrast, a significant difference in performance was observed between different normalization methods for the DAL dataset (one-way ANOVA, p = 0.00012), with a posthoc Tukey's Honest Significant Difference (HSD) test revealing that MRN performed significantly better than either NoNo, TMM, or TPM **(Figure 2 B) (Supplemental Table 4).** While no difference was seen in normalization methods for the tissue classification, it was decided to use MRN for the remaining analyses due to its increased performance with the age model.

In addition to evaluating different normalization methods, random forest model parameters were tested with different values to identify optimal model performance. Parameter sets for both the DAL and tissue datasets were tested in a similar manner. For tissue classification the optimal parameters were 'n_estimators': 1500, 'min_samples_leaf': 3, 'max_features': 'sqrt', 'max_depth': 48, 'bootstrap': False (**Supplemental Table 5**). For the age model, the optimal parameters were *'n_estimators'*: 1500*, 'min_samples_leaf'*: 3, *'max_features'*: 'sqrt', *'max_depth'*: 48, *'bootstrap'*: False (**Supplemental Table 6**). However, it was found that reducing

‘*n_estimators’* to 300 had negligible impact on the r^2 score while dramatically reducing runtime (**300:** mean r^2 0.555, std 0.051, **1500:** mean r^2 0.563 std 0.049, ANOVA: F-Statistic: 1.0875, P-value 0.2987). Therefore, the remaining analyses used parameters *'n_estimators'*: 300*, 'min_samples_leaf'*: 3, *'max_depth'*: 7, *'bootstrap'* False. Furthermore, it was decided to set *max_depth* at 7 for both models, as this lower value performed nearly as well as higher parameters and reduces the chance of overfitting the model. The dramatic difference in performance for *max_features* for the age dataset (**Figure 2 B**)--and to a lesser degree in the tissue dataset (**Figure 2 A**)--warranted additional investigation. A GridSearch optimizing for *'max_features'* revealed optimal *'max_features'* = 700 for tissue (**Supplemental Figure 7 A**) and *'max_features'* = 1000 for DAL (**Supplemental Figure 7 B**). These parameters take into account the tradeoffs between model performance and model speed, which has an impact when training Boruta models in later steps.

Optimal Parameter Performance

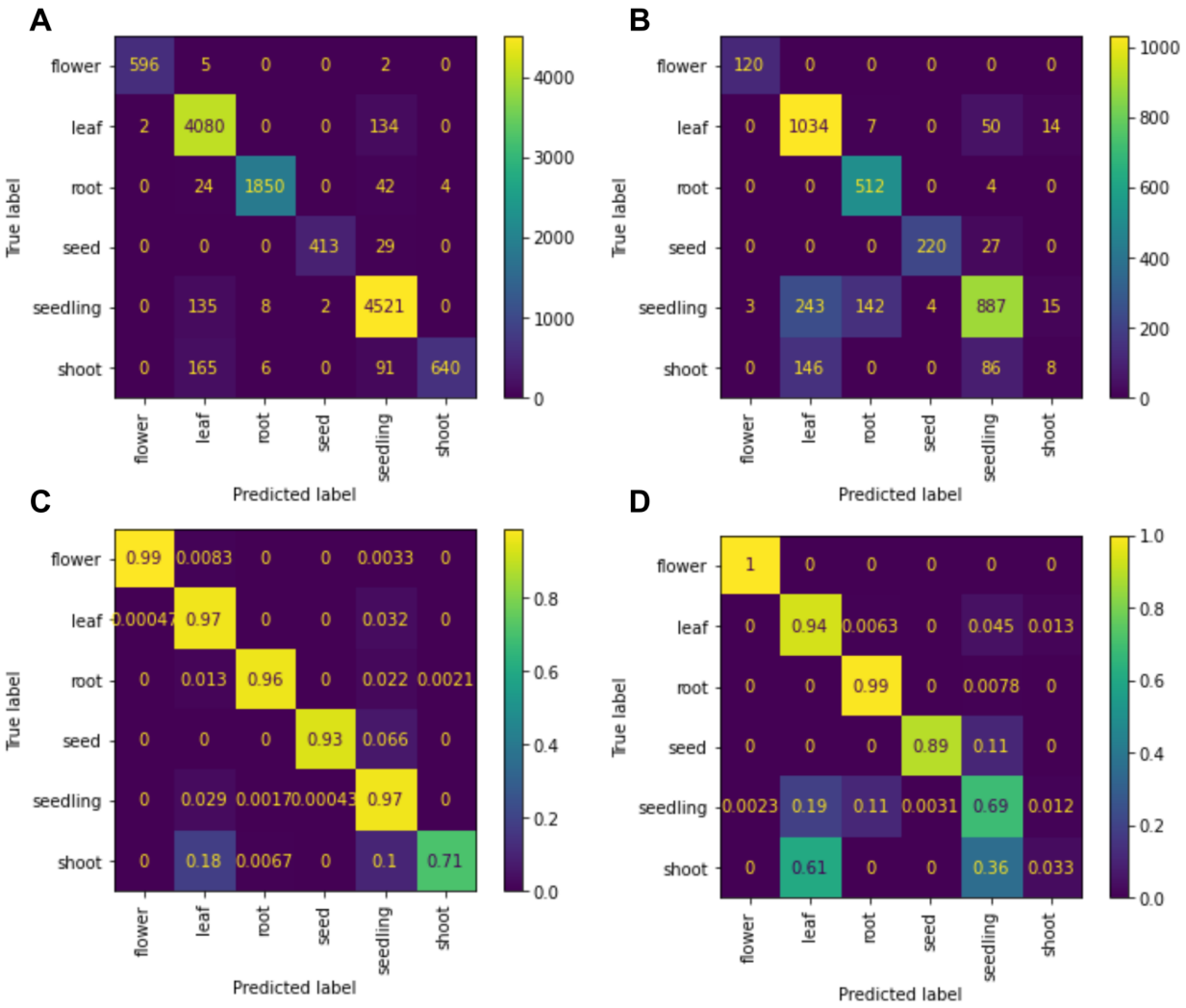


Figure 3: Confusion Matrices of Tissue Data. **A)** Unnormalized (values represent actual count) train **B)** Unnormalized test **C)** Normalized train (values in each row and column sum to 1) **D)** Normalized test.

Tissue classification using optimal parameters showed an F1 score of 0.942 and a model accuracy of 0.948 for the training dataset and an F1 score of 0.739 and an accuracy of 0.792 for the testing dataset. Confusion plots demonstrating the results of each classification are visualized in **Figure 3**, with subplots **A** and **B** showing the training and testing data performance, respectively. Values along the diagonal are the

number of correctly predicted labels and those non-diagonal values are incorrect predictions. Subplots **C** and **D** show the same results but with counts normalized such that the sums of columns and rows are 1. These confusion matrices show that labels for “flower”, “leaf”, “root” and “seed” performed the best, whereas labels for “seedling” and “shoot” did not perform as well, likely due to their ambiguity, as we saw some annotations labeled as “seedling” which were plants that should have one or more pairs of true leaves and “shoot” is difficult to define except in plants before they develop primary leaves. Therefore, an additional model was created excluding “seedling” and “shoot” (referred to as tissue-4). The tissue-4 model had an F1 score of 0.996 and a model accuracy of 0.995 for the training dataset and an F1 score of 0.995 accuracy 0.994 for testing. A confusion matrix showing the breakdown of categories is shown in **Figure 4**.

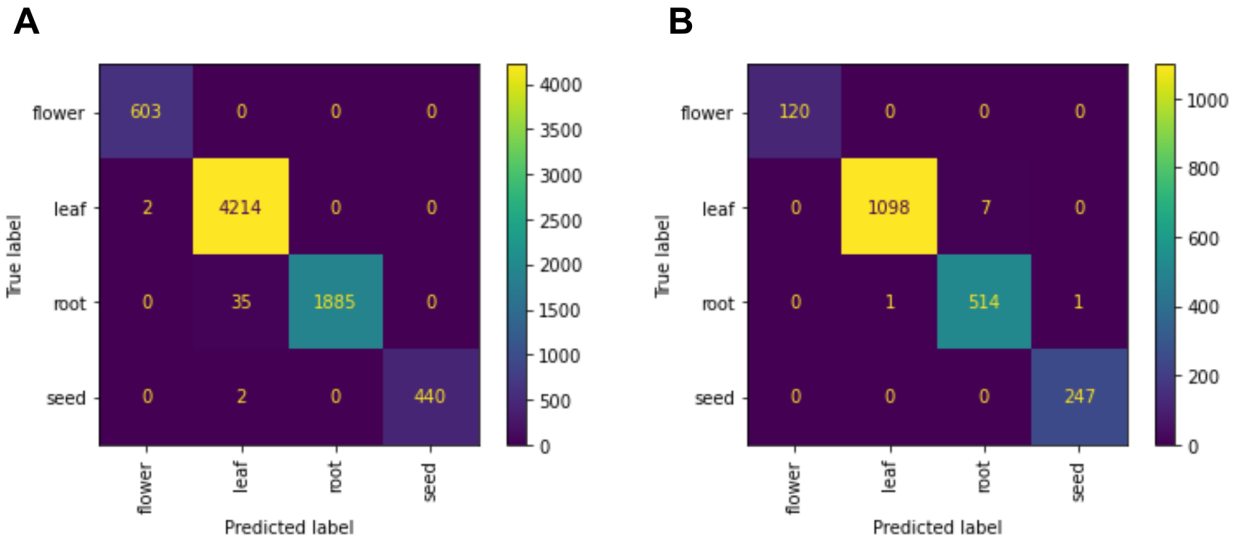


Figure 4: Confusion matrices of tissue-4 models. **A)** Training dataset accuracy. **B)** Testing dataset accuracy.

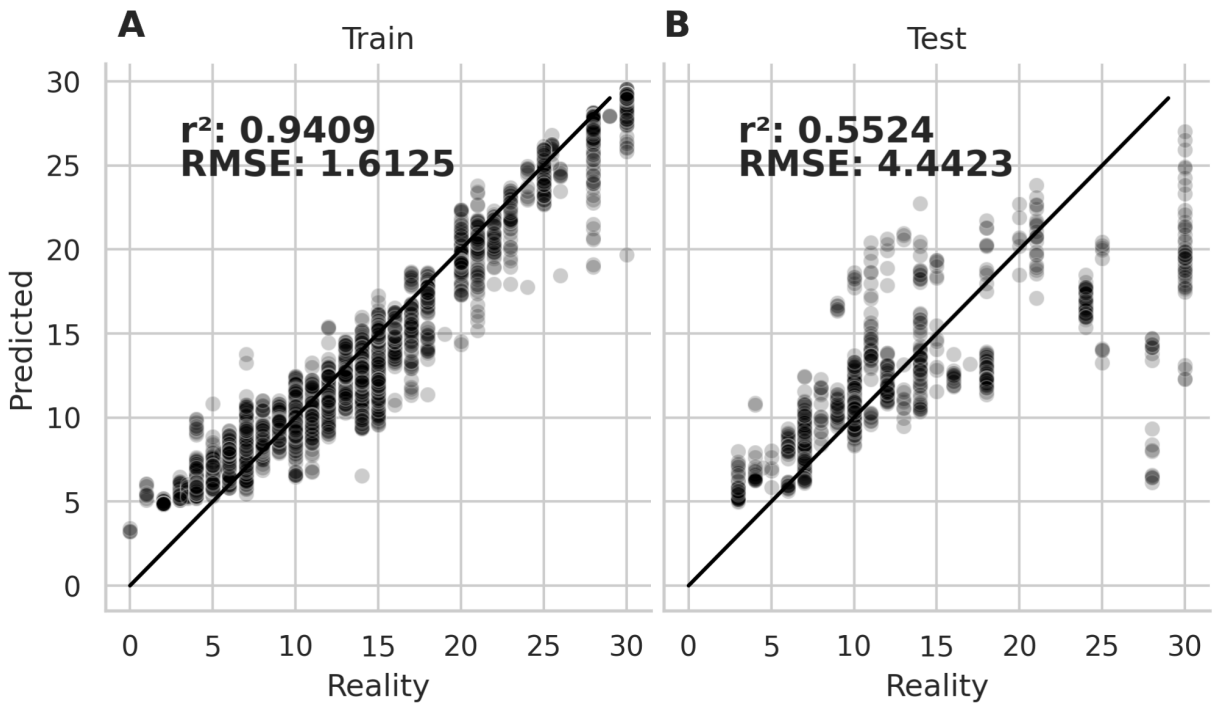


Figure 5: Model performance for DAL model. **A)** Results from the training dataset, and **B)** results from the testing dataset. The x-axis represents the actual age while the y-axis represents the age predicted by the model.

Age prediction using optional parameters in the regression model for the DAL model had an r^2 of 0.9409 and Root Mean Square Error (RMSE) of 1.6125 for the training dataset and an r^2 of 0.5524 and RMSE of 4.4423 for the testing dataset (**Figure 5**). Two other models were created using data from the complete “day” dataset. The first model was for predicting age from samples that were labeled as Days After Germination (DAG) and the second was for Days After Sowing (DAS). The DAG model had an r^2 of 0.9983 and RMSE of 0.2992 for the training, and an r^2 of 0.4493 and RMSE of 0.46904 for the testing. The DAS model had an r^2 of 0.9995 and RMSE of 0.1231 for the training, and an r^2 of -0.5712 and RMSE of 3.6729 for the testing. Both the DAG and DAS datasets had substantially fewer samples (distribution visualized in **Supplemental**

Figure 9 A and **D** respectively) than the DAL model. The DAS testing model only had testing data for 4 time points which allowed it to achieve an unwarranted better RMSE than the DAL model (which the very poor r^2 of -0.5712 revealed). We used the above models (tissue-6, tissue-4, and DAL) to predict annotations for the entire RNA-seq dataset, including previously unannotated samples. This table of predictions is available as **Supplemental Table 7,** which also includes information about the train test splits used for each model in this paper.

The performance of the DAG and DAS models was also performed using the DAL dataset as testing data (**Supplemental Figure 10**). The model performance of the DAG and DAS models on the DAL dataset was lower than the DAL model. This lower score indicates that the DAG and DAS datasets may be overfit due to the limited data.

An additional model was created using the DAL dataset supplemented using SMOTE. SMOTE synthesizes new data from minority classes (i.e., time points with fewer samples). It does this by drawing lines between random samples of closely related features of the minority class in the feature space. A random point along this line is taken which results in a new synthetic sample that has a resemblance to the two parent samples (Chawla et al., 2002). SMOTE is effective because it creates new data points which have a plausible feature space as compared to the class they were created from (Chawla et al., 2002). The SMOTE over-sampled DAL dataset contained 16112 samples, over three times the amount of the original training dataset size of 5062 (**Supplemental Figure 6**). The DAL SMOTE model offered a modest improvement over the DAL model, with an r^2 of 0.977 and RMSE of 1.2335 for the training dataset and an r^2 of 0.563 and RMSE of 4.3896 for the testing dataset (**Supplemental Figure 11**).

Assessing Annotation Accuracy

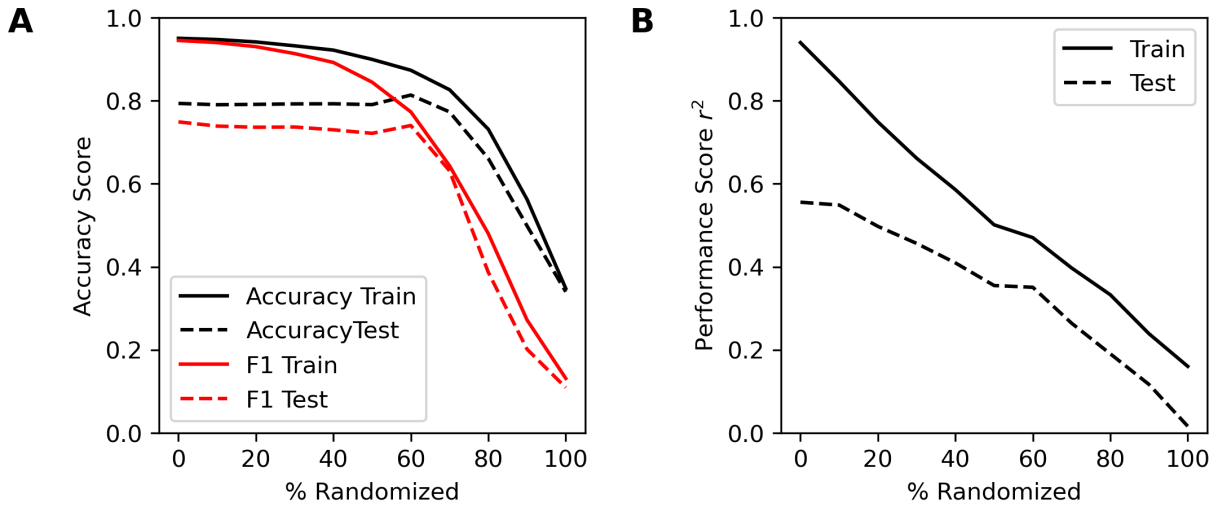


Figure 6: Model accuracy over an increasing amount of randomization for **A)** tissue, measured using both accuracy and F1 score, and **B)** age, measured using r^2. As the percent of samples with randomized tissue labels or age values increased, accuracy in the model both with the training and testing datasets decreased.

Annotation accuracy was assessed using randomization. For both tissue classification and DAL prediction, between 0 and 100% of the tissue labels or age values for the training datasets were randomly shuffled. These randomized training datasets were used to create a model, which was then evaluated on the testing dataset. Both models saw decreases in the performance for both training and testing with increasing randomization (**Figure 6**). The tissue classification did not lose substantial performance until higher amounts of randomization were introduced (**Figure 6 A**). The DAL model showed decreasing performance over increasing randomization, reaching nearly an r^2 of 0 at 100% randomization (**Figure 6 B**).

Number of Samples Required

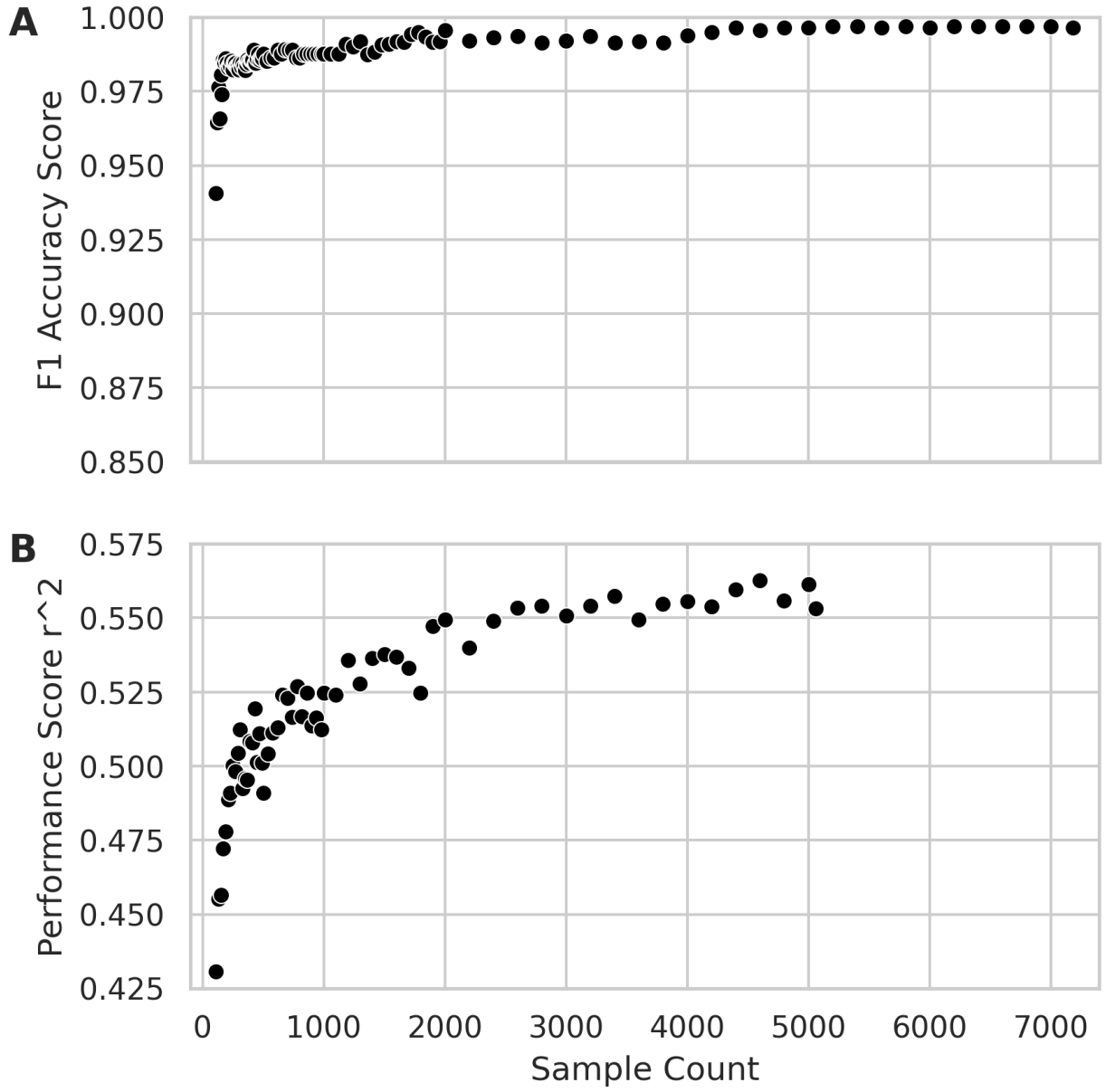


Figure 7: Model performance for different sample counts. Samples were added randomly from the training set, and a new model was created for each sample count. **A)** Model performance for tissue-4 classification dataset **B)** Model performance for DAL regression dataset. Both **A** and **B** are set at a y-axis range of 0.15, which excludes low-count models in preference of readability. For the full range which includes low-count model performance please see **Supplemental Figure 12**. Performance is based on testing datasets. Model performance for the tissue-6 dataset is included as **Supplemental Figure 13**.

The number of samples required for maximum accuracy was assessed for both models. Samples were randomized using two different approaches: either according to BioProject or randomly from the training datasets. The intent of this experiment was to determine the number of samples required to reach a plateau of model performance.

Performance results from the model where samples were added randomly from the training datasets for the tissue-4 classification models are shown in **Figure 7 A** for a F1 range from 0.850 to 1 (for the full range see **Supplemental Figure 12 A**). For this randomization approach, the model's accuracy was at near maximum after adding only a few hundred samples. This is in contrast with the regression DAL model, which reached its maximum only after a few thousand samples **Figure 7 B** (**Supplemental Figure 12 B** for the full range).

For the randomization approach accounting for BioProject, curves looked similar but were slightly delayed when compared to randomly sampling data (**Supplemental Figure 5**). This is likely due to BioProjects adding only 1 or a few time points. Curves reached the same plateaus as randomly adding samples.

Boruta and Gene Feature Importance

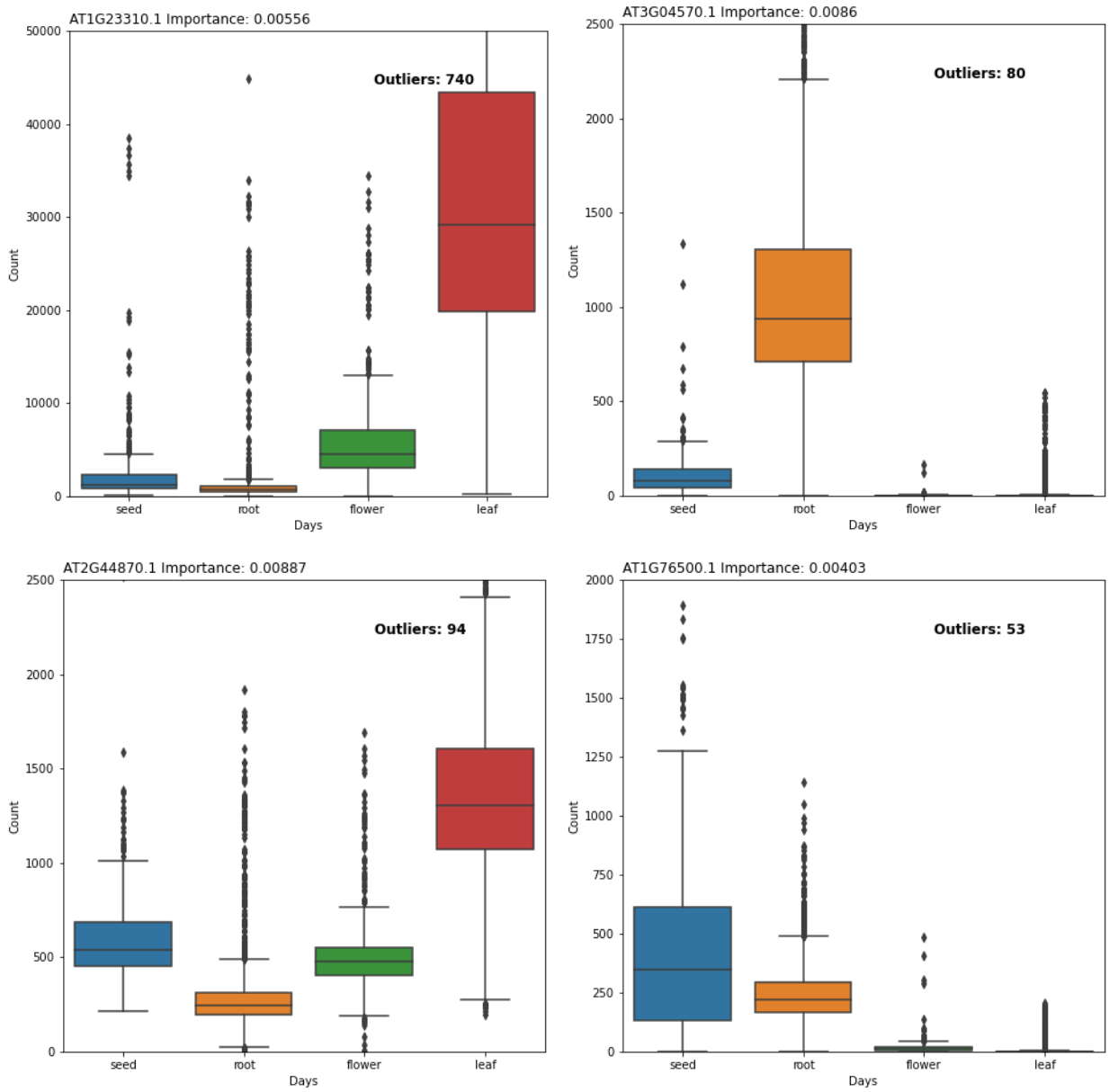


Figure 8: Four of the top-ranked genes from the tissue-4 classification model. Plots show large differences in expression between tissue labels. The number in the panel title next to the gene name is the feature importance of that gene in the model.

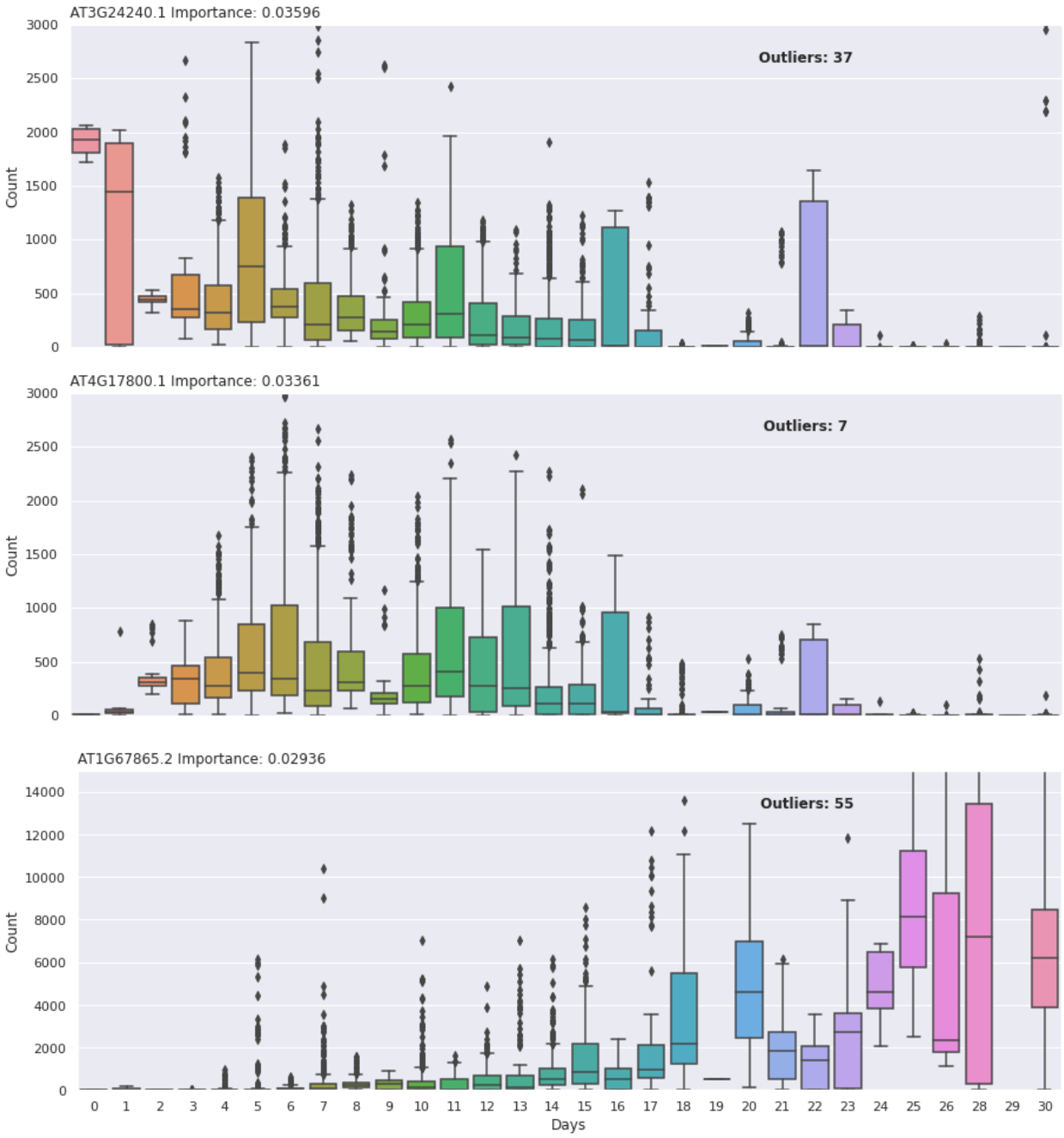


Figure 9: Top three genes for the DAL model. The x-axis represents days, and the y-axis represents MRN normalized counts. Ages falling between days (e.g. 5.5) were rounded up for the sake of plotting. For plotting, some outliers were removed to conserve y-axis space. Complete graphs for each of these genes are available in **Supplemental File 1**.

Boruta feature selection was performed using both the DAL, tissue-4, and tissue-6 datasets to identify genes that were better at predicting their respective labels and age values better than a randomized "shadow" feature of themselves. Boruta creates “shadow” features by randomizing the values of a gene across all samples. These “shadow” features are included during model creation, and if the true feature performs worse than this shadow feature, it is eliminated (Kursa et al., 2010). For the tissue-6 dataset, 20017 genes were shown to perform better than randomized features (accepted) and 23207 were rejected after 200 runs. For the tissue-4 dataset,7837 genes were accepted, and 34788 were rejected after 200 runs. For the DAL dataset, 15024 genes were accepted and 26394 were rejected after 200 runs. A reduced GEM was then made for both tissue-6, tissue-4, and DAL by keeping only those genes that were accepted. These reduced datasets represent the genes that perform better than random, however, Borutadoes not rank how well they perform. To rank these genes we used these reduced datasets to perform random forest feature selection. The list of genes and their rankings are available as **Supplemental Table 8** for tissue-6, **Supplemental Table 9** for tissue-4, and **Supplemental Table 10** for the DAL datasets. **Figure 8** shows four of the top genes selected as most important for classifying the tissue-4 of the sample. Note the difference in expression between the different samples. **Figure 9** shows three of the top genes selected for prediction in the DAL model. Ages falling between days (e.g. 5.5) were rounded up for the sake of plotting. **Supplemental Figure 14** shows 4 of the top genes for the tissue-6 model.

Discussion

For this study, we created a massive Arabidopsis RNA-seq dataset and included metadata such as categorical tissue labels and quantitative age values and used it to explore answers to two questions important for predictive biomarkers traits using gene expression data. These questions are: first, how many samples are required for such models, and second, how do normalization methods affect model performance, Towards this end, we randomized subsets of the massive Arabidopsis dataset to explore how performance changes as data size increases. We investigated four different normalization approaches. To maximize model performance we performed random forest parameter selection. To reduce the variables in the data, we applied Boruta and removed genes that were never predictive of the outcome.

Normalization Method

Four normalization methods were evaluated, TMM, MRN, TPM, and NoNo. In the tissue-6 dataset, the normalization method had no impact on model performance (**Figure 2 A**). This contrasted with the DAL dataset where a significant increase in model performance was seen for the MRN dataset over the other three methods (**Figure 2 B**). MRN normalization takes the geometric mean of each gene across samples and then uses this to calculate a ratio. The median of these ratios taken across each sample is the normalization factor for that sample (Anders & Huber, 2010). This technique takes into account all samples and all genes in the dataset when calculating these normalization factors. Thus, MRN shares information across samples to identify genes that are not differentially expressed (i.e., stable) and uses those as normalization

factors. TMM also does this, but with a weighted mean of log ratios-based strategy (Robinson & Oshlack, 2010), whereas TPM does not attempt to identify stable genes and normalizes gene counts in a sample to one million. NoNo data is unnormalized, and differences in sample size can have a major impact on relative expression levels to other datasets. In our experiment, the normalization method did not matter in the classification problem because the genes selected by the model had dramatic differences between tissue types (**Figure 8**). It seems, therefore, that it does not matter what normalization method is used if genes contributing to a label (such as tissue type) tend to be expressed only in their respective class labels. This contrasted with the age-in-days regression-based models, which relied on continuous, subtle differences between days to create a reliable predictive model (**Figure 9**). In the age case, if genes are not accurately normalized, the importance of genes can be reduced or not detected. We suspect some causes may be related to sequencing depth and the influence of highly differential genes. We also suspect that the differences in model performance between MRN and TMM normalizations are due to MRN more dramatically changing the underlying data, whereas TMM does not change the counts as much (as illustrated in **Supplemental Figure 2**). Whereas the difference between models may not be very noticeable when normalizing a single dataset (Evans et al., 2018), normalizing across our massive conglomerate dataset favored the more aggressive normalization of MRN. Further testing will need to be performed to determine if MRN is the best method of normalization for RNA-seq data in other situations.

The differences in how these normalization methods impact the overall gene count of all samples is illustrated in **Supplemental Figure 2**. To summarize, TPM

results in all samples having a million counts, TMM results in smoothing when compared to NoNo, and MRN results in more drastic smoothing compared to TMM.

Accuracy and Model Performance

Accurate annotations are important for constructing meaningful models. Ensuring accurate annotations of samples in large, conglomerate datasets can be difficult. Annotations used in this paper came from over 2000 independent scientific experiments, collected by an untold number of researchers. This results in a high likelihood there exist differences in experimental design (e.g. differences in temperature impacting development but not age), chances for errors in reporting (e.g. information being entered incorrectly or incorrect samples being uploaded to NCBI), and differences in opinion on how a certain annotation class should be reported (e.g. should age be reported as days after sowing or days after germination?). For conglomerate datasets such as our Arabidopsis data, manual effort was required to reduce the impacts of these annotation irregularities.

The tissue-6, tissue-4, and DAL models were performed with increasing amounts of randomized annotations to assess if the annotations reflected actual biology. The premise was that if the models were being overfit and not reflective of true biological variation then we would not see a decrease in model accuracy. However, we saw in all three cases that annotation prediction accuracy decreased. In the DAL models, there was a linear decrease (**Figure 6 B**) whereas in the tissue-6 (**Figure 6 A**) and tissue-4 models (**Supplemental Figure 15**) there was a delay in model accuracy decline. This delayed decline in accuracy in the tissue datasets is partially due to the actual amount

or randomization being different than the stated amount. This difference is due to the small number of possible tissue types resulting in the same annotation being randomly assigned to a sample (**Supplemental Figure 16)**. However, it also shows the robustness of random forest classification models. In these models the true transcriptomic signal needed to accurately predict the testing dataset remained in the model even under modest randomization because the training dataset separated out incorrectly annotated signals into other branches of the forest (what we refer to as “bad branches”. These “bad branches” accurately predicted poorly annotated labels, but did not impact the good branches needed to predict the testing dataset, allowing for a high training accuracy. This highlights an advantage of random forest classification that it can be fairly robust to poorly annotated data if the majority of data is accurately classified and there is true biological variation present. Random forest regression is also recalcitrant to outliers and poorly annotated data, but because it is predicting a continuous variable via averages it shows a linear decrease in performance with increasing randomness.

The two tissue models, tissue-6 and tissue-4, had excellent accuracy. However, tissue-6 illustrates the issue in large conglomerate datasets of annotation specificity. Of the 240 samples in the testing set labeled as “shoot” only 8 of them were accurately labeled as such by our prediction model (3.3% accuracy) (**Figure 3 B** and **D**). The vast majority of “shoot” labeled samples were predicated as either “seedling” or “leaf”. This is likely because “shoot” is an ambiguous label when considering Arabidopsis plants. Whereas a young seedling (stage 0-1) (Boyes et al., 2001) has an obvious shoot, an older Arabidopsis plant does not. This results in the shoot label sharing a large amount

of biological similarity with either the seedling label or the leaf label (which includes full rosette) making them difficult to distinguish. The other poorly defined term is “seedling”, which also has subjectivity built into it. There is not a well-defined annotation for what an Arabidopsis seedling is, so we decided to use it before stage one which is around 6 days post-germination (Boyes et al., 2001). However, many of the annotations were sparse and we were not able to distinguish if a “seedling” annotation fit our criteria. It is likely that a large amount of “seedling” labeled samples do not fit our definition, which is illustrated by some of the available annotations being, “leaf (20-day seedling)”, “15-day-old seedling” and “21 day old seedling”. These should be plants that are well beyond this stage, and in some instances with several sets of true leaves. The reduced tissue-4 model mitigates these issues by removing the ambiguous labels, and instead concentrates solely on four well-defined categories, resulting in excellent accuracy. However, it trades this increased accuracy for a reduced number of annotation categories.

