# ABSTRACT


| | |
|---|---|
| Title of Thesis: | GENOME ASSEMBLY AND COMPARATIVE RNA-SEQ VISUALIZATION OF RED RASPBERRY, *RUBUS IDAEUS* |
| | Haley Melissa Wight, Master of Science, 2019 |
| Thesis Directed By: | Dr. Sridhar Hannenhalli, Department of Cell Biology and Molecular Genetics |



The red raspberry, *Rubus idaeus*, is a major commercial fruit valued for its taste and high antioxidant content. Genomic resources are needed to improve research and breeding of this high value crop. Using a hybrid approach of both long and short sequence reads, a *de novo* genome was assembled consisting of 2,145 scaffolds with a genome completeness of 95.3% and an N50 score of 638 KB. 35,566 protein coding genes were annotated with a transcriptome completeness score of 97.2%. This resource contributes to phylogenetic and comparative studies of the agriculturally valuable Rosaceae family. Comparing the raspberry and strawberry transcriptomes reveal numerous transcription factors, both shared and unique, with roles in meristem regulation. Further, a comparative electronic fluorescent pictograph (eFP) web viewer was constructed to enable visualization of RNA expression patterns of orthologs across Rosaceae species.

# GENOME ASSEMBLY AND COMPARATIVE RNA-SEQ VISUALIZATION OF RED RASPBERRY, *RUBUS IDAEUS*

by

Haley Melissa Wight

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Masters of Science
2019

Advisory Committee:
Dr. Sridhar Hannenhalli, Chair
Dr. Stephen Mount
Dr. Najib El-Sayed

Acknowledgements

This work was primarily completed in the lab of Zhongchi Liu. I would like to thank Dr. Zhongchi Liu, Dr. Sridhar Hannenhalli, Dr. Stephen Mount, and Dr. Najib El-Sayed for providing me with an abundance of guidance throughout my graduate school education. I would like to acknowledge Dr. Junhui Zhou and Dr. Kelsey Galimba for performing total RNA and DNA extractions for sequencing. I would like to acknowledge Muzi Li for anchoring the genomic sequences to genetic maps.

The work is supported by NSF grants (IOS1444987) to Zhongchi Liu, Sridhar Hannenhalli, Ann Callahan and Chris Dardick. Financial support was also provided by the NSF Computation and Mathematics for Biological Networks Research Traineeship (NSF_NRT 1632976).

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1: Introduction

The red raspberry, *Rubus idaeus*, is widely distributed in all temperate regions of Europe, Asia, and North America and has been used as food and medicine since 4$^{th}$ century AD [1]. Often dubbed "European red raspberry", *Rubus idaeus* is an extremely important and globally commercialized specialty fruit crop with a large number of commercial varieties, high price, and increasing consumer demands. Owing to its health promoting value, unique flavor, and attractive appearance *Rubus idaeus* sales have recently climbed with world production over 795,000,000 tons (FAOSTAT; Red Raspberry Production). In contrast, the black raspberry, *Rubus occidentalis* (a species native to North America), is limited to the United States' Pacific Northwest region [4]. In addition to its economic and health-promoting value, the red raspberry plants possess interesting and sometimes unique biological characteristics such as cold hardiness, aggregate fruits, perennial roots and biennial canes, either summer-bearing or ever-bearing flowering/fruiting, and large numbers of hybrids and cultivars. However, red raspberry breeding and research has fallen behind relative to other special fruit crops due to several genetic problems including apomixes, pollen incompatibility, and poor seed germination [5]. In addition, there is a lack of genomic resources such as no genome sequencing and limited transcriptome data [6]. With the recent publication of a high quality black raspberry (*Rubus occidentalis)* genome [7], it is timely that a red raspberry genome be published for furthering comparative genomics, genetic breeding and gene identification.

*Rubus idaeus* is a member of the *Rosaceae* family, an economically important family that also includes rose, peach, apple, cherry, pear, almond, strawberry, and blackberry. Up to now, the genomes of several *Rosaceae* family members have been sequenced, including *Rubus occidentalis* (black raspberry) [4,7]*, Malus domestica* (apple) [8,9], *Prunus persica* (peach) [10,11], *Pyrus bretschneideri* (Chinese pear) [12], *Fragaria vesca* (woodland strawberry) [13,14], *Potentilla micranthia* (mock strawberry) [15], and *Rosa chinensis* (Chinese rose) [16,17]. Due to the small genome

size, wide variety of fruit types (pomes, drupes, achenes, hips, follicles and capsules), and plant growth habits (ranging from herbaceous to cane, bush and tree forms), *Rosaceous* genomes offer one of the best systems for comparative studies in genome evolution and development [18]. The availability of whole-genome sequences of key diploid species such as *Rubus idaeus* in this family will be crucial to these efforts. For instance, both raspberry and strawberry develop aggregate achenes (ovaries) on an enlarged receptacle. They were hypothesized to evolve from an ancestral fruit with relatively flat receptacle [18]. Whether raspberry and strawberry independently evolved an enlarged receptacle is not known.

Here, I report a draft genome assembly of red raspberry, *Rubus idaeus,* using long reads of single-molecular real-time (SMRT) Pacific Biosciences sequencing as well as high coverage Illumina short reads. The resulting draft genome is 299 Mbp in size and contains 2,145 scaffolds with a N50 of 638 Kb. Using RNA-seq data from dissected fruit tissues at two developmental stages, I annotated the genome yielding 35,566 protein coding genes with a BUSCO-calculated transcriptome completeness score of 97.2%. The genome was anchored to two previously published high density linkage maps of *Rubus idaeus* [19], facilitating future marker development, breeding, and identification of genes controlling useful trait characteristics.

Comparative transcriptomics between raspberry and strawberry receptacle reveals both shared and unique transcription factors with roles in meristem regulation, suggesting possible independent evolution of raised receptacle in raspberry and strawberry. Additional comparative analysis, bolstered by this reference sequence, will allow the study of the complex evolution of morphological diversity in fleshy fruits of *Rosaceae*. To aid this comparative analysis, the lab of Zhongchi Liu has generated tissue- and stage-specific transcriptome profiles of peach, apple, raspberry and strawberry throughout early fruit development. To make this data accessible, I have developed a comparative electronic fluorescent pictograph (eFP) web viewer. The function of an eFP is to display a cartoon image depicting various tissue types, each colored in with a hue indicating the level of expression of a queried gene. The

eFP software, widely used in model organisms such as mouse, has been employed here to display the fruit transcriptome data within individual species as well as expression patterns of orthologs across the four fruit crops. The creation of this web interface will make the transcriptome data available to the *Rosaceae* community for cross-species comparisons and within species visualization of early stage fruit development.

# Chapter 2: Genome Assembly and Annotation of Red Raspberry *Rubus Idaeus*

*Genome assembly*

*Rubus idaeus* is a diploid species (2n=2x=14) with an estimated genome size of 293 Mbp based on flow cytometry analysis [5]. I first sequenced the *Rubus idaeus* genome using 120X Illumina coverage. A k-mer distribution analysis estimated the *Rubus idaeus* genome to be 303 Mbp. The bimodal distribution of 31-mers (Figure 1) suggests heterozygosity in the genome.

