

RESEARCH

Open Access

# RNA-Seq analysis and annotation of a draft blueberry genome assembly identifies candidate genes involved in fruit ripening, biosynthesis of bioactive compounds, and stage-specific alternative splicing

Vikas Gupta<sup>1,2†</sup>, April D Estrada<sup>1†</sup>, Ivory Blakley<sup>1</sup>, Rob Reid<sup>1</sup>, Ketan Patel<sup>1,4</sup>, Mason D Meyer<sup>1</sup>, Stig Uggerhøj Andersen<sup>2</sup>, Allan F Brown<sup>3</sup>, Mary Ann Lila<sup>3</sup> and Ann E Loraine<sup>1\*</sup>

## Abstract

**Background:** Blueberries are a rich source of antioxidants and other beneficial compounds that can protect against disease. Identifying genes involved in synthesis of bioactive compounds could enable the breeding of berry varieties with enhanced health benefits.

**Results:** Toward this end, we annotated a previously sequenced draft blueberry genome assembly using RNA-Seq data from five stages of berry fruit development and ripening. Genome-guided assembly of RNA-Seq read alignments combined with output from *ab initio* gene finders produced around 60,000 gene models, of which more than half were similar to proteins from other species, typically the grape *Vitis vinifera*. Comparison of gene models to the PlantCyc database of metabolic pathway enzymes identified candidate genes involved in synthesis of bioactive compounds, including bixin, an apocarotenoid with potential disease-fighting properties, and defense-related cyanogenic glycosides, which are toxic. Cyanogenic glycoside (CG) biosynthetic enzymes were highly expressed in green fruit, and a candidate CG detoxification enzyme was up-regulated during fruit ripening. Candidate genes for ethylene, anthocyanin, and 400 other biosynthetic pathways were also identified. Homology-based annotation using Blast2GO and InterPro assigned Gene Ontology terms to around 15,000 genes. RNA-Seq expression profiling showed that blueberry growth, maturation, and ripening involve dynamic gene expression changes, including coordinated up- and down-regulation of metabolic pathway enzymes and transcriptional regulators. Analysis of RNA-seq alignments identified developmentally regulated alternative splicing, promoter use, and 3' end formation.

**Conclusions:** We report genome sequence, gene models, functional annotations, and RNA-Seq expression data that provide an important new resource enabling high throughput studies in blueberry.

**Keywords:** RNA-Seq, Genome, Transcriptome, Blueberry, Alternative splicing, Fruit ripening, Metabolic pathways

\* Correspondence: aloraine@uncc.edu

†Equal contributors

<sup>1</sup>Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, North Carolina Research Campus, Kannapolis, NC 28081, USA

Full list of author information is available at the end of the article

## Background

A diet rich in blueberries can help protect against diabetes [1], cardiovascular disease, and age-related cognitive decline [2,3]. Molecular or biochemical mechanisms underlying these positive health benefits are not known, but most research has thus far focused on the antioxidant and anti-inflammatory properties of polyphenolic phytochemicals that accumulate as the fruit ripen. Blueberry fruit are an especially rich source of polyphenolic anthocyanin pigments, which give blueberries their distinctive color. Of these, malvidin, delphinidin, and peonidin are the most abundant by weight [4]. The relative abundance of anthocyanin species can differ between genotypes [5,6], and *in vivo* research has shown that different anthocyanins affect biological systems in different ways [7-9], suggesting that berry varieties may offer distinct health benefits. Blueberries also contain relatively large amounts of quercetin [4], another polyphenolic that may have beneficial effects in Alzheimer's disease [10] and inflammation-related disorders [11]. Berries may also contain other as-yet undiscovered beneficial phytochemicals that could interact with anthocyanins or other compounds to potentiate biological efficacy [12,6]. Genomic studies that catalog the full genetic repertoire of blueberry could enable greater understanding of bioactive compounds, a necessary step toward developing new varieties bred for health benefits.

Blueberries are in the Cyanococcus section of family Ericaceae, genus *Vaccinium*, which also includes cranberry (*V. macrocarpon*), lingonberry (*V. vitis-idaea*), and more than 400 other species [13]. Commercially harvested blueberry species in North America include lowbush (wild) blueberry *Vaccinium angustifolium*, a low, spreading shrub grown in managed stands in the Northern US and Canada, and the highbush blueberry species *Vaccinium corymbosum* and *Vaccinium ashei*, which are larger shrubs grown in orderly rows in orchards that require annual pruning to maintain productivity. Of the two highbush species, *V. corymbosum* is the most widely grown, while *V. ashei* is grown solely in the Southern US. *V. corymbosum* was first domesticated in the early 20th century by US Department of Agriculture scientist Fredrick Coville working with New Jersey farmer Elizabeth White, who recruited local pickers to locate wild berry plants with unusually large fruit. Coville's breeding these early wild selections produced varieties suitable for commercial production, some of which are still grown today. Both lowbush and *V. corymbosum* highbush blueberries are deciduous and require a period of low temperatures during the winter season to induce flowering the following spring. To expand the range where highbush blueberries can be grown commercially, breeding programs have selected varieties with reduced chilling requirement, leading to development of

sub-varieties called 'southern highbush' because they require fewer days of colder temperatures to trigger flowering. Ploidy levels of berry species range from diploid to hexaploid, and most varieties of highbush berry contain genetic material introduced from diverse genotypes and species, including *V. darrowii* Camp (evergreen blueberry) and *V. arboreum* Mar. (sparkleberry), as well as rabbiteye and lowbush blueberry. Although there is a great diversity across varieties, highbush blueberry plants within the same cultivar are highly uniform, as all are clones propagated from a single selection. Thus sequences collected from individuals from the same cultivar are expected to be highly homogenous with few differences between individuals.

Estimates based on flow cytometry predict that a haploid blueberry genome is around 600 million bases, five times the size of the *Arabidopsis thaliana* genome [14]. In a related study, a draft genome assembly of a diploid northern highbush blueberry was generated using HiSeq Illumina reads [15]; the unassembled sequences are available from the Short Read Archive under accession SRA053499. This draft assembly consists of 225,479 contigs organized into 13,757 scaffolds with an N50 scaffold size of 145 kb, meaning that at least half of the sequence data is organized into scaffolds of 145 kb or larger. Plant genes are typically smaller than 2 kb, and intergenic regions are often smaller, which means that a 145 kb or larger contig could accommodate 50 or more genes. Although the genome assembly is still a work in progress, its large N50 make this draft assembly an important new resource for RNA-Seq analysis and gene discovery in blueberry.

To date, blueberry improvement efforts have focused on agronomic traits, such as ability to withstand mechanical harvesting, or consumer-focused traits, such as berry size, flavor, and mouth feel. Due to rising consumer interest in the health-protective effects of blueberries and other fruits and vegetables, breeding for nutritional and health-protective qualities may become practical in the near future. Breeding a more healthful berry will require more complete knowledge of genes encoding enzymes of secondary metabolism as well as their putative regulators. Toward this end, we performed high-throughput transcriptome sequencing (RNA-Seq) and differential gene expression analysis of five stages of berry development and ripening. Genome data, RNA-Seq expression profiles, and functional annotations have been made publicly available and will provide an important new resource for interpretation of high-throughput data from blueberry species.

## Data description

### Berry collection and RNA extraction

Blueberry samples were collected from the field from 4- or 5-year-old blueberry plants growing at the North

Carolina Department of Agriculture Piedmont Research Station. Plants were labeled by row and position within the row; for example, plant 2-41 occupied position 41 within row 2. All samples intended for RNA extraction were flash-frozen on liquid nitrogen in the field immediately after collection and stored at -80°C until use. For RNA extraction, whole berry samples were ground to powder in a mortar and pestle with liquid nitrogen and total RNA was extracted using the Spectrum Total Plant RNA Kit (Sigma, St. Louis, USA). Extracted RNAs were treated with DNase I prior to library construction using RNAase-Free DNase (catalog number 79254) from Qiagen.

#### **454 library construction and sequencing for May 2009 samples**

Two libraries were prepared from samples of green and ripe fruit, respectively, from plants of the O'Neal variety of southern highbush blueberry (*V. corymbosum*). The green fruit library was prepared from a mix of unripe, green fruits of varying sizes harvested on 18 May 2009 from plant 2-41. The ripe fruit library was prepared from ripe fruits harvested on 15 June 2009 from plants 2-40, 2-41, and 2-42, also of the O'Neal variety. The libraries for sequencing were constructed using the SMART PCR cDNA synthesis kit from CloneTech, Mountain View, USA. The 3' and 5' primers used in first strand cDNA synthesis were aagcagtggatcaacgcagagtact(30)VN and aagcagtggatcaacgcagagtacgctggg, respectively, where V was a, g, or c and N was any nucleotide. The products of first strand cDNA synthesis were amplified using a poly(A) disruption PCR primer designed to introduce non-A bases in the poly(A) tail region of the cDNA, because homopolymeric sequences are difficult to sequence using the 454 technology. The poly(A) disruption primer sequence was attctagaggccgagggcggcgcacatgt(4)gtct(4)gttctgt(3)ct(4)VN, where numbers in parentheses indicate the number of times the preceding base appeared in the sequence and V was a, g, or c. The sequence of the 5' primer used for second strand cDNA synthesis was aagcagtggatcaacgcagagt. The two libraries were sequenced in two sectors of the same plate on a 454-GS FLX Titanium sequencer (454 Life Sciences, Roche Diagnostics, USA) at the David H. Murdock Research Institute. Sequence data are available from the Short Read Archive under accession SRP039977.

#### **454 library construction and sequencing for May 2010 samples**

Green and ripe berries were harvested from O'Neal variety plant 2-42 on 29 April 2010 and 26 May 2010, respectively. For sequencing and library construction, samples of total RNA were sent to the North Carolina State University Genome Sciences Laboratory (GSL). Each sample was used to synthesize two libraries, which

were sequenced on the same plate of a 454-GS FLX Titanium sequencer. Libraries were synthesized at the GSL following the protocol reported previously [16]. Sequence data are available from the Short Read Archive under accession SRP039977.

#### **Berry collection, library synthesis, and sequencing of berry development samples**

Berries from five stages were selected from three plants (3-33, 2-41, and 2-42) of the O'Neal variety of southern highbush blueberry *V. corymbosum*. We designated the five stages as pads, cups, green, pink, and ripe. The 'pads' and 'cups' stages corresponded approximately to stages S1/S2 (pads) and S3/S4 (cups) described in [17]. Green fruit were fully rounded green berries, pink berries were partially pigmented but still firm, and ripe berries were fully colored and soft. Samples were collected during the growing season of 2011. Pads were collected on 4 April, cups on 19 April, mature green fruit on 28 April, pink fruit on 20 May, and ripe fruit on 2 June. Following RNA extraction and DNase I treatment (as described in the previous section) libraries were synthesized using the TruSeq A kit (catalog number FC-121-1001) from Illumina (San Diego, USA) following the manufacturer's instructions. Libraries were synthesized using different TruSeq adapters to allow multiplexing, combined and then sequenced in three lanes with five libraries per lane using a HiSeq sequencer from Illumina. Sequence data are available from the Short Read Archive under accession SRP039977.

#### **Berry collection, library synthesis, and sequencing of berry cultivars**

Fully ripe and green berries from four berry cultivars (Pamlico, Lenoir, O'Neal and Ozark Blue) were harvested in 2009 from plants growing at the Piedmont Research Station (same field as above) and frozen on liquid nitrogen in the field. RNA was extracted as described in the preceding section. Libraries were synthesized using the mRNA-Seq Sample Preparation kit (catalog number RS-930-1001) from Illumina following manufacturer instructions. Libraries were sequenced using paired-end, 76 cycle sequencing at the UNC Chapel Hill Lineberger Cancer Research Center on a GAIIx sequencer from Illumina, San Diego, USA. Sequence data are available from the Short Read Archive under accession SRP039971.

#### **Sequence processing and alignment**

Prior to alignment, all sequences were trimmed to remove low-quality bases at the 5' and 3' ends of sequences. Single-end Hiseq reads (100 bp) from the berry fruit development series were trimmed to 85 bases to remove lower quality bases. Five bases were trimmed from the 3' end and 10 bases were trimmed from the 5' end

of each read using the FASTX-Toolkit from Galaxy [18]. Similarly, the 76 bases long paired-end GAIx sequences were trimmed to 61 bases by removing ten bases on the 5' end and three bases on the 3' end of each sequence. For 454-generated sequences, ten bases were removed on the 5' end only. The Illumina sequences were aligned onto the blueberry draft genome assembly [15] using TopHat2 [19] and Bowtie2 [20] using default parameters, except for the maximum intron size parameter, which was set to 6,000 bases consistent with typical intron size distributions for plant genes. 454 sequences were aligned onto the reference genome sequence using GMAP [21] with default parameters except for intron length, which was set to 6,000 bp.

#### Gene model generation, filtering, and protein sequence assignment

Cufflinks [22] was used to generate transcript models using paired-end and single-end Illumina sequences alignments with maximum intron set to 6,000 bases. Three *ab initio* gene-finding programs (Augustus [23], GlimmerHMM [24], and GeneMark [25]) were used to generate gene models from genomic sequence. *Arabidopsis*-trained parameters were used for GlimmerHMM and default parameters were used for Augustus and GeneMark. Because many of the resulting *ab initio* and Cufflinks-based gene models covered the same genomic regions, a step-wise filtering protocol (Additional file 1: Figure S1) was applied to reduce redundancy in the final gene set. First, all genes generated by Cufflinks were selected for inclusion in the final gene set, because these were predicted from the RNA-Seq expression data and thus the level of evidence supporting them was high. Any *ab initio* gene finder-predicted gene that overlapped one of the Cufflinks-predicted genes was eliminated from the candidate gene list. Next, all remaining candidate genes that were predicted by GeneMark were considered. If a candidate GeneMark gene had homology to a known protein (by BLASTX) or overlapped with an expressed sequence alignment, it was added to the final gene set and, as before, overlapping gene models were removed from the remaining candidate gene set. The process was repeated for genes predicted by Glimmer, Augustus, and GMAP. The selection order of priority for *ab initio* gene prediction programs was based on visual inspection of predicted gene models and consistency with alignments of full-length blueberry sequences available from GenBank. Protein coding sequences were detected in the gene models using the TAU program [26]. Gene models were formatted into a BED-detail (BED14) file (V\_corymbosum\_scaffold\_May\_2013.bed) in which field 4 listed the transcript (gene model) name, field 13 listed the gene name, and field 14 listed the name of the program that generated the gene model. Note that many genes

predicted by Cufflinks were associated with multiple transcripts, typically products of alternative splicing. Thus, in the case of Cufflinks-predicted genes, the gene field contained a name such as CUFF.11 while the gene model field contained a name such as CUFF.11.1, the first reported transcript generated from gene CUFF.11. Gene models in bed format were added to the project bitbucket repository [27] in subdirectory GeneModelAnalysis/results.

#### Functional annotation of gene models using BLASTX against the NCBI non-redundant protein database

Sequences of spliced transcripts were searched against the non-redundant protein database from NCBI (nr) using a database that was downloaded in June 2013. Virtual cDNA sequences obtained from the gene model annotations were searched against the nr database using BLASTX. The resulting matches were used for annotation if the e-value was  $10^{-4}$  or better, the alignment length was at least 30 amino acids, and the percent identity was 30% or higher. For each gene model with at least one high-quality match meeting these criteria, an annotation string was generated containing the GenBank identifier and FASTA header for the best-matching sequence, information about the BLASTX-generated alignment, and the program that was used to generate the gene model. The annotation string was transferred onto the corresponding blueberry gene model by replacing field 14 in the BED-detail file described above. Field 14 data for gene models that did not have high-quality matches meeting the criteria described above were not modified. The modified gene model file was added to the bitbucket repository subdirectory titled BlastxAnalysis/results as V\_corymbosum\_scaffold\_May\_2013\_wDescr.bed.gz.

#### Functional annotation of gene models using Blast2GO

Blast2Go version 2.7.0 (Build 05122013) was obtained via Java Web Start and used to associate Gene Ontology (GO) terms and Enzyme Commission (EC) numbers with individual gene models from blueberry. BLASTP and InterProScan search results were loaded into Blast2GO and the Blast2GO function Annotation > Perform Annotation Step menu was used to perform GO annotation. BLASTP results were obtained by searching predicted blueberry protein sequences against the nr protein database using an e-value cutoff of  $10^{-3}$  and reporting a maximum of five 'hit' sequences per query. IPRScan version 4.8 was run using default parameter settings. Databases current as of July 2013 were searched. MySQL databases ('b2\_g\_sep13') from Blast2Go.com were downloaded and installed on a local server to enable faster processing. Blast2GO results were saved in plain text format (Additional file 2) and in Blast2GO format and added to the bitbucket in a subdirectory named Blast2GO.



### Functional annotation of gene models using BLASTX against PlantCyc enzymes

FASTA-format sequence files containing amino acid sequences for enzymes in Plant Metabolic Network species-specific databases (e.g., AraCyc and GrapeCyc) were downloaded in July 2013 and combined into one database. The blueberry cDNAs were used as queries in a BLASTX search of the combined database. Hits to PlantCyc enzymes were then filtered so that hits with at least 60% subject coverage, 45% identity or higher, and e-value of 0.001 or less were retained. Details of how this was done are explained in Markdown file MakingAnnotationFiles.Rmd in the bitbucket repository for the paper. The best hit for each blueberry gene model was identified by e-value and used to generate annotation text, which was inserted into field 14 of the gene models bed file and saved to subdirectory PlantCyc/results as V\_corymosum\_scaffold\_May2013\_wDescrPwy.bed. To enable keyword searching in Integrated Genome Browser, this annotation text included the best matching PlantCyc enzyme and a list of PlantCyc pathway identifiers when available.

### Analyses

#### Building and filtering blueberry gene models

To characterize gene expression in blueberries, Illumina and 454/Roche sequencing of green and ripe blueberry cDNA was done, including a developmental time course experiment in which three biological replicates of five berry fruit stages were sequenced using 100-base, single-end Illumina HiSeq sequencing. Table 1 summarizes the sequencing strategies used and amount of sequence obtained, which totaled around 800 million sequence RNA-Seq 'reads' and more than 75 billion bases. Most sequence data were from the developmental time course experiment, which surveyed berry samples collected from three individuals of the O'Neal cultivar, an early ripening variety of southern highbush blueberry that is widely grown in North Carolina and other southern US states. The second largest RNA-Seq dataset was from paired-end, 75-base-pair Illumina sequencing of ripe and green berries collected from the O'Neal, Arlen, Lenoir and Pamlico southern highbush cultivars. Analysis of genetic differences between the cultivars will be described

elsewhere, but the data were included here to enable a more complete transcriptome assembly and analysis. In addition, two full plates of 454 sequencing of ripe and unripe berries from the O'Neal cultivar were done using berries collected during spring and summer of 2009 and 2010. Because of the abundance of data and availability of easy-to-use software for generating gene models from Illumina-based RNA-Seq data, we used the Illumina sequence data to generate a genome-guided transcriptome assembly and reserved the 454 data for gene model validation and assessment.

To generate berry gene models from the RNA-Seq data, the Illumina sequence reads were aligned onto the May 2013 reference blueberry genome using the spliced alignment program TopHat [19] and then merged into gene models using Cufflinks [28]. This step produced 64,666 transcript models representing 57,925 genes. Of the multi-exon genes, 24% were predicted to generate multiple transcript variants due to alternative promoter use or alternative splicing. To assess how realistic this level of transcript variation was, we compared the frequency of alternative transcripts in blueberry with alternative transcription rates in *Arabidopsis* and soybean, using the genome annotations and assembly releases that were available in July 2013 for those species. The alternative transcription rates among multi-exon genes for soybean and *Arabidopsis* were 25% and 26% respectively; these rates were similar to blueberry, indicating that the frequency of transcript variation found in the blueberry gene models was reasonable and not likely to be an artifact of incorrect transcript model assembly.

Because the Cufflinks gene models were based on berry fruit RNA-Seq data, some genes that were expressed primarily in other sample types (e.g., roots) might be missed. To complement the Cufflinks-generated gene models, *ab initio* gene prediction programs were used to generate additional gene models from the reference genomic sequence (Additional file 1: Figure S1A). The *ab initio* gene finding programs generated more than 185,000 gene models, of which 75% overlapped with genes identified in the RNA-Seq data by Cufflinks and therefore were likely to be redundant. To eliminate duplicates and create a non-redundant collection of blueberry gene models, a stepwise filtering protocol was applied

**Table 1 Summary of sequencing strategies and sequences obtained**

Sequencing strategy	Samples	Number of reads (millions)	Bases (billions)
Single-end Illumina, 100 bp reads	Five fruit stages from O'Neal cultivar	485	48.6
Paired-end Illumina, 76 bp reads	Ripe, unripe berries from O'Neal, Lenoir, Ozark blue, and Pamlico cultivars	377	28.3
454	Ripe and unripe berries from 2009, 2010 harvests, O'Neal cultivar	2.16	0.73
Total		864.16	77.63

(Additional file 1: Figure S1B) that retained gene models based on their relative level of evidence. To start with, all gene models based on the aligned RNA-Seq data were added to the non-redundant set. Next, gene models with homology to known proteins were selected. Following this step, gene models were kept according to which *ab initio* gene finding program produced them. We found through manual inspection in Integrated Genome Browser (IGB) that some programs produced more realistic models when visually compared with the 454 alignments, so we configured the filtering pipeline based on these observations. Note that this final filtering step could be further optimized through systematic comparative evaluation of *ab initio* models to RNA-Seq based gene models, but due to the limitations of time and data availability, we did not do this. Ultimately, the final non-redundant set of gene models included 70,581 gene models and 63,840 genes, of which most were based on the RNA-Seq data (Table 2).

Next, an open reading frame (ORF)-finding program (TAU [26]) was used to identify and annotate a conceptual translation for each gene model. Altogether, there were 57,079 gene models from 51,515 genes that could be annotated with a protein sequence of at least 30 residues, the expected lower range of protein sizes based on *Arabidopsis* annotations. The remaining transcript models for which no longer protein sequence could be predicted had a broad range of sizes and exon numbers, ranging from short, single-exon genes to longer models with multiple exons, suggesting that in many cases, the gene models were simply incomplete or the sequence itself contained insertions, deletions or other problems that prevented conceptual translation. Another possibility was that many of these genes were non-coding genes whose primary products were RNA and not protein. A third possibility was that many of these were actually pseudogenes. An analysis of *Arabidopsis* RNA-Seq data [29] found that many non-coding genes, including both pseudogenes and non-protein coding genes, were detectable in libraries that were prepared using the same protocol as with the blueberry RNA-Seq libraries, and so these latter two possibilities may have been correct in many cases.

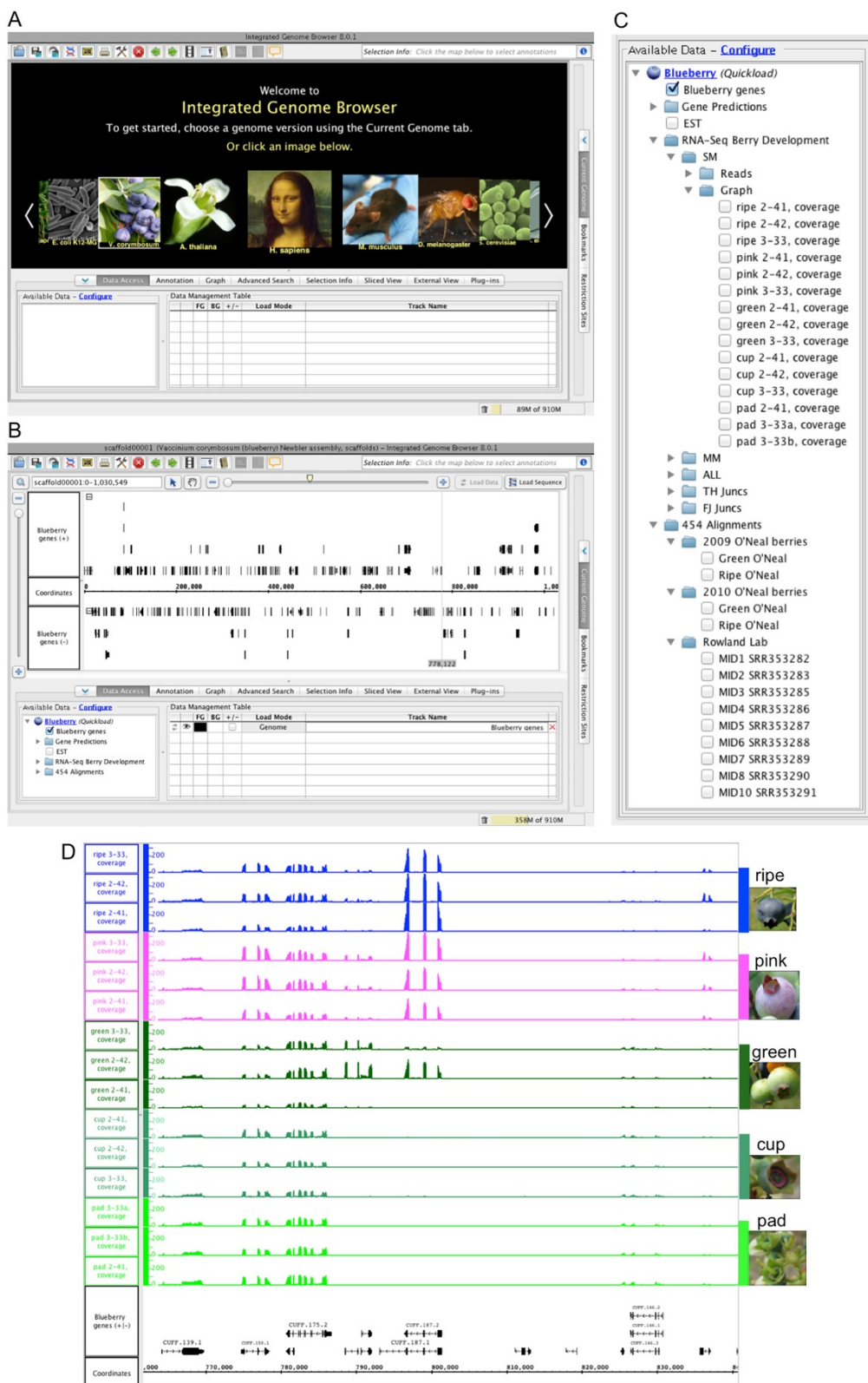
**Table 2 Genes predicted by Cufflinks RNA-Seq analysis or *ab initio* gene-finding programs**

Method	Number of genes
Cufflinks (RNA-Seq)	56,087
GeneMarkhmm	4,794
Augustus	1,745
GlimmerHMM	933
454 Scaffolds (454 RNA-Seq)	281

It is important for researchers to have access to the blueberry data in formats that facilitate exploration, visualization, and analysis of individual sequences, genes, and gene families. To enable better use of the data, the blueberry genome assembly, annotations, and RNA-Seq data have been made available in the Integrated Genome Browser, an open source, desktop genome visualization platform. IGB uses a companion IGBQuickLoad Web site to distribute datasets in formats that enable both interactive visualization and larger-scale analysis. For example, users can download datasets in formats that are amenable to bioinformatic analysis, such as the BAM and bed formats, or instead can simply view a listing of the datasets in IGB and open them there. To open the blueberry genome in IGB, users first select the blueberry image shortcut from the IGB start screen (Figure 1). Clicking the blueberry image triggers loading of the non-redundant gene dataset described above along with a listing of blueberry genome scaffold and their sizes (in the 'Current Genome' tab) and a catalog of available datasets (in the 'Data Access' tab). As of this writing, the available datasets included the Illumina and 454 RNA-Seq data described here as well as reference datasets from other studies, including a 454-based sequencing dataset that included multiple sample types and berry fruit ripening stages [16]. Thus IGB allows researchers to view and explore multiple datasets from diverse sources, not just the current study.

