Analysis of gene expression data in transgenic and non-transgenic soybean cultivars using bioinformatics tools

Kei Chin Cheng

Department of Plant Science
McGill University
Montreal, Quebec, Canada

August, 2007

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

Abstract

Current safety assessment for novel crops, including transgenic crops, uses a targeted approach, which determines crop safeness by assessing the content of a few specific chemical components. However, microarray technology can simultaneously assess the whole transcriptome and can therefore be used to analyze target genes as well as unintended effects. In this study, we used this technique as a non-targeted approach. Gene expression data from a microarray experiment with five soybean cultivars was analyzed using bioinformatics. Two cultivars were transgenic (RoundUp®) and three were non-transgenic. We show that the variation in gene expression between transgenic and non-transgenic soybean is less than that between non-transgenic cultivars. A MySQL database coupled with CGI web interfaces was developed to store and present the results (http://thor.agrenv.mcgill.ca/cgi-bin/soy/soybean.cgi). By integrating the microarray data with gene annotations and other soybean data, a comprehensive view of differences in gene expression can be explored between cultivars.

Résumé

Les méthodes actuelles d'évaluation du risque pour des cultures nouvelles, incluant les cultures transgéniques, utilisent une approche ciblée; elles évaluent le contenu en composés chimiques spécifiques. La technologie des micropuces étant maintenant disponible, il est possible d'évaluer la totalité du transcriptome. Nous avons utilisé cette technologie comme approche non-ciblée. Dans la présente étude, les données d'expériences de micropuces comparant l'expression des gènes de cinq cultivars de soja sont analysées par des méthodes bioinformatiques. Deux de ces cultivars sont des soja transgéniques RoundUp® et trois sont non-transgéniques. Nous montrons que la variation de l'expression des gènes entre soja transgéniques et non-transgéniques est moins grande qu'entre des cultivars non-transgéniques. Une base de données MySQL et une interface web CGI ont été développées pour entreposer et récupérer les données. L'intégration avec d'autres données sur le soja a rendu possible l'exploration de données génétiques globales entre cultivars en terme de fonctions biologiques.

Dedication

This thesis is dedicated to my parents who offered me unconditional love,
support and understanding throughout all these years of education.
Thank you for everything.

Acknowledgements

I would like to thank my supervisors Dr. Stromvik and Dr. Marc Fortin for their guidance and inspiration, especially Dr. Stromvik for her incredible amount of patience and extra working hours to help me finish this thesis. Thank you Dr. Fortin for initiating and establishing the microarray experiment. Also, I would like to thank my advisory committee members, Dr. Philippe Seguin and Dr. Kevin Wade, for the discussions and valuable advice. I must also thank Julie Beaulieu for generating the gene expression data. In addition, I would like to thank my laboratory mates Francois Fauteux for solving problems in analyzing data; Frederic Latour and Andre Lessard for the technical support; Annie Archambault for the emotional support and translation of the abstract into French; and Hanaa Saeed for her friendship and bringing me good memories of the conference trip to Saskatoon, field visit to Quebec and the enjoyable coffee breaks.

This study was funded through the support of AFMnet, FQRNT, NSERC, and Centre Sève. I would also like to thank the Scholarships Committee of the Faculty of Agricultural and Environmental Sciences for the award of the Frederick Dimmock Memorial Fellowship, Centre Sève and Agri-Food for the award of two travel grants to present my posters at the 3rd and 4th Canadian Plant Genomics Workshop. Thank you very much.

Table of contents

List of tables

List of figures

List of abbreviations

Abbreviation	Term
BLAST	Basic Local Alignment Search Tool
CBRI	Center for Biomedical Research Informatics
CDF	Chip Description File
cDNA	Complementary Deoxyribonucleic Acid
CEL	Cell intensity file
cRNA	Complementary Ribonucleic Acid
DAG	Directed Acyclic Graph
dChip	DNA-Chip analyzer
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
EC	Enzyme Commission
EMBL	European Molecular Biology Laboratory
EPSPS	5-Enolpyruvylshikimate-3-Phosphate Synthase
EST	Expressed Sequence Tag
ESTIMA	Expressed Sequence Tag Information Management and Annotation
GB	Gigabyte
GO	Gene Ontology
GRAS	Generally Recognized As Safe
ID	Identifier
int	Integer
KEGG	Kyoto Encyclopedia of Genes and Genomes
LIMMA	Linear Models for Microarray Data
LIS	Legume Information System
MAS	Microarray Suite software
MeSH	Medical Subject Headings
mRNA	Messenger Ribonucleic Acid
NCBI	National Center for Biotechnology Information

OS	Operating System
PC	Personal Computer
PCA	Principal Component Analysis
RAM	Random Access Memory
RMA	Robust Multichip Average
RNA	Ribonucleic Acid
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SCN	Soybean Cyst Nematode
SGMD	Soybean Genomics and Microarray Database
SIB	Swiss Institute for Bioinformatics
SQL	Structured Query Language
TAIR	The Arabidopsis Information Resource
TC	Tentative Concensus
T-DNA	Transferred Deoxyribonucleic Acid
TIGR	The Institute of Genomic Research
varchar	Variable Character
XML	Extensible Markup Language

1 Introduction

Soybean (*Glycine max* (L.) Merr.) is one of the most economically important crops in North America and worldwide, providing abundant proteins and vegetable oil for human and livestock consumption. Being a member of the *Fabaceae* family, soybean is also one of the major contributors to the global nitrogen cycle. In order to improve performance in different climates, many new varieties of soybean are developed every year using traditional breeding and/or genetic engineering. These plants with novel traits have for instance improved seed quality, cold tolerance, and disease, pest and herbicide resistance (Dunwell, 2005). Before commercialization, all crops with novel traits have to undergo safety assessment in order to assure that it is safe for human and animal consumption. In this study, the overall gene expression profiles of five soybean cultivars are compared using microarray technology. Two of the cultivars are transgenic for resistance to the herbicide glyphosate (RoundUp®) and three cultivars are conventional counterparts. The aim of this project is to develop a computational environment where soybean gene expression, whether from transgenic or conventional cultivars, can be objectively compared. To our knowledge, there is no published report to compare 'substantial equivalence' in transgenic soybean at the gene expression level and this is the first database specially designed to assist researchers in crop safety assessment. Our objectives are to develop a database with web tools to store and retrieve the results of the microarray experiment, and link to other soybean genomic information in order to assist researchers to identify changes in gene expression and determine whether these changes alter biological processes in soybean.

1.1 Safety assessment of transgenic crops

Concerns have been raised for crops with novel traits developed by transgenic techniques that the incorporation of foreign genes into organisms may produce unintended toxins or allergens, or nutrients and essential gene products may be down-regulated. For example, one report found that genetically modified potato with the insertion of a modified soybean glycinin gene unexpectedly over-expressed a toxin glycoalkaloid, although this did not negatively affect rats fed the transgenic potato (Hashimoto *et al.*, 1999). This showed that unintended effect can occur in transgenic crops. However, it is also important to point out that, because of genetic recombination, unintended effect can also occur in novel crops developed by traditional plant breeding. Therefore all crops with novel traits must undergo crop safety assessment.

Current safety assessment of new crops is based on the concept of 'substantial equivalence'. If the chemical component in the new crop is substantially similar to their counterpart that means it possesses no health risk and the new trait is safe for commercialization (Millstone *et al.*, 1999). However, this is a target approach, which only analyzes a certain number of specific compounds known to be a safety concern. Unknown and unintended effects on metabolism outside of those specific compounds are not assessed on a regular basis. Therefore, a non-target, profiling approach using microarray technology can be used to evaluate potential changes on global gene expression (Kuiper *et al.*, 2001). A number of new studies have shown that microarray technology, based on non-target and unbiased approaches, is a potential tool to detect unintended effects (Gregersen *et al.*, 2005, Ouakfaoui and Miki, 2005, Baudo *et al.*, 2006). The whole transcriptome, representing all genes in the plant of study, is spotted on a glass slide, or printed on Affymetrix microarray GeneChips (Affymetrix, 2001). Then, by hybridizing the microarray with mRNA from the plant samples, global gene expression profiles can be generated. The gene expression profiles from a sample of novel crops can be compared to the gene expression profiles from a conventional crop in order to detect any potential differential gene expression. If information such as annotations of the sequences and their functional identification of the gene products is available, this may give further insight into biological functions and metabolic pathways that are relevant for crop safety assessment (Cellini *et al.*, 2004). Knowing in which crop each gene is up- or down-regulated, will also allow a targeted assessment of individual genes of particular interest. Thus, both targeted and non-targeted effects on gene expression can be assessed with microarray technology.

1.2 Rationale of this study

The rationale of this study was to analyze gene expression data in transgenic and conventional soybean cultivars and to develop a database for soybean transcriptome and ancillary data that can be applied as a part of crop safety assessment strategies. Five replicates of each of five soybean cultivars, totally consisting of data from twenty-five microarray hybridizations were obtained (Beaulieu, 2005). Two of the cultivars were transgenic and the other three cultivars were non-transgenic soybeans. The first scope of our study was to survey changes in gene expression among these five different cultivars in order to evaluate the range of variation within and between the groups of transgenic and/or non-transgenic cultivars. The second scope was to compare gene expression profiles of transgenic soybeans to non-transgenic soybeans to obtain lists of differentially expressed genes.

Since the hybridization on the whole transcriptome is carried out on one microarray chip, there is massive amount of data to process within one single experiment and bioinformatics tools are necessary to analyze the data. Most importantly, we need to integrate the gene expression data with other available information such as gene function to provide a comprehensive description of the experiment data. Therefore, a database is essential to handle, organize and interpret the results from the microarray experiment. With the integration of available soybean information into a database, it is possible to interpret and analyze the microarray experiment using meaningful terms, which describe the biological functions. Therefore, our third scope was to develop a database with a suite of web interfaces to allow users to easily retrieve data and results of the microarray experiment with cross-reference annotations of the expressed sequence tags (EST) and hyperlinks to external public databases. This environment makes it possible to explore differences in gene expression, if any, between transgenic and non-transgenic soybean cultivars and to interpret the results based on gene functional annotations to determine any changes that could alter biological processes.

1.3 Hypotheses

Our hypotheses are:

A. Transgenic and non-transgenic soybean genotypes can be distinguished by the analysis of global gene expression profiles using bioinformatics tools;

B. Gene expression profiles are more accurately compared on a gene function level than on a single gene level.

1.4 Objectives

Our specific objectives were:

1. to create a database for storing and analyzing soybean EST and gene expression data;
2. to combine gene expression data with information on gene product, molecular functions and metabolic pathways, to give biological meanings to EST sequences and microarray probes;
3. to establish methods and tools to compare gene expression profiles of different soybean cultivars based on a single gene level and on a biological functional level, such as gene ontology (GO) terms, to reveal possible changes in biological processes;
4. to create a web interface to analyze and visualize the data and results.

2 Literature review

2.1 Assessment of transgenic crops

Genetically modified (transgenic) plants have been designed to improve crop resistance to disease, pest, herbicide or abiotic stress; or to improve the crop's qualities and nutritional values (Dunwell, 2005). However, commercialized genetically modified crop has raised concerns relating to the safety of consumption. Since the insertion of transgenes (genes from another species) into plant DNA can cause disruption at the integration site. If the disruption site is a gene-rich region, it can induce mutation due to loss of gene function (Koncz *et al*., 1992). New gene products could potentially be produced due to the rearrangement of sequences. In addition, transgenes might interfere with other genes by unpredictable gene-gene interactions or gene regulation (Kuiper *et al*., 2002). The expression of one gene can also regulate the activity of other pathways due to metabolic crosstalk between them (Tattersall *et al*., 2001). For example, over-expression of phytoene synthase in transgenic canola not only increased the production of carotenoids, but also altered the production of other unexpected metabolites such as tocopherol, chlorophyll, fatty acyl composition and phytoene (Shewmaker *et al*., 1999). Therefore transgenic crops have the potential risk of expressing unintended toxic or allergenic proteins. However, traditional plant breeding also alters the genotype by introducing genetic recombination and therefore unintended effects are not limited to transgenic crops.

One desired crop trait is herbicide resistance, so that herbicide application in the field does not damage the crop. Glyphosate (RoundUp®) is a popular herbicide that inhibits the production of 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) and inhibits the shikimate pathway to produce aromatic amino acids and essential components of protein productions (Steinrucken and Amrhein, 1980). Since the shikimate pathway is very important for plant growth, application of glyphosate is lethal to plants. In RoundUp® ready soybean, the gene for a glyphosate tolerant EPSPS that is found in

Agrobacterium sp. (strain CP4) is inserted into the soybean genome to make soybean survive glyphosate applications (Padgette *et al*., 1995).

The current safety assessment of crops with novel traits (including transgenic plants) is based on the concept of 'substantial equivalence'. New crops are compared to a group of conventional crops (i.e. GRAS, Generally Recognized As Safe) that have a long history of safe use by compositional analysis (FAO/WHO, 2000). This is a "target approach" to assess the intended effect due to transgenes activity in known metabolic pathways. It also analyses major compounds such as essential nutrients and anti-nutrients, and naturally occurring toxins and allergens. It is assumed that if the chemical components in the new crop are substantially similar to the GRAS, it is as safe as the crops in our market and therefore the new trait is safe for commercialization (Millstone *et al*., 1999). However, the target approach has its limits to assess unexpected pleiotropic effects, because the key compounds being analyzed are restricted to a certain number of known compounds. Therefore, a non-target approach is needed to assess unpredictable, unintended effects.

One non-target approach for the assessment of unintended effects is the use of a profiling method (Kuiper *et al*., 2001). The idea is to screen for potential changes at the genome, transcriptome (gene expression), protein or metabolic pathway level (Kok and Kuiper, 2003). Currently only the whole genome or transcriptome of the species that have been sequenced can be used for profiling. The whole proteome or metabolome of any species has yet to be measured. The development of microarrays has allowed the simultaneous analysis of many thousands of genes (the transcriptome) (Baudo *et al*., 2006). Therefore microarray technology is an important tool for evaluating the pleiotropic effects on global gene expression (Ouakfaoui and Miki, 2005). Gene expression profiles can be generated from mRNA of a transgenic crop and compared to mRNA of a conventional crop. Thus, targeted and non-targeted changes to gene expression can be detected, which can lead to further study to explore their effects on the metabolic pathways and the food safety.

2.2 Previous studies to compare transgenic with non-transgenic plants

A few studies have been performed to compare gene expression profiles of transgenic plants with their non-transgenic counterpart, but to date no report present comparisons of gene expression profiles in transgenic soybean. One of the most remarkable studies using microarray technology was done on transgenic Arabidopsis plants generated with simple T-DNA constructs with the marker genes *nptll* and the reporter *uidA*, and subjected to various environmental stresses (Ouakfaoui and Miki, 2005). Gene expression data was analyzed using the Affymetrix Microarray Suite software (MAS 5.0), and only genes assigned by the software to have a two fold change (increase or decrease) were considered significantly differentially expressed. When comparing the global gene expression in the transformed lines to the control line under optimal growth conditions, only a small number of differentially expressed genes were found (varying between 39 and 180). These represented a very small portion (0.17%-0.8%) of the genes screened using the Affymetrix ATH1 Arabidopsis GeneChip (22,500 probe sets, representing the Arabidopsis transcriptome). The results showed that the insertion of the commonly used marker genes *uptll* and *uidA* has minimal effect on the global gene expression levels in transgenic Arabidopsis under optimal growth conditions, and that the T-DNA insertion of the transgenes leads to very little functional disturbance to the genomes of transgenic plants. More importantly, the number of genes affected by the insertion of transgenes was significantly lower than the number of genes affected by common abiotic stresses such as heat, cold, salt and drought (varying between 1080 and 4406). Also, when the gene expression profiles of transgenic lines were compared with the profile of the control line under abiotic stresses, the stress response was not different, meaning that the transgenes did not affect the stress response. The conclusion was that transgenic plants generated with simple T-DNA constructs containing common marker genes are equivalent to non-transgenic plants (Ouakfaoui and Miki, 2005).

Two studies on wheat reported the comparison of substantial equivalence of transgenic and non-transgenic crops at the transcriptome level (Gregersen *et al*., 2005, Baudo *et al*., 2006). The first study compared the gene expression profiles of the developing seeds in transgenic wheat transgenic for an *Aspergillus fumigatus* phytase with wild type wheat (Gregersen *et al*., 2005). A 9K cDNA microarray was employed to evaluate the use of microarray-based technique for detecting potential unintended effects in transgenic plants (Gregersen *et al*., 2005). The study applied the LIMMA software package (Smyth *et al*., 2005) from the Bioconductor package (Gentleman *et al*., 2004) for diagnostic plots and statistical analysis of the gene expression data, and real time RT-PCR analysis to validate the differentially expressed genes. The *lmFit* function of the LIMMA package was used to fit a linear model to the gene expression data, then followed by constructing a design matrix and a contrast matrix. The *eBayes* function was used in order to generate a list of differentially expressed genes based on the p-value calculated by a modified *t*-test. The estimated log2 fold changes (M-score) and log odds values (B-score) were also calculated by the *eBayes* function. The comparison between two wheat lines showed very slight variations for the three sampling time points (8, 16, 32 days after pollination) but the differentially expressed genes could not be confirmed by real time RT-PCR. The authors concluded that the expression of *A. fumigatus* phytase had no significant effects on the global gene expression pattern in the developing seeds of transgenic wheat (Gregersen *et al*., 2005).

The second study reported the comparison of gene expression profiles of transgenic and conventionally bred wheat lines that over-express genes encoding high molecular weight subunits of glutenin (Baudo *et al*., 2006). The same 9K cDNA microarray was used for pair-wise comparisons between the transgenic wheat line, conventionally bred wheat sister line and the non-transgenic control line. Gene expression data was analyzed using the commercial software GeneSpring (GeneSpring 6.2, Silicon Genetics, USA) and GenStat (GenStat 7th Edition, GenStat Procedure Library, Release PL15, Lawes Agricultural Trust, Rothamsted, Harpenden, UK). Genes with significant differential expression ($p < 0.05$ and 1.5 fold change) were identified. The numbers of differentially expressed genes in the comparison between transgenic line and the non-transgenic line at 8, 14, 28 days post-anthesis were 6, 5 and 2 respectively. It only represented a small proportion (0.06%, 0.05% and 0.02 %) of the genes spotted on the microarray. In the comparison between conventional bred line and non-transgenic line, the number of differentially expressed genes varied from 26 to 527 (0.27% to 5.59%). In the comparison between transgenic line and conventional bred line, the number of differentially expressed genes varied from 4 to 154 (0.04% to 1.63%). The results showed that transgenic manipulation led to very small changes in expression profiles. Most importantly, there were greater differences in gene expression due to conventional breeding than genetic modification in transgenic wheat. This implied that the presence of the transgene and associated T-DNA with marker and reporter genes has smaller impact on global gene expression patterns than gene recombination thru conventional breeding. As with the previous study, the conclusion is that a single transgene has minimal effects on the transcriptome and a transgenic crop can be substantially equivalent to the control non-transformed line (Baudo *et al*., 2006).

All three studies demonstrate the use of microarray technology in the comparison between transgenic and non-transgenic plants and indicate that microarray is a potential tool to determine substantial equivalence in crop safety assessment. Also, the comparisons between transgenic and non-transgenic control lines imply that transgenesis has minimal effects on the global gene expression patterns in these transgenic plants.

2.3 Soybean microarray technology

Although the soybean genome sequence is not completed, there are many expressed sequenced tags (EST) available in public databases, which can represent the transcriptome (a collection of all transcribed genes). For example, about 330,000 EST sequences were generated by the Public EST Project for soybean (Shoemaker *et al*., 2002) and the Functional Genomics Program for Soybean (Vodkin *et al*., 2004) together. Global gene expression profiles can be studied with cDNA microarrays containing around 30,000 representative soybean genes (Vodkin *et al*., 2004) or with Affymetrix GeneChip arrays (Affymetrix, 2001).

The Affymetrix Soybean GeneChip contains 35,611 soybean transcripts (Affymetrix 2004-5). It also contains probes for the transcriptomes of two pathogens: the fungal pathogen *Phytophthora sojae* (represented by 15,421 probe sets) and the cyst nematode *Heterodera glycines* (represented by 7,431 probe sets). An Affymetrix probe set represents a transcript or a gene and consists of 11 oligonucleotide probe pairs, each 25 nucleotides long and spanning regions of each gene. The probe pair contains two probes, a perfect match probe (perfectly matches its target sequence) and a mismatch probe (where the 13th nucleotide is a mismatch), in order to also assess non-specific hybridization (Affymetrix, 2001). Including control probes, there are a total of 61,170 probe sets on the soybean Affymetrix GeneChips, consisting of over 1,340,000 probes.

2.4 Previous studies using Affymetrix Soybean GeneChip

Several studies have been made using soybean cDNA arrays, but to date very few have used the Affymetrix GeneChips. Two examples of the experiments using Affymetrix Soybean GeneChip were the analysis of gene expression profiles of host and pathogen in nematode-infected soybean (Ithal *et al*., 2007) and the study of changes in gene expression affected by the Asian soybean rust disease (Panthee *et al*., 2007). In the first study, the root tissues of soybean cyst nematode (*H. glycines*) infected and uninfected soybean were compared at three time points (2, 5 and 10 days post-infection) (Ithal *et al*., 2007). GeneChip Operating Software version 1.0 (GCOS v. 1.0) was used for statistical analysis. An F test followed by converting the p-values to q-values (Ithal *et al*., 2007), identified genes with q-values less than 0.05 and 1.5 fold change as significantly differentially expressed. Four hundred and twenty nine differentially expressed genes were identified among the 35,611 soybean genes present on the array; and 1,850 differentially expressed genes identified among the 7,431 *H. glycines* genes present on the same array. The soybean EST sequences corresponding to the identified genes were used as a query and search for the *Arabidopsis* orthologs in the TAIR database using WU-BLAST2 search. The top hit with an e-value less than 10^{-3} was used to annotate the soybean genes. Among the 429 differentially expressed genes, 320 of them were assigned with putative functions. The putative annotations were not experimentally confirmed, but serve as useful information for selecting genes of interest for further studies.

The second study used Affymetrix soybean GeneChip for gene expression profiling of soybean with Asian soybean rust disease that caused by *Phakopsora pachyrhizi* (Panthee *et al*., 2007). Gene expression data was analyzed using ArrayAssist Software from Stratagene. Differentially expressed genes were identified using unpaired *t*-test by a cut-off p-value < 0.05. The functions of putative encoded proteins were assigned to the differentially expressed genes using ExPasy protein database (htty://us/expasy.org). There were 112 differentially expressed genes found in 3-weeks old leaves (V2 growth stage) in response to *P. pachyrhizi* infection after 72 hours of inoculation. Most of the upregulated genes are identified as being involved in defense- and stress-related responses (Panthee *et al*., 2007). Both studies demonstrate that the Affymetrix Soybean GeneChip is a reliable and comprehensive platform to perform hybridization experiment for the soybean transcriptome.

2.5 Public data resources for soybean gene annotations

Since one single hybridization experiment produces massive amounts of data, handling, processing and analyzing pose a challenging task. Thus, the application of bioinformatics to microarray data analysis is essential. Before analysis, data has to be stored and organized in a database, which can serve as a repository. The database has to be extensible and flexible to compare data from different microarray experiments. It can incorporate statistic methods and algorithms to allow the detection of unintended effects on gene expression. It is feasible to combine transcript data with information on genetics, homology, functions, metabolic regulations and toxicology. Thereafter, it can correlate gene expression data with known biological processes and pathways, and hence, allow us to understand what the expression data tells us about the difference in transgenic and non-transgenic soybean.

2.5.1 Soybean expressed sequence tag (EST) data

An Expressed Sequence Tag (EST) is a short sequence (tag) of a transcribed gene. A collection of ESTs is sequenced from cDNA libraries constructed from mRNA extracted from different tissue and organ systems at various developmental stages. The EST sequence is obtained from the raw chromatograms generated by the DNA sequencer and subsequently processed into high-quality sequences for publication by deleting the cloning vectors, poly-tails and short repeat sequences. Assembly software such as Phrap (www.phrap.org/) is used to align and assemble them into longer consensus sequences (contigs), which represent a gene. The Public EST Project For Soybean (Shoemaker *et al*., 2002) produced over 300,000 5' ESTs from around 80 cDNA libraries representing many different tissues and physiological conditions of the soybean plant. A concurrent project, The Functional Genomics Program For Soybean (Vodkin *et al*., 2004) selected around 35,000 of the 5' ESTs and sequenced the corresponding 3' sequence to construct cDNA microarrays. The data is housed in the Soybean Genome Initiative (http://soybean.ccgb.umn.edu/) within the BioData system (http://biodata.ccgb/umn/edu/) at Center for Biomedical Research Informatics (formerly the Center for Computational Genomics and Bioinformatics) at University of Minnesota. This system contains the EST sequences, contig data, BLAST (Basic Local Alignment Search Tool) (Altschul *et al*., 1990) results, and statistics information, which are organized in a data file system. It displays very detailed information about each library, includes cloning method, cultivar, tissue type, developmental stage and number of ESTs. It also contains GenBank Accession number, raw and filtered sequences for each EST. There is also a graph to show the quality of each sequence and its metadata. Protein annotation can also be obtained from the BLAST results.

The Legume Information System (LIS) (http://www.comparative-legumes.org) is a comparative resource for legumes includes soybean, *Medicago truncaula* and *Lotus japonicus* (Gonzales *et al*., 2005). LIS gathered transcript data from legume plants and Arabidopsis from NCBI High Throughput Genomic division (Benson *et al*., 2007). These unfinished sequences were generated from large-scale genomic projects and are undergoing various stages of assembly processing. LIS takes these sequences and their constituent contigs from GenBank, and obtains their consensus sequences using a sliding window of 10,000 bp with an overlap of 3,000 bp. Then, these consensus sequences are analyzed using different sequence databases and provide protein name, protein blocks, and motif information. Also, LIS developed their EST database from NCBI raw EST and cDNA data. The raw EST data are screened for quality and contamination. Then, the cleaned EST data are assembled using Phrap (htp://www.phrap.org). These consensus sequences are then annotated with protein names, blocks, motif information. LIS also integrates genetic maps, physical maps and pathway information from collaborate projects such as SoyBase (http://www.soybase.org/) and Southern Illinois University soybean genome project (http://soybeangenome.siu.edu/).