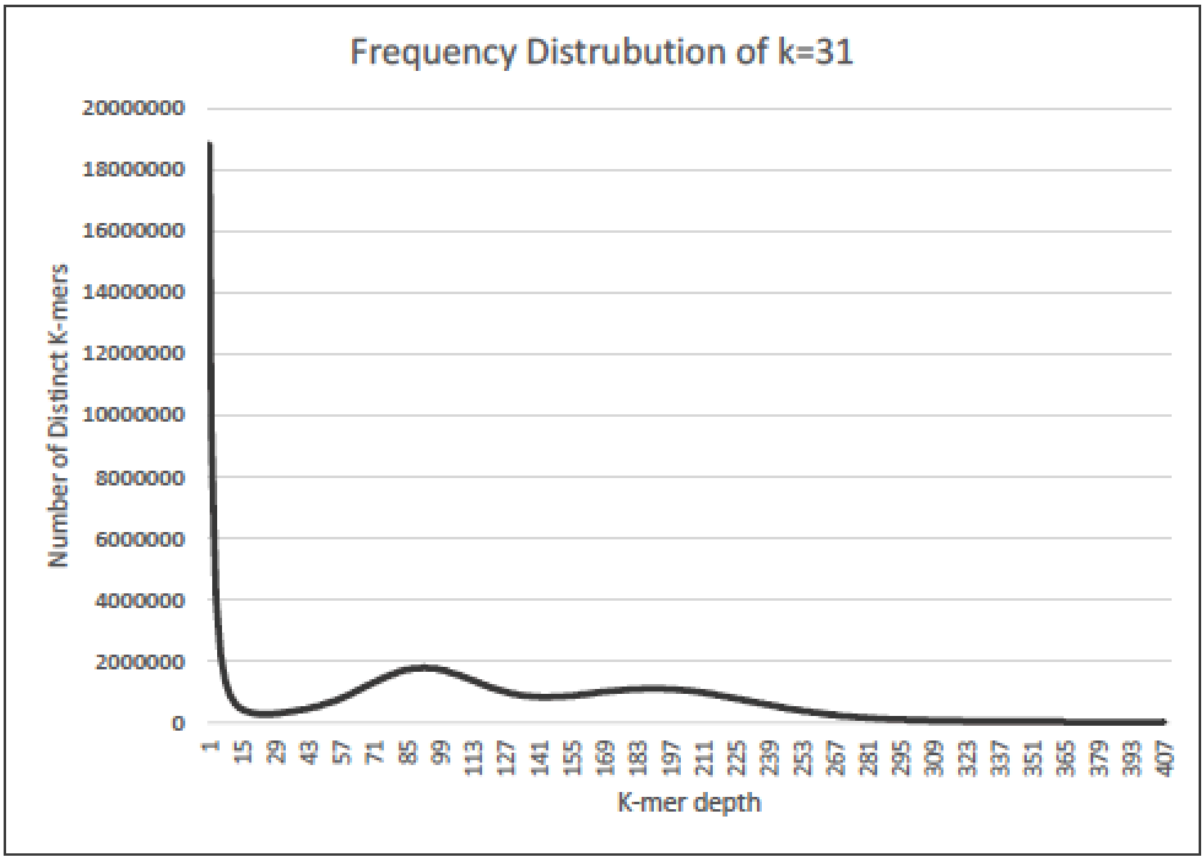


**Figure 1**. Bimodal K-mer distribution of *Rubus idaeus* (variety Joan J.) genome

31-mer distribution of *Rubus idaeus* genome using jellyfish with 150-bp paired-end whole genome sequencing data.

To overcome the issue of heterozygosity for genome assembly, a hybrid genome assembly approach was used taking advantage of both the sequencing depth and accuracy offered by the Illumina platform (at 120X coverage) and the sequence length offered by the PacBio platform (at 26X overage). The pipeline of the assembly is outlined in Figure 2. I used Redundans [20] and Haplomerger2 [21] tools to correct for heterozygosity. A comparative genomic approach [22,23] was used as part of the genome assembly. Specifically, the most recently assembled genomes of closely related species *Potentilla micranthia* [15]*, Rubus occidentalis* [4], and *Fragaria vesca* [13] were leveraged to improve scaffolding using MeDuSa [22]. The resulting *R. idaeus* genome assembly is 299 Mbp in size containing 2,145 scaffolds with a N50 of 638 Kb (Table 1). To assess the completeness of the genome, BUSCO v.3.0.2 [24] was used to locate the presence or absence of the embryophyta_odb9 (plant) dataset. The BUSCO Completeness Score reached 95.3% (Table 1), which validates the good assembly quality.

**Table 1**. Statistics of genome and transcriptome assemblies. Single (S), Duplicated (D), Fragmented (F) and Missing (M) single-copy orthologs are reported alongside the BUSCO completeness score.

| | |
|---|---|
| Total length | 299,945,700 bp |
| Scaffold N50 | 638,152 bp |
| Contig N50 | 250,294 bp |
| Smallest Scaffold | 501 bp |
| Largest Scaffold | 4,458,320 bp |
| N's | 174,429 bp (.000582%) |
| Sequence GC's | 37.9% |
| % Repeats | 43.35% |
| Busco Completeness Score (Genome) | 95.3% (S:86.1%, D:9.2%), F:1.5%, M:3.2% |
| Number of Annotated Protein Coding Genes | 35,566 |
| Busco Completeness Score (Transcriptome) | 97.2% (S:92.9%,D:4.3%), F:1.1%, M:1.7% |

The genome assembly pipeline is shown in Figure 2. A mixture of Illumina short reads and Pacbio long reads were assembled into contigs using MaSuRCA; an assembler which combines the efficiency of de Bruijn graph and Overlap-Layout-

Consensus approaches [25]. The specific settings used in the configuration file other than default were PE= pe 180 20, JF_SIZE = 200000000 and SOAP_ASSEMBLY=0. Subsequently, Redundans was used to remove heterozygous contigs using an all versus all BLAT approach. The Redundans pipeline also performed scaffolding using a mixture of Illumina short reads and Pacbio long reads [20]. The genomes of *Potentilla micranthia* [15]*, Rubus occidentalis* [4], and *Fragaria vesca* [13] were leveraged to improve scaffolding using MeDuSa [22]. Scaffolds with less than 10X coverage were removed and scaffolds with more than 500 consecutive N's were split. Bowtie2 version 2.3.0 [26] was used to map the Illumina reads back onto the genome prior to Pilon with maximum fragment length to be 1000 and default settings elsewise. The mapping rate was 97.8% which further validates assembly quality. Pilon [27] was then used for one iteration to correct bases, fix misassembly and fill assembly gaps using the diploid parameter. Repeats were then softmasked by first creating a custom repeat library with RepeatModeler (Smit et al., 2015) using the NCBI engine option and then using RepeatMasker (Smit et al., 2013). Lastly, Haplomerger2 [21] split the resulting assembly into two sub-assemblies to further remove hetereozygosity.

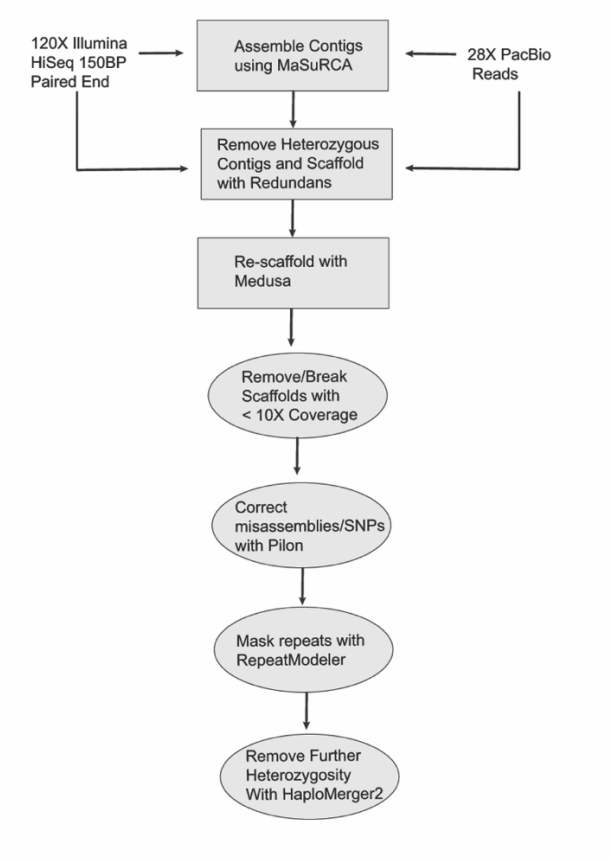


**Figure 2**. **Genome assembly pipeline.**

Flowchart represents all steps of the genome assembly process upstream of anchoring to the linkage map.

*Anchoring scaffolds to genetic maps*

The following section was created by Muzi Li. The scaffolds were anchored onto pseudochromosomes (Figure 3) taking advantage of two previous genetic linkage maps. They are respectively the 'Heritage' and 'Tulameen' variety-based linkage maps that collectively contained 4225 markers. As a result, the pseudochromosomes contain 80.1% of the assembly (ie. at 240 Mb). The average magnitude of the Pearson correlation coefficient between the physical and map locations is 0.92 showing a high consistency between the genome and previously published linkage maps (Figure 3).

BLAT was run with default settings to identify unique and complete matches to each marker [30]. After preparing the input files from BLAT, pseudochromosomes were then constructed using ALLMAPS with default parameters [31]. Each genetic map was given a weight of 1. Chimeric scaffolds were manually broken at positions with low coverage, correcting many misassemblies. The seven pseudochromosomes were then constructed by integrating 98% of the markers from the genetic map.

