LINKING GENES AND PHYSIOLOGY UTILIZING GENOMICS AND TRANSCRIPTOMICS APPROACHES IN SWEET CHERRY (*Prunus avium*)

By

TYSON ANDREW KOEPKE

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

Doctor of Philosophy

WASHINGTON STATE UNIVERSITY
Molecular Plant Sciences Graduate Program

DECEMBER 2012

© Copyright by TYSON ANDREW KOEPKE, 2012 All Rights Reserved
To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of TYSON ANDREW KOEPKE find it satisfactory and recommend that it be accepted.

_____________________________________
Amit Dhingra, Ph. D., Chair

_____________________________________
Matthew Whiting, Ph. D.

_____________________________________
Anantharaman Kalyanaraman, Ph. D.

_____________________________________
Nnadozie Oraguzie, Ph. D.

_____________________________________
Andrew McCubbin, Ph. D.
ACKNOWLEDGMENTS

I would like to thank Dr. Amit Dhingra for the opportunity to work in his lab and the opportunities he has provided for me to grow as a scientist. I would also like to extend my gratitude to my Ph. D. committee: Dr. Matthew Whiting, Dr. Anantharaman Kalyanarman, Dr. Nnadozie Oraguzie and Dr. Andrew McCubbin. Their advice and guidance has been essential for me to progress and further my understanding of their respective fields. I also appreciate the support of WSU faculty who’ve supported me including Dr. David Kramer, Dr. Diter von Wettstein, and Dr. Kim Kidwell.

I appreciate the help and support of the other members of the Dhingra Lab. These include undergraduates: Matt Allan, Christina Duncan, Jennifer Hartwig, Steven Hartwig, and Brittany Urso who helped with many facets of different projects and were great students to help mentor. The other graduate students in the Dhingra lab, especially Scott Schaeffer, Christopher Hendrickson, Rick Sharpe, and Dr. Derick Jiwan have been extremely valuable for their input, assistance and support. All of the many others whom I have not mentioned directly have also been very valuable to my growth.

To my many co-authors on the various manuscripts, thank you for the opportunity to collaborate with and for all of the hard work that has gone into the publications presented here and the many others soon to follow.

I also want to thank the growers who have so graciously shared their knowledge with me and given me the opportunity to see and understand the orchard systems and problems that are critical to their operations.

I want to acknowledge the Achievement Rewards for Collegiate Scholars (ARCS) and the NIH Protein biotech for their fellowships. WSU, the Washington Tree Fruit Research Commission, and USDA SCRI funds have also enabled the production of the research presented in this dissertation.

Last, but not least, I want to thank my entire family and all of my friends for supporting me through this Ph. D. The past several years have included some rough times, but knowing you have all had my back has allowed me to persevere. Thank you!
Sweet cherry is an important horticultural crop in Washington. Sweet cherry trees are grown as composite plants where the scion is grafted onto a rootstock. Rootstock genotype influences tree yield potential (among other traits) largely by affecting floral bud numbers. A transcriptomics approach (differential display and 3’UTR sequencing) was used to examine the genetic mechanism of the rootstock’s effect on floral bud initiation.

The first section of this dissertation outlines the development of key genomic resources for sweet cherry. Utilizing both reference mapping and de novo genome assembly for the sweet cherry variety ‘Stella’, 1.5 million single nucleotide polymorphisms were identified and 180Mb of genome were assembled respectively. The identified polymorphisms were screened and roughly 300,000 are located in peach genes and could be used as gene-based markers for future research. From this set, several genes involved in fruit ripening were found to contain nonsense mutations and could be the underlying cause for non-climacteric ripening in sweet cherry.

Examining gene expression in developing floral buds of the scion, rootstock effects were identified through differential display and 3’UTR analyses. Differential display identified 207 transcripts that were putatively controlled by the rootstock genotype while 3’UTR identified 115 differentially expressed gene fragments. Several transcription factors including some with domains previously documented to control flowering were identified in these datasets and their expression levels were examined via qRT-PCR for verification.
These studies have created vast amounts of information from both the sweet cherry genome and transcriptome. Many single nucleotide polymorphisms were identified and leveraged as DNA markers for screening sweet cherry populations. This demonstrated the rapid application of a small subset to aid sweet cherry breeding and population screening. Additionally, a computational program was developed to assess restriction sites in large datasets and used to convert the polymorphism information into cleaved amplified polymorphic sequence markers.

Overall, the work described in this dissertation has laid the foundation for sweet cherry genomics research and identified several directions for future studies.
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>xii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1. OVERVIEW</td>
<td>1</td>
</tr>
<tr>
<td>2. BACKGROUND</td>
<td></td>
</tr>
<tr>
<td>A. Invited Review: Rootstock Scion Somatogenetic Interactions in Perennial Composite Plants</td>
<td>10</td>
</tr>
<tr>
<td>i. Abstract</td>
<td>10</td>
</tr>
<tr>
<td>ii. Introduction</td>
<td>11</td>
</tr>
<tr>
<td>iii. Physiological Traits</td>
<td>12</td>
</tr>
<tr>
<td>iv. Physical Mechanisms</td>
<td>21</td>
</tr>
<tr>
<td>v. Molecular Mechanisms</td>
<td>31</td>
</tr>
<tr>
<td>vi. Conclusions and Future Prospects</td>
<td>36</td>
</tr>
<tr>
<td>B. Impact of Advances in Sequencing Technology</td>
<td>50</td>
</tr>
<tr>
<td>3. GENOMICS</td>
<td></td>
</tr>
<tr>
<td>A. Comparative Genomics of Prunoideae</td>
<td>59</td>
</tr>
<tr>
<td>i. Abstract</td>
<td>59</td>
</tr>
<tr>
<td>ii. Introduction</td>
<td>60</td>
</tr>
<tr>
<td>iii. Results</td>
<td>61</td>
</tr>
<tr>
<td>iv. Discussion</td>
<td>66</td>
</tr>
<tr>
<td>v. Conclusions</td>
<td>70</td>
</tr>
</tbody>
</table>

vi
vi. Methods

B. RADshredder: A tool enabling RAD data utilization in de novo genome assembly

   i. Abstract
   ii. Background
   iii. Methods
   iv. Results and Discussion
   v. Conclusions

4. TRANSCRIPTOMICS

   A. Examining rootstock scion interactions of sweet cherry via transcriptome profiling

      i. Abstract
      ii. Introduction
      iii. Methods
      iv. Results and Discussion
      v. Conclusions

   B. Rapid gene-based SNP and haplotype marker development in non-model eukaryotes using 3’UTR sequencing

      i. Abstract
      ii. Background
      iii. Methods
      iv. Results and Discussion
      v. Conclusions

   C. CisSERS: Customizable in silico Sequence Evaluation for Restriction Sites

      i. Abstract
ii. Introduction 173

iii. Materials and Methods 175

iv. Results 179

v. Discussion 182

vi. Conclusions 183

5. CONCLUSIONS and FUTURE WORK 198
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2A.1</td>
<td>Recent representative reviews covering different aspects of rootstock scion interaction</td>
<td>45</td>
</tr>
<tr>
<td>2A.2</td>
<td>Physiological traits affected by rootstocks summarized in this review</td>
<td>46</td>
</tr>
<tr>
<td>2A.3</td>
<td>Physical mechanisms proposed to be underlying rootstock scion interactions</td>
<td>47</td>
</tr>
<tr>
<td>2A.4</td>
<td>Molecular analyses performed to unravel rootstock scion interactions</td>
<td>48</td>
</tr>
<tr>
<td>3A.1</td>
<td>Raw sequencing data</td>
<td>77</td>
</tr>
<tr>
<td>3A.2</td>
<td>Nonsense mutation analysis statistics</td>
<td>78</td>
</tr>
<tr>
<td>3A.3</td>
<td>Classifications of polymorphisms</td>
<td>79</td>
</tr>
<tr>
<td>3A.4</td>
<td>Classification of polymorphisms in coding regions</td>
<td>80</td>
</tr>
<tr>
<td>3A.5</td>
<td>Sk locus analysis</td>
<td>81</td>
</tr>
<tr>
<td>3A.6</td>
<td>Number of genes containing mutation for each EC number</td>
<td>82</td>
</tr>
<tr>
<td>3A.7</td>
<td>Number of genes for each sample containing nonsense SNPs for each EC number</td>
<td>83</td>
</tr>
<tr>
<td>4A.1</td>
<td>Comparison of transcriptome profiling approaches</td>
<td>130</td>
</tr>
<tr>
<td>4A.2</td>
<td>A summary of the differential display results</td>
<td>131</td>
</tr>
<tr>
<td>4A.3</td>
<td>Blast summary for differential display fragments</td>
<td>132</td>
</tr>
<tr>
<td>4A.4</td>
<td>cDNA sequencing data using 454 GS Titanium chemistry</td>
<td>133</td>
</tr>
<tr>
<td>4A.5</td>
<td>3’UTR sequencing results</td>
<td>134</td>
</tr>
<tr>
<td>4A.6</td>
<td>Seeded PaCE Output Summary</td>
<td>135</td>
</tr>
<tr>
<td>4B.1</td>
<td>Summary of 3’UTR sequencing results</td>
<td>161</td>
</tr>
<tr>
<td>4B.2</td>
<td>Experimental assessment of SNPs</td>
<td>162</td>
</tr>
<tr>
<td>4B.3</td>
<td>Transcriptome assembly results</td>
<td>163</td>
</tr>
<tr>
<td>4B.4</td>
<td>Summary of transitions and transversions across Bing and Rainier</td>
<td>164</td>
</tr>
<tr>
<td>4C.1</td>
<td>Restriction Analysis Program Comparison</td>
<td>187</td>
</tr>
<tr>
<td>4C.2</td>
<td>CisSERS computer run time comparisons</td>
<td>188</td>
</tr>
<tr>
<td>4C.3</td>
<td>Calculated Time to completion and Actual Time to Completion values</td>
<td>189</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

1.1 Sweet cherry bearing acres 7

1.2 Sweet cherry utilized production and value 8

2A.1 Overview of review topics 49

3A.1 GO-term composition of nonsense SNP-containing datasets by molecular function 84

3A.2 Venn diagram of nonsense SNPs 85

3A.3 KEGG pathway of genes containing nonsense SNPs 86

3A.4 KEGG pathway of genes containing nonsense SNPs 87

3A.5 Venn diagram of genes containing nonsense mutations in almond 88

3B.1 Example RAD data at a single locus of a genome 104

3B.2 Expected comparison of reads and RADshredder pseudo-reads 105

3B.3 Depiction of RADsplitter input and results 107

3B.4 Actual depth reduction image 108

3B.5 NGen Assembly Results 109

3B.6 Contig length analysis of NGen assembly N2-454 and N4-454 110

3B.7 PstI analysis of N2-454 contigs 111

3B.8 CLC Genomics Workbench Assembly Results 112

3B.9 CLC Genomics Workbench Assembly Results 113

4A.1 Illustration representing three distinct physiological stages of sweet cherry growth 136

4A.2 Mature floral buds on the four sample combinations 137

4A.3 Differential display gel from two primer combinations 138

4A.4 Overlapping circle diagram of the differential display results 139

4A.5 Overlapping circle diagram of the significant 3’UTR results 140

4A.6 Bar chart of the qRT-PCR $2^{\Delta\Delta CT}$ values 141

4B.1 General schema for rapid identification of SNPs 165
4B.2 Analysis of variation of identified SNPs via high resolution melting (HRM) curves 166
4B.3 Pedigree relationships of the 8 cultivars used in this study 167
4B.4 Four primer sets with the HRM curves 168
4B.5 Screenshot of the SeqMan (DNASTAR) visualization of contig456 169

4C.1 Experimental flow chart 190
4C.2 Screenshot of the CisSERS cut site selection process 191
4C.3 Screenshot of the CisSERS processing tab 192
4C.4 Performance tests across two different computer platforms 193
4C.5 Time to completion of a 3’ restriction site analysis 194
4C.6 Gel visualization functionality test 195
4C.7 CisSERS predicted gel image vs. wet-lab gel visualization test 1 196
4C.8 CisSERS predicted gel image vs. wet-lab gel visualization test 2 197
Dedication

This dissertation is dedicated first and foremost to my family: my parents Mike and Sheila, Brother Drew, Sister Ashley, grandparents Eleanor, Don, Rosalie, Gary, Ella Mae, and John, all of my Aunts, Uncles and Cousins, especially William Gary Levis. It is with your love and support that this dissertation has become a reality and I can continue to pursue my dreams. Thank you,

I also wish to dedicate this dissertation to Dr. Virginia G. Lee (April 17, 1986 - December 31, 2010) and her family. Virginia impacted my life so deeply I simply cannot express it in words. May this dissertation be a beginning of my part of helping to fulfill her dream to ‘Change the World’ and make it a better place for all.
CHAPTER 1:
OVERVIEW

Sweet cherry, *Prunus avium* L., is an important fruit crop in plant family Rosaceae. Other members of this economically important family are apples, pears, peaches, almonds, and strawberries. Together, the Rosaceae family accounts for $9-10 billion worth of revenues annually over the past three years in the United States (NASS 2012). Sweet cherries account for $0.5-0.8 billion of this value with a majority of the production, $0.2-0.5 billion, being produced in the state of Washington. In the past 13 years, production acreage of sweet cherries in the US has increased by 30,000 acres, representing more than 50% increase as shown in Figure 1. Expectedly, utilized production has also increased by roughly the same percentage from an average of ~220,000 tons per year from 1997-1999 to ~340,000 tons per year average in 2009-2011 (Figure 2). Interestingly, the value of sweet cherry production over the same period has increased more than 2.7 fold. This increase could be a result of the improvements to human health attributed to sweet cherry consumption. Sweet cherry consumption has potential preventative benefits to diseases such as Alzheimer’s, cancer, and inflammation related diseases (McCune et al. 2010). Due to the economic importance of sweet cherry and its potential to benefit human health, it is vital to research and improve this crop to maintain global competitiveness for the growers in Washington.

Sweet cherry production in the Pacific Northwest varies significantly with many varieties being produced throughout the growing season and across the region. The cherry fruits vary in quality attributes such as color, flavor, and most importantly size. Orchards range from a few acres to several thousand acres in size. Some orchards are based on large, older trees while others are pedestrian orchards with trees less than ten years old. Trees are grown in various architectures from open vase and central leader to Kym Green Bush (KGB) and trellised systems like the double axis, V-trellis, and Upright Fruit Offshoots (UFO) systems. Each production system has optimal row and tree spacing and pruning requirements to facilitate optimal growth and yield. The production system used for an orchard directly affects the potential value of the orchard with higher density plantings being more productive (Whiting et al. 2005b).
While there are considerable differences in production practices, there are several critical aspects of sweet cherry production that remain fairly consistent. Sweet cherry scions are typically self-incompatible and do not produce true-to-type seed necessitating clonal production through propagation. In addition to self-incompatibility, sweet cherry has an extended juvenility period, sometimes greater than four years, before the trees flower. This delay limits breeding efforts as many years are required after a cross to determine the outcome and cross those offspring. These issues have also delayed progress in understanding the genetics of sweet cherry.

Due to the inability of conventional breeding to rapidly improve all characteristics of sweet cherry, rootstocks have been used to overcome soil-borne diseases and plant vigor instead of traditional breeding approaches used in many other crop species. Grafting can affect many aspects of scion growth and physiology which along with potential mechanisms controlling these changes are discussed in detail in Chapter 2A (Page 10). One of the major traits controlled by the rootstock in sweet cherry is tree size, or vigor. Through the use of dwarfing rootstocks such as Gisela® 5 and 6, trees can be planted closer together promoting high density plantings. As mentioned above, these high density orchards are becoming more popular as they tend to be more productive.

An important trait that needs to be controlled in sweet cherry production is the leaf area per fruit (Whiting and Lang 2004). As trees are planted closer together and on dwarfing rootstocks, the number of fruit per tree also needs to be reduced to maintain effective leaf area to fruit ratios. However, since there are more trees per acre, the overall yields can still be increased with proper management. Sweet cherry value as a crop is determined primarily by fruit size and as fruit numbers increase, the fruit size decreases. This creates an emphasis on crop load management and optimization. The goal of this management is to obtain a maximal number of large fruit while too few really large fruit and too many small fruit are each damaging to the value of the crop. To manage crop load, producers either use pruning, hand, chemical, or mechanical thinning which add costs to the production system. Additionally, dwarfing rootstocks tend to have higher flower numbers with yield being affected up to ten-fold (Whiting personal communication). Understanding how the rootstock is controlling flowering, specifically floral bud initiation, in sweet
cherry could reveal mechanisms that can be leveraged to more optimally control flower numbers. There are two main components to this phenomenon. One is represented by the signal emanating from the rootstock and second is the impact it has on target genes that results in the reprogramming of the developmental program at the site of action. One of the steps to bridge this gap in knowledge is to determine which genes in the scion are modulated by the rootstock. The pursuit to fill this gap is the goal of this dissertation.

The best approach to complete this research is the analysis of gene expression through transcriptomics methods. A major caveat to most transcriptomics approaches is the need for a reference data set to utilize before starting the experiments. This reference data set could be a list of genes to interrogate or complete genome and transcriptome sequences. RT-PCR, Northern, microarrays and qRT-PCR all require some sort of a priori knowledge in order to either perform the technique or to analyze the data. Differential display and cDNA-AFLP, however, are unbiased approaches. The downfall is that these procedures require cloning and sequencing and verification of quantitative differences with another method, typically qRT-PCR which can present issues as outlined in Chapter 4A (Page 115) on transcriptomics. The advent of new sequencing technologies has enabled RNA sequencing (RNAseq) and 3’UTR sequencing. The details of these sequencing advances and their influence on transcriptomics are the subject of Chapter 2B (Page 50). These new methods also benefit greatly from a reference, specifically a genome sequence from either the target organism or a closely related species.

At the onset of this project in 2007, no species from the Rosaceae family had an available genome sequence and sweet cherry had 21 sequences available on NCBI. In 2010, a draft of the heterozygous apple was published (Velasco et al. 2010) and a peach genome draft was released shortly thereafter (IPGI 2010). In the meantime, the Sweet Cherry Genome Sequencing Consortium led by the Dhingra Genomics program was formed as an international collaboration to develop a genome sequence for sweet cherry. Additionally, the Rosaceae White paper in 2008 established that it was important for each genus in Rosaceae to have a minimum of one genome sequenced (Shulaev et al. 2008). The reason for this call to action was for an overall improvement in genetics, genomics and breeding for these crops. Additionally,
the development and utilization of molecular markers to advance the traditional breeding of sweet cherry was suggested in this white paper.

In general, genome sequences are of high utility to researchers. As mentioned, many transcript level experiments require knowledge of the genes for a useful conclusion. Research on the primary plant model *Arabidopsis thaliana* has led to many improvements in understanding how plants function. Specifically, Benfey et al. (2010) discuss the benefits of the Arabidopsis genome on root biology. They discuss how the genome aided in reverse genetics, quantitative and association studies. With the reference genome, transcriptomic approaches were enabled to identify genes and networks of genes related to specific stimuli and responses. This is one example of how a genome sequence can facilitate drastic progress in a field of study and the same is expected by the production of a sweet cherry genome sequence.

As sequencing technologies have greatly influenced the ability to perform high-throughput transcriptomic analyses, the same technological improvements have had similarly drastic effects on genome sequencing. Given the ability to generate these large amounts of data, the onus has shifted to the assembly of the short, random fragments into a cohesive genome sequence. Overall, there are two primary methodologies for assembling genomes from these sequences: 1. Reference guided assembly also called reference mapping and 2. *De novo* assembly. Both of these methods are significantly more rapid than the historical methods of primer walking and BAC by BAC sequencing with first generation sequencing data.

Reference mapping has the benefit of speed and the influence of the reference to piece together reads. However, reference guided assembly is limited by the reference used because genes in the species of interest that are not in the reference will not be obtained. Similarly, genomic rearrangements will not be observed via reference mapping as the contigs are oriented according to the reference. These potential problems with reference mapping are reduced as the reference species is more closely related to the species of interest.
De novo assembly is slower and more difficult to accomplish. The small reads do not span repeats causing breaks in assemblies that require a physical map to orient the contigs. The advantages of de novo assembly are that there is no bias for or against sequences in the reference.

Due to the distinct advantages of both assembly approaches, both have been utilized to obtain the best possible version of the sweet cherry genome. Chapter 3A (Page 59) examines the mapping of sweet cherry sequencing data against the peach genome. This work includes similar analysis for four almond genotypes enabling global comparisons of sequence differences among sweet cherry, almond and peach. Alternatively, the de novo approach was pursued in Chapter 3B (Page 90). In this section, specialized sequencing data around restriction sites were utilized to bridge gaps in the sequences and to coalesce the contigs to improve contig length. The restriction-site associated DNA (RAD) data are delivered as raw Illumina reads and pre-assembled RAD contigs. Tools to coordinate these data sets to optimize de novo assembly did not exist therefore the section also focuses on the production and utility of the RADshredder program.

The results from both of these approaches helped to provide the references that were needed for the transcriptomic analyses of the original question: which scion genes involved in floral bud initiation are controlled or influenced by the rootstock. This question is addressed using both differential display and 3’UTR sequencing in Chapter 4A (Page 115).

The results of the transcriptomics project led to several other outcomes. First, the 3’UTR sequences were assembled and yielded many nucleotide level differences, primarily single nucleotide polymorphisms or SNPs. These SNPs were examined, converted into DNA markers, and analyzed in Chapter 4B (Page 143). SNPs are important molecular features that can impact gene function and thereby plant physiology. Haplotypes represent individual alleles and are comprised of physically linked SNPs in this work. The variation of these SNP markers was also tested in a breeding population of sweet cherry (Fernandez i Marti et al. 2012).

From both the genome and transcriptome projects, many sequences have been obtained. Analysis of the genomic sequences for effective restriction sites cutting the sweet cherry genome and 3’UTR
needed to be completed. Existing tools to perform this analysis were not capable of analyzing tens and hundreds of thousands of sequences. To overcome this lack of computational resources, a custom program to perform high-throughput analysis of restriction sites was developed. This tool, CisSERS, is described in detail in Chapter 4C (Page 172). As mentioned previously, many of these sequences contain genetic differences in the form of SNPs. These SNPs have a large likelihood of modifying a restriction site when all potential restriction enzymes are analyzed. As an added utility to the restriction site identification in CisSERS, the ability to visualize how sequences are cut by the restriction enzyme was included to facilitate the conversion of SNPs into DNA markers, specifically Cleaved Amplified Polymorphic Sequences or CAPS markers. These markers can be used to improve breeding efforts if they co-segregate with a trait of interest.

Overall, the goal of the work described in this dissertation was to identify rootstock controlled genes involved in the floral development of sweet cherry scions. Because of the immense lack of resources in sweet cherry, including genetic, genomic, and transcriptomic knowledge, a major focus of this work was to develop these resources before the original goal could be pursued. The following chapters detail the projects from developing the sweet cherry genome through analyses of the transcriptome in developing floral buds of sweet cherry. Each chapter section was prepared for publication in various journals as detailed in italics under the author list for each section. These works combine to enable linking of genes to the unique physiological phenomenon generated by the interaction of the rootstock and scion genotypes in sweet cherry production.
Figure 1: Sweet cherry bearing acres in the United States (adapted from USDA NASS Non-citrus fruit and nut summaries 1999-2012) (NASS 2012).
Figure 2: United States sweet cherry utilized production in blue shown in tons and production value in red shown in $1,000. (adapted from USDA NASS Non-citrus fruit and nut summaries 1999-2012) (NASS 2012).
Works Cited:


CHAPTER 2A: BACKGROUND

Invited Review: Rootstock Scion Somatogenetic Interactions in Perennial Composite Plants

Tyson Koepke¹ and Amit Dhingra¹

In Review at Plant Cell Reports

¹Department of Horticulture, Washington State University
149 Johnson Hall, Pullman WA 99164

tkoepke@wsu.edu, phone 509-335-6586, fax 509-335-8690
adjingra@wsu.edu, phone 509-335-3625, fax 509-335-8690

Abstract

The ancient plant production practice of grafting which instantly imparts new physiological properties to the desirable scion still remains shrouded in mystery. Yet, grafting remains a widely used technique in the production of several horticultural species. In a composite grafted plant, rootstocks control many aspects of scion growth and physiology including yield and quality attributes as well as disease, cold and physiological disorder resistance. Broadly, physical, physiological, biochemical and molecular mechanisms have been reviewed to develop an integrated understanding of this enigmatic process that challenges existing genetic paradigms. This review summarizes the reported mechanisms underlying some of the economically important traits and identifies several key points to consider when conducting rootstock scion interaction experiments. Study of the somatogenetic interactions between rootstock and scion is a field that is ripe for discovery and vast improvements in the coming decade. Further, utilization of rootstocks based on a better understanding of the somatogenetic interactions is highly relevant in the current agricultural environment where there is a need for sustainable production practices. Rootstocks may offer a rapid and non-transgenic approach to rapidly respond to the changing environment and expand agricultural production of annual and perennial crops in order to meet the global food and fuel demands of the future.

Keywords

Rootstock, scion, somatogenetic interactions, composite, perennial
Introduction:

Grafting is widely used in the production of several horticultural species. For many economically important woody perennials such as tree fruits, nuts and grapes, modern production systems are primarily dependent on grafting of a variety or scion on a rootstock that may belong to the same or different species to control important traits including root-related diseases, tree size and yield (Figure 1). These advantages have recently been leveraged in the production of herbaceous dicots. The combination of genotypes in a single composite plant has produced a unique and interesting biological model that transcends the known genetic paradigms. The instant physiological modification of desirable traits in the scion, that mimic genetic changes, is mediated by several hypothesized agents derived from the rootstock. Trait improvement bypasses the reproductive cycle eliminating years of selection during breeding and is a result of cellular or genetic interactions in the somatic cells. We use the phrase ‘somatogenetic interactions’ to represent this phenomenon in composite plants.

The first verifiable report of grafting was in 412 B.C. when grafting appeared to have already been a common technique though some believe that grafting has been documented as early as 1800 B.C. (Mudge et al., 2009). Disease resistance and cold tolerance are traits that are linked to rootstock control and were primary uses of rootstocks as discussed by Rogers and Beakbane (1957). There are many more specific scion traits that are controlled, at least in part, by the rootstock and fall into several general categories that are detailed in the first section of this review.

While knowing which traits are affected by the rootstock is important, unraveling the underlying mechanisms for these interactions is critical to predictably control these traits in production systems. Rogers and Beakbane (1957) also discussed the ‘hypothetical’ mechanisms of rootstock/scion interactions of: nutrients, transport, and growth regulating substances. A review by Webster (2004) discussed many of the potential changes in scion growth that can result in the dwarf phenotype also noting the same potential mechanisms. Several other recent reviews on herbaceous plants examine physical interaction of the rootstock and scion (Martinez-Ballesta et al., 2010) and the hormonal conversation between the two genotypes (Aloni et al., 2010). These reviews form the foundation for the examination of the
physiological mechanisms explaining the rootstocks’ control of scion traits in perennial crops in the second section of this review.

Lastly, molecular evidence for these interactions has begun to be examined and is discussed in the third section. It is expected that the changes in physiology are controlled by gene expression and protein function. Identifying which genes and proteins are at the core of the physiological changes would considerably improve our understanding of how grafting works. These interactions, as will be detailed later, include global changes to gene expression, protein activity changes, and even transport of RNA molecules across the graft union. The effect of RNA transport across the graft union also has a review dedicated to it solidifying the importance of this area of investigation (Harada, 2010). The connections between the physiology, gene expression, and protein function in these plants and how the rootstocks modify each of these is a very intriguing field with many insights in recent years and much remaining to be revealed of these somatogenetic interactions.

Most of the recent reviews on rootstock scion interactions have been focused on herbaceous dicots, specifically grafting in vegetable production (Table 1). This review examines some of the same traits and mechanisms and how those mechanisms function in perennial composite plants (Figure 1). The perennial nature of the species focused on here provides both challenges, such as grafting woody plants, and opportunities, like the ability to evaluate the effects of the rootstocks years after the initial grafting procedure and related stress. Comparisons across the two types of composite plant systems are discussed to add perspective and clarity where experimental data from perennials are less clear. While much has been learned about the many interactions of the rootstock and scion, this field of research is primed for further advancements. Understanding of somatogenetic interactions in composite plants is extremely relevant in these times as we look to conquer marginal lands. Rootstocks may offer a rapid and non-transgenic approach to expand agricultural production of annual and perennial crops.

**PHYSIOLOGICAL TRAITS:**

Grafting has continued throughout the ages due to the documented control of physiological traits. There has been much research detailing and examining the various characteristics that are improved in a
scion based on the rootstock used. Traits related to growth, yield, quality, disease resistance, physiological disorders, and cold tolerance are discussed in this section. The literature reviewed in this work highlights the role rootstocks play in improving scion traits of perennial crop plants. Also included are several important examples from herbaceous annual composite plants (Table 2) to provide a larger perspective.

**Growth (Tree size):** There are several key aspects of scion growth and physiology that are mediated by the rootstock. Though these traits are quite diverse, they are all connected to the composite plant’s vigor. As such, plant vigor can, and has been, evaluated through many physical measures determining the amount of vegetative growth. While high or low vigor is desired based on the species and location, low vigor typically corresponds to dwarf plants and higher flower numbers while vigorous plants generally contain larger amounts of vegetative growth with lower production.

Total plant size is an important scion trait that is controlled by the rootstock and has been shown in many plant families. In 2004, Santos et al. (2004) examined rootstock-mediated dwarfing in sweet cherry and found that trunk cross sectional area (TCSA), final shoot length and final node number were significantly affected by different rootstocks. Similarly, when measuring girth expansion of peach trunks, rootstock genotype demonstrated up to a 2.5 fold effect (Tsipouridis and Thomidis, 2005). Whiting et al. (2005) examined the growth characteristics of the sweet cherry variety ‘Bing’ on a standard, a semi-dwarfing, and a dwarfing rootstock. TCSA varied significantly among the rootstocks starting in the fourth year after planting and expanded to almost 2 fold differences by the end of the 9th season. Likewise, 7 year-old ‘Hedelfinger’ sweet cherry on 3 rootstocks showed significant differences in TCSA (Robinson et al., 2006).

Grafting of a clemetine scion onto several interspecific citrus rootstocks, tree height, canopy diameter, circumference and tree volume varied significantly (Bassal, 2009). The clementine trees on the most dwarfing rootstock were approximately 10 percent smaller in each of these measures. Gijon et al. (2010) showed that pistachio leaf area, leaf and stem dry weights varied based on rootstock. In pear, the
number of growth points and total length of annual growth varied based on the rootstock used (North and Cook, 2008).

Pruning weight and specific leaf area (leaf area per gram of dry mass) are indicators of plant vigor that were controlled by the rootstock in the wine grape scion Cabernet-Sauvignon, though rootstock effects were not the focus of the study (Koundouras et al., 2008). Also, the grape variety Shiraz (Syrah) showed significant differences in pruning weight in each of the two years tested (Soar et al., 2006).

**Yield:** One of the primary reasons grafting is utilized is to control yield, the leading agronomic property controlled by the rootstock. Yield is known to be controlled by the rootstock of various perennial species including the rubber tree (Cardinal et al., 2007), grape (Main et al., 2002; Soar et al., 2006), peach (Tsiouridis and Thomidis, 2005), mango (Smith et al., 2003), sweet cherry (Robinson et al., 2006; Whiting et al., 2005), pear (Bertelsen and Callesen, 2001; North and Cook, 2008), clementine (Bassal, 2009) and apricot (Hernández et al., 2010). The changes in yield can range up to 6.5 fold (Whiting et al., 2005) depending on the crop and diversity of rootstocks examined. The underlying traits included in the overall yield of these perennial species are: precocity, flower number, fruit set and biennial bearing.

Precocious rootstocks were observed by Smith et al. (2003) in mango where early year production contributed to the 2.4 fold cumulative yield difference of ‘Kensington Pride’ mango on 9 different rootstocks. In pears, Bertelsen and Callesen (2001) identified precocity and biennial bearing effects of the rootstock with the ‘Clara Frijs’ scion having a several year range for the start of production. Higher flower and fruit densities correlated with higher overall yields even though there were also correlations to decreased fruit weights. The precocity effect of the rootstocks is also true in apricot according to Hernandez et al. (2010). The cumulative yield data support the conclusion that some rootstocks caused earlier production though the yearly production data to support this were not provided. Hernandez et al. (2010) continued to document how the fruit number per branch and overall fruit set were also strongly controlled by the rootstock in this apricot trial.

Whiting et al. (2005) has shown control of productivity and precocity on sweet cherry scions by the rootstock. Robinson et al. (2006) also showed that differences in yield of three sweet cherry scions
was controlled by the rootstock but fruit size showed no significant differences. Since fruit size was therefore not a contributing factor to the differences in yield, fruit number must have been the cause. Whiting (personal communication) later found that floral bud number was linked to the nearly 10 fold yield differences found on ‘Bing’ sweet cherry on multiple rootstocks.

In grapes as examined by Soar et al. (2006), rootstocks producing greater canopy surface area and pruning weight were correlated to greater yield. This could be due to pruning of each plant to the same number of buds during winter thus allowing more vigorous rootstocks to be more productive. Also in grapes, Main et al. (2002) discovered that rootstock control of ‘Chardonel’ berry weight contributed to changes in cluster weight and eventually yield differences on the different rootstocks. The impact of berry weight on yield in this study was furthered by the observations that the number of clusters per vine was the same for all rootstocks. Likewise, Bassal et al. (2009) showed that clementine fruit weight was unchanged by rootstock leaving fruit number to be inferred as the cause for the yield differences.

North et al. (2008) showed the rootstocks’ control over pear yield on a per tree basis which correlated strongly with the number of fruit per tree while fruit weight did not correlate to overall yield. Also in pear, Sugar et al. (2005) identified significant differences in fruit set based on the rootstock, though yearly variation was also a significant factor. Yield was not directly reported but presented as yield efficiency based on TCSA which varied greatly by rootstock so the changes in efficiency could be solely due to TCSA differences. Similarly, bloom and crop densities per TCSA were controlled by the rootstock.

In contrast, Russo et al. (2007) demonstrated that the tested apple rootstocks significantly controlled fruit weight and yield though fruit number were not reported. Similarly, peach fruit weight and total yield were controlled by the rootstock and fruit number were not shown (Tsipouridis and Thomidis, 2005).