Once the genes are loaded, users can search for genes by name or by annotation key word using the 'Quick Search' (top left) or 'Advanced Search' tab; the latter also enables searching for sequence motifs. Gene models are associated with descriptive text assembled from homology searches against BLAST databases from NCBI and PlantCyc (described below), but users can investigate individual gene models in greater detail using the BLAST features available through IGB. Right-clicking a gene model displays a menu that displays options to run BLASTX or BLASTP searches against the nr protein database at NCBI. Selecting one of these options opens a web browser window, which shows results from the search. In addition to the BLAST search feature, the right-click menu (also called a 'context menu') offers the option to show the sequence of a genome model in a separate window, which in turn enables selecting and copying the protein sequence. Using this feature, users can easily copy and paste the protein sequence into other web-based search and analysis tools, such as the InterproScan search tool, which can identify conserved motifs in protein sequences.

Users can also load the RNA-Seq datasets and use the observed pattern of expression to gain insights into the function of individual genes. To illustrate, the image shown in Figure 1 depicts a region from scaffold00001



**Figure 1** Blueberry genes in Integrated Genome Browser. (A) IGB start screen showing shortcut to the blueberry genome. (B) Clicking the image labeled *Vaccinium corymbosum* triggers loading of blueberry genome assembly May 2013 and the non-redundant gene set. (C) Blueberry datasets available in IGB. (D) Coverage graphs showing read density on a region of scaffold00001.

showing gene models and RNA-Seq coverage graphs from five berry developmental stages. Because the overall read counts for each track were similar, one can visually estimate and compare expression levels for genes provided the coverage graphs are put onto the same scale, which is possible using the IGB 'Graph' tab. In this example, comparing graph heights between stages highlights ripening related expression of gene CUFF.187, which is homologous to genes in many other plant species but has no known function. The stage-specific expression of CUFF.187 demonstrates the power and usefulness of visualizing gene models alongside expression data within a visualization environment that also enables rapid exploration of genes and their functions.

#### Functional annotation of blueberry gene models and comparison with other plant species

To identify putative functions of newly annotated blueberry genes, we searched for homologous proteins using BLASTX searches of the non-redundant nr database from GenBank. We found that most blueberry genes (58%) had significant homology (e-value  $\leq 0.001$  and percent identity  $\geq 30\%$ ) with at least one protein in the nr protein database. Typically, the highest scoring, best match to a blueberry protein was a protein from the wine grape *V. vinifera* (Figure 2A). The abundance of best-scoring hits from grape was probably due to two factors: the similarity between grape and blueberry [30] and the large number of grape sequences that were present in the nr database at the time this analysis was done. The nr database contained nearly 80,000 proteins from grape, many of which were conceptual translations from the sequenced grape genome, which has been available since 2007 [31]. The tea plant *Camellia sinensis*, also represented in Figure 2A, is more closely related to blueberry than grape, but had fewer best hits overall primarily because there were fewer tea sequences in the databases. We found that in general, if a blueberry gene had a match to a tea protein, the tea protein match had a better score and higher percent identity than the corresponding best match from grape.

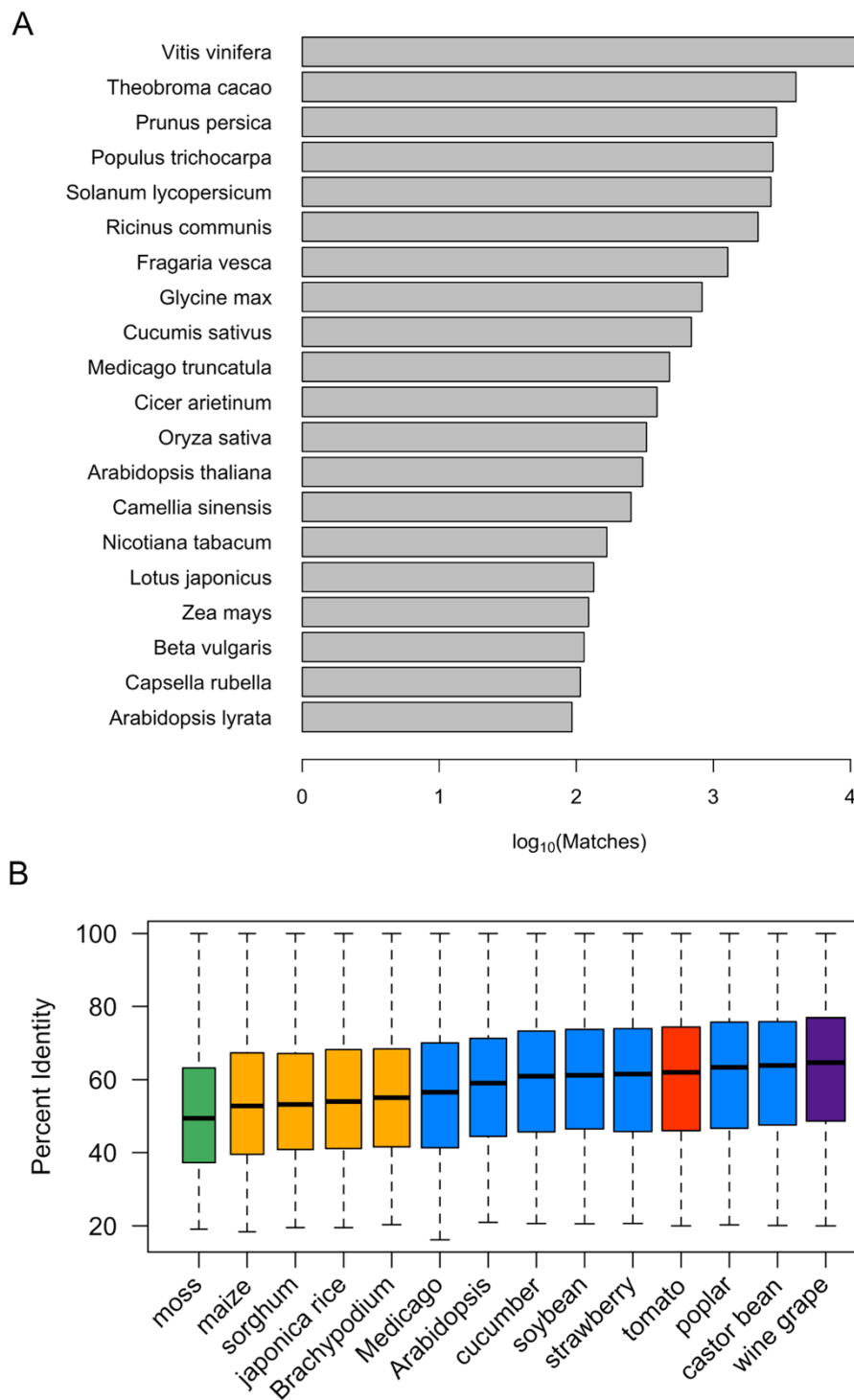
Nonetheless, these results prompted us to explore relationships between blueberry and other plant genomes. For this, we used plant-specific RefSeq databases from GenBank, including databases for grape, tomato, *Arabidopsis*, and several other plant species where a close-to-complete, well-annotated proteome was available [32]. To characterize relationships between blueberry and these other plants, we searched the blueberry transcripts against these close-to-complete plant RefSeq databases and identified the best scoring protein from each genome for each blueberry gene model. Figure 2B shows the distribution of percent identity scores obtained for the best-matching sequences from each species. Grape

proteins had the highest median percent identity scores, followed by poplar, castor bean, and tomato. Blueberry and other *Ericales* species are part of the asterid clade of flowering plants, and are more closely related phylogenetically to tomato than to rosids *Populus trichocarpa* (poplar) and *Ricinus communis* (castor bean) [33]. The BLAST results do not contradict this relationship but instead highlight how sequence similarity may reflect similarities in physical or biochemical characteristics. Grape and blueberry are deciduous, berry-producing plants with long generation times. Poplar is also a deciduous, woody plant with a long generation time. Biochemical or morphological similarities that could explain the similarity between blueberry and castor bean are less obvious, however. These results suggest that in-depth comparison between grape, blueberry, and castor bean seed transcriptomes could lead to new insights into developmental programs at work in berries of the three species.

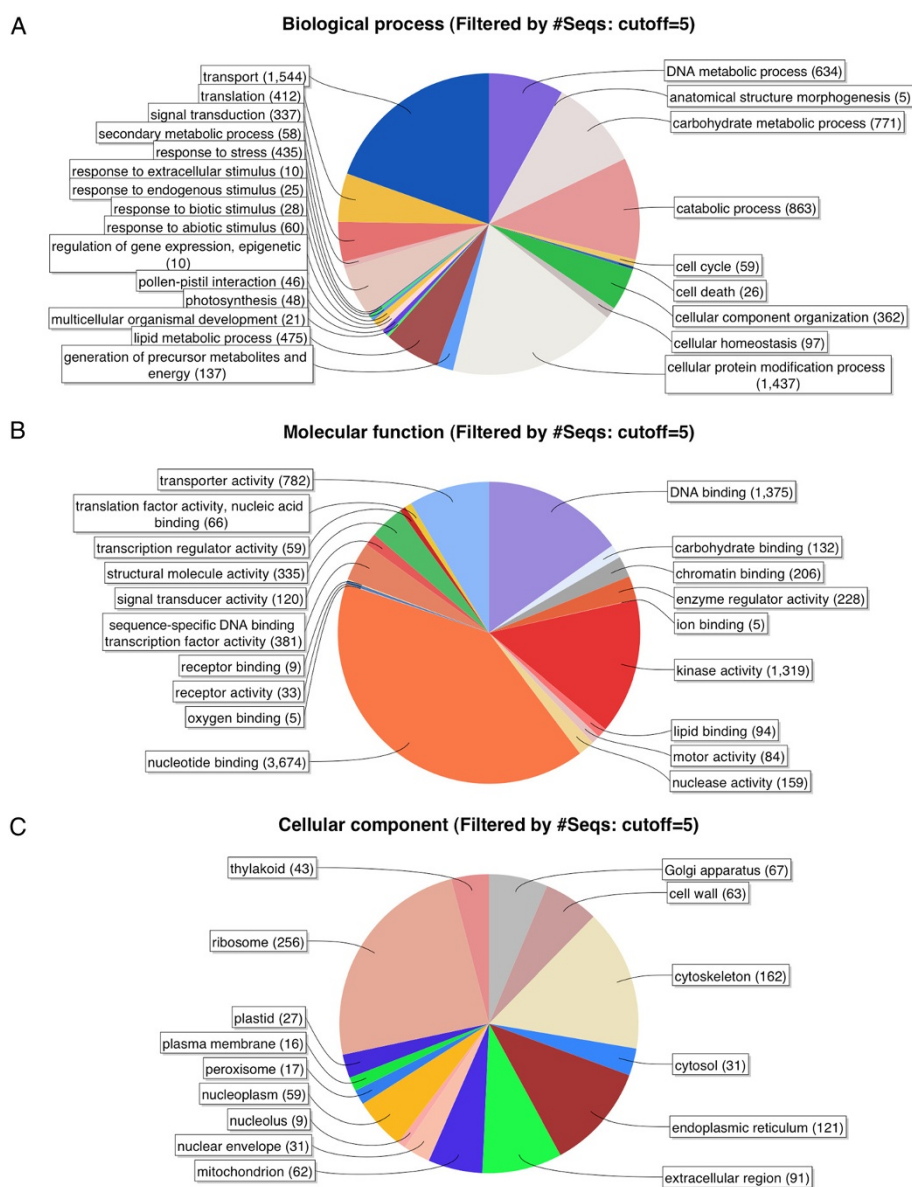
#### Blast2GO annotation

Blast2GO is a popular annotation platform that uses results from homology searches to associate sequence with GO terms and other functional annotations [34], a critical step toward enabling analysis of high-throughput gene expression studies. To facilitate analysis of the blueberry developmental time course data and also provide a resource for other blueberry researchers, we used Blast2GO to annotate the blueberry proteins with GO terms. Blueberry sequences were searched against the nr protein database and the InterPro database [35] and the results were imported into the Blast2GO program's graphical user interface, which assigned GO terms to 18,143 transcripts representing 15,308 genes. Plots summarizing GO terms and the number of gene products assigned to each term are shown in Figure 3, and a file containing the GO term mappings is provided as Additional file 3. Similar to *Arabidopsis* [36] and other plants, a large number of genes were annotated with terms related to transcription factors. Around 1,300 transcripts received GO annotation 'DNA binding,' of which around 400 were also annotated with the more specific GO term 'sequence-specific DNA binding transcription factor activity.' These sequences likely represented expressed transcription factors involved in regulation of gene expression during fruit development and ripening. More than 500 proteins were associated with terms related to primary and secondary metabolism, lipid metabolic processes, or carbohydrate metabolic processes, reflecting the dynamic processes underway in fruit development. Thus the GO annotation identified a large set of candidate genes likely to be involved in regulation and synthesis of bioactive secondary metabolites.