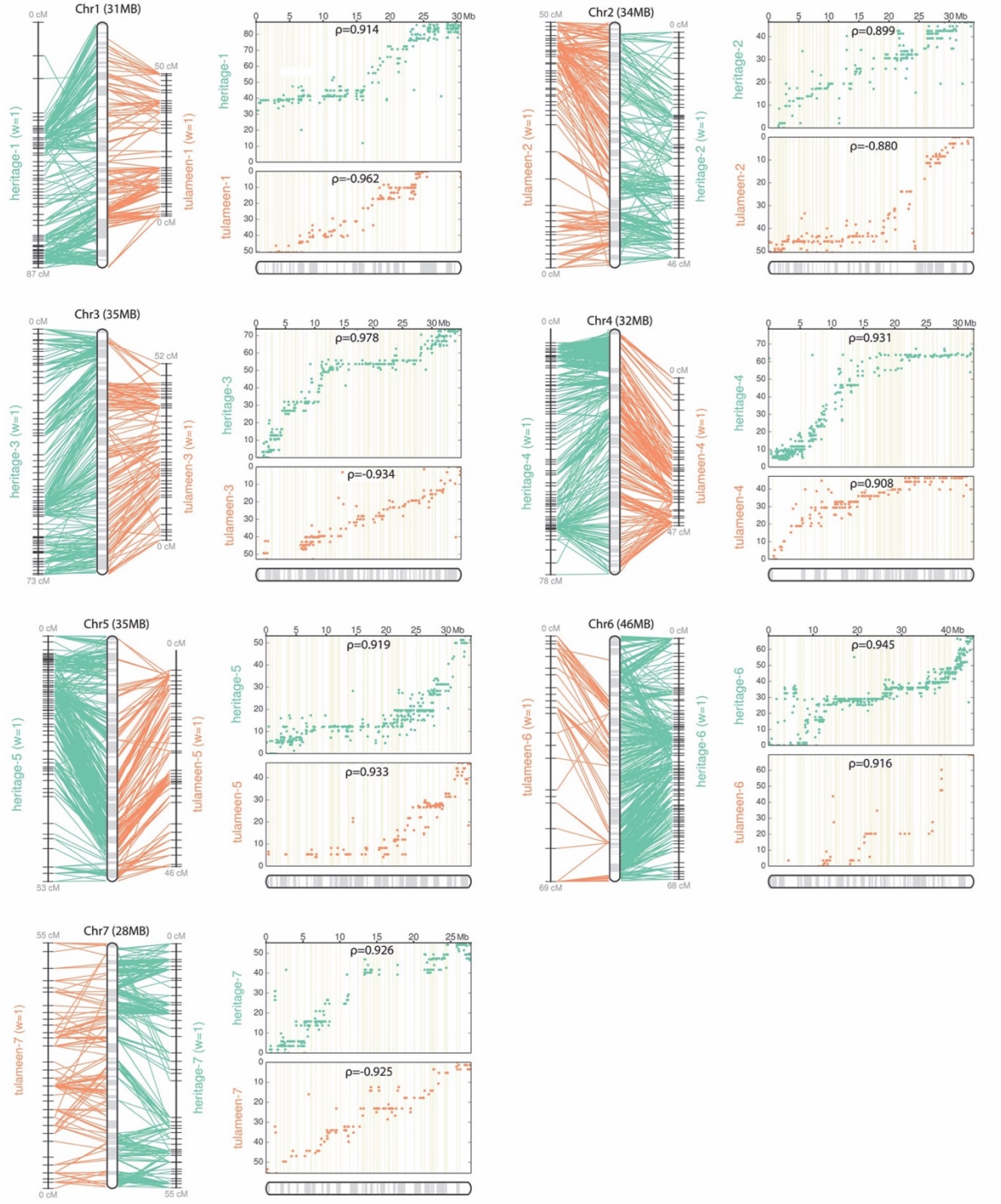


**Figure 3. The correlation between physical map and the linkage maps of seven chromosomes.**

This figure was created by Muzi Li. For each chromosome, the left figure illustrates the connections between physical positions on the assembled pseudomolecule and the two flanking linkage maps colored in orange and teal respectively. The orange

coloring represents the tulmaneen linkage map whereas the teal represents the heritage linkage map [19]. On the right is the scatter plot with dots representing the physical position on the chromosome (*x* axis) versus the map position (*y* axis). Rho ($\rho$) is the Pearson correlation coefficient (right panel). Each panel represents distinct chromosome.

*Genome annotation*

To annotate the *Rubus idaeus* genome, a transcriptome was generated from 1,057,377,357 Illumina RNA-seq reads; they are pooled from 18 RNA-seq libraries derived from three different fruit tissues (ovary wall, ovule/seed, receptacle) at two developmental stages (0 and 12 Days Post-Anthesis or DPA) multiplied by three biological replicates. A combination of *ab initio* and alignment guided assembly was employed to annotate the soft-masked genome. This resulted in 35,566 protein coding genes with a BUSCO-calculated transcriptome completeness score of 97.2% (Table 1). The high completeness score indicates that majority of the gene regions have been sequenced. Finally, Blast2GO was used to associate Gene Ontology (GO) terms to the annotated genes.

Repeat Masker was used with a custom repeat library built with Repeat Modeler. The Illumina Reads of RNA-seq data described above were trimmed with Trimmomatic [32]. RNA-Seq reads were mapped onto the draft genome sequence using Bowtie2 [33]. The bam file obtained was used to generate the training set for the gene prediction of BRAKER1 pipeline [34]. Candidate transcripts containing no known protein domains by Interproscan5 [35] were removed from the final set (13.96% percent decrease).

Trinity was then used to assemble the transcriptome on both genome guided and *de novo* settings [36]. Prior to trinity assembly, reads were normalized using the perl script provided by Trinity and aligned using Bowtie2 [33,36]. Trinity assemblies were amassed into a comprehensive transcriptome database using PASA [37]. Lastly, cd-hit-v4.6.8 [38] was used to cluster transcriptome assemblies from the resulting PASA and BRAKER1 assemblies with over 95% identity into unigenes. Unigenes that did

not map to the genome, had no RNA-seq evidence, and had no known protein domains or orthologues were removed.

Blast2GOPro version 5.1.1 was used to associate Gene Ontology (GO) terms to the resulting transcripts. Protein sequences were searched against the non-redundant (nr) database protein database from NCBI using BLASTP with an e-value cutoff of 1.0E-3 [39]. InterProScan was run using default databases in order to assign putative domains to each transcript.

*Analysis of Rubus idaeus transcriptome*

The dynamics of gene expression values throughout fruit development was studied within the fruit tissue of *Rubus idaeus* that forms the edible fruit (the seed and wall). Global analysis of the RNAseq libraries was performed by Principal Component Analysis (PCA). The PCA plot shows a clear separation of tissue type (Figure 4D). The severe heterogeneity between tissue type reinforces the demand for a spatial approach when studying fruit development.

Both seed and wall had the highest number of differential gene expression at stage 9 and stage 6, respectively (Figure 4A and B). The drastic level of differential expression at stage 6 and stage 9 within wall and seed tissue motivated a Gene Ontology (GO) analysis to be performed on these genes. Genes responsible for this volatility were identified to be involved in lignin polymers synthesis, consistent with the morphological changes of the seed coat and stone formation during 6 DPA (Figure 4C).

*Rubus idaeus* is a stone fruit in which an outer fleshy part surrounds a single shell of hardened endocarp with a seed inside, such as a peach pit. The stone is composed mostly of lignin, a structural polymer that is also known to create the rigid structure of bark. There is a high overlap of 18 GO enriched terms between the wall and seed tissue (Figure 4C). Many of the terms in this overlapping set refer to lignin metabolism, carbohydrate metabolism, and cell wall rearrangement. The formation of

lignin is catalyzed by peroxidases which is another functional GO term enriched at stage six in both the wall and seed. This is consistent with the formation of the stone endocarp. Although tissue libraries exhibited unmistakable clustering with regards to principal component analysis, there seems to be a concerted action between tissue types to form a protective barrier around the seed during fruit development.

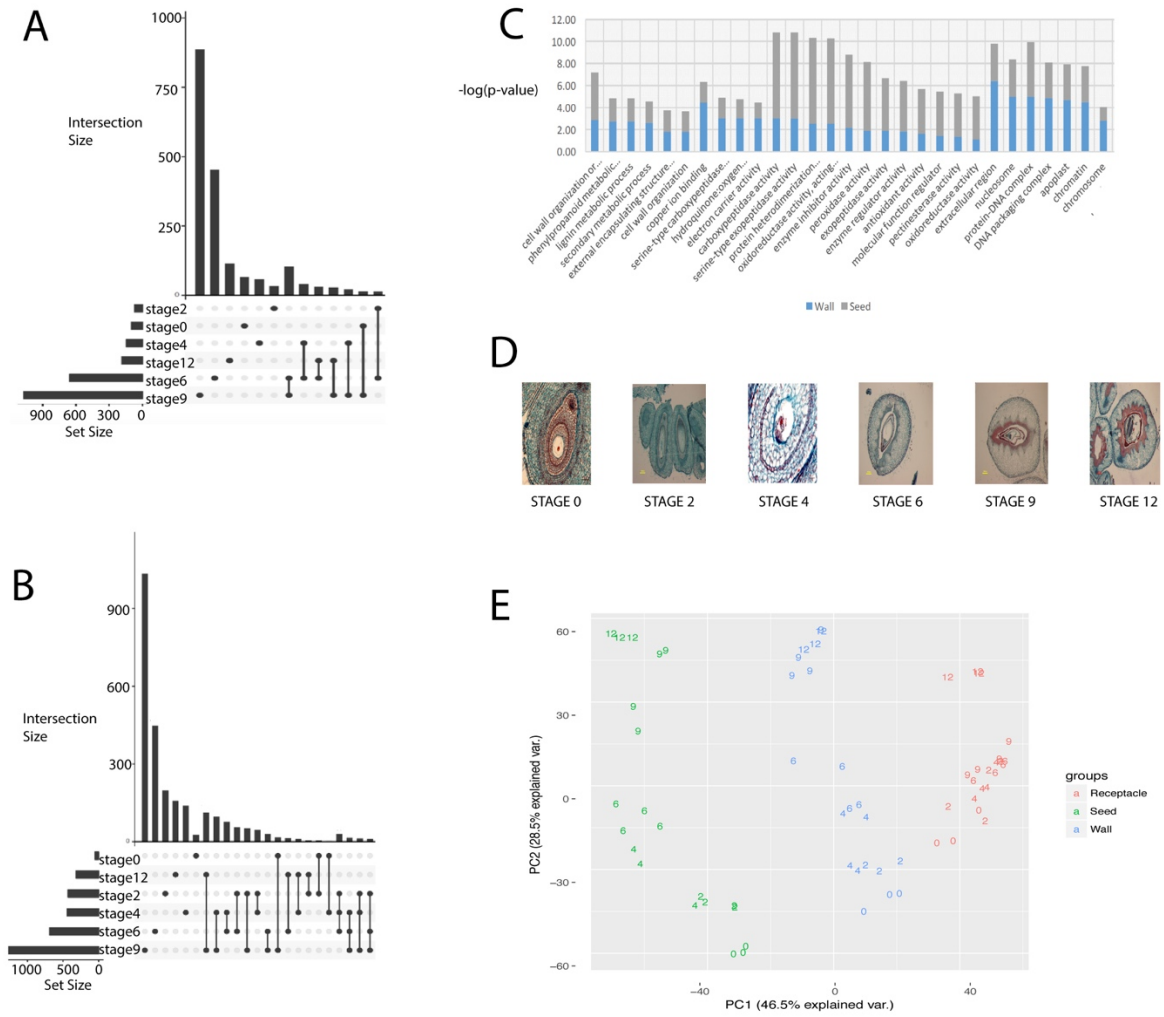


**Figure 4. RNA-seq and Differential Expression Analysis**

(A) Intersections of upregulated genes of **wall** tissue among different stages. The intersections are plotted as a matrix where each column corresponds to an intersection and each row corresponds to the set considered within the intersection. Filled cells indicate the set is included in the intersection plotted whereas light grey cells indicate the set is not included in the intersection. (B) Intersections of upregulated genes of

**seed** tissue among different stages. (C) Gene Ontology (GO) terms of upregulated genes in stage six of seed and wall tissue. (D) Tissue dissection of raspberry throughout development shows stone formation begins to form at stage 6 and is fully developed by stage 9. This dissection was done by Junhui Zhou. (D) Principal Component Analysis of RNA-seq samples. The tissue is indicated by color and is separated primarily on PC1; stage is indicated by number and is separated on PC2.

# Chapter 3: Comparative Genomics and Transcriptomics Within the Rosaceae Family

*Comparative Genomics*

Orthologous gene groups were established from 10 angiosperms using OrthoFinder-1.1.2 [40]; these include 9 members of the *Rosaceae* family (*Prunus persica*, *Pyrus communis*, *Malus domestica*, *Rosa chinesis*, *Rosa multiflora*, *Rubus occidentalis*, *Rubus idaeus*, *Fragaria vesca*, *Potentilla micrantha*) and the model organism *Arabidopsis thaliana* as the outlier species to root the tree. The resulting phylogenetic tree (Figure 5A) is consistent with previously published phylogenetic analyses of the *Rosaceae* family [18]. In total 25,193 orthogroups were established. As shown in Figure 5A, 10,205 orthogroups contained proteins from all 9 *Rosaceae* species as well as *Arabidopsis*. Interestingly, many specific orthogroups (1,878) are unique to *Malus domestica* and *Pyrus communis*. Both species belong to the subfamily *Maleae*, which has undergone whole genome duplications at its origin [8,18,41]. The large number of orthogroups shared between *Malus domestica* and *Pyrus communis* suggests that substantial diversification occurred after whole genome duplication (WGD) within the *Maleae* subfamily, which may have contributed to the subfamily's pome fruit type [9,18]. Expectedly, all members of the *Rosaceae* family share many orthogroups (1,420) that are distinct from *Arabidopsis thaliana*. Members of the same genus also show a high number of common gene families. Specifically, there are 1,071 and 775 orthogroups limited to the *Rosa* and *Rubus* genera, respectively (Figure 5A). As *Rubus* is one of the largest and most morphologically diverse genus in the *Rosaceae* family [42], I examined the enrich GO terms among the 775 *Rubus*-specific orthogroups. Significantly enriched GO terms include chromatin assembly, RNA-splicing, and fungal-type cell wall organization, suggesting that *Rubus*-specific genes are involved in gene regulation and defense.

Strawberry and raspberry share the same base chromosome number (n=7); they have diverged probably ~75 million years ago. *Rubus occidentalis* and *Rubus idaeus*, on the other hand, are closely related species. Syntenic blocks revealed a high

collinearity between *Rubus idaeus* and *Rubus occidentalis* and between *Rubus idaeus* and *F. vesca* (Figure 5B and C). *R. occidentalis* had 25,289 gene pairs represented within 1,596 collinear blocks with *R. idaeus. F. vesca* and *R. idaeus* shared 17,769 syntenic gene pairs within 887 collinear blocks. This high degree of synteny helps validate the *Rubus idaeus* assembly. When compared with the more distant peach genome, *Prunus persica*, which has a different base chromosome number (n=8), collinearity decreases slightly: *P persica* and *R. idaeus* share 17,064 gene pairs on 877 collinear regions. Although there is lower collinearity, there are strikingly large conserved syntenic blocks. For example, a large portion of *Rubus idaeus* chromosome 7 corresponds with Peach chromosome 2 while a smaller portion of *Rubus idaeus* chromosome 7 correlates with peach 7 (Figure 5D).

