

# High-throughput sequencing, assembly, and annotation of expressed sequences from blueberry fruit

Ketan Patel<sup>1,2</sup>, Gad Yousef<sup>1</sup>, Mary Grace<sup>1</sup>, Flaubert Mbeunkui<sup>1</sup>, Allan Brown<sup>1</sup>, Mary Ann Lila<sup>1</sup>, Ann Loraine<sup>1,2</sup>

<sup>1</sup>Plants for Human Health Institute, 600 Laureate Way, Suite 1329, North Carolina Research Campus, NC State University, Kannapolis, NC 28081

<sup>2</sup>Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223

## Abstract

Blueberries contain beneficial compounds that protect against disease, support neural function in the face of aging or environmental insult, and enhance vascular health. We are studying metabolic pathways responsible for production of health-protective bioactive compounds in blueberries using approaches from natural products chemistry and next-generation sequencing. To identify genes required for production and regulation of beneficial compounds in berry fruit, we performed deep sequencing of unripe and mature blueberry fruit cDNAs using the Roche FLX-Ti platform. We performed a single plate of sequencing, yielding more than 440,000 reads from each sample type, with median and average reads lengths of about 300 bases respectively for each. Clustering using Roche software yielded 5,700 EST clusters with consensus sequences sized 800 bases or larger. Comparable clustering of the 5,134 blueberry ESTs currently available from dbEST yielded around 821 clusters and 2,509 singleton ESTs. This new data set represents a significant advance toward discovery and characterization of the expressed gene repertoire of blueberry. Results from homology-based annotation of the new 454 blueberry ESTs and cluster consensus sequences will be presented.

## Methods

### Plant material

Fruits of O'Neal<sup>1</sup>, a Southern Highbush Blueberry, were used as starting material for transcriptome sequencing. Two different cDNA libraries were sequenced, corresponding to a late green fruit growth stage (immature) and >75% blue (fully ripe, harvest) fruit growth stage.

### RNA isolation and cDNA synthesis

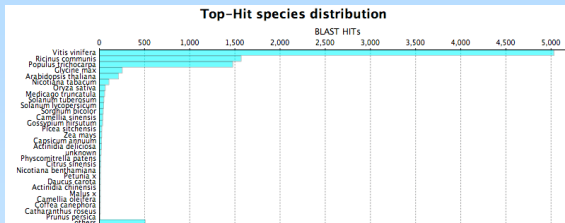
Total RNA was extracted from unripe and mature blueberry fruit using the Spectrum Plant Total RNA Kit from Sigma-Aldrich. Contaminating genomic DNA was removed by DNase I treatment. Total RNA was quantified using a Nanodrop and RNA integrity was assessed by running 1-2 ug on 1.2% formaldehyde-agarose gel.

Poly A+ RNA was purified using the Oligotex mRNA Minikit from Qiagen and used for first strand synthesis. The SMART PCR cDNA Synthesis Kit from Clontech was used to construct the ds cDNA library with the following modifications. A modified 3' SMART CDS Primer IIA primer was used for first strand synthesis. Second strand synthesis and cDNA amplification were performed using the 5' PCR Primer II A and a modified 3' PCR to break up the poly A tail. cDNA fragment length and quality were checked on a 1.2% agarose gel before submitting for 454 pyrosequencing<sup>2</sup>.

### Data assembly and functional annotation

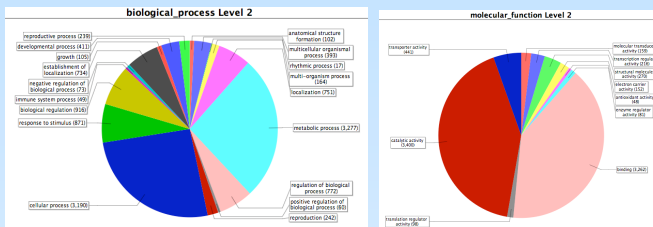
454 cDNA reads were assembled using Roche GS De Novo Assembler Software and assemblies from 454 cDNA reads for functional annotation was performed using Blast2GO<sup>3</sup>.

## Annotation Results



**Figure 1. Top-Hit species distribution**

Contigs (>500bp) generated by merging both libraries were searched against the NCBI non-redundant (nr) protein database using blastx algorithm. 92% (9748/10,575) had at least one significant alignment to existing proteins in the protein database (E-value cutoff, 1.0E-4), with grape having the most "hits".



**Figure 2. GO categories from blueberry fruit libraries**

Functional annotation of the contigs is based on Gene Ontology (GO) vocabulary using Blast2GO and shows to cover a broad range of GO categories.

### Contigs with ESTs mostly from ripe berries

Contig Name	U/R ESTs	Annotation
contig29530	0.10	NC domain-containing, C:mitochondrion
contig07886	0.12	F:beta-amylase activity
contig29517	0.15	HMG Co-A reductase, P:isoprenoid biosynthetic process
contig28563	0.17	tubulin-like protein
contig01320	0.18	F:transcription factor activity
contig27698	0.18	HMG Co-A reductase, P:isoprenoid biosynthetic process
contig27608	0.19	P:response to stress P:response to chemical stimulus
contig26946	0.19	pathogenesis-related thaumatin-like protein
contig27220	0.19	proline dehydrogenase
contig25610	0.19	F:oxidoreductase activity
contig02417	0.20	non-symbiotic hemoglobin class 1
contig27837	0.21	alpha-expansin 4, P:plant-type cell wall organization
contig28216	0.22	embryo-specific protein 1
contig29113	0.22	band 7 family protein
contig03049	0.22	C:extracellular region C:cell wall
contig28266	0.22	P:response to water deprivation P:regulation of transcription
contig29621	0.22	chloroplast small heat shock protein
contig03062	0.22	cytochrome c oxidase polypeptide vc
contig10312	0.22	proline dehydrogenase
contig09025	0.23	P:response to chitin P:transcription
contig00152	0.23	protein kinase

**Table 3. Relative transcript abundance**

Relative abundance ratios are determined by counting the number of reads for each contig from the two libraries.

### Contigs with ESTs mostly from unripe berries

Contig Name	U/R ESTs	Annotation
contig09145	9.6	auxin-responsive protein
contig14086	6.4	zinc finger - F:nucleic acid binding P:regulation of transcription
contig15924	6.3	cellulose synthase
contig10196	5.8	F:hydrolase activity
contig23994	5.3	wr3 (wound-responsive 3) nitrate transmembrane transporter
contig06978	5.3	F:desacetoxyvindoline 4-hydroxylase activity
contig01156	5.1	nadh dehydrogenase subunit 1
contig09146	4.7	P:reductive pentose-phosphate cycle F:phosphoribulokinase activity C:chloroplast
contig06806	4.7	P:chlorophyll biosynthetic process C:magnesium chelatase complex
contig03567	4.5	F:asparaginase activity P:glycoprotein catabolic process F:peptidase activity
contig08264	4.5	amino acid binding
contig00621	4.5	F:serine-type endopeptidase activity P:plant-type cell wall modification
contig08939	4.4	P-protein, F:glycine dehydrogenase (decarboxylating) activity
contig01327	4.4	F:glutathione transferase activity F:acylglutathione lyase activity
contig01328	4.4	F:oxidoreductase activity
contig06705	4.3	P:oxidation reduction F:peroxidoxin activity C:chloroplast thylakoid
contig09549	4.3	RNA-binding protein, P:defense response to bacterium P:response to cold
contig29427	4.3	glycine hydroxymethyltransferase
contig01736	4.3	PREDICTED: hypothetical protein [Vitis vinifera]
contig04309	4.2	cysteine protease
contig01870	4.1	lysine histidine transporter
contig26228	4.1	chloroplast cu zn superoxide dismutase
contig02367	4.1	beta-galactosidase protein 1
contig07927	4.1	chloroplast inner envelope
contig26581	4.2	beta xylosidase

## Conclusion

This work characterizes a blueberry fruit transcriptome using 454 sequencing. Comparative sequencing of two different stages of fruit development cDNA libraries has provided information on variation of gene expression in addition to the identification of many genes potentially involved in the regulation/production of bioactive compounds. Functional annotation shows that the 454 sequences cover a broad range of GO categories. Additional 454 sequencing and metabolic profiling should lead to the identification of genes and metabolic pathways responsible for the production and regulation of beneficial bioactives in blueberries.

## References

<sup>1</sup>Plant material was provided by Dr. Jim Ballington (NCSU) and plants were grown at NCDA&CS Piedmont Research Station (Salisbury, NC)

<sup>2</sup>454 pyrosequencing was performed at David H Murdock Research Institute (DHMRI) – <http://www.dhMRI.org/>

<sup>3</sup>Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón and Montserrat Robles Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005 21: 3674-3676

This work was funded by University of North Carolina General Administration.