In cotton, Dong et al. (2008) showed that the yield was significantly controlled by the rootstock. The rootstock also had significant effects on the boll weight and number of bolls per plant and seed cotton per plant which accounts for the changes in yield.
These are some examples of the role of the rootstock on reproductive traits underlying yield in perennial plants. In combination, it is clear that flowering, fruit set, and fruit quality aspects related to yield are modified by the rootstock. Flowering is known to be controlled by many environmental variables including water and nutrient stress. The interaction of the rootstock on these variables as well as hormone levels could control these reproductive traits and are discussed in detail later in this review. It is also important to consider that yield based assessments are commonly confounded by “typical” pruning and management practices. This is clearly seen in Whiting et al.’s (2005) work on sweet cherry where the rootstock had a significant effect on yield, but the training system was also quite essential to the productivity of the system.

**Quality:** Fruit quality is another economically relevant trait controlled by rootstocks in many different species. As the types of products from perennial composite plants vary greatly, so do the quality attributes of interest, some of which reveal changes in most of these scion/rootstock assessments.

Fruit weight was widely noted to be controlled in many species including peach (Tsipouridis and Thomidis, 2005), grape (Koundouras et al., 2009; Main et al., 2002), grapefruit (McCollum et al., 2002), pear (Bertelsen and Callesen, 2001; North and Cook, 2008; Sugar et al., 2005), apricot (Hernández et al., 2010), and sweet cherry (Whiting et al., 2005). Alternatively, Bassal (2009) found no change to fruit qualities such as fruit size, weight, juice %, and peel thickness of clementine, but acidity was strongly controlled by the rootstock and while firmness and ascorbic acid content showed differences, the environment exerted greater control than the rootstock.

Likewise, the acidity of grapefruit, along with the brix were reported to be rootstock controlled in ‘Marsh’ by McCollum et al. (2002). Fruit brix, pH and titratable acidity were also found to be significantly effected in rootstock trials with the grape variety ‘Chardonel’ and the resulting wine (Main et al., 2002), while Koundouras et al. (2009) found no differences in skin anthocyanins flavan-3-ol monomers, or phenol-free glycosyl-glucose (PFGG) by the rootstocks of ‘Cabernet Sauvignon’. Tsipouridis and Thomidis (2005) also found peach total acids along with soluble solids and flesh firmness to be rootstock related traits.
Fruit firmness and color of sweet cherry fruits were found to be controlled primarily by the scion although rootstock did have a significant effect (Gonçalves et al., 2006). Also in sweet cherry, Whiting et al. (2005) identified a significant effect of the rootstock on fruit size. Flesh color, soluble solid concentration and flesh firmness were identified as rootstock controlled in kiwifruit (Thorp et al., 2007). In apricot, Hernandez et al. (2010) found flesh firmness, fruit shape, pulp thickness, stone weight, percent pulp, soluble solids, and fruit color were all rootstock effected though only one year of data were presented leading to the possibility that these traits may be highly environmentally regulated as seen in studies on other crops. Similar to these results, North et al. (2008) identified a significant role of the rootstock on fruit firmness, background color, and total soluble solids of pears. Also, Sugar et al. (2005) showed the amount of russet, a physiological alteration of the pear epidermis that becomes corky, was effected by the pear rootstock.

A much more drastic example of quality modification by the rootstock is presented in tobacco grafted on tomato where the production of nicotine is nearly eliminated (127µg/g dry weight) compared to self-grafted tobacco plants (6,720µg/g dry weight) (Ruiz et al., 2005). This suggests that the amount of control of scion qualities documented to date in perennials may by a small fraction of the potential impact rootstocks may have on these traits.

In these assessments of fruit quality characteristics, the fruits were tested based on the same date of harvest. This ignores data demonstrating that maturity is modulated by the rootstock genotype. Days to full bloom and days to fruit maturity varied slightly due to the rootstock genotype for Peach varieties (Reighard et al., 2011). DeJong et al. (2004) found fruit maturity differences were also considered large enough to modify sampling procedures when comparing Peach on several rootstocks. Additionally, Barden and Marini (1992) used an apple fruit maturity index to identify maturity differences caused by different rootstock genotypes. Future investigations into the effect of rootstock on fruit quality attributes may be well suited, and far more reproducible and widely applicable to the community, by assessing and harvesting fruit based on maturity as these examples have shown, rather than picking all fruit on the same date.
**Disease Resistance:** The ability to control disease, as mentioned in the introduction, is one of the primary historical advantages of rootstocks. The vast majority of the disease resistance, however, is based on soil-borne diseases that infect the roots. These diseases can kill the plants before they come into production or cause significant damage and yield reductions. While resistance and tolerance to these diseases is critical, it does not depend on the interplay between the rootstock and scion and are therefore not covered in detail in this review. This includes the reported fire blight resistance in apple (Russo et al., 2007) where the resistance prevents the rootstock from dying but does not aid the scion’s resistance, and pear decline (Seemüller et al., 2009) where the rootstock genotypes varied in severity of symptoms but did not prevent scions from obtaining the causal bacteria. Still, there are a couple of examples where the rootstock influences the disease characteristics of the scion.

While Russo et al. (2007) were unable to show a rootstock effect on fire blight resistance, Jensen et al. (2003) report a difference in ‘Gala’ apples on several rootstocks. The difference is noticed only when the canker length is expressed as a percentage of the shoot length. This presents dwarfing rootstocks as more susceptible on a percent of shoot length basis while the mean canker lengths are not significantly different among any of the rootstock/scion combinations. The distribution of canker length, however, does support the conclusion that there is a difference among the rootstocks ability to control fire blight though more extensive examination will need to be completed to fully assess this trait.

A second example is anthracnose development in avocado, a serious post-harvest fungal disease in avocado that leads to rotting of the fruit. After post-harvest ripening at 22°C, fruit of ‘Hass’ scion from 3.5 year old and 8 year old trees on each of two different rootstocks demonstrated significant differences in both disease incidence and severity (Willingham et al., 2001). Most importantly, the approximately 30% decrease in disease incidence and severity on one rootstock corresponded to about 50% more acceptable fruit. To understand the mechanism for this resistance, the authors examined diene concentrations in the same trees and fruits since several dienes had previously been linked to fungal disease resistance. A significantly higher amount of dienes was found in the leaves of trees on the more resistant rootstock suggesting a role of the dienes in fruit resistance to anthracnose infection of avocado.
Leaf nutrient levels were also examined and showed significant differences thereby suggesting that scion nutrition may be the trait indirectly controlling anthracnose resistance.

Vegetable grafting for disease has also been recently reviewed by King et al. (2007) though, as seen in perennials, there are few examples of non-soil borne, scion diseases. Most of the resistance to these non-soil borne, non-root related diseases were explained simply as an effect of vigor allowing the plants to sustain production while diseased. The contrary example mentioned in that review is that Edelstein et al. (1999) showed spider mite resistance was imparted to Cucurbita sp. scions from the rootstock.

Overall, these are several examples of the rootstock controlling scion diseases though they are not near as prevalent as would be expected based on the role of grafting to control root diseases. The total effect of rootstocks on scion and fruit diseases may be larger than currently documented as the results may be so far removed from the rootstock that observations have not been completed.

**Physiological Disorders:** In addition to controlling diseases, physiological disorders of the fruit borne by the scion are controlled by the rootstock. Because these disorders are not caused by an external organism, they are covered separately here.

Ritenour et al. (2004) examined stem-end rind breakdown (SERB) in citrus where the fruit develops a very narrow ring of unaffected tissue immediately around the stem surrounded by a dark, irregularly-shaped region of collapsing peel tissue, a condition known to be influenced by both pre- and post-harvest conditions. For ‘Valencia’ oranges on eight rootstock genotypes in 2001, Ritenour et al. (2004) found SERB ranged from 56 to 21 percent of fruit after 41 days at 70°F (21°C) and 54 to 26 percent after 101 days at 38°F (3°C). In 2002, SERB was far less severe but still displayed significant differences with the same rootstock being worst in both years. Rootstock was also shown to affect SERB in navel oranges. ‘Ray Ruby’ grapefruit, however, showed no significant SERB while a rootstock effect on post-harvest decay was shown. ‘Oroblanco’, a hybrid of pummel and grapefruit, displayed similar results to ‘Ray Ruby’ with SERB being insignificant and the rootstock having a large impact on post-harvest decay.
In kiwifruit, differences in physiological pitting, sun damage, and stem end browning were observed on a single scion variety on 8 different rootstocks (Thorp et al., 2007). Thorp et al. continued to test these differences which ranged from 76.6 percent for sun damage to 34.5 percent for stem end browning but were unable to find a significant factor. Physiological pitting, however, ranged up to 45.5 percent based on rootstock and was negatively correlated to Mg concentrations though it is unclear if Mg levels significantly differed among rootstocks.

As alluded to by Thorp et al. (2007), physiological disorders may result from nutrient differences caused by the rootstocks. Additionally, these disorders could both be drastically controlled by the physiological characteristics of the fruit which, as mentioned in the fruit quality section, can be modulated by the rootstock. These physiological disorders may be caused by abiotic stresses, such as water shortage and nutrition, which are covered in detail in the physical mechanisms section.

**Cold Tolerance:** Another economically vital agronomic trait is cold hardiness. Cold hardy rootstocks have enabled production of many perennial crops in severe climates and rootstock cold hardiness is a critical evaluation trait for rootstock selection. While it is clear that different genotypes have different cold hardiness attributes, being able to impart differential cold hardiness on the scion is of particular interest.

Tsipouridis and Thomidis (2005) found that the rootstocks effected the frost resistance of ‘May Crest’ peach with the percent of flowers that were damaged from frost ranging from 0 to 55. Similarly in Pecan, Smith et al. (2001) found that trees on some rootstocks were more susceptible to early frost damage than others. Data were pooled for all cultivars on the rootstocks making the rootstock/scion interaction unrecognizable, possibly diluting the effect of the rootstock on more susceptible varieties. The differences ranged from 0-9 percent of trees on a given rootstock that were expected to die due to the extensive damage resulting from the early frost. In pistachio, Epstein et al. (2004) found that frost killed 41 percent of pistachio trees with 72 percent of the surviving trees being damaged while another rootstock showed zero death and only 4 percent of trees were damaged. Chilling injury was also markedly controlled by rootstock for the grapefruit variety ‘Marsh’ (McCollum et al., 2002).

20
These chilling injury studies are the result of actual environmental anomalies making them nearly impossible to reproduce. Controlled growth rooms and greenhouses, however, could be used to simulate these types of rapid temperature changes at distinct stages of growth to determine if these effects are maturity-based susceptibility. Care would need to be taken to modify the air temperature without significantly changing the soil temperature below the surface to more fully replicate the environmental shock the perennial composite plants experience in the field.

**Summary:** These results show the widespread effect of rootstocks on scion physiology. While these traits were presented individually, they are highly integrated and changes in one likely alter many of the others. This is most obvious when considering fruit number and weight where excess crop load leads to decreased size of the fruit. Additionally, reductions in plant vigor are associated with decreased growth and increased reproductive characteristics. Rootstock-controlled plant vigor may be related to resistance to disease (King et al., 2007) and physiological disorders of perennial scions. The interactions of these traits, however, are also likely modulated through one or more of following mechanisms.

**PHYSICAL MECHANISMS**

As has been clearly established, there are many traits in perennial species scions that are controlled by the rootstock they are grown on. To understand how the rootstock causes these effects, many experiments have been conducted examining multiple facets that can control plant growth and physiology (Table 3). These range from microscopic analyses of the graft unions to testing nutrient and water passage to the scion. The following subsections detail research on the mechanisms for the rootstock to control scion physiology in the areas of: graft union repair, water usage and transport, hormones, and nutrition.

**Graft Union Repair:** In herbaceous plants, the formation and function of the graft union has become a point of interest. The healing and formation of the graft union was also described in *Arabidopsis* to have four stages by Flaishman et al. (2008). Recently, Yin et al. (2012) expanded this to six major steps: 1. Wound-induced response, 2. Cell debris clean, 3. Graft union cellular communication, 4. Auxin accumulation and responses, 5. Graft union cell division and differentiation, 6. Vascular reconnection.
Yin et al. (2012) also identified that mRNA signals are changed in the graft union within one day after grafting. By the second day after grafting, auxins had stimulated cell division and differentiation. On the third day after grafting, transport was already functioning across the graft union.

As with herbaceous composite plants, the development of a functioning graft union in perennial crops is a critical stage during the production of a composite plant. Many studies to document and understand the development of the graft union in perennials have been completed including grafting of woody tissues as well as micrografting of herbaceous tissues of these plant species.

Examining the development of the graft union in olive through hydraulic resistivity and hydraulic conductance experiments, Gasco et al. (2007) showed a much longer period of healing is required for woody perennial plants. Gasco et al. (2007) measured hydraulic resistance in the roots, across the graft union, and in the shoot to determine the percentage that each comprised of the total plant hydraulic resistance. At day 30, the graft union of the heterografts, grafts between different genotypes, accounted for roughly 85 percent of the total resistivity. By day 90, these had dropped to around half of the total hydraulic resistance. Measurements at 360 days after grafting showed that the graft union was a minor part of the total resistance and continued to drop to about 3 percent of the total by the 480th day after grafting. In addition, there were no distinguishable differences in graft union hydraulic resistance when comparing the reciprocal grafts of a dwarfing and vigorous variety. These data provide evidence that in olive, and probably most perennials, the timing of graft union examination and hydraulic studies needs to be carefully considered when interpreting the results. This is because in tests soon after grafting the graft union will have a much larger influence than later in the growth of these plants. Also, since the graft union resistivity ended up decreasing in impact to be nearly negligible, the graft union hydraulic properties do not explain sustained dwarfing. This conflicts with the data from herbaceous composite plants where grafts were functional days after grafting. This discrepancy could result from graft union healing affecting the growth of the plant during that year while Gasco et al. (2007) were able to isolate graft healing from the hydraulic measurements over a year after the grafting stress. A report of micrografting in peach by Zhang et al. (2012), however, shows that the graft union is functionally transporting
RNAs within 4-10 days of grafting suggesting that the stage, species, and type of grafting used can significantly modify the graft union healing process.

In contrast, Atkinson et al. (2003) examined an apple scion on three rootstocks one year after grafting for their root and stem hydraulic conductivity. Analysis of percentage functional xylem revealed that the graft tissue and scion stem vasculature varied based on the rootstock. The graft tissue conductivity increased very strongly with vigor and suggested that the graft union vasculature is a critical part of vigor control by the rootstock. However, since measurements were not completed to demonstrate that the graft union had fully healed at the time of the experiment, these results could arise if the graft union was still unhealed.

To understand early vascular development in the grafts in sweet cherry, xylem vessel anatomy was probed in combinations of the Rainier scions and non-dwarfing, semi-dwarfing and dwarfing rootstocks six months after bud grafting (Olmstead et al., 2006a). Through self-grafts of the rootstocks and the Rainier/rootstock heterografts, these experiments clearly demonstrated that graft union xylem vessel number increased and lumen volume decreased with the increasing dwarfing nature of the rootstock. Changes in the scion tissue were also discovered though they are much less drastic than the graft union differences. The amount of wounding-induced callus at the graft union was also found to differ among the combinations. Olmstead et al. continued to note that the formation of xylem rays in the graft union callus tissue was abnormal, forming non-linearly, thereby reducing the continuity of xylem across the graft union.

Using intact two year old nursery budded trees extracted from the field, dye uptake and transport was correlated with decreased vigor studies of the sweet cherry scion ‘Lapins’ on a non-dwarfing and dwarfing rootstock (Olmstead et al., 2006b). The dye showed a decreased flow into the graft union and was reduced further in the scion just above the graft union in plants grafted on the dwarfing rootstock compared to those grafted on vigorous rootstock which was associated to xylem vessel size as seen in their work on Rainier grafting combinations.
Tombesi et al. (2010) also examined xylem characteristics in composite plants using peach as their model including vigorous, semi-dwarfing and dwarfing rootstocks. Xylem was analyzed above and below the graft union(s) of six and seven year old trees and demonstrated that higher rootstock xylem conductance correlated with higher vigor. Additionally, the rootstock genotype did not significantly modify the xylem characteristics of the scion and the vigorous/dwarfing/scion combination displayed an intermediate phenotype. Tombesi et al.’s data lead to the conclusion that xylem restriction in the root, stem, or scion may have resulted in a dwarf phenotype in peach.

Several of the studies above examined the repair of the graft union directly through dye uptake and microscopic examination while the others utilized the resistance to water transport across the union. It appears that the state of the tissue when the graft is performed, being herbaceous or woody, may have a significant effect on the rate of graft union healing and subsequent effects on the growth and physiology of the composite plants. Though they are separate sections in this review, the connections between water usage, hydraulic conductivity and healing of the graft union are nearly impossible to separate.

**Water Usage and Transport:** Water usage is commonly examined through the measurement of water potential, hydraulic conductivity, or hydraulic resistivity in the grafted plants. Intuitively, increases in hydraulic resistance lead to decreases in hydraulic conductivity which, when severe, can display as drought-like symptoms.

Peach tree daily growth and stem water potential were examined by Basile et al. (2003). The stem relative extension rate was found to be significantly different for many time points during many of the days examined. Stem water potential showed more time points with significant differences and followed the general trend of decreasing as relative extension rate was increasing and vice versa. Additionally, a strong correlation of daily relative extension rate was shown with the mean air temperature.

Clearwater et al. (2007) examined kiwifruit rootstocks of different vigor for their water pressure in various parts of the plants. Through these experiments, Clearwater et al. (2007) identified that spring root pressure increases were rootstock dependent and a delay in increasing pressure was correlated to reductions in scion vigor. Root pressures also correlated to scion leaf pressures suggesting that these
vigor-reducing rootstocks of kiwifruit acted by causing water deficit stress in the scion. Koundouras et al. (2008) compared the water potential of the grape varietal Cabernet-sauvignon on two different rootstocks and identified a significant effect of the rootstock in non-irrigated conditions but not with deficit or full irrigation. The supply of deficit or full irrigation reducing the significance of the rootstock on water pressure suggests an important role for grape rootstocks in drought or low water conditions.

In olives, Nardini et al. (2006) tested reciprocally grafted and ungrafted dwarfing and vigorous scions. Using the dwarf as a rootstock lead to a reduction of leaf surface area by about half and correlated with the 2.5 fold decrease in hydraulic conductance of the dwarfing root system. Additionally, they identified that the root system accounted for 60-70 percent of total plant hydraulic resistance while the graft union had a negligible effect on hydraulic resistance. This finding is congruent with the graft union healing presented in the previous subsection since the olive plants were tested at 360 and 450 days after grafting, when the graft union of olive is nearly and fully functional respectively according to Gasco et al. (2007).

Solari et al. (2006) showed that hydraulic conductance in the rootstock, scion and graft union of peach trees were all controlled significantly by the rootstock with the larger conductance on the larger trees. The graft union, however, only account for 10 percent of the overall hydraulic resistance on each one year post grafting peach tree, also consistent with the results in olives by Nardini et al. (2006) and Gasco et al (2007).

In a similar publication on water relations of rootstock/scion interactions, Solari and DeJong (2006) used exogenously increased root hydraulic pressure on composite peach trees. By pressurizing the roots, they were able to demonstrate increases in growth rate, leaf transpiration and conductance, and net CO₂ exchange of peach on both dwarfing and vigorous rootstocks. In both rootstock types, increases in pressure resulted in the mentioned measures in a very linear fashion. Additionally, by recording the relative shoot extension rate during root pressurization, the increased root pressure caused a short-term stimulation of growth that was nearly 500 fold increased within minutes which the authors partially attributed to the elastic expansion of cells. This response then decreased to a new steady state within 15
minutes after the pressure treatment started and the steady state rate of elongation for the pressurized plants was higher than that of non-pressurized control plants. These results clearly demonstrate that root hydraulic pressure is a significant controlling factor of scion vigor in peach.

Since observations consistent with this finding have been present in many of the perennial species employing grafting, it is likely that root hydraulic pressure is a major player in vigor control of each of these species. Similar experiments testing the hormone and nutrient concentrations of the increased hydraulic flow would be interesting to identify if there is a change in absolute amounts being delivered to the apical meristems for growth during the pressure events.

**Hormones:** Understanding the hormonal relationships among the rootstock and scion are critical to develop a concept of how these genotypes interact. A recent review of hormonal relationships of grafted crops discussed in detail the effect of exogenous auxins and cytokinins on the establishment of the graft union vasculature (Aloni et al., 2010). Auxins, cytokinins, ethylene, reactive oxygen species, and abscisic acid (ABA) were also discussed with regard to graft incompatibility, senescence, and salt stress. Aloni et al. (2010) continued to dissect the role of hormones on many of the physiological traits discussed previously in this review from the perspective of herbaceous composite plants.

Hormone examinations have also been reported in perennial grafted plants. Kamboj et al. (1999a) analyzed grafted and ungrafted apple cultivar ‘Fiesta’ shoot xylem sap for amounts of zeatin and zeatin riboside, the two primary cytokinins in apple xylem. In the ungrafted rootstock varieties, xylem sap cytokinin levels did show a correlation with vigor; however, the role of the scion in the interaction muted these differences to where statistical significance was not found. Additionally, the more vigorous rootstock contained a higher proportion of zeatin riboside than zeatin and the opposite was true for the two dwarfing rootstocks though the authors state that this observation could be a result of differential timing of growth between the rootstocks. Using a similar design, Kamboj et al. (1999b) analyzed ABA and the auxin indole acetic acid (IAA) in the xylem sap of non-grafted rootstocks as well finding small correlations but also lacking statistical significance. These correlations were that mean ABA concentrations decreased with increasing vigor of the rootstocks. The differences in IAA were not
significant among the tested population of rootstocks. While these results are primarily within non-grafted
rootstocks, they demonstrate differences in the natural state of the rootstock genotypes on their hormone
production.

Scion levels of zeatin riboside and IAA amounts were significantly controlled by the rootstock
when Sorce et al. (2002) examined the hormone interactions of the peach variety ‘Armking’ on three
interspecific rootstocks. Vigor of the plants was positively correlated to zeatin riboside concentrations and
negatively with IAA concentrations. Due to the examination of ungrafted control rootstocks, the authors
were able to note that IAA and zeatin riboside were balanced in the ungrafted plants. That balance was
not seen in the grafted plants suggesting the hormonal conversation between the rootstock and scion was
altered which could explain the changes in vigor.

Soar et al. (2006) examined ABA levels to correlate ABA with differential drought responses of
the wine grape Shiraz (Syrah) on seven different rootstocks. The sap levels of ABA did not provide
significant evidence that xylem transported ABA was controlling the different drought characteristics
imparted by the tested rootstocks. Although there was a negative correlation of ABA concentrations with
stomatal conductance, the correlation varied greatly between the two years. While ABA levels did not
completely explain the differences, the authors recommended that sap screening of ABA concentrations
may be a rapid screen to assess relative drought tolerance of rootstocks. Additionally, Soar et al. (2006)
state that differences in water stress may have produced the changes in ABA rather than ABA causing the
observed changes in water stress.

Several studies in apple have been reported to examine the hormonal landscape produced by
different rootstocks. Tworkoski and Miller (2007) showed that increased auxin to cytokinin ratios
corresponded to decreased bud break in removed branches from mature trees. Similarly, Van Hooijdonk
et al. (2010) showed that exogenous GA_{4+7} had an equilibrating effect on final node number of stems of
apple scions on dwarfing and vigorous rootstocks during the first seven months after grafting. On one
year old apple trees, Li et al. (2012) found that increased zeatin in leaves and roots of scions correlated
with increased vigor. IAA content was decreased in the scions by the dwarfing rootstock with significant
differences found in leaves, branch bark, and the roots. The graft union, however, was found to have no obvious influence on hormone transport as the IAA interaction was within one genotype and not the graft union.

Dong et al. (2008) described the effects of self and reciprocal grafting on early and late senescence lines of cotton. Higher ABA concentrations were found in plants on the early senescing rootstock and their chlorophyll content and net photosynthesis decreased. Concentrations of the cytokinins zeatin plus zeatin riboside, dihydrozeatin plus dihydrozeatin riboside, and isopentenyl plus isopentenyl adenine were consistently higher in the leaves of plants on the late senescing rootstock. Overall, this work shows that both ABA and cytokinins are controlled by root genotype and both control senescence.

The role of the rootstock in hormone manipulations in the scions has been shown for auxins, cytokinins, and ABA directly and GA$_{1+7}$ indirectly. The crosstalk between auxins and cytokinins seem to control scion vigor with auxins being negative regulators and cytokinins contributing positively. Interestingly, differences in the type of cytokinin, specifically the ratios zeatin versus zeatin riboside (Kamboj et al., 1999a), were implicated to control vigor though the absolute amounts of each were not reported so it may be a change in one that significantly disturbed the reported ratios. Additionally, higher cytokinins were correlated to decreased senescence which could be a mechanism for explaining how high cytokinins can increase overall growth. The role of ABA inducing senescence is consistent with previous examinations of this hormone. The origin of ABA is likely to be from the root as Dong et al. (2008) showed export of ABA from the roots through the xylem sap and is consistent with the recommendation of Soar et al. (2006) to test rootstocks for xylem ABA levels as a preliminary drought tolerance screen. Combining the information from Dong et al. (2008) and Kamboj et al. (1999b) showed that ABA was negatively correlated to vigor which could result from earlier senescence. ABA caused decreases in stomatal conductance which yields decreased photosynthesis and may promote the onset of senescence. The interplay of all of these hormones in rootstock scion interactions needs to be examined more fully. By measuring the levels of these hormones in multiple graft combinations across the developmental time
course, these interactions may be more fully revealed enabling more concrete assertions to their roles in rootstock scion interactions to be made.

**Nutrition:** Since the roots are responsible for absorbing vital minerals and nutrients for the entire plant, it is expected that the rootstock would have a significant contribution to the nutritional status of the composite plant. Here, the results of analyses of scion mineral content, salt stress tolerance, and photosynthesis are reviewed.

*Scion mineral composition:* ‘May Crest’ peach was grown on 14 different rootstocks and the leaf mineral composition was analyzed. P, Ca, Mg, Mn, Cu, and Fe all showed 2-6 fold differences in the percent composition (Tsipouridis and Thomidis, 2005). North et al. (2008) found percent dry mass of N, P, and Mg in pear leaves was rootstock controlled while K and Ca did not change. Similarly, Fallahi et al. (2001) examined leaf mineral content on a percent dry weight bases of ‘bc-2 Fuji’ apples on three rootstocks. Ca, Mg, K, Fe, Zn, Cu, and Mn varied by year and N concentrations were only significant based on rootstock for one of the two years of the experiment. Since each of these tests based the amount of each mineral on the total % composition, a significant change in one ion may have influenced the total percentages of all of them. The authors also noted differences in fruit production during the sampling year which could have significantly contributed to the differences in leaf nutrient levels.

Cl, K, Ca and Mg were all found to accumulate at different rates in leaves of mango when two rootstocks were reciprocally grafted by Schmutz and Ludders (1999) while Na assimilation was not affected by the rootstock genotype. Almansa et al. (2002) also examined the contribution of the rootstock to accumulation of minerals in lemon leaves during salt stress finding that the absolute amounts of K, Na, Fe, Mn, Cu, and Zn did not change due to the rootstock. Meanwhile the Cl levels were changed significantly by the rootstock, especially in the high salt treatment, which correlates with Schmutz and Ludder’s findings suggesting that chloride and not sodium may be the toxic element of salt stress.

Papadakis et al. (2004) examined the effect of two rootstocks on ‘Naveline’ orange and their responses to high boron concentrations. They found that the rootstocks absorbed the boron at different rates which lead to the differences in scion boron concentration and subsequent boron toxicity levels. The
effects of the rootstock were most apparent on the higher concentration of boron though the resulting physiological changes were not significant between the rootstocks.

Salt stress: Schmutz and Ludders (1999) showed that the rootstock genotype was significant for the salt tolerance phenotypes of mango. By reciprocal grafting of two rootstocks and applied salt stress, reductions in total growth, CO₂ assimilation, and transpiration were identified. Mickelbart et al. (2002) tested three avocado rootstocks in combination with salt stress to examine the effects on the ‘Hass’ avocado scion. The rootstocks had significant effects on the amount of Na⁺ and Cl⁻ ions in the leaves which were most extreme at the highest tested level of salinity. The amount of accumulated ions also correlated with decreases in shoot extension and individual leaf area which were rootstock related traits in this experiment. Additionally, leaf necrosis occurred at the highest level of salinity though primarily in the rootstock that accumulated the highest amount of ions. Chlorophyll content also decreased with increases in NaCl accumulation demonstrating an interaction of salt tolerance with photosynthetic capacity.

Examination of Chinese, Japanese, and European pears on two Asian pear rootstocks allowed Okubo et al. (2000) to show that the rootstock had large differences in salt tolerance. Fresh weight, scion defoliation and mortality each had significant differences based on the rootstock. The differences in the traits being tested were even more significant at higher salt concentrations. Some of the scions also clearly showed differences in chloride concentrations and smaller changes in sodium concentrations. These results correspond well with Martinez-Rodriguez et al. (2008) which showed that leaf sap Na concentrations were significantly different based on the tomato rootstocks.

Photosynthesis: In addition to leaf mineral composition, Fallahi et al. (2001) found that net photosynthesis was slightly affected by the rootstocks on bc-2 Fuji apple and was not correlated to overall vigor. Stomatal conductance, intercellular CO₂ concentration, transpiration, and leaf area were all significantly controlled by the rootstock though yearly variation was also apparent for transpiration and leaf area. Averaging measurements for three sweet cherry varieties, net CO₂ assimilation, stomatal conductance, and intracellular CO₂ concentration were lower on dwarfting rootstocks than on semi-
dwarfing and vigorous rootstocks (Gonçalves et al., 2006). Conversely, intrinsic water use efficiency was higher in these dwarfing rootstocks compared to the others.

**Summary:** As seen in the physiological traits section, the physical mechanisms also interact highly. Hydraulic pressure is expected to modulate the amount of hormones and nutrients moved into the scion. This will also be affected by the graft union during the healing process. Nutrient levels, water content and hormones also control the photosynthetic capacity and subsequently the growth. Nutrient accumulation differences due to the rootstock are most likely due to adsorption differences of the roots. This was demonstrated by the differential accumulation of ions during salt stress treatments. Though these mechanisms can explain a significant amount of the rootstock controlled traits, little is understood for how these mechanisms are developed and controlled by the rootstock and forms the basis for the final section of this review.

**MOLECULAR MECHANISMS**

The previously discussed studies have revealed many of the physiological aspects that control rootstock scion interactions including several mechanisms with very strong supporting evidence. It is critical to learn what genetic differences underlie those physical mechanisms and how those physical mechanisms further manipulate the gene expression in the scion and ultimately alter the traits in a grafted scion. Recently, biochemical analyses on protein functions and activity have shed some light on a couple of important pathways in composite plant growth and development. In the past decade, there have been drastic improvements to molecular techniques improving the ease of use, ease of analysis, throughput, and overall costs of gene expression experiments. These advances have moved expression studies from cDNA-AFLP and differential display to microarrays and now to RNA sequencing. Through expression analyses of composite perennial plants, genes putatively controlling plant physiology have been revealed. The following section details molecular analyses to understand the somatogenetic interactions of the rootstock and scion in perennial composite plants (Table 4).

**Biochemical Analysis:** Examining lemon tree leaves on different rootstocks, Almansa et al. (2002) showed that the superoxide dismutase (SOD) activity in the leaves was rootstock controlled and varied.
more by rootstock than the salt levels tested. Another protein activity study was reported by Agbaria et al. (1998) on grafted rose plants. Two scion varieties were grown reciprocally grafted and one interspecific rootstock was also used for both varieties. The activities of glutamine synthetase and nitrate reductase in the leaves were significantly modulated by the rootstock though total nitrate levels did not differ significantly because of the rootstock. As discussed in the disease section, increased production of dienes by Hass avocado led to decreased anthracnose infection of the avocado fruit demonstrating an ability of the rootstock to control biochemical composition of a scion and its fruit (Willingham et al., 2001).

**Expression Analysis:** One way the rootstock can control many aspects of scion growth including protein activity is through the modulation of scion gene expression. As mentioned in the hormone section, Li et al. (2012) found significant differences in kinetin in two year old scions. Additionally, they found that root expression of IPT3, an important gene in cytokinin synthesis, correlated with increased cytokinin levels and vigor. Separately, expression of \textit{PIN1}, a gene known to be involved in polar auxin transport in model plants, was far decreased in the trees with a dwarfing genotype used as an interstock. This led to the hypothesis that basipetal IAA transport is significantly reduced in the dwarfing genotype leading to whole plant hormonal alterations and potentially causes the dwarfing effect.

Global gene expression patterns of three year old ‘Gala’ apple scions on dwarfing and semi-dwarfing rootstocks were investigated by Jensen et al. (2003). Their cDNA-AFLP approach to examine shoot tip gene expression identified 43 bands up-regulated on the semi-dwarfing rootstock correlating to 36 unique genes and 95 bands up-regulated on the dwarfing rootstock relating to 56 unique genes. Of the 92 genes, ontologies from 26 percent were transcription/translation related and 31 percent had no homology or no known function. Many of the identified genes showed potential roles consistent with physiological differences in photosynthesis, tree size, stress tolerance, and flowering. Jensen et al. discussed many of the identified genes and their potential roles including a gene with similarity to the \textit{Arabidopsis BAK1} (brassinosteroid insensitive I (BRI)-associated protein kinase gene) which was up-regulated in the dwarfing rootstock compared to the semi-dwarfing genotype. Additionally, a sorbitol-6-phosphate dehydrogenase (SDH) was up-regulated in the scion on the dwarfing rootstock. Several
identified gene expression differences were confirmed with qRT-PCR, however, no functional gene analysis was performed.

The mechanism of rootstock-induced dwarfing in sweet cherry was examined by Prassinos et al. (2009) using ‘Bing’ on a dwarfing and a semi-dwarfing rootstock. Their results show that the dwarfing rootstock caused shoot elongation of one and two year old scions to stop earlier than that of scions on the semi-dwarfing rootstock. As a result of earlier growth cessation, the number of nodes and total length of the branches were decreased in scions on the dwarfing rootstock as was trunk cross section area. Prassinos et al. (2009) then used cDNA-AFLP and microarray approaches to examine the gene expression in multiple tissues from the one year old scions. RNA was extracted below the graft union, at the graft union, and above the graft union before, during and after the differences in growth cessation. Comparing across rootstocks through the cDNA-AFLP experiment, 49 and 136 transcript derived fragments (TDFs) were differentially expressed and when examined further most of the genes these fragments were identified as regulatory proteins with unknown roles. The microarray analysis identified that 99 of the 1040 TDFs were differentially expressed in the shoot and 56 more in the graft union. These included many transcription factors. Up-regulation of BAKI in the dwarf combination was found which is consistent with the results in apple on dwarfing rootstocks reported by Jensen et al. (2003) suggesting a possible role of brassinosteroid signaling in the dwarfing control of scions.