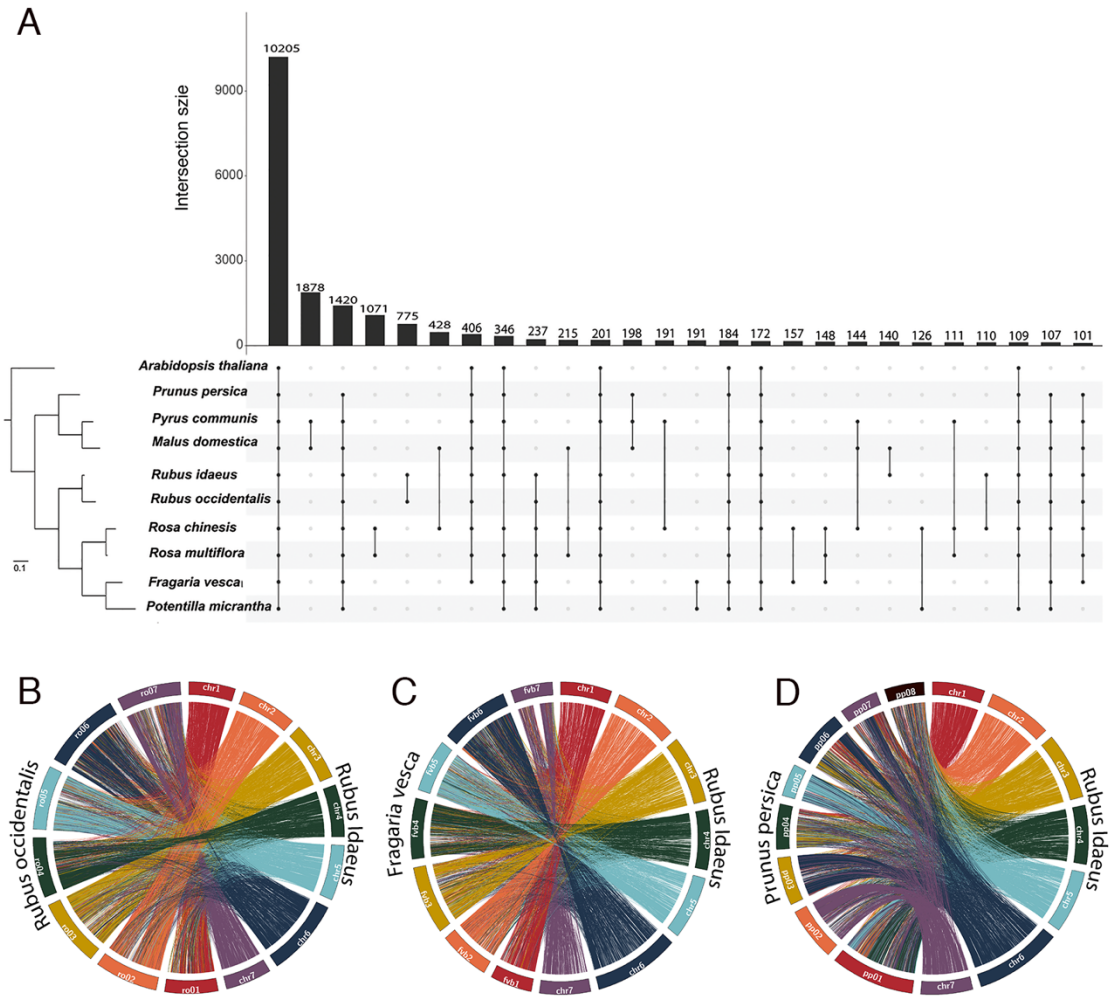

**Figure 5. The distribution of shared gene families among nine *Rosaceae* species and *Arabidopsis thaliana.***
(A)The left panel describes the phylogeny among the species. The branch length distances represent substitutions per site. The right panel is an UpSet plot [43]: an alternative representation of a venn diagram with intersections (shared genes) greater than 100. The species described in each intersection is represented by the dotted lines, the size of the intersection is described by the bar chart above. (B) Circos plots [44] displaying macrosynteny between the genomes of *Rubus idaeus* and *Rubus occidentalis*. (C) Macrosynteny between *Rubus idaeus* and *Fragaria vesca.* (D) Macrosynteny between *Rubus idaeus* and *Prunus persica*. For B to D, each connecting line represents an orthologous gene pair and the right half of each circle consists of the seven *Rubus ideaus* chromosomes colored by the spectral order in the rainbow.

*Comparative Transcriptomics of Rubus idaeus and Fragaria vesca*

*The Rosaceae* family members produce many economically important fruits with diverse fruit. Of particular interest is the contrasting fruit morphology between raspberry and strawberry (Figure 6A). Both raspberry and strawberry possess a raised receptacle covered by numerous aggregate ovaries, yet their fruit flesh is derived from different tissues: ovaries form the fleshy fruit in raspberry while the receptacle enlarges into the fleshy fruit in strawberry. A hypothetical ancestral fruit to both fruits may have existed ~75 million years ago and may possess a flat receptacle topped with numerous aggregated ovaries (achenetum) [18]. It was proposed that the receptacle in both fruits may have evolved to independently become raised and enlarged.

I compared receptacle-expressed genes between raspberry and strawberry, taking advantage of RNA-seq data derived from three tissues: receptacle, ovule, and ovary wall at 0 DPA (when the flower just opens) from both raspberry and strawberry. Interestingly, raspberry receptacle has the highest number of receptacle-up-regulated genes or DEGs (differentially expressed genes) when compared with the ovary wall and the ovule (Figure 6B), suggesting *R. idaeus* receptacle a highly complex tissue. In addition, between species comparison revealed significant overlap between raspberry

and strawberry receptacle-up-regulated DEGs according to the hypergeometric test (Figure 6B). I examined the transcription factors (TFs) among the overlapping set of genes and observed a large number of homeobox genes whose *Arabidopsis* homologs were shown to function in meristem maintenance such as STM, KNAT6, and Pound-Foolish (Figure 6C) [45,46]. The high prevalence (~47.7%) of known meristem regulators among the shared receptacle-specific TFs is consistent with previous suggestion that receptacles possess meristem-like activities [47,48] . The analysis here expands this observation further to include the raspberry receptacle.

In addition to the overlapping set of DEGs between strawberry and raspberry receptacles, a large number of genes (~1500 genes) are unique to raspberry receptacle. TFs extracted from this group of genes include several noted families of regulators (Figure 6C): (1) genes involved in regulating the abaxial and adaxial polarity such as *CRC, YABBY5,* and *[49]*; (2) known meristem regulatory genes *NAC/NAM* as well as *REM* [50,51]; (3) *AP2-EREBP* genes involved in ethylene signaling or other diverse biological processes [52].The expression of these unique TF families in raspberry receptacle suggest the likelihood of independent evolution of raised receptacle in raspberry and strawberry, leading to convergent use of shared TFs such as STM and independent use of unique TFs such as NAC/NAM. An alternative explanation of these raspberry unique TFs is that they evolve to provide additional properties to the receptacle including resistance to stress.

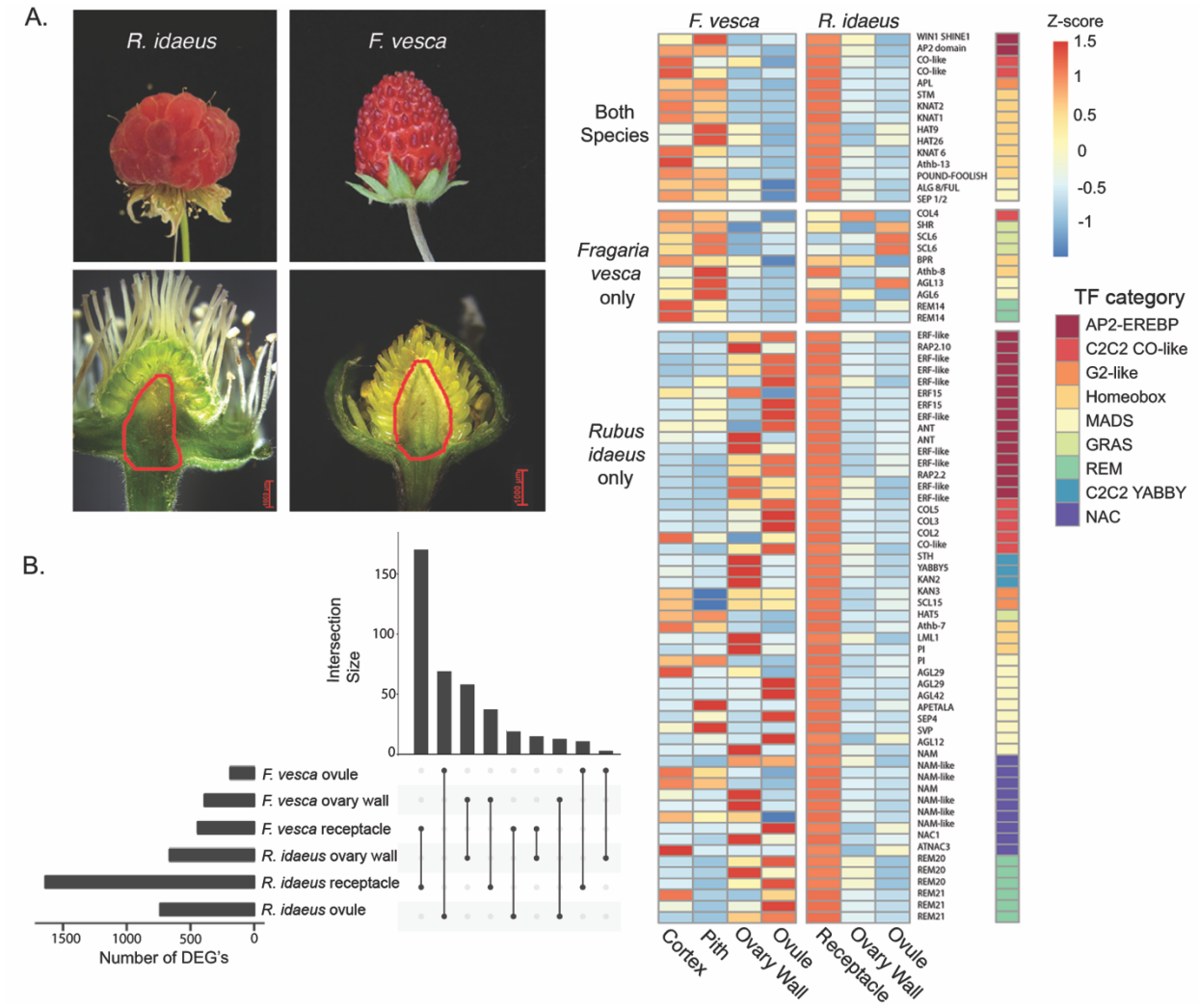


**Figure 6. Comparative transcriptome analysis of receptacle in raspberry and strawberry**

(A) Ripe fruit (top panel) of a red raspberry *R. idaeus* and a wild strawberry *F. vesca*, highlighting the juicy achenes in raspberry and dry achenes in strawberry. The dry achenes are embedded in the juicy receptacle of the strawberry. Beneath the ripe fruits are the young receptacle (at 0 DPA) in *R idaeus* and *F. vesca*. The red outline indicates the receptacle tissue used for RNA-seq. In strawberry, the receptacle is further divided into pith (center) and the cortex. These images were provided by Zhongchi Liu. (B) An UpSet plot displaying overlapping tissue-specific genes between the two species. The highest overlap between the two species is in the receptacle, which is significant based on the hypergeometric test. (C) Heatmap of relative expression level of transcription factors (TFs) across tested tissues in the two species. The TFs are grouped based on they are specifically expressed in the receptacle of both species (top panel), specifically expressed in the receptacle of *F. vesca* only (middle panel), and specifically receptacle of *R. idaeus* only (bottom). Similar TFs or TFs acting in related processes are categorized and coded with similar color and shown to the right of the heatmap.

# Chapter 4: Electronic Fluorescent Pictograph (eFP) and Comparative Viewer

*Single species viewer*

Oftentimes publicly available RNA-seq data goes underutilized because of the enormous volume of the data and the inaccessibility of the data format due to limited computational knowledge of many biologists. There is considerable value, therefore, in a simple, web-based interface that allows existing transcriptome data sets to be explored in a visual manner. The first eFP was created for non-bioinformatics scientist to interact with Arabidopsis and mouse model datasets [53]. The user is first presented with a form to submit a gene ID of interest. Upon submission, the varying plant tissues at varying timepoints are colored according to the expression level of the gene of interest. The tool is intended as a quick and easy means of identifying significant changes of gene expression and can be used to facilitate hypothesis generation.

The *Rubus idaeus* eFP encompasses data collected from ovule, seed, ovary wall, and receptacle tissue at 0, 2, 4, 6, 9 and 12 days post anthesis (DPA). After the user enters a gene ID (Figure 7A), they are redirected to the result page. This result page includes the eFP images (Figure 7B), bar graph representation (Figure 7C) of log(tpm) expression, a table of the raw data, and a list of the gene's orthologs in other species of Rosaceae. The eFP image and bar graph are downloadable to a png image and the table of raw data can be downloaded as an excel file. I have also created identical single species viewers for *Malus domestica* and *Prunus persica*. The *Malus domestica* dataset encompasses ovule, seed, ovary wall, hypanthium, and pedicel tissue at 0, 6, 12 and 20 DPA. The *Prunus persica* dataset encompasses ovule, seed, ovary wall, hypanthium, and pedicel tissue at 0, 5, 12 and 18 DPA.

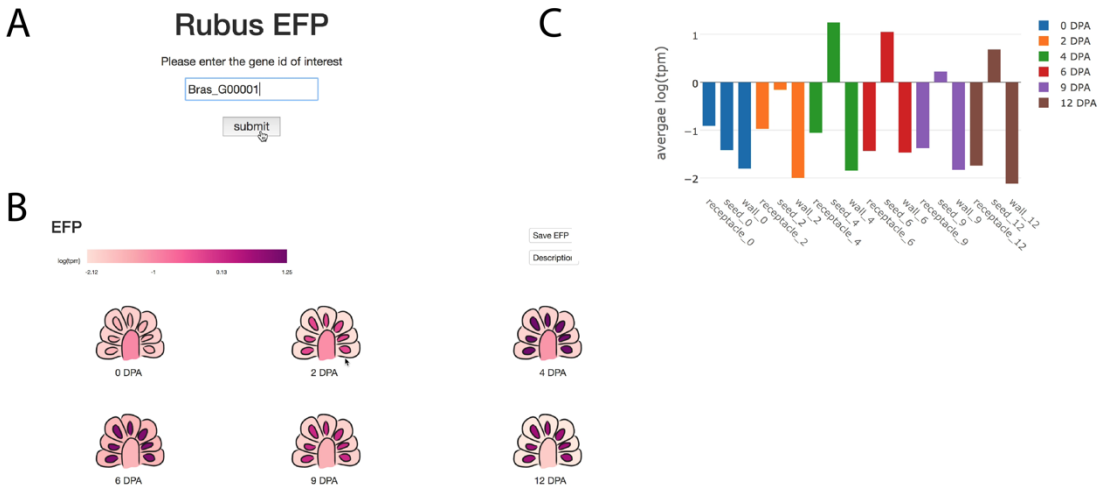


**Figure 7. Example user interface elements for gene number Bras_G00001**
(A) User entry form (B) eFP Browser output image (C) Bar graph representation of expression data

In extension to the single species viewer, the eFP website also has capabilities to visualize multiple genes within the same species through a heatmap. The user enters gene IDs in a comma separated list format (Figure 8A). The response page then contains a heatmap of every gene entered in the list (Figure 8B). The scale of this heatmap can be dynamically decided to be either be log(tpm) or z-score. On mouse hover of every box reveals the exact expression value. This feature is particularly helpful for multigene families.

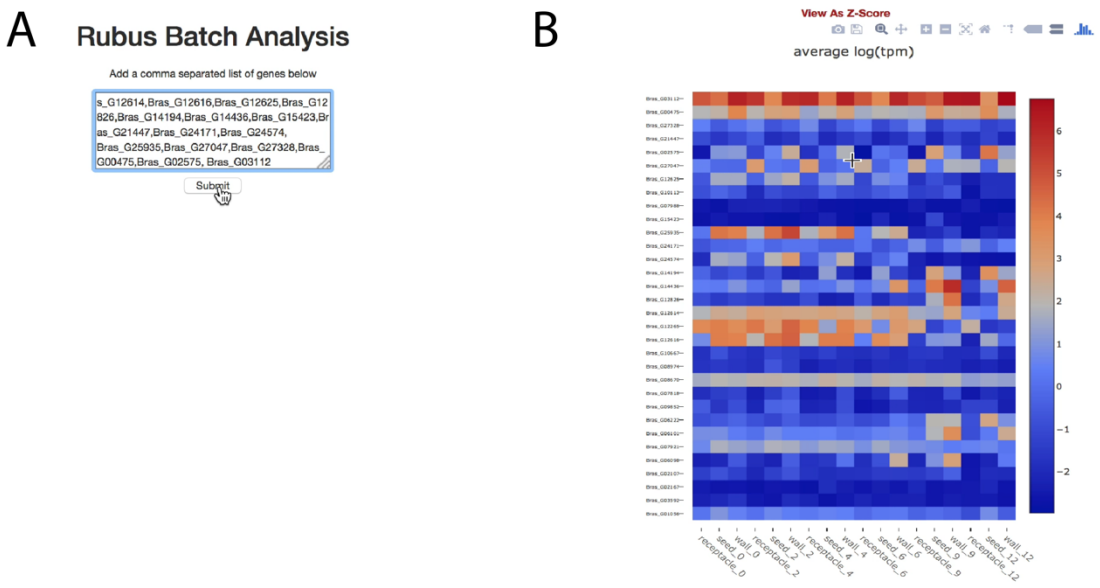


**Figure 8. Example user interface elements for *Rubus idaeus* batch viewer**
(A) User entry form (B) heatmap output of log(tpm) expression values

The eFP browser is implemented using a MEAN (MongoDB/ Express.js/ AngularJS/ Node.js) stack. The primary motivation for this was to use the same language, Javascript, across both the front and backend development of the web application. Backend queries are made to MongoDB, a non-relational and document-oriented database that is much faster at handling large sets of unstructured data than its SQL counterparts. The queries made within the batch viewer utilize AJAX (Asynchronous JavaScript and XML). AJAX allows the result page to be updated asynchronously through JSON documents received from the backend server. This allows for the user to dynamically add and remove genes of interest without having to reload every data entry in the query. After the data is queried, a JSON document is sent to the HTML response page to display the results. For each tissue, the engine loops through every pixel of the image and replaces all instances of the tissue's color with the color reflecting the expression intensity. Thus, image rendering is completed client side using HTML canvas. The create_efp.js script performs most of the functions for the generation of the output image. The bar graph and heatmap are constructed with the python's plotly library. The fruit diagrams in the eFP were drawn using Adobe

Illustrator. Orthologs were precalculated using Orthofinder on default settings [54]. Expression values were obtained by using Salmon, a pseudoalignment tool, to all current transcriptomes of the appropriate species [55–57].

*Comparative viewer*

While most such analyses are focused on biological processes occurring within a species, comparative analysis of similar biological processes across different species is emerging as a powerful approach in achieving novel insights into the development and evolution of morphological and functional diversity. Viewing only one species at a time is limiting when a comprehensive, comparative dataset is available comparative dataset. Therefore, to facilitate visualization of gene expression across members of the Rosaceae family (*Rubus idaeus, Malus domestica, Prunus persica and Fragaria vesca*) I have created a novel approach to the eFP: a comparative viewer. An example of how this visualization is helpful could be to distinguish paralogs from homologs for hypothesizing function.

As stated earlier, the *Rubus idaeus* dataset encompasses samples collected from ovule, seed, ovary wall, and receptacle tissue at 0, 2, 4, 6, 9 and 12 days post anthesis (DPA) (Figure 9B). The *Malus domestica* dataset encompasses ovule, seed, ovary wall, hypanthium, and pedicel tissue at 0, 6, 12 and 20 DPA (Figure 9D). The *Prunus persica* dataset encompasses ovule, seed, ovary wall, hypanthium, and pedicel tissue at 0, 5, 12 and 18 DPA (Figure 9C). *Fragaria vesca*, is a previously published dataset and eFP viewer that contains embryo, ghost, seed, ovule, pith and cortex tissue at stages 1,2,3,4,5 [58] (Figure 9A).

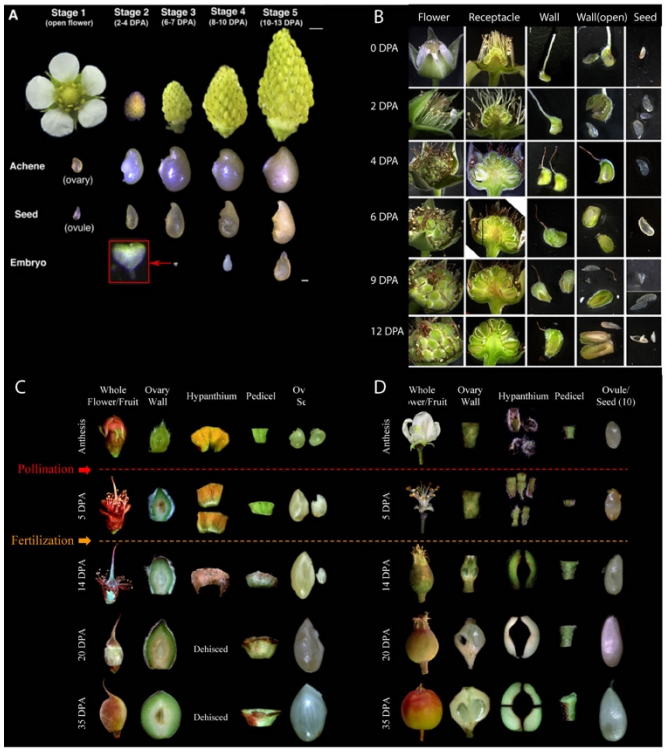


**Figure 9. Comparative Morphology of Rosaceae Fruit Tissue sampled for RNA-seq dataset**

These images were provided by Chunying Kang, Junhui Zhou and Kelsey Galimba. (A) *Fragaria vesca* dissected into pith/cortex (receptacle), ovule, embryo, ghost (seed coat), and wall tissues at five timepoints between 0 and 13 days post anthesis (DPA) (B) *Rubus idaeus* dissected into receptacle, seed, and wall at 0,2,4,6,9, and 12 DPA (C) *Prunus persica* dissected into wall, hypanthium, pedicel and seed at 0, 5, 14, 20 and 35 DPA (D) *Malus domestica* dissected into wall, hypanthium, pedicel and seed at 0, 5, 14, 20 and 35 DPA

The user is first presented with a form to submit a gene ID of interest and its corresponding species. Upon submission, the eFP image of the gene of interest appears alongside the eFP's of predetermined orthologs. The user can also download an excel file of raw transcripts per million (tpm) of each sample directly from the webpage. The scale of this eFP can either be log(tpm) or z-score which is able to be adjusted by clicking on the legend fixed to the top of the webpage.

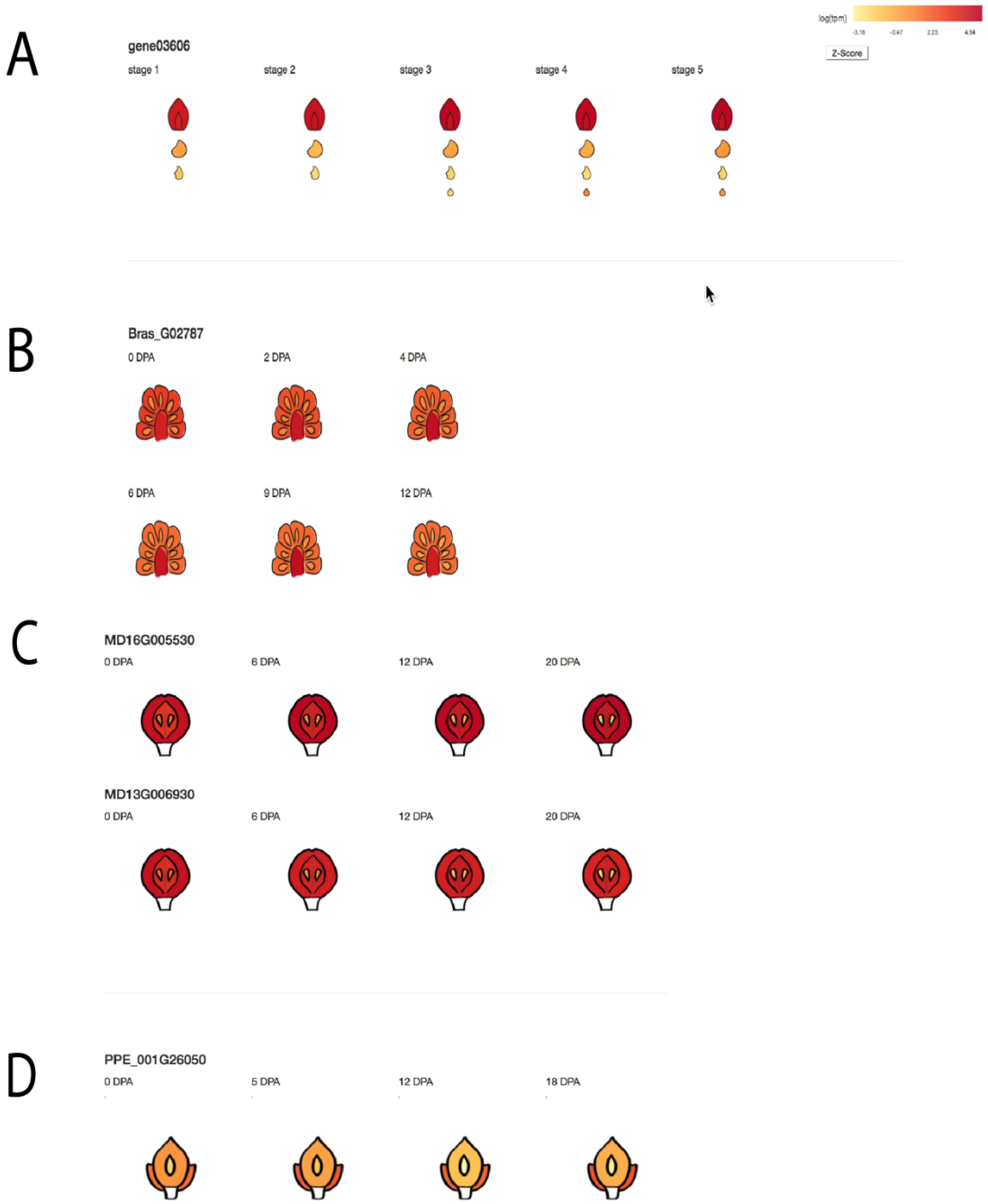


**Figure 10. Example user interface for comparative viewer of gene number gene03606**

(A) eFP depiction of *Fragaria vesca* user input gene03606 (B) eFP depiction of *Rubus idaeus* orthologs (C) eFP depiction of *Malus domestica* orthologs (D) eFP depiction of *Prunus persica* orthologs

The comparative browser is implemented identically to the single species viewer using a MEAN stack. After the user inputs a gene of interest, a JSON document is returned with the gene of interest, its orthologs, and corresponding expression levels for each gene. The comparative_viewer.js dynamically creates individual eFPs and vertically aligns them on the eFP page. Orthologs were pre-calculated using Orthofinder on default settings [54]. Expression values were obtained by using Salmon, a pseudoalignment tool, to all current transcriptomes of the appropriate species [55–57].

# Bibliography

1. Miller, Timothy, Libbey, Carl. Selectivity of Valent Herbicides in Red Raspberries [Internet]. 2007. Available from: http://ir4.rutgers.edu/FoodUse/PerfData/2310.pdf

2. Food and Agriculture Organization of the United Nations Statistics Division (FAOSTAT) [Internet]. Available from: http://www.fao.org/faostat/en/#data/QC/visualize

3. Red Raspberry Production [Internet]. Available from: https://extension.psu.edu/red-raspberry-production

4. VanBuren R, Bryant D, Bushakra JM, Vining KJ, Edger PP, Rowley ER, et al. The genome of black raspberry ( *Rubus occidentalis* ). The Plant Journal. 2016;87:535–47.

5. Graham J, Woodhead M. 24. Raspberries and Blackberries: The Genomics of Rubus. 2009.

6. Hyun TK, Lee S, Rim Y, Kumar R, Han X, Lee SY, et al. De-novo RNA Sequencing and Metabolite Profiling to Identify Genes Involved in Anthocyanin Biosynthesis in Korean Black Raspberry (Rubus coreanus Miquel). PLoS One [Internet]. 2014 [cited 2018 Oct 21];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3914977/

7. VanBuren R, Wai CM, Colle M, Wang J, Sullivan S, Bushakra JM, et al. A near complete, chromosome-scale assembly of the black raspberry (Rubus occidentalis) genome. Gigascience [Internet]. 2018 [cited 2018 Oct 21];7. Available from: https://academic.oup.com/gigascience/article/7/8/giy094/5069394

8. Daccord N, Celton J-M, Linsmith G, Becker C, Choisne N, Schijlen E, et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nature Genetics. Nature Publishing Group; 2017;49:1099–106.

9. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, et al. The genome of the domesticated apple (Malus × domestica Borkh.). Nature Genetics. Nature Publishing Group; 2010;42:833–9.

10. Ahmad R, Parfitt DE, Fass J, Ogundiwin E, Dhingra A, Gradziel TM, et al. Whole genome sequencing of peach (Prunus persica L.) for SNP identification and selection. BMC Genomics. BioMed Central; 2011;12:569.

11. Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, et al. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic

diversity, domestication and genome evolution. Nature Genetics. Nature Publishing Group; 2013;45:487–94.

12. Chagné D, Crowhurst RN, Pindo M, Thrimawithana A, Deng C, Ireland H, et al. The Draft Genome Sequence of European Pear (Pyrus communis L. 'Bartlett'). Tinker NA, editor. PLoS ONE. 2014;9:e92644–e92644.

13. Edger PP, VanBuren R, Colle M, Poorten TJ, Wai CM, Niederhuth CE, et al. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. GigaScience. Oxford University Press; 2018;7:1–7.

14. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, et al. The genome of woodland strawberry (Fragaria vesca). Nature Genetics. Nature Publishing Group; 2011;43:109–16.

15. Buti M, Moretto M, Barghini E, Mascagni F, Natali L, Brilli M, et al. The genome sequence and transcriptome of Potentilla micrantha and their compari ... [Internet]. GigaScience. 2018 [cited 2018 Feb 12]. Available from: http://gigadb.org/dataset/100407

16. Raymond O, Gouzy J, Just J, Badouin H, Verdenaud M, Lemainque A, et al. The Rosa genome provides new insights into the domestication of modern roses. Nature Genetics. Nature Publishing Group; 2018;50:772–7.

17. Hibrand Saint-Oyant L, Ruttink T, Hamama L, Kirov I, Lakhwani D, Zhou NN, et al. A high-quality genome sequence of Rosa chinensis to elucidate ornamental traits. Nature Plants. Nature Publishing Group; 2018;4:473–84.

18. Xiang Y, Huang C-H, Hu Y, Wen J, Li S, Yi T, et al. Evolution of Rosaceae Fruit Types Based on Nuclear Phylogeny in the Context of Geological Times and Genome Duplication. Molecular biology and evolution. Oxford University Press; 2017;34:262–81.

19. Ward JA, Bhangoo J, Fernández-Fernández F, Moore P, Swanson J, Viola R, et al. Saturated linkage map construction in Rubus idaeus using genotyping by sequencing and genome-independent imputation. BMC Genomics. BioMed Central; 2013;14:2.

20. Pryszcz LP, Gabaldón T. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic acids research. Oxford University Press; 2016;44:e113.

21. Huang S, Kang M, Xu A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. Bioinformatics. Oxford University Press; 2017;33:2577–9.

22. Bosi E, Donati B, Galardini M, Brunetti S, Sagot M-F, Lió P, et al. MeDuSa: a multi-draft based scaffolder. Bioinformatics. 2015;31:2443–51.

23. Pop M, Kosack DS, Salzberg SL. Hierarchical scaffolding with Bambus. Genome research. Cold Spring Harbor Laboratory Press; 2004;14:149–59.

24. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

25. Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. Oxford University Press; 2013;29:2669–77.

26. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:357–9.

27. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Wang J, editor. PLoS ONE. Public Library of Science; 2014;9:e112963.

28. Smit, AFA, Hubley R. RepeatModeler [Internet]. 2015 [cited 2018 Aug 5]. Available from: http://www.repeatmasker.org

29. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. http://www.repeatmasker.org; 2013.

30. Kent WJ. BLAT--the BLAST-like alignment tool. Genome research. Cold Spring Harbor Laboratory Press; 2002;12:656–64.

31. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, et al. ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biology. BioMed Central; 2015;16:3.

32. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England). Oxford University Press; 2014;30:2114–20.

33. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. BioMed Central; 2009;10:R25.

34. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. Bioinformatics. 2016;32:767–9.

35. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics (Oxford, England). Oxford University Press; 2014;30:1236–40.

36. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. NIH Public Access; 2011;29:644–52.

37. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic acids research. Oxford University Press; 2003;31:5654–66.

38. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. Oxford University Press; 2006;22:1658–9.

39. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. Oxford University Press; 2005;21:3674–6.

40. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology. BioMed Central; 2015;16:157.

41. Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, et al. The genome of the pear (Pyrus bretschneideri Rehd.). Genome research. Cold Spring Harbor Laboratory Press; 2013;23:396–408.

42. Alice LA, Campbell CS. Phylogeny of Rubus (rosaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. Am J Bot. 1999;86:81–97.

43. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. Oxford University Press; 2017;33:2938–40.

44. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome research. Cold Spring Harbor Laboratory Press; 2009;19:1639–45.

45. Long JA, Moan EI, Medford JI, Barton MK. A member of the KNOTTED class of homeodomain proteins encoded by the STM gene of Arabidopsis. Nature. 1996;379:66–9.

46. Ung N, Lal S, Smith HMS. The Role of PENNYWISE and POUND-FOOLISH in the Maintenance of the Shoot Apical Meristem in Arabidopsis1[W][OA]. Plant Physiol. 2011;156:605–14.

47. Hollender CA, Kang C, Darwish O, Geretz A, Matthews BF, Slovin J, et al. Floral Transcriptomes in Woodland Strawberry Uncover Developing Receptacle and Anther Gene Networks. Plant Physiology. 2014;165:1062–75.

48. Shahan R, Zawora C, Wight H, Sittmann J, Wang W, Mount SM, et al. Consensus Coexpression Network Analysis Identifies Key Regulators of Flower and Fruit Development in Wild Strawberry. Plant Physiology. 2018;178:202–16.

49. Kumaran MK, Bowman JL, Sundaresan V. YABBY Polarity Genes Mediate the Repression of KNOX Homeobox Genes in Arabidopsis. The Plant Cell. 2002;14:2761–70.

50. Vroemen CW, Mordhorst AP, Albrecht C, Kwaaitaal MACJ, de Vries SC. The CUP-SHAPED COTYLEDON3 Gene Is Required for Boundary and Shoot Meristem Formation in Arabidopsis. Plant Cell. 2003;15:1563–77.

51. Mantegazza O, Gregis V, Mendes MA, Morandini P, Alves-Ferreira M, Patreze CM, et al. Analysis of the arabidopsis REM gene family predicts functions during flower development. Ann Bot. 2014;114:1507–15.

52. Riechmann JL, Meyerowitz EM. The AP2/EREBP family of plant transcription factors. Biol Chem. 1998;379:633–46.

53. Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ. An "Electronic Fluorescent Pictograph" Browser for Exploring and Analyzing Large-Scale Biological Data Sets. PLoS One [Internet]. 2007 [cited 2019 Jan 16];2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1934936/

54. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology. 2015;16:157–157.

55. Daccord N, Celton J-M, Linsmith G, Becker C, Choisne N, Schijlen E, et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nature Genetics. 2017;49:1099–106.

56. Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, et al. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nature Genetics. 2013;45:487–94.

57. Edger PP, VanBuren R, Colle M, Poorten TJ, Wai CM, Niederhuth CE, et al. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. GigaScience. 2018;7:1–7.

58. Hawkins C, Caruana J, Li J, Zawora C, Darwish O, Wu J, et al. An eFP browser for visualizing strawberry fruit and flower transcriptomes. Horticulture Research. 2017;4:17029.