



Michigan Technological University
Create the Future Digital Commons @ Michigan Tech

Dissertations, Master's Theses and Master's
Reports - Open

Dissertations, Master's Theses and Master's
Reports

2012

POPLAR GENE EXPRESSION DATA ANALYSIS PIPELINES

Xiang Li

Michigan Technological University

Follow this and additional works at: <https://digitalcommons.mtu.edu/etds>

 Part of the [Computer Sciences Commons](#)


Copyright 2012 Xiang Li

Recommended Citation

Li, Xiang, "POPLAR GENE EXPRESSION DATA ANALYSIS PIPELINES", Master's report, Michigan Technological University, 2012.

<https://doi.org/10.37099/mtu.dc.etds/603>

Follow this and additional works at: <https://digitalcommons.mtu.edu/etds>

 Part of the [Computer Sciences Commons](#)

POPLAR GENE EXPRESSION DATA ANALYSIS PIPELINES

By

Xiang Li

A REPORT

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Computer Science

MICHIGAN TECHNOLOGICAL UNIVERSITY

2012

©2012 Xiang Li

This report has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Computer Science.

Department of Computer Science

Report Advisor: Dr. Hairong Wei

Committee Member: Dr. Laura Brown

Committee Member: Dr. Zhenlin Wang

Committee Member: Dr. Victor Bursov

Department Chair: Dr. Charles Wallace

Contents

Chapter 1.....	1
Introduction to Poplar Gene Analysis Pipeline	1
1.1 Background	1
1.2 Goal and objectives.....	2
1.3 Design.....	2
1.4 Pipeline flowchart	3
1.5 Roles and tasks.....	5
Chapter 2.....	6
Poplar Gene Expression Data Analysis Pipeline.....	6
2.1 Identification of DEGs	6
2.1.1: Trimming part of genes.....	6
2.1.2: Calculating p-values for each gene	6
2.1.3: Trimming genes with larger p-values	8
2.1.4: P-value correction.....	8
2.1.5: Selection of DEGs.....	8
2.2 Identification of Differentially Expressed Pathways	10
2.2.1: Calculate the fold change for each poplar gene.....	11
2.2.2: Call SAM_GS program to evaluate all the pathways defined in pathway matrix file	12
2.2.3: Calculate the average fold change for each pathway	12
2.3 Domain Enrichment Analysis	13
2.3.1: Calculate genome size, genome number, sample size, and sample number	13
2.3.2: Calculate Enrichment Factor.....	14
2.3.3: Calculate Enrichment Score.....	14
2.3.4: Calculate Average Fold Change	14
2.4 GO enrichment Analysis	16
2.4.1: Generate a list of Arabidopsis genes of DEGs in each time point	17
2.4.2: Call TermFinder to generate GO terms given the list of Arabidopsis genes ..	18
2.4.3: Mapping the Arabidopsis gene list of each GO back to poplar genes, and calculating the genome size, genomic gene number, sample size, sample number	18

2.4.4: Appending the annotation of Arabidopsis genes to the result	19
2.4.5: Identification of common GO terms and Unique GO terms.....	20
Chapter 3.....	23
GO Hierarchy Analysis.....	23
3.1 Different types of GO relations.....	23
3.2 Algorithm used to generate GO tree	24
3.3 Usage of online tool – GO tree	26
Chapter 4.....	28
GO Hierarchy Analysis Poplar Gene Expression Data Analysis On-line Tool	28
4.1 Introduction to poplar gene expression data analysis on-line tool	28
4.2 User Registration and Portion	28
4.2.1 Register	28
4.2.2 Duplicate username	29
4.2.2 Duplicate Email	30
4.3 User Login Portion	31
4.3.1 Login.....	31
4.3.2 Password retrieval	32
4.4 Data analysis Pipeline	33
4.4.1 Identification of DEGs	33
4.4.2 Pathway Enrichment Analysis.....	40
4.4.3 Domain Enrichment Analysis	42
4.4.4 GO Enrichment Analysis.....	44
Conclusion and Future Work	47
Reference:	48

Abstract

Analyzing large-scale gene expression data is a labor-intensive and time-consuming process. To make data analysis easier, we developed a set of pipelines for rapid processing and analysis popular gene expression data for knowledge discovery. Of all pipelines developed, differentially expressed genes (DEGs) pipeline is the one designed to identify biologically important genes that are differentially expressed in one of multiple time points for conditions. Pathway analysis pipeline was designed to identify the differentially expression metabolic pathways. Protein domain enrichment pipeline can identify the enriched protein domains present in the DEGs. Finally, Gene Ontology (GO) enrichment analysis pipeline was developed to identify the enriched GO terms in the DEGs.

Our pipeline tools can analyze both microarray gene data and high-throughput gene data. These two types of data are obtained by two different technologies. A microarray technology is to measure gene expression levels via microarray chips, a collection of microscopic DNA spots attached to a solid (glass) surface, whereas high throughput sequencing, also called as the next-generation sequencing, is a new technology to measure gene expression levels by directly sequencing mRNAs, and obtaining each mRNA's copy numbers in cells or tissues.

We also developed a web portal (<http://sys.bio.mtu.edu/>) to make all pipelines available to public to facilitate users to analyze their gene expression data. In addition to the analyses mentioned above, it can also perform GO hierarchy analysis, i.e. construct GO trees using a list of GO terms as an input.

Chapter 1

Introduction to Poplar Gene Analysis Pipeline

1.1 Background

The DNA sequences of genes carry the codes for synthesis of functional gene product – protein. When genes are expressed, messenger RNAs (mRNAs) are transcribed by using DNA as templates, and are subsequently translated into various functional proteins. We refer to this process as gene expression. However, due to the genetic programs and various environmental conditions, not all genes are expressed at the same time and at the same levels. Different gene expression profiles can lead to the various gene dosages that define phenotypes. Different genes may participate in different biological process, and function in different tissues, and various environmental conditions. Hence analyzing gene expression data plays an essential role in getting better understanding how life functions.

DNA microarrays, or DNA chips is the technology that facilitate the parallel execution of experiments on all genomic genes simultaneously ^[1]. Such technology measures mRNAs in cells or tissues at any moment. In the microarray experiment, Gene-specific DNA fragments are immobilized on a glass slide. Then mRNAs extracted from control or treated tissues are labeled with fluorescence dyes, cy5 and cy3, and used to hybridize to the array under the law of base-pairing rule. After hybridization is done, we can measure the quantity of each mRNA species from the same genes by measuring the spot lit on the chip, such quantity yields raw value of gene expression. This technology is called microarray data. The data we used is yielded from Affymetrix poplar genome chips that contain 61413 gene-specific fragments for interrogating gene expression of 61413 genes

Due to the advent of new DNA sequencing technologies, high throughput sequencing technology, also called next-generation sequencing technologies are now emerged as an alternative technology to microarray for gene expression analysis^[2]. The development of high throughput sequencing is driven by the high demand for low-cost sequencing. Such technique can produce thousands or millions of sequences at once^{[4] [5]} by taking advantage of parallelizing the sequencing process. High throughput sequencing is a technique that can be used directly sequencing mRNA species. Several high throughput sequencing methods are taking advantage of the power of massively-parallel sequencing to map short DNA sequence on long genome with large scale. Typically they will count the abundance of each mRNA species in cell or tissues.

Analyzing microarray or high throughput sequencing data is tedious and very time-consuming when the data are yielded from various tissues and in a time series. To reduce the workload for biologists, we developed multiple gene expression pipelines and assembled them together to facilitate the analysis of gene expression data. We focus on the gene expression data from a bioenergy species, poplar. Poplar is not only a top-quality source of biofuel, but also the first tree species to have its entire genome sequenced. Our gene expression package as shown (<http://sys.bio.mtu.edu>) will not only help biologists save a lot of time but also identify important genes involved in different biological processes.

1.2 Goal and objectives

The goal of this project is to develop a set of tools for rapid analyzing gene expression data for model species, poplar. The tools can be used to identify DEGs, pathways, enriched protein domains and GO terms. The input data type can be either microarray data or high throughput data.

The objectives of the pipeline tools are to:

1. Analyze gene expression data for identifying DEGs and gene sets by applying different statistical methods.
2. Evaluate the results from DEG pipeline; identify the enriched protein domains and GO terms. Calculate the enrichment factors, enrichment scores and average fold changes for enriched protein domains and GO terms.
3. Perform GO hierarchy analysis of a given list of GO terms. Then construct the GO tree based on the relations among the given GO terms.
4. Implement web application of all developed pipelines with friendly user interfaces to facilitate the use of the poplar gene analysis pipeline.

My work concentrated on the development of both command line package and web application. In the web application, we allow users to upload their data obtained from their experiments. We will run multiple analyses in background and send the result links via e-mail to notify the user.

1.3 Design

Our pipeline line tools can take gene expression data from a time series or multiple comparisons. The tools analyze data from each time point or treatment

independently. We use a data set of multiple time points as an example to demonstrate the usages of the developed pipelines. We have the following restrictions for gene expression files uploaded:

1. At each time point, expression file must contain control (untreated) replicates and treatment replicates. In most cases control replicates represent the information from wild type plants.
2. Due to the statistical methods we use, each time point must have at least 2 control replicates and 2 treatment replicates.
3. Data from each time point may have different number of control replicates and treatment replicates but the numbers of control replicates and treatment replicates must be consistent throughout all time points.

DEG pipeline takes the gene expression data as input and try to identify the DEGs for each time point. Here we used Rank Product method ^[3]to identify the DEGs for microarray data. Pathway (or DEG set) analysis attempts to identify the DEG set by taking the gene expression data and the pathway matrix files as input. Pathway analysis is applied to gene expression data across all the time points. . The pathway matrix is a file in which each row is a gene and each column is a pathway, and a number in each intersection is either 0 or 1. A value of one indicates this gene is in this pathway, zero otherwise. Domain and GO enrichment analyses are applied to the DEGs that were output from DEG pipeline. After getting the DEGs from each time points, we extract them separately and try to identify the corresponding protein domains and GO terms. Calculate the enrichment factors, enrichment scores as well as the average fold changes of the genes included.

1.4 Pipeline flowchart

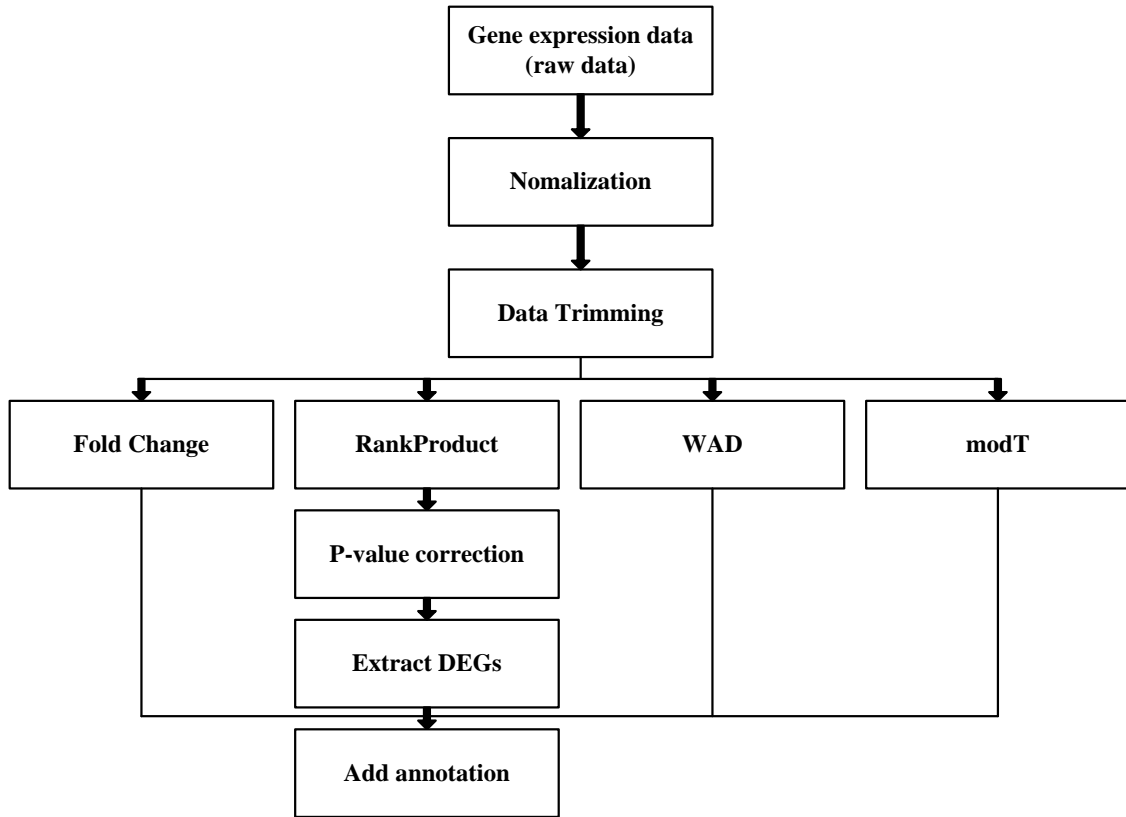


Figure 1: DEG analysis flowchart

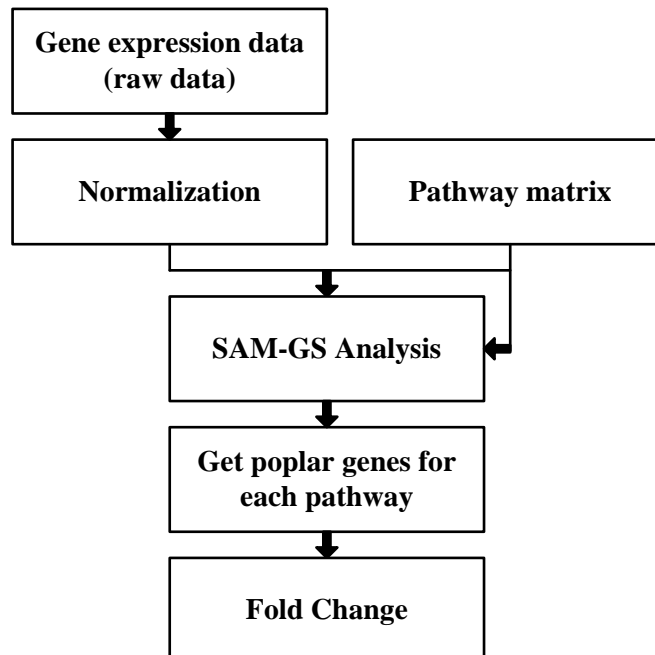


Figure 2: Pathway Analysis Flowchart

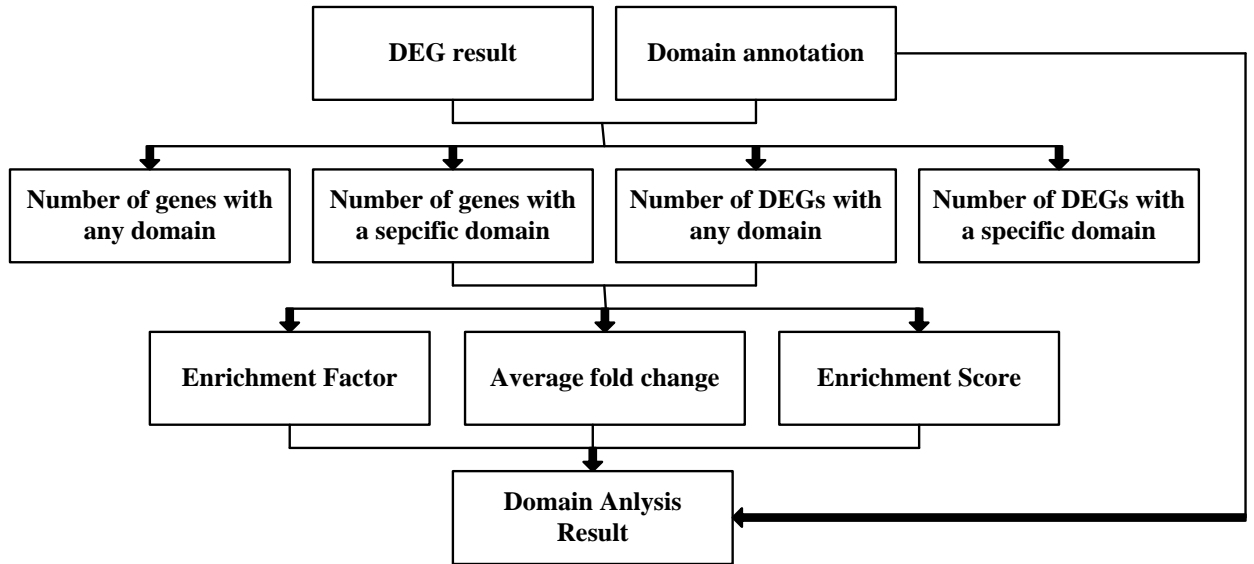


Figure 3: Domain Enrichment Flowchart

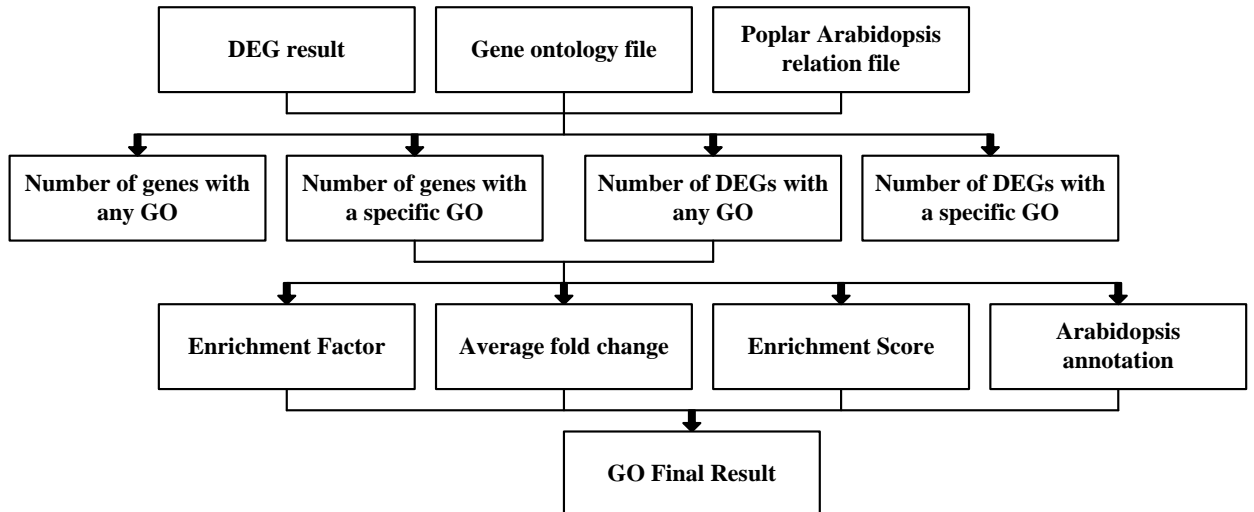


Figure 4: GO Enrichment Analysis Flowchart

1.5 Roles and tasks

I was responsible for the development of command-line pipeline programs and the web tools. I also performed the data analyses after getting results from pipeline program. Chapter 2 introduces the usage and functions of various pipeline programs I developed. Chapter 3 shows the GO hierarchy analysis. Chapter 4 provides the usage of our online tool.

