BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database

Volume 9(17)

# BBGD454: A database for transcriptome analysis of blueberry using 454 sequences

**Omar Darwish[1], L Jeannine Rowland[2] & Nadim W Alkharouf[1]***

[1]Department of Computer and Information Sciences, Towson University, Towson, MD 21252, USA; [2]Genetic Improvement of Fruits and Vegetables Laboratory, USDA-ARS, Beltsville, MD 20705, USA; Nadim W Alkharouf - Email: nalkharouf@towson.edu; *Corresponding author

Received September 30, 2013; Accepted September 30, 2013; Published October 16, 2013

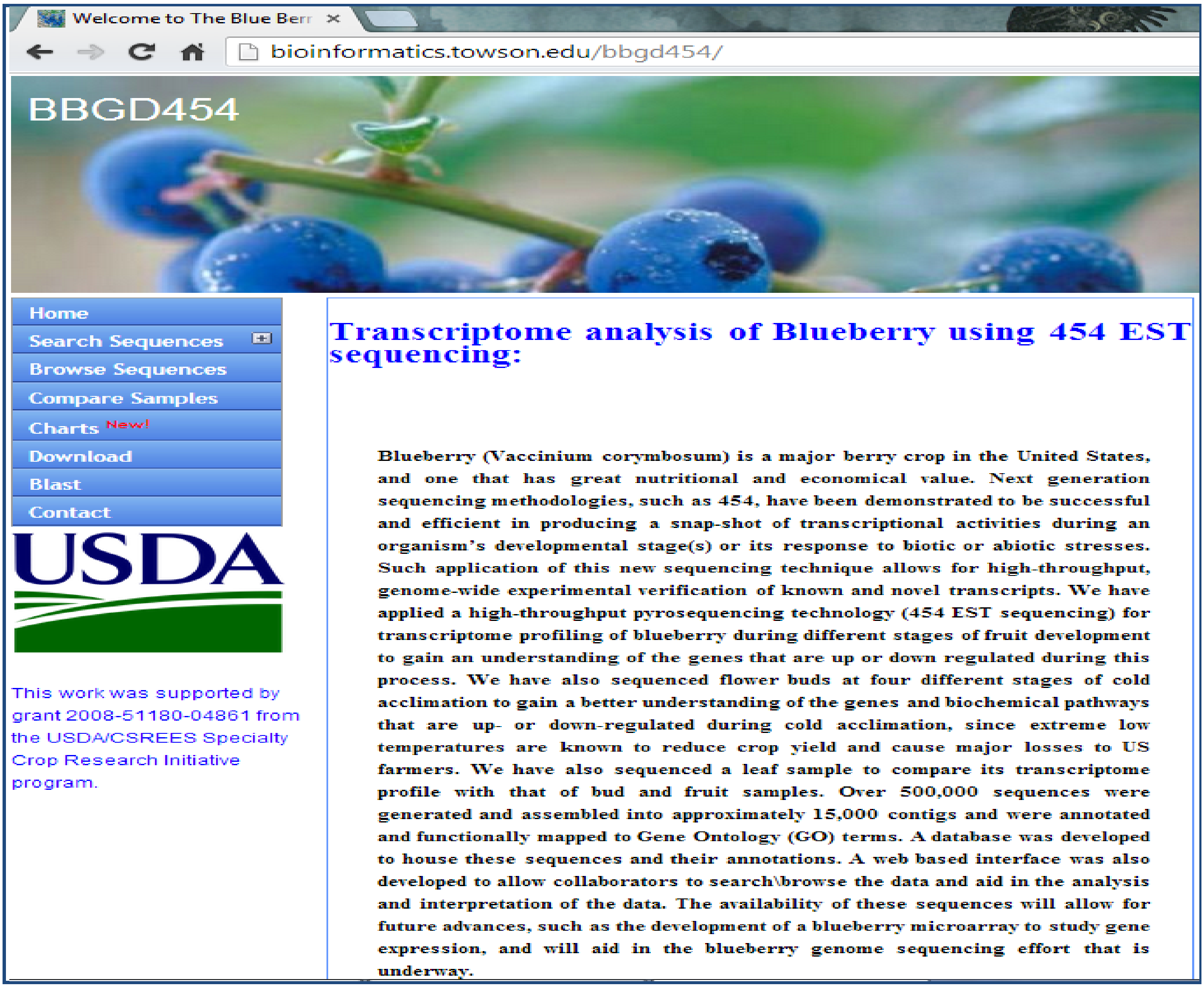
**Abstract:**
Blueberry is an economically and nutritionally important small fruit crop, native to North America. As with many crops, extreme low temperature can affect blueberry crop yield negatively and cause major losses to growers. For this reason, blueberry breeding programs have focused on developing improved cultivars with broader climatic adaptation. To help achieve this goal, the blueberry genomic database (BBGD454) was developed to provide the research community with valuable resources to identify genes that play an important role in flower bud and fruit development, cold acclimation and chilling accumulation in blueberry. The database was developed using SQLServer2008 to house 454 transcript sequences, annotations and gene expression profiles of blueberry genes. BBGD454 can be accessed publically from a web-based interface; this website provides search and browse functionalities to allow scientists to access and search the data in order to correlate gene expression with gene function in different stages of blueberry fruit ripening, at different stages of cold acclimation of flower buds, and in leaves.

**Availability**: It can be accessed from: http://bioinformatics.towson.edu/BBGD454/

**Key words:** Blueberry, database, 454, transcriptomics.

## Background:

Blueberry (*Vaccinium corymbosum*) is one of the major berry crops grown in the United States **[1]**. North America, in fact, is the world's leading blueberry producer, accounting for nearly 90% of world production at the present time. Total area devoted to growing commercial blueberries in North America is approximately 74,000 hectares. Blueberry is a high value crop, often times grown in acidic and imperfectly drained soils that would otherwise be considered unfit for agricultural production **[1].** Blueberry is also an important fruit crop because of its nutritional value. Of all fresh fruits and vegetables, blueberries are one of the richest sources of antioxidants **[2, 3, 4].** Blueberry is a model organism for the heath family *Ericaceae*, which also includes the economically important, closely related cranberry as well as the economically important, more distantly related ornamentals, rhododendron, azalea, and mountain laurel. For all these related species, genomic studies, including EST generation and microarray analyses, are lacking or completely absent. Functional genomic studies on berry crops are lacking, especially studies dealing with the molecular impacts of low temperature on berry crop yield and chilling accumulation. Low temperature extremes reduce blueberry yields and impact the profitability and competitiveness of U.S. producers. Enhanced cold tolerance during the winter and early spring of elite varieties would be of great value to the blueberry industry. New low-chilling varieties are also desirable for the southern U.S. BBGD454 is a public database that houses 454 transcriptome sequences from numerous blueberry cDNA library samples. The ultimate goal of these experiments is to better understand the genetic control of cold hardiness, chilling requirement, and various fruit quality traits and apply the information to develop new cultivars for the blueberry industry.


Welcome to The Blue Berr
bioinformatics.towson.edu/bbgd454/
BBGD454
Home
Search Sequences
Browse Sequences
Compare Samples
Charts New!
Download
Blast
Contact
USDA
This work was supported by grant 2008-51180-04861 from the USDA/CSREES Specialty Crop Research Initiative program.

Transcriptome analysis of Blueberry using 454 EST sequencing:

Blueberry (Vaccinium corymbosum) is a major berry crop in the United States, and one that has great nutritional and economical value. Next generation sequencing methodologies, such as 454, have been demonstrated to be successful and efficient in producing a snap-shot of transcriptional activities during an organism's developmental stage(s) or its response to biotic or abiotic stresses. Such application of this new sequencing technique allows for high-throughput, genome-wide experimental verification of known and novel transcripts. We have applied a high-throughput pyrosequencing technology (454 EST sequencing) for transcriptome profiling of blueberry during different stages of fruit development to gain an understanding of the genes that are up or down regulated during this process. We have also sequenced flower buds at four different stages of cold acclimation to gain a better understanding of the genes and biochemical pathways that are up- or down-regulated during cold acclimation, since extreme low temperatures are known to reduce crop yield and cause major losses to US farmers. We have also sequenced a leaf sample to compare its transcriptome profile with that of bud and fruit samples. Over 500,000 sequences were generated and assembled into approximately 15,000 contigs and were annotated and functionally mapped to Gene Ontology (GO) terms. A database was developed to house these sequences and their annotations. A web based interface was also developed to allow collaborators to search\browse the data and aid in the analysis and interpretation of the data. The availability of these sequences will allow for future advances, such as the development of a blueberry microarray to study gene expression, and will aid in the blueberry genome sequencing effort that is underway.

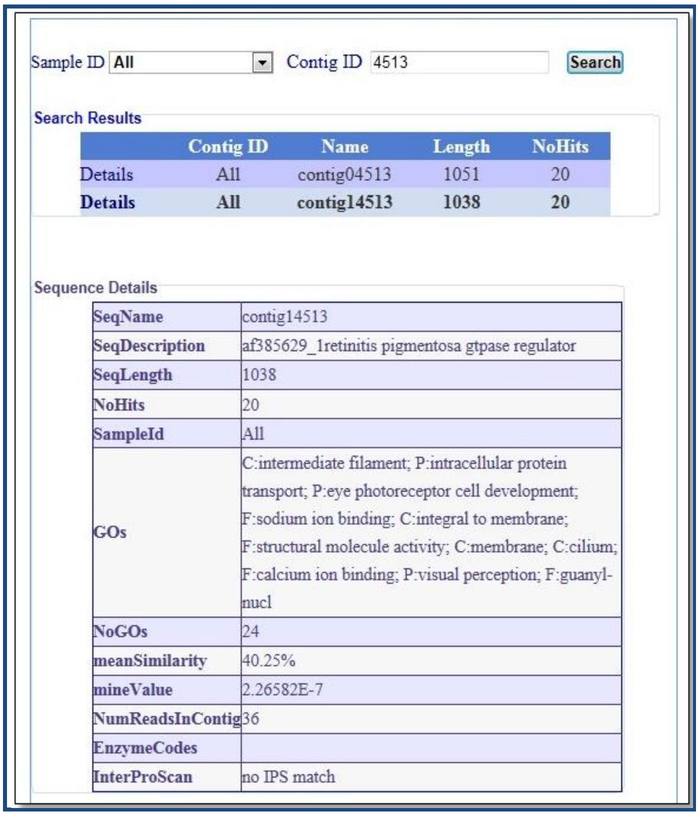
**Figure 1:** A snapshot of BBGD454 main web page, shows a quick summary about the project and the functions this website provides.

**Methodology:**

BBGD454 database was designed, implemented and hosted using Microsoft SQL Server 2008 Enterprise Edition. Microsoft Visual Studio 2008 was used to design and implement the web pages, which were programmed using ASP.NET framework 2.35 with C# programming language. Real time charts were built using a charting library called Highcharts written in pure JavaScript. Both the database and the website are on the same server at Towson University in Baltimore, MD, USA. This server is running Microsoft Windows Server 2003 and Internet Information Services (IIS V6.0). BBGD454 stores unique putative transcripts (contigs and singletons) that resulted from assembling the normalized, corrected and filtered 454 sequences using the GS DeNovo Assembler (454 Life Sciences, Roche). In addition to the sequences, BBGD454 also houses information on taxonomy, gene function, tissue specificity, gene ontology, and blast results of those contigs with high number of reads in one library against all the other libraries. All blast results were obtained using Blast2Go **[5].**

**Utility to the biological community:**

The database contains 454 sequences from nine cDNA libraries that were constructed from mRNA from various organs collected from plants of the highbush blueberry (*V. corymbosum*) variety Bluecrop. Organs included young, fully expanded leaves, flower buds collected at various stages of cold acclimation (0, 397, 789, and 1333 chill units), and fruit collected at various stages of ripening (green, white, pink, and blue) **[6].** The nine libraries (leaves, buds at four different stages of cold acclimation, and fruit at four different stages of ripening) were multiplexed and sequenced on two plate runs of the 454-GS FLX Titanium platform. A summary of the sequencing and assembly results is shown in **Table 1 (see supplementary material)**. Overall, 1,348,819 reads were generated, with an average read length of 287 nucleotides (nt). This yielded a total of ~390 megabases of cDNA sequence. The BBGD454 web-accessible interface **(Figure 1)** provides an easy way to search, browse and download the sequences and expression data stored in the database. The following are the main functions the website provides:



Sample ID All Contig ID 4513 Search

Search Results

| | Contig ID | Name | Length | NoHits |
|---|---|---|---|---|
| Details | All | contig04513 | 1051 | 20 |
| **Details** | **All** | **contig14513** | **1038** | **20** |

Sequence Details

| | |
|---|---|
| SeqName | contig14513 |
| SeqDescription | af385629_1retinitis pigmentosa gtpase regulator |
| SeqLength | 1038 |
| NoHits | 20 |
| SampleId | All |
| GOs | C:intermediate filament; P:intracellular protein transport; P:eye photoreceptor cell development; F:sodium ion binding; C:integral to membrane; F:structural molecule activity; C:membrane; C:cilium; F:calcium ion binding; P:visual perception; F:guanyl-nucl |
| NoGOs | 24 |
| meanSimilarity | 40.25% |
| mineValue | 2.26582E-7 |
| NumReadsInContig | 36 |
| EnzymeCodes | |
| InterProScan | no IPS match |

**Figure 2:** A snapshot of the Search by Contig ID page. Users can use the Search pages (By Contig Id, By Sequence Description) to search for the sequences using any part of their IDs or description.

### *(i) Search*

Users can search by contig number, or by sequence description **(Figure 2).** Partial numbers and/or characters can be used if one is not sure of the contig number or full gene name.  Both the search by contig and search by description return their results in a nice tabular format that allows the user to select any record of the returned search results to see details about that specific sequence. The information includes sequence name, sequence description, sequence length, number of hits, gene ontolgy, RPKM **[7]**, number of reads, and more.

### *(ii) Browse*

The webpage displays a table of all the blueberry cDNA samples. The user can click on any of the samples to see all sequences of that specific sample in a table that can be sorted based on many sequence attributes (id, length, description, etc). Sequence details can also be seen from this page by clicking on details next to any sequence id (contig number).

### *(iii) Compare samples*

Users can compare the expression of genes between two different libraries from the web site. This should help identify genes that are potentially differentially expressed during cold acclimation and chilling accumulation and during fruit development.

### *(iv) Blast*

We have a Blast application on the site that provides the functionality of blasting a sequence of interest against the sequences stored in BBGD454, the Sanger ESTs stored in BBGD (http://bioinformatics.towson.edu/bbgd/) and/or a collection of sequence data comprising EST and genomic sequences from all plant species, kingdom Viridiplantae.

## Caveats:

The sequences were de-novo assembled, which may have resulted in a number of false positives. Once the genome of blueberry is completely sequenced, one can redo the analysis, using these sequences, and aligning them to the genome. This would yield more accurate results.

## Future Development:

A web browser aligning the reads to the genome of blueberry would be extremely beneficial. This is something we might consider implementing once the genome is published and made available.

## Authors' contributions:

Omar Darwish, a doctoral student in IT, designed and developed the database and user interface, under the guidance of Nadim Alkharouf at Towson University. L. Jeannine Rowland was the PI of the Blueberry transcriptomic project at the USDA and was in charge of library preparation as well as the overall design of the experiments. All authors contributed to the writing of the manuscript.

## References:

**[1]** Yarborough D, Factors contributing to the increase in productivity in the wild blueberry industry. Small Fruits Rev 2004 **3**: 33

**[2]** Nile SH *et al. Nutrition*. 2013 **9007**: 00220 [PMID: 24012283]

**[3]** Cho E *et al. Arch Ophthalmol*. 2004 **122:** 883 [PMID: 15197064]

**[4]** Kalt W *et al. Journal-American Pomological Society*. 2007 **61**: 151

**[5]** Conesa A *et al. Bioinformatics* 2005 **21**: 3674 [PMID: 16081474]

**[6]** Rowland L *et al. BMC plant biology*. 2012 **12**: 46 [PMID: 22471859]

**[7]** Mortazavi A *et al. Nature methods*. 2008 **5**: 621 [PMID: 18516045]

**Edited by P Kangueane**

**Citation**: **Darwish *et al.*** Bioinformation 9(17): 883-886 (2013)

## Supplementary material:

**Table 1:** Summary of the sequence assemblies housed on BBGD454

| Sample | Total number of reads assembled | Total number of reads in contigs | Number of contigs/singletons | Average contig/singleton length (nt)[a] |
|---|---|---|---|---|
| Flower bud 0′ (low temperature exposure)[b] | 69,943 | 43,073 | 2,675 / 26,870 | 804 / 323 |
| Flower bud 397′ | 74,169 | 39,999 | 2,751 / 34,170 | 760 / 306 |
| Flower bud 789′ | 69,874 | 37,681 | 2,645 / 32,193 | 785 / 319 |
| Flower bud 1333′ | 72,733 | 41,836 | 2,421 / 30,897 | 796 / 302 |
| **All flower bud samples** | **291,342** | **228,938** | **10,350 / 62,404** | **898 / 280** |
| Green fruit | 73,168 | 46,708 | 2,241 / 26,460 | 720 / 284 |
| White fruit | 69,260 | 42,682 | 2,029 / 26,578 | 700 / 298 |
| Pink fruit | 68,767 | 43,975 | 1,964 / 24,792 | 750 / 297 |
| Blue/Ripe fruit | 59,622 | 37,615 | 1,941 / 22,007 | 819 / 311 |
| **All berry samples** | **259,527** | **199,643** | **6,726 / 59,884** | **818 / 267** |
| Leaves and stems | 62,465 | 36,763 | 1,781 / 25,702 | 771 / 298 |
| **All samples** | **614,028** | **490,517** | **14,764 / 123,511** | **933 / 253** |