Following up on their earlier cDNA-AFLP work on rootstock related gene expression, Jensen et al. (2010) leveraged microarrays to examine expression of 55,230 apple transcripts with the ‘Gala’ scions on seven different rootstocks. RNA from actively growing shoots tips of young and mature plants were examined via microarray and the results identified 116 transcripts that were correlated to tree size. In this experiment, sorbitol dehydrogenases were expressed at higher levels in the larger trees as expected since apples utilize sorbitol as a transport molecule rather than sucrose as is seen in many other species. While the expression of SDH explained 34 percent of the tree size differences, these results contrast with the 2003 results where higher expression of SDH was linked to the dwarfing rootstock (Jensen et al., 2003). This suggests possible differences in the experimental design that may have changed the way these plants
responded, especially if the samples for RNA analysis were taken at different stages of growth between the two experiments which is unclear. Additionally, one third of the putative SDH transcripts in apple were similarly expressed though they did not meet the filtering criteria (Jensen et al., 2010). This suggests that it may be allelic expression differences that led to the up-regulation of one form seen from the dwarfing rootstock in the 2003 experiment. This group continued to expand on this experiment by testing a second microarray and identified 39 transcripts with expression levels correlating to fireblight resistance {Jensen, 2012 #3499}.

**Graft union transport: RNA:** Using tissue culture grown and grafted apples, Kanehira et al. (2010) sequenced cDNAs from the phloem cells. Analyzing these phloem cell cDNAs from non-grafted plants, they identified multiple sequences that were known to be expressed in phloem, especially metallothionin-like genes. In addition, an ABA biosynthetic gene and an auxin/aluminum responsive protein were captured in the phloem. In vitro grafting, or micro-grafting, of the apple scion ‘Fuji’ onto an interspecific rootstock allowed the expression of *gibberellin acid insensitive* (*gai*), a gene whose mRNA is known to be phloem transported. Due to sequence differences between the two grafted species, a rootstock specific allele was tested through nested RT-PCR. The results clearly show that the RNA from the rootstock allele is being transported across the graft union and into the phloem of the scion which was confirmed via in *situ* hybridization.

Likewise, Xu et al. (2010) inspected *gibberellic acid insensitive* (*gai*) mRNA expression in the roots and shoots of interspecific grafts of tissue cultured apple. As with Kanehira et al. (2010), the rootstock form of *gai* was found in the shoots. Additionally, the scion form of *gai* was identified in the roots. The transfer of these mRNAs was also seen on the fourth day after grafting while the mRNAs were not found in the xylem even at the 35th day after grafting.

Furthermore, Zhang et al. (2012) used the same approach to perform experiments in Chinese pear micro-grafted onto a wild pear rootstock. Their results are exactly as seen in apple with the mRNA of *gai* able to transfer across the graft union by the fourth day after grafting and no presence identified in the xylem. Additionally, they examined 2 year old grafted trees finding the transport of the rootstock *gai*
mRNA up to 40cm above the graft union. From the data provided, it appears that the 2 year old trees were grafted conventionally and not micro-grafted which would lead to an increased period of healing prior to transport of mRNAs across the graft union. It would be interesting to examine older trees and perform the same experiment to identify if there is a limit to the distance the mRNAs from the rootstock can travel in the scion or if this was an artifact of the slowly healing graft union in woody grafts. Leveraging the knowledge that *gai* is transported through the phloem, Zhang et al. (2012) furthered their experiments by testing the expression of NPTII and GAI:NPTII in transgenic tobacco. These transgenic lines were used as rootstocks and showed that the Pyrus *gai* transcript enabled the phloem based transport of NPTII into the scion while NPTII mRNA expressed without *gai* was not mobile.

RNA transport in composite plants was reviewed by Harada (2010) with a specific focus on RNA molecule transfer between the rootstock and scion. The experiments reviewed utilized Arabidopsis, tobacco, tomato, potato, melons, pumpkins and apple to demonstrate the long distance transport of RNA molecules, being mRNA, miRNA, or siRNA, through the graft union into the scion. Harada (2010) discusses the potential mechanisms for RNA molecules being moved through the phloem including RNA-binding proteins and chaperone proteins as well as some potential conserved motifs in the RNA sequences.

Kassai et al. (2011) have since shown that post-transcriptional gene silencing of scion genes by the rootstock is possible in tobacco, opening another regulatory window for rootstock scion interactions. They also showed much stronger results utilizing a phloem specific virus, commelina yellow mottle virus (CoYMV), promoter than when the cauliflower mosaic virus (CMV) promoter is used.

*Other:* Golecki et al. (1998) studied the ability of structural phloem proteins to be transmitted across the graft union of Cucurbits. The appearance of the proteins from the grafted individual depended partly on the graft method where more direct alignment of the phloem yielded faster translocation as expected. This transfer requires that the vasculature across the graft union to be repaired prior to the translocation of the phloem proteins. Some genotypes were identified as acceptors of the proteins while others were donors.
Stegemann and Bock (2009) were able to also show the transfer of either large portions of plastid DNA or entire plastids through the graft union of tobacco, though the transfer was only successful over several cell layers at the graft union. The transfer of DNA across the graft union, even if only a few cell layers, may produce unique cells in the composite plant that could result in the production of unique products and physiology.

**Summary:** These reports documenting differential scion gene expression, molecular transport, and biochemical activity start to reveal the foundation for rootstock scion interactions. Since many transcription factors have been shown to be modulated, large changes of global gene expression could easily result causing the physiological differences noted previously. The transport of RNA molecules, proteins and potentially plastids from the rootstock into the scion is a critical development in the understanding of these somatogenetic interactions. It remains to be shown how the expression of the transported RNA molecules in the scion can modulate global gene expression. Lack of localized tissue remains one of the challenges in unraveling molecular underpinnings of these somatogenetic interactions. Nodes, internodes, graft unions and growing shoot tips represent general tissues where the impacts of grafting are observable; however, it is not clear what cell types participate in receiving the signal from the rootstock and modify scion traits.

**CONCLUSIONS and FUTURE PROSPECTS:**

It is clear that our understanding of rootstock scion interactions and their underlying mechanisms have vastly improved in recent years. The control of scion traits by the rootstock has been highly documented in almost every area of perennial plant growth and physiology.

One major concern must be the evaluation of traits while the plants are not at the same stage of growth or maturity. Clearwater et al. (2007) showed in kiwi that the rootstock controls when sap flow begins which can lead to changes on in-season maturity. While common horticultural practice assumes that the fruit from a scion will be exactly the same on any rootstock, it is likely that the rootstock can affect harvest maturity and therefore all aspects related to harvest date. The role of the rootstock
controlling this trait in other species will provide critically important information to both the research and production communities.

While much evidence had suggested the role of the graft union as the critical point of rootstock scion interactions, the recent work by Gasco et al. (2007) along with concurring evidence from Clearwater et al. (2007), Nardini et al. (2006), Solari and DeJong (2006), and Solari et al. (2006) refutes the role of the graft union restricting nutrient and water flow to the scion after the graft union is fully healed. In the annual composite plants, the role of graft union healing may be more critical because the separation of the wound and healing response from the other characteristics is not possible. This is also seen in the micro-grafting experiments performed on the herbaceous growth of the perennial plants where the graft union heals and becomes functional much more rapidly than seen in typical wood based grafting techniques; however, hydraulic conductance experiments on micro-grafted plants need to be completed to prove this unequivocally. Examining these processes in perennials allowed the disjunction of the healing from the overall control and is a critical advantage of studying perennial grafted plants. From the differences in these works, the strict requirement for future research to document the time from grafting is critical to enable the studies to be compared.

Since the graft union is not causing the sustained differences in plant growth control, the clear effect of root hydraulic pressure is intriguing when considering vigor and other aspects of the composite plant. Higher amounts of water pressure would be expected to transport nutrients and hormones more effectively to the scion. Experiments either examining the natural correlation among these measures or manipulating the root pressure (Solari and DeJong, 2006) would reveal another layer of the complex interaction between these genotypes. Additionally, improved water content would significantly improve photosynthetic efficiency leading to better carbon fixation and faster growth rates. As with Prassinos et al. (2009), analysis of gene expression at different hydraulic pressures and different periods of growth will significantly improve the knowledge of the gene expression pathways involved in these interactions. Understanding the full relationship between water requirements, hydraulic conductance, and the effects
The recent findings that multiple types of RNA molecules are actively transported across the graft union provide a putative mechanism for the rootstock to control scion gene expression, and vice versa. However, it is unclear that the mere movement of RNA molecules in the phloem across the graft junction has any regulatory role. Higher-throughput gene expression studies utilizing RNA sequencing approaches may enable the identification of gene networks involved in the complex pathways controlled by the interaction of the genotypes. Examination of gene expression changes in micro-grafted plants from day 4 onward, as shown by Kanehira et al. (2010), Xu et al. (2010), and Zhang et al. (2012) to be the start of RNA transport into the scion, may help reveal some of the core genes and transcription factors that then trigger the larger, widespread alterations to physiology. Expression analyses can be used with highly variable rootstocks which may result in the ability to identify genes related to the unchanged traits thereby revealing critical genes for the physiology of each fruit.

The vast effects of rootstocks on scions may also be important to consider during breeding efforts for both rootstock and scions. Complementation of the combination may ‘rescue’ seemingly suboptimal genotypes into excellent production partners. This is true for both rootstocks and scions and the genetic information from DNA markers and DNA sequencing may be useful for predicting these optimal combinations.

Developing worldwide research standards will be critical as the field of rootstock scion interactions matures. These standards must include the age of the graft union due to its previously stated significance. Also, phenotyping fruit at specific maturities for each species would enable more direct comparisons across sample sites and among different studies. While sampling at specific maturities will place a larger burden on the researcher, the added value to the data obtained would be significant enough to warrant this change to research methodology. Another consideration for future projects is that most existing studies include one genotype of each vigor class whereas a full range of rootstocks may more effectively reveal useable differences. This approach would be similar to bulked segregant analysis where
expression levels of many genes are changed and few are consistent across the various rootstocks. This can help reduce the initial gene set for further investigations. In conclusion, the field of rootstock scion interactions in perennial and annual has made major headway yet many courses for further future discovery remain.
Works Cited:


Table 1. Recent representative reviews covering different aspects of rootstock scion interaction.

<table>
<thead>
<tr>
<th>Title</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A History of Grafting</td>
<td>(Mudge et al., 2009)</td>
</tr>
<tr>
<td>Vigour mechanisms in dwarfing rootstocks for temperate fruit trees</td>
<td>(Webster, 2004)</td>
</tr>
<tr>
<td>Physiological aspects of rootstock-scion interactions</td>
<td>(Martinez-Ballesta et al., 2010)</td>
</tr>
<tr>
<td>Hormonal signaling in rootstock-scion interactions</td>
<td>(Aloni et al., 2010)</td>
</tr>
<tr>
<td>Grafting and RNA transport via phloem tissue in horticultural plants</td>
<td>(Harada, 2010)</td>
</tr>
<tr>
<td>Grafting for disease resistance</td>
<td>(King et al., 2007)</td>
</tr>
</tbody>
</table>
Table 2. Physiological traits affected by rootstocks summarized in this review.

<table>
<thead>
<tr>
<th>Traits</th>
<th>Vigor/Tree Size</th>
<th>Yield</th>
<th>Quality</th>
<th>Disease resistance</th>
<th>Physiological disorders</th>
<th>Cold tolerance</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clementine</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Bassal, 2009)</td>
</tr>
<tr>
<td>Grape</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Koundouras et al., 2008)</td>
</tr>
<tr>
<td>Sweet cherry</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Santos et al., 2004)</td>
</tr>
<tr>
<td>Pistachio</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Gijon et al., 2010)</td>
</tr>
<tr>
<td>Peach</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>(Tspouridis and Thomidis, 2005)</td>
</tr>
<tr>
<td>Sweet cherry</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Whiting et al., 2005)</td>
</tr>
<tr>
<td>Sweet cherry</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Robinson et al., 2006)</td>
</tr>
<tr>
<td>Grape</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Soar et al., 2006)</td>
</tr>
<tr>
<td>Pear</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(North and Cook, 2008)</td>
</tr>
<tr>
<td>Rubber tree</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Cardinal et al., 2007)</td>
</tr>
<tr>
<td>Mango</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Smith et al., 2003)</td>
</tr>
<tr>
<td>Apricot</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Hernández et al., 2010)</td>
</tr>
<tr>
<td>Pear</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Bertelsen and Callesen, 2001)</td>
</tr>
<tr>
<td>Pear</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Sugar et al., 2005)</td>
</tr>
<tr>
<td>Grape</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Main et al., 2002)</td>
</tr>
<tr>
<td>Grape</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Koundouras et al., 2009)</td>
</tr>
<tr>
<td>Sweet cherry</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Gonçalves et al., 2006)</td>
</tr>
<tr>
<td>Grapefruit</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>(McCollum et al., 2002)</td>
</tr>
<tr>
<td>Tobacco</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Ruiz et al., 2005)</td>
</tr>
<tr>
<td>Kiwi</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Thorp et al., 2007)</td>
</tr>
<tr>
<td>Avocado</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>(Willingham et al., 2001)</td>
</tr>
<tr>
<td>Pear</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Seemüller et al., 2009)</td>
</tr>
<tr>
<td>Apple</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Jensen et al., 2003)</td>
</tr>
<tr>
<td>Annuals</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(King et al., 2007)</td>
</tr>
<tr>
<td>Cucurbita</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>(Edelstein et al., 1999)</td>
</tr>
<tr>
<td>Citrus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Ritenour et al., 2004)</td>
</tr>
<tr>
<td>Pistachio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>(Epstein et al., 2004)</td>
</tr>
<tr>
<td>Pecan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>(Smith et al., 2001)</td>
</tr>
</tbody>
</table>
Table 3. Physical mechanisms proposed to be underlying rootstock scion interactions.

<table>
<thead>
<tr>
<th>Species</th>
<th>Graft union repair</th>
<th>Water usage and transport</th>
<th>Hormones</th>
<th>Nutrition</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Flaishman et al., 2008)</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Yin et al., 2012)</td>
</tr>
<tr>
<td>Olive</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>(Gasco et al., 2007)</td>
</tr>
<tr>
<td>Peach</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Zhang et al., 2012)</td>
</tr>
<tr>
<td>Sweet cherry</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Olmstead et al., 2006a)</td>
</tr>
<tr>
<td>Sweet cherry</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Olmstead et al., 2006b)</td>
</tr>
<tr>
<td>Peach</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Tombesi et al., 2010)</td>
</tr>
<tr>
<td>Peach</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Basile et al., 2003)</td>
</tr>
<tr>
<td>Apple</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Atkinson et al., 2003)</td>
</tr>
<tr>
<td>Kiwi</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Clearwater et al., 2007)</td>
</tr>
<tr>
<td>Grape</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Koundouras et al., 2008)</td>
</tr>
<tr>
<td>Olives</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Nardini et al., 2006)</td>
</tr>
<tr>
<td>Peach</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Solari et al., 2006)</td>
</tr>
<tr>
<td>Peach</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Solari and DeJong, 2006)</td>
</tr>
<tr>
<td>Review</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Aloni et al., 2010)</td>
</tr>
<tr>
<td>Apple</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Kamboj et al., 1999a)</td>
</tr>
<tr>
<td>Apple</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Kamboj et al., 1999b)</td>
</tr>
<tr>
<td>Peach</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Soar et al., 2006)</td>
</tr>
<tr>
<td>Grape</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Tworkoski and Miller, 2007)</td>
</tr>
<tr>
<td>Apple</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Van Hooijdonk et al., 2010)</td>
</tr>
<tr>
<td>Apple</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>(Li et al., 2012)</td>
</tr>
<tr>
<td>Cotton</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Dong et al., 2008)</td>
</tr>
<tr>
<td>Peach</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Tsipouridis and Thomidis, 2005)</td>
</tr>
<tr>
<td>Apple</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Fallahi et al., 2001)</td>
</tr>
<tr>
<td>Mango</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Schmutz and Ludders, 1999)</td>
</tr>
<tr>
<td>Lemon</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Almansa et al., 2002)</td>
</tr>
<tr>
<td>Orange</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Papadakis et al., 2004)</td>
</tr>
<tr>
<td>Avocado</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Mickelbart and Arpaia, 2002)</td>
</tr>
<tr>
<td>Pear</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Okubo et al., 2000)</td>
</tr>
<tr>
<td>Tomato</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Martinez-Rodriguez et al., 2008)</td>
</tr>
<tr>
<td>Sweet cherry</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(Gonçalves et al., 2006)</td>
</tr>
<tr>
<td>Pear</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>(North and Cook, 2008)</td>
</tr>
</tbody>
</table>
Table 4. Molecular analyses performed to unravel rootstock scion interactions.

<table>
<thead>
<tr>
<th>Species</th>
<th>Biochemical Analysis</th>
<th>Expression Analysis</th>
<th>Graft union transport</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemon</td>
<td>X</td>
<td></td>
<td></td>
<td>(Almansa et al., 2002)</td>
</tr>
<tr>
<td>Rose</td>
<td>X</td>
<td></td>
<td></td>
<td>(Agbaria et al., 1998)</td>
</tr>
<tr>
<td>Apple</td>
<td>X</td>
<td></td>
<td></td>
<td>(Jensen et al., 2003)</td>
</tr>
<tr>
<td>Sweet cherry</td>
<td></td>
<td>X</td>
<td></td>
<td>(Prassinos et al., 2009)</td>
</tr>
<tr>
<td>Apple</td>
<td></td>
<td></td>
<td>X</td>
<td>(Jensen et al., 2010)</td>
</tr>
<tr>
<td>Apple</td>
<td></td>
<td></td>
<td>X</td>
<td>(Kanehira et al., 2010)</td>
</tr>
<tr>
<td>Apple</td>
<td></td>
<td></td>
<td>X</td>
<td>(Xu et al., 2010)</td>
</tr>
<tr>
<td>Pear</td>
<td></td>
<td></td>
<td>X</td>
<td>(Zhang et al., 2012)</td>
</tr>
<tr>
<td>Review</td>
<td></td>
<td></td>
<td>X</td>
<td>(Harada, 2010)</td>
</tr>
<tr>
<td>Tobacco</td>
<td></td>
<td></td>
<td>X</td>
<td>(Kasai et al., 2011)</td>
</tr>
<tr>
<td>Cucurbit</td>
<td></td>
<td></td>
<td>X</td>
<td>(Golecki et al., 1998)</td>
</tr>
<tr>
<td>Tobacco</td>
<td></td>
<td></td>
<td>X</td>
<td>(Stegemann and Bock, 2009)</td>
</tr>
</tbody>
</table>
Figure 1: Overview of review topics. On the top portion of the figure, major categories of rootstock-controlled scion traits are displayed in green boxes. Major physical and molecular factors involved in rootstock scion interactions are shown on the sides with the general direction of transport when provided in the text.
12.5. Impact of Advances in Sequencing Technology

12.5.1 Introduction

While non-sequencing based transcriptomic methods such as microarrays produce meaningful results, they require a priori knowledge of the genes being examined. As advances in sequencing technologies have improved genomics research since 2005, the field of transcriptomics has also been significantly enhanced. The transcriptomic methods previously discussed in this chapter rely either on model system data or sequences developed through Sanger sequencing. Since the advances discussed here are still recent, stone fruit research using the newer methods have yet to be published; therefore, this section focuses on the technological improvements pertaining to transcriptomic research and approaches in other species are presented to provide insight on the future direction of transcriptomics.

12.5.2 Sequencing Technologies and Their Applications for Transcriptomics

During the past decade, sequencing technologies have developed that far exceed the capabilities of Sanger sequencing for many applications. These new systems, or next generation sequencers, have recently been renamed second generation sequencers with the advent of third generation machines.

12.5.2.1 1st Generation

Sanger Sequencing: Incorporation of capillary electrophoresis in Sanger sequencing that uses the chain termination method was the first major advancement in technology. This enhancement reduced the number of reactions per sequence from 4 to 1 and enabled higher throughput than sequencing gels by running the DNA through a capillary tube yielding the flow gram and sequence output. Though the newer technologies have much higher throughput, the high length and quality of Sanger reads remain
unmatched. Transcriptomic approaches using Sanger sequencing are limited due to the necessity for individual reactions per gene being investigated.

Sequencing of cDNA libraries has been the primary method to develop ESTs and full-length cDNA sequences. When entire cDNA libraries are sequenced, most of the transcriptome from a tissue is identified. There are, however, limitations of cloning that may result in the loss of non-clonable genes or generate representation bias during colony picking for subsequent sequencing. Since genes have different levels of expression, it can be costly and time consuming to thoroughly investigate a library to the depth required for identification of genes expressed at low levels. This approach was employed on three types of peaches that resulted in the identification of 10,847 unique sequences or genes. This information was further used to identify 1,819 simple sequence repeats (SSRs) that could be used in molecular genetics studies in Prunus (Vecchietti et al. 2009). Another method based on Sanger sequencing is differential display (Liang and Pardee 1998) which helps to identify genes that are expressed differently among the tested samples. This method was used in stone fruits for unraveling Phytoplasma induced gene expression changes in Prunus armeniaca (Carginale et al. 2004).

12.5.2.2 2nd Generation

Second-generation sequencing (SGS) technologies started becoming available in 2005 and have initiated the bulk of the advancements in sequencing based transcriptomics. 454, Illumina, and SOLiD are the three original SGS technologies and each uses a sequencing by synthesis methodology. Though Ion Torrent is a much more recently developed system, the majority of the workflow and technology is very similar to 454 including sequencing by synthesis so it is included here as an SGS technology. The SGS technologies generally have the common themes of creating cDNA, PCR amplifying the cDNA, and sequencing by synthesis. Bias can be introduced during both the cDNA preparation and PCR amplification but have been generally discounted due to a lack of alternatives. This potential bias also increases the need for confirmatory qRT-PCR on any quantitative sequencing results pertaining to a given gene. The high-throughput of these SGS methods has led to easier deciphering of splice junctions, discovery of novel transcripts, quantitative analyses and identification of allelic variation. The methods
behind the technologies give each platform unique advantages and disadvantages when comparing their utility for transcriptome analysis and these are described below in order of decreasing sequence read length generated by individual platforms.

**454**: 454 sequencing is a pyrosequencing method where photo-capture of the light emitted as a result of the nucleotide addition is analyzed and converted into sequence. 1 million reads with an average length of 400 bases are being produced using the Titanium chemistry, and reads of ~800bp have recently been commercialized. Limitations in the light capture create potential issues with sequencing homopolymer (presence of 7 or more A, T, G or C) runs which can cause open reading frames to be disrupted if not resolved properly. The long read lengths of 454, >400bp, are useful for developing de novo transcriptome assemblies. The major drawback of using 454 for transcriptomic analyses is the relatively low number of reads produced per run.

One of the first uses of 454 for transcriptomics examined prostate cancer cell lines and found 10,000 genes expressed as predicted, 25 novel splicing events and an additional 10,000 novel ESTs with average reads of 102bp (Bainbridge et al. 2006). The same year, a method for using 454 to sequence the 5’ end of transcripts was developed as an improvement on SAGE based techniques and tested on maize (Gowda et al. 2006). In this experiment, multiple translation start sites were identified in the sequenced transcripts though only 34% of the sequences mapped back to the maize genome at the time of their analysis. Later, a quantitative transcriptome profiling method was developed by selectively sequencing the 3’ UTR (Eveland et al. 2008). In their work, Eveland et al. (2008) identified 14,822 transcripts and found expression differences between the wild type and their mutant as well as expression differences between alleles with up to 98% sequence identity in maize. During the past year, sequencing of the phloem RNAs from ash species yielded 58,673 ESTs of which only 55% had homology in the non-redundant GenBank database (Bai et al. 2011). This study also identified 1,272 SNPs and 980 SSRs of which 7 out of 25 tested SSR loci were polymorphic in the test species. Deep Super SAGE, an ultra-deep sequencing version of Super SAGE, was developed and used to examine chickpea gene expression after salt stress (Molina et al. 2011). This process identified 469 genes that were differentially regulated in both
roots and nodules of chickpea 2 hours after salt treatment. Using the 3’UTR strategy, 2243 SNPs were identified in sweet cherry with many being polymorphic in a screen of some major parental varieties (Koepke et al. 2012).

**Ion Torrent:** Similar to the 454 workflow, the Ion Torrent sequencing platform creates up to 5 million reads of 200 bp using pH detection on a semiconductor instead of an enzymatic cascade to produce light (http://www.iontorrent.com/). Transcriptome and small-RNA sequencing can be completed on this system and the smaller system allows individual projects to be run without requiring coordination with other labs to fill the larger output systems runs. Detailed transcriptome analysis protocols and citations are not yet readily available since the platform became publically accessible in December 2010.

**Illumina:** The newest Illumina platform, the High-Seq 2000, produces up to 6 billion reads of 100bp for paired end sequencing. A major difference between Illumina and Ion Torrent is the bridge PCR performed on the sequencing flow cell which yields clusters of similar sequences compared to the emulsion PCR for amplification with the other SGSs. This platform sequences by synthesis through many cycles of addition of one dye incorporated nucleotide, detection and cleavage of the dye. The transition from GAII to High-Seq 2000 has nearly increased the output 10-fold and further increases in read length are expected. This high sequencing depth typically yields very high confidence in the consensus sequences from Illumina data.

A quantitative method for sequencing cDNA to estimate gene expression was developed and first tested on human liver and kidney RNA samples and showed high correlation to a microarray on the same samples (Marioni et al. 2008). This method has been widely applied and was recently used to dissect transcriptome variations during berry development of grapes (Zenoni et al. 2010). RNA-seq of the grape berries revealed 6,695 of 17,324 expressed genes exhibit stage specific expression including stage specific alternate splicing in 210 genes. Zenoni et al.’s project also identified 85,870 SNPs in addition to the quantitative profile of gene expression for the berry stages tested. Illumina based RNA-seq has also been modified to specifically sequence miRNAs. One miRNA project found that 249 miRNAs, 236 previously known, are expressed in safflower leaf, seed, and petal and that 116-133 were differentially
expressed when comparing the tissues (Li et al. 2011a). An alternative to the traditional RNA to cDNA to sequencing paradigm of transcriptome sequencing is flowcell reverse-transcription sequencing, FRT-seq (Mamanova et al. 2010). As the name suggests, the RNA is bound to the flowcell and reverse transcribed into DNA on the flow cell instead of prior to binding. FRT-seq showed a stronger correlation between samples from the same library than the standard RNA-seq protocol.

**SOLiD:** The SOLiD platform creates a unique file format for the 700 million reads of up to 75 bp which can be obtained in a single run (http://www.appliedbiosystems.com/). Called sequencing by ligation, the ligation of an 8 base probe with the first two nucleotides dye labeled to the growing strand, detection of incorporation and the last 3 bases cleaved is cycled with 5 different primers to read each base with its 5’ and 3’ neighbor. Since each base in a read is sequenced twice, the accuracy is claimed to be higher than other systems.

A method for transcriptome sequencing, short quantitative random RNA sequencing (SQR-RNA seq), that identified 31-37% of transcripts were not from known exons or predicted genes (Cloonan et al. 2008). SOLiD sequencing of total small RNAs and immunoprecipitated small RNAs from human embryonic stem cells yielded the identification of 818 new miRNAs including many that were less conserved than the consensus miRNA pattern (Goff et al. 2009). For single cell transcriptomic investigations, a protocol was developed that amplifies the transcripts of a cell enough for use with SOLiD’s RNA-seq application (Tang et al. 2010). Comparing to microarrays, this method identified 5,260 more genes in the single cell and found 1,753 new splice sites. Additionally, 8-19% of genes known to have multiple isoforms were found to have multiple isoforms expressed in the oocyte or blastomere. These data also revealed that knockout lines of miRNA processing genes showed large expression differences in a majority of transcripts compared to the wild type.

**Combinations:** Since each system is unique in its strengths and weaknesses, combinations have been applied to supplement each other. Using Illumina and SOLiD, grape transcripts were examined and 401 transcripts showing C to U RNA editing events were identified with 28% of these having significant tissue specific expression differences (Picardi et al. 2010). Compared to the individual platform results,
the combined data included a lower percentage of false positives. Using both 454 and Illumina for non-overlapping samples, mitochondrial RNA expression was profiled in mosquito revealing species and stage specific differences but not sex specific changes in expression (Neira-Oviedo et al. 2011).

12.5.2.3 3rd Generation

All the previously available sequencing platforms rely on cDNA preparation and PCR amplification prior to sequencing. Both of these steps introduce biases in interpretation of data related to quantitative expression. Third generation sequencing (TGS) systems are based on single molecule sequencing. This allows the direct sequencing of RNA molecules, thus removing the cDNA preparation and amplification steps that could introduce bias into the results. These newest methods aim to use electron microscopy to visualize the actual molecular sequence but are not yet commercialized. The different TGS technologies are reviewed more generally by Shadt (2010) including how they compare to SGS.

PacBio: The PacBio system uses single-molecule real time (SMRT) technology to obtain a maximum of 750,000 reads of ~1kb in total length by anchoring the polymerase to a plate to visualize which dye incorporated bases are added (http://www.pacificbiosciences.com/). While direct RNA sequencing is still in development, single molecule sequencing of cDNAs will not prevent the RNA to cDNA errors but will eliminate the amplification step of SGS thereby reducing the potential for errors in the sequences. The potential for sequencing the RNA directly may reduce bias or loss during cDNA preparation. Research using the PacBio platform for transcriptomics has not yet been published.

Helicos: Helicos offers a ligation free, amplification free process to sequence cDNAs where up to billions of 25-35 bp reads are generated in a single run with this true single molecule sequencing (tSMS; http://www.helicosbio.com/). The technology relies on incorporating a fluorescently labeled nucleotide, detecting the addition through light capture, removing the label and repeating. This capture of light from single base additions eliminates the homopolymer problems seen with 454’s light capture. Removing bias from ligation and amplification is proposed to improve the quantitative prowess of this method for transcriptome profiling.
To perform transcriptomics on the Helicos machine, single-molecule sequencing direct gene expression (smsDGE) was used on first strand cDNA from yeast (Lipson et al. 2009). smsDGE demonstrated the ability to highly correlate to microarray data and identify both SNPs and splice variants. The next development for Helicos was direct RNA sequencing (DRS) as reported where the cDNA synthesis used for smsDGE was bypassed (Ozsolak et al. 2009). Later, low quantity RNA sequencing (LQ-RNA-seq) was tested on 400ng and 250pg of starting RNA and yielded very similar results between the samples (Ozsolak et al. 2010).

12.5.3 Conclusions

Overall, the sequencing advances of the past decade and the continued progress in sequencing technology are drastically changing the field of transcriptomics. As costs continue to diminish and throughput increases in the coming years, the ability for stone fruit researchers to conduct transcriptomic analyses will facilitate a new level of understanding for important biological processes in these species. With these advances, generating quality transcriptomic data will no longer be the bottleneck in the research pipeline, as shown in other systems. Data analysis, however, is progressing more slowly and each sequencing platform has its own error types that need to be handled separately.
Works Cited:


Comparative Genomics of Prunoideae


In Review at Plant Biotechnology Journal

The authors wish it to be known that the first two authors would like to be regarded as co-first authors.

Abstract

Prunus is an economically important genus with a wide range of physiological and biological variability. Using the peach genome as a reference, sequencing reads from four accessions of almond and one sweet cherry variety were used for comparative analysis of these three Prunus species. Reference mapping enabled the identification of many biological relevant polymorphisms within the individuals. Examining depth of polymorphisms and the overall scaffold coverage highlighted many regions of potential interest including hundreds of small scaffolds with no coverage from any individual. Of the polymorphisms, nonsense mutations account for about 70,000 of the identified single nucleotide polymorphisms (SNPs). Blast2GO analyses on these nonsense SNPs revealed several interesting results. First, nonsense SNPs were not evenly distributed across all gene ontologies. Specifically, sweet cherry is found to have nonsense SNPs in two 1-aminocyclopropane-1-carboxylate synthase (ACS) genes and two 1-aminocyclopropane-1-carboxylate oxidase (ACO) genes which may be at the evolutionary root of the non-climacteric ripening of sweet cherry.
Introduction:

Genetic and genomic diversity arises through multiple mechanisms including whole genome duplication, gene copy and transposable elements. However, when concerning closely related species, and even more varieties, single nucleotide polymorphisms (SNPs) play a large role in contributing genetic variation. These differences in closely related species, as with varieties, determine the phenotypic diversity observed in those species. While large scale rearrangements, duplications and deletions contribute to genetic changes, SNPs and insertion and deletions (indels) can have direct effects on gene function. SNPs and indels can be rapidly assessed through high-throughput sequencing and re-sequencing efforts and are becoming widely used as genetic markers in breeding programs (Ahmad et al., 2011; Ganal et al., 2009; Hyten et al., 2010; Kulheim et al., 2009).

While most previously identified polymorphisms have been the result of intraspecific analyses, the genetic changes contributing to the phenotypic variation across species of a genus are also of interest. Prunus, a diverse genus in the Rosaceae family with economic products from ornamentals, fruits, to the seeds, and wood based products, is a good candidate genus for this type of analysis. Prunus contains species that are diploid with n=x=8 and have estimated genome sizes between 225Mb and 300Mb (IPGI, 2010; Shulaev et al., 2008; Zhebentyayeva et al., 2008), relatively small for the Rosaceae family. This Rosaceae white paper also established peach as the reference genotype for Prunus due to the vast genomic resources available for peach including many ESTs, markers, and linkage maps (Zhebentyayeva et al., 2008). The recently completed draft genome sequence of peach is 220-230Mb (IPGI, 2010) though this was lower than the previous estimates suggesting the sequenced length of the related genomes may also be lower than estimated.

Peach, almond, and sweet cherry production was valued over 3.6 billion dollars in the US in 2010 (NASS, 2011) demonstrating the economic importance of this genus and the value of understanding the genomic structure of these species. Each of these are Rosaceous crops in the Prunus genus, each producing stone fruits, a perennial in growth habit, and a prolonged juvenility stage that has hindered the rate of progress of conventional breeding and understanding their genetic systems.
While these species are closely related, they have different traits that are important to production. In almond, the primary trait of interest is the difference between bitter and sweet almonds though flowering time and shell hardness are also important. Bitterness in almonds is driven by the production of amygdalin, a cyanogenic diglucoside and its degradation products benzaldehyde and cyanide (Sánchez-Pérez et al., 2012; Sánchez-Pérez et al., 2008). This trait has been found to be controlled by a single, dominant gene called *Sweet kernel (Sk)* that produces sweet almonds (Dicenta and Garcia, 1993). SSR markers have placed *Sk* on linkage group 5 of the “R1000” and “Desmayo Largueta” almond maps (Sánchez-Pérez et al., 2010). The position of these SSRs on the almond “Texas” x peach “Earlygold”, or (TxE) place the *Sk* locus between 11Mb and 14.6Mb on peach scaffold 5 (Jung et al., 2008). In sweet cherry, polymorphisms are desired for the development of a linkage map to enable identification of useful DNA markers and placement of these markers on the current peach reference genome. Several targets for DNA markers in sweet cherries are: size, firmness, pedicel-fruit retention force (PFRF), and powdery mildew resistance. Since the peach genome is available and intra-specific polymorphism analyses have already been concluded in peach (Ahmad et al., 2011), this work focuses on the genomic differences of almond and cherry in respect to peach.

Here, a reference mapping approach using the peach genome v1.0 as the reference and high-throughput sequencing from four almond genotypes and one sweet cherry genotype were used to identify regions of increased and decreased conservation in *Prunus*. Detailed analysis of SNPs and indels was completed to build a resource for future inquiries into these species. Additionally, preliminary analysis of the *Sk* locus in almond was completed identifying 238 SNP candidates for the *Sk* gene. The collective polymorphism dataset provides several regions of interest that have lower polymorphism rates and may be essential to the shared characteristics of these *Prunus* species.

**Results:**

**Sequencing data acquisition:** Four almond genotypes were chosen for sequencing including 2 sweet varieties, Ramillete and Lauranne, and 2 bitter selections, D05-187 and S3067. Shotgun sequencing of these four almond genotypes produced 142 million 76-base Illumina reads. Each of the individual
almond genotypes was sequenced at 8-13x coverage, or 2.1-3.3 Gb of sequence, and combined to yield a 10.8Gb dataset or 43x coverage (Table1). The sweet cherry cultivar Stella was chosen for genomic sequencing, and through 454 single-end reads, 454 paired-end reads, and Illumina paired-end reads, 1.6 Gb of sequence or roughly 7x coverage was acquired. Transcriptome sequencing of sweet cherry produced an additional 460Mb of sequence of single end 454 reads from Bing and Rainier cultivars.

**Assembly:** A reference based assembly of the reads onto v1.0 of the peach genome (IPGI, 2010) was completed to identify regions of conservation and divergence in the *Prunus* genus. 56% of the combined Illumina reads mapped to the peach genome while 44% of the total Illumina reads were mapped to the peach chloroplast, and 0.2% were not mapped to either. The eight primary scaffolds of peach were covered between 0.4 and 6.3x as shown in ‘3A:Additional File 1 Coverage Report’ which contains the coverage statistics for each scaffold and data set. These scaffolds were covered an average of 4.94x for the combined cherry data and a 3.14x average for the almond genotypes. Overall, 162 of the 334 scaffolds contained zero reads from cherry or almond while an additional 24 were not mapped by the cherry data. Also, mapping data show that 96-99% of peach genes were mapped to with these datasets (Table 2).

**Polymorphism Analyses:** Overall, 13,126,567 initial polymorphisms were identified between each individual genotype and peach. Potential polymorphisms were initially identified and parsed to 9,751,035 after filtering to retain only sites with at least 3 reads supporting the difference as previously described by (Deschamps and Campbell, 2010; Hyten et al., 2010; Koepke et al., 2012; Kulheim et al., 2009). These polymorphisms were then further identified based on their position to identify 6,138,404 total polymorphic sites.

**Polymorphism type and region identification:** Based on the reference genome annotations, the polymorphisms passing the filtering criteria were classified by their location (Table 3) yielding an average of 260,000 in the CDS for the almond cultivars and greater than 300,000 in sweet cherry. The exon based polymorphisms in the almond cultivars average 52.1\% (155,010), 43.3\% (128,778), and 4.5\% (13,454) for sense, missense, and nonsense mutations respectively (Table 4). Additionally, 0.1\% (342) of the CDS
SNPs are read-through mutations, mutations modifying a stop codon into an amino acid yielding C terminus extension (Zirn et al., 2005). Sweet 1, however, had a much higher rate of nonsense mutations at 10.6% while the other 3 varieties were 2-3% non-sense mutations. The insertions and deletions in the exons averaged ~3,000 each for the four almond genotypes. In the cherry genomic dataset, exonic SNPs were 50.4% (162,662) sense, 42.8% (137,976) missense, 6.6% (21,234) nonsense, and 0.1% (335) read-through mutations along with 16,155 indels (Table 4).

Polymorphism depth analyses: The passed filtering dataset was also used to analyze the occurrence patterns of the polymorphisms. For scaffold 1 (3A:Additional File 2 scaffold depth images), it is clear that there are several regions of interest containing significantly more or less than the average number of polymorphic sites. Similar mapping of the number of genes in these regions of the peach scaffold reveal low gene density regions with high polymorphism rates. Statistical analyses reveal 346 sections that significantly differ from the mean number of polymorphisms in each 50kb region on each individual scaffold (3A:Additional File 3 significant regions). 95 of these section combine to make 31 regions are greater than 100kb in length with the longest region containing significantly higher polymorphisms being a 600kb block in almond from 20.45Mb-30Mb on scaffold 1. This region in cherry contains two 50kb blocks and one 100kb block that are also significantly higher in polymorphism rate. These genomic regions potentially lie at the heart of divergence from other members of Prunoideae.

Analysis of Sk locus: Further filtering of the almond polymorphisms around the Sk locus was completed to identify putative candidates for the Sk gene and causative mutation for the bitter/sweet phenotype. Using the BPPCT017 (11Mb) and BPPCT038 (14.6Mb) markers flanking the Sk locus as the bounds reduced the 311,497 polymorphisms identified on scaffold 5 to 56,155 between the markers (Table 5). Subsequent reduction of this dataset was completed by removing polymorphisms that were not homozygous in both sweets and within both bitter cultivars. Also, the homozygous polymorphisms were required to be different between the sweets and bitters yielding 6,304 polymorphisms of which 228 caused codon changing mutations. These missense, nonsense, and read-through SNPs as well as the indels comprise the reduced set of putative candidates for future screening and analysis.
**Blast2GO Global Analysis**: A global comparison of putative nonsense mutations within cherry and the four selected cultivars of almond reveal a similar distribution of mutations across various gene ontologies. This can be seen in gene ontologies relating to biological process, molecular function as well as cellular component (Figure 1). Response to stress, protein modification process, catabolic process, and transport each comprised at least 10% each of the total biological process gene ontologies for each tested dataset. With respect to molecular function, over 35% of annotated genes containing nonsense SNPs are involved in nucleotide binding with around 15% having kinase activity and slightly under 15% having DNA binding activity. Finally, with respect to cellular component, about 25% of all annotated genes were predicted to be localized to the plastid, with both the mitochondrion and plasma membrane comprising 15% of all annotated genes. As there appeared to be little variation in the GO term composition of the five datasets Blast2GO analysis of the entire peach gene set was performed and compared to datasets mapping back to nonsense-SNPs. A chi-square test revealed that several GO-terms have statistically higher or lower GO terms than predicted (3A: Additional File 4 Chi Table). Nonsense mutations map back to a total of 133 unique KEGG pathways, with Bitter 1 mapping to 121, Bitter 2 to 119, Sweet 1 to 124, Sweet 2 to 127, and Cherry to 127 KEGG pathways, respectively (3A: Additional File 5 GlobalKEGGComparison). The cherry nonsense SNP-containing dataset contains members participating in atrazine degradation, chlorocyclohexane and chlorobenzene degradation, fluorobenzoate degradation, synthesis and degradation of ketone bodies, toluene degradation while none of the investigated almond cultivars did. Conversely, all four almond cultivars contained predicted nonsense mutations within genes involved in butirosin and neomycin biosynthesis, D-alanine metabolism, D-arginine and D-ornithine metabolism, and glucosinolate biosynthesis all of which lacked participating genes with putative nonsense mutations in cherry.

Comparison of the nonsense-containing genes within the five data sets reveals that a large subset of the genes, 1,191 in total, is shared between all members (Figure 2). Additionally, some nonsense SNPs are unique to individual samples. The largest of these sets, 2,535 genes, are the nonsense mutations unique to
cherry. 1,276 genes containing putative nonsense SNPs are present within each individual almond cultivar and absent from the cherry analysis.

**Blast2GO Targeted Pathway Analysis:** The most abundant biological process gene ontology represented in the datasets, “Response to stress” was selected as a GO of interest to further investigate. Further breakdown of this category reveals that its members are involved in a total of 92 KEGG pathways within the 5 investigated datasets (3A: Additional File 6 Response To Stress). While all datasets contain genes with putative nonsense SNPs in numerous pathways, only cherry contains putative nonsense SNPs related to stress in C5-branched dibasic acid metabolism, chlorocyclohexane and chlorobenzene degradation, indole alkaloid biosynthesis, isoquinoline alkaloid biosynthesis, naphthalene biosynthesis, N-glycan biosynthesis, nicotinate and nicotinamide metabolism, primary bile acid biosynthesis, retinol metabolism, steroid degradation, steroid hormone biosynthesis, toluene degradation, tropane, piperidine and pyridine alkaloid biosynthesis and valine, leucine and isoleucine biosynthesis. Alternately, all four cultivars of almond contain potential nonsense SNPs in alanine, aspartate and glutamate metabolism, benzoate degradation, caprolactam degradation, fatty acid elongation, geraniol degradation, monoterpenoid biosynthesis, sulfur metabolism, and vitamin B6 metabolism while cherry lacks nonsense mutation in these pathways.

Further investigation into nonsense mutations present within members of the cyanogenamino acid metabolism was performed as this pathway leads to amygdalin synthesis and catabolism. Blast2GO analysis performed through searching for the keywords “Prunasin” and “Amygdalin” revealed the presence of 4 isoforms of peach prunasin beta-glucosidase and amygdalin beta-glucosidase with nonsense mutations in cherry (ppa003891m, ppa016583, ppa003856m, and ppa003831m), one in Bitter 1 (ppa003831m), three in Bitter 2 (ppa003856m, ppa003891m, ppa003831m), four in Sweet 1 (ppa003891m, ppa016583, ppa003856m, and ppa003831m), and four in Sweet 2 (ppa003891m, ppa016583, ppa003856m, and ppa003831m). Based upon annotation, numerous other members within this pathway contained putative nonsense mutations and additional members with potential prunasin beta-glucosidase or amygdalin beta-glucosidase activity were detected (Figure 3 and Table 6).
These data were further probed to identify differences in nonsense SNPs involved in ripening, another biological process of interest with respect to the *Prunus* genus. Within the 25 genes in all peach sequences mapping to the ripening GO term, cherry surprisingly contains 8 which are predicted to contain putative nonsense SNPs (two putative multidrug resistance genes, two 1-aminocyclopropane-1-carboxylate synthase (ACS) genes, two 1-aminocyclopropane-1-carboxylate oxidase (ACO) genes, one polygalacturonase gene and one glycosyl hydrolase family 9 protein gene) (Figure 5). Bitter 1 and Bitter 2 almond datasets each contain two nonsense SNP-containing genes (one ACO gene and one putative multidrug resistance gene) while Sweet 1 and Sweet 2 almonds each contain three (two ACO gene and one putative multidrug resistance gene). Further analysis of genes with supposed SNP in the KEGG pathway “cysteine and methionine biosynthesis” reveals additional nonsense SNPs that may affect ethylene synthesis, but are not mapped to the ripening GO term (Figure 4 and Table 7).

The final biological process of interest, abscission, has very few members containing nonsense mutations compared to the other gene ontologies, in fact only 23 genes in the entire peach gene set map to this term. Sweet 1 and Sweet 2 almonds are predicted to have a single gene related to abscission containing a nonsense SNP in a gene encoding btb poz ankyrin repeat protein. Bitter 1 almond sequences had no nonsense mutations in any predicted gene sequences, while Bitter 2 almond has a nonsense mutation in a gene encoding a probable adp-ribosylation factor gtpase-activating protein (agd5-like). Cherry has nonsense SNPs present both in the gene encoding adp-ribosylation factor gtpase-activating protein (agd5-like) as well as the btb poz ankyrin repeat protein.

**Discussion:**

**Uneven distribution of sequencing reads:** Reference based assemblies are built upon the presumption that the sequenced genomes are highly similar to the reference. When differences exceed the threshold of the mapping software, reads from these highly divergent regions are not mapped. 7.5% of reads from *indica* rice cultivars did not map to the Nipponbare rice genome (Subbaiyan et al., 2012). In our data, 0.2% of the Illumina reads were unmapped to the peach genome strengthening the assertions that peach is a credible reference genome for *Prunus*.
In addition to unmapped reads, the data produced in this work identified 162 and 186 scaffolds of peach that were not covered by any reads from almond and sweet cherry respectively. One explanation for this is that these smaller, un-anchored scaffolds may be unique to the peach genome. Alternatively, as the peach genome was built using Sanger sequencing, these could be repetitive regions and the shorter reads used here were placed in the dual location during the mapping. In either case, these regions provide insights into genome structure differences that need to be further evaluated to fully understand the differences among these Prunus species.

**Polymorphism Analyses:** Our analyses show 48% of the mutations in the CDS to be non-synonymous in almond and 50% in sweet cherry. This is comparable to the 57% found in rice cultivars (Subbaiyan et al., 2012), however it is interesting that this interspecies comparison identified a higher percentage of synonymous SNPs than the intraspecific comparison in rice. While it is possible that the heterozygous nature of the almond and sweet cherry cultivars caused polymorphisms to be screened out during filtering, it is unlikely that this would have significantly shifted the representation of synonymous and non-synonymous mutations.

It is important to note that the read-through mutations could be discussed as non-sense mutations of the almond gene in peach; therefore, discussion of read-through and non-sense mutations is limited by the perspective of the analysis which, in this case, is in respect to the peach reference genome. At first glance, the 0.1% generation rate of read-through mutations suggests that these mutations may be highly deleterious with strong selection against them as they occur at ~1/50th of the rate that non-sense mutations arise. A closer examination, however, reveals that while the probability of a stop codon mutation causing a read-through mutation is 85%, there is only one stop codon per protein. This contrasts significantly with the 4.2% chance of a random SNP causing a nonsense mutation multiplied by the 403 amino acids found in the average CDS in the peach genome. Calculating for the distribution of amino acids yields one polymorphism having a 4.18% or 0.21% chance of causing a nonsense or read-through mutation respectively in the average peach gene. The data from this work show a 2.5 fold change from expected
providing intrigue but requiring further evaluation regarding the effect of these mutations on gene function.

346 regions with higher and lower rates of polymorphism were identified in this work. Higher rates could result from genomic duplications or from low conservation yielding more divergence. Similarly, regions with lower than average polymorphisms could be the result of either low divergence where few polymorphisms arose, or of very high amounts of differentiation preventing the mapping of the sequencing reads to these locations. Subbaiyan (2012) revealed similar regions of lower polymorphism rates in six inbred lines of rice with several being greater 100kb in length. The 600kb region in almond is particularly interesting as it may represent a larger region of diversion between almond and peach and may contain genes related to the nature of these two species.

**Analysis of Sk locus:** The combination of the existing DNA markers, the reference sequence, and genotype specific sequencing yielded 228 candidate mutations for the sweet kernel trait in almond. Since this work was completed using only 2 bitter and 2 sweet genotypes, reductions in this candidate set would be expected if more genotypes were examined. However, whole genome sequencing of further genotypes is not necessary at this time as site specific testing of the genotypes for the identified mutations are expected to identify the allele responsible for the difference between these types of almonds. As the major and highly critical trait, developing a gene based marker for the sweet kernel gene will provide a drastic benefit to the almond community by rapidly identifying the undesirable bitter genotypes. As suggested by Michelmore (1991), bulked segregant analysis can function in an obligate outcrossing species. The results shown here demonstrate the ability of the approach to produce a small candidate list from a large region of interest. Adding more individuals to the bulks in this work would allow the marker placement to be independently confirmed as well, though using 2 individuals of each phenotype was possible due to the previously developed markers for the sweet kernel locus.

**Blast2GO comparisons:** The global distribution of GO terms within the nonsense SNP-containing genes was similar among all samples tested. This suggests two potential options regarding the presence of nonsense SNPs: 1.) Certain gene ontologies have accrued nonsense SNPs at similar levels
across species in *Prunus* or 2.) Nonsense SNPs simply occur randomly throughout the genome and each gene ontology contains similar numbers of genes among the samples investigated. In order to assess this, comparison of the observed number of members for each GO term was performed against expected values generated from the entire peach predicted gene set using a chi-square test. This test displayed that numerous GO categories contained statistically significant higher as well as lower numbers of nonsense SNPs than expected (3A: Additional File 4 Chi Table). This suggests that many GO terms are linked to an increased likelihood to generate nonsense SNPs in *Prunus*, while others GO terms appear to be more conserved in the genus supporting option 1 above. Interestingly, the GO terms associated with significantly higher nonsense SNPs (p-value < 1E-10) include the biological processes “DNA metabolic process” (GO:0006259), “cellular protein modification process”(GO:0006464), “signal transduction” (GO:0007165), and “pollen-pistil interaction” (GO:0009875); the cellular components “mitochondrion”(GO:0005739), “cytoskeleton” (GO:0005856), and “plastid” (GO:0009536); and the molecular functions “nucleotide binding” (GO:0000166) and “kinase activity” (GO:0016301). GO terms associated with significantly lower nonsense SNPs (p-value < 1E-10) include the biological processes “response to biotic stimulus” (GO:0009607), “response to abiotic stimulus” (GO:0009628), “anatomical structure morphogenesis” (GO:0009653 )and “response to endogenous stimulus” (GO:0009719); the cellular component “cytosol” (GO:0005829); and the molecular functions “chromatin binding” (GO:0003682), “sequence-specific DNA binding transcription factor activity” (GO:0003700) and “structural molecule activity” (GO:0005198). While a connection between GO term and occurrence of nonsense SNPs appears to exist, this does not disprove the option 2 stated above.

Concerning the GO term “response to stress”, there appears to be significant genetic variability with respect to nonsense SNPs. In fact, sequences containing detected nonsense SNPs mapped to this GO term more than any other GO term investigated in the biological process domain. This gene ontology is of high agricultural importance as breeding and genetic modification of plants resistant to both biotic and abiotic stresses is a large focus in both industry and academia. Previous studies have used gene-based SNPs detected through interspecific comparisons to identify, verify and attach function to SNPs which
may be involved in stress response (Parida et al., 2012). These putative nonsense SNPs represent a preliminary dataset within *Prunus* which may be used in similar studies.

Basic differences exist in the ripening patterns of members of the *Prunus* genus. Peach, apricot, and plum fruits are climacteric, meaning that a burst of ethylene occurs quickly followed by an increase in respiration production. Cherry and almond, on the other hand exhibit non-climacteric ripening, an outlier in the genus. The identification of non-sense mutations in several versions of ACS and ACO could significantly damage the ethylene production pathway in cherry rendering it nearly unable to provide the burst seen in other fruits in this genus. While cherry is non-climacteric, the cherry fruit color and maturation is modulated by application of exogenous ethylene (Koepke and Dhingra unpublished).

While these results enable the identification of targets for gene linked marker screening, it is important to realize the limitations of this project. First of all, nonsense SNPs do not necessarily equate to loss of function of a protein. Additionally, as these sequences were aligned to a predicted peach data set, the true sequence of genes of interest may be biased. Potential splice variants may have the ‘non-sense’ mutation in an exon that is not utilized in these species. Also, the presence of a single non-sense mutation may not be deleterious at all and could be sufficiently complemented by the other allele especially in a genus where very few self-compatible varieties exist leading to high amounts of heterosis. Gene duplications or those genes unique to almond or cherry may not be represented in these data; alternatively, they may be represented as SNPs while they are actually different alleles.

**Conclusions:**

Using reference based assemblies of four almond genotypes and one sweet cherry variety, we were able to begin interspecific comparative genomic analysis of Prunoideae. Over 99 percent of the raw reads mapped to the peach genome though nearly 44 percent mapped to the chloroplast. Identifying hundreds of smaller scaffolds in the peach genome that were not mapped to by either the almond or sweet cherry data identifies many potentially peach-specific regions of interest for further investigation. The 6.1 million putative SNPs provide a resource for gene based investigations. While many of the SNPs and indels are in non-coding regions, 250 to 300 thousand SNPs are located in the coding regions of annotated
peach genes. These SNPs should prove to be useful in expanding our knowledge of genetics and genomics in these species through their use as molecular markers and gene based interrogations. The coverage depth images revealed 31 regions that have significantly different amounts of SNPs.

A keystone goal of genomics is to identify genes responsible for specific traits. Here, we examined the bitterness trait of almond and identified 228 codon-changing mutations near the previously identified Sk locus. Additionally, we provide the first report in plants of nonsense SNP abundance in a genus being linked to specific GO terms. A global analysis of SNPs has also revealed several candidate mutations of interest for different physiological properties of these species including response to stress, ripening and abscission. Combined, these data should provide a foundation for further genomics and genetics research in Prunoideae.

Methods:

**Sequencing data acquisition:** *Almond* – D05-187 (Bitter1) and S3067 (Bitter2) are homozygous bitter selections from the CEBAS-CSIC and Ramillete (Sweet1) and Lauranne (Sweet2) are each homozygous sweet cultivars of almond. Approximately 10x coverage was obtained for each of the four genotypes with 76bp Illumina paired-end reads.

*Cherry* – The sweet cherry genome project has developed roughly 7x coverage of Stella, an important parental cultivar. These data were derived mostly through single-end 454 with some paired-end 454 and Illumina paired-end sequencing. Both 454 GS-FLX and 454 GS-FLX+ versions were used to acquire these sequences. Also, 454 transcriptome data from Bing and Rainier cultivars of sweet cherry were obtained and used in the analyses.

*Peach Genome* – The peach genome version 1.0 was obtained from phytozome.net for use as the reference sequence throughout this project. The chloroplast and mitochondrial genomes were excluded from the assembly initially and the chloroplast was later used to screen the unassembled reads.

**Assembly:** A reference based assemblies of both the cherry genomic 454 and cherry transcriptomic 454 data was assembled using the NGen assembler (DNASTar) version 3.1.0 with the peach genome version 1.0 as the reference and using the following 454 default parameters: mersize =21,
merSkipQuery = 3, minMatchPercent = 85, MaxGap = 15, minAlignedLength = 50. Similarly, all Illumina data from the 4 almond genotypes and sweet cherry were assembled using the peach genome as a reference with the Illumina default parameters: mersize = 21, minMatchPercent = 93, mismatchPenalty = 20, MaxGap = 6, minAlignedLength = 35. For each assembly, the different genotypes were input separately to enable the unique SNP information to be attained for each individual.

**Polymorphism Analyses**—Assembled data were imported into SeqMan (DNAStar) where SNP reports were created. A custom script was used to remove polymorphisms with less than 3 reads confirming each non-reference call similar to previous SNP reporting works (Deschamps and Campbell, 2010; Hyten et al., 2010; Koepke et al., 2012; Kulheim et al., 2009). These filtered SNPs were then imported into ArrayStar (DNAStar) to enable further analyses.

**Polymorphism type and region identification**: Custom computational comparisons of the base calls from the sequenced individuals against the peach genome were completed to determine the base changes involved. Similarly, polymorphism regions were identified by analyzing the reference position against the annotation of the peach genome. These SNPs were classified as 5' UTR, intron, exon, 3'UTR or intergenic. Exonic polymorphisms were further classified as sense, nonsense, missense or read-through mutations based on the resulting amino acid compared to the peach genome annotation. Read-through mutations were defined as the SNPs causing a stop codon to be changed into an amino acid thereby elongating the C terminus of the protein with respect to the peach gene (Zirn et al., 2005).

**Polymorphism depth analyses**: To visualize the depth of the polymorphisms across the 8 main scaffolds of the peach reference, the total polymorphisms in each discreet 50kb window were analyzed and displayed as a single pixel wide bar one pixel high for each 20 polymorphisms. The graphs for each individual were then compiled into a single image per scaffold. The composite polymorphism set, where each unique SNP was counted once for each species, was also analyzed in this manner. The distribution of polymorphism counts per 50kb window was analyzed to identify regions of the peach reference that had a polymorphism depth greater than 2 standard deviations from the mean of that scaffold.
**Analysis of Sk locus:** The total almond SNP report was filtered to retain only the polymorphic sites near the Sk locus. Since the markers BPPCT017 and BPPCT038 are located at ~11Mb and 14.6Mb on peach linkage group 5 respectively, they were used for the bounds around the Sk locus. All polymorphisms that were conserved within a group but contrasting between the two types were retained as both bitter and both sweet genotypes are homozygous for the trait. Further screening reduced the data set to only contain codon changing polymorphisms that make up the candidate gene set.

**Blast2GO comparisons:** Nucleotide sequences for all predicted *Prunus persica* genes containing nonsense mutations in sweet almond, bitter almond, cherry transcriptome, and cherry genome contigs were imported into Blast2GO (Conesa et al., 2005; Gotz et al., 2008). BLASTX was performed against the nr database for all sequences using an ExpectValue cutoff of 1.0E-3 and a HSP length cutoff of 33. After GO-Mapping, annotation using was performed using an E-Value-Hit-Filter of 1.0E-6, Annotation CutOff of 55, GO Weight of 5, and an Hsp-Hit Coverage CutOff of 0. Blast2GO default evidence code weights were used. Following GO annotation, an Interpro scan (Quevillon et al., 2005) was performed and results were merged to the GO annotations. Annotation augmentation was performed using ANNEX (Myhre et al., 2006), followed by GO-slim with the goslim_plant.obo database. As coverage of peach genes was low in cherry transcriptome and cherry genome, all cherry data were combined into a single Blast2GO file and were analyzed as a group entitled “Cherry”.

Analyses were performed at a global level to identify Gene Ontology information for biological processes, cellular components, and molecular functions for all predicted peach genes predicted to have a nonsense mutation. A Blast2GO analysis was also performed on the entire peach gene dataset. These results were used to determine expected GO values for each dataset investigated. A ratio of the number of members mapping to a particular GO term to the total number of members was calculated. This ratio was multiplied by the total number of GO terms per sample investigated to determine expected GO-term value. A chi-square test was performed to determine if the observed GO distribution of nonsense SNP-containing genes was significantly different from the expected. Custom scripts were used to compare datasets to determine which contained unique or shared entries. Finally, KEGG pathway maps and
corresponding information were downloaded from the KEGG Pathway Database through Blast2GO (http://www.genome.jp/kegg/pathway.html) (Kanehisa, 2002; Kanehisa et al., 2012).

**Author contributions:**

RS, FM, BM and RH initiated the project, collected samples, and performed DNA sequencing. TK, SS, and AD designed the analyses. TK and AH completed the reference mapping, mutation analyses and analysis of the Sk locus. SS performed the BLAST2GO analysis and processing. TK and SS performed statistical analyses. SS, TK, and AD wrote the initial manuscript. All authors approved the final manuscript.
Works Cited:


Table 1. Raw sequencing data. Total data acquired for each species.

<table>
<thead>
<tr>
<th></th>
<th>Cherry</th>
<th>Almond</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Transcriptome</td>
<td>454</td>
</tr>
<tr>
<td>Sequences:</td>
<td>1225030</td>
<td>3742780</td>
</tr>
<tr>
<td>Total Bases:</td>
<td>467371273</td>
<td>1020571812</td>
</tr>
<tr>
<td>Mean:</td>
<td>381</td>
<td>272.6774782</td>
</tr>
</tbody>
</table>
Table 2. Nonsense mutation analysis statistics

<table>
<thead>
<tr>
<th></th>
<th>Cherry</th>
<th>Bitter 1 Almond</th>
<th>Bitter 2 Almond</th>
<th>Sweet 1 Almond</th>
<th>Sweet 2 Almond</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Peach Genes Mapped Against</td>
<td>27576 (96.20%)</td>
<td>28332 (99.33%)</td>
<td>28420 (99.64%)</td>
<td>28488 (99.38%)</td>
<td>27590 (96.73%)</td>
</tr>
<tr>
<td>Total Genes with Predicted Nonsense Mutation(s)</td>
<td>5384</td>
<td>4016</td>
<td>5110</td>
<td>5302</td>
<td>5529</td>
</tr>
<tr>
<td>Unique to Cultivar With GO info</td>
<td>2535</td>
<td>190</td>
<td>409</td>
<td>467</td>
<td>624</td>
</tr>
</tbody>
</table>
Table 3. Classifications of polymorphisms identified in each data set based on their location relative to peach genes.

<table>
<thead>
<tr>
<th></th>
<th>Cherry Transcripts</th>
<th>Cherry Illumina</th>
<th>Cherry 454</th>
<th>Almond Sweet1</th>
<th>Almond Sweet2</th>
<th>Almond Bitter1</th>
<th>Almond Bitter2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Total</td>
<td>247818</td>
<td>39338</td>
<td>701534</td>
<td>678587</td>
<td>774780</td>
<td>483905</td>
<td>646641</td>
</tr>
<tr>
<td>mRNA Total</td>
<td>266340</td>
<td>40978</td>
<td>739110</td>
<td>707564</td>
<td>808159</td>
<td>504028</td>
<td>673423</td>
</tr>
<tr>
<td>CDS Total</td>
<td>194279</td>
<td>22881</td>
<td>305314</td>
<td>274904</td>
<td>290793</td>
<td>207911</td>
<td>266252</td>
</tr>
<tr>
<td>3' Total</td>
<td>32317</td>
<td>1083</td>
<td>42486</td>
<td>26924</td>
<td>30330</td>
<td>19015</td>
<td>25398</td>
</tr>
<tr>
<td>5' Total</td>
<td>10788</td>
<td>402</td>
<td>13967</td>
<td>13309</td>
<td>14687</td>
<td>10102</td>
<td>12891</td>
</tr>
<tr>
<td>Intergenic</td>
<td>43137</td>
<td>58654</td>
<td>797699</td>
<td>1789323</td>
<td>2263108</td>
<td>1208734</td>
<td>1734295</td>
</tr>
</tbody>
</table>
Table 4. Classification of polymorphisms in coding regions. Read-through mutations are stop codons in the peach reference that have been mutated into non-stop codons.

<table>
<thead>
<tr>
<th></th>
<th>Cherry Transcripts</th>
<th>Cherry Illumina</th>
<th>Cherry 454</th>
<th>Almond Sweet1</th>
<th>Almond Sweet2</th>
<th>Almond Bitter1</th>
<th>Almond Bitter2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sense</td>
<td>99947</td>
<td>12702</td>
<td>149960</td>
<td>132747</td>
<td>156693</td>
<td>196849</td>
<td>133753</td>
</tr>
<tr>
<td>Mis-sense</td>
<td>80131</td>
<td>10688</td>
<td>127288</td>
<td>118544</td>
<td>151659</td>
<td>106056</td>
<td>138853</td>
</tr>
<tr>
<td>Non-sense</td>
<td>3261</td>
<td>443</td>
<td>20791</td>
<td>32810</td>
<td>8109</td>
<td>5446</td>
<td>7452</td>
</tr>
<tr>
<td>Read-through</td>
<td>113</td>
<td>23</td>
<td>312</td>
<td>355</td>
<td>404</td>
<td>257</td>
<td>354</td>
</tr>
<tr>
<td>Deletions</td>
<td>5114</td>
<td>6415</td>
<td>107</td>
<td>2415</td>
<td>3399</td>
<td>3549</td>
<td>4028</td>
</tr>
<tr>
<td>Insertions</td>
<td>4820</td>
<td>9533</td>
<td>100</td>
<td>2299</td>
<td>3295</td>
<td>3412</td>
<td>3761</td>
</tr>
</tbody>
</table>
Table 5. $Sk$ locus analysis demonstrating the effect of the various parameters on the reduction in potential targets related to bitterness in almond.

<table>
<thead>
<tr>
<th>Region/Filter</th>
<th>Number of target polymorphisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Chromosome 5</td>
<td>311497</td>
</tr>
<tr>
<td>B. A + 11-14.6MB</td>
<td>56155</td>
</tr>
<tr>
<td>C. B + fitting genetic patterns</td>
<td>6304</td>
</tr>
<tr>
<td>D. C + with codon change</td>
<td>228</td>
</tr>
</tbody>
</table>
Table 6: Lists the number of genes containing mutation for each EC number represented in Figure 3.

<table>
<thead>
<tr>
<th>EC Number</th>
<th>Enzyme</th>
<th>Cherry</th>
<th>Bitter 1</th>
<th>Bitter 2</th>
<th>Sweet 1</th>
<th>Sweet2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1.21 (Red)</td>
<td>Beta-glucosidase</td>
<td>8</td>
<td>7</td>
<td>13</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>3.5.5.4 (Yellow)</td>
<td>Cyanoalanine nitrilase</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3.5.5.1 (Green)</td>
<td>Nitrilase</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6.3.1.1 (Orange)</td>
<td>Aspartate-ammonia ligase</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4.1.2.10 (Brown)</td>
<td>(R)-mandelonitrile lyase</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>3.2.1.118 (Blue)</td>
<td>Prunasin beta-glucosidase</td>
<td>12</td>
<td>5</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>3.2.1.117 (Pink)</td>
<td>Amygdalin beta-glucosidase</td>
<td>14</td>
<td>7</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>3.5.1.1 (Grey)</td>
<td>Asparaginase</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1.14.13.68 (Purple)</td>
<td>4-hydroxyphenylacetaldehyde oxime monooxygenase</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2.1.2.1 (Cyan)</td>
<td>glycine hydroxymethyltransferase</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 7: Lists the number of genes for each sample containing nonsense SNPs for each EC number in Figure 4. Only the lower half of the pathway was imaged.