Chapter 2

Poplar Gene Expression Data Analysis Pipeline

2.1 Identification of DEGs

To identify DEGs, the pipeline perform the following procedures:

2.1.1: Trimming part of genes

Genes with high expression values and standard deviations usually produce significant biological effect. Thus we exclude those genes with lower expression values and smaller standard deviation at the very beginning.

The DEG pipeline ranks all the genes by their average expression values and standard deviations separately in descending order, and then combines these two ranks and calculates the average rank for each gene. The pipeline does the trimming according to this average rank.

Trimming gene is optional. Users can specify the cut-off value represents the trimming percentage. For instance, 0.2 represents trimming 20 percent genes.

2.1.2: Calculating p-values for each gene

We integrate Rank Product into our pipeline for identifying DEGs. Rank Product is a non-parametric statistic method for the detection of DEGs for microarray data^[3]. It is based on the rankings of fold changes. The formula for calculating the rank product is as follows:

$$\text{RankProduct}(g) = (\prod_{i=1}^k r_{g,i})^{1/k}$$

Where g represents a specific gene, n represents the number of all the genes; k represents the number of replicates; $r_{g,i}$ represents the rank of i th replicate of gene g .

This method assumes that the orders of all the genes are random, i.e. the probability of finding a specific gene among the top r of n genes list equals to r/n .

Rank Product method is fast and simple. It provides a straightforward and statistically stringent way to determine the significance level for each gene and reliable in high noisy data ^[3].

Besides Rank Product, we also implemented two additional methods to identify the DEGs, they are Weight Average Difference (WAD) and moderated T-statistics (modT).

WAD is a fold-change-based method. It uses the average difference and relative average signal intensity so that DEGs are highly ranked on the average for the different conditions. ^[6]

For an individual gene i in time point j , we need to calculate the average difference of this gene first. The average difference of gene i in time point j is calculated as followings:

$$AD_{ij} = AVG_{ij}^{control} - AVG_{ij}^{treatment}$$

$AVG_{ij}^{control}$ represents the average log signal of all the control replicates belong to gene i in time point j , $AVG_{ij}^{treatment}$ represents the average log signal of all the treatment replicates belong to gene i in time point j .

Average log signal intensity w_{ij} is used to evaluating the average difference of all the replicates of gene i in time point j .

$$w_{ij} = \frac{AVG_{ij}^{vector} - \min(vector)}{\max(vector) - \min(vector)}$$

AVG_{ij}^{vector} represents the average log signal of all the replicates belonging to gene i in time point j . Vector is used to store the all the replicates of gene i in time point j , and the result is calculated as follows:

$$WAD_{ij} = AD_{ij} * w_{ij}$$

The larger the WAD_{ij} , the more significant a differentially expressed a gene is. WAD statistics can be an alternative of rank product in defining DEGs.

The modT is considered as reformulated posterior odds statistics, in which posterior residual standard deviations are used in place of ordinary standard deviations. ^[7] Moderated t-statistic follows a t-distribution with augmented degrees of freedom, and it generates a more predictable ranking of genes compared with ordinary t-statistics.

We also implemented the modT method to calculate the rankings of poplar genes in each time point. The larger the modT value, the more significant a differentially expressed gene is. It is also an alternative of Rank Product in defining DEGs.

Other than these three methods, we also attempted to use ANOVA to identify DEGs. However, this method is too time-consuming and we finally decided not to integrate it.

2.1.3: Trimming genes with larger p-values

The selection of DEGs is based on the result of Rank Product method. We get p-value for each remaining genes in step 1. We sort genes based on p-values in ascending order and trim those genes with p-value larger than 0.1.

2.1.4: P-value correction

As dataset becomes larger, false positives get increased in DEG list. In order to make the false positive genes under control, we had to apply a multiple testing correction and calculate the corrected p-values.

Multiple methods are available for multiple testing of p-values. We chose Benjamini and Hochberg false discovery rate correction method^[8] in our program. This p-value correction method is the least stringent compared to other methods such as Bonferroni and Westfall and young Permutation.^[9]

Benjamini Hochberg FDR multiple correction method calculates corrected p-value as follows:

1. Sort the original p-values of all genes in ascending order.
2. Calculate corrected p-value by using following formula:

$$\text{Corrected p-value} = \text{Original p-value} * n / r$$

r stands for the rank of the gene and n stands for the total number of genes.

2.1.5: Selection of DEGs

After calculating corrected p-values, we can select DEGs. In our program, all the genes with corrected p-value smaller than 0.05 are defined as DEGs by default. This threshold is believed to be reasonable and also suggested by Benjamini Hochberg FDR correction method. Thus, the false discovery rate of the DEG list is limited to 5%. However, users can also specify their own threshold.

User must normalize their microarray expression data before using the pipeline to do analysis. Our pipeline tool takes the normalized microarray expression gene data file as input. The format of an example microarray expression gene data file is shown in Figure 5:

	A	B	C	D	E	F	G	H	I
1	probename	CTRL_6h_R1	CTRL_6h_R2	Treatment_6h_R1	Treatment_6h_R2	CTRL_12h_R1	CTRL_12h_R2	Treatment_12h_R1	Treatment_12h_R2
2	AFFX-BioB-3_at	8.709300654	8.690448642	9.477913874	9.307432709	9.043666658	9.136690919	9.306020406	9.254956336
3	AFFX-BioB-5_at	8.986986038	8.989446105	9.715366668	9.564624114	9.272237526	9.393854803	9.467609128	9.512218423
4	AFFX-BioB-M_at	9.326124856	9.349060966	10.03872345	9.936833056	9.721698344	9.763426925	9.953524721	9.875506813
5	AFFX-BioC-3_at	10.34546754	10.31057665	10.86547667	10.70983684	10.61196179	10.64099234	10.80724605	10.72618771
6	AFFX-BioC-5_at	10.06232235	10.05244904	10.63761711	10.49284954	10.31836684	10.36495202	10.57774787	10.43505492
7	AFFX-BioDn-3_at	12.42750849	12.37194945	12.82074274	12.75536662	12.69338156	12.69280768	12.79844035	12.64881905
8	AFFX-BioDn-5_at	11.51735598	11.45565915	11.81260521	11.76362891	11.69114992	11.73194164	11.83901016	11.66887207
9	AFFX-CreX-3_at	14.14194971	14.16672413	14.39177104	14.35525731	14.33023465	14.30735931	14.33073104	14.29702249
10	AFFX-CreX-5_at	13.95962628	13.93151555	14.10195357	14.09420777	14.05011463	14.07783052	14.11875069	14.07588817
11	AFFX-DapX-3_at	11.39782308	11.45030993	11.787602	11.91064919	11.81070897	11.70630603	11.82606465	11.43002282
12	AFFX-DapX-5_at	9.093997381	9.278619274	9.705430637	9.820820771	9.423781322	9.362969174	9.697945392	8.960581687
13	AFFX-DapX-M_at	10.3795747	10.47830552	10.75214754	10.85187263	10.69340295	10.57420421	10.8655395	10.31340146
14	AFFX-LysX-3_at	8.811129448	8.921840653	9.270084502	9.458066805	9.160851181	9.157210807	9.331408405	9.006251668
15	AFFX-LysX-5_at	7.211123999	7.416167668	7.904150299	7.944852436	7.644807949	7.624423627	7.850218621	7.414317792
16	AFFX-LysX-M_at	7.606008967	7.762413176	8.311319412	8.2749727	8.029024316	7.89919643	8.205869618	7.790080702
17	AFFX-PheX-3_at	8.534526523	8.712549709	9.129501083	9.198972398	8.957239501	8.955130787	9.087914799	8.807582747
18	AFFX-PheX-5_at	8.18267572	8.346424662	8.760239089	8.760476857	8.486902618	8.554234024	8.776510193	8.392394591
19	AFFX-PheX-M_at	8.191901705	8.255325767	8.696550883	8.793217768	8.545014066	8.603840228	8.695606526	8.319786778
20	AFFX-Ptp-actin-3_s_at	12.55082559	12.60700709	12.68307883	12.81992169	12.71560682	12.77042789	12.44716136	12.62100135

Figure 5: Microarray Gene Expression file

We have the following restrictions for the input microarray gene expression file:

- The input file should be a tab-delimited file.
- If we have multiple time points in the microarray gene expression file, the number of control and treatment replicates should be the same.
- In each time point, we should have at least two replicates for control and also for treatment.
- Row 1 provides the name of each column. The names of control replicates should start with “c” or “C” and the names of the treatment replicates should start with “t” or “T”.

Running command:

```
$ perl ./script/DEG_pipeline.pl -p gene -i file1.txt -g file2.txt -t 0.2 -tp 6
```

- p gene/pathway: ‘gene’ indicates we are doing DEG analysis and ‘pathway’ indicates we want to do pathway analysis.
- i <file>: Specifies the microarray gene expression file.
- g <file>: Specifies the microarray gene annotation file.
- t (0-1): Specifies the trimming percentage of the original data introduced in 2.1.1; the value should be between 0 and 1.
- tp: Specifies the time point number, this number should be positive integer.

The resulting file of a differentially expressed genes sample is shown in Figure 6:

	A	B	C	D	E	F	G	H	I	J
1	ProbeID	1 Avg_ctrl	1 Avg_treat	1 FoldChange	1 Rank_modT	1 Rank_WAD	1 RP_origPvalue	1 RP_correctedPvalue	1 Rank_RP	1 DEG
2	PtpAffx.53295.1.S1_at	41.50906116	38.40445535	-1.08083973	43010	37953	0.3472			nonDEG
3	PtpAffx.224655.1.S1_s_at	18.34824367	18.14094689	-1.01142701	58154	55539	0.4466			nonDEG
4	PtpAffx.102427.1.A1_s_at	384.6017434	243.880899	-1.57700642	6414	6901	0.0453	0.137303637	9048	nonDEG
5	Ptp.303.2.A1_at	38.72076724	13.93961806	-2.77774951	1248	7735	0.0034	0.04283728	2078	DEG
6	PtpAffx.206393.1.S1_s_at	3889.644684	673.9855423	-5.771109971	2617	58	1.00E-04	0.01149562	423	DEG
7	PtpAffx.218496.1.S1_at	48.51263417	64.07073023	1.320701943	22131	21249	0.1461			nonDEG
8	PtpAffx.224971.1.S1_at	16.62052848	20.81897111	1.252605845	29675	30664	0.2022			nonDEG
9	Ptp.2155.1.S1_at	846.3182373	3036.502323	3.587896597	302	222	0	0	20	DEG
10	PtpAffx.82275.1.A1_at	1198.03164	477.001173	-2.511590553	4219	857	0.0047	0.049281158	2460	DEG
11	PtpAffx.71036.1.S1_at	76.5303036	62.25285574	-1.229346071	24886	23578	0.17			nonDEG
12	PtpAffx.11727.1.A1_at	132.4413342	640.0912355	4.833017117	2913	378	0	0	28	DEG
13	PtpAffx.10911.1.S1_at	1585.534821	1236.5875	-1.28218571	18556	11044	0.1354			nonDEG
14	PtpAffx.204364.1.S1_at	433.2910001	329.6348552	-1.314457477	29399	11286	0.1068			nonDEG
15	PtpAffx.20863.1.A1_at	7742.467194	7004.539907	-1.105349858	39874	20105	0.3026			nonDEG
16	Ptp.2730.1.S1_at	71.25802974	123.5225711	1.733454763	8700	8391	0.0126	0.094314354	5324	nonDEG
17	PtpAffx.94956.1.A1_at	53.34345119	60.3863354	1.132029032	51771	35721	0.3154			nonDEG
18	PtpAffx.215889.1.S1_s_at	43.02002638	31.27219326	-1.375663869	18659	21501	0.103			nonDEG
19	Ptp.4063.2.S1_a_at	1850.156695	200.3893541	-9.232809312	27	50	0	0	43	DEG
20	PtpAffx.84732.1.S1_s_at	555.0758243	647.7839622	1.167018872	28139	20192	0.3664			nonDEG

Figure 6: Sample DEG result file

Row 1 contains the column name. From column 2, the integer in the front of each column name indicates which time point that result belongs to. Column 2 provides the sum of unlogged expression values of all control replicates in that time point. Column 3 stands for the sum of unlogged expression values of all treatment replicates in that time point. Column 4 presents the fold change. Column 5 provides the rank for each sample from the moderated T method. Column 6 contains the ranks from the WAD method. Column 7 contains the p-value results from the Rank Product method. Column 8 provides the corrected p-value. Column 9 provides the ranks from the Rank Product method. Column 10 indicates whether the current gene is DEG or not in a selected time point.

2.2 Identification of Differentially Expressed Pathways

Metabolic pathways are series of chemical reactions occurring within a cell. In each pathway, a precursor chemical is modified by a series of chemical reactions to produce one or multiple chemicals. Usually enzymes encoded by multiple genes catalyze these reactions coordinately.

Gene set analysis evaluates the expression of a set of genes catalyzing biological pathways.^[10] In this analysis, users need to upload their matrix file that is used to provide the relations between pathways and genes. Figure 7 shows the format of a pathway matrix file.

probeID	beta;-alanine biosynthesis I	beta;-alanine biosynthesis II	beta;-alanine biosynthesis III	cis-zeatin biosynthesis
Ptp.6223.1.S1_s_at	1	0	0	0
PtpAffx.24142.1.A1_at	1	0	0	0
PtpAffx.208466.1.S1_s_at	0	1	0	0
PtpAffx.208467.1.S1_at	0	1	0	0
PtpAffx.218950.1.S1_at	0	1	0	0
PtpAffx.223598.1.S1_at	0	1	0	0
PtpAffx.3156.1.A1_s_at	0	1	0	0
PtpAffx.51307.1.S1_s_at	0	1	0	0
PtpAffx.144250.1.S1_s_at	0	0	1	0
PtpAffx.144250.2.S1_at	0	0	1	0
PtpAffx.219043.1.S1_at	0	0	1	0
PtpAffx.95543.1.A1_at	0	0	1	0
PtpAffx.129568.2.A1_a_at	0	0	0	1
PtpAffx.126487.1.S1_s_at	0	0	0	0
PtpAffx.201244.1.S1_at	0	0	0	0
PtpAffx.205240.1.S1_at	0	0	0	0
PtpAffx.206898.1.S1_at	0	0	0	0
PtpAffx.249.331.S1_a_at	0	0	0	0
PtpAffx.249.331.S1_at	0	0	0	0

Figure 7: pathway matrix file

In the matrix file as shown, each row represents a poplar gene and each column represents a pathway. In this file 0 indicates the gene is not present in this pathway and 1 indicates the gene belongs to this pathway. In order to enable the pathway analysis using sample power, we combined the sample replicates for all the time points together prior to the analysis.

Significance Analysis of Microarray for Gene Sets (SAM-GS) was integrated into our pipeline for pathway enrichment analysis.^[7] Normally when we analyze whether a set of genes are associated with a phenotype, we measure to check whether the gene set consists of genes whose expression values continuously changed in one direction with the phenotype. Compared to GSEA (Gene Set Enrichment Analysis) method proposed by Mootha et al.^[11], SAM-GS is more sensitive, leading to the result of statistically significant gene sets more reliable.

We identified the differentially expressed pathways in the following steps:

2.2.1: Calculate the fold change for each poplar gene

Calculate the fold change for each individual poplar gene appears in the gene expression file in each time point separately. The calculation of fold change (FC) for given poplar gene g exists in time point i is as follows:

1. Calculate the sum of control replicates and treatment replicates in time point i . We refer them as $sum_control$, $sum_treatment$.

2. If $sum_treatment \geq sum_control$, $FC = sum_treatment / sum_control$.

If $sum_treatment < sum_control$, $FC = -1 * sum_control / sum_treatment$.

Negative FC indicates gene G in time point i is down-regulated, otherwise it is up-regulated.

2.2.2: Call SAM_GS program to evaluate all the pathways defined in pathway matrix file

SAM_GS program calculates the p-value and corrected p-value for all the pathways base on the members in current pathway. The smaller the corrected p-value, the more significant a pathway is.

2.2.3: Calculate the average fold change for each pathway

Suppose pathway P contains n genes, denoted as gene-1, gene-2, gene-3...gene-n.

- For each pathway gene, if FC of such gene is larger than 1, subtract it by 1; if it is smaller than -1, add 1.
- Add the resulting values of all the genes together. And then divided the sum by n.
- If the result for the previous step is a negative value, add -1. Otherwise add 1. This resulting value is average fold change of the pathway.

Running command:

```
$perl ./script/DEG_pipeline.pl -p pathway -i file1.csv -g anno/file2.csv
```

- p gene/pathway: the option of 'pathway' indicates the analysis is to identify differentially expressed rather than gene
- i <file>: Specifies the microarray gene expression file, this file should be CSV format, i.e. comma delimited file.
- g <file>: Specifies the pathway matrix file, this file should be CSV format, i.e. comma delimited file.

The result file of pathway analysis is shown in Figure 8:

	A	B	C	D	E	F	G
1	GS.stats.GS.name	GS.stats.GS.size	GS.stats.GS.p_value	GS.stats.GS.q_value	Sample genes	Avg_FC_1	Avg_FC_2
2	beta..alanine.biosynthesis.I	4	0.22	0.073552975	Ptp.6223.1.S1_s_at PtpAffx.24142.	1.123079328	-1.10390131
3	beta..alanine.biosynthesis.II	19	0	0	PtpAffx.208466.1.S1_s_at PtpAffx.	-1.168863948	-1.07359295
4	beta..alanine.biosynthesis.III	4	0.04	0.019609118	PtpAffx.144250.1.S1_s_at PtpAffx.	-2.132689441	1.184874118
5	cis.zeatin.biosynthesis	2	0.04	0.019609118	PtpAffx.129568.2.A1_a_at PtpAffx.	1.112594437	-1.136496832
6	de.novo.biosynthesis.of.uridine.5..monophosphate	8	0.05	0.023712113	PtpAffx.126487.1.S1_s_at PtpAffx.	-1.384438885	-1.061765606
7	arginine.biosynthesis.II..acetyl.cycle.	5	0.01	0.008639661	PtpAffx.220320.1.S1_at PtpAffx.21	-1.355483741	1.06476692
8	ent.kaurene.biosynthesis	3	0.1	0.040027787	PtpAffx.201731.1.S1_at PtpAffx.20	1.034604294	-1.052352464
9	S.adenosyl.L.methionine.cycle	11	0.04	0.019609118	Ptp.49.1.S1_at Ptp.52.1.S1_at Ptp.	-2.225956179	1.203897029
10	methionine.biosynthesis.II	11	0.04	0.019609118	Ptp.49.1.S1_at Ptp.52.1.S1_at Ptp.	-1.927389267	1.204607464
11	trans.lycopene.biosynthesis	3	0.01	0.008639661	PtpAffx.222709.1.S1_at PtpAffx.53	-1.095462514	1.099340811
12	trans.zeatin.biosynthesis	1	0.03	0.01678088	PtpAffx.204470.1.S1_at	-1.028142986	1.008461927
13	X1_4.dihydroxy.2.naphthoate.biosynthesis.II..plants.	1	0.02	0.012927493	PtpAffx.216312.1.S1_at	-1.021671263	1.091815285
14	X2.ketoglutarate.dehydrogenase.complex	4	0.06	0.026987807	Ptp.4041.2.S1_s_at PtpAffx.151462	-1.222636506	1.067167905
15	abscisic.acid.biosynthesis	18	0.13	0.049107683	PtpAffx.129974.1.S1_s_at PtpAffx.	1.14851028	-1.199176464
16	IAA.biosynthesis.I	3	0.09	0.037044584	PtpAffx.204443.1.S1_s_at PtpAffx.	-1.365869846	-1.032888906
17	abscisic.acid.glucose.ester.biosynthesis	16	0.1	0.040027787	PtpAffx.146922.2.A1_at PtpAffx.20	-2.425752449	1.097892291
18	cytokinins.O.glucoside.biosynthesis	30	0.04	0.019609118	Ptp.6751.1.A1_s_at Ptp.6831.1.S1	-2.007517721	-1.011844321
19	cytokinins.9.N.glucoside.biosynthesis	30	0.04	0.019609118	Ptp.6751.1.A1_s_at Ptp.6831.1.S1	-2.007517721	-1.011844321
20	cytokinins.7.N.glucoside.biosynthesis	30	0.04	0.019609118	Ptp.6751.1.A1_s_at Ptp.6831.1.S1	-2.007517721	-1.011844321

Figure 8: pathway analysis result

In the output file, each row represents one pathway. Column A contains the pathway name. Column B gives the number or poplar genes within this pathway.

Column C contains the p-value. Column D gives the q-value (corrected p-value). Column E shows the list of poplar genes current pathway corresponds to. The following column provides the average fold change of pathway in each time point.

2.3 Domain Enrichment Analysis

A protein domain is a conserved part or structure of protein sequence that can function independently from the rest of protein. Domains vary in length from between about 25 amino acids up to 500 amino acids in length. We downloaded the Interpro-scanner and associated databases to our server and ran it to analyze 73013 protein sequences and identified all protein domains contained in these sequences. Each domain in our analyses is represented by an Interpro ID. Domain annotation file in Figure 9 provides the corresponding relation between protein domains and poplar genes.

	A	B	C
1	domain	description	ProbeID
2	IPR000005	Helix-turn-helix,sAraCstype	PtpAffx.223505.1.S1_at
3	IPR000005	Helix-turn-helix,sAraCstype	PtpAffx.218921.1.S1_at
4	IPR000005	Helix-turn-helix,sAraCstype	PtpAffx.215626.1.S1_at
5	IPR000005	Helix-turn-helix,sAraCstype	PtpAffx.220281.1.S1_at
6	IPR000005	Helix-turn-helix,sAraCstype	PtpAffx.215016.1.S1_at
7	IPR000007	Tubby,sC-terminal	PtpAffx.200078.1.S1_at
8	IPR000007	Tubby,sC-terminal	Ptp.1282.1.S1_at
9	IPR000007	Tubby,sC-terminal	Ptp.1963.1.S1_at
10	IPR000007	Tubby,sC-terminal	Ptp.2238.1.S1_at
11	IPR000007	Tubby,sC-terminal	Ptp.2238.1.S1_s_at
12	IPR000007	Tubby,sC-terminal	Ptp.5835.1.S1_at
13	IPR000007	Tubby,sC-terminal	PtpAffx.124217.1.S1_s_at
14	IPR000007	Tubby,sC-terminal	PtpAffx.146084.1.S1_at
15	IPR000007	Tubby,sC-terminal	PtpAffx.26405.3.A1_at

Figure 9: protein-domain annotation file

One protein domain is usually contained by multiple poplar genes. In this file Column A shows the Interpro ID of the protein, Column B gives the description of certain protein domain, Column C provides the poplar genes contain current domain.

The outcome of biological events in the cells and tissues is determined by the protein-protein interaction through domains. The purpose of domain enrichment analysis is to identify the gene families that are thriving or over-represented in the data. Domain enrichment analysis is applied on the DEGs output from DEG pipeline. .

According to the protein domain annotation file we have, nearly half of the poplar genes have protein domains. We evaluate the enriched protein domains for each DEG list from various time points separately. The procedures in detail are as follows:

2.3.1: Calculate genome size, genome number, sample size, and sample number

In the program for protein domain enrichment analysis, we calculate enrichment score and enrichment factor to evaluate the degree of enrichment of a protein domain. To do this, we need to obtain four parameters.

1. Genome size

Genome size represents the total number of poplar genes that have at least one protein domain. It can be obtained by scanning the domain annotation file.

2. Genome number

Genome number represents the number of poplar genes that have a specific protein domain.

3. Sample size

Sample size represents the number of DEGs in current time point that have at least one protein domain.

4. Sample number

Sample number represents the number of DEGs that have a specific protein domain.

The genome size is fixed once the domain annotation file is given. The sample size is fixed given the specific time point. The values of genome number and sample number vary in domains and time points.

2.3.2: Calculate Enrichment Factor

The enrichment factor is calculated by the following formula:

$$EF_{domain} = \frac{num_sample/num_genome}{sample_size/genome_size}$$

Enrichment factor is one statistics to measure the enrichment level.

2.3.3: Calculate Enrichment Score

Enrichment score is the other criterion to evaluate the degree of a protein domain is enriched. The enrichment score of a specific domain is calculated by the quintile function for hypergeometric distribution – phyper:

$$ES_{domain} = \text{phyper}(num_{sample}, num_{genome}, size_{genome} - num_{genome}, size_{sample})$$

2.3.4: Calculate Average Fold Change

The average FC of a set of genes in a domain is also an important parameter for us to evaluate the activities.

Calculating the average FC of a domain is similar with calculating the average fold change of pathway. Suppose domain D is corresponding to n DEGs ($DEG_1, DEG_2, \dots, DEG_n$), their fold changes are FC_1, FC_2, \dots, FC_n . The fold change can only be up-regulated, i.e. the value is larger than 1 or down regulated, i.e. the value is smaller than -1.

At first, we calculate the sum of fold changes of all the DEGs by implementing the following formula:

$$SUM_{FC} = \sum_i^{N_{up}} (FC_{up_i} - 1) + \sum_j^{N_{down}} (FC_{down_j} + 1)$$

Then we use the following formula to calculate the average fold change:

$$AVG_{FC-D} = \begin{cases} \frac{SUM_{FC}}{N_{up} + N_{down}} + 1, & SUM_{FC} > 0 \\ \frac{SUM_{FC}}{N_{up} + N_{down}} - 1, & SUM_{FC} < 0 \end{cases}$$

For example, if domain D is corresponding to 2 DEGs and there fold changes in certain time point are 2.5, -3. $SUM_{FC} = (2.5 - 1) + (-3 + 1) = -0.5$, average fold change $AVG_{FC-D} = \left(-\frac{0.5}{2}\right) - 1 = -1.25$.

Running command:

\$perl script/DomainAnalysis_pipeline.pl -i DEG_result.txt -g domain_anno.txt -tp n

- -i <file>: Specifies the DEG result file.
- -g <file>: Specifies the protein domain annotation file.
- -tp n: Specifies the time point number.

The result of domain analysis pipeline is shown in Figure 10:

A	B	C	D	E	F	G	H	I	J	K
Domain	Description	Time_appear	GenomeSize_1	NumGenome_1	SampleSize_1	NumSample_1	EnrichFactor_1	AverageFC_1	GenesinDomain_1	EnrichScore
IPR000010	Proteinase inhibitor I25,	1	NA	NA	NA	NA	NA	NA	NA	NA
IPR000023	Phosphofructokinase	4	34315	16	1704	2	2.51723885	-6.271506086	PtpAffx.161291.1.S1_at	0.04216383
IPR000047	Helix-turn-helix motif, la	3	34315	21	1704	1	0.958948133	-2.775838886	Ptp.7526.1.S1_at	0.28031393
IPR000056	Ribulose-phosphate 3-ep	2	34315	6	1704	2	6.712636933	-3.714266754	Ptp.1231.2.S1_s_at Ptp.	0.00218284
IPR000058	Zinc finger, AN1-type	2	34315	21	1704	2	1.917896266	2.759536933	PtpAffx.132656.1.A1_at	0.08350795
IPR000061	SWAP/Surp	1	34315	14	1704	2	2.8768444	2.045560642	Ptp.5855.1.S1_at PtpAff	0.02949416
IPR000070	Pectinesterase, catalytic	2	34315	92	1704	5	1.094451674	-3.657422156	Ptp.1705.1.A1_at PtpAff	0.30674419
IPR000089	Biotin/lipoyl attachment	1	34315	28	1704	3	2.1576333	-4.249747079	Ptp.6434.1.S1_s_at PtpA	0.04798496
IPR000092	Polyprenyl synthetase	1	NA	NA	NA	NA	NA	NA	NA	NA
IPR000095	PAK-box/P21-Rho-bindin	1	34315	20	1704	1	1.00689554	-2.585415797	PtpAffx.211787.1.S1_s_e	0.2615705
IPR000101	Gamma-glutamyltranspe	1	NA	NA	NA	NA	NA	NA	NA	NA
IPR000109	TGF-beta receptor, type	3	34315	101	1704	8	1.595082043	-2.696081064	PtpAffx.110895.1.S1_at	0.06383012
IPR000114	Ribosomal protein L16	1	34315	5	1704	1	4.02758216	2.019522077	Ptp.7926.1.S1_at	0.02228981
IPR000116	High mobility group, HM	2	34315	5	1704	1	4.02758216	-2.846044397	PtpAffx.22228.2.S1_at	0.02228981
IPR000118	Granulin	1	34315	11	1704	2	3.661438327	-4.438407483	PtpAffx.160901.1.S1_s_e	0.01493791
IPR000121	PEP-utilizing enzyme	1	NA	NA	NA	NA	NA	NA	NA	NA
IPR000133	ER lumen protein retaini	1	NA	NA	NA	NA	NA	NA	NA	NA

Figure 10: Result of protein domain analysis

In the output of protein domain analysis, Column A contains the Interpro ID. Column B gives the description of domain specified in column 1. Column C contains the number of times a certain domain's enrichment score is no larger than 0.05 among all time points. In the following columns one can find the

values of genome size, genome number, sample size, sample number, enrichment factor, enrichment score and average fold change. A number of DEGs corresponds to the each domain is also shown in the result. If NA is list in all the parameters in a certain domain, then such domain is not possessed by any DEG in current time point.

We can identify the common domains and unique domains from the result of the domain analysis. Common domains refer to those protein domains with enrichment score smaller than 0.05 at all the time points. Unique domains refer to those protein domains with enrichment score smaller than 0.05 at only one time point.

2.4 GO enrichment Analysis

GO is a major bioinformatics initiative to standardize the representation of gene and gene product attributes across all species.^[12]

In more details, the GO project aims at:

1. Maintain and develop its controlled vocabulary of gene and gene product attributes;
2. Annotate genes and gene products, and assimilate and disseminate annotation data;
3. Develop tools in order to facilitate the creation, maintenance as well as the usage of ontologies;

Since there is no universal standard terminology in biology and related domains, and term usages may be specific to a species, research area or even a particular research group. This makes communication and data sharing more difficult. The GO provides ontology of defined terms representing gene product properties. The ontology covers three domains:

1. Cellular component, the parts of a cell or its extracellular environment, where a gene product functions.
2. Molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis;
3. Biological process, operations or sets of molecular events with a defined beginning and end involved by a gene product, pertinent to the functioning of integrated living units: cell, tissues, organs, and organisms.

Each of these domains is represented by a directed acyclic graph with one source. Cellular component, molecular function, and biological process are the sources of these three graphs. Each node in the graph is a term that represents a sub

domain. When we have a gene list, we can examine the GO terms associated with the short gene lists and get some ideas about which molecular function, cellular component or biology process is enriched in the associated gene list.

In the GO enrichment analysis, we tried to identify the GO terms related to DEGs in each time point. However, the current databases do not contain the relations between GO terms and poplar genes. However in the GO database, there is an *Arabidopsis thaliana* annotation file – TAIR. This file provides the relations between GO terms and *Arabidopsis* genes, shown in Figure 11:

TAIR	locus:2193997	P40	GO:0000028	P	AT1G72370	AT1G72370 P40 AP40	protein	TAIR:locus:2193997
TAIR	locus:2058324	AT2G04390	GO:0000028	P	AT2G04390	AT2G04390 T1O3.20 T1	protein	TAIR:locus:2058324
TAIR	locus:2051229	AT2G05220	GO:0000028	P	AT2G05220	AT2G05220 F5G3.12 F5	protein	TAIR:locus:2051229
TAIR	locus:2084988	RPSAb	GO:0000028	P	AT3G04770	AT3G04770 RPSAb 40s	protein	TAIR:locus:2084988
TAIR	locus:2075735	AT3G10610	GO:0000028	P	AT3G10610	AT3G10610 F13M14.10	protein	TAIR:locus:2075735
TAIR	locus:2175428	AT5G04800	GO:0000028	P	AT5G04800	AT5G04800 MUK11.13	protein	TAIR:locus:2175428
TAIR	locus:2185485	AT5G14850	GO:0000030	F	AT5G14850	AT5G14850 T9L3.150 T9	protein	TAIR:locus:2185485
TAIR	locus:504954755	PNT1	GO:0000030	F	AT5G22130	AT5G22130 PNT1 PEA4	protein	TAIR:locus:504954755
TAIR	locus:2201786	FATB	GO:0000036	F	AT1G08510	AT1G08510 FATB fatty	protein	TAIR:locus:2201786
TAIR	locus:2206300	mtACP2	GO:0000036	F	AT1G65290	AT1G65290 mtACP2 m	protein	TAIR:locus:2206300
TAIR	locus:2042331	MTACP-1	GO:0000036	F	mitochondrial	AT2G44620 MTACP-1 M	protein	TAIR:gene:2042330
TAIR	locus:2042331	MTACP-1	GO:0000036	F	AT2G44620	AT2G44620 MTACP-1 A	protein	TAIR:locus:2042331
TAIR	locus:2090285	FaTA	GO:0000036	F	AT3G25110	AT3G25110 AtFaTA Fa	protein	TAIR:locus:2090285
TAIR	locus:2123256	AT4G13050	GO:0000036	F	AT4G13050	AT4G13050 F25G13.140	protein	TAIR:locus:2123256
TAIR	locus:2181216	ACP5	GO:0000036	F	AT5G27200	AT5G27200 ACP5 acyl	protein	TAIR:locus:2181216
TAIR	locus:2142908	AT5G35930	GO:0000036	F		AT5G35930 AT5G35930	protein	TAIR:gene:2142907
TAIR	locus:2168968	mtACP3	GO:0000036	F	mitochondrial	AT5G47630 AT5G47630	protein	TAIR:gene:4010713302
TAIR	locus:2168968	mtACP3	GO:0000036	F	mitochondrial	AT5G47630 AT5G47630	protein	TAIR:gene:2168967

Figure 11: Gene Association file – TAIR

After obtaining the relations between GO terms and *Arabidopsis* genes, we map *Arabidopsis* genes to poplar genes according to our poplar gene annotation file. Figure 12 shows the mapping relations between poplar genes and *Arabidopsis* genes.

A	B
AFFX-BioB-M_at	AT2G43360
AFFX-Ptp-gapdh-3_at	AT1G12900
AFFX-Ptp-r2-Ec-bioB-3_s_at	AT2G43360
AFFX-r2-Ec-bioB-M_at	AT2G43360
Ptp.1003.1.A1_at	AT1G67060
Ptp.1006.1.A1_at	AT1G74750
Ptp.101.1.S1_at	AT1G61667
Ptp.1011.1.S1_s_at	AT3G11780
Ptp.1042.1.A1_at	AT1G79660

Figure 12: Relations between poplar genes and Arabidopsis genes

Arabidopsis genes and poplar genes are homologous, i.e. they are functionally similar. Thus we can generate a mapping table to reveal their relationships. GO enrichment analysis evaluate GO terms of DEGs at each time point. The procedures in details are as follows:

2.4.1: Generate a list of *Arabidopsis* genes of DEGs in each time point

By consulting the Arabidopsis – poplar homolog gene chart released by Phytozome.com, in 2012, we generate a file containing a list of Arabidopsis genes corresponds to DEGs in each time point. Two or more poplar genes may have the same Arabidopsis homolog gene.

2.4.2: Call TermFinder to generate GO terms given the list of Arabidopsis genes

We use TermFinder^[13] to identify the enriched GO terms from a list of DEGs. The module can be downloaded from CPAN.

TermFinder outputs a set of GO terms based on the list of Arabidopsis genes. The result is based on the background database, gene numbers in the list, the frequency of GO terms are annotated across the provided gene list. Like the domain analysis, the P-value is also calculated by the function of the hypergeometric distribution.^[13] The method also calculates the corrected p-value for each GO by applying Bonferroni correction method. The result of TermFinder is shown in Figure 13:

-- 1 of 1419 --									
GOID	GO:0008152								
TERM	metabolic process								
CORRECTED P-VALUE	9.36E-70								
UNCORRECTED P-VALUE	6.60E-73								
FDR_RATE	0.00%								
EXPECTED_FALSE_POSITIVES	0								
NUM_ANNOTATIONS	753 of 1144 in the list, vs 11823 of 29645 in the genome								
The genes annotated to this node are:									
AT4G22890, AT2G20610, AT2G36570, AT3G12710, AT3G59890, AT5G46290, AT5G16990, AT4G20020, AT2G14260, AT4G36360, AT2G30150									

Figure 13: Result from TermFinder

2.4.3: Mapping the Arabidopsis gene list of each GO back to poplar genes, and calculating the genome size, genomic gene number, sample size, sample number

From the sample shown in Fig 2-9 we can see a list of Arabidopsis genes associated with a GO term. Now we use this list map back to poplar genes, and then calculate the average fold changes for each GO. We also need to get the genome size, genome number, sample size and sample number as domain analysis to calculate the p-value for each GO.

1. Genome size

Genome size represents the total number of poplar genes exist in DEG result file that corresponds to at least one GO term listed in the result of TermFinder in current time points.

2. Genomic gene number

Genomic gene number represents the number of poplar genes that associated with current GO term.

3. Sample size

Sample size represents the number of DEGs corresponds to at least one GO term specified in the result of TermFinder in current time point.

4. Sample number

Sample number represents the number of DEGs associated with the current GO term.

Once we get these four numbers, we can calculate the enrichment factor and enrichment score as what we did for protein domain analysis. We can also calculate the average fold change by mapped poplar genes associated with a specific GO term. Here we calculate two average fold changes of all genes associated with each GO in each time point. Average fold change 1 is calculated by mapping the Arabidopsis genes back to DEGs in current time point.. Average fold change 2 is calculated by mapping the Arabidopsis genes back to whole poplar gene set. The calculation is also the same as which specified in domain enrichment analysis.

2.4.4: Appending the annotation of Arabidopsis genes to the result

In GO analysis, we also append the annotation of Arabidopsis genes to the result. The Arabidopsis annotation file is shown in Figure 14:

A	B	
Name	Type	Short_description
AT1G01010	protein_coding	ANAC001 (Arabidopsis NAC domain containing protein 1); transcription factor
AT1G01020	protein_coding	ARV1
AT1G01030	protein_coding	NGA3 (NGATHA3); transcription factor
AT1G01040	protein_coding	DCL1 (DICER-LIKE 1); ATP-dependent helicase/ double-stranded RNA binding / protein binding / ribonuclease III
AT1G01046	mirna	MIR838a; miRNA
AT1G01050	protein_coding	AtPPa1 (Arabidopsis thaliana pyrophosphorylase 1); inorganic diphosphatase
AT1G01060	protein_coding	LHY (LATE ELONGATED HYPOCOTYL); DNA binding / transcription factor
AT1G01070	protein_coding	nodulin MtN21 family protein
AT1G01073	protein_coding	unknown protein
AT1G01080	protein_coding	33 kDa ribonucleoprotein, chloroplast, putative / RNA-binding protein cp33, putative
AT1G01090	protein_coding	PDH-E1 ALPHA (PYRUVATE DEHYDROGENASE E1 ALPHA); pyruvate dehydrogenase (acetyl-transferring)
AT1G01100	protein_coding	60S acidic ribosomal protein P1 (RPP1A)
AT1G01110	protein_coding	IQD18 (IQ-domain 18)
AT1G01115	protein_coding	unknown protein

Figure 14: Annotation file of Arabidopsis genes

2.4.5: Identification of common GO terms and Unique GO terms

A common GO term is the one enriched in all time points. A unique GO term is defined as the one enriched in only one time point. After getting the results from each time point, we do a summation and generate two additional files for common GO terms and unique GO terms.

In order to show the result of common GO terms and unique GO terms more, we also use the barplot function in R to render the average fold change. Figure 15 shows average fold changes of unique GO terms in one time point generated by vertical barplot.

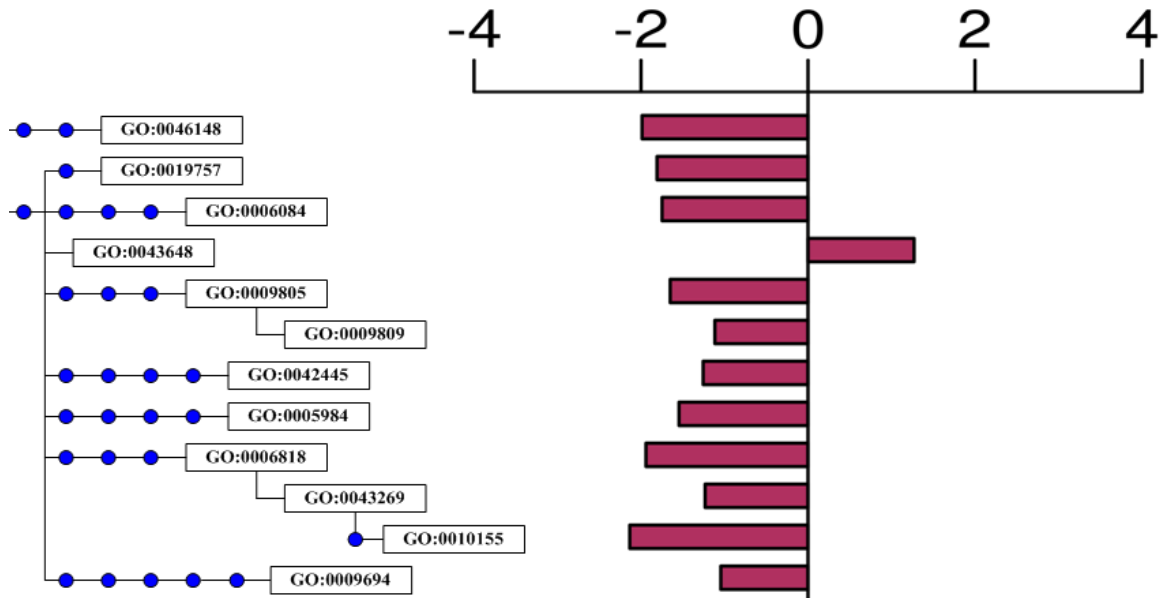


Figure 15: Unique GO terms rendered by vertical barplot

Running command:

```
$ perl ./script/GO-Term_pipeline.pl -i DEG_result.txt -g relation.txt -tp n -e TAIR.txt -n gene_ontology.obo
```

- -i <file>: Specifies the DEG output file
- -g <file>: Specifies file provides relations between poplar genes, Arabidopsis genes and GO terms
- -e <file>: Specifies the annotation file of Arabidopsis genes
- -n <file>: Specifies the GO ontology file
- -tp n: Specifies the time point number

The output of GO enrichment analysis is shown in Figure 16:

Timepoint	GOID	GenomeSize	NumGenome	SampleSize	Numsample	Sample Poplar genes	Enrichment Factor	Enrichment Score	Term_Type
1	GO:0008152	22087	1341	1276	139	Ptp.5042.3.S1_s_at Ptp	1.794205069	3.96E-12	P
1	GO:0009987	22087	22	1276	1	PtpAffx.144144.1.S1_s	0.786798233	0.365729	P
1	GO:0005975	22087	552	1276	80	PtpAffx.6111.2.S1_a_at	2.508632048	8.24E-15	P
1	GO:0044237	22087	90	1276	13	PtpAffx.135595.1.S1_at	2.500269941	0.00062891	P
1	GO:0006091	22087	146	1276	2	PtpAffx.220191.1.S1_at	0.237117276	0.9917614	P
1	GO:0006006	22087	25	1276	9	PtpAffx.308.2.A1_x_at	6.231442006	5.88E-07	P
1	GO:0055114	22087	1527	1276	176	Ptp.6442.1.S1_at Ptp.6	1.995077118	1.06E-19	P
1	GO:0006007	22087	138	1276	22	PtpAffx.135595.1.S1_at	2.759495252	4.14E-06	P
1	GO:0042221	22087	4	1276	3	Ptp.4373.1.S1_x_at Ptp	12.98217085	1.11E-05	P
1	GO:0019752	22087	19	1276	2	Ptp.3400.1.S1_s_at Ptp	1.822059066	0.09348406	P
1	GO:0006520	22087	82	1276	12	PtpAffx.156807.1.S1_at	2.533106507	0.000808222	P
1	GO:0046686	22087	733	1276	122	Ptp.4195.1.S1_at PtpAf	2.880991075	2.86E-27	P
1	GO:0000096	22087	77	1276	8	PtpAffx.83673.1.A1_at	1.798395961	0.03304748	P
1	GO:0008652	22087	118	1276	13	PtpAffx.83673.1.A1_at	1.906985548	0.00815794	P

Figure 16: GO enrichment analysis result

In the output of GO enrichment analysis, Column 1 shows the time point to which this GO term belongs to. Column 2 gives the GOID. Column 3 provides the genome size. Column 4 provides the genomic gene number. Column 5 provides the sample size. Column 6 provides the sample number. Column 7 provides the list of DEGs correspond to the certain GO in current time point, followed by enrichment factor and enrichment score. Column 10 shows the type of current GO, “P” stands for biological process, “C” stands for cellular component, and “F” stands for molecular function. We still have several columns that are not shown up, which provide the p-value, corrected p-value and two averaged fold changes.

The Common GO result file is shown in Figure 17:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
GO	Type	Term	Definition	Timepoint 1	Timepoint 2	Timepoint 3	Timepoint 4	Timepoint 5	Timepoint 6	Timepoint 7	Timepoint 8	Timepoint 9	Timepoint 10	Timepoint 11	Timepoint 12
GO:0009607	Biological Process	response to Any process		2.21898523	-2.5825452	2.55400403	-2.0249169	-1.9707369	1.178478	-2.50429	5.760706	1.828988	4.342654	1.808944	-1.34673
GO:0006810	Biological Process	transport The direct		-2.4807625	-3.1761805	-1.0436531	-2.9307502	-1.7663082	-5.62438	-2.80612	1.908426	1.594501	-1.18023	1.186858	-1.55063
GO:0006950	Biological Process	response to Any process		-1.1064286	2.65277155	2.01745641	-1.4625205	2.61855304	-2.07774	-3.12935	4.170214	1.775278	2.623059	1.310914	1.035952
GO:0055114	Biological Process	oxidation-ri A metabolic		-3.2822484	2.29997772	1.59875757	2.90702868	1.20993129	1.551408	-3.18921	2.420716	1.708979	2.611416	1.162618	1.651754
GO:0051707	Biological Process	response to Any process		-2.9325188	3.43227625	1.07758614	6.0541319	-1.1451184	-1.95449	-2.50233	5.760706	1.822592	4.370659	1.808944	-1.34673
GO:0009698	Biological Process	phenylprop The chemici		-3.6223225	1.31330276	-2.1849499	3.06616661	-1.8697484	3.820421	-3.22545	7.580631	-1.24299	5.336261	-1.37046	2.36673
GO:0048046	Cellular Components	apoplast The cell me		-4.3315426	4.84655959	1.4580909	1.65987549	-1.3546381	-1.72272	-4.33154	4.84656	1.458091	1.659875	-1.35464	-1.72272
GO:0005576	Cellular Components	extracellularThe space e		-5.4037481	1.43531534	1.83447673	2.34141026	-1.9120985	-3.01804	-4.49456	3.606107	1.462853	1.841239	-1.43769	-1.91702
GO:0005618	Cellular Components	cell wall The rigid or		-4.2932383	4.44663995	1.42204436	2.48132468	-2.0762415	-1.63939	-4.13838	4.476576	1.26394	2.290695	-2.06295	-1.68014
GO:0003824	Molecular Function	catalytic act Catalysis of		-3.5012338	5.70548398	1.64631331	2.31646287	2.13699584	-1.5176	-3.23428	2.978361	1.122578	1.57915	-1.29557	-1.22188
GO:0016491	Molecular Function	oxidoreduc Catalysis of		-3.6439985	3.3042237	1.33008578	3.55277769	2.01260152	2.016899	-3.2423	3.958537	1.648365	3.286023	1.156487	1.683157
GO:0043169	Molecular Function	cation bindi Interacting		-5.6682144	2.72442937	1.02450768	-1.091873	1.16336427	-2.508	-3.40821	4.319059	1.182232	2.431853	1.115393	1.247292
GO:0046872	Molecular Function	metal ion bi Interacting		-2.2817767	2.41174993	-1.2701693	-1.2647791	3.73785295	-2.45113	-3.15879	4.387965	1.214114	2.557244	1.09442	1.32043

Figure 17: Result file of Common GO terms

The result file of common GO terms provides the type, the description as well as the two average fold changes among all the time points of a certain common GO. Users can check whether it is up-regulated or down-regulated in each time point conveniently.

The Unique GO result file is shown in Figure 18:

GO	type	Timepoint	Term	Definition	Average_FC1	Average_FC2
GO:0042023	Biological Process	3	DNA endoreduplication	Regulated re-replication	-2.62973378	-2.62973378
GO:0019760	Biological Process	1	glucosinolate metabolism	The chemical reactions associated with the	-3.1213787	-3.65014878
GO:0010564	Biological Process	3	regulation of cell cycle	Any process that modulates the cell cycle	-2.90131872	-2.67449966
GO:0008610	Biological Process	1	lipid biosynthetic process	The chemical reactions associated with the	-3.76623028	-3.56326147
GO:0006006	Biological Process	1	glucose metabolic process	The chemical reactions associated with the	-2.48303573	-3.55240421
GO:0009913	Biological Process	4	epidermal cell differentiation	The process in which a cell differentiates into	-2.25315269	-1.76916467
GO:0006094	Biological Process	1	gluconeogenesis	The formation of glucose from non-carbohydrate	-3.72571474	-3.72571474
GO:0006091	Biological Process	1	generation of precursor metabolites and energy	The chemical reactions associated with the	2.149351458	-3.74688293
GO:0006817	Biological Process	4	phosphate ion transport	The directed movement of phosphate ions	-2.25210164	-2.25210164
GO:0006098	Biological Process	1	pentose-phosphate metabolic process	The process in which glucose is converted into	-3.07718786	-3.07099927
GO:0080167	Biological Process	1	response to karrikin	Any process that results in a response to	-2.61763418	-2.61763418

Figure 18: Result file of Unique GO terms

In the result file, column 3 specifies which time point the current GO belongs to, followed by the short description and definition of current GO. This file also gives two average fold changes of current GO.

Chapter 3

GO Hierarchy Analysis

3.1 Different types of GO relations

The ontologies of GO terms are structured as a directed acyclic graph with only one source, with GO terms represented by nodes and relations denoted by arcs as nodes in the graph. Like all the terms are categorized into three different classes, the relationships between GO terms can be classified into the following categories; part of relation; has part relation; regulates relation, negatively regulatory relation and positively regulatory relation.

1. is a relation: If we say *A is a B*, we indicate that *A is a subtype of B*. For example, fatty acid binding is a lipid binding, mucosal tolerance induction is a mucosal immune response, or lyase activity *is a* catalytic activity.
2. part of relation: The part of relation is used to represent part-whole relationships in the GO. A part of relation would only be added between A and B if B is necessarily part of A, i.e. the presence of the B implies the presence of A. For example, autophagic vacuole assembly is part of macroautophagy.
3. regulates relation: If one process directly affects the manifestation or quality of another process, then we say the former process regulates the latter process. The target of the regulation can be a pathway, an enzymatic reaction, or it may be a quality, such as cell size or PH.
4. positively regulates relation & negatively regulates relation: These two relations are two sub-relations of regulates relation.

For a given GO, its relations to other GO terms can be found in the obo file. For example:

[Term]
id: GO:0032741
name: positive regulation of interleukin-18 production
namespace: biological_process
def: "Any process that activates or increases the frequency, rate, or extent of interleukin-18 production." [GOC:mah]
synonym: "activation of interleukin-18 production" NARROW []
synonym: "positive regulation of IL-18 production" EXACT []
synonym: "stimulation of interleukin-18 production" NARROW []
synonym: "up regulation of interleukin-18 production" EXACT []
synonym: "up-regulation of interleukin-18 production" EXACT []
synonym: "upregulation of interleukin-18 production" EXACT []
is_a: GO:0001819 ! positive regulation of cytokine production
is_a: GO:0032661 ! regulation of interleukin-18 production
relationship: positively_regulates GO:0032621 ! interleukin-18 production

Figure 19: GO term example

From Figure 19, we can see that GO:0032741 “is a” GO:0001819. GO:0032741 is a GO:0032661. GO:0032741 “positively regulates” GO:0032621.

There still exist some other GO relationships besides the fundamental ones we mentioned above, e.g. located in relation; adjacent to relation; has participant relation and so on. Our GO hierarchy analysis only takes the following 5 basic GO relations into consideration:

1. is a relation
2. part of relation
3. regulates relation
4. negatively regulates relation
5. positively regulates relation

In our GO ontology analysis, we try to build a tree-structure graph to reveal the relations for a given list of GO terms. We make the source of the acyclic graph to be the root of the tree. And we define the parent-child relationship between each pair of GO terms with the basic GO relationships mentioned above. The parent-child relationship is defined as follows:

1. If GO1 “is a” GO2, we define GO1 as the parent of GO2
2. If GO1 is “part of” GO2, we define GO2 as the parent of GO1
3. If GO1 “regulates” GO2, we define GO1 as the parent of GO2
4. If GO1 “positively regulates” GO2, we define GO1 as the parent of GO2
5. If GO1 “negatively regulates” GO2, we define GO1 as the parent of GO2



Figure 20: GO relation example

In Figure 3-2, we can see a part of tree represent the GO relations specified in Figure 20.

3.2 Algorithm used to generate GO tree

We design an algorithm to generate GO trees for biological process, cellular component and molecular function separately. Here we select the biological process as an example to introduce our algorithm:

Our algorithm contains two parts. Part 1 figures out all the GO terms and their relations exist in the tree, i.e. vertices and edges exist in the tree. Part 2 figures out the layer information, i.e. the structure of the tree.

Part 1:

Step 1. Create hash tables for “is a” relation, “part of” relation, “regulates” relation, “positively regulates” relation and “negatively regulates” relation respectively. We hash all these five relationships of biological process GO terms exist in ontology file into corresponding hash tables. Keys represent the parent and values represent the children.

Step 2. Read the GO list, extract all the biological process GO terms and push them into an array - @GO_p.

Step 3. Each time we pop an element e from the head of @GO_p, check the all the basic relations of e, get a list of all the target GO terms. Scan the target GO list, and push those GO terms that did not appear before into the array @GO_p.

Step 4. Repeat step 3 until the @GO_p is empty. Then we get the integral list of GO relation pairs as well as the GO terms appear in the GO tree.

Part 2:

Step 1. Read the list of GO relation pairs, push all the GO terms with children in an array - @parents, push all the GO terms with parents in another array - @children

Step 2. Determine the root of the whole tree. We know a GO term is the root iff it exists in array @parent but does not appear in the array @children.

Step 3. Push the root of the tree into a new created array - @array. Create a hash - %hash_layer to store the layer of all GO terms. Initialize a variable - \$current_layer which used to specify the current layer number to be 1. The root is at layer 1.

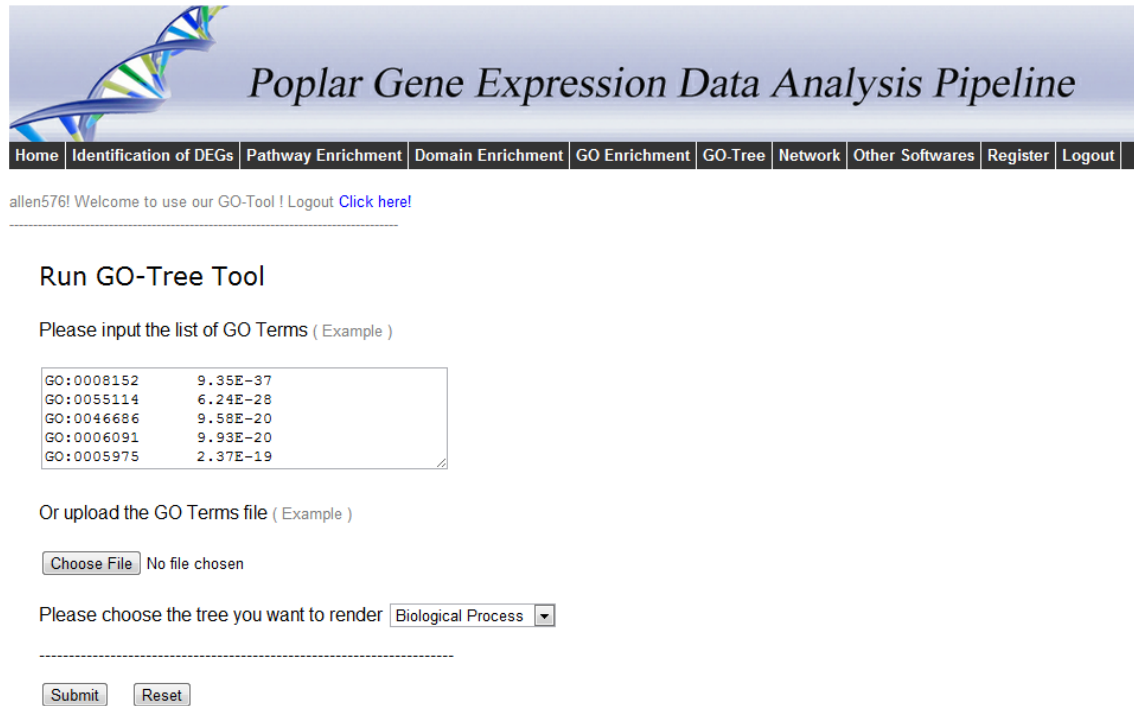
Step 4. Each time adds the value of \$current_layer by 1, pop out all the existing elements in @array, identify their children and store in array - @array2. Scan @array2 and hash all the members in @array2 to %hash_layer with value \$current_layer. Then override @array with @array2.

Step 5. Repeat step 4 until the @array is empty. Now we gain all the layer information in the %hash_layer.

3.3 Usage of online tool – GO tree

There are two ways to plot the GO tree by using our online tool.

Method 1: Paste a list of a GO term and their p-values, make sure they are separated by a tab symbol. Then select the type of GO tree (biological process, cellular component or molecular function) you want to plot at the radiobox. Click “submit” button.



The screenshot displays the 'Poplar Gene Expression Data Analysis Pipeline' web application. The header includes a navigation bar with links: Home, Identification of DEGs, Pathway Enrichment, Domain Enrichment, GO Enrichment, GO-Tree, Network, Other Softwares, Register, and Logout. Below the header, a welcome message from 'allen576l' is shown. The main section is titled 'Run GO-Tree Tool' and contains two input methods: 'Please input the list of GO Terms (Example)' and 'Or upload the GO Terms file (Example)'. The first method shows a text area with a table of GO terms and p-values. The second method shows a file upload button and a dropdown menu for selecting the tree type to render. At the bottom, there are 'Submit' and 'Reset' buttons.

GO:0008152	9.35E-37
GO:0055114	6.24E-28
GO:0046686	9.58E-20
GO:0006091	9.93E-20
GO:0005975	2.37E-19

Please choose the tree you want to render Biological Process ▼

Submit Reset

Figure 21: Paste the list of GO term along with their pvalues

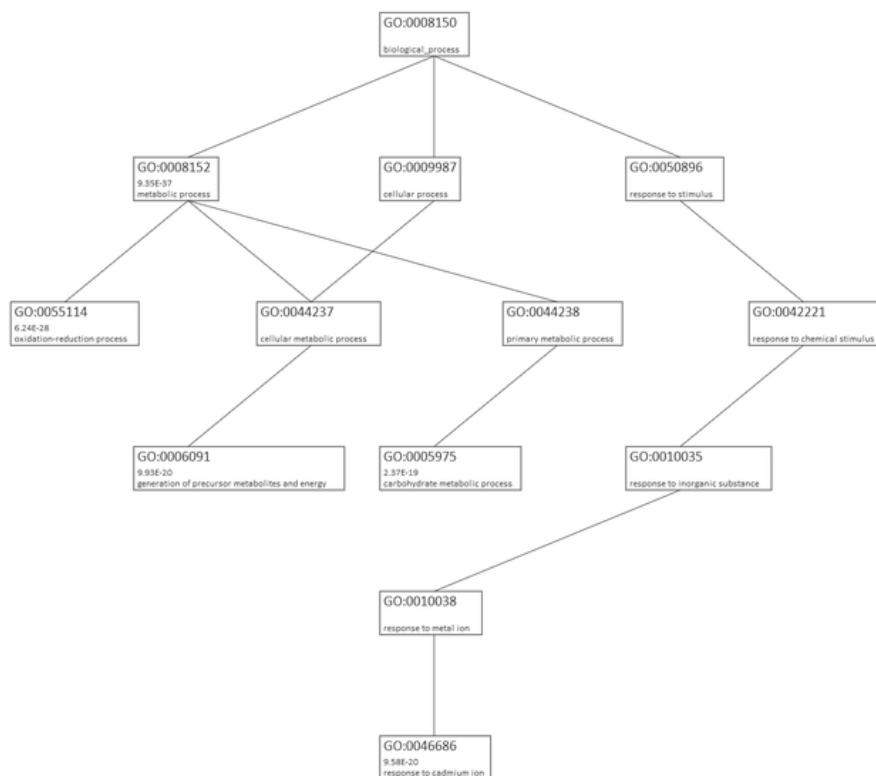


Figure 22: GO-Tree result (biological process)

Method 2: Upload a two column tab-delimited file that is resulting from you GO enrichment analysis. Column 1 contains GO terms, Column 2 provides the p values of each GO term. The sample input file is specified in Figure 23. Then select the type of GO tree (biological process, cellular component or molecular function) one wants to plot at the radio box. Click “submit” button.

	A	B
1	GO:0008152	9.35E-37
2	GO:0055114	6.24E-28
3	GO:0046686	9.58E-20
4	GO:0006091	9.93E-20
5	GO:0005975	2.37E-19

Figure 23: Upload GO-Tree file format

Chapter 4

GO Hierarchy Analysis Poplar Gene Expression Data Analysis On-line Tool

4.1 Introduction to poplar gene expression data analysis on-line tool

Poplar Gene Expression Data Analysis Pipeline is an online tool designed for analyzing gene expression data from poplar. The URL is “<http://sys.bio.mtu.edu>”. This tool is the web application of our pipeline program. It can analyze both microarray gene expressions and high-throughput gene expressions. For high-throughput gene expression, we only take the latest data version – version 3. Here is an introduction of our online tool.

4.2 User Registration and Portion

4.2.1 Register

Users must register before using web tool. The user registration page is shown in Figure 24:



New User Register

User Name *	<input type="text"/>
Password:(>=6) *	<input type="password"/>
Confirm Password *	<input type="password"/>
Country	<input type="text" value="United States"/>
Institute: *	<input type="text"/>
Email Address: *	<input type="text"/>
<input type="button" value="Submit"/> <input type="button" value="Reset"/>	

The fill marked * is required

Figure 24: User registration webpage

4.2.2 Duplicate username

We store the user registration information in our database. The user name must be unique. We will check the username in our database once a user submits a user ID. If we find this user name is already registered, the user will be directed to the webpage shown in Figure 25:

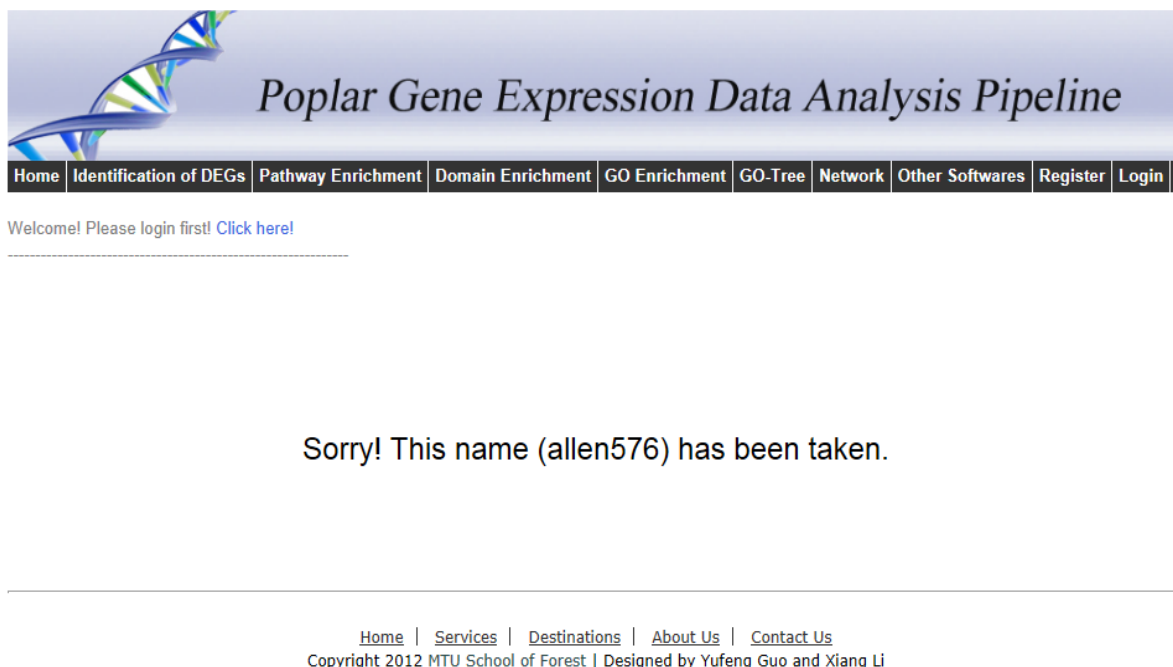


Figure 25: Duplicate username error

4.2.2 Duplicate Email

The Email address also has to be unique. After completing the analyses requested, we will send the URL links of results to users via email. Our tool can automatically search the database to check if a given email address is already registered. If a given email has already been registered, users will be directed to the webpage shown in Figure 26:



Welcome! Please login first! [Click here!](#)

Sorry! This email (xli5@mtu.edu) has been registered by other users

[Home](#) | [Services](#) | [Destinations](#) | [About Us](#) | [Contact Us](#)
Copyright 2012 [MTU School of Forest](#) | Designed by Yufeng Guo and Xiang Li

Figure 26: Duplicate email error

4.3 User Login Portion

4.3.1 Login

Each time users want to use the online tool, they have to login first. Users can either click the “Login” on the main menu or the hyperlink on up-right of each page, as shown in Figure 27:

Welcome! Please login first! [Click here!](#)

Figure 27: Login hyperlink

The login webpage is shown in Figure 28:

Please login first

Username:

Password:

[Forgot Password](#)

Figure 28: User login webpage

Make sure the username and password are correct. Otherwise an alert box will pop out, which is shown in Figure 29:

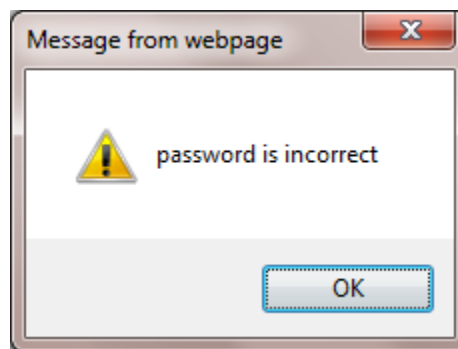


Figure 29: Alert box indicates the password is incorrect

4.3.2 Password retrieval

If users forgot their password, click the link - “Forgot Password”, then website will redirect users to a find password webpage, it is shown in Figure 30:

Find password, please input username and email!

Username:

Email:

Figure 30: Finding password webpage

Users can retrieve their password by entering the usernames and registered email addresses in this text bar. Once the users click “Submit”, our server will search

the database immediately. If a match is found, the server extracts the password and sends it to the registered email address.

4.4 Data analysis Pipeline

After the user logs in, he/she can do the data analysis by using the online tool. All the analyses are specified in the main menu bar, which is shown in Figure 31:



Figure 31: Menu Bar of pipeline Analysis

In Fig 4-8, we can see Identification of DEGs – DEG Analysis, Pathway Enrichment Analysis, Domain Enrichment Analysis, GO Enrichment Analysis and GO-Tree. These analyses all specified in pervious chapters. Here we only introduce the usage of online tool.

4.4.1 Identification of DEGs

In perform DEG analysis, users need to click “Identification of DEGs” in the main menu bar, the webpage is shown in Figure 32:

← → ↻ sys.bio.mtu.edu/deg.php 🔍 ☆ 🌐

Identification of Differentially Expressed Genes(DEGs)

1. Select data platform type:

Microarray ▼

2. Upload the gene expression data file: (Example: Microarray data) (Example: High-throughput sequencing data)

Choose File No file chosen

Note that: (1) Controls column names must begin with C or c, whereas treatment column names begin with T or t; (2) For each time point/comparison, control data are listed at left, and treatment data at right.

3. Enter the NUMBER of all time points or comparisons (e.g. treatments vs control):

2

4. Microarray Data (only): Percentage for trimming (lowly expressed genes) (e.g. 0.1 - 10%; 0 - do not trim):

0.2

5. High Throughput Sequencing Data (only) : Enter the minimum values allowed:(Set values smaller than this to this value. Default minimum value is 1):

1

5. Enter cut-off corrected p-value(FDR) for selecting DEGs (e.g. 0.05):

0.05

6. Do you want to run the following analyses upon running DEG analysis? (you are encouraged to run them at one time. However, you can always examine the DEG output first and then come back to load the DEG output and run the following analyses at a later time)

☐ Metabolic Pathway Enrichment Analysis (Gene Set Enrichment Analysis)

☐ Protein Domain Enrichment Analysis

☐ Gene Ontology (GO) Enrichment Analysis

Next

Figure 32: DEG analysis webpage

In this webpage, users can specify the following parameters:

- 1) Data type of the uploaded data, either “microarray” or “high-throughput”.
- 2) Gene expression data file, click “Choose File” button and upload the gene expression data file from local PC. Make sure the format of uploaded file is correct. You can click the example link to check the sample file.
- 3) Time point number
- 4) For microarray data, specify the trimming percentage of the gene expression data.
- 5) For high throughput sequencing data, specify the minimum normalized value.
- 6) Cut-off threshold of corrected p-value. We will classify the popular genes with corrected p-value smaller than this threshold as the DEGs
- 7) Analyses one wants to do after the DEG analysis in the checkboxes. There are three checkboxes - DEG sets Analysis (Pathway Analysis), Domain Analysis and GO Term Analysis. These checkboxes provide huge convenience in gene expression analysis. Users can choose all of them and then the online tool will do all the analyses once in the background. Once the analyses are completed, server will send all the results to the users using registered email addresses.

Click “next” and go to the confirmation page, as shown in Figure 33:

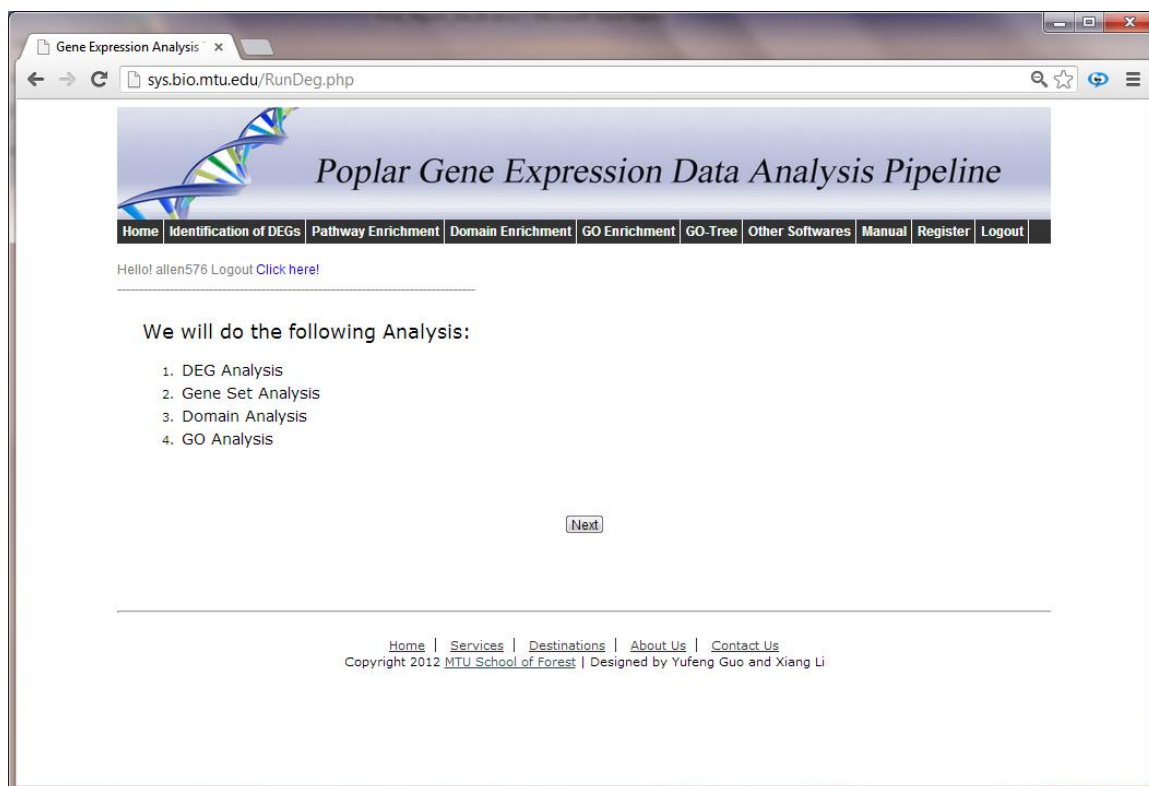


Figure 33: DEG pipeline Confirmation (Analyses)

Server will show all the analyses we will do in this page, if nothing is wrong, click “next” button. Then the users are redirected to web page that used to confirm the parameters, as shown in Figure 34:

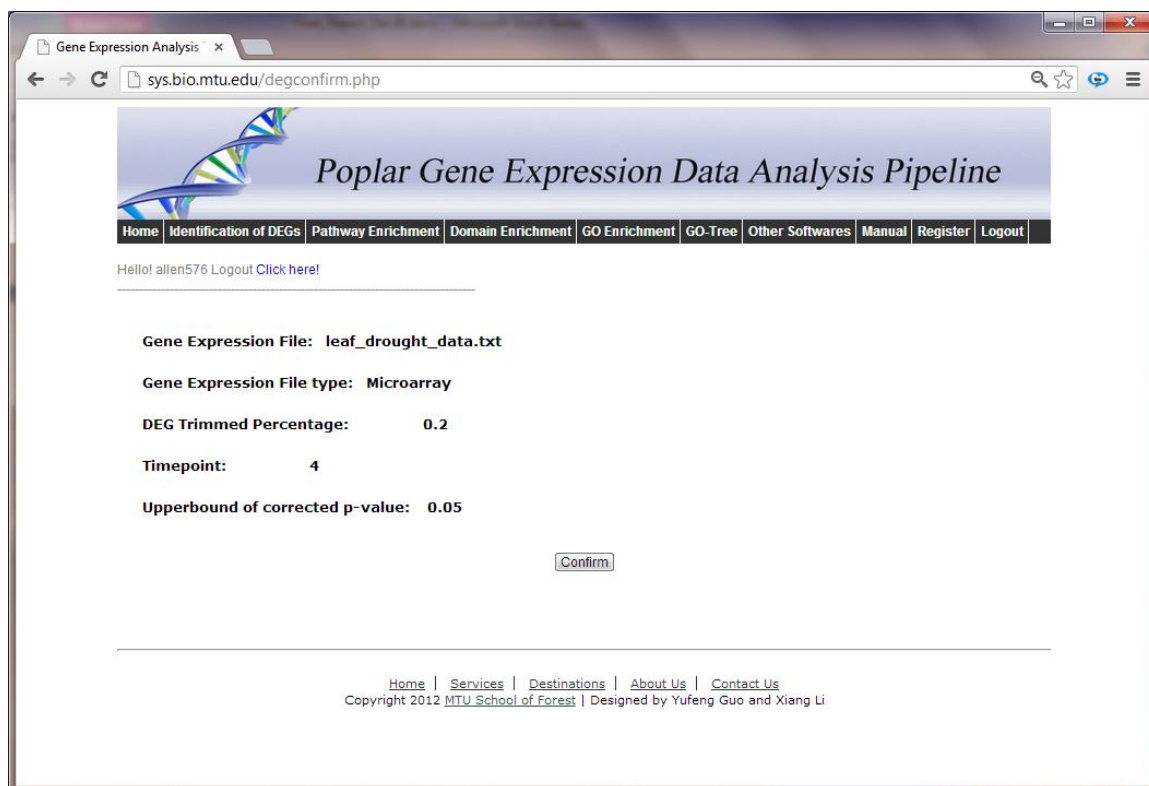


Figure 34: DEG pipeline Confirmation (parameters)

In the web page shown in Figure 34, one will find all the parameters he entered before. If all the parameters are correct, click “confirm” button. Then he will be redirected to the final webpage, as shown in Figure 35:

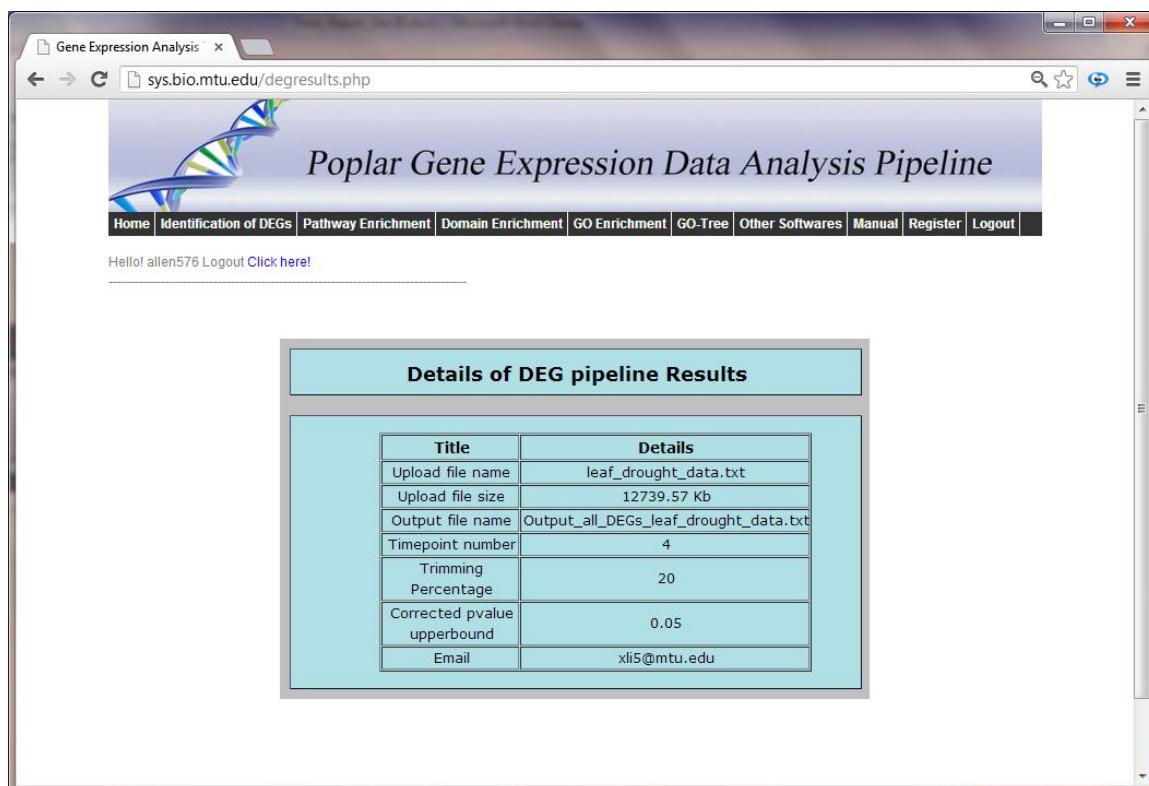


Figure 35: Final page of DEG pipeline

In Figure 34, users can get a form that specifies how the server will run the program. We can see the form contains the upload file name, upload file size, output file name, time point number, trimming percentage, upper bound value of corrected p-value which used to classify the DEGs, and the email address which the result will be sent to. Once users see this webpage, it indicates the tool has already got started to run the analysis. Users can logout and check their email after several hours.

Although user can do each pipeline analysis separately, we highly recommend users use this pipeline to do all the analyses at once. This provides some advantages in using the resource. Our online tool also checks the input files. Users will be directed to error warning webpage if they make mistakes of the file format. For example, Figure 36 shows the error warning webpage when user enters wrong time point number:

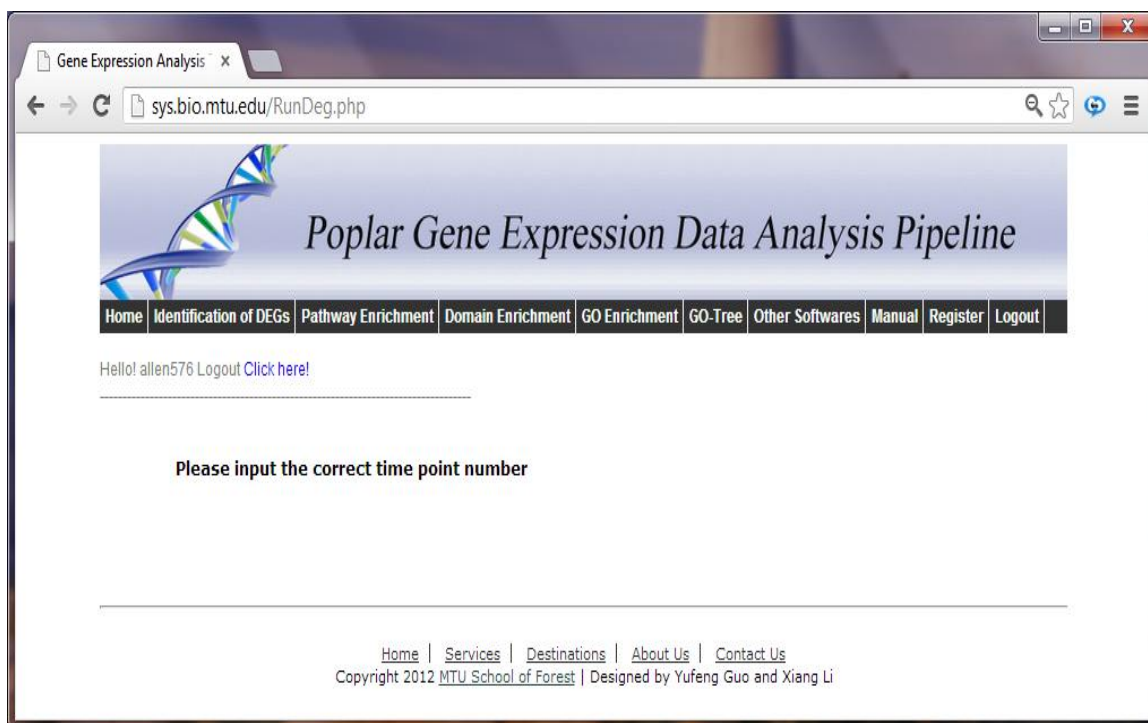


Figure 36: Wrong time point number warning

Our server uses this email– sysbiomtu@gmail.com. The output files will be sent to user’s email address. A few example messages from our server are shown in Figure 37:

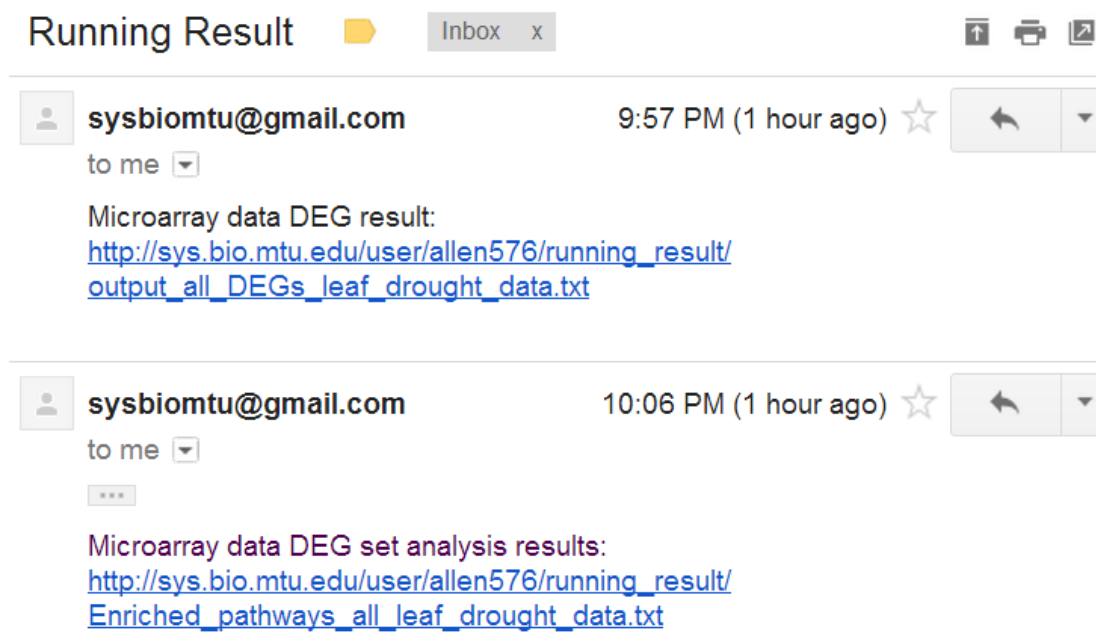
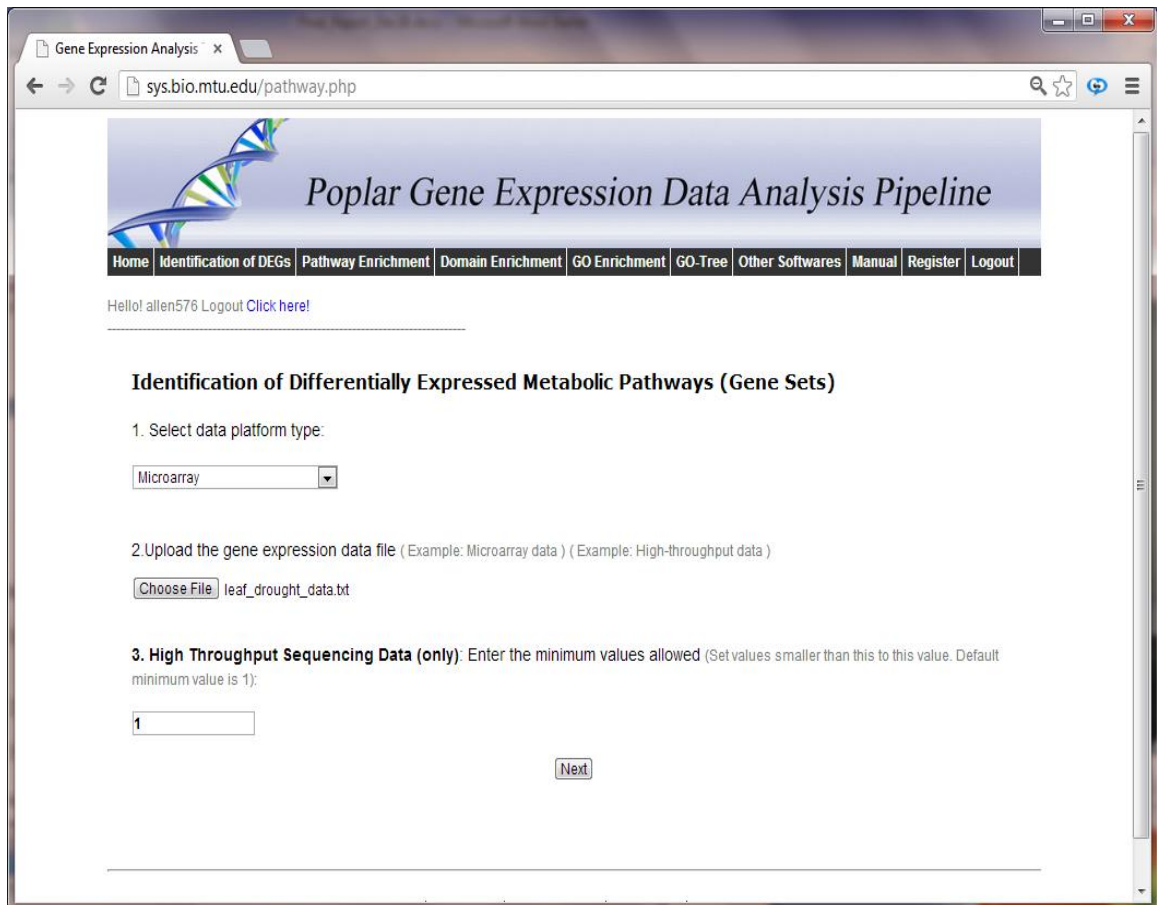


Figure 37: Results get from server

Users can download the results by clicking the links in the email.

4.4.2 Pathway Enrichment Analysis

If one wants to do the pathway enrichment analysis, he needs to click the “Pathway Enrichment” in the main menu bar.



The screenshot shows a web browser window with the address bar displaying 'sys.bio.mtu.edu/pathway.php'. The page title is 'Gene Expression Analysis'. The main heading is 'Poplar Gene Expression Data Analysis Pipeline'. Below the heading is a navigation menu with links: Home, Identification of DEGs, Pathway Enrichment, Domain Enrichment, GO Enrichment, GO-Tree, Other Softwares, Manual, Register, and Logout. A user greeting 'Hello! allen576 Logout Click here!' is visible. The main content area is titled 'Identification of Differentially Expressed Metabolic Pathways (Gene Sets)'. It contains three steps: 1. Select data platform type: A dropdown menu is set to 'Microarray'. 2. Upload the gene expression data file: A 'Choose File' button is followed by the filename 'leaf_drought_data.bt'. 3. High Throughput Sequencing Data (only): A text box contains the value '1'. A 'Next' button is located at the bottom right of the form.

Figure 38: Pathway Enrichment Analysis

In this webpage, users can:

- 1) Choose the data type of the uploaded data, either “microarray” or “high-throughput”.
- 2) Click “Browse” button and upload the gene expression data file from local PC.
- 3) For high throughput sequencing data, enter the minimum normalized value.

After entering all the parameters, click “next”. We will see the confirmation webpage shown in Figure 39:

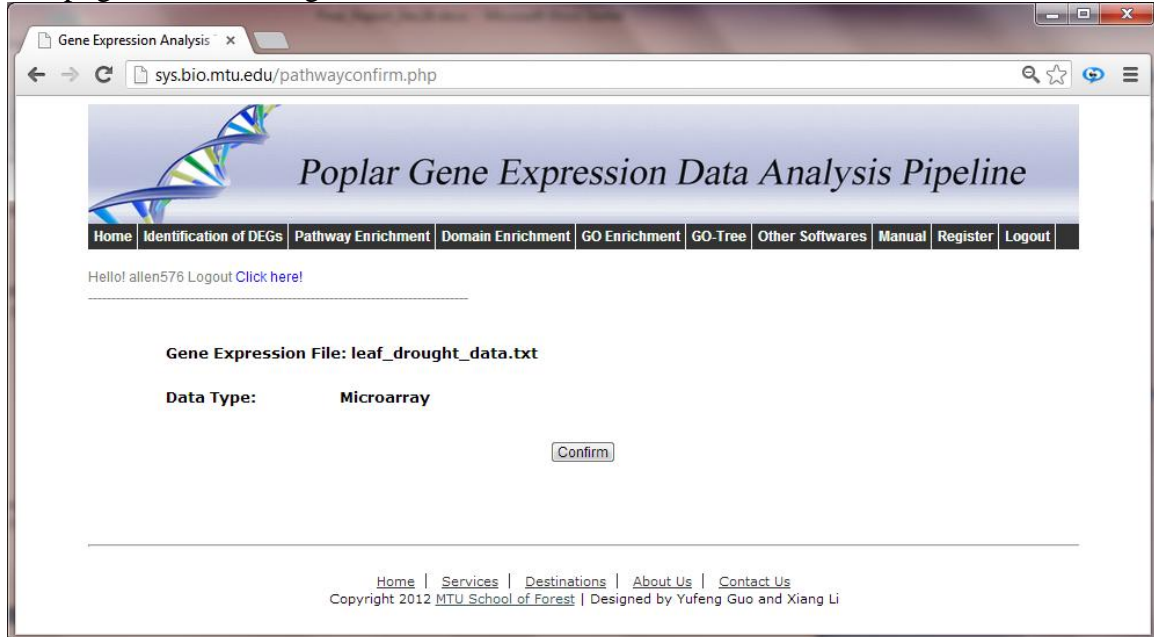


Figure 39: Confirmation page of pathway enrichment analysis

After confirming the parameters entered, click “confirm” button and make the server run the program. The final webpage of pathway enrichment analysis is shown in Figure 40.

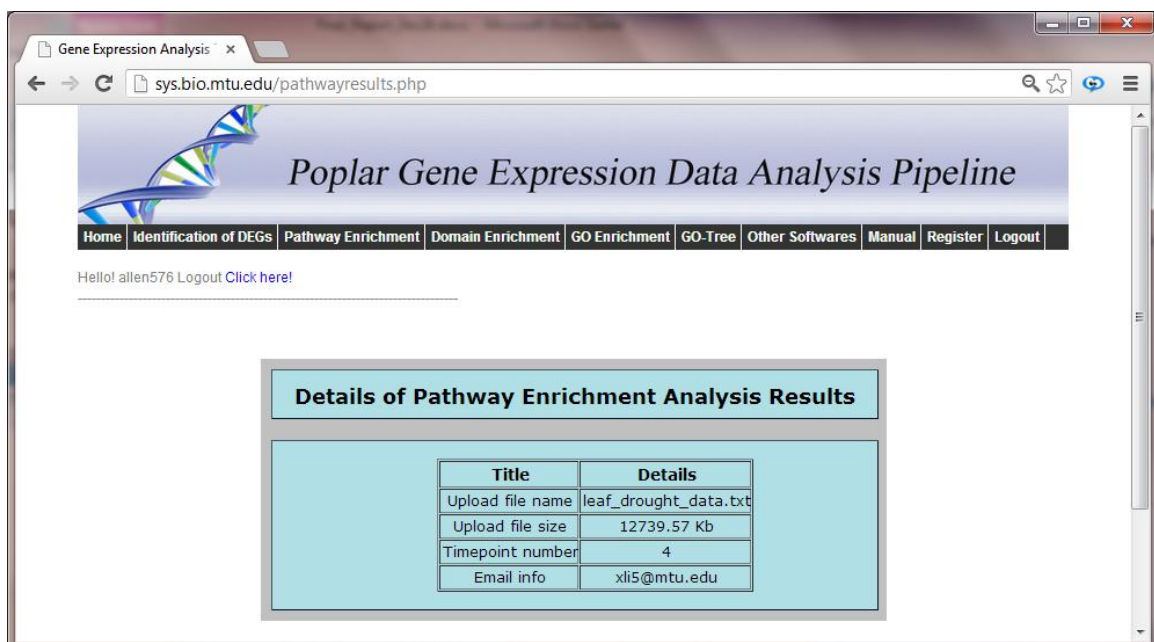


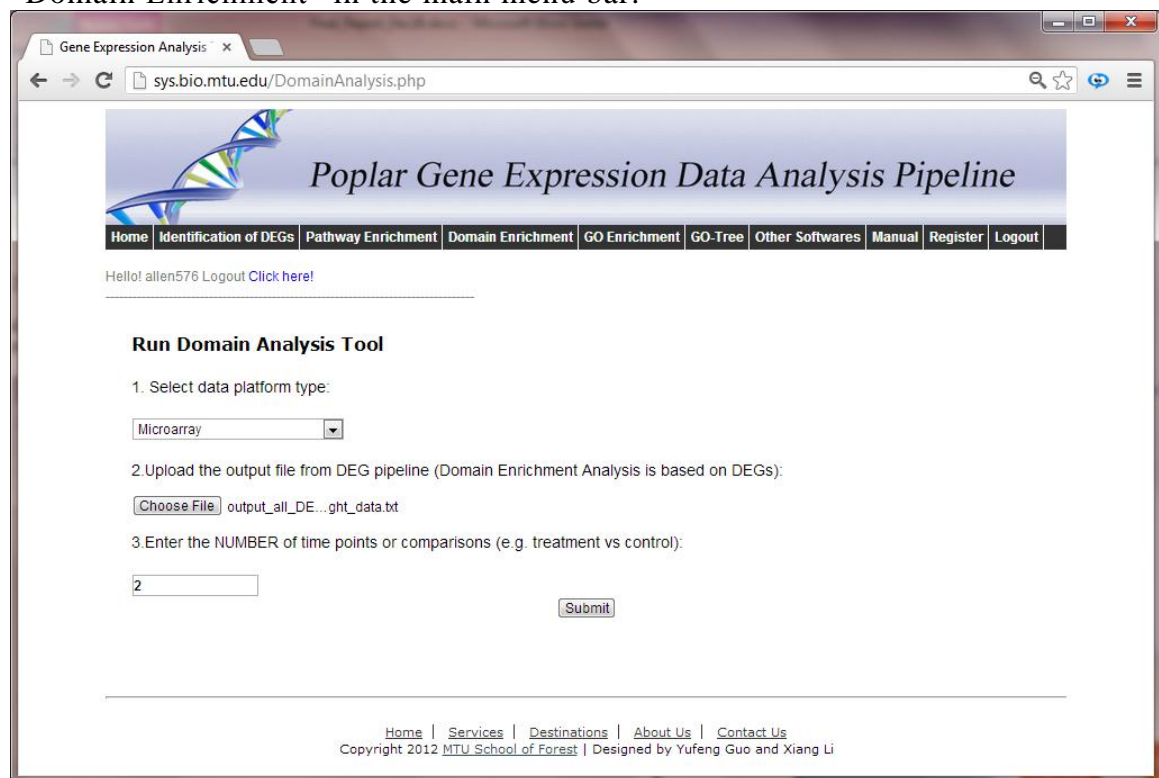
Figure 40: Final result of pathway enrichment

Webpage in Figure 40 shows the details of upload file, time point number and the email address that the server will send the outputs to.

For the pathway enrichment analysis, we currently do not allow user to upload their pathway matrix file. Our server has its own pathway matrix file for poplar. It is composed of 340 pathways.

4.4.3 Domain Enrichment Analysis

If one wants to identify enriched domains in a list of DEGs, first clicks the “Domain Enrichment” in the main menu bar.



The screenshot shows a web browser window with the address bar displaying 'sys.bio.mtu.edu/DomainAnalysis.php'. The page title is 'Gene Expression Analysis'. The main header features a DNA double helix graphic and the text 'Poplar Gene Expression Data Analysis Pipeline'. Below the header is a navigation menu with links: Home, Identification of DEGs, Pathway Enrichment, Domain Enrichment (highlighted), GO Enrichment, GO-Tree, Other Softwares, Manual, Register, and Logout. A user greeting 'Hello! allen576 Logout Click here!' is visible. The main content area is titled 'Run Domain Analysis Tool' and contains three steps: 1. Select data platform type: A dropdown menu is set to 'Microarray'. 2. Upload the output file from DEG pipeline (Domain Enrichment Analysis is based on DEGs): A 'Choose File' button is followed by the filename 'output_all_DE...ght_data.txt'. 3. Enter the NUMBER of time points or comparisons (e.g. treatment vs control): A text input field contains the number '2'. A 'Submit' button is located below the input field. At the bottom of the page, there is a footer with links: Home, Services, Destinations, About Us, Contact Us, and copyright information: Copyright 2012 MTU School of Forest | Designed by Yufeng Guo and Xiang Li.

Figure 41: Domain Enrichment Analysis

From Figure 41 we can see users need to specify the following parameters to run the domain enrichment analysis:

- 1) Choose the data type of the uploaded data, either “microarray” or “high-throughput”.
- 2) Click “Choose Files” button and upload the output file of DEG pipeline.

Similar with pathway enrichment analysis, users will be directed to the confirmation page that contains the upload file name and time point number.

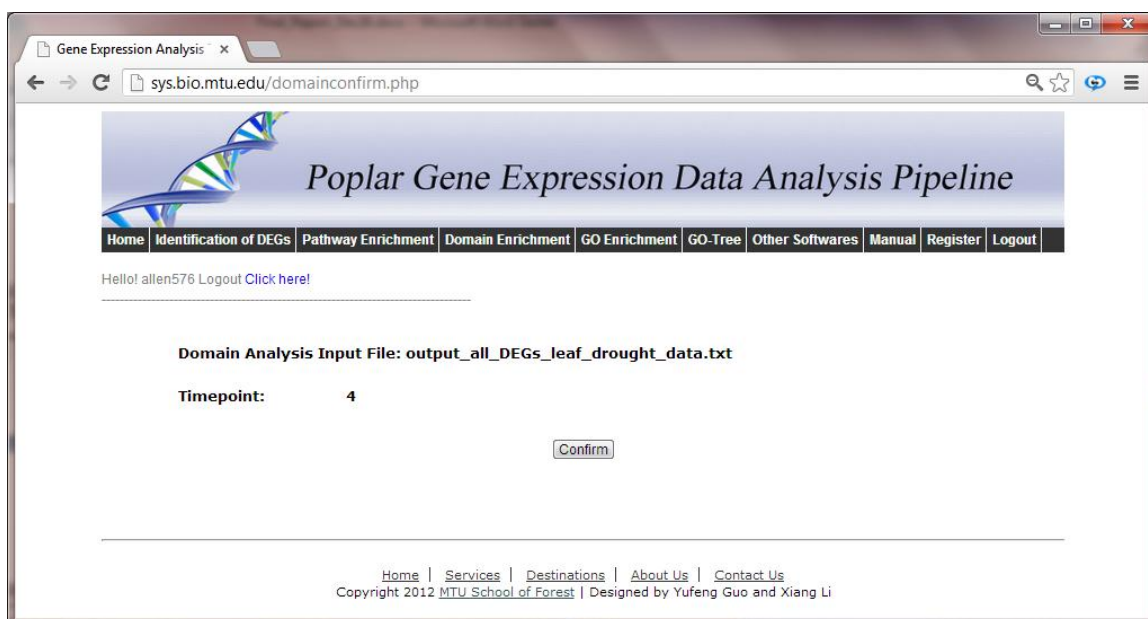


Figure 42: Confirmation page of domain enrichment analysis

After confirming the input file and time point number, click “confirm” button to run the program. The final webpage of pathway enrichment analysis is shown in Figure 43.

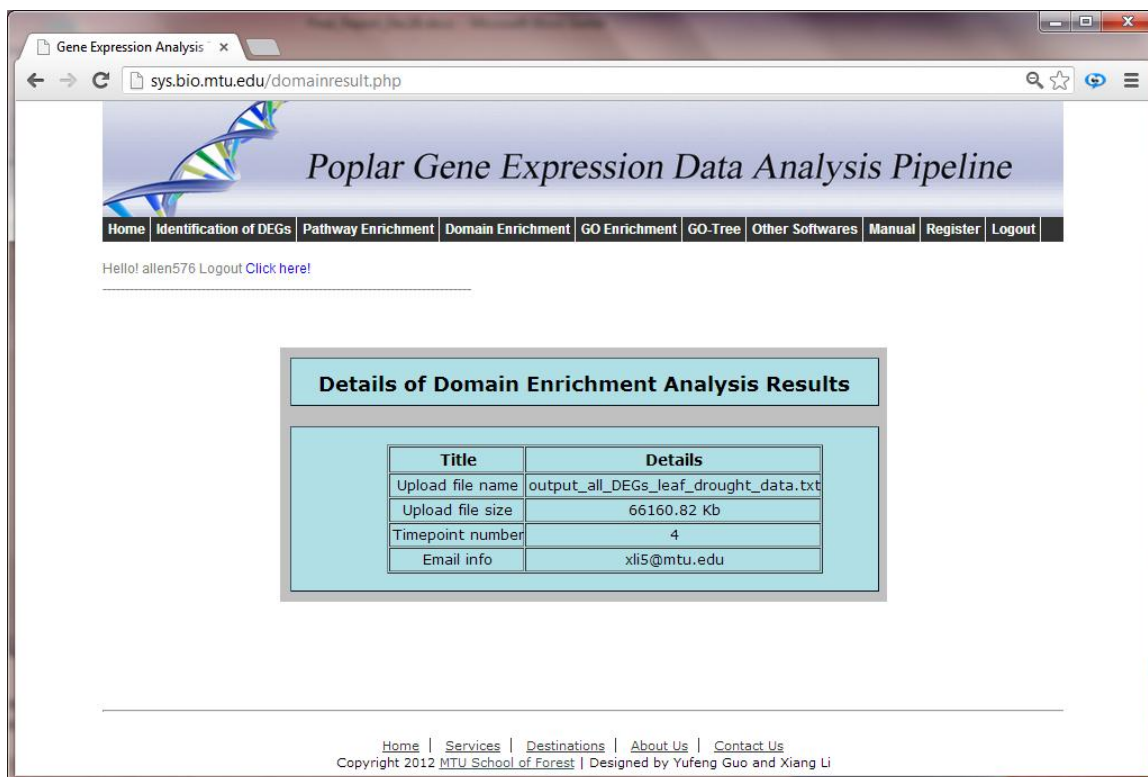
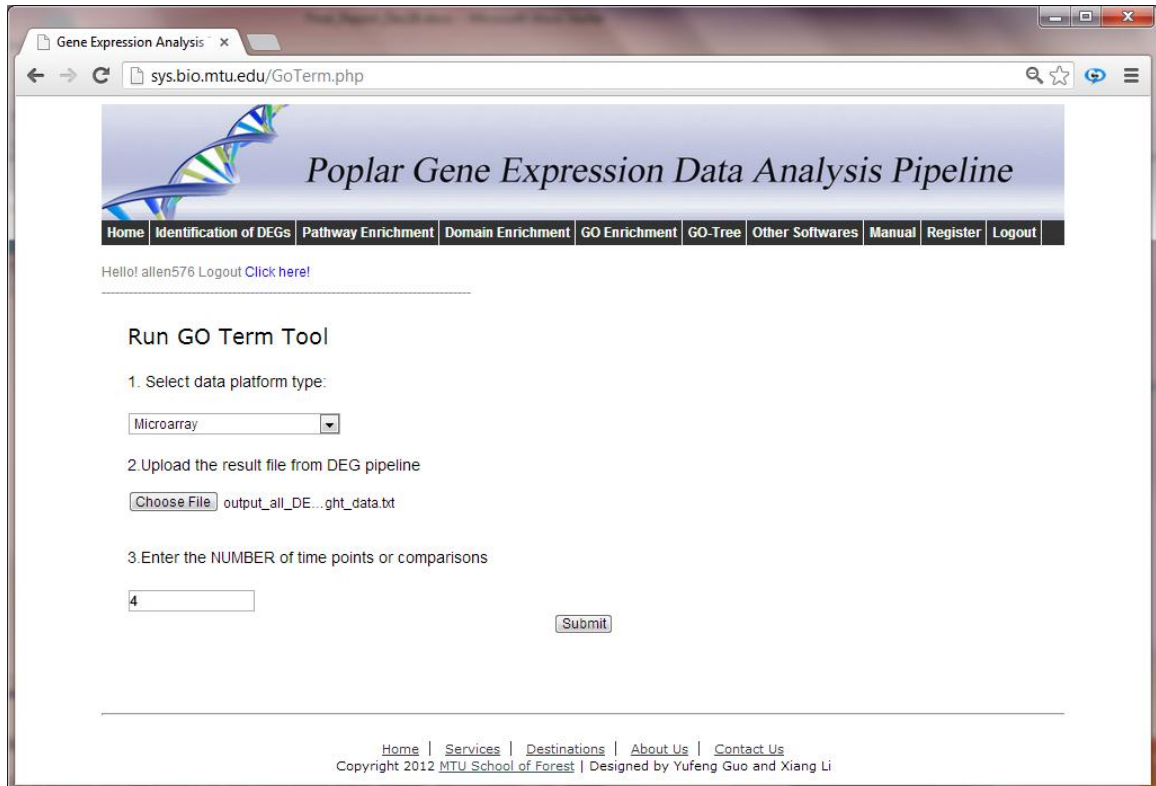


Figure 43: Final result of domain enrichment

For domain enrichment analysis, our tool already has the protein domain annotation file in place. Users only need to upload the protein domain annotation file. We also keep track of the latest domain annotation file and renew the annotation file to produce the results up to date.

4.4.4 GO Enrichment Analysis

If one wants to do the GO enrichment analysis, first click the “GO Enrichment” in the main menu bar.



The screenshot shows a web browser window with the address bar displaying 'sys.bio.mtu.edu/GoTerm.php'. The page title is 'Gene Expression Analysis'. The main heading is 'Poplar Gene Expression Data Analysis Pipeline'. Below the heading is a navigation bar with links: Home, Identification of DEGs, Pathway Enrichment, Domain Enrichment, GO Enrichment (selected), GO-Tree, Other Softwares, Manual, Register, and Logout. A user greeting 'Hello! allen576 Logout Click here!' is visible. The main content area is titled 'Run GO Term Tool' and contains three steps: 1. Select data platform type: A dropdown menu is set to 'Microarray'. 2. Upload the result file from DEG pipeline: A 'Choose File' button is next to the filename 'output_all_DE...ght_data.txt'. 3. Enter the NUMBER of time points or comparisons: A text box contains the number '4'. A 'Submit' button is located below the text box. At the bottom of the page, there is a footer with links: Home, Services, Destinations, About Us, Contact Us, and copyright information: Copyright 2012 MTU School of Forest | Designed by Yufeng Guo and Xiang Li.

Figure 44: GO Enrichment Analysis

In this webpage, users have to:

- 1) Choose the data type of the uploaded data, either “microarray” or “high-throughput”.
- 2) Click “Choose Files” button and upload the output file from DEG pipeline.
- 3) Enter the number of time points in the text box.

Similar with pathway enrichment analysis and domain enrichment analysis, users are directed to the confirmation page that contains the upload file name and time point number.

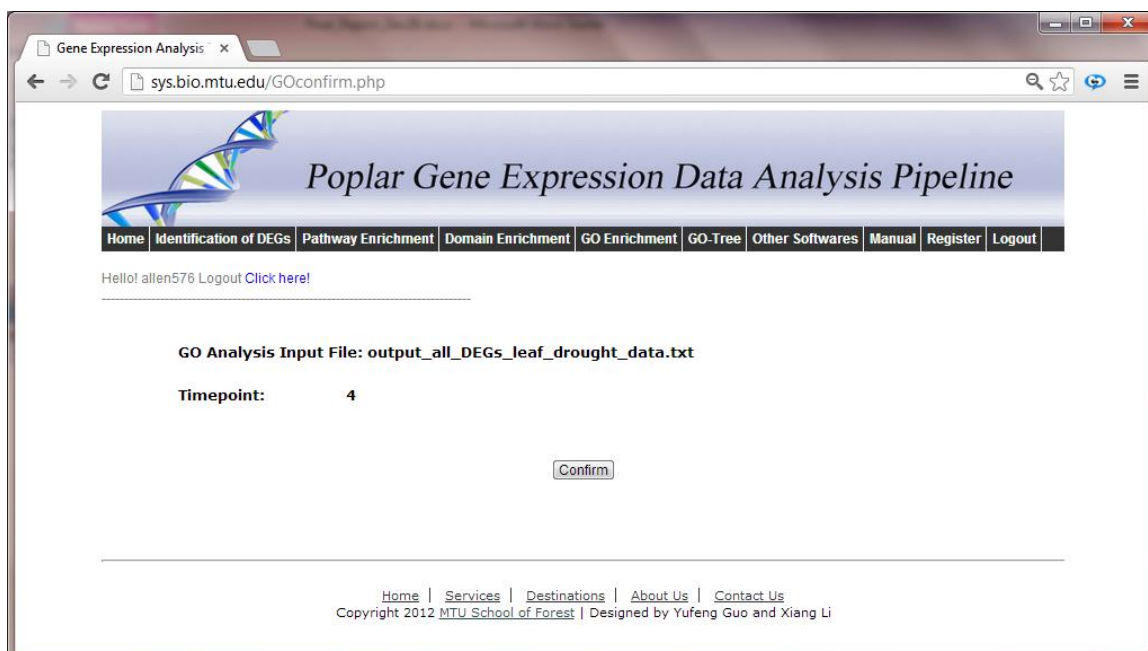


Figure 45: Confirmation page of GO enrichment analysis

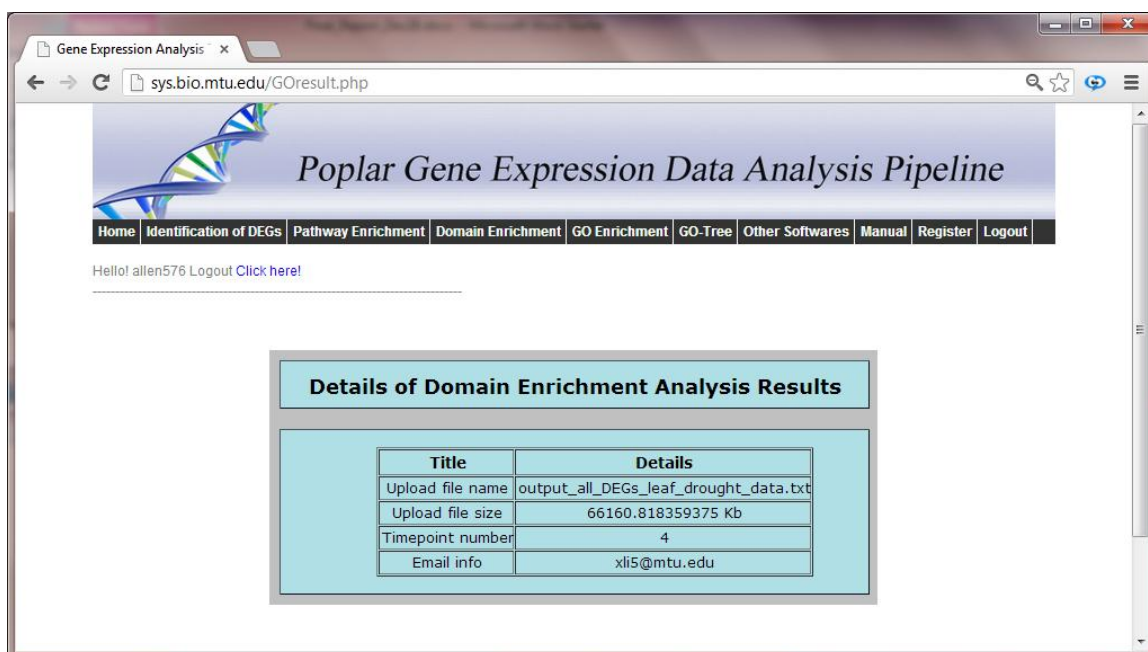


Figure 46: Final result of GO enrichment

For the GO enrichment analysis, users only need to upload the output file resulting from DEG pipeline. Our tool already has the GO ontology file and annotation files, Arabidopsis genes' annotation file and GO annotation file. We

keep track of the latest lease of annotation files and GO annotation file and update of files once the new annotation become available.

Conclusion and Future Work

The purpose of this project is to develop a set of tools for rapid analyzing gene expression data for poplar genes. The pipelines can be used to identify of DEGs, enriched pathways, enriched domains and enriched GO terms. We also developed an on-line tool to enable users do the analyses of their data easily. Such tool can be used to analyze both microarray data and high throughput data.

Some potential contents that can be integrated to improve our pipeline tool are as follows:

1. Integrate multiple methods to identify the DEGs, add an option to the methods so that users can choose proper methods for their data.
2. Add more automated visualization schemes, for example heat-map or bar-plot to automatically show the average fold change or other information we are interested in.
3. Add more species, currently our tool can only analyze poplar gene expression data.
4. Add methods to build gene network and automatically identify regulatory modules
5. Add time execution time bar to let users know how much time it needs to get the running results.

Reference:

1. Wolfgang Huber. "Analysis of microarray gene expression data "
2. Paul J.Hurd and Christopher J.Nelson " Advantages of next-generation sequencing versus the microarray in epigenetic research"
3. Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P.(2004) "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments."
4. Hall N (May 2007). "Advanced sequencing technologies and their wider impact in microbiology". *J. Exp. Biol.* **210** (Pt 9): 1518-25.
5. Church GM (January 2006). "Genomes for all". *Sci. Am.* **294** (1): 46–54.
6. Kadota, K., Y. Nakai, and K. Shimizu, *A weighted average difference method for detecting differentially expressed genes from microarray data*. Algorithms Mol Biol, 2008.3: p. 8.
7. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments*. Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
8. Benjamini, Yoav; Hochberg, Yosef (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society, Series B (Methodological)* **57** (1): 289–300.
9. Genetics, S., *Multiple Testing Corrections*. 2003.
10. Irina Dinu, John D Potter, Thomas Mueller, Qi Liu, Adeniyi J Adewale, Gian S Jhangri, Gunilla Einecke, Konrad S Famulski, Philip Halloran and Yutaka Yasui (2007) "Improving gene set analysis of microarray data by SAM-GS"
11. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, et al.: "PGC-specific genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes" *Nat Genet* 2003,34:267-273
12. The Gene Ontology Consortium (January 2008). "The Gene Ontology project in 2008". *Nucleic Acids Res.* **36** (Database issue)
13. Elizebeth I.Boyle, Shuai Weng, Jere Gollub, Heng Jin, David Botstein, J.Michael Cherry, Gavin Sherlock, "GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes" *Bioinformatics*. 2004 Dec 12;20(18):3710-5. Epub 2004 Aug 5.