The Institute of Genomic Research (TIGR) is the major institute participating in the Human Genome Projects. They also collected publicly available EST sequences (including soybean ESTs) to assemble into tentative concensus (TC) sequences, which represent genes (Quackenbush *et al*., 2000). TIGR developed bioinformatics tools to assemble EST sequences and assign annotations to TCs. Soybean EST and TC information can be retrieved from their Gene Indices web page (http://compbio.dfci.harvard.edu/tgi/plant.html). TC number is also widely used by the scientific community and it is also mapped to Gene Ontology (GO) terms to obtain biological functional terms.

2.5.2 Nucleotide data and resources

The NCBI (National Center for Biotechnology Information) public genome database GenBank (Benson *et al*., 2007) is part of the International Nucleotide Sequence Database Collaboration. Comprehensive DNA sequence information is collected from genome projects around the world and can be retrieved from the world wide web (http://www.ncbi.nlm.nih.gov/). The data is further organized into different databases such as dbEST for transcripts data and UniGene for gene-oriented clusters of transcript sequences. Also, other databases include whole genome sequences, three-dimensional macromolecular structures, taxonomy, single nucleotide polymorphism, chemical molecules and substances, protein domains, microarray data, cancer and disease related chromosomes, and journals. These databases provide comprehensive information about a gene or protein of interest and a web tool BLAST (Basic Local Alignment Search Tool) to compare nucleotide or protein sequences by sequence similarity searches (Altschul *et al*., 1990). Therefore it is widely used by researchers to obtain information, and has become the core data system for the scientific community. Including GenBank Accession numbers as identifiers facilitates communication between different databases.

2.5.3 Protein data and resources

Swissprot is a protein knowledgebase maintained by The Swiss Institute for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI) (http://ca.expasy.org/sprot/). It integrates protein sequences with updated biological knowledge and manually curated entries (Boeckmann *et al*., 2003). The core data consists of amino acid sequence, protein name, taxonomic data and citation information. Each protein entry is provided with high-quality annotation on protein function, enzymatic information e.g., enzyme commission (EC) number, secondary and quaternary structure, etc. The nomenclature is standardized to facilitate communication across different databases. It is designed especially to closely follow the format of other EBI databases. Therefore the SwissProt identifier is excellent to use for making links to other important databases such as Gene Ontology (http://www/geneontology.org) and Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/).

2.5.4 Functional annotations

Gene and protein sequences require annotations to describe their functions. Gene Ontology (GO) (http://www.geneontology.org/) provides a standard set of terminology to describe gene products across different databases consistently. Gene products are classified according to their biological processes, molecular functions and sub-cellular location (The Gene Ontology Consortium, 2000). The GO terms are organized into a tree-like structure called directed acyclic graphs (DAGs) to resemble a hierarchy. This allows a more specialized child term to have one or more less specialized parent terms. All the child terms inherit all the properties of their parent terms. Therefore, when a gene product is annotated with a child term, then all the parent terms also apply to that gene product. GO terms are written, maintained and curated by the GO collaborators. They also make association between GO and other genomic and proteomic public databases such as Swissprot and TIGR, thus it can facilitate uniform queries across them.

In order to understand the metabolic function of the gene products, the curators of KEGG (Kyoto Encyclopedia of Genes and Genomes) have organized information of metabolic pathways manually entered from published materials (Kanehisa *et al*., 2004). The pathway database (http://www.genome.jp/kegg/pathway.html) integrates current knowledge on molecular interaction networks and biological processes. The reference maps of metabolic network were generated to show protein interaction, for example, direct protein-protein interaction, gene expression relation, and enzyme-enzyme relation. All these enzymes are assigned with EC numbers.

2.6 Previous study on interpreting gene expression data using functional terms

Biological interpretation of microarray experiments is needed to provide biological knowledge and facilitate communication among different laboratories and across platforms. The list of differentially expressed genes resulted from microarray analysis are usually translated into functional annotations by searching through literature and multiple public databases gene-by-gene manually (Draghici *et al*., 2003). However, this is a tedious and slow process. Therefore, several tools for automatically assigning functional annotations (such as GO terms) to microarray experiment have been developed, such as GOStat (Beissbarth and Speed, 2004) and FatiGO (Al-Shahrour *et al*., 2007). However, annotations are only provided for a limited set of organisms, for instance, yeast, human, mouse, *Drosophila* and *Arabidopsis*. Most of all, the frequency of occurrence of a functional annotation from differentially expressed genes may be misleading because the number of genes involved in different gene classes (represented by GO terms) are different, and thus, the probability to observe each GO term varies. In order to measure significance of observed GO terms, both tools apply hypergeometric distribution to calculate the probability for the observed numbers of each GO term resulting from random distribution (Draghici *et al*., 2003). A χ^2 test or Fisher's Exact Test is used to compare the expected probability with the observed probability, and provides p-values for ranking the list of GO terms.

2.7 EST and microarray databases

ESTIMA, Expressed Sequence Tag Information Management and Annotation project, is an open-source database system designed to organize EST data from multiple high-throughput EST sequencing projects such as honeybee, cattle, and songbird (Kumar *et al*., 2004). The database, which was developed for the Oracle database management system (www.oracle.com), includes cDNA library information, EST sequences and their metadata, contig information, and gene function annotations such as Gene Ontology terms and homolog ID through BLAST search. The web interface allows users to access the database and retrieve results (http://titan.biotec.uiuc.edu/ESTIMA/).

There are several database projects that combine EST data with microarray data, for example, SGMD and BarleyBase.

SGMD (the Soybean Genomics and Microarray Database) stores EST and microarray data to explore the interaction of soybean with the major pest, soybean cyst nematode (SCN) (http://psi081.ba.ars.usda.gov/SGMD/Default.htm). The database stores over 50 million rows of DNA microarray data and around 20,000 EST data (Alkharouf and Matthews, 2004). Relevant EST information is stored in the database such as cloning information, GenBank accession number, BLAST results and links to PubMed to view relevant journal citations. The web interfaces are embedded with analytical tools, for example, analysis of variance (ANOVA), *t*-tests and K-means clustering to show the result and its significance and reproducibility of measurement.

The SGMD web interface provides on-the-fly statistics analysis to compare cDNA microarray data. However, the SGMD database only contains around 20,000 EST data from the soybean root libraries. There are only GenBank IDs and BLASTX reports to show the homology of genes and proteins, and no annotations are provided to give information of the biological function and metabolic pathway.

BarleyBase (www.barleybase.org) is developed as a database for cereal microarray data (Shen *et al*., 2005) but has the capacity to included data from other plants as well. It houses raw and normalized microarray data from Affymetrix Barley and Arabidopsis GenChips with comprehensive annotations. It is also integrated with analysis and visualization tools to explore and compare microarray experiments. The database stores data of the microarray chip, experimental protocols, raw and normalized gene expression data, and annotations such as plant ontologies, BLAST hits, Gene Ontology, pathway and gene family information. The analysis tools are very flexible allowing data analyses based on experiment, between hybridizations or treatments, gene-centric expression profiles, (for example, the analysis a subset of data with certain biological criteria such as annotation keywords, gene family or KEGG pathway.) The data can be analyzed using the R statistical toolbox (Hornik, 2007). The results can be displayed in a tabulated format with profile plots and heatmaps. Gene lists can also be exported in tab-delimited text files. There are many visualization tools including box plots and histograms for expression values, and scatter plots to show reproducibility and variability of experiment comparison. BarleyBase also links to other public plant databases.

BarleyBase is an excellent database system to analyze and visualize microarray data. It is planning to expand to support multiple species experiment including maize, rice, wheat and soybean. Currently, it contains Affymetrix soybean GeneChip data with annotations. The soybean annotations include BLAST hits against UniProt/TrEMBL, TIGR, Barley GeneChip, and Arabidopsis GeneChip sequences. UniProt/TrEMBL is a bigger set of data contains all the computer-annotated translations of The European Molecular Biology Laboratory (EMBL) (in nucleotide sequences but not in SwissProt.) To date, there is no microarray data and no functional annotation, for example, GO or

KEGG for soybean in BarleyBase. In order to investigate unintended effect of transgenic soybean, we need additional information on the metabolic process and toxicology. However, currently no bioinformatics system is set up for this type of integrated soybean data.

2.8 Conclusion

A non-targeted method based on microarray technology is needed to compare cultivars from the same crop to assess the significance of changes as a result of trait modification. Two studies have been done to examine substantial equivalence between transgenic and non-transgenic wheat for safety assessment and demonstrated that these transgenic plants were substantially equivalent to their conventional counterparts. However, no published journal paper reported on transgenic soybean using microarray technology. The Affymetrix Soybean GeneChip has been used to study differences in gene expression due to infection by pathogens and demonstrated that it is a reliable platform to screen for changes in the whole transcriptome. Several EST and microarray databases were developed, however no database is specially designed for interpreting biological knowledge in the comparison between transgenic and non-transgenic soybean.

3 Materials and methods

3.1 Data processing and database construction

All computations were performed on a Mac Power PC G5 (dual) running Mac OS X operating system version 10.4.9 equipped with 8 GB RAM. Perl (version 5.8.6) (www.perl.com) scripts were written for parsing data files and to load data into a MySQL relational database (version 5.0.18) (http://www.mysql.com). Perl CGI (http://search.cpan.org/dist/CGI.pm/) scripts were used to create the web-interfaces, the perl module DBI and DBD::mysql (http://dev.mysql.com/downloads/dbi.html) was used to connect CGI scripts to our database and the CGIwithR package (Firth, 2003) was used for running R (Hornik, 2007) statistical analysis within the CGI script.

3.1.1 Soybean sequence data

Three types of information were obtained before being organized and stored into our database: soybean EST sequence data, corresponding annotations, and microarray data. We obtained soybean EST sequences in the BioData file format from 84 cDNA libraries constructed by the groups of Dr. R. Shoemaker (Iowa State University) and Dr. L. Vodkin (University of Illinois) (Shoemaker *et al*., 2002, Vodkin *et al*., 2004) courtesy of Dr. Ernest Retzel, CBRI (formerly Center for Computational Genomics and Bioinformatics) University of Minnesota. A total of 279,714 5' EST sequences were obtained from "A Public EST Project for Soybean" (Shoemaker *et al*., 2002) and 29,772 3' EST sequences, sequenced from clones chosen from among the 5' EST sequences, were obtained from "A Functional Genomics Program for Soybean" (Vodkin *et al*., 2004). An additional 8,936 EST sequences from two re-rack cDNA libraries of the two soybean EST projects were obtained from NCBI GenBank. In addition, 61,673 soybean (mRNA) sequences other than the above projects were also obtained from NCBI GenBank (downloaded in Sept 2005). All the essential information from these soybean sequences such as EST ID, sequence length, sequence processing information (e.g., location of repeats, trim, polyA-tail), clone ID, library and vector information, and their corresponding GenBank accession and gi number, were stored in our database. Raw and processed soybean EST sequences and mRNA sequences downloaded from NCBI were organized in our computer file system, which was adapted from the University of Minnesota BioData system (http://biodata.ccgb.umn.edu/).

3.1.2 Protein annotation using SwissProt

Sequence similarity searches were performed on all the soybean EST sequences against 168,297 SwissProt protein sequences (http://ca.expasy.org/sprot/) (Gasteiger *et al.*, 2003) using the BLASTX program (Altschul *et al.*, 1990) to obtain corresponding protein annotations. SwissProt protein sequences were downloaded in Feb 2005 and formatted into a BLAST target database. BLASTX was done using the standalone BLAST program (version 2.2.10) downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/. Corresponding protein annotations such as protein ID and definition, blast search hit score and e-value were obtained from the BLASTX results and stored into our database.

3.1.3 Functional annotation using Gene Ontology

The Gene Ontology databases (The Gene Ontology Consortium, 2000) including the MySQL tables: term, term_definition, term2term, and graph_path were downloaded in January 2005 from http://www.geneontology.org/GO.downloads.database.shtml and directly reproduced in our database. The GO terms link to our soybean sequences through the BLAST results with the SwissProt protein ID. The associations between SwissProt and GO terms were obtained from the file: "UniProt GO Annotations" downloaded from http://www.geneontology.org/GO.current.annotations.shtml/ and were integrated into the database.

3.1.4 Functional annotation using the Enzyme Commission numbers

A list of recommended enzyme names and EC numbers were obtained in Jun 2005 from the Enzyme Nomenclature site http://www.chem.qmul.ac.uk/iubmb/enzyme/

and integrated into the database. In additional, enzyme names and EC numbers were also extracted from MeSH (Medical Subject Headings, National Library of Medicine) (http://www.nlm.nih.gov/mesh). 2005 MeSH files were downloaded in May 2005 from http://www.nlm.nih.gov/mesh/filelist.html. The associations between Enzyme EC numbers and SwissProt protein IDs were obtained from ExPASy Enzyme nomenclature database (version 36) http://ca.expasy.org/enzyme/ in Jan 2005 (Gasteiger *et al*., 2003). All EC numbers and enzyme names were integrated into the database and linked to the soybean sequences through SwissProt protein IDs.

3.1.5 Functional annotation using KEGG pathways

Metabolic and regulatory pathways were downloaded from the ftp site at KEGG (Kyoto Encyclopedia of Genes and Genomes) http://www.genome.jp/anonftp/ (Kanehisa *et al*., 2006). Enzymes identities within each pathway were obtained by extracting EC numbers from each of the pathways (downloadable XML files from the ftp KGML/map folders, version 0.6 Mar 2005). EC numbers, pathway names and map numbers where extracted and integrated into the database. By linking soybean EST ID through SwissProt protein ID through EC number to the metabolic pathway, an EST sequence, representing a gene coding for an enzyme involved in a particular metabolic pathway can be retrieved from our database. Diagrams of the pathway maps were also downloaded from KEGG and were organized in our computer file system.

3.1.6 Sequence annotation using TIGR

Information of 31,928 tentative consensus (TC) sequences was downloaded from The TIGR (The Institute for Genomic Research) Glycine max Gene Index Project (Release 12.0) (Quackenbush *et al*., 2000). TC numbers, GenBank accession numbers of the member ESTs of the TCs and GO annotations were integrated into our database and linked to our soybean sequences through GenBank accession numbers.

3.1.7 Soybean microarray data

Twenty-five raw data files (CEL files) of a microarray experiment using Affymetrix Soybean GeneChip (Affymetrix, 2004-5) were obtained from Dr. Marc Fortin (Beaulieu, 2005) and used as a starting point for the processing and analysis in this thesis. The data analysis of this microarray experiment is described in the next section. All raw data (e.g., probe intensities), pre-processed data (e.g., normalized probe-set intensities), and results from statistical analysis (e.g., fold change, statistics scores and p-value) were organized and stored in the database. Information about the microarray GeneChip such as probe sequences, probe location of the chip, and corresponding GenBank accession number of the probes were integrated into the soybean EST and annotation database to describe the microarray data. Probe identifiers were linked to GenBank accession numbers, which were further linked to the SwissProt protein identities and functional annotations. Information on the probes (such as soybean probe sequences, consensus sequences of the probes, probes' locations on the chip) was downloaded from Affymetrix website http://www.affymetrix.com/support/technical/byproduct.affx?product=soy.

3.2 Microarray analysis

3.2.1 Soybean biological samples

The microarray experiment was designed and mRNA extracted by Julie Beaulieu under the supervision of Dr. Marc Fortin (Beaulieu, 2005). The hybridizations and scans were carried out at the Genome Quebec and McGill University Genome Centre. Five biological replications were performed for each of the five cultivars for a total of twenty-five microarray hybridizations (CEL files) using the Affymetrix Soybean GeneChips (Affymetrix, 2004-5).

Two of the soybean varieties (2601R and PS46RR Monsanto Canada Inc. Guelph) are transgenic (resistant to the herbicide glyphosate (RoundUp®)), whereas three cultivars (Mandarin Ottawa, S03W4 and Bayfield) are conventional. The soybean cultivars were selected because of similarities such as maturity group, field trial performances (yield and days to maturity), and biochemical content (Beaulieu, 2005).

Plants were grown in a growth chamber under optimal growth condition: 16-hours photoperiod and 25/19° day/night temperatures. The first trifoliate leaves were harvested at the V2 stage when they were completely unrolled. RNA extraction was done using the RNeasy Plant Mini Kit (Qiagen). Quality assessment was tested by Agilent 2100 bioanalyzer (Palo Alto, CA). Affymetrix GeneChip hybridization and processing were done at the McGill University and Genome Quebec Innovation Center Microarray platform (Beaulieu, 2005).

3.2.2 Microarray pre-processing

Data analysis was done using R (Hornik, 2007) and the BioConductor packages (Gentleman *et al*., 2004) such as *affy, limma, cluster* and *made4*. Quality assessment of the microarray data was done using *affyRNAdeg* function from the *affy* package. All 61,170 probe sets (138,734 probes) including control probes and probes for *Glycine max*, *Phytophthora sojae* and *Heterodera glycines* were pre-processed and normalized together. The microarray data was pre-processed using three different normalization methods: RMA (Irizarry *et al*., 2003), MAS5 (Affymetrix 2002) and dChip (Li and Wong, 2001).

3.2.3 Analysis method 1: gene expression at the gene level

Soybean cultivar specific pattern of gene expression (classifying them into groups) was examined by principle component analysis (PCA) analysis for the twenty-five non-processed chips and unsupervised hierarchical clustering for the twenty-five normalized chips was carried out using Euclidean distance and average linkage. The closest related non-transgenic soybean to each transgenic soybean is defined from the clustering where the distance between two cultivars is the shortest.

The comparison of gene expression profiles was based on two approaches: (1) pair-wise comparison of a transgenic cultivar with its closest related non-transgenic counterpart (as per group clustering analysis); and (2) one transgenic cultivar compared with a group of non-transgenic cultivars based on the concept of "substantial equivalence".

To evaluate the variation of gene expression between different soybean cultivars, pair-wise comparison for every two cultivars was done using LIMMA (Linear Models for Microarray Data) (Smyth *et al*., 2005) at p-value < 0.01 and fold change > 2. The data from each transgenic cultivar was compared with the data from the closest related non-transgenic cultivar to differentiate gene expression. The RMA processed data are in log2 base; MAS5 and dChip processed data are in log10 base. Before using the LIMMA package, MAS5 and dChip processed data were transformed to log2 base for statistical analysis. However, for calculating the differences of intensity in two samples by fold change, RMA processed data were transformed to log10 base.

The possibility of applying the concept of "substantial equivalence" in microarray experiment was evaluated in our second approach by grouping the data from the three non-transgenic cultivars as the reference group and compare the gene expression with each of our two transgenic cultivars using LIMMA at p-value < 0.01 and fold change > 2.

All microarray data, including raw intensities from CEL files, pre-processed data using three normalizations and summarized methods, log transformed intensities, t-scores and p-values from LIMMA analysis and information about the probes, were stored and integrated into the soybean database.

3.2.4 Analysis method 2: gene expression at the functional term level

To analyze gene expression by functional groups (based on GO terms) rather than by individual probes, the parent GO terms (describing the biological function in a more general term) were traced back from the child GO terms and associated with the probes.

All probes were first filtered in order to prevent uninformative probes averaging out the changes. Probes that had larger than two-fold change in any pair-wise comparisons were included. Each probe was linked with GO terms and parent GO terms, and then (normalized) intensities of probes that shared the same GO terms were averaged as the combined intensity for each GO term. The combined intensities were then transformed into log2 base and used for LIMMA analysis to analyze gene expression at a functional-term level. Pair-wise comparisons between each of the two transgenic soybeans to non-transgenic soybean Bayfield were done. A list of GO terms that were distinguished at p-value < 0.01 was obtained. GO terms, t-scores and p-values from LIMMA analysis for each comparison were stored in the database.

4 Results

4.1 Database development

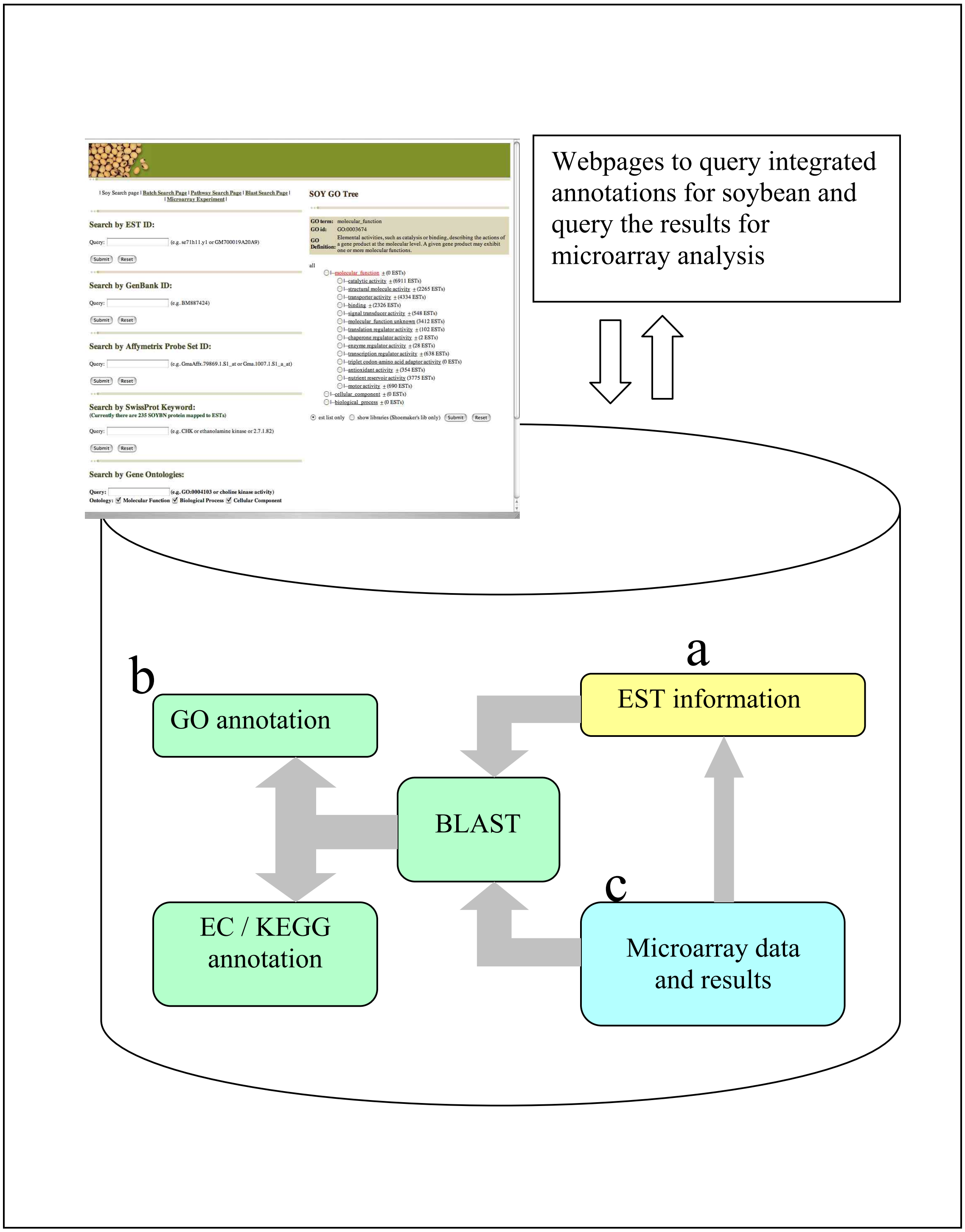
4.1.1 Database description: soybean transcripts data

We have integrated soybean EST sequences with functional annotations and microarray data and coupled the database with web interfaces to access and display the information, shown as an overview in **Figure 4.1**. The core of our database is the 318,422 EST sequences and ancillary information about the soybean cDNA libraries from which they were obtained. **Figure 4.2** shows the tables that organize the sequence information. The table DNA_SEQUENCE specifies the sequence ID, length and location where the sequence is stored in our file system (the path) whereas the ancillary information about the EST sequences such as the locations of the clone vector, polyA-tail, repeat sequence and the trim site are stored in the tables VECTOR, TAIL, REPEATS and TRIM, respectively. The cDNA library information including the library ID, tissue type, and growing conditions are stored in the table LIBRARY. The library information is linked to the sequence information through table SEQ_ACCESSION, which maps the sequence ID, library ID and GenBank accession number. The SEQ_ACCESSION table can link to the BLAST table by using sequence ID as the query ID for linking to the BLASTX search results. The GenBank accession number from the SEQ_ACCESSION table can link to the TIGR contig information through GenBank accession number to obtain the corresponding contig ID for the EST sequences (from table TIGR_GB) and the GO terms associated with each contig (from the table TIGR_GO). Other information for the additional 8,936 EST sequences downloaded from NCBI websites are stored in the table GB_ACCESSION, which also links to the BLAST table by using the GenBank accession number as the query ID.

Figure 4.1. Overview of the soybean database structure.

a) Soybean EST and contig information obtained from public EST projects, GenBank and TIGR were organized in tables modified from the ESTIMA database schema to include EST sequencing pipeline data
b) Functional annotations such as GO terms, EC numbers and KEGG molecular pathways link to the EST and microarray data through the protein names obtained from the BLASTX results
c) Microarray data and results link to the annotations through BLASTX results and GenBank accession numbers of the EST sequences.