**Figure 2 Results from searching nr and RefSeq protein databases. (A)** Number of best-matching proteins by species. Blueberry transcript sequences were searched against the non-redundant protein database from NCBI and the best-matching protein for each blueberry sequence was identified. The plot shows the number of blueberry genes whose best matching protein was from the indicated species. **(B)** Distribution of percent identity scores by plant RefSeq database. Blueberry transcript sequences were used to search RefSeq protein databases for plants with close-to-complete, annotated genomes. Boxplots show the distribution of percent identity scores by species. RefSeq databases include wine grape (*Vitis vinifera*), castor bean (*Ricinus communis*), poplar (*Populus trichocarpa*), tomato (*Solanum lycopersicum*), strawberry (*Fragaria vesca*), soybean (*Glycine max*), cucumber (*Cucumis sativus*), *Arabidopsis* (*Arabidopsis thaliana*), *Medicago* (*Medicago truncatula*), *Brachypodium* (*Brachypodium distachyon*), rice (*Oryza sativa*), sorghum (*Sorghum bicolor*), corn (*Zea mays*), and a moss (*Selaginella moellendorffii*).



**Figure 3** Gene Ontology annotations for the three sub-trees of GO. (A) Biological process; (B) molecular function; (C) cellular component. Numbers indicate the number of protein-coding blueberry transcripts assigned to each category. Plots were made using Blast2GO. The number of genes assigned to each category can be obtained from Additional file 3.

### Metabolic pathway annotation using PlantCyc enzyme database

The PlantCyc database is a collection of curated and computationally predicted enzymes, enzymatic reactions, and metabolic pathways for 18 plant species, including grape, *Arabidopsis*, cassava, poplar, and rice [37]. The PlantCyc databases are closely tied with the Pathway Tools software package [38], a visualization and database system for pathways and biochemical reactions. At the time of this writing, the pathways database for *V. vinifera* (GrapeCyc version 3.0) was one of the most

complete, containing annotations for 432 pathways. The similarity between grape and blueberry sequences suggested that comparing the blueberry sequences to enzymes in GrapeCyc and other PlantCyc databases would identify candidate genes encoding enzymes of primary or secondary metabolism. To identify enzymatic functions for blueberry genes, the BLASTX algorithm was used to search for matching PlantCyc sequences. To eliminate matches arising from alignments between domains that occur in many different sequences (e.g., ATP-binding cassette), only matches that covered at least 60%

of the subject sequence with at least 45% identity were considered. Under these criteria, transcripts from more than 7,100 blueberry genes were found to match at least one PlantCyc enzyme sequence, and there were over 450 pathways from PlantCyc that had at least one enzyme matching a blueberry sequence. As before, grape proteins were typically the best matches for blueberry sequences.

#### **In-depth analysis of ethylene biosynthetic pathway gene expression**

Ethylene is a gaseous plant hormone that controls many aspects of plant development and physiology, especially ripening. In climacteric fruits such as tomato and banana, a burst of ethylene biosynthesis triggers ripening, and post-harvest treatment with exogenous ethylene can control ripening onset and progression. By contrast, blueberries do not appear to undergo a burst of ethylene synthesis prior to ripening, and post-harvest application of ethylene has far less effect on the ripening process. Nonetheless, ethylene can influence aspects of flowering and ripening, as shown by experiments with ethephon, a horticultural chemical that when applied to foliage is enzymatically converted to ethylene. When applied to blueberry plants during the harvest, ethephon accelerates and synchronizes ripening, and when applied in the fall (autumn), ethephon delays flowering the following spring and increases the number of flower buds (reviewed in [39]). This suggests that ethylene is important in flowering and fruit development but its role is likely to be very different than in climacteric fruits.

Ethylene is synthesized from L-methionine via three reactions that are catalyzed by SAM synthase (methionine adenosyltransferase), ACC synthase (1-aminocyclopropane-1-carboxylate synthase), and ACC oxidase (1-aminocyclopropane-1-carboxylate oxidase). In most plants, large multi-gene families encode all three enzymes. The PlantCyc database reports that the grape genome contains 34 SAM synthase genes, 20 ACC synthase genes, and 12 ACC oxidase genes. Consistent with this, the PlantCyc-based annotation of blueberry identified similarly large numbers of genes encoding ethylene biosynthetic genes, including 22 genes encoding SAM synthase, six genes for ACC synthase, and seven genes encoding ACC oxidase (Additional file 1: Figure S2). Interestingly, the expression patterns for SAM synthases were highly variable and at least one SAM synthase gene was highly expressed (more than 1,000 reads per kilobase of transcript per million reads mapped, RPKM) in each stage of berry fruit development and ripening. Expression of genes encoding ACC synthase, a key control point for ethylene and ripening in tomato, was also highly variable. Some genes encoding ACC synthases were expressed at very low levels (close to 0 RPKM), while others were

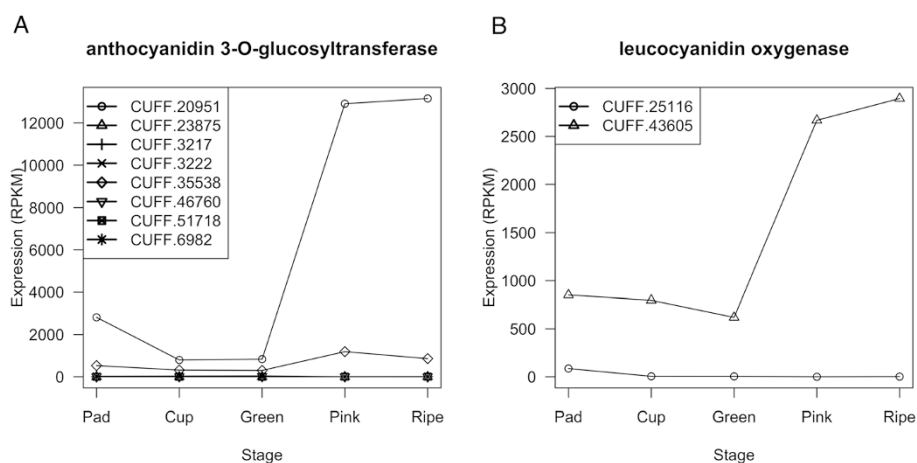
expressed at 200 RPKM or higher, which was around the 95th percentile of gene expression as measured in RPKM. Two ACC oxidase genes were expressed at 150 RPKM or higher at each stage, while others were expressed at much lower levels. Interestingly, one ACC oxidase gene (CUFF.81159) was highly expressed during ripening and reached more than 3,000 RPKM in ripe fruit. The uniformly high expression (>150 RPKM) of ethylene biosynthesis genes at every stage, combined with the extremely high expression of ACC oxidase gene CUFF.81159, suggested that ethylene is produced throughout berry fruit development and that these levels are likely to peak in ripe fruit.

#### **Anthocyanin biosynthetic pathways**

Anthocyanins are the 3-O-glycosylated forms of anthocyanidins, which consist of a polyphenolic ring substituted with -H, -OH, and -OCH<sub>3</sub> groups at different positions. The substituted group and its location on the polyphenolic ring determine the type of anthocyanidin and may also dictate aspects of biological activity. In blueberry, the anthocyanins containing malvidin, delphinidin, and peonidin aglycones are especially abundant by weight [4]. Another level of diversity arises from the type of sugars attached, and these sugar groups may influence bioavailability in the mammalian digestive tract [40,41]. Thus it is of interest to identify enzymes that catalyze steps in anthocyanin biosynthesis and profile their expression pattern and relative abundance in ripe fruit.

Comparison with PlantCyc enzymes identified 31 genes with homology to enzymes involved in anthocyanin synthesis, including two genes (CUFF.20951 and CUFF.43605) that were highly expressed in pink and ripe fruit in comparison with other genes in the pathway (Figure 4). CUFF.20951 was the most highly expressed; it reached more than 12,000 RPKM in pink and ripe fruit, up from around 2000 RPKM in green fruit. The best matching PlantCyc enzyme for CUFF.20951 was an enzyme from poplar that was annotated as anthocyanidin-3-O-glucosyltransferase, which transfers a glucose sugar onto the 3-O position of the anthocyanidin polyphenolic ring. Searching the nr database using BLASTP identified even better matches, including predicted flavonoid galactosyl and glucosyltransferases from a variety of plant species. Based on its homology to other sequences, the CUFF.20951 protein likely is involved in glycosylation of flavonoids, but its preferred aglycone and sugar substrates are impossible to determine based on sequence data alone. However, judging from the extremely high expression of this gene in blueberry, the CUFF.20951 protein is likely to be responsible for a high percentage of anthocyanin (or other flavonoid) production in ripe berry fruit.

The other highly expressed gene was CUFF.43605, which reached a peak of around 2000 RPKM in pink



**Figure 4 Expression pattern of anthocyanin biosynthetic genes.** Expression pattern for genes annotated to enzymatic reactions in anthocyanin-related pathways in PlantCyc are shown. **(A)** Genes encoding leucoanthocyanidin dioxygenase, catalyzing EC 1.14.11.19, dihydroquercetin synthesis from leucoanthocyanidin or cyaniding synthesis from leucoanthocyanidin. **(B)** Genes for enzymes catalyzing 3-O-glucosylation of delphinidin (EC 2.4.1.115). Both are annotated to PlantCyc pathway anthocyanin biosynthesis (delphinidin 3-O-glucoside).

and ripe fruit, up from around 900 RPKM in the earlier stages. The best PlantCyc match for CUFF.43605 was a protein from *Brassica rapa* (Bra019350, 78% identity) annotated as a leucocyanidin oxygenase, also called anthocyanidin synthase (ANS) or leucocyanidin dioxygenase (LDOX). Homologous enzymes from *Arabidopsis* [42,43] and rice [44] convert leucoanthocyanidin to anthocyanidins, precursors for anthocyanins, but they can also catalyze formation of other bioactive flavonoid precursors, notably dihydroquercetin, the precursor of quercetin. Anthocyanins and quercetin are both abundant in berries, but quercetin has greater bioavailability and therefore may be a more potent bioactive compound in berry fruit [45,46]. If the preferred end product of the CUFF.43605 reaction is indeed dihydroquercetin, then its remarkably high expression likely has a positive effect on quercetin concentration in berry fruit. However, if its major end product is cyanidin, then it likely acts to decrease quercetin levels by consuming leucoanthocyanidin, the dihydroquercetin precursor. As with CUFF.20951, sequence and expression analysis alone is likely insufficient to distinguish these possibilities. Nonetheless, the high expression of CUFF.43605 makes this gene a fruitful candidate for investigating genetic control of anthocyanin and quercetin abundance in berries.

#### Other pathways - bixin and dhurrin

Bixin is the primary component of annatto, a commonly used food dye collected from the seeds of *Bixa orellana*, which grows in the tropics and is also known as the lipstick plant. Annatto is used in folk medicine and for body decoration and has also been investigated as a plant-based treatment for diabetes [47], cancer [48], and

microbial infections [49]. We found that four blueberry genes matched two of three annotated enzymes of the pathway (Additional file 1: Figure S3). Expression of all four genes peaked in the mature green stage of fruit development, around the time when seeds were developing. Grape seed contains bixin [50], and, as discussed previously, the grape and blueberry proteomes were remarkably similar. These results suggested that blueberries may contain a bixin-like compound.

Blueberry genes similar to enzymes from a potentially harmful pathway were also found. Blast analysis identified putative blueberry homologs for each of three biosynthetic enzymes involved in synthesis of dhurrin in sorghum [51-53]; the blueberry genes shared between 45 and 60% identity with their putative homologs from sorghum. Dhurrin, which reaches up to 10% dry weight in sorghum seedlings, is one of a large class of cyanogenic glycoside (CG) defense compounds plants synthesize from amino acid precursors as a form of chemical warfare against insects and other herbivores [54,55]. Stored as inactive glycosides, mechanical damage to cells (such as from chewing) activates endogenous glycosidases that remove the sugar group, triggering production of toxic hydrogen cyanide (HCN) due to instability of the sugar-free aglycone or to the activity of other catabolic enzymes. Interestingly, the putative blueberry homologs were most highly expressed in the green fruit stages (Additional file 1: Figure S3D), suggesting that green berries synthesize a cyanogenic glycoside that discourages insects and mammals from eating unripe berries.

Homology searches also identified putative berry homologs of enzymes involved in two CG catabolic pathways, one that removes the glycosyl group from the CG



leading to production of HCN, and another pathway that detoxifies CGs by converting them to harmless byproducts. Seven genes had significant similarity to enzymes in the cyanogenic catabolic pathway, including four genes resembling dhurrinase, which deglycosylates dhurrin, and three genes that were similar to hydroxynitrile lyase, which catabolizes the dhurrin aglycone to HCN. Similar to the biosynthetic enzymes, these catabolic, cyanogenic enzymes were most highly expressed in unripe fruit. We also identified a blueberry candidate gene encoding nitrilase 4, an enzyme that detoxifies dhurrin and possibly other cyanogenic glycosides by converting them to aspartic acid and asparagine [56]. Searching with nitrilase 4 from *A. thaliana* (AtNIT4, AT5G22300) identified blueberry gene CUFF.32314, which shared 83% identity with the *Arabidopsis* protein. Plant nitrilase 4 enzymes are highly conserved [56], typically sharing 60 to 70% identity at the amino acid level, suggesting that CUFF.32314 indeed encodes nitrilase 4. The putative blueberry nitrilase CUFF.32314 gene was highly expressed in young fruit and ripe fruit, reaching more than 70 RPKM in ripe berries, but had much lower expression (<40 RPKM) in mature green berries. Thus the expression profile of the candidate gene involved in CG detoxification was roughly complementary to profiles of candidate genes involved in CG synthesis and cyanide release. This suggests that immature green berries produce CG compounds that discourage herbivory, but that berry NAT4-like genes then detoxify these CGs as part of the ripening process.

#### Differential gene expression during blueberry development and ripening

Fruit development and ripening are dynamic processes involving cell division, expansion, differentiation, synthesis of pigments and other secondary metabolites, as well as import of sugars into the fruit (reviewed in [57]). To better understand these processes in blueberry, we used the developmental time course RNA-Seq data to examine global patterns of gene expression throughout fruit development, maturation, and ripening. Statistical analysis of gene expression identified nearly 19,000 genes that were differentially expressed between at least two stages, even when a relatively stringent false discovery rate cutoff of 0.001, corresponding to roughly 1 in 1,000 false discoveries, was applied. Table 3 reports the number of genes that were differentially expressed between stages, broken down by up- and down-regulation. Except for the pink versus ripe fruit comparison, there were more up-regulated than down-regulated genes in every comparison between a later stage and an earlier stage, reflecting the greater diversity of genes that were expressed in the later stages as tissues continued to develop and differentiate and as the seeds formed within

**Table 3 Up- and down-regulated genes between sample types**

	Pad	Cup	MG	Pink	Ripe
Pad	-	566	2,390	6,010	6,077
Cup	526	-	1,221	5,270	5,003
MG	1981	919	-	3,583	3,726
Pink	6,765	6,594	3,038	-	18
Ripe	6,821	6,466	5,159	56	-

Each cell in the table represents a comparison between two sample types. Cells above the diagonal report the number of genes that were up-regulated in the later of the two compared stages. Cells below an empty cell report the number of genes that were down-regulated in the later of the two compared stages. Cutoff for differential expression was false discovery rate of 0.001. Stages are shown in order of earlier (Pad) to later (Ripe).

the fruit. Viewed globally, the gene expression changes that occurred during progression of fruit growth, maturation, and ripening were extensive and dynamic.

Figure 5 reports the number of genes that were differentially expressed (DE) between adjacent stages and highlight the physiological transitions as fruits enlarge from small, pad-like structures on the base of the pedicel to round, mature berries harboring seeds, flesh, and sugars. The most similar of the adjacent stages were the pink and ripe fruit, collected 14 days apart. Fewer than 100 genes were DE between pink and ripe fruit. Of these, most (75%) were down-regulated. The up-regulated genes included genes of unknown function and genes involved in enzymatic activity, such as UDP-glucosyltransferases, an alcohol dehydrogenase, and hormone biosynthetic genes. Down-regulated genes included many enzymes, such as the putative dhurrin biosynthetic enzymes, as well as other genes of unknown function. By contrast, the pad and cup stages, which were collected 10 days apart, had more than ten times as many DE genes. GO enrichment analysis of genes differentially expressed between the pad and cup stages found that a significant number of DE genes were annotated with terms related to transcriptional regulation, regulation of enzymatic activity, and cell wall biosynthesis (Additional file 4). The type and high number of DE genes in the pad-to-cup versus the pink-to-ripe comparison reflects the extent of developmental changes underway during early, rapid growth stages of berry development.

However, the transition from fully rounded green fruit to pink fruit involved by far the largest number of differences. The chief physical difference between the two stages was color; pink berries had some reddish color indicating the onset of ripening, but other than this, green and pink berries were similar. Gene Ontology enrichment analysis identified around 40 terms as being significantly enriched among genes differentially expressed between the green and pink fruit. Two terms were related to photosynthesis and likely reflected down-regulation of

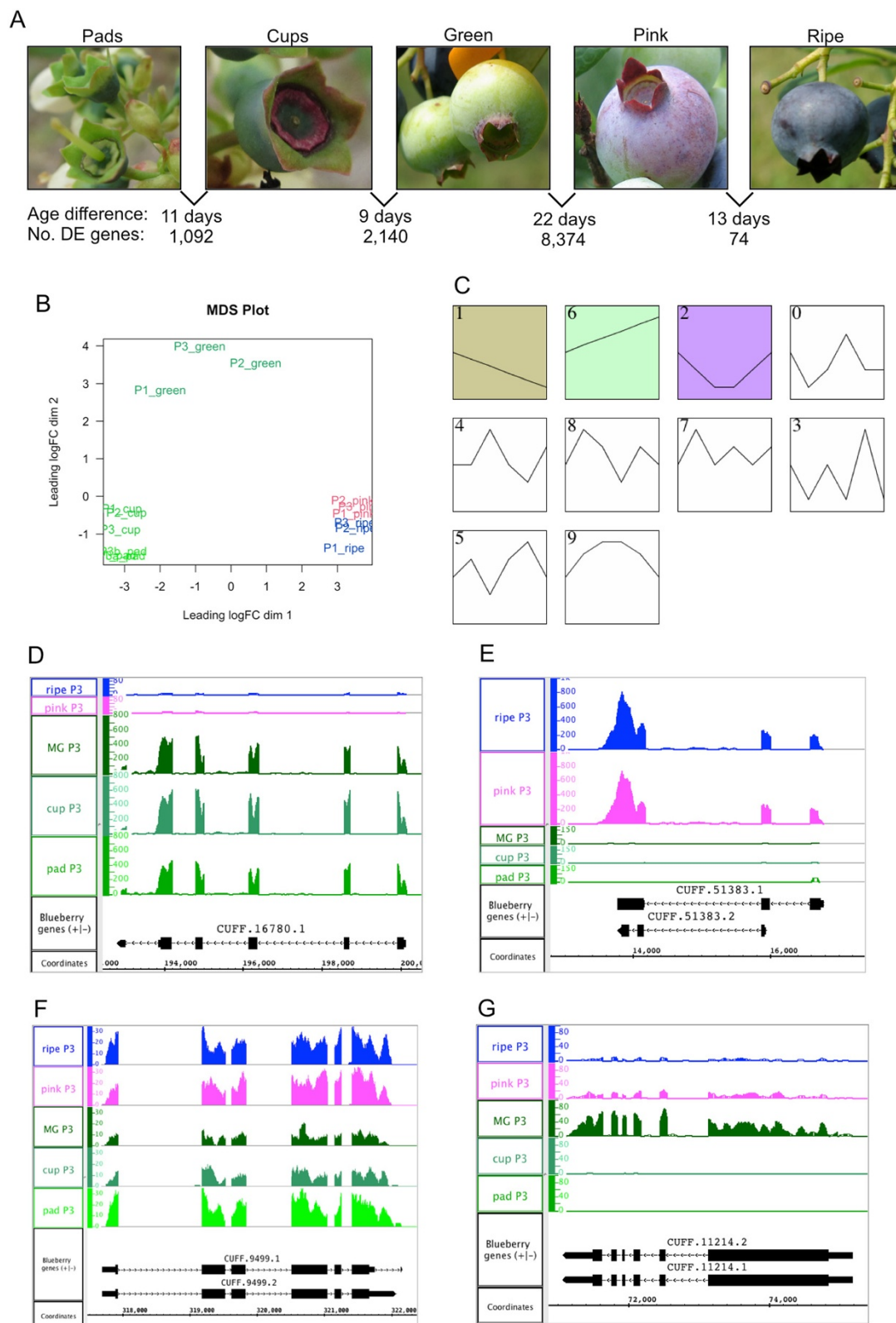


Figure 5 (See legend on next page.)

(See figure on previous page.)

**Figure 5 RNA-Seq analysis of gene expression changes during berry fruit development and ripening.** (A) Photographs of berries exemplifying five stages of berry fruit development and ripening and the number of differentially expressed (DE) genes between adjacent stages. (B) Multi-dimensional scaling (MDS) plot of expression values from five sample types and three plants (P1-P3). Samples that are near each other in the two-dimensional space are more similar with respect to gene expression. (C) Gene expression profiles detected by clustering program Short Time-series Expression Minor (STEM). The x-axes represent stage (earlier to the left) and the y-axes represent expression level. Clusters 0 through 9 are listed in order of statistical significance, with clusters 1, 6, 2, and 0 being the most significant. (D-G) Cufflinks-predicted genes from clusters 1, 6, 2, and 0 encoding proteins similar to (D) leucoanthocyanidin reductase, (E) MYB-like transcription factor up-regulated in ripening fruit, (F) tryptophan synthase (indole-salvaging), and (G) abscisic acid signaling transcription factor ABI3, respectively. In (D-G), graphs represent the number of reads overlapping genomic positions indicated on the 'coordinates' axis. RNA-Seq coverage graphs from P3 (plant 3-33) are shown.

photosynthetic functions, despite the fact that when the pink berries were collected, they typically were only partly pigmented and much of the berry surface was still green. Of 43 genes annotated to the term 'photosynthesis,' 27 were differentially expressed and all were down-regulated in pink fruit. Other significant terms were related to metabolic pathway functions, including the terms metabolic process, catalytic activity, catabolic process, hydrolase activity, carbohydrate metabolic process, lipid metabolic process, transferase activity, and biosynthetic process. There was also enrichment in functions related to transcriptional regulation of gene expression, including chromatin binding, nucleic acid binding, sequence-specific DNA binding transcription factor activity, and signal transduction.

Multi-dimensional scaling was used to cluster samples according to their overall similarity of gene expression pattern (Figure 5B). Samples from similar stages formed clusters. Pink and ripe fruit formed a cluster, cup and pad stages formed another cluster, and mature green fruit formed another cluster separate from the others. Interestingly, one ripe fruit sample (plant 2-41, P1) clustered slightly distant from the other ripe fruit samples. Visualization of RNA-Seq coverage graphs in IGB showed that some genes were up- or down-regulated in this sample relative to the other ripe fruit samples. Genes that were down-regulated in P1 relative to the other ripe berry samples had functions related to cell wall degradation or modification, whereas genes that were up-regulated had functions related to stress responses. This suggests that the P1 ripe berry sample was undergoing a stress relative to the other samples. Nonetheless, the possible presence of a possible confounding factor related to stress did not significantly reduce power to identify DE genes between samples types.

A time series clustering program (STEM [58]) was used to identify groups of genes that varied in concert over time. Scaled, averaged expression values for genes found to be DE between any two fruit stages were provided as inputs to the STEM program, which identified several gene expression profiles depicted graphically in Figure 5C. Two of the most statistically significant profiles had complementary patterns (clusters labeled

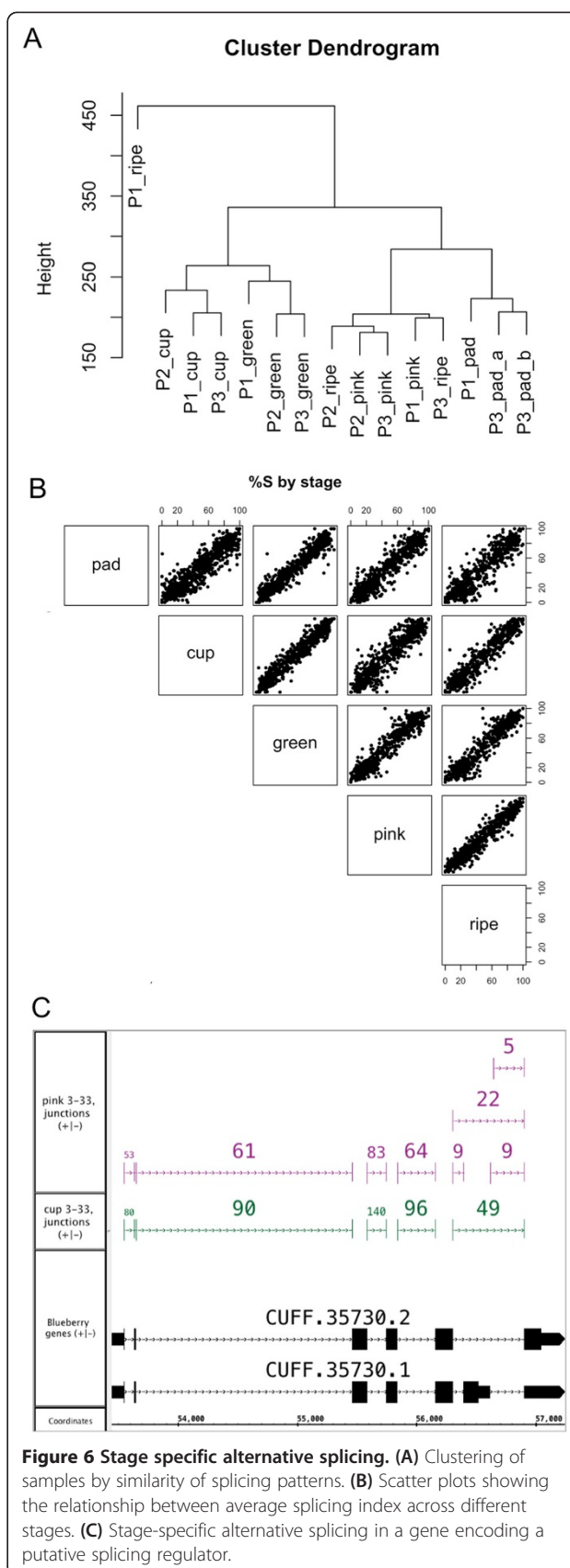
1 and 6). The 'early high' profile (number 1 in Figure 5C) contained genes with high expression in early fruit stages and lower expression in later stages. Gene Ontology enrichment analysis showed that this early high cluster contained an unusually large number of genes annotated with GO terms related to photosynthesis, catalytic activity, kinase activity, cell cycle, and DNA binding (Additional file 4). The 'late high' profile (number 6), contained genes with expression patterns complementary to the early high profile, with genes reaching their peak expression in the pink and ripe fruit stages. The genes in this cluster were enriched with GO terms related to transport, sugar metabolism, and catalytic activity, but had a lower than expected proportion of genes annotated to term DNA metabolic process. Only 1% of DNA metabolic process genes were in the late high cluster. These two complementary profiles highlight the diverse biological processes underway in ripe, fully mature fruit versus immature, rapidly growing fruit [57,59-61]. In the early stages, cell growth and cell division are underway, while in later stages, sugars and other metabolites are being imported into the fruit and cell growth happens largely through cell expansion, not cell division.

The next two most significant clusters also had complementary patterns. The 'green high' (number 0) cluster included around 1,200 genes that peaked in expression in the mature green stage. This cluster contained unusually many genes annotated with terms related to metabolic processes, transport, DNA binding, and secondary metabolism, suggesting that the green fruit stage represents a developmental transition and also is rich in production of secondary metabolites (Additional file 5). The 'green low' cluster (number 2) contained around 1,500 genes but was significantly enriched with only one term: catalytic activity. Interestingly, genes annotated with this term were unusually prevalent among all genes in all four profiles, suggesting that fruit development is rich in biosynthetic processes. However, the nature of these biosynthetic processes varies among stages, beginning with photosynthetic processes in the early stages and transitioning to secondary metabolism in the later stages. Figure 5D-G show RNA-Seq coverage graphs for genes exemplifying the four profiles.

### Transcript variation during fruit development and ripening

As described above, nearly one quarter of multi-exon genes were associated with multiple gene models corresponding to putative products of alternative promoters, alternative splicing, or alternative polyadenylation sites. Any or all three of these mechanisms of alternative transcription can play a role in developmental regulation of gene expression. To test whether alternative transcription was developmentally regulated during berry fruit development and ripening, we used the CuffDiff program, a companion program to Cufflinks, to identify genes whose patterns of splicing, promoter use, or polyadenylation changed during the development time series. Cuffdiff reported more than 700 genes as undergoing some type of differential transcript variation between pairs of conditions. To assess the results, a subset of the highest confidence CuffDiff results was selected at random and manually inspected using IGB. RNA-Seq read alignments were loaded into the viewer alongside the gene models and read support for alternative transcripts was compared between samples. In most cases involving alternative splice sites, the region that was different between alternative transcripts had small numbers of supporting reads, typically fewer than ten reads per splice junction. By contrast, examples of alternative promoter use (Additional file 1: Figure S4A) and alternative 3' ends (Additional file 1: Figure S4B) had much stronger support. In these instances, typically there were dozens of reads supporting one or both alternatives in each sample. Thus, the CuffDiff analysis was able to identify differential use of alternative promoters and alternative transcription sites but was not as good at distinguishing differential splicing.

We therefore used an alternative approach based on the ArabiTag algorithm [62] to test specifically for differential splicing between stages. For this, a splicing score was calculated that represented the percentage of spliced reads supporting alternative splice site choices from differentially spliced regions. Hierarchical clustering of splicing scores found that ripe and pink berry samples formed a cluster, mature green berries formed a cluster, while the cups and pad stages were intermixed (Figure 6A). Interestingly, the P1 ripe fruit sample was an outlier and formed a distinct cluster apart from the others; this was consistent with previous results in which P1 clustered apart from P2 and P3 in an MDS plot (Figure 5B). Nonetheless, pairwise comparisons of average splicing score found for most alternatively spliced genes, the relative abundance of splice forms was consistent between stages, with some outliers (Figure 6B), and annotated spliced variants were co-expressed. Statistical testing of the splicing score supported this observation, identifying around 90 genes with developmentally



**Figure 6 Stage specific alternative splicing.** (A) Clustering of samples by similarity of splicing patterns. (B) Scatter plots showing the relationship between average splicing index across different stages. (C) Stage-specific alternative splicing in a gene encoding a putative splicing regulator.



regulated differential splicing, including some with predicted functions related to splicing. These included CUFF.35730 (Figure 6C), which was similar to splicing-related transformer-SR ribonucleoproteins from many plant species. The best *Arabidopsis* match (AT4G35785) is one of two transformer-like genes in *Arabidopsis*; both genes (AT4G35785 and SR45a) contain alternatively spliced 'toxic exons' that introduces a premature stop (termination) codon (PTC), and splicing of the toxic exon in SR45a is sensitive to stress [63]. Differential inclusion of a PTC-containing toxic exon appears to be conserved in blueberry, as CUFF.35730 also contained a 'toxic exon' that introduced a stop codon. According to the RNA-Seq data, the full-length, exon-skipped form represented a higher percentage of the splice variants in cup and mature green fruit stages, while the exon-included form was less abundant in pink fruit. Thus splicing patterns in blueberry during fruit development and ripening vary by stage, similar to overall gene expression levels.

#### Comparison with previous sequencing studies of blueberry

Before now, several studies have been published that used Sanger [64,17], 454 [16], or Illumina sequencing [65] to characterize blueberry genes and profile gene expression changes. Two studies used *de novo* transcriptome assembly procedures to assemble transcript models, which they then annotated using similar methods to those described here. At the time of writing, only a subset of the data reported in the Illumina-based study was publicly available [65], but the others have made their raw sequence reads available as part of dbEST or the Short Read Archive. To provide a resource to researchers and also assess the completeness of the new transcriptome assembly reported here, we aligned publicly available EST, 454 and Illumina sequences onto the blueberry assembly and compared the alignments to the non-redundant set of gene models generated by Cufflinks and the *ab initio* gene prediction programs. Of the 22,415 Sanger ESTs from dbEST, 19,967 (89%) were aligned onto the draft genome assembly. A comparable percentage of sequences from the much larger 454 sequencing study [16] were also aligned. More than 85% of these 454-based ESTs aligned to the genome, and around 75% of the 454 ESTs overlapped over one or more berry gene models. Similar results were obtained for data from the Illumina-based study [65]. We obtained around 15.5 million read pairs from the Short Read Archive using object number SRR942391. Around 55.4% were mapped as proper pairs while another 13.28% were mapped as singletons. Of the mapped reads on the genome, more than 95% overlapped a blueberry gene model. Thus the non-redundant gene models presented here encompass the

vast majority of previously generated sequences from blueberry.

The Sanger-EST-based study published by Zifkin and co-workers reported in-depth analysis of several genes involved in anthocyanin and proanthocyanidin biosynthesis, two distinct branches of flavonoid biosynthesis [17]. Proanthocyanidins (PAs), also called condensed tannins, are colorless, have a bitter taste, and may play a role in defense against disease and in discouraging consumption of immature berries. Consistent with this, Zifkin et al. found that PA genes were highly expressed in early stages of berry development, while genes specific to anthocyanin biosynthesis were most highly expressed following onset of ripening. Moreover, genes required for synthesis of precursors common to both pathways exhibited a bi-phasic expression profile, where expression was highest in early and late stages and lowest in the intermediate stages. To assess the degree of correspondence between this earlier, EST-based study and the current RNA-Seq-based study, EST accessions were obtained from supplementary data [17], their sequences were aligned onto the blueberry genome, and the alignments were used to identify the corresponding blueberry gene models. Alignments were also made available as part of the IGBQuickLoad data repository to enable searching and browsing of EST data. Mappings between ESTs and blueberry gene models are shown in Table 4.

In general, the RNA-Seq results were consistent with the EST-based study. As shown in Additional file 1: Figure S5, RNA-Seq based expression patterns for genes encoding VcANR (anthocyanidin reductase), VcDFR (dihydroflavonol reductase), and VcUFGT (UDP-Glc: flavonoid-3-*O*-glycosyltransferase) were similar to those reported previously. Interestingly, the gene corresponding to VcUFGT was CUFF.20951, which was also the most highly expressed gene in the anthocyanin biosynthesis pathway according to the RNA-Seq data (Figure 4A). Zifkin et al. [17] also identified ESTs encoding a Myb family transcription factor with biphasic expression and showed it was able to activate an ANR promoter from poplar, demonstrating it is a likely regulator of the PA pathway in blueberry. The EST reported for this gene (JK664730) mapped to two different genes in the blueberry genome assembly, only one of which (CUFF.51789) exhibited the previously reported biphasic expression pattern. The other gene (CUFF.14288) had an expression profile more similar to the 'high early, low late' cluster 1 in Figure 5C, suggesting that it may co-regulate PA biosynthetic genes in early fruit stages.

#### Discussion

In recent years high-throughput sequencing technologies have been used to investigate the transcriptomes of

**Table 4 Correspondence between blueberry flavonoid biosynthesis ESTs and blueberry gene models**

Gene	ESTs	Gene Id
VcANR	CV091176, EG026069	CUFF.29797
anthocyanidin reductase		
VcDFR	JK655064, JK655458, JK664855, JK666169	CUFF.50634
dihydroflavonol reductase		
VcUFGT	JK664803, JK666412, JK661182, JK663155	CUFF.20951
UDP-Glc:flavonoid-3-O-glycosyltransferase		
VcMYBPA1	JK664730, JQ085966	CUFF.51789
R2R3 MYB transcription factor	JK664730, JQ085966	CUFF.14288
VcF3'H	JK654488	CUFF.39752
	(JK663426, JK666734, JK667006, JK667133, JK667135, JK655710 not found)	
VcF3'5'H	JK666610	Gene.g10884.t1
flavonoid 3',5'-hydroxylase	JK666329, JK664585	CUFF.51728
	JK665366, JK665058	CUFF.53209
	JK664236	CUFF.51711

Previously identified ESTs encoding flavonoid biosynthesis enzymes and regulators were aligned to the blueberry genome. Alignments were visualized alongside the blueberry gene models in Integrated Genome Browser to identify newly annotated genes for each EST.

numerous plant species, including many for which no reference genome sequence is available (reviewed in [66]). As sequencing accuracy improves and the cost of sequencing drops, more such projects will become feasible. However, making use of these new data can be challenging when a reference genome is not available to guide the assembly of sequence reads into contigs that are long enough to support functional annotation. Previously, we attempted *de novo* assembly of RNA-Seq reads, but technical difficulties related to the large amount of repetitive sequences in the data hindered these attempts (data not shown). Although the blueberry genome assembly used here is a draft, it was nonetheless complete enough to enable formation of high-quality gene models that could be annotated with GO terms, pathways, and protein homologies. Moreover, the gene models obtained were similar in structure and number to those of many other well-annotated transcriptomes, with one exception, which was that many apparently single-exon genes with no convincing homology to known protein sequences were obtained. Determination of the function of these single-exon, apparently non-protein-coding genes is beyond the scope of the current study, but it seems likely that many of these genes may represent long non-coding RNAs, pseudogenes, miRNAs, or other genes whose primary product is RNA and not protein. This possibility does not seem unlikely as

other non-coding RNAs as well as pseudogenes have been observed to be expressed in other similarly prepared RNA-Seq libraries [29,67].

Interestingly, blueberry sequences had highest overall percent identity to sequences from the grape vine *V. vinifera*, a basal eudicot with a slow mutation rate. The next most similar species in terms of percent identity was the castor bean *R. communis*, followed by poplar, and then by tomato, the only other asterid species in the RefSeq database at the time this analysis was done. The 454-based transcriptome analysis published by Rowland and co-workers observed an almost identical result [16], as did another group that annotated a *de novo* transcriptome assembly for the tea plant *C. sinensis*, another *Ericales* species [68]. These results seem to suggest that blueberry and perhaps even other *Ericales* species are closer to grape than to tomato; however, sequence similarity does not necessarily imply phylogenetic closeness, and a more likely explanation is that grape has diverged at a slower rate than tomato and so is more similar to the last common ancestor of grape and the *Ericaceae*. It is also a woody, deciduous plant that makes berry-like fruit and so shares some morphological characteristics with blueberry. The same explanation may also apply to poplar, also a slow-evolving, woody, deciduous plant. However, neither of these explanations serve to explain why castor bean, an herbaceous plant with a short generation time, appears to be more closely related to blueberry at the sequence level. More recently, phylogenetic analysis of mitochondrial and chloroplast sequences from cranberry (*Vaccinium macrocarpon*) confirmed that cranberry belongs in the asterid grouping and is clearly more closely related to other asterids than grape [69]. It is beyond the scope of this paper to determine correct placement of blueberry relative to these other groups. Instead, the data provided here may provide a means for experts in phylogenetic relationships to investigate these and other questions related to the evolution of flowering plants.

Another important contribution is that for the first time, we provide necessary resources for performing high-throughput gene expression studies in blueberry. The gene models provided here are a key resource for processing and analysis of RNA-Seq expression data, and to maximize their usefulness, we made the RNA-Seq data analysis described here available as part of an open source repository of data files and source code [27]. To complement the gene structures, we provided Gene Ontology and PlantCyc-based enzyme annotations for more than half the protein-coding gene models. The GO and enzyme annotations will provide critical resources for future studies, especially high-throughput studies such as RNA-Seq in which statistical analysis typically identifies many hundreds or even thousands of

differentially expressed genes. As shown here, being able to identify categories of over- or under-represented in large gene lists enables deeper understanding of biological differences. Moreover, the pathway annotations may further enable interpretation as well as functional prediction. Prior analysis of the AraCyc database together with a compendium of microarray expression data showed that enzymes belonging to the same metabolic pathway are often highly co-expressed, a property that enables identification of missing players in metabolic or regulatory pathways through large-scale analysis of expression data [70,71]. This approach has been used in many studies that used co-expression to identify candidate genes, leading to a deeper understanding of genes involved in metabolic pathways. Therefore, one important long-term result from the RNA-Seq analysis presented here, along with the pathway predictions, is that we have laid the groundwork for further studies identifying other enzymes in high value pathways. This work may be particularly relevant to anthocyanin biosynthesis, since different types of anthocyanins have different effects on mammalian biology, possibly due to differences in bioavailability. Another benefit of these data is that because the RNA-Seq data were from an autotetraploid blueberry, aligned onto a diploid genome sequence, it may be possible to use these data to investigate allele-specific gene expression, using SNPs or other polymorphisms to distinguish alleles. This and other studies may become possible thanks to the databases and datasets presented there, and our future work will focus on improving these resources to enable better understanding of metabolic pathways that are active in blueberry.

## Methods

### Analysis code availability

R Markdown files, python scripts, and shell scripts used in data analysis and data processing are version-controlled in a git-based, publicly accessible repository [27]. The repository contains files documenting which subsections of the repository were used in the analyses described above. Note that we will continue to update and modify the code repository to meet the needs of users; however, readers interested in retrieving older versions of the repository that existed at the time of publication can do so using the git program, which is well documented at many sites.

### Expression analysis

The number of reads per gene was counted using the samtools view -c command. All single-mapping reads that overlapped a gene region were counted, and gene regions were defined as the smallest start and largest end position of transcripts annotated to the gene. For

each comparison, the method 'exactTest' with tagwise dispersion from Bioconductor library edgeR [72] was used to identify differentially expressed genes. False discovery rate for each comparison was calculated using the p.adjust method in R, with option 'BH' for Benjamini and Hochberg. For clarity, all results reported here applied an FDR cutoff of 0.001. Code used to perform differential expression analysis is available in the project bitbucket repository; readers interested in exploring other cutoff options can obtain copies of source code files, edit parameters in the file, and re-run the analysis. Gene Ontology terms that were unusually enriched among differentially expressed genes were identified using GOSec [73], also from Bioconductor.

### Analysis of alternative splicing with CuffDiff

The non-redundant gene set was used as input to cuffcompare, which produced a file (cuffcmp.combined.gtf) suitable for use as input to cuffdiff. The cuffdiff was then run using the full set of alignments (both multi- and single-mapping) obtained for the blueberry development and ripening time course dataset with cuffcmp.combined.gtf file as reference gene models. For analysis of splicing differences, the file splicing.diff output by cuffdiff was used. It was observed that the most significant splicing differences reported by cuffdiff all had the same low p-value ( $10^{-5}$ ) and so these were selected for further analysis. A random subset of the splicing differences from genes that were annotated by the BLASTX step above was selected for visual inspection in IGB.

### Analysis of alternative splicing with ArabiTag

The non-redundant gene set, BAM files with read alignments, and junction files produced by the FindJunctions program (Integrated Genome Browser tools package) were provided to the AltSpliceAnalysis software [74], which is based on the ArabiTag algorithm described in [62]. The software identifies annotated alternative splicing events and then uses the RNA-Seq data to count reads supporting alternative splicing choices associated with each event. Data analysis code written in R was used to calculate the percentage of support (%S) for each variant, using gapped reads to support alternative donor or acceptor sites and non-gapped reads to support retention of introns. To identify differential splicing between conditions, a *t*-test was performed on splicing scores. Differential splicing was tabulated and highest confidence differences were manually inspected using Integrated Genome Browser.

### Availability and requirements

**Project name:** BlueberryGenome.

**Project home page:** <https://bitbucket.org/lorainelab/blueberrygenome>.

**Operating system(s):** Platform independent.

**Programming language:** R, python.

**Other requirements:** RStudio, 0.98.507 or higher, R 3.0.3 or higher.

**License:** The MIT License.

### Availability of supporting data

RNA-Seq data are freely available for visualization in Integrated Genome Browser, and analysis code is available from a publicly available git repository [27].

All sequence data are available in the Short Read Archive under accessions SRP039977 and SRP039971. Files containing alignments, RNA-Seq coverage graphs, and output from TopHat2 are available from a publicly accessible IGBQuickLoad site [75] configured to serve data for visualization in Integrated Genome Browser. Data files including gene models and related annotations are available from the git source code repository. The git source code repository also contains R code used to analyze data. Archived snapshots of these files are also available from the *GigaScience* repository, GigaDB [76].

### Additional files

**Additional file 1: Figure S1.** Data processing and analysis pipeline. The computational workflow used to generate blueberry gene models and coding sequences is shown. **(A)** Sources of gene annotations. **(B)** Step-wise selection protocol used to create the nonredundant gene set. **Figure S2.** Expression pattern of candidate ethylene biosynthesis genes. Expression of **(A)** SAM synthase, **(B)** ACC synthase and **(C)** ACC oxidase genes are shown in RPKM. **(D)** Enzymes and intermediates in ethylene biosynthesis pathway. **Figure S3.** Expression pattern for candidate bixin and dhurrin biosynthetic genes. **(A)** Bixin biosynthesis as annotated in PlantCyc as pathway id PWY-5305. The first step is catalyzed by a lycopene cleavage oxygenase, which was not found in the blueberry annotations. Subsequent reactions are catalyzed by enzymes BADH (bixin aldehyde dehydrogenase) and nBMT (norbixin methyltransferase), which were found in blueberry. **(B)** and **(C)** Expression of blueberry genes encoding BADH and nBMT are shown on two plots with different y-axis scales as one gene (CUFF.4231) is highly expressed in fruit. **(D)** Synthesis and breakdown of cyanogenic glycoside dhurrin. **(E-J)** Expression of genes encoding putative dhurrin synthesis and breakdown enzymes. **Figure S4.** Stage-specific alternative promoter and three prime ends. **(A)** Read alignments from mature green and ripe fruit showing an alternative promoter in CUFF.24092, which encodes a conserved protein of unknown function. **(B)** Coverage graphs showing read density at locus CUFF.32196, encoding a conserved protein of unknown function. Graphs represent the number of reads that overlap the base positions indicated on the coordinates axis track. **Figure S5.** RNA-Seq coverage graphs illustrating expression of genes involved in synthesis of proanthocyanidins, anthocyanins, or both. Each image shows alignments of ESTs alongside blueberry gene models. **(A)** CUFF.29797, encoding VcANR, anthocyanidin reductase **(B)** CUFF.20951, encoding VcUFGT **(C)** VcDFR **(D)** VcMYBPA1a **(E)** VcMYBPA1b.

**Additional file 2: Pathway and enzyme annotations for blueberry genes and transcripts.**

**Additional file 3: Gene Ontology (GO) annotations for blueberry genes and transcripts.**

**Additional file 4: GO term enrichment analysis of differentially expressed genes.**

**Additional file 5: GO term enrichment analysis of genes from STEM clusters 0, 1, 2, and 6.**

### Abbreviations

ACC: 1-aminocyclopropane-1-carboxylate; BLAST: Basic Local Alignment Search Tool; BLASTP: Basic Local Alignment Search Tool protein-protein search; BLASTX: Basic Local Alignment Search Tool search against protein databases using translated nucleotide query; cDNA: Complementary deoxyribonucleic acid; CG: Cyanogenic glycoside; DE: Differentially expressed; DNA: Deoxyribonucleic acid; EC: Enzyme Commission; EST: Expressed Sequence Tag; GMAP: Genomic Mapping and Alignment Program; GO: Gene Ontology; HCN: Hydrogen cyanide; IGB: Integrated Genome Browser; MDS: Multi-dimensional scaling; NCBI: National Center for Biotechnology Information; nr: Non-redundant; PA: Proanthocyanidin; RNA: Ribonucleic acid; RPKM: Reads per kilobase of transcript per millions mapped; SAM synthase: Methionine adenosyltransferase.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

AEL designed the study, analyzed data, and wrote the manuscript. ADE collected berry samples, prepared libraries for sequencing, analyzed data, and helped write the manuscript. IB collected berry samples and analyzed data. KP collected blueberry samples, prepared libraries for sequencing, and analyzed data from the 454 sequencing experiments. VG and SUA developed the genome annotation pipeline used to generate blueberry gene models. VG made blueberry gene models, analyzed data, annotated blueberry genes with Gene Ontology terms, and helped write the manuscript. MDM analyzed data. RR and ARB provided early access to the blueberry genome sequence. All authors read and helped edit the manuscript. All authors read and approved the final manuscript.

### Acknowledgments

Jim Ballington and the North Carolina Department of Agriculture Piedmont Research Station provided blueberry plants. Lisa J Rowland donated genomic DNA used in genomic sequencing. Shira Stav and Tyler Estrada helped collect berries used to prepare libraries for RNA sequencing. We thank Nate Watson, Ra'ad Ghabaraih, Adam Baxter, and Cory Brouwer for contributions to early *de novo* transcriptome assembly efforts not reported here. A grant (to AEL) from UNC General Administration provided financial support. Integrated Genome Browser software used in the study was supported by NIH grant 1R01RR032048-01 (to AEL). We also express our thanks to the reviewers whose suggestions enabled us to improve the manuscript.

### Author details

<sup>1</sup>Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, North Carolina Research Campus, Kannapolis, NC 28081, USA.

<sup>2</sup>Centre for Carbohydrate Recognition and Signaling, Department of Molecular Biology and Genetics, Aarhus University, Gustav Wieds Vej 10, 8000 Aarhus C, Denmark. <sup>3</sup>Plants for Human Health Institute, North Carolina State University, North Carolina Research Campus, Kannapolis, NC 28081, USA.

<sup>4</sup>NMRC-Frederick, 8400 Research Plaza, Ft. Detrick, MD 21702, USA.

Received: 2 September 2014 Accepted: 22 January 2015

Published online: 13 February 2015

### References

1. Babu PV, Liu D, Gilbert ER. Recent advances in understanding the anti-diabetic actions of dietary flavonoids. *J Nutr Biochem*. 2013;24(11):1777–89.
2. Del Rio D, Rodriguez-Mateos A, Spencer JP, Tognolini M, Borges G, Crozier A. Dietary (poly)phenolics in human health: structures, bioavailability, and evidence of protective effects against chronic diseases. *Antioxid Redox Signal*. 2013;18(14):1818–92.
3. Vauzour D. Dietary polyphenols as modulators of brain functions: biological actions and molecular mechanisms underpinning their beneficial effects. *Oxid Med Cell Longev*. 2012;2012:914273.
4. US Department of Agriculture. USDA database for the flavonoid content of selected foods, Release 3.1. <http://www.ars.usda.gov/nutrientdata/flav>. Accessed 14 Jan 2015.
5. Yousef GG, Brown AF, Funakoshi Y, Mbeunkui F, Grace MH, Ballington JR, et al. Efficient quantification of the health-relevant anthocyanin and



- phenolic acid profiles in commercial cultivars and breeding selections of blueberries (*Vaccinium* spp.). *J Agric Food Chem*. 2013;61(20):4806–15.
6. Routray W, Orsat V. Blueberries and their anthocyanins: factors affecting biosynthesis and properties. *Compr Rev Food Sci Food Saf*. 2011;10(6):303–20.
  7. Grace MH, Ribnicky DM, Kuhn P, Poulev A, Logendra S, Yousef GG, et al. Hypoglycemic activity of a novel anthocyanin-rich formulation from lowbush blueberry, *Vaccinium angustifolium* Aiton. *Phytomedicine*. 2009;16(5):406–15.
  8. Kay CD, Hooper L, Kroon PA, Rimm EB, Cassidy A. Relative impact of flavonoid composition, dose and structure on vascular function: a systematic review of randomised controlled trials of flavonoid-rich food products. *Mol Nutr Food Res*. 2012;56(11):1605–16.
  9. Prior RL, Wu X. Anthocyanins: structural characteristics that result in unique metabolic patterns and biological activities. *Free Radic Res*. 2006;40(10):1014–28.
  10. Ansari N, Khodagholi F. Natural products as promising drug candidates for the treatment of Alzheimer's disease: molecular mechanism aspect. *Curr Neuropharmacol*. 2013;11(4):414–29.
  11. Spagnuolo C, Cerella C, Russo M, Chateauvieux S, Diederich M, Russo GL. Quercetin downregulates Mcl-1 by acting on mRNA stability and protein degradation. *Br J Cancer*. 2011;105(2):221–30.
  12. Rimando AM, Kalt W, Magee JB, Dewey J, Ballington JR. Resveratrol, pterostilbene, and piceatannol in vaccinium berries. *J Agric Food Chem*. 2004;52(15):4713–9.
  13. Galletta G, Ballington J. Blueberries, cranberries, and lingonberries. In: Janick J, Moore J, editors. *Fruit breeding vol. II. Vine and small fruits*. NY: Wiley; 1996. p. 1–107.
  14. Costich DE, Ortiz R, Meagher TR, Bruederle LP, Vorsa N. Determination of ploidy level and nuclear DNA content in blueberry by flow cytometry. *Theor Appl Genet*. 1992;83:1001–6.
  15. Bian Y, Ballington J, Raja A, Brouwer C, Reid R, Burke M, et al. Patterns of simple sequence repeats in cultivated blueberries (*Vaccinium* section *Cyanococcus* spp.) and their use in revealing genetic diversity and population structure. *Mol Breeding*. 2014;34(2):675–89.
  16. Rowland LJ, Alkharouf N, Darwish O, Ogden EL, Polashock JJ, Bassil NV, et al. Generation and analysis of blueberry transcriptome sequences from leaves, developing fruit, and flower buds from cold acclimation through deacclimation. *BMC Plant Biol*. 2012;12:46.
  17. Zifkin M, Jin A, Ozga JA, Zaharia LI, Scherthaner JP, Gesell A, et al. Gene expression and metabolite profiling of developing highbush blueberry fruit indicates transcriptional regulation of flavonoid metabolism and activation of abscisic acid metabolism. *Plant Physiol*. 2012;158(1):200–24.
  18. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A. Manipulation of FASTQ data with Galaxy. *Bioinformatics*. 2010;26(14):1783–5.
  19. Kim D, Perteau G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36.
  20. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–9.
  21. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. 2005;21(9):1859–75.
  22. Trapnell C, Roberts A, Goff L, Perteau G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7(3):562–78.
  23. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res*. 2004;32(Web Server issue):W309–12.
  24. Majoros WH, Perteau M, Salzberg SL. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics*. 2004;20(16):2878–9.
  25. Borodovsky M, Lomsadze A. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Curr Protoc Bioinformatics*. 2011;Chapter 4:Unit 4.6.1–10.
  26. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, et al. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res*. 2009;20(1):45–58.
  27. Source code repository for blueberry RNA-Seq data analysis and genomic studies. <http://www.bitbucket.org/lorainelab/blueberrygenome>.
  28. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511–5.
  29. Loraine AE, McCormick S, Estrada A, Patel K, Qin P. RNA-Seq of *Arabidopsis* pollen uncovers novel transcription and alternative splicing. *Plant Physiol*. 2013;162(2):1092–109.
  30. Smith SA, Beaulieu JM, Stamatakis A, Donoghue MJ. Understanding angiosperm diversification using small and large phylogenetic trees. *Am J Bot*. 2011;98(3):404–14.
  31. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*. 2007;449(7161):463–7.
  32. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res*. 2011;40(Database issue):D130–5.
  33. Group TAP. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc*. 2009;161(2):105–21.
  34. 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*. 2005;21(18):3674–6.
  35. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, et al. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*. 2012;40(Database issue):D306–12.
  36. Mitsuda N, Ohme-Takagi M. Functional analysis of transcription factors in *Arabidopsis*. *Plant Cell Physiol*. 2009;50(7):1232–48.
  37. Zhang P, Dreher K, Karthikeyan A, Chi A, Pujar A, Caspi R, et al. Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol*. 2010;153(4):1479–91.
  38. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, et al. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform*. 2009;11(1):40–79.
  39. Retamales JB, Hancock JF. Blueberries. *Crop production science in horticulture series*, vol 21. Wallingford, Oxfordshire, UK. Cambridge, MA: CAB; 2012.
  40. McGhie TK, Walton MC. The bioavailability and absorption of anthocyanins: towards a better understanding. *Mol Nutr Food Res*. 2007;51(6):702–13.
  41. Yi W, Akoh CC, Fischer J, Krewer G. Absorption of anthocyanins from blueberry extracts by caco-2 human intestinal cell monolayers. *J Agric Food Chem*. 2006;54(15):5651–8.
  42. Wilmouth RC, Turnbull JJ, Welford RW, Clifton IJ, Prescott AG, Schofield CJ. Structure and mechanism of anthocyanidin synthase from *Arabidopsis thaliana*. *Structure*. 2002;10(1):93–103.
  43. Turnbull JJ, Nagle MJ, Seibel JF, Welford RW, Grant GH, Schofield CJ. The C-4 stereochemistry of leucocyanidin substrates for anthocyanidin synthase affects product selectivity. *Bioorg Med Chem Lett*. 2003;13(21):3853–7.
  44. Reddy AM, Reddy VS, Scheffler BE, Wienand U, Reddy AR. Novel transgenic rice overexpressing anthocyanidin synthase accumulates a mixture of flavonoids leading to an increased antioxidant potential. *Metab Eng*. 2007;9(1):95–111.
  45. Erlund I, Freese R, Marniemi J, Hakala P, Alfthan G. Bioavailability of quercetin from berries and the diet. *Nutr Cancer*. 2006;54(1):13–7.
  46. Manach C, Williamson G, Morand C, Scalbert A, Remesy C. Bioavailability and bioefficacy of polyphenols in humans. I. Review of 97 bioavailability studies. *Am J Clin Nutr*. 2005;81(1 Suppl):230S–42.
  47. Takahashi N, Goto T, Taimatsu A, Egawa K, Katoh S, Kusudo T, et al. Bixin regulates mRNA expression involved in adipogenesis and enhances insulin sensitivity in 3T3-L1 adipocytes through PPARgamma activation. *Biochem Biophys Res Commun*. 2009;390(4):1372–6.
  48. Tibodeau JD, Isham CR, Bible KC. Annatto constituent cis-bixin has selective antimyeloma effects mediated by oxidative stress and associated with inhibition of thioredoxin and thioredoxin reductase. *Antioxid Redox Signal*. 2010;13(7):987–97.
  49. Galindo-Cuspinera V, Rankin SA. Bioautography and chemical characterization of antimicrobial compound(s) in commercial water-soluble annatto extracts. *J Agric Food Chem*. 2005;53(7):2524–9.

50. Siva R, Doss FP, Kundu K, Satyanarayana VSV, Kumar V. Molecular characterization of bixin—An important industrial product. *Ind Crop Prod*. 2010;32(1):48–53.
51. Busk PK, Moller BL. Dhurrin synthesis in sorghum is regulated at the transcriptional level and induced by nitrogen fertilization in older plants. *Plant Physiol*. 2002;129(3):1222–31.
52. Tattersall DB, Bak S, Jones PR, Olsen CE, Nielsen JK, Hansen ML, et al. Resistance to an herbivore through engineered cyanogenic glucoside synthesis. *Science*. 2001;293(5536):1826–8.
53. Krothapalli K, Buescher EM, Li X, Brown E, Chapple C, Dilkes BP, et al. Forward genetics by genome sequencing reveals that rapid cyanide release deters insect herbivory of *Sorghum bicolor*. *Genetics*. 2013;195(2):309–18.
54. Vetter J. Plant cyanogenic glycosides. *Toxicon*. 2000;38(1):11–36.
55. Barceloux DG. Cyanogenic foods (cassava, fruit kernels, and cycad seeds). *DM Dis Mon*. 2009;55(6):336–52.
56. Jenrich R, Trompetter I, Bak S, Olsen CE, Moller BL, Piotrowski M. Evolution of heteromeric nitrilase complexes in Poaceae with new functions in nitrile metabolism. *Proc Natl Acad Sci U S A*. 2007;104(47):18848–53.
57. Gillaspay G, Ben-David H, Gruissem W. Fruits: a developmental perspective. *Plant Cell*. 1993;5(10):1439–51.
58. Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*. 2006;7:191.
59. Birkhold KT, Koch K, Darnell RL. Carbon and nitrogen economy of developing rabbiteye blueberry fruit. *JASHS*. 1992;117(1):139–45.
60. Castrejón ADR, Eichholz I, Rohn S, Kroh LW, Huyskens-Keil S. Phenolic profile and antioxidant activity of highbush blueberry (*Vaccinium corymbosum* L.) during fruit maturation and ripening. *Food Chem*. 2008;109(3):564–72.
61. Jaakola L, Maatta K, Pirttila AM, Torronen R, Karenlampi S, Hohtola A. Expression of genes involved in anthocyanin biosynthesis in relation to anthocyanin, proanthocyanidin, and flavonol levels during bilberry fruit development. *Plant Physiol*. 2002;130(2):729–39.
62. English AC, Patel KS, Loraine AE. Prevalence of alternative splicing choices in *Arabidopsis thaliana*. *BMC Plant Biol*. 2010;10:102.
63. Gulledge AA, Roberts AD, Vora H, Patel K, Loraine AE. Mining Arabidopsis thaliana RNA-seq data with Integrated Genome Browser reveals stress-induced alternative splicing of the putative splicing regulator SR45a. *Am J Bot*. 2012;99(2):219–31.
64. Dhanaraj AL, Alkharouf NW, Beard HS, Chouikha IB, Matthews BF, Wei H, et al. Major differences observed in transcript profiles of blueberry during cold acclimation under field and cold room conditions. *Planta*. 2007;225(3):735–51.
65. Li X, Sun H, Pei J, Dong Y, Wang F, Chen H, et al. De novo sequencing and comparative analysis of the blueberry transcriptome to discover putative genes related to antioxidants. *Gene*. 2012;511(1):54–61.
66. Strickler SR, Bombarely A, Mueller LA. Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *Am J Bot*. 2012;99(2):257–66.
67. Bhargava A, Clabaugh I, To JP, Maxwell BB, Chiang YH, Schaller GE, et al. Identification of cytokinin-responsive genes using microarray meta-analysis and RNA-Seq in Arabidopsis. *Plant Physiol*. 2013;162(1):272–94.
68. Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, et al. Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics*. 2011;12:131.
69. Polashock J, Zelzion E, Fajardo D, Zalapa J, Georgi L, Bhattacharya D, et al. The American cranberry: first insights into the whole genome of a species adapted to bog habitat. *BMC Plant Biol*. 2014;14:165.
70. Srinivasasainagendra V, Page GP, Mehta T, Coulibaly I, Loraine AE. CressExpress: a tool for large-scale mining of expression data from Arabidopsis. *Plant Physiol*. 2008;147(3):1004–16.
71. Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, et al. Transcriptional coordination of the metabolic network in Arabidopsis. *Plant Physiol*. 2006;142(2):762–74.
72. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–40.
73. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*. 2010;11(2):R14.
74. Alternative splicing analysis source code repository. <http://www.bitbucket.org/lorainelab/altspliceanalysis>.
75. Blueberry IGBQuickLoad site. <http://www.igbquickload.org/blueberry>.
76. Gupta V, Estrada AD, Blakley IC, Reid R, Patel K, Meyer MD, et al. Supporting material for RNA-Seq analysis and annotation of a draft blueberry genome assembly. *GigaScience Database* 2015. <http://dx.doi.org/10.5524/100117>

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

