DEVELOPMENT OF TOOLS FOR SOYKB: SNPviz, DIFFERENTIAL EXPRESSION AND
PROTEIN BIOVIEWER

A Thesis

Presented to

The Faculty of the Graduate School

At the University of Missouri

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

By

HONGXIN ZHANG

Dr. Dong Xu, Advisor

JULY 2014

The undersigned, appointed by the dean of the Graduate School, have examined the
thesis entitled

DEVELOPMENT OF TOOLS FOR SOYKB: SNPviz, DIFFERENTIAL EXPRESSION AND
PROTEIN BIOVIEWER

Presented by Hongxin Zhang

A candidate for the degree of

Master of Science

And hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Dong Xu

Dr. Trupti Joshi

Dr. Toni Kazic

ACKNOWLEDGEMENTS

Firstly, I want to sincerely thank my advisor Dr. Dong Xu for his constant help during my Master's program. I really appreciate his patience, especially when he answers my questions with full details, clarity and profundity. Because of his kind offer to let me join the Digital Biology Lab, I found my academic and career interest, and friendship with lab members. He taught me how to think critically and how to investigate hypotheses.

I also would love to wholeheartedly thank Dr. Trupti Joshi, who endowed me with abundant bioinformatics knowledge. During my Master's program, for most of the time, I was working with her on SoyKB development, which is the main focus of this paper. She is my friend and also my mentor of bioinformatics. Not being afraid to ask question is the most important thing she taught me.

I also want to express my gratitude to Dr. Toni Kazic for giving me confidence when I doubted myself. Her care rescued me from my self-abnegation brought by language barriers and reestablished my self-esteem and determination to overcome obstacles.

Besides, I would like to thank all Digital Biology Lab members. I respectfully send my thanks to Shiyuan Chen, Yang Liu, Qiuming Yao, and Li Song who helped me answer questions in my study and shared their valuable working and life experience with me.

My genuine thanks would be extended to Mrs. Jodette Lenser as well for her academic advices and information about the program.

At the end, I would love to thank my parents and my friends who accompany me through the peaks and valleys, and encourage me to become a cutting-edge developer.

TABLE OF CONTENTS

List of figures	v
List of tables.....	viii
1 . Introduction	1
2 . Gene Model	3
3 . SNPviz	7
3.1 Introduction	8
3.2 Rationale Behind Features	8
3.2.1 <i>Input Options</i>	8
3.2.2 <i>Display Options</i>	10
3.2.3 <i>Phylogeny Clustering</i>	11
3.2.4 <i>Integration of Datasets</i>	11
3.2.5 <i>Output Options</i>	11
3.3 Demonstration	12
3.3.1 <i>Stand-alone version</i>	12
3.3.2 <i>Web Version</i>	20
3.4 Design Methods	23
3.4.1 <i>Genetic Distance between Samples</i>	23
3.4.2 <i>Tree Drawing Methods</i>	23
3.4.3 <i>Double Nucleotide Transfer to Single Nucleotide</i>	24
3.4.4 <i>Sample Combining</i>	24
3.5 Testing and Validation	27
3.5.1 <i>White Box Testing</i>	27
3.5.2 <i>Black Box Testing</i>	29
3.5.3 <i>Validation</i>	31
3.6 Result	32
4 . Differential Expression	34
4.1 Introduction	34
4.2 Functions.....	34
4.2.1 <i>Gene Lists</i>	35

4.2.2	<i>Venn Diagram</i>	36
4.2.3	<i>Volcano Plot</i>	37
4.2.4	<i>Function Analysis and Pathway Analysis</i>	39
4.2.5	<i>Pathway Information</i>	40
4.2.6	<i>Enrichment Analysis</i>	41
4.3	Methods	44
5	Protein Bioviewer	46
5.1	Introduction	46
5.2	Requirement Analysis	47
5.3	System Design	52
5.4	Demonstration	57
6	Summary	59
	Bibliography	60

LIST OF FIGURES

FIGURE 2.1 (A) NEW ADDED GENE IN VERSION 9.0 (B) DELETED GENE IN VERSION 9.0	5
FIGURE 2.2 CHANGED GENE IN VERSION 1.0.....	6
FIGURE 2.3 CHANGED GENE IN VERSION 9.0.....	6
FIGURE 3.1 STAND-ALONE SNPviz WORKFLOW	12
FIGURE 3.2 MAIN INTERFACE OF STAND-ALONE SNPviz.....	13
FIGURE 3.3 STAND-ALONE SNPviz INPUT DATASET INTERFACE.....	14
FIGURE 3.4 STAND-ALONE SNPviz MODIFY SAMPLE NAMES ONE.....	15
FIGURE 3.5 STAND-ALONE SNPviz MODIFY SAMPLE NAMES TWO	15
FIGURE 3.6 STAND-ALONE SNPviz DELETE SNP DATASETS INTERFACE.....	16
FIGURE 3.7 STAND-ALONE SNPviz CLUSTERING INTERFACE	17
FIGURE 3.8 STAND-ALONE SNPviz CLUSTERING RESULT WITHOUT NUCLEOTIDE BASES ON SAMPLES. SHOW GENE INFORMATION SELECTED, CHOSE FROM 31GENOMES AND SOJA DATASETS, RANGE: 90000-100000000. THREE-COLOR SCHEMA SELECTED.	18
FIGURE 3.9 STAND-ALONE SNPviz CLUSTERING RESULT WITH NUCLEOTIDE BASES ON SAMPLES. SHOW GENE INFORMATION, CHOSE FROM 31GENOMES AND SOJA DATASETS, RANGE: 90000-100000000. MULTI-COLOR SCHEMA SELECTED.	19
FIGURE 3.10 WEB SNPviz MAIN INTERFACE.....	20
FIGURE 3.11 AVAILABLE DATABASES ON SOYKB WITH WEB SNPviz.....	21
FIGURE 3.12 A) WEB SNPviz WITH THREE SAMPLES. B) WEB SNPviz WITH SEVEN SAMPLES	22
FIGURE 3.13 UPGMA TESTING EXAMPLE 1, EDWARDS RESULT[8]	28
FIGURE 3.14 UPGMA TESTING EXAMPLE 1, SNPviz RESULT.....	28
FIGURE 3.15 UPGMA TESTING EXAMPLE 2, OPPEERDOES[21] RESULT	29
FIGURE 3.16 UPGMA TESTING EXAMPLE 2, SNPviz RESULT.....	29

FIGURE 4.1 DIFFERENTIAL EXPRESSION – DATASET SELECTION PAGE.....	35
FIGURE 4.2 DIFFERENTIAL EXPRESSION CONDITION SELECTION PAGE	35
FIGURE 4.3 DIFFERENTIAL EXPRESSION GENE LISTS TAB VIEW	36
FIGURE 4.4 DIFFERENTIAL EXPRESSION VENN TAB VIEW	37
FIGURE 4.5 DIFFERENTIAL EXPRESSION VOLCANO TAB VIEW	38
FIGURE 4.6 DIFFERENTIAL EXPRESSION FUNCTION TAB ANNOTATION DATASETS SELECTION VIEW	39
FIGURE 4.7 DIFFERENTIAL EXPRESSION FUNCTION TAB WITH PFAM OUTPUT	40
FIGURE 4.8 DIFFERENTIAL EXPRESSION PATHWAY TAB VIEW	41
FIGURE 4.9 DIFFERENTIAL EXPRESSION ENRICHMENT ANALYSIS TAB VIEW	41
FIGURE 4.10 ENRICHMENT ANALYSIS INPUT PAGE.....	42
FIGURE 4.11 GENEMERGE INPUT FILE	43
FIGURE 4.12 USER ACCOUNT JOBS LIST	43
FIGURE 4.13 ENRICHMENT ANALYSIS RESULT DISPLAY PAGE.....	43
FIGURE 5.1 PROTEIN BIOVIEWER MAIN PAGE	47
FIGURE 5.2 AMINO ACID WITH THREE-LETTER AND ONE-LETTER MAPPING TABLE [33]	48
FIGURE 5.3 3D PROTEIN SECONDARY STRUCTURES	49
FIGURE 5.4 THE INSPIRATION OF PROTEIN BIOVIEWER	49
FIGURE 5.5 AMINO ACIDS CHARACTERISTICS MAPPING TABLE.....	50
FIGURE 5.6 GLYMA15G08990.1 TMHMM WEB OUTPUT.....	51
FIGURE 5.7 PROTEIN BIOVIEWER WORKFLOW	52
FIGURE 5.8 SECONDARY STRUCTURE ANALYSIS WORKFLOW.....	53
FIGURE 5.9 INPUT FILE	54
FIGURE 5.10 OUTPUT FILE.....	54

FIGURE 5.11 SCHEMATIC REPRESENTATION OF TRANSMEMBRANE PROTEINS: 1. A SINGLE TRANSMEMBRANE A-HELIX (BITOPIC MEMBRANE PROTEIN) 2. A POLYTOPIC TRANSMEMBRANE A- HELICAL PROTEIN 3. A POLYTOPIC TRANSMEMBRANE B-SHEET PROTEIN.....	55
FIGURE 5.12 TMHMM ANALYSIS WORKFLOW	56
FIGURE 5.13 GLYMA15G08990.1 OUTPUT. A) ACROSS TRANSMEMBRANE INTO PROTEIN B) ACROSS TRANSMEMBRANE LEAVE PROTEIN	57
FIGURE 5.14 PROTEIN BIOVIEWER GENE INPUT PAGE	58
FIGURE 5.15 THE PROTEIN BIOVIEWER LINK FROM GENE PAGE	58