<table>
<thead>
<tr>
<th>EC Number</th>
<th>Enzyme</th>
<th>Cherry</th>
<th>Bitter 1</th>
<th>Bitter 2</th>
<th>Sweet 1</th>
<th>Sweet2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6.1.57  (Red)</td>
<td>Aromatic-amino-acid transaminase</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4.4.1.15  (Yellow)</td>
<td>1-aminocyclopropane-1-carboxylate synthase</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1.14.17.4 (Green)</td>
<td>Aminocyclopropanecarboxylate oxidase</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6.3.1.1  (Orange)</td>
<td>Methionine synthase</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4.1.1.50 (Lime Green)</td>
<td>Adenosylmethionine decarboxylase</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2.6.1.5  (Blue)</td>
<td>Tyrosine transaminase</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2.1.1.10 (Pink)</td>
<td>Homocysteine S-methyltransferase</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2.1.1.37 (Grey)</td>
<td>DNA (cytosine-5')-methyltransferase</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>2.5.1.6 (Purple)</td>
<td>Methionine adenosyltransferase</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2.1.1.14 (Cyan)</td>
<td>5-methyltetrahydropteroylglutamate-homocysteine S-methyltransferase</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 1: GO-term composition of nonsense SNP-containing datasets separated by molecular function, biological process, and cellular component. Blast2GO was used to assign function to sequences predicted to have nonsense mutations. GO-terms were separated by percent composition for each dataset including the entire peach dataset. Comparison to the entire peach identifies GO-terms which may have higher or lower frequencies of developing nonsense-SNPs.
Figure 2: Venn diagram displaying presence of nonsense SNPs present within the five investigated datasets mapped against peach predicted genes. A comparison of the composition of putative nonsense SNP containing genes between the four investigated cultivars of almond and the combined cherry data set reveals the presence of a large set, 1191, of nonsense containing homologues across all members. Additionally, each sample has a unique set of genes containing putative nonsense SNPs, most notably cherry with 2535 genes.
Figure 3: KEGG Pathway of genes containing nonsense mutations within cherry and almond samples involved in Cyanoamino acid metabolism. Table 6 lists the number of genes containing mutation for each EC number represented.
Figure 4: KEGG pathway of genes containing nonsense mutations within cherry and almond samples participating in cysteine and methionine metabolism. Genes containing nonsense mutations were mapped to EC numbers using Blast2GO and mapped on the cysteine and methionine metabolism KEGG map closest to ethylene production. Table 7 lists the number of genes for each sample containing nonsense SNPs for each EC number. Only the lower half of the pathway was imaged.
Figure 5: Venn diagram of Peach genes containing nonsense mutations detected within the four investigates cultivars of almond. Sequences corresponding to mutations in the peach predicted genes were recorded for each investigated cultivar of almond. Datasets were cross-compared to identify sequences containing nonsense mutations unique to each cultivar and made into a Venn diagram using Venny (Oliveros, 2007).
**Additional Files**

3A.**Additional File 1 Coverage Table:** excel file of the mapping coverage for each scaffold of the peach genome for each sample. Blank entries are the result of no mapping.

3A.**Additional File 2 Scaffold Depth Images:** PPT file with a compressed bar graph depicting polymorphism rate in each 50kb window for each sample.

3A.**Additional File 3 significant regions:** Excel file of the 50kb regions with significantly higher or lower polymorphism depth.

3A.**Additional File 4 Chi Table:** Chi-square test of observed Gene Ontology distribution amongst datasets. Gene ontologies were determined for the entire set of peach genes to test if the observed number of gene ontologies for the cherry and almond nonsense SNP-containing genes were significantly different than expected. Gene ontology IDs with P-values lower than 0.001 were highlighted and noted with either higher or lower representation than expected.

3A.**Additional File 5 GlobalKEGGComparison:** KEGG pathways with members predicted to have nonsense SNPs. Blast2GO was used to assign EC numbers to genes in each dataset containing putative nonsense SNPs. These EC numbers were mapped back to KEGG maps. The table below lists the KEGG pathways and records the presence of members with nonsense SNP with an X.

3A.**Additional File 6 Response to Stress:** KEGG pathways with members in “Response to Stress” gene ontology. All sequences with the parent GO term “Response to Stress” were selected and mapped to KEGG maps. Datasets containing a nonsense SNP mapping back to a KEGG map are indicated with an “X” suggesting a potential loss of function in an aspect of the metabolic process.
RADshredder: A tool enabling RAD data utilization in *de novo* genome assembly

Tyson Koepke¹, Artemus Harper¹, Jonathan Maldonado², Lydia Paradiso³, Rick Nipper⁴, Jason Boone⁴, Herman Silva², Anantharaman Kalyanaraman⁵, Amit Dhingra¹

*Prepared for submission to BMC Genomics*

1. Department of Horticulture, Washington State University, PO Box 646414, Pullman, WA 99164-6414
2. Facultad de Ciencias Agronómicas Universidad de Chile Av. Santa Rosa 11315, 8820808 La Pintana Casilla 1004 Santiago, Chile
3. New Jersey Institute of Technology, University Heights, Newark, NJ 07102-1982
4. Floragenex, Portland OR 97201
5. School of Electrical Engineering and Computer Science, Washington State University, PO Box 642752, Pullman WA 99164-2752

Corresponding author

Email addresses:

TK: tkoepke@wsu.edu

AD: adhingga@wsu.edu
Abstract

Background

De novo genome assemblers typically assume even coverage across the entire genome which hinders the utilization of reduced representation sequence data for genome assemblies. Restriction associated DNA (RAD) sequencing is a reduced representation technique that enables the highly confident assembly of small portions of genomes. De novo genome assemblers typically treat the raw RAD reads as repeats regions due to excess coverage and RAD contigs as normal quality reads rather than the high quality consensus they represent.

Findings

Using our new processing tool RADshredder, raw RAD read data and the corresponding RAD contigs can be processed into pseudo-reads that maximize the utility of RAD data in de novo genome assembly. RADshredder was tested on data generated for the diploid sweet cherry genome and demonstrates an improvement in assembly when comparing RADshredder output to the raw reads or the contigs in the assembly. This improvement is characterized by a 36 and 62 base increases in N50 length compared to assemblies using the RAD contigs or RAD raw reads respectively. The best assembly contigs were analyzed and found to represent ~96% of Arabidopsis flowering proteins.

Conclusions

This tool provides a method to consistently modify high-confidence reduced representation but high coverage data into pieces that will assemble with available software. Though this tool was developed for RAD data, it may prove to be more useful for integrating other reduced representation sequencing data into de novo genome assemblies. The built in flexibilities of the tool will enable RADshredder to be individually optimized for each specific dataset to help ensure the best possible results.
Findings

Background

Genome sequences greatly improve the ability to understand the genetic mechanisms behind interesting and important physiological traits. In sweet cherry (*Prunus avium*), investigations of many economically important traits including flowering, rootstock scion interactions, and fruit-pedicel abscission for mechanical harvesting can be greatly aided by obtaining a comprehensive genome assembly of sweet cherry. Model organisms have benefitted from genome sequence information for several years now and have facilitated the development of many techniques that can be leveraged in non-model species. Like sweet cherry, most organisms do not have genome sequences and are therefore dependent on the genome of the closest related model organism.

Reference based assemblies against completed genomes can yield large amounts of quality information for these related, un-sequenced species. These data can reveal coding change mutations in genes of interest and more globally can identify highly conserved or divergent regions among different species (Koepke, Schaeffer et al. in preparation). A caveat of this approach is the loss of information in areas where the genomes may differ which escalates as the distance from the reference organism increases.

*De novo* genome assembly is an alternative to reference based assembly and is an unbiased method to assemble a genome. *De novo* assembly has advanced rapidly in the past decade with the advent of second generation sequencing technologies and the large datasets produced by them. 454, Illumina and Solid technologies each have distinct advantages and disadvantages when utilized for *de novo* genome assemblies (Metzker 2010). As these technologies have evolved to provide longer reads with increased accuracy, *de novo* assemblies have become increasingly reliable (Miller et al. 2010). Additionally, the ability to produce single and paired end sequencing data from these platforms has aided assembly of contigs and scaffolds. As explained by Miller et al. (Miller et al. 2010), each of the multiple algorithms used for *de novo* assembly of second generation sequencing data face similar challenges including: repeat regions, non-uniform sequencing coverage, and a need for high power computers to run many of them. Of
these potential pitfalls of *de novo* genome assembly, repeat regions consistently pose the greatest challenge. To overcome this difficulty, many assembly programs treat over-represented sequenced regions as repeats and may remove them from the assembly.

In contrast to expensive whole genome assembly, many researchers are utilizing reduced representation sequencing techniques to obtain high depth of coverage on specific areas of interest with lower costs. These methods include exome sequencing (Ng et al. 2009), RNA sequencing (Snyder et al. 2009), genotyping by sequencing (Elshire et al. 2011) and restriction associated DNA (RAD) sequencing (Chutimanitsakun et al. 2011). Each of these methods is highly useful for identifying differences between samples, potentially identifying putative causes for phenotypic changes in populations. These techniques typically avoid the non-coding, repetitive regions of the genome, thereby simplifying analysis and reducing overall costs.

RAD sequencing leverages the predictable randomness of restriction enzymes to specifically sequence regions of the genome around the recognition sites (Baird et al. 2008). Paired-end sequencing library preparations create DNA molecules anchored on one end at the restriction site with the fragment length varying within a specified range. This allows the sequences at the restriction site to overlap completely with the paired reads having partial overlapping at the other end. The partially overlapped “paired end” is then assembled into contigs, yielding a contig on each side of the restriction site as seen in Figure 1. RAD data have been extensively utilized for developing SNP based markers and for linkage mapping (Chutimanitsakun et al. 2011; Baird et al. 2008; Pfender et al. 2011). RAD sequencing data, however, are not currently optimal for inclusion in genome assemblies because the sequencing is not random and therefore not uniform across the genome.

Since reduced representation sequencing data are being generated for various projects and completed genomes are highly desired, it would be advantageous to utilize this data for *de novo* genome assembly in species lacking a genome sequence. To assemble a draft genome for sweet cherry, *Prunus avium*, single and paired end 454 sequences, Illumina paired-end and RAD data were generated. In this article, we present the tool RADshredder that was designed to overcome the non-uniformity of RAD data
and improve de novo genome assembly through the utilization of RAD sequences. Because flowering is an important trait of interest in sweet cherry, the best assembly was analyzed for the presence of known flowering genes to assess the quality of the assembly. Overall, RADshredder produces an improved de novo assembly and nearly all flowering related protein sequences were identified within the best assembly.

**Methods:**

**DNA prep.** Immature leaves from the sweet cherry variety Stella were collected from a virus free tree at the Pear Acres research plot in Prosser, WA. Leaves were rinsed with 70% ethanol and frozen in liquid nitrogen. Frozen leaves were then ground into a fine powder using a FreezerMill (Spex Sample Prep). DNA was then extracted using a modified CTAB protocol.

**Sequencing/data generation. 454.** Stella DNA was prepared into both 454 FLX and Titanium single end (454SE) libraries according to manufacturer’s instructions (Roche). The FLX library was sequenced on 8 regions of 454 FLX sequencing plates and the Titanium library was sequenced on 3 large regions of 454 Titanium sequencing plates, the equivalent of 1.5 full runs. Stella DNA was also sent to Roche where four 8kb insert and two 20kb insert paired end libraries (454PE) were created. Each paired end library was sequenced on 1 of 8 regions on a single 454 Titanium sequencing run by Roche. All raw 454 data were produced by image analysis using the gsRunProcessor. The resulting sequence flowgram format (sff) files were then processed using Pyrobayes (Quinlan et al. 2008). Either the sffs or the Pyrobayes fasta and associated quality files were then used for assembly depending on the assembler.

**Illumina Shotgun (SHOT).** An Illumina paired end library was prepared from Stella DNA according to protocol (Illumina). During size selection, fragments of 500bp to 700bp were retained to create a smaller distribution of insert sizes. The prepared library was multiplexed and sequenced on one lane of an Illumina GAII 2x60 sequencing run at the University of Oregon. The 4bp index sequence was trimmed from the reads by Floragenex and the trimmed fastq was used for assembly.

Due to poor initial assembly results, an additional 76x Illumina data were obtained from 2x100 standard Illumina HiSeq 2000 sequencing at Michigan State University. These libraries were multiplexed
and the fastq was obtained after initial sorting and filtering of the data via Illumina’s standard data processing.

**RAD.** Stella DNA was digested and prepared as described in Chutimanitsakun et al. (Chutimanitsakun et al. 2011). PstI was used to digest the DNA due to its typical effectiveness for previous RAD projects. The RAD library was also multiplexed and sequenced on Illumina GAII 2x60 sequencing run. The raw Illumina reads obtained from the RAD libraries were filtered based on the indexing sequence, index removed, and grouped into clusters based on exact matches of the first 45bp of the single read end and formatted into compact RAD format (CRF). Each CRF cluster was then assembled using Floragenex’s proprietary software and contigs greater than 130bp were exported into a new file. Contigs shorter than 130bp and crf clusters with a single read depth of less than 35x or greater than 750x were not included in the RAD assembly file.

**RADshredder.** To create overlapping, paired-end reads to navigate the realm between the single contig and the ultra-deep reads, we developed the RADshredder program. RADshredder incorporates the mapping abilities of the Burrows Wheeler Aligner (BWA) (Li and Durbin 2009) to map the reads to the contigs from the RAD assembly file. The BWA results are then processed with samtools (Li et al. 2009) to create a pileup output. The pileup depth and reference base are then converted into pseudo-reads (Figure 2) via a custom algorithm. This algorithm converts the actual depth at each nucleotide to be recorded proportionally to the user inputs.

The user decides on the two main factors of the algorithm which are the depth of the read reduction, and the length of the pseudo-reads desired. Here, depths of 5, 10 and 15 were used with pseudo-read lengths of 100 bases and 200 bases. The output pseudo-read fasta file was then ready to be processed through RADsplitter as described below. All of this RADshredder flow has been integrated into a single perl script which can be run on the command line or through a java based GUI that has been developed to run the script. The final outputs of RADshredder are the composite samtools pileup for every contig and a fasta file with the contig name, which includes the single read sequence.
**RADsplitter:** Both the RAD assembly file and the RADshredder output fasta are formatted in fasta format with the single read sequence in the header and the associated sequence as the read. To split this data into a single fasta file with the single read and paired contig/pseudo-read as the first and second reads in a pair, each file was parsed separately using the RADsplitter custom perl script. The output of RADsplitter is a fasta file that uses the /1 and /2 naming convention (Figure 3) employed by Illumina for paired end reads so they can be processed by assembly programs correctly. The resulting RADsplitter fasta files were then used for assembly as the RAD contigs and RAD pseudo-reads.

**Assembly. NGen:** After the data were created and processed, assemblies were completed using DNAstar’s NGen software (labeled N#). Using default parameters in the run script with an estimated genome size of 225Mb and turning off the small contig removal, data were assembled in the following combinations:

1. 454SE + 454PE + SHOT + PacBio;
2. 454SE + 454PE + SHOT + RAD contigs;
3. 454SE + 454PE + SHOT + RAD reads;
4. 454SE + 454PE + SHOT + RAD reads + RAD contigs;
5. 454SE + 454PE + SHOT + D5 100bp pseudo-reads;
6. 454SE + 454PE + SHOT + D5 200bp pseudo-reads;
7. 454SE + 454PE + SHOT + D10 100bp pseudo-reads;
8. 454SE + 454PE + SHOT + D10 200bp pseudo-reads;
9. 454SE + 454PE + SHOT + D15 100bp pseudo-reads;
10. 454SE + 454PE + SHOT + D15 200bp pseudo-reads;

These combinations were chosen to compare the depth and length parameters and how each affected the final assembly. Additionally, the default assembly parameters for the software were designed for either 454 or Illumina data and not the combination attempted here. To overcome this complication, each dataset was assembled using the 454 and Illumina default setting separately and labeled N#-454 or N#-Ill respectively. Total contigs, total contig bases, average contig size, N50 and contigs greater than 2k were
analyzed to compare the assembly results. Additionally, analysis of the PstI restriction site abundance and locations was completed utilizing the high-throughput restriction site analysis tool: CisSERS (Koepke, Sharpe et al. in preparation) to determine how many PstI sites were located near the end for RAD data to improve the assembly.

**CLC:** These datasets, minus the PacBio data, were also assembled using CLC’s Genomics Workbench assembly software (labeled C#). After the initial round of assembly, modifications to the assembly parameters were made and a subset of these groups was re-processed incorporating the additional 76x Illumina data with chloroplast screening to test the effectiveness of RADshredder data in assemblies with a more typical amount of sequencing data (labeled C#b). Since the initial assembly of the data was not affected by the inclusion of the RADshredder pseudo-reads, the effect on scaffolding was tested by enabling the scaffolding portion of the CLC’s Genomics Workbench assembler on these secondary assemblies.

**Post Assembly Analyses:** To examine the utility of the best assembly, analysis was completed to identify sequences of flowering related genes. These genes are of particular interest in sweet cherry due to the drastic ability of the rootstock to control flower number and due to the fact that flower number directly correlates with fruit yield. Using the complete list of Arabidopsis flowering proteins and genes from NCBI, the contigs from assembly C10b were analyzed via local blast. The results from blastx or blastn were compiled to determine how many of the Arabidopsis flowering related protein and mRNA sequences were present in this assembly. E values of 10-10 were used for an initial screening of high-quality matches while the e value limit was also removed to find any matches based on the default blast settings.

**Results and Discussion:**

**Sequencing data generation:** The sweet cherry genome is estimated to be 200-250 Mb and 225Mb was used as the benchmark (IPGI 2010; Shulaev et al. 2008; Zhebentayeva et al. 2008). By this standard, the initial sequencing data produced 4.53x 454 and 9.06x Illumina sequencing data as seen in Chapter 3A:Table 1. Pyrobayes analysis reduced the raw 454 total by 4872 bases. Sub-assembly of the
RAD data resulted in a decrease to 0.12x coverage. These reductions in starting data leave ~7x depth which is below the 18x for 454 and 50x for Illumina used for de novo assembly on the NGen software when compared to the amount of data used during benchmarking (http://www.dnastar.com/ngen_benchmarks.aspx). 26Mb of PacBio sequences greater than 50bp were produced equating to 0.1x coverage. These data were used for the initial assembly tests.

The second batch of Illumina sequencing produced paired end reads of 100 bases for a total of 17.2 Gb. This additional 76x sequencing data was used in the final CLC’s Genomics Workbench assemblies (C#b).

**RADshredder/RAD splitter.** The conversion of RAD contigs and the raw RAD reads into pseudo-reads for use in de novo assembly was accomplished by the RADshredder computer program. RADshredder effectively created pseudo-reads that correspond to between 1/5 and 1/2 of the read depths depending on the region analyzed (Figure 4). This was expected at a depth reduction level of X=5 because in order to retain significant overlaps to enable assembly, the reduction is slightly less than X (Figure 2). The single read end was reduced from 138 to 18 reads. When run with the RAD dataset on a Linux machine with 16Gb RAM, RADshredder took approximately 9.5 hours to complete. The major bottleneck in the RADshredder computation is the individual alignment and pileup for each contig which was 153,620 alignments for this dataset. Running RADshredder with depth reduction values of 5, 10, and 15 and the pseudo-read length at 100bp and 200bp produced 6 fasta files for further analysis. These files contained from 2,531,274 (D5_100bp) to 470,130 (D15_200bp) pseudo-reads that were added to the assemblies. While RADshredder is specific to this application, the batch processing through bwa and samtools may be sections that could have further utility in other applications.

**Assembly.** NGen: The assembly statistics from the 20 different assemblies were compiled into 3B.AdditionalFile.1. Figure 5 depicts the primary assembly statistics for each assembly for comparison. It is very clear that the 454 default setting produced a more desirable assembly as there are less contigs with increases in average contig length, N50 and contigs >2kb. Comparing the datasets within the 454 assembled statistics reveals that the RADshredder data did not yield an increase in average contig size or
N50 nor did it reduce total contig number as expected compared to either assembly N2 containing RAD contigs or assembly N4 with the raw RAD reads plus the RAD contigs. RADshredder pseudo-reads, therefore, did not improve the NGen assembly based on any of these statistical measures which could be a result of the longer 454 reads included in the assembly. Since the raw RAD reads included assemblies (N3 and N4) yielded more contigs greater than 2500bp than the other assemblies, it appears that the NGen algorithm may be sufficiently handling the raw RAD data and utilizing it to improve the overall assembly results. Analysis of the average contig length and extremely large number of contigs produced in assembly N4 revealed that 8,701,287 of the contigs have lengths less than 100 bases. This significantly added to contig number while drastically reducing average contig length and N50. A majority of these contigs, 4,788,939, were 55 bases long. These likely resulted from the RAD single end reads creating a contig but failing to extend into a longer contig. This is supported by the fact that these 55base contigs were less represented in the assemblies not containing the raw RAD reads with only 513,493 in N2. The distribution of the other contig sizes appears relatively the same between assemblies of datasets N2 and N4 (Figure 6).

Analysis of the PstI recognition sites in the assembly results of dataset N2 with CisSERS (Koepke, Sharpe et al. in preparation) revealed that only 27.5% of the sequences larger than 1000 bases contained at least one PstI site (Figure 7). Further, only 8.88% and 3.62% of these sequences had a site within 500bp of the 5’ end or 3’ end respectively. This subset decreases to a total of about 7.1% when analyzing the 300bp on each end. This range near the end of the contigs is of interest because the RAD data was constructed around PstI restriction sites and yield the RAD assembled contigs of 100-400 bases. For these RAD contigs to bridge gaps in the assembly, the PstI site must be located near enough to the end of the contig for the RAD assembly data to successfully aid the assembly. Since PstI has a six base recognition site, PstI sites are expected to be located every once every 167,936 bases. In N2 contigs greater than 1000 bases, the 46,540 contigs had an average contig length of 1731 bases. In these 80,571,495 total bases, CisSERS identified 22,897 total pstI sites yielding a rate of one in every 3518 bases. This is slightly higher than expected which could be due to the RAD data assisting assembly of
these sites and non-proportionally increasing the lengths of PstI containing contigs. To check this, the rate
in the contigs between 1000 and 200 bases (170,674,338 bases) was found to be 1 in 7147 bases. The
total cut sites in contigs greater than 200 bases is 46777 at a final rate of 1 in 5371 bases which is a
slightly lower rate than random prediction. Overall, the greater occurrence rate of PstI sites in the larger
contigs suggests that the RAD data facilitated improved assembly around the PstI sites.

**CLC:** Assemblies of the 10 datasets were completed using CLC’s Genomics Workbench. The
reporting statistics of most interest from Genomics Workbench include contig number, average contig
length, N50 and total contig bases (Figure 8 and 3B.AdditionalFile.1 Assembly Stats). Analysis of the
initial 10 assemblies shows similar results to the NGen assemblies when analyzing the effect of adding
the RAD pseudo-reads to the assembly.

After the additional 50x Illumina data were obtained, Genomics Workbench assembly results
were analyzed for datasets: 2, 4, 5, 7, 9, and 10. These assemblies utilized the peach chloroplast genome
to screen out chloroplast DNA reads to ease assembly and provide statistics only on the genomic data.
From all of the CLC datasets, C10b provided the best results summarized by the higher average and N50
contig lengths (Figure 9 and 3B.AdditionalFile.1 Assembly Stats). The reduction in total contig bases is
also expected as similar regions are collapsed into single contigs and contig lengths increase. C5b, C7b,
and C9b assemblies each are of lower quality than C2b though C7b is an improvement on the C4b
assembly which includes the raw RAD data. This shows that RADshredder can be functional, though it
will take optimization for the individual dataset being analyzed. The overall improvement of C10b over
C2b in N50 and average contig length are 36 and 22 bases respectively. This is a subtle difference in these
averages though the maximum contig length is 345 bases longer in C10b than C2b.

Though the total bases input for each assembly is different, the major comparisons between RAD
reads, contigs, and RADshredder pseudo-reads focus on variable depths at ~1% of the genome since each
data set represents the same loci across the genome. There is a difference in the way the CLC’s Genomics
Workbench and DNAStar’s NGen handled the RADshredder pseudo-reads as Genomics Workbench
worked best using the D5_100bp with D15_200bp being the worst set while NGen produced the best
results using the D15_200bp and D5_100bp was the poorest. These classifications are based on the assembly statistics and may not represent the accuracy of the assembly which needs to be confirmed biologically.

Post Assembly Analyses: 199 flowering related protein accessions from Arabidopsis thaliana were analyzed for their presence in the final CLC bio assembly C10b. Eleven of these accessions represent the same protein description leaving 188 unique proteins. Of these 188 protein sequences analyzed, 95.7% (180) showed homology with some of the sweet cherry contigs at an e value less than $10^{-10}$. 118 of these showed either long homology (>200aa) or multiple exons on the same contig adding confidence in the likely annotation of these protein sequences. The 8 protein sequences were not identified to have homology with an e value <$10^{-10}$ each have homology at lower levels with several containing multiple smaller exons that are too short to match the $10^{-10}$ criteria while representing solid matches to sequences in the sweet cherry assembly.

Analysis of Arabidopsis thaliana flowering mRNAs present in the sweet cherry genome was also completed. Using an e value cut off of $10^{-10}$, only 35 of 375 analyzed flowering mRNA sequences were identified in sweet cherry. When no e value cut off is employed, 337 of 375 mRNA sequences (264 of 299 unique) still have no homology to the contigs in this sweet cherry genome assembly. These homology rates are expectedly lower than seen with the protein sequences as codon usage is expected to be different due to the evolutionary distance between sweet cherry and Arabidopsis.

Conclusions

Overall, RADshredder functioned as designed, and, in one dataset, the pseudo-reads enhanced the assembly and scaffolding of the sweet cherry genome. This positive result was not possible without the additional Illumina sequencing data for the sweet cherry genome. Additionally, the small number of PstI sites near the edges of contigs decreased the likelihood that the RAD contigs and raw RAD reads would span the gaps and create full contigs around the PstI sites.

While it was previously expected that the depth of coverage from the raw RAD reads would create repeat-like regions that would disrupt the overall assembly, this was not observed in the sweet
cherry dataset. Higher depths of RAD, GBS, or other reduced representation sequencing may cause more assembly issues than observed here and could potentially benefit from RADshredder processing. While RADshredder was able to successfully improve the genome assembly of sweet cherry, the additional processing time and parameter tests may make modification of the assembler’s parameters a more fruitful use of the computer resources depending on the amount, source, and quality of the reduced representation sequence data to be processed through RADshredder.

The best version of the sweet cherry assembly from this work had sequences with homology to 96% of Arabidopsis flowering protein sequences. This shows that this version of the sweet cherry genome assembly can be utilized as a draft that can be leveraged to improve research on other important traits in this important horticultural plant. Future efforts to complete the genome, however, still need to be undertaken. One path toward completing the genome will include optimization of the assembly parameters. Secondly, additional sequencing data, especially including sufficient numbers of long reads from PacBio technology may be extremely helpful in bridging contigs. Lastly, the creation of a linkage map for sweet cherry is expected to anchor these contigs and enable to structuring of chromosome length scaffolds.

Tyson Koepke, Scott Schaeffer, Derick Jiwan, Christopher Hendrickson, Artemus Harper, Jonathan Maldonado, Lydia Paradiso, Rick Nipper, Jason Boone, Herman Silva, Anantharaman Kalyanaraman, Amit Dhingra

**Author Contributions:** AD and HS are primary contributors to the Sweet Cherry Genome Sequencing Consortium. TK, SS, DJ, and CH processed DNA and performed 454 sequencing runs. RN and JB provided the RAD sequencing analysis. TK, RN, JB, and AD developed the concept of RADshredder. TK developed the perl code for RADshredder and AH developed the java GUI. AH completed NGen assemblies of the data and JM completed CLC assemblies. TK, LP, and AK developed methods to test the assemblies. TK wrote the manuscript.
Works Cited:


Figure 1 - Example RAD data at a single locus of a genome. The blue lines on top represent the genomic DNA containing a PstI restriction site. Immediately below the genomic lines are a representation of the RAD contig data comprised of the individual pairs of reads shown in green.
A hypothetical contig of length ~350bp is shown (A) and an approximation of the reads that would map to this contig (B) demonstrates the original read depth. C represents how the RADshredder pseudo-reads should overlap and account for the variable depth along the contig where the red section of the bar designates bases whose read depth will be decreased while the blue section remains unmodified. The graph (D) provides a view of the depth at every 50bp for the reads (orange) and the pseudo-reads (blue).

**Detailed workflow:** RADshredder uses the RAD assembly file and the compact RAD format (crf) file and reference maps all reads matching the first 45bp of the single read with a given contig. The mapping is
completed using the BWA index, aln, and samse commands. The resulting .sam file converted into a
useable pileup format through samtools’ faidx, import, sort, and pileup –cf commands. The pileup is then
parsed for the base sequence and the depth at each base is calculated. The input variable “read depth
factor”, or X, initializes the compression rate that will be used. The depth of the first 50bp is reduced by
X and trimmed until the first base’s depth is ≥1. Then the next 100bp are printed as the next pseudo-read,
first 50bp depth’s reduced by 2X, and this cycle repeated until the remaining portion of the contig is
<100bp. The adjustment to a 2X reduction for the non-tail section was made to account for the 50bp
minimum overlap between pseudo-reads and avoid over-representing the center section of contigs. At the
end of the contig, the final base depth is reduced by X each cycle until it is <1. When the final base depth
is <1, it is trimmed until the last base depth is ≥1 where the printing, reducing and trimming cycle repeats
or the remaining number of bases is <50 which triggers the next contig to be processed.
Figure 3 - Depiction of RADsplitter input and results. A: The RAD assembly fasta that serves as input for RADsplitter. B: RADsplitter output where the header of the input contig 1 is used for the output contig 1/1 and 1/2 and the single read sequence from the header is used as the /1 sequence and the contig for the /2 sequence.
Figure 4 – Actual depth reduction image. Actual depth of A. RAD raw reads and B. RADshredder pseudo-reads D5_100bp. Images from DNASTAR SeqMan Pro strategy view of datasets mapping to contig 9527 from assembly C10b. C. Actual (green) and represented depth (blue).
Figure 5 – NGen Assembly Results. In all four graphs, the x axis represents the data set utilized as described in the methods. Blue are 454 assemblies and Red are Illumina default settings. Overall, the 454 defaults tend to yield more, large contigs which is an improvement of the assembly.
Figure 6 - Contig length analysis of NGen assembly N2-454 and N4-454. Contig length analyses demonstrating the primary differences between RS2 (blue) and RS4 (red) are located between 0 and 400 bases long. A: logarithmic representation of depth for the first 3000 bases with two significant peaks in boxes. B: an expanded view of all contigs greater than 300bp. This shows that RS2 and RS4 are essentially the same though the density of the red marks covers the blue marks which are underneath in this graph. C: Logarithmic view of the yellow region in A. D: expanded linear representation of the number of contigs at each length with RS2 having more contigs at each of these lower depths than are found in RS4.
Figure 7 - PstI analysis of N2-454 contigs. This chart depicts the total percentage of contigs containing a pstI restriction site within range depicted from the 5’ end (0-#), entire contig (all), or the 3’ end (-#0).
Figure 8 – CLC Genomics Workbench Assembly. Results from datasets 1-10 as described in the methods.
Figure 9 – CLC Genomics Workbench Assembly Results. Results from B datasets as described in the methods using the additional Illumina data with the chloroplast reads removed.
Additional File

3B.AdditionalFile.1. Assembly Stats.xlsx – Excel workbook with the Ngen and CLC assembly statistics.

This file also contains the contig length analysis and PstI site analysis.
Examining rootstock scion interactions of sweet cherry via transcriptome profiling


Abstract

Rootstocks are known to control floral bud development and overall yield of sweet cherry, an economically important crop. To understand what genes in the scion are modulated by the rootstock, the transcriptomics approaches of differential display and 3’UTR sequencing were used. Gene expression was measured in the developing floral buds of two scion varieties each grafted onto two rootstock genotypes. Differential display identified 207 putatively differentially expressed gene fragments while 3’UTR sequencing and subsequent analyses revealed 115 differentially expressed transcripts. Additionally, near full length transcriptome sequencing was completed to enable clustering and more effective annotations. Blast analysis revealed that several of the differentially expressed genes in the May sampling due to the rootstock were transcription factors or DNA binding domain containing proteins. Several of these candidates were analyzed via qRT-PCR to confirm the identified expression differences.
Introduction

Sweet cherry, Prunus avium, is an economically important fruit crop with a value of $890 million in 2011 in the U.S (NASS 2012). The value of the sweet cherry crop is dictated by a combination of total yield and fruit size. Size, unfortunately for producers, is negatively correlated with fruit numbers causing a perpetual balancing act between number of fruit and their size at retail. To control fruit numbers and subsequently fruit size, growers utilize hand or chemical thinning treatments. Additionally, different rootstocks are used in production of sweet cherries that cause the scions to produce different numbers of flowers and can control yield up to 10 fold (Whiting personal communication). Because the rootstock controls the number of floral buds produced by scions, the rootstock is expected to modulate gene expression of flowering genes in the sweet cherry scions.

Flowering is an important trait for most agricultural crops and has been studied in depth in multiple model plant systems. From these model species, a complex network of genes has been identified that integrate many environmental signals to control when and how plants flower (Amasino and Michaels 2010). Genes involved in circadian rhythm and photoreception that are involved in flowering have also been identified (Amasino and Michaels 2010). Non-coding RNAs, specifically miRNAs, are also critical players in the pathway controlling flowering in model systems and trees (Wang et al. 2011). These model systems, however, have been annual plants with several key differences in their flowering physiology compared to sweet cherry. Sweet cherry is a perennial plant with a prolonged juvenility stage that lasts for several years before the plant become competent to flower. Once competent, sweet cherry trees produce floral buds that over-winter and open the following spring. Perennial biology along with the complexity of having two genotypes interacting at a somatic level to regulate flowering numbers presents a unique biological model that defies existing genetic paradigms. Change in flower number in sweet cherry is dependent on the type of rootstock used. The axillary meristems that transition to reproductive tissues represent the site of action of the molecular signal emanating from the rootstock. There are two components to this phenomenon which we describe as somatogenetic interactions. One is the signal and the other is the target genes or the proteins that are modified in their action in response to interaction with
the signal. It is important to understand both aspects of this interaction at a gene-based functional level to unravel the phenomenon and regulate flower numbers and thereby the fruit numbers.

Transcriptome profiling or study of gene expression at a global level is a strategy that has been used to understand gene expression changes underlying specific traits. Because this strategy is very effective at identifying strong candidate genes that may underlie a given process, many methods to examine gene expression have been developed (Table 1). Each of these methods plays a unique role due to highly variable requirements such as prior knowledge of the sequences to be examined and differ greatly in the type and amount of information that is generated.

Out of these, differential display leverages anchored poly-T primers to specifically amplify the 3’ end of transcripts (Liang and Pardee 1992). PCR amplification of the 3’ end of transcripts is performed with a radiolabeled nucleotide and expression differences are visualized on a polyacrylamide gel. Differentially expressed bands can then be extracted and sequenced to ascertain their identity. Further, quantitative reverse transcription PCR (qRT-PCR) has been established as the primary method to confirm expression levels of an individual gene including those identified through differential display (Rajeevan et al. 2001). While microarrays provide a global view of gene expression, they are limited to the known gene sequences in the species of interest and more particularly the genes that are represented on the array. Recently, next-generation sequencing technologies have improved the capabilities of examining transcriptomes past microarrays by enabling the discovery of novel transcripts as well as gene splicing. These next-generation platforms produce large amounts of data that improve confidence in the accuracy of the data; however, the large amount of data greatly complicates analysis since previous computational and bio-informatics tools are not optimized for handling the data efficiently and effectively.

Across next-generation platforms, there are several methods for quantifying gene expression. Illumina based RNA-seq uses a shotgun approach where regions from any part of a transcript are sequenced and matched to a reference (Marioni et al. 2008). A count of reads per transcript is then normalized based on the length of the transcript (Mortazavi et al. 2008). The 3’ untranslated region (UTR) profiling protocol described by Eveland et al. (2008) provides an alternative method for quantitative RNA
sequencing and was originally developed to leverage the longer reads (100bp) generated by the 454 platform. For this method, the 3’UTR of poly-A tailed transcripts are selected and sequenced, and the number of reads per transcript is normalized by the number of reads per sample.