Figure 4.1



| Soy Search page | Batch Search Page | Pathway Search Page | Blast Search Page |
| Microarray Experiment |

Search by EST ID:

Query: (e.g. se71h11.y1 or GM700019A20A9)

Submit Reset

Search by GenBank ID:

Query: (e.g. BM887424)

Submit Reset

Search by Affymetrix Probe Set ID:

Query: (e.g. GmaAffx.79869.1.S1_at or Gma.1007.1.S1_a_at)

Submit Reset

Search by SwissProt Keyword:
(Currently there are 235 SOYBN protein mapped to ESTs)

Query: (e.g. CHK or ethanolamine kinase or 2.7.1.82)

Submit Reset

Search by Gene Ontologies:

Query: (e.g. GO:0004103 or choline kinase activity)
Ontology: ☑ Molecular Function ☑ Biological Process ☑ Cellular Component

SOY GO Tree

GO term: molecular_function
GO id: GO:0003674
GO Definition: Elemental activities, such as catalysis or binding, describing the actions of a gene product at the molecular level. A given gene product may exhibit one or more molecular functions.

```
all
  ○ |--molecular_function + (0 ESTs)
      ○ |--catalytic activity + (6911 ESTs)
      ○ |--structural molecule activity + (2265 ESTs)
      ○ |--transporter activity + (4334 ESTs)
      ○ |--binding + (2326 ESTs)
      ○ |--signal transducer activity + (548 ESTs)
      ○ |--molecular_function unknown (3412 ESTs)
      ○ |--translation regulator activity + (102 ESTs)
      ○ |--chaperone regulator activity + (2 ESTs)
      ○ |--enzyme regulator activity + (28 ESTs)
      ○ |--transcription regulator activity + (638 ESTs)
      ○ |--triplet codon-amino acid adaptor activity (0 ESTs)
      ○ |--antioxidant activity + (354 ESTs)
      ○ |--nutrient reservoir activity (3775 ESTs)
      ○ |--motor activity + (690 ESTs)
  ○ |--cellular_component + (0 ESTs)
  ○ |--biological_process + (0 ESTs)
```

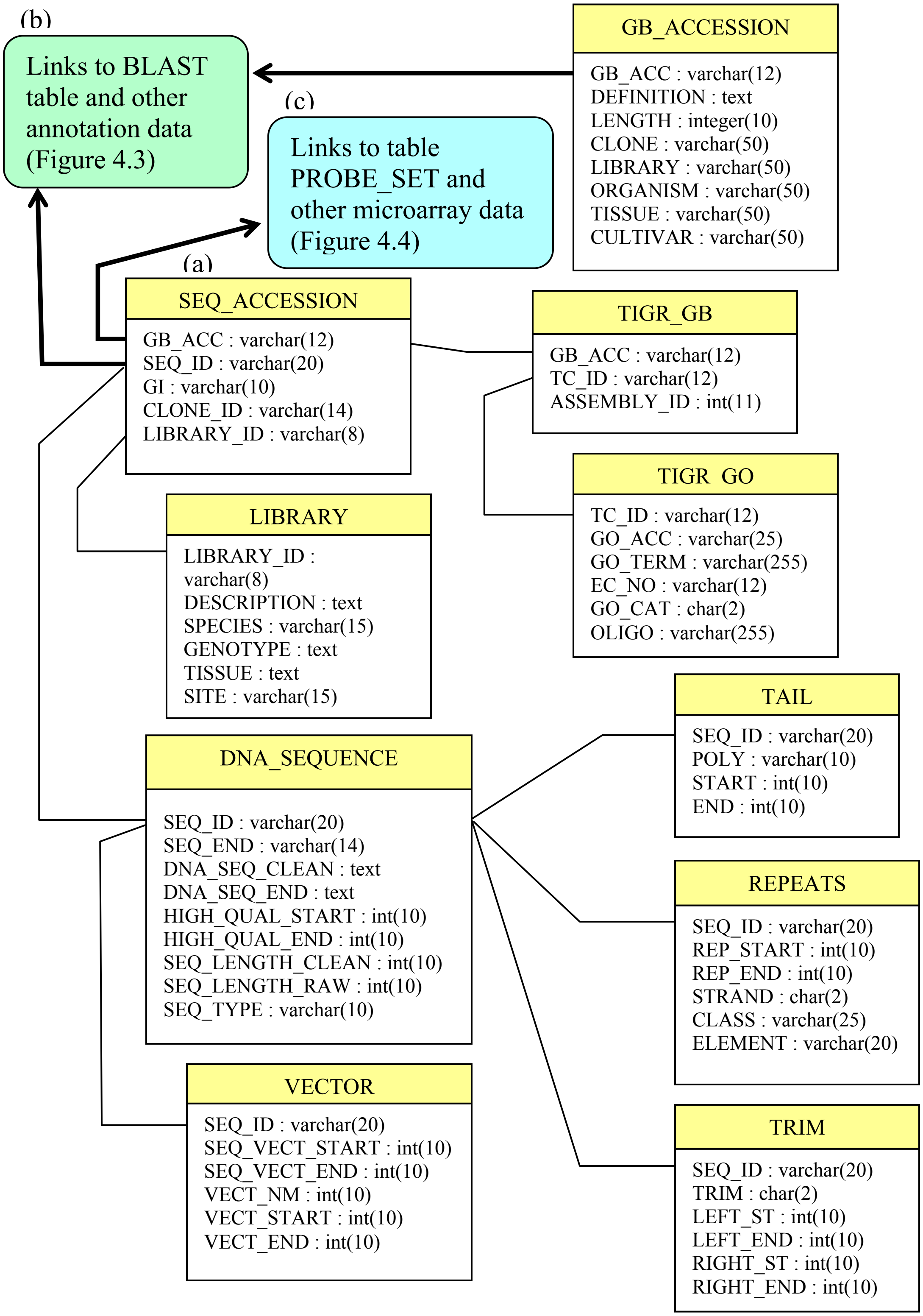
● est list only ○ show libraries (Shoemaker's lib only) Submit Reset

Webpages to query integrated annotations for soybean and query the results for microarray analysis
a
b
c
EST information
GO annotation
BLAST
EC / KEGG annotation
Microarray data and results

Figure 4.2. Database structure of the section for gene transcript (sequence) information.

a) Tables for EST sequences (DNA_SEQUENCE, LIBRARY, REPEATS, SEQ_ACCESSION, TAIL, TRIM, VECTOR), table for mRNA sequences downloaded from GenBank (GB_ACCESSION) and tables for TIGR contigs data (TIGR_GB, TIGR_GO)
b) Tables GB_ACCESSION and SEQ_ACCESSION link sequence data to annotation data
c) Table SEQ_ACCESSION links sequence data to microarray data.

Figure 4.2


(b)
Links to BLAST table and other annotation data (Figure 4.3)
(c)
Links to table PROBE_SET and other microarray data (Figure 4.4)
(a)
GB_ACCESSION
GB_ACC : varchar(12)
DEFINITION : text
LENGTH : integer(10)
CLONE : varchar(50)
LIBRARY : varchar(50)
ORGANISM : varchar(50)
TISSUE : varchar(50)
CULTIVAR : varchar(50)
SEQ_ACCESSION
GB_ACC : varchar(12)
SEQ_ID : varchar(20)
GI : varchar(10)
CLONE_ID : varchar(14)
LIBRARY_ID : varchar(8)
TIGR_GB
GB_ACC : varchar(12)
TC_ID : varchar(12)
ASSEMBLY_ID : int(11)
TIGR_GO
TC_ID : varchar(12)
GO_ACC : varchar(25)
GO_TERM : varchar(255)
EC_NO : varchar(12)
GO_CAT : char(2)
OLIGO : varchar(255)
LIBRARY
LIBRARY_ID : varchar(8)
DESCRIPTION : text
SPECIES : varchar(15)
GENOTYPE : text
TISSUE : text
SITE : varchar(15)
TAIL
SEQ_ID : varchar(20)
POLY : varchar(10)
START : int(10)
END : int(10)
DNA_SEQUENCE
SEQ_ID : varchar(20)
SEQ_END : varchar(14)
DNA_SEQ_CLEAN : text
DNA_SEQ_END : text
HIGH_QUAL_START : int(10)
HIGH_QUAL_END : int(10)
SEQ_LENGTH_CLEAN : int(10)
SEQ_LENGTH_RAW : int(10)
SEQ_TYPE : varchar(10)
REPEATS
SEQ_ID : varchar(20)
REP_START : int(10)
REP_END : int(10)
STRAND : char(2)
CLASS : varchar(25)
ELEMENT : varchar(20)
VECTOR
SEQ_ID : varchar(20)
SEQ_VECT_START : int(10)
SEQ_VECT_END : int(10)
VECT_NM : int(10)
VECT_START : int(10)
VECT_END : int(10)
TRIM
SEQ_ID : varchar(20)
TRIM : char(2)
LEFT_ST : int(10)
LEFT_END : int(10)
RIGHT_ST : int(10)
RIGHT_END : int(10)

4.1.2 Database description: soybean protein data

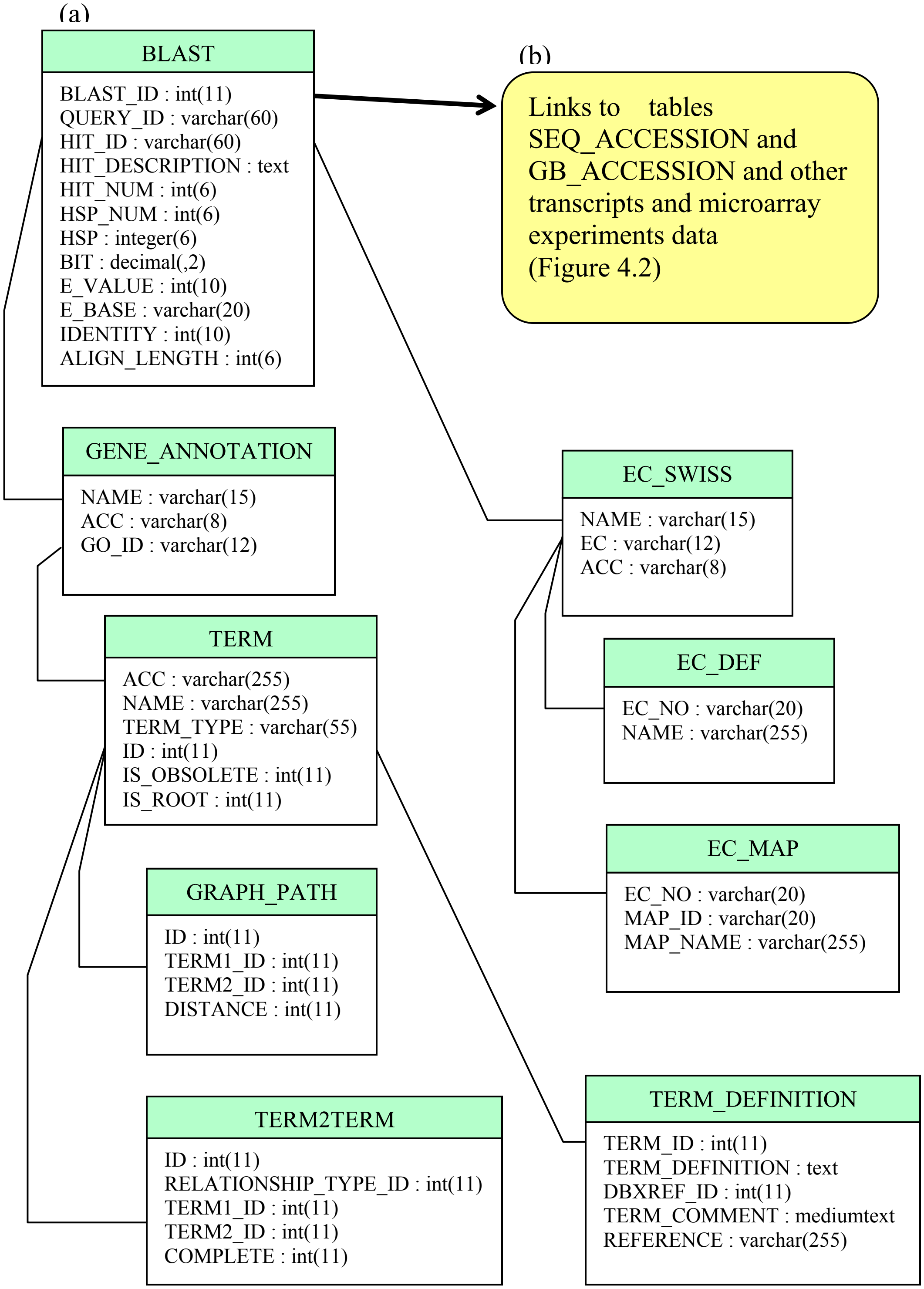
The BLASTX analysis against SwissProt allowed us to assign protein annotations to 175,910 ESTs (over half of the 318,422 EST sequences). **Figure 4.3** shows the annotation section of the database. The BLAST table contains the BLASTX search results and links our EST data to their corresponding protein information. Of the 37,637 soybean probe sequences on the Affymetrix GeneChip, we assigned protein annotations to 8,667 sequences. These BLASTX search results are also incorporated into the BLAST table and link to other protein and functional annotations. The SwissProt protein names are stored as the hit IDs. Other information about the proteins such as the protein descriptions, hit scores and e-values are also stored in the BLAST table. The SwissProt protein IDs link to other functional annotations such as gene ontology (GO terms) and KEGG molecular pathways through the GENE_ANNOTATION and EC_SWISS tables. The protein descriptions that describe the enzymes with appropriate EC (enzyme commission) numbers are linked to the KEGG pathways (stored as tables EC_DEF, and EC_MAP) through EC_SWISS table. There are 73,996 EST sequences assigned with EC numbers, around 23% of the EST sequences were enzymes. By linking the transcript sequences data to protein SwissProt annotations through BLASTX search result in the BLAST table, we can map the transcript sequences to their corresponding functional annotations such as GO terms and KEGG molecular pathways providing a more comprehensive description of the soybean data.

Figure 4.3. Database structure of the section for protein and functional annotations.

a) Table for BLASTX search results (BLAST); tables for gene ontology terms information (GENE_ANNOTATION, GRAPH_PATH, TERM, TERM2TERM, TERM_DEFINITION) and tables for KEGG pathways with the enzyme commission numbers (EC_DEF, EC_SWISS, EC_MAP)

b) BLAST table links protein annotation data to transcript sequence information and hence links to microarray experiments data.

Figure 4.3


(a)
BLAST
BLAST_ID : int(11)
QUERY_ID : varchar(60)
HIT_ID : varchar(60)
HIT_DESCRIPTION : text
HIT_NUM : int(6)
HSP_NUM : int(6)
HSP : integer(6)
BIT : decimal(,2)
E_VALUE : int(10)
E_BASE : varchar(20)
IDENTITY : int(10)
ALIGN_LENGTH : int(6)
(b)
Links to tables SEQ_ACCESSION and GB_ACCESSION and other transcripts and microarray experiments data (Figure 4.2)
GENE_ANNOTATION
NAME : varchar(15)
ACC : varchar(8)
GO_ID : varchar(12)
EC_SWISS
NAME : varchar(15)
EC : varchar(12)
ACC : varchar(8)
TERM
ACC : varchar(255)
NAME : varchar(255)
TERM_TYPE : varchar(55)
ID : int(11)
IS_OBSOLETE : int(11)
IS_ROOT : int(11)
EC_DEF
EC_NO : varchar(20)
NAME : varchar(255)
EC_MAP
EC_NO : varchar(20)
MAP_ID : varchar(20)
MAP_NAME : varchar(255)
GRAPH_PATH
ID : int(11)
TERM1_ID : int(11)
TERM2_ID : int(11)
DISTANCE : int(11)
TERM_DEFINITION
TERM_ID : int(11)
TERM_DEFINITION : text
DBXREF_ID : int(11)
TERM_COMMENT : mediumtext
REFERENCE : varchar(255)
TERM2TERM
ID : int(11)
RELATIONSHIP_TYPE_ID : int(11)
TERM1_ID : int(11)
TERM2_ID : int(11)
COMPLETE : int(11)

4.1.3 Database description: soybean microarray experiment data

The section of the database that organizes the microarray data is shown in **Figure 4.4.** Data for the Affymetrix Soybean GeneChip, for example the probe IDs, the sequences of the probes, and the locations of the probes on the chip are stored in the table CDF_FILE. The whole transcript sequences representing the genes with the corresponding probe IDs and GenBank accession number are stored in the table PROBE_SEQ. The PROBE_SET table contains the probe IDs, GenBank accession number, and the corresponding sequence and clone IDs to map to our soybean EST data and hence, associates the microarray data with their corresponding transcript, protein and functional annotations. Also, the microarray data can directly link to the BLAST table by using probe ID as the query ID to provide biological information for our microarray experiment.

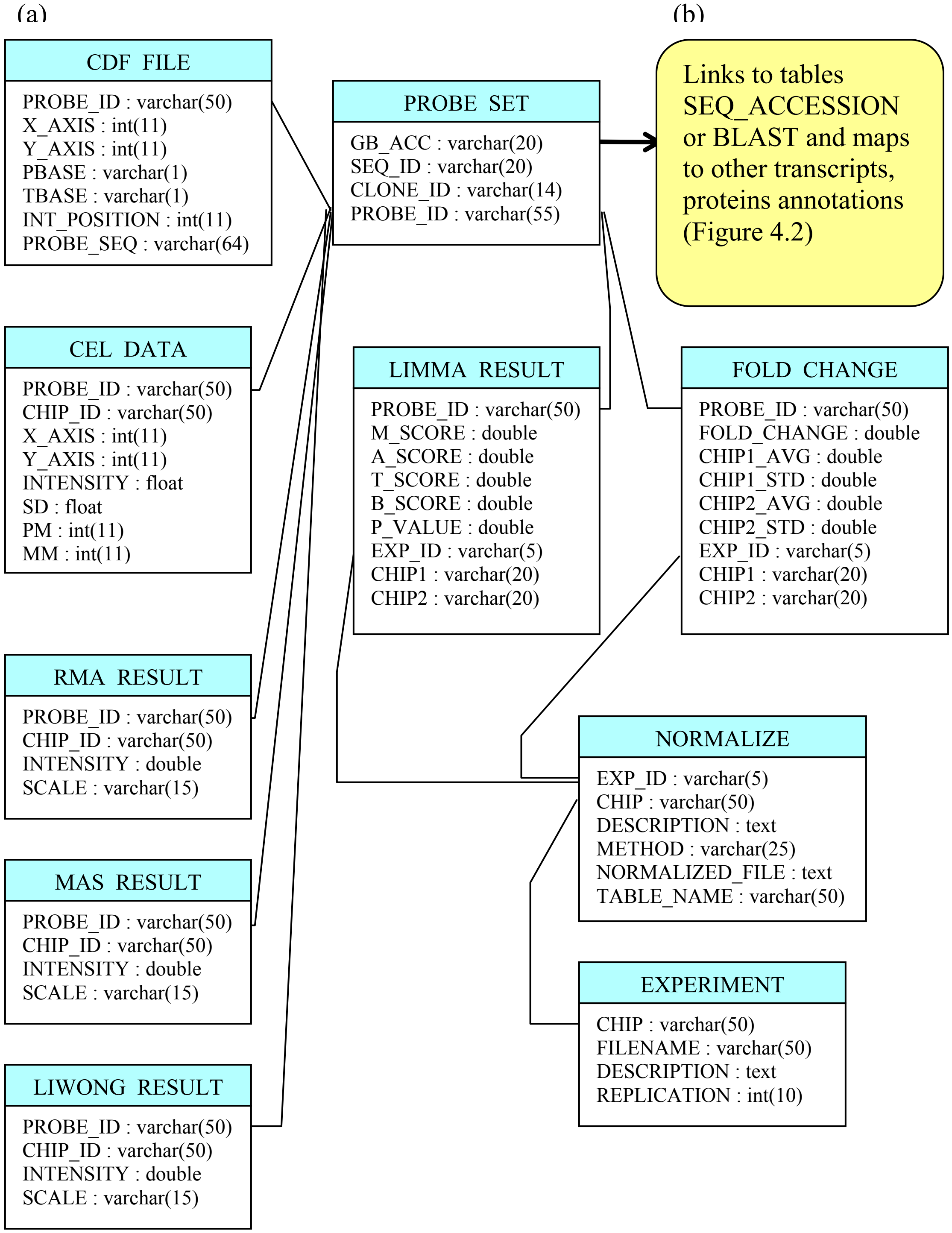
The raw data for our microarray experiment are stored in the table CEL_DATA, which contains the information for every chip, for example, the chip IDs, probe IDs, and the intensity of each probe. The processed data for our microarray experiment using three normalization methods RMA, MAS, dCHIP are stored in three tables RMA_RESULT, MAS_RESULT and LIWONG_RESULT respectively. All the raw and processed microarray data is linked to the PROBE_SET table by the probe IDs. For the analyzed results, the EXPERIMENT table describes which chips are used for the pair-wise comparison. The NORMALIZE table describes which normalization method are used in each pair-wise comparison. The microarray results for each pair-wise comparison analyzed by the LIMMA package are stored in the LIMMA_RESULT table. It includes the scores and p-value from the statistical test for each probe in all pair-wise comparisons. Also, the fold change and average intensity for each probe in all pair-wise comparisons are stored in the table FOLD_CHANGE. All the analyzed microarray results are linked to the PROBE_SET table and hence integrated with the soybean transcript, protein and functional annotations that can provide insight into biological and functional differences between samples.

Figure 4.4. Database structure of the section of the section for microarray experiment data.

a) Tables for chip information (CDF_FILE, PROBE_SET); table for raw data (CEL_DATA), tables for normalized data (LIWONG_RESULT, MAS_RESULT, RMA_RESULT); and tables for analyzed results (EXPERIMENT, FOLD_CHANGE, LIMA_RESULT, NORMALIZE)

b) PROBE_SET table links microarray data to transcript sequence information and protein annotations data.

Figure 4.4



CDF FILE

PROBE_ID : varchar(50)
X_AXIS : int(11)
Y_AXIS : int(11)
PBASE : varchar(1)
TBASE : varchar(1)
INT_POSITION : int(11)
PROBE_SEQ : varchar(64)

PROBE SET

GB_ACC : varchar(20)
SEQ_ID : varchar(20)
CLONE_ID : varchar(14)
PROBE_ID : varchar(55)

Links to tables SEQ_ACCESSION or BLAST and maps to other transcripts, proteins annotations (Figure 4.2)

CEL DATA

PROBE_ID : varchar(50)
CHIP_ID : varchar(50)
X_AXIS : int(11)
Y_AXIS : int(11)
INTENSITY : float
SD : float
PM : int(11)
MM : int(11)

LIMMA RESULT

PROBE_ID : varchar(50)
M_SCORE : double
A_SCORE : double
T_SCORE : double
B_SCORE : double
P_VALUE : double
EXP_ID : varchar(5)
CHIP1 : varchar(20)
CHIP2 : varchar(20)

FOLD CHANGE

PROBE_ID : varchar(50)
FOLD_CHANGE : double
CHIP1_AVG : double
CHIP1_STD : double
CHIP2_AVG : double
CHIP2_STD : double
EXP_ID : varchar(5)
CHIP1 : varchar(20)
CHIP2 : varchar(20)

RMA RESULT

PROBE_ID : varchar(50)
CHIP_ID : varchar(50)
INTENSITY : double
SCALE : varchar(15)

NORMALIZE

EXP_ID : varchar(5)
CHIP : varchar(50)
DESCRIPTION : text
METHOD : varchar(25)
NORMALIZED_FILE : text
TABLE_NAME : varchar(50)

MAS RESULT

PROBE_ID : varchar(50)
CHIP_ID : varchar(50)
INTENSITY : double
SCALE : varchar(15)

EXPERIMENT

CHIP : varchar(50)
FILENAME : varchar(50)
DESCRIPTION : text
REPLICATION : int(10)

LIWONG RESULT

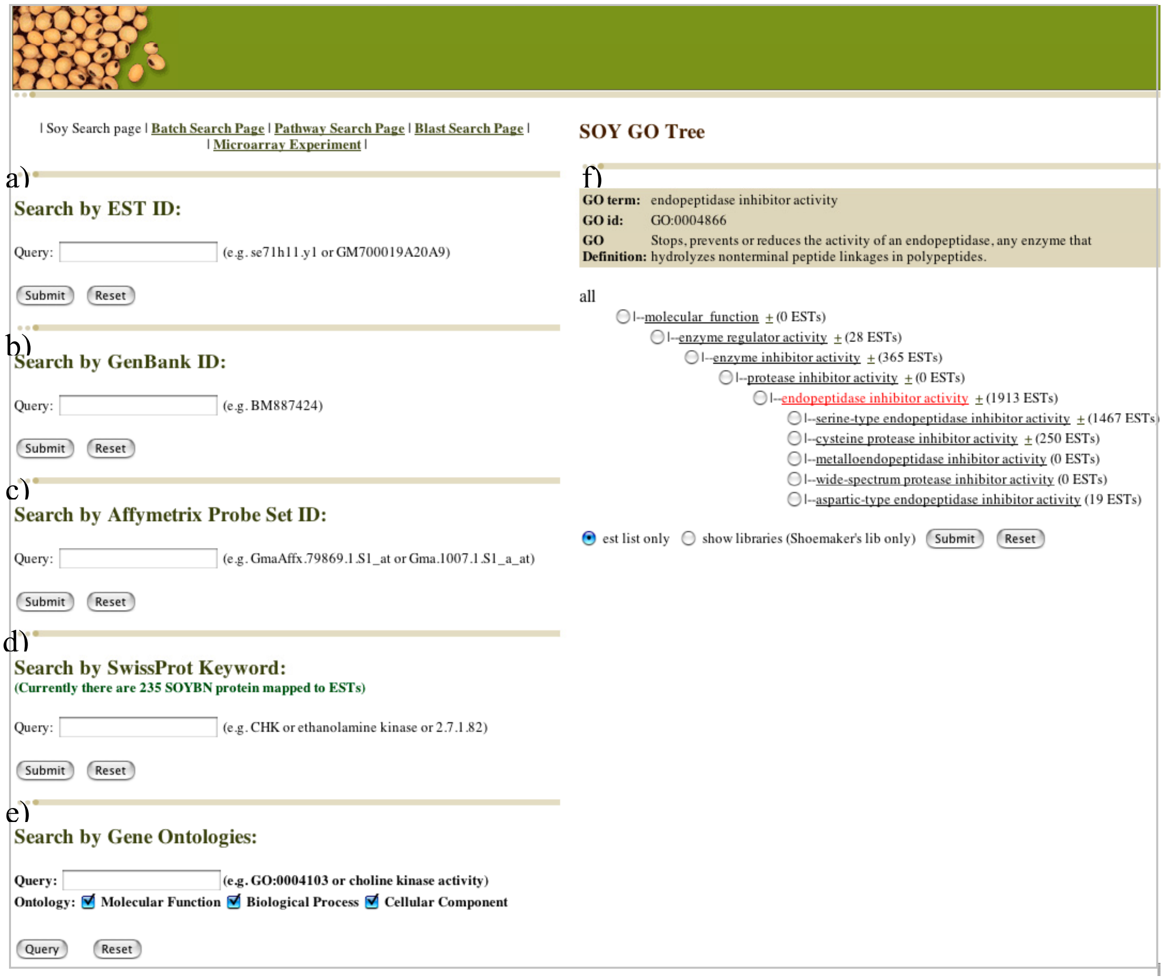
PROBE_ID : varchar(50)
CHIP_ID : varchar(50)
INTENSITY : double
SCALE : varchar(15)

4.1.4 Web interfaces allow navigation and querying of the database

Aside from using SQL (Structured Query Language) with a command line shell to retrieve data from the database, a couple of web interfaces were developed to access the database and display the data in a tabular format. **Figure 4.5** shows the SOY Search Page to retrieve all available IDs and annotations for a soybean transcript or a group of transcripts that share similar protein name or function from our database. Data can be retrieved by entering any EST ID, GenBank accession number, Affymetrix probe ID, SwissProt protein ID/name, EC enzyme number or GO term/number. A clickable GO tree that illustrates the hierarchy structure of the ontology is available to select a GO term for searching the associated IDs and annotation for the corresponding soybean sequences from the database.

Figure 4.6 shows the SOY Search Result page for displaying all available IDs and annotations for the query IDs from the SOY Search Page. After receiving the query ID, the corresponding EST sequence, Affymetrix probe sequence and TIGR TC contig will be retrieved from our database. The BLASTX results for the EST and the Affymetrix probe sequences, such as the SwissProt protein IDs and descriptions, BLAST scores and e-values, are displayed in the EST and AFFY tables. The associated GO numbers, GO terms and EC enzyme number are also displayed. The TC table displays the information for the TIGR contig, such as TC ID, the IDs and GenBank accession numbers for the EST that involved in the assembling of that contig, the associated GO number/term and EC enzyme number. All these IDs are hyperlinked to the original public databases, such as the Soybean Genomics Initiative website (http://soybean.ccgb.umn.edu/) for the information from the soybean EST projects, the GenBank (http://www.ncbi.nlm.nih.gov) for the sequence information, the SwissProt protein database (http://ca.expasy.org) for the protein information, the Gene Ontology (http://amigo.geneontology.org) for the functional annotations, the KEGG database (http://www.genome.jp) for the biological pathway maps and enzyme information and the TIGR Gene Index for the contig sequence and information (http://compbio.dfci.harvard.edu), to facilitate detailed database searches for the soybean search results.

Figure 4.5



| Soy Search page | **Batch Search Page** | **Pathway Search Page** | **Blast Search Page** |
| **Microarray Experiment** |

Search by EST ID:

Query: (e.g. se71h11.y1 or GM700019A20A9)

Submit Reset

Search by GenBank ID:

Query: (e.g. BM887424)

Submit Reset

Search by Affymetrix Probe Set ID:

Query: (e.g. GmaAffx.79869.1.S1_at or Gma.1007.1.S1_a_at)

Submit Reset

Search by SwissProt Keyword:
(Currently there are 235 SOYBN protein mapped to ESTs)

Query: (e.g. CHK or ethanolamine kinase or 2.7.1.82)

Submit Reset

Search by Gene Ontologies:

Query: **(e.g. GO:0004103 or choline kinase activity)**
Ontology: ☑ **Molecular Function** ☑ **Biological Process** ☑ **Cellular Component**

Query Reset

SOY GO Tree

GO term: endopeptidase inhibitor activity
GO id: GO:0004866
GO Definition: Stops, prevents or reduces the activity of an endopeptidase, any enzyme that hydrolyzes nonterminal peptide linkages in polypeptides.

```
all
    ○ |--molecular_function + (0 ESTs)
        ○ |--enzyme regulator activity + (28 ESTs)
            ○ |--enzyme inhibitor activity + (365 ESTs)
                ○ |--protease inhibitor activity + (0 ESTs)
                    ○ |--endopeptidase inhibitor activity + (1913 ESTs)
                        ○ |--serine-type endopeptidase inhibitor activity + (1467 ESTs)
                        ○ |--cysteine protease inhibitor activity + (250 ESTs)
                        ○ |--metalloendopeptidase inhibitor activity (0 ESTs)
                        ○ |--wide-spectrum protease inhibitor activity (0 ESTs)
                        ○ |--aspartic-type endopeptidase inhibitor activity (19 ESTs)
```

● est list only ○ show libraries (Shoemaker's lib only) Submit Reset

Figure 4.5. The main page of the soy database. (http://thor.agrenv.mcgill.ca/cgi-bin/soy/soybean.cgi) Users can submit queries to the database to retrieve all available IDs and annotations for the soybean transcript of interest. Queries can be made using:

a) EST ID
b) GenBank accession number
c) Affymetrix probe ID
d) SwissProt protein ID or name
e) GO number or term
f) a clickable GO tree to assist searching for a GO term from the gene ontology hierarchy structure.

Figure 4.6. Database result page, showing information about an Affymetrix probe ID: Gma.1137.1.S1_x_at.

Data are organized into three tables:

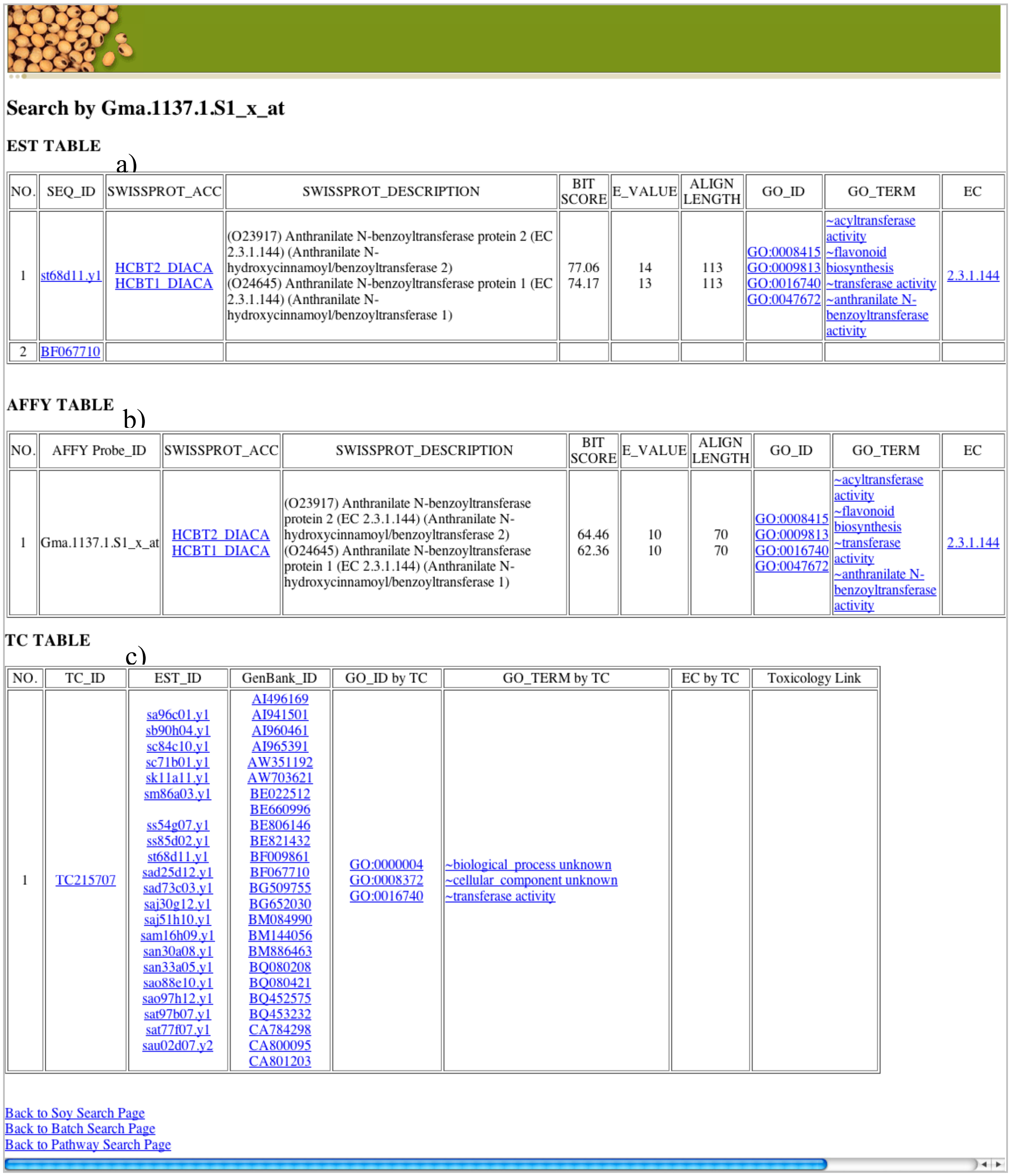
a) EST table displays the BLASTX result with scores and e-value for the EST sequence

b) AFFY table displays the BLASTX result with scores and e-value for the Affymetrix probe sequence

c) TC table displays the information for the corresponding TIGR contig

The SwissProt protein ID/description, GO number/terms and EC enzyme numbers are displayed to the corresponding EST ID, GenBank accession number or Affymetrix probe ID with hyperlinks to the original public database.

Figure 4.6



Search by Gma.1137.1.S1_x_at

EST TABLE

a)

NO.	SEQ_ID	SWISSPROT_ACC	SWISSPROT_DESCRIPTION	BIT SCORE	E_VALUE	ALIGN LENGTH	GO_ID	GO_TERM	EC
1	st68d11.y1	HCBT2_DIACA HCBT1_DIACA	(O23917) Anthranilate N-benzoyltransferase protein 2 (EC 2.3.1.144) (Anthranilate N-hydroxycinnamoyl/benzoyltransferase 2) (O24645) Anthranilate N-benzoyltransferase protein 1 (EC 2.3.1.144) (Anthranilate N-hydroxycinnamoyl/benzoyltransferase 1)	77.06 74.17	14 13	113 113	GO:0008415 GO:0009813 GO:0016740 GO:0047672	~acyltransferase activity ~flavonoid biosynthesis ~transferase activity ~anthranilate N-benzoyltransferase activity	2.3.1.144
2	BF067710								

AFFY TABLE

b)

NO.	AFFY Probe_ID	SWISSPROT_ACC	SWISSPROT_DESCRIPTION	BIT SCORE	E_VALUE	ALIGN LENGTH	GO_ID	GO_TERM	EC
1	Gma.1137.1.S1_x_at	HCBT2_DIACA HCBT1_DIACA	(O23917) Anthranilate N-benzoyltransferase protein 2 (EC 2.3.1.144) (Anthranilate N-hydroxycinnamoyl/benzoyltransferase 2) (O24645) Anthranilate N-benzoyltransferase protein 1 (EC 2.3.1.144) (Anthranilate N-hydroxycinnamoyl/benzoyltransferase 1)	64.46 62.36	10 10	70 70	GO:0008415 GO:0009813 GO:0016740 GO:0047672	~acyltransferase activity ~flavonoid biosynthesis ~transferase activity ~anthranilate N-benzoyltransferase activity	2.3.1.144

TC TABLE

c)

NO.	TC_ID	EST_ID	GenBank_ID	GO_ID by TC	GO_TERM by TC	EC by TC	Toxicology Link
1	TC215707	sa96c01.y1 sb90h04.y1 sc84c10.y1 sc71b01.y1 sk11a11.y1 sm86a03.y1 ss54g07.y1 ss85d02.y1 st68d11.y1 sad25d12.y1 sad73c03.y1 saj30g12.y1 saj51h10.y1 sam16h09.y1 san30a08.y1 san33a05.y1 sao88e10.y1 sao97h12.y1 sat97b07.y1 sat77f07.y1 sau02d07.y2	AI496169 AI941501 AI960461 AI965391 AW351192 AW703621 BE022512 BE660996 BE806146 BE821432 BF009861 BF067710 BG509755 BG652030 BM084990 BM144056 BM886463 BQ080208 BQ080421 BQ452575 BQ453232 CA784298 CA800095 CA801203	GO:0000004 GO:0008372 GO:0016740	~biological_process unknown ~cellular_component unknown ~transferase activity		

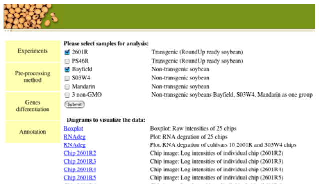
Back to Soy Search Page
Back to Batch Search Page
Back to Pathway Search Page

The results for the microarray experiment can be retrieved from the database through a special section of the interfaces. An overview of the query flow is presented in **Figure 4.7**. **Figure 4.8a** shows the SOY Microarray Analysis webpage where samples (any of the five soybean cultivars) can be selected for pair-wise comparison. Diagrams to assess the quality of the data, such as boxplot of the intensities of the chips, RNA degradation plot and the individual chip image are visualized. After selecting two samples, the web page allows a choice of normalization method for pre-processing the raw data as shown in **Figure 4.8b**. Diagrams such as boxplot, PCA analysis and hierarchical clustering are available to visualize the pre-processed data. After selecting the pre-processing method, the webpage allows selecting the cut-off p-value and fold change for differentially expressed genes from the results of the statistical analysis (**Figure 4.8c**). The list of differentially expressed genes for the pair-wise comparison is displayed by the probe IDs (**Figure 4.8d**). Statistical scores such as t-score, p-value and fold change are also displayed. A hyperlink is provided to display a plot of the intensities of an individual probe against five soybean cultivars. Check boxes are also available to submit a list of probe IDs to the Soybean Search Page to retrieve all the available IDs and annotations for those probes. It links to the annotation view by clicking on the Annotated Probe List button on the left panel. The annotation view (**Figure 4.8e**) displays the associated SwissProt protein ID/description and the GO number/term with the fold change and p-value for the list of differentially expressed genes. All these IDs are hyperlinked to the original public databases to facilitate detailed database searches. To retrieve results from the gene class analysis based on GO term annotations, similar query pages are developed (**Figure 4.7**). The list of GO terms (represents changes of the gene class) is displayed in the result pages (**Figure 4.8f**) with the statistical scores and the number of the genes involved in each gene class. The intensities of the individual genes of each identified GO terms can be displayed with log2 fold change (**Figure 4.8g**), which can allow users to identify whether the genes were regulated in a similar pattern.

Figure 4.7. Flowchart of the microarray web interface to access the database for displaying differentially expressed genes or functional gene class based on GO term.

Figure 4.7

A/ Build the query



Please select samples for analysis:

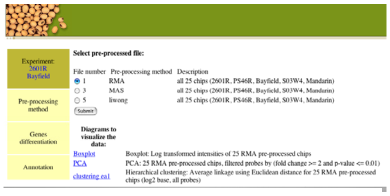
Experiments	☑ 2601R	Transgenic (RoundUp ready soybean)
	□ PS46R	Transgenic (RoundUp ready soybean)
Pre-processing method	☑ Bayfield	Non-transgenic soybean
	□ S03W4	Non-transgenic soybean
	□ Mandarin	Non-transgenic soybean
Genes differentiation	□ 3 non-GMO	Non-transgenic soybeans Bayfield, S03W4, Mandarin as one group
Annotation		

Submit

Diagrams to visualize the data:

Boxplot	Boxplot: Raw intensities of 25 chips
RNAdeg	Plot: RNA degration of 25 chips
RNAdeg	Plot: RNA degration of cultivars 10 2601R and S03W4 chips
Chip 2601R2	Chip image: Log intensities of individual chip (2601R2)
Chip 2601R3	Chip image: Log intensities of individual chip (2601R3)
Chip 2601R4	Chip image: Log intensities of individual chip (2601R4)
Chip 2601R5	Chip image: Log intensities of individual chip (2601R5)

Step 1: select biological samples (Figure 4.8a)



Experiment: 2601R Bayfield

Pre-processing method

Genes differentiation

Annotation

Select pre-processed file:

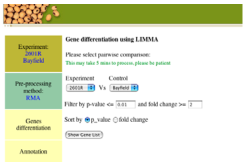
File number	Pre-processing method	Description
● 1	RMA	all 25 chips (2601R, PS46R, Bayfield, S03W4, Mandarin)
○ 3	MAS	all 25 chips (2601R, PS46R, Bayfield, S03W4, Mandarin)
○ 5	liwong	all 25 chips (2601R, PS46R, Bayfield, S03W4, Mandarin)

Submit

Diagrams to visualize the data:

Boxplot	Boxplot: Log transformed intensities of 25 RMA pre-processed chips
PCA	PCA: 25 RMA pre-processed chips, filtered probes by (fold change >= 2 and p-value <= 0.01)
clustering ea1	Hierarchical clustering: Average linkage using Euclidean distance for 25 RMA pre-processed chips (log2 base, all probes)

Step 2: select preprocessing method (Figure 4.8b)



Experiment: 2601R Bayfield

Pre-processing method: RMA

Genes differentiation

Annotation

Gene differentiation using LIMMA

Please select pairwise comparison:
This may take 5 mins to process, please be patient

Experiment: 2601R Vs Control: Bayfield

Filter by p-value <= 0.01 and fold change >= 2

Sort by ● p_value ○ fold change

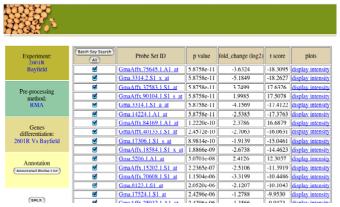
Show Gene List

Step 3: select cut-off values for statistical analysis (Figure 4.8c)

Probe level analysis

Functional level analysis

B/ Display result view



Experiment: 2601R Bayfield

Pre-processing method: RMA

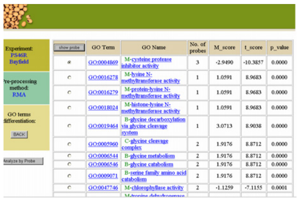
Genes differentiation: 2601R Vs Bayfield

Annotation

Annotated Probe List

BACK

Batch Soy Search All	Probe Set ID	p value	fold_change (log2)	t score	plots
☑	GmaAffx.75645.1.A1_at	5.8758e-11	-3.6324	-18.3095	display intensity
☑	Gma.3314.2.S1_x_at	5.8758e-11	-5.1849	-18.2627	display intensity
☑	GmaAffx.37583.1.S1_at	5.8758e-11	3.7499	17.6326	display intensity
☑	GmaAffx.90104.1.S1_s_at	5.8758e-11	1.9985	17.5078	display intensity
☑	Gma.3314.1.S1_a_at	5.8758e-11	-4.1569	-17.4122	display intensity
☑	Gma.14224.1.A1_at	5.8758e-11	-2.5385	-17.3763	display intensity
☑	GmaAffx.84169.1.A1_at	1.2220e-10	2.3786	16.6879	display intensity
☑	GmaAffx.40133.1.S1_at	2.4572e-10	-2.7063	-16.0631	display intensity
☑	Gma.17306.1.S1_s_at	8.9814e-10	-1.9139	-15.0461	display intensity
☑	GmaAffx.18584.1.S1_x_at	1.8866e-09	-2.6738	-14.4623	display intensity
☑	Gma.5206.1.A1_at	5.0701e-08	2.4126	12.3057	display intensity
☑	GmaAffx.15202.1.S1_at	2.2365e-07	-2.5106	-11.3919	display intensity
☑	GmaAffx.70608.1.S1_at	1.1504e-06	-3.3199	-10.4486	display intensity
☑	Gma.8123.1.S1_at	2.0520e-06	-2.1207	-10.1043	display intensity
☑	Gma.17524.1.S1_at	2.4296e-06	-1.2788	-9.9530	display intensity



Experiment: PS46R Bayfield

Pre-processing method: RMA

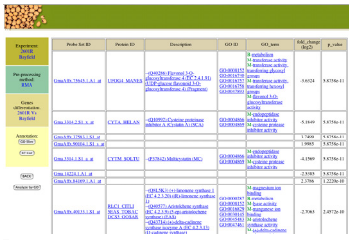
GO terms differentiation:

BACK

Analyze by Probe

show probe	GO Term	GO Name	No. of probes	M_score	t_score	p_value
●	GO:0004869	M-cysteine protease inhibitor activity	3	-2.9490	-10.3857	0.0000
○	GO:0016278	M-lysine N-methyltransferase activity	1	1.0591	8.9683	0.0000
○	GO:0016279	M-protein-lysine N-methyltransferase activity	1	1.0591	8.9683	0.0000
○	GO:0018024	M-histone-lysine N-methyltransferase activity	1	1.0591	8.9683	0.0000
○	GO:0019464	B-glycine decarboxylation via glycine cleavage system	1	3.0713	8.9038	0.0000
○	GO:0005960	C-glycine cleavage complex	2	1.9176	8.8712	0.0000
○	GO:0006544	B-glycine metabolism	2	1.9176	8.8712	0.0000
○	GO:0006546	B-glycine catabolism	2	1.9176	8.8712	0.0000
○	GO:0009071	B-serine family amino acid catabolism	2	1.9176	8.8712	0.0000
○	GO:0047746	M-chlorophyllase activity	2	-1.1259	-7.1155	0.0001

Step 4: display gene list or GO list (Figure 4.8d, Figure 4.8f)



Experiment: 2601R Bayfield

Pre-processing method: RMA

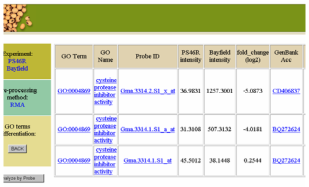
Genes differentiation: 2601R Vs Bayfield

Annotation: GO Slim EC List

BACK

Analyze by GO

Probe Set ID	Protein ID	Description	GO ID	GO_term	fold_change (log2)	p_value
GmaAffx.75645.1.A1_at	UFOG4_MANES	--(Q40286) Flavonol 3-O-glucosyltransferase 4 (EC 2.4.1.91) (UDP-glucose flavonoid 3-O-glucosyltransferase 4) (Fragment)	GO:0008152 GO:0016740 GO:0016757 GO:0016758 GO:0047893	B-metabolism M-transferase activity M-transferase activity, transferring glycosyl groups M-transferase activity, transferring hexosyl groups M-flavonol 3-O-glucosyltransferase activity	-3.6324	5.8758e-11
Gma.3314.2.S1_x_at	CYTA_HELAN	--(Q10992) Cysteine proteinase inhibitor A (Cystatin A) (SCA)	GO:0004866 GO:0004869	M-endopeptidase inhibitor activity M-cysteine protease inhibitor activity	-5.1849	5.8758e-11
GmaAffx.37583.1.S1_at					3.7499	5.8758e-11
GmaAffx.90104.1.S1_s_at					1.9985	5.8758e-11
Gma.3314.1.S1_a_at	CYTM_SOLTU	--(P37842) Multicystatin (MC)	GO:0004866 GO:0004869	M-endopeptidase inhibitor activity M-cysteine protease inhibitor activity	-4.1569	5.8758e-11
Gma.14224.1.A1_at					-2.5385	5.8758e-11
GmaAffx.84169.1.A1_at					2.3786	1.2220e-10
GmaAffx.40133.1.S1_at	RLC1_CITLI 5EAS_TOBAC DCS3_GOSAR	--(Q8L5K3) (+)-limonene synthase 1 (EC 4.2.3.20) ((R)-limonene synthase 1) --(Q40577) Aristolchene synthase (EC 4.2.3.9) (5-epi-aristolochene synthase) (EAS) --(Q43714) (+)-delta-cadinene synthase isozyme A (EC 4.2.3.13) (D-cadinene synthase)	GO:0000287 GO:0008152 GO:0016829 GO:0030145 GO:0045483 GO:0047461	M-magnesium ion binding B-metabolism M-lyase activity M-manganese ion binding M-aristolochene synthase activity M-(+)-delta-cadinene	-2.7063	2.4572e-10



Experiment: PS46R Bayfield

Pre-processing method: RMA

GO terms differentiation:

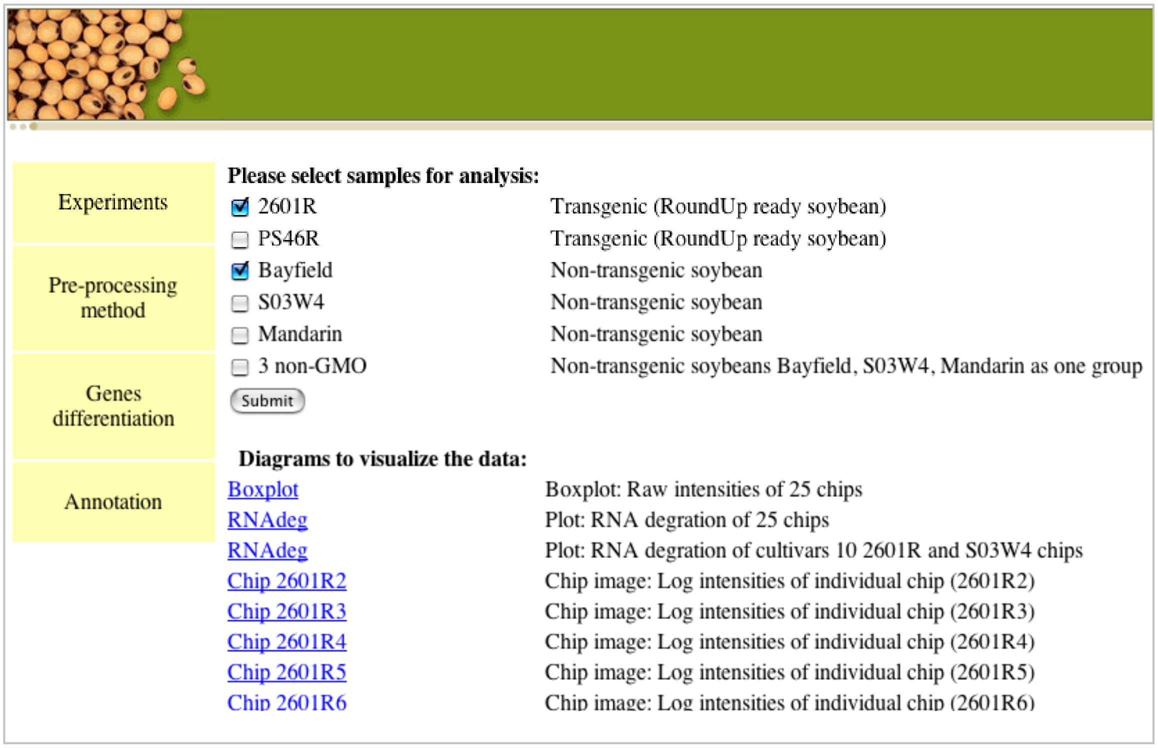
BACK

Analyze by Probe

GO Term	GO Name	Probe ID	PS46R intensity	Bayfield intensity	fold_change (log2)	GenBank Acc
GO:0004869	cysteine protease inhibitor activity	Gma.3314.2.S1_x_at	36.9831	1257.3001	-5.0873	CD406837
GO:0004869	cysteine protease inhibitor activity	Gma.3314.1.S1_a_at	31.3108	507.3132	-4.0181	BQ272624
GO:0004869	cysteine protease inhibitor activity	Gma.3314.1.S1_at	45.5012	38.1448	0.2544	BQ272624

Step 5: display annotations for GO list or individual genes for one GO term (Figure 4.8e, Figure 4.8h)

Figure 4.8a



Please select samples for analysis:

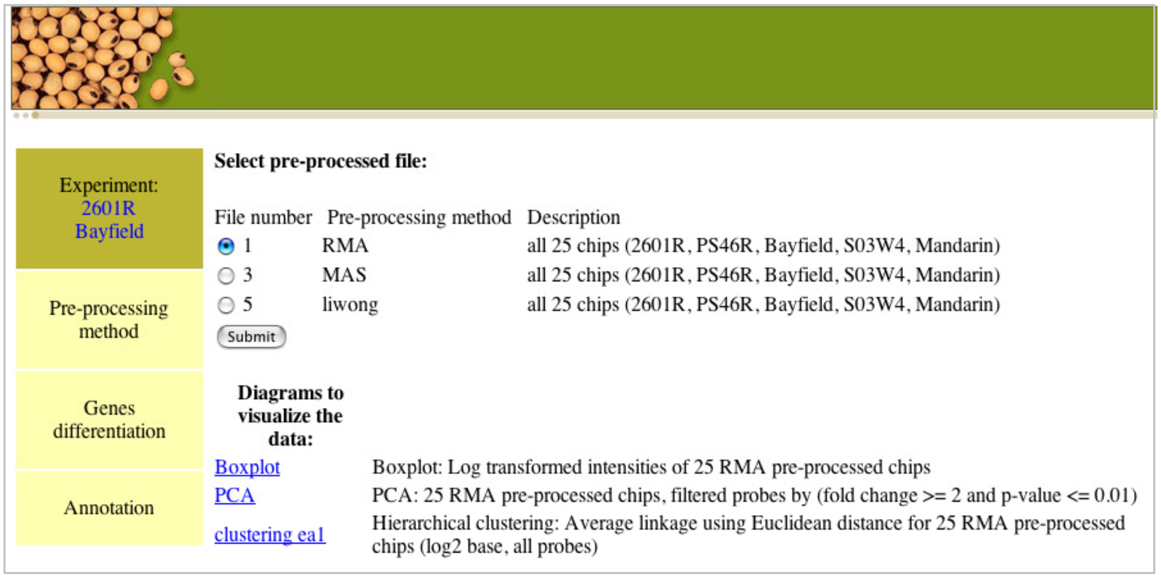
☑ 2601R	Transgenic (RoundUp ready soybean)
□ PS46R	Transgenic (RoundUp ready soybean)
☑ Bayfield	Non-transgenic soybean
□ S03W4	Non-transgenic soybean
□ Mandarin	Non-transgenic soybean
□ 3 non-GMO	Non-transgenic soybeans Bayfield, S03W4, Mandarin as one group

Diagrams to visualize the data:

Boxplot	Boxplot: Raw intensities of 25 chips
RNAdeg	Plot: RNA degration of 25 chips
RNAdeg	Plot: RNA degration of cultivars 10 2601R and S03W4 chips
Chip 2601R2	Chip image: Log intensities of individual chip (2601R2)
Chip 2601R3	Chip image: Log intensities of individual chip (2601R3)
Chip 2601R4	Chip image: Log intensities of individual chip (2601R4)
Chip 2601R5	Chip image: Log intensities of individual chip (2601R5)
Chip 2601R6	Chip image: Log intensities of individual chip (2601R6)

Figure 4.8a. Step 1 of the microarray analysis webpage: to select samples to compare. The first view is used to select samples for pair-wise microarray experiment and to display links to diagrams for visualization the raw data.

Figure 4.8b



Select pre-processed file:

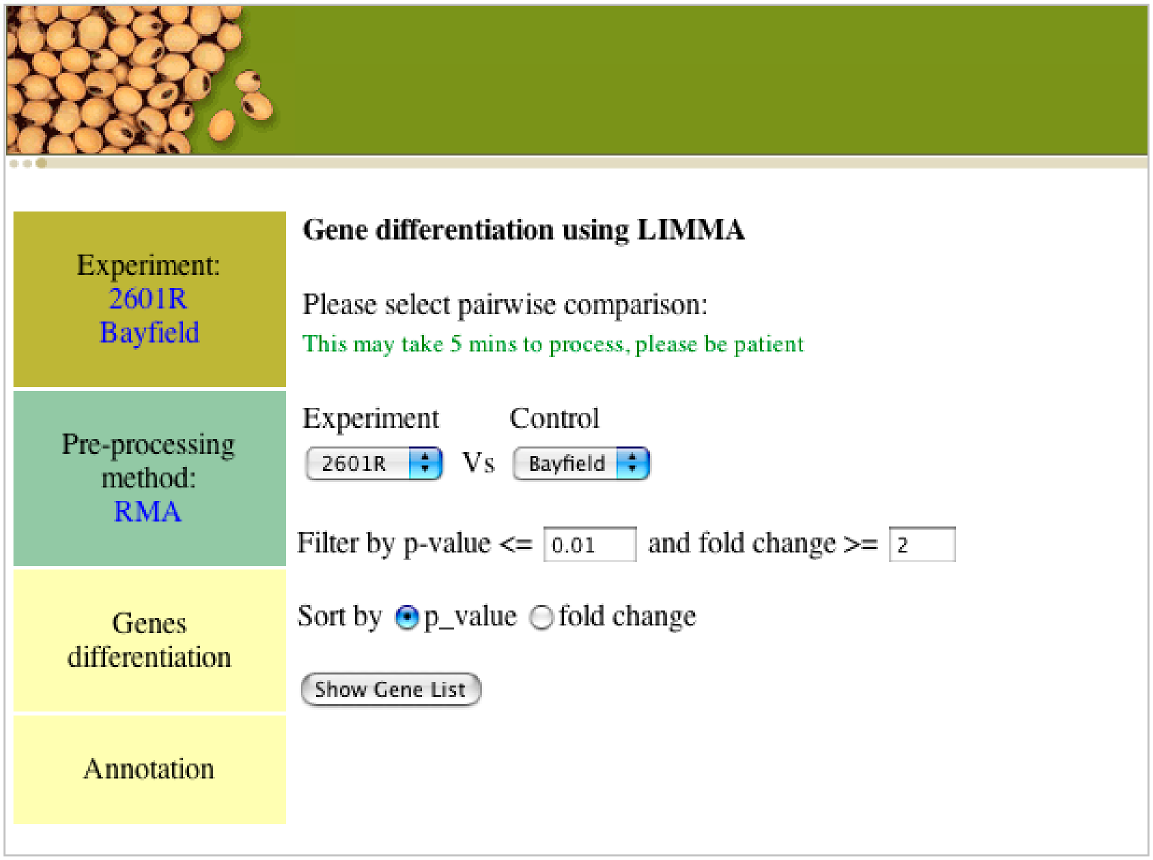
File number	Pre-processing method	Description
● 1	RMA	all 25 chips (2601R, PS46R, Bayfield, S03W4, Mandarin)
○ 3	MAS	all 25 chips (2601R, PS46R, Bayfield, S03W4, Mandarin)
○ 5	liwong	all 25 chips (2601R, PS46R, Bayfield, S03W4, Mandarin)

Diagrams to visualize the data:

Boxplot	Boxplot: Log transformed intensities of 25 RMA pre-processed chips
PCA	PCA: 25 RMA pre-processed chips, filtered probes by (fold change >= 2 and p-value <= 0.01)
clustering ea1	Hierarchical clustering: Average linkage using Euclidean distance for 25 RMA pre-processed chips (log2 base, all probes)

Figure 4.8b. Step 2 of the microarray analysis webpage: to select pre-processing method The second view is used to select pre-processing method for the microarray data and to display links to diagrams for visualization the pre-processed data.

Figure 4.8c



Experiment:
2601R
Bayfield

Pre-processing method:
RMA

Genes differentiation

Annotation

Gene differentiation using LIMMA

Please select pairwise comparison:
This may take 5 mins to process, please be patient

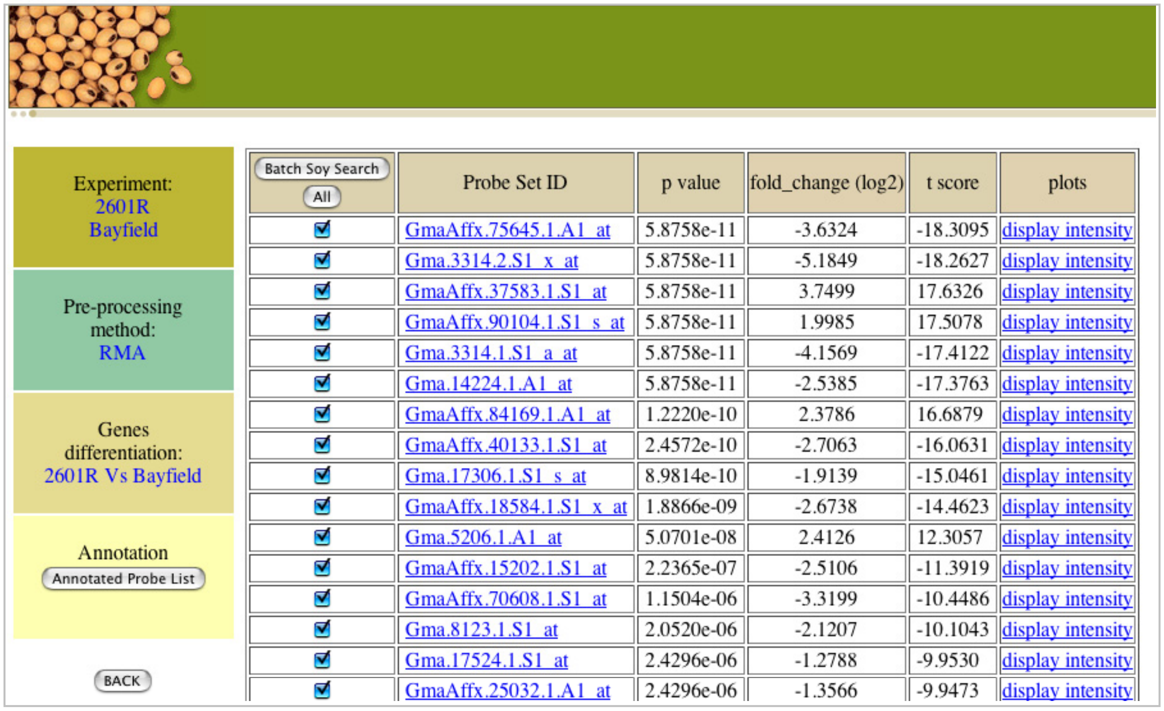
Experiment Control
2601R Vs Bayfield

Filter by p-value <= 0.01 and fold change >= 2

Sort by ● p_value ○ fold change

Figure 4.8c. Step 3 of the microarray analysis webpage: to define cut-off p-value and fold change.
The third view is used to select cut-off p-value and fold change for displaying differentially expressed genes using LIMMA package.

Figure 4.8d



Experiment:
2601R
Bayfield

Pre-processing method:
RMA

Genes differentiation:
2601R Vs Bayfield

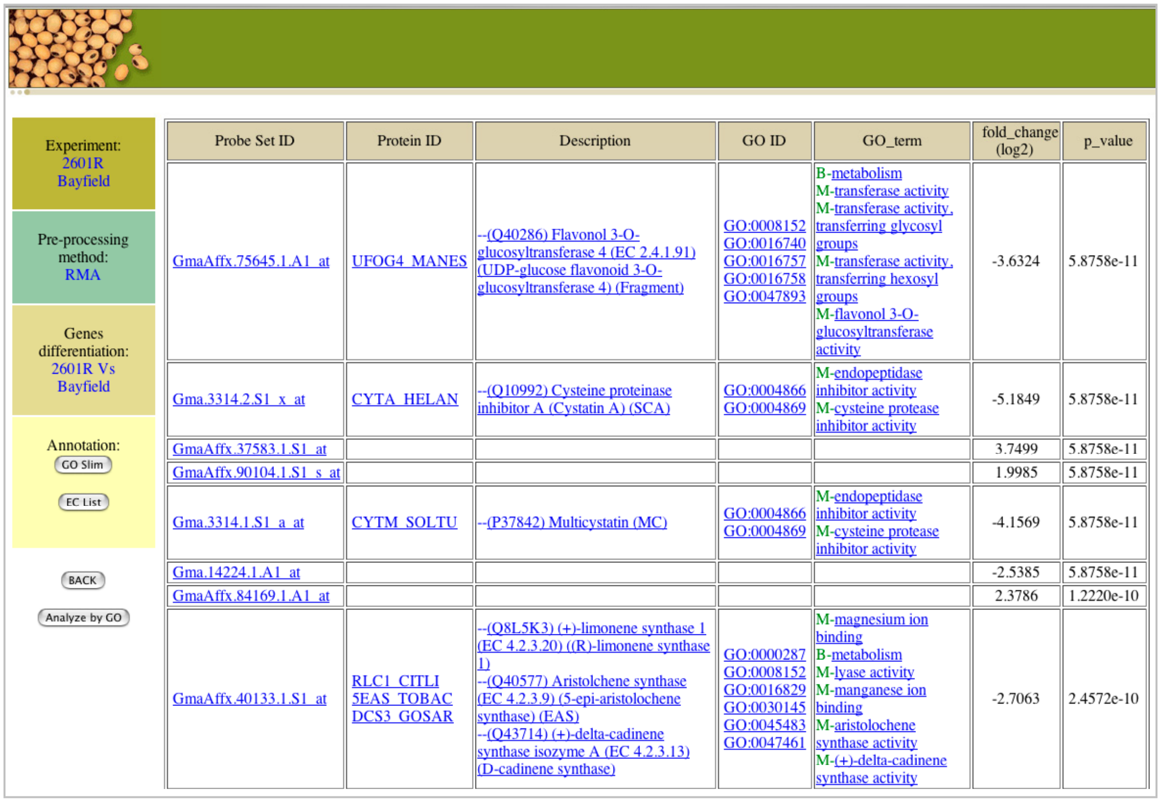
Annotation
Annotated Probe List

BACK

Batch Soy Search All	Probe Set ID	p value	fold_change (log2)	t score	plots
☑	GmaAffx.75645.1.A1_at	5.8758e-11	-3.6324	-18.3095	display intensity
☑	Gma.3314.2.S1_x_at	5.8758e-11	-5.1849	-18.2627	display intensity
☑	GmaAffx.37583.1.S1_at	5.8758e-11	3.7499	17.6326	display intensity
☑	GmaAffx.90104.1.S1_s_at	5.8758e-11	1.9985	17.5078	display intensity
☑	Gma.3314.1.S1_a_at	5.8758e-11	-4.1569	-17.4122	display intensity
☑	Gma.14224.1.A1_at	5.8758e-11	-2.5385	-17.3763	display intensity
☑	GmaAffx.84169.1.A1_at	1.2220e-10	2.3786	16.6879	display intensity
☑	GmaAffx.40133.1.S1_at	2.4572e-10	-2.7063	-16.0631	display intensity
☑	Gma.17306.1.S1_s_at	8.9814e-10	-1.9139	-15.0461	display intensity
☑	GmaAffx.18584.1.S1_x_at	1.8866e-09	-2.6738	-14.4623	display intensity
☑	Gma.5206.1.A1_at	5.0701e-08	2.4126	12.3057	display intensity
☑	GmaAffx.15202.1.S1_at	2.2365e-07	-2.5106	-11.3919	display intensity
☑	GmaAffx.70608.1.S1_at	1.1504e-06	-3.3199	-10.4486	display intensity
☑	Gma.8123.1.S1_at	2.0520e-06	-2.1207	-10.1043	display intensity
☑	Gma.17524.1.S1_at	2.4296e-06	-1.2788	-9.9530	display intensity
☑	GmaAffx.25032.1.A1_at	2.4296e-06	-1.3566	-9.9473	display intensity

Figure 4.8d. Step 4 of the microarray analysis webpage: to display the list of differentially expressed genes.
The fourth view is used to display the list of differentially expressed genes based on individual probes ID. The p-value, fold change and t-score are displayed. A link to a plot to illustrate the intensities of an individual probe in five soybean cultivars is also provided. The check boxes submit the list of probe IDs to the SOY Search Page to display available IDs and annotations.

Figure 4.8e



Experiment:
2601R
Bayfield

Pre-processing method:
RMA

Genes differentiation:
2601R Vs
Bayfield

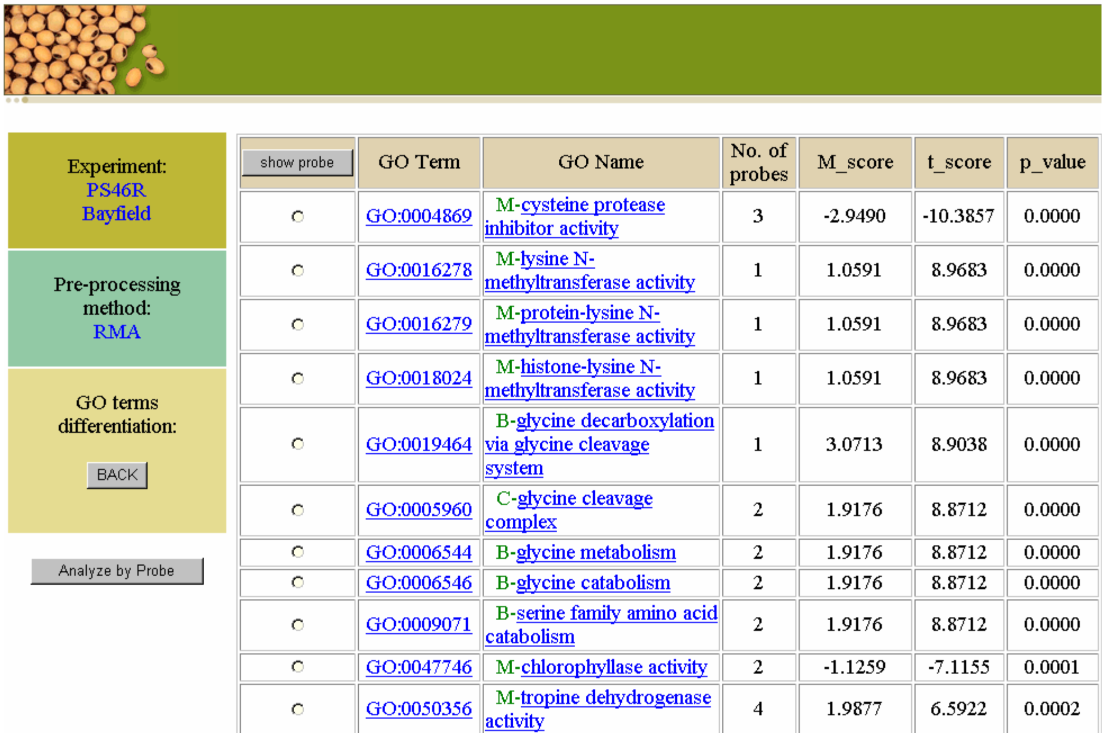
Annotation:

Probe Set ID	Protein ID	Description	GO ID	GO_term	fold_change (log2)	p_value
GmaAffx.75645.1.A1_at	UFOG4_MANES	--(Q40286) Flavonol 3-O-glucosyltransferase 4 (EC 2.4.1.91) (UDP-glucose flavonoid 3-O-glucosyltransferase 4) (Fragment)	GO:0008152 GO:0016740 GO:0016757 GO:0016758 GO:0047893	B-metabolism M-transferase activity M-transferase activity, transferring glycosyl groups M-transferase activity, transferring hexosyl groups M-flavonol 3-O-glucosyltransferase activity	-3.6324	5.8758e-11
Gma.3314.2.S1_x_at	CYTA_HELAN	--(Q10992) Cysteine proteinase inhibitor A (Cystatin A) (SCA)	GO:0004866 GO:0004869	M-endopeptidase inhibitor activity M-cysteine protease inhibitor activity	-5.1849	5.8758e-11
GmaAffx.37583.1.S1_at					3.7499	5.8758e-11
GmaAffx.90104.1.S1_s_at					1.9985	5.8758e-11
Gma.3314.1.S1_a_at	CYTM_SOLTU	--(P37842) Multicystatin (MC)	GO:0004866 GO:0004869	M-endopeptidase inhibitor activity M-cysteine protease inhibitor activity	-4.1569	5.8758e-11
Gma.14224.1.A1_at					-2.5385	5.8758e-11
GmaAffx.84169.1.A1_at					2.3786	1.2220e-10
GmaAffx.40133.1.S1_at	RLC1_CITLI 5EAS_TOBAC DCS3_GOSAR	--(Q8L5K3) (+)-limonene synthase 1 (EC 4.2.3.20) ((R)-limonene synthase 1) --(Q40577) Aristolchene synthase (EC 4.2.3.9) (5-epi-aristolochene synthase) (EAS) --(Q43714) (+)-delta-cadinene synthase isozyme A (EC 4.2.3.13) (D-cadinene synthase)	GO:0000287 GO:0008152 GO:0016829 GO:0030145 GO:0045483 GO:0047461	M-magnesium ion binding B-metabolism M-lyase activity M-manganese ion binding M-aristolochene synthase activity M-(+)-delta-cadinene synthase activity	-2.7063	2.4572e-10

Figure 4.8e. Step 5 of the microarray analysis webpage: to display the result list with annotations.

The fifth view is used to display the annotation view of the list of differentially expressed genes. The SwissProt protein ID/description, GO number/term, fold change and p-value for the probes are displayed. The protein and GO IDs are hyperlinked to the original public databases.

Figure 4.8f



Experiment:
PS46R
Bayfield

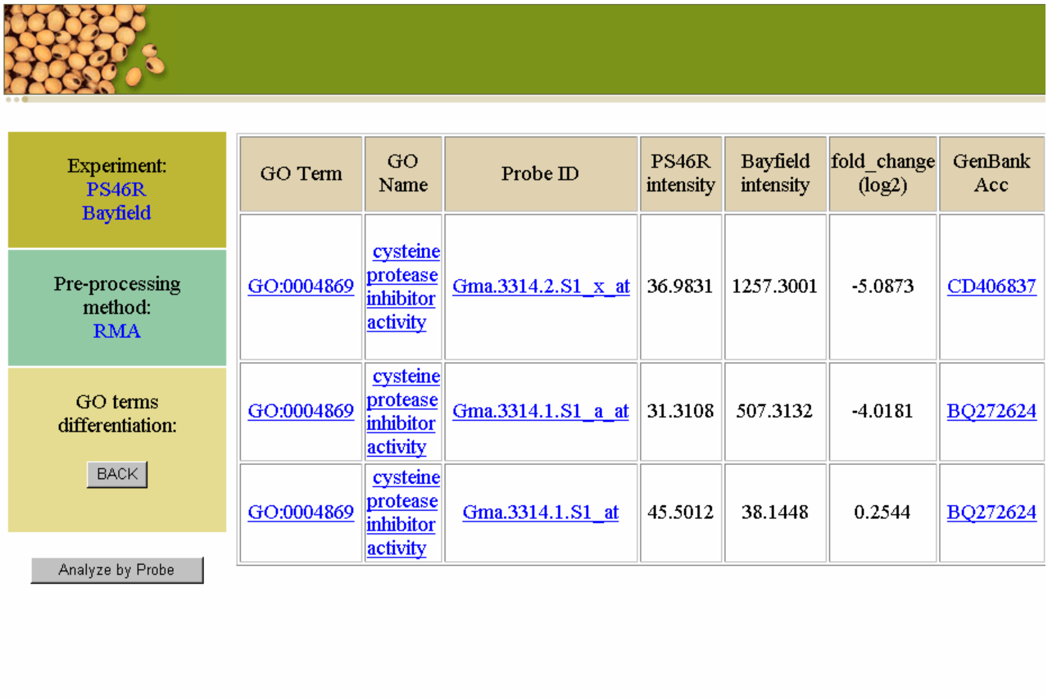
Pre-processing method:
RMA

GO terms differentiation:

show probe	GO Term	GO Name	No. of probes	M_score	t_score	p_value
○	GO:0004869	M-cysteine protease inhibitor activity	3	-2.9490	-10.3857	0.0000
○	GO:0016278	M-lysine N-methyltransferase activity	1	1.0591	8.9683	0.0000
○	GO:0016279	M-protein-lysine N-methyltransferase activity	1	1.0591	8.9683	0.0000
○	GO:0018024	M-histone-lysine N-methyltransferase activity	1	1.0591	8.9683	0.0000
○	GO:0019464	B-glycine decarboxylation via glycine cleavage system	1	3.0713	8.9038	0.0000
○	GO:0005960	C-glycine cleavage complex	2	1.9176	8.8712	0.0000
○	GO:0006544	B-glycine metabolism	2	1.9176	8.8712	0.0000
○	GO:0006546	B-glycine catabolism	2	1.9176	8.8712	0.0000
○	GO:0009071	B-serine family amino acid catabolism	2	1.9176	8.8712	0.0000
○	GO:0047746	M-chlorophyllase activity	2	-1.1259	-7.1155	0.0001
○	GO:0050356	M-tropine dehydrogenase activity	4	1.9877	6.5922	0.0002

Figure 4.8f. Result page 1 of the functional gene class analysis showing list of GO terms. The webpage is used to display the GO terms (that represent functional gene classes) have changes in microarray analysis. The GO number/term, number of genes involved in this function, statistical test score such as M-score (log2 fold change), t-score and p-value are displayed. The GO IDs are hyperlinked to the original public databases.

Figure 4.8g



Experiment:
PS46R
Bayfield

Pre-processing method:
RMA

GO terms differentiation:

GO Term	GO Name	Probe ID	PS46R intensity	Bayfield intensity	fold_change (log2)	GenBank Acc
GO:0004869	cysteine protease inhibitor activity	Gma.3314.2.S1_x_at	36.9831	1257.3001	-5.0873	CD406837
GO:0004869	cysteine protease inhibitor activity	Gma.3314.1.S1_a_at	31.3108	507.3132	-4.0181	BQ272624
GO:0004869	cysteine protease inhibitor activity	Gma.3314.1.S1_at	45.5012	38.1448	0.2544	BQ272624

Figure 4.8g. Result page 2 of the functional gene class analysis showing probes involved in the gene class (GO:0004869).

The webpage is used to display the genes belong to the functional group (cysteine protease inhibitor activity) showing up or down regulation in each gene. The GO number/term, intensities of each gene of the two compared cultivars, log2 fold change and GenBank accession number are displayed. The GO IDs and GenBank accession number are hyperlinked to the original public databases. The probe IDs also link back to the SOY search page.

4.2 Identification of cultivars based on microarray data

4.2.1 Quality assessment of microarray data

Initial RNA quality assessment (Beaulieu, 2005) showed that the integrity of the twenty-five cRNA samples was preserved with no obvious degradation (Beaulieu, 2005). However, when the quality assessment was repeated using *affyRNAdeg* functions from the R-Bioconductor *affy* package without transforming the intensities of probes to log2 base, as presented in **Figure 4.9,** the RNA degradation plot using all 61,170 probes for twenty-five microarray chips showed that there are two degradation patterns. Five samples are shown to have steeper slope than the other twenty samples. This indicates that these five samples (S03W4-1, S03W4-3, 2601R-2, 2601R-4 and 2601R-6) have a higher degree of RNA degradation. The plot shows the average intensity of the probes ordered by their proximity to the 5' end of the gene, from left to right. Since RNA degradation usually starts at the 5' end, probes close to the 5'end of the gene have lower intensity than that of the probes closer to the 3' end. If the difference in intensity between the 5' and 3'end of the probes is larger (steeper slope), it indicates poorer quality of RNA material (The GEPAS team 2005).

4.2.2 Group classification to define variations in 25 samples

Multivariate analysis was used to identify groups of cultivars based on gene expression. In **Figure 4.10**, the results of group classification using principal component (PCA) analysis for 25 non-processed chips is shown. It demonstrates that Mandarin form a group separate from the other samples. The five samples shown to have higher degree of RNA degradation (S03W4-1, S03W4-3, 2601R-2, 2601R-4 and 2601R-6) group together, and the remaining 15 samples from transgenic and non-transgenic form one heterogeneous group. These results show that PCA clustering does distinguish a cultivar if it is different enough from the others, but does not distinguish transgenic cultivars from non-transgenic ones. Furthermore it is clear that the RNA quality has a great impact on the clustering analysis.

Hierarchical clustering using Euclidean distance and average linkage for the 25 RMA normalized arrays is presented in **Figure 4.11**. Confirming the results of the PCA analysis, all five samples of the Mandarin cultivar form a separate cluster, the five samples (S03W4-1, S03W4-3, 2601R-2, 2601R-4 and 2601R-6) that were shown to have a higher degree of RNA degradation in **Figure 4.9** cluster into a separate group and the other 15 samples of non-transgenic and transgenic soybeans form one cluster (**Figure 4.11**). This indicates that the variations in gene expression between the four (non-Mandarin) cultivars are very small. The results also show that quality of cRNA targets for GeneChip hybridization might skew the results of cultivar classification in hierarchical clustering. This experimental error can not be corrected using our pre-processing tools. It also demonstrates that our pre-processing and quality assessment tools can detect the differences between poor quality samples and real differences in gene expression profiles among different cultivars.

Although these five samples (S03W4-1, S03W4-3, 2601R-2, 2601R-4 and 2601R-6) show differences in the RNA degradation plot and form a closer group in PCA analysis (for raw data) and hierarchical clustering (for RMA normalized data), due to weak statistical power, it would be difficult to use only two 2601R chips in the statistical analysis. We continued our analysis using all 25 chips and took notice that RNA degradation might have an effect within each S03W4 and 2601R cultivars, and the genes that were degraded in these two cultivars might not be identified as differentially expressed genes in the comparisons.

Except for the poor quality samples, all samples of the transgenic 2601R and PS46RR group together with Bayfield and S03W4 in both the PCA analysis and hierarchical clustering. These two transgenic cultivars do not cluster into a separate group from the non-transgenic cultivars. Therefore, the transgenic cultivars cannot be said to be different from the non-transgenic cultivars in a group-classification based on gene expression.

The results from the hierarchical clustering analysis (**Figure 4.11**) show that the five Bayfield samples, the five PS46RR samples and the two good quality 2601R samples cluster together. Therefore, it is likely that of the three conventional cultivars used in this study, Bayfield is the most closely related non-transgenic cultivar to the two transgenic cultivars.

Figure 4.9

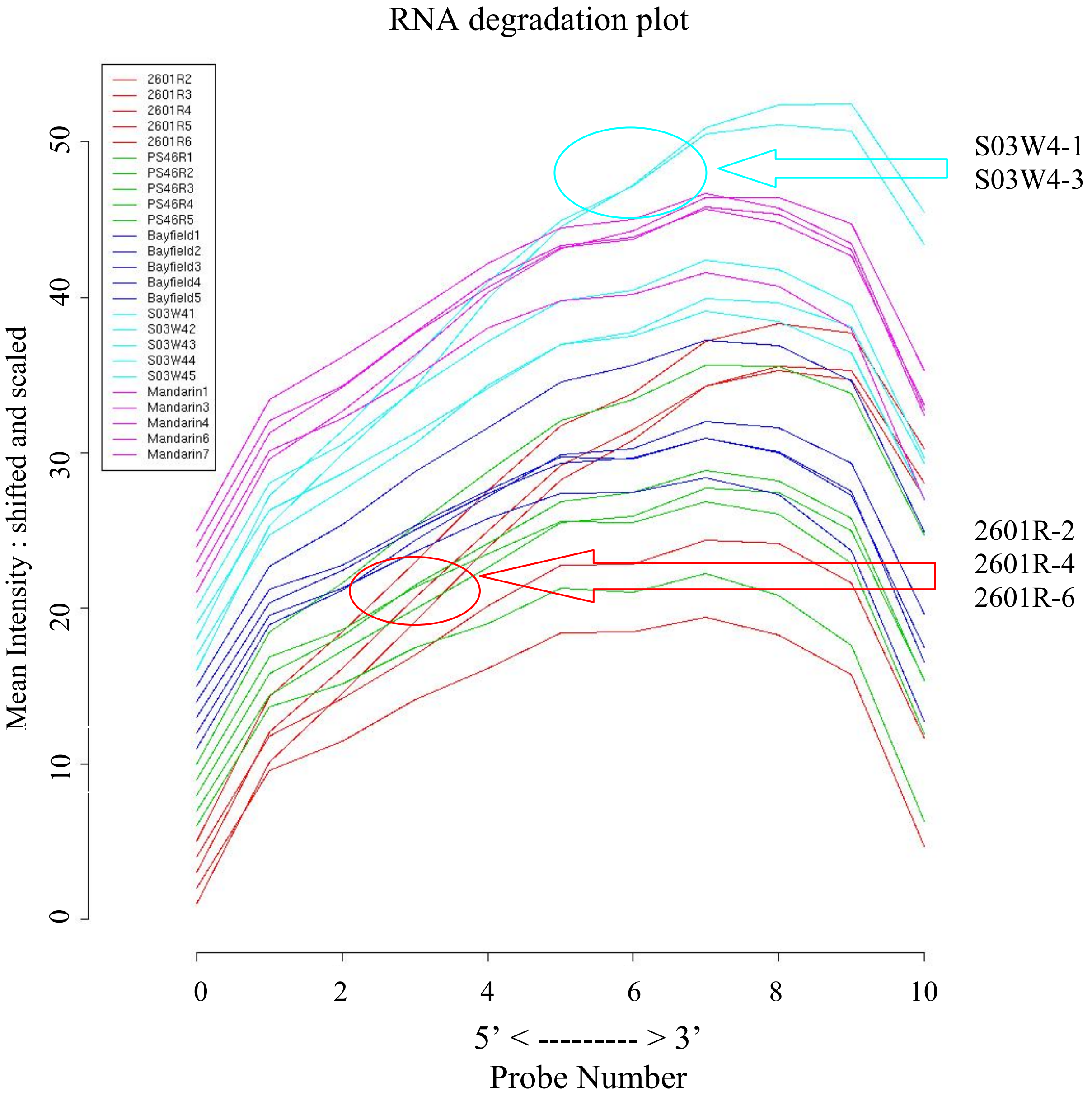
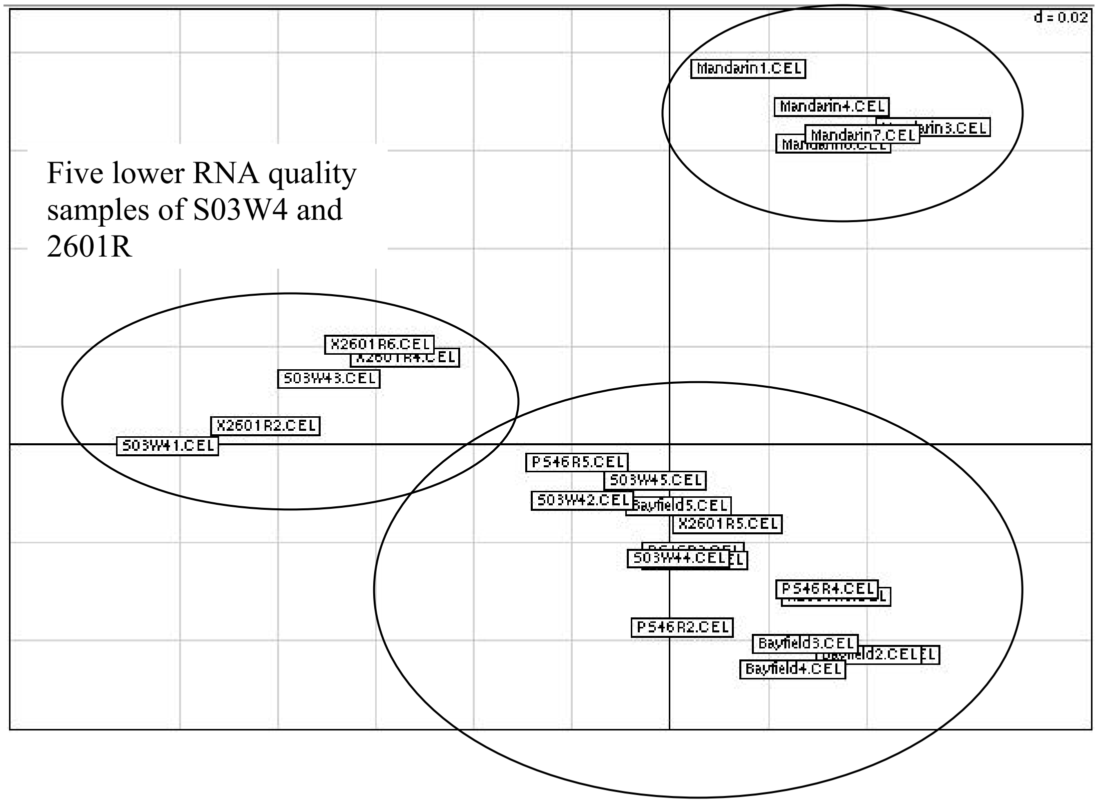


Figure 4.9. RNA degradation plot of 25 cRNA targets for Affymetrix GeneChip hybridization.

The average intensities for most samples have similar degradation patterns except for samples from two cultivars: S03W4 (light blue) and 2601R (red).

Figure 4.10

Five samples of Mandarin



Five lower RNA quality samples of S03W4 and 2601R

15 samples of good RNA quality transgenic (2601R, PS46RR) and non-transgenic (Bayfield, S03W4) soybeans

Figure 4.10. PCA analysis of the 25 microarray (raw data) from the five cultivars. Mandarin forms a separate group from the other soybean cultivars. Soybean cultivars cannot be classified into independent gene expression groups, based solely on whether they are transgenic or not (distance=0.02).

Figure 4.11

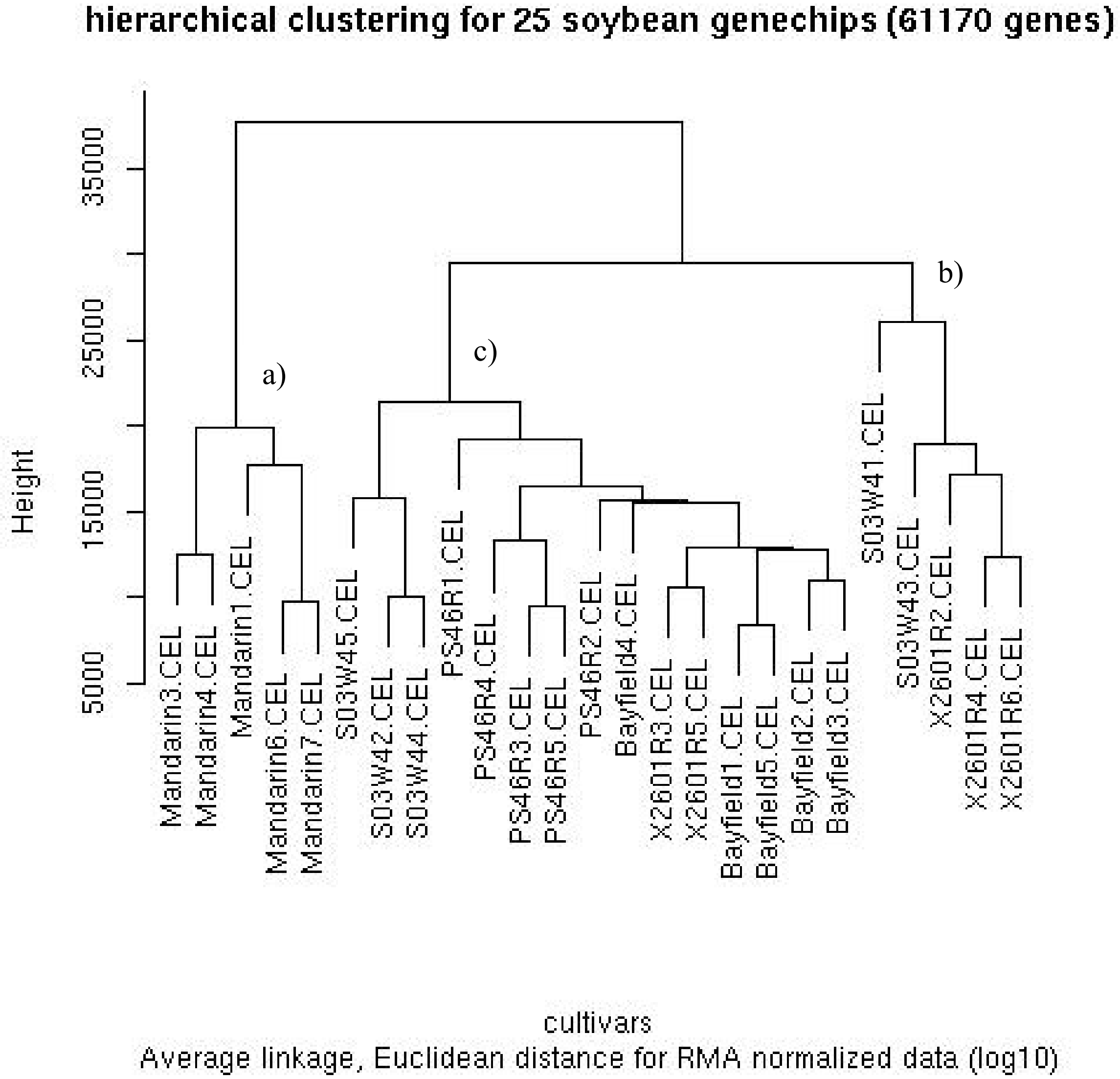


Figure 4.11. Hierarchical clustering for 25 samples of five cultivars using all probes on the arrays.

Three main clusters are:

a) Mandarin samples
b) poor quality samples of S03W4 and 2601R
c) the remaining 15 transgenic and non-transgenic samples of 2601R, PS46R, S03W4 and Bayfield.

4.2.3 Pair-wise comparison to define variations in five cultivars

In order to make exhaustive inter-cultivar comparisons, pair-wise comparisons using LIMMA was carried out between each soybean cultivar. Data preprocessed with different methods (RMA, MAS5 and dCHIP) show similar results. Mandarin has the highest numbers of differentially expressed genes among all five cultivars (**Figure 4.12**). After processing with RMA or MAS5 normalization methods, more than 1000 genes out of 37,583 total soybean genes on the chip are differentially expressed at p-values less than 0.01 and intensities greater than two-fold-change in the comparisons between Mandarin and any other cultivar. While the other four cultivars Bayfield, S03W4, 2601R and PS46RR are less different from each other (with less than 350 differentially expressed genes out of a total of 37,583 genes). After RMA preprocessing, only 44 genes are differentially expressed between Bayfield and 2601R, while 109 genes are differentially expressed when comparing Bayfield to PS46RR. The number of differentially expressed genes between Bayfield and each of the two transgenic cultivars is less than the number of differentially expressed genes between the two transgenic cultivars (137 differentially expressed genes). In the comparison of the other non-transgenic cultivar S03W4 to both transgenic cultivars, there are 248 genes differentially expressed when comparing S03W4 to 2601R; and 290 genes are differentially expressed when comparing S03W4 to PS46RR. The differences between transgenic and non-transgenic soybeans are less than the differences between two non-transgenic soybeans (332 differentially expressed genes). Based on the fewest differentially expressed genes, Bayfield is again shown to likely be the closest related non-transgenic cultivar to each of the transgenic cultivars.

Figure 4.12

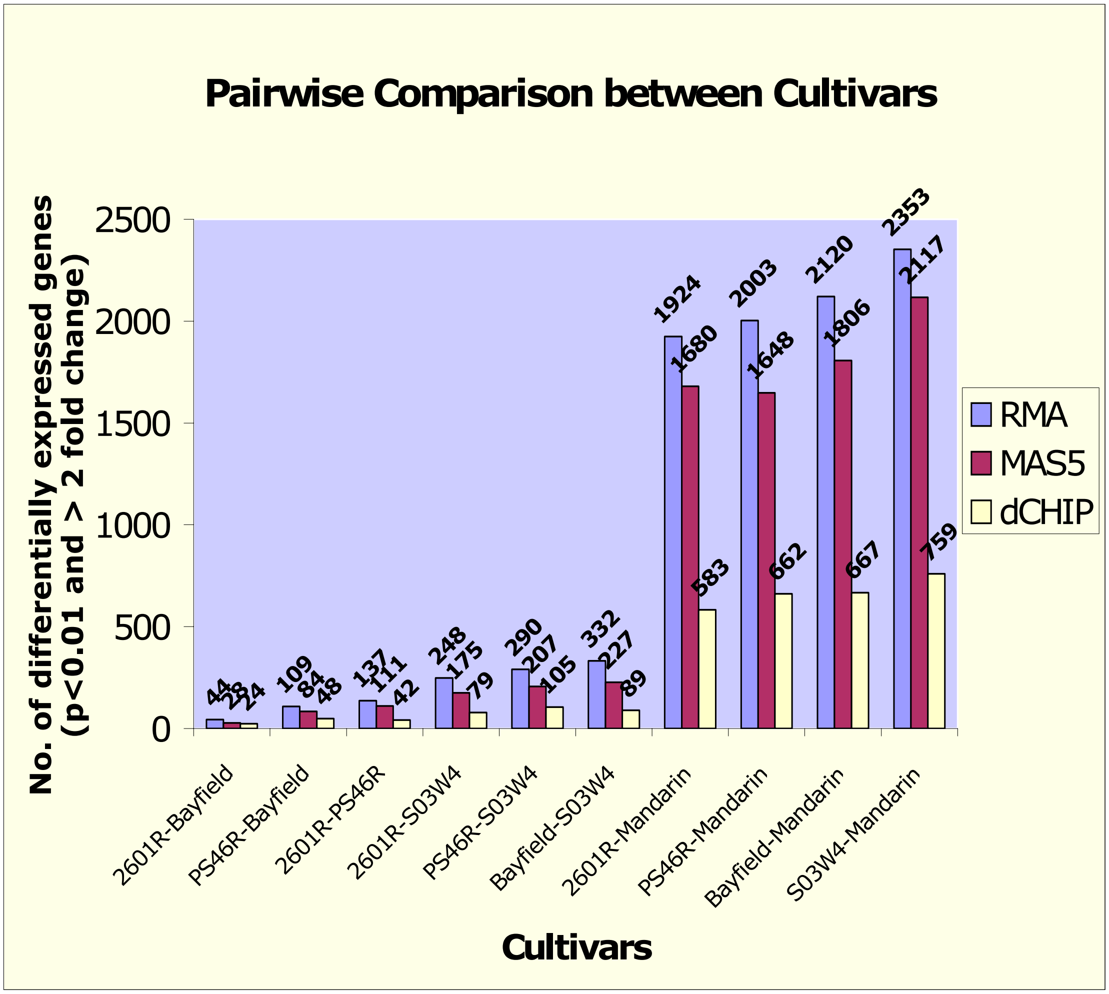


Figure 4.12. Pair-wise comparison between five different soybean cultivars.

LIMMA analysis on three sets of different pre-processed microarray data (using RMA, MAS5 or dCHIP). The numbers of differentially expressed genes (p-value < 0.01, intensities greater than 2 fold change) are located above the bars.

4.2.4 Resolving differences in gene expression at the probe level: comparison of one transgenic cultivar to one non-transgenic cultivar

Bayfield, being the cultivar found to likely be the closest conventional relative to the two transgenic cultivars, was compared individually with the transgenic cultivars 2601R and PS46RR to detect differential gene expression. Within the 44 and 109 differentially expressed genes found in the previous pair-wise comparison of 2601R and PS46RR to Bayfield (using RMA pre-processing method, cut-off at p-value < 0.01 and fold change > 2), only eight genes are differentially expressed in common in both transgenic cultivars (**Table 4.1**). Only three of these genes are annotated with Gene Ontology (GO) terms. Two genes that were down-regulated are in the category "cysteine protease inhibitor activity" and one gene that was down-regulated is in the category "dihydroflavonol-4-reductase activity". One of the up-regulated genes belongs to a TC (tentative contig, www.tigr.org) annotated with the category "cinnamoyl-CoA reductase activity".

In order to understand the molecular function of these genes, pair-wise comparisons in each of the transgenic cultivars are interpreted in terms of GO molecular function (using parent terms that describe the functions in more general annotations). In the comparison between 2601R and Bayfield using the RMA preprocessing method (cut-off at p-value < 0.01 and fold change > 2): two genes are identified as involved in "endopeptidase inhibitor activity"; five genes are involved in "transferase activity"; five genes are involved in "binding"; and one gene is involved in each of the functions "lyase activity", "signal transducer activity", "isomerase activity", "oxidoreductase activity", "transporter activity" and "hydrolase activity" (**Table 4.2**). The results are similar in the comparison of PS46RR to Bayfield using the same method: two genes are involved in "endopeptidase inhibitor activity"; thirteen genes are involved in "transferase activity"; eleven genes are involved in "binding"; five genes are involved in "hydrolase activity"; five genes are involved in "oxidoreductase activity"; three genes are involved in "signal transducer activity"; one gene is involved in "transporter activity and "nutrient reservoir activity" (**Table 4.3**).

Table 4.1. Differentially expressed genes common to both transgenic soybean cultivars (2601R and PS46R), compared with non-transgenic soybean Bayfield.

Probe Set ID	**Protein ID / Contig ID**	**Protein / Contig Description**	**GO terms**	**Fold change** (log2) (2601R)	**P-value** (2601R)	**Fold change** (Log2) (PS46R)	**P-value** (PS46R)
Gma.3314.1.S1_a_at	CYTM_SOLTU	(P37842) Multicystatin	(GO:0004866) endopeptidase inhibitor activity (GO:0004869) cysteine protease inhibitor activity	-4.157	5.88E-11	-4.018	6.12E-12
Gma.3314.2.S1_x_at	CYTA_HELAN	(Q10992) Cysteine proteinase inhibitor A	(GO:0004866) endopeptidase inhibitor activity (GO:0004869) cysteine protease inhibitor activity	-5.185	5.88E-11	-5.087	1.43E-11
GmaAffx.18584.1.S1_x_at				-2.674	1.89E-09	-2.370	1.29E-08
Gma.5206.1.A1_at	TC209225	Rev interacting protein mis3-like (Partial 19%)		2.413	5.07E-08	1.922	1.38E-06
GmaAffx.78465.1.S1_s_at	TC226919	Cinnamoyl CoA reductase-like protein (Partial 52%) (EC 1.2.1.44)		1.686	0.0003	1.387	0.0023
GmaAffx.57421.1.S1_at	TC221352			1.103	0.0011	1.249	0.0001
GmaAffx.52672.1.S1_at	TC217896	Replication factor C 110 kDa subunit (Partial 18%)		2.153	0.0029	1.758	0.0039
Gma.15664.1.S1_at	DFRA_VITVI	(P51102) Dihydroflavonol-4-reductase (EC 1.1.1.219)	(GO:0009813) flavonoid biosynthesis (GO:0016491) oxidoreductase activity (GO:0045552) dihydrokaempferol 4-reductase activity	-1.449	0.0030	-1.355	0.0033

Table 4.2. GO term interpretation of differentially expressed genes in transgenic soybean 2601R, compared with non-transgenic soybean Bayfield.

GO parent term	Specific GO term	Probe ID	Fold change(log2) / p-value
Endopeptidase inhibitor activity (GO 0004866)	Cystenine protease inhibitor activity (GO:0004869)	* **Gma.3314.1.S1_a_at**	-4.157 / 5.88E-11
		* **Gma.3314.2.S1_x_at**	-5.185 / 5.88E-11
Transferase activity (GO:0016740)	Flavonol 3-0-glucosyltransferase activity (GO:0047893)	* **GmaAffx.75645.1.A1_at**	-3.632 / 5.88E-011
	Indole-3-acetate beta-glucosyltransferase activity (GO:0047215)	* **GmaAffx.70608.1.S1_at**	-3.320 / 1.15E-06
	Zeatin O-beta-D-xylosyltransferase activity (GO:0050404)	* **GmaAffx.70608.1.S1_at**	-3.320 / 1.15E-06
	Glutathione transferase activity (GO:0004363)	Gma.8123.1.S1_at	-2.121 / 2.05E-06
	Two-component sensor molecular activity (GO:0000155)	* **Gma.17524.1.S1_at**	-1.279 / 2.43E-06
	Acyltransferase activity (GO:0047672)	Gma.6037.1.S1_at	2.172 / 0.0089
	Anthranilate N-benzoyltransferase activity (GO:0047672)	Gma.6037.1.S1_at	2.172 / 0.0089
Binding (GO:0005488)	Magnesium ion binding (GO:0000287)	GmaAffx.40133.1.S1_at	-2.706 / 2.46E-10
	Manganese ion binding (GO:0030145)	GmaAffx.40133.1.S1_at	-2.706 / 2.46E-10
	Translation elongation factor activity (GO:0003746)	Gma.8123.1.S1_at	-2.121 / 2.05E-06
	ATP binding (GO :0005524)	* **Gma.17524.1.S1_at**	-1.279 / 2.43E-06
	Cyclosporin A binding (GO:0016018)	GmaAffx.70933.2.S1_at	1.069 / 0.0011
	Unfolded protein binding (GO:0051082)	GmaAffx.70933.2.S1_at	1.069 / 0.0011
	Iron ion binding (GO:0005506)	Gma.1674.2.S1_at	1.096 / 0.0012
Lyase activity (GO:0016829)	Aristolochene synthase activity (GO:0045483)	GmaAffx.40133.1.S1_at	-2.706 / 2.46E-10
	(+)-delta-cadinene synthase activity (GO:0047461)	GmaAffx.40133.1.S1_at	-2.706 / 2.46E-10
Signal transducer activity (GO:0004871)	G-protein coupled photoreceptor activity (GO:0008020)	* **Gma.17524.1.S1_at**	-1.279 / 2.43E-06
Isomerase activity (GO:0016853)	Peptidyl-prolyl cis-trans isomerase activity (GO:0003755)	GmaAffx.70933.2.S1_at	1.069 / 0.0011
Oxidoreductase activity (GO:0016491)	Dihydrokaempferol 4-reductase activity (GO:0045552)	Gma.15664.1.S1_at	-1.449 / 0.0031
Transporter activity (GO:0005215)	Antiporter activity (GO:0015297)	GmaAffx.59053.1.S1_at	-1.171 / 0.0034
	Drug transporter activity (GO:0015238)	GmaAffx.59053.1.S1_at	-1.171 / 0.0034
Hydrolase activity (GO:0016787)	Glucan endo-1,3-beta-D-glucosidase activity (GO:0042973)	Gma.5205.1.A1_at	1.064 / 0.0042

* Indicates the probes were also differentially expressed in the comparison using a reference group of three non-transgenic cultivars Bayfield, S03W4 and Mandarin.

Table 4.3. GO term interpretation of differentially expressed genes in transgenic soybean PS46RR, compared with non-transgenic soybean Bayfield.

GO parent term	Specific GO term (molecular function category)	Probe ID	Fold change(log2) / p-value
Endopeptidase inhibitor activity (GO 0004866)	Cysteine protease inhibitor activity (GO:0004869)	* Gma.3314.2.S1_x_at * Gma.3314.1.S1_a_at	-5.087 / 1.43E-11 -4.018 / 6.12E-11
Transferase activity (GO:0016740)	Protein serine/threonine kinase activity (GO 0004674)	Gma.12743.1.A1_at Gma.9637.1.S1_at * GmaAffx.54889.1.S1_at * Gma.8137.1.S1_at GmaAffx.27802.1.S1_at Gma.17557.1.A1_at	3.922 / 1.18 E-14 2.117 / 7.23E-06 1.962 / 1.16E-05 1.330 / 0.0002 1.208 / 0.0029 1.086 / 0.0099
	Limonoid glucosyltransferase activity (GO:0050645)	Gma.13340.1.S1_at GmaAffx.72198.1.A1_at GmaAffx.51466.1.A1_at	1.466 / 8.40E-05 -1.411 / 0.0013 1.360 / 0.0066
	Hydroquinone glucosyltransferase activity (GO:0050505)	Gma.13340.1.S1_at GmaAffx.58893.1.S1_at	1.466 / 8.40E-05 -1.521 / 8.73E-05
	Flavonol 3-O-glucosyltransferase activity (GO:0047893)	GmaAffx.58893.1.S1_at GmaAffx.72198.1.A1_at	-1.521 / 8.73E-05 -1.411 / 0.0013
	Glutathione transferase activity (GO:0004364)	GmaAffx.88997.1.S1_s_at	3.225 / 2.10E-13
	Acyltransferase activity (GO:0008415)	* GmaAffx.86027.1.S1_at	1.664 / 2.88E-06
	Anthranilate N-benzoyltransferase activity (GO:0047672)	* GmaAffx.86027.1.S1_at	1.664 / 2.88E-06
	Histone-lysine N-methyltransferase activity (GO:0018024)	* GmaAffx.55247.1.S1_at	1.063 / 1.38E-06
	Protein-tyrosine kinase activity (GO:0004713)	GmaAffx.27802.1.S1_at	1.208 / 0.0029
	Indole-3-acetate beta-glucosyltransferase activity (GO:0047215)	GmaAffx.51466.1.A1_at	1.360 / 0.0066
Hydrolase activity (GO:0016787)	Chlorophyllase activity (GO:0047746)	Gma.13454.1.A1_at GmaAffx.40675.1.S1_at	-1.096 / 1.16E-05 -1.189 / 0.0025
	Nucleoside-triphosphatase activity (GO:0017111)	* GmaAffx.54889.1.S1_at	1.962 / 1.16E-05
	Acid phosphatase activity (GO:0003993)	GmaAffx.16212.1.S1_at	-1.584 / 0.0081
	Glucan endo-1,3-beta-D-glucosidase activity (GO 0042973)	Gma.7852.1.S1_at	1.591 / 0.0091

Table 4.3 continues on next page

Table 4.3 continued.

GO parent term	Specific GO term (molecular function category)	Probe ID	Fold change(log2) / p-value
Binding (GO :0005488)	ATP binding (GO:0005524)	* Gma.4564.1.A1_at	5.911 / 6.84E-25
		Gma.12743.1.A1_at	3.922 / 1.18E-14
		* Gma.9827.1.S1_at	3.157 / 4.36E-09
		Gma.9637.1.S1_at	2.117 / 7.23E-06
		* GmaAffx.54889.1.S1_at	1.962 / 1.16E-05
		* Gma.8137.1.S1_at	1.330 / 0.0002
		GmaAffx.27802.1.S1_at	1.208 / 0.0029
		Gma.17557.1.A1_at	1.086 / 0.0099
	Sugar binding (GO :0005529)	Gma.12743.1.A1_at	3.922 / 1.18E-14
		Gma.17557.1.A1_at	1.086 / 0.0099
	DNA binding (GO:0003677)	Gma.5949.1.A1_at	-2.693 / 6.62E-08
		* GmaAffx.54889.1.S1_at	1.962 / 1.16E-05
	Zinc ion binding (GO:0008270)	* GmaAffx.55247.1.S1_at	1.063 / 1.38E-06
	Steroid binding (GO:0005496)	Gma.9637.1.S1_at	2.117 / 7.23E-06
	NAD binding (GO:0051287)	Gma.2096.3.S1_s_at	1.602 / 7.62E-05
Oxidoreductase activity (GO:0016491	-	Gma.2096.1.S1_at	2.761 / 0.0034
	Tropine dehydrogenase activity (GO:0050356)	Gma.2096.3.S1_s_at	2.602 / 7.62E-05
		Gma.2096.2.S1_a_at	2.811 / 0.0005
		Gma.2096.2.S1_x_at	2.821 / 0.0034
	Tropinone reductase activity (GO:0050358)	Gma.2096.2.S1_a_at	2.811 / 0.0005
		Gma.2096.2.S1_x_at	2.821 / 0.0034
	3-oxoacyl-[acyl-carrier protein] reductase activity (GO:0004316)	Gma.2096.3.S1_s_at	2.602 / 7.62E-05
	Dihydrokaempferol 4-reductase activity (GO:0045552)	Gma.15664.1.S1_at	-1.355 / 0.0033
Signal transducer activity (GO:0004871)	Receptor activity (GO:0004872)	Gma.12743.1.A1_at	3.922 / 1.18E-14
		Gma.9637.1.S1_at	2.117 / 7.23E-06
	Transmembrane receptor activity (GO:0004888)	* GmaAffx.54889.1.S1_at	1.962 / 1.16E-05
Transporter activity (GO:0005215)	-	* Gma.15686.1.A1_at	-2.207 / 3.02E-07
Nutrient reservoir activity (GO:0045735)	-	GmaAffx.16212.1.S1_at	-1.584 / 0.0081

* Indicates the probes were also differentially expressed in the comparison using a reference group of three non-transgenic cultivars Bayfield, S03W4 and Mandarin

4.2.5 Resolving differential gene expression at the probe level: comparison of one transgenic cultivar to a group of non-transgenic cultivars

In the comparison of one transgenic soybean to a reference group (that represents the GRAS to assess substantial equivalence) of three non-transgenic soybeans instead of comparing to one non-transgenic soybean (Bayfield), the numbers of differentially expressed genes are reduced from 44 to ten genes in 2601R, and 109 to 49 genes in PS46RR. There are only five genes differentially expressed in common in both 2601R and PS46RR. However, only two of them are assigned with GO annotation, both involved in "cysteine protease inhibitor activity". **Table 4.4** shows the ten differentially expressed genes in 2601R. Three of them (protein sequences similar to flavonol 3-O-glucosyltransferase, Phytochrome A, and Zeatin O-xylosyltransferase or Indole-3-acetate beta-glucoxyltransferase, respectively) are involved in "transferase activity". Phytochrome is also involved in "binding" and "signal transducer activity". Two other genes (protein sequences similar to Cysteine proteinase inhibitor A and Multicystatin) are involved in "endopeptidase inhibitor activity". Nine of these ten genes are also differentially expressed in the comparison using only Bayfield as the comparator except the probe GmaAffx.52838.1.S1_at, which has a p-value of 0.06 in the comparison with Bayfield. However, it is significantly down-regulated in the comparison with Mandarin and S03W4 at p-value smaller than 0.0002. Unfortunately, there is no similar sequences found from the BLAST search using e-value < 0.01, therefore, there is no annotation for this gene and no information of its molecular function is thus provided. **Table 4.5** shows twelve (out of 49) differentially expressed genes that have GO terms annotations in the comparison of PS46RR using a reference group of three non-transgenic soybeans (Bayfield, Mandarin and S03W4). Six of the differentially expressed genes are involved in the molecular function "binding"; four genes involved in "transferase activity"; and two genes are involved in "cysteine protease inhibitor activity". Some of the genes have multiple functions such as GmaAffx.55247.1.S1_at and GmaAffx.54889.1.S1_at, which are involved in both "transferase activity" and "binding". Most of these twelve genes are also differentially expressed when using only Bayfield for comparison, except in the cases of Gma.16328.1.S1_at and Gma.2590.1.A1_s_at, whose fold changes between PS46RR and Bayfield were slightly below 2 (i.e. 1:1.92 and 1:1.82 respectively) and consequently can not be said to be differentially expressed with the same strict criteria.

Table 4.4. Differentially expressed genes in transgenic soybean 2601R, compared with a reference group of three non-transgenic soybean cultivars (Bayfield, S03W4 and Mandarin).

Probe Set ID	Protein ID / Contig ID	Protein Description	GO terms	Fold change (log2)	P_value
GmaAffx.75645.1.A1_at	UFOG4_MANES	(Q40286) Flavonol 3-O-glucosyltransferase 4 (EC 2.4.1.91)	(GO:0016740) transferase activity (GO:0047893) flavonol 3-O-glucosyltransferase activity	-3.723	1.36E-13
GmaAffx.84169.1.A1_at				2.367	4.79E-11
GmaAffx.70608.1.S1_at	ZOX_PHAVU IAAG_MAIZE	(P56725) Zeatin O-xylosyltransferase (EC 2.4.2.40) (Q41819) Indole-3-acetate beta-glucoxyltransferase (EC 2.4.1.121)	(GO:0016740) transferase activity (GO:0050404) zeatin O-beta-D-xylosyltransferase activity (GO:0047215) indole-3-acetate beta-glucosyltransferase activity	-3.059	1.26E-05
Gma.13345.1.S1_at				-1.260	1.38E-05
GmaAffx.57421.1.S1_at	TC221352			1.074	0.0002
Gma.17524.1.S1_at	PHYA_SOYBN	(P42500) Phytochrome A	(GO:0016740) transferase activity (GO:0000155) two-component sensor activity (GO:0005524) binding (GO:0005524) ATP binding (GO:0004871) signal transducer activity (GO:0008020) G-protein coupled photoreceptor activity	-1.229	0.0002
Gma.3314.2.S1_x_at	CYTA_HELAN	(Q10992) Cysteine proteinase inhibitor A (Cystatin A)	(GO:0004866) endopeptidase inhibitor activity (GO:0004869) cysteine protease inhibitor activity	-4.344	0.0022
Gma.13860.1.A1_at				1.419	0.0031
Gma.3314.1.S1_a_at	CYTM_SOLTU	(P37842) Multicystatin (MC)	(GO:0004866) endopeptidase inhibitor activity (GO:0004869) cysteine protease inhibitor activity	-3.385	0.0065
GmaAffx.52838.1.S1_at				-1.102	0.0077

Table 4.5. Differentially expressed genes that have GO term annotations in transgenic soybean PS46R, compared with a reference group of three non-transgenic soybean cultivars (Bayfield, S03W4 and Mandarin).

Probe Set ID	Protein ID	Protein Description	GO terms (molecular function category)	Fold change (log2)	P_value
Gma.4564.1.A1_at	RGA2_SOLBU	(Q7XBQ9) Disease resistance protein RGA2 (RGA2-blb) (Blight resistance protein RPI)	(GO:0005524) ATP binding	5.909	1.20E-28
Gma.16328.1.S1_at	ROC1_NICSY	(Q08935) 29 kDa ribonucleoprotein A, chloroplast precursor (CP29A)	GO:0003723) RNA binding	-1.064	1.16E-08
GmaAffx.55247.1.S1_at	SUVH9_ARATH	(Q9T0G7) Probable histone-lysine N-methyltransferase, H3 lysine-9 specific 9 (EC 2.1.1.43) (Histone H3-K9 methyltransferase 9) (H3-K9-HMTase 9) (Suppressor of variegation 3-9 homolog 9) (Su(var)3-9 homolog 9)	(GO:0018024) histone-lysine N-methyltransferase activity (GO:0008270) zinc ion binding	1.031	3.54E-08
GmaAffx.86027.1.S1_at	HCBT3_DIACA	(O23917) Anthranilate N-benzoyltransferase protein 2 (EC 2.3.1.144) (Anthranilate N-hydroxycinnamoyl/benzoyltransferase 2)	(GO:0008415) acyltransferase activity (GO:0047672) anthranilate N-benzoyltransferase activity	1.804	3.12E-07
GmaAffx.784.1.A1_at	CWF26_SCHPO	(O94417) Cell cycle control protein cwf26	(GO:0000398) nuclear mRNA splicing via spliceosome	-1.437	5.43E-07
GmaAffx.54889.1.S1_at	WRK52_ARATH	(Q9FH83) Probable WRKY transcription factor 52 (WRKY DNA-binding protein 52) (Disease resistance protein RRS1) (Resistance to Ralstonia solanacearum 1 protein)	(GO:0003677) DNA binding (GO:0005524) ATP binding (GO:0004888) transmembrane receptor activity (GO:0004674) protein serine/threonine kinase activity (GO:0017111) nucleoside-triphosphatase activity	2.076	5.75E-07
Gma.15686.1.A1_at	PIP22_ARATH	(P43287) Aquaporin PIP2.2 (Plasma membrane intrinsic protein 2b) (PIP2b) (TMP2b)	(GO:0005215) transporter activity	-1.760	2.40E-05
Gma.2590.1.A1_s_at	RS24_ARATH	(Q9SS17) 40S ribosomal protein S24	(GO:0003735) structural constituent of ribosome	-1.123	0.0001
Gma.8137.1.S1_at	BAK1_ARATH	(Q94F62) BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase 1 precursor (EC 2.7.1.37) (BRI1-associated receptor kinase 1) (Somatic embryogenesis receptor-like kinase 3)	(GO:0004674) protein serine/threonine kinase activity (GO:0005524) ATP binding	1.128	0.0002
Gma.3314.2.S1_x_at	CYTM_HELAN	(Q10992) Cysteine proteinase inhibitor A (Cystatin A) (SCA)	(GO:0004869) cysteine protease inhibitor activity	-4.247	0.0003
Gma.9827.1.S1_at	RGA2_SOLBU	(Q7XBQ9) Disease resistance protein RGA2 (RGA2-blb) (Blight resistance protein RPI)	(GO:0005524) ATP binding	2.125	0.0018
Gma.3314.1.S1_a_at	CYTM_SOLTU	(P37842) Multicystatin	(GO:0004869) cysteine protease inhibitor activity	-3.246	0.0026

4.2.6 Resolving differential gene expression at the functional-term level: comparison of one transgenic cultivar to one non-transgenic cultivar

In order to explore whether there are broad trends in gene expression for genes that share the same biological function, we averaged the gene expression levels (intensities) of all probes that were assigned with the same GO terms as the gene expression value for each GO terms and made comparisons between samples based on individual GO terms instead of individual probes. When comparing differentially expressed genes between 2601R and Bayfield by their GO terms, there are 27 GO terms (in Molecular function category) shown to be different at p-value < 0.01 (**Table 4.6**). The GO terms can be defined as a gene class having a group of genes with related functions. The gene class with molecular function "cysteine protease inhibitor activity" is shown to be significantly down-regulated in 2601R. In addition, the bigger gene classes consisting of a group of genes annotated with the parent GO terms describing more general functions such as "endopeptidase inhibitor activity", "protease inhibitor activity", "enzyme inhibitor activity", and "enzyme regulator activity" is also significantly down-regulated. Gene classes are bigger when they are more general, and hence, the numbers of genes belonging to these gene classes for the statistical analyses increased from three to 41. Within these 41 genes for gene classes "enzyme regulator activity", only two (Gma.3314.2.S1_x_at: cysteine proteinase inhibitor A and Gma.3314.1.S1_a_at: multicystatin) correspond to genes that are significantly down-regulated when analyzed based on individual probes instead of GO terms. However, the differences of expression in these two genes were very large. Even after combining the intensities of other (26 down-regulated and 13 up-regulated) genes, gene classes annotated with the parent GO terms of "cysteine protease inhibitor activity" are significantly down-regulated. The gene class "carbon-oxygen lyase activity, acting on phosphates", which it is the parent term of "(+)-delta-cadinene synthase activity", "aristolochene synthase activity" and "casbene synthase activity", is significantly down-regulated. Both "(+)-delta-cadinene synthase activity" and "aristolochene synthase activity" are also significantly different in the analysis based on individual probes. The sub-class (represented by a child term) "zeatin O-bega-D-xylosyltransferase activity" and its parent classes (represented by parent terms)

"UDP-xylosyltransferase activity" and "xylosyltransferase activity" are also down-regulated. Although only one of the three genes (GmaAffx.70608.1.S1_at: zeatin O-xylosyltransferase) included in the analysis is significantly differentially expressed in the individual probe level analysis, all three genes are down-regulated in the gene class analysis. In addition, genes belonging to the gene class annotated with child terms "indole-3-acetate beta-glucosyltransferase activity" and "dihydrokaempherol 4-reductase activity" are also shown to be significantly different in the analysis based on individual probes.

In the comparison between PS46RR and Bayfield, gene classes annotated with the GO term "cysteine protease inhibitor activity" are significantly down-regulated just as in the comparison between 2601R and Bayfield (**Table 4.7**). Two of the three genes that are included in the analysis for this GO term are also differentially expressed based on the comparison using individual probes (Gma.3314.2.S1_x_at: cysteine proteinase inhibitor A and Gma.3314.1.S1_a_at: multicystatin). However, the gene classes annotated with the parent terms are not significantly differentially expressed after combining the intensities of other probes for genes that are involved in the same functions, because the intensities for the probes of the up-regulated genes averaged out the difference in gene expression due to the down-regulated probes Gma.3314.2.S1_x_at and Gma.3314.1.S1_a_at. Gene classes annotated with the GO term "histone-lysine N-methyltransferase activity" with its parent terms "protein-lysine N-methyltransferase activity" and "Lysine-lysine N-methyltransferase activity" are up-regulated in this analysis. Many probes are annotated with the parent terms of "cystein proteinase inhibitor activity" as opposed to the parent terms of "histone-lysine N-methyltransferase activity", for which there are only one annotated probe. Therefore, no other probe would decrease or increase the intensity of up-regulation of this one gene within the parent gene classes of "histone-lysine N-methyltransferase activity". Other gene classes annotated with child terms are differentially expressed based on GO term analysis and also individual probe analysis such as "chlorophyllase activity", "tropine dehydrogenase activity" and "tropinione reductase activity".

Table 4.6. GO terms distinguished in the comparison between 2601R and Bayfield.

GO ID	GO term (molecular function category)	No. of probes *	P-value	Fold change (Log2)
GO:0004869	Cysteine protease inhibitor activity	3	1.8528E-07	-3.215
GO:0047461	(+)-delta-cadinene synthase activity	2	3.1972E-06	-2.072
GO:0045483	Aristolochene synthase activity	3	1.1554E-06	-1.696
GO:0047215	Indole-3-acetate beta-glucosyltransferase activity	4	0.00015	-1.277
GO:0004866	Endopeptidase inhibitor activity	18	0.00021	-0.756
GO:0030414	Protease inhibitor activity	18	0.00021	-0.756
GO:0050404	Zeatin O-beta-D-xylosyltransferase activity	2	0.01718	-2.287
GO:0035252	UDP-xylosyltransferase activity	3	0.00264	-1.580
GO:0042285	Xylosyltransferase activity	3	0.00264	-1.580
GO:0016838	Carbon-oxygen lyase activity, acting on phosphates	5	0.00264	-1.283
GO:0050449	Casbene synthase activity	1	0.00268	-0.944
GO:0008863	Formate dehydrogenase activity	2	0.00295	0.600
GO:0046547	Trans-aconitate 3-methyltransferase activity	1	0.00297	-1.085
GO:0004857	Enzyme inhibitor activity	24	0.00299	-0.609
GO:0008134	Transcription factor binding	3	0.00299	0.660
GO:0008605	Protein kinase CK2 regulator activity	2	0.00784	-0.630
GO:0045552	Dihydrokaempherol 4-reductase activity	5	0.00883	-0.717
GO:0030234	Enzyme regulator activity	41	0.00889	-0.484

* Indicates the numbers of probes that were involved in each GO term gene class.

Table 4.7. GO terms distinguished in the comparison between PS46RR and Bayfield.

GO ID	GO term (molecular function category)	No. of probes *	P-value	Fold change (Log2)
GO:0004869	Cysteine protease inhibitor activity	3	9.7603E-07	-2.949
GO:0016278	Lysine N-methyltransferase activity	1	1.9876E-06	1.059
GO:0016279	Protein-lysine N-methyltransferase activity	1	1.9876E-06	1.059
GO:0018024	Histone-lysine N-methyltransferase activity	1	1.9876E-06	1.059
GO:0047746	Chlorophyllase activity	2	7.5607E-05	-1.126
GO:0050356	Tropine dehydrogenase activity	4	0.00023	1.988
GO:0050358	Tropinone reductase activity	3	0.00063	1.800
GO:0050513	Glycoprotein 2-beta-D-xylosyltransferase activity	1	0.00085	0.725

* Indicates the number of probes that were involved in each GO term gene class.

5 Discussion

5.1 Effect of transgenes on global gene expression is within the natural range of variation of their conventional counterparts

In this study we are comparing the global gene expression profiles of leaves from five different soybean cultivars. The results of the study demonstrate that the insertion of a transgene has minimal effects on global gene expression. The conventional cultivar Mandarin Ottawa is the cultivar most different from the others as defined by a higher number of differentially expressed genes in pair-wise comparisons and in cluster analysis. This is not surprising, as Mandarin is an older cultivar, released in 1934 (Kumudini *et al.*, 2001, Beaulieu, 2005), and has a longer history of commercialization than the other four soybean cultivars. Although Mandarin is a major ancestor of North American soybean cultivars and has contributed 11-22% to the genomes of present-day northern soybean elite lines (Kisha *et al.*, 1998, Sneller, 2003), its contribution to the northern gene pool has been reduced in the past 10-15 years (Sneller, 2003). Therefore the more ancient soybean cultivar might be more distant genetically compared to the recently developed cultivars, which are more inbred. However, it is somewhat surprising that the expression profiles of leaves of two different, though not remotely related, cultivars can vary by that many genes (over 1,000), and a study of even more distant conventional cultivars could be expected to show that the natural range of variation at the gene expression level in soybean is quite large. Four of the other soybean cultivars are very similar in global gene expression pattern. Our hierarchical clustering analysis could not distinguish the group of transgenic soybean cultivars (2601R and PS46RR) from the other (non-Mandarin) non-transgenic cultivars (Bayfield and S03W4), and less than 332 genes (>1% of the total soybean genes arrayed) differed significantly (p-value < 0.01) with expression levels higher than twofold in any pair-wise comparisons among these four cultivars. Most strikingly, the number of differentially expressed genes between non-transgenic cultivars (Bayfield/S03W4) was higher than the number of differentially expressed genes between transgenic and non-transgenic soybeans. Ouakfaoui and Miki had already demonstrated a single insertion of T-DNA and common reporter genes did not affect gene expression level in transgenic plants (Ouakfaoui and Miki, 2005). Our result implies that the insertion of a transgene into a plant genome does not have great impact on global gene expression in plants. The result is similar to a previous finding that the expression of *A. fumigatus* phytase had minimal effect on the gene expression pattern in the transgenic wheat seedlings (Gregersen *et al*., 2005) and also similar to a recent cDNA microarray study in wheat lines expressing genes encoding high molecular weight subunits of glutenin (Baudo *et al*., 2006) that suggested the presence of transgene has less impact on the transcriptome than conventional breeding. However, we could not distinguish transgenic and non-transgenic soybeans based on the minimal differences between them; therefore, we reject our first hypothesis that transgenic and non-transgenic soybean genotypes (cultivars) can be distinguished by their global gene expression profiles. Cultivars can be distinguished from others, if they are sufficiently distant genetically (e.g. Mandarin and Bayfield), however the dataset available is too limited to determine the boundaries.

In our microarray experiment, there was a problem of choosing the appropriate comparators to assess only the transgenic effects rather than the genetic diversity among these soybean cultivars. The ideal comparator would have been the near isogenic parental line grown under identical conditions (FAO/WHO, 2000). However, such comparators are difficult to obtain in practice, since companies rarely reveal their breeding programs. Both transgenic soybean cultivars that were used in this study are derived from the same line, 40-3-2, and the same insertion event, although we do not know where in the genome the transgene is integrated. This is the soybean cultivar A5403 that has been transformed with the transgene 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase found in the CP4 strain of *Agrobacterium* and which confers tolerance to the herbicide glyphosate (RoundUp®) (Padgette *et al*., 1995). Line 40-3-2 is used in various breeding programs to develop new cultivars with the RoundUp® ready gene to adapt to the northern soybean growing area (Delannay *et al*., 1995). Using the parent line A5403 for comparison might have been more appropriate. However, A5403 is a southern cultivar from Asgrow Seed Company (Padgette *et al*., 1995, Sneller, 2003), which may not have the same gene expression characteristics as our northern soybean cultivars under the same growing condition. Therefore, we identified the closest non-transgenic cultivar as the conventional counterpart for the comparison to determine substantial equivalence between transgenic and non-transgenic crops. The five soybean cultivars in our study were carefully chosen based on field trial reports and literature paper (Beaulieu, 2005). Our analyses demonstrate that our examined non-transgenic cultivars (Bayfield and S03W4) are very close in terms of global gene expression profile to our transgenic cultivars, Bayfield being the closest. Each transgenic cultivar is closer to Bayfield than to each other. Therefore, Bayfield was used as the comparator in the pair-wise transgenic versus non-transgenic comparisons.

In the comparison between each of the two transgenic cultivars (2601R and PS46RR) and Bayfield, only eight genes are differentially expressed in common in both transgenic soybean cultivars. It is possible that these eight genes are affected by the insertion of the transgene (which again, is the same insertion event in the parent 40-3-2) resulting in intended or unintended effects, but it may also be that the differences are due to the variation of the plant genotypes themselves. Only four of the eight genes have GO annotations, the other genes have no similar sequence found from the BLASTX search result having. Two of them (Gma.3314.1.S1_a_at: multicystatin, Gma.3314.2.S1_x_at: cysteine proteinase inhibitor A) are down regulated and annotated as having cysteine protease inhibitor activity. One of the genes (Gma.15664.1.S1_at: dihydroflavonol-4-reductase) involved in dihydrokaempferol 4-reductase activity and flavonoid biosynthesis is down regulated. Further experiment such as quantitative real-time RT-PCR will be needed to validate the results.

In addition, we applied the concept of substantial equivalence to investigate if a group of conventional breed cultivars (GRAS, Generally Recognized As Safe) could be used as the control (FAO.WHO, 2000) in gene expression experiments to assess whether our two transgenic cultivars are within the natural range of variation of their conventional counterpart cultivars that have similar performance and phenotype. The comparisons between each transgenic cultivar to the group of three non-transgenic cultivars show similar results, which implies that differentially expressed genes can be identified using this approach.

5.2 Annotation database integrated with biological functional terms provides information to predict unintended effects

In order to obtain biological information from the gene expression data, many researchers translate a list of differentially expressed genes to relevant biological processes and pathways manually through literature and public databases searches (Draghici, 2003). However, this is a tedious and time-consuming process. Integrating nucleotide information for the soybean genes on the microarray with BLAST search results (SwissProt protein IDs), GO terms annotation and KEGG pathways in our database, minimizes the time and effort for retrieving all these cross-references gene-by-gene manually. Also, with the help of the database, we can interpret the differentially expressed genes based on functional annotations in terms of gene ontology molecular function category.

A GO term does not only provide functional annotation, but it also represents a gene class whose members share the same biological function. We observed many of the differentially expressed genes assigned to the parent terms “transferase activity” and “binding” from the cross-references we obtained from the database. However, GO terms are organized in a hierarchy(tree)-like structure, so that a gene assigned with a child term is also associated with the parent terms that describe the function of that gene in a more general term (The Gene Ontology Consortium, 2000). Therefore, many genes can be assigned to one parent term as opposed to the child term, which describes a very specific function of only a few of genes assigned to it. Thus, we do not know if observing the high frequency in “transferase activity” or “binding” is due to real significantly regulated biological processes or because these are random events (since they represent very large gene classes that have a high probability to be observed). Currently, there is a statistical method to calculate the probability that a certain GO term occurs several times just by chance in the list of differentially regulated genes (Draghici *et al*., 2003). This approach makes use of the cumulative hypergeometric or binomial distribution and χ^2 or Fisher's exact test to identify significantly over-represented GO terms (Khatri and Draghici, 2005). However, this approach only calculates the probability of random events but it does not consider the expressing levels of the other genes on the same microarray chip. Therefore, by the help of our database, we generated a value of the combined intensities of all the genes assigned to each GO term in each microarray chips and used these values for statistical analysis to identify which GO term is more relevant. And hence, by averaging the intensities of the genes belonging to each GO term, none of the protein classes with "binding" function is shown in our GO term list. This is because the number of genes assigned to each binding function is large and the other genes do not have a consistent up or down regulation pattern. Therefore expression levels of other genes averaged out the expression differences of the individual differentially regulated genes. These results show that identifying differentially expressed biological functions (i.e. obtaining GO terms) by averaging intensities can identify gene classes that consist of genes expressing consistent patterns. This provides better ranking of the functional gene classes. However, we could not accept our second hypothesis that this method is more accurate than obtaining GO terms from the annotated list of differentially expressed genes, because genes within the same functional group might not necessarily be co-regulated within the same tissue at the same time. Further investigation into individual genes within the gene class has to be done based on scientific knowledge, and our web tools can provide individual expression intensities for each genes to assist investigators to identify whether the expression pattern is biological relevant or not.

6 Conclusion

We have created a database to integrate existing genomics data for the soybean nucleotides, transcripts, proteins and results from our microarray experiment, and developed web interfaces to retrieve and display these data. By mapping information in the soybean database, results from the microarray experiments are associated with corresponding protein names and functional (GO) terms, which provide insight into functional differences between samples and enhance the prediction of unintended effects in transgenic soybean cultivars.

Although we could not distinguish transgenic cultivars from non-transgenic cultivars due to minimal difference between our transgenic and non-transgenic cultivars, our microarray analysis shows that the analysis of gene expression profiles of transgenic crops and their conventional counterparts can identify differentially expressed genes under similar growth condition. The pair-wise analysis in the comparison of transgenic soybean to the closest conventional counterparts produced a list of differentially expressed gene and revealed that, in both transgenic cultivars, genes involved in cysteine protease inhibitor activity and dihydroflavonol-4-reductase activity were down-regulated. It may reflect an effect of the insertion event, an effect of the transgene product and thus a real unintended effect, or a natural variation of the parent genotype. Further investigations in the laboratory will be needed to assess effects like this.

We could not show that analysis based on functional gene class comparison is more accurate than analysis based on individual genes, because there is no laboratory data to validate our results. Most importantly, genes within the same gene class might not be necessarily co-regulated in the same tissue at the same developmental stage. However, by combining intensities of genes within the same gene class, we could provide better ranking of the functional terms by average out general terms that have high probability to be observed randomly. Also, the gene class with genes showing consistent expression patterns can be moved up on the list to reveal biological relevant event. Our web tool provides functions to display individual intensities for each gene, which assist research to observe if there is any consistent expression pattern within the same functional gene class and identify relevant biological processes.

Our study demonstrated the use of microarray technology and the development of database with web interface as a tool for crop safety assessment. It is important to point out that obtaining a target gene list cannot conclude whether the transgenic soybean is absolutely safe or not, since gene expression might not necessarily influence metabolite accumulation. Furthermore, it has been agreed that the concept of substantial equivalence was developed as a practical approach to the safety assessment process, but it is not a safety assessment itself. It was not established to characterize the hazard; rather it is used as a starting point that is to lead to further safety assessment (FAO/WHO, 2000).

In conclusion, we have shown that the insertion of a transgene in our examined transgenic soybean cultivars has minimal effect on gene expression, and we demonstrated the screening of unintended effects by analyzing gene expression data using bioinformatics tools and the development of a database for obtaining relevant biological information on the differentially expressed genes. Hence, we provided a tool for easier prediction of the molecular functions and pathways likely to be influenced by the transgene insertion or gene product.

7 Recommendation for future studies

A general problem for a database project is the burden of version control that requires regular download and update of the new release data from the public databases. Since new data for soybean is not released as often as for species like human or mouse, our database is still considered to be comprehensive to provide information for our transgenic and non-transgenic soybeans comparison. In addition, a common problem for many sequence projects is that the existing annotations are incomplete. Therefore even though we obtained a list of differentially expressed genes, only a subset of those genes is annotated with biological functions. Therefore any differentially expressed genes and gene class we find from our analysis need to undergo further investigation for safety assessment, and regular updates of the database have to be done.

In the attempt of making use of GO terms for gene class differentiation, we only performed a simple calculation to average all the intensity of the genes assigned to one GO term. However, genes with high intensities may mask the expression values of the low intensities genes, although the fold-change of the high intensities may not be larger and more significant than the low intensity genes. A use of mean or median could be used to normalize the intensity between these genes. Therefore further development of the statistical method has to be done.

Currently, there are groups using very sophisticated statistical method for functional class group testing, such as functional class score (Pavlidis *et al*., 2004, Mootha *et al*., 2003) or global test (Goeman *et al*., 2004). However, due to the incompleteness of our soybean data, functional gene class analysis is still difficult even if we used the most sophisticate statistical analysis. Therefore, better annotation of soybean has to be done in order to interpret biological functions related to the gene expression data.

Also, it is very important to do a follow-up experiment to validate the differentially expressed genes, using for instance real time RT-PCR technique. In the end, to see if there are any phenotypic effects of the gene expression, it would be necessary to also identify whether the differences in gene expression level significantly correlate with the production of protein and metabolite components.

Further studies can be carried out to understand the intended effect of the transgene that encode EPSP synthase. Since the glyphosate herbicide targets the shikimate pathway (Steinrucken and Amrhein, 1980), it would be expected to affect the downstream metabolic pathways that produce aromatic amino acids (phenylalanine, tyrosine and tryptophan), and their secondary products isoflavones (genistein, daidzein, bound coumestrol and biochanin) (Taylor *et al*., 1999). In addition, further comparisons of soybean with and without the treatment of glyphosate herbicide can also be done to assess whether the application of herbicide would affect the change at gene expression level. However in the previous compositional analysis of transgenic soybeans treated with glyphosate herbicide demonstrated that these treated soybeans were comparable to the parental soybean cultivar and other conventional soybeans (Taylor *et al*., 1999). Since full proteome and metabolome profiling method is not available yet (Metzdorff *et al*., 2006), microarray technology is the only available tool for analyzing the full transcriptome and it does have the potential to be a useful tool for screening for unintended effects in transgenic crops.

References

Affymetrix. (2001) Technical notes: GeneChip arrays provide optimal sensitivity and specificity for microarray expression analysis. *Documentation from Affymetrix website*. 1-4. Accessed on August 7, 2007.
https://www.affymetrix.com/support/technical/technotes/25mer_technote.pdf

Affymetrix. (2002) Statistical algorithms description document. *Documentation from Affymetrix website.* 1-28. Accessed on August 7, 2007.
http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf

Affymetrix. (2004-2005) Data Sheet: GeneChip Soybean Genome Array. *Documentation from Affymetrix website*.: 1-2. Accessed on August 7, 2007.
http://www.affymetrix.com/support/technical/datasheets/soybean_datasheet.pdf

Al-Shahrour F, Minguez P, Tarraga J, Medina I, Alloza E, Montaner D and Dopazo J. (2007) FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory modifs and interaction data with microarray experiments. *Nucleic Acids Research*. 35 (Web Server issue): W91-W96.

Alkharouf NW and Matthews BF. (2004) SGMD: The soybean genomics and microarray database. *Nucleic Acids Research.* 32 (Database issue): D398-D400.

Altschul S.F, Gish W., Miller W., Myers E.W and Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*. 215: 403-410.

Baudo MM, Lyons R, Powers S, Pastori GM, Edwards KJ, Holdsworth MJ and Shewry PR. (2006) Transgenesis has less impact on the transcriptome of wheat grain than conventional breeding. *Plant Biotechnology Journal*. 4: 369-380.

Beaulieu J. (2005) *Exploration of high-density oligoarrays as tools to assess substantial equivalence of genetically modified crops*. Master Thesis. McGill University, Canada. 1-64.

Beissbarth T and Speed TP. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*. 20(9): 1464-1465.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL. (2007) GenBank. *Nucleic Acids Research.* 35 (Database issue): D21-D25.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*. 31(1): 365-370.

Cellini F, Chesson A, Colquhoun I, Constable A, Davies HV, Engel KH, Gatehouse AMR, Karenlampi S, Kok EJ, Leguay JJ, Lehesranta S, Noteborn HPJM, Pedersen J and Smith M. (2004) Unintended effects and their detection in genetically modified crops. *Food and Chemical Toxicology*. 42: 1089-1125.

Delannay X, Bauman TT, Beighley DH, Buettner MJ, Coble HD, DeFelice MS, Derting CW, Diedrick TJ, Griffin JL, Hagood ES, Hancock FG, Hart SE, LaVallee BJ, Loux MM, Lueschen WE, Matson KW, Moots CK, Murdock E, Nickell AD, Owen MDK, Paschal II EH, Prochaska LM, Raymond PJ, Reynolds DB, Rhodes WK, Roeth FW, Sprankle PL, Tarochione LJ, Tinius CN, Walker RH, Wax LM, Weigelt HD and Padgette SR. (1995) Yield evaluation of a glyphosate-tolerant soybean line after treatment with glyphosate. *Crop Science*. 35: 1461-1467.

Dunwell JM. (2005) Transgenic crops: the current and next generations. *Methods in Molecular Biology.* 286: 377-398.

Draghici S, Khatri P, Martins RP, Ostermeier C and Krawetz SA. (2003) Global functional profiling of gene expression. *Genomics* 81: 98-104.

FAO/WHO. (2000) Safety aspects of genetically modified foods of plant origin. *Report of a Joint FAO/WHO Expert Consultation on Foods Derived from Biotechnology.* 1-35. Accessed on August 7, 2007.
http://www.who.int/foodsafety/publications/biotech/en/ec_june2000_en.pdf

Firth D. (2003) CGIwithR : Facilities for processiong web forms using R. *Journal of Statistical Software.* 8(10): 1-8.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD and Bairoch A. (2003) ExPASy : the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research.* 31: 3784-3788.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY and Zhang J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*. 5(10): article R80, 1-16.

Goeman JJ, Vande Geer SA, De Kort F and Van Houwelingen HC. (2004) A global test for groups of genes : testing association with a clinical outcome. *Bioinformatics*. 20(1): 93-99.

Gonzales MD, Archuleta E, Farmer A, Gajendran K, Grant D, Shoemaker R, Deavis WD and Waugh ME. (2005) The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Research.* 33 (Database issue): D660-D665.

Gregersen PL, Brinch-Pedersen H and Holm PB. (2005) A microarray-based comparative analysis of gene expression profiles during grain development in transgenic and wild type wheat. *Transgenic Research*. 14: 887-905.

Hashimoto W, Momma K, Katsube T, Ohkawa Y, Ishige T, Kito M, Utsumi S and Murata K. (1999) Safety assessment of genetically engineered potatoes with designed soybean glycinin: compositional analyses of the potato tubers and digestibility of the newly expressed protein in transgenic potatoes. *Journal of the Science of Food and Agriculture*. 79: 1607-1612.

Hornik. (2007) The R FAQ. *Documentation from R-Project website*. (ISBN 3-900051-08-9). Accessed on August 7, 2007. http://www.r-project.org/

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 4(2): 249-264.

Ithal N, Recknor J, Nettleton D, Hearne L, Maier T, Baum TJ and Mitchum MG. (2007) Parallel genome-wide expression profiling of host and pathogen during soybean cyst nematode infection of soybean. *Molecular Plant-Microbe Interactions*. 20(3): 293-305.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M and Hirakawa M. (2006) From genomics to chemical genomics : new developments in KEGG. *Nucleic Acids Research.* 34 (Database issue): D354-D357.

Khatri P and Draghici S. (2005) Ontological analysis of gene expression data : current tools, limitations, and open problems. *Bioinformatics*. 21(18): 3587-3595.

Kisha TJ, Diers BW, Hoyt JM and Sneller CH. (1998) Genetic diversity among soybean plant introductions and north american germplasm. *Crop Science*. 38: 1669-1680.

Kok KJ and Kuiper H. (2003) Comparative safety assessment for biotech crops. *Trends in biotechnology.* 21(10): 439-444.

Koncz C, Nemeth K, Redie GP and Schell J. (1992) T-DNA insertional mutagenesis in Arabidopsis. *Plant Molecular Biolology*. 20(5): 963-976.

Kuiper HA, Kleter GA, Noteborn HPJM and Kok EJ. (2001) Assessments of the food safety issues related to genetically modified foods. *The Plant Journal*. 27(6): 503-528.

Kuiper HA, Kleter GA, Noteborn HPJM and Kok EJ. (2002) Substantial equivalence-an appropriate paradigm for the safety assessment of genetically modified foods? *Toxicology.* 181-182: 427-431.

Kumar CG, LeDuc R, Gong G, Roinishivili L, Lewin HA and Liu L. (2004) ESTIMA, a tool for EST management in a multi-project environment. *BMC Bioinformatics*. 5: article 176, 1-10.

Kumudini S, Hume DJ and Chu G. (2001) Genetic improvement in short season soybeans : I. dry matter accumulation, partitioning, and leaf area duration. *Crop Science.* 41: 391-398.

Li C and Wong WH. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*. 98(1): 31-36.

Metzdorff SB, Kok EJ, Knuthsen P and Pedersen J. (2006) Evaluation of a non-targeted “omic” approach in the safety assessment of genetically modified plants. *Plant Biolology.* 8: 662-672.

Millstone E, Brunner E and Mayer S. (1999) Beyond 'substantial equivalence'. *Nature*. 401: 525-526.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, AltshulerD and Groop LC. (2003) PGC-1aplha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*. 34(3): 267-273.

Ouakfaoui SE and Miki B. (2005) The stability of the Arabidopsis transcriptome in transgenic plants expressing the marker genes nptII and uidA. *The Plant Journal*. 41: 791-800.

Padgette SR, Kolacz KH, Delannay X, Re DB, LaVallee BJ, Tinius CN, Rhodes WK, Otero YI, Barry GF, Eichholtz DA, Peschke VM, Nida DL, Taylor NB and Kishore GM. (1995) Development, identification, and characterization of a glyphosate-tolerant soybean line. *Crop Science*. 35: 1451-1461.

Panthee DR, Yuan JS, Wright DI, Marois JJ, Mailhot D and Stewart Jr CN. (2007) Gene expression analysis in soybean in response to the causal agent of Asian soybean rust (*Phakopsora pachyrhizi* Sydow) in an early growth stage. *Functional and integrative genomics*. (doi:10.1007/s10142-007-0045-8). 1-11.

Pavlidis P, Qin J, Arango V, Mann JJ and Sibille E. (2004) Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research*. 29(6): 1213-1222.

Quackenbush J, Lian F, Holt I, Pertea G and Upton J. (2000) The TIGR Gene Indics: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research*. 28: 141-145.

Shen L, Gong J, Caldo RA, Nettleton D, Cook D, Wise RP and Dickerson JA. (2005) BarleyBase-an expression profiling database for plant genomics. *Nucleic Acids Research.* 33 (Database issue): D614-D618.

Shewmaker CK, Sheehy JA, Daley M, Colburn S and Ke DY. (1999) Seed-specific overexpression of phytoene synthase: increase in carotenoids and other metabolic effects. *The Plant Journal*. 20(4): 401-412.

Shoemaker R, Keim P, Vodkin L, Retzel E, Clifton SW, Waterston R, Smoller D, Coryell V, Khanna A, Erpelding J, Gai X, Brendel V, Raph-Schmidt C, Shoop EG, Vielweber CJ, Schmatz M, Pape D, Bowers Y, Theising B, Martin J, Dante M, Wylie T and Granger C. (2002) A compilation of soybean ESTs: generation and analysis. *Genome.* 45: 329-338.

Smyth GK, Michaud J and Scott H. (2005) The use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics.* 21(9): 2067-2075.

Sneller CH. (2003) Impact of transgenic genotypes and subdivision on diversity within elite North American soybean germplasm. *Crop Science.* 43: 409-414.

Steinrucken HC and Amrhein N. (1980) The herbicide glyphosate is a potent inhibitor of 5-enolpyruvyl shikimic acid-3-phosphate synthase. *Biochemical and Biophysical Research Communications*. 94(4): 1207-1212.

Tattersall DB, Bak S, Jones PR, Olsen CE, Nielsen JK, Hansen ML, Hoj PB and Moller BL. (2001) Resistance to an herbivore through engineered cyanogenic glucoside synthesis. *Science.* 293(5336): 1826-1828.

The Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics.* 25: 25-29.

The GEPAS team. (2005) Expresso: Affymetrix normalization using Expresso. *Documentation from GEPAS website*. Accessed on August 7, 2007. http://gepas.bioinfo.cipf.es/cgi-bin/tutoX?c=expresso/expresso.config

Taylor NB, Fuchs RL, MacDonald J, Shariff AR and Padgette SR. (1999) Compositional analysis of glyphosate-tolerant soybeans treated with glyphosate. *Journal of Agriculture and Food Chemistry*. 47: 4469-4473.

Vodkin LO, Khanna A, Shealy R, Clough SJ, Gonzalez DO, Philip R, Zabala G, Thibaud-Nissen F, Sidarous M, Stromvik MV, Shoop E, Schmidt C, Retzel E, Erpelding F, Shoemaker RC, Rodiguez-Huete AM, Polacco JC, Coryell T, Kleim P, Gong G, Lui L, Pardinas J and Schweitzer P. (2004) Microarrays for global expression constructed with a low redundancy set of 27,500 sequenced cDNAs representing an array of developmental stages and physiological conditions of the soybean plant. *BMC Genomics*. 5(1): article 73, 1-18.