LIST OF TABLES

TABLE 3.1 MULTI-COLOR SCHEMA COLOR MAPPING BETWEEN SYMBOLS AND COLORS.....	19
TABLE 3.2 SAMPLE COMBINING STEP 1	25
TABLE 3.3 SAMPLE COMBINING STEP 2	25
TABLE 3.4 SAMPLE COMBINING STEP 3	25
TABLE 3.5 SAMPLE COMBINING STEP 4	26
TABLE 3.6 SAMPLE COMBINING STEP 5	26
TABLE 3.7 SAMPLE COMBINING STEP 6, INDEX AND POSITION MAPPING BEFORE SORTING	26
TABLE 3.8 SAMPLE COMBINING STEP 6, INDEX AND POSITION MAPPING AFTER SORTING	26
TABLE 3.9 SAMPLE COMBINING STEP 7	26
TABLE 3.10 DECISION TABLE IN BLACK BOX TESTING	31
TABLE 5.1 CHARACTERISTICS AND COLOR MAPPING RELATIONSHIP WITH OUTPUT EXAMPLE.....	50
TABLE 5.2 SSPRO4 OUTPUT LETTER AND SECONDARY WEB OUTPUT MAPPING RELATIONSHIP	54
TABLE 5.3 GLYMA15G08990.1 TRANSMEMBRANE HELIX INFORMATION	56

ABSTRACT

Soybean Knowledge Base (SoyKB) is a comprehensive website, which contains lots of web sources for soybean. By integrating multi omics data and solving data storage problem, SoyKB is able to build connection between multiple datasets. Lots of hands-on tools has been designed and developed into SoyKB. And three of them, SNPviz, Differential Expression and Protein BioViewer, will be introduced in detailed methodology and function demonstration.

SNPViz is a simple-to-use SNP haplotype viewer. Its variety of options to view haplotypes from designated gene areas to larger chromosomal areas makes users able to scrutinize on different levels with ease. Its dataset selection function increases its convenience to use as it allows users to put in SNP-array files along with annotation files. SNPviz collects all the samples and composes a phylogeny tree according to the identity of haplotypes.

Differential Expression is for visualizing all kinds of data including Transcriptomics RNA-Seq, Microarray, Metabolites and Proteomics data, which all come from public sources and collaborators. It features six functions: Gene Lists, Venn Diagram, Volcano Plot, Function Information, Path Information and Enrichment Analysis. They will be used based on thresholds settings (p-value/q-value, fold change) by users' input.

Protein BioViewer is designed to show the primary protein sequence, predicted secondary structures, phosphorylation positions, amino acid characteristics, functional domain positions and transmembrane helix positions.

1. INTRODUCTION

Web based data analysis is becoming one of most efficient ways to help bioinformatics researchers understand their data. Different tools on different websites offer various ways to process and display data by either giving plain text analysis result or introducing data visualization tools to users. And at same time, more and more datasets are generated focusing on different purposes, but they all will benefit researchers. Motivated by these reasons, Soybean Knowledge Base (SoyKB) [11, 12] was conceptualized and developed. It integrates all different kinds of data including transcriptomics, proteomics, metabolomics, miRNA, SNP, sRNA, CDNA, generated by soybean research groups. SoyKB also gives abundant tools to let users choose suitable way to analyze data, like Affymetrix Probe ID Mapper, Gene Family Browser and In Silico Breeding Program. SoyKB also built directly connection between Soybase and Phytozome websites for updating the newest public dataset to make them available to users. As one of most well known websites about soybeans, SoyKB has effectively helped hundreds of biologists and biology groups, especially those who are interested in soybean analysis and would like to understand their data thoroughly.

There are three tools will be introduced, which have become very useful and widely accessible tools on SoyKB.

SNPViz is short name of Signal Nucleotide Polymorphism Visualization. It is designed to help analyze SNP information by providing grouping results, which will be transformed into pictorial output in formats of Joint Photographic Experts Group (JPEG)

and Portable Network Graphics (PNG). It has stand-alone and web versions. The performance of stand-alone version will be affected by the quality of computer components, which namely is the capacity CPU and Memory. The maximum number of combinable datasets is also limited in this version. But in the web version, there is no such limitation. Users need to contact Dr. Trupti Joshi or Dr. Dong Xu to upload their own datasets. There are two public datasets available on SoyKB. They are 31 wild and cultivated soybean genomes [17] and whole-genome sequencing of Glycine soja [14]. The user can use these two datasets as examples to familiarize themselves with the features of the web version of SNPviz.

Differential Expression is a multi-functional tool for analyzing different categories data among Transcriptomics RNA-Seq, Microarray, Metabolites and Proteomics data for both private and public purposes. Helping biologists bring the data to the front is the mainly purpose of this tool. It significantly shortens the analyzing process by using some public available tools and self-coded tools. Users do not need to program their own function to visualize their data. Differential Expression does the work of coding for people and even more. It also gives users an opportunity to present their data to the public in the future.

Demonstrating different kinds of information of genes is the main goal of the Protein BioViewer. It analyzed all genes in gene model v1.0, which will be explained in Gene Model section. This tool will show primary protein sequence, predicted secondary structure, phosphorylation positions, protein characteristics, functional domain positions and transmembrane helix positions.

2. GENE MODEL

Gene model versions will be mentioned in detailed introduction to tools on the website in this paper. All information related to gene model is offered below.

- Gene identifier

The gene identifiers are adopted in the researches of soybean genes to describe soybean genes of which some will be used in the function demonstration of all tools later in this paper. The gene identifier is formatted as Glyma##g#####, in which the ## indicates the chromosome number and the ##### is the numeric index along with whole chromosome (www.phytozome.net).

- Gene annotation external databases

Some gene annotation databases will be used in following introduction. Especially Pfam [23], PANTHER [19], KOG [15] and GO [1] will be associated with Differential Expression tool and Protein BioViewer. Genes can be classified into groups according to their family, molecular function, biological process and pathway. These databases are helping to classify selected gene into groups by the data they have and their own algorithms. Please see more details with their papers.

- v1.0 and v1.1

The Soybean (Glycine max) gene annotation files, which used in SoyKB, were downloaded from Phytozome (www.phytozome.net). v1.0 and v1.1 are two versions of soybean gene annotation models. They are produced on the base of Large-scale shotgun sequencing [14]. v1.0 is based on partial gene set Glyma0, which has been improved in

December 2008 as Glyma1. v1.1 is based on Glyma1, which is the most current assembly (www.phytozome.net).

- v9.0 and Gma_189

v1.1 gene annotation files were downloaded under Phytozome version 9. Specially designed for soybean gene research, Phytozome use Gma_189 to illustrate the difference between different versions of Phytozome on soybeans since gene annotation files are continually updating. Gma in Gma_189 stands for Glycine max and the number 189 in all file names is a Phytozome internal identifier for the current release of this genome annotation file. This number can be disregarded (www.phytozome.net).

- Version 9.0 and Version 1.0

In two gene models, the number of genes has changed since there are new added (Figure 2.1 a), deleted (Figure 2.1 b), unchanged and some changed genes between two versions (Figure 2.2 and Figure 2.3). In v1.0, the total number of genes is 75778 and in v1.1, the number changes to 73320. In order to incorporate both soybean gene models, SoyKB is transferring infrastructure from version 1.0 to version 9.0 by supporting both gene model in gene card page. SoyKB is planning to add multi-models supports in all related tools and pages. In gene card page, if selected gene exists in both gene models, a clickable link will be shown under Gene Version. Users can switch gene model if needed.

Figure 2.1 (a) New added gene in Version 9.0 (b) Deleted Gene in Version 9.0

Figure 2.2 Changed gene in Version 1.0

Figure 2.3 Changed gene in Version 9.0

3. SNPviz

Within the same species or among biologically related species, the same position in the DNA sequence may be different because of nucleotides variations (A, T, C or G). When this phenomenon happened, we call these species the ones having Single-nucleotide polymorphism (SNP). SNP is prevalent in human genome. In general, one SNP occurs among 500-1000 base pairs in average. Most cases of SNP happen during base transition, insertion and deletion. Since these SNP mostly happen in intron, the SNP cases, which happens in exon, becomes important research resources for biologists. Now that a large number of datasets about SNP cases are generated, analyzing these datasets is becomes one of the difficult problems as processing and reviewing data demands tremendous efforts in the first place, especially for comparing multiple samples among different datasets or species. To solve this problem, SNPviz offers a convenient way to group and visualize samples in one or cross multiple datasets.

Besides helping biologists to analyze an enormous amount of data efficiently and completely, collaborator also wants to reveal that the number of haplotypes will evaluate the consistency of haplotypes. SNPviz was original designed for soybean research. Later, the designer expanded the fields that the software could be applied to. And as result, SNPviz has been published on PLoS ONE [18].

The goal of SNPviz is to provide pictorial output after grouping collected samples by using the ClusterW method. Multiple samples in dataset(s) with selected region can be

displayed with SNP data and gene information including exon locations found in this region.

3.1 Introduction

SNPViz is designed as an easy-to-use JAVA software with stand-alone and web versions. The web version has been integrated into SoyKB. The stand-alone version is also available on this website. Both versions are coming with detailed README files, and the standalone version also has a user guide to use of functions step by step. In order to give users more personal experience, both versions provide users with a wide range of features that could individualize the tools, including flexibility for combining up to three datasets with different samples, availability of different pictorial output color schemas etc. Users are allowed to combine up to three datasets together in the stand-alone version. If they want to merge more than three datasets, they can contact Dr. Trupti Joshi or Dr. Dong Xu to upload their data to the SoyKB server and then do the SNP analysis by using the web version.

3.2 Rationale Behind Features

3.2.1 Input Options

SNPViz requires an SNP-array or GBS file inputted as a tab-delimited text file to be run. SNP data is included in the first line of the file while column identifiers are excluded. In the first column, the chromosome number is placed. The sequence position is placed in the second column. The third column is assumed by the software to be the

reference sequence. The rest of the columns should be filled with all sample SNP data, which is a single nucleotide or a double nucleotide. By referencing to the International Union of Pure and Applied Chemistry (IUPAC), double nucleotides will be translated into single letter code.

SNPViz has a comprehensive set of functions including inputting datasets, renaming samples, and deleting datasets. Header file in XML format, carrying all sample names, are needed when one adds new datasets to the program. SNP-array or GBS files are linked with the header files in SNPviz. The folder that shares the same name with the dataset is the only place that the header files and SNP text files should exist. The sample names, which appear in the header files, will show up spontaneously in the SNPviz clustering window upon the input of a SNP text chromosome file if related head files are added to the program.

To connect SNP positions and gene regions and exon locations, a general feature format (GFF3/GFF2) or general transfer format (GTF) annotation file could be incorporated. By having GFF/GTF files inputted, the program will be able to pop out a window to alert the users to select identifying features of GFF files, including names, locus_id, or mRNA. Therefore, the identification of label correlating with the gene and exon locations could be realized. The gene positions will be missing from the output pictorial if no GFF/GTF file is available or provided by users. For the web version on SoyKB, gene annotations for either Glyma1.0 or Glyma1.1 are automatically drawn from the database.

Only integer values will be allowed for input when users specify chromosome starting and ending positions. Other formats are invalid. A warning message will jump out upon the input of values in wrong format. The annotation files and available computer memory will decide the size of the region on chromosome to be viewed and printed.

3.2.2 Display Options

Entering the SNP-array or GBS files of the target chromosome, the nucleotide position starting and ending points, and optionally, the related annotation files are the prerequisites for users modifying output display options. A variety of colors are available to demonstrate SNP data, enough to represent every variant base positions existing in each genome. In the cases of default color setting which employs black, white and grey for representation, black implies the difference of a base at the position in comparison to the reference sequence and white suggests the base at that position being same with the reference while grey indicates the unavailability of data because base at that position has a “-” or “N” rather than a nucleotide. As mentioned, more colors are available for more customized visualization of the SNP data. For example, red, green, yellow, and blue can be used to indicate A, G, C, and T, respectively. Other colors appear when more than one nucleotide occurs at a specific position. The output could be graphed in either the default color setting or multiple color setting with or without the nucleotide. For viewers demanding for greater clarity in the output, the width-predetermined space can be added to increase the room between two neighboring

samples if users select the “show space” option. Various ways of visual representation allows the users to investigate datasets with more freedom and comprehensiveness.

3.2.3 Phylogeny Clustering

The distance matrix data calculated by using ClustalW, which was explained in Thompson et al. 1994 [6], made allocating samples into haplotype groups possible. Categorizing samples and a establishing phylogeny tree are realized by employing Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [7].

3.2.4 Integration of Datasets

Three is the maximum number for combination of different datasets concerning the same chromosome so far as its SNP file is uploaded for each dataset. As multiple datasets are mingled with each other, the first sample will be automatically chosen to be the reference for each dataset in SNPviz. In case some SNP positions don't exist in all merging datasets, the reference SNP is made for those positions. SNP positions lacking data, shown as “-” or “N”, are excluded when graphing the phylogeny tree and the genetic distance is calculated based on the number of mismatched data points divided by the number of total data points.

3.2.5 Output Options

Joint Photographic Experts Group (JPEG/JPG) and portable network graphics (PNG) are formats that the phylogeny tree can be saved as. All the SNP positions contained in the phylogeny tree can be saved as a text file. So is the distance matrix information. The

gene and exon information for the haplotype region will also exist in the text file if a GFF/GTF was employed in previous steps.

3.3 Demonstration

3.3.1 Stand-alone version

Figure 3.1 Stand-alone SNPviz Workflow

Stand-alone version is a JAVA application, developed in NetBeans. The Design Swing GUI in NetBeans is an easy-to-use designing tool that allows dragging and positioning GUI components into application canvas. All related interface functions were completed in NetBeans. Figure 3.1 gives the basic functionality of stand-alone SNPviz. Each function will be introduced in detail in following paragraphs.

3.3.1.1 Preloaded Datasets

The downloaded stand-alone SNPviz from <http://soykb.org/SNPviz/> comes with two publicly available datasets. Each dataset will come with a header file and all twenty chromosome SNP files. The header file is structured in one line including all sample names. If users want to use their own datasets, they need to create a similar file to put into SNPviz. Details about self-created files will be introduced later. Two public dataset headers are already preloaded into SNPviz and do not need to be added to the program by using the Input Dataset function. To use these datasets, users only need to select specific chromosome SNP file in the SNPviz Clustering window.

3.3.1.2 Main Interface

Figure 3.2 Main Interface of stand-alone SNPviz

Figure 3.2 is the main interface of stand-alone version SNPviz. It shows the basic introduction and instruction. Main interface aggregates all functions including Input

Dataset, Modify Datasets, Delete Datasets, Run Clustering, Open User Guide, which will be introduced in detail in following sections.

3.3.1.3 *Input Dataset*

The “Input Dataset” function helps users adding their data into SNPviz, which is shown in Figure 3.3. When the user uses non-preloaded datasets for the first time, he or she needs to select the header file and name to the dataset. The name of the dataset will be treated as a unique ID during clustering process. A notice will be given to users if inputted names have already existed. Dataset information, including dataset names and sample names, is stored in XML file in the same folder of the software. Once users fully understand the structure of XML file, they can easily edit this file without using the Input Dataset Function. The XML file can be edited in a lots of modern text editor such as the TextWrangler in Mac OS and the Notepad++ in Windows.

Figure 3.3 Stand-alone SNPviz Input Dataset Interface

The header file should be written under strict rules. It should only have one line, and keywords in the header file should be separated by tab. SNPviz will recognize all

keywords from the third keyword as sample names. The “Input Dataset” function links the header file to the inputted SNP-array or GBS file so that the sample names automatically appear when a SNP chromosome text file is added to SNPviz.

3.3.1.4 *Modify Sample Names*

The “Modify Sample Names” function enables users to view all datasets that are added into SNPviz. Two interface examples showing in Figure 3.4 and Figure 3.5. Users can then select one dataset at a time to view all samples available in this dataset and edit them by directly changing the name showing in white text boxes.

Figure 3.4 Stand-alone SNPviz Modify Sample Names one

Figure 3.5 Stand-alone SNPviz Modify Sample Names Two

3.3.1.5 Delete Datasets

“Delete Datasets” function, showing in Figure 3.6, allows the user to view all available datasets in SNPviz. Users can delete one or more datasets by selecting corresponding datasets. After users deleted datasets, SNPviz no longer retains dataset information, including dataset names and sample names. If users reselect deleted datasets, a notice will be shown to remind the users that dataset in request have already been deleted.

Figure 3.6 Stand-alone SNPviz Delete SNP Datasets Interface

3.3.1.6 Run SNPviz Clustering

“Run SNPviz Clustering” function is the main interface of SNPviz and allows users to create a clustering pictorial of user-defined regions from one to three SNP-array or GBS datasets. To create this clustering pictorial, the SNP data undergoes multiple sequence alignment in ClustalW. Then samples are grouped, and a phylogeny tree is constructed using UPGMA.

Figure 3.7 Stand-alone SNPviz Clustering Interface

From the interface, as Figure 3.7 shows, users can select from a variety of display options. Haplotypes can be viewed with two color schemas. Three-colors uses white to represent SNPs that are identical to the reference, black standing for SNPs that are different from the reference, and grey for positions that contains no SNP data (Figure 3.8). Alternatively, users can view same SNP information in multi-colors schema (Figure 3.9). Table 3.1 shows the mapping relationship between letters and colors. The nucleotide base can be included in the image if 'Show SNP Nucleotide' option is selected (Figure 3.9 shows the output with nucleotide base, Figure 3.8 shows without). Users can also include an annotation file so that both gene and exon locations will be shown in selected region (Figure 3.8).

Figure 3.8 Stand-alone SNPviz clustering result without nucleotide bases on samples. Show gene information selected, chose from 31genomes and soja datasets, range: 90000-10000000. Three-color schema selected.

Figure 3.9 Stand-alone SNPviz clustering result with nucleotide bases on samples. Show gene information, chose from 31genomes and soja datasets, range: 90000-10000000. Multi-color schema selected.

Symbol	Base	Color
A	A	Red
G	G	Blue
C	C	Green
T	T	Yellow
K	G/T	Dark Magenta
M	A/C	Brown
R	A/G	Lime
S	C/G	Dark Red

Symbol	Base	Color
W	A/T	Cyan
Y	C/T	Pink
B	Not A	Sky Blue
D	Not C	Green Yellow
H	Not G	Light Blue
V	Not T	Magenta
N	Any base	Orange
-/.	Missing or no data	White

Table 3.1 Multi-color schema color mapping between symbols and colors

3.3.1.7 Help

This Feature has two sub functions, ‘README’ and ‘User Manual’. ‘README’ will lead users to a new interface, which gives brief introduction of the SNPviz and basic function description. And after clicking on the ‘User manual’, a file in Portable Document Format (.pdf) will be opened. It covers all aspects of SNPviz including general information, program features, operating instructions and scenarios for integrating datasets.

3.3.2 Web Version

SNPViz

Quick Search Gene Card Go

Please choose any data you want to compare Download Stand-Alone Version README

Please Choose Chromosome: Gm01 Please input starting point: Please input ending point: Clustering

☒ Glyma v1.0 ☐ Glyma v1.1 ☐ Three Color ☒ Multi Color ☐ UPGMA ☐ Show SNP Nucleotide ☒ Show Space

GWAS (Public) Soja (Public) NAM_41 (Private) BGI_25 (Private)

Ref (W82) W01 Beijing4 W02 Donggou16 W03 Wuha4 W04 Yanjin3 W05 Mengjin1 W06 Jidong5
W07 Fengcheng20 W08 Jixian11 W09 Kalyuan21 W10 Shuangcheng4 W11 Talyuan67 W12 Anqing18
W13 Yimeng15 W14 Yimeng19 W15 Yanjin4 W16 Hailun19 W17 Zhangwu20 C01 Wenfeng7 C02 Tiefeng8
C08 Union C12 Jindou21 C14 Brazil10 C16 Tainong1 C17 Zigongdongdou C19 Jilinxioli1 C24 Gandou4
C27 Cangdou5 C30 Yudou12 C33 Hefeng25 C34 Gui199 C35 Guangzhoudali

Select All Clear All

If you use SNPviz in published work, please cite:
Langewisch T, Zhang H, Vincent R, Joshi T, Xu D, Bilyeu K. (2014) Major Soybean Maturity Gene Haplotypes Revealed by SNPviz Analysis of 72 Sequenced Soybean Genomes. PLoS ONE 9(4): e94150. doi:10.1371/journal.pone.0094150

Figure 3.10 Web SNPviz Main Interface

The basic program architecture and core algorithm of web version is exactly the same with stand-alone version in order to keep consistency between the versions. The User Interaction Interface will be introduced with a main focus on identifying the difference between the web version and stand-alone-version. Figure 3.10 shows the main page of web version.

3.3.2.1 Dataset, samples selection

Apart from two public available dataset, 31 genomes and soja, users of SoyKB can upload their own data into SNPviz web version. This function is only available to SoyKB users. In figure 3.11, NAM_41 and BGI_25 are private datasets, which were uploaded as items in different groups. All users belong to these groups are able to combine samples in these dataset with other public samples.

Figure 3.11 Available Databases on SoyKB with Web SNPviz

3.3.2.2 Competitive Edge

As to the clustering methods, sample combining method and display options, both versions are sharing exactly same code for consistency. But there are some advantages for using web version.

- **Worldwide data accessible.** Users do not bring the data with themselves all the time. Once uploaded into SoyKB, the data would be very easy to maintain and retrieve when one has Internet access. Along with national basic information infrastructure development, high range Wi-Fi connection, large mobile data commutation coverage and wide availability of the Ethernet are making Internet more accessible.
- **Unlimited combining datasets.** In order to make the interface of stand-alone SNPviz more organized, after considering the performance of JAVA program, only up to

three datasets could be selected for combining. But for the web version, there is no limit on the number of dataset that could be combined together.

- **Keep consistent output result.** Due to the limitation of computer display performance, user may not be able to get the best output result since the width for each bar is calculated automatically based on the characteristics of display screen. In web version, all samples will have same bar width, see Figure 3.12.

3.4 Design Methods

3.4.1 Genetic Distance between Samples

The biggest challenge for clustering samples properly is how to find the genetic distance between each sample and the reference. According to Schauser [24], the genetic distance is described as the number of mismatched data points divided by the number of total data points. Positions without data are completely removed from the calculation. If the comparing samples are AT-C-G-CA and TTC--G-CA, the genetic distance will be four divided by five, or $4/5$. The one is the number of mismatches because of the A is not matched with T at the location 1. Number 5 is the total number of positions, where both positions have gene strands. According to Schauser [24], ClustalW was described as distance calculation method, which will be used in SNPviz for calculating the distance between all samples and the reference. ClustalW was originally described in Thompson et al. 1994 [6].

3.4.2 Tree Drawing Methods

After being generated, the distance matrix will be treated as input file for drawing methods. In SNPviz, Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Neighbor Joining (NJ) are adopted as tree drawing algorithm. To keep the result accurate, two published and open sources were used which were implemented by Peter Sestoft, an adjunct professor of life sciences at the University of Copenhagen (1999), and remains mostly the same to meet the demand of this project. The original code can

be found at <http://www.itu.dk/people/sestoft/bsa/Match7/java> [28]. It was merely changed for fitting SNPviz output.

For publishing purpose [18], NJ has been hidden from the interface. But the code is available for users to develop on their own for non-commercial research purposes.

3.4.3 Double Nucleotide Transfer to Single Nucleotide

Most organisms have two sets of chromosomes, which contain genes in same positions. Every position has 2 alleles, which can be same (homozygous AA or GG) or different (heterozygous AG). Sometimes it can also be different from reference when it is called SNP. So it is possible for double nucleotides, which are translated into International Union of Pure and Applied Chemistry single letter code, to occur in one position when we deal with SNP data. More detail can be found on [http://www.ncbi.nlm.nih.gov/staff/tao/tools/tool_lettercode.html]

3.4.4 Sample Combining

When samples from different datasets are combined, the positions that the datasets have information about in the range defined by the numbers users input on chromosomes may be different. The combining method adopted in following examples will be explained with detailed steps.

R1, S1 are chosen from dataset A, and R2, S2 are chosen from dataset B in the range 0 – 1000. A has five positions: P1, P2... P5. B has four positions: p1, p2... p4. P2 and p2 are known to be in the same position on the same chromosome. A two dimensional array will be created to store all SNP data.

Step 1: Add all R1 and S1 data into the array. For combining purpose, SNPviz assumes R2 and S2 have positions P1, P2...P5 that they are void of in fact. The fabricated positions are filled by copying the reference samples R1 in P1, P2 ... P5 to R2 and S2 at positions P1, P2 ... P5.

	P1	P2	P3	P4	P5	p1	p2	p3	p4
R1	A	T	G	G	C				
S1	A	A	G	C	C				
R2	A	T	G	G	C				
S2	A	T	G	G	C				

Table 3.2 Sample Combining Step 1

Step 2: Add data of R2, S2 at position p1 ... p4 into array.

	P1	P2	P3	P4	P5	p1	p2	p3	p4
R1	A	T	G	G	C				
S1	A	A	G	C	C				
R2	A	T	G	G	C	C	T	A	A
S2	A	T	G	G	C	G	A	T	A

Table 3.3 Sample Combining Step 2

Step 3: Copy the reference data R2 at p1...p4 to R1 and S1 at position p1...p4.

	P1	P2	P3	P4	P5	p1	p2	p3	p4
R1	A	T	G	G	C	C	T	A	A
S1	A	A	G	C	C	C	T	A	A
R2	A	T	G	G	C	C	T	A	A
S2	A	T	G	G	C	G	A	T	A

Table 3.4 Sample Combining Step 3

Step 4: Find the same positions between P1, P2 ... P5 and p1, p2 ... p4. In this example, P2 and p2 are in fact the same position on the chromosome. Therefore, the SNP data in R2, S2 in p2 is the real value and should overwrite the original value in P2.

	P1	P2	P3	P4	P5	p1	p2	p3	p4
R1	A	T	G	G	C	C	T	A	A
S1	A	A	G	C	C	C	T	A	A
R2	A	T	G	G	C	C	T	A	A
S2	A	A	G	G	C	G	A	T	A

Table 3.5 Sample Combining Step 4

Step 5: Delete all sample data in position p2.

	P1	P2	P3	P4	P5	p1	p3	p4
R1	A	T	G	G	C	C	A	A
S1	A	A	G	C	C	C	A	A
R2	A	T	G	G	C	C	A	A
S2	A	A	G	G	C	G	T	A

Table 3.6 Sample Combining Step 5

Step 6: Sort out positions and record the original index.

0	1	2	3	4	5	6	7
P1	P2	P3	P4	P5	p1	p3	p4

Table 3.7 Sample Combining Step 6, Index and position mapping before sorting

0	5	6	1	2	3	7	4
P1	p1	p3	P2	P3	P4	p4	P5

Table 3.8 Sample Combining Step 6, Index and position mapping after sorting

Step 7: Based on the index array, which was created in last step, rearrange all the data.

	0	5	6	1	2	3	7	4
	P1	p1	p3	P2	P3	P4	p4	P5
R1	A	C	A	T	G	G	A	C
S1	A	C	A	A	G	C	A	C
R2	A	C	A	T	G	G	A	C
S2	A	G	T	A	G	G	A	C

Table 3.9 Sample Combining Step 7

3.5 Testing and Validation

In order to make SNPviz publishable, it has been tested thoroughly. As one part of validation, SNPviz testing needs to be designed to satisfy validation criteria. It needs to meet the original design purpose, implement all requirements of collaborators and work properly with precise result. Since SNPviz has two versions, both versions should be tested individually because the using environment is different. SNPviz testing will be introduced in two directions: shared code testing and individual using environment testing. For the special characteristic of testing, white box testing was used for testing shared code, and black box testing was used for using environment.

3.5.1 White Box Testing

White box testing is more focusing on unit, integration and system testing. In this way, SNPviz was tested more on the result conducting process. All methods introduced in section 3.4 were tested by hand. Based on the description of Schauer[24], SoyKB created own code to implement this method, the result has been verified by hand calculation. SoyKB adopted UPGMA from Sestoft[28], so this method needs to validate by input public available examples. Two examples [8, 21] were used to test the phylogeny tree output. The results of SNPviz were exactly same with examples', showed in Figure 14-17.

As one of most important testing in White Box Testing, Unit testing was conducted to examine each component of SNPviz. As described before, stand-alone version SNPviz has six functions included. Each function as one component should work without

any effect from other functions. Each function has been tested to work individually and all worked well.

Figure 3.13 UPGMA Testing Example 1, Edwards result[8]

```
double[][] ds = new double[][] {
    {0.0,0.0,0.0,0.0,0.0,0.0,0.0 },
    {19.0,0.0,0.0,0.0,0.0,0.0,0.0 },
    {27.0,31.0,0.0,0.0,0.0,0.0,0.0 },
    {8.0,18.0,26.0,0.0,0.0,0.0,0.0 },
    {33.0,36.0,41.0,31.0,0.0,0.0,0.0 },
    {18.0,1.0,32.0,17.0,35.0,0.0,0.0 },
    {13.0,13.0,29.0,14.0,28.0,12.0,0.0 }
};
String[] tmpName = new String[]{"A","B","C","D","E","F","G"};
upgmaTree = new UPGMA(tmpName, ds,choice, countArray);
```

Figure 3.14 UPGMA Testing Example 1, SNPviz result

The pairwise evolutionary distances are given by the following distance matrix:

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

Figure 3.15 UPGMA Testing Example 2, Opperdoes[21] result

```

double[][] ds = new double[][] {
    {0.0,0.0,0.0,0.0,0.0,0.0},
    {2.0,0.0,0.0,0.0,0.0,0.0},
    {4.0,4.0,0.0,0.0,0.0,0.0},
    {6.0,6.0,6.0,0.0,0.0,0.0},
    {6.0,6.0,6.0,4.0,0.0,0.0},
    {8.0,8.0,8.0,8.0,8.0,8.0}
};
String[] tmpName = new String[]{"A","B","C","D","E","F"};
upgmaTree = new UPGMA(tmpName, ds,choice, countArray);
  
```

Figure 3.16 UPGMA Testing Example 2, SNPviz result

3.5.2 Black Box Testing

To examine the functionality of both versions, black box testing was used. Black box testing cases are built with specific requirement input and expected output. SoyKB conducted testing based on predicted users behaviors and expected outcomes. Classical black box testing, boundary testing, error guessing and decision table testing, were used.

- **In boundary testing**, the limitation about selectable chromosome region with or without gene annotation information for stand-alone testing was tested. Dr. Langewisch and I tested on Mac OS and Windows 7 individually. We found that the result is highly relevant with computer memory. Because it is impossible to know the users' computer information, it is unreasonable to limit the range of chromosome region that users can select from. SNPviz stand-alone version gives pop-up message if the excessive amount of data in selected range leads to shortage of memory.
- **Error guessing** also was used for testing predicted user behaviors. For example, there are two position input boxes on both versions. User may type in string, negative value by mistake, or the inputted start position may bigger than end position. And on stand-alone version, when users select samples from different datasets, the selected files may not belong to same chromosome. These situations all will stop the program running. To identify these predictable inputs, some pop-up messages also will be shown to users accordingly.
- **Decision Table** was used to test selectable options on both SNPviz versions. As described in section 3.2.2, selectable options need to be tested to validate the SNPviz results are same with expected results. 32 test cases were generated for both versions individually. Giving selected choice, both versions all generated expected result. Table 3-10 shows the mapping relationship between selectable options and selected choice.

Choice	With Space (0)/ Without (1)	UPGMA-0/ NJ-1	Three Color-0/ Multi Color-1	ID (0)/ NAME (1)	With SNP Nucleotide (0)/ Without Nucleotide (1)
0	0	0	0	0	0
1	0	0	0	0	1
2	0	0	0	1	0
...	...	...	...	...	...
31	1	1	1	1	1

Table 3.10 Decision table in Black Box Testing

- **Cross Species Testing** was also used to check the feasibility of stand-alone SNPviz. Rice, Arabidopsis and maize have been tested with data and gene annotation files. Data description will be introduced in Validation and all gene annotation files were downloaded from Phytosome.

3.5.3 Validation

Dr. Langewisch and I conducted validation by comparing our genes of interest (E genes) in soybean to homologs of these genes in other species (rice, Arabidopsis, and maize). In addition to the E genes, we also tried looking at some important flowering time genes and their homologs. We just thought it would be interesting to look at allelic variation of similar genes with potentially the same function.

The flowering time genes in soybean are based to an extent on the genes that are homolog in Arabidopsis. The Arabidopsis SALK set was from the 1001 Arabidopsis Genomes Project (<http://1001genomes.org/datacenter/>).

We examined NAM population in maize and some landraces and teosinte lines. The dataset is available on http://www.panzea.org/lit/data_sets.html#NAM_map.

The last dataset was the rice hapmap3. We divided this dataset into Indica, Japonica, and wild lines.

The reason we want to do across species validation is because it would be interesting to look at allelic variation of similar genes with potentially the same function. The flowering time genes in soybean are based to an extent on the genes that are homolog in Arabidopsis. We thought it would make the paper [18] more interesting to compare across species. The result does demonstrate that the program could be used for other species besides soybean. But after generating all phylogeny trees of these three species, we found it is difficult to assign alleles based on the haplotypes. So we did not include this validation in published paper.

We also did validation about E1, E2, E3, E4 and Dt1 genes. The result is available on paper [18] Result section.

3.6 Result

The goal of SNPviz is to produce phylogenetic trees of soybean samples and visualizations of haplotypes by sourcing SNP data from entire genome re-sequencing. It made the examination of the haplotype diversity and allelic variation for soybean genes possible for biologists. Dr. Tiffany Langewisch, the first author of the scholarly work *Major Soybean Maturity Gene Haplotypes Revealed by SNPviz Analysis of 72 sequenced Soybean Genomes*, was using SNPviz to analyze soybean mature genes. In the work, it reads “E1, E2, E3, and E4 along with Dt1gene for plant growth architecture were analyzed in an effort to determine the number of major haplotypes for each gene, to evaluate the consistency of the haplotypes with characterized variant alleles, and to

identify evidence of artificial selection.”¹ This work has been composed and submitted on September 2013. After employing the second version SNPviz and revising details in the paper, PLOS One accepted this work on March 2014 and published on April 2014 [18].

¹ T. Langewisch, H. Zhang, R. Vincent, T. Joshi, D. Xu, and K. Bilyeu, "Major Soybean Maturity Gene Haplotypes Revealed by SNPviz Analysis of 72 Sequenced Soybean Genomes," PLoS One, vol. 9, p. e94150, 2014.

4. DIFFERENTIAL EXPRESSION

4.1 Introduction

Bio-related information is attracting more and more attention in our life now. With enormous number of expression datasets are coming out for analyzing and research purposes in different fields. The soybean research field is no exception. SoyKB, the soybean research website, aims to build a series of easy-to-use tools that are crucial to improving how biologists understand their data, especially the way they analyze expression datasets. Differential Expression was designed and developed in SoyKB group. For now, there are many multi-omics datasets integrated into this tool, including transcriptomics (both RNA-Seq and micro-array), proteomics and metabolomics, to accommodate the needs of for different groups.

4.2 Functions

Base on biologists' requests and suggestions, six functions were integrated into Differential Expression. Figure 4.1 shows the datasets selections interface. Most of the datasets belong to specific group and are only open for private use. Once the omic category, specific datasets and data generated method are chosen, all conditions will be showed for combining purposes. (Figure 4.2) To protect the exclusiveness of certain datasets, the public dataset, 4 genotypes Control (C) & PAMP Treatment (P) [25] will be used as an example in the following function demonstration.

Differential Expression

Please select the data type, dataset and method

-----Data Format-----

Please set up filtering parameters:

Fold Change:
2

p-value:

Submit

-----Data Format-----
☒ TranscriptomicsRPKM RNA-Seq
☐ TranscriptomicsReadCount RNA-Seq
☐ Microarray
☐ Metabolomics
☐ Proteomics

Figure 4.1 Differential Expression – Dataset selection page

Differential Expression Quick Search Gene Card Go

Please select the data type, dataset and method

TranscriptomicsRPKM RNA-Seq Dataset6-7 Root Hair/Stripped Root Heat Stress cuffdiff

☐ Roothair_T0_25C_vs_T0_40C
☐ Roothair_T6_25C_vs_T0_40C
☐ Roothair_T24_25C_vs_T0_40C
☐ Strippedroot_T3_25C_vs_T0_40C
☐ Strippedroot_T12_25C_vs_T0_40C

☐ Roothair_T3_25C_vs_T0_40C
☐ Roothair_T12_25C_vs_T0_40C
☐ Strippedroot_T0_25C_vs_T0_40C
☐ Strippedroot_T6_25C_vs_T0_40C
☐ Strippedroot_T24_25C_vs_T0_40C

Please set up filtering parameters:

Fold Change: This dataset based on natural logarithm (ln(x))
2

p-value:
0.05

Submit

Gene Lists Venn Diagram Volcano Plot Function Pathway Enrichment Analysis

Figure 4.2 Differential Expression Condition Selection Page

4.2.1 Gene Lists

In Gene list tab, SoyKB offers an effective way to view attributes, which are available in the selected dataset. Figure 4.3 shows partial genes, which satisfy the criteria the user inputted on p-value (or q-value) and fold change. In the first selected condition (control_11272_vs_control_11268), fold change and p-value/q-value have

been rounded by keeping three decimals. Users can download gene information in selected condition by clicking download button under the condition description.

control_11272_vs_control_11268:
A total of 61 records have been found. [Hide Entire list](#)

Download Data

ID	GeneID	locus	status	fold_change	p_value	q_value
1	Glyma01g02080.1	Gm01:1602932-1604459	OK	INF	0	0
2	Glyma02g02880.1	Gm02:2166499-2167848	OK	4.104	0	0.004
3	Glyma02g03220.1	Gm02:2488349-2488989	OK	INF	0	0
4	Glyma02g03230.1	Gm02:2494912-2496440	OK	12.094	0	0
5	Glyma02g03280.1	Gm02:2540910-2541720	OK	-7.438	0	0

Figure 4.3 Differential Expression Gene Lists Tab View

4.2.2 Venn Diagram

After filtering all genes based on inputted criteria, filtered genes in each condition presented in the Venn diagram may have intersections. Using Venn diagram to do condition comparison has become a well-known method of gene clustering in bioinformatics. For analyzing purposes, SoyKB also adopts a public method from Oliveros, J.C. (2007) VENNY [20]. An interactive tool using Venn Diagrams for comparing lists, shows in Figure 4.4. In each intersection, the number indicates the overlapping genes from two or more conditions. When user clicks on the number, the name of overlapping genes will be shown in the bottom text box. User is able to download overlapping genes by clicking on the 'Download' button. A text file will be generated for users to download. At the right side, filtered genes in each condition will be shown individually. The Venn diagram also provides users with an option to save the diagram as a PNG file. Oliveros, the VENNY creator, also produced several useful functions for customizing image output.

graphics in R is widely used in academic circle. Most of bio-related researchers understand and buy the results from R. This has become the reason that SoyKB chose to use this method in Differential Expression Tool.

Figure 4.5 Differential Expression Volcano Tab View

Same to Cui X, Churchill GA (2003) [5], the negative log10-transformed p-values of the selected condition are plotted against the log ratios (log2 fold change). For some datasets, $\ln(x)$ may be used as x axis unit.

In Figure 4.5, all genes in selected condition are displayed. Dark blue points stand for genes that do not satisfy the criteria. Contrarily, the red points stand for genes that satisfy the criteria. All Volcano Plots are displayed in Venn Diagram Tab based on user's selections.

4.2.4 Function Analysis and Pathway Analysis

Figure 4.6 Differential Expression Function Tab Annotation Datasets Selection View

In Function Analysis and Pathway Analysis tab, users are able to view the categorized function annotation on pie chart for gene families, Pfam [23], PANTHER [19] and KOG [15]. Users can select a specific condition of a certain database they want to know about. Figure 4.6 showed the selection tab. After selecting the categories, all genes in selected category will be shown in pie chart (Figure 4.7). Open Flash Chart [30], which is an open source that offers free use of source code, is used to generate pie charts. The result will be split into two sections, Unannotated and Annotated. In the Annotated section, all functions found in the selected category will be shown with a total number of annotated genes, the percentage of this category in the total annotated genes and the percentage of the this category in total genes. 'Show Entire List' will display all matched genes in specific groups. Users are also able to download function names, gene lists and other information separately (Inositol 1, 3, 4-trisphosphate 5/6-linase, shows in Figure 4.7 Annotated section, for example). Downloading entire list will be also available for users.

Figure 4.7 Differential Expression Function Tab with Pfam output

4.2.5 Pathway Information

Metabolic pathways are serious reaction occurring in cells. A lot of public available datasets have found out how particular genes co-work with each other in specific pathway. SoyKB got all Glycine max (Soybean) related pathways from KEGG [13] including mapping genes. In order to keep consistency, SoyKB also adopted some displaying pattern similar to that in Gene List Tab showing in Figure 4.8. Users can show or hide the entire list. All genes all classified into different groups based on pathway information.

control_11272_vs_control_11268:
A total of 119 pathways have been found. [Hide Entire list](#)
[Download Data](#)

1.1.10_Nucleotide.sugars.metabolism - 9 records						
Glyma05g00220	Glyma1655s00200	Glyma16g27440	Glyma05g00220	Glyma1655s00200	Glyma16g27440	Glyma05g00220
Glyma1655s00200	Glyma16g27440					

1.1.11_Pyruvate.metabolism - 15 records						
Glyma05g00220	Glyma06g43590	Glyma12g09750	Glyma1655s00200	Glyma16g27440	Glyma05g00220	Glyma06g43590
Glyma12g09750	Glyma1655s00200	Glyma16g27440	Glyma05g00220	Glyma06g43590	Glyma12g09750	Glyma1655s00200
Glyma16g27440						

1.1.12_Glyoxylate.and.dicarboxylate.metabolism - 18 records						
Glyma05g00220	Glyma06g43590	Glyma12g09750	Glyma1655s00200	Glyma16g27440	Glyma19g06340	Glyma05g00220
Glyma06g43590	Glyma12g09750	Glyma1655s00200	Glyma16g27440	Glyma19g06340	Glyma05g00220	Glyma06g43590
Glyma12g09750	Glyma1655s00200	Glyma16g27440	Glyma19g06340			

Figure 4.8 Differential Expression Pathway Tab View

4.2.6 Enrichment Analysis

Gene Lists	Venn Diagram	Volcano Plot	Function	Pathway	Enrichment Analysis
control_11272_vs_control_11268					Perform Enrichment Analysis
control_LDX_vs_control_11272					Perform Enrichment Analysis
treated_11272_vs_treated_11268					Perform Enrichment Analysis
treated_LDX_vs_treated_11272					Perform Enrichment Analysis

Figure 4.9 Differential Expression Enrichment Analysis Tab View

In Enrichment Analysis Tab, SoyKB adopted GeneMerge [3] to do gene set enrichment analysis. Users can directly click the ‘Perform Enrichment Analysis’ button, shows in Figure 4.9. This function can be used as stand-alone tool under “Tools” in SoyKB bar. After inputting wanted gene list, users are able to select soybean gene model, name current job and do enrichment analysis as shown in Figure 4.10.

In order to use GeneMerge, three input files need to be generated: 1) population set file (Figure 4.11 .pop file); 2) gene association file (Figure 4.11 .assoc file); 3) description file (Figure 4.11 .desc file). SoyKB supports enrichment analysis of four

categories. Soybean will show each category accordingly based on selected gene models.

Gene association file on gene model v1.1 was collected from Phytozome. Gmax 189.

Enrichment Analysis

Please select gene model

☐ Glyma_v1.0 ☒ Glyma_v9.0

Please Name Your Job (Option, 100 character maximum)

control_11272_vs_cor

Please input your interested genes(one gene per line) (eg. Glyma06g45700.1)

Glyma02g05210.1
Glyma02g07070.1
Glyma03g24320.1
Glyma03g37930.1
Glyma04g04910.1
Glyma04g39320.1
Glyma05g00220.1
Glyma05g07640.1
Glyma05g24860.1
Glyma05g26360.1
Glyma05g36750.1
Glyma06g43590.1
Glyma06g44830.1
Glyma06g44870.1
Glyma07g12750.1

Submit

Figure 4.10 Enrichment Analysis Input Page

All description files were obtained from Pfam [23], PANTHER [19], KOG [15] and GO [1] datasets. For GeneMerge algorithm, please check Cristian et al. [3]. After the preparation of all input files, Enrichment Analysis tool is ready to use. Users can input the gene list they are interested in and each gene should be inputted in separate lines. After choosing gene model, naming the job and clicking on the “submit” button, a message will be pop out informing the users of the job ID and the result link. Since enrichment analysis will be run on SoyKB server and will take several minutes, the job will be automatically saved in users’ account panel (Figure 4.12). In user account panel, users can click on the Job IDs, and the results will be shown in another browser tab (Figure 4.13). Users can download enrichment result either in user job panel or on the result page.

Glyma01g00210.1	Glyma0196s130.1 PF05695
<u>Glyma01g00241.1</u>	Glyma0196s50.1 PF01040
Glyma01g00270.1	<u>Glyma01g00241.1</u> PF00076
Glyma01g00280.1	Glyma01g00270.1 PF02776
Glyma01g00291.1	Glyma01g00300.1 PF10551;PF04434
Glyma01g00300.1	Glyma01g00311.1 PF02721
Glyma01g00311.1	Glyma01g00321.2 PF04434;PF03101;PF10551
.pop file	.assoc file

PF00071 Ras family
 PF00072 Response regulator receiver domain
 PF00075 RNase H
PF00076 RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
 PF00078 Reverse transcriptase (RNA-dependent DNA polymerase)
 PF00079 Serpin (serine protease inhibitor)

.desc file

Figure 4.11 GeneMerge Input File

User Submitted Jobs List					
Job ID	Job Name	Job Type	Created Time	Status	
1396315787	Monk_0.5h_vs_Ethylene_0.5h	Enrichment Analysis	04/01/2014 01:29:47 am UTC	Done	Download Result
1397526278	Empty_Vector_vs_Glyma02g14880_Overexpression	Enrichment Analysis	04/14/2014 07:44:38 pm MDT	Processing	
1398027129	Williams_82_Control_vs_Williams_82_Chitin_octomer	Enrichment Analysis	04/20/2014 02:52:09 pm MDT	Done	Download Result
1400387557	control_11272_vs_control_11268	Enrichment Analysis	05/17/2014 10:32:37 pm MDT	Processing	

Figure 4.12 User Account Jobs List

Download Result							
PFAM							
KOG							
Pfam_ID	Pop_freq	Pop_frac	Study_frac	raw_Pval	corr_Pval	Description	Genes_with_term
KOG0626	0.001	53/73320	8/794	0.000	0.000	"Beta-glucosidase, lactase phlorizinhydrolase, and related proteins"	Glyma11g13810.1 Glyma11g13820.1 Glyma11g13820.1 Glyma11g13830.1 Glyma11g13850.1 Glyma12g05800.1 Glyma12g15620.1 Glyma11g15420.1
KOG2344	0.001	44/73320	5/794	0.000	0.010	Exocyst component protein and related proteins	Glyma12g08020.1 Glyma14g37840.1 Glyma19g35880.1

Figure 4.13 Enrichment Analysis Result Display Page

4.3 Methods

No matter whether the doctor wants to know which gene is the main cause of lung cancer, or the soybean researcher wants to know how a specific gene affects the flowering time of soybeans, researchers always have to identify the special genes they need to look at. Differential expression analysis empowers the biologists to find the target genes by comparing the difference between expressed levels under various conditions to see whether conditions are up-regulated or down-regulated. After acquiring the information, one could proceed to the next step.

In many omics, Differential Expression Analysis is considered one of the most important methods to vindicate hypotheses and assumptions. For example in transcriptomics, DNA microarray has been used for years to generate differential expression data that quantifies the abundance of mRNA corresponding to different genes. And in recent years, high-throughout sequencing of cDNA (RNA-Seq) also appears to be a good way to analyze gene expression. With the decreasing cost of sequencing, both methods all allow researchers to generate more datasets for well-deigned experiments, which leads to more precise result rather than analyzing few replicates. However, having more datasets means researchers need to spend more time analyzing data. For speed up the analyzing process, SoyKB offers users the Differential Expression tool, especially soybean researchers. Users will input well-known measurements such as p-value, q-value (if have) and fold change for filtering results in specific dataset. And

for RNA-Seq datasets specifically, datasets generating methods may be provided with cuffdiff or edgeR methods.

5. PROTEIN BIOVIEWER

Protein is an important material to maintain biological functions of our bodies. Topics about protein are hot among researchers. The project, Protein BioViewer which is introduced here, consists of all soybean gene secondary sequences analysis, TMHMM analysis and image output. It is developed based on Linux, Apache, MySQL and PHP (LAMP). Before writing this paper, first version of Protein BioViewer, which was generated based on gene model v1.0, has been integrated into the website of SoyKB.

5.1. Introduction

Designed by using CSS, HTML5 and PHP graphics, the Protein BioViewer demonstrates primary protein sequence and projects secondary structures, amino acid properties, phosphorylation positions, trans-membrane helix positions and functional domain positions. Because SSpro4 only runs in the Linux system, we need some scripts written in C language to manage secondary sequence data. Most of the time, SSpro4 in this research was used to analyze protein secondary sequence and transmembrane helix. In gene model v1.0, Soybean has 75778 genes. By running SSpro4 [4], SoyKB got all gene secondary sequences. And by running TMHMM [16], SoyKB got all prediction of transmembrane helices in soybean proteins in gene model v1.0. Figure 5.1 shows the main page of Protein BioViewer by using Glyma1508990.1 as example.

Figure 5.1 Protein BioViewer Main Page

5.2. Requirement Analysis

Protein structure can mainly be separated into four different level structures. The primary structure is the linear amino acids sequence. As we know, 20 different amino acids residues constitute protein sequence. 20 kinds of amino acids will appear in standard amino acid abbreviations, see Figure 5.2. By folding and docking, the primary structure will become to secondary structure. Two main types of secondary structures are alpha helix and the beta strand. Both of them are defined by the hydrogen bonds between the main chains. Before designing this project, there wasn't any public available datasets for soybean related secondary structure and transmembrane protein information. Since understanding the protein basic knowledge is very important to researchers, SoyKB decided to construct a tool that includes available information in one page and is very nice for researching purpose. Secondary structure analysis, protein

characteristics information, domain information, transmembrane helix prediction will be introduced one by one in following paragraphs.

Full name	Three-letter abbreviation	One-letter abbreviation
Alanine	Ala	A
Cysteine	Cys	C
Aspartic acid	Asp	D
Glutamic acid	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V

Figure 5.2 Amino Acid with Three-letter and One-letter Mapping Table [33]

- **The suggestion to predict Secondary structure prediction** goes way back to 1990.

The tool, which has been used to do the secondary structure analysis, is SSpro4 [4].

SSpro4's analysis is based on information about protein evolution and homologous

protein's secondary structure. For software detail, please refer to J. Cheng et al. 2005

[4] and G. Pollastri et al. (2002) [22].

Figure 5.3 3D Protein Secondary Structures

SoyKB already has secondary structure viewer in 3D version (Figure 5.3), but in order to combine with other information, it has to be converted into 2D version.

Figure 5.4 is the expected output for Secondary Structure Predictions.

Figure 5.4 The Inspiration of Protein BioViewer

- **Amino acids characteristics** are crafted based on the information in table 5.1. All characteristics are grouped based on their attributes, see table 5.5. Visualize all characteristics in one page greatly shorten the time for information search.

Characteristic	Color	Output
Acidic	Red
Basic	Green	
Neutral	Grey	
Hydrophobic	Dark blue
Hydrophilic	Blue	
Large	Red
Medium	Green	
Small	Dark blue	
Surface	Yellow
Buried	Maroon	
Positive	Red
Negative	Brown	
Aliphatic	Orange
Aromatic	Purple	
Acyclic	Blue
Cyclic	Red	

Table 5.1 Characteristics and Color Mapping Relationship with Output Example

Amino Acid Table																				
Residues	ala	arg	asn	asp	cys	glu	gln	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
acidic				•		•														
acyclic	•	•	•	•	•	•	•	•		•	•	•	•				•	•		•
aliphatic	•							•		•	•									•
aromatic									•					•					•	•
basic		•							•			•								
buried	•				•					•	•		•	•				•	•	•
charged		•		•		•			•			•								
cyclic									•						•	•			•	•
hydrophobic	•							•		•	•	•	•	•	•			•	•	•
large		•				•	•		•	•	•	•	•	•				•	•	•
medium			•	•	•										•		•			•
negative				•		•														
neutral	•		•		•		•	•	•	•	•		•	•	•	•	•	•	•	•
polar		•	•	•	•	•	•		•			•				•	•			
positive		•							•											
small	•							•									•			
surface		•	•	•	•	•	•	•	•			•				•	•	•	•	•

Figure 5.5 Amino Acids Characteristics Mapping Table

- **Domain** information is also a part of the Protein BioViewer. For gene model v1.0, COIL, GENE3D, PANTHER [19], Pfam [23], PIR [32], PRINTS [2], ProDom [27], PROSITE [29], SEG, SMART [26], SUPERFAMILY [9] and TIGRFAMs [10] are available at this time.

The data sources have been introduced in Section 4.3.4. Each domain existing in a specific kind of protein will be marked by an ellipse with a link to resource databases.

- **Transmembrane helix analysis** in Protein BioViewer adopted TMHMM [16] as prediction method to predict all genes in gene model v1.0. Genes were grouped, predicted online and transferred into SoyKB database by self-coded scripts. Figure 5.6 shows the Glyma15g08990.1 transmembrane helix result from TMHMM. The outside, TMhelix and inside position range are stored in databases.

```
# Glyma15g08990.1 Length: 597
# Glyma15g08990.1 Number of predicted TMHs: 2
# Glyma15g08990.1 Exp number of AAs in TMHs: 44.14975
# Glyma15g08990.1 Exp number, first 60 AAs: 0.0008
# Glyma15g08990.1 Total prob of N-in: 0.03202
Glyma15g08990.1 TMHMM2.0      outside    1  263
Glyma15g08990.1 TMHMM2.0      TMhelix     264 283
Glyma15g08990.1 TMHMM2.0      inside     284 467
Glyma15g08990.1 TMHMM2.0      TMhelix     468 490
Glyma15g08990.1 TMHMM2.0      outside    491 597
```

Figure 5.6 Glyma15g08990.1 TMHMM Web Output

5.3. System Design

Since PHP was used as SoyKB programming language, by keeping consistency, LAMP framework was used to build Protein BioViewer system. Additionally, in order to analyze all secondary data from SSpro4 and html data from TMHMM efficiently, C language was used to analysis data and PHP was used to store data in database. Figure 5.7 shows the basic Protein BioViewer workflow. The workflow of each component of Protein BioViewer will also be introduced in following paragraphs.

Figure 5.7 Protein BioViewer Workflow

Figure 5.8 Secondary Structure Analysis Workflow

Secondary structure prediction was finished in one month since there are some limitations of SSpro4. Workflow of analyzing is demonstrated in Figure 5.8. Before predicting the secondary structure of all 75778 genes in SSpro4, some preprocessing steps need to be taken first. Due to the limitation of SSpro4, the input total number of

file lines cannot be bigger than 900. Then 75778 genes were divided into 150 files with 450 genes each. And the input files also need to be cleaned up because of some special characters, which SSpro4 cannot accept in input files. After this step, gene sequences are ready to be inputted into SSpro4 with testing script. The analyzing process relies on six CPUs server. Six analyzing threads were used and each took approximately 10 hours. Figure 5.9 and 5.10 give the input and output file examples.

```
>Glyma15g08990.1
MENSGLALLAFRARITYDPFNALVNWNPNDPCKWLGVHCVDGKVQMM
>Glyma15g09000.1
MTHMSTTKEVASYQMHTKSFRDSWGITTNLFCFFIFF
```

Figure 5.9 Input File

```
>Glyma15g08990.1
MENSGLALLAFRARITYDPFNALVNWNPNDPCKWLGVHCVDGKVQMM
CCCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCEEECCCCCCCC
>Glyma15g09000.1
MTHMSTTKEVASYQMHTKSFRDSWGITTNLFCFFIFF
CCCCCCHHHHHHHHHCCCCCCCCCCCCCHEHHHEEC
```

Figure 5.10 Output File

After getting the secondary sequence, information processing needs to be added to Protein BioViewer. On web page, we use different images to represent each letter, as showed in Table 5.2.

Letter	-	C	H	E
Output	—

Table 5.2 SSpro4 Output Letter and Secondary Web Output Mapping Relationship

Transmembrane protein has two types, Alpha-helical and Beta-barrels showing in Figure 5.11. Since Beta-barrels so far found only in outer membranes of Gram-negative bacteria, cell wall of Gram-positive bacteria, and outer membranes of

mitochondria and chloroplasts. SoyKB only predict Alpha-helical in Protein BioViewer. Transmembrane helix (TMH) prediction has been done years ago and TMHMM already achieved a great result, so all 75778 genes were predicted by TMHMM. As introduced in requirement analysis, all prediction results were stored at local system and were analyzed by self-coded scripts into SoyKB databases. The analysis process based on workflow showed in Figure 5.12.

Figure 5.11 Schematic representation of transmembrane proteins: 1. a single transmembrane α -helix (bitopic membrane protein) 2. a polytopic transmembrane α -helical protein 3. a polytopic transmembrane β -sheet protein

Figure 5.12 TMHMM Analysis Workflow

As example, Glyma15g08990.1 related TM Helix information showed in table 5.3 and correspondingly Figure 5.13 shows the Glyma1508990.1 TM Helix result in Protein BioViewer. The helix positions are enclosed in the red box. The representation of the data in the output, Figure 5.13 was inspired by Figure 5.11.

ID	Name	Start	End
Glyma15g08990.1	outside	1	263
Glyma15g08990.1	TMhelix	264	283
Glyma15g08990.1	inside	284	467
Glyma15g08990.1	TMhelix	468	490
Glyma15g08990.1	outside	491	597

Table 5.3 Glyma15g08990.1 Transmembrane Helix Information

Figure 5.13 Glyma15g08990.1 Output. a) Across Transmembrane into protein b) Across Transmembrane leave protein

5.4. Demonstration

As an individual tool, users can use Protein BioViewer under “Tools” in header bar on SoyKB directly. Figure 5.14 shows the gene selection page. It also is integrated into other pages. For example, in gene page showed in Figure 5.15, user can access Protein BioViewer by clicking on the button ‘Open Protein BioViewer’.

Figure 5.14 Protein BioViewer Gene Input Page

Figure 5.15 The Protein BioViewer Link from Gene Page

6. SUMMARY

There are future works related to three tools introduced above. SoyKB is trying to merge new gene model into all tools and also keeps old version working at same time. For development purpose, all three tools need to support new gene model. The SNPviz and Differential Expression are used base on user selection about the specific dataset. Both tools need to identify which gene model are used to generate the selected data. In this case, mapping table showing the relationships between gene models and datasets, which are available on SoyKB, needs to be created. For Protein BioViewer, it needs to predict all secondary structure and transmembrane helix information with new gene model. TMHMM 2.0 has been announced with stand-alone version. It will accelerate analysis time. SSpro 4.0 will also be used to predict gene secondary structure.

Three tools has been integrated into SoyKB. Improvement and future development of all three tools will continue under the guidance of users' feedbacks. With modern technology used in SoyKB, researchers do not need to spend more time to make user-friendly output by themselves. Asking SoyKB group for help makes data analysis easier and more convenient. And this is one of the purposes to build SoyKB website. Three tools are developed in cooperation with biologists and engineers. With their efforts and new tools, SoyKB is becoming more and more handy.

BIBLIOGRAPHY

- [1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-9, May 2000.
- [2] T. K. Attwood, P. Bradley, D. R. Flower, A. Gaulton, N. Maudling, A. L. Mitchell, *et al.*, "PRINTS and its automatic supplement, prePRINTS," *Nucleic Acids Res*, vol. 31, pp. 400-2, Jan 1 2003.
- [3] C. I. Castillo-Davis and D. L. Hartl, "GeneMerge--post-genomic analysis, data mining, and hypothesis testing," *Bioinformatics*, vol. 19, pp. 891-2, May 1 2003.
- [4] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi, "SCRATCH: a protein structure and structural feature prediction server," *Nucleic Acids Res*, vol. 33, pp. W72-6, Jul 1 2005.
- [5] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biol*, vol. 4, p. 210, 2003.
- [6] T. J. D., H. D. G., and G. T. J., "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, 1994.
- [7] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Building phylogenetic trees. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.*, 1st ed. United Kingdom: Cambridge University Press, 1998.
- [8] R. J. Edwards. (2013). Available:
<http://www.southampton.ac.uk/~re1u06/teaching/upgma/>
- [9] J. Gough, K. Karplus, R. Hughey, and C. Chothia, "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure," *J Mol Biol*, vol. 313, pp. 903-19, Nov 2 2001.

- [10] D. H. Haft, J. D. Selengut, and O. White, "The TIGRFAMs database of protein families," *Nucleic Acids Res*, vol. 31, pp. 371-3, Jan 1 2003.
- [11] T. Joshi, M. R. Fitzpatrick, S. Chen, Y. Liu, H. Zhang, R. Z. Endacott, *et al.*, "Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding," *Nucleic Acids Res*, vol. 42, pp. D1245-52, Jan 2014.
- [12] T. Joshi, K. Patil, M. R. Fitzpatrick, L. D. Franklin, Q. Yao, J. R. Cook, *et al.*, "Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics," *BMC Genomics*, vol. 13 Suppl 1, p. S15, 2012.
- [13] KEGG. Available: [http://www.genome.jp/kegg-bin/show_pathway?gmx00030]
- [14] M. Y. Kim, S. Lee, K. Van, T. H. Kim, S. C. Jeong, I. Y. Choi, *et al.*, "Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome," *Proc Natl Acad Sci U S A*, vol. 107, pp. 22032-7, Dec 21 2010.
- [15] E. V. Koonin, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, D. M. Krylov, K. S. Makarova, *et al.*, "A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes," *Genome Biol*, vol. 5, p. R7, 2004.
- [16] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *J Mol Biol*, vol. 305, pp. 567-80, Jan 19 2001.
- [17] H. M. Lam, X. Xu, X. Liu, W. Chen, G. Yang, F. L. Wong, *et al.*, "Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection," *Nat Genet*, vol. 42, pp. 1053-9, Dec 2010.
- [18] T. Langewisch, H. Zhang, R. Vincent, T. Joshi, D. Xu, and K. Bilyeu, "Major Soybean Maturity Gene Haplotypes Revealed by SNPviz Analysis of 72 Sequenced Soybean Genomes," *PLoS One*, vol. 9, p. e94150, 2014.

- [19] H. Mi, A. Muruganujan, and P. D. Thomas, "PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees," *Nucleic Acids Res*, vol. 41, pp. D377-86, Jan 2013.
- [20] J. C. Oliveros. (2007, VENNY. An interactive tool for comparing lists with Venn Diagrams. Available: <http://bioinfogp.cnb.csic.es/tools/venny/index.html>
- [21] F. Opperdoes. (1997). Available: <http://www.icp.ucl.ac.be/~opperd/private/upgma.html>
- [22] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles," *Proteins*, vol. 47, pp. 228-35, May 1 2002.
- [23] M. Punta, P. C. Coghill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, *et al.*, "The Pfam protein families database," *Nucleic Acids Res*, vol. 40, pp. D290-301, Jan 2012.
- [24] L. Schausser. March 12, 2013). Multiple alignment: Heuristics. Available: http://www.daimi.au.dk/~schauser/genome_analysis_F03/lectures_F03/clustalW.pdf
- [25] R. J. Schmitz, Y. He, O. Valdes-Lopez, S. M. Khan, T. Joshi, M. A. Urich, *et al.*, "Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population," *Genome Res*, vol. 23, pp. 1663-74, Oct 2013.
- [26] J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting, "SMART, a simple modular architecture research tool: identification of signaling domains," *Proc Natl Acad Sci U S A*, vol. 95, pp. 5857-64, May 26 1998.
- [27] F. Servant, C. Bru, S. Carrere, E. Courcelle, J. Gouzy, D. Peyruc, *et al.*, "ProDom: automated clustering of homologous domains," *Brief Bioinform*, vol. 3, pp. 246-51, Sep 2002.
- [28] P. Sestoft. Programs for biosequence analysis. *Chapter 7: Phylogenetic trees*. Available: <http://www.itu.dk/people/sestoft/bsa/Match7.java>

- [29] C. J. Sigrist, E. de Castro, L. Cerutti, B. A. Cucho, N. Hulo, A. Bridge, *et al.*, "New and continuing developments at PROSITE," *Nucleic Acids Res*, vol. 41, pp. D344-7, Jan 2013.
- [30] teethgrinder. *Open Flash Chart*. Available: <http://teethgrinder.co.uk/open-flash-chart/>
- [31] H. Wickham, *ggplot2: elegant graphics for data analysis*: Springer New York, 2009.
- [32] C. Wu and D. W. Nebert, "Update on genome completion and annotations: Protein Information Resource," *Hum Genomics*, vol. 1, pp. 229-33, Mar 2004.
- [33] S. Z. Yu Tian Wang, Changiz Taghibiglou, "Inhibitors of phosphatase and tensin homolog (pten) compositions, uses and methods," N. a. a. o. t. s. L.-a. a. c. f. i. n. o. peptides, Ed., ed, 2011.