Using the 3’ UTR has several advantages including discerning allele biased expression (Eveland et al. 2008). The uniqueness of the 3’UTR, which provides many advantages, is also a disadvantage without a reference sequence since the high specificity greatly reduces the ability of the sequence to align to genes with known functions from other species. To overcome this, a reference transcriptome assembly needs to be completed to assign putative functions to the unique 3’UTRs that otherwise have limited homology to sequences from other species.

Here, both differential display and 3’UTR sequencing were utilized to examine the transcriptome of developing floral buds of sweet cherry. The high producing scion Rainier and the low producing scion Bing were grown on the highly productive rootstock Gisela®6 and the low producing rootstock Mazzard. These combinations produced four rootstock scion combinations and two scion genetic backgrounds to examine the effect of the rootstock on gene expression during floral bud development and were examined at two developmental time points. Differential display produced 173 differentially expressed fragments in Bing based on the rootstock. 3’UTR sequencing revealed 115 transcripts with significant differences in expression. Each of these datasets included multiple transcription factors, several of which were analyzed via qRT-PCR to confirm the differential expression.

Methods

Plant Materials and Sample Preparation: Bing and Rainier trees were planted in Washington State University’s Tukey orchard in Pullman, WA in 2001. Both scion genotypes were grafted on Gisela®6 and Mazzard rootstocks in sets of five in a single, randomized block. In 2007, developing spurs containing floral buds were excised from the branches of these trees, leaf and bark material removed, and flash frozen in liquid nitrogen. The three central spurs from first year wood (the red color region in Figure 1) were selected. First year wood was selected because the tissue is no longer juvenile, being competent to respond to the flowering initiation signal, and is presumably receiving the signal for the first time. For
each sampling, 12 spurs buds from each side of two trees were excised, prepared, collected and flash frozen in liquid nitrogen. Some example spurs are shown in Figure 2. Sampling was conducted on May 19th 2007 and September 12th, 2007. The tally of samples from May includes: Bing/Mazzard (BM5), Bing/Gisela®6 (BG5), Rainier/Mazzard (RM5), and Rainier/Gisela®6 (RG5). Similarly, September samples are: Bing/Mazzard (BM9), Bing/Gisela®6 (BG9), Rainier/Mazzard (RM9), and Rainier/Gisela®6 (RG9). The frozen tissues were pooled for each rootstock/scion/date combination and ground in a SPEX SamplePrep 6870 FreezerMill (Spex SamplePrep, Metuchen, NJ) with 5 cycles of: four minutes grinding and two minutes cooling, with grinding at 15 counts per second (Koepke et al. 2012). Total RNA was extracted from the ground samples with the RNeasy Plant RNA Extraction Kit (Qiagen, Germany). RNA quality was verified via RNA gel electrophoresis to examine the ribosomal RNA bands for degradation.

**Differential Display:** Differential display was performed on the four Bing samples using the RNAimage differential display kit (GenHunter, Nashville, TN) according to the manufacturer’s protocol. First, reverse transcription was performed using ~200ng of total RNA and the H-T11A primer. The H-T11A primer is a poly-T primer anchored with a 3’ A to provide specific binding at the end of poly A tails and not the middle. These first strand cDNAs were then amplified in eight separate reactions one of the H-AP 193-200 primers and H-T11A. These reactions included α[33P]dATP in addition to non-radiolabeled nucleotides. The reactions were incubated using the following thermocycler condition: 40 cycles of 94°C for 30 seconds, 40°C for 2 minutes, 72°C for 30 seconds; 72°C 5 minutes. The radiolabeled products were electrophoresed on a 6% urea, polyacrylamide gel in 1xTBE for three hours at 35watts. Samples were loaded in duplicate with one set being rootstocks adjacent and the second with the sampling dates adjacent as seen in Figure 3. Gels processed and autoradiographed and bands showing differential expression were marked and labeled for extraction and sequencing. The X-ray film was overlaid on the polyacrylamide gel and the desired bands were cut out with a scalpel and placed in 1.5mL tubes. 40µL of H20 was added to the tube containing the gel, plastic wrap, filter paper, and piece of X-ray film. The tube was vortexed, centrifuged briefly, and boiled for 15 minutes. Tubes were placed on ice to
cool and the liquid was removed from the tube into a fresh 1.5 mL tube. A total of 7µL of the extracted product was re-amplified using 4µL of the appropriate H-AP primer (2µM), 4µL H-T11/A (2µM), 3.2µL of 250mM dNTPs, 4µL 10x buffer and 0.4µL of Taq DNA polymerase in a 40µL reaction. These reactions were amplified and visualized with agarose gel electrophoresis to verify proper amplification. Products were then ligated into the pGem®T easy (Promega) vector. The plasmids were transformed into E. coli, cultured, and extracted via alkaline lysis. Insert containing plasmids were sequenced via Sanger sequencing using M13 primers. Sequences were trimmed of plasmid and primer sequences and analyzed via blast to identify a transcriptomic contig match for further annotation. Matched transcripts were compared against the non-redundant nucleotide and protein databases via local blast.

**nfl-cDNA sequencing and analyses:** Near full-length cDNA (nfl-cDNA) synthesis was performed via the Super-SMART cDNA synthesis kit (CloneTech). SMART cDNAs from *Prunus avium* were pooled to equal concentration and were processed for 454 sequencing as per the 454 Titanium protocol (Roche) altered with the following modifications. 10µg of nfl-cDNA were nebulized at 30psi for 75 seconds. 5µg of the non-nebulized nfl-cDNA pool and the nebulized 10µg were run on a low-melting agarose gel. The portion of each nfl-cDNA in the range of 500-800bp were excised and purified from the gel using the MinElute Gel Extraction Kit (Qiagen). These purified nfl-cDNAs from each species were pooled and further processed and sequenced as per 454 Titanium protocol (Roche). The reads from one Titanium sequencing plate were assembled using MIRA version 3.4 (Chevreux et al. 2004). Fasta files of assembled contigs were converted to single line and spaces removed from the header by PreprocessPaCE.pl (Kalyanaraman et al. 2003).

**3′UTR Sequencing:** The Ambion aRNA synthesis kit and a biotinylated poly-T primer for 3′UTR profiling described by Eveland et al. (2008) was used to convert 5µg of RNA into cDNA. The Ambion cDNAs were bound to previously prepared streptavidin Dyna beads (Invitrogen) according to manufacturers’ directions. These streptavidin bound cDNAs were digested with a combination of the identified enzymes. The digested cDNAs were ligated to barcoded, or multiplexed, A adaptors as published (Koepke et al. 2012) according to the 3′UTR protocol (Eveland et al. 2008). Single stranded
template (sst) cDNAs were eluted from the streptavidin beads and sequenced on the 454 FLX system according to protocol (Roche).

3’UTR analysis: The 3’UTR reads are processed to a single line with PreprocessPaCE as above. Then the reads were screened based on the sample barcode, the barcode removed and the read appropriately named using a custom MidRemoval script (4A.AdditionalFile1_MidRemoval_Generic). Using the assembled transcripts from the nfl-cDNA sequencing as seeds or references, the 3’ UTR reads were clustered onto the nfl-seeds using a modified version of PaCE (Kalyanaraman et al. 2003). A file listing the members and cluster name is the output from PaCE and is analyzed using ClusterCounter.pl (4A.AdditionalFile2_ClusterCOunter.pl) to determine the number of reads from each sample, or the ‘counts’, are present in the cluster. This data is processed into a comma separated value (csv) file by the same script. Chi-square analysis of the clusters between biologically interesting combinations was used to identify clusters with differential expression. The nfl-transcripts for these clusters were annotated using Blast+ (Camacho et al. 2009).

qRT-PCR: To confirm the differential expression of candidate transcripts identified from the differential display and 3’UTR sequencing methods, the candidates were first screened based on potential functions. Since global gene expression changes are typically the result in expression changes of transcription factors, transcripts containing homology to known transcription factors were selected. Primers were designed based on the near full length transcript sequences to amplify a ~100-150 bp region. For several candidates, the transcripts aligned to several regions of individual genomic contigs allowing the primer design to span an intron to enable detection of gDNA contamination screening to occur.

Three samples of RNA were extracted from the ground tissue using a modified Qiagen RNeasy Plant RNA extraction column preparation protocol. First, the ground tissues were suspended in buffer RLC, vortexed, and shaken at 4°C for 10 minutes in a tube shaker at 1400 rpm. Then the sample was centrifuged at 1500rpm for 5 minutes and the supernatant loaded onto the lilac column. The original tube was then spun at full speed for 2 minutes and the remaining supernatant was also loaded onto the
corresponding lilac column. The Qiagen RNeasy protocol was followed from this point with two elutions with 32µL of RNase free water. The RNA was quantified using a Nanodrop.

50µL of the RNA was treated with DNase using the DNA-free™ kit (Ambion) according to the manufacturer’s protocol. After DNase treatment, the RNA was quantified again and ~80ng was electrophoresed on an agarose gel to check its quality. After positive RNA quality was confirmed, 14µL of RNA was used in first strand cDNA synthesis via the VILO kit (Invitrogen) according to manufacturer’s protocol. 1µL of the produced cDNA was then used to test for Beta-tubulin, the proposed reference gene, and to examine the DNase treatment by amplifying with primer set TK-48 which spans an intron.

After positive control gene amplification, cDNA was diluted according to the VILO kit instructions to prepare for qRT-PCR. qRT-PCR reactions were performed using iTAQ with ROX and SYBR (BioRad) and 20µL reactions were prepared as per the recommendations outlined by BioRad. 1µL of cDNA diluted to 100ng/µL RNA equivalents was used per reaction with 5µL H2O, 2µL of each primer (10µM), and 10µL of iTAQ SYBR® Green Supermix with ROX. The qRT PCR reactions were performed on a Stratagene MX3005 using the following parameters: 95°C 5min; 50 cycles of 95°C 30sec, 57°C 30 sec, 72°C 30sec; 72°C 5 min. Fluorescence readings were taken at the end of each elongation step. A melting step was performed at the conclusion of the cycles at 95°C for 30 seconds, 54°C for 30 seconds and ramp up to 95°C to produce a dissociation curve. ΔΔCT values and 2^ΔΔCT were computed from the outputs and used to compare the expression of the genes in the samples.

Results and Discussion:

Differential Display: The autoradiographs of the polyacrylamide gels revealed many differences in gene expression among the Bing samples. The primer/sample combinations of HAP-194+ H-T11A +BM9 and HAP-200+ H-T11A +BM9 did not have sufficient product so the entire primer sets were excluded from further analysis. In the remaining 6 sets, 207 differentially expressed gene fragments were identified. A portion of the radiograph from HAP-195+ H-T11A and HAP-198+ H-T11A is shown in Figure 3. Several gene fragments displaying consistent expression levels across the sampling date and
rootstock used provide the baseline for comparisons of the fragments with large variations. Overall, 8 fragments were differentially expressed based on the rootstock used but showed no temporal variation, 58 fragments exhibited temporal differences and remained unchanged by the rootstock, and the remaining 141 exhibited both temporal and rootstock based differential expression (Figure 4 and Table 2). 46 of these interesting gene fragments were extracted, cloned, and sequenced (Table 3). 29 of the sequenced clones matched a cDNA reference sequence and, since some fragments matched to the same transcriptome sequence, the number of unique cDNAs identified was 19. This could have been caused by improper annealing of a primer, allelic-variation or repetitive sequences in the genes themselves. This is commonly observed in differential display experiments (Cin et al. 2005). Allelic variation is possible only if there is specific expression of different forms in these samples since they all are taken from genetically identical scion material. In total, 7 of the 19 unique cDNA sequences had homology to fully descriptive annotations. Many of the cDNAs and fragments, however, did not have significant sequence identity to the databases on NCBI. The descriptive annotations were for: Acetyl-CoA carboxylase, Rubisco subunit binding-protein beta subunit, GAPDH, Cytochrome c oxidase polypeptide, low molecular weight heat-shock protein, MYBR domain class transcription factor, and vacuolar processing enzyme from various species as seen in 4A. Additional File3_Differential Display results. Of these, MYBR is an interesting candidate due to its role as a transcription factor. The other annotations do not have a clear connection with controlling gene expression or flowering. However, it cannot be concluded if the observed change in expression is due to the signal interacting with these proteins directly or transcription factors that regulate these genes.

Transcriptome sequencing: cDNAs produced using the SMRT cDNA syntheses were sequenced on the 454 GS-FLX sequencer using Titanium chemistry (Roche). The single titanium sequencing run produced 1,228,439 reads at an average length of 395 bases (Table 4). This 485Mb of data were processed using PyroBayes to improve de novo assembly (Quinlan et al. 2008). After processing, the reads were assembled using MIRA (Chevreux et al. 2004). The resulting 59,279 contigs comprised a total transcriptomic space of 24.5 Mb, 10-12% of the estimated genome size. Statistics on the assembled
contigs are highlighted by the average contig length of 413 bases. Similar to these results, a 454 GS-FLX based transcriptome assembly of Glanville fritillary butterfly (*Melitaea cinxia*) produced 48,354 contigs with an average length of 192 bases. While the results of these works are comparable, further sequencing depth of the sweet cherry transcriptome is expected to coalesce these contigs into a smaller number or longer contigs that will be more useful for future research.

**3’UTR Sequencing:** After library preparation following a previously published protocol (Eveland et al. 2008), samples were run on individual lanes of GS-FLX sequencing plate. The eight lanes yielded 580,455 reads (Table 5) with an average read length of 85bp. This low read length was expected for the GS-FLX platform in combination with the 3’UTR library preparation method used here. Barcode, or index, sequences were removed from the reads and the sequence headers renamed according to sample name.

After the reads were prepared and processed through PreprocessPaCE.pl, they were clustered with a modified version of PaCE (Kalyanaraman et al. 2003). The transcriptome sequences produced above were input as seeds and matches between the 3’UTR reads and the seeds were analyzed. An unexpected consequence of this approach allowed seeds to be coalesced when a read would match both. Because these transcripts contained enough similarity to share reads, clusters containing more than one transcript seed were removed from further analysis. The PaCE outputs (4A.Additional File 4_PaCEoutput.639729) were processed through the custom script ClusterCounter.pl to create comma separated values (.CSV) files of the counts for each sample type and transcript for import into a spreadsheet program. The files were processed in Microsoft Excel to create expected counts for each transcript normalized for the actual number of reads produced per sample. All clusters were required to have a minimum of 10 members to enable further statistical analyses while reducing the dataset. Clusters with multiple references or too low reads were output into separate CSV files for reference (4A.Additional Files 5-7) but were not used further. Actual and expected counts for each rootstock specific comparison, BM5 and BG5 for example, were used to compute Chi² values (4A.Additional File 8_3UTR analysis). When sum of reads in each comparison was set at a minimum of ten, 119 chi square
values were identified p-value <0.01 in at least of the comparisons corresponding to 115 unique transcripts (Figure 5). 95 of the 115 were identified from the September samples. Prassinos et al. (2009) showed that the rootstock controls the cessation of terminal shoot growth in Bing sweet cherry on Gisela®5 and Gisela®6 rootstocks. They also identified 99 transcript derived fragments showing differential expression via cDNA-AFLP analysis on the shoot tips. Since the rootstock related changes in expression identified in this work are more readily apparent in the closing stages of the growth cycle, the changes could originate from the modified growth cessation program described by Prassinos et al. (2009). Additionally, Mazzard is more vigorous than Gisela®6 which is more vigorous than Gisela®5. Prassinos et al. (2009) also showed that growth cessation in Gisela®5 occurred 6-13 days earlier than Gisela®6 which would be expected to stop before Mazzard if following this paradigm where growth cessation controls overall plant height. While it was not examined in this study, identifying the approximate time of growth cessation of Bing and Rainier scions on Mazzard and Gisela®6 rootstocks could provide insights that could enable further direct comparison among the results of these two studies.

When the 115 differentially expressed transcripts are analyzed, 48 of them have descriptive blast hits via blastx, blastn, or tblastn against the downloaded databases from NCBI. Another 19 have no useful annotations being BAC clones, ESTs or matching an unclassified genomic region while the last 48 do not share any homology with the database entries. The fact that 42 percent of the transcripts do not match known sequences is noteworthy. Since the peach genome is available and is model for the Prunus family, some level of protein and nucleotide homology would be expected with sweet cherry. Alternatively, the transcriptome assembly itself may require more depth to build longer contigs and improve the ability of blast to find homology. Interestingly, annotation of the cDNA-AFLP fragments by Prassinos et al. was far more successful with 86 of 99 fragments being annotated. It is not clear why there is such a large disparity in the annotation frequencies between these two experiments though it is possible that developing floral buds represent a less characterized tissue increasing the difficulty of identifying strong homology.

**qRT-PCR:** qRT-PCR is an irreplaceable tool for validation of differential expression from differential display and sequencing based transcriptome profiling approaches. Five transcripts from the
differential display and six from the 3’UTR were chosen to be confirmed via qRT-PCR. Beta-tubulin has been used previously as a control gene for sweet cherry qRT-PCR (Koepke/Killian/Dhingra unpublished) and was therefore used as a control.

When visualized on an agarose gel, ribosomal RNA was intact and did not show any degradation. Three replicate samples, each from a separate RNA extraction of the same tissue samples were analyzed via qRT-PCR and the results are shown in 4A. Additional File 9_qRT-PCR_results. The beta-tubulin control product varied more than desired among the samples with an overall ΔCT standard deviation of 1.16 which is above the acceptable range of 1.0 (Rieu and Powers 2009). Substantial differences are noted in every qRT-PCR primer set with a couple of genes showing up-regulation in Bing/Gisela combinations compared to Bing/Mazzard (Figure 6). Interestingly, each primer set showed lower expression rates during the September sampling of Rainier/Gisela compared to the Rainier/Mazzard samples. Though these results represent the means of three replicates, re-sampling and testing of these genes in a different growth year will be important to demonstrate their consistent role in determining floral bud number.

As seen with the differential display, it is possible that multiple clusters or fragments belong to a single transcript or gene family. Thereby, when performing qRT-PCR, the entire family is tested and constant or the differential expression previously found is actually a split of the expression into more than one grouping leading one grouping to display significant differences. Additionally, recent assemblies of the sweet cherry genome have identified that sweet cherry is roughly 60% AT which may provide multiple loci per transcript for a poly-T based primer to anneal other than the poly-A tail as desired.

**Conclusions:**

Both differential display and 3’UTR identified many differentially expressed transcripts based on the rootstock in Bing and Rainier developing floral buds. This adds to previous work showing scion gene expression changes in the scion due to the rootstock in apple (Jensen et al. 2003; Jensen et al. 2010) and sweet cherry (Prassinos et al. 2009). Since qRT-PCR on several candidates tested in this work support the results from the differential display or 3’UTR, the tested candidates represent a small subset of candidates for functional analyses.
As mentioned in Koepke and Dhingra (2013), the modification of growth stage and maturity by the rootstock should result in large expression changes. Phenotyping the growth stage of the floral buds along with the corresponding global gene expression would be a very useful extension of the results presented here. Additionally, 3’UTR sequencing has been overwhelmed by the increases in sequencing depth and read length seen by the Illumina sequencing system. RNA-seq methods using Illumina sequencing now derive the transcriptome sequences in a single experiment and don’t rely on a reference sequence. Separately, mapping the 3’UTR reads obtained here to the peach reference transcriptome may be useful as the peach transcripts may be more readily annotated. There have also been several new approaches to normalizing and analyzing RNA-seq type data including SERE (Schulze et al. 2012) that could be applied to the 3’UTR dataset to see if other differentially expressed candidates can be identified.

Overall, while the results from these experiments need further enquiry, many genes have been identified that are putatively controlled by the rootstock genotype. Since many aspects of tree physiology are controlled by the rootstock (Koepke and Dhingra 2013), further analysis needs to be completed to both verify the differential expression and to confirm a role in floral bud development. The result presented here represent an important first step toward unraveling the mechanism for how rootstocks control floral bud number in sweet cherry scions.

**Author Contributions:** TK, MW, and AD designed the studies and biological sampling. TK performed the sampling and RNA extractions. TK, SS, and AD performed the differential display. TK, BU, JH, and MA cloned and sequenced the differentially expressed bands. TK and SS created 454 sequencing libraries and performed the sequencing. AK and TK created the modified clustering program for analysis. TK, AH, and VK completed post-clustering computational analyses of the 3’UTR sequencing data. TK and CH performed the qRT-PCR verification. AD supervised the research. TK wrote the first draft of the manuscript. All authors read and approved the manuscript.
Works cited:


Table 1: Comparison of transcriptome profiling approaches

<table>
<thead>
<tr>
<th>Method</th>
<th>Quantitative</th>
<th>High Throughput</th>
<th>Requires Reference Gene/Genome</th>
<th>Can Identify Unknown Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT-PCR</td>
<td>~</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>Northern</td>
<td>~</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>Differential Display</td>
<td>~</td>
<td>~</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>Microarrays</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>SAGE</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X?</td>
</tr>
<tr>
<td>Q-PCR</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3’UTR</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>POST</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>X</td>
</tr>
</tbody>
</table>
Table 2. A summary of the differential display results:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Differences</td>
<td>207</td>
</tr>
<tr>
<td>Rootstock Dependent</td>
<td>8</td>
</tr>
<tr>
<td>Time Dependent</td>
<td>58</td>
</tr>
<tr>
<td>Rootstock and Time Dependent</td>
<td>141</td>
</tr>
</tbody>
</table>
Table 3. Blast summary for differential display fragments.

<table>
<thead>
<tr>
<th>Sequenced</th>
<th>Match full length cDNA</th>
<th>No cDNA match</th>
<th>Unique cDNAs</th>
<th>cDNAs with blast hits</th>
<th>cDNAs with informative blast hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
<td>29</td>
<td>17</td>
<td>19</td>
<td>14</td>
<td>7</td>
</tr>
</tbody>
</table>
Table 4. nfi-cDNA sequencing data using 454 GS Titanium chemistry

<table>
<thead>
<tr>
<th>Species</th>
<th>Prunus avium</th>
</tr>
</thead>
<tbody>
<tr>
<td># bases</td>
<td>485,830,995</td>
</tr>
<tr>
<td># reads</td>
<td>1,228,439</td>
</tr>
<tr>
<td># contigs</td>
<td>59,279</td>
</tr>
<tr>
<td>Total Contig bases</td>
<td>24,532,780</td>
</tr>
<tr>
<td>Avg. contig size</td>
<td>413.85</td>
</tr>
</tbody>
</table>
Table 5. 3’UTR sequencing results. Total number of reads produced for each sample.

<table>
<thead>
<tr>
<th>Sample</th>
<th>BM5</th>
<th>BG5</th>
<th>BM9</th>
<th>BG9</th>
<th>RM5</th>
<th>RG5</th>
<th>RM9</th>
<th>RG9</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Reads</td>
<td>72,325</td>
<td>105,128</td>
<td>80,160</td>
<td>46,069</td>
<td>105,510</td>
<td>71,286</td>
<td>40,238</td>
<td>59,736</td>
<td>580,455</td>
</tr>
</tbody>
</table>
Table 6. Seeded PaCE Output Summary: Results from the PaCE output and ClusterCount.pl. Clusters only consider single reference clusters.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># Seed sequences</td>
<td>59279</td>
</tr>
<tr>
<td># clusters &gt;=10 members</td>
<td>1436</td>
</tr>
<tr>
<td># clusters &lt;10 members</td>
<td>22997</td>
</tr>
<tr>
<td># significant clusters (chi2 p&lt;0.01)</td>
<td>382</td>
</tr>
</tbody>
</table>
Figure 1: Illustration representing three distinct physiological stages of sweet cherry growth. A (white).
New growth, vegetative. B (red). First year wood, making floral buds for the first time. C (yellow).
Second year wood, fruiting for the first time. Sampling for this project came from section B (red).
Figure 2. Mature floral buds on the four sample combinations. Bing and Rainier are the scions while Mazzard and Gisela®6 rootstocks.
Figure 3. Differential display gel from two primer combinations. Sample names are listed at the top. The left two sets of four display the HAP-195xHT_{11}A while the right two sets are HAP-198xHT_{11}A. Red lines highlight a few of the fragments displaying rootstock related expression differences, blue lines highlight temporal differences and the yellow line highlights a consistently expressed fragment.
Figure 4. Overlapping circle diagram of the differential display results. In red are the number fragments displaying rootstock related changes in expression and blue are the temporally dependent gene fragment counts.
Figure 5. Overlapping circle diagram of transcripts with significantly different expression levels due to the rootstock from 3’UTR sequencing. Yellow = Bing in May, Red = Bing in September, Green = Rainier in May, Blue = Rainier in September. Blank regions of the diagram have no transcripts.
Figure 6. Bar chart of the qRT-PCR $2^{-\Delta\Delta CT}$ values when compared to the beta-tubulin controls (#12). Analyses are shown for rootstock differing pairs (ie. Bing/Mazzard 5 vs Bing/Gisela 5) and the Mazzard sample of each pair was set as the sample control. Bars above 2 and below 0.5 represent two-fold expression differences and are found in each gene set as expected.
Additional Files:

4A.AdditionalFile1_MidRemovalGeneric.pl: Perl script for the removal of multiplexing identifier tags from sequences.

4A.AdditionalFile2_ClusterCounter.pl: Perl script to analyze clusters generated by PaCE to create comma separated value files.

4A.AdditionalFile3_Differential Display results: Excel workbook containing qualitative assessment of the differential display bands and sequencing results of the cloned fragments.

4A.AdditionalFile4_PaCEoutput.639729: The clustering output from the 3’UTR transcriptome analysis.

4A.AdditionalFile5_Post_SPaCE_5_4_2011.best: CSV formatted file with single reference clusters containing at least 10 members.

4A.AdditionalFile6_Post_SPaCE_5_4_2011.MultiRef: CSV formatted file with single reference clusters containing at least 10 members.

4A.AdditionalFile7_Post_SPaCE_5_4_2011.SingleRefLowReads CSV formatted file with single reference clusters containing at least 10 members.

4A.AdditionalFile8_3UTR analysis: Excel file with the 3’UTR analysis including chi square test and blast annotations.

4A.AdditionalFile9_qRT-PCR_results: Excel file with the raw CT values, ΔCT, ΔΔCT, and 2^ΔΔCT analyses.
CHAPTER 4B: TRANSCRIPTOMICS

Rapid gene-based SNP and haplotype marker development in non-model eukaryotes using 3’UTR sequencing

Tyson Koepke, Scott Schaeffer, Vandhana Krishnan, Derick Jiwan, Artemus Harper, Matthew Whiting, Nnadozie Oraguzie and Amit Dhingra

As Published in BMC Genomics, January 12, 2012. 13:18

Abstract

Background

Sweet cherry (Prunus avium L.), a non-model crop with narrow genetic diversity, is an important member of sub-family Amygdaloideae within Rosaceae. Compared to other important members like peach and apple, sweet cherry lacks in genetic and genomic information, impeding understanding of important biological processes and development of efficient breeding approaches. Availability of single nucleotide polymorphism (SNP)-based molecular markers can greatly benefit breeding efforts in such non-model species. RNA-seq approaches employing second generation sequencing platforms offer a unique avenue to rapidly identify gene-based SNPs. Additionally, haplotype markers can be rapidly generated from transcript-based SNPs since they have been found to be extremely utile in identification of genetic variants related to health, disease and response to environment as highlighted by the human HapMap project.

Results

RNA-seq was performed on two sweet cherry cultivars, Bing and Rainier using a 3’ untranslated region (UTR) sequencing method yielding 43,396 assembled contigs. In order to test our approach of rapid identification of SNPs without any reference genome information, over 25% (10,100) of the contigs were
screened for the SNPs. A total of 207 contigs from this set were identified to contain high quality SNPs. A set of 223 primer pairs were designed to amplify SNP containing regions from these contigs and high resolution melting (HRM) analysis was performed with eight important parental sweet cherry cultivars. Six of the parent cultivars were distantly related to Bing and Rainier, the cultivars used for initial SNP discovery. Further, HRM analysis was also performed on 13 seedlings derived from a cross between two of the parents. Our analysis resulted in the identification of 84 (38.7%) primer sets that demonstrated variation among the tested germplasm. Reassembly of the raw 3’UTR sequences using upgraded transcriptome assembly software yielded 34,620 contigs containing 2243 putative SNPs in 887 contigs after stringent filtering. Contigs with multiple SNPs were visually parsed to identify 685 putative haplotypes at 335 loci in 301 contigs.

**Conclusions**

This approach, which leverages the advantages of RNA-seq approaches, enabled rapid generation of gene-linked SNP and haplotype markers. The general approach presented in this study can be easily applied to other non-model eukaryotes irrespective of the ploidy level to identify gene-linked polymorphisms that are expected to facilitate efficient Gene Assisted Breeding (GAB), genotyping and population genetics studies. The identified SNP haplotypes reveal some of the allelic differences in the two sweet cherry cultivars analyzed. The identification of these SNP and haplotype markers is expected to significantly improve the genomic resources for sweet cherry and facilitate efficient GAB in this non-model crop.
**Background**

Sweet cherry (*Prunus avium* L.), a non-model crop, is an important non-climacteric member of sub family Amygdaloideae where other members like peach and plum demonstrate climacteric fruit ripening. Sweet cherry is a diploid (2n = 16) and is estimated to be slightly larger than peach, 225-300MB (Arumuganathan and Earle 1991; Peach Genome v1.0, International Peach Genome Initiative 2010). Sweet cherry underwent a recent breeding-related genetic bottleneck that reduced the diversity present in the germplasm (Mariette et al. 2010). Genetic variability can be utilized to screen for resistance to diseases and improve the efficiency of selecting desirable genotypes through breeding especially in sweet cherry where natural diversity is lacking. Types of variation at the nucleotide level are: microsatellites or simple sequence repeats (SSRs), single nucleotide polymorphisms (SNPs), insertions and deletions (indels) and genomic rearrangements (Li et al. 2011c). Identification of genetic diversity in species which lack significant genomic resources has typically been a time-consuming and laborious process.

SSR markers have been used extensively for population genetics and genome mapping studies in several members of Rosaceae (Celton et al. 2009; Aranzana et al. 2003). SSR identification techniques are typically costly and time consuming (Rafalski and Tingey 1993; Zeid et al. 2009; Zane et al. 2002). Most published SSRs are located in the intergenic regions (Li et al. 2011c). A recent study in *Populus* attempted to identify SSRs in exons or expressed gene fragments. The abundance of microsatellites within the coding region was three-fold lower than intergenic regions and, when present, microsatellites do not show useful allelic variability. Further, the authors concluded that candidate gene approach for development of microsatellites may not be the best strategy (Li et al. 2011c). While SSRs remain difficult to develop, SNP identification and validation has rapidly improved in past years mostly due to reduction of sequencing costs. Previously, direct sequencing of a gene of interest related to supernodulation was used to identify SNPs (Kim et al. 2005). Similar studies in non-model species lacking such resources require sequence information from related species. SNPs have also been used for anchoring a linkage map and bovine genome (Nilsen et al. 2008). Ganal et al. (2009) reviewed recent
SNP identification methods including DNA arrays, amplicon sequencing, mining existing EST resources, and using sequence data generated with second generation sequencing technologies. Compared to other methods, re-sequencing applications were determined to produce a higher percentage of validated SNPs, while non-reference based next-generation sequencing, or de novo, approaches required the least amount of a priori genetic/genomic information. A major caveat of using second generation sequencing de novo is the ability to acquire sufficient depth to accurately identify SNPs. Therefore, a reduced representation sequencing approach was recommended. Many reduced representation methods integrating high throughput sequencing are discussed by Davey et al. (2011) and the authors further elaborated on the utility of SNP-based molecular markers.

Continued improvements in second generation DNA sequencing technologies have increased the ability to obtain significant sequencing depth in a rapid and cost efficient manner, compared to Sanger sequencing approaches (Shendure and Ji 2008). Bundock et al., (Bundock et al. 2009) performed amplicon sequencing on genes of interest with 454 technology to produce a large number of reliable SNPs from two parents of a QTL mapping population of sugar cane finding high success rates for SNP verification (93%). Recently, next generation technologies have been widely utilized for sequencing transcriptomes of various species (Folta et al. 2010; Isom et al. 2010; Cantacessi et al. 2010). Eveland et al. (2008) reported a quantitative transcriptomics approach based on selective sequencing of the 3’UTR of mRNA from Zea mays. Their work demonstrated a clear ability to resolve the expression of nearly identical genes (99% nucleotide identity) based on variation in the 3’UTR (97% nucleotide identity). Through comparison with sequences in multiple maize databases, 93.8% of the SNPs identified by Eveland et al. were confirmed (Eveland et al. 2008). Use of a 3’UTR directed approach exploits the higher number of variations found in the 3’UTR region compared to the coding region of a gene. Higher sequence variation, combined with physical linkage to a specific gene, increases the potential impact of 3’UTR polymorphisms in connecting genetics and functional genomics studies especially in non-model eukaryotes. This is in contrast to current approaches where intergenic polymorphisms are used for scoring a segregating phenotype without the associated gene-related information. The method presented here
utilized the positive aspects of 3’UTR sequencing, as a reduced representation approach, to facilitate rapid gene-linked SNP identification.

In addition to identifying polymorphisms, current research in human genomics has demonstrated the utility of developing haplotype information as a way to more fully understand genotype to phenotype relationships, especially in context of health, disease and response to environmental cues (Tewhey et al. 2011; Suk et al. 2011; Deloukas and Bentley 2003). Generally, haplotypes are comprised of allelic variants on each of the two chromosomes at the same locus, though the definition and utilization varies in application from linking multiple polymorphisms across several loci down to multiple polymorphisms in a single gene (Johnson et al. 2001). Additionally, haplotype determination has been aided by DNA strand specific or genomic phase-based information generated using second generation sequencing technologies since each sequencing read is from only one homologous chromosome and not a consensus of the two (He et al. 2010). Similarly, next generation RNA-seq and 3’UTR sequencing has the ability to reveal haplotypes within a gene (Snyder et al. 2009) and thus enable identification of allele specific sequence and its expression simultaneously.

Here we present our approach that utilizes 3’UTR sequencing to rapidly develop SNP and haplotype markers in sweet cherry, a species without a published genome sequence and a non-model crop. Through de novo assembly of 454 generated-3’UTR sequencing reads and strict filtering, we initially identified a putative set of contigs containing SNPs. Primer sets designed to amplify the regions of these contigs with putative SNPs were developed and used for High Resolution Melting (HRM) analysis among eight currently utilized parental cultivars of sweet cherry and 13 hybrid seedlings derived from a cross between two of the parental cultivars, respectively. We determined that 68 out of 223 (30.5%) and 65 out of 217 (30.0%) of the tested primer pairs are able to detect genetic variability. From these polymorphic sites, 685 haplotypes were identified from 301 contigs containing multiple SNPs.
Methods

RNA Extraction and cDNA preparation: Tissue samples from developing floral buds of two commercially important cherry cultivars, Bing and Rainier, were excised from the trees and flash frozen in liquid nitrogen. The frozen tissues were pulverized uniformly in a SPEX SamplePrep 6870 FreezerMill (SPEX SamplePrep, Metuchen, NJ) for five cycles each with cooling for two minutes and grinding at 15 counts per second for four minutes. Total RNA from each sample was extracted using the RNeasy Plant DNA Extraction Kit (Qiagen, Germany). First strand cDNA was then synthesized using the Ambion aRNA synthesis kit with a biotinylated poly-T B-adaptor [see 4B.Additional File 1 for adaptor sequences] for 3’UTR profiling as described by Eveland et al. (2008). Second strand cDNA was created, cleaved with MspI, and ligated to modified A-adaptors containing indexing tags [see 4B.Additional File 1 for adaptor sequences] as per the Eveland protocol.

Sequencing and assembly: The 3’UTR libraries were sequenced as per the 454 FLX protocol (Roche, USA) on a single LR-70 sequencing plate. After sequencing, the 454 produced reads were processed using a custom script [see 4B.Additional File 2] to remove the multiplexing barcode and rename each read with its appropriate sample name at the end of the header. All of the modified reads were then assembled using SeqMan from the Lasergene 7 suite (DNASTAR, Madison, WI).

SNP Identification: For method development, a total of 10,100 contigs were examined for the presence of putative SNPs using Lasergene 7’s SeqMan (DNASTAR, Madison, WI). The high confidence SNPs have at least two alleles represented by a minimum depth of three reads per nucleotide call per allele. Primer pairs flanking potential SNP loci were designed using the PRIMER3 program (Rozen and Skaletsky 2000) to amplify 50-100 base pair amplicons. This yielded 223 primers from regions of 207 contigs for HRM analyses.

Population Variation Screen: Eight sweet cherry cultivars: Bing, Chelan, Emperor Francis, New York 54, Regina, Selah, Stella and Cowiche used as parental material in the Washington State University (WSU) Sweet Cherry Breeding Program (Prosser, WA) were used to test the polymorphisms of the
identified SNP loci across Bing and Rainier cultivars. For segregant analysis, 13 seedlings from an F$_1$
mapping population of Selah x Cowiche were used. Leaves of these accessions were collected from the
WSU Irrigated Agriculture Research & Extension Center in Prosser, WA and DNA was extracted from
dried leaves using a CTAB extraction protocol (Doyle and Doyle 1990). The reaction mixture for HRM
analysis consisted of 0.6 μL of each primer (10µM), 12.0 μL SYBR® Green, 5 ng of genomic DNA and
autoclaved nanopure water to a total volume of 20 μL. The Cultivar panel comprised of 223 primer sets
tested on all eight parental cultivars and the Seedling panel included 217 primers sets tested on one
reaction of each parent, Cowiche and Selah, and one of each hybrid seedling. Analyses were performed
on the LightCycler® 480 System (Roche Branford, CT) using the following PCR cycling and HRM
conditions. Initial melting for 10 minutes at 95°C was followed by 45 cycles of 95°C for 10 seconds,
57°C for 15 seconds, and 72°C for 15 seconds, then heated to 95°C for 1 minute and cooled to 40°C. High
Resolution Melting analysis was then automatically initiated whereby the amplicons were heated from
60°C to 90°C with 25 acquisitions per degree. As the temperature slowly increased, the dye fluorescence
was recorded, plotted and later analyzed using the LightCycler® 480 Gene Scanning Software. Since the
$T_m$ can vary based on the HRM reaction conditions, curve shapes were visually examined and the number
of distinct curve profiles was identified for each primer set.

**Secondary Assembly and SNP reporting:** After the HRM analysis, a second assembly using SeqMan
NGen v3.0 (DNASTAR, Madison, WI) was performed due to its improved algorithm and the results were
used for SNP reporting on the entire data set. This assembly was completed using the default parameters
for NGen 3.0’s *de novo* transcriptome assembly of: 85% match, match size 21, genome length 225MB.
The whole SNP report was initially filtered to retain the HRM confirmed SNPs using a minimum total
depth of 10 reads at the polymorphic base and at least 20% variance from the consensus. Further filtering
into high confidence SNPs was performed by screening for at least two alleles represented by a minimum
depth of three reads per nucleotide call per allele. This minimum depth per allele for each SNP equals or
exceeds the published depths using either 454 data (Emrich et al. 2007; Kulheim et al. 2009) or Illumina
data (Hyten et al. 2010; Varala et al. 2011). Additionally, SNPs resulting from the first or last five bases
of reads were rejected. The transition and transversions ratio (R value) was determined by summing all of the transitions (C/T and A/G) and transversions (A/C, A/T, C/G, and G/T).

**Haplotype Identification:** Haplotypes were identified visually by analyzing the combined transcriptome assembly generated using NGen 3.0 in SeqMan (DNASTAR). Similar to the SNP screening, at least three reads of an allele spanning two SNP loci were required to link SNPs into a haplotype. When two or more haplotypes were present at one locus, they were differentiated and recorded as separate haplotypes for their use as haplotype markers.

**Results and Discussion**

**Method Overview:** The general method presented in this study is based on four steps as outlined in Figure 1. The first step of sample preparation involves identification of appropriate individuals across whom genetic polymorphism needs to be determined. In our study, we used two closely related sweet cherry cultivars to test our approach. However, it is recommended that phenotypically diverse individuals should be chosen. Additionally, the number of individuals can be increased as desired keeping in mind the expected transcriptome size and the number of sequencing reads expected to be generated by the next generation sequencing platform that will be employed for transcriptome sequencing in step 2. This parameter is critical for strict filtering of data for identification of SNPs in step 3. Total RNA needs to be extracted from tissues which are representative of the phenotypic diversity between the samples. Developing reproductive buds used in this study were derived from Bing and Rainier each grafted onto two rootstocks Mazzard and Gisela 6. Bing and Rainier grafted on Gisela 6 yielded fruit that was 656% to 212% more than the same cultivars grafted on Mazzard (Whiting et al. 2005a). The RNA is converted into cDNA and further processed for selection of 3’UTRs (Eveland et al. 2008). In step 2, after extensive quantification of the 3’UTR libraries, samples are pooled in equimolar ratios and sequenced using next-generation sequencing platforms. At the time, we performed pyrosequencing on the 454 GS FLX instrument since it provided the longest read lengths. However, at present, such a method would benefit greatly from Illumina or SOLiD platforms since the read lengths have greatly improved (Xiong et al.
Depending on the sequencing platform the raw sequence data needs to be pre-processed by trimming of tags and adaptor sequences prior to moving to step 3 of data processing where the sequence data is assembled. We used the NGen v3.0 (DNASTAR, Madison, WI) assembler and the output was visualized using SeqMan which generated a SNP report. The final set of SNPs was selected using strict parameters as outlined in the materials and methods. In step 4, putative SNPs were tested for variability across 8 parental cultivars and 13 progeny derived from a cross between two cultivars using HRM analysis. Utilization of SNPs for screening variability in population has been well documented in literature (Chaisan et al.; Li et al. 2011b; Mader et al. 2010). Subsequently, for SNP validation, barcoded amplicon sequencing for a very large number of markers (SNP or haplotype) across a large array of progeny in a segregating population or genetic collection would be an efficient approach. For smaller number of samples or for initial confirmation of variation, techniques such as HRM may be more appropriate as utilized in this case. Rapid identification of gene-linked polymorphisms as proposed in this method can facilitate efficient Gene Assisted Breeding (GAB), genotyping and population genetics studies in non-model eukaryotes.

**Sequencing and Assembly of 3'UTRs:** Pyrosequencing of 3'UTR libraries from Bing and Rainier on a single 454 GS FLX sequencing plate produced a total of 580,455 reads (Table 1). The reads had an average length of 85 bp which is as expected from the 454 GS FLX sequencing platform and the 3'UTR library preparation. The reads were processed with a custom script to trim index sequences and label the headers appropriately [see 4B. Additional File 2]. Transcriptome assembly of the trimmed sequences with SeqMan 7.0 (Lasergene Suite 7.0, 2009) yielded 43,380 contigs.

**Initial SNP identification:** To test our experimental approach, analysis of a subset of the assembled contigs was performed to identify SNPs within the dataset. The 100 contigs with the highest number of reads and contigs 1-10,000 as produced by SeqMan 7 (Lasergene Suite 7.0, 2009) were analyzed yielding 600 contigs containing at least one high confidence SNP. These high confidence SNPs have at least two alleles represented by a minimum depth of three reads per nucleotide call per allele. Since false polymorphism of indels can be high (Margulies et al. 2005), indels were not included in this analysis to
avoid identification of false polymorphisms as previously described (Barbazuk et al. 2007). The total number of SNPs in this dataset was not calculated as only the described subset was examined. A total of 223 primer sets were designed from 207 contigs with PRIMER3 (Rozen and Skaletsky 2000) to amplify the small regions around the identified SNPs [see 4B, Additional File 3 for primer sequences and associated contigs].

**Population Variation Screening:** The automated genotype calling of the LightCycler 480 analysis software v1.5.0 identified only a few SNPs with more than one allele. Modifying the analysis parameters did not provide significant improvement of the automated analysis (data not shown). However, manual analysis identified multiple curve types for many primer sets as well as heritable patterns between Cowiche, Selah, and their seedlings. While the differences in melting curve shape are small, homozygotes and heterozygotes were visibly distinguishable with many of the primer sets (Figures 2A & B). It is unclear why the HRM curves presented in this manuscript differ from those shown by Wu et al. (Wu et al. 2008). These smaller changes in the derivative plots could be due to the larger amplicon size (~ 150 bp).

Manual analysis of the Cultivar and Seedling tests indicated that 68 out of 223 (30.5%) and 65 out of 217 (30.0%), respectively, of the designed primer pairs displayed variation with 49 pairs showing variation in both tests (Table 2). This is expected as it is recommended to design three primer sets for each SNP of interest according to ABI’s guide to HRM (Applied-Biosystems 2010) analysis which suggests that a success rate of 33% is typical. A total of 23 primer sets from the Cultivar panel and 19 from the Seedling panel were considered non-variant for this experiment since they displayed indiscernible variation (Figure 2C). Additionally, it became evident during the analysis that multiple SNPs in an amplified region made distinction more difficult, though it was still possible in the best cases (Figure 2D-F). Eight of the non-variant primer sets were shared between the two panels. Reactions which did not produce a curve in either panel were labeled ‘failed’. Some of the failed primer sets produced amplicons on one of the two panels suggesting amplification issues. Ten primer sets failed in both panels, most likely due to an error either in the contig sequences or the primer design. Overall, 84 of the 217 (38.7%) primer sets used on both panels showed variation in one or both sets. The remaining 61.3% of the SNPs did not have
detectable variation in the individuals tested. One explanation for this is that the tested cultivars mathematically only represent 12.5% of the alleles from Rainier’s paternal parent, Van, based on the pedigrees of the tested cultivars (Figure 3). Alternatively, lack of detection may be a result of the amplicon length hindering the ability to visualize the melting differences between variants. This variation detected by HRM was far lower than the detection from amplicon sequencing of sugar cane though the sugarcane work focused on genes of interest whereas we used a de novo approach (Bundock et al. 2009). The authors had screened for SNPs in polyploid parents and the resulting progeny. It is critical to note that in this work, we identified SNPs from two cultivars and then validated them across 8 parental cultivars, 6 of which are not closely related. Additionally, the progeny used for SNP variation screening are far removed from the genotypes used for initial SNP discovery. Most importantly, sweet cherry has a narrow genetic diversity further reducing the possibility of identifying a large number of SNPs. Our work clearly illustrates that sequencing and assembly based method for identification of SNPs is highly effective and that the HRM screen is likely a limiting step. Heritability of the curve types can also reveal cultivars that are homozygous at a given locus (AA x AA) or heterozygous (AB x AB) (Figure 4A & B). Additionally, the Seedling HRM curves can confirm that one parent is homozygous and the other heterozygous with an approximately 1:1 ratio (8:5) of curve types matching the two parents (Figure 4C). Though higher numbers of individuals need to be tested to obtain statistical significance, noting that these patterns are distinguishable through HRM provides a foundation for the use of this method to screen progeny or parents to determine their allelic composition.

**Secondary Assembly and SNP reporting:** DNA assembly programs continued to improve since the initial assembly which was used to design primers and analysis of population variability. Subsequently, the trimmed reads were re-assembled using NGen v3.0 (DNASTAR, 2011). This assembly produced 34,620 contigs [see 4B.Additional File 4] with an average length of 149 bp (Table 3). Since the aim was to obtain high depth of coverage of around 100 bp upstream of the poly-A tail, the longer contigs were unexpected. Analysis of this issue confirmed that the poly-T sequence containing primer used for first strand cDNA synthesis annealed to some poly-A regions in coding regions of the transcripts as well.
While not all the sequencing reads were from the direct vicinity of the poly-A tail, the contigs remain gene-linked due to their cDNA origin. This could reduce the total number of identified SNPs since genic regions have a greater selection against mutations when compared to the 3'UTR as previously described (Eveland et al. 2008).

After filtering the new SNP report for a minimum read depth of 10 and 20% variance from the consensus an initial list of SNPs was derived. These SNPs were examined to remove any SNPs resulting from the ends of the reads and filtered requiring a minimum of three confirming reads per base call per allele. A total of 2243 putative SNPs were identified in 887 contigs after this filtering [see 4B,Additional File 5]. These data, consisting of contigs and SNPs, have been uploaded to NCBI (GenBank JP376615-JP382830 and dbSNP NCBI ss# 469992783-469995036 except 469992784, 469992792, 469992801, 469992809, 469992818, 469992823, 469992825-7, 469992834-5, 469992842, 469992851, 469992853-4, 469992859, 469992866-7). Analysis of the “failed” HRM primer sets on contigs obtained from the NGen assembly showed that none had a significant change in the contig consensus. However, three of the 10 primers did show multiple possibilities for primer binding which could decrease PCR effectiveness.

The 2243 putative SNPs identified in the assembled gene space (expressed sequences) of 5.19 Mb yields a SNP frequency of 1 in 2,315 bp (0.43 SNPs per kb of gene space). The sweet cherry gene space of 5.19 MB generated in this study represents approximately 2.3 – 1.7% of the estimated genome size of 225 – 300 MB. Previous studies utilizing whole genome sequence have reported a frequency of 1 SNP in 114bp (8.8 SNPs per kb) and 1 SNP in 208 bp (4.8 SNPs per kb) in almond *Prunus armenica* (genome size = ~200MB) and apple *Malus x domestica* (genome size = 740MB) respectively (Wu et al. 2008; Velasco et al. 2010). The recent genetic bottleneck and Bing being a parent of Rainier reduces the number of potential alleles present in the dataset to 3 whereas the almond and apple studies examined 25 and 5 cultivars respectively.

As mentioned earlier, coding regions of genes were also sequenced inadvertently since the poly T primer annealed to regions other than the 3'UTR region, thereby further reducing the number of polymorphic sites in the sequenced regions. Analysis of the putative sweet cherry SNPs for transitions
(C/T and A/G) and transversions (A/C, A/T, C/G, and G/T) yields a transition to transversion ratio (R value) of 1.14/1 (Table 4). This is nearly identical to the 1.16:1 ratio found across 25 almond cultivars (Wu et al. 2008) and differs slightly from the 1.27:1 ratio in *Prunus mume*, Japanese apricot (Fang et al. 2006).

**Haplotype Identification:** From the final SNP report, contigs possessing more than one high quality SNP were analyzed for the presence of haplotypes. The sequence and base position for each distinguishable haplotype of the contig were detailed [see 4B.Additional File 6]. An example of a haplotype containing 10 SNPs at a single locus (Figure 5) demonstrates clear differences between the two haplotypes. In total, 301 contigs contained at least two haplotypes and 15 had more than two haplotypes in a region. Additionally, 34 contigs had multiple, unlinked haplotype regions that likely represent sections of haplotypes that, due to lack of read length or inadequate depth of sequence reads and the previously stated requirements, were not able to be linked in this analysis. In total, this amounts to 685 unique haplotypes over 335 loci in 301 contigs. Development of these haplotype blocks is expected to greatly benefit sweet cherry breeding efforts specifically, but warrant consideration for future phylogenetic and comparative genomic studies in other related species as well. As haplotypes, these SNP blocks also represent loci that may be extremely useful for development of molecular markers like CAPS. Since these haplotyped SNPs are inherited as a block, future studies would benefit from a higher depth of coverage to ensure complete linkage of haplotype blocks. It is acknowledged that the linked SNPs are very close in the short contigs, and they tend to be haplotypes due to low probability of recombination between them. However, such haplotypes are highly relevant to the current short read sequencing platforms where shorter reads of 50 to 100 bp can be utilized to accurately identify an allele in a diploid or polyploid sample or detect mutations that may occur individually creating a new haplotype.

**Access to Sequence and SNP data:** Due to the nature of the contigs and SNPs, many of them did not fit the requirements for typical submission to NCBI. All contigs and high quality SNPs are available as additional files to this manuscript. All of the raw sff files were uploaded to NCBI’s Sequence Read Archive (SRA046001.1). Contigs greater than or equal to 200bp in length were added to GenBank’s
Transcriptome Sequence Assemblies (TSA) database (GenBank JP376615-JP382830) as *Prunus avium* assemblies and SNPs corresponding to these sequences larger than 200bp were uploaded to dbSNP (NCBI ss# 469992783-469995036 except 469992784, 469992792, 469992801, 469992809, 469992818, 469992823, 469992825-7, 469992834-5, 469992842, 469992851, 469992853-4, 469992859, and 469992866-7).

**Conclusions**

A method for developing gene-linked SNP and haplotype markers through high-throughput 3’UTR sequencing for species lacking genome sequences was demonstrated. Through this process, 2243 putative SNPs were identified and 34,620 contig sequences were obtained and added to NCBI database for use by the plant research community. To our knowledge, the 685 haplotypes developed in this study are the largest set of reported SNP-based haplotypes in sweet cherry and demonstrates that haplotypes can be identified using 3’UTR sequencing. These haplotypes can be utilized for the development of CAPS markers to resolve allelic differences in 301 sites on the sweet cherry genome. These genomic resources represent a large advance in sweet cherry genomics. Potential applications of these SNPs may involve high-throughput amplicon sequencing with these primer sets using next generation sequencing technologies to obtain digital or sequence-based information in genetics studies. This is in contrast to the SNP-arrays that produce an analog signal in genotyping experiments and represent mostly intergenic polymorphisms derived from a few individuals limiting its potential applicability beyond the included polymorphisms. This methodology is expected to be of great utility in polyploid species where allele-specific haplotypes can be highly informative.

As sequencing costs plummet, the general approach reported here could be broadly implemented in identifying gene-linked polymorphisms amongst parental individuals which can then be rapidly utilized in segregation studies of a desirable set of phenotypes in the derived progeny. Polymorphisms that co-segregate with the phenotype are expected to represent the gene or set of genes that regulate the said phenotype. Establishment of these correlations is expected to open avenues for directly linking genetic
and functional genomics approaches with phenomics, an emerging discipline focused on understanding genotype-phenotype relationships.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TK, SS, NO, MW and AD designed the study. TK and SS prepared samples for 3’UTR sequencing. TK and VK performed assemblies and computational analysis. DJ guided and designed primers. TK analyzed the HRM curves. AH aided the computational analyses and managed data upload to NCBI. MW aided in designing primary sample collection. NO provided sweet cherry parental cultivars, segregating progeny, and guided HRM analysis. AD supervised the research and guided data interpretation. TK and AD wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

This project was supported by WSU startup funds to AD and NO and Washington Tree Fruit Research Commission funds to AD and MW. TK and SS acknowledge support received from NIH Protein Biotechnology Training Program and ARCS fellowship. VK and AH were supported by US Department of Agriculture National Research Initiative (USDA-NRI) grant 2008 –35300-04676 to AD. Technical assistance of Dr. Kahraman Gurcan for preliminary SNP analysis, primer design and performing of HRM experiments is gratefully acknowledged. We are thankful to David Rockefeller and Devaswa Bhagawati for their assistance in SNP analysis and primer design. Authors are thankful to Dr. Ted Kisha, USDA-ARS and Dr. Katherine Evans at Tree Fruit Research and Extension Center, WSU for useful discussions and critical reading of the manuscript. We are thankful to anonymous reviewers for their excellent suggestions.
Works Cited:


2. Peach Genome v1.0, International Peach Genome Initiative [http://www.phytozome.net/peach.php#A]


Table 1 – Summary of 3’UTR sequencing results

The table represents number of bases, reads and average read lengths generated for Bing and Rainier cultivars.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Bing</th>
<th>Rainier</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of bases</td>
<td>25409323</td>
<td>23893350</td>
<td>49302673</td>
</tr>
<tr>
<td>Number of reads</td>
<td>303684</td>
<td>276771</td>
<td>580455</td>
</tr>
<tr>
<td>Avg. read length</td>
<td>83</td>
<td>86</td>
<td>85</td>
</tr>
</tbody>
</table>
Table 2 – Experimental assessment of SNPs

A total of 223 predicted SNP sites were tested via HRM in 8 cultivars and 217 predicted SNP sites were tested in 13 seedlings derived from two of the cultivars. The table represents the results of these HRM analyses.

<table>
<thead>
<tr>
<th>Type</th>
<th>Cultivar panel</th>
<th>Seedling panel (Selah x Cowiche)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number tested</td>
<td>Number with variation</td>
</tr>
<tr>
<td>SNP primer sets</td>
<td>223</td>
<td>68</td>
</tr>
</tbody>
</table>
Table 3 – Transcriptome assembly results

Statistics on the transcriptome assembly for the sequence data generated for Bing and Rainier developing floral buds.

<table>
<thead>
<tr>
<th>Assembly Version</th>
<th><strong>NGen v3.0</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of contigs</td>
<td>34,620</td>
</tr>
<tr>
<td>Avg. contig length</td>
<td>149</td>
</tr>
<tr>
<td>Median contig length</td>
<td>118</td>
</tr>
<tr>
<td>Total contig bases</td>
<td>5,191,475</td>
</tr>
<tr>
<td>Number of putative SNPs</td>
<td>2,243</td>
</tr>
</tbody>
</table>
Table 4 – Summary of transitions and transversions across Bing and Rainier

Each SNP was classified based on the base change that occurred. The total number of transitions (a sum of the C/T SNPs and A/G SNPs) is 1193 which is marginally greater than total number of transversions (A/C + A/T + C/G + G/T) which is 1046. The R value (transitions/transversions) is 1.14 as expected within a species.

<table>
<thead>
<tr>
<th>Transitions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C/T</td>
<td>598</td>
</tr>
<tr>
<td>G/A</td>
<td>595</td>
</tr>
<tr>
<td>Total</td>
<td>1193</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transversions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A/C</td>
<td>242</td>
</tr>
<tr>
<td>A/T</td>
<td>348</td>
</tr>
<tr>
<td>C/G</td>
<td>158</td>
</tr>
<tr>
<td>G/T</td>
<td>298</td>
</tr>
<tr>
<td>Total</td>
<td>1046</td>
</tr>
</tbody>
</table>
Figure 1 – General schema for rapid identification of SNPs. The method consists of four stages 1. Sample preparation, 2. Sequence data generation, 3. Data processing and 4. Variation screening or validation of polymorphism. Content in parentheses denotes the materials, software and methods used in this study. The variable polymorphic regions can facilitate efficient gene assisted breeding (GAB), genotyping and population genetics studies.
Figure 2 - Analysis of variation of identified SNPs via high resolution melting (HRM) curves generated on 8 cultivars used in this study. HRM derivative plots, -(d/dT) fluorescence as a function of temperature, of several primer sets when analyzing 8 sweet cherry cultivars representing the common patterns observed during analysis. Comparisons outside one frame are not meaningful and the frames are not to scale with each other as the curve shape is the focus. **A-C** are from primers amplifying a region expected to contain 1 putative SNP while **D** contains 2, **E** has 3, and **F** contains 5. **A.** Primer set 121 produces a single curve pattern denoted by an arrow representing either a homozygous locus across all 8 cultivars tested or a heterozygous locus shared by all 8 tested cultivars. **B.** Primer set 100 has three distinct curve patterns highlighted as 1, 2 and 3 representing three allelic forms at the sampled locus. **C.** Primer set 115 has an indiscernible pattern. **D-F.** Each demonstrates variation in the population; however, the more SNPs present in the amplified region the smaller the differences among the melt curves.
Figure 3 - Pedigree relationships of the 8 cultivars used in this study. Pedigree of the sweet cherry cultivars used for SNP development (blue box) and those used for HRM analysis of the SNPs (bold and all caps). The maternal parent is marked by a red line and the paternal parent by a blue line.
Figure 4 – Four primer sets with the HRM curve for the two parents, Cowiche and Selah, on the left and the 13 seedlings on the right. HRM derivative plots, \(-\frac{d}{dT}\) fluorescence as a function of temperature. A-C contain 1 putative SNP while D. contains 2 putative SNPs. A. Primer set 131 shows no variation as expected for crossing two of the same homozygotes. Note the single curve profile in both the parents and seedlings. B. Primer set 189 shows a single curve profile in the parent panel which differentiates into a 1:2:1 (3:6:4) genotype ratio as expected for a heterozygous x heterozygous cross in the seedlings. It is represented by three different curve profiles where profile number 1 corresponds to the heterozygous parental profiles. C. Primer set 100 shows a 1:1 (8:5) genotype ratio as expected in a homozygous and heterozygous cross. The curve profiles are similar between the seedlings and the parents. D. Primer 92 shows 2 parental curve types. However, the seedlings show several distinct curve types which is not unexpected due to the presence of 2 high quality SNPs in this region.
Figure 5 - Screenshot of the SeqMan (DNASTAR) visualization of contig456 showing 2 alleles at a single locus. Boxes 1 and 2 represent unique haplotypes obtained from the NGen 3.0 assembly of the 454 reads from Bing and Rainier according to the filtering parameters described in the methods. These haplotypes differ at each of the bases labeled in green on one of the haplotypes for a total of 10 SNPs between these haplotypes. Haplotype 1 consists of 11 reads from Bing and 1 of Rainier while haplotype 2 is entirely Rainier.
Additional files

4B.Additional_File_1
File format: XLS
Title: Adaptor sequences for 3’UTR sequencing
Description: Sequences of adaptors used in the 3’UTR sequencing of cDNA. AMID-B is an oligo-dT primer with a biotinylated 5’end. Adaptors AMID-1A to AMID12-B represent complementary oligonucleotide pairs with embedded barcode sequences. Column A is the primer name and B is the sequence.

4B.Additional_File_2
File format: PL
Title: sup2.pl
Description: Custom script used to remove index sequences and rename the header with the appropriate sequence.

USAGE: sup2.pl {Reads FASTA format file} {Primers/MIDs/Barcodes with corresponding headers in csv format} {# bases from start of primer to the beginning of the barcode} {New FASTA filename to be written into}

Example:
Input (fasta file):
>1300_8769_5430 length=258 urmand=JHSK987KJSH2KJHJK8777
AGTCCCCGGGGTTTTAAAGGGGCCCCCTTTTAAAAAAGTCGTCAATGCGGTAGTCTGCAAAAA
AATTTCGCCCGGGGGGGGGGTAGCCGTATGCA

Input (MIDs csv file):
Sample1,ATAGTGA
Sample2,ATGCATG

Output: A fasta file of the remaining sequence after removing the primer/bar code/MID with corresponding header attached as specified in the input "MIDs csv" file

4B.Additional_File_3
File format: XLS
Title: Primers and HRM analysis
Description: The table represents contig number (column B), predicted amplicon length (column C), number of SNPs (column D), forward and reverse primers for each set (column E and F) used for HRM analysis. Included in the table is the Cultivar number of curve profiles (column G), number of Cowiche x Selah curve types (column H) and the Seedling number of curve profiles (column I).
4B.Additional_File_4
File format: FAS
Title: Contig sequences
Description: A fasta file containing the 34,620 contigs from NGen v3.0

4B.Additional_File_5
File format: XLS
Title: Filtered SNP report
Description: This table is modified output generated from NGen v3.0 and SeqMan. The contig number and all details about the SNP are given including number of calls for each base at the given position from Columns B-L. Column M is the 5’ flanking sequence. Column N is the polymorphism. Column O is the 3’ flanking sequence. Columns M and O have been provided to enable rapid analysis of other germplasm.

4B.Additional_File_6
File format: XLS
Title: Haplotypes identified in sweet cherry
Description: The table presents different haplotypes identified in each contig. Some contigs have multiple positions indicated as A, B or C positions. Nucleotides corresponding to a given position in an allele are presented. Cells are merged when the differences between alleles are no longer traceable. A questions mark (?) symbolizes incomplete depth for a confirmed call at this base.
CHAPTER 4C: TRANSCRIPTOMICS

CisSERS: Customizable in silico Sequence Evaluation for Restriction Sites

Tyson Koepke¹², Richard M. Sharpe¹³, Artemus Harper², John Grimes⁴, Marco Galli⁷, Mio Satoh-Cruz⁵, Anantharaman Kalyanaraman⁴, Katherine Evans², David Kramer⁵, and Amit Dhingra¹²*

In review at Nucleic Acids Research

¹ Molecular Plant Sciences Graduate Program, Washington State University, Pullman, WA, 99164, USA
² Department of Horticulture, Washington State University, Pullman, WA, 99164, USA
³ School of Biological Sciences, Washington State University, Pullman, WA, 99164, USA
⁴ Electrical Engineering and Computer Science, Washington State University, Pullman, WA, 99164, USA
⁵ MSU-DOE Plant Research Laboratory, Michigan State University, East Lansing, MI, 48824, USA

* To whom correspondence should be addressed. Tel: 509-335-6586; Fax: 509-335-8690; Email: adhingra@wsu.edu

The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors.

ABSTRACT

High-throughput sequencing has produced an immense volume of data containing biologically relevant nucleotide polymorphisms. Tools that leverage these polymorphisms and resulting changes to restriction sites in a high-throughput manner have yet to be created. CisSERS was developed to analyze large sequence datasets as a standalone tool and provide biologists with DNA digest information including site locations and gel predictions. These data enable the user to analyze and make decisions for designing genotyping by sequencing, reduced representation sequencing, 3'UTR sequencing, and cleaved amplified polymorphic sequence (CAPS) markers for large sequence populations. Using a fasta-formatted file as input, CisSERS has several functionalities available to the user and utilizes the REBASE enzyme database. CisSERS is a java based graphical user interface built around a perl backbone. Wet-lab confirmation was done for several of the applications of CisSERS including CAPS marker development. Here, we present the tool CisSERS, the usage algorithm, and results from in-silico and wet-lab analyses demonstrating that CisSERS is a tool that will facilitate efficient data utilization in genetics and genomics studies.
INTRODUCTION

In the past few years, high-throughput sequencing technologies have generated vast amounts of data that have overwhelmed many of the currently available data analysis tools. The lack of computational tools specifically created to process large quantities of sequence information has limited the translation of these countless data into useful knowledge for addressing biological questions. Existing sequence data can be harnessed for nucleotide polymorphism information, ascertaining genetic diversity in a population, and reduced representation sequencing. The consequences of nucleotide polymorphisms are diverse. They might result in altering the phenotype if there is a change in an amino acid or alterations in the regulatory regions. Alternatively, these may be inconsequential mutations. Biologists endeavor to first identify and then utilize the polymorphic information to establish causal relationships between the genotype and the phenotype in genomics and genetics approaches.

There are several approaches in use that exploit nucleotide polymorphism information. On a global genomic scale, reduced representation sequencing (1-3), genotyping by sequencing (4,5), and SNP arrays (6,7) are commonly utilized. Genotyping by sequencing and reduced representation sequencing can both utilize restriction site information during sequencing library preparation for genomes (2,3) and transcriptomes (8-10). An example is Restriction-site Associated DNA (RAD) sequencing that enables identification of polymorphisms which are subsequently used as DNA markers for population analysis (11,12). Whole genome analysis of restriction sites can provide better information to help guide decisions for enzyme selection in digesting DNA for RAD sequencing libraries as well as BAC library production and sub-cloning. Specific sequencing of 3’UTRs was enabled through the utilization of restriction enzyme digests in maize (8) and sweet cherry (9).

At the resolution of a single gene or several genes, the afore-mentioned approaches are overkill. For such instances, methods such as high resolution melting (9), allele specific PCR (13), single locus restriction fragment length polymorphisms (RFLP), amplified fragment length polymorphisms (AFLP)
and cleaved amplified polymorphic sequences (CAPS) (14) are used. Of these, methods based on restriction enzyme digestion are relatively easy, widely used, and cost-effective to perform and analyze due the reduced need for specialized equipment and expertise. For genetics applications, restriction enzyme based molecular markers are commonly employed. Historically, these molecular markers have been developed from polymorphisms among a limited set of individuals. With the availability of high-throughput sequencing technologies, the onus has shifted to identifying multiple site specific polymorphisms across large populations. CAPS markers, also known as PCR based Restriction Fragment Length Polymorphic markers (PCR-RFLP), are routinely used for mapping traits in populations and for enabling efficient breeding (14). CAPS markers are popular due to their relatively low cost and general ease of use through the reliance on the common, simple molecular biology tools of PCR, enzymatic digests and gel electrophoresis (15). These strategies, however, rely on a priori knowledge including the location and sequence of the restriction sites. In addition to polymorphism screening, many types of molecular biology methods utilize restriction site information and therefore require an efficient tool to analyze large sequence datasets.

Several current restriction site analysis tools, summarized in Table 1, have been designed to handle one or several small sequences for targeted analysis (16-22), not the thousands that are now available with high-throughput sequencing. NEBcutter is the most generalized restriction site analysis tool with the others focusing primarily on developing DNA markers. NEBcutter provides useful functionalities from cloning analysis to gel predictions based on different types of gels (22). Version 2.0 of this web tool handles a single sequence of less than 300kb for analysis regarding any number of enzymes from the REBASE database. The analysis pipeline enables most molecular functionalities but is not set up for high-throughput analysis and only shows enzymes with a specific number of recognition sites in the given sequence. This tool also enables multiple digests by incorporating buffer specificities.

Like NEBcutter, most of the previous marker developing tools are web-server based, limiting functionality for high-throughput analysis depending on the users’ internet connection and the tool’s
server availability. The need to upload large datasets to these webserver-based programs can cause a significant bottleneck. Since the marker tools are mainly designed to develop CAPS markers, several of these tools have added primer design for amplification of a region around a polymorphism modified restriction site. Additionally, several of these tools include levels of automated decision making that aids primer set and enzyme selection although this reduces user control and preferences.

The tool we present here, CisSERS: Customizable in silico Sequence Evaluation for Restriction Sites, was developed to enable high-throughput analysis of restriction sites. While CisSERS incorporates the REBASE database for restriction enzyme information like many of the other programs in Table 1, CisSERS downloads the list for use as a standalone version. CisSERS provides the digest information, including a gel prediction, to facilitate the user’s decision making process for selection of the most appropriate enzyme for the project application. CisSERS retains project files to provide easy access to the predicted cut site information to reduce time when the project requires multiple interactions, additional analysis or when two projects require comparison. Overall, CisSERS was designed to bridge the gap between sequence acquisition and valuable implementation of these diverse approaches to addressing biological questions.

MATERIAL AND METHODS

CisSERS overview:

CisSERS (4C.AdditionalFile.3.CisSERS.jar) was developed to provide complete processing of fasta files for restriction site analyses including tables of counts and predicted gel image outputs. As a standalone program, CisSERS requires a onetime download of Perl, the java archive (jar) and a default enzyme list. Using hundreds of thousands of sequences or a single sequence in a single fasta file, selected enzyme sites are identified, displayed and analyzed through a java based graphic user interface (GUI). After analysis is completed, the outputs are displayed in the program including tables describing the cut counts and locations for each restriction site and dynamically created predicted gel images. Figure 1
provides a graphical overview of the CisSERS workflow. 4C.Additional.File.1_CisSERS_user_manual contains program usage instructions.

User Input: For data entry into CisSERS, the user selects the fasta-formatted DNA sequence file for processing using a standard folder/file selection. The file is analyzed to verify that it is in the proper format and CisSERS will warn the user if irregularities are identified once it is started. For transcriptome analysis approaches where poly-A trimming and 5’ to 3’ orienting are desired, selection of the ‘Sequences have Poly-A tails’ check box enables this pre-processing. If poly-A is selected, identification and subsequent trimming of the poly-A is completed. Sequences without a recognized poly-A sequence are placed in a separate file and are not processed further. The user can also determine how far into a sequence the poly-A tail search will continue. Poly-A tail identification, based on the approach found in EMBOSS program ‘trimest’ (23), occurs if 4 or more consecutive A’s are found within the specified range. The poly-A tail search will be extended until more than 1 non-A character is identified. If a search limit has been specified, the algorithm will extend the poly-A tail beyond the range to most effectively trim the sequence. While the poly-A identification is running, the algorithm is also applied to identify poly-T heads which are reverse complemented and placed in the proper 5’ to 3’ orientation for all further analyses. The trimmed files, containing all sequences in the 5’ to 3’ orientation, and non-trimmed files generated during this process are saved for further use if desired. Without the Poly-A selection, all reads are processed in their input orientation and must be oriented in the 5’ to 3’ direction for use in non-transcriptomic approaches.

To enable the analysis of portions of transcripts from the 5’ or 3’ end for either side specific preparations or enable to analysis of least degraded transcript portions, the second step allows the user to choose the desired ‘cut site’ area. The user can choose the area of predicted cut sites from either end or along the entire length of the sequence. The 3’ and 5’ options enable the selection of a range of the sequence to be analyzed from the desired end. The user enters this information by dragging the slider or entering the exact positional information in the dialog boxes below the diagram (Figure 2).
The third and final input required prior to initiating analysis by CisSERS is the enzyme selection. The master list of enzymes is retrieved from the REBASE database (24) and can be periodically updated by the user from within CisSERS. Enzyme selection can be done through a checkbox tree, a filtering window, or through a name/site search. This allows the user to choose enzymes meeting a number of criteria including custom lists. Some of the criteria include recognition site length and overhang. Additionally, if an enzyme has any selection for or against methylation, the enzymes are labeled to inform the user to verify the appropriateness of the enzyme for their application. User defined enzymes and recognition sequences can also be added to the database if desired. The options menu also allows the user to select which outputs are desired prior to starting the analysis.

Analysis: After all user inputs are entered, selecting 'Run' at the bottom of the screen begins the analysis process. During this time, a ‘processing’ tab (Figure 3) is shown depicting the progress of the analysis through each stage. Once the analysis is complete, this tab will disappear and ‘Summary’, ‘Best’ and ‘Top’ tabs will appear when the ‘Sequences have Poly-A tails’ box is selected; when the ‘Sequences have Poly-A tails’ box is not selected, an additional ‘Gel Visualization’ tab appears.

Results/Outputs: The primary outputs of CisSERS are: Summary, Best, Top tables and Predicted Gel Visualization as mentioned previously. Each of these outputs is shown on an individual tab and can be saved individually by CisSERS. Projects can be saved and reloaded to eliminate the need to reprocess the data.

The ‘Summary’ table displays the total results for each enzyme. This includes the total number of sequences which contain at least one occurrence of that enzyme’s recognition sequence, the total number of cut sites in the fasta file and the percentage of sequences that are cut by this enzyme. The ‘Best’ table is used for finding combinations of enzymes that cut the most sequences. For 3’UTR sequencing, a minimum number of enzymes that cut nearly every transcript in the desired range are ideal. Using a greedy approach, where sequences cut by the best enzyme are removed and the next best enzyme is identified, the enzymes are listed with the combined percentage of sequences cut. Additionally, since there are minimum lengths for processing DNA through some applications, the 'Best' table also displays
cut sites in the 'Pre-cut Area', the area between the beginning of the sequence and the beginning of the desired cut area. The last table produced is the ‘Top’ table. This table is a filtered version of the ‘Summary’ table that only displays the enzymes cutting a minimum percentage of the sequences. This setting defaults to 95% and is adjustable through the options menu. These data tables combine to inform the user of the restriction site information necessary to enable many biological approaches.

The predicted ‘Gel Visualization’ tab facilitates cloning analyses, CAPS marker decision making and selection of electrophoresis strategy based on size of fragments generated for a given enzyme or a combination of enzymes. This tab allows the user to choose between two viewing modes. The ‘Standard Mode’ allows multiple sequences to be analyzed with one or more enzymes. Each sequence is shown in a single lane and the selected enzymes are processed as a multiple digest (double digest for two enzymes). Standard mode also allows the user to scroll through single enzymes one at a time to identify those that create differential patterns among the selected sequences. The ‘Single Digest Mode’ enables comparison of different individual enzyme digestions on a single sequence. In this mode, the predicted gel image is created by processing one sequence and each lane corresponds to a different enzyme.

Other Files: Since many enzymes recognize the same site, each site is only processed once and the enzymes are compiled in a ‘Neo(Iso)schizomers list’. This list is viewable to identify alternative enzymes recognizing the same sequence to enable substitutions since all recognition site data are identical among each group of enzymes. Intermediate files and a full list of cut sites for each sequence/enzyme combination are also available as outputs when desired.

Requirements: CisSERS can be run on Windows or Linux operating systems running Java 5 Standard Edition or greater and Mac running OS 10.5 or higher. Perl version 5.8.8 or greater is also required.
RESULTS

Performance Testing

To compare how CisSERS performs on different platforms, a laptop and a desktop computer were used to process different types of input files; these computer’s specifications are described in Table 2. The following files were utilized to ensure code accuracy, performance and time to completion analysis: 4 base RE, 6 base RE, and 8 base RE, Galli sequences and the E. coli strain K-12 substrain DH10B accession number NC_010473.1, with results shown in Table 2. The non E. coli input files are available on NAR online as 4C.AdditionalFile.2_Datasets. The 6-base cutter RE test file was used to analyze the time CisSERS requires to completely process files with different quantities of sequences comprising 60 bases per sequence. The size of the 6-base cutter RE test fasta files ranged from 1,101 to 49,235 sequences. Time to completion ranged from 6.3 seconds for 1,101 sequences on the desktop computer to 5,291.54 seconds, or 88.2 minutes, on the laptop computer for 49,235 sequences. The relationship between the number of sequences and the time to completion from both computers is shown in Figure 4. Additional time to completion trials were performed using low coverage 454FLX/Titanium sequenced contig datasets. For each dataset, CisSERS was configured to detect poly-A tails and test a 3’ restriction digest with all sequences. The range of the sequence to be analyzed was set from 80 to 600 bases from the 3’ end of the transcript for all 238 purchasable restriction enzymes. The datasets utilized ranged between 1983 contigs with an average of 417 bases per contig to 8126 contigs with an average of 320 bases per contig. Time to completion ranged from 2.38 minutes on the desktop for the dataset containing the least number of contigs to 32.88 minutes on the laptop for the largest dataset as seen in Figure 5. To ensure large data set functionality, the set of 6 base cutter enzymes and a dataset consisting of 130,503 contigs with an average length of 718 bases completed in 8.7 hours and 13.5 hours on the desktop and laptop respectively.
**In-silico Analysis and Wet-lab Validation**

**Gel Visualization:** To test the predicted gel visualization functionality, the Enterobacteria Lambda phage DNA sequence, accession number NC_001416.1 retrieved from the NCBI database, was digested *in silico* with HindIII. CisSERS does not utilize a gel matrix translation algorithm and reports band size in a distance dependent manner. This can be seen clearly in Figure 6 where the CisSERS linear representation of the band sizes matches the sizes in the wet-lab gel though there are differences in the band spacing. The restriction site prediction algorithm reports the beginning nucleotide position and will report the size of fragments by the difference of the beginning of the restriction site and where the restriction enzyme actually cuts in regards to other restriction site prediction programs that report fragment sizes. This can be seen by the one nucleotide size difference in the 4632nt and the 23129nt band sizes CisSERS reports and the corresponding band lengths of 4631nt and 23130nt lengths from the same cut sites reported by NEBcutter.

**CAPS marker development: Test case 1**- ATPC1 is one of the two nuclear encoded genes in *Arabidopsis* for the γ subunit of the chloroplast ATP synthase (25). The coupling factor quick recovery (cfq) mutant of *Arabidopsis* was identified as a point mutation in the ATPC1 gene and reduces overall photosynthetic capabilities (26). To screen a population of *Arabidopsis* for the cfq mutation, the wild type (WT) and mutant sequences (cfq) were processed using CisSERS. The analysis revealed TaqI as an enzyme that generates discernible differences due to this point mutation (Figure 7a). This result was validated biologically by amplifying the ATPC1 gene from a wild type, a cfq mutant, and a heterozygous individual. The amplified product was then digested with TaqI for 1 hour at 65°C and electrophoresed on a 10% TBE-Acrylamide gel (Bio-Rad) and stained with ethidium bromide (Figure 7b). The banding pattern in the biological gel matches the predicted gel image produced by CisSERS and verified the utility of this tool for an enzyme selection for CAPS analysis of this mutation. *Arabidopsis* represents a diploid test case with a very well defined genome. Such analyses can get complicated in the case of a sample with a higher ploidy as illustrated in the next test case.
Test case 2 - A gene putatively involved in bitter-pit disorder of apple was identified in previous work (Schaeffer and Dhingra, unpublished). Apple is an allotetraploid with a recently published genome (27). Cloning of this gene from eight apple cultivars varying in degree of disorder prevalence was completed. Sequencing of selected clones yielded 15 sequences which were manually trimmed to remove plasmid and primer sequences (Galli sequences). These sequences were then analyzed with CisSERS to identify an enzyme which separates the major alleles present in these cultivars. Upon evaluation, Cac8I was chosen due to its potential to differentiate five of the alleles across these eight cultivars (Figure 8A). Wet-lab validation was completed by amplifying the region and digesting with Cac8I for 3 hours at 37°C. The resulting DNA fragments were electrophoresed on a 2% agarose gel and visualized (Figure 8B). The wet-lab gel and the predicted gel image exhibited some disagreements. These differences highlight plausible genetic complexity in polyploid species. However, further analysis provides a resolution to some of the differences. CisSERS predicts the banding only of the sequence provided, and, for heterozygous organisms, the banding pattern for a restriction enzyme digest will be representative of all alleles as seen for the ‘Macintosh’ predicted digest and the actual digest (Fig. 8A first 3 lanes and Fig. 8B lane 1b). To resolve the latter difficulty, CisSERS can link two or more sequences together to provide more accurate predictive gel visualization. This is demonstrated with the ‘Red Gravenstein’ samples in Figure 8A where the ‘Red_Grav4’ and ‘Red_Grav10’ alleles were linked to produce the ‘Red Grav’ composite. The linked predicted gel visualization matches the wet-lab gel from ‘Red Gravenstein’ in lane 3b of Figure 8B. The predicted gel visualization for cloned sequences from ‘Haralson’ and the actual gel demonstrates the possibility that the ‘Haralson’ cultivar is homozygous for this allele and further in-depth investigation is required. The wet-lab gel digests demonstrate the remaining cultivars display a combination of two alleles indicating these are heterozygous and require sequencing of additional clones to capture the other allele. In this case, CisSERS output successfully identified an enzyme that was effective at resolving different genotypes and alleles. This potential CAPS marker is currently being developed for deployment in a larger number of varieties to determine its usefulness as a marker for bitter-pit disorder in apple.
DISCUSSION

Performance Test

Performance time is dependent on the platform and selected application when running CisSERS. These analyses demonstrate the usability of CisSERS without high powered computing resources. Additionally, while the 130,503 sequence dataset took hours to complete, analyzing these one at a time and compiling the results using existing tools would be time prohibitive.

Wet-lab Validation

The wet-lab tests demonstrate multiple functions of CisSERS. First, the predicted gel visualization is accurate although it is based on a linear size to distance relationship instead of the logarithmic relationship seen in an agarose gel. To ensure this accuracy, the predicted image band sizes needed to match the actual banding pattern of the gel. The size fractionation of the actual gel bands, as seen with the genomic Enterobacteria Lambda phage DNA digestion, matched the predicted gel band sizes. Secondly, identifying enzymes for CAPS markers was highly effective in the cfq example. This was less effective in the Galli sequences; however, the resulting information about the probable heterozygosity of the locus is critical and enables problem analysis prior to establishing CAPS markers.

High-throughput sequencing technologies are expensive and experimental design is a major component prior to sequencing. Based on the effective identification of restriction sites in standard and custom sequences, the identification of enzymes for reduced representation sequencing is also expected to be accurate and help ensure quality experimental design prior to sequencing. Combined, these experiments confirm the biological applicability of CisSERS as a highly effective addition to researcher’s toolkits.
Limitations and Future Improvements

While CisSERS is a comprehensive and useful tool, there are several limitations. The cut site is represented as the first base of the restriction site though this also allows the grouping of identical recognition sequences regardless of the location of the cleavage within that site. This grouping speeds the overall processing of CisSERS. Extremely large datasets may require extensive amounts of RAM or lengthy run times when processing all enzymes and gel visualization of these datasets may cause noticeable computer lag. These limitations are overcome by decreasing the amount of input sequences or the number of enzymes being processed. At this point, the methylation identification is simply based on any change in function of the enzyme by methylation including: requiring methylation, requiring no-methylation and any partial specificity. Currently, each of these enzymes must be analyzed by the user to make sure the chosen enzyme fits their project and the type of DNA they are processing. Though there are several limitations to this implementation of CisSERS, the overall utility and processing speed overcome these limitations to provide users with the ability to analyze large datasets for restriction sites.

CONCLUSIONS

As a program for biologists, CisSERS enables the identification of restriction sites in large multi-sequence data files. This is expected to facilitate more rapid development and deployment of CAPS markers for breeding and restriction enzyme selection for mutation identification that leverage the polymorphisms present in populations. Genotyping by sequencing and reduced representation sequencing approaches commonly utilize a restriction enzyme and CisSERS provides an efficient platform that will aid in the decision making process for users to determine the number of sites across the genome or transcriptome of interest. Overall, CisSERS is a front end tool for maximum utilization of next-generation sequencing data as the science begins to shift focus from how much data can be obtained to how we best utilize these data.
Additional Files

Supplementary Data: 4C.Additional.File.1_CisSERS_Users_Manual.pdf, 4C.Additional.File.2_Datasets, 4C.Additional.File.2_CisSERS.jar

**FUNDING**

This work was supported by WSU Agriculture Research Center Hatch funds to AD and KE. US Department of Agriculture National Research Initiative (USDA-NRI) grant 2008-35300-04676 to AD and AK supported AH and JG. TAK acknowledges the National Institutes of Health Protein Biotechnology Training Grant T32GM008336 and ARCS Fellowship Program. RMS work was supported by the National Science Foundation under Grants IOS 0641232 and MCB 1146928 and by Civilian Research and Development Foundation Grant RUB1-2982-ST-10. Work related to characterization of the cfq mutant was supported by US Department of Energy, Office of Science, Basic Energy Sciences Program DE-FG02-04ER15559 to DMK.

**ACKNOWLEDGEMENT**

The authors would like to thank Vandhana Krishnan, IBEST, University of Idaho for useful discussions.

**Author Contributions:** TK and AD developed the original concept. TK developed the original code for primary functions. AK facilitated the programing and development of the initial version and facilitated discussions on functional testing. RS and TK drove the development of the final features and functions. JG created the final perl code underlying many of the functions of CisSERS. AH developed the java GUI to facilitate input and analysis. MG and KE were responsible for the apple test data and results. MS and DK identified and produced the Arabidopsis validation example. TK and RS wrote the first draft of the manuscript. AD supervised the research. All authors reviewed the final manuscript.
Works Cited:


Table 1. Restriction Analysis Program Comparison. This comparison of some essential traits of CisSERS and other restriction site analysis programs highlights the advantages of CisSERS and some of the shared components with previously available tools. Many of these tools were designed for CAPS marker or derived CAPS (dCAPS) marker development and each has varying limitations.

<table>
<thead>
<tr>
<th>Program</th>
<th>Web-based</th>
<th>Automated decision making</th>
<th>Primer design</th>
<th>Data input type</th>
<th>Enzyme list</th>
<th>Primary Functions</th>
<th>Predicted gel image</th>
<th>Major limitation</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CisSERS</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Fasta or multi-fasta</td>
<td>REBASE with customization</td>
<td>multiple digest site analyses</td>
<td>Yes</td>
<td>Processing resource limited</td>
<td></td>
</tr>
<tr>
<td>dCAPS Finder 2.0</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Requires 2 sequences</td>
<td>Preset database</td>
<td>CAPS or dCAPS design</td>
<td>No</td>
<td>2-60 base sequences</td>
<td>(Neff et al. 2002)</td>
</tr>
<tr>
<td>BlastDigestester</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>multi-fasta</td>
<td>unknown</td>
<td>CAPS design</td>
<td>No</td>
<td>Limited by Blast</td>
<td>(Ilic et al. 2004)</td>
</tr>
<tr>
<td>SNP2CAPS</td>
<td>No</td>
<td>Partial</td>
<td>No</td>
<td>Alignment file</td>
<td>User input</td>
<td>CAPS design</td>
<td>No</td>
<td>Multiple alignments</td>
<td>(Thiel et al. 2004)</td>
</tr>
<tr>
<td>CapsID</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Alignments</td>
<td>unknown</td>
<td>CAPS design</td>
<td>Yes</td>
<td>unknown</td>
<td>(Taylor and Provart 2006)</td>
</tr>
<tr>
<td>SNP cutter</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>dbSNP or preformatted SNP file</td>
<td>Premade lists using REBASE</td>
<td>CAPS or dCAPS design</td>
<td>No</td>
<td>Format dependent</td>
<td>(Zhang et al. 2005)</td>
</tr>
<tr>
<td>SNP-RFLPing</td>
<td>Yes</td>
<td>partial</td>
<td>Yes</td>
<td>SNPs</td>
<td>REBASE</td>
<td>CAPS or dCAPS design</td>
<td>No</td>
<td>Human and Rat only</td>
<td>(Chang et al. 2006)</td>
</tr>
<tr>
<td>NEBcutter</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Fasta</td>
<td>REBASE</td>
<td>Comprehensive digest site analysis</td>
<td>Yes</td>
<td>1 sequence, max 300kb</td>
<td>(Vincze et al. 2003)</td>
</tr>
</tbody>
</table>
Table 2. CisSERS computer run time comparisons. Datasets comprising sequences with each restriction enzyme recognition site, unknown restriction enzyme recognition sites and a common prokaryotic genome run on two computers with different compute resources Table 2:

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of Sequences/Number of Total K bases</th>
<th>Win Vista Home Premium Intel Core 2 Duo CPU T6400 2Ghz 4GB RAM 64bit</th>
<th>Win7 Pro, AMD Phenom II x4 965 3.4Ghz 12GB RAM 64 bit</th>
<th>Number of unique Restriction Enzymes used in processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 base RE</td>
<td>16/1.0</td>
<td>378 msec</td>
<td>187 msec</td>
<td>16</td>
</tr>
<tr>
<td>6 base RE</td>
<td>84/5.0</td>
<td>700 msec</td>
<td>517 msec</td>
<td>84</td>
</tr>
<tr>
<td>8 base RE</td>
<td>6/0.4</td>
<td>370 msec</td>
<td>212 msec</td>
<td>6</td>
</tr>
<tr>
<td>Galli Sequences</td>
<td>15/8.0</td>
<td>671 msec</td>
<td>490 msec</td>
<td>238</td>
</tr>
<tr>
<td><em>Escherichia coli</em> strain K-12 substrain DH10B</td>
<td>1/4,690</td>
<td>216 sec</td>
<td>163 sec</td>
<td>238</td>
</tr>
</tbody>
</table>
Table 3. Calculated Time to Completion and Actual Time to Completion Values.

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Number of Sequences</th>
<th>Number of Bases</th>
<th>Desktop Actual Time</th>
<th>Laptop Actual Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>6bpCutLarge1File</td>
<td>1,101</td>
<td>66,060</td>
<td>6.30</td>
<td>9.58</td>
</tr>
<tr>
<td>6bpCutLarge2File</td>
<td>1,283</td>
<td>76,980</td>
<td>7.42</td>
<td>11.85</td>
</tr>
<tr>
<td>6bpCutLarge3File</td>
<td>1,543</td>
<td>92,580</td>
<td>9.71</td>
<td>15.66</td>
</tr>
<tr>
<td>6bpCutLarge4File</td>
<td>1,774</td>
<td>106,620</td>
<td>12.13</td>
<td>19.73</td>
</tr>
<tr>
<td>6bpCutLarge5File</td>
<td>2,550</td>
<td>153,000</td>
<td>17.55</td>
<td>28.89</td>
</tr>
<tr>
<td>6bpCutLarge6File</td>
<td>5,183</td>
<td>310,980</td>
<td>74.97</td>
<td>122.37</td>
</tr>
<tr>
<td>6bpCutLarge7File</td>
<td>9,116</td>
<td>546,960</td>
<td>131.83</td>
<td>220.95</td>
</tr>
<tr>
<td>6bpCutLarge8File</td>
<td>12,332</td>
<td>739,920</td>
<td>222.22</td>
<td>375.84</td>
</tr>
<tr>
<td>6bpCutLargeTriple8File</td>
<td>49,235</td>
<td>2,954,100</td>
<td>3354.68</td>
<td>5291.54</td>
</tr>
</tbody>
</table>
Figure 1. Experimental flow chart. A graphical depiction of the three phases of the CisSERS process and their subsections.
Figure 2. Screenshot of the CisSERS cut site selection process for the 5’ UTR trimming (top) and 3’UTR trimming (bottom). The yellow triangles can be selected and dragged or the exact location can be entered in the appropriate text box. The triangles and color schemes change based on the button selection made.
Figure 3. Screenshot of the CisSERS processing tab depicting the minor differences between standard analysis (left) and the poly-A trimming processing tab (right).
Figure 4. Performance tests across two different computer platforms. Time to completion of identifying restriction sites of 6-base cutter enzymes among fasta files with varying amounts of sequences was completed to illustrate CisSERS processing dependence on computer platform performance.
Figure 5. Time to completion of a 3’ restriction site analysis of four low coverage (1983 to 8126 contigs) 454FLX/Titanium sequenced databases. CisSERS was configured to analyze the datasets for poly-A tail sequences utilizing all 238 purchasable restriction enzymes across the two evaluated computer platforms.
Figure 6. Gel visualization functionality test. Enterobacteria Lambda phage DNA, accession number NC_001416.1, was in silico digested with HindIII through CisSERS and a commercially available Enterobacteria Lambda phage DNA was digested with the restriction enzyme HindIII for electrophoretic gel visualization comparison to the predicted gel image. Fragment sizes shown were calculated from the CisSERS reported cut locations and the manufacture’s publicized fragment sizes when digesting Enterobacteria Lambda phage DNA with the restriction enzyme HindIII.
Figure 7. Test Case 1: CisSERS predicted gel image vs. wet-lab gel visualization test with the *Arabidopsis* ATPC1, cfq mutant sequences. The two were linked to create the “F1 het” lane image while the F1 heterozygous plant DNA was analyzed and labeled “F1 het” in the wet-lab validation image. The banding patterns of all three samples of the CisSERS prediction match the wet-lab validation confirming the effectiveness of CisSERS to determine effective CAPS marker enzymes. In the wet-lab validation image, MM = NEB 100bp ladder.
Figure 8. Test Case 2: CisSERS predicted gel image vs. wet-lab gel visualization test with the Galli sequences. A. CisSERS predicted gel image of 12 identified alleles from 8 apple cultivar’s cDNA clones, and 2 linked gel images (Gold_Del, and Red_Grav). B. Wet-lab electrophoresed gel image of amplified products (#a) and corresponding restriction digest (#b); 1. ‘Macintosh’, 2. ‘Winesap’, 3. ‘Red Gravenstein’, 4. ‘Haralson’, 5. ‘Cox’s Orange Pippin’, 6. ‘Braeburn’, 7. ‘HoneyCrisp’, 8. ‘Golden Delicious’, MM = 100bp DNA molecular marker.

Analysis of the individual cultivars (A: Haralson 2 and B: 4b) suggest that ‘Haralson’ is homozygous for the sequenced allele; (A: Macintosh 9, Macintosh 2, Macintosh 5 and B: 1b) indicates that each allele present in ‘Macintosh’ is not yet sequenced; and (A: Cox_Org 10, Cox_Org 5 and B: 5b) also indicates that each allele of ‘Cox’s Orange Pippin’ has not yet been sequenced.
CONCLUSIONS AND FUTURE WORK

The work in this dissertation represents foundational genomic and transcriptomic knowledge and tools to facilitate further research in sweet cherry physiology, molecular biology and molecular genetics. Studies on other important physiological aspects of sweet cherry growth will benefit directly from these resources including examining the development of the fruit-pedicel abscission zone which is an ongoing collaborative project between the Dhingra, Whiting and Oraguzie programs at WSU. With RNA-seq experiments already underway, having a reference genome from this species is expected to greatly aid in annotation and selection of genes of interest for follow-up studies. Improvements to the genome sequence are also ongoing especially through the high throughput sequencing of pooled BACs to anchor and facilitate closing of gaps in the sequence.

As the cost of sequencing continues to decrease per base, the ability to begin developing haplotype sequences rather than genome sequences will become important. Currently, genome sequences represent an amalgam of the two chromosome sequences in diploid individuals. This presents the opportunity for only one allele to be represented at each locus which in homozygous individuals is sufficient. For crops like sweet cherry in which most varieties are self-incompatible, developing a homozygous line is not feasible. For apple and pear, double haploid lines have been obtained and are being sequenced in the Dhingra program to help assemble the genomes even though diploid draft versions are already available. The development of methods to procure phase specific sequences and assemble genome haplotypes may prove to be nearly as significant as the initial development of genome sequences. This is most important if large insertions/deletions are present in some chromosomes. If an assembler takes the shortest path to completion of the genome, the assembler is biased for the deletion region and can lead to genome assemblies representing fewer bases than estimated by flow cytometry. This has been seen recently with the peach genome where the estimated genome size was ~250 Mb (Arumuganathan and Earle 1991) while the assembled genome added up to only 220-230 Mb (IPGI 2010). Though this is
within 10% of the estimate, 10% of a genome may have major consequences for the understanding of how any species works.

In addition to the gains made via the de novo genome sequencing approach, the Prunus reference mapping results also yielded some interesting information. These results mirror the biological reality that sweet cherry and peach fruit have significant differences in ripening. Sweet cherry is non-climacteric, lacking a burst of ethylene and respiration that precedes ripening. The expected loss or decrease in function of four genes in sweet cherry involved in the production of ethylene, two ACS and two ACO genes, could help explain the evolution of non-climacteric development of this species. These and other genes involved in the ripening pathway could be investigated further as more data accumulates from other fruit producing species in the context of fruit ripening mechanisms prevalent in angiosperms. Ongoing research on sweet cherry abscission has demonstrated that exogenous application of ethylene effects fruit ripening and the development of the fruit-pedicel abscission zone. Additionally, varieties exhibit a differential response which may be controlled by the amount of endogenous ethylene produced by those varieties. The variety Stella, which was sequenced here, needs to be examined to document the abscission response to ethylene treatment for comparison to the known categories. It is important to document the physiological nature of the sequenced individual because hypotheses built on the genome of this variety need to be based on its own physiological properties. It would also be interesting to sequence these genes and the other forms of ACS and ACO in sweet cherry to identify if these genes correlate to differential ripening and abscission characteristics. Due to ACS and ACO being gene families in sweet cherry, the reference genome will be very helpful for designing allele specific primers to enable cloning and qRT-PCR to examine expression levels. Also from the reference mapping project, a set of 500 sweet cherry SNPs located in peach genes were selected for use in a linkage map of sweet cherry. In this way, the reference mapping work will help facilitate completion of the de novo genome assembly.

The transcriptomics projects have added to the foundational knowledge for sweet cherry research. Over two-thousand SNPs were identified from sweet cherry mRNAs which is 100 fold more than the number of sweet cherry ESTs publically available at the onset of this research (Koepke, Schaeffer et al.)
2012). These SNPs have been used to identify differences between closely related cultivars of sweet cherry (Fernandez i Marti, Athanson et al. 2012). The work demonstrated a higher degree of resolution in identifying sweet cherry genotypes with the SNP markers than the current standard of performing similar analysis using SSR markers. These results make an important contribution to the development of SNPs as gene-linked markers not only for evolutionary evaluation of varieties but also for breeding purposes.

The advances in sequencing technology are going to continue to have a transformative effect on how transcriptomics based experiments will be completed. These technologies have made it relatively inexpensive for data to be produced for many samples; therefore, utilizing an RNA sequencing based bulked segregant analysis will be the most effective approach to identify causal genes for different physiological characteristics. Bulked segregant analysis of recombinant inbred lines or near isogenic lines where multiple individuals at the extremes of each phenotypic characteristic are examined will reduce the background genetic differences among the individuals. Unfortunately, recombinant inbred lines and near isogenic lines are not in the near future of sweet cherry research. To perform these analyses in sweet cherry, many individuals from a breeding population will be needed. Alternatively, when the trait can be modulated by a rootstock, multiple rootstocks could be used to examine gene expression in identical genetic backgrounds as performed here. The unfortunate effect of this design is that the rootstocks are typically highly divergent and typically change many aspects of physiology rather than just the trait of interest. Again, as with bulked segregant analysis, the influence of the rootstock on non-related traits can be reduced by increasing the number and diversity of rootstocks tested.

While transcriptome profiling will continue to be important to plant biology, miRNA analyses could become just as common and critical. As seen in the model flowering pathway, most gene regulation appears to be modulated by miRNA intermediates where the actions were originally presumed to be more direct (Wang, Park et al. 2011). Though this could be a flowering specific regulatory anomaly, it is more likely that miRNA intermediates are involved in the control of many developmental processes.

The sequences obtained through both the genomic assemblies and transcriptome profiling projects are a great resource to leverage to develop molecular markers for improving the breeding of
sweet cherry. Due to the high-throughput nature of the approaches we undertook to develop these sequences, it was not amenable to use the existing tools to analyze these data. To address this issue, the tool CisSERS was developed that equips a biologist or a breeder to convert SNPs in a gene or several genes of interest into useful CAPS markers. CisSERS is a tool that has the potential to widely benefit biologists. The combination of high-throughput analysis with the ability to visualize simulated gel images will enable the rapid conversion of SNPs from sequencing research into usable markers.

Through the projects completed as part of this dissertation, great advances have been made in the understanding of the genomic composition of sweet cherry. The main question regarding how rootstocks control scion traits is yet unresolved. As described, the results presented here provide a foundation for the rootstock scion interaction experiments that will potentially provide answers to this question. The technological advances of the past decade will also greatly influence the ability of researcher to undertake these global and complex questions in plant biology.

Another bottleneck for future research and improvements to sweet cherry is the lengthy juvenility stage. As more is known about the transition of plants from juvenility to adult phase, potential targets for manipulation arise. Some intriguing possibilities are raised when this is added to the recent developments demonstrating transport of RNA species across the graft union into the scion. Based on the current state of knowledge, it appears that flower inducing transcripts could be expressed in a transgenic rootstock to induce sweet cherry trees to flower earlier, possibly saving several years in each cycle. This technology could also enable the analysis of fruit characteristics at a younger plant age. Early testing could identify plants that are not desired for the industry and they could be culled, saving years of establishment costs.

Being able to genetically modify sweet cherry rootstocks would also enable the specific analysis of root gene expression on scion traits. The importance of the functional aspect to candidate gene discovery and validation cannot be overstated. Testing genes in heterologous systems can provide much information; however, analyzing the gene function in the native genetic background is the most effective way to ensure the genes are interacting normally. Several key pieces need to be resolved during these experiments such as: 1. Is the transgene transported to the scion? This is important because there may be
species and variety specific transport sequences involved in the transport of RNA molecules in the phloem and changes in transport may lead to further refinement of the transport motifs (Harada 2010; Kanehira, Yamada et al. 2010). 2. Once in the scion, where does the transcript get transported to? 3. Does the transcript function in the scion alone or does it interact with its allelic counterparts? The transcripts may be transported but if they are immediately degraded their function will be short lived. Though direct transgenic scions may prove to be a necessary intermediate while root transgenics are optimized to control scion traits, transgenic rootstocks could be more marketable to the general public due to their lack of pollen production and the fruit will still be non-transgenic.

While floral bud numbers have been the focus of this dissertation, there are many equally important scion traits that have been shown to be rootstock controlled. Since the rootstock can control onset of bud break (Clearwater, Blattmann et al. 2007) and vegetative growth cessation (Prassinos, Ko et al. 2009), it is equally likely that the rootstock can effect flowering date and harvest maturity of fruit. Since the highest value for cherry producers comes from the earliest and latest cherries, rootstocks may be able to influence maturity to add value to their crop while providing more, high quality fruit to consumers. In conclusion, the research presented in this dissertation has significantly contributed to the ability of sweet cherry researchers to link important physiology to genes due to the genomics and transcriptomics foundation that has been developed.
Works Cited:


