Running head: Insights into strawberry flower networks

Corresponding Author:
Zhongchi Liu
Department of Cell Biology and Molecular Genetics
University of Maryland
College Park, MD 20742
Email: zliu@umd.edu
Phone: (301) 405-1586
Fax: (301) 314-1248

Research Area:
Genes, Development and Evolution
Floral transcriptomes in woodland strawberry uncover developing receptacle and anther gene networks

Courtney A. Hollender¹,* Chunying Kang¹,* Omar Darwish², Aviva Geretz¹, Benjamin F. Matthews³, Janet Slovin⁴, Nadim Alkharouf², and Zhongchi Liu¹,#

1. Dept. of Cell Biology and Molecular Genetics
University of Maryland
College Park, MD 20817

2. Dept. of Computer and Information Sciences
Towson University
Towson, MD 21252

3. USDA-ARS
Soybean Genomics and Improvement
Beltsville, MD, 20705

4. USDA-ARS
Genetic Improvement of Fruits and Vegetables
Beltsville, MD, 20705

*: Equal contribution

Summary

The receptacle, which subtends floral organs, is a poorly studied yet important floral organ.

Genome wide gene expression analyses in strawberry flowers identified key regulatory genes of the developing receptacle.
Footnotes

This work was supported by National Science Foundation Grant MCB0923913 to ZL, JS, and NA, by Maryland MAES Hatch Project (MD-CBMG-0525 and MD-CBMG-0738) to ZL, and by Hokensen graduate fellowship to CH.

Present address of Courtney A. Hollender: USDA-ARS, Appalachian Fruit Research Station, Kearneysville, WV 25430

Corresponding author: Zhongchi Liu (zliu@umd.edu)
Abstract

Flowers are reproductive organs and precursors to fruits and seeds. While the basic tenets of the ABCE model of flower development are conserved in angiosperms, different flowering plants exhibit different and sometimes unique characteristics. A distinct feature of strawberry flowers is the development of several hundreds of individual apocarpous (unfused) carpels. These individual carpels are arranged in a spiral pattern on the subtending stem tip, the receptacle. Therefore, the receptacle is an integral part of the strawberry flower and is of significant agronomic importance, being the precursor to strawberry fruit. Taking advantage of next-generation-sequencing and Laser Capture Microdissection (LCM), we generated different tissue- and stage- transcriptomic profiling of woodland strawberry flower development. Using pair-wise comparisons and weighted gene co-expression network analysis (WGCNA), we identified modules of co-expressed genes and hub genes of tissue-specific networks. Of particular importance is the discovery of a developing receptacle-specific module exhibiting similar molecular features to those of young floral meristems. The strawberry homologs of number of meristem regulators including LOST MERISTEM and WUSCHEL are identified as hub genes operating in the developing receptacle network. Further, almost 25% of the F-box genes in the genome are transiently induced in developing anthers at the meiosis stage, indicating active protein degradation. Together, this work provides important insights into the molecular networks underlying strawberry’s unique reproductive developmental processes. This extensive floral transcriptome dataset is publicly available and can be readily queried at the project website, serving as an important genomic resource for the plant biology research community.
**Introduction**

Flowers are reproductive organs and precursors to fruits and seeds. Past work in model systems, mainly *Arabidopsis* and *Antirrhinum*, led to the formulation of the ABCE model of flower development (Coen and Meyerowitz, 1991; Krizek and Fletcher, 2005). Although the essential tenets of the ABCE model for flower development are conserved in angiosperms, each family of flowering plants can exhibit different and sometimes unique characteristics. In *Fragaria vesca* (*F. vesca*), a diploid wild strawberry, flowers differ from those of *Arabidopsis* in several aspects (Hollender et al., 2012). The most striking distinction is the formation of several hundreds of independent carpels arranged in a spiral pattern on an enlarged stem tip, the receptacle. These carpels are apocarpous, meaning they do not fuse with one another. Molecular communication from the fertilized carpels to the subtending receptacle triggers the receptacle growth into the edible flesh fruit (Nitsch, 1950; Kang et al., 2013). Therefore, flower development in strawberry is intimately tied to strawberry fruit development.

As part of developing *F. vesca* into a model plant for the *Rosaceae* family (Shulaev et al., 2011), we previously described in detail the morphological and developmental progression of *F. vesca* flower development (Hollender et al. 2012), dividing flower development from stage 1, floral primordial initiation, to stage 12, the completion of flower development. This work laid the foundation for the current work in detailing transcriptomes of various floral tissues at different stages. Tissue- and stage- specific transcriptomes enable in-depth molecular studies of flower development and have a wide applicability across the economically important members of the *Rosaceae* family including strawberry, apple, peach, raspberry, and rose.

Taking advantage of next-generation-sequencing and laser capture microdissection (LCM), two dimensional (tissue and stage) transcriptome data for flower development in *F. vesca* were generated. Using k-means clustering and weighted gene co-expression network analysis (WGCNA), tissue- and stage-specific gene clusters and network modules are identified. A number of key meristem regulatory genes, including *FveLOM* (*LOST MERISTEM*) and *FveWUS* (*WUSCHEL*), are identified as hub genes operating in the developing receptacle network. In addition, we found a transient surge of transcript levels of a large number of F-Box (FBX) genes in anthers of stage 9 flowers, at which stage microspore mother cells are undergoing meiosis.
Network analysis of the anther module at meiosis stage identified highly connected hub genes that are homologous to genes with roles in protein degradation. Finally, transcripts of putative *F. vesca* ABCE floral homeotic genes are shown to accumulate in expected floral organs, suggesting conserved functions of the ABCE genes. Together, the genome scale gene expression profiling described here places the foundation for further biochemical and functional analysis of strawberry flower development. The data is publicly available and can be readily queried at the project website “Strawberry Genome Resources (SGD)” (http://bioinformatics.towson.edu/Strawberry) (Darwish et al., 2013).

**Results**

**Global analysis of RNA-seq data**

RNA-seq data were generated from 15 different floral samples at different developmental stages (Fig. 1A; Fig. S1A, B). Stages 1 to 7 floral samples were isolated by LCM (Fig. S1B; Materials and Methods), while stages 7 to 12 floral samples were isolated by hand dissection (HD) under a microscope. Microspores from stage 10 flowers were also isolated by LCM (Fig. S1B). All samples are named “tissue_stage”, where the stage refers to the flower developmental stage defined in Hollender et al. (2012). For LCM isolated samples, the average number of raw reads per library was about 27 million (Data S1); 30 to 40% of these raw reads mapped to the CDS, while 50 to 66% of these raw reads mapped to the gene (200 bp upstream+exons+introns+200 bp downstream). Since LCM-derived sample preparation utilized a strategy other than polyA selection to capture non-ribosomal RNAs (see Materials and Methods), the higher percentage of mapped reads to the gene than to CDS may reflect noncoding RNAs from introns, 5’UTR or 3’UTR. For HD-isolated samples, the average number of raw reads per library was 31 million (Data S1); 61 to 72% of these raw reads mapped to CDS while slightly higher percentages (71 to 76%) mapped to the gene. Mapped reads against CDS were used in all subsequent analyses.

Normalized read counts (RPKM) for each gene were calculated and genes with RPKM lower than 0.5 were considered “not expressed” (Kang et al. 2013). The *F. vesca* genome was predicted to have 34,809 genes (Shulaev et al., 2011); 34,527 genes were found to be expressed in all floral tissues combined (Data S1). Pollen has the least number of expressed genes (at 11,548) (Fig. 1B; Data S1). Surprisingly, LCM samples have more expressed genes than HD
samples by about 10,000, yet LCM samples have more genes expressed at a low level (1 to 10 RPKM; Fig. 1B). Two factors may have contributed to this difference. First, LCM-derived RNAs were processed differently including a non-ribosomal RNA capture step and a cDNA amplification step (see Materials and Methods). This may result in detection of genes expressed at levels not detectable by standard RNA-seq method used for preparing HD samples. Second, LCM samples are mostly younger tissues with more complex transcriptomes. To distinguish these two possibilities, we isolated stage 10 carpels using HD as well as LCM. The carpel_10 HD sample was processed as other HD samples, while the carpel_10 LCM sample was processed as other LCM samples. The carpel_10 LCM sample showed about 10,000 more expressed genes than the carpel_10 HD sample (Fig. 1B; Data S1), indicating that the difference in the number of expressed genes is due to tissue isolation and RNA processing methods.

The overall relatedness of transcriptomes of different tissues was shown by the cluster dendrogram done for the HD and LCM samples. For HD samples (Fig. 1C), the two biological replicates (A and B) show good correlation, and mature pollen transcriptome is closely related to that of stage 12 anthers, which contain mature pollen. For the most part, different stage anthers are similar to each other and different stage carpels are similar to each other, except that the mature carpel (stage 12) transcriptome is more similar to that of the leaf. For LCM samples (Fig. 1D), stages 6-7 receptacle and stages 1-4 flowers cluster closely to each other, supporting similar molecular networks (see later). Biological replicates of LCM samples are not as closely paired as the HD replicates. The laser capture from sections of a smaller number of flowers may result in higher variability between biological replicates.

**Identification of gametophyte and sporophyte genes**

Pairwise comparisons were made with either DESeq or Baggerley’s test in CLC Genomics Workbench (CLC) (Fig. S1A, S2, Data S2). For the four LCM samples, the transcriptomes of stages 5-6 perianth, stages 6-7 anthers, and stages 6-7 receptacles were respectively compared with that of stages 1-4 flowers using the Baggerley’s test in CLC. For the HD carpels, each carpel stage is compared with its immediate earlier stage using DESeq. Likewise, the transcriptomes of HD anthers were compared between successive stages using DESeq. Up- or down-regulated genes with 2-fold cutoff were identified for each pairwise comparison (Data S2).
The numbers of overlapping up- or down-regulated genes between each comparison for LCM samples (Fig. S2A), HD carpels (Fig. S2B), and HD anthers (Fig. S2C) are shown by the Venn diagram. Few up- or down-regulated genes are shared among the comparisons. For instance, not a single gene is up-regulated continuously in all HD anther stages (stages 9 through 12), and only 4 genes are continuously up-regulated in all three HD carpel stages (stages 9, 10, and 12).

We identified gametophyte-specific genes by determining genes in common between microspore_10 preferential and pollen preferential gene lists. To facilitate comparisons between LCM and HD samples, the original expression values were further normalized using the quantile method within CLC, leading to Normalized RPKM (NRPKM). Comparison between stage 10 anther and stage 10 microspore genes using NRPKM as input in CLC resulted in 3498 genes preferentially expressed in stage 10 anthers and 2257 genes preferentially expressed in stage 10 microspores (with 2-fold cutoff) (Data S3). Likewise, stage 12 anther data was compared with that of pollen by DESeq using read counts as input, which identified 4287 genes preferentially expressed in stage 12 anthers and 3177 genes preferentially expressed in pollen (Data S3). 862 genes were found in common between stage 10 anther preferential and stage 12 anther preferential lists and thus considered “male sporophyte-abundant genes” (Fig. 2A; Data S3). GO analysis revealed enriched metabolic processes (small molecule, organic acid, cellular ketone, and others) among these male sporophyte-abundant genes (Fig. 2B). Similarly, 119 genes were identified as “male gametophyte-abundant genes” due to their presence in both stage 10 microspore-preferential and pollen-preferential lists (Fig. 2C; Data S3).

Transcription factor GaMYB (gene23946), ADIPOR-like receptor (gene31121), and CNG channel alpha3 (gene13563) are among the highest expressed “male-gametophyte-abundant genes”. To examine if these genes are also abundantly expression in the male organ of other plant species, we were able to identify the Arabidopsis and rice homologs of GaMYB (MYB101/AT2G32460 and Os01g0812000) and found that the pMYB101::GUS is specifically expressed in the pollen grain and the Arabidopsis triple mutant, myb101 myb97 myb120 exhibits defects in pollen function (Liang et al., 2013). There is no detailed study of the ADIPOR-like receptor homologs in Arabidopsis or rice, nor could we reliably identify a homolog of the F. vesca CNG channel alpha3 in Arabidopsis or rice.
Identification of temporal or spatial expression trends across *F. vesca* floral transcriptomes

Thirty-five K-means clusters (containing 9043 genes) with the most obvious tissue and stage-specific expression trends were selected (see Materials and Methods). Of these, 19 clusters representing 3,931 genes are shown in Fig 3A. Carpel-specific gene clusters (C1 and C22) and anther-specific gene clusters (C2, C3, C4, C5, C7, and C25) were identified (Fig. 3A). Pollen has a highly unique transcriptome; almost 50% of the genes with tissue specific expression were found in the pollen (C23, C28, C31, C32, C33, C34, C35). Clusters C2, C3, and C7 are shared between microspore_10 and anthers in floral stages 9 to 11, as microspores are contained within these anthers. Similarly, six clusters (C25, C28, C30, C31, C34, and C35) are shared between mature pollen and stage 12 anthers, which house the pollen.

Clusters with similar expression trends were further combined into 10 superclusters (Fig. 3A; Data S4). MapMan bins (Thimm et al., 2004) were used to identify specific categories of genes or pathways enriched in specific superclusters (Fig. 3C). The top enriched MapMan bin for the carpel-specific supercluster (supercluster 1) is “DNA synthesis/chromatin structure”. The top MapMan bin for stage 9 anthers (supercluster 4) is “protein degradation”, consistent with the up-regulation of a large number of FBX genes in anthers of stage 9 flowers (discussed later). The top enriched MapMan bins for pollen (supercluster 9) are “regulation of transcription” and “cell wall” including synthesis, degradation, and modification.

Among the 3,931 genes that make up the above 10 superclusters (Fig. 3A), 161 are transcription factors belonging to 29 families (Fig. 3B; Data S4). These transcription factors show distinct stage or tissue-specific expression patterns mirroring each of the ten superclusters to which they belong. The dynamic and stage- or tissue-specific expression patterns of these transcription factors probably reflect the key functions they play. For example, the young carpel-specific (C1) cluster consists of five bHLH genes, among which is *FveHECATE1* (*gene23830*), a known regulator of carpel development in *Arabidopsis* (Gremski et al., 2007). Three *WERKY*s and six *Ethylene Response Factor* genes with potential roles in stress responses are among the pollen-specific transcription factors (Data S4).
The remaining 16 clusters of the 35 clusters representing 5112 genes are interesting in that they show either LCM-specific or HD-specific expression (Fig. S3A, B; Data S4). For LCM-specific clusters, MapMan bins of “DNA replication”, “photosynthesis”, and “regulation of transcription” are enriched (Fig. S3C), suggesting actively dividing cells as well as green tissue (sepal) development. In contrast, metabolic processes are more enriched for HD-specific clusters including “amino acid metabolism”, “nucleotide metabolism”, “minor and major CHO metabolism”, “lipid metabolism”, and “redox” (Fig. S3C). Therefore, to certain degrees, the LCM- and HD-specific clusters reflect the younger vs. older tissues and their respective activities.

Co-expression network analysis with WGCNA
An alternative analysis tool named “weighted gene co-expression network analysis (WGCNA)” was adopted (Langfelder and Horvath, 2008). WGCNA is a systems biology approach aimed at understanding networks instead of individual genes. In this study, co-expression networks are constructed on the basis of pairwise correlations between genes in their common expression trends across all sampled tissues. Modules are defined as clusters of highly interconnected genes; genes within the same cluster have high correlation coefficients among one another. This analysis resulted in 23 distinct modules (labeled by different colors) shown by the dendrogram (Fig. 4A), in which each tree branch constitutes a module, and each leaf in the branch is one gene (Fig. 4A; Data S5). The module eigengene is the first principal component of a given module and can be considered a representative of the module’s gene expression profile. The 23 module eigengenes for the 23 distinct modules were each correlated with distinct tissue types due to eigengenes’ tissue-specific expression profiles (Fig. 4B). Notably, 12 out of 23 co-expression modules are comprised of genes that are highly expressed in a single floral tissue type (r > 0.8, P < 10⁻³) (Fig. 4B; Data S5). Therefore, each of these 12 modules identifies (or correlates with) a specific tissue or stage cluster of genes. For example, the blue module identifies 4584 pollen-specific genes (Fig. 4A, B), and the turquoise module identifies 5791 LCM-specific genes (Fig. 4A, B). This is consistent with earlier analysis using K-means clustering, where 1954 pollen-specific genes (Fig. 3A) and 3824 LCM-specific genes (Fig. S3A) were found. Despite different algorithms, WGCNA and K-means clustering both identified pollen- and LCM-specific clusters as the two largest clusters/modules.
Of particular interest is the identification of a developing receptacle-specific module (light yellow, 123 genes) as well as a young floral (stages 1-4) module (dark red; 86 genes) (Fig. 4A, B). These two small modules were not detected by the earlier K-means clustering method (Fig. 3A). Fig. 5A and B show the eigengene expression for the dark red (flower_1-4) module and light yellow (receptacle_6-7) module respectively. A heatmap showing the relative NRPKM of each gene from the receptacle module and the stages 1-4 flower module revealed that many of the receptacle module genes are weakly expressed in flower stages 1-4 (Fig. 5C; Data S6), indicating molecular similarities between the developing receptacle and young flowers. This close relationship between receptacle_6-7 and flower_1-4 was also revealed in earlier global analysis (Fig. 1D).

In addition, some genes are co-expressed in the carpel_7-8 and the receptacle_6-7 modules (Fig. 5B, C). At floral stages 7-8, small carpel primordia cover the outer surface of the developing receptacle; hand dissection of the carpel primordia off the receptacle surface may be imprecise and hence may include a small amount of receptacle tissues. The three highest co-expressed transcription factor genes in carpel_7-8 and receptacle_6-7 are knotted1-like (gene31590), zinc finger protein (gene20982), and bHLH93 (gene05560) (Data S6), which are however also highly expressed in flower_1-4, making tissue contamination a less likely explanation. Alternatively, these co-expressed genes in carpel_7-8 and receptacle_6-7 are biologically relevant; they may underscore the developmental relationship between these two floral tissues.

WGCNA can also be employed to construct gene networks, in which each node represents a gene and the connecting lines (edges) between genes represent co-expression correlations. Hub genes are those that show most connections in the network. In the receptacle module network (Fig. 5D; DataS7), 27 of the 111 genes encode transcription factors. Strikingly, many of the hub genes are meristem regulators (Fig. 5D) including three WUS/WOX genes (FveWUS1/gene30464; FveWOX3/gene28935; FveWOX4/gene14133), one FveAIL3/gene20828, two GRAS (FveLOM3/gene31749; FveSCL18/gene01184), one FveLFY/gene30750, three NAC (FveCUC3/gene31474; FveCUC2/gene30749; FveANAC018/gene34006), and five TALE/BELL/Knotted1-like genes (such as FveSTM/gene19507) (Data S6). Most of these
transcription factor genes are up-regulated in both developing receptacle and stages 1-4 flowers (Data S6) and may underlie meristematic activities operating in the developing receptacle as well as stages1-4 flowers.