The DAL age dataset was also subject to inaccuracies in labeling. This dataset consisted of only samples which were annotated with the word “day” or equivalent. It is unclear if the meaning of each of these labels is referring to “days after sowing”, “days after germination” or “days after stratification”. It could be that these dates should be shifted anywhere from 0-6 days if they are to represent the same timescale. This uncertainty in actual date precision is captured in our DAL model, with the testing dataset having an RMSE of 4.4423, which can be thought of as the average number of days the prediction deviates from the true age. The DAL model was able to capture

~55% of the variability in the DAL dataset, which is good considering the issues with annotations.

To explore if we could mitigate the issue of precision with the age annotations, we tested two additional datasets: DAG and DAS. The DAG and DAS age categories were more precise in how age was reported on NCBI, specifying whether the measurements are referring to time after germination or sowing respectively. Unfortunately, the more specific annotations for DAG and DAS failed to achieve higher model accuracy when compared to DAL (**Supplemental Figure 9**). This is because we did not have as many samples from as many BioProjects to create models. However, if we compare the performance of the DAG dataset (r^2 = .45, **Supplemental Figure 9 A,B,C)** to that of the DAL model created using a comparable number of BioProjects (45 BioProjects)(r^2 = .28, **Supplemental Figure 4 C**) we see better performance of the DAG dataset than the DAL. This illustrates how improved precision in age can potentially make better models, and that insufficient accurately annotated samples are currently available. We encourage researchers to submit as accurate information as possible when submitting BioSample data about their projects to NCBI.

Number of Samples Required

One of the main goals of this project was to assess how many samples are required for annotation model creation of RNA-seq datasets and if samples from disparate experiments could be used together. This was explored for both the categorical tissue datasets (tissue-6 and tissue-4) as well as the quantitative age dataset (DAL). Tissue classification rapidly achieved high performance after only a few hundred samples were

added, whereas age prediction performance only reached a plateau after 2-3 thousand samples (**Figure 7 A** and **B**). This is likely due to the differences in the complexity of the models. The tissue classification problem was only attempting to classify samples into 6 or 4 categories, with each of these being distinct tissues of Arabidopsis. It is known that there are a number of genes that are only active in certain tissues (Shi et al., 2021). In contrast, the age regression model was over a continuous range of 0-30 days. In addition to the large number of ages of the sample (0-30 days) the model also had to take into account all of the other latent variables present within these samples. These latent variables include differences in genotype, temperature, moisture level, nutrient availability, gene knockouts, growth media, pest pressure, and tissue type. All of these factors may have an impact on gene expression in a manner semi-independent of gene expression related to age. This means that age models must either identify genes unimpacted by these latent variables or independently predict age by taking into account these latent variables. In reality, it is most likely a trade-off between these two cases, where the model includes genes mostly predictive of only age and genes that take into account other variables that may impact the age variable **Figure 9**. While tissue-related expression patterns are also impacted by these latent variables, it appears that tissue type differentiation has unique enough expression patterns to accurately differentiate between them (tissue-4 **Figure 8**, tissue-6 **Supplemental Figure 14**).

Supplementing the DAL dataset using SMOTE resulted in a modest increase in annotation accuracy. SMOTE is intended to be used on classification datasets, not continuous datasets, so in order to use SMOTE on the DAL dataset we rounded the

days and converted them to distinct categories. After converting back to numerical values, we saw a modest increase in model performance (**Supplemental Figure 11**). This increased performance as a result of balancing by SMOTE illustrates the importance of balanced datasets for model prediction. Balancing unbalanced datasets is a currently active area of research, especially for continuous variable models (Yang et al., 2021), and a consensus on how to balance tabular datasets, such as ours, does not have an agreed-upon solution. Here we show that using SMOTE on our continuous variable age had a modest increase in model performance. Ultimately, this increase was not dramatic enough to warrant the decrease in model interpretability resulting from introducing artificially generated data.

Here we do not propose a definitive cutoff for the number of samples required to create an accurate predictive model using RNA-seq datasets. Differences in datasets, experimental design, biological impact, and sample annotation quality will have an impact on model accuracy, so broad recommendations are warranted. As already stated, tissues tend to have genes that are uniquely expressed thus this variable represents a "simple" model, or one with fewer multi-functional, co-dependent, or conditional relationships amongst genes. Thus, we expect the tissue model to represent a lower bound in terms of the number of required samples. We point readers to **Figure 7** and suggest that biomarker experiments should include at least a few hundred samples for accurate prediction. The age models indicate that for more complicated traits (such as for environmentally impacted traits), a few thousand samples may be required.

Conclusion

The research presented here provides foundational knowledge about possible sample size requirements and the effects of RNA-seq normalization for transcriptomic biomarker model development. Results provide guidance on the minimum number of samples and normalization methods that may be needed for accurate models. Such guidance has applications for the development of predictive transcriptomic biomarkers which are gaining popularity in precision medicine and specialty agricultural crops. Finally, we encourage researchers submitting RNA-seq samples to NCBI, or other public repositories, to provide correct and comparable annotations for their samples.

REFERENCES

Acharjee, A., Kloosterman, B., Visser, R. G. F., & Maliepaard, C. (2016). Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics*, *17 Suppl 5*(Suppl 5), 180.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106.

Azodi, C. B., Pardo, J., VanBuren, R., de Los Campos, G., & Shiu, S.-H. (2020). Transcriptome-Based Prediction of Complex Traits in Maize. *The Plant Cell*, *32*(1), 139–151.

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K. D., Resenchuk, S., Tatusova, T., Yaschenko, E., & Ostell, J. (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research*, *40*(Database issue), D57–D63.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The Arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *Genesis* , *53*(8), 474–485.

Bostanci, E., Kocak, E., Unal, M., Guzel, M. S., Acici, K., & Asuroglu, T. (2023). Machine Learning Analysis of RNA-seq Data for Diagnostic and Prognostic Prediction of Colon Cancer. *Sensors* , *23*(6). https://doi.org/10.3390/s23063080

Boyes, D. C., Zayed, A. M., Ascenzi, R., McCaskill, A. J., Hoffman, N. E., Davis, K. R., & Gorlach, J. (2001). Growth stage-based phenotypic analysis of Arabidopsis: A model for high throughput functional genomics in plants. *The Plant Cell*, *13*(7), 1499.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *The Journal of Artificial Intelligence Research*, *16*(1), 321–357.

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1), 270.

Evans, C., Hardin, J., & Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, *19*(5), 776–792.

Favre, L., Hunter, D. A., O'Donoghue, E. M., Erridge, Z. A., Napier, N. J., Somerfield, S. D., Hunt, M., McGhie, T. K., Cooney, J. M., Saei, A., Chen, R. K. Y., McKenzie, M. J., Brewster, D., Martin, H., Punter, M., Carr, B., Tattersall, A., Johnston, J. W., Gibon, Y., … Brummell, D. A. (2022). Integrated multi-omic analysis of fruit maturity identifies biomarkers with drastic abundance shifts spanning the harvest period in "Royal Gala" apple. *Postharvest Biology and Technology*, *193*, 112059.

Federhen, S., Clark, K., Barrett, T., Parkinson, H., Ostell, J., Kodama, Y., Mashima, J., Nakamura, Y., Cochrane, G., & Karsch-Mizrachi, I. (2014). Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Standards in Genomic Sciences*, *9*(3), 1275–1277.

Feng, X., Zhang, R., Liu, M., Liu, Q., Li, F., Yan, Z., & Zhou, F. (2019). An accurate

regression of developmental stages for breast cancer based on transcriptomic biomarkers. *Biomarkers in Medicine*, *13*(1), 5–15.

Gapper, N. E., Rudell, D. R., Giovannoni, J. J., & Watkins, C. B. (2013). Biomarker development for external CO2 injury prediction in apples through exploration of both transcriptome and DNA methylation changes. *AoB Plants*, *5*, 1–9.

Hadish, J. A., Biggs, T., Shealy, B., Wytko, C., Smith, M., Feltus, F. A., & Ficklin, S. P. (2021, July 30). *GEMmaker: Automated Scalable Processing of Large RNA-Seq Datasets*. Bioinformatics Open Source Conference (BOSC).

Hatoum, D., Hertog, M. L. A. T. M., Geeraerd, A. H., & Nicolai, B. M. (2016). Effect of browning related pre- and postharvest factors on the “Braeburn” apple metabolome during CA storage. *Postharvest Biology and Technology*, *111*, 106–116.

Herman, G. R., & Schumacher, R. S. (2018). “Dendrology” in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation. *Monthly Weather Review*, *146*(6), 1785–1812.

Homola, D., & Beanico, M. (2019). *BorutaPy* (Version 0.3) [Computer software]. https://github.com/scikit-learn-contrib/boruta_py

Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta - A system for feature selection. *Fundamenta Informaticae*, *101*(4), 271–285.

Leisso, R. S., Gapper, N. E., Mattheis, J. P., Sullivan, N. L., Watkins, C. B., Giovannoni, J. J., Schaffer, R. J., Johnston, J. W., Hanrahan, I., Hertog, M. L. A. T. M., Nicolaï, B. M., & Rudell, D. R. (2016). Gene expression and metabolism preceding soft scald, a chilling injury of “Honeycrisp” apple fruit. *BMC Genomics*, *17*(1), 1–23.

Lemaitre, G., Nogueira, F., & Aridas, C. (2017). Imbalanced-learn: A Python Toolbox to

Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*. https://github.com/scikit-learn-contrib/imbalanced-learn

Limeta, A. (2020). *BioSampleParser*. https://github.com/angelolimeta/BioSampleParser

Li, W. V., & Li, J. J. (2018). Modeling and analysis of RNA-seq data: a review from a statistical perspective. *Quantitative Biology (Beijing, China)*, *6*(3), 195–209.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Meshcheryakov, G. (2021). *conorm* (Version 1.2.0) [Computer software]. https://gitlab.com/georgy.m/conorm

NCBI Resource Coordinators. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *44*(D1), D7–D19.

Pedregosa, F., Varoquaux, G., & Gramfort, A. (2011). Scikit-learn: Machine learning in Python. *Of Machine Learning …*. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , *26*(1), 139–140.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25.

Rokach, L. (2016). Decision forest: Twenty years of research. *An International Journal on Information Fusion*, *27*, 111–125.

Shi, D., Jouannet, V., Agustí, J., Kaul, V., Levitsky, V., Sanchez, P., Mironova, V. V., &

Greb, T. (2021). Tissue-specific transcriptome profiling of the Arabidopsis inflorescence stem reveals local cellular signatures. *The Plant Cell*, *33*(2), 200–223.

Smith, A. M., Walsh, J. R., Long, J., Davis, C. B., Henstock, P., Hodge, M. R., Maciejewski, M., Mu, X. J., Ra, S., Zhao, S., Ziemek, D., & Fisher, C. K. (2020). Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics*, *21*(1), 119.

Somssich, M. (2019). *A short history of Arabidopsis thaliana (L.) Heynh. Columbia-0* (No. e26931v5). PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.26931v5

Supplitt, S., Karpinski, P., Sasiadek, M., & Laczmanska, I. (2021). Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. *International Journal of Molecular Sciences*, *22*(3). https://doi.org/10.3390/ijms22031422

Van Der Maaten, L., Postma, E., & Van Den Herik, J. (2009). Dimensionality Reduction: A Comparative Review. *Tilburg Centre for Creative Computing*. https://members.loria.fr/moberger/Enseignement/AVR/Exposes/TR_Dimensiereductie.pdf

Yang, Y., Zha, K., Chen, Y.-C., Wang, H., & Katabi, D. (2021). Delving into Deep Imbalanced Regression. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2102.09554

Supplemental Materials

Supplemental Tables

Supplemental Table 1: Arabidopsis RNA-seq SRA RunInfo Retrieved from NCBI November 2022.
Included as a separate file.

Supplemental Table 2: Arabidopsis BioSample data retrieved from NCBI.
Included as a separate file.

Supplemental Table 3: Summary of the splits for the Tissue Dataset. Shows total sample counts and number of BioProjects in each. Most BioProjects consisted of a single type of tissue type, but a few consisted of multiple which is why a sum of the BioProjects Train does not match with total number of BioProjects.

	Samples in Train	BioProjects in Train	Samples in Test	BioProjects in Test	Samples Ratio	BioProject Ratio
flower	603	66	120	13	0.199005	0.19697
leaf	4216	288	1105	76	0.262097	0.263889
root	1920	159	516	43	0.26875	0.27044
seed	442	46	247	13	0.558824	0.282609
seedling	4666	362	1294	78	0.277325	0.21547
shoot	902	59	240	20	0.266075	0.338983

Supplemental Table 4: Tukey HSD between the different normalization methods for Age. MRN was significantly higher than the other normalization methods.
There are significant differences among the groups.
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
MRN	NoNo	-0.0303	0.0028	-0.0526	-0.008	True
MRN	TMM	-0.0372	0.0001	-0.0595	-0.0149	True
MRN	TPM	-0.0248	0.022	-0.0471	-0.0026	True
NoNo	TMM	-0.0069	0.8549	-0.0292	0.0154	False
NoNo	TPM	0.0055	0.9217	-0.0168	0.0278	False
TMM	TPM	0.0124	0.4807	-0.0099	0.0347	False

Supplemental Table 5: Parameter Optimization Results for RandomForest Classification Tissue.
Included as a separate file.

Supplemental Table 6: Parameter Optimization Results for Random Forest Regression Age.
Included as a separate file.

Supplemental Table 7: Predictions for 32432 RNA-seq Samples. Prediction columns for the three models are: 'tissue_6_prediction', 'tissue_4_prediction','DAL_prediction'. Ground truth annotation columns are 'tissue', 'days_age', 'age_category', and 'age_category_full_name'. Note that ground truth columns do not have values for every RNA-seq sample, as all samples do not have ground truth information about their annotations. The first three columns are information about the RNA-seq sample name: 'experiment', BioProject: 'bioproject_name', and BioSample: 'biosample_name'. Also has information about if the sample was in the training or testing set for each model: 'tissue_4_train_test', 'tissue_6_train_test', 'DAL_train_test'.

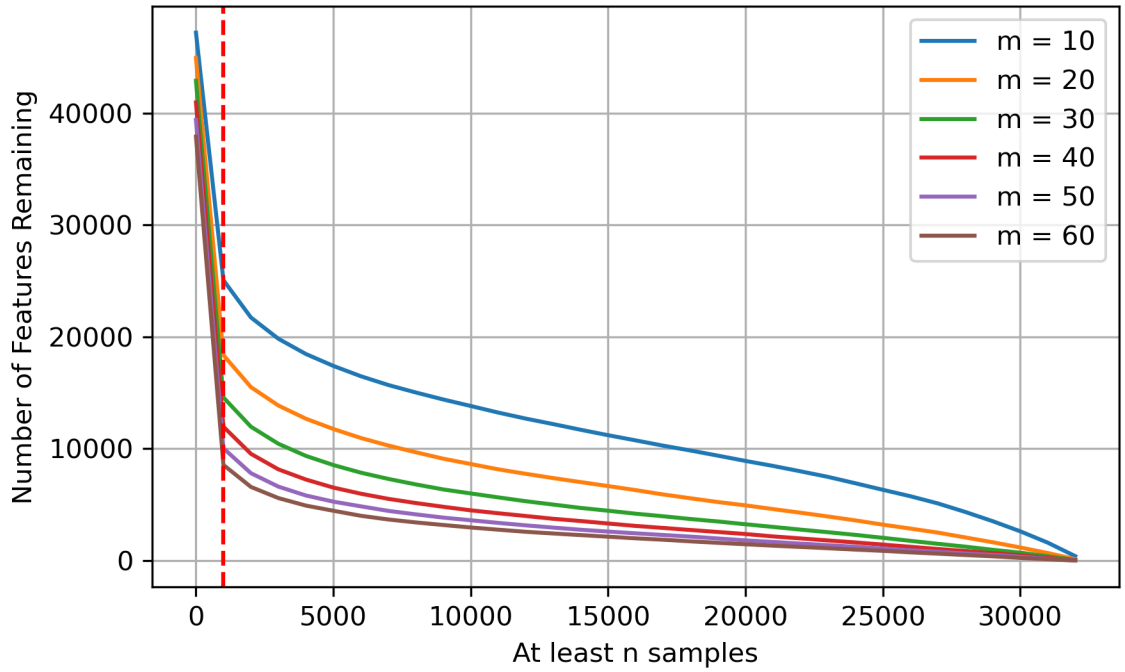
Supplemental Table 8: Feature Importance Tissue-6 After Boruta.
Included as a separate file.

Supplemental Table 9: Feature Importance Tissue-4 After Boruta.
Included as a separate file.

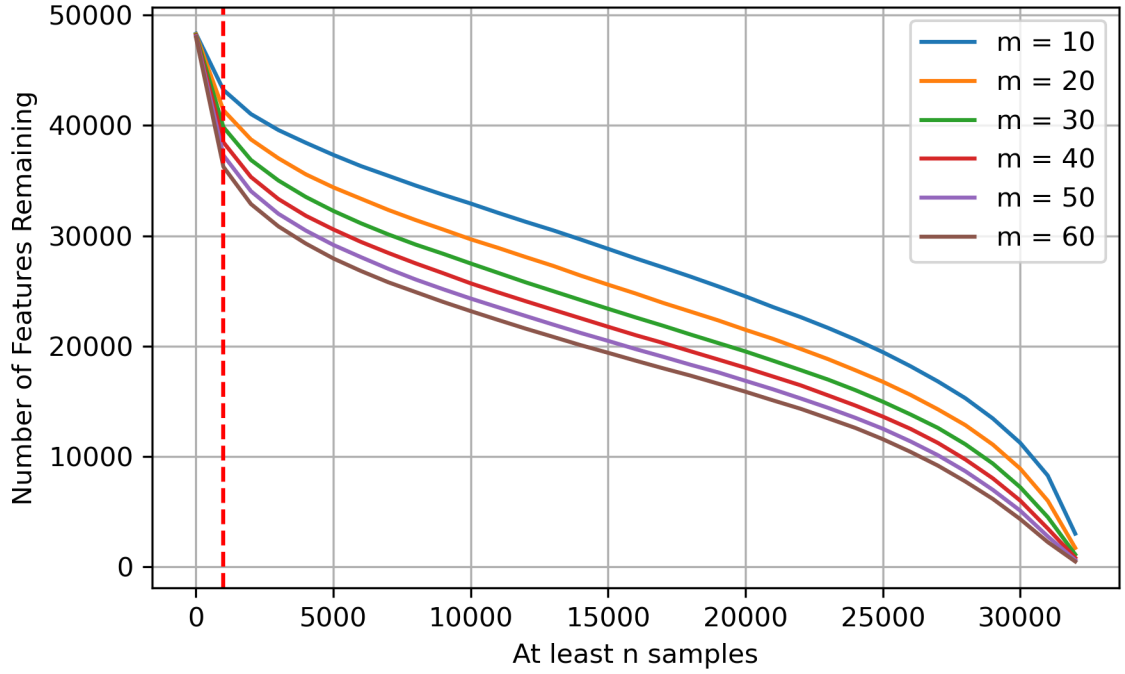
Supplemental Table 10: Feature Importance DAL After Boruta.
Included as a separate file.

Supplemental Figures

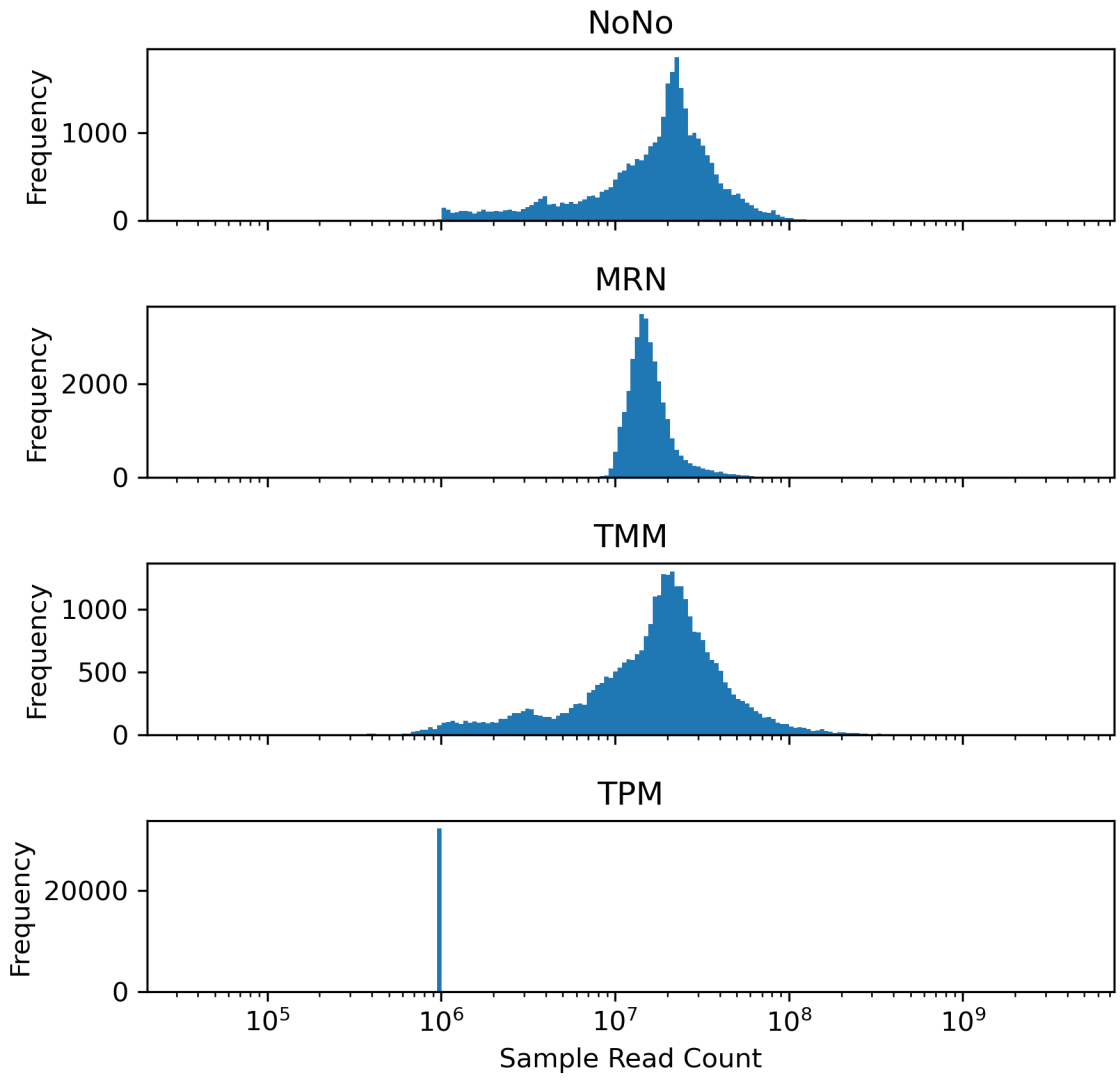
A



B

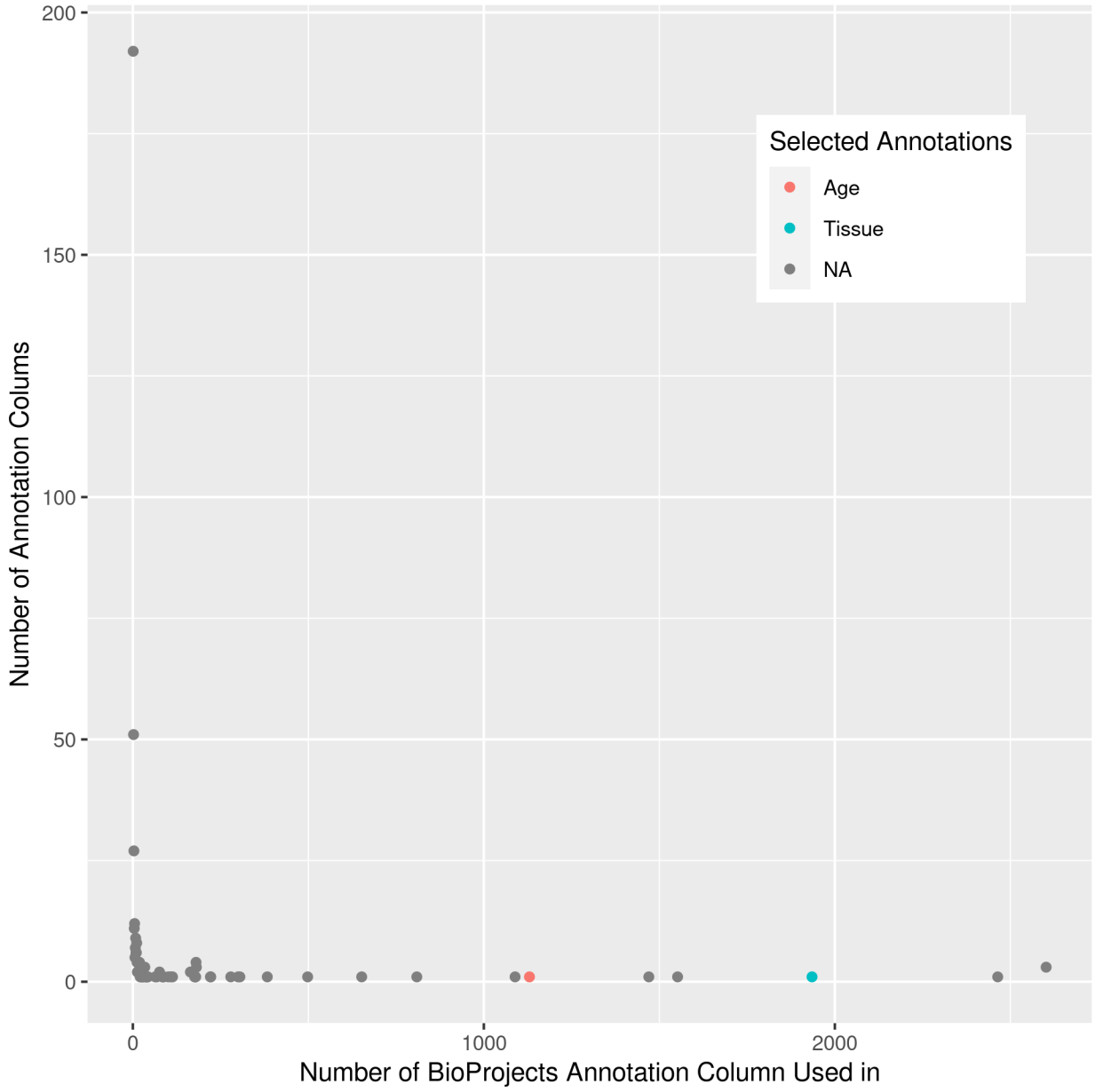


Supplemental Figure 1: Number of features remaining that are over m count for n samples (x-axis). **A)** Number of genes remaining for TPM with different minimum counts (m), **B)** Same except for NoNo. Red Line is at 1000 samples. Final thresholding was based on the NoNo dataset, with all other datasets (MRN, TMM, and TPM) containing the same genes.

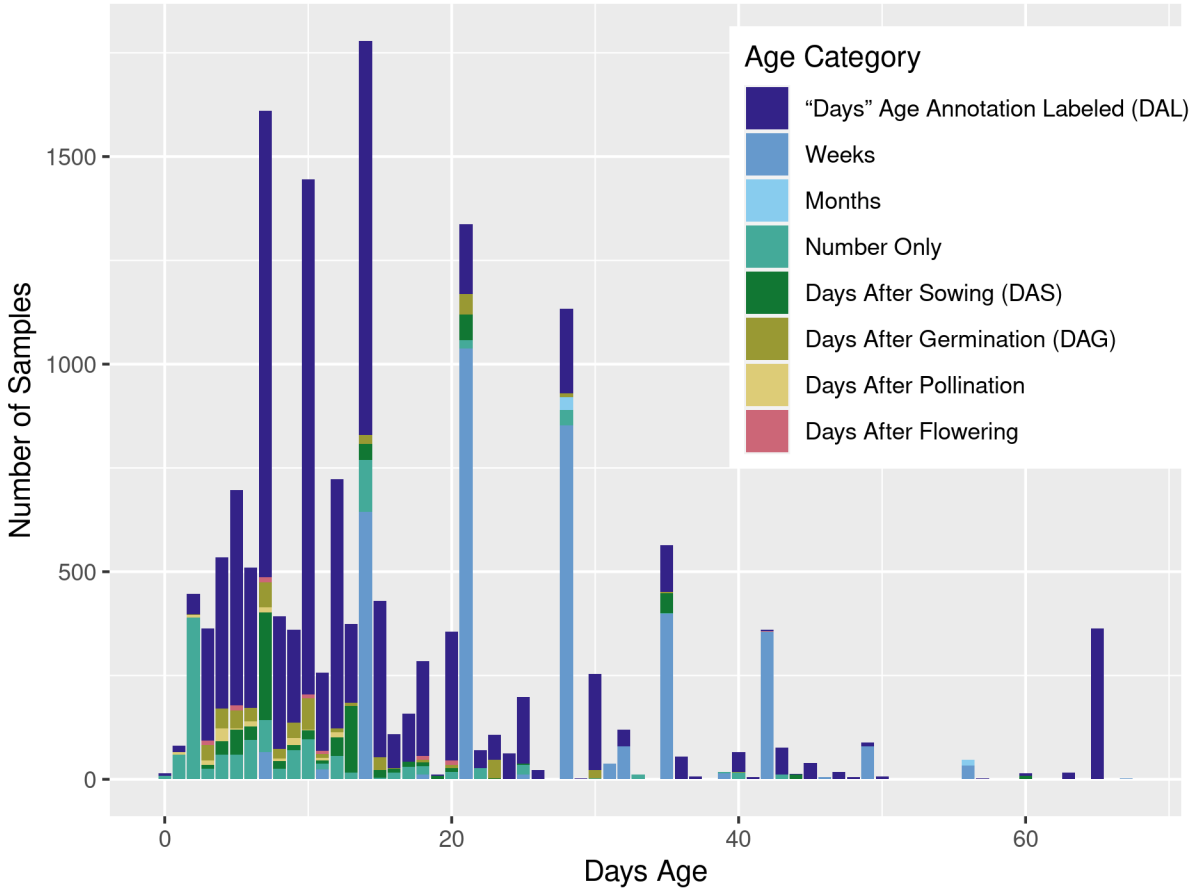


Supplemental Figure 2: Histogram of total sample read count for the four normalization processes used. The x-axis is log2 transformed and is the same for each plot. The y-axis maximum value is different in each plot. *Abbreviations:* Trimmed Mean

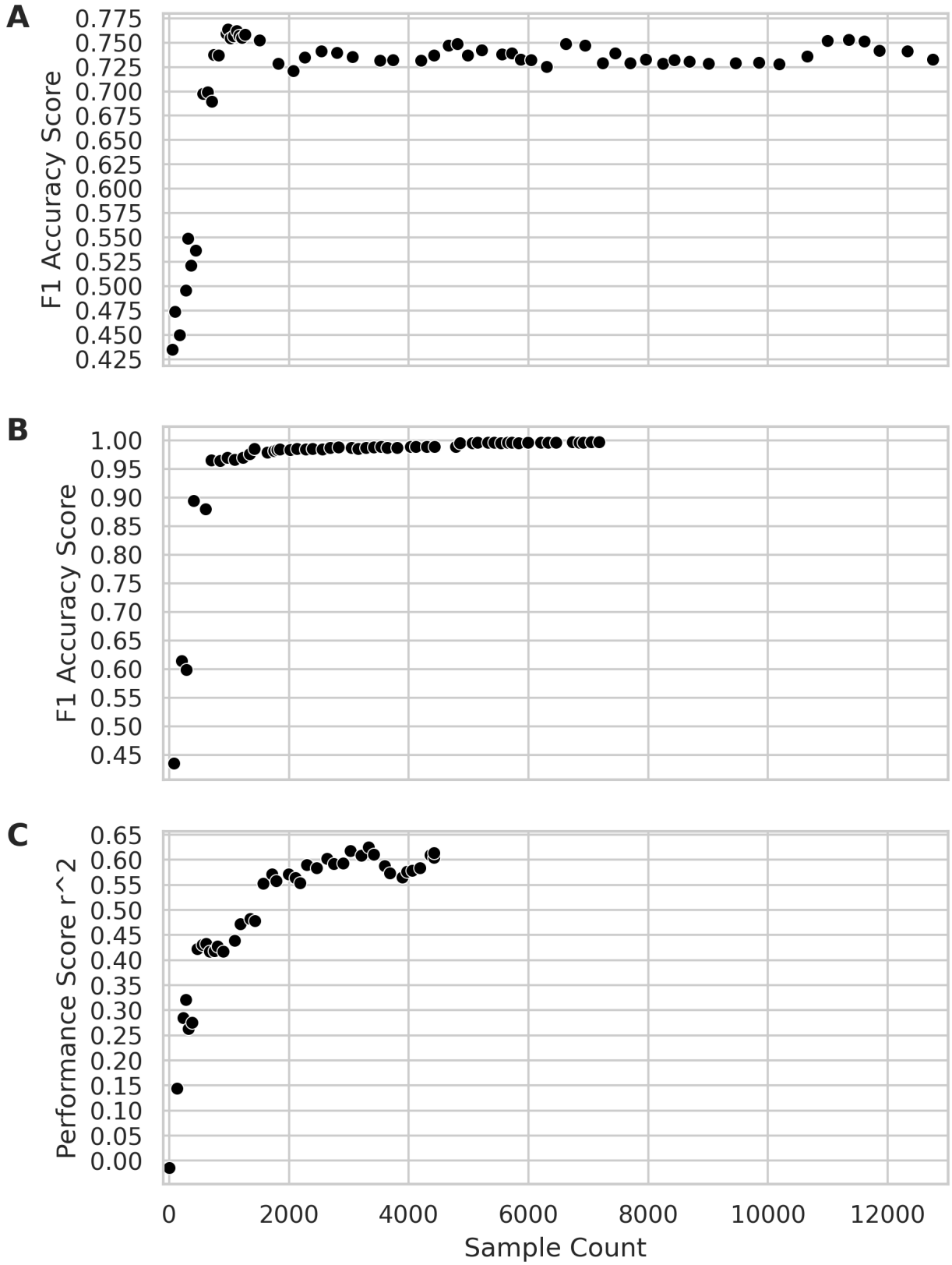
of M values (TMM), Median of Ratios Normalization (MRN), Transcripts per kilobase million (TPM), unnormalized count data (NoNo).



Supplemental Figure 3: Diagram showing the sparsity of data. The annotations retrieved are very sparse, with many annotation columns only used for one or a few BioProjects. The annotations (Tissue and Age) we use for this project are highlighted in color. The x-axis represents the number of BioProjects which use an annotation, and the y-axis represents the number of annotations which are at that category. We were able to highlight tissue and age because they are the only annotation which is present for that number of BioProjects.

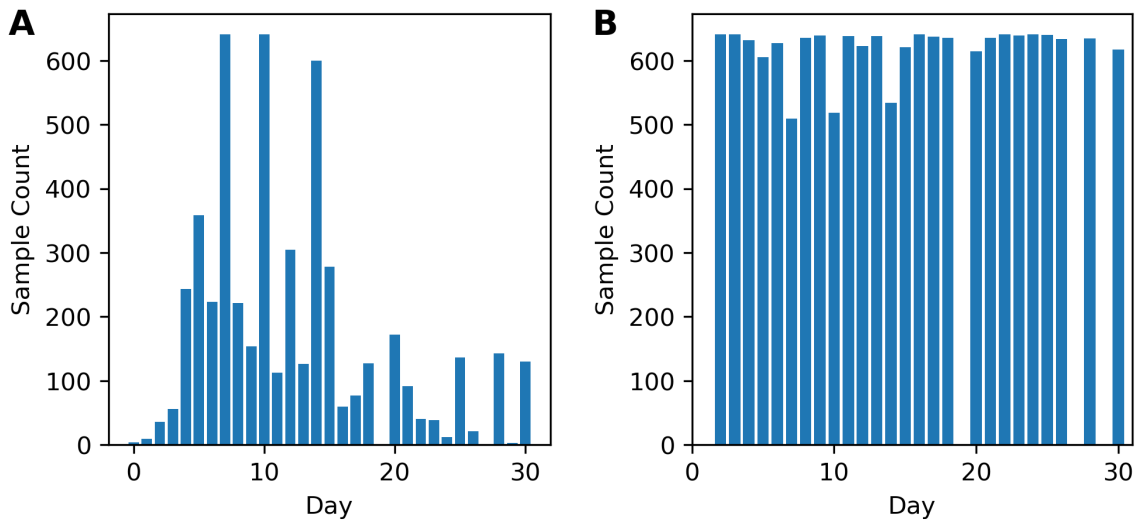


Supplemental Figure 4: Age categories over time. Annotations for age were split into categories based on how the BioSample information for age was reported (i.e. if sample age information was reported as “10 Days after Germination” It would be reported as “Days after Germination (DAG)” whereas if it was just a number “10”, then it would be reported as “Number Only”). Categories were: ""Days" Age Annotation Labeled (DAL)", "Weeks", "Months", "Number Only", "Days After Sowing (DAS)", "Days After Germination (DAG)", "Days After Pollination", and "Days After Flowering". Additional information can be found in Table 2 in the text.

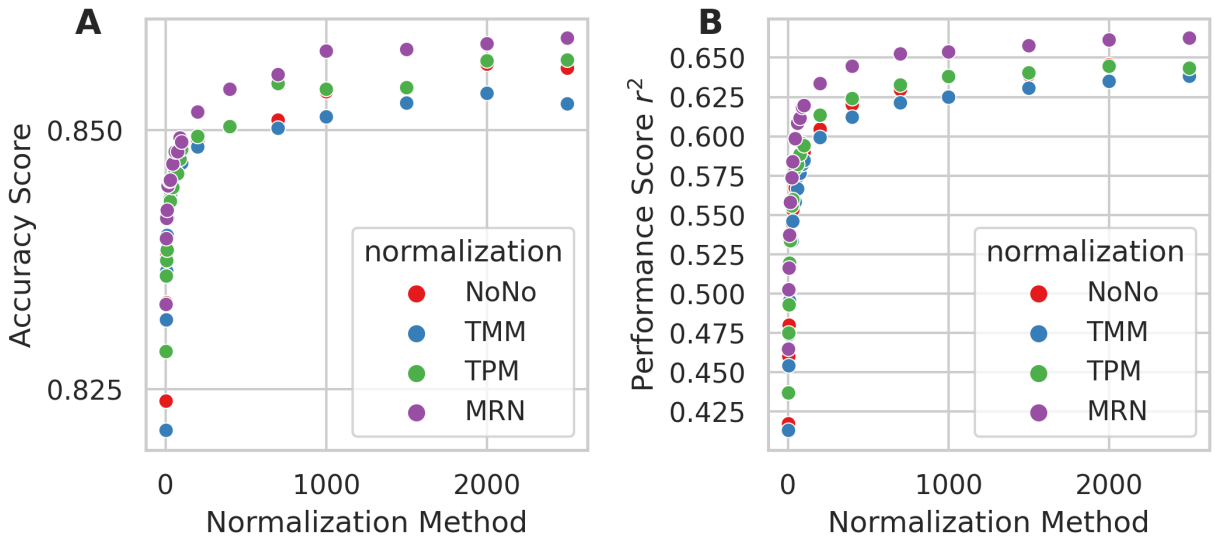


Supplemental Figure 5: Model Performance Adding by BioProjects. BioProjects were

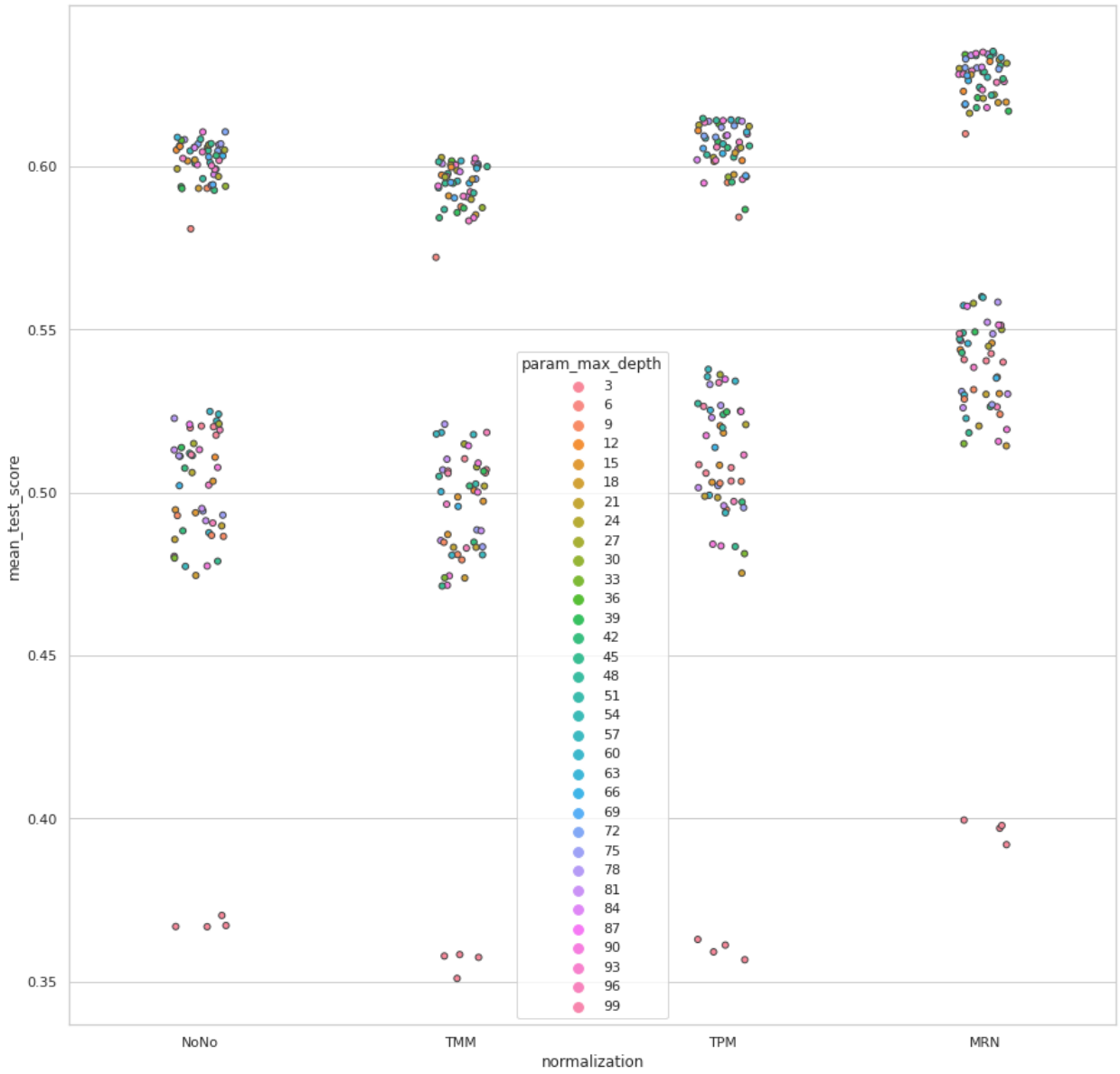
added 10 at a time. Performance is always measured on the respective testing dataset. **A)** Tissue-6 Performance. **B)** Tissue-4 performance **C)** DAL Performance. The x-axis are the same for each model.



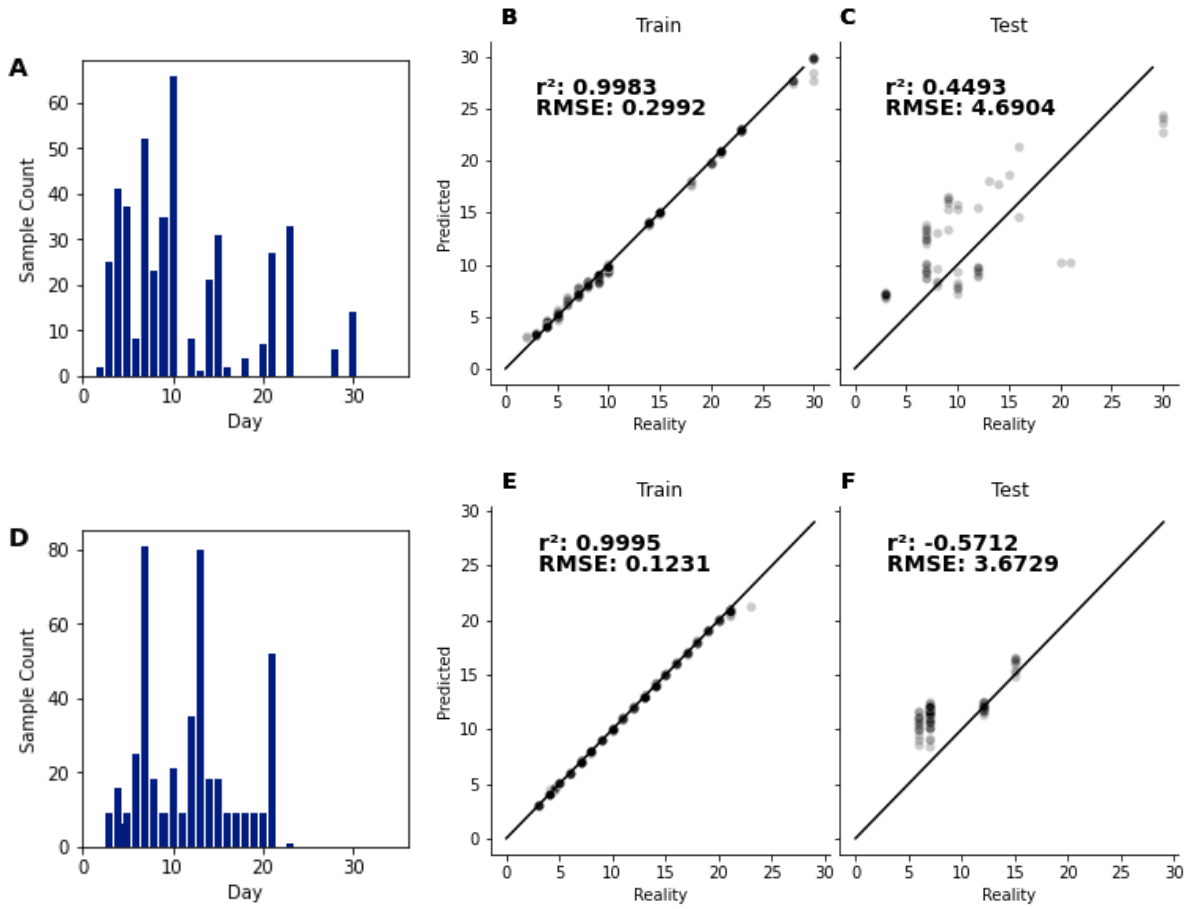
Supplemental Figure 6: SMOTE Resampling of Regression Data. Dates with half days were rounded up to the nearest date. Dates with sample count under 10 samples were removed before running SMOTE (0,1,19,27, 29). **A)** Sample distribution before SMOTE **B)** Sample distribution after SMOTE



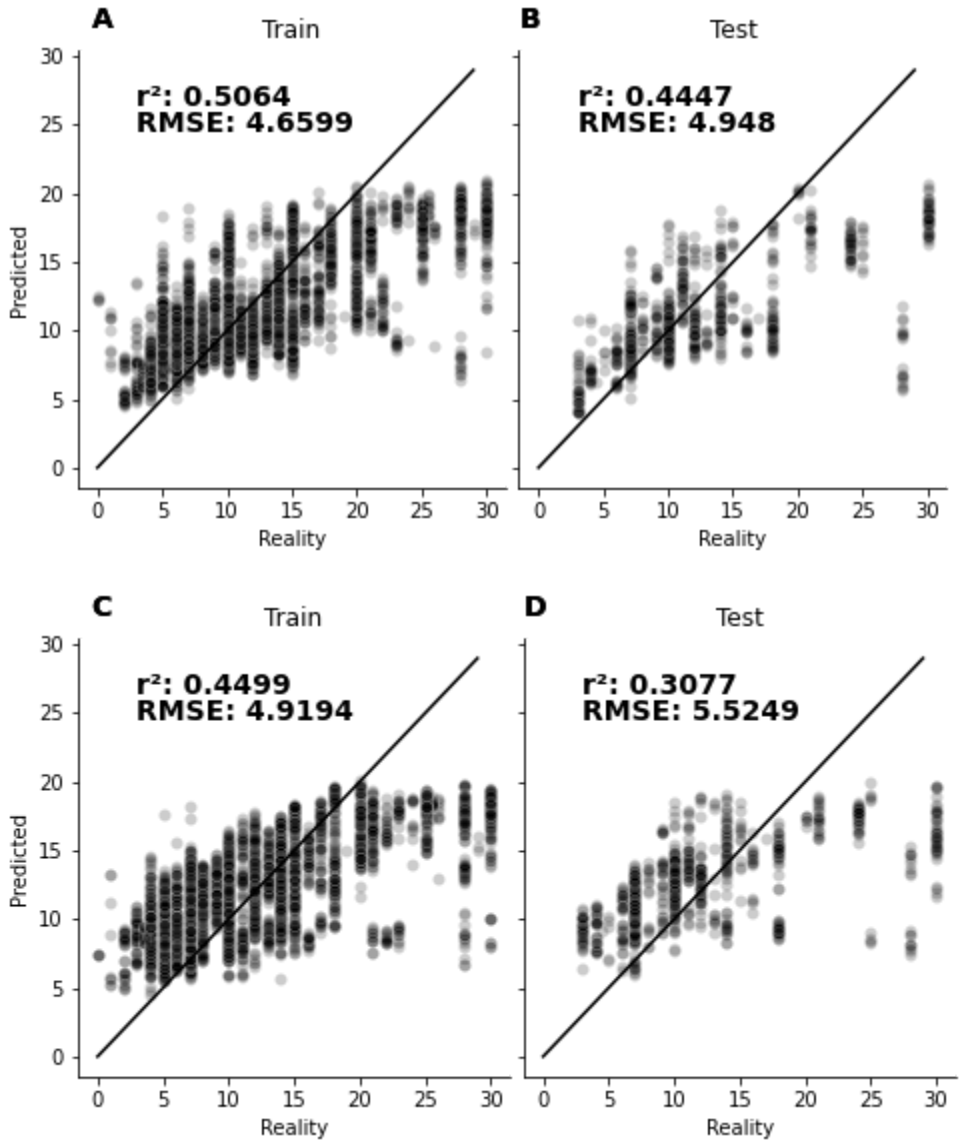
Supplemental Figure 7: GridSearchCV of different *max_feature* depths (x-axis) for the **A)** tissue-6 classification model and **B)** DAL regression model. Note that y-axis is scaled at 0.025 between ticks for both plots.



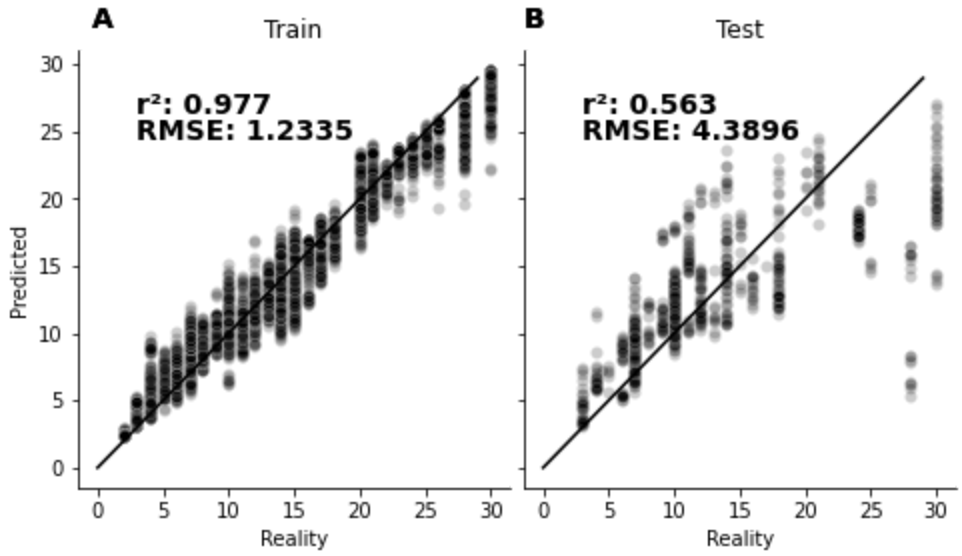
Supplemental Figure 8: Coloration of Figure B as “max_depth” to illustrate that the very low points are just too low of “max_depth”



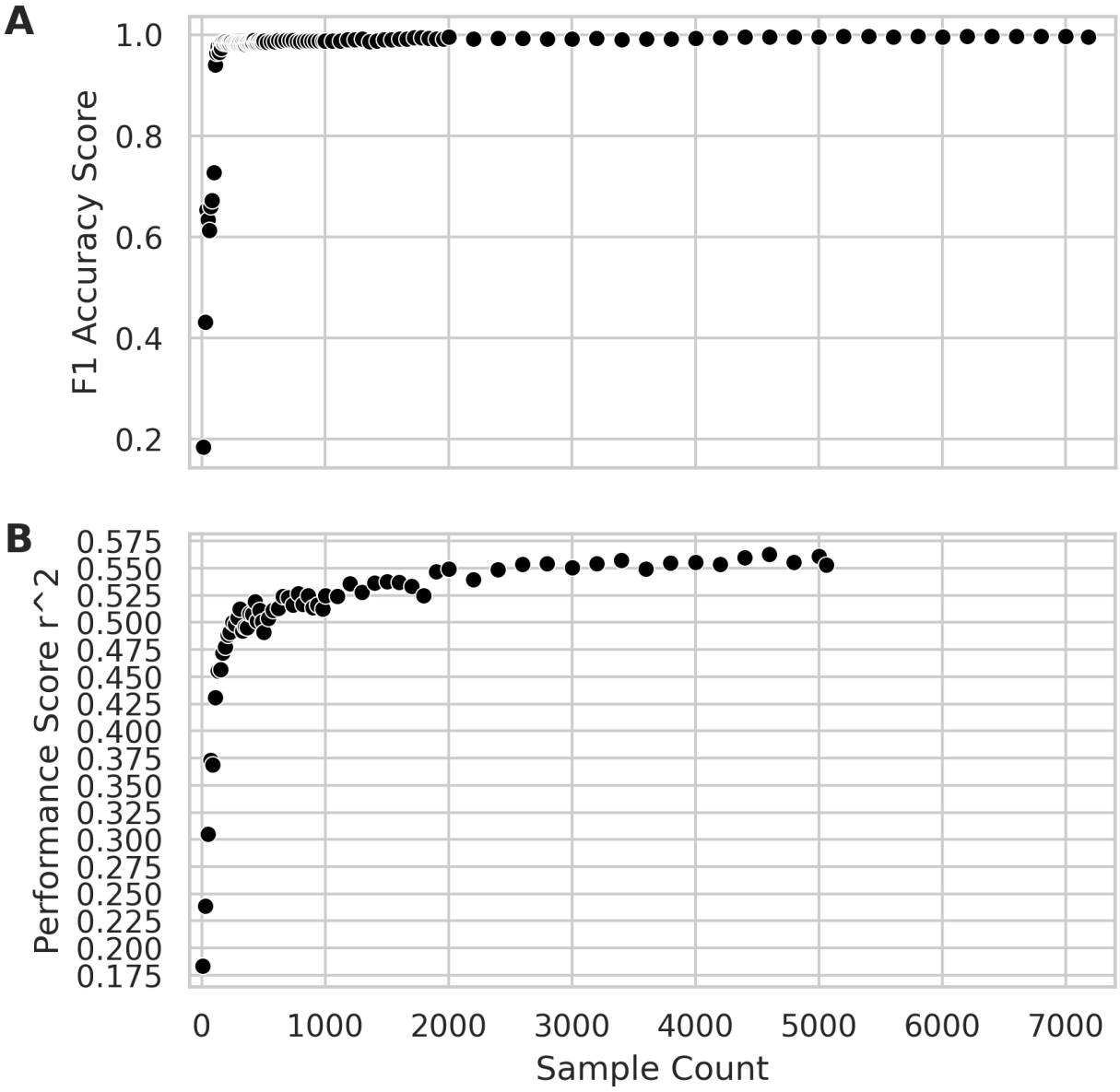
Supplemental Figure 9: Age models using samples classified as Days After Germination (DAG)(**A,B,C**) and Days After Sowing (DAS)(**D,E,F**)**.** Plots **A** and **B** represent the distribution of samples for DAG and DAS respectively. Plots **B** nd **C** represent the training and testing performance (r^2) for DAG and **E** and **F** represent the training and testing performance for DAS. DAG model was created using 574 samples across 52 BioProjects, and the DAS model was created using 804 samples across 19 BioProjects.



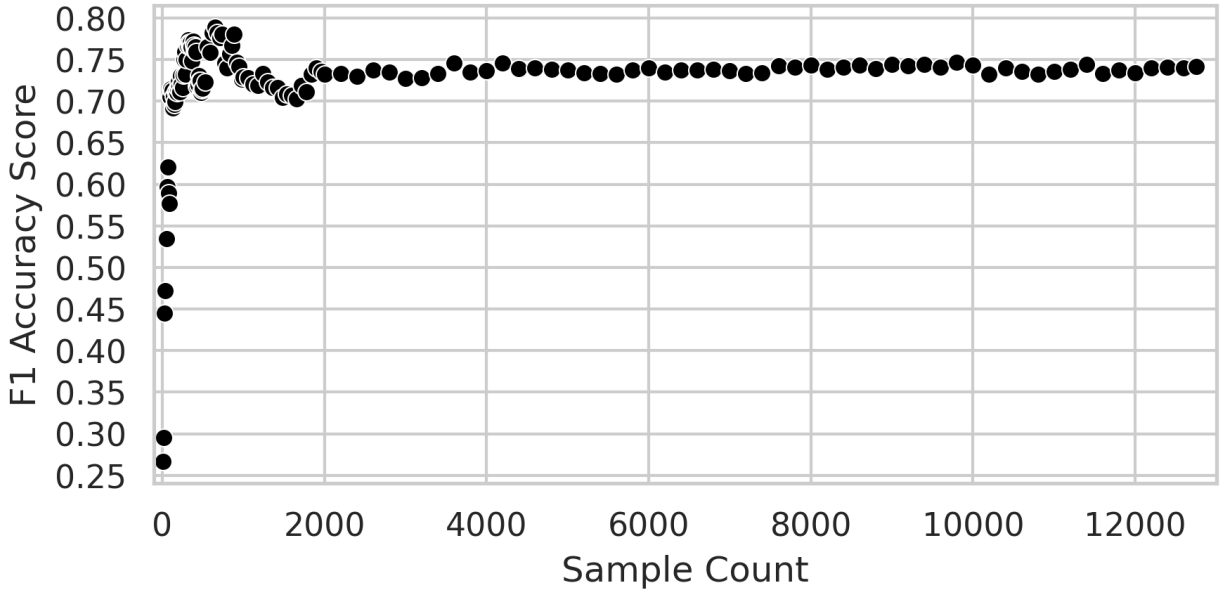
Supplemental Figure 10: DAG model (A and B) and the DAS model (C and D) performance on the DAL dataset. DAL dataset is split between the same train and test splits used for the DAL model. DAG and DAS model performance was lower than the DAL model on the DAL dataset.



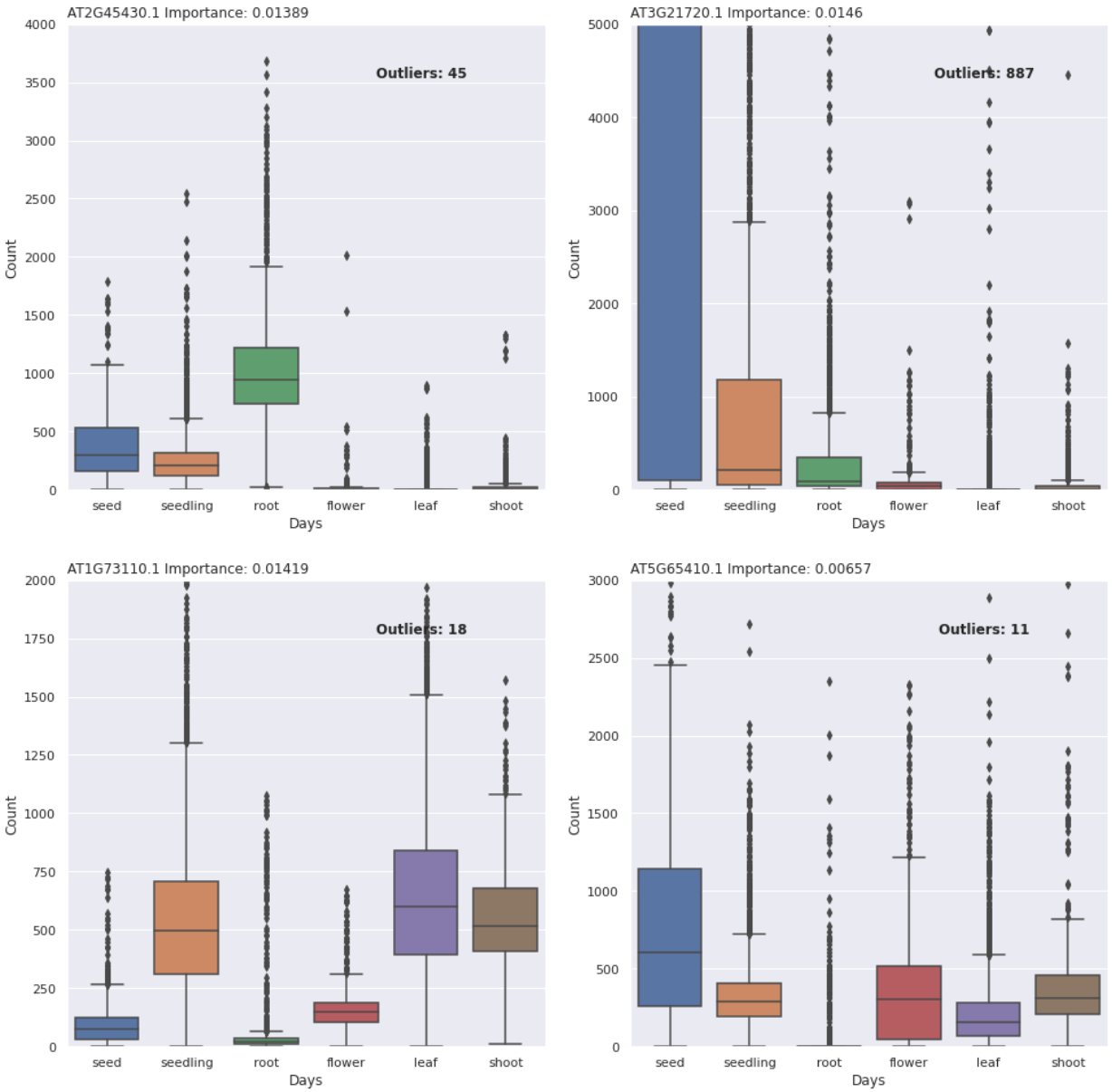
Supplemental Figure 11: SMOTE DAL dataset model performance. **A** is training performance, and **B** is testing performance.



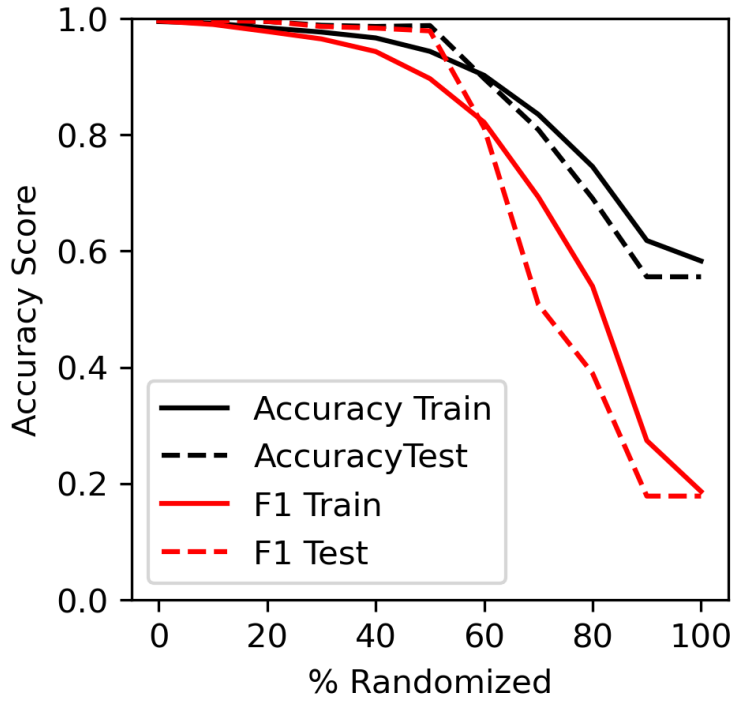
Supplemental Figure 12: Full range model performance for different sample counts. This is the same figure as **Figure 7** except the entire y-axis is shown. **A)** Model performance for tissue-4 classification dataset **B)** Model performance for DAL regression dataset.



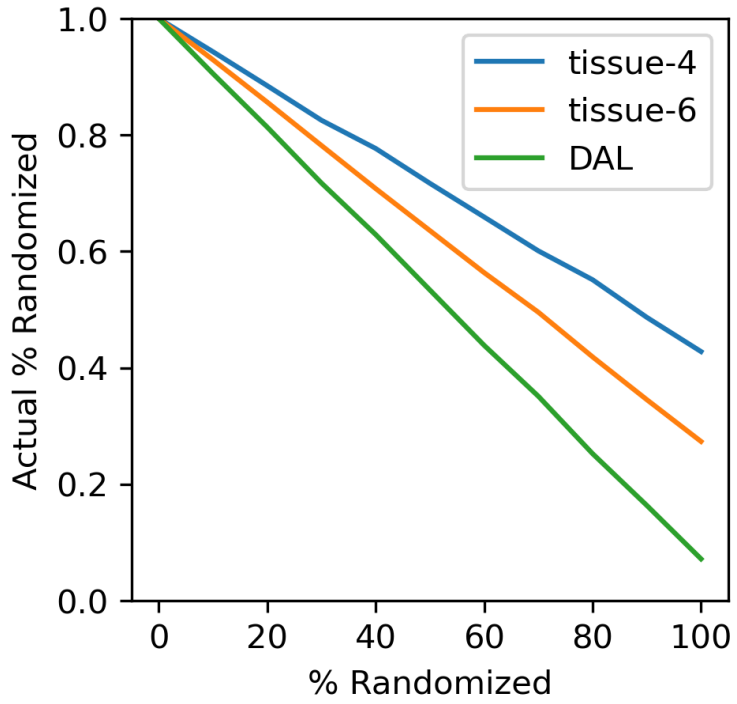
Supplemental Figure 13: Model Performance of tissue-6 dataset randomly added. This is complementary to **Figure 7** which shows tissue-4 and DAL datasets.



Supplemental Figure 14: Four of the top genes (features) from the tissue-6 dataset.



Supplemental Figure 15: Accuracy results for tissue-4 randomization



Supplemental Figure 16: Actual Random Plot for DAL, tissue-4 and tissue-6 datasets. Due to tissue-4 and tissue-6 having only 4 and 6 categories respectively, during randomization, there is still a relatively large chance that the same label is assigned. The x-axis represents what percent of the data is randomized, and the y-axis shows the

actual percent, once randomly assigning the same variable to self is taken into account. Notice that DAL is more or less a 45-degree line, whereas tissue-6 and tissue-4 have less actual percent randomized.

CHAPTER FIVE:

PHILOSOPHY OF THE SCIENCE

A Conclusion For, and Reflection Of My Ph.D. Research

The first sequenced plant genome was that of *Arabidopsis thaliana* (Arabidopsis), completed in mid-2000 (Arabidopsis Genome Initiative, 2000). This herculean accomplishment was the culmination of a decade of planning and work by dozens of researchers from labs spanning America and Germany. To mark this accomplishment, a workshop was held to discuss the achievement and prioritize future goals. This workshop, titled “Functional Genomics and the Virtual Plant” (Chory et al., 2000), resulted in the creation of a new working group that would continue research on Arabidopsis. Their mission statement was:

> “to exploit the revolution in plant genomics by understanding the function of all genes of a reference species within their cellular, organismal, and evolutionary context by the year 2010” (Chory et al., 2000)

In retrospect, this was an incredibly lofty mission statement. It still has not been completely realized in 2023; 13 years past their target date. The complexities of organisms mean that more information than genome sequences is required to fully understand function. However, I like this mission statement because it highlights something that we see happen again and again. It highlights the idea that new technology--a new way to sense, measure, and process--will solve all of our problems. By understanding the parts we believed we would be able to figure out the whole.

Sequenced genomes and sequencing technologies have undoubtedly provided us with a fantastic resource for understanding how organisms work and were a huge leap forward, but they are not the final destination.

At the time of the completion of this first plant genome sequence, I was only 7 years old, running about in my parent's backyard catching fireflies and playing with my brother. My concept of plant science was that plants have silly-sounding scientific names my father would tell us during walks through the woods and that it was possible to make maple syrup from the trees in our backyard. I knew nothing of genomes, and if I had heard of DNA it was probably only in the context that it had an interesting-looking spiral and that it is very small. I did not know that I would grow up to pursue a career in plant science, and that ultimately I would be working on the same questions being dreamt of at the advent of the plant sequencing revolution.

The majority of my Ph.D. research has been in *Malus domestica* (apple) with an emphasis on transcriptomics and postharvest biology. My research has been completely in-silico, with my friends joking that I work on "virtual apples". My research has concentrated on using sequencing resources to deduce functions and biomarkers for apples by both transferring information from model organisms and working to use new apple resources to investigate these questions from the ground up. This has been in part possible through past sequencing efforts. The first apple genome was of the cultivar 'Golden Delicious' published in 2010 through an international collaborative effort of researchers from Italy, New Zealand, Belgium, and the United States--including several research labs from Washington State University (Velasco et al., 2010). Like genomes that came before it, this was an extraordinary resource, but on its own would

not answer all of the questions we have about apple biology. My research has concentrated on using these genomic resources in conjunction with transcriptomic resources. Whereas the genome of an organism remains relatively constant over the course of an organism's life, the transcriptome changes from moment to moment. By taking many measurements over different conditions we are able to generate a dataset that captures this diversity of function. However, measuring on its own is not enough. New technologies on their own are nothing without proper data processing. We must develop and apply bioinformatic techniques to filter and process this data to its meaningful components to deduce important potential functions for further investigation. This top-down approach allows us to investigate questions of how apples behave at the molecular level. We ask these questions to learn more about basic biology--how apple molecular mechanisms differ from that of model organisms--but also in terms of applied biology through the investigation of biomarkers that can potentially be used for improving fruit end-use quality.

Contribution of my Research

The research in this dissertation investigated the use of transcriptomic resources for the identification of potential neo-functionalizations of pome fruit genes, the development of models for the identification of postharvest biosignatures, and the investigation of massive datasets to better understand outstanding questions of transcriptomic model development. The central theme has been to use the existing technology of RNA-seq to answer biologically relevant questions in non-model pome fruit systems for the elucidation of novel molecular functions as well as the betterment of the pome fruit industry.

Chapter 2 of this dissertation used a transcriptomic approach to investigate the hypoxia response of postharvest apples. Most of the research about hypoxia responses in plants has been done over a relatively short time scale in model organisms, as longer periods of time result in cell death. This contrasts dramatically with pome fruit which can be stored for up to a year in hypoxic conditions without significant degradation in quality. The research presented in this chapter has intellectual merit with respect to the plant hypoxic response. We were able to show that through genome duplication events, genes central to the hypoxia response have diverged and developed novel transcriptomic expression patterns in response to hypoxia. This directly expands our knowledge about the apple hypoxia response and provides new hypotheses for gene functions that can be verified using molecular techniques. Additionally, this research has broader impact implications for the pome fruit industry. A deeper understanding of the mechanisms behind how apples are responding to controlled atmosphere and 1-Methylcyclopropene during storage is important as the apple industry continues to modernize. We showed how the hypoxic response with and without ethylene has an impact on the apple fruit transcriptome. This better understanding of how fruit postharvest response has application for deciding on how and when to apply postharvest treatments. Additional investigation of the hypoxia response in other apple cultivars may help explain why different cultivars respond to postharvest treatments in different ways, and in the far future may assist with breeding of new apple cultivars which respond favorably to postharvest storage regimes.

Chapter 3 of this dissertation investigated using predictive transcriptomic markers for monitoring apple phenotypic traits. This has a direct broader impact on the fruit

industry, which has expressed interest in new tools for monitoring difficult-to-measure traits in packing houses. Accurate monitoring of postharvest fruit has a large economic and environmental impact, as unexpected loss in quality can result in millions of dollars of loss and dramatic food waste. This study investigated the feasibility of using predictive transcriptomic biomarkers as a way to monitor fruit in a variety of postharvest conditions. This is a preliminary study that concentrated on a relatively easy-to-measure trait (firmness) which lays the groundwork for future investigation of difficult-to-measure traits that are not readily detectable with current measurements at the time of harvest but have dramatic phenotypic impact after pro-longed storage (maturity, environmental impacts, disorders). Detection allows for pre-emptive decision-making about storage and marketing decisions about different batches of apples. In addition, this chapter investigated different predictive model techniques for transcriptomic data.

Chapter 4 of this dissertation sought to answer open questions about modeling using RNA-seq data. This ultimately arose from the investigation of transcriptomic modeling in Chapter 3 and our realization that it is important for future experimental design. While several papers have used transcriptomic data to model phenotypic traits, there has been little exploration of the required samples or the proper method of normalization for accurate transcriptomic data modeling. This has implications for considering how many samples are necessary during experimental design in transcriptomic modeling experiments, with applications going beyond the *Arabidopsis thaliana* dataset we created to perform these experiments. Additionally, we showed how normalization methods can have a significant impact on final results in massive conglomerate RNA-seq datasets. By investigating these open questions, we hope to

provide guidance for the development of transcriptomic models in other species, which will help improve predictive modeling for agricultural applications.

Appendix One of this dissertation is a tool developed in the Ficklin lab which I was the lead author and developer on by the name of GEMmaker (Hadish et al., 2022). GEMmaker is a nextflow workflow that processes RNA-seq reads to count data using several popular alignment tools. What sets it apart from other related workflows is its ability to manage and process massive amounts of publicly available RNA-seq data from NCBI on high-performance computing clusters (HPC) without overrunning available resources. This was especially necessary for Chapter 4 of this dissertation which was only possible by processing thousands of datasets using Washington State University's HPC Kamiak over several months. Additionally, GEMmaker can be easily used for processing smaller datasets as evidenced in the other chapters of this dissertation. GEMmaker is currently being used by some of my colleagues in their research, and I hope that it will continue to be a valuable tool for them.

In addition to the research presented in these chapters, I have been involved in several other projects using transcriptomic approaches to investigate pome fruit biology. First was a project investigating the molecular mechanisms of how postharvest temperature modulation reduces superficial scald in *Malus domestica* (apple) variety ‘Granny Smith’. I created co-expression networks and performed differential expression to identify genes that are activated and repressed by this treatment (L. A. Honaas et al., Jan 12-16 2019). This has a broader impact on the apple industry as producers are interested in managing postharvest disorders using organic techniques so that their crops can fetch a higher profit. Intermittent warming is a technique that is a good

substitute for diphenylamine (DPA)(which is banned in the European Union), but the molecular mechanisms of why it works were largely unknown.

Second was a project investigating the impact of canopy architecture on maturity in *Pyrus communis* (Pear) variety ‘d’Anjou’ where I developed a co-expression technique that can take into account phased edges (L. Honaas et al., 2021). This technique I developed was able to parse apart gene co-expression interactions which were different between the two canopy conditions but were not able to be detected by existing condition-specific techniques. These so-called “phased edges” involve genes that are likely involved in the differences in maturity and development seen between pear fruit on the internal vs. external portion of the canopy.

Third is a project concentrating on predicting apple maturity from transcriptomic data (Unpublished). This project has many parallels with Chapter 3 of this dissertation, with notable differences being the inclusion of four apple varieties (‘Granny Smith’, ‘WA38’, ‘Red Delicious’, and ‘Honey Crisp’). The inclusion of these four varieties meant that syncing date measurements between them was necessary for the development of predictive transcriptomic biomarkers capable of being used across all of them. I worked on developing new techniques for syncing up the biological age of these varieties based on their gene expression. Biological age contrasts with horticultural maturity, as different apple varieties are harvested at different times dependent on desired phenotypic characteristics (e.g. ‘Granny Smith’ are typically harvested early compared to other varieties). Whereas Chapter 3 of this dissertation concentrated on using an easily measured trait (firmness) for the sake of investigating the feasibility of predictive transcriptomic biomarkers in apples, this maturity project is predicting a trait that is more

difficult to measure. This is a long-term project, with additional years of data currently being collected. The preliminary analytical techniques I have developed will assist future lab members on this project as more data is collected.

Finally, I was a co-author of the paper describing the techniques behind the tool Knowledge Independent Network Construction (KINC) developed by the Ficklin lab (Burns et al., 2022). KINC is a tool that attempts to identify condition-specific transcriptomic (or other omics technology) co-expression networks through the use of Gaussian mixture models. I contributed intellectually to this project through the development of a method to address an issue where ‘spokes‘ of nodes formed around improper hub genes.

Observations and Potential Future Direction

I would like to mention a few of the difficulties associated with non-model organism research that need to be addressed going forward. The most pressing of these issues is that of consistent nomenclature. Unlike the central TAIR database for Arabidopsis (Berardini et al., 2015), the apple community does not have a central repository where gene names are standardized. Whereas Arabidopsis researchers are expected to use accepted gene name or their standard abbreviations (e.g. RELATED TO AP2 2 equates to RAP2.2), there is no such consistency in apple biology. Names of genes in apple papers are typically based on their homology to Arabidopsis genes, but this can rapidly become confusing because these genes are seldom 1:1 due to genome duplication events. Additionally, there are usually differences in how researchers identify these homologies, which can result in the same locus being called different names. This problem is further complicated by the fact there are multiple apple genomes. These

genomes do not have standardized chromosomal gene ids or a standardized method to map between them. To understand something you must know its name, and currently, it can be difficult to know which apple gene is being referenced in the literature. While there currently exists efforts for standardizing the location of genome resources (Jung et al., 2019), consistent gene naming must also be pursued so that the scientific literature can be easily compared. Until this standardization happens, researchers must report both the genome they are using, the locus id, and the annotation version in addition to any gene name they choose to use. This issue is not unique to apples, and will need to be addressed in all organisms as genome sequencing efforts increase.

For future big picture direction, as sequencing continues to decrease in price its expansion as a tool for investigating gene function and for monitoring of traits will become ubiquitous. It is easy to imagine a world a few decades from now where handheld sequencing machines are used for monitoring pome fruit (or any other crop plant) by producers, with results being available in real time. RNA-seq has been a wonderful tool for the research world, and its expansion into food production monitoring is inevitable. We are still in the beginnings of the era of big data, with the quantity of data available for the life sciences continuing to rapidly expand. However, with this continually expanding data comes the need for new ways of managing and interpreting results. Biology has entered an era of massive information, and new techniques will need to be developed to handle these data in a manner where they can be used to their full potential. This necessitates the training of biologists in bioinformatic techniques as well as developing collaborations with mathematicians and statisticians.

What I Would Tell Myself

In my junior year of my undergraduate, I told my academic advisor Dr. Dawn Reding that I was thinking of pursuing graduate school after the completion of my undergraduate degree. I had enjoyed my courses in genetics, genomics, and plant taxonomy and had fun working in the lab doing PCR reactions and other undergraduate chores. I had also recently been accepted into a summer NSF REU program at the Donald Danforth Plant Science Center (DDPSC) in St. Louis, which I was looking forward to. Plant science was exciting to me, and I believed that it would be a rewarding career. When I told Dr. Reding about my decision to pursue a Ph.D., her advice startled me. She looked me in the eyes and asked me if there was any other career that I was interested in. She said that if there was a different career that I thought I might enjoy--even just a little bit--that I should pursue that rather than a Ph.D.

If I could go back and give myself advice it would be the same advice that Dr. Reding gave me that day. From the outside, a Ph.D. seems like a magical thing, getting to pursue scientific knowledge and help humanity advance. Sure, I knew there would be setbacks, that experiments would not always go as planned, and that there would be long days of work. I knew that pay would be low, and that I would cherish free lunches provided at seminars. I knew all of these things--but I did not know what I was signing up for. Now as I conclude my Ph.D., I do not believe that it is possible to know what you are signing up for when you decide to go to graduate school. I was prepared in terms of all the technical aspects--good grades, lab experience, passion--but I was not prepared for how diving this deep into research can impact you. Becoming so invested in the minutiae of a project, to the point where it feels it has become part of you is an

exhausting experience that can result in pain if you let it. I remember that I used to think that this was a ridiculous concept. There was a scientific talk I went to when I was at the DDPSC where during questions and answers a researcher from the audience got into a fierce debate with the presenter. At the time I thought it was ridiculous and mildly amusing that two grown adults could be so invested in a molecular mechanism that they could get to a point of fury just short of a yelling match in a public forum. It is only after being in research that you can know this passion, that you can know what it means to feel like an idea belongs to you, that you have developed and cherished it, and that you have the evidence to back it up. The emotional toll that this can have is exhausting, and part of my growth as a researcher during these years has been coming to grips with how to deal with these feelings that I believe all researchers have felt at some point. A complete separation of research passion and self-identity is not possible, but maintaining a healthy relationship between them is an important skill.

With these ideas of research passion, I have also learned to appreciate the role of adversarial interactions in the advancement of research. This is something that is fundamentally intertwined with research that I did not fully appreciate coming in. The interaction between a reviewer of a manuscript and the person being reviewed is one such adversarial interaction that is fundamental to scientific research. A good reviewer is not one who only corrects a few grammar errors and gives a thumbs up, but is rather someone who takes the time to understand the research, recognize its flaws, and asks difficult questions that challenge the person being reviewed to defend their research with scientific backing. A proper review can sometimes be hard to deal with, and this process is one which many people hate. It is too easy to feel personally attacked, and

harbor remorse towards the dreaded “reviewer number 2” but this process is essential for proper science. A good review will make you uncomfortable, but it will also make you a better researcher. These adversarial relationships permeate throughout the rest of scientific research as well, being ingrained in the training of graduate students through committee defenses, through interactions with your boss over wording in a manuscript, or through the process of tenure. This trial by fire is what makes academia work, but it is important to understand how to not take it personally.

Adapting to the mentality of academia and research is what defines the graduate school experience. Undergraduate research opportunities give you a taste of how scientific research works, but it only touches the periphery. While undergraduates may have their own project, it is always under tightly controlled conditions where a mentor is checking and verifying results and ideas. It is only once you get to graduate school that you gain more freedom and the time required to truly do your own work and develop new ideas. This freedom is what transforms an undergraduate who has good technical laboratory abilities into a researcher who is able to reason and expand scientific ideas. This transformation can be exhausting, but it is what ultimately makes science work. Creating independent thinkers who can think critically, come up with new ideas and execute their plans is the purpose of graduate school training.

I say I would give myself the same advice my undergraduate advisor gave me because I know it would not have changed my decision. After asking me these questions, she informed me that she was not trying to discourage me, but was just trying to make sure that I truly wanted to pursue research. This advice was intended to be a warning, that research is difficult in ways you can not imagine, and in order to be

successful you must be passionate and persistent. This is an idea that I think can not be stated enough, as hearing it allows you to realize that this is what you really want to be doing. I know that I would still have chosen this path and that the hardships which are unknowable until you have been through them are worthwhile. Despite setbacks I have faced and a pandemic weathered there is nothing else in the world that I want to pursue besides the pursuit of knowledge. To do science is to know pain, but any other career is boring in comparison. The excitement of learning something truly new, no matter how small, is a thrill that makes it worth it. The process of graduate school is to teach you the ways of research, and how to be happy and content while engaged in the pursuit of knowledge.

REFERENCES

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796–815.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The Arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *Genesis* , *53*(8), 474–485.

Burns, J. J. R., Shealy, B. T., Greer, M. S., Hadish, J. A., McGowan, M. T., Biggs, T., Smith, M. C., Feltus, F. A., & Ficklin, S. P. (2022). Addressing noise in co-expression network construction. *Briefings in Bioinformatics*, *23*(1). https://doi.org/10.1093/bib/bbab495

Chory, J., Ecker, J. R., Briggs, S., Caboche, M., Coruzzi, G. M., Cook, D., Dangl, J., Grant, S., Guerinot, M. L., Henikoff, S., Martienssen, R., Okada, K., Raikhel, N. V., Somerville, C. R., & Weigel, D. (2000). National Science Foundation-Sponsored Workshop Report: “The 2010 Project” Functional Genomics and the Virtual Plant. A Blueprint for Understanding How Plants Are Built and How to Improve Them. *Plant Physiology*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1539254/

Hadish, J. A., Biggs, T. D., Shealy, B. T., Bender, M. R., McKnight, C. B., Wytko, C., Smith, M. C., Feltus, F. A., Honaas, L., & Ficklin, S. P. (2022). GEMmaker: process massive RNA-seq datasets on heterogeneous computational infrastructure. *BMC Bioinformatics*, *23*(1), 1–11.

Honaas, L. A., Hargarten, H., Ficklin, S. P., Hadish, J., Wafula, E. K., dePamphilis, C. W., Mattheis, J., & Rudell, D. (Jan 12-16 2019). *Co-Expression networks provide insight into postharvest fruit physiology*. International Plant and Animal Genome

Conference, San Diego, CA.

Honaas, L., Hargarten, H., Hadish, J., Ficklin, S. P., Serra, S., Musacchi, S., Wafula, E., Mattheis, J., dePamphilis, C. W., & Rudell, D. (2021). Transcriptomics of Differential Ripening in “d”Anjou’ Pear (Pyrus communis L.). *Frontiers in Plant Science*, *12*, 609684.

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

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.

APPENDIX ONE:

GEMMAKER: PROCESS MASSIVE RNA-SEQ DATASETS ON HETEROGENEOUS COMPUTATIONAL INFRASTRUCTURE

Hadish, J. A., Biggs, T. D., Shealy, B. T., Bender, M. R., McKnight, C. B., Wytko, C., Smith, M. C., Feltus, F. A., Honaas, L., & Ficklin, S. P. (2022). GEMmaker: process massive RNA-seq datasets on heterogeneous computational infrastructure. *BMC Bioinformatics*, *23*(1), 1–11.

Authors:

John A. Hadish[1], Tyler D. Biggs[2], Benjamin T. Shealy[3], M. Reed Bender[4], Coleman B. McKnight[5], Connor Wytko[6], Melissa C. Smith[3], F. Alex Feltus[4,5,7], Loren Honaas[8] and Stephen P. Ficklin[1,2,*]

[1] Molecular Plant Sciences Program, Washington State University, Pullman, WA, USA.

[2] Department of Horticulture, Washington State University, Pullman, WA, USA.

[3] Department of Electrical and Computer Engineering, Clemson University, Clemson, SC, USA.

[4] Biomedical Data Science and Informatics, Clemson University, Clemson, SC, USA.

[5] Department of Genetics and Biochemistry, Clemson University, Clemson, SC, USA.

[6] Department of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, USA.

[7] Center for Human Genetics, Clemson University, Greenwood, SC, USA.

[8] USDA Agricultural Research Service, Wenatchee, WA, USA.

* Corresponding author

https://doi.org/10.1186/s12859-022-04629-7

Attributions

JAH and **SPF** wrote the manuscript. **JAH, TDB, BTS, SPF,** and **CW** developed the GEMmaker workflow. **MCS, FAF, LH** and **SPF** obtained funding for GEMmaker and supervised the work of **JAH, BTS, TDB, CBM** and **MRB. JAH** performed testing of 475 Oryza sativa data using GEMmaker on local and HPC systems. **RMB** and **CBM** performed testing on Kubernetes systems. **SPF** and **BTS** performed testing of the 26 K Arabidopsis dataset. All authors read and approved the final manuscript.

Abstract

Background: Quantification of gene expression from RNA-seq data is a prerequisite for transcriptome analysis such as differential gene expression analysis and gene co-expression network construction. Individual RNA-seq experiments are larger and combining multiple experiments from sequence repositories can result in datasets with thousands of samples. Processing hundreds to thousands of RNA-seq data can result in challenges related to data management, access to sufficient computational resources, navigation of high-performance computing (HPC) systems, installation of required software dependencies, and reproducibility. Processing of larger and deeper RNA-seq experiments will become more common as sequencing technology matures.

Results: GEMmaker, is a nf-core compliant, Nextflow workflow, that quantifies gene expression from small to massive RNA-seq datasets. GEMmaker ensures results are highly reproducible through the use of versioned containerized software that can be executed on a single workstation, institutional compute cluster, Kubernetes platform or the cloud. GEMmaker supports popular alignment and quantification tools providing results in raw and normalized formats. GEMmaker is unique in that it can scale to process thousands of local or remote stored samples without exceeding available data storage.

Conclusions: Workflows that quantify gene expression are not new, and many already address issues of portability, reusability, and scale in terms of access to CPUs. GEM-maker provides these benefits and adds the ability to scale despite low data storage infrastructure. This allows users to process hundreds to thousands of RNA-seq samples even when data storage resources are limited. GEMmaker is freely available and fully documented with step-by-step setup and execution instructions.

Keywords: RNA-seq, Workflows, Gene expression matrix, Gene co-expression network, Differential gene expression, Nextflow

Background

Transcriptome sequencing (RNA-seq) is used in the life sciences to explore gene–gene and gene-trait relationships (Z. Wang et al., 2009). The full workflow for an RNA-seq experiment consists of several steps including experimental design, RNA collection, cDNA library construction sequencing, read cleaning, transcript mapping and gene expression quantification. Downstream computational analyses vary depending on the research goal, and can include differential gene expression (DGE) (Love et al., 2014;

Robinson et al., 2010), gene regulatory network construction (Delgado & Gómez-Vela, 2019; Mochida et al., 2018), eQTL analysis (Sun & Hu, 2013; Zhu et al., 2016), and gene co-expression network (GCN) analysis (Langfelder & Horvath, 2008; Shealy et al., 2019).

Individual RNA-seq experiment increasingly include hundreds to thousands of samples. These experiments are often made available on public repositories–such as the National Center for Biotechnology Information (NCBI) (NCBI Resource Coordinators, 2016)–allowing them to be mined for new knowledge. To prepare RNA-seq data for downstream computational analysis, expression levels must first be quantified, which is the process of converting raw RNA-seq reads to count data. Count data is stored as a gene expression matrix (GEM) which is an *n x m* matrix of n genes and m samples with values representing gene expression levels. Quantification of gene expression levels is performed using popular tools such as HISAT2 (Kim et al., 2015), Salmon (Patro et al., 2017), kallisto (Bray et al., 2016), or STAR (Dobin et al., 2013). Examples of ancillary tools include the SRAToolkit (Ncbi, 2014) for data retrieval from the NCBI SRA, Trimmomatic (Bolger et al., 2014) for contaminant and quality trimming (HISAT2/STAR workflows), SAMtools (Li et al., 2009) for storing alignments, Stringtie (Pertea et al., 2015) for read counting (HISAT2/ STAR workflow) and quality analysis reports such as FastQC (Andrews, 2010) and MultiQC (P. Ewels et al., 2016).

Several automated RNA-seq workflows have been created to ease the burden of managing the steps of RNA-seq processing. These include Pipelines in Genomics (PiGx) (Wurmus et al., 2018), Visualization Pipeline for RNA sequencing analysis (VIPER) (Cornwell et al., 2018), handy parameter-free pipeline for RNA-Seq analysis

(hppRNA) (D. Wang, 2018), Closha (Ko et al., 2018), the Transparent Reproducible and Automated PipeLINE (TRAPLINE) (Wolfien et al., 2016) and the nf-core/ rnaseq workflow (P. Ewels et al., 2019).

A popular advancement in workflow construction is the use of framework software to construct and then manage execution of the workflow. Popular examples include Galaxy (Afgan et al., 2018), Kepler (Ludäscher et al., 2006), Nextflow (Di Tommaso et al., 2017) and Snakemake (Koster & Rahmann, 2012). Workflow managers simplify workflow construction and ensure automation with reproducible results, and often provide automatic execution on a variety of computing platforms. For example, Nextflow can manage execution of workflows on desktop computers or HPC systems such as Grid Engine (Gentzsch, 2001), Portable Batch System (PBS) (Feng et al., 2007), HTCondor (Thain et al., 2005), SLURM (Jette et al., 2003), Kubernetes (VMware, 2017), popular commercial cloud platforms, and others. Nextflow also uses containers, such as Docker (Merkel, 2014) and Singularity (Koster & Rahmann, 2012) to encapsulate dependent software for the workflow, eliminating the need for installation of software and managing interdependencies. Containerization ensures that software versions are consistent, ensuring reproducible results even when the workflow is executed on different computing platforms. One benefit of workflow frameworks is when larger datasets are used, researchers are not required to rewrite a workflow when moving to a different computing platform. Additionally, workflows built with containerized software can run simultaneously on multiple platforms.

To assist bioinformaticians in the development of portable, standards-based reproducible workflows, the nf-core framework (P. A. Ewels et al., 2020) was developed

which provides workflow construction standards, peer-review and best-practice recommendations for workflows constructed using Nextflow. The nf-core provides an interactive community of developers accessible via online communication tools to assist others in development of workflows. It consists of many released workflows and a variety of others that are under construction. These include the RNA-seq workflow: nf-core/rnaseq.

Here we introduce an RNA-seq workflow named GEMmaker. Despite the existence of other workflows, it grew from the need to process 26,055 SRA runs from 17,018 SRA experiments. Unfortunately, the nf-core/rnaseq workflow was not able to scale to this large dataset as it would exhaust available storage. When thousands of RNA-seq samples are used, intermediate files can exceed available compute storage as is the case of the HISAT2 tool which can quickly consume terabytes of storage when hundreds or thousands of samples require processing. Other gene quantification tools such as Salmon (Patro et al., 2017) and kallisto (Bray et al., 2016) require less data storage but can also exhaust storage depending on the number of samples.

The inability to scale without overrunning user data storage is a limitation of Nextflow rather than the nf-core/rnaseq workflow, which could overrun user storage—especially for large datasets. There are two key factors inhibiting scaling. First, Nextflow does not currently support cleanup of intermediate files. Second, Nextflow tends to execute all instances of the same step (e.g., downloading of SRAs from NCBI) before moving to the next step (e.g., quantification with kallisto) compounding the challenge of cleanup of intermediate files since cleanup cannot occur until later steps are completed.

Until the time that Nextflow supports a file cleanup strategy, a solution is needed to support RNA-seq workflows that need to scale without overrunning storage. Ideally, the solution would be to contribute code to the nf-core/rnaseq workflow to support file cleanup, but the nf-core standards require that workflows only support native Nextflow functionality. GEMmaker, therefore, exists to provide a workflow that supports massive scaling of RNA-seq processing when storage is limited. GEMmaker v2.1 is fully nf-core compatible and can be used in the same manner as any nf-core workflow. It provides much of the functionality of the nf-core/rnaseq workflow as well as the portability and reproducibility benefits inherit with Nextflow and nf-core workflows. GEMmaker is not better than other workflows in terms of accuracy of results or improved computational time, so we do not compare it to other workflows. Rather, it is meant to process increasingly large datasets without overrunning storage using the same steps that are common in other RNA-seq workflows. The following describes the implementation of GEMmaker and provides storage performance results.

Implementation

GEMmaker uses Nextflow and is a combination of Groovy scripts for interfacing with Nextflow, Python scripts for wrangling intermediate data, and Bash scripts for execution of each software tool in the workflow. Nextflow was selected as the framework because it is widely used, is well supported, has a robust community of workflow creators in the life sciences, supports multiple computing platforms and supports containerization systems such as Docker and Singularity. Nextflow allows for execution of workflows from a command-line interface, which is common with most HPC platforms. These attributes make GEMmaker relatively easy to use. The following is an example

command-line for execution of GEMmaker on a local machine using Singularity (for containerization), quantification using Salmon, and a file containing a list of SRA run IDs for Arabidopsis thaliana Illumina datasets:

```
nextflow run systemsgenetics/gemmaker -profile singularity \
 --pipeline salmon \
 --salmon_index_path Arabidopsis_thaliana.TAIR10.salmon.indexed \
 --sras SRAs.txt
```

GEMmaker adopts the nf-core recommendations and standards to provide consistency in functionality with other popular nf-core workflows.

