Exploration of high-density oligoarrays as tools to assess substantial equivalence of genetically modified crops

Julie Beaulieu
Plant Science Department
McGill University, Montreal

August 2005

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

Your file Votre référence
ISBN: 978-0-494-24614-6
Our file Notre référence
ISBN: 978-0-494-24614-6

I. ACKNOWLEDGEMENTS

As thesis supervisor, **Dr. Marc G. Fortin**, guided me through this challenging research project while giving me the freedom to explore the many aspects of gene expression profiling. **Fabrice Langlois** provided indispensable advice and support in information systems-related technology. Thanks to my family, **Isabelle, Eric** and **Samantha Beaulieu, Cecile Brazeau, Claude Beaulieu, Diane Labelle**, and **Lorraine Nagy** for their support and encouragements. Thanks to lab mates **Karine Thivierge, Sophie Cotton, Genevieve Morin** and **Philippe Dufresne** for their support and encouragements. Special thanks to open source software developers (R, Bioconductor, Linux).

Technical contributions

Christine Ide, technician in the Plant Molecular Genetics laboratory, extracted RNA and performed PCR. The microarrays were processed at the McGill University and Génome Québec Innovation Centre.

Financial contributions

This work was supported by a grant from the Advanced Food and Materials Network.

II. ABSTRACT

Since the early 1990s, the concept of substantial equivalence has been a guiding principle of the Canadian Food Inspection Agency and Health Canada's regulatory approach toward products of plant biotechnology destined for the food and livestock feed markets. To assess substantial equivalence in terms of chemical composition, genetically modified (GM) plants are compared to conventional counterparts at the level of macro- and micro-nutrients, allergens and toxicants. Such targeted comparative analyses are limited in their scope and their capacity to detect unintended changes in chemical composition. There is a need to develop more effective testing protocols to improve the substantial equivalence assessment of GM crops. The objective of this thesis was to explore high-density oligoarrays as tools to assess substantial equivalence of Roundup Ready™ soybean. Three conventional and two GM soybean varieties were selected according to the similarity of their performance in field trials. Total RNA was extracted from first trifoliate leaves harvested from soybean plants grown in a controlled environment until the V2 stage. To annotate the 37 776 soybean probesets present on the multi-organism Soybean Affymetrix GeneChip™, consensus sequences were aligned with TIGR Soybean Gene Index tentative consensus sequences using BLASTN. After redefining the chip description file to exclude non-soybean probesets, the effects of three different normalization methods (Robust Multichip Average (RMA), Microarray Analysis Suite (MAS 5.0) and Model-Based Expression Index) were compared and Significance Analysis of Microarrays (SAM for R-Bioconductor) was applied to detect differential gene expression between conventional and GM soybean varieties. Eleven candidate genes were selected for further studies.

III. RÉSUMÉ

Depuis le debut des années 1990, le concept de l'équivalence en substance est au coeur de la réglementation introduite par l'Agence canadienne d'inspection des aliments et Santé Canada à l'égard des produits de la biotechnologie végétale. Afin d'évaluer l'équivalence en substance de plantes génétiquement modifiées (GM) en matière de composition chimique, la composition chimique des plantes GM est comparée à celle de plantes similaires issues de techniques d'amélioration génétique conventionelles en matière de macronutriments et micronutriments, allergènes et toxines. Parce qu'elles sont ciblées, de telles analyses comparatives sont limitées dans leur portée et leur capacité à identifier des modifications non désirées dans la composition chimique des plantes GM. Des protocoles d'analyse plus poussés doivent être mis en place afin d'améliorer l'évaluation de l'équivalence en substance des plantes GM. L'objectif de cette thèse était d'explorer la possibilité d'appliquer la technologie des microréseaux d'ADN à l'analyse de l'équivalence en substance du soya Roundup Ready®. Afin de comparer les variétés de soya conventionelles aux variétés GM, cinq variétés furent sélectionnées sur la base de leur performance lors d'essais en champ. Les plants de soya furent cultivés en chambre de croissance jusqu'au stade V2 et les ARN furent purifiés des premières feuilles trifoliées. Afin d'identifier les 37 776 séquences d'ADN de soya représentées sur les microréseaux Soybean Affymetrix GeneChips®, une recherche de similitudes entre ces séquences et les séquences de soya repertoriées dans le TIGR Soybean Gene Index fut effectuée en ayant recourt à l'algorithme BLASTN. Trois méthodes de normalisation des données (Robust Multichip Average, Microarray Analysis Suite et Model-Based Expression Index) furent comparées et la technique Significance Analysis of Microarrays fut appliquée afin d'identifier des gènes exprimés à différents niveaux entre les variétés conventionnelles et les variétés GM. Onze gènes candidats furent retenus pour fins d'études subséquentes.

IV. LIST OF TABLES

V. LIST OF FIGURES

VI. LIST OF ABBREVIATIONS

dChip	Model-based expression index algorithm
EPSPS	5-enolpyruvyl-shikimate-3-phosphate synthase
EST	Expressed sequence tag
FDR	False discovery rate
GM	Genetically modified
IM	Ideal mismatch
MAS	Affymetrix Microarray Analysis Suite 5.0
MBEI	Model-Based Expression Index
MM	Mismatch oligonucleotide probe
mRNA	Messenger RNA
MvA	Minus *vs.* Average plot
PM	Perfect match oligonucleotide probe
PNT	Plant with novel trait
RMA	Robust Multichip Average
SAM	Significance Analysis of Microarrays

1. INTRODUCTION

1.1 General introduction

A recurrent issue in the safety assessment of genetically modified (GM) crops is the paucity of analytical methods to detect unintended or unexpected outcomes of genetic modification. The aim of safety assessment is to evaluate the *substantial equivalence* of the GM crop and a conventional counterpart [*i.e.* identify similarities and differences between the GM crop and a comparator which benefits from a history of safe use] in regard to agronomic, physiological and compositional characteristics. There are two analytical approaches to defining the composition of plants and plant products which are either based on *targeted* or *profiling* analysis methods. The widely accepted and applied *targeted* analysis methods quantify predefined classes of compounds (*e.g.* macro- and micro-nutrients) as well as known endogenous toxins and allergens. *Profiling* methods, on the other hand, provide indiscriminate analysis of gene expression (mRNA), protein and secondary metabolite composition. Although the latter approach tends to be more comprehensive and hence, better suited to the detection of unexpected effects of genetic modification, it is not currently applied because profiling technologies are still in development.

1.2 Research hypothesis

Gene expression profiling in the form of high-density oligoarray data analysis can be applied to the assessment of substantial equivalence of GM crops.

1.3 Objectives

The main objective of this thesis was to use a novel *profiling* method to assess the substantial equivalence of a GM crop. The specific objectives of this thesis were to:

- Compare gene expression profiles of conventional and GM soybean (*Glycine max* L. Merrill) using high-density oligoarrays as a gene expression profiling tool.
- Apply three different pre-processing methods (Robust Multichip Average, Microarray Analysis Suite 5.0, and Model-Based Expression Index) to extract gene expression measures from high-density oligoarrays.
- Apply statistical analyses to identify differentially expressed genes in GM soybean.

2. LITERATURE REVIEW

2.1 Unintended effects of genetic modifications

Advances in plant molecular biology and genetics in the past 10 years have led to the introduction of GM crops into the food and feed supply. Crops are modified through the transfer or alteration of genes using recombinant DNA technologies (bioballistics, electroporation, *Agrobacterium*-mediated transformation). Typical gene cassettes used in such transformations include a strong promoter (*e.g.* CaMV 35S), a selectable marker gene and the gene of interest which may be altered and/or in reversed orientation.

In the past, the aims of genetic modifications in crops have been to increase tolerance to specific pests (*e.g.* crops expressing the *Bacillus thuringiensis* toxin) or broad-spectrum herbicides such as glyphosate (*e.g.* Roundup Ready crops). A growing trend in the agbiotech industry is to alter more complex traits such as nutritional value (*e.g.* vitamin A rice, high-flavonol tomatoes) and resistance to abiotic stresses (*e.g.* salt or drought tolerant wheat). These modifications have the potential to fundamentally alter plant metabolism resulting in intended as well as unintended changes in chemical composition (Cellini et al., 2004; Konig et al., 2004a). Since the current methods of plant transformation do not offer control over the insertion site, the number of copies transferred, or the integrity of the gene cassette, unintended effects may result from disruption of a functional gene at the point of insertion, rearrangements of the gene cassette or ectopic co-expression of neighboring genes (Windels et al., 2001; Kohli et al., 2003; Cellini et al., 2004; Committee on Identifying and Assessing Unintended Effects of Genetically Engineered Foods on Human Health, 2004; Williams and Bowles, 2004; Dunwell, 2005). Modifications such as the overexpression of transcription factors, introduction or alterations of biosynthetic pathways, expression of transgenes to increase tolerance to biotic or abiotic challenges, all carry the potential for unexpected interactions between gene products as well as increases

or decreases in the availability and activity of other plant biochemicals (Conner and Jacobs, 1999). If these alterations affect the performance of the crop or increase the concentration of a known endogenous toxin, it is expected that the breeding program of the GM line will be halted and thus, these effects will not be reported (Beachy et al., 2002; Cellini et al., 2004). However, unintended alterations of agronomic traits or chemical composition following genetic modification of crop plants have been reported in the scientific literature. Examples are given in Table 1.

2.2 Assessment of substantial equivalence

In 1995, Canadian regulatory agencies began approving GM crops for release into the environment and use as food and livestock feed. Prior to regulatory approval, GM crops considered as Plants with Novel Traits (PNTs) undergo safety assessments (Macdonald and Yarrow, 2003). The Canadian Food Inspection Agency (CFIA), responsible for livestock feed and environmental safety assessments, has determined that one of the main objectives of its GM crop safety assessment is to evaluate "the relative phenotypic expression of the PNT compared to a similar counterpart, where differences are anticipated" based on species-specific biology documents as well as experimental data submitted by the applicant (CFIA, 2004). According to the *Guidelines for the Safety Assessment of Novel Foods,* Health Canada, responsible for the safety of food products, also requests information regarding how the composition of the GM crop compares to that of the unmodified crop (Health Canada, 1994). In addition, an evaluation of the potential for "secondary" effects on biochemistry, physiology and secondary metabolism of the GM crop is conducted as part of the safety assessment. Thus, in the course of safety assessments of GM crops, Canadian regulatory agencies evaluate the *substantial equivalence* of the GM crop and a conventional "counterpart". In other words, regulatory agencies conduct an assessment of substantial equivalence to identify similarities and differences between the GM crop and a "generally recognized as safe" comparator in regard to ecological, agronomic, physiological and compositional characteristics.

Table 1. Examples of unintended effects of genetic modifications in crop plants.

Crop	Genetic modification	Reported unintended effects	Authors
Apple	Overexpression of a fruit-specific polygalacturonase gene	"Profound effects on leaf morphology, plant water relations, stomatal structure and function, and leaf attachment."	(Atkinson et al., 2002)
Canola	Overexpression of a phytoene-synthase gene	Alteration of chlorophyll and tocopherol levels, alteration of fatty acid composition	(Shewmaker et al., 1999)
Potato	Expression of bacterial levansucrase	Lower yield, smaller tubers, modified starch granule morphology	(Gerrits et al., 2001)
Red spring wheat	Expression of 5-enolpyruvyl-shikimate-3-phosphate synthase (EPSPS) derived from *Agrobacterium* spp. CP4	Increased shikimic acid levels in kernels following glyphosate treatments	(Bresnahan et al., 2003)
Rice	Expression of soybean glycinin	Increased vitamin B6 content	(Momma et al., 1999)
Tomato	Expression of two maize transcription factors to increase flavonoid synthesis	"[…]levels of valine and γ-aminobutyric and citric acids, sucrose, nucleosides and nucleotides, phenylalanine, cinnamic derivative 1, and U1 were higher in control tomatoes […]. Similarly, ANOVA confirmed that transgenic tomatoes contain significantly more glutamine, asparagine, flavonoid glycosides, and trigonelline."	(Gall et al., 2003)

When evaluating compositional characteristics, the "counterpart" and the type of analytical methods selected to assess substantial equivalence are key determinants of the outcome of the safety assessment (Kok and Kuiper, 2003). Currently, the recommended comparator of the GM line is the direct parent line (Kok and Kuiper, 2003; Committee on Identifying and Assessing Unintended Effects of Genetically Engineered Foods on Human Health, 2004). However, if a consistent difference in composition is identified between the GM and the parent line, the risk assessment proceeds to evaluate the impact of the alteration by referring to the range of "generally recognized as safe" levels of the particular analyte in other commercial varieties of the crop under review (OECD, 1993). In other words, the risk assessment focuses on differences that are beyond the range of natural variation for a specific analyte in a specific crop (Konig et al., 2004a). To standardize risk assessments, reference to compositional databases has been advocated (Kok and Kuiper, 2003; EFSA, 2004). For example, one such database, the International Life Science Institute Crop Composition Database reports on lectins, isoflavones, trypsin inhibitors, ash, carbohydrates, crude protein, moisture and fat in soybean grains collected in Ontario in 2002 (ILSI, 2005).

In terms of analytical methods, substantial equivalence of GM crops is assessed by Canadian regulatory agencies using *targeted* analyses with an aim to identify and quantify predefined compounds such as lipids, carbohydrates, amino-acids, known toxicants and allergens. The outcomes of these assessments are summarized in publicly available "Decision Documents" (for example: CFIA, 1995a; Health Canada, 1996, 2004). The potential of such *targeted* analyses to identify unintended effects of genetic modification has been put into question (Cellini et al., 2004; Committee on Identifying and Assessing Unintended Effects of Genetically Engineered Foods on Human Health, 2004; Corpillo et al., 2004; Konig et al., 2004b). The integration of *profiling* analyses among the analytical tools currently employed to assess substantial equivalence has been proposed to allow for broader coverage of potential unintended effects (Kuiper et al., 2003; Cellini et al., 2004).

Profiling analyses are based on techniques developed for transcriptomics, proteomics and metabolomics (Fiehn et al., 2001). By applying these techniques to the assessment of substantial equivalence, the idea is to compare molecular profiles and to identify mRNA, proteins or metabolites showing a different pattern of expression/biosynthesis in GM crops (Committee on Identifying and Assessing Unintended Effects of Genetically Engineered Foods on Human Health, 2004). As opposed to the well established and internationally recognized *targeted* methods, *profiling* technologies are still in development (Kuiper et al., 2003). In 2000, a European Union working group, Entransfood (GMOCARE), was commissioned to develop *profiling* analysis methods to assess the substantial equivalence of GM crops (Cellini et al., 2004). In their report, the participants acknowledged that there are still important hurdles in the application of the *profiling* approach to the safety assessment of GM crops (Konig et al., 2004b). However, proteomic and metabolomic techniques, more precisely two-dimensional electrophoresis (Corpillo et al., 2004; Lehesranta et al., 2005) and nuclear magnetic resonance spectroscopy (Gall et al., 2003; Charlton et al., 2004; Manetti et al., 2004), have been successfully applied to the characterization of GM crops. The objective of this thesis was to explore the use of a gene expression profiling technique for the substantial equivalence assessment of GM soybean.

2.3 Gene expression profiling

Gene expression profiling is arguably the most comprehensive profiling method currently available (Lehesranta et al., 2005). Gene expression profiling is used to estimate the relative levels of mRNA species throughout the RNA population of cells and tissues and to explore patterns of transcription (Hughes and Shoemaker, 2001). The importance of transcriptional activity in cellular biology is best described in this excerpt:

"The transcription of genomic DNA to produce mRNA is the first step in the process of protein synthesis, and differences in gene expression are responsible for both morphological and phenotypic differences as well as indicative of cellular responses to environmental stimuli and perturbations. Unlike the genome, the transcriptome is highly dynamic and changes rapidly and dramatically in response to perturbations or even during normal cellular events such as DNA replication and cell division. In terms of understanding the function of genes, knowing when, where and to what extent a gene is expressed is central to understanding the activity and biological roles of its encoded protein. In addition, changes in the multi-gene patterns of expression can provide clues about regulatory mechanisms and broader cellular functions and biochemical pathways" (Lockhart and Winzeler, 2000).

Given that they provide powerful insight into transcriptional activity, expression profiling techniques are increasingly relied upon in plant molecular biology in which they have received wide application. For example, gene expression profiling has been used to characterize changes in gene expression induced by overexpression of transcription factors (Zik and Irish, 2003), to monitor responses to biotic and abiotic challenges (Whitham et al., 2003; Seki et al., 2004), and to study differential gene expression associated with plant growth and development (Zhang et al., 2005).

The integration of gene expression profiling techniques among tools to assess substantial equivalence is a product of the assumption that genetic modification has the potential to alter the transcriptome of a GM plant in unexpected ways. Examples of such pleiotropic effects were observed in the course of a cDNA-AFLP study on the RNA profiles of *Nicotiana benthamiana* plants agro-infiltrated with two commonly used reporter genes *uidA* (encoding ß-glucuronidase) and *gfp* (encoding green fluorescent protein) (Page and Angell, 2002). In the case of *gfp*, pleiotropic effects (in the form of differential gene expression) were noticeable

irrespective of cellular localization of GFP while, in the case of *uidA*, differential gene expression occurred when the protein products were targeted to the endoplasmic reticulum. These findings were surprising given that *uidA* and *gfp* are of bacterial and fish origin, respectively, and are not known to interfere with transcriptional activity in plants (Jefferson et al., 1987; Stewart, 2001). In a microarray-based study, overexpression of a transcription factor involved in floral organogenesis in *Arabidopsis thaliana* loss-of-function mutants was shown to induce differential gene expression of hundreds of genes, many of which with unknown function (Gomez-Mena et al., 2005). Finally, introduction of the entire cyanogenic glucoside (dhurrin) biosynthesis pathway from *Sorghum bicolor* into acyanogenic *Arabidopsis* showed very little impact on global gene expression using both focused and global microarrays (Kristensen et al., 2005). However, considerably more differential gene expression was observed using the focused microarrays when only part of the pathway was present.

A variety of methods have been developed to capture gene expression "snapshots". The completion of large scale EST sequencing efforts coupled to the centralization of facilities and standardization of protocols has led to the predominant use of microarrays for gene expression profiling (Meyers et al., 2004). In the plant community, microarrays are the choice platform for large scale gene expression profiling projects such as the European Union Compendium of Arabidopsis Gene Project, Arborea (poplar), Medicago, rice, and maize microarray projects.

There are many types of two-dimensional microarrays. The first eukaryote DNA microarrays, developed in the early 1990s, contained plant gene sequences attached to a microscope slide (Schena et al., 1995). To obtain a gene expression profile, these probes were hybridized to fluorescently-labeled messenger RNA (mRNA). This technology is still in use today in the form of cDNA polymerase chain reaction (PCR)-amplified products or 30 to 70-mer oligonucleotides printed on glass slides as spots and hybridized to a mixture of two cDNA samples

labelled with a red (Cyanine 5) or green (Cyanine 3) fluorescent dye (Schulze and Downward, 2001). A second innovation in DNA microarrays came in the form of high-density synthesis of short oligonucleotides on glass wafers using a photolithographic masking technique (Fodor et al., 1991). Today such high-density oligoarrays, known as GeneChips, are manufactured by Affymetrix, and contain more than one million 25-mer oligonucleotide probes within a 1.28 cm^2 quartz wafer (Schulze and Downward, 2001). The main advantages of the Affymetrix technology reside in the high reproducibility of *in situ* synthesis of oligonucleotides and the incorporation of biotin-labelled nucleotides as opposed to fluorescent cyanine dyes during target preparation. Inconsistent fluorescence of the Cyanine 5 (red) dye, which varies according to ambient ozone levels, was shown to affect the reproducibility of experiments conducted with cDNA microarray (Fare et al., 2003). Also, *in situ* synthesis of oligoarrays offers more control on the selection of probe sequences, thus limiting the effects of noise due to cross-hybridization as well as the potential for probe tracking mistakes (Wang et al., 2003; Alba et al., 2004). Finally, differential expression observed with GeneChips was validated more readily by quantitative real-time PCR (qRT-PCR) in cross platform comparisons (Li et al., 2002; Mah et al., 2004; Park et al., 2004; Larkin et al., 2005).

Large scale genome and expressed sequence tags (EST) sequencing projects were among the necessary precursors of the Affymetrix GeneChip technology. When the full length sequence of a genome has not been characterized, consensus sequences derived from large scale EST sequencing projects serve as a template for probe design. ESTs are obtained by sequencing the 5' and/or 3' end of double stranded cDNA fragments inserted into cloning vectors (Adams et al., 1991). The cDNA fragments are prepared from randomly isolated mRNA, and thus only represent genes that were expressed in the originating tissues (Alba et al., 2004). Non-redundant consensus sequences are defined by aligning ESTs according to pre-determined parameters (Lee et al., 2005). On GeneChips, each consensus sequence is represented by a collection of 11 to 20 complementary 25-mer

oligonucleotide probes (called "probesets") designed to be as sequence specific as possible to minimize the potential for cross-hybridization and uniform enough to conserve a constant guanosine triphosphate (GTP) and cytosine triphosphate (CTP) content (Lipshutz et al., 1999). For each perfect match probe (PM), there is a corresponding mismatch probe (MM) with a single homomeric base change at the 13th position meant to serve as a control for non-specific binding. The scattering of probe pairs across the array also reduces the potential for location-based imbalances. The location of each probe on the chip, the type (PM or MM), total GTP and CTP content, and probeset membership is detailed in a chip description file (CDF).

For target preparation, total RNA is extracted from homogenized tissue and converted to double stranded cDNA (ds cDNA) (Affymetrix, 2004). Double-stranded cDNA is obtained through reverse transcription of mRNA transcripts using an oligo(dT) primer for first strand synthesis. The T7 RNA-polymerase promoter appended to the oligo(dT) primer allows for *in vitro* transcription incorporating biotinylated nucleotides (CTP and UTP). Each target cRNA preparation is fragmented and hybridized to a separate array. Target binding is detected by staining with a fluorescent dye coupled to streptavidin. A laser scanner captures the fluorescence emitted at the location of each probe (composed of millions of identical 25-mer oligos) and converts the 75th percentile of pixel intensity into a signal that is used to calculate relative mRNA abundance in the original samples. The results (signal intensities) are recorded in *.CEL files.

2.4 Pre-processing steps for GeneChip data

Experiments in which known concentrations of cRNAs were "spiