IDENTIFYING THE GENETIC MECHANISM OF ETHYLENE-INDUCIBLE FRUIT ABSCISSION IN SWEET CHERRY

By

BENJAMIN RICHARD KILIAN

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

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
Program in Molecular Plant Sciences

MAY 2017

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of BENJAMIN RICHARD KILIAN find it satisfactory and recommend that it be accepted.

Amit Dhingra, Ph.D., Chair

Anantharaman Kalyanaraman, Ph.D.

Helmut Kirchhoff, Ph.D.

Matthew Whiting, Ph.D.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my graduate study supervisors, specifically, Dr. Amit Dhingra for his support, guidance, and enduring patience with me during my time as a PhD student. I have learned a great deal from him and he has been a great mentor and friend. My thesis has been greatly supported by my committee members and I could not have done this without them. Each provided me with excellent support and advice throughout my studies.

I appreciate the help and support of all the members of the Dhingra Lab. Each one has been gracious with their input, assistance and support. You have been invaluable to my growth as a scientist.

Without the love and support of my family, this thesis was going to be just a dream for me. I acknowledge my parents for their love and support throughout my life. I would like to thank my wife, Abby, for her sacrifice, support, and encouragement during my PhD studies and being by my side through the both the tears and joy. Finally, I would like to thank my two lovely daughters, Aeslyn and Ellie who joined our family during the course of my program, for giving me great joy and a deeper meaning to my life.

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

IDENTIFYING THE GENETIC MECHANISM OF ETHYLENE-INDUCIBLE FRUIT ABSCISSION IN SWEET CHERRY

Abstract

by Benjamin Richard Kilian, Ph.D.
Washington State University
May 2017

Chair: Amit Dhingra

Sweet cherry (*Prunus avium*, L.) belongs to the Amygdaloideae subfamily in the Rosaceae family and is related to Peach (*Prunus persica*, L.) which presents a classical example of climacteric ripening along with members in other sub-families such as Apple (*Malus x domestica* Borkh). While the fruit ripening process in sweet cherry is non-climacteric, a novel response to ethylene has been observed in a subset of cultivars, e.g., ‘Bing’, where a pre-harvest canopy-wide application of exogenous ethylene results in the formation of a distinct pedicel-fruit abscission zone. This decreases the force required to separate fruit from pedicel, presenting the possibility of harvesting the fruit mechanically. Some cultivars, e.g., ‘Chelan’, maintain a high pedicel fruit retention force, while the abscission zone in ‘Skeena’ develops without exogenous ethylene. This natural spectrum of phenotypes for ethylene-inducible abscission zone formation presents an interesting forward genetics model to unravel the underlying molecular mechanism. This is critical for a perennial crops species where reverse genetics approaches are not feasible due to the lack of genetic resources available in annual model system crops. Understanding the

causal molecular and genetic underpinnings of the inducible abscission zone formation offers the potential to expand the basic knowledge of ethylene's role in a non-climacteric fruit crop. Practical applications from this work include precision in timing the application of chemicals that induce abscission zone formation in other non-climacteric horticultural crops to enable mechanical harvesting, efficiently manage labor and enhance safety. My research used a genomics approach to refine limited genomic data for the 'Stella' genotype by resequencing five additional genotypes: 'Bing', 'Kimberly', 'Glory', 'Staccato' and 'Sweetheart'. Genes involved in ethylene inducible fruit abscission were then characterized using a developmental time-course transcriptomic approach. The results are presented in the following chapters.

Chapter 1: A international collaboration established the reference genome for sweet cherry and outlines the rationale for strategies used.

Chapter 2: Multiple approaches to identify polymorphisms in genomic data were evaluated.

Chapter 3: The results of the developmental time course, genotypic, and tissue specific transcriptome analysis that enabled the identification of co-expressed sequences during ethylene-induced abscission are described.

TABLE OF CONTENTS

DEVELOPMENTAL TIME COURSE TRANSCRIPTOME ANALYSIS OF ETHYLENE INDUCIBLE PEDICEL-FRUIT ABSCISSION FORMATION IN SWEET CHERRY (*PRUNUS AVIUM* L.)

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

Foreword

The Rosaceae family encompasses a diverse group of agriculturally important crops including apple, pear, peach, cherry, strawberry, raspberry, almond, and rose. Members of this family provide high-value nutrition and are consumed in various forms including fresh, dried, juice, and processed products. Rosaceae fruits are an important dietary source of phytochemicals, such as flavonoids and other phenolic compounds including molecules that have demonstrated effective anti-cancer properties [1, 2]. Rosaceae fruits, especially sweet cherry, are considered a high value commodity with great potential for an increased market share over the next decade. This upward trend in sweet cherry's market share can be further supported through ensuring superior fruit quality which can be achieved by deeper understanding of vital biological processes that occur within the fruit, most notably ripening and abscission.

Sweet cherry, a non-climacteric fruit, does not exhibit a characteristic burst in respiration or ethylene biosynthesis as ripening progresses. Rather, the presence of ethylene does not appear to be necessary for ripening to occur [3, 4], implying that ripening occurs through other methods/signals. This alternative, non-climacteric pathway appears to bypass the known ethylene synthesis and perception signal transduction commonly seen in climacteric fruits. Interestingly, several non-sense mutations have been discovered in several sweet cherry ACS and ACO genes [5] which further support the non-climacteric nature of ripening in sweet cherry. While this is the predominant school of thought, it has also been proposed that the ethylene burst may happen much earlier during fruit development as the fruit changes color.

In climacteric fruits there is evidence that ethylene stimulates a molecular regulatory cascade leading to the physiological responses associated with ripening processes [6]. Interestingly, in some sweet cherry varieties such as 'Bing', exogenous application of ethylene to the whole tree induces the formation of a pedicel fruit abscission zone at the time of harvest [7]. In a fruit, ripening and abscission are temporally associated developmental stages. As a fruit ripens, many complex developmental processes take place that eventually lead to the textural and constitutional changes in the fruits determining their final composition. The metabolic changes include altering cell structure, changing cell wall thickness, altering permeability of plasma membrane, decreasing structural integrity of the cells and increasing the intracellular spaces [8]. These textural changes leading to the softening of fruit are the result of enzymatic action altering the compositional structure of cell wall, partially or completely solubilizing the major classes of cell wall polysaccharides such as pectin and cellulose, hydrolyzing starch and other polysaccharides. Abscission, the programmed shedding of plant organs, is the culmination of cellular responses to both external (environmental) and internal (hormonal, genetic, etc.) cues within plant tissues. Although a commercially important process, abscission has been primarily used as a tool for unraveling phytohormone physiology rather than an important developmental process deserving careful study in its own right [9]. In sweet cherry, the unique ethylene-inducible, genotype-specific abscission at the pedicel fruit junction is a novel system to study abscission in a non-climacteric plant species.

Previous studies have evaluated the development of the pedicel-fruit abscission zone of sweet cherry using exogenously applied ethylene at the time of harvest [7, 10], however, the underlying genetic regulation and control remain uncharacterized. The approach undertaken in this dissertation to unravel the complex genetic regulation of abscission in sweet cherry combined the advantages of an established physiological model of genotype-dependent abscission zone development and next-generation genome and transcriptome analysis. The recent increase of genetic and genomic information can now complement the foundational work in anatomy and physiology. One of the most important areas of current and future work is to decipher gene regulation. Genetic control over developmental processes and vital agricultural traits is an area that should be pursued in order to solve problems in existing cultivars via horticultural or chemical approaches and utilize the genetic information for developing superior varieties in the future.

Although genomic data are fundamentally important, they must be generated within a genetic, physiological or developmental context. The primary aim of the research described in this dissertation is generating valuable genomics resources for sweet cherry. These resources provide a framework for understanding the ethylene-inducible genotype-dependent regulation of pedicel-fruit abscission zone development in non-climacteric sweet cherry. The result of the research is presented in three chapters.

Chapter 1: Sweet Cherry Genome - Strategies & tools for sequencing, assembling and annotation of the sweet cherry genome

This chapter provides an overview and background to the technological advances and techniques that have been used to generate sweet cherry draft assemblies and gene annotations. Additionally, it gives specific detail and parameters for the 'Stella' sweet cherry genome assembly.

Chapter 2: Evaluation of multiple genomics approaches to identify genome wide polymorphisms in *Prunus avium*.

A multi-pronged genomics approach, consisting of gel-based, sequencing-based or a hybrid of the two methods, to acquire polymorphism information has been evaluated. The gel-based method included the Target Region Amplification Polymorphism (TRAP) approach [11]. The hybrid approach to identifying polymorphism consists of utilizing the sweet cherry

SNParray developed as part of the RosBREED consortia [12]. Digital methods include TRAPseq (a modified reduced representation sequencing approach using TRAP PCR) and whole genome sequencing from multiple next-generation sequencing platforms and subsequent SNP analysis. Each of the approaches has distinct advantages and drawbacks which have been highlighted in this chapter. Further, several data analysis approaches have also been evaluated to identify the most optimal one for identification of genome-wide polymorphisms.

Chapter 3: Developmental time-course transcriptome analysis of ethylene inducible pedicel-fruit abscission formation in non-climacteric sweet cherry (*Prunus avium* L.)

While the physiological response to exogenous ethylene in the formation of pedicel fruit abscission zone has been well characterized [7], causal genetic underpinnings are poorly understood. To correlate established phenotypic responses of abscission zone development to the genetic regulation leading to its formation, a time-course transcriptome experiment was implemented. Expression of differentially expressed transcripts was validated with qRT-PCR. To our knowledge, this is the first investigation describing the transcriptomic analysis of the ethylene inducible abscission zone formation within Rosaceae.

Focus of Research

Overall, the goal of this dissertation is to describe the basis of the unique genotype-dependent phenotypic response to exogenous ethylene which leads to the formation of an abscission zone at the pedicel-fruit junction. In order to accomplish this goal, genomic sequences from five sweet cherry genotypes were generated and a developmental time-course RNAseq of the pedicel-fruit abscission zone following ethephon application on three genotypes was performed to identify the putative genes involved in this novel process.

Results from this work corroborate the overall trends from previous studies in model organisms such as Arabidopsis, however, several new and hitherto unannotated and differentially expressed transcripts were identified. Further studies are needed to evaluate the functional protein products of the genes that were identified as differentially expressed due to ethylene application. Characterization of the metabolomic, proteomic and physiological responses to ethylene and their effects on pedicel-fruit abscission zone development are expected to provide greater context and understanding of direct regulation of these important processes. A broader discussion of the results and future directions is presented in the conclusion section of this dissertation.

References

1. Mazur, W.M., et al., *Phyto-oestrogen content of berries, and plasma concentrations and urinary excretion of enterolactone after a single strawberry-meal in human subjects.* The British journal of nutrition, 2000. **83**: p. 381-387.

2. Egan, A.N., J. Schlueter, and D.M. Spooner, *Applications of next-generation sequencing in plant biology.* American Journal of Botany, 2012. **99**: p. 175-185.

3. Seymour, G.B., et al., *Fruit Development and Ripening.* Annual review of plant biology, 2013: p. 1-23.

4. Liu, M., et al., *Ethylene Control of Fruit Ripening: Revisiting the Complex Network of Transcriptional Regulation.* Plant physiology, 2015. **169**: p. 2380-90.

5. Koepke, T., et al., *Comparative genomics analysis in Prunoideae to identify biologically relevant polymorphisms.* Plant Biotechnology Journal, 2013. **11**: p. 883-893.

6. Osorio, S., F. Scossa, and A.R. Fernie, *Molecular regulation of fruit ripening.* Frontiers in plant science, 2013. **4**: p. 198.

7. Smith, E. and M. Whiting, *Effect of ethephon on sweet cherry pedicel-fruit retention force and quality is cultivar dependent.* Plant Growth Regulation, 2010. **60**: p. 213-223.

8. Ampopho, B., et al., *The Molecular Biology and Biochemistry of Fruit Ripening.* 2013: p. 1-216.

9. Taylor, J.E. and C.A. Whitelaw, *Signals in abscission.* New Phytologist, 2001. **151**: p. 323-339.

10. Zhao, Y., et al., *Pedicel-fruit retention force in sweet cherry (Prunus avium L.) varies with genotype and year.* Scientia Horticulturae, 2013. **150**: p. 135-141.

11. Hu, J. and B.A. Vick, *Target Region Amplification Polymorphism: A Novel Marker Technique for Plant Genotyping.* Plant Molecular Biology Reporter, 2003. **21**: p. 289-294.

12. Peace, C., et al., *Development and Evaluation of a Genome-Wide 6K SNP Array for Diploid Sweet Cherry and Tetraploid Sour Cherry*. PloS one, 2012. **7**.

CHAPTER 1

Sweet Cherry 'Stella' Genome - Strategies & tools for sequencing, assembling and annotation

Benjamin Kilian[1,2], Christopher Hendrickson[1,#] Tyson Koepke[1,2*], Scott Schaeffer[1,2$], Artemus Harper[1], Herman Silva[3], Lee Meisel[4], Nnadozie Oraguzie[5], Matthew Whiting[5] and Amit Dhingra[1,2,§]

[1]Department of Horticulture, Washington State University, Pullman WA 99164 USA

[2]Molecular Plant Sciences Graduate Program, Washington State University, Pullman, WA 99164 USA

[3]Herman Silva, Universidad de Chile, Faculty of Agricultural Sciences, Department of Agricultural Production, Av. Santa Rosa 11315, 8820808 La Pintana, Santiago, Chile

[4]Lee Meisel, Universidad de Chile, Instituto de Nutrición y Tecnología de los Alimentos (INTA), El Líbano 5524, 7830490, Macul, Santiago, Chile

[5]Irrigated Agriculture Research and Extension Center, Washington State University, Prosser, WA 99350 USA

# Present address: Department of Mathematics and Natural Sciences, National University, La Jolla CA 92037

* Present address: Phytelligence Inc., 615 NE Eastgate Blvd #3, Pullman, WA 99163

$ Present address: USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030-2600

[§]Corresponding author

Email addresses:

AD: adhingra@wsu.edu

Introduction

Access to complete genome sequences has long held the promise of enabling one to solve both large and small scale biological problems. Genetic sequencing provides information needed to identify and subsequently characterize key genes that are involved in various biological processes. However, the knowledge gleaned from genome sequences often introduces more complex questions about genetic interactions and the cross talk among various gene networks. An ever-increasing number of plant genomes are being sequenced due to high-throughput sequencing technologies. Key parameters must be carefully considered when sequencing the genome of a new species, including overall genome size, the presence of duplications and repetitive genomic regions, as well as ploidy, the most important variable tied to genome assembly efficiency. Horticultural crop species possess large genomes with areas of sequence repeats and duplication. However, a combination of new sequencing platforms is enabling sequencing and resequencing of the genomes of these species in order to improve the quality of the assembled genomes.

Genome sequencing technologies

A technological transformation over the past fifteen years has occurred in genome sequencing processes and technologies. Sanger sequencing, the main sequencing technology used over the past thirty years, produces high quality reads up to 1 kb in length and still considered the standard for accuracy [1-3]. The Sanger sequencing method has led to many significant biological achievements including completing the human draft genome as well as many plant and animal genomes. No technology is perfect, and the limitations of this technology are what led to the need for improved ways of sequencing many complex genomes [4].

DNA sequencing throughput has increased approximately 100,000-fold in the years following the human genome sequencing project. This unprecedented advancement was primarily made possible through the development of next-generation sequencing (NGS) technologies, which have increased exponentially in capability and availability while simultaneously decreasing costs. These new technologies have been dubbed "second generation" sequencing technologies as well as the more recent appearance of third generation technologies [5, 6].

The new sequencing technologies can be conveniently clustered into three primary categories. These include: sequencing by synthesis, sequencing by ligation, and single-molecule sequencing, the latter being known as the "third-generation" sequencing platform [7]. The primary advantages of NGS include enhanced speed and throughput combined with reduced manpower and cost [8]. The development of parallelized processing of raw sequence signal has enhanced the speed at which sequencing occurs and large datasets are generated. Depending on the technology used, greater than one billion sequencing reads can be generated per run on an instrument. These technologies have enabled the development of complex experimental designs where one can investigate the genome organization via DNA sequencing and, developmental, steady state and temporal changes in genome-wide expression using RNA sequencing.

The next-generation technologies include Roche 454 pyrosequencing (2005), Solexa/Illumina (2006), SOLiD (2007), and Helicos single-molecule sequencing (2008). These all take advantage of massive parallel sequencing by concurrently sequencing millions of DNA fragments that have been bonded to a surface. The sequenced fragments are relatively short, generally ranging from 25 to 400 base pairs. Even though NGS approaches are still used in many projects, the use of third generation sequencing is increasing due to its longer reads, relatively short run times, and resolution of single DNA molecules. Third generation technologies have become a useful supplement, serving as scaffolds for shorter reads and filling in gaps and repeat regions from the large datasets produced by second-generation technology. Companies producing third-generation sequencing technologies include Pacific Biosciences (PacBio) and Life Technologies Starlight, Oxford (Nanopore) along with several other new and upcoming approaches. These different sequencing platforms have recently been systematically reviewed for their underlying mechanisms and strengths/limitations [7, 9].

Since each sequencing technology has unique strengths and weaknesses, combinations of technologies have been applied to specific projects in order to complement and optimize accuracy and efficiency of data output. A recent study evaluating coverage bias has demonstrated that the bias can be corrected by combining second and third-generation sequence technologies [10]. The authors of this study concluded that there is no universal 'best' platform for all purposes. Rather, each of the sequencing technologies has characteristic or inherent errors that need to be dealt with using platform specific methods (Table 1).

Overall, advancements in sequencing technology over the past fifteen years are having a transformative impact on the field of genomics and transcriptomics [11]. As costs continue to decrease and overall throughput rises, the ability to conduct genomic and transcriptomic analyses will enable an increased level of understanding of important and novel biological questions. Already, the generation of large amounts of high quality genomic and transcriptomic data is no longer the limitation in the research pipeline; rather, data analysis has become the bottleneck.

Strategies to assemble data from next generation sequencing

Genome sequencing studies do not generally result in reconstruction of the complete chromosome sequence. They lead to a fragmented, yet usable draft genome. Genome assembly software, not unlike working on a puzzle, matches reads and builds contiguous stretches of DNA (contigs). The resulting draft genome is often sufficient to identify relevant genes and regulatory elements in the species being sequenced. The most commonly used NGS technologies have limitations that encumber reconstruction of full chromosomes. This often includes sequencing errors and areas of large repeats in the sequence. Assembling genomes require a specific minimum threshold combination of genome coverage, read lengths and read quality to efficiently accomplish the task [12].

Assembly programs use various heuristic approaches to assemble read data into contiguous sequence, resulting in output variability, leading one to question how the quality of an assembly can and should be assessed. The primary genome assembly strategies used by currently available sequence assemblers can be organized in one of several major groups including: overlap-layout-consensus (OLC), de Bruijn graph, and string graph [13]. The characteristics of the read data being assembled determine the choice of approach. For instance, highly accurate, short reads have been successfully assembled via de Bruijn graph based approaches. This is the primary assembler program used by CLC Genomics Workbench. Specifically, the de Bruijn graph approach holds an advantage over the overlap-layout-consensus and string graph because it possesses an inherent self-correction mechanism using accurate reads to refine and correct assembly errors as it progresses through the algorithm [14].

The overlap-layout-consensus (OLC) assembly approach begins by identifying every read pair that has a sufficient overlap and then arranges them in a graph with nodes representing reads and edges between read pairs overlapping by sequence. This graphical organization permits complex assembly algorithms to account for the comprehensive relationship identified between reads. The global overlap graph is simplified by removing redundant data. The Celera

Assembler [15] made this approach popular and until the second-generation sequencers emerged it dominated the genome assembly approaches. The overlap-layout-consensus approach has been limited by computational complexity until a new, efficient string indexing approach was introduced in the form of the SGA assembler [16].

De Bruijn graph assembler programs function by using read sequence to improve its internal graph structure and advance graph correction methods. These processes occur prior to and iteratively throughout the actual assembly which is a vital step in achieving a high-quality assembly. The Euler assembler [17] made this approach popular, and has influenced design of modern assemblers that are primarily focused on short-read data. These recent assemblers include Velvet [18], SOAPdenovo [19, 20] and ALLPATHS [21]. The efficiency of the De Bruijn graph assembler program is limited by sequencing errors and will most likely decrease further with third-generation reads that have an even higher error rate.

The string graph approach is similar to a de Bruijn graph in concept, but use the full-length read instead of forming k-mers. What makes this approach distinct is that it removes transitively inferred overlapping sequences. Pacific Biosciences (Menlo Park, CA 94025, USA) has produced FALCON, an assembly program for diploid species using string graph assembly principles [22].

An ideal genome assembly is able to fully capture and reassemble the genome representing its native size and structure. With NGS approaches, total genomic DNA is isolated from tissues and sheared into pieces of variable size, depending on the platform. NGS platforms acquire sequence from size-selected portions of the genome, and produce sequence reads of a targeted length. Like Sanger sequence reads, each position’s nucleotide base is assigned a quality score such as the Phred score, which is a score calculated by analyzing DNA sequence chromatogram files and assigning quality scores to each base call [23]. This score is based on the base calling program developed in 1998 by Phil Green and Brent Ewing (USA). It is used to estimate sequence error probability for each base-call, as determined by specific parameters computed from tracer data. The confidence of bases called across the length of a read tends to follow a predictable pattern, with reduced confidence at the 5' and 3' ends of each read [24]. Data yield and mean quality (cumulative confidence across a read) can be incorporated into algorithms in the assembly of a genome using *de novo* or reference-guided approaches. The approaches taken depend on the characteristics of the data being assembled. Repetitive sequences provide some of the biggest technical challenges for modern, short read sequencing technology. A genome can be assembled using a *de novo* approach where the assembler recreates the genome based on the provided statistical and computational parameters. Another way to assemble a genome is via a reference-guided approach where the data from a new species or another genotype of the same species is 'fitted' onto an existing genome.

Each of these approaches has significant strengths and weaknesses. While *de novo* genome assembly minimizes the risk of inaccurate assembly from the reference, it requires tremendous read-depth and computational resources as the topology of all pieces are considered against each other. As the read count of each portion of a genome increases, the confidence in its local topology also increases. However, due to inherent biases of short read sequencing approaches, amplification and capture of genome regions is not uniform, meaning read-depth will vary from zero to well over 1000X coverage. This often occurs in repeat regions of the genome where the assembler cannot easily identify the correct location for reads mapping equally well to multiple genomic locations. Thus, *de novo* genome assembly can, and often does, incorrectly reconstruct reads into larger contiguous sequence regions. Downstream *in silico* processing can compound this inaccuracy as homology between contigs is used to construct larger genome ‘scaffolds’ that can span many open reading frames. Sequencing read depth can also significantly impact reliability of allelic variation or polymorphisms such as single nucleotide polymorphisms (SNPs), microsatellite repeats, SSRs, and insertions or deletions (InDels) [25]. In addition to read depth, integrating a combination of sequence platforms into the experimental approach may be a viable option to address repeat regions. Applying the best features of each technology is expected to result in a more complete, higher quality assembly; however, this hybrid approach is limited in efficiency as it demands increased labor and consumables cost associated with additional library preparation and machine run-time [26].

Different runs with varied parameters of any *de novo* assembler program result in slightly different outcomes. This variance comes from multi-threading, or the processor’s ability to execute multiple process or threads simultaneously, combined with an artifact from probabilistic data structures being generated by the assembly. Running the assembly with a single thread will produce the same results every time. The trade-off, however, is that the assembly process will be greatly accelerated using multiple parallel processing threads. Each thread constructs contigs in an order that follows the order of the thread execution. This element is not user controlled. A contig’s size and relative position can be shifted if the contig start site is being built from two distinct starting points. This means that nearly identical assembly runs inevitably lead to differing outcomes, conditional upon the order of thread execution. Although the output of each run varies slightly, the overall assembly content retains its essential nature.

Together with known weaknesses in various NGS sequencing platforms, these algorithms cannot accurately reconstruct genomic regions with poor sequence depth or which contain solely repeat regions. The Pacific BioSciences RS NGS platform can yield long reads up to 50 kbp length with uniform base-call confidence (88% confidence) [27]. *De novo* genome assemblies can be improved through the presence of kilobase-sized reads that span gaps from short-read paired-end sequencing. As an output of single molecule sequencing technology, these reads have not undergone amplification and are therefore not affected by PCR-based artifacts such as GC-bias and unpredictable amplification [9]. The long reads can span the repeat regions, however sufficient sequencing depth and genome representation is required. The Roche 454 and more recent Ion Torrent NGS platforms can provide up to 1 kbp reads, but cannot accurately call bases in homopolymeric regions of 8 nucleotides or greater (Table 1). The Illumina platform offers improved confidence in base calls and increased data yield, but at the cost of generating short read lengths (from 35-150 bp) which leads to other problems for assemblers to overcome [28]. One of these problems is systematic error– meaning that many individual base-call errors tend to occur at similar genomic locations [29]. This error type can be especially problematic for SNP identification because the error occurs with cross-genotype consistency. With these strengths and weaknesses in mind, the combination of NGS platforms and assemblers must be carefully considered prior to the reconstruction of a new genome. Resequencing projects provide critical information regarding polymorphic regions. These regions can be converted to molecular markers which may be useful for marker-assisted breeding and barcoding projects that are based on high-throughput genotyping of breeding stock to cull progeny not carrying a desirable haplotype [30-32].

It is feasible, and becoming increasingly more cost-effective, to construct quality draft assemblies completely from the relatively low-cost short reads produced by the standard next-generation sequencing platform, Illumina. A standard sequencing project ought to focus on generating relatively deep coverage (≥ 30×) of paired-end sequences from short DNA fragments

(500-1000 bp), and combine this with additional coverage (10-20×) from longer DNA reads (3-10 kbp) [33]. When designing experiments, users can control parameters such as the sequencing platform, the read length, and the size/number of mate-pair libraries. Each choice affects the chosen assembler's ability to correctly reconstruct the sample's genome. Read length has an important influence on the complexity of the assembly. Longer reads lead to fewer repeats in the genome that tend to obfuscate the assembly algorithm. Read length can be the most difficult parameter to 'tune' as it depends fundamentally on the upper length limit defined by the specific technology used.

Assessing assembly quality

Determining whether an assembly is correct and comparing assembly quality is difficult because the correct answer or 'standard' is unknown (otherwise an assembly would not be needed). The assembler output is usually fragmented, containing errors ranging from small nucleotide substitutions to copy number changes in tandem repeats to large-scale genome structure rearrangements.

The quality of a genome assembly has traditionally been represented as an N50 value, calculated using the count and average size of the contigs. The N50 value is a length in base pairs from which contigs of that length or greater capture 50% of all captured length in all contigs. Generally larger N50 values are associated with greater genome or transcriptome sequencing depth in a given project, and will rise as assembled contigs or scaffolds begin to resemble the chromosomes [28, 33]. Short read data will however, quickly reach a coverage point where little incremental information can be gained. Since the short reads do not span repeat regions, the assemblers cannot infer the sequence. This is where the longer reads of third-generation sequencing technologies are a huge benefit to genome assembly [26, 34].

While the N50 value is an accepted descriptive statistic to assess the relative quality of an assembly, the genomics community has sought more robust means to describe assembly accuracy [35]. Length statistics alone can be misleading and generally uninformative because 1) the percentage of Ns in scaffolds may be very high and 2) there is always contamination from organellar DNA and other species in draft assemblies calculated using the count and average size of the contigs. Contamination from organellar DNA can be avoided by working with DNA isolated directly from nuclei.

One of the main computational concerns in *de novo* genome assembly is the possibility of improper collapsing of built contigs into a supercontig or consensus sequence. This can take two or more unique contigs that may contain unique alleles of a gene or unique gene members of a family and 'collapse' them into one. In the process, the unique genetic or allelic information is lost. There are a few options available to assess what is a 'best' framework in which a researcher allows an assembler(s) to run beyond iterative assemblies and comparison of summary statistics [36]. Further tools are needed to address an assembly's capture and representation of the genome. Tools such as RAMPART have begun to emerge that address the manual inputs needed to conduct such approaches, but comprehensive computational methods to address the 'quality' issue of genome assemblies remain elusive [37].

Historical overview and genetic improvement of sweet cherry (**Prunus avium** ***L.)***

Sweet cherry (*Prunus avium* L.) is thought to have been originally domesticated in the Caucasus region between the Black and Caspian seas [38]. Greek and Roman civilizations most likely acquired and transported cherries from the Anatolian region of modern Armenia. Cherries were cultivated extensively in Europe by the beginning of the 17th century [39-41].

Modern sweet cherry cultivars are roughly grouped into two main categories: dark-fleshed and light-fleshed (‘mahogany’ and ‘blush’). The former includes popular cultivars such as ‘Sweetheart’, ‘Kiona’, ‘Cowiche’, ‘Lapins’, ‘Van’ and ‘Bing’. The development of these cultivars reflects the evolving disease pressures and market demands in the industry which translates to identifying cultivars that mature at different time points to extend the market. They are characterized by a dark mesocarp and exocarp due to anthocyanin production, similar to their wild cherry relatives. However, variability exists among cultivar’s sweetness, tartness, size, pedicel-fruit retention force, resistance to diseases and disorders, ethylene responses, and fruit set. ‘Bing’ is perhaps the industry’s most popular cultivar, exhibiting a flavor profile coveted in the market along with many other desirable traits including high firmness and sugar content, and excellent shelf-life. ‘Rainier’ is the most popular among few light-fleshed or “blush” cultivars. ‘Rainier’ fruit are less tart than mahogany types and generally higher sugar content-commanding a premium price in the market. The blush cultivars are challenging to grow given their sensitivity to rain-induced cracking [42] and their high susceptibility to mechanical damage (bruising) during harvest and postharvest processing.

These advanced germplasm, in concert with improved production practices, have allowed for significant market growth of fresh and processed cherry industries throughout the Americas, Europe and Asia. To date, the sweet cherry industries in these regions have ranged from 295,000-418,000 tons in annual production in recent years generating $772-843 million annually in the United States [41]. Adding to global production figures, Chile and Argentina are rapidly expanding production acreage. Fresh cherry inventory imported into American markets (from South American production) has expanded the timeline for U.S. consumer access and concurrently increased the demand for the fruit. Despite annual variability, long term forecasts suggest that the sweet cherry market will continue expanding, thanks in part to the increased export market and improved harvesting technologies.

Increasingly, consumers have sought access to fresh sweet cherry products rather than secondary or processed products. Thus, the processed fruit market sector has diminished as the overall industry has expanded. This shift in demand has helped drive cultivar and rootstock development towards optimized production efficiency. Modern cherry producers seek rootstock and cultivars with precocity, balanced fruit set, minimal blind wood and sucker production, resistance to viral, fungal and bacterial pathogens, no fruit cracking and reduced vigor, among numerous other traits. Traditional breeding programs and phenotyping of progeny have enabled significant gains in available cultivars for large-scale sweet cherry production throughout the world. To address the expanding needs of today's growers, further germplasm improvement has applied genomics-based resources [43, 44]. Densely populated linkage maps have provided insight into the relative location of loci controlling many commercially desirable traits [45, 46]. Studying these traits has revealed a broad multi-loci genetic regulatory effect. These findings limit the proposed utility of marker-based approaches in targeted improvement of sweet cherry due to the complexity of the genetic control of phenotypes. Direct gene-trait correlations can help address the limitations of random molecular marker-trait associations by unraveling the underlying biological correlations at the transcriptional, proteomic or metabolic level in sweet cherry ontogeny. In parallel, extensive and standardized phenotyping of the breeding germplasm can aid in establishing these relationships.

In order to accelerate genomics-assisted crop improvement there is a need to have access to a complete sweet cherry genome. A genome is expected to enable the identification of alleles, their association with desirable traits and subsequent movement into breeding material that addresses evolving market needs. Until recently, this resource has not been available. Instead, genomics information from model systems such as peach for which a genome is available [47] has been used. This strategy has allowed for some improvement of sweet cherry cultivars; it is, however, limited in its ability to address horticultural concerns unique or dominant in sweet cherry. A draft 'Stella' sweet cherry genome was recently released as part of an international collaborative effort between groups in the United States and Chile. Sequencing of additional sweet cherry cultivars has since been initiated to establish a critical mass of genomic information which can serve as a foundation for efficient sweet cherry improvement.

Sweet cherry fruit displays non-climacteric ripening. However, there is another school of thought which proposes that the climacteric stage occurs much earlier in fruit development [48]. Despite this difference in opinion, sweet cherry is unique compared to many other Rosaceae species including peach, the commonly used model for genetic studies in Rosaceae. A recent study using two sweet cherry genotypes, 'Bing' and 'Rainier', suggested that non-sense mutations in several members of the ethylene biosynthesis genes ACS and ACO could be responsible for disrupting the ethylene production pathway [49]. These mutations could render sweet cherry without the respiratory burst commonly observed in other fruits of the genera after harvest. A limiting caveat to this result is that non-sense SNPs do not automatically imply protein loss of function. Additionally, the mutations observed in this study could contain errors derived from the fact that the sweet cherry reads were aligned to the peach genome rather than to sweet cherry. It is possible that respiration may be regulated through a mechanism distinct from ethylene receptors and subsequent signal transduction [50].

Access to the sweet cherry genome sequence is expected to provide information on upstream regulatory regions of genes (promoters and enhancers), aid in identifying gene families and genes not discovered through EST sequencing projects, and, provide a contrast to the peach reference genome, a classical climacteric member of the Rosaceae family.

A high-quality genome assembly in sweet cherry is a critical resource that can complement genetic and gene expression studies for establishing cause-effect relationships at the gene level. Instead of facilitating further probability-based information for breeding purposes, direct gene-trait relationships yield sequence-based information from which biotechnological and chemical-based solutions can be developed. A recent gene expression-based study reported 2,000 SNPs between the dark mahogany ‘Bing’ sweet cherry and blush ‘Rainier’ sweet cherry and suggested that variable expression of the transcription factor MYB10 may be a central regulator of cherry skin and flesh coloration and follow up studies confirmed this finding [51-53]. Future breeding, biotechnological and chemical genomics-based strategies can use this gene-trait relationship to modulate desired sweet cherry skin coloration in newly developed sweet cherry lines. Similarly, reverse-genetics approaches have applied small molecule screens to identify novel chemistries that modulate target protein activity [54]. Combined with gene expression analyses, these approaches provide a rapid workflow (relative to perturbative and laborious mapping projects) to identify specific phenotypes by understanding the underlying biological reasons for the trait at the genome level.

To fully understand the role of a gene’s control of a complex trait in sweet cherry, it is necessary to frame its expression in a plant system in which upstream, downstream and interacting genetic elements are described. Environmental influence can significantly alter gene expression profiles, and strongly impact how the fruit responds to stress. Development of a high-quality genome sequence in cherry can provide critical sequence-based information on genes and their promoter or other regulatory regions. Promoters are critical in the spatiotemporal regulation of gene expression [55]. Expression of gene family members or an individual gene can be differentially regulated between vegetative and fruit tissues, or between different regions of the fruit itself. This resource opens up the field for utilization of native sequences and improving crops using precision breeding techniques, most notably genome editing technologies such as CRISPR/Cas9 [56-58] and related technologies [59]. These technologies complement traditional breeding methods and allow for additional possible trait introduction to the plant breeder. Burgeoning knowledge of cis-acting regulatory elements in promoter regions from non-model plants has enabled improved accuracy in predictive modelling to identify environmental and transcriptional activators that influence gene expression [60]. Understanding of cherry promoter regions, combined with gene expression analysis can yield a robust model of a traits' presence in a genetic background and set of environmental conditions.

Fully-descriptive high-throughput phenotyping data is required to leverage the wealth of genomics resources in Rosaceae crops. In apple, recent efforts have begun to develop high-throughput standardized phenotyping of complex apple sensory traits such as the flavor, aroma and metabolite profiles [61, 62], texture [63, 64], and even annual bearing [65, 66]. These efforts have extended to growth habit and architecture as well with genotype-phenotype correlation, though extensive high-throughput techniques are still in development [67, 68]. Similar approaches have been applied to cherry [69]. Recent research from Washington State University has begun broad phenotyping of sweet cherry PFRF from which genetic associations are being developed [70, 71]. These studies provide an excellent model for future research to follow in addressing the critical need for high-throughput descriptive phenotyping in cherry. Only through such complementary research will the breadth of genomics resources in cherry be utilized in accelerated crop improvement.

Sequencing, assembling and annotation of the 'Stella' sweet cherry genome

Sweet cherry is a member of the Rosaceae family. Like other fruit species in the Rosaceae, it is thought to be native to central Asia. However, the modern cultivated sweet cherry displays unique characteristics compared to related species. It ripens according to a non-climacteric pattern, in contrast to peach, which exhibits climacteric ripening within the Rosaceae family [47]. Ethylene appears to play an important role in fruit ripening regardless of the ripening regime, primarily by influencing and regulating gene expression and the cascading responses to the phytohormone and/or developmental stage signal (Figure 1). As is the case for many tree fruits in Rosaceae, genome duplication events among cherry species have led to functional diversification of gene families, such as in the allotetraploid tart cherry [72]. Sweet cherry is an outcrossing-dependent diploid with a genome of $2n = 16$ estimated to be 200-250 Mb [44, 73].

The 'Stella' sweet cherry cultivar was chosen as the sweet cherry genotype for genome sequencing primarily because 'Stella' is extensively used in sweet cherry breeding programs worldwide. This is because it is the original source of self-compatibility due to the presence of a mutated S_4 allele (S_4'). The four base deletion mutation in the *S*-allele originated from irradiated 'Napoleon' pollen that was used to fertilize 'Emperor Francis' in the early 1950s [74, 75]. The use of 'Stella' in breeding programs to impart self-fertility reduces reliance on bees for pollination, eliminates the need for pollinizers and may even boost yield. The 'Stella' sweet cherry genome was sequenced through international collaboration using multiple sequencing platforms to generate the sequence data. The data were primarily generated via Illumina sequencing platform where 76× Illumina data were obtained from 2x100 standard Illumina HiSeq 2000 sequencing. Read files were obtained after initial sorting and filtering of the data via

Illumina's standard data processing. Additional sequencing data were generated on the PacBio and 454 sequencing platforms accounting for 21.2 Mb and 1.18 Gb data respectively, Table 2.

Significant resources have been used to develop SNP markers in major tree fruits with many millions of SNPs identified in peach and apple cultivars in recent years that have been used to assess lineage and genotypes of individuals in established breeding populations and in novel germplasm for future breeding use [76-78]. This has resulted in significantly faster and cheaper SNP identification in tree fruits for use in downstream analysis [79]. Emergence of NGS technologies have enabled identification of polymorphisms in expressed genes. Recent release of the apple [80], Chinese white pear [81], European pear [82], strawberry [83], and peach [47] genomes have provided additional resources with which to address biological questions.

Sweet cherry genome assembly

De novo genome assembly is an unbiased alternative to reference based assembly. The algorithms for genome assembly begin with processed reads and rebuild the genome digitally while minimizing error introduction to the assembly. Errors are often introduced due to sequencing bias from the sequencing platform used. Second generation sequencing read lengths (50-150 bp) are generally shorter than those produced by Sanger sequencing (800-900 bp). However, short read lengths present a difficulty for assembly because repetitive regions in the genome create assembly difficulties that increasing the sequence depth may not be able to overcome [13]. If the repeat length in a genome is *N*, assembly will be greatly improved if at least some read lengths are longer than *N*. Any repeat regions in the genome longer than read length generate gaps in the putative *de novo* assembly. Combined with the short length of second generation sequences, this means that any genomes assembled from these sequences alone are highly fragmented and more prone to misassembled rearrangements.

A reference-based assembly of cherry genomic Illumina and 454 data was performed using CLC Genomics assembler v 7.0 with the peach genome v 2.0 as the reference and using the default parameters: length fraction = 0.5, Similarity fraction = 0.8. Additionally, a *de novo* based assembly was generated based on Illumina reads using CLC Genomics v 7.0 with the default Illumina assembly parameters: Minimum contig length = 200, mismatch cost = 2, Insertion cost = 3, Deletion cost = 3, Length fraction = 0.5, Similarity fraction = 0.8. This assembly generated 96,080 contiguous sequences with an N50 of 4,130 from a total of 136,453,160 high quality reads, Table 3.

Accessing draft Sweet Cherry Genome

The raw read data were submitted to NCBI SRA (accession numbers: SAMN05414729, SAMN05414730). Prior to release of the draft genome via NCBI, data are available via the genomics lab portal at WSU (http://gmod.wsu.edu/portal/).

Annotation strategies

Genomics research aims to identify causal genetic locations, or DNA/RNA sequences which control the presence or absence of a phenotype. Extracting a functional role of these locations or sequences and their interactions is critical to understand the biology of complex quantitative traits. This step can be challenging given the hundreds of millions of base pairs produced in genomic or transcriptomic assemblies.

With a genome assembly in place, a variety of ORF-finding and gene prediction tools can be employed. These established tools such as FGENESH [84] and/or AUGUSTUS [85], can define CDS, splicing and regulatory domains for genes to catalog functional proteins. However, these tools are often run in isolation and fail to utilize other *a priori* information from a given sequence query to build a robust and accurate prediction of functional role [86]. Further, the predictive nature of the tools often leads to extensive, compounding errors. Inaccurate start codon identification, erroneous splicing models and additional mistakes can lie undiscovered in predicted gene sets. Targeting peptide and cis-acting element prediction in tree fruits is only beginning to be understood. Thus, regulatory prediction tools can also introduce spurious results into initial genome or transcriptome analyses. It is also important to consider the lack of representation of environmental influence which can result in alternative splicing of a given gene.

Gene predicting programs find the single most likely coding sequence (CDS) of a gene and do not report untranslated regions (UTRs) or alternatively, spliced variants. Gene prediction is therefore a somewhat misleading term. On the other hand, gene annotations include UTRs and alternative splice isoforms by synthesizing all available evidence [86]. Gene annotation is thus a more complex task than gene prediction. Any pipeline for genome annotation must deal with heterogeneous data sources including expressed sequence tags (ESTs), RNAseq data, protein homologies and gene predictions [87]. Additionally, it must efficiently synthesize all of these data into coherent gene models and produce an output describing the results in sufficient detail for these outputs to be suitable inputs for various genome browsers and annotation databases.

Among other tools, perhaps the most popular functional annotation workflow is the Blast2GO (B2G) software suite [88]. Using B2G, a researcher can assign functional roles to protein coding and noncoding RNAs and can mine existing sequence for putative functionality using BLAST, InterPro Scan, Pfam, KEGG mapping, and additional tools [89]. Gene Ontology (GO) terms are assigned to sequences following these analyses to categorize queries into functional classes. This tool empowers researchers to extract meaningful data from genome and transcriptome-based data sets.

Future for genome sequencing

The ultimate genome sequencing platform would incorporate single DNA or RNA molecules without any pre-amplification step or wet lab pre-processing such as fragmentation, without using secondary detection steps, have read lengths of Mb to Gb in size, not be affected by GC bias of the genome, high read accuracy and would be flexible enough to generate as many sequence reads as are necessary for requested coverage depth [9]. Additionally, it should be economical to both acquire and run, easy to operate, have short run times and simple or no library pre-preparation steps. This idealized sequencing platform does not exist currently, but progress is being made towards achieving this ideal. Constant innovation is happening in the chemistry and efficiency to existing approaches (Illumina MiSeq; Nanopore), as well as, to provide diverse options to the consumer to fit the widest variety of needs possible. Electronic BioSciences is developing a nanopore system with a very fast sequencing rate (~50kb/s). Genia (Roche) is involved in the nanopore sequencing market and has expertise in analog-to-digital sensors on integrated circuits. This distinctive approach could set them apart due to its increased sensitivity compared to the passive chips of other companies. Many of the recent startup companies are focusing on medical diagnostic markets where cheap, portable sequencing machines are most needed and effective.

Conclusions and outlook

Sequencing of the non-climacteric sweet cherry ‘Stella’ genome has been performed using a combination of next-generation sequencing approaches to generate the highest quality assembly possible with the resources available. ‘Stella’ reads were both *de novo* assembled and mapped to the Rosaceae reference *P. persica* (Peach) genome v2.0. The data generated from the genome assembly and annotation have been made available to the scientific community, NCBI accession number: SAMN05414730.

As the influx of genome sequence data increases, challenges in genomics research have shifted from data generation to analysis. No longer does it take years and hundreds of thousands to millions of dollars to generate a genome from a single genetic background of one organism. Instead, complete draft genome sequences from dozens if not hundreds of genetic backgrounds can be generated in weeks. Blast2GO, RAMPART and similar tools are capable of running on desktop computer environment. Alone, this represents a tremendous stride forward for genomics research capacity in sweet cherry. However, sequencing of additional genotypes is required to enable cross-cultivar and cross-species comparisons needed for future cherry improvement. As genome sequences of additional cultivars begin to emerge, the capacity for multi-genome comparison will also increase.

Although biologists are not able to sequence and assemble complex plant genomes with a single push of a button, it is possible and affordable to sequence and assemble a wide variety of interesting non-model plant genomes and obtain highly useful draft genome assemblies. This is only efficient if biologists remain aware of and up-to-date with the myriad of technology and algorithmic challenges involved [13]. The next frontier for plant genomics is likely the characterization of the diversity of genomic variations across large populations, accurately annotate their functional elements, and develop predictive quantitative models relating genotype to phenotype. Improved sequencing technology and advancement in assembly software are certain to play a large role in these studies, and we think a tight relationship between biology, technology and analytics are vital to enhance the field for many years to come [90].

References

1. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proceedings of the National Academy of Sciences, 1977. **74**: p. 5463-5467.
2. Rothberg, J.M. and J.H. Leamon, *The development and impact of 454 sequencing.* Nature biotechnology, 2008. **26**: p. 1117-24.
3. Sanger, F., et al., *Nucleotide sequence of bacteriophage ΦX174 DNA.* Nature, 1977. **265**: p. 687-695.
4. Barba, M., H. Czosnek, and A. Hadidi, *Historical perspective, development and applications of next-generation sequencing in plant virology.* Viruses, 2014. **6**: p. 106-36.
5. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Human Molecular Genetics, 2010. **19**: p. 227-240.
6. Lee, H., et al., *Third-generation sequencing and the future of genomics.* bioRxiv, 2016: p. 048603.
7. Egan, A.N., J. Schlueter, and D.M. Spooner, *Applications of next-generation sequencing in plant biology.* American Journal of Botany, 2012. **99**: p. 175-185.
8. Zhang, J., et al., *The impact of next-generation sequencing on genomics.* J Genet Genomics, 2011. **38**: p. 95-109.
9. Buermans, H.P.J. and J.T. den Dunnen, *Next generation sequencing technology: Advances and applications.* Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2014. **1842**: p. 1932-1941.
10. Ross, M.G., et al., *Characterizing and measuring bias in sequence data.* Genome Biology, 2013. **14**.

11. Trainotti, L., et al., *Functional Genomics,* in *Genetics, Genomics and Breeding of Stone Fruits*, C. Kole and A.G. Abbott, Editors. 2012, CRC Press. p. 292-322.

12. Schatz, M., *Assembly of large genomes using second-generation sequencing.* Genome Research, 2010. **20**: p. 1165-1173.

13. Nagarajan, N. and M. Pop, *Sequence assembly demystified.* Nature reviews. Genetics, 2013. **14**: p. 157-67.

14. Chaisson, M.J.P., R.K. Wilson, and E.E. Eichler, *Genetic variation and the de novo assembly of human genomes.* Nature Reviews Genetics, 2015. **16**: p. 627-640.

15. Myers, E., et al., *A Whole-Genome Assembly of Drosophila.* Science, 2000. **2196**.

16. Simpson, J.T. and R. Durbin, *Efficient construction of an assembly string graph using the FM-index.* Bioinformatics, 2010. **26**: p. 367-373.

17. Pevzner, P.A., H. Tang, and M.S. Waterman, *An Eulerian path approach to DNA fragment assembly.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**: p. 9748-53.

18. Zerbino, D.R. and E. Birney, *Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.* Genome Research, 2008. **18**: p. 821-829.

19. Li, R., et al., *De novo assembly of human genomes with massively parallel short read sequencing.* Genome research, 2010. **20**: p. 265-72.

20. Luo, R., et al., *SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.* GigaScience, 2012. **1**: p. 18.

21. Butler, J., et al., *ALLPATHS: De novo assembly of whole-genome shotgun microreads.* Genome Research, 2008. **18**: p. 810-820.

22. Chin, C.-s., et al., *Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing.* 2016.

23. Ewing, B., L.D. Hillier, and M.C. Wendl, *Base-Calling of Automated Sequencer Traces Using Phred.* Genome Research, 1998. **8**: p. 186-194.

24. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions.* Nat. Rev. Genet., 2011. **13**: p. 36-46.

25. Davey, J.W., et al., *Genome-wide genetic marker discovery and genotyping using next-generation sequencing.* Nature reviews. Genetics, 2011. **12**: p. 499-510.

26. Bleidorn, C., *Third generation sequencing: technology and its potential impact on evolutionary biodiversity research.* Systematics and Biodiversity, 2016. **14**: p. 1-8.

27. Rhoads, A. and K.F. Au, *PacBio Sequencing and Its Applications.* Genomics, Proteomics and Bioinformatics, 2015. **13**: p. 278-289.

28. Utturkar, S.M., et al., *Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences.* Bioinformatics, 2014. **30**: p. 2709-2716.

29. Meacham, F., et al., *Identification and correction of systematic error in high-throughput sequence data.* BMC Bioinformatics, 2011. **12**: p. 451.

30. Iwata, H., et al., *Potential assessment of genome-wide association study and genomic selection in Japanese pear Pyrus pyrifolia.* Breeding science, 2013. **63**: p. 125-40.

31. Yang, H., et al., *Application of next-generation sequencing for rapid marker development in molecular plant breeding: a case study on anthracnose disease resistance in Lupinus angustifolius L.* BMC genomics, 2012. **13**: p. 318.

32. Dirlewanger, E., et al., *Comparative mapping and marker-assisted selection in Rosaceae fruit crops.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**: p. 9891-9896.

33. Bradnam, K.R., et al., *Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species.* GigaScience, 2013. **2**: p. 10.

34. Miyamoto, M., et al., *Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes.* BMC Genomics, 2014. **15**: p. 1-8.

35. Salzberg, S.L., et al., *GAGE: A critical evaluation of genome assemblies and assembly algorithms.* Genome Research, 2012. **22**: p. 557-567.

36. Gurevich, A., et al., *QUAST: Quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**: p. 1072-1075.

37. Mapleson, D., N. Drou, and D. Swarbreck, *RAMPART: A workflow management system for de novo genome assembly.* Bioinformatics, 2015. **31**: p. 1824-1826.

38. Smith, E., *Abscission, Storability, and Fruit Quality of Mechanically Harvested Fresh Market Stem-Free Sweet Cherry*. 2009.

39. Thorpe, T., *History of plant tissue culture.*, in *Methods in molecular biology (Clifton, N.J.)*. 2012. p. 9-27.

40. Services, N.A.S. and U.s.A.S. Board, *Cherry Production*. 2015, USDA, National Agricultural Statistics Service. p. 1-4.

41. USDA, *Noncitrus Fruits and Nuts 2014 Summary*, in *United States Department of Agriculture - National Agricultural Statistics Service*. 2015. p. 1-97.

42. Long, L. and J. Olsen, *Sweet Cherry Cultivars for Brining, Freezing, and Canning in Oregon*. 2013. p. 1-6.

43. Rios, J.C., et al., *Association between the concentration of n-alkanes and tolerance to cracking in commercial varieties of sweet cherry fruits.* Scientia Horticulturae, 2015. **197**: p. 57-65.

44. Carrasco, B., et al., *Breeding in peach, cherry and plum: From a tissue culture, genetic, transcriptomic and genomic perspective.* Biological Research, 2013. **46**: p. 219-230.

45. Cabrera, A., et al., *Rosaceae conserved orthologous sequences marker polymorphism in sweet cherry germplasm and construction of a SNP-based map.* Tree Genetics and Genomes, 2012. **8**: p. 237-247.

46. Klagges, C., et al., *Construction and Comparative Analyses of Highly Dense Linkage Maps of Two Sweet Cherry Intra-Specific Progenies of Commercial Cultivars.* PLoS ONE, 2013. **8**: p. 1-10.

47. Arus, P., et al., *The peach genome.* Tree Genetics and Genomes, 2012. **8**: p. 531-547.

48. Nunez-Elisea, R. and T.L. Davenport, *Abscission of Mango Fruitlets as Influenced by Enhanced Ethylene Biosynthesis.* Plant physiology, 1986. **82**: p. 991-994.

49. Koepke, T., et al., *Comparative genomics analysis in Prunoideae to identify biologically relevant polymorphisms.* Plant Biotechnology Journal, 2013. **11**: p. 883-893.

50. Gong, Y., X. Fan, and J.P. Mattheis, *Responses of 'Bing' and 'Rainier' Sweet Cherries to Ethylene and 1-Methylcyclopropene.* J. Am. Soc. Hortic. Sci., 2002. **127**: p. 831-835.

51. Wunch, A., *Differential expression of cherry MYB10 in white and red varieties is responsible for anthocyanin levels*, in *7th International Rosaceae Genomics Conference, Seattle, WA*. 2014.

52. Jin, W., et al., *The R2R3 MYB transcription factor PavMYB10.1 involves in anthocyanin biosynthesis and determines fruit skin colour in sweet cherry (Prunus avium L.).* Plant Biotechnology Journal, 2016: p. n/a-n/a.

53. Starkevic, P., et al., *Expression and anthocyanin biosynthesis-modulating potential of sweet cherry (Prunus avium L.) MYB10 and bHLH genes.* PLoS ONE, 2015. **10**.

54. Blackwell, H.E. and Y. Zhao, *Chemical genetic approaches to plant biology.* Plant physiology, 2003. **133**: p. 448-455.

55. Dutt, M., et al., *Temporal and spatial control of gene expression in horticultural crops.* Horticulture Research, 2014. **1**: p. 14047.

56. Belhaj, K., et al., *Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR/Cas system.* Plant methods, 2013. **9**: p. 39.

57. Liu, L. and X. Duo, *CRISPR – Cas system : a powerful tool for genome engineering.* Plant molecular biology, 2014. **85**: p. 209-218.

58. Zhang, H., et al., *The CRISPR/Cas9 system produces specific and homozygous targeted gene editing in rice in one generation.* Plant biotechnology journal, 2014. **12**: p. 797-807.

59. Chen, K. and C. Gao, *Targeted genome modification technologies and their applications in crop improvements.* Plant cell reports, 2014. **33**: p. 575-83.

60. McGuire, A. and G. Church, *Predicting regulons and their cis-regulatory motifs by comparative genomics.* Nucleic acids research, 2000. **28**: p. 4523-30.

61. Biasioli, F., et al., *Towards high throughput phenotyping of the multisensory space of apple quality*, in *6th Rosaceous Genomics Conference, Mezzocorona, San Michele all'Adige, Italy*. 2012.

62. Rowan, D.D., et al., *High throughput metabolic phenotyping of apple fruit.* Acta Horticulturae, 2012. **945**: p. 213-218.

63. Gálvez-López, D., et al., *Texture analysis in an apple progeny through instrumental, sensory and histological phenotyping.* Euphytica, 2012. **185**: p. 171-183.

64. Schmitz, C.A., et al., *Fruit Texture Phenotypes of the RosBREED U.S. Apple Reference Germplasm Set.* Hort Science, 2013. **48**: p. 296-303.

65. Durand, J.B., et al., *New insights for estimating the genetic value of segregating apple progenies for irregular bearing during the first years of tree production.* Journal of Experimental Botany, 2013. **64**: p. 5099-5113.

66. Guitton, B., et al., *Genetic analysis and QTL detection for biennial bearing in apple.* Acta Horticulturae, 2012. **929**: p. 65-72.

67. Segura, V., C.E. Durel, and E. Costes, *Dissecting apple tree architecture into genetic, ontogenetic and environmental effects: QTL mapping.* Tree Genetics and Genomes, 2009. **5**: p. 165-179.

68. Segura, V., et al., *Phenotyping progenies for complex architectural traits: A strategy for 1-year-old apple trees (Malus x domestica Borkh.).* Tree Genetics and Genomes, 2006. **2**: p. 140-151.

69. Chavoshi, M., et al., *Phenotyping Protocol for Sweet Cherry (Prunus avium L.) to Enable an Understanding of Trait Inheritance*, in *7th International Rosaceae Genomics Conference, Seattle, WA*. 2014.

70. Whiting, M.D., et al., *A total systems approach to developing a sustainable , stem-free sweet cherry production , processing and marketing system.* Acta Horticulturae, 2012. **965**: p. 1-29.

71. Zhao, Y., et al., *Pedicel-fruit retention force in sweet cherry (Prunus avium L.) varies with genotype and year.* Scientia Horticulturae, 2013. **150**: p. 135-141.

72. Beaver, J.A. and A.F. Iezzoni, *Allozyme Inheritance in Tetraploid Sour Cherry (Prunus Cerasus L).* J Amer Soc Hort Sci, 1993. **118**: p. 873-877.

73. Arumuganathan, K. and E.D. Earle, *Nuclear DNA content of some important plant species.* Plant Molecular Biology Reporter, 1991. **9**: p. 208-218.

74. Lewis, D. and L.K. Crowe, *Structure of the incompatability gene, IV. Types of mutation in Prunus avium L.* Heredity, 1954. **8**: p. 357-363.

75. Ikeda, K., et al., *Molecular markers for the self-compatible S-4'-haplotype, a pollen-part mutant in sweet cherry (Prunus avium L.).* J. Am. Soc. Hortic. Sci., 2004. **129**: p. 724-728.

76. Ahmad, R., et al., *Whole genome sequencing of peach (Prunus persica L.) for SNP identification and selection.* BMC Genomics, 2011. **12**: p. 569.

77. Montanari, S., et al., *Identification of Pyrus Single Nucleotide Polymorphisms (SNPs) and Evaluation for Genetic Mapping in European Pear and Interspecific Pyrus Hybrids.* PLoS ONE, 2013. **8**: p. 1-11.

78. Scalabrin, S., et al., *A catalog of molecular diversity of Prunus germplasm gathered from aligning NGS reads to the peach reference sequence: Bioinformatic Approaches and Challenges.* Acta Horticulturae, 2013. **976**.

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

80. Botton, A., et al., *Signaling Pathways Mediating the Induction of Apple Fruitlet Abscission.* Plant Physiology, 2011. **155**: p. 185-208.

81. Wu, J., et al., *The genome of the pear (<i>Pyrus bretschneideri</i> Rehd.).* Genome Research, 2013. **23**: p. 396-408.

82. Chagné, D., et al., *The draft genome sequence of European pear (Pyrus communis L. 'Bartlett').* PLoS ONE, 2014. **9**: p. 1-12.

83. Shulaev, V., et al., *The genome of woodland strawberry (Fragaria vesca).* Nature Genetics, 2011. **43**: p. 109-116.

84. Salamov, A.A. and V.V. Solovyev, *Ab initio Gene Finding in Drosophila Genomic DNA Ab initio Gene Finding in Drosophila Genomic DNA.* Genome Research, 2000. **10**: p. 516-522.

85. Stanke, M. and B. Morgenstern, *AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints.* Nucleic Acids Research, 2005. **33**: p. 465-467.

86. Yandell, M. and D. Ence, *A beginner's guide to eukaryotic genome annotation.* Nature reviews. Genetics, 2012. **13**: p. 329-42.

87. Pérez-de-Castro, a.M., et al., *Application of genomic tools in plant breeding.* Current genomics, 2012. **13**: p. 179-95.

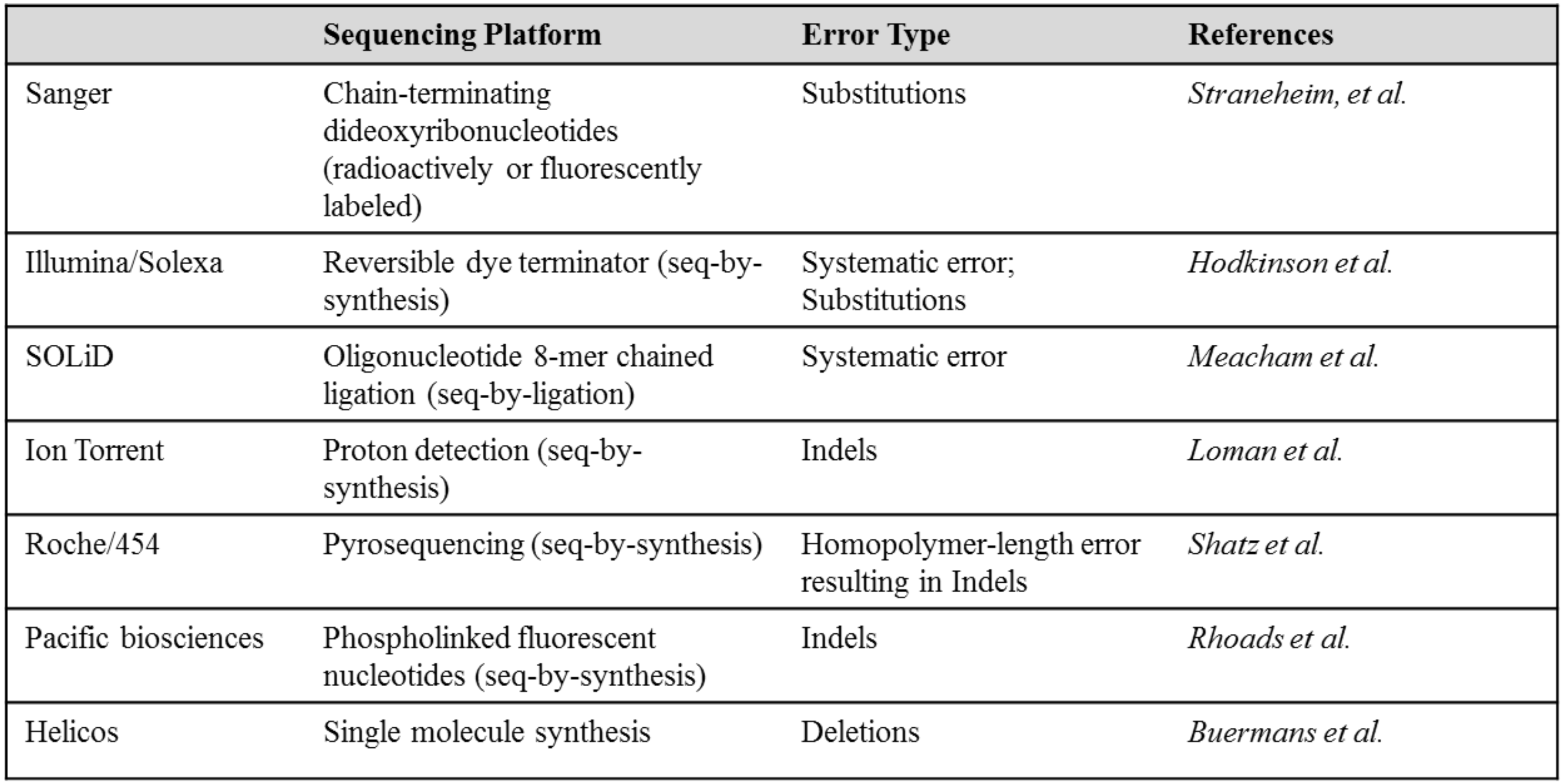
88. Conesa, A. and S. Götz, *Blast2GO: A comprehensive suite for functional analysis in plant genomics.* International journal of plant genomics, 2008. **2008**: p. 619832.

89. Götz, S., et al., *High-throughput functional annotation and data mining with the Blast2GO suite.* Nucleic acids research, 2008. **36**: p. 3420-35.

90. Schatz, M.C., J. Witkowski, and W.R. McCombie, *Current challenges in de novo plant genome sequencing and assembly.* Genome Biology, 2012. **13**: p. 243.

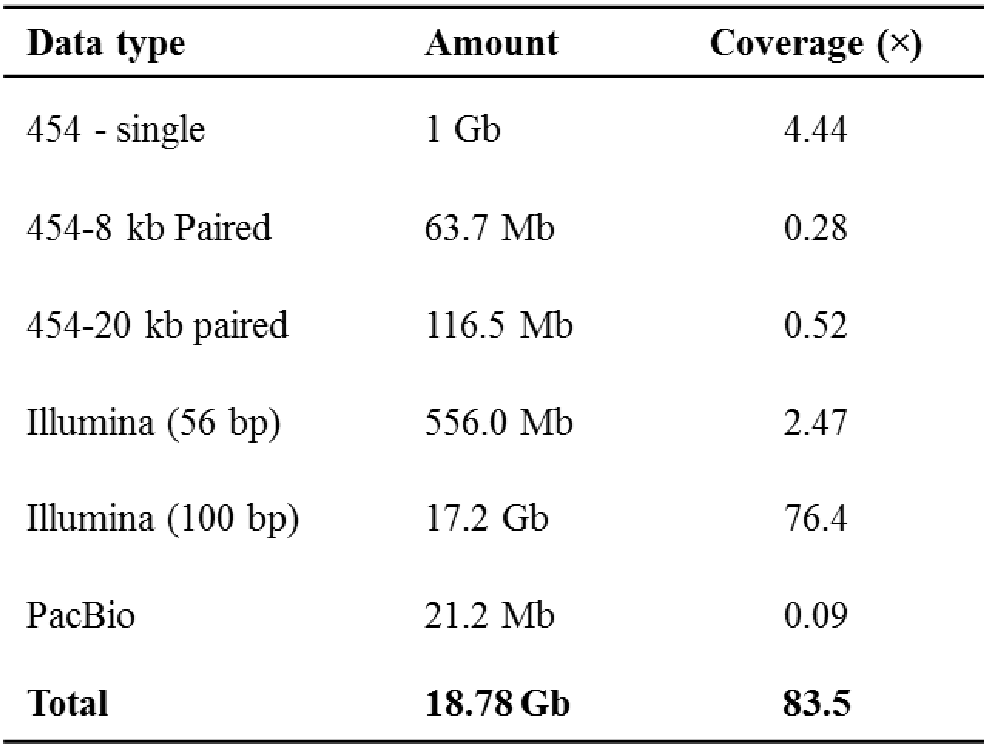
Tables

Table 1. Sequencing technologies and associated error types**.**



	Sequencing Platform	**Error Type**	**References**
Sanger	Chain-terminating dideoxyribonucleotides (radioactively or fluorescently labeled)	Substitutions	*Straneheim, et al.*
Illumina/Solexa	Reversible dye terminator (seq-by-synthesis)	Systematic error; Substitutions	*Hodkinson et al.*
SOLiD	Oligonucleotide 8-mer chained ligation (seq-by-ligation)	Systematic error	*Meacham et al.*
Ion Torrent	Proton detection (seq-by-synthesis)	Indels	*Loman et al.*
Roche/454	Pyrosequencing (seq-by-synthesis)	Homopolymer-length error resulting in Indels	*Shatz et al.*
Pacific biosciences	Phospholinked fluorescent nucleotides (seq-by-synthesis)	Indels	*Rhoads et al.*
Helicos	Single molecule synthesis	Deletions	*Buermans et al.*

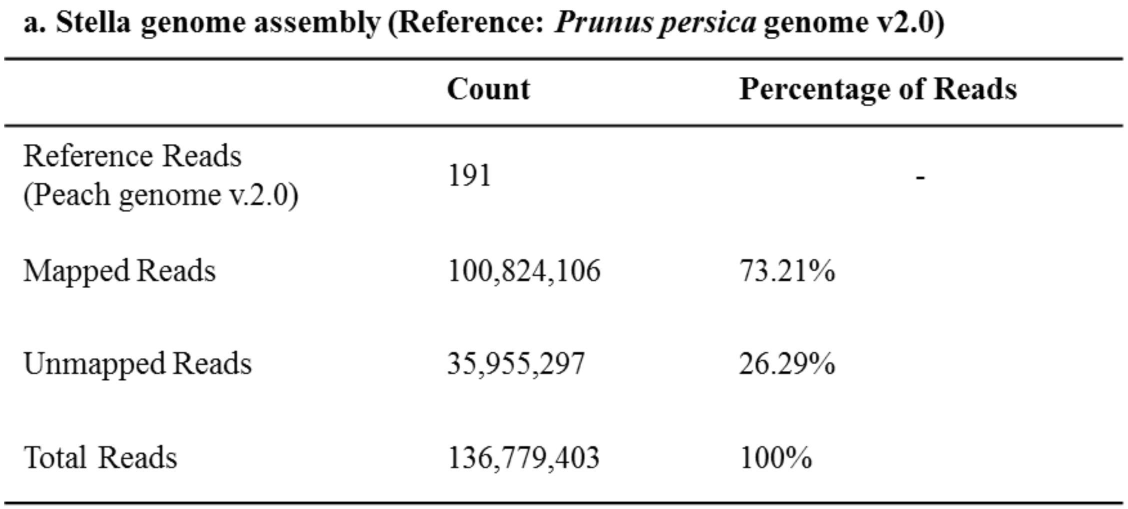
Table 2. ‘Stella’ sweet cherry genome sequencing data. The data has been provided by the sweet cherry International Consortium headed by Amit Dhingra (Washington State University, USA) and Herman Silva (Universidad de Chile, Chile).



Data type	Amount	Coverage (×)
454 - single	1 Gb	4.44
454-8 kb Paired	63.7 Mb	0.28
454-20 kb paired	116.5 Mb	0.52
Illumina (56 bp)	556.0 Mb	2.47
Illumina (100 bp)	17.2 Gb	76.4
PacBio	21.2 Mb	0.09
Total	**18.78 Gb**	**83.5**

See, http://gmod.wsu.edu/portal/

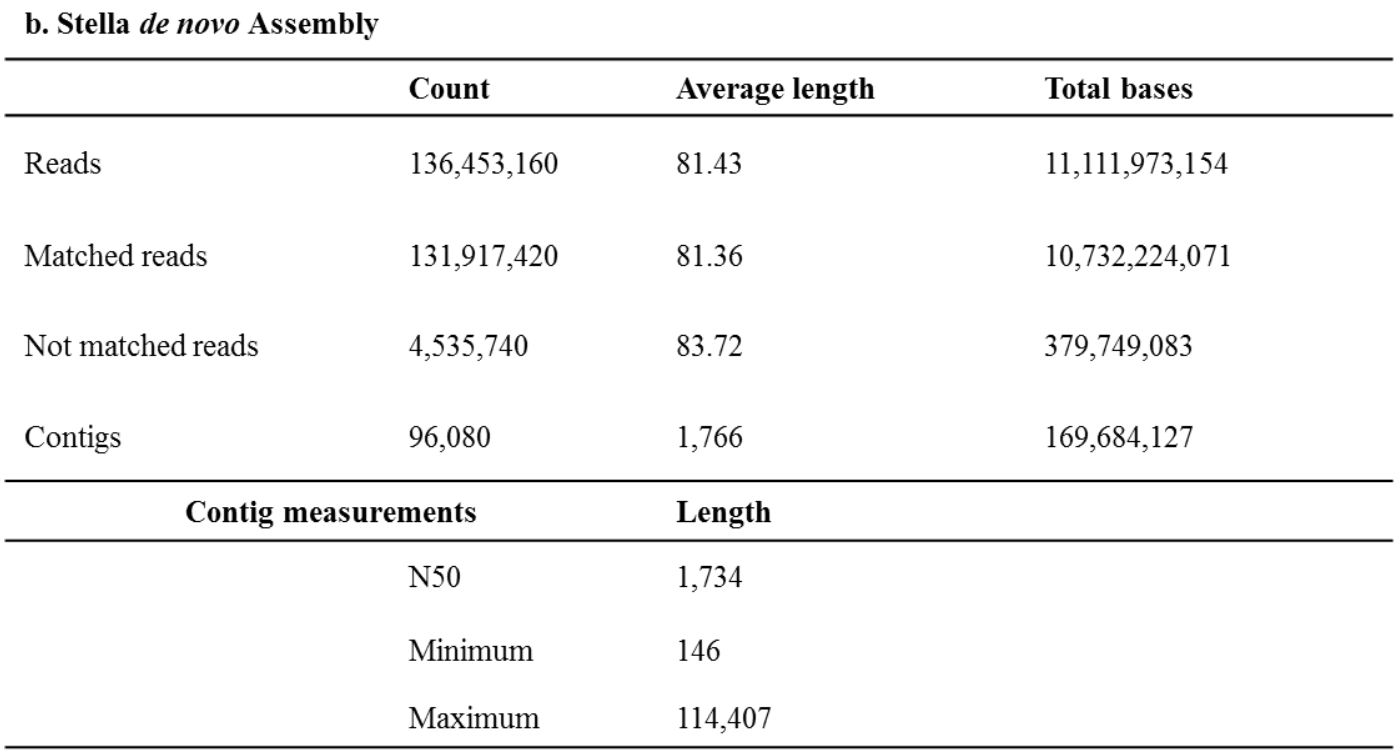
Table 3. Sweet cherry ‘Stella’ genome assembly descriptive statistics. A) ‘Stella’ genome assembly that was mapped to the *P. persica* reference genome. B) ‘Stella’ *de novo* assembly using next-generation sequencing data.



a. Stella genome assembly (Reference: *Prunus persica* genome v2.0)

	Count	**Percentage of Reads**
Reference Reads (Peach genome v.2.0)	191	-
Mapped Reads	100,824,106	73.21%
Unmapped Reads	35,955,297	26.29%
Total Reads	136,779,403	100%

NCBI accession number: SAMN05414730



b. Stella *de novo* Assembly

	Count	Average length	Total bases
Reads	136,453,160	81.43	11,111,973,154
Matched reads	131,917,420	81.36	10,732,224,071
Not matched reads	4,535,740	83.72	379,749,083
Contigs	96,080	1,766	169,684,127
Contig measurements		**Length**	
	N50	1,734	
	Minimum	146	
	Maximum	114,407	

NCBI accession number: SAMN05414729

Figures

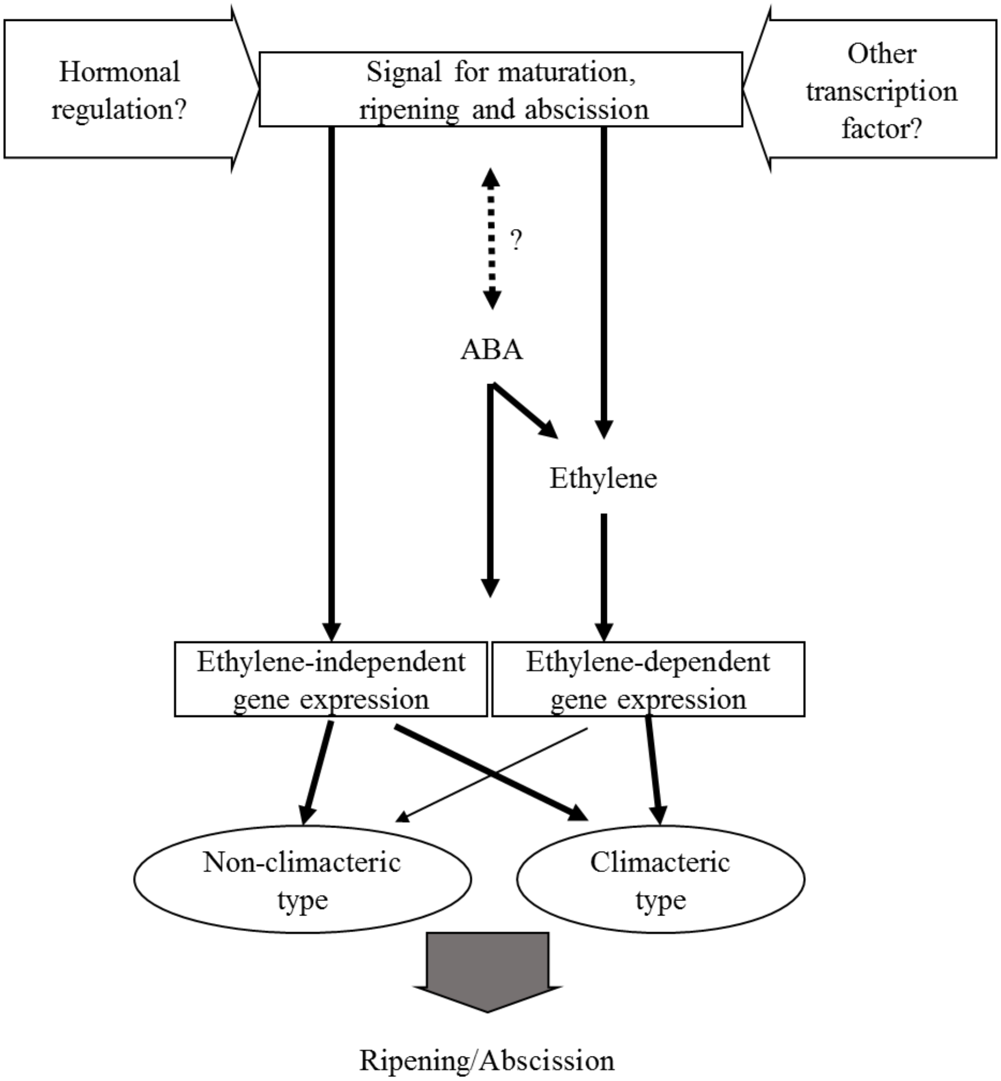

Hormonal regulation?
Signal for maturation, ripening and abscission
Other transcription factor?
?
ABA
Ethylene
Ethylene-independent gene expression
Ethylene-dependent gene expression
Non-climacteric type
Climacteric type
Ripening/Abscission

Figure 1. Generalized schematic summary of the ripening regulation in climacteric and non-climacteric fruits. Adapted from *Fruit Ripening: Physiology, Signalling and Genomics*, edited by Nath, Bouzayen, Mattoo, and Pech. 2014. page 7**.**

CHAPTER 2

Evaluation of multiple genomics approaches to identify genome wide polymorphisms in sweet cherry (*Prunus avium* L.)

Seanna Hewitt[1,2,ξ], Benjamin Kilian[1,2,ξ], Ramyya Hari[1,ξ] Richard Sharpe[1,2], Tyson Koepke[1,2,*], , Amit Dhingra[1,2§]

Invited submission for *Computational and Structural Biotechnology Journal.*

[1]Department of Horticulture, Washington State University, Pullman WA 99164-6414

[2]Molecular Plant Sciences Graduate Program, Washington State University, Pullman, WA 99164 USA

* Present address: Phytelligence Inc., 1615 NE Eastgate Blvd #3, Pullman, WA 99163

[ξ]co-first authors

[§]Corresponding author

Email address: AD: adhingra@wsu.edu

Abstract

Sweet cherry (*Prunus avium* L.) is an economically important non-climacteric tree fruit crop in the Rosaceae family. It has undergone an evolutionary bottleneck resulting in limited genetic diversity in the germplasm. Identification of genetic polymorphisms and subsequent development of molecular markers is important for marker assisted breeding for developing superior sweet cherry varieties. A gel-based molecular marker approach (TRAP), a 6k cherry SNParray, modified reduced representation sequencing (TRAPseq) and whole genome sequencing approaches were evaluated in the identification of genome-wide polymorphisms in sweet cherry cultivars. Genome-wide polymorphisms were detected among the genotypes using all platforms with variable efficiency.

Overall, whole genome sequencing and gel-based approach successfully detected polymorphisms in two of the five genotypes where SNParray and reduced representation sequencing failed to detect genetic differences. A combination of several approaches is necessary for efficient polymorphism identification, especially between closely related cultivars of a species. The information generated in this study provides a valuable resource for future genetic and genomic studies in sweet cherry, and the approaches evaluated here can be utilized in other closely related species with limited genetic diversity in the germplasm.

Key Words

Polymorphisms, *Prunus avium*, Next-generation sequencing, Target Region Amplification Polymorphism (TRAP), genetic diversity, SNParray, Reduced Representation Sequencing, Whole genome sequencing

1. Introduction

Application of various molecular tools and access to plant genomes has facilitated identification of genome-wide polymorphisms and development of molecular markers. Efforts to establish an association with polymorphisms and desirable traits which are then expected to be used in efficient development of desirable genotypes are ongoing. Next-generation sequencing (see Chapter 1) now allows for development of genomic information even for non-model plant systems accelerating the development of molecular markers and genetics research [1].

Sweet cherry (*Prunus avium* L.) is a member of the Rosaceae family which includes many other important crop species including apple (*Malus domestica*, Borkh.), peach (*Prunus persica*), plum (*Prunus domestica*), almond (*Prunus dulcis*), strawberry (*Fragaria spp.*), raspberry (*Rubus idaeus x R. strigosus*) and rose (*Rosa spp.).* Sweet cherry has an estimated genome size of 225-330 Mb [2, 3], but it is lacking in genomic information when compared to other Rosaceae members including peach or apple. Linkage maps and molecular markers have been developed for peach and almond, two other members of the sub-family Prunoideae [4]. A comprehensive and advanced draft of the peach genome is also available which serves as the foundation for several comparative studies [5]. Recently, a draft genome of sweet cherry cultivar 'Stella' was released as well [6]. In order to conduct diversity and genetics-related studies, efforts were made to evaluate the transferability of the molecular markers from peach to sweet cherry with mixed success [4].

Furthermore, domesticated sweet cherry genotypes exhibit a genetic bottleneck comprising only three chloroplast haplotypes despite a large number of wild land races [7, 8]. Given the genetic bottleneck, it can be difficult to differentiate between different varieties which are expected to be closely related. Previously, a study compared and evaluated the utility of seven simple sequence repeat (SSR) molecular markers versus 40 single nucleotide polymorphism (SNP) molecular markers to determine the genetic diversity and relatedness in 99 cultivated genotypes of sweet cherry [9]. SSRs were found to generate a higher average number of alleles per locus, mean observed heterozygosity, expected heterozygosity, and polymorphic information content values, however, the SNPs allowed for finer resolution of a closely related genotype which was indistinguishable with SSRs. Both sets of markers produced a similar genetic relatedness for all the accessions tested [9].

In this study we evaluated the utility of different genotyping approaches to differentiate between five closely related genotypes. These included three well established varieties, ‘Sweetheart’, ‘Bing’ and ‘Staccato’ and two new genotypes, ‘Glory’ and ‘Kimberly’. The latter were serendipitous selections made by farmers upon observing distinct flowering phenotypes. A gel-based, Targeted Region Amplified Polymorphism (TRAP) approach, a reduced representation TRAPseq approach, a SNParray and a whole genome sequencing approach were evaluated. All approaches resulted in the identification of polymorphic loci across the five genotypes. Comparison of ‘Glory’ and ‘Staccato’ yielded two polymorphic regions using the gel-based TRAP approach and thirty-four putative polymorphic sequence regions with the TRAPseq approach. Interestingly, SNParray had limited success in the identification of polymorphic regions amongst the closely related genotypes while the whole genome sequencing analyses approaches produced varied results. Comparing high quality contigs across genotypes using Seqman Pro NGen software (DNASTAR, Inc., Madison, WI) generated approximately five hundred high quality putative SNPs for each pair-wise genotypic comparison (Figure 4). However, other approaches such as DiscoSNP and Stacks generated an average of 250,000 and 500,000 predicted polymorphisms, respectively across all compared genotypes.

2. Methods

2.1. Plant material source and preparation

The five sweet cherry genotypes used in this study, ‘Bing’, ‘Sweetheart’, ‘Staccato’, ‘Glory’, and ‘Kimberly’, were obtained from VanWell Nursery, East Wenatchee, WA. Emerging leaf tissues were collected for each genotype and flash frozen in liquid nitrogen. All samples were pulverized under liquid nitrogen using SPEX SamplePrep® FreezerMill 6870 (Metuchen, NJ USA) and kept frozen at -80°C prior to processing.

2.2. Genomic DNA extraction

Total genomic DNA was extracted from young leaf tissue using cetyltrimethylammonium bromide (CTAB) phenol chloroform extraction method [10]. A total of 1 mL of CTAB buffer (0.8M guanidinium thiocyanate, 0.4 M ammonium thiocyanate, 0.1M sodium acetate pH 5.0, 5% w/v glycerol, and 38% v/v water saturated phenol) was added to approximately 100 mg frozen leaf tissue powder, shaken to evenly mix sample and incubated at room temperature for 5 min. Chloroform (200 μl) was added to the sample and shaken vigorously and incubated at room temperature, 3 min. Samples were centrifuged at 17,000 × g at 4°C for 15 min and the aqueous phase was collected and moved into to a clean 1.5 mL microcentrifuge tube to which 600 μL of isopropanol was added, and the tubes gently rocked 5-6 times and incubated at room temperature for 10 min. Samples were centrifuged 17,000 × g at 4°C for 10 minutes and the supernatant decanted, while retaining the pellet. The pellet was washed with 1mL of ethanol (75% v/v), vortexed for 10 seconds and centrifuged 9,500 × g at 4°C for 5 minutes. Extracted DNA pellets were air dried and suspended in 50 μl of RNase free water and incubated at 37°C with RNase free DNaseI for 30 minutes. DNaseI was inactivated by incubating the tubes at 65°C for 10 minutes. Extracted genomic DNA (2 µl) was electrophoresed on a 1% agarose gel and compared to Lambda DNA dilution series (100, 80, 60, 40, 20, 10 ng) to estimate quality and quantity.

2.3. TRAP – Target Region Amplified Polymorphism

The TRAP assay is a PCR-based technique that uses one fixed primer targeting a conserved DNA sequence across the genome and one or two arbitrary primers with either an AT- or GC-rich core that anneals with an intron or exon, respectively [11]. The arbitrary primers are fluorescently labeled at the 5’-end to enable laser mediated detection of DNA fragments on the LI-COR 4300 DNA Analyzer during electrophoresis and subsequent analysis. PCR was conducted in a final reaction volume of 10 µL with the following components: 2 µL of the 30-50 ng/µL DNA sample, 1.5 µL of 10× reaction buffer (Qiagen), 1.5 µL of 25 mM $MgCl_2$, 1 µL of 5 mM dNTPs, 3 pmol each of 700 and 800-IR dye-labeled arbitrary primers, 10 nmol of the fixed primer, and 1 U of DNA polymerase (Biolase). PCR was carried out by initially denaturing the template DNA at 94°C for 2 min. Five cycles of 94°C for 45 s, 40°C for 45 s, and 72°C for 1 min, followed by 35 cycles at 94°C for 45 s, 50°C for 45 s, and 72°C for 1 min were performed. The final extension step was at 72°C for 7 min. The product was then electrophoresed on a LI-COR 4300 DNA Analyzer (LI-COR Biosciences, Lincoln, NE) for visualization. A 6.5% polyacrylamide gel (KB-PLUS, LI-COR) was cast, the reactions loaded, and run at 1500 V for 2.5 hours and images were collected automatically in the computer file. The images were then analyzed using LI-COR 4300 DNA Analyzer image software to identify polymorphisms.

2.4. TRAPseq

A modified reduced representation sequencing method was implemented for Glory and Staccato. This approach was derived from the TRAP molecular marker approach discussed previously (2.3, TRAP). Genomic DNA (~1 µg) was isolated from 'Glory' and 'Staccato' young leaf tissue. This was followed by a TRAP PCR using fixed (MADS, PPR1 and PPR2) and arbitrary (ODD15 and GA5) primers (Table 4). The TRAP PCR parameters used were identical to the TRAP protocol, described above. Following TRAP amplification, the PCR product was purified using Qiagen PCR purification kit per protocol. The reduced representation sample library was prepared using a modified NEBNext® Fast DNA Library Prep Set. Libraries were sheared with NEB Next Fragmentase per standard protocol. After heat disabling the fragmentase, 0.05 µL dATP (100mM), 0.2µL Taq polymerase (5U/µL), 0.65µL MgCl2 (50mM), and 2.1µL of 10x Taq polymerase buffer were added to directly to each fragmented reaction. The mix was incubated at 72°C for 20 minutes for A-tailing. Complementary, custom adaptors were then annealed to the sheared DNA, the annealed product purified and extracted according the NEBNext FastDNA Library Prep protocol. The libraries were quantified, pooled, and sequenced using the Ion Torrent PGM (Thermo Fisher Scientific, Inc., Waltham, MA). The sequencing run included 850 flows on a 318C chip producing single reads of various lengths.

2.5. SNParray

The sweet cherry SNParray is a 6K Infinium II array designed with SNPs from both diploid sweet cherry (*P. avium*) and allotetraploid sour cherry (*P. cerasus*) [12]. The array contains 5696 predicted SNPs, obtained through re-sequencing of sixteen sweet and eight sour cherry accessions. The array includes 4214 SNPs representing the sweet cherry genome and 1482 representing the sour cherry genome. For this study, 'Bing', 'Sweetheart', 'Glory', 'Kimberly', 'Staccato' and 'Stella' sweet cherry cultivars were analyzed. The output data were analyzed with GenomeStudio v. 1.0, Genotyping module (Illumina, Inc., San Diego, CA). The software determines cluster positions of the AA/AB/BB genotypes for each putative SNP which were manually parsed, interpreted and categorized. Default quality metrics for GenomeStudio were used in the assay: GenTrain score ≥ 0.5, minor allelic frequency (MAF) ≥ 0.15 and call rate of > 80%.

The data show pair-wise comparisons between each cultivar for each specific SNP. These comparisons were used to identify potential SNPs between each genotypic set. Following the SNParray analysis, a subset of the predicted SNPs was evaluated *in silico* by using BLAST to compare twenty SNPs from NCBI with our *de novo* assembly from each genotype. All twenty SNPs tested were confirmed using this method (Table 7).

2.6. Genome sequencing, assembly and SNP identification

A combination of sequencing platforms was used to generate sequence data for the sweet cherry genotypes. For two genotypes, 'Glory' and 'Staccato', approximately 40× coverage of sequence was generated using PacBio RS II (v2 SMRT cell) (Pacific Biosciences, Menlo Park, CA) with a mean read length of eight kilobases per read. The reads were assembled into contiguous sequences using the SMRT (HGAPII) program from Pacific Biosciences [13]. For all of the genotypes, approximately 25× coverage sequence data represented by 2×100 paired end reads were generated with the Illumina Hi Seq 2000 sequencing platform. The reads were quality filtered; trimmed; merged; and individually mapped to the PacBio generated references of both 'Glory' and 'Staccato' assemblies using CLC Genomics v7.0. In addition to the reference based assembly, each of the genotypes were independently assembled *de novo* using just Illumina paired end reads.

SNP analysis was performed using SeqMan Pro software (DNASTAR, Inc., Madison, WI). Initially, four reads were used to call the consensus base at a given locus. Additional layers of analysis were performed with up to ten required consensus reads to reduce the potential for false positive SNPs. Predicted SNPs from each parameter were passed through a series of bioinformatics filters to validate the polymorphic nature of the specific loci [29–31]. These bioinformatics filters included: 1) Removing contiguous sequences that were shorter than 201 nucleotides to reduce putative SNPs that were too close to the 5' or 3' end of a given contig, 2) Removing contiguous sequences containing ambiguous bases for each selected threshold. As an example, if the minimum threshold number of reads is set at eleven, then any position along the called consensus having ten or fewer reads would be called *N* (ambiguous)), 3) Removing sequences containing more than four SNPs in the same contig. If SeqMan output contained multiple SNPs particularly in a sequence, the whole sequence was eliminated from further analysis due to the potential for of faulty alignment in the pairwise comparison of homologous contigs.

2.7. Analysis of Illumina sequencing data using DISCOSNP

DISCOSNP was used to identify SNPs and small indels from sweet cherry Illumina data [14]. The input files for DISCOSNP evaluation were Illumina read files and three independent modules were progressively run to generate high quality putative polymorphisms from raw read data. The first, *kissnp2,* detected SNPs by comparing the sample reads. *Kissreads2* enhanced *kissnp2* results by adding mean read coverage and average quality of the reads forming the polymorphism. Finally, a .vcf file was generated from the kissnp2/kissreads2 outputs. A primary script derived from the published DISCOSNP user guide was used to run the three modules (Supplemental File 5). The output of DISCOSNP consisted of a multi-fasta file sorted by ranking of putative SNPs in descending order of probability. Additional information concerning read coverage, average PHRED quality for each input dataset, lengths of unambiguous left and right extensions, and Phi coefficient for each SNP was also generated.

2.8 Comparative analysis of polymorphisms across samples using Stacks

Stacks [15] was used to identify SNPs from the sweet cherry short-read sequence genomic data. This was accomplished through building artificial loci from the raw data (‘stacks’ of reads). An internal module (`Process_shortreads)` was used to filter reads with uncalled bases, discard reads with low quality scores and remove any traces of remaining inline barcodes. Thereafter, the dataset was processed by running the denovo map wrapper, which includes ustacks, cstacks, sstacks, populations (Figure 7). Ustacks built stacks and formed loci and searched for SNPs. Cstacks merged the alleles. Sstacks formed a set of stacks that searched against catalog stacks (cstacks) [16]. Each stacks set generated individual loci and SNPs were detected at each locus using a maximum likelihood framework by iteratively comparing loci for each sweet cherry genotype in a pairwise comparison against other genotypes.

3. Results and Discussion

3.1. Pedigree information regarding the sweet cherry genotypes

Given the lack of genetic diversity within sweet cherry, it is important to know the pedigree information regarding the five primary genotypes used in this study namely, ‘Bing’, ‘Sweetheart’, ‘Staccato’, ‘Glory’ and ‘Kimberly’. ‘Sweetheart’ is the maternal parent of ‘Staccato’ while the paternal parent is unknown as it was developed via open pollination. ‘Van’ and ‘Newstar’ (pollinator) are the parents of ‘Sweetheart’, but ‘Sweetheart’ and ‘Staccato’ have no known connection to the other four genotypes used in this study. Previously published SNP marker analysis has shown the paternal parent of ‘Bing’ to likely be ‘Napoleon’ [17]. ‘Napoleon’ is also the paternal grandparent of ‘Stella’ (Figure 1). Therefore, ‘Bing’ and the genomics reference ‘Stella’ share Napoleon in their pedigree as a paternal parent and grandparents respectively. ‘Kimberly’ and ‘Glory’ were serendipitous discoveries in orchards based on their distinct flowering time phenotype and therefore have unknown lineage. Sweet cherry varieties are also categorized based on their S-allele genotype [18, 19]. The S-allele genotype of the five varieties used in this study is provided in Table 2.

3.2. Evaluation of gel-based approach, TRAP

Target Region Amplification Polymorphism (TRAP) [11] was used to identify genetic differences by visualizing differential amplification patterns unique to ‘Glory’, ‘Staccato’. ‘Bing’ was included as a positive control. A fixed primer designed to amplify the MADS box gene family was implemented in this study because this diverse gene family is predicted to contain polymorphic regions even in closely-related plant cultivars. An additional cohort of sweet cherry specific fixed primers were designed from intronic regions of flowering locus genes to improve specificity for sweet cherry (Table 3) [20]. TRAP was used to identify polymorphic regions specifically between the ‘Glory’ and ‘Staccato’ cultivars due to the fact that although they are observed to have differing flowering dates, this flowering phenotype has not led to the evaluation of any molecular markers to date. This approach identified two putative polymorphic regions within the MADS gene family (Figure 2) out of a total of 45 amplified loci.

3.3. TRAPseq – modified reduced representation sequencing to identify polymorphisms

A modified reduced representation sequencing approach was used to identify polymorphic regions, specifically between ‘Glory’ and ‘Staccato’ cultivars where no polymorphisms were detectable using SNParray. The reduced representation of the genome was achieved by performing TRAP PCR (Table 4), followed by generating NGS sequence data from the amplified products (Ion Torrent PGM, Thermo Fisher Scientific, Inc., Waltham, MA). This experiment generated 133 Mb sequence data for ‘Glory’ and ‘Staccato’ genotypes. These data were separately assembled into 1,577 and 1,404 contigs for ‘Glory’ and ‘Staccato’, respectively, using CLC Genomics Workbench, v.7.5.

The sequence reads for both genotypes were reciprocally mapped to the ‘Glory’ and ‘Staccato’ *de novo* assemblies which were derived from Illumina sequence data. Following read mapping, Fixed Ploidy Variant Detection (CLC Genomics Workbench v7.5) was used to identify predicted variants from the reference genome for each of the genotypes by comparing the variant tracks associated with each of the genotypes. From these analyses 19 single nucleotide variants (SNV), one multi-nucleotide variant (MNV), 13 deletions and 8 insertions for a total of 41 predicted polymorphisms were detected (Table 5).

3.4. Evaluation of cherry SNParray

The SNPs represented on the array are spread relatively evenly across each chromosome, but the finite number of the putative polymorphisms indicates that only a representative subset of potential SNPs can be examined from the sweet cherry genome. The SNParray has been used as a genotyping tool for sweet cherry population structure analysis [21] as well as to determine allelic identity of SSR fragments [22].

Polymorphisms were identified using the SNParray results from ‘Bing’, ‘Sweetheart’, ‘Glory’, ‘Kimberly’, ‘Staccato’ and ‘Stella’. The data obtained from the SNParray was analyzed and genotype specific frequencies of SNPs for each genotype were calculated (Table 6). Interestingly, when ‘Glory’ and ‘Staccato’, were compared to ‘Bing’ and ‘Stella’, approximately 600 SNPs were identified. However, only 70 SNPs were identifiable when the two genotypes were compared to ‘Sweetheart’ and ‘Kimberly’. The SNParray failed to detect any SNPs between ‘Glory’ and ‘Staccato’.

The SNParray represents only a limited number of SNPs which are derived from the originally represented genotypes. Due to these inherent limitations, a SNParray may not be able to identify polymorphisms present in new and closely related genotypes. In such cases, a gel-based or reduced representation or whole genome sequencing based approach becomes necessary.

3.5. Whole genome sequencing to identify polymorphisms

A reference based assembly of the Illumina reads mapped against v2.0 of the peach genome was completed using CLC Genomics Workbench v7.5 to identify regions of conservation and divergence in the *Prunus* genus. A total of 79% of the combined Illumina reads mapped to the peach genome while 18% of the total Illumina reads were mapped to the peach chloroplast, and 3% did not map to either. The eight primary scaffolds of peach had a genome coverage between 37.8 and 61.8 × (Supplemental File 1 contains the coverage statistics for each scaffold and data set). These scaffolds were covered an average of 47.8× for the combined cherry data. These data demonstrated that our reads were of high quality and consistent with published *Prunus* data and can therefore be used in polymorphism analysis.

Five sweet cherry genotypes were selected for whole genome sequencing. Each genotype generated 22.2× average coverage, or 4.6 - 5.5 Gb of sequence. Two sweet cherry genotypes (‘Glory’ and ‘Staccato’) were also sequenced using NGS technology (PacBio RS II, v2 SMRT cell). Paired-end Illumina reads from each sweet cherry genotype were mapped back to the PacBio assemblies of ‘Glory’ and ‘Staccato’. Due to the relatively high error rate of PacBio [13, 23], only the high quality contigs were exported for further analysis. Long-read data for these genotypes provided a scaffold upon which the more accurate short read Illumina data was mapped.

3.5.1. Identification of polymorphisms using Seqman Pro

Following genome assembly, SNPs were identified between genotype pairs using Seqman Pro (DNASTAR, Madison, Wisconsin USA) based on identical contig alignment. Mapping Illumina reads to the long contig scaffolds from PacBio reads reduced the total number of contigs to be compared from more than 100,000 per genotype (from Illumina de novo assembly) to 18,833 and 19,163 for ‘Staccato’ and ‘Glory’, respectively. This approach allowed for a relatively quick and efficient comparison of individual contigs and resulted in high quality SNP identification between genotypes.

The sequence surrounding an identified SNP, 100 bp on each side of the SNP (201 nucleotides), was used to query protein databases (NCBI) to determine the likelihood of the SNP is associated with a known gene. Local blast databases of sequences were made containing the SNP at the minimum threshold of twenty reads and queried with the SNP sequences of the other combinations to check the veracity of the high quality SNPs.

Over 2,000 putative SNPs were identified among the sweet cherry genotypes that were evaluated through stringent filtering steps. These SNPs can be considered as targets for the development of positive identification markers and genotyping assays. These data indicate that while highly similar over the majority of the genome, there are significant differences among the genotypes that can be exploited for genotyping purposes.

The advantages to using this approach include the ability to efficiently compare all assembled sequences between two genotypes. With five genotypes to iteratively compare, all comparisons were performed in a reasonable time-frame. A drawback of this approach is the fact that the comparison is made with consensus assembly sequence. This means read variants depend on assembly parameter stringency.

3.5.2. Identification of polymorphisms using DISCOSNP

To avoid the bias associated with draft reference genomes [24-27], a reference-free approach was used as an alternative method to detect polymorphisms using the Illumina reads. DISCOSNP detects both heterozygous and homozygous isolated SNPs and indels from single or multiple read sets, with no reference genome required [14].

Illumina read data from all comparisons of the samples were analyzed using DISCOSNP and the total number of SNPs for each genotype was compared to 'Stella' (Figure 5). A total of 1,239,949 polymorphisms were identified, with an average of 247,990 per genotype. 'Bing' had the highest percentage (50.3%) of high quality polymorphisms when compared head-to-head with each of the other genotypes, 'Sweetheart', 'Staccato', 'Glory', and 'Kimberly' (Figure 6). This suggests that of the five cultivars in question, 'Bing' is the most genetically divergent. At the same time 'Glory' and 'Staccato' generated 117,796 putative SNPs – the lowest number of all the genotypic comparisons using this method, providing further evidence that 'Glory' and 'Staccato' are the most similar genotypes in this study.

3.5.3. Identification of genome-wide polymorphisms using Stacks

SNPs were identified using Stacks [15, 16] by generating loci from short read Illumina data. Polymorphisms were identified in genotype-specific loci. Overall, 575,008 putative polymorphisms were identified among the compared genotypes (Supplemental file 7).

Following populations analysis using Stacks, the data were filtered to remove loci with missing values, resulting in 9,029 total loci (for each allele) that were used in STRUCTURE analysis. Five iterations were performed in STRUCTURE to analyze K values 1-5. From these analyses, two primary populations emerged. 'Bing' was characterized as being genetically distinct from the other genotypes using this approach and within the non‘Bing’ population, ‘Glory’ and ‘Staccato’ formed their own subpopulation.

Pairwise SNP count data was used to generate a dendrogram using R “plot” and “hclust” functions (Figure 7). This dendrogram is analogous to that generated by NTSys and as with NTSys, the UPGMA method of hierarchical clustering was employed. These data confirmed previous STRUCTURE analysis, demonstrating that ‘Glory’ and ‘Staccato’ are a closely-related subpopulation within the non‘Bing’ population.

4. Conclusion

In this work various polymorphism detection approaches were evaluated using five genotypes of sweet cherry. The TRAP method was used to generate a profile comparable to Amplified Fragment Length Polymorphism (AFLP). TRAP differs from AFLP in that it takes advantage of available sequence information, using the known sequence of a candidate gene as the fixed primer in addition to one or two arbitrary primers to amplify putative candidate gene regions [11]. Similar to other PCR-based approaches, problems with primer specificity leads to concerns about reproducibility for the TRAP approach. However, in contrast to other PCR methods, TRAP uses relatively long primers which support consistency and reproducibility of the approach [11].

The cherry SNParray used in this study represented 4,214 putative SNPs derived from twenty-four accessions [12]. Since this is a hybridization based approach, only those SNPs that were originally represented on the array will be detectable. The array is unable to accurately identify polymorphisms in the context of genomic location in genotypes that were not represented originally on the array.

The TRAPseq approach generated data that was used to identify 41 polymorphisms between ‘Glory’ and ‘Staccato’ using variant detection software. Therefore, TRAPseq is an effective method to generate quality data to identify polymorphisms. It is, however, limited as a reduced representation sequencing approach - it does not provide breadth across the genome and cannot therefore provide a complete profile of sequence-based differences among genotypes.

Whole genome sequencing approach is advantageous in that it provides genome wide coverage and can be easily implemented in species with little or no genetic information. Whole genome sequencing is limited by the depth of coverage and assembly methodology which may be difficult especially around polymorphic repeat regions of the genome. Often other methods can be combined with the short-read approach to arrive at a reasonable solution.

The sweet cherry genotypes ‘Glory’ and ‘Staccato’ are very closely related genotypes. We have demonstrated that no SNPs were detected for these genotypes using 6k cherry SNParray which evaluates 4,214 sweet cherry SNPs. The TRAP assay positively identify a pair of SNPs between ‘Glory’ and ‘Staccato’ in a pairwise comparison (Figure 2). TRAPseq identified 41 putative SNPs between ‘Glory’ and ‘Staccato’. Additionally, whole genome sequencing approaches showed wide variation in identifying high quality polymorphisms using Seqman Pro, DISCOSNP, and Stacks software.

When comparing all five sweet cherry genotypes from this study to the peach reference, 310 regions with higher and lower rates of polymorphism were identified. Higher rates could result from genome duplications or from low conservation yielding more genetic divergence [28]. Similarly, regions with lower than average polymorphisms could be the result of either low divergence where few polymorphisms exist, or of very high rates of genetic differentiation, preventing the mapping of the sequencing reads to these locations.

Both gel-based and sequence based approaches were evaluated for their utility in identifying polymorphisms. These approaches used for discovery and cross-validation of SNPs included TRAP, SNParray, and long and short-read genome sequencing. These approaches each successfully identified polymorphisms. Taken as a whole they provided a robust dataset of predicted polymorphisms as the limitations and strengths of each approach are complementary.

In this study, gel-based and DNA sequence-based approaches toward SNP identification combined to provide the most practical approach to identify these genetic differences. These SNPs are useful in expanding our knowledge of genetics and genomics in Rosaceae species through their use as molecular markers and gene-based interrogations.

References

1. Koepke, T., et al., *Rapid gene-based SNP and haplotype marker development in non-model eukaryotes using 3'UTR sequencing.* BMC genomics, 2012. **13**: p. 18.

2. Arumuganathan, K. and E.D. Earle, *Nuclear DNA content of some important plant species.* Plant Molecular Biology Reporter, 1991. **9**: p. 208-218.

3. Carrasco, B., et al., *Breeding in peach, cherry and plum: From a tissue culture, genetic, transcriptomic and genomic perspective.* Biological Research, 2013. **46**: p. 219-230.

4. Koepke, T., et al., *Comparative genomics analysis in Prunoideae to identify biologically relevant polymorphisms.* Plant Biotechnology Journal, 2013. **11**: p. 883-893.

5. International Peach Genome, I., et al., *The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution.* Nat Genet, 2013. **45**(5): p. 487-94.

6. Dhingra, A., et al. *Sweet Cherry Genome Project*. 2013; Available from: https://genomics.wsu.edu/sweet-cherry-genome-project/.

7. Mariette, S., et al., *Population structure and genetic bottleneck in sweet cherry estimated with SSRs and the gametophytic self-incompatibility locus.* BMC Genet, 2010. **11**: p. 77.

8. Campoy, J.A., et al., *Genetic diversity, linkage disequilibrium, population structure and construction of a core collection of Prunus avium L. landraces and bred cultivars.* BMC Plant Biol, 2016. **16**: p. 49.

9. Fernandez i Marti, A., et al., *Genetic Diversity and Relatedness of Sweet Cherry (Prunus Avium L.) Cultivars Based on Single Nucleotide Polymorphic Markers.* Frontiers in Plant Science, 2012. **3**: p. 1-13.

10. Healey, A., et al., *Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species.* Plant Methods, 2014. **10**: p. 21.

11. Hu, J. and B.A. Vick, *Target Region Amplification Polymorphism: A Novel Marker Technique for Plant Genotyping.* Plant Molecular Biology Reporter, 2003. **21**: p. 289-294.

12. Peace, C., et al., *Development and Evaluation of a Genome-Wide 6K SNP Array for Diploid Sweet Cherry and Tetraploid Sour Cherry.* PloS one, 2012. **7**.

13. Roberts, R.J., M.O. Carneiro, and M.C. Schatz, *The advantages of SMRT sequencing.* Genome biology, 2013. **14**: p. 405.

14. Uricaru, R., et al., *Reference-free detection of isolated SNPs.* Nucleic Acids Research, 2015. **43**: p. e11.

15. Catchen, J.M., et al., *Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences.* Genes|Genomes|Genetics, 2011. **1**: p. 171-182.

16. Catchen, J., et al., *Stacks: An analysis tool set for population genomics.* Molecular Ecology, 2013. **22**: p. 3124-3140.

17. Rosyara, U.R., et al., *Identification of the Paternal Parent of 'Bing' Sweet Cherry and Confirmation of Descendants Using Single Nucleotide Polymorphism Markers.* J Amer Soc Hort Sci, 2014. **139**: p. 148-156.

18. Hedhly, A., et al., *Paternal-specific S-allele transmission in sweet cherry (Prunus avium L.): the potential for sexual selection.* J Evol Biol, 2016. **29**(3): p. 490-501.

19. Ipek, A., et al., *Determination of self-incompatibility groups of sweet cherry genotypes from Turkey.* Genet Mol Res, 2011. **10**(1): p. 253-60.

20. Castède, S., et al., *Mapping of Candidate Genes Involved in Bud Dormancy and Flowering Time in Sweet Cherry (Prunus avium).* PloS one, 2015. **10**: p. e0143250.

21. Castede, S., et al., *Genetic determinism of phenological traits highly affected by climate change in Prunus avium: Flowering date dissected into chilling and heat requirements.* New Phytologist, 2014. **202**: p. 703-715.

22. De Franceschi, P., et al., *Cell number regulator genes in Prunus provide candidate genes for the control of fruit size in sweet and sour cherry.* Molecular breeding : new strategies in plant improvement, 2013. **32**: p. 311-326.

23. Ferrarini, M., et al., *An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome.* BMC genomics, 2013. **14**: p. 670.

24. Bradnam, K.R., et al., *Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species.* GigaScience, 2013. **2**: p. 10.

25. Ekblom, R. and J.B.W. Wolf, *A field guide to whole-genome sequencing, assembly and annotation.* Evolutionary Applications, 2014: p. n/a-n/a.

26. Hunt, M., et al., *REAPR: a universal tool for genome assembly evaluation.* Genome biology, 2013. **14**: p. R47.

27. Utturkar, S.M., et al., *Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences.* Bioinformatics, 2014. **30**: p. 2709-2716.

28. Clevenger, J., et al., *Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations.* Molecular Plant, 2015. **8**: p. 831-846.

Tables

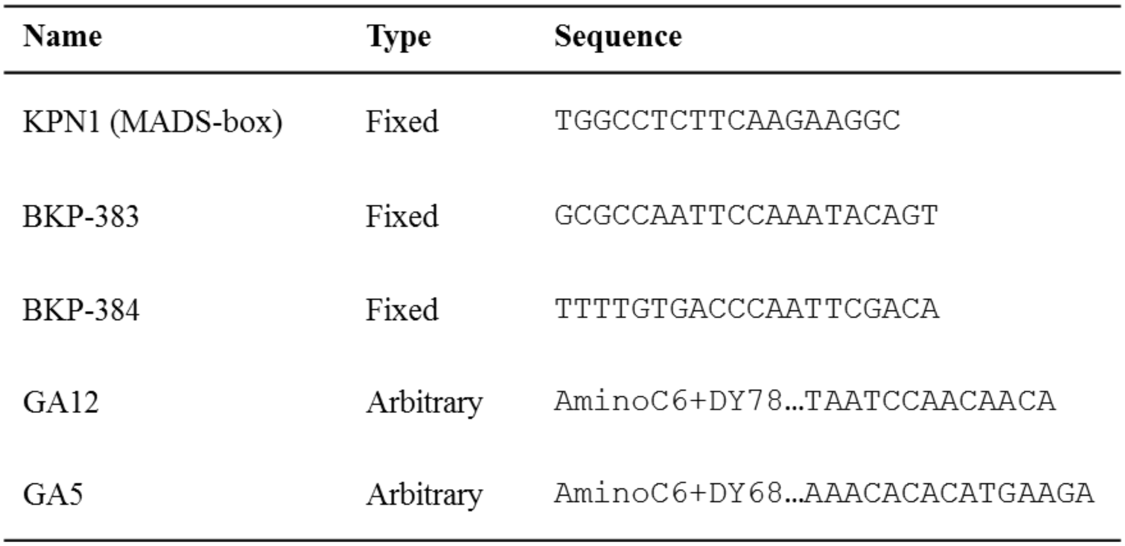
Table 1. Sweet cherry genome assembly statistics. *De novo* assembled using CLC Genomics Workbench v. 8.0.

	Bing	**Sweetheart**	**Staccato**	**Glory**	**Kimberly**
Total Clean Reads	55571037	57636693	64192322	56010744	51053859
Total Clean Nucleotides	5150832924	5346338137	5972602461	5151578386	4728316659
N percentage	1.8%	1.7%	2.2%	2.3%	2.6%
GC percentage	36.6%	36.7%	36.4%	36.4%	36.2%
Contig Number	72246	70143	66479	69452	71072
Contig Total Length (nt)	163419540	165662299	169722627	163687825	162087977
Contig Mean Length (nt)	1338	2362	2553	2356	2281
Contig N50	3574	6699	7456	6466	6267

Table 2. Sweet cherry S-alleles. Incompatability genotypes of sweet cherry cultivars that were used in this study. Group SC indicates that these cultivars are Self Compatible (SC) and can pollinate flowers with the same S-alleles. Group III cultivars are incompatible with themselves and require cross-pollination with another group for optimal fruit set.

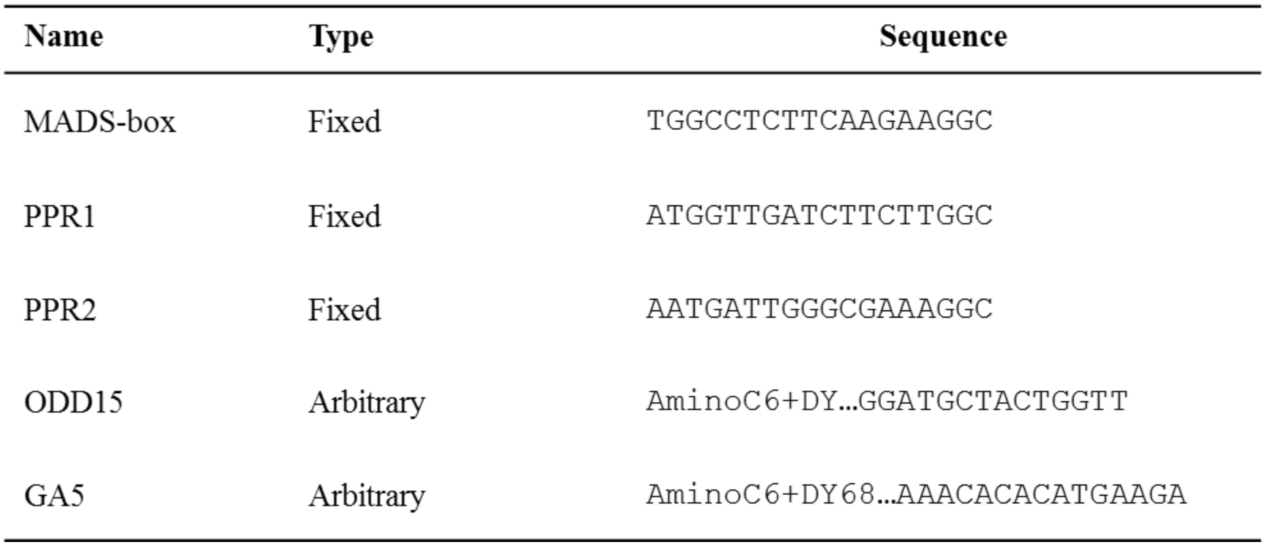
Cultivar	S-Allele	Group
Bing	S_3S_4	III
Sweetheart	S_3S_4'	SC
Staccato	S_3S_4'	SC
Glory	S_3S_4'	SC
Kimberly	S_3S_4	III
Stella	S_3S_4'	SC

Table 3. TRAP Primers. The name, sequence and type of primer used in the TRAP Assay.



Name	Type	Sequence
KPN1 (MADS-box)	Fixed	TGGCCTCTTCAAGAAGGC
BKP-383	Fixed	GCGCCAATTCCAAATACAGT
BKP-384	Fixed	TTTTGTGACCCAATTCGACA
GA12	Arbitrary	AminoC6+DY78…TAATCCAACAACA
GA5	Arbitrary	AminoC6+DY68…AAACACACATGAAGA

Table 4. TRAPseq Primer details. Primer details including name, sequence and type of primer used in the TRAP PCR for TRAPseq polymorphism detection. A fixed primer was used in conjunction with both of the arbitrary primers for amplification of each of the genotypes examined.



Name	Type	Sequence
MADS-box	Fixed	TGGCCTCTTCAAGAAGGC
PPR1	Fixed	ATGGTTGATCTTCTTGGC
PPR2	Fixed	AATGATTGGGCGAAAGGC
ODD15	Arbitrary	AminoC6+DY…GGATGCTACTGGTT
GA5	Arbitrary	AminoC6+DY68…AAACACACATGAAGA

Table 5. Polymorphisms derived from TRAPseq approach. Variants were identified using the Fixed Ploidy Variant Detection tool (CLC Genomics Workbench v7.5). This experiment compared 'Glory' reads to a 'Staccato' *de novo* assembly. Nineteen single nucleotide variants (SNV), a single multi-nucleotide variant (MNV), thirteen deletions and 8 insertions for a total of 41 predicted polymorphisms were detected.

Mapping Contig Name	Ref. Position	Polymorphism Type	Length	Reference	Allele	Count	Coverage	Frequency	Fwd/Rvs balance
Staccato_contig_7	262	SNV	1	C	A	166	167	99.401	0.024
Staccato_contig_13	8	SNV	1	T	G	11	12	91.667	0.273
Staccato_contig_13	14	SNV	1	C	A	13	14	92.857	0.385
Staccato_contig_25	375	Deletion	1	A	-	384	405	94.815	0.266
Staccato_contig_28	721	Deletion	1	A	-	99	105	94.286	0.394
Staccato_contig_29	213	Deletion	1	A	-	11	11	100.000	0.455
Staccato_contig_34	93	Insertion	1	-	G	3875	4043	95.845	0.376
Staccato_contig_34	97	Deletion	1	T	-	4166	4272	97.519	0.381
Staccato_contig_34	100	Insertion	1	-	C	4187	4272	98.010	0.382
Staccato_contig_44	338	Deletion	1	C	-	15	15	100.000	0.067
Staccato_contig_44	342	Deletion	1	G	-	14	14	100.000	0.071
Staccato_contig_53	946	Deletion	1	C	-	51	51	100.000	0.275
Staccato_contig_100	436	Insertion	1	-	G	12	13	92.308	0.250
Staccato_contig_100	438	Insertion	3	-	AAA	12	12	100.000	0.250
Staccato_contig_100	440	Insertion	2	-	TA	12	12	100.000	0.250
Staccato_contig_164	530	Deletion	1	C	-	42	42	100.000	0.167
Staccato_contig_164	540	Deletion	1	C	-	38	40	95.000	0.158
Staccato_contig_166	26	Insertion	1	-	G	77	77	100.000	0.403
Staccato_contig_166	27	SNV	1	T	G	77	80	96.250	0.403
Staccato_contig_181	19	SNV	1	A	C	12	13	92.308	0.417
Staccato_contig_181	22	SNV	1	C	T	12	13	92.308	0.417
Staccato_contig_251	444	SNV	1	A	C	13	13	100.000	0.308
Staccato_contig_438	323	Deletion	1	A	-	11	12	91.667	0.364
Staccato_contig_438	363	Deletion	2	GC	-	11	11	100.000	0.364
Staccato_contig_506	14	Deletion	1	T	-	10	10	100.000	0.300
Staccato_contig_704	83	SNV	1	T	C	11	11	100.000	0.455
Staccato_contig_746	113	SNV	1	C	T	33	33	100.000	0.455
Staccato_contig_746	119	MNV	2	TG	CA	25	26	96.154	0.440
Staccato_contig_746	125	SNV	1	G	T	23	23	100.000	0.478
Staccato_contig_795	122	Deletion	1	A	-	13	14	92.857	0.462
Staccato_contig_943	167	SNV	1	A	T	17	17	100.000	0.294
Staccato_contig_1041	370	Insertion	1	-	A	2120	2187	96.936	0.286
Staccato_contig_1061	136	Insertion	1	-	T	42	42	100.000	0.333
Staccato_contig_1116	374	SNV	1	T	A	32	35	91.429	0.438
Staccato_contig_1243	354	SNV	1	G	A	57	57	100.000	0.386
Staccato_contig_1243	356	SNV	1	C	T	57	57	100.000	0.386
Staccato_contig_1243	369	SNV	1	G	A	108	115	93.913	0.463
Staccato_contig_1243	378	SNV	1	G	T	121	124	97.581	0.479
Staccato_contig_1243	380	SNV	1	C	T	121	124	97.581	0.479
Staccato_contig_1243	384	SNV	1	G	A	135	141	95.745	0.444
Staccato_contig_1299	119	SNV	1	A	T	19	20	95.000	0.421

Table 6. The percentage and total number of unique SNPs on the sweet cherry SNParray. The bottom-left half of the table shows the percentages of SNPs held in common between two compared genotypes. The upper-right half of the table show the actual count of SNPs called between each genotypic comparison. Using SNParray there were found no unique polymorphisms between ‘Staccato’ and ‘Glory’ cultivars as well as the ‘Kimberly’ and ‘Sweetheart’ cultivars, however, unique SNPs were found between all other comparisons.

	Bing	**Sweetheart**	**Staccato**	**Glory**	**Kimberly**	**Stella**	
Bing	-	566	607	606	565	515	**Count of unique SNPS in RosBreed cherry SNP Array**
Sweetheart	9.9%	-	68	67	0	504	
Staccato	10.7%	1.2%	-	0	68	539	
Glory	10.6%	1.2%	0.0%	-	67	537	
Kimberly	9.9%	0.0%	1.2%	1.2%	-	503	
Stella	9.0%	8.8%	9.5%	9.4%	8.8%	-	
	Percentage of unique SNPs in RosBreed cherry SNP Array						

Table 5. Verification of TRAPseq derived Polymorphisms. A subset of SNP sequences from the 6k cherry SNParray [12] were found on NCBI and blasted against the de novo assembly (Bing). Each SNP was visually identified in the BLAST results. The subset of SNPs was randomly selected from across 5 chromosomal locations and all 20 tested contained the predicted SNP.

NCBI SS#	Original Full Name in 6K SNP Array	SNP Sequence (NCBI)	Sequence (BLAST sequence, Bing)	Predicted SNP	Actual SNP	Chromosome	Physical position (bp^a)	Estimated genetic position #1 (cM)	Estimated genetic position #2 (cM)
ss490545369	RosBREED_snp_sweet_cherry_Pp1_00094685	TTCAGTAGAGCTCCCTGGGGCTTGC [A/C] AAGTTTGTCAAACTCAATCAACTTG	TTCAGTAGAGCTCCCTGGGGCTTGCCAAGTTTGTCAAACTCAATCAACTTG	C	C	1	94685	0.21	0.00
ss490545372	RosBREED_snp_sweet_cherry_Pp1_00114487	AACAACTACCATAACAGACCTTCGA [C/T] GATAGTATACAACAAACGAAACCAC	AACAACTACCATAACAGACCTTCGATGATAGTATACAACAAACGAAACCAC	T	T	1	114487	0.26	0.00
ss490545375	RosBREED_snp_sweet_cherry_Pp1_00154827	TTGCTTTGTAGACCTTGTCCATCTA [C/T] CTTGTAGTTTTCTTCTTTTCATTAA	TTGCTTTGTAGACCTTGTCCATCTATCTTGTAGTTTTCTTCTTTTCATTAA	T	T	1	154827	0.35	0.00
ss490545378	RosBREED_snp_sweet_cherry_Pp1_00196391	TTTGGAATGTTCTTGCTCCAATTGA [G/T] CTTCTCAGCTTTGGCAGTAGTTCTC	TTTGGAATGTTCTTGCTCCAATTGATCTTCTCAGCTTTGGCAGTAGTTCTC	T	T	1	196391	0.44	0.00
ss490548745	RosBREED_snp_sweet_cherry_Pp2_00911589	ACACCCATACCCAAGTTCTTCCGTC [A/G] GAAGATGATGATGGTGCTAAGAATA	ACACCCATACCCAAGTTCTTCCGTCGGAAGATGATGATGGTGCTAAGAATA	G	G	2	911589	1.94	0.00
ss490548749	RosBREED_snp_sweet_cherry_Pp2_00979200	AAGACGGATAGCCAGGGTGAAAAAA [A/C] CTTGCCAAGTAACTAATTAATAGCA	AAGACGGATAGCCAGGGTGAAAAAACCTTGCCAAGTAACTAATTAATAGCA	C	C	2	979200	2.08	0.00
ss490548753	RosBREED_snp_sweet_cherry_Pp2_01029506	GGTTTTGTAAGGATGGTATACCTTA [C/T] TGGGAAAAGCAATTCTGCACTTTGG	GGTTTTGTAAGGATGGTATACCTTATTGGGAAAAGCAATTCTGCACTTTGG	T	T	2	1029506	2.19	0.00
ss490548757	RosBREED_snp_sweet_cherry_Pp2_01123211	TGTGTTTAACAACTTTGTCCTTGCA [A/C] AGTTTAACTGGGCAACAATATACTG	TGTGTTTAACAACTTTGTCCTTGCACAGTTTAACTGGGCAACAATATACTG	C	C	2	1123211	2.39	0.00
ss490552278	RosBREED_snp_sweet_cherry_Pp4_00355217	TAACTTCTTGCATCTTGAGGAAAAC [A/G] GGTGGTGGTAACCTAATCTCGCTGC	TAACTTCTTGCATCTTGAGGAAAACGGGTGGTGGTAACCTAATCTCGCTGC	G	G	4	355217	2.47	0.92
ss490552281	RosBREED_snp_sweet_cherry_Pp4_00394639	TTATTGCGCCAGAATCTGAGCTGAG [A/C] CGAGACGAGACTTGCCTATGGTTCA	TTATTGCGCCAGAATCTGAGCTGAGCCGAGACGAGACTTGCCTATGGTTCA	C	C	4	394639	2.58	1.07
ss490552284	RosBREED_snp_sweet_cherry_Pp4_00430644	CTGCAAAAACAACCAGCTCCGTGAA [G/T] ACATAAACACGCAGCATCCAAATGC	CTGCAAAAACAACCAGCTCCGTGAATACATAAACACGCAGCATCCAAATGC	T	T	4	430644	2.69	1.19
ss490552287	RosBREED_snp_sweet_cherry_Pp4_00473163	TCATCGGATTGGATTACCTCTCGTT [C/T] GAGTCTGAGGTGAAGGTTTATAGCC	TCATCGGATTGGATTACCTCTCGTTTGAGTCTGAGGTGAAGGTTTATAGCC	T	T	4	473163	2.81	1.35
ss490555002	RosBREED_snp_sweet_cherry_Pp6_00940675	TTCTTTACTCAGTTCTTGGTCACTG [A/C] AAAGTTCATCCGACTCTTGGTGCAC	TTCTTTACTCAGTTCTTGGTCACTGCAAAGTTCATCCGACTCTTGGTGCAC	C	C	6	940675	9.34	0.00
ss490555005	RosBREED_snp_sweet_cherry_Pp6_00981908	GATAAGTGCTGGTGAGGTTTTACAT [A/G] TCATAATATACCTGGTCTGTTTCGT	GATAAGTGCTGGTGAGGTTTTACATGTCATAATATACCTGGTCTGTTTCGT	G	G	6	981908	9.40	0.00
ss490555008	RosBREED_snp_sweet_cherry_Pp6_01023194	GGTCTATTGATTCTGGAAATGCTGC [A/G] AATGGTCAGATCCAATCTGAGCGCT	GGTCTATTGATTCTGGAAATGCTGCGAATGGTCAGATCCAATCTGAGCGCT	G	G	6	1023194	9.46	0.00
ss490555011	RosBREED_snp_sweet_cherry_Pp6_01062002	ATAGGCAACATTAAAAATATTTAAT [A/G] GGAGGCTGTTAATTCTTGCAATGCT	ATAGGCAACATTAAAAATATTTAATGGGAGGCTGTTAATTCTTGCAATGCT	G	G	6	1062002	9.51	0.00
ss490558164	RosBREED_snp_sweet_cherry_Pp8_13263864	TCTGGAAGAACTTGAAGGAATTCAC [A/G] GGTTTCAAATCCAAGAGCGGGTTGA	TCTGGAAGAACTTGAAGGAATTCACGGGTTTCAAATCCAAGAGCGGGTTGA	G	G	8	13263864	24.29	35.79
ss490558167	RosBREED_snp_sweet_cherry_Pp8_13297952	TACTAGTTTCTTCTTTTCTTTGGTC [A/G] GCATCTTCCTTGCATTCCTTATAGT	TACTAGTTTCTTCTTTTCTTTGGTCGGCATCTTCCTTGCATTCCTTATAGT	G	G	8	13297952	24.33	35.95
ss490558170	RosBREED_snp_sweet_cherry_Pp8_13376142	ATCTTTGGTGCTTTCACCATTTGAC [A/G] TGGAGAGGCTCTCTTTCTCCTCCTT	ATCTTTGGTGCTTTCACCATTTGACGTGGAGAGGCTCTCTTTCTCCTCCTT	G	G	8	13376142	24.43	36.31
ss490558173	RosBREED_snp_sweet_cherry_Pp8_13397168	TTGCAATCACTCTAATGCTTCTACT [A/C] TTTCTTGCAGATGGCCTGATCTGGA	TTGCAATCACTCTAATGCTTCTACTCTTTCTTGCAGATGGCCTGATCTGGA	C	C	8	13397168	24.46	36.41

Figures

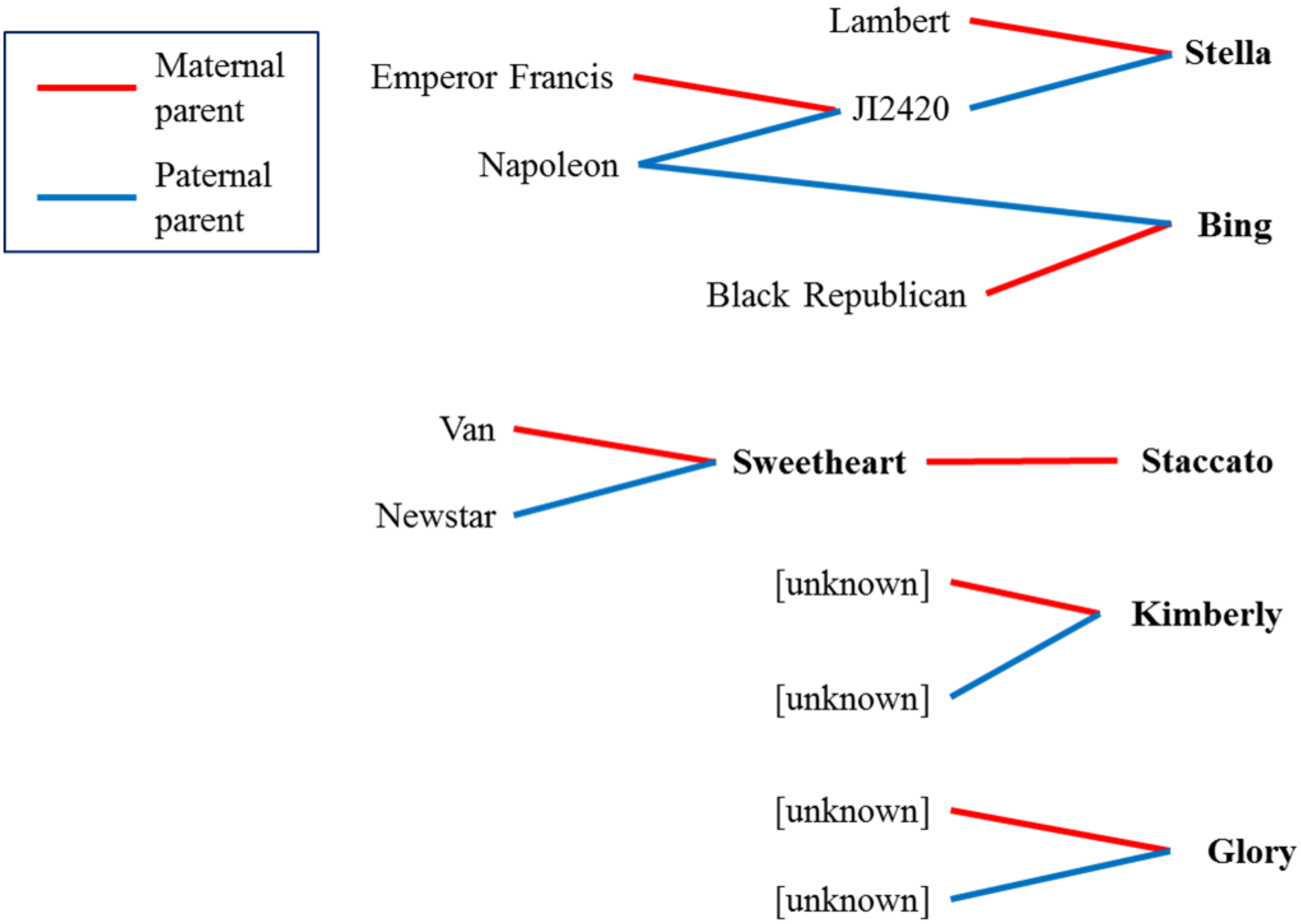


Figure 1– Pedigree relationships of six of the sweet cherry cultivars described in this study (bold). Pedigree of the sweet cherry cultivars used for SNP development. The maternal parent is marked by a red line and the parental parent by a blue line.

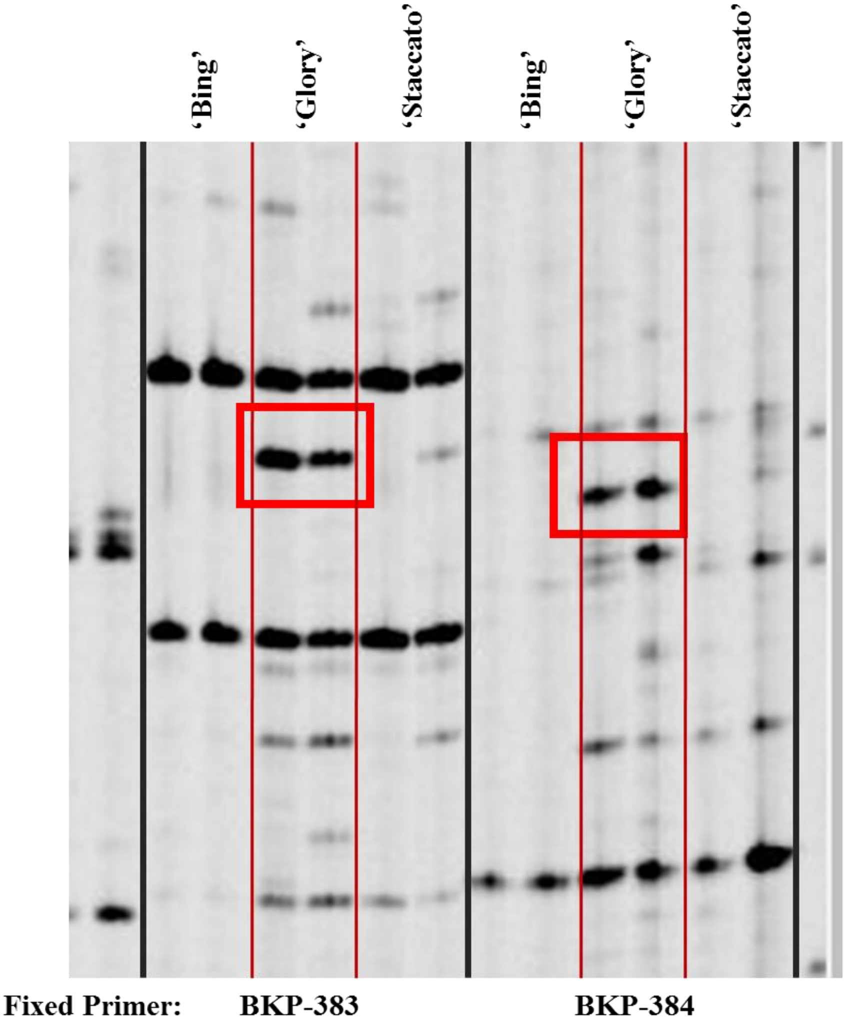
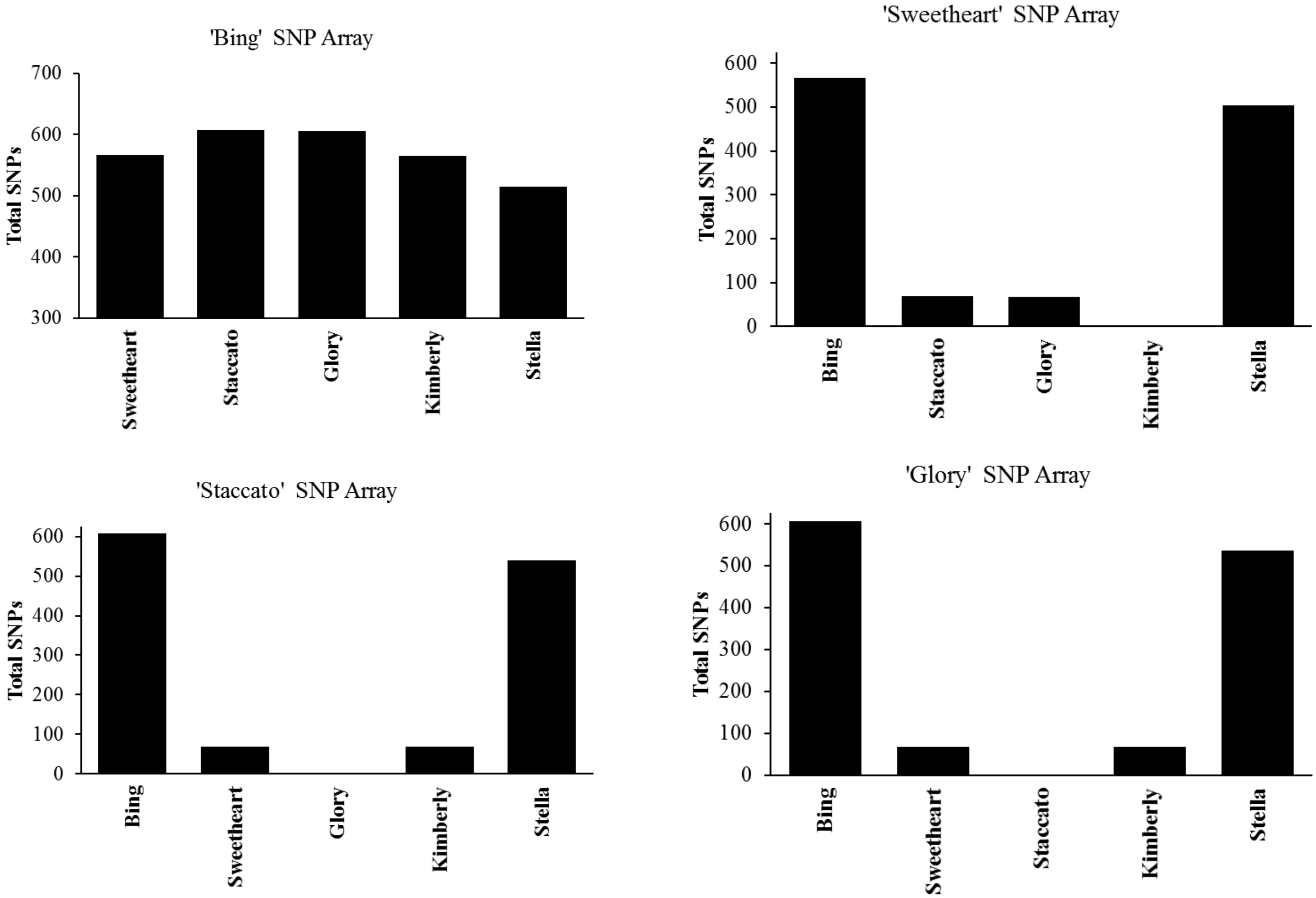


Figure 2. TRAP Assay of Bing, Glory and Staccato sweet cherry cultivars. Assays replicated in duplicate. Primer screen was performed using fixed primers BKP-383, 384 and arbitrary primers SA12, GA5. Primer sequences are available (Table 3). Red boxes are indicative of putative TRAP markers. The size of the BKP-383 and BKP-384 'Glory' markers are approximately 336 and 330 bp, respectively.


'Bing' SNP Array
Total SNPs
700
600
500
400
300
Sweetheart
Staccato
Glory
Kimberly
Stella
'Sweetheart' SNP Array
Total SNPs
600
500
400
300
200
100
0
Bing
Staccato
Glory
Kimberly
Stella
'Staccato' SNP Array
Total SNPs
600
500
400
300
200
100
0
Bing
Sweetheart
Glory
Kimberly
Stella
'Glory' SNP Array
Total SNPs
600
500
400
300
200
100
0
Bing
Sweetheart
Staccato
Kimberly
Stella

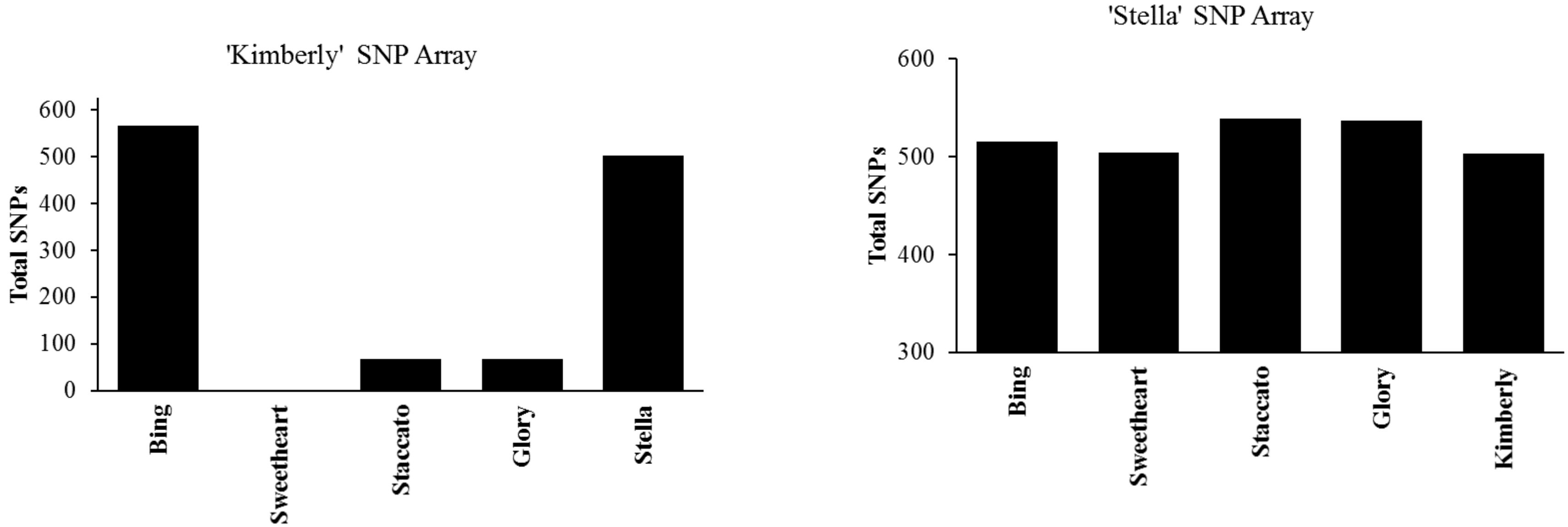


Figure 3. Individual genotype comparisons of SNPs using SNParray. The title of each subfigure indicates the reference by which the listed genotypes were compared.

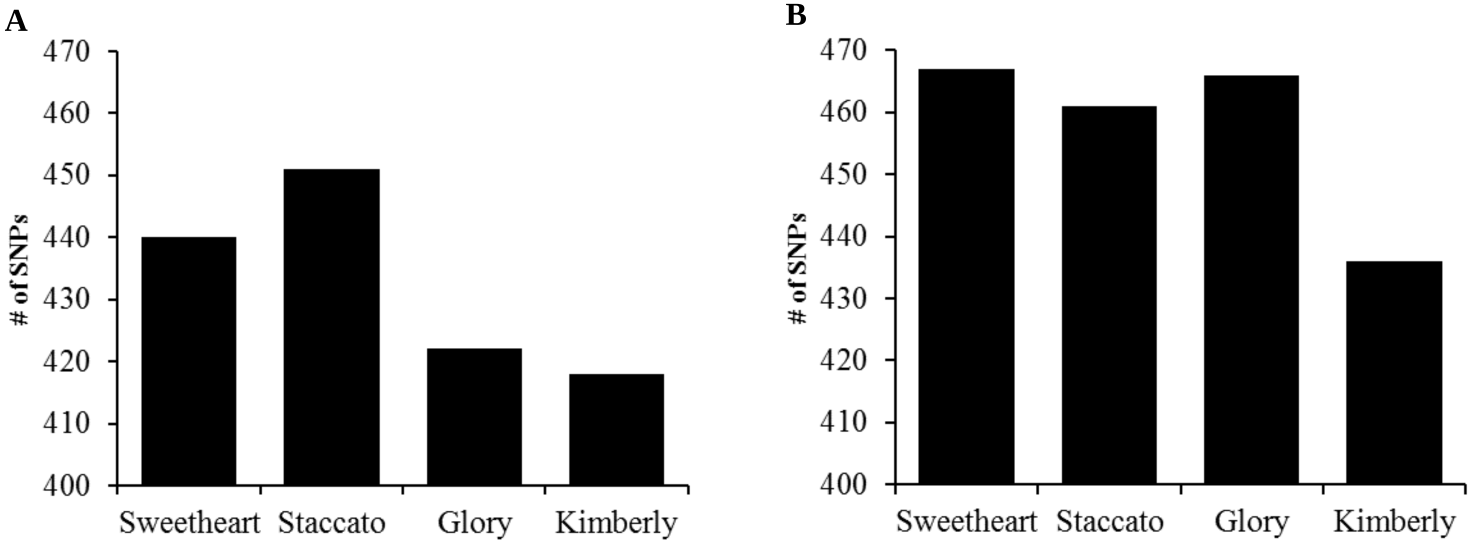


Figure 4. NGen-Seqman putative SNP detection. These data are based on mapping Illumina reads to the Glory (A) and Staccato (B) PacBio *de novo* assemblies. The putative polymorphisms from ‘Bing’ sweet cherry were compared to each genotype to provide a baseline comparison from which all genotypes could be compared. The polymorphisms shown were stringently filtered to only retain those polymorphisms with a minimum threshold of twenty reads called for a given polymorphism.

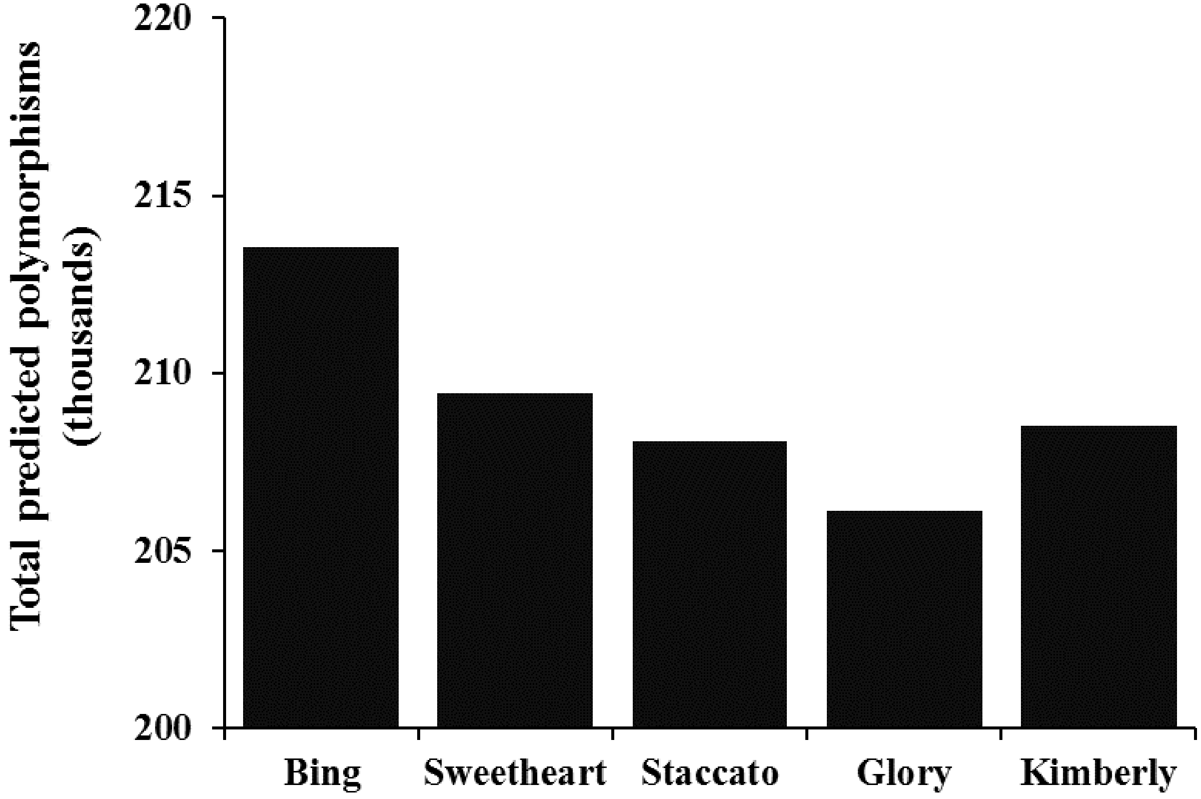


Figure 5. Total predicted polymorphisms using DISCOSNP. Each of the genotypes was compared against the 'Stella' sweet cherry genome. The total number of polymorphisms corroborate similar trends using the other polymorphism detection methods used in this study.

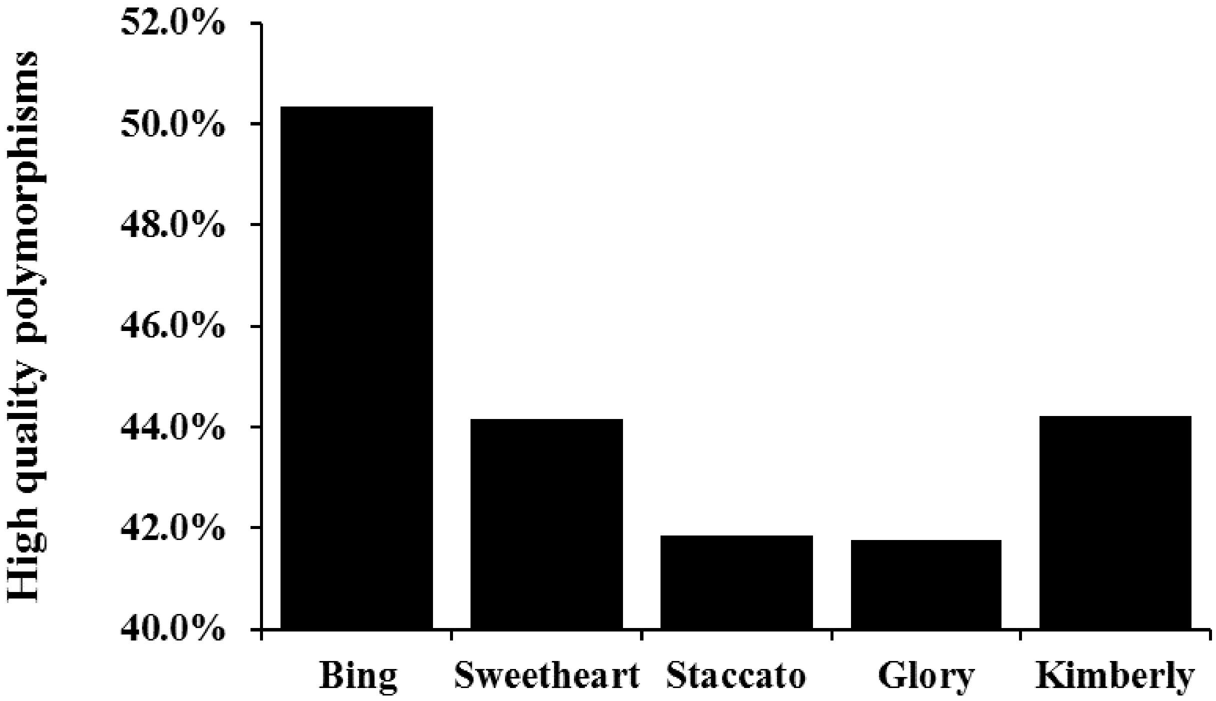


Figure 6. Percentage of high quality polymorphisms for each genotype using DiscoSNP. High quality polymorphisms were determined by keeping predicted polymorphisms with the score of 1.0, indicating that each read analyzed contained the polymorphism.

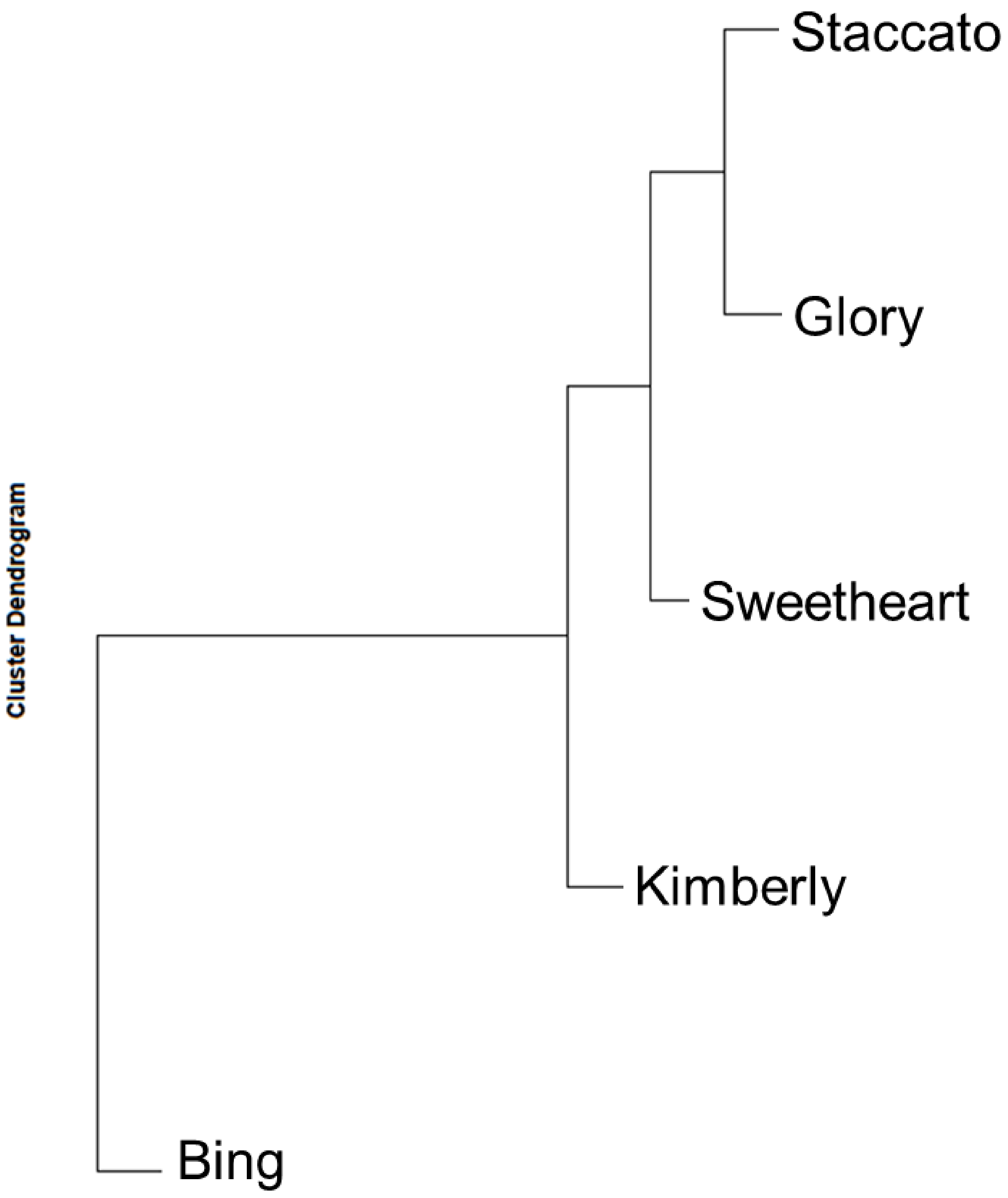


Figure 7. Stacks cluster dendrogram of genotypic relationships. Two primary populations were identified between ‘Bing’ and the other four genotypes. Within these populations, ‘Staccato’ and ‘Glory’ are a distinct subpopulation.

Chapter 2 - Supplemental Files

Supplemental File 1a. ‘Stella’ mapped to peach genome (.fasta)

Supplemental File 1b. ‘Bing’ mapped to peach genome (.fasta)

Supplemental File 1c. ‘Sweetheart’ mapped to peach genome (.fasta)

Supplemental File 1d. ‘Staccato’ mapped to peach genome (.fasta)

Supplemental File 1e. ‘Glory’ mapped to peach genome (.fasta)

Supplemental File 1f. ‘Kimberly’ mapped to peach genome (.fasta)

Supplemental File 2a. ‘Stella’ *de novo* genome assembly (.fasta)

Supplemental File 2b. ‘Bing’ *de novo* genome assembly (.fasta)

Supplemental File 2c. ‘Sweetheart’ *de novo* genome assembly (.fasta)

Supplemental File 2d. ‘Staccato’ *de novo* genome assembly (.fasta)

Supplemental File 2e. ‘Glory’ *de novo* genome assembly (.fasta)

Supplemental File 2f. ‘Kimberly’ *de novo* genome assembly (.fasta)

Supplemental File 3a. ‘Bing’ mapped to PacBio assembly, ‘Glory’ (.fasta)

Supplemental File 3b. ‘Sweetheart’ mapped to PacBio assembly, ‘Glory’ (.fasta)

Supplemental File 3c. ‘Staccato’ mapped to PacBio assembly, ‘Glory’ (.fasta)

Supplemental File 3d. ‘Glory’ mapped to PacBio assembly, ‘Glory’ (.fasta)

Supplemental File 3e. ‘Kimberly’ mapped to PacBio assembly, ‘Glory’ (.fasta)

Supplemental File 4. SNParray data (.xlsx)

Supplemental File 5. DiscoSNP User Guide (.pdf)

Supplemental File 6. NGen-Seqman SNPs (.xlsx)

Supplemental File 7. Stacks SNPs output (.xlsx)

Competing interests

The authors declare no competing interests.

Authors' contributions:

Conceived and designed the experiments: BK, SH, TK, AD

Performed the experiments: SH, BK, RH, TK

Analyzed the data: SH, BK, AD

Contributed reagents/materials/analysis tools: TK, AD

Wrote the paper: BK, SH, AD

Acknowledgements

SNParray collaboration with Iezonni lab, MSU. Initial SNParray analysis performed by Dr. Yunyang Zhao. Illumina & PacBio sequence data donated to Dhingra lab by VanWell Nurseries, Inc.

CHAPTER 3

Developmental time course transcriptome analysis of ethylene inducible pedicel-fruit abscission zone formation in non-climacteric sweet cherry (*Prunus avium* L.)

Benjamin Kilian[1,2], Tyson Koepke[1,2*], Jonathan Abarca[1], Rick Sharpe[1,2], and Amit Dhingra[1,2§]

[1]Department of Horticulture, Washington State University, Pullman WA 99164-6414

[1]Department of Horticulture, Washington State University, Pullman WA 99164 USA

[2]Molecular Plant Sciences Graduate Program, Washington State University, Pullman, WA 99164 USA

* Present address: Phytelligence Inc., 1615 NE Eastgate Blvd #3, Pullman, WA 99163

[§]Corresponding author

Email address:

AD: adhingra@wsu.edu

Abstract

Background: As a non-climacteric fruit, sweet cherry does not produce the defining ethylene burst characteristic of climacteric species. However, previous work has shown that exogenous ethylene treatment of the tree, when the fruit has attained 80% maturity prior to harvest, elicits an abscission response at pedicel-fruit junction in a genotype-dependent manner. The genetic mechanism underlying this novel ethylene response in sweet cherry remains to be unraveled.

Results: A developmental time-course transcriptome analysis was performed on the fruit-pedicel abscission zone following an exogenous ethylene treatment. Three genotypes, ‘Chelan’, ‘Bing’ and ‘Skeena’ representing the range of observed phenotypic responses were used in this study. Phenotypic data and abscission zone samples were excised and collected prior to ethylene (Ethephon) application (240 ppm), 6 hours post treatment, 7 days post-treatment, and 14 days post-treatment. Transcript data were assembled into a representative transcriptome and relative expression values were calculated. Various transcription factors, abscission-related genes and transcripts of unknown function exhibited differential expression in a genotype dependent manner over the developmental time-course. Differential expression was confirmed using quantitative reverse transcription PCR. Gene ontology and pathway information also identified gene-network components involved in the abscission process.

Conclusions: This work has identified some of the genetic regulation components that are induced by ethylene in the formation of the pedicel fruit abscission zone in a non-climacteric plant species where ethylene is not indicated to participate in the fruit ripening process. The information gleaned from this work may offer insight into the molecular mechanisms underlying abscission in non-climacteric systems. These data are expected to be used for identification of

allelic variation in the abscission-related genes which can be utilized in marker-assisted breeding of novel varieties that align with new and developing fruit harvest technologies.

Key Words

Ethephon, ethylene, RNA-seq, transcriptome, *Prunus avium*, abscission, non-climacteric, qRT-PCR, gene expression, Rosaceae

Background

Organ abscission and senescence in Angiosperms are coordinated by a complex network of endogenous and exogenous signaling pathways. Fruit ripening, a subset of senescence processes, manifests as biochemical and physiological changes affecting appearance, texture, flavor and aroma of the fruit [1]. Developmental changes accomplish organ disintegration primarily for seed dissemination purposes. Abscission is the targeted senescence of specific cell layers leading to fruit separation from the main plant body [2]. Determinate plant structures such as leaves, flowers, corollas, and fruit often abscise following their functional lifespans [3, 4]. Organ separation from the plant results from adhesion loss between cells caused by middle lamella dissolution. This occurs through the action of hydrolytic enzymes such as polygalacturonase and cellulose [5]. The plant hormones auxin and ethylene are primary regulators of abscission processes. These two plant hormones generate a precisely balanced regulatory mechanism, controlling the development of cell size and shape in the separation layer [6] which sets up abscission zones through progressive structural changes. Minute shifts to internal hormonal balance initiate molecular signals resulting in morphological changes. If the ratio of auxin to ethylene is disturbed in favor of auxin, ethylene biosynthesis tends to be suppressed, effectively reducing all ethylene dependent responses. If the ratio is reversed, ethylene is upregulated in a feed-forward manner. In the context of abscission, ethylene binds to specific receptors in AZ cells, initiating a signal transduction pathway leading to cell death, reduction in cell wall adhesion, and eventually, organ abscission [4].

As the first gaseous phytohormone identified, ethylene (C_2H_4) regulates numerous developmental processes such as climacteric fruit ripening, senescence, abscission, and even volatile (aroma) production [7, 8]. Ethylene has been shown to be instrumental in other phases of plant development as well, from germination to flowering. The pathway for ethylene transduction has upstream components common to many ethylene signaling responses [9]; therefore, the superficially simple nature of the plant ethylene signaling pathway (see chapter 1, Figure 1) does not adequately explain the diverse ethylene responses. The different responses to ethylene are regulated by ethylene response factor (ERF) transcription factors. These genes are encoded by a large plant transcription factor family, and therefore optimally confer a large diversity and specificity of ethylene responses. This includes, presumably, the abscission zone formation response. Understanding control mechanisms that underlie ethylene action specificity requires revealing the myriad components that mediate ethylene function that is specific to a variety of developmental processes. For example, identifying ripening-associated transcriptional regulators through gene expression correlation has enhanced our understanding of mechanisms that control ripening in fruits [10, 11].

While evidence directly linking ethylene production and leaf abscission remains elusive [13], ethylene has been shown to accelerate corolla and flower abscission [14]. For many species, not only is endogenous ethylene concentration important, but ethylene sensitivity of tissue determines the hormone’s effect [12, 14]. Although there are many biochemical similarities among leaf, flower and fruit abscission, regulatory pathways differ dramatically and are determined by such diverse factors as the species, the organ and the physical location of the AZ [13, 15]. For instance, fruitlet abscission in citrus, mango, apple and cherry has been a valuable resource in abscission research, demonstrating the connection between a change in ethylene concentration in various fruit tissue and whole fruit abscission [16-19]. Ethylene has been demonstrated to induce the formation of an abscission zone in *Prunus cerasus* (sour cherry) at the pedicel-stem junction, loosening the fruit and allowing for efficient mechanical harvesting [20].

Enzymes including pectinase, cellulase and polygalacturonase have been shown to be regulated by ethylene and correlated with plant abscission [21]. Others such as peroxidases, chitinases, uronic acid oxidase, and β-1,3-glucanase are associated with abscission processes but are not induced by ethylene, suggesting that although ethylene may regulate a subset of genes, an alternate mechanism may control other abscission-related genes [2].

Sweet cherry (*Prunus avium* L.), a member of the Rosaceae family has been classified as a non-climacteric fruit. There is some evidence, however, that the characteristic respiratory burst may occur earlier in development than when humans harvest the fruit [22]. This leads to the hypothesis that sweet cherry could be considered climacteric with a divergent climacteric response timeline. Further studies evaluating this hypothesis in sweet cherry could have profound implications in the understanding of angiosperm development and fruit ripening.

In this study, we begin with the assumption of non-climacteric ripening for sweet cherry and there are underlying genetic reasons for the different ripening responses. A recent report investigating sweet cherry polymorphisms identified several missense mutations in ACS and ACO genes (involved in ethylene biosynthesis) implying the possibility of underlying causality for non-climacteric ripening pattern [23]. In sweet cherry, a genotype-specific reduction in pedicel-fruit retention force (PFRF) in response to ethephon application prior to ripening has been demonstrated [24]. Ethephon (2-Chloroethylephosphonic acid) is a commercially available plant growth regulator that is rapidly metabolized to ethylene upon foliar application. For sweet cherry cv. ‘Bing’, ethephon application approximately ten days before harvest has been shown to be the minimal temporal threshold required for subsequent mechanical harvest [25]. Under the same regime, cv. ‘Chelan’ PFRF remained higher than the level required for mechanical harvest (400 g of pedicel pull force). ‘Skeena’ PFRF naturally declines below the threshold for mechanical harvest regardless of ethephon application. A sweet cherry phenotyping study has shown no significant correlation between PFRF and fruit quality across a wide variety of commercial sweet cherry cultivars and F_1 seedlings [25]. PFRF responses remained consistent across multiple years, indicating a significant genetic effect on the phenotype and suggesting that the phenotype is genetically stable and can perhaps manipulated at the genetic level.

In this work, PFRF data and pedicel-fruit abscission zone transcriptome data from the three above-mentioned sweet cherry cultivars were used to decipher molecular mechanisms of ethylene inducible pedicel-fruit abscission zone formation. Gene expression patterns of ethylene-responsive genetic elements in abscission zone inducible genotype ‘Bing’ were identified and validated using qRT-PCR.

Results and Discussion

Phenotyping of pedicel-fruit abscission zone (PFAZ)

Pedicel-fruit retention force has been used as an indirect measurement of the viability of the PFAZ (pedicel-fruit abscission zone) [24, 25]. As average PFRF values decrease below 400 g the fruit are predicted to be able to be mechanically harvested with minimal loss. Although force is typically measured in Newtons [N = kg m/s^2], in our experiment N are not used as the defining measurement of force because distance and time are not included in the digital measuring device. Although a measurement of mass, grams are appropriate for indirectly measuring sweet cherry PFAZ. Quantitative observations of a basically binary phenotype such as PFAZ (at harvest the AZ is present or not present) allow the researcher to estimate the state of the AZ and correlate specific gene expression in the mechanically excised AZ with the physiological state of genotype dependent AZ development. However, because the individual biological samples were pooled, in this experiment it was not possible to determine a direct correlation between individual gene expression over the developmental time course and the specific AZ state of an individual sweet cherry fruit. The pooling of the samples, however, beyond providing the physical experimental advantage of adequate biological material, provided a baseline of gene expression for each genotype over an optimal sample size.

Phenotypic data (diameter, color, and PFRF) were measured for each genotype at each time point during the 2010, 2013 and 2014 sweet cherry seasons (Supplemental File 1a-c). A summary of PFRF data (calculated mean, standard deviation and 95% confidence interval) are presented in Table 1. For each of the three seasonal replications of the experiment, there was no statistically significant correlation between row and color, row and PFRF, or color and PFRF in any intra-genotypic comparison of treatment and control. This observation is justified biologically because at each time point the fruits were at very similar developmental stages and randomly sampled from the trees. Any discrepancies or fluctuations in color and row do not correlate with PFRF values, possibly because other environmental factors such as water availability and temperature fluctuations are more important for phenotypic variation at the time of harvest.

For 2010 and 2013 sampling, ethephon was applied 14 days before the predicted harvest date for each of the genotypes. The reason for this approach was because previous data suggested ethephon applications for ‘Bing’ should be made 14 days before harvest at 1.2 L ha^{-1} (1 pt A^{-1}) concentration. This regime was shown to efficiently remove fruit and minimize potential deleterious effects on fruit quality [24]. However, a need was felt to normalize the treatment and sampling time to represent equivalent developmental stage in each genotype because each genotype has a different maturation timeline for the fruit after bloom and response to exogenous ethylene may vary at the genetic level across the three genotypes used in this study. A representative developmental timeline was established for each genotype based on flowering and harvesting dates in the literature. Ethephon was applied at 80% completion of that timeline for each of the genotypes. This percentage coincided with 14 d before harvest for ‘Bing’, 12 d before harvest for ‘Chelan’ and 16 d before harvest for ‘Skeena’.

The modification to ethephon application period was implemented for the 2014 season experiment. Although ‘Chelan’ showed statistically significant PFRF response to ethephon application at the time of harvest for the 2010 and 2013 seasons, Table 2, average PFRF was not reduced to the threshold required for efficient mechanical harvest, Figure 1a. These physiological data support anecdotal observations that ‘Chelan’ forms neither a developmental nor ethylene-induced PFAZ. Gene expression data comparisons with ‘Chelan’ to ‘Bing’ and ‘Skeena’ may indicate a genotype-specific difference in key ethylene-responsive regulatory genes.

With no alterations in the timing for ethephon application for the 2014 season, ‘Bing’ predictably showed no significant PFAZ variation in ethylene treated samples PFRF following the schedule modification, Figure 1b. The higher average PFRF observed in the 2010 treatment data may be the result of weather discrepancies or errors in predicting the optimal harvest window for treatment and data collection.

With the change in ethephon application in 2014 sample collection, the PFRF data for ‘Skeena’ shifted significantly in that the expected phenotype of both Treatment and Control PFRF values were below the 400 g threshold recommended for mechanical harvest, Figure 1c. This shift could be due to a more congruent alignment of developmental ethylene responses among genotypes and a positive result of the 80% application, or it could be simply from other uncontrolled variables including serendipitously optimal weather conditions around the harvest time point.

Developmental time course transcriptome analysis

General steps in the construction of *P. avium* abscission zone transcriptomes are outlined in a workflow illustrated in Figure 2. Briefly, raw reads were processed through a quality check protocol and all low quality and contaminating reads were removed from each sample. A master transcriptome dataset (*PaRef*) was generated from the combined read data from ‘Bing’, ‘Chelan’, and ‘Skeena’ genotypes at each time point (Table 3). To evaluate pedicel-fruit abscission in sweet cherry, eight transcriptome libraries for each of the three sweet cherry genotypes (control and ethephon-treated samples for each of the four time points) were generated by mapping individual sample reads back to the *PaRef* assembly, Supplemental File 3a-x. Differential expression was then calculated between control and ethephon treated samples at the four distinct time points using a RNAseq module in CLC Genomics Workbench as described in the corresponding methods section.

Transcriptome Assembly and Annotation

Assembled contigs passed the filter criteria of greater than 200 base length combined with five or greater average read coverage. Read quality trimming reduced the total number of assembled contigs and increased individual base resolution and average read length. In summary, 84,260 contigs were assembled from 1,075,449,571 total trimmed reads with an N50 of 1,102 bases for the *PaRef* assembly (Table 3). This represents transcripts present in four developmental time points (0 h, 6 h, 7 d and 14 d post-treatment) and two experimental conditions (ethephon treatment and water control). The overall contig number found in this experiment may diverge from other reported plant transcriptomes due to greater base calling resolution between SNP alleles, orthologous and/or paralogous gene copies in the three distinct genotypes under consideration [26]. The *de novo* approach to transcriptome assembly is effective, but stringent read trimming and filtering parameters may lead to RNA copy loss in extracted RNA pools from minute but widespread errors attributed to enzymatic and mechanical processes used in the production of the final *in silico* read sequence. The errors introduced through amplification processes lower the total number of authentic RNA read occurrences thereby reducing statistical power provided by the high number and high quality reads produced by Illumina sequencing technology. Mapping each individual Illumina read dataset (from each genotype, each time point, and each treatment) to the assembled reference transcriptome sequence was performed to correct for bias introduced by stringent trimming and assembly regime. Mapping the reads back to the assembled reference with less stringent parameters better approximates the number of transcripts represented in the overall transcriptome. Sequencing errors introduced through this workflow will include reads that pass the homology tests, but any polymorphisms from this will not be incorporated into the reference assembly. These individually mapped files are used to generate normalized read expression values for each of the assembled contigs. The reference is used for mapping (here the reference is the *de novo* transcriptome of all three genotypes) to correlate contigs by name and easily compare expression values across genotypes for given time points.

The assembled sweet cherry unique contigs were first annotated through homologous search against different protein databases. A total of 38,874 (46.6%), unique contigs had significant hits (E-value 1e-5) in the GenBank non-redundant (nr) database. Consistent with previous reports, the percentage of annotated genes was positively correlated with gene length (Figure 3). Sweet cherry unique contigs were further annotated by assigning them with gene ontology (GO) terms. A total of 28,002 unique contigs (72.0%) were assigned with at least one GO term, among which 18,883 (67.4%), 22,684 (81.0%), 19,368 (69.2%) were assigned in the biological process, molecular function, and cellular component category, respectively. The unique contigs were further classified into different functional categories. The top twenty-five groups in the biological process and molecular function categories are shown in Figure 4. Annotated RNAseq datasets with five-fold differentially expressed genes from time point two (six hours following ethephon application) were analyzed for GO term enrichment using Fisher's Exact Test, Blast2GO [27]. Interestingly, 'Bing' and 'Skeena' had a similar profile of GO terms being enriched in the dataset, but 'Chelan' had almost no GO terms enriched, indicating the possibility that similar gene profiles were being activated or repressed in the former genotypes, but 'Chelan' has a unique gene expression response to ethephon application.

Transcriptomics comparison, differentially expressed genes (DEGs)

RNAseq projects have several challenges in analysis to overcome, specifically in the detection of differential expression. The first challenge comes from bias inherent in the sequencing technology itself; the second is laboratory or experimental errors (common to many new technologies) inducing technical variation across samples. The third and most important challenge is that the current costs associated with RNAseq prohibit generation of optimal biological replicates, retarding consistent statistical analysis. Experiments with small sample size and biological replication often lead to high false discovery rates (FDR) for differential gene expression [28]. The expression quantification of short reads using RNAseq data depends on the length of the feature; longer features tend to have an increased number of reads associated with them. The expression values must be normalized to mitigate bias associated with unbalanced sampling (preference for longer reads). This normalization can be performed in a number of different ways including as described by Mortazavi et al, resulting in the expression value known as Reads Per Kilobase per Million reads (RPKM) [29]. This normalization factor was used to eliminate read length bias and more closely approximate relative gene expression. A closely related normalization factor, Transcripts Per Million (TPM) has been shown to reduce transcript variability among samples within a given experiment [30]. Overall, TPM correlates with RPKM for each of the candidate genes evaluated, and the amplitude between time points, while variable, show clear differences between samples and time points, specifically for ethylene related (Supplemental File 5c).

Due to inherent limitations of the *de novo* transcriptome assembly, gene expression analysis was restricted to the contig consensus sequence annotation. It is difficult to differentiate between variant alleles or gene family members of highly similar sequence, without the use of subsequent molecular techniques [31]. However, each sample (genotype, treatment, time point) was mapped back to the ‘Stella’ reference genome (*de novo* assembly) (Supplemental Files 4a-x).

Mapping the ‘Skeena’ reads from time point three, or seven days after ethephon application, to *PaRef* resulted in unresolvable low quality contigs generated. This was most likely the result of initial library construction errors prior to Illumina sequencing because the total read count that passed initial quality assessment for ‘Skeena’ Control sample from the time point in question had a dramatically reduced total read count which leads to highly skewed normalized values (Figure 6).

Global differential gene expression

A transcriptome-wide picture of the gene expression differences between each genotype at each time point, the average RPKM value for each time point was calculated and the ratio of ethephon treatment to control shows a significant response in the ethylene responsive genotype, ‘Bing’ (Figure 4). The fact that the mean transcript expression levels of ‘Bing’ alone increases directly following ethylene application arises from one of two general options. One option is a small but broad increase in gene expression generally across the transcriptome or, alternatively, a few key genes dramatically shifting transcript abundance in ‘Bing’. An overall change in transcript ratio in ‘Bing’ supports the observed phenotypic changes in response to ethylene. Additionally, this observation does not rule out the possibility that ‘Chelan’ and ‘Skeena’ genotypes may have a more directed or specific ethylene response rather than a generalized one in the PFAZ.

In order to detect differentially expressed genes (DEGs), CLC Genomics Workbench (v8.5) was used to statistically evaluate control and treated samples. Differential expression testing of RNAseq data generally requires multiple replicates per sample, allowing for calculations of means and variance. Statistically significant differences in gene expression between samples were identified through CLC Genomics RNAseq Analysis workflow.

Transcription factors are important upstream regulatory proteins and play critical roles in various plant developmental processes and plant responses to abiotic and biotic stresses. In the present study, from the sweet cherry unique contigs, we identified a total of 1,296 transcription factors that were classified into 66 different families. The largest group of transcription factors was the AP2/ERF-ERF family (88, 6.8%) followed by C2H2 (82, 6.3%), NAC (81, 6.3%), bHLH (78, 6.0%), MYP-related (74, 5.7%), MYB (70, 5.4%), C3H (61, 4.7%), and WRKY (58, 4.5%) families. These eight families represented approximately half (46%) of the transcription factors identified in the unique sweet cherry transcripts. Specifically, ethylene responsive transcription factors (ERTF) were shown to be highly expressed in 'Bing' samples in response to ethylene treatment six hours following application. For the same time point, both 'Chelan' and 'Skeena' ERTF genes were not significantly differentially expressed. This result implies that the transcription factor family is altered in its expression by some unknown mechanism in 'Bing' which is missing in 'Chelan' and 'Skeena'. The sequences representing each of these genes *in silico* exhibited no obvious polymorphisms between the three sweet cherry genotypes. ERTF-1, showing highest differential expression, was PCR amplified from genomic DNA for each genotype, cloned, and sequenced. Comparative sequence analysis consistently showed that the genetic sequence was identical between species. This does not obviate the possibility that the upstream regulatory regions of this gene may be polymorphic across the three genotypes. It is possible that undetected regulating microRNAs [32] or epigenetic differences contribute to the observed expression response.

Additionally, well-studied gene families in the ethylene biosynthesis pathway such as 1-aminocyclopropane-1-carboxylate synthase (ACS) and 1-aminocyclopropane-1-carboxylate oxidase (ACO) showed no statistically significant differential expression in response to ethylene application in any of the genotypes examined (Table 7). Although previous work has shown non-sense SNPs within several members of each of these gene families [23], it did not translate in our work as a differential gene expression response. This could be for multiple reasons, including a failure to capture gene expression at the optimal developmental period for ethylene response or perhaps a related, but undetected gene family member is responsible for the response.

Validation of RNAseq derived differential gene expression via qRT-PCR

Differentially expressed transcripts between ethylene treated and control conditions were confirmed by quantitative real-time PCR (qRT-PCR). qRT-PCR is a valuable tool to validate digital differential expression studies supported by sequencing-based transcriptome profiling approaches because, while impractical for high-throughput screening, it remains the gold standard for individual gene expression studies. qRT-PCR validation was performed with samples collected in 2014. As described above, the timing of ethylene application in the experimental design was modified in 2014 to approximate the 'Bing' developmental time course response. However, the genotypes, number of time points, sampling methods and the concentration and volume of the experimental application remained consistent with experiments from earlier seasons. In this study, twenty-one relevant candidate genes were selected for subsequent analysis with qRT-PCR (Supplemental File 4). The selection of these genes was determined by a literature-based understood connection with ethylene combined with the results of the RNAseq differential expression studies. qRT-PCR fold-change analysis shows close correlation with gene expression trends of RNAseq transcriptome analysis (Figure 6).

An ethylene receptor gene increased twofold six hours after ethylene application in both 'Bing' and 'Chelan' and did not significantly change in 'Skeena'. However, the same gene had the twofold expression ratio at harvest for 'Bing' but was no longer detected in 'Chelan' (Table

7). ‘Skeena’ showed no change in expression of the ethylene receptor throughout the time course.

Polygalacturonase, which has a major role in cell wall degradation and fruit softening, was not detectable at the time of ethylene treatment (14 d prior to Harvest) but in ‘Bing’ and ‘Chelan’ it increases dramatically with the application of ethylene: twenty-fold in ‘Bing’ and greater than 200-fold in ‘Chelan’. This result means that exogenous ethylene changes the transcript abundance in ‘Chelan’ even if it does not correlate with the formation of an abscission zone. This implies some separation in the pathways that lead to the two distinct, but often linked processes of abscission and overall fruit ripening.

As described above, the gene families most often associated with ethylene biosynthesis, ACS and ACO, did not appear to be significantly altered by the application of ethylene at each of the three time points measured (Table 7). This indicates that there may be an alternative mechanism responding to the exogenous ethylene and altering physiology at the pedicel-fruit abscission zone.

Conclusions

In the current study, sweet cherry PFAZ transcript induction response to ethephon application in a genotype-specific manner is evidenced. However, variable expression patterns observed in this study demonstrate the regulatory and physiological complexity underlying sweet cherry PFAZ development and suggest that greater mechanistic understanding of sweet cherry PFAZ has yet to be achieved. While several of the findings support prior work in sweet cherry and other non-climacteric fruits, others are novel and present new regulatory candidates for investigation of putative roles in PFAZ development.

In this study, a potential molecular mechanism for pedicel fruit abscission zone development was uncovered. Up-regulation of ethylene response transcription factors may be responsible for stimulating abscission zone formation. ERTF expression was greater in ‘Bing’ isolated abscission zones six hours post ethylene application compared to Chelan and Skeena of the same transcript. ERTF responds to the exogenous ethylene by increasing transcript abundance. ERTF and associated genes may be molecular keys to initiate the cascade of signals and enzymatic reactions that eventually lead to the physiological development of an abscission zone. ‘Skeena’ forms the abscission zone without the requirement of exogenous ethylene; either ERTF is not required in the development of the ‘Skeena’ PFAZ, or it is regulated and expressed in an alternative timeline than can be observed in the experimental design of this project, i.e. expressed between 0-6 hours post ethephon application or after the six-hour time point, but prior to seven days post ethephon. ‘Chelan’ may have a similar story, albeit with a different outcome in that it does not result in the PFAZ development. Identification of PFAZ regulating genes leads to development of a molecular approach to assay and predict the abscission zone development of any sweet cherry genotype. Moreover, our transcriptome data provides a useful resource for gene mining of ethylene responsive and abscission related processes in sweet cherry.

Transcriptome data do not take post-transcriptional regulation and modifications into account. However, there is evidence for post-transcriptional regulation playing an important role in tomato, Arabidopsis and citrus abscission [33-35]. This is an area of acute interest for future work in cherry because common, yet diverse modifications have the potential to reveal much about the biological regulation of the abscission zone development.

Methods

Plant material – ethephon treatment, phenotypic measurements and sample collection

All sweet cherry trees were located at Washington State University's Roza Farm, about 10 km north of Prosser, Washington, USA (46.2°N, 119.7°). All trees were irrigated weekly from bloom to leaf senescence with low-volume under-tree microsprinklers and grown using standard orchard management practices. Each trial was arranged with four single-tree replications per treatment. The trees had an in-row spacing of 2.44 m (8 ft) and between row spacing of 4.27 m (14 ft). Rows were planted in a north-south orientation and trained to a Y-trellis architecture.

Abscission zones and PFRF data were collected during three separate seasons (2010, 2013, and 2014). Each replication was performed in the same orchard block, but in distinct trees within the cherry block depending on availability with other projects.

Two treatments were used (Ethephon and H_2O) for each replication. The treatments were applied via hand spray directly on foliage and developing fruit directly following data and sample collection to be used as control. This was done early in the morning (between 0600 and 0800 hours) to reduce the effects of ethylene evolution from warm temperatures and wind as previously described [24]. Ethephon, 240 ppm, was dissolved in tap water immediately prior to being sprayed on the foliage and fruit of four trees (designated Treatment trees) and control (H_2O) sprayed in the same manner on four different trees (designated Control trees for the duration of the experiment).

Data were collected for four time points, the same trees at each time point. (1) 12 days before expected harvest for 'Chelan', 14 days before expected harvest date for 'Bing', and 16 days before expected harvest for 'Skeena'; (2) 6 hours after the application of ethephon and H_2O; (3): 7 days before expected harvest date, and (4) on the expected harvest date. Samples were collected at approximately 1200 hrs (time points 2, 3, 4) to avoid diurnal effect variability on gene expression data. Additionally, the data were collected as quickly as possible to minimize environmental and genetic effects distinct from treatment differences. Temperature and local weather condition also has a significant effect on PFRF. Generally, cool, wet weather tends to increase the average pedicel-fruit retention force for each cherry genotype regardless of developmental stage.

Ten cherry fruits were randomly selected from each of four trees per genotype per treatment for a total of 40 samples. Size (row) and color data were collected using commonly available methods such as a color palate chart and a pass-through row calculation device. Pedicel-fruit retention force (PFRF) were measured using a modified digital force gauge (Imaga). This was accomplished by holding the stem of each fruit to be measured and placing the fruit in the modified adaptor of the digital force gauge. The stem was then pulled by hand with a quick, steady tug to remove it from the fruit. The digital force gauge displays the highest force achieved for each event. A single user measured PFRF for each time point to minimize user-derived variability in the data.

The average of the color and row measurements were used to gather ten randomly spatially distributed fruits that were similar to the same color and row as the average of previously collected fruits. These ten fruits were then sliced with six quick motions using a standard single-edge razor blade. The first cut was below the stem approximately 0.5 cm. This cut was between the stem and the internal stone. This left the pedicel and a thin disc of fruit/skin attached. Next, two sets of parallel cuts were made downward on the cherry fruit disc on either side of the stem, effectively making a cube piece of fruit 3mm x 3mm x 3mm attached to the pedicel. Finally, the pedicel was cut off directly above the fruit and the cube of fruit tissue consisting of the abscission zone and some pedicel tissue was placed in a 15ml falcon tube and kept frozen throughout processing.

Total RNA extraction

Sweet cherry abscission zone tissue was ground via SPEX SamplePrep® FreezerMill 6870 (Metuchen, NJ USA). It was kept frozen in liquid nitrogen throughout processing. The freezer mill pulverized and homogenized the excised tissue derived from forty fruits from four trees for each time point into a single sample. The pulverized samples were stored at -80°C.

Total RNA was extracted using an acid guanidinium thiocyanate phenol chloroform extraction method similar to that described by Chomczynski (1987). 1mL of 0.8M guanidinium thiocyanate, 0.4M ammonium thiocyanate, 0.1M sodium acetate pH 5.0, 5% w/v glycerol, and 38% v/v water saturated phenol were added to approximately 100 mg powdered tissue, shaken to evenly mix sample and incubated at room temperature (RT) for 5 minutes. 200µL chloroform was added and shaken vigorously until the entire sample became homogenously cloudy and then was incubated at RT, 3 minutes. Samples were then centrifuged at 17,000 x g at 4°C for 15 minutes and the aqueous upper phase was transferred to a clean 1.5mL microcentrifuge tube. To this, 600µl 2-propanol was added, inverted 5-6 times and incubated at RT for 10 minutes. Samples were centrifuged 17,000 x g at 4°C for 10 minutes and the supernatant decanted. 1 mL 75% DEPC ethanol was added to the pellet, vortexed for 10 seconds and centrifuged 9,500 x g at 4°C for 5 minutes. Pellets were then suspended in RNase free water and incubated at 37°C with RNase free *DNaseI* for 30 minutes and *DNaseI* inactivated at 65°C for 10 minutes.

RNA quality check

Extracted RNA was quality checked with the Bio-Rad (Hercules, CA) Experion system using the Experion RNA High Sensitivity Analysis kit or the Agilent (Santa Clara, CA) 2100 Bioanalyzer system using the RNA NanoChip and Plant RNA Nano Assay Class.

Illumina Sequencing

RNA samples that passed the quality threshold of RIN 8.0 were sent to Michigan State University for library preparation and Illumina sequencing. At the end of each sequencing cycle, there was a single-base extension. The cycle was then repeated 50 to 100 times, making the read-lengths 50 to 100 bp. It is possible to get longer reads, but it is more likely to get higher error rates due to substitution errors [36].

cDNA library preparation and transcriptome sequencing

The Illumina Hi Seq 2000 sequencing platform (San Diego, CA.) was used to sequence 2x100 PE reads from the cDNA libraries generated from the above RNA extractions at Michigan State University's Research Technology Support Facility. cDNA and final sequencing library molecules were generated with Illumina's TruSeq RNA Sample Preparation v2 kit (San Diego, CA.) and instructions with minor modifications. Modifications to the published protocol include a decrease in the mRNA fragmentation incubation time from 8 minutes to 30 seconds to create the final library proper molecule size range. Additionally, Aline Biosciences' (Woburn, MA) DNA SizeSelector-I bead-based size selection system was utilized to target final library molecules for a mean size of 450 base pairs. All libraries were then quantified on a Life Technologies (Carlsbad, CA) Qubit Fluorometer and qualified on an Agilent (Santa Clara, CA) 2100 Bioanalyzer (Dr. Jeff Landgraf, personal communication).

RNAseq read datasets were processed with the CLC Create Sequencing QC Report tool to assess read quality. The CLC Trim Sequence process was used to trim the first thirteen bases due to GC ratio variability and for a Phred score of 30. All read datasets were trimmed of ambiguous bases. Illumina reads were then processed through the CLC Merge Overlapping Pairs tool and all reads were *de novo* assembled to produce contiguous sequences (contigs). Trimmed reads used for the assembly were mapped back to the assembled contigs, mapped reads were used to update the contigs, and contigs with no mapped reads were ignored. Consensus contig sequences were extracted as a multi-fasta file. The individual genotype specific read datasets, original non-trimmed reads, were mapped back to the assembled contigs to generate individual time course and water or ethephon treated sample reads per contig and then normalized for sequencing depth and gene length with Reads Per Kilobase per Million reads (RPKM) method [29]. Additionally, the reads were normalized using Transcripts Per kilobase Million (TPM) method.

Differential expression

Differential expression between genes was identified using the CLC Genomics RNA-Seq Analysis tool. RPKM and TPM data were both tested using Kal et al.'s test [37] which compares samples (time points for the current study) and considers proportions rather than raw read counts. This provides the user with expression fold-change data between sample combinations of interest as well as a two-tailed p-value for the statistical test. Additionally, the test includes an FDR correction for the p-value which is important for sorting results efficiently. By filtering the p-values and fold-change in the dynamic spreadsheet one can make the experiment more stringent and obtain more biologically relevant genes. In this study, FDR corrected p-value of ($p < 0.001$) and fold-change of greater than five was used.

Quantitative real-time PCR Overview

Quantitative real-time PCR (qRT-PCR) was used to validate the RNAseq derived expression patterns of selected RNA transcripts in different sweet cherry varieties. Genomic DNA contamination was removed by treatment with TURBO DNA-free (DNAse I) according to manufacturer's methods (Life technologies, Carlsbad, CA USA). RNA quality was verified using a denaturing gel and BioAnalyzer 2100 (Agilent, CA USA). For each sample, 500 ng of total RNA were used to generate 1st strand cDNA using the Invitrogen VILO kit (Life Technologies, Carlsbad, CA USA). cDNA preparations were then diluted to uniform concentration of 50 ng/μl. Initial qRT-PCR technical replicate reactions were prepared for each of the 25 genes using the iTaq Universal SYBR Green Supermix (BioRad, Hercules, CA). Reactions were prepared according to manufacturer's protocols with 100ng template cDNA and optimized thermal cycle conditions (Supplemental file 2).

RNA quality and integrity

RNA quality was initially determined by a visual inspection of the ribosomal subunits in a 1% agarose gel. This quick check eliminated poor quality RNA from passing on to cDNA library preparation.

Reverse-transcription

VILO cDNA synthesis kits were used to generate three technical replicates of cDNA for each RNA isolation. The cDNA from the three technical replicates were pooled into a single sample (50 ng/ul) which was used to perform qRT-PCR.

qRT-PCR primer design

To confirm the differential expression of candidate transcripts identified from the RNAseq method, the candidate genes were first screened based on potential functions. Because global gene expression changes typically result from expression changes in transcription factors, transcripts with homology to known transcription factors were selected. Primers were designed based on the near full length transcript sequences to amplify an approximately 100-150 bp region in the 3' region of the transcript. 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.

Three independent extractions of RNA from each sample of the ground tissue using the general CTAB RNA extraction protocol were performed. First, ground tissue was suspended in 600 µl RNA extraction buffer (CTAB, 2% final concentration; NaCl, 1.4M; 0.5 M EDTA, pH 8.0 20 mM; 1M TRIS, pH 8.0 100 mM; Polyvinylpyrolidone (PVP40), 2% final concentration; Water to final volume) including 2-mercaptoethanol (1%) and vigorously vortexed. Thereafter, 600 µl chloroform was added and samples mixed thoroughly. Each sample was centrifuged at 14,000rpm for 2 minutes and the supernatant transferred to a new tube. A second chloroform extraction (600µl) was performed and supernatant was transferred to a new tube. Then an equal volume of isopropanol was added and the samples were thoroughly mixed, and centrifuged at 14,000rpm for 15 mins. The supernatant was carefully decanted from the tube and 600 µl 70% ethanol was added to the pellet. The tube was flicked vigorously to wash the pellet and then it was centrifuged for 2 min at 14,000rpm. The supernatant was again carefully decanted and the pellet allowed to air dry for 10 mins. The pellet was suspended in 90 µl DEPC-treated H_2O. The tubes were incubated at 65°C for 15 mins (dry heating dock) and then spun down for 2 mins to remove impurities. The supernatant was transferred to new tubes and then 30 µl LiCl was added to precipitate the RNA and the tubes were placed at -20°C for 30-60 mins. Following this, the tubes were centrifuged at 20,000rpm, 4°C, for 30 mins. The supernatant was decanted and 100 µl 70% ethanol added to the tube to wash the pellet The pellet was air dried and 35 µl DEPC water was added to the pellet and the tubes were incubated at 65°C for 15 mins. The tubes were then centrifuged at 14,000 rpm for 5 mins and the supernatant transferred to a clean tube.

RNA quality was checked by loading 1 µl RNA in 0.8% agarose gel and visualizing ribosomal RNA bands. A total of 30ul of the RNA was treated with DNase using the DNA-free™ kit (Ambion) according to the manufacturer's protocol. Following DNase treatment, 1 µl RNA was electrophoresed on a 0.8% agarose gel to verify quality (distinct rRNA bands). Following visual quality confirmation, 9 µl RNA and 5 µl Luciferase RNA (10 pgµl^{-1}) were included in first strand cDNA synthesis via the VILO kit (Invitrogen) according to manufacturer's protocol.

Reference genes

In this study, rbcL was found to be consistently expressed across all time points and genotypes (within 0.5 standard deviations) so it was used as a reference gene to compare the tested genes. In addition to this reference gene, the bacterial derived *luciferase* gene was used as a "spiked" reference, 50 ng/reaction (Figure 7).

qRT-PCR

Following positive control gene amplification, and individual sample quantification, the cDNA was diluted according to the VILO kit instructions. qRT-PCR reactions were performed using iTAQ with ROX and SYBR (BioRad) and 20µL reactions were prepared as per the recommendations outlined by BioRad. A total of 2µL of cDNA diluted to 50ng/µL RNA equivalents was used per reaction with 5µL H_2O, 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.

PCR efficiency

In order to capture PCR efficiency in the data, Cq values and efficiencies were calculated for each reaction using the LinRegPCR tool [38, 39]. Cq values resulting from efficiencies below 1.80 or 2.20 were judged unacceptable and were treated as unsuccessful or undetected amplifications. Cq values with efficiency values that were within expected parameters, but exceeded (or equaled) 40.00 were also deemed unacceptable and disregarded in downstream analysis. In the same manner, Cq values between (35.00-39.99) were determined to be of *low confidence* and were marked for special consideration in downstream analysis.

Fold-change expression was determined from Cq values of all gene targets (among all replicates of all samples) among the ‘Bing’, ‘Chelan’ and ‘Skeena’ genotypes using the Pfaffl method [40]. Expression values were determined in reference to rbcL and the luciferase “spiked” gene.

Competing interests

The authors declare no competing interests.

Authors' contributions

Conceived and designed the experiments: AD, TK and BK

Performed the experiments: TK, JA, BK

Analyzed the data: BK, AD

Contributed reagents/materials/analysis tools: AD

Wrote the paper: BK and AD

References

1. Giovannoni, J.J., *Genetic Regulation of Fruit Development and Ripening.* The Plant Cell, 2004. **16**: p. 170-181.
2. Sexton, R. and J.A. Roberts, *Cell Biology of Abscission.* Annual Review of Plant Physiology, 1982. **33**: p. 133-162.
3. Ito, Y. and T. Nakano, *Development and regulation of pedicel abscission in tomato.* Frontiers in Plant Science, 2015. **6**: p. 1-6.
4. Roberts, J.a., K.a. Elliott, and Z.H. Gonzalez-Carranza, *Abscission, dehiscence, and other cell separation processes.* Annual review of plant biology, 2002. **53**: p. 131-58.
5. Taylor, J.E., et al., *Polygalacturonase expression during leaf abscission of normal and transgenic tomato plants.* Planta, 1991. **183**: p. 133-138.
6. Taylor, J.E. and C.A. Whitelaw, *Signals in abscission.* New Phytologist, 2001. **151**: p. 323-339.
7. Alexander, L. and D. Grierson, *Ethylene biosynthesis and action in tomato: a model for climacteric fruit ripening.* Journal of experimental botany, 2002. **53**: p. 2039-2055.
8. Ju, C. and C. Chang, *Mechanistic Insights in Ethylene Perception and Signal Transduction.* Plant Physiology, 2015. **169**: p. pp.00845.2015.
9. Johnson, P.R. and J.R. Ecker, *The ethylene gas signal transduction pathway: a molecular perspective.* Annual review of genetics, 1998. **32**: p. 227-54.
10. Ampopho, B., et al., *The Molecular Biology and Biochemistry of Fruit Ripening.* 2013: p. 1-216.
11. Seymour, G.B., et al., *Fruit Development and Ripening.* Annual review of plant biology, 2013: p. 1-23.

12. Paul, V., R. Pandey, and G.C. Srivastava, *The fading distinctions between classical patterns of ripening in climacteric and non-climacteric fruit and the ubiquity of ethylene-An overview.* Journal of Food Science and Technology, 2012. **49**: p. 1-21.

13. Brown, K.M. and K.M. Brown, *Ethylene and abscission.* Physiologia Plantarum, 1997. **100**: p. 567-576.

14. Van Doorn, W.G., *Effect of ethylene on flower abscission: A survey.* Annals of Botany, 2002. **89**: p. 689-693.

15. McManus, M.T., *Further examination of abscission zone cells as ethylene target cells in higher plants.* Annals of botany, 2008. **101**: p. 285-92.

16. Botton, A., et al., *Signaling Pathways Mediating the Induction of Apple Fruitlet Abscission.* Plant Physiology, 2011. **155**: p. 185-208.

17. Eccher, G., et al., *Roles of Ethylene Production and Ethylene Receptor Expression in Regulating Apple Fruitlet Abscission.* Plant Physiol, 2015. **169**(1): p. 125-37.

18. Nunez-Elisea, R. and T.L. Davenport, *Abscission of Mango Fruitlets as Influenced by Enhanced Ethylene Biosynthesis.* Plant physiology, 1986. **82**: p. 991-994.

19. Ruperti, B., et al., *Ethylene biosynthesis in peach fruitlet abscission.* Plant, Cell and Environment, 1998. **21**: p. 731-737.

20. Wittenbach, V.A. and M.J. Bukovac, *Cherry Fruit Abscission, Evidence for Time of initiation and the involvement of Ethylene.* Plant Physiology, 1974. **54**: p. 494-498.

21. Chang, C. and A.B. Bleecker, *Ethylene biology. More than a gas.* Plant physiology, 2004. **136**: p. 2895-2899.

22. Eccher, T. and N. Noe, *Respiration of cherries during ripening.* Acta horticulturae, 1998. **464**.

23. Koepke, T., et al., *Comparative genomics analysis in Prunoideae to identify biologically relevant polymorphisms.* Plant Biotechnology Journal, 2013. **11**: p. 883-893.

24. Smith, E. and M. Whiting, *Effect of ethephon on sweet cherry pedicel-fruit retention force and quality is cultivar dependent.* Plant Growth Regulation, 2010. **60**: p. 213-223.

25. Zhao, Y., et al., *Pedicel-fruit retention force in sweet cherry (Prunus avium L.) varies with genotype and year.* Scientia Horticulturae, 2013. **150**: p. 135-141.

26. Bräutigam, A., et al., *Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C 3 and C 4 species.* Journal of Experimental Botany, 2011. **62**: p. 3093-3102.

27. Conesa, A., et al., *Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.* Bioinformatics (Oxford, England), 2005. **21**: p. 3674-6.

28. Lorenz, D.J., et al., *Statistical Analysis of Next Generation Sequencing Data.* 2014: p. 25-50.

29. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature Methods, 2008. **5**: p. 621-628.

30. Wagner, G.P., K. Kin, and V.J. Lynch, *Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.* Theory in biosciences = Theorie in den Biowissenschaften, 2012. **131**: p. 281-5.

31. O'Neil, S.T. and S.J. Emrich, *Assessing De Novo transcriptome assembly metrics for consistency and utility.* BMC genomics, 2013. **14**: p. 465.

32. He, L. and G.J. Hannon, *MicroRNAs: small RNAs with a big role in gene regulation.* Nature reviews. Genetics, 2004. **5**: p. 522-531.

33. Meir, S., et al., *Microarray Analysis of the Abscission-Related Transcriptome in the Tomato Flower Abscission Zone in Response to Auxin Depletion1[C][W][OA]*. Plant Physiology, 2010. **154**: p. 1929-1956.

34. Seymour, G.B., et al., *Regulation of ripening and opportunities for control in tomato and other fruits.* Plant biotechnology journal, 2012: p. 269-278.

35. Sundaresan, S., et al., *Abscission of flowers and floral organs is closely associated with alkalization of the cytosol in abscission zone cells.* Journal of Experimental Botany, 2015. **66**: p. 1355-1368.

36. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nature biotechnology, 2008. **26**: p. 1135-45.

37. Kal, a.J., et al., *Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources.* Molecular biology of the cell, 1999. **10**: p. 1859-1872.

38. Ramakers, C., et al., *Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data.* Neuroscience Letters, 2003. **339**: p. 62-66.

39. Ruijter, J.M., et al., *Amplification efficiency: Linking baseline and bias in the analysis of quantitative PCR data.* Nucleic Acids Research, 2009. **37**.

40. Pfaffl, M.W., *A new mathematical model for relative quantification in real-time RT-PCR.* Nucleic acids research, 2001. **29**: p. e45.

Tables

Table 1. Mean PFRF values. The mean Pedicel fruit retention force (PFRF) values and standard deviation for each genotype and treatment over three separate seasons (2010, 2013, 2014).

		2010				**2013**				**2014**			
	Time Point:	**0**	**0.25**	**7**	**14**	**0**	**0.25**	**7**	**14**	**0**	**0.25**	**7**	**14**
Chelan Treatment	Average PFRF (kg)	1.101	1.103	0.926	0.532	1.295	1.128	0.717	0.427	1.603	1.509	0.432	0.832
	StdDev	0.239	0.261	0.278	0.141	0.286	0.266	0.134	0.187	0.302	0.167	0.211	0.199
	95% Confidence Interval	0.074	0.081	0.086	0.044	0.089	0.082	0.041	0.058	0.094	0.052	0.065	0.062
Chelan Control	Average PFRF (kg)	1.048	1.089	1.036	0.652	1.335	1.282	0.845	0.469	1.658	1.518	0.580	1.037
	StdDev	0.257	0.272	0.265	0.164	0.347	0.318	0.202	0.144	0.404	0.260	0.199	0.213
	95% Confidence Interval	0.080	0.084	0.082	0.051	0.108	0.098	0.063	0.045	0.125	0.081	0.062	0.066
Bing Treatment	Average PFRF (kg)	1.148	1.056	0.696	0.480	1.595	1.467	0.519	0.254	1.606	0.748	0.462	0.215
	StdDev	0.221	0.253	0.176	0.129	0.346	0.285	0.290	0.105	0.316	0.379	0.251	0.153
	95% Confidence Interval	0.069	0.078	0.055	0.040	0.107	0.088	0.090	0.032	0.098	0.117	0.078	0.047
Bing Control	Average PFRF (kg)	1.156	1.064	0.856	0.669	1.564	1.458	0.602	0.521	1.575	0.799	0.619	0.418
	StdDev	0.246	0.258	0.148	0.172	0.391	0.267	0.234	0.211	0.307	0.379	0.253	0.194
	95% Confidence Interval	0.076	0.080	0.046	0.053	0.121	0.083	0.072	0.065	0.095	0.118	0.078	0.060
Skeena Treatment	Average PFRF (kg)	1.317	1.232	0.562	0.320	0.661	0.660	0.496	0.315	1.196	1.070	0.255	0.156
	StdDev	0.304	0.265	0.132	0.123	0.275	0.261	0.154	0.096	0.261	0.270	0.142	0.081
	95% Confidence Interval	0.094	0.082	0.041	0.038	0.085	0.081	0.048	0.030	0.081	0.084	0.044	0.025
Skeena Control	Average PFRF (kg)	1.319	1.263	0.996	0.680	0.701	0.634	0.823	0.652	1.184	0.962	0.333	0.290
	StdDev	0.332	0.294	0.293	0.126	0.342	0.280	0.239	0.204	0.334	0.301	0.139	0.130
	95% Confidence Interval	0.103	0.091	0.091	0.039	0.106	0.087	0.074	0.063	0.104	0.093	0.043	0.040

Table 2. T-test of PFRF for Chelan, Bing, and Skeena. T-test showing significant responses between ethylene treatment and control PFRF values for three genotypes and three seasonal replications at harvest time point (~14 d after treatment). ‘Bing’ and ‘Skeena’ showed statistically significant responses every year. While Chelan also showed a statistically significant response to the ethylene treatment for 2010 and 2014, in both instances mean PFRF was not reduced to mechanically harvestable values indicative of PFAZ formation. In 2013, Chelan did not significantly respond to the ethylene treatment.

	2010	2013	2014
Bing	3.62E-07	3.68E-10	1.56E-06
Chelan	2.09E-03	2.65E-01	2.85E-05
Skeena	5.87E-21	1.62E-14	1.14E-06

Table 3. Summary of Bing, Chelan, Skeena *de novo* transcriptome assembly. Illumina reads used in assembly generation. This assembly was used as the reference for mapping individual genotype, time point, and treatment read files to generate subsequent differential gene expression data.

Contig Measurements	Length (including scaffold regions)	Length (excluding scaffold regions)
N75	505	477
N50	1268	1102
N25	2401	2162
Minimum	131	114
Maximum	16838	15998
Average	754	708
Count	84260	89547
Total	63535148	63387120

Summary Statistics	Count	Average length	Total Bases
Reads	1075449571	83.27	89552327949
Matched	984093648	83.07	81747855601
Not matched	91355923	85.43	7804472348
Contigs	84260	754	63535148
Reads in pairs	685702776	253.1	NA
Broken paired reads	163462695	253.1	NA

Table 4a. Summary of Bing *de novo* transcriptome assembly. This assembly was performed using 100 bp paired end data on the Illumina sequencing platform.

Contig Measurements	**Length (including scaffold regions)**	**Length (excluding scaffold regions)**
N75	593	542
N50	1512	1326
N25	2542	2335
Minimum	144	115
Maximum	15305	15305
Average	835	777
Count	59760	64129
Total	49902045	49796872

Summary Statistics	**Count**	**Average length**	**Total Bases**
Reads	377364374	82.72	31217168329
Matched	353700448	82.64	29229305064
Not matched	23663926	84	1987863265
Contigs	59760	835	49902045
Reads in pairs	25661724	257.79	NA
Broken paired reads	52820437	75.68	NA

Table 4b. Summary of Chelan *de novo* assembly.

Contig Measurements	Length (including scaffold regions)	Length (excluding scaffold regions)
N75	577	535
N50	1471	1295
N25	2515	2308
Minimum	131	100
Maximum	16784	16784
Average	824	769
Count	63646	68070
Total	52424808	52316885

Summary Statistics	Count	Average length	Total Bases
Reads	390084738	84.22	32854433281
Matched	359208487	84.05	30192611090
Not matched	30876251	86.21	2661822191
Contigs	63646	823	52424808
Reads in pairs	262690278	257.2	NA
Broken paired reads	47527268	76.53	NA

Table 4c. Summary of Skeena *de novo* assembly.

Contig Measurements	Length (including scaffold regions)	Length (excluding scaffold regions)
N75	616	563
N50	1558	1681
N25	2580	2384
Minimum	111	24
Maximum	15862	15862
Average	853	796
Count	55638	59518
Total	47453854	47372520

Summary Statistics	Count	Average length	Total Bases
Reads	311083226	84.23	26203909702
Matched	287466860	84.06	24163685236
Not matched	23616366	86.39	2040224466
Contigs	55638	852	47453854
Reads in pairs	203556052	244.78	NA
Broken paired reads	38793291	75.98	NA

Table 5. Gene expression of ethylene responsive sequences. Data were collected in ‘Bing’, ‘Chelan’, and ‘Skeena’ before ethylene application, 6 hours after ethephon application and at harvest (100% development) for each of the three genotypes. Fold change values were calculated between treatment and control for each gene at a given time point.

Gene Name	**Bing**			**Chelan**			**Skeena**		
	0 h	**6h**	**Harvest**	**0 h**	**6h**	**Harvest**	**0 h**	**6h**	**Harvest**
Ethylene Responsive TF	-1.20	3.51	-1.39	1.13	3.10	1.27	-1.47	3.66	1.21
Cell Wall Invertase	1.06	2.13	-2.28	-2.22	-4.99	-6.92	-2.06	-4.23	-3.76
WRKY TF 5	-1.78	2.73	3.73	-1.14	1.20	1.27	-1.88	1.99	2.60
Class I chitinase	7.73	1.01	1.93	2.69	-2.73	-4.23	-1.18	-1.80	-2.30
unknown	14.93	6.73	8.51	-1.47	2.50	6.92	1.58	1.01	-1.01
Patatin group a-3-like	n.d.	-4.08	13.09	13.27	-9.32	-8.94	-3.94	n.d.	n.d.
Spermine synthase	n.d.	n.d.	3.53	-8.34	-4.72	15.45	-3.23	-2.43	-1.54
unknown	1.18	-2.01	3.20	1.03	-1.01	-1.79	-1.62	-2.58	n.d.
Ethylene receptor	n.d.	2.30	2.66	1.28	2.38	n.d.	-1.15	-1.45	-1.06
Methylthioribose kinase	1.49	10.42	n.d.	n.d.	n.d.	-1.79	-1.06	1.46	n.d
Expansin	1.55	1.75	-2.97	1.46	-9.00	-2.11	1.01	1.20	-1.52
Polygalacturonase	n.d.	1.00	20.97	n.d.	-4.44	221.32	-2.28	-2.14	-1.30
Expansin (a1)	n.d.	4.89	1.13	-1.23	2.22	1.88	1.05	1.26	-1.26
Ap2 erf TF	1.27	-4.79	-1.51	-1.15	-2.30	-2.62	-1.87	1.06	8.11
Kda class IV HSP	1.55	n.d.	1.13	1.01	-1.40	-1.39	1.05	1.26	-3.32
Polyneuridine-aldehyde esterase	2.53	5.17	2.17	1.12	-1.16	16.11	-1.95	3.71	2.77
Ap2 erf TF-2	1.72	-1.02	-1.67	-2.04	-2.60	-2.36	-1.68	1.89	3.94
Endochitinase pr4	1.41	1.29	-1.61	-1.52	-2.77	1.01	-2.33	1.04	2.79
SAM-dependent methyltransferase	1.71	3.48	-2.91	-3.27	-2.68	-1.39	-4.53	-1.13	n.d.
ACS	1.12	1.48	1.01	-1.55	-1.93	-2.83	-2.69	1.03	-1.26
ACO	1.09	-1.56	-3.53	1.78	-3.16	-1.67	-3.18	1.33	-2.06
n.d. – not detected									

Figures

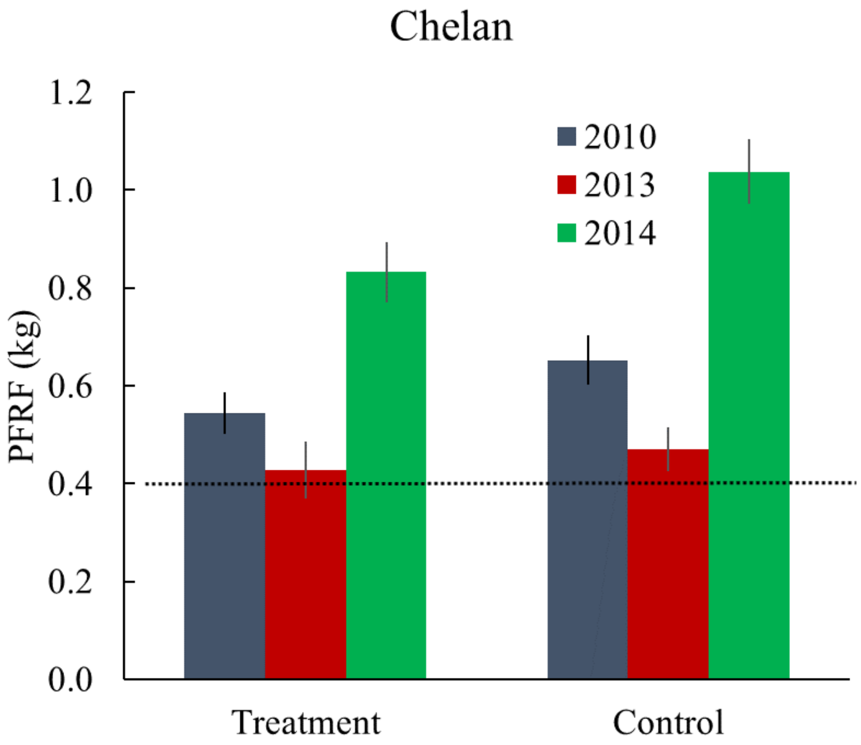


Figure 1a. Pedicel-fruit retention force (PFRF) at harvest for Chelan. This shows all three seasons (2010, 2013, 2014) that data was collected. Error bars represent 95% CI values (Supplemental File 1). The dotted line indicates the threshold PFRF value required for mechanical harvest.

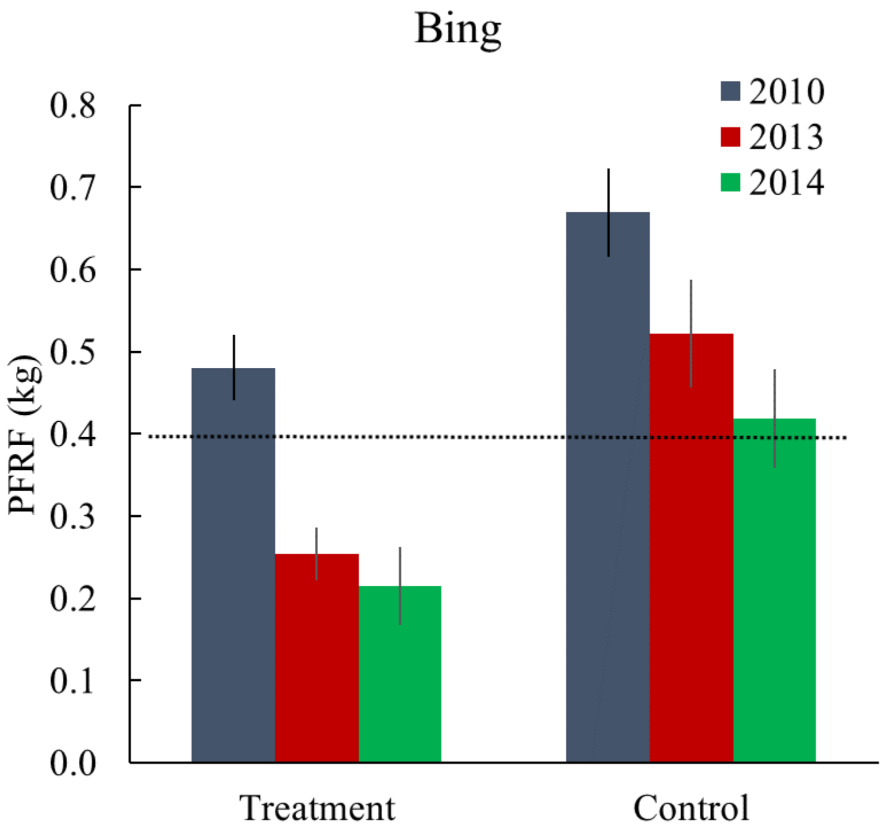


Figure 1b. Pedicel-fruit retention force (PFRF) at harvest for Bing. This shows all three seasons (2010, 2013, 2014) that data was collected. Error bars represent 95% CI values, see Supplemental File 1. The dotted line represents the threshold PFRF for mechanical harvest.

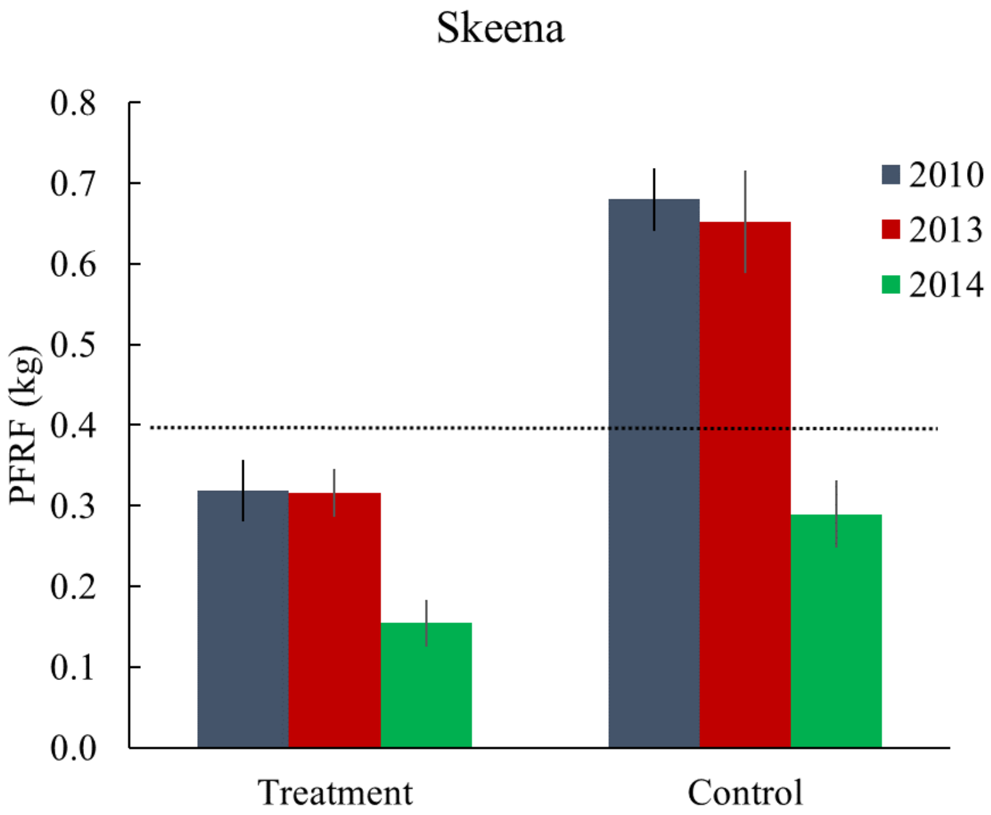


Figure 1c. Pedicel-fruit retention force (PFRF) at harvest for **Skeena**. This shows all three seasons (2010, 2013, 2014) that data was collected. Error bars represent 95% CI values, see Supplemental File 1. The dotted line represents the threshold PFRF required for mechanical harvest.

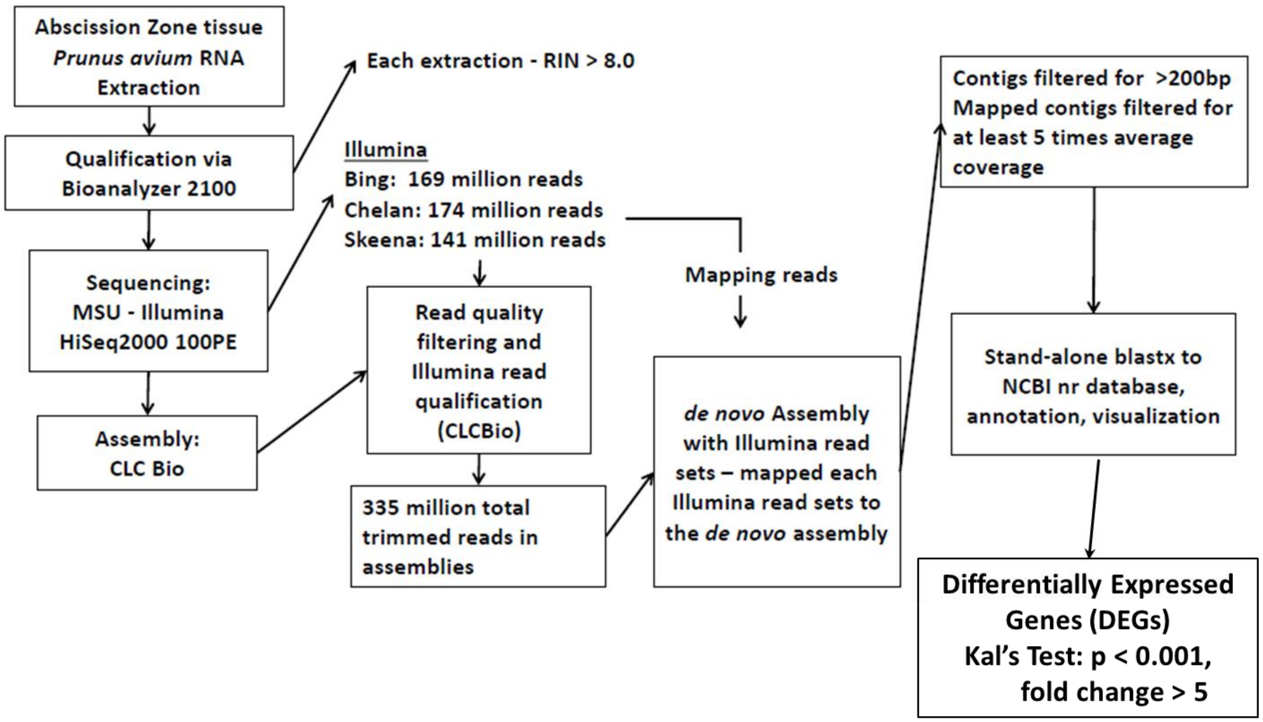


Figure 2. *de novo* Transcriptome development workflow. Total RNA from four developmental time points in three *P. avium* genotypes was sequenced on the HiSeq2000 platform. Read sequences were assembled with CLC Bio's Genomic Workbench and annotated with the NCBI nr database version 2.2.29+. Differentially expressed genes were analyzed regarding biological significance.

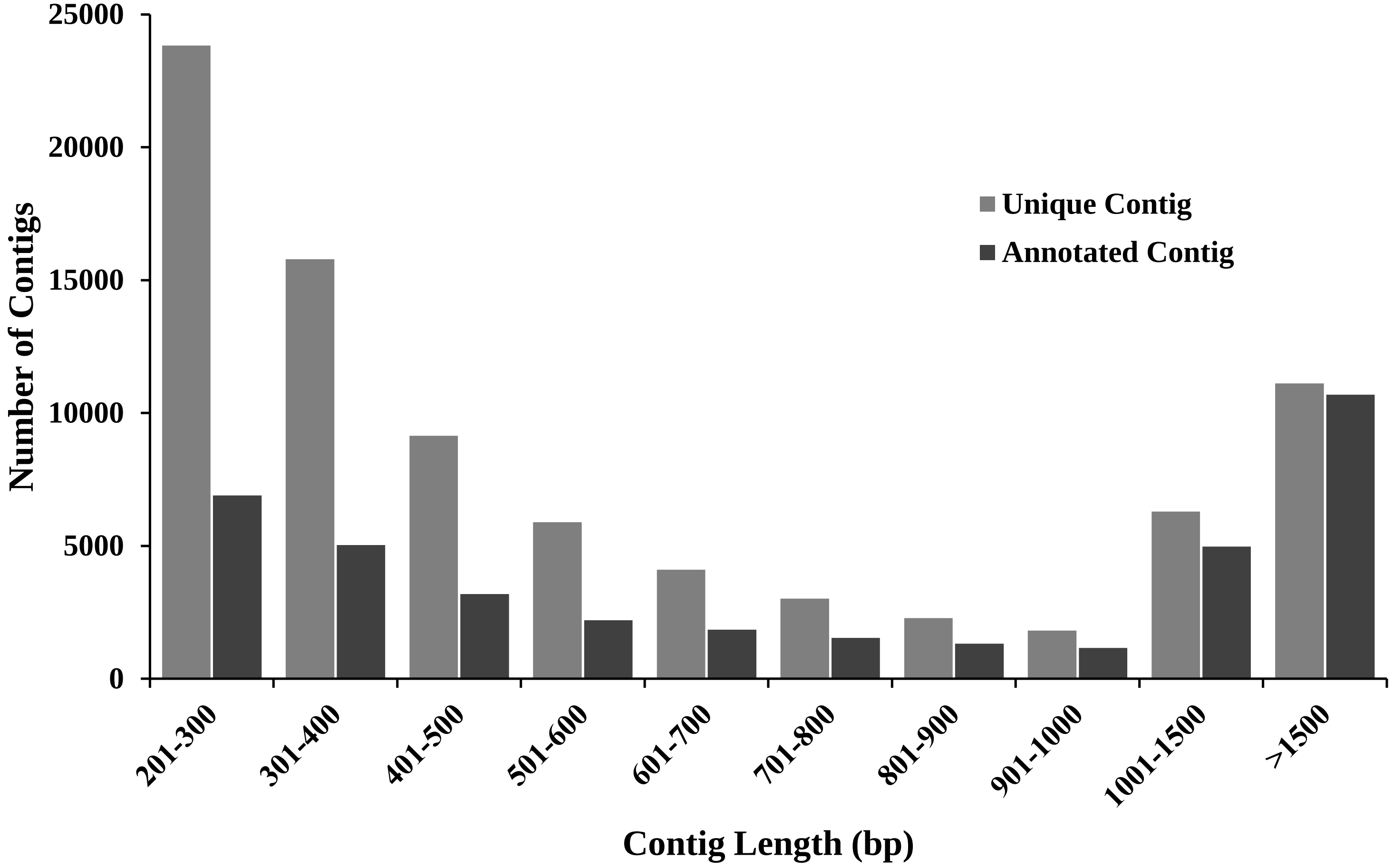


Figure 3. Length distribution of sweet cherry unique contigs. The percentage of annotated contigs increased with contig length.

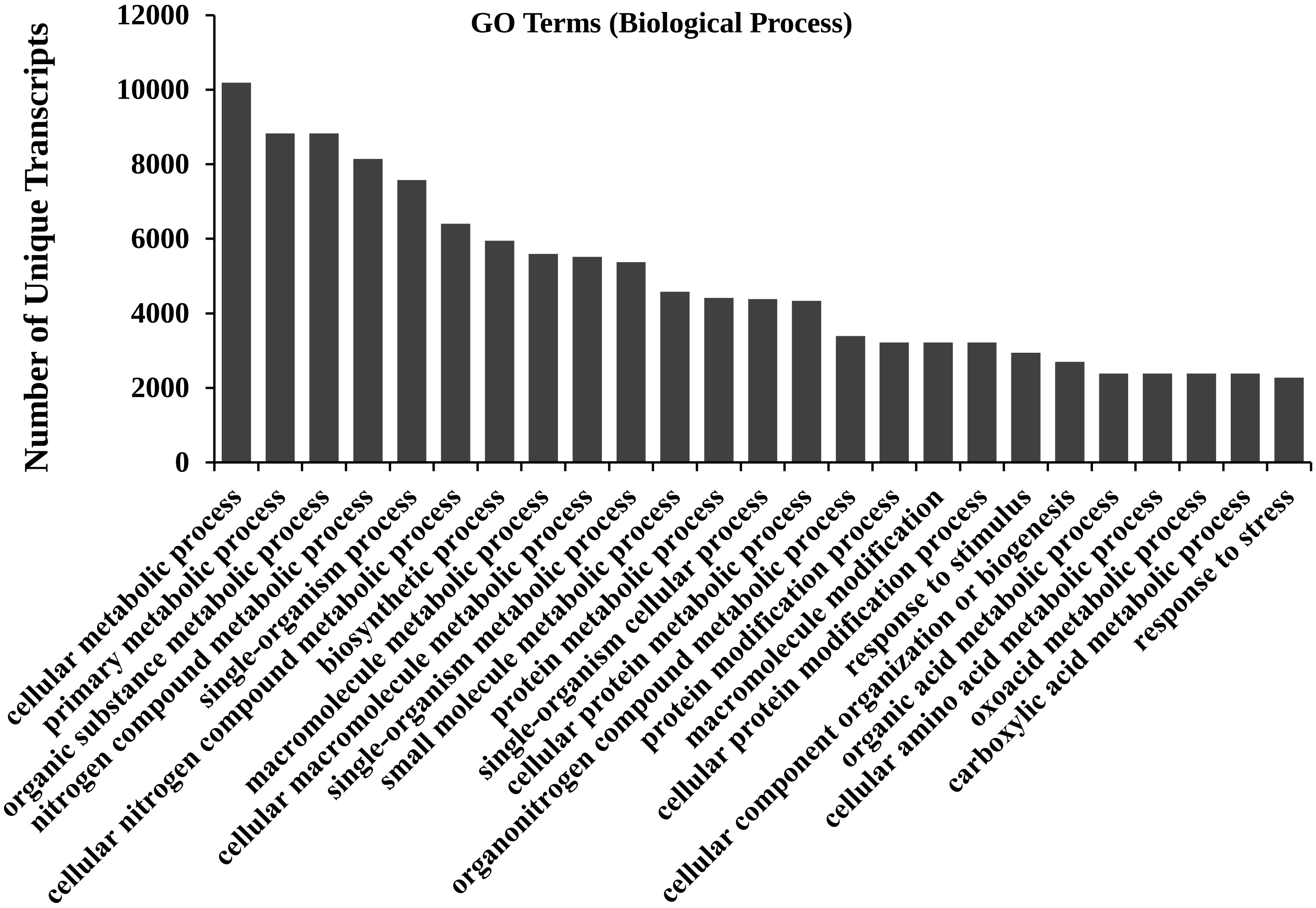


Figure 4. Functional classification of unique contigs in sweet cherry. The number of unique sweet cherry contigs associated with the GO terms for biological process in order of decreasing magnitude.

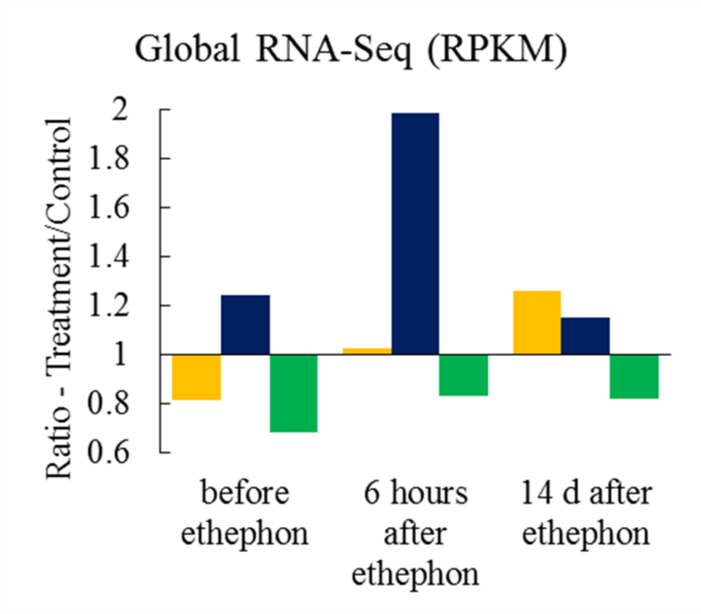


Figure 5. Mean gene expression across all transcriptomes. RPKM ratio of the treatment value was divided by the control expression value for every contig. Yellow represents ‘Chelan’, Blue represents ‘Bing’, and Green represents ‘Skeena’ genotypes. Data demonstrates a significant increase in the gene expression ratio 6 hours after ethylene treatment, indicating that ‘Bing’ is uniquely responding at the transcript level to the treatment.

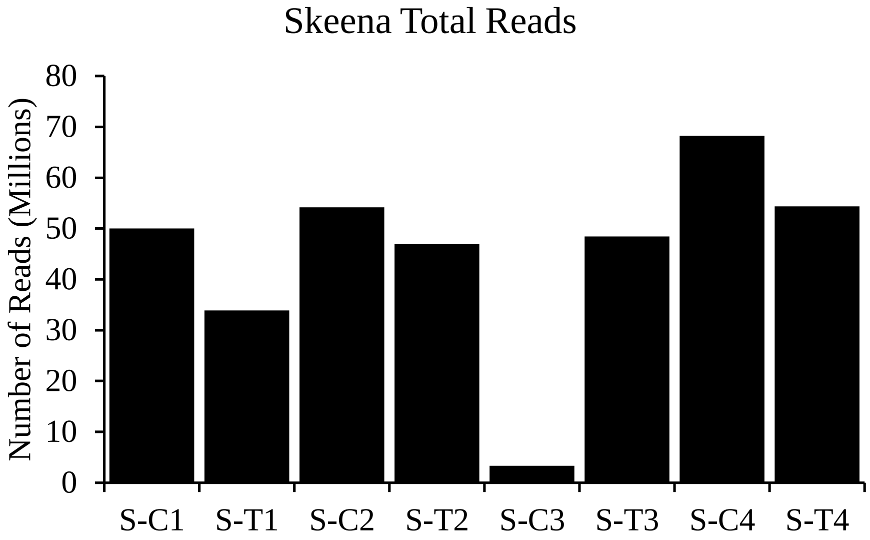


Figure 6. Number of RNAseq reads for 'Skeena' (2010). Total number of reads generated for 'Skeena' shows that there was a problem with 'Skeena', Control, Time point 3 (S-C3) sample. Failure in initial library construction is the most likely source for the error in this sample.

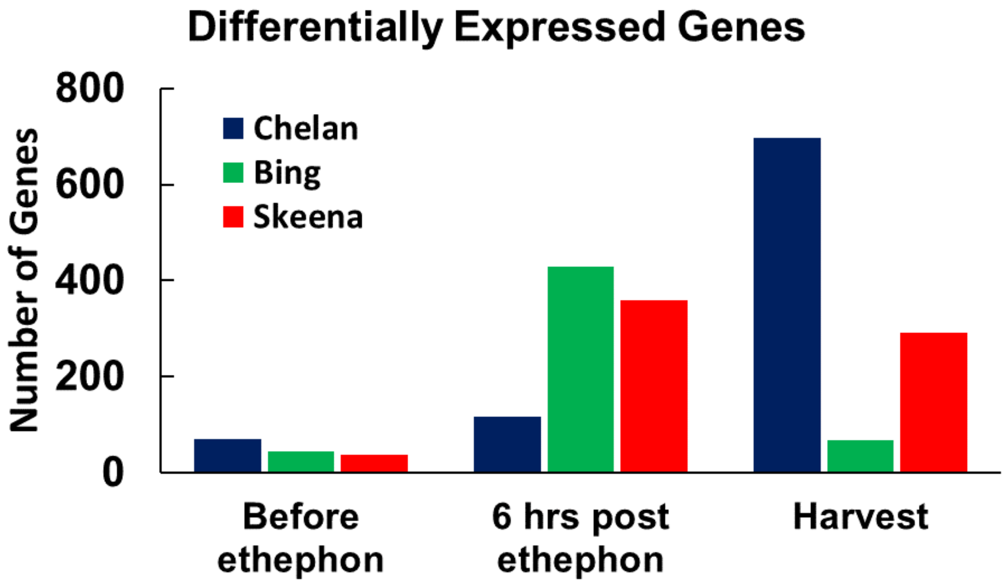
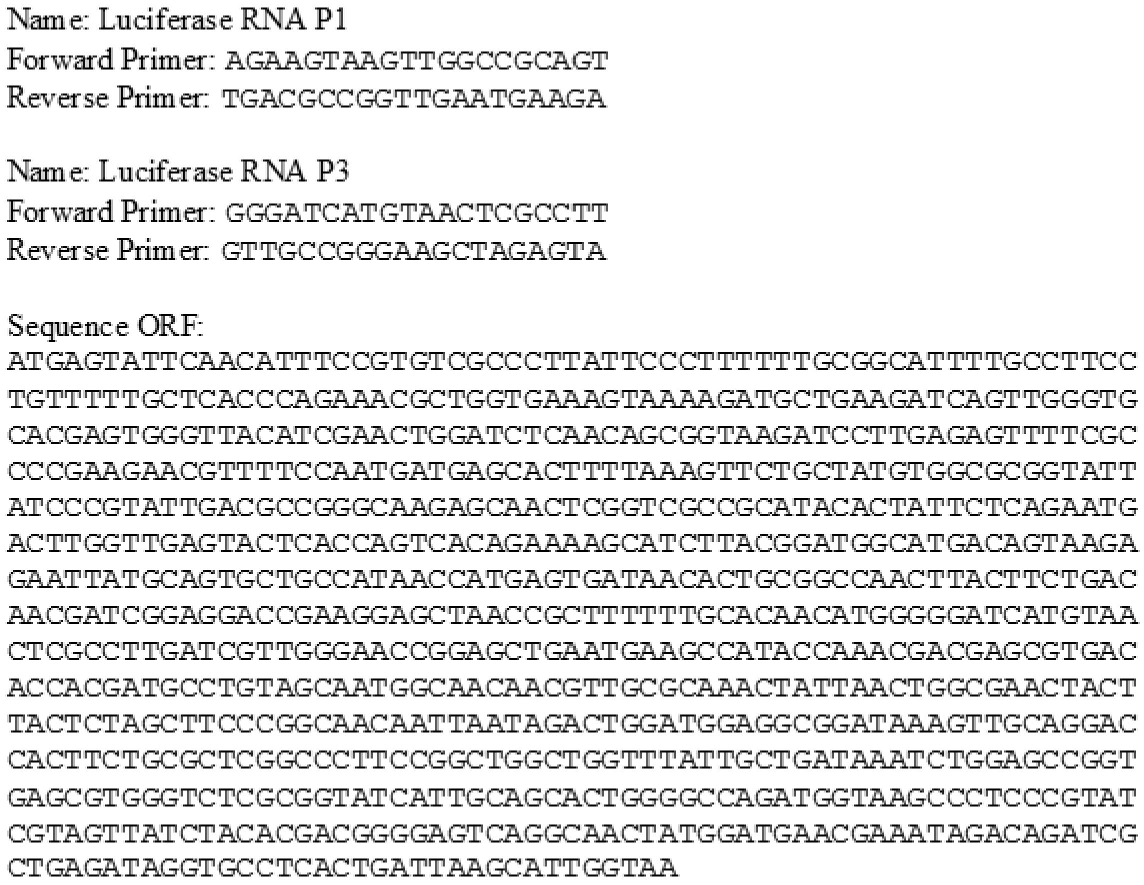


Figure 7. Total number of differentially expressed genes generated through the RNAseq workflow (CLC Genomics Workbench). These were generated by five-fold difference between ethephon treatment and control as well as Kal's test (0.95).



Name: Luciferase RNA P1
Forward Primer: AGAAGTAAGTTGGCCGCAGT
Reverse Primer: TGACGCCGGTTGAATGAAGA

Name: Luciferase RNA P3
Forward Primer: GGGATCATGTAACTCGCCTT
Reverse Primer: GTTGCCGGGAAGCTAGAGTA

Sequence ORF:

```
ATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCC
TGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTG
CACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGC
CCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATT
ATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATG
ACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGA
GAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGAC
AACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAA
CTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGAC
ACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACT
TACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGAC
CACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGT
GAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTAT
CGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCG
CTGAGATAGGTGCCTCACTGATTAAGCATTGGTAA
```

Figure 8. Luciferase gene sequence. This gene was used as the reference gene that was spiked into all the cDNA samples used in the qRT-PCR experiment.

Chapter 3 - Supplemental Files

Supplemental File 1a. 2010 PFAZ phenotypic data, including row, color, and PFRF raw data.

Supplemental File 1b. 2013 PFAZ phenotypic data, including row, color, and PFRF raw data.

Supplemental File 1c. 2014 PFAZ phenotypic data, including row, color, and PFRF raw data.

Supplemental File 2a. Read mapping table; Chelan reads mapped to *PaRef* assembly.

Supplemental File 2b. Read mapping table; Bing reads mapped to *PaRef* assembly.

Supplemental File 2c. Read mapping table; Skeena reads mapped to *PaRef* assembly.

Supplemental File 3a. Sequence file of Bing, Control, Time point 1 mapped to *PaRef* assembly.

Supplemental File 3b. Sequence file of Bing, Control, Time point 2 mapped to *PaRef* assembly.

Supplemental File 3c. Sequence file of Bing, Control, Time point 3 mapped to *PaRef* assembly.

Supplemental File 3d. Sequence file of Bing, Control, Time point 4 mapped to *PaRef* assembly.

Supplemental File 3e. Sequence file of Bing, Treatment, Time point 1 mapped to *PaRef* assembly.

Supplemental File 3f. Sequence file of Bing, Treatment, Time point 2 mapped to *PaRef* assembly.

Supplemental File 3g. Sequence file of Bing, Treatment, Time point 3 mapped to *PaRef* assembly.

Supplemental File 3h Sequence file of Bing, Treatment, Time point 4 mapped to *PaRef* assembly.

Supplemental File 3i. Sequence file of Chelan, Control, Time point 1 mapped to *PaRef* assembly.

Supplemental File 3j. Sequence file of Chelan, Control, Time point 2 mapped to *PaRef* assembly.

Supplemental File 3k. Sequence file of Chelan, Control, Time point 3 mapped to *PaRef* assembly.

Supplemental File 3l. Sequence file of Chelan, Control, Time point 4 mapped to *PaRef* assembly.

Supplemental File 3m. Sequence file of Chelan, Treatment, Time point 1 mapped to *PaRef* assembly.

Supplemental File 3n. Sequence file of Chelan, Treatment, Time point 2 mapped to *PaRef* assembly.

Supplemental File 3o. Sequence file of Chelan, Treatment, Time point 3 mapped to *PaRef* assembly.

Supplemental File 3p. Sequence file of Chelan, Treatment, Time point 4 mapped to *PaRef* assembly.

Supplemental File 3q. Sequence file of Skeena, Control, Time point 1 mapped to *PaRef* assembly.

Supplemental File 3r. Sequence file of Skeena, Control, Time point 2 mapped to *PaRef* assembly.

Supplemental File 3s. Sequence file of Skeena, Control, Time point 3 mapped to *PaRef* assembly.

Supplemental File 3t. Sequence file of Skeena, Control, Time point 4 mapped to *PaRef* assembly.

Supplemental File 3u. Sequence file of Skeena, Treatment, Time point 1 mapped to *PaRef* assembly.

Supplemental File 3v. Sequence file of Skeena, Treatment, Time point 2 mapped to *PaRef* assembly.

Supplemental File 3w. Sequence file of Skeena, Treatment, Time point 3 mapped to *PaRef* assembly.

Supplemental File 3x. Sequence file of Skeena, Treatment, Time point 4 mapped to *PaRef* assembly.

Supplemental File 4. Quantitative RT-PCR primer names, sequences and sequenced amplicons.

Supplemental File 5a. RNAseq data – RPKM values

Supplemental File 5b. RNAseq data – TPM values

Supplemental File 6a. Quantitative RT-PCR reaction conditions and thermal profile.

Supplemental File 6b. LinRegPCR input as fluorescence readings from qRT-PCR instrument and PCR cycle, with resulting efficiency and Cq output with regression statistics.