GEMmaker uses a variety of software tools for gene expression-level quantification and quality control that can be selected by the user. These software are listed in Table 1 and the step-by-step flow of the workflow using these tools is shown in Figure 1. There are four primary paths for gene expression quantification within GEMmaker: STAR, HISAT2, Salmon and kallisto. The STAR and HISAT2 paths include read trimming via Trimmomatic, SAMtools for storing alignments and Stringtie for quantification. Salmon and kallisto do not require those steps. All paths provide a MultiQC report to help endusers explore the quality of results from the workflow.

As mentioned previously, GEMmaker is designed to scale. It can scale to process increasingly larger experiments (or large numbers of samples from public repositories) that can include hundreds to thousands of RNA-seq samples without intermediate files overrunning available compute storage. It supports execution on a large variety of computational platforms such that researchers can take full advantage of the compute

facilities available to them including local desktop workstations, institutional clusters, national-funded resources such as XSEDE (Towns et al., 2014), the Pacific Research Platform (Smarr et al., 2018), and commercial clouds.

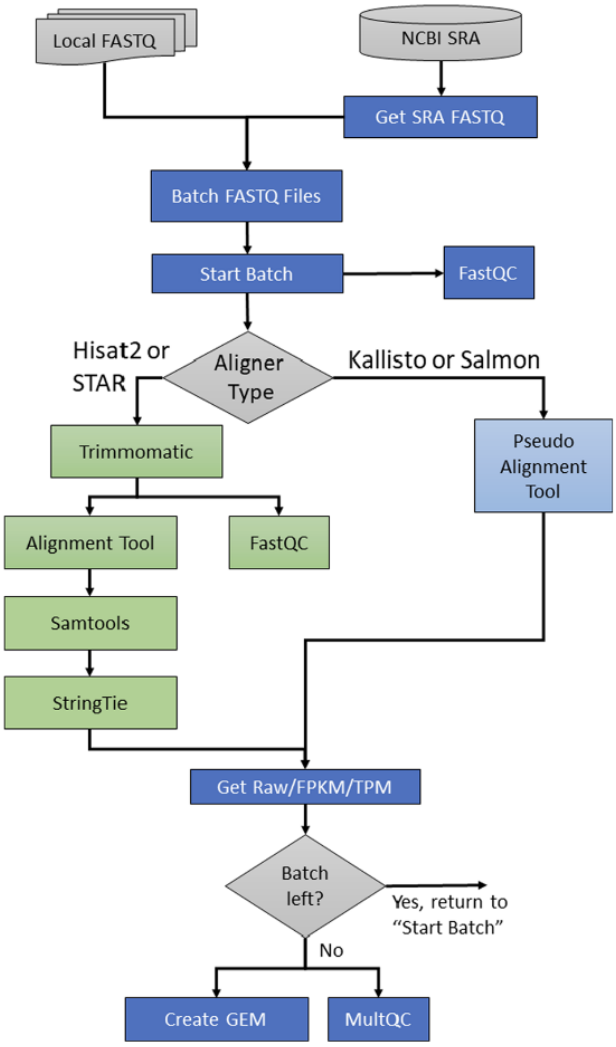


Figure 1 GEMmaker workflow diagram. GEMmaker supports the inclusion of both local and remote RNA-seq data files and offers four different alignment tools for gene expression quantification: Hisat2, STAR, Kallisto, and Salmon

To ensure storage requirements are not exceeded, GEMmaker moves input FASTQ files between three folders: “stage”, “processing” and “done”. Initially all samples are

placed in the “stage” folder and GEMmaker will move into the “processing” folder as many samples as there are CPUs available. The user sets the number of CPUs that the workflow can use with the –max_cpus argument. On a compute cluster, this could be tens to hundreds. Nextflow is then instructed to automatically begin processing any samples that appear in the “processing” folder. As usual, Nextflow will process samples in parallel, using all CPUs, by first executing the first step for all samples, then the second for all samples, and so forth. However, because GEMmaker limits the number of samples to the number of CPUs, when a sample completes a step, it will move to the next step because Nextflow does not see any samples waiting. When a sample fully completes all steps, GEMmaker will then move the sample from the “processing” folder into the “done” folder and will move one sample from the “stage” folder into the “processing” folder. Nextflow sees this new sample in the “processing” folder and immediately begins processing that sample through each step. There is no lag between the time one sample finishes, and another begins and Nextflow should keep all CPUs consistently busy processing samples in parallel. As the workflow progresses for each sample, GEMmaker will cleanup unwanted intermediate files. This ensures space is cleaned before more samples begin processing. If the user specifies a –max_cpu size that does not exceed the resources of the computational platform, then GEMmaker can successfully process hundreds to thousands of samples.

While GEMmaker, by default, cleans all intermediate files, there are arguments that can be provided, as described in the online documentation, to control which intermediate files are removed. Users can keep downloaded SRA and FASTQ files, trimmed FASTQ files, SAM and BAM alignment files, and kallisto and Salmon

pseudoalignment files. If any of these files are needed for downstream analyses they can be retained.

The speed at which the samples are processed depends on the number of processors and available memory of the compute nodes. Users with limited CPUs or RAM may need more time to process all samples. If users set the –max_cpus setting higher than storage will support, then GEMmaker may not be able to cleanup intermediate files before overrunning storage. It is difficult to recommend a value which maximizes the trade-off between the number of CPUs and storage requirements because RNA-seq samples and genomic reference sequences can be dramatically different in size, resulting in different sized intermediate files. However, using averaged values from the sample data reported here, we provide a rough recommendation that users have about 30 times the storage of an average sample size, times the number of CPUs when using HISAT2. For an average sample size of 2.5GBs this would require 75 GB per CPU. For kallisto and Salmon we recommend 7 times the storage of an average sample per CPU (17GBs).

To ensure portability between HPC systems, GEMmaker makes use of containerized software. This alleviate the burden of installing the same software versions on every computational system on which it is run. All GEMmaker dependent software are provided in the GEMmaker docker image and their versions are listed in Table 1. GEMmaker retrieves this Docker image from Docker Hub the first time it is run—users need not install any software other than Nextflow and a containerization software (Singularity or Docker). Thus, a GEMmaker workflow can be performed on any computational system and results will be reproducible and consistent.

Results

We tested GEMmaker on WSU's Kamiak cluster which uses the SLURM scheduler (Jette et al., 2003), Clemson University's Palmetto cluster which uses the PBS scheduler (Feng et al., 2007), the Rodeo Kubernetes cluster at the Texas Advanced Computing Center (TACC) which contains homogenous set of compute nodes, and the Pacific Research Platform's Nautilus cluster which contains a heterogenous set of compute nodes. In all platforms GEMmaker successfully completed. Because data storage usage is of most importance, GEMmaker was tested using two different datasets: a publicly available 475-sample Oryza sativa (rice) RNA-seq dataset (NCBI SRA accession PRJNA301554) (Wilkins et al., 2016), and the Arabidopsis thaliana 26,055-runs from NCBI.

The 475 rice dataset consists of samples from two subspecies of rice, subdivided into 4 genotypes, grown in a hydroponic environment that underwent treatments of heat stress, drought stress and control. Measurements were taken every 15 min for several hours with 2 replicates. We selected this dataset to demonstrate execution of a large single experiment on a typical stand-alone workstation that researchers may have available to them. The Arabidopsis 26,055 dataset was selected using all Illumina RNA-seq datasets available at the time the list was collected. An SRA experiment can contain multiple runs which resulted in 17,018 SRA experiments. This included both paired and non-paired RNA-seq runs for Arabidopsis thaliana sequenced using the Illumina platform. The list of SRA run IDs is provided as Additional file 1: Data 1. We selected all RNA-seq data to test massive scale processing on a typical institutional HPC cluster. The 475 rice dataset was tested on Washington State University's HPC

cluster, Kamiak. To simulate execution on a stand-alone workstation, the job was limited to 16 CPUs and 6 GB of RAM (a reasonable set of resources for a performant workstation). The compute node contained Intel(R) Xeon(R) Gold 6138 CPU @ 2.00 GHz processors, had 256 GB of RAM (although, only 6 GB were requested) with access to 650 TB of network attached storage to allow for as much expansion of storage as needed (although, this large storge size is not required as shown in Figure 2). GEMmaker was executed twice for each quantification tool (STAR, HISAT2, kallisto and Salmon) once with cleanup of intermediate files turned on and again turned off. Because the primary performance metric of concern is storage usage, a monitoring script tracked the storage space consumed. Results of the test are found in Figure 2. With the option to clean intermediate files enabled, all the quantification tools consumed less than 1 Tb of storage. At maximum, HISAT2 consumed 680 GB, kallisto 322 GB, Salmon 342 GB, and STAR 701 GB. When intermediate files were not cleaned, both Salmon and kallisto consumed approximately 12 TB of storage, HISAT2 38 TB and Star 41 TB. Salmon and kallisto took less time (~ 3 days) than STAR (4 days), or HISAT2 (~ 5.5 days) to run. Compute time is strongly dependent on each computer's hardware and the queue size. Therefore, this test could have run quicker if the number of CPUs were increased. The range of storage space (between 322 and 680 GB) required to execute GEMmaker on this set of 475 samples, with intermediate file cleaning enabled, is commonly available on stand-alone workstations.

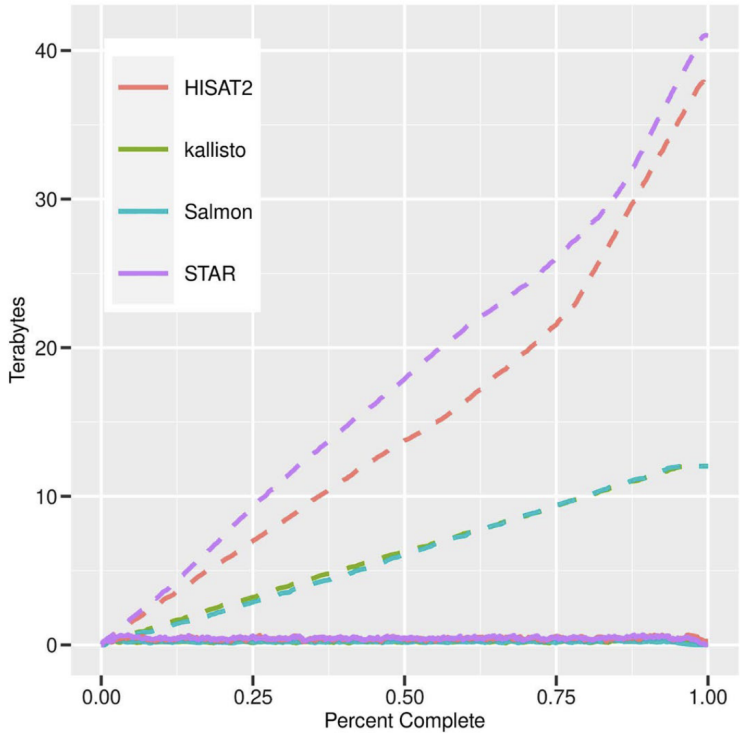


Figure 2 Storage usage comparison. Storage sizes for processing the 475-sample time-series rice dataset is shown. Dashed lines indicate tests in which GEMmaker was configured to not cleanup of intermediate files between batches, while solid lines indicate that a cleanup was performed.

To demonstrate processing of tens of thousands of RNA-seq datasets, the 26 K SRA runs were processed on WSU's Kamiak HPC cluster with a –max_cpus setting of 120 (i.e., 120 currently running jobs in parallel). We used the kallisto pipeline, and GEMmaker completed processing the 26 K runs over 28 days. We designed GEMmaker so that if a dataset is corrupted, or if information was incorrectly entered into NCBI that it would report these and then continue with other samples. This reduces downtime and allows the user to look at these files manually. GEMmaker reported that of the 26 K runs, 19 SRA files had no metadata available via NCBI web services and could not be

retrieved; 179 had missing download URLs; 3 samples were corrupted after download; and 1 failed to download due to a network timeout. Just as with the rice data, GEMmaker was instructed to clean intermediate files (SRA files, FASTQ files, kallisto index files, etc.) and keep only raw and TPM count files, but actual storage usage was not measured during runtime. The results folder consumed 48 GB of storage.

Limitations

Despite the advantages that GEMmaker affords, it has limitations. First, we could not include every quantification tool made to date; users who need other tools are encouraged to request features on the GEMmaker GitHub issue queue. Second, if GEMmaker is preempted before it completes, as was the case with the 26 K Arabidopsis dataset, then there may be working directories that do not get cleaned. Because GEMmaker is a Nextflow workflow, it can resume execution where it left off. However, Nextflow creates new working directories for each step of the workflow for each sample and when it is resumed it creates new working folders—the folders with failed steps remain. When a sample completes a step, then GEMmaker can clean up the working directories that were successful but there is not a mechanism in Nextflow to know about the directories with failed results so that they can be cleaned. As a result, if a high –max_cpus is used (e.g., 120) and Nextflow is preempted this may result in higher storage usage from directories with failed jobs. Third, related to usability, GEMmaker does not have a graphical user interface (GUI). Users familiar with the UNIX command line will not see this as an issue, but those who have limited experience may find this difficult. Finally, GEMmaker was not designed for data security. Users with

sensitive data will need to coordinate with data security experts to ensure processing is executed in a secure facility.

Conclusion

GEMmaker addresses issues of scale for processing massive RNA-seq experiments with hundreds to thousands of samples (although it can be used for small datasets as well). While automated RNA-seq workflows already exist, GEMmaker is unique in that it does not overrun data storage facilities yet provides similar functionality to that of gold-standard RNA-seq workflows. GEMmaker allows researchers to take advantage of existing smaller computing infrastructure which can be beneficial if there is limited access to larger facilities. GEMmaker returns count data in various formats (e.g., raw and normalized) so that results can be used in downstream transcriptome analyses such as differential gene expression, regulatory network construction and gene co-expression analysis.

Availability and requirements

Project name: GEMmaker

Project home page: https://github.com/ SystemsGenetics/GEMma ker

Operating systems(s): Platform independent

Programming language: Nextflow Groovy, Python and bash

Other requirements: Nextflow and Java. Docker or singularity are optional but suggested Any restrictions to use by non-academics: GPL v2.0 license.

REFERENCES

Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Ech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537–W544.

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

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

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527.

Cornwell, M., Vangala, M., Taing, L., Herbert, Z., Köster, J., Li, B., Sun, H., Li, T., Zhang, J., Qiu, X., Pun, M., Jeselsohn, R., Brown, M., Shirley Liu, X., & Long, H. W. (2018). VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. *BMC Bioinformatics*, *19*. https://doi.org/10.1186/s12859-018-2139-9

Delgado, F. M., & Gómez-Vela, F. (2019). Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artificial Intelligence in Medicine*, *95*, 133–145.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature*

Biotechnology, *35*(4), 316–319.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21.

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*(3), 276–278.

Ewels, P., Hammarén, R., Peltzer, A., Moreno, D., Garcia, M., Rfenouil, Colin, M., Panneerselvam, S., F, Sven, Jun-Wan, Alneberg, J., Aanil, Haglund, S., Di Tommaso, P., Jemt, A., Kochtobi, & Veeravalli, L. (2019). *nf-core/rnaseq*. https://doi.org/10.5281/zenodo.1400710

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* , *32*(19), 3047–3048.

Feng, H., Misra, V., Rubenstein, D., Feng, H., Misra, V., & Rubenstein, D. (2007). PBS: a unified priority-based scheduler. *Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems - SIGMETRICS '07*, *35*, 203.

Gentzsch, W. (2001). Sun Grid Engine: Towards creating a compute power grid. *Proceedings - 1st IEEE/ACM International Symposium on Cluster Computing and the Grid, CCGrid 2001*, 35–36.

Jette, M. A., Yoo, A. B., & Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management. *Job Scheduling Strategies for Parallel Processing, Lecture*

Notes in Computer Science, *2862*, 44–60.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360.

Ko, G., Kim, P.-G., Yoon, J., Han, G., Park, S.-J., Song, W., & Lee, B. (2018). Closha: bioinformatics workflow system for the analysis of massive sequencing data. *BMC Bioinformatics*, *19*(S1), 43.

Koster, J., & Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* , *28*(19), 2520–2522.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , *25*(16), 2078–2079.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E., Tao, J., & Zhao, Y. (2006). Scientific workflow management and the Kepler system: Research Articles. *Concurrency and Computation: Practice & Experience*, *18*(10), 1039–1065.

Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*, *2014*(239), 2.

Mochida, K., Koda, S., Inoue, K., & Nishii, R. (2018). Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets.

Frontiers in Plant Science, *871*(November), 1–7.

Ncbi. (2014). *SRA Handbook [Internet] - Aspera Transfer Guide*. https://www.ncbi.nlm.nih.gov/books/NBK242625/

NCBI Resource Coordinators, N. R. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *44*(D1), D7–D19.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , *26*(1), 139–140.

Shealy, B. T., Burns, J. J. R., Smith, M. C., Alex Feltus, F., & Ficklin, S. P. (2019). GPU Implementation of Pairwise Gaussian Mixture Models for Multi-Modal Gene Co-Expression Networks. *IEEE Access*, *7*, 160845–160857.

Smarr, L., Crittenden, C., DeFanti, T., Graham, J., Mishin, D., Moore, R., Papadopoulos, P., & Würthwein, F. (2018). *The Pacific Research Platform*. 1–8.

Sun, W., & Hu, Y. (2013). eQTL Mapping Using RNA-seq Data. *Statistics in Biosciences*, *5*(1), 198–219.

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: the Condor experience: Research Articles. *Concurrency and Computation: Practice &*

Experience, *17*(2-4), 323–356.

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G. D., Roskies, R., Scott, J. R., & Wilkens-Diehr, N. (2014). XSEDE: Accelerating scientific discovery. *Computing in Science and Engineering*, *16*(5), 62–74.

VMware. (2017). *DEMYSTIFYING KUBERNETES Overcoming Misconceptions About Container Orchestration*.

Wang, D. (2018). hppRNA-a Snakemake-based handy parameter-free pipeline for RNA-Seq analysis of numerous samples. *Briefings in Bioinformatics*, *19*(4), 622–626.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63.

Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M.-M. M., Pham, G. M., Nicotra, A. B., Gregorio, G. B., Krishna Jagadish, S. V., Septiningsih, E. M., Bonneau, R., Purugganan, M., Plessis, A., Gregorio, G. B., Purugganan, M., Pham, G. M., Jagadish, S. V. K., Nicotra, A. B., Septiningsih, E. M., Hafemeister, C., & Wilkins, O. (2016). EGRINs (Environmental gene regulatory influence networks) in rice that function in the response to water deficit, high temperature, and agricultural environments. *The Plant Cell*, *28*(10), 2365–2384.

Wolfien, M., Rimmbach, C., Schmitz, U., Jung, J. J., Krebs, S., Steinhoff, G., David, R., & Wolkenhauer, O. (2016). TRAPLINE: A standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. *BMC Bioinformatics*, *17*(1), 1–11.

Wurmus, R., Uyar, B., Osberg, B., Franke, V., Gosdschan, A., Wreczycka, K., Ronen, J., & Akalin, A. (2018). PiGx: reproducible genomics analysis pipelines with GNU Guix. *GigaScience*, *7*(12). https://doi.org/10.1093/gigascience/giy123

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., & Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487.