The hub gene with the highest edge number (81 edges) is FveLOM3 (gene31749), a member of the GRAS transcription factor family. In Arabidopsis, triple mutants of lom1; lom2; lom3 exhibited arrested shoot apical meristem, axillary meristem, and root meristem (Schulze et al., 2010; Engstrom et al., 2011), thus highlighting FveLOM3 as a possible key regulator of the developing receptacle. Other highly connected hub genes include a B3 domain transcription factor FveVRN1 (gene29444), a Myb transcription factor FveAPL (gene13545), and FveWUS1 (gene30464). Interestingly, a second WUS homolog in F. vesca, FveWUS2 (gene14621) was identified. While highly similar to FveWUS1 in amino acid sequence alignment (Fig. S4A, B), FveWUS2 is expressed in developing anthers between stages 7-8 and 11 (Fig. S4C).

**Anther stage 9-specific module transiently expresses a large number of FBX genes**

Pairwise comparisons (Fig. S2) revealed 1453 up-regulated genes in stage 9 anthers when compared with stages 7-8 anthers. Among these 1453 up-regulated genes, 211 (14%) encode FBX-containing genes (Fig. 6A; Data S2 and S8). Since the F. vesca genome has 820 FBX genes representing 2% of the genome, the much higher percentage (14%) of FBX genes among the stage 9 anther up-regulated genes suggests a heightened protein degradation activity in meiosis stage anthers. This also underlies the enriched MapMan category “protein degradation” in the anther_9 super-cluster (Fig. 3A, C). Interestingly, as revealed by the heatmap (Fig. 6B), this surge of FBX gene expression in stage 9 anthers is transient; many of them are subsequently down-regulated at stage 10 and 11 anthers. Specifically, 95 and 77 FBX genes are down-regulated in stage 10 and stage 11 anthers, respectively (Fig. 6A, B).

In total, 296 FBX genes were differentially expressed, either up or down, between successive anther stages (Fig. 6B; Data S8). The top six most abundant FBX subfamilies among the 296 differentially expressed FBX genes were DUF295, FBD, Kelch repeat, LRR, FBA-1, and FBA-3 (Fig. 6C). The DUF295 class is particularly enriched among the anther_9 up- and anther_10 down-regulated FBX genes (Fig. 6C). WGCNA analysis was carried out on the stage 9 anther
module to identify hub genes with many connecting edges (Fig. 6D-E; Data S8, S9). Five genes have an edge number higher than 200, and four of these five appear to have a function in protein degradation (Data S8; Fig. 6E). Gene11150 codes for an F-box/kelch-repeat protein, gene32754 encodes a WD repeat-containing protein 42A, which interacts with Cul4-Ddb1 E3 ligase and is a nucleocytoplasmic shuttling protein (Wu et al., 2012), gene07916 is a homolog of 26S protease regulatory subunit 4, gene01189 encodes histone 3.3, and gene20570 encodes an aspartic proteinase (precursor of nepenthesin-1). The identity of these hub genes underscores active protein degradation activities during stage 9 anther development.

Has a similar phenomenon been observed in other plant species? Developmental transcriptomic profiling of anther development was previously reported in *Brassica rapa* (Dong et al., 2013), *Oryza sativa* (Fujita et al., 2010), and *Zea mays* (Ma et al., 2008), all of which employed microarrays. We applied the same statistical standard using one-way ANOVA analysis in MeV (p-value < 0.001) and detected only 3 up-regulated genes in the F2 stage of *Brassica rapa* whole flower bud collection, 87 up-regulated genes in the Q stage of maize anther collection, and 458 up-regulated probes in M2 and M3 stage rice anthers (see Materials and Methods). These equivalent comparisons at meiotic anther stages yielded a much lower number of DE genes than *F. vesca* stage 9 anthers, which have 1453 up-regulated genes (Fig. S2). Therefore, significant data variation could be contributed by different platforms and tissue collection/comparison methods.

We further analyzed the rice transcriptomic data because it gave the highest number of DE genes as described above. Z-score normalized values of the 458 up-regulated probes at M2 and/or M3 anther stages were used to generate the heatmap in MeV (Fig. S5A; Data S10). An M2-specific cluster showed transient induction at the M2 stage. In contrast, most M3 stage expressed genes are also expressed at the P1 stage, the stage immediately following the M3 stage. Among the 644 F-box genes in the rice genome, 356 are on the Affymetrix chip, representing 1.5% of the high quality probes on the chip. Only 6 F-box genes are among the 338 M2-M3 up-regulated genes (1.8%). Therefore, we did not observe an enrichment of F-box genes among up-regulated genes in rice M2-M3 anthers. Loraine et al. (2013) compared gene expression profiling using microarray vs. RNA-seq and showed that microarray could only detect highly expressed genes.
Because the majority of our up-regulated F-box genes at stage 9 *F. vesca* anthers were not expressed at a high level (Fig. S5C), they would not have been identified by earlier microarray studies, thus illustrating the power of RNA-seq over microarray in identifying gene expression trends.

**F. vesca ABCE genes**

The ABCE classes of genes are the most well known and well studied flower developmental genes (Krizek and Fletcher, 2005), and their RNAs are expressed in highly stereotypic pattern in a flower. We identified *F. vesca* homologs of the A, B, C and E genes (Fig. S6) and showed that their expression patterns are consistent with the predictions of the ABCE model. The A gene *FveAP1* is highly expressed in stages 1-4 flowers, receptacle, and perianth (Fig. 7; Data S11). The class B genes, *FvePla, FvePlb* and *FveAP3*, are expressed strongly in anthers and weakly in stages 1-4 flowers and stages 5-6 perianth, both of which contain only a small fraction of petal primordia. The class C gene *FveAG* is highly expressed in carpels and anthers, as well as the receptacle, from which carpel primordia arise, consistent with our earlier *in situ* hybridization data (Hollender *et al.* 2012). Class E gene *FveSEP3* and *FveSEP4* are expressed in all floral tissues including receptacle with *FveSEP3* at a much higher level than *FveSEP4*. *FveSEP1* and *FveSEP-LIKE1* are expressed at low levels only in stages 1-4 flowers and stages 6-7 receptacle. None of the A, B, C, and E genes is expressed in leaves or seedlings except *FveAP2* (Fig. 7), which is expressed in all tissues and may have a broader function (Maes *et al.*, 2001; Würschum *et al.*, 2006).

**Discussion**

Strawberry is an important specialty crop with unique floral structure and accessory fruit. The diploid strawberry, *F. vesca*, is an excellent model for the cultivated strawberry and other Rosaceae plants. Using LCM and HD, different *F. vesca* floral organs and tissues were isolated and RNA-seq was performed. Two different bioinformatic methods were used to analyze the RNA-seq data, revealing several novel insights including (1) the identification of a developing receptacle-specific module and hub transcription factors, (2) the discovery of a large network of FBX genes at stage 9 anthers and hub genes, and (3) the molecular similarity between the transcriptome of young floral meristem and that of developing receptacle. The systems approach
in data mining via WGCNA was particularly fruitful in identifying tissue-specific modules and important hub genes for future biochemical and functional studies.

Developing receptacle and young floral bud are similar in their transcriptomes

By botanical definition the receptacle of flowering plants is the area at the stem tip that bears flower organs or groups of flowers. Some examples of enlarged receptacles in economically important plants are those subtending sunflower head, the seed-containing chamber of a lotus flower, and the fleshy fruits of strawberry. Although most receptacles are not consumed as fruit some are as in the case of strawberry. However, almost nothing is known about the molecular underpinnings of this important plant organ, despite tremendous progress in understanding the molecular mechanism of standard floral organ development.

While the botanical definition of a receptacle describes a static structure subtending the flowers or floral organs, the developing receptacle, as analyzed here, is a dynamic and growing structure. The apex of the developing receptacle dome in the stage 6-7 *F. vesca* flower is still giving rise to carpel primordia (Hollender et al., 2012) and thus is the terminating floral meristem, but the basal part of the receptacle dome has finished giving rise to carpel primordia and has become the receptacle. Hence, the term “developing receptacle” is used to describe this dynamic structure at floral stages 6-7. During subsequent stages the receptacle continues to enlarge in size even after the receptacle apex terminates its floral meristem activity (Hollender et al., 2012).

Our work reported here provided an initial molecular insight into the development of this important plant organ. We found that the transcriptome of a developing receptacle is most similar to that of young floral (stages 1 to 4) primordia with overlapping sets of genes specifically expressed in both tissues. Hub genes, such as *FveLOM* and *FveWUS*, were identified by their large numbers of connecting edges within the receptacle network. Future work such as in situ hybridization showing expression zones of these meristem regulators may further shed light on the unique aspects of the developing receptacle. Such knowledge should aid future manipulations of selected meristem regulators to prolong meristematic activities in the developing receptacle so as to achieve desirable fruit size and shape.
Anthers of floral stage 9 transiently express a large number of FBX genes

During the course of analyzing transcriptomes of strawberry flowers, we noted a significant enrichment of MapMan bin “Protein Degradation” in the stage 9 anther supercluster (Fig. 3C). Further examination of the anther_9 up-regulated gene list (Data S2) identified a large number of genes, including FBX genes, with predicted functions in protein degradation. Hence we further analyzed these anther_9 expressed FBX genes (Fig. 6A, B), which revealed a transient up-regulation at stage 9 anthers (Fig. 6A, B). Our earlier work indicated that at early stage 9, microspore mother cells enter meiosis, the resulting tetrads are tightly connected in locules. The middle layer has degenerated. At late stage 9, meiosis is completed, the tetrads of microspores become loose in each locule, but tapetum degeneration is not yet initiated (Hollender et al. 2012). Hence, the induction of FBX genes in anthers of stage 9 flowers is unlikely related to tapetum degeneration. Possibly, these FBX genes are required for meiosis or for middle layer degeneration.

FBX proteins were first characterized as a subunit of the Skp1-Cullin-F-box (SCF) ubiquitin-ligase complex (Bai et al., 1996; Kipreos and Pagano, 2000) and can be classified into 19 groups based on the presence of additional domains such as DUF (domain of unknown function), LRR, kelch repeats, and PPR (Gagne et al., 2002; Kuroda et al., 2002). Many of the up-regulated anther_9 FBX proteins belong to the DUF295 family (Fig. 6C). DUF295 family FBX proteins play diverse functions from posttranslational regulation of the SET domain polycomb group protein (Jeong et al., 2011) to regulating ascorbic acid biosynthesis (Zhang et al., 2009). While we do not yet know the function of these FBX genes in the stage 9 anther, this is an important observation that will form the basis of future functional studies.

We investigated if the observed phenomenon was also reported in other plant species. Transcriptomic studies have been conducted in pollen, sperm, and anthers in diverse plant species including Arabidopsis, rice, brassica, maize, and cotton (Honys and Twell, 2004; Ma et al., 2007; Ma et al., 2008; Chen et al., 2010; Fujita et al., 2010; Deveshwar et al., 2011; Yang et al., 2011; Ma et al., 2012; Dong et al., 2013; Wuest et al., 2013). While some studies were focused on tapetum and male gametophytes, others compared anthers of male sterile mutants with anthers of wild type. One study noted enrichment of GO term “protein metabolism” in post-
meiotic stage anthers and attributed it to tapetum degeneration (Deveshwar et al. 2011). Another study showed enriched FBX genes in pollen and unicellular microspores (Wuest et al., 2013). None reported enriched FBX gene expression at early stage anther development as observed in this study. Nevertheless, different FBX gene families may play important functions at different developmental stages.

We subsequently focused our analysis on microarray-based transcriptomic data from three prior studies that are most similar to our work (Ma et al., 2008; Fujita et al., 2010; Dong et al., 2013). We found that the rice transcriptome showed a reasonable number (338) of up-regulated genes at meiotic anthers, a stage similar to *F. vesca* stage 9 anthers, yet we could not detect an enrichment of FBX genes among up-regulated genes in rice meiotic anthers (Fig. S5B). Since RNA-seq is significantly more sensitive than microarray in detecting genes expressed at low levels (Loraine et al., 2013), microarray-based transcriptomes may have missed the FBX genes. Alternatively, the observed surge of FBX genes in anther_9 could be specific to strawberry and less relevant to the general regulatory mechanisms of anther development. Further comparative genomic studies that employ identical gene expression profiling method are required to distinguish these possibilities.

**Optimization of the LCM procedure and the data analyses**

We adopted the LCM technique to isolate young floral tissues too difficult to access via hand dissection. LCM posed several challenges due to extremely low RNA yields (in the range of 5-10ng) and poorer RNA quality due to tissue fixation. Several improvements were made in tissue fixation and handling (Fig. S7) including the adoption of acetone fixation (Ohtsu et al., 2007), the minimization of section ribbon expansion time (in H2O) on slides, and the use of only freshly dewaxed slides for laser capture.

The NuGEN Ovation RNA-Seq System V2 was employed to amplify LCM-derived RNA. This system employs the Ribo-SPIA technology, so the cDNA amplification is initiated at the 3’ end as well as randomly throughout the whole transcriptome, allowing amplification of poly A+ as well as non-polyadenylated transcripts and overcoming some of the challenges of poorer RNA quality. The RNA processing method described above for LCM samples may be more sensitive
in detecting genes expressed at low levels as well as noncoding RNAs, thus contributing to the apparent increase in expressed genes found in the LCM samples (Fig. 1B). To normalize the variation in the number of expressed genes, mapped reads from HD and LCM were normalized by quantile method in CLC, resulting in Normalized RPKM or NRPKM. This normalization appeared to work well when LCM samples are compared with HD samples.

Conclusions
The genome-scaled approach utilized here to investigate strawberry flower development is surprisingly informative in uncovering novel networks and hub genes. The work provides important insights into the molecular events underlying strawberries’s unique reproductive developmental process as well as valuable resources to the plant biology community in areas of flower development, gamete formation, and fruit development. Our work highlights the effectiveness of WGCNA analysis tool in uncovering small clusters of genes and networks and calls for cautions in comparative genomics when the data are generated by different methods or techniques.

Materials and Methods
Floral tissue isolation
A 7th generation inbred line of *F. vesca*, Yellow Wonder 5AF7 (Slovin et al., 2009), was grown in a growth chamber with 12 h light at 25°C and 12 h dark at 20°C. Anthers and carpels were hand-dissected from stage 7-12 flowers under a dissecting microscope. For pollen, mature anthers were placed in a petri dish containing drierite for a few hours to release pollen, which was washed in dH2O and collected by centrifugation. Each sample has two biological replicates and each replicate contains tissues from 2-12 flowers.

Floral tissues (stages 1-7) were isolated by LCM (Fig. S1B). Before LCM, tissues were fixed in ice-cold 100% acetone (Ohtsu *et al.* 2007), exchanged for 50:50 acetone:Hemo-De (Scientific Safety Solvents), and then exchanged for 100% Hemo-De. Fixation vials were placed at 60°C and paraplast Plus® chips (Leica Microsystems; www.leica-microsystems.com) were added gradually to the Hemo-De to slowly infiltrate the tissue with wax. Subsequently, samples were incubated in 100% molten wax for 2-3 days with several changes of fresh 100% wax. After
solidifying the wax in boats at room temperature, 10-12μm thick sections were prepared using a rotary microtome, and sections were floated on drops of RNAse-free water on Leica PEN-Membrane 2.0 μm slides (Cat # 11505158), which were placed on a slide warmer at 40°C for 15-30 minutes. Slides were dewaxed based on the Leica protocol, and tissues were laser captured on the same day. An average of ~2 x 10^6 μm^2 of tissue (ranging from 1 x 10^6 to 8.7 x 10^6 μm^2) were captured with a Leica Laser Microdissection Microscope-ASLMD, collected in the cap of a 0.2 ml PCR tube containing the Extraction Buffer from the Arcturus® PicoPure® RNA isolation kit (Life Technologies; www.lifetechnologies.com), and stored at -80°C. Each LCM-derived sample contains pooled sections from 2 to 6 flowers.

**RNA isolation and sequencing**

Total RNA from LCM or HD samples (except anther and pollen) was extracted with the Arcturus® PicoPure® RNA isolation kit in conjunction with Turbo RNase-free DNase (Life Technologies). Total RNA from stage 9-12 anthers and from pollen was extracted using RNeasy Plant Mini Kit (Qiagen, www.qiagen.com) in conjunction with RNase-free DNase (Qiagen). For HD samples, a minimum of 150 ng total RNA was sent to the Weill Cornell Medical College Genomics Resources Core, where poly-A selection and Illumina Truseq RNA-seq library preparation were conducted. Four to six bar-coded libraries were pooled and sequenced in one lane with Single Read Clustering and 51 Cycles on the Illumina HiSeq2000.

For LCM-derived samples, approximately 2 x 10^6 μm^2 of LCM collected tissues typically yielded 10-30 ng total RNA. Between ~3.5 ng to 38 ng total RNA was converted to cDNA and then amplified, using Ovation® RNA-Seq System V2 (Nugen Technologies; http://www.nugen.com/nugen/). Next, RNA-seq libraries were made using NuGEN’s Encore® NGS Multiplex Library System I. Four barcoded LCM sample libraries were pooled in equal ratios for sequencing in one lane at Cornell Weill Medical College as described above.

**Global and differential gene expression analysis of RNA-Seq data**

51bp single end raw reads were mapped against the *F. vesca* CDS and genes (www.rosaceae.org/species/fragaria/fragaria_vesca/genome_v1.0) (Shulaev et al. 2011) using CLC Genomics Workbench (CLC bio; http://www.clcbio.com/products/clc-genomics-
workbench) (Anders and Huber, 2010). 7 nucleotides from 5’ terminus were trimmed due to low quality scores. The RNA-seq Analysis tool in CLC was used for mapping allowing 3 mismatches. Mapping statistics are shown in Data S1. To compare LCM samples with HD samples, the original expression value was further normalized by quantile method within CLC, the resulting value is the Normalized RPKM or NRPKM.

For Fig. 1C-D, the dendrograms were plotted using the variance stabilizing transformed read counts (built in the DESeq package) by Euclidean distance measure. To identify differentially expressed genes (Fig. S2 and Fig. 2), pairwise comparisons between LCM samples or between LCM and HD samples were made by Baggerley’s test built in CLC with NRPKM values as input. DESeq version 1.10.1 was used to compare between HD samples following DESeq’s vignette; mapped read counts without normalization were used as input. The cutoff for pairwise comparisons was set at fold change > 2, FDR \( p \)-value < 0.01.

**K-means clustering analysis**

Before identifying specific gene expression clusters (Fig. 3), 24,836 differentially expressed (DE) genes were identified by first comparing (1) all the anther samples (2) all the carpel samples, (3) all the LCM samples, and (4) different tissues from the same stage, using DESeq by multiple factor design with mapped read counts as input. The resulting DE genes (adjp < 0.0001) from above comparisons were combined. Then, genes expressed at NRPKM < 1 were removed and genes whose NRPKM in leaf or seedling contributed to > 40% of total reads across all tissues and stages were also removed. After these filtering steps, 23,424 DE genes remain.

The “relative NRPKM value” (\( \frac{\text{NRPKM}_{\text{geneX in tissueX}}}{\text{averaged NRPKM}_{\text{geneX across all tissues}}} \)) was calculated for each of these 23,424 DE genes to facilitate the identification of genes with low absolute NRPKM values (such as transcription factors) but with significant expression changes in specific tissues or stages. K-means clustering with Euclidean distance in MeV2.8 (Saeed et al., 2003) yielded 100 clusters based on inputs of log2 “relative NRPKM values”. 11275 genes that reside in those clusters with obvious tissue- or stage-specific expression trends were selected and then used to generate fifty K-means clusters in MeV4.8. Finally, 9043 genes that reside in clusters with even more obvious tissue- or stage-specific
expression trends were selected to make the final 35 K-means clusters. Only 19 out of the 35 clusters are shown in Fig. 3, and the remaining 16 clusters are shown in Fig. S3. Transcription factors were extracted from the respective clusters (Fig. 3B and Fig S3B) based on the *F. vesca* TF table (Kang et al. 2013). The log2 “relative NRPKM” of each TF was used to yield heatmaps in MeV4.8 (Fig. 3B and Fig. S3B).

**Assignment of MapMan bins and GO terms**

MapMan bins of *F. vesca* genome were assigned by the Mercator pipeline for automated sequence annotation (http://mapman.gabipd.org/web/guest/app/mercator) (Thimm et al. 2004); 100 was used as the BLAST_CUTOFF. GO ontologies were assigned using Blast2GO (Conesa et al., 2005), and the GO annotation file for ~20,003 *F. vesca* genes was previously shown (Kang et al. 2013). GO enrichment was derived with Fisher’s exact test and a cutoff of FDR < 0.05; the genome annotation file described above was used as the reference. Only GO terms for Biological Process are shown.

**Identification of F-box protein**

1185 annotated F-box genes were obtained by word search from strawberry genome (www.rosaceae.org/species/fragaria/fragaria_vesca genome_v1.0). The protein sequences were batch searched in Pfam server (http://pfam.sanger.ac.uk/search) with default settings against only Pfam-A families (E-value = 1.0). Any protein with a hit to PF00646--F-box, PF12937--F-box-like, or PF13013--F-box-like_2 was defined as a putative F-box protein. 815 out of 1185 genes remained. These 815 F-box proteins were grouped into subfamilies based on their domain organization, including DUF295 (PF03478), FBD (PF08387), FBA-1(PF07734), FBA-3 (PF08268), and the Kelch-repeat subfamily (both Kelch_3 (PF13415) and Kelch_6 (PF13964). LRR_4 (PF12799), LRR_6 (PF13516), and LRR_8 (PF13855) were combined as LRR subfamily.

**Analysis of the rice anther microarray data**

Microarray data were downloaded from Gene Expression Omnibus in NCBI (accession number GSE12579 for *Zea mays*, GSE47665 for *Brassica rapa*, and GSE14304/dataset GDS3957 for *Oryza sativa*). For the rice dataset, DE genes were identified by one-way ANOVA between
samples M1, M2 and M3 in MeV_4.8.1 (p-value < 0.001). Rice version_7.0 annotation files were downloaded from “MSU Rice Genome Annotation Project’ website. Identification of F-box proteins was based on the presence of interpro F-box domain. Gene names corresponding to probes in the microarray were obtained by blastn (match length > 50bp, p-value < 1.00E-15, number of targets <= 3). Although 458 probes showed M2-M3 up-regulated expression (Fig. S5A), after removing low-quality probes (defined as those hybridizing to more than three target genes) and blastn searches, 338 genes were identified as the M2-M3 up-regulated genes (Fig. S5B). The normalized value shown in Fig. S5A was obtained in MeV based on the following formula: Value = [(Value) – Mean(Row)]/[Standard deviation(Row)].

**Phylogenetic tree construction of ABCE genes**

Protein sequences of *Arabidopsis ABCE* genes were BLASTed against translated protein sequences of the strawberry genome (hybrid gene models version 1.0) at the Plant and Food Research server (strawberrygenome.org). Candidate strawberry homologs were BLASTed against NCBI flowering plant protein database for verification. Additional validations included the identification of conserved protein domains (MADS and AP2 domains) and intron/exon structure. The protein sequence alignment was made using Clustal Omega ([http://www.ebi.ac.uk/Tools/msa/clustalo](http://www.ebi.ac.uk/Tools/msa/clustalo)). The unrooted phylogenetic tree (Fig. S6) was constructed using MEGA 5.05 ([http://www.megasoftware.net/](http://www.megasoftware.net/)) with the neighbor–joining statistical method and bootstrap analysis (1000 replicates).

**Gene network construction and visualization**

Co-expression networks were constructed using the WGCNA (v1.29) package in R (Langfelder and Horvath, 2008). Among the 23,424 genes used in K-means clustering analysis, 21,517 genes with averaged NRPKM from two replicates higher than 2 were used for the WGCNA unsigned co-expression network analysis; the average NRPKM was imported into WGCNA. The modules were obtained using the automatic network construction function “blockwiseModules” with default settings except that the power = 16, TOMType is “signed”, minModuleSize = 30, and mergeCutHeight = 0.25. The eigengene value was calculated for each module and used to test the association with each tissue type. The total connectivity and intramodular connectivity (function “softConnectivity”), kME, and kME-p-value were calculated for the 20954 genes,
which were clustered into 23 tissue-specific modules. The other 563 genes were outliers (grey module) and not shown in Data S6. The networks were visualized using Cytoscape v.3.0.0.

**Accession Number**

Illumina reads of all 30 samples (two biological replicates of 15 different floral tissues/stages) have been submitted to the Sequence Read Archive (SRA) at NCBI (http://www.ncbi.nlm.nih.gov/sra). The accession number is SRP035308.

**Supplemental Materials**

Figure S1. Summary of tissues and stages profiled in this study

Figure S2. Venn diagrams of differentially expressed genes derived from pairwise comparisons

Figure S3. Gene clusters specifically expressed in LCM or HD samples

Figure S4. *FveWUS1-2* sequence alignment with *Arabidopsis WUS* and their RNA-seq expression

Figure S5. Rice meiotic stage anthers did not show increased expression of large numbers of F-box genes

Figure S6. Phylogenetic tree of floral homeotic genes

Figure S7. Electrophoresis histograms of RNAs isolated from *F. vesca* floral tissues showing RNA quality after different treatments.

Data S1. RNA-seq statistics and expressed gene number

Data S2. Pairwise comparisons between different tissues

Data S3. Male gametophyte and sporophyte abundant genes

Data S4. K-means clustering results

Data S5. Data of WGCNA analysis

Data S6. Flower_1-4 and receptacle_6-7 network analysis

Data S7. Light yellow module network file

Data S8. F-box genes in anther_9 network

Data S9. Pink module network file

Data S10. Results from analysis of rice anther microarray data

Data S11. *F. vesca* ABCE genes
Author contributions
CH, CK, AG, and JS performed the experiments; CK, ZL, OD and NA analyzed the data; BM provided essential equipment and training; ZL, CH, and CK wrote the manuscript; ZL, CK, CH, BM, NK, and JS commented on the manuscript.

Acknowledgements
We thank Dr. Chris Dardick for help with the initial data analysis.
Literature Cited


Figure legends

Figure 1. Floral tissue collection and global analysis of the floral transcriptomes
(A) A diagram illustrating the four floral tissue types, perianth, anther, carpel, and developing receptacle, collected for the RNA-seq. (B) Number of expressed genes and respective expression levels (in each sample type based on the average RPKM of two biological replicates. (C) Cluster dendrogram showing global relationship between biological replicates and among different HD samples. The y-axis is the degree of variance. (D) Cluster dendrogram showing global relationship among different LCM samples.

Figure 2. Identification and analysis of male gametophyte and sporophyte genes.
(A) Diagrams illustrating bi-lobed anther filled with microspores. Red color indicates preferential expression of sporophyte (left)- or gametophyte (right)-abundant genes. (B) Top ten enriched GO terms of the 862 sporophyte-abundant genes. (C) Hierarchical clustering of 119 male-gametophyte abundant genes based on their log2NRPKM values.

Figure 3. K-means clustering reveals unique tissue and/or stage-specific expression trends.
(A) 19 clusters representing 3931 genes are shown with distinct stage and tissue-specific expression patterns. Averaged log2 “relative NRPKM” of all the genes in each cluster was used to generate the heatmap. These clusters are further grouped into ten superclusters shown to the right. (B) Clustering based on log2 “relative NRPKM” of 161 transcription factors. (C) Enriched MapMan bins shown to the right for each of the ten superclusters. Significant enrichment is indicated by low p-values.

Figure 4. Weighted Gene Network Co-expression Analysis (WGCNA) of genes in floral tissues.
(A) Hierarchical cluster tree showing co-expression modules identified by WGCNA. Each leaf in the tree is one gene. The major tree branches constitute 23 modules labeled by different colors. (B) Module-tissue association. Each row corresponds to a module. The number of genes in each module is indicated on the left. Each column corresponds to a specific tissue. The color of each cell at the row/column intersection indicates correlation coefficient between the module and the
tissue type. A high degree of correlation between a specific module and the tissue type is indicated by dark red.

**Figure 5. Developing receptacle-specific genes and networks**

(A) Eigengene expression profile for the dark red (stage 1-4 flower) module in different tissues. y-axis indicates the value of the module eigengene; x-axis indicates sample type. The number of genes in the module is indicated in parenthesis. (B) Eigengene expression profile for the light yellow (stages 6-7 receptacle) module in different tissues. (C) Heatmap showing the relative NRPKM of each gene from the dark red (stage 1-4 flower) module and the light yellow (stages 6-7 receptacle) module. (D) The correlation network of light yellow (stages 6-7 receptacle) module. 111 genes with the edge weight higher than 0.1 are visualized by Cytoscape. 27 transcription factors are shown by larger circles.

**Figure 6. Over 200 FBX genes are transiently induced in anthers of stage 9 flowers**

(A) Number of FBX genes that are up (red)- or down (blue)-regulated at each anther stage, when each is compared with the immediate earlier stage. Number of expected up (grey) - or down (patterned grey)-regulated genes is derived by multiplying number of up or down regulated genes at the respective stage by the percent FBX genes in the genome (2.35%). (B) Hierarchical clustering of 296 differentially expressed FBX genes at the four anther stages. Log2(fold change) compared with the immediate earlier stage is shown. (C) Percentage of specific classes of FBX genes among the anther_9 up- and anther_10 down-regulated genes. Percent expected FBX class (grey) is derived by multiplying the number of anther_9 up or anther_10 down genes with the percent respective class of FBX genes in the genome. (D) Eigengene expression profile for the pink (stage 9 anther) module. (E) The correlation network of the pink (stage 9 anther) module with the edge weight higher than 0.6 visualized by Cytoscape. 37 FBX genes are indicated by larger circles among the 232 genes shown in the network.

**Figure 7. Expression of A, B, C, E class genes in strawberry flowers**

Heatmap showing gene expression levels of *F. vesca* A, B, C, E class genes. Numbers are the NRPKM value for each gene in respective tissues. Left most column indicates the A, B, C, and E classes. The right most column indicates the name of each gene, *FveAP1* (*gene04562*),
FveAP2 (gene23876), FveAP3 (gene14896), FvePla (gene11267), FvePlb (gene11268), FveAG (gene24852), FveSEP1 (gene04229), FveSEP3 (gene07201), FveSEP4 (gene26118), and FveSEPLIKE1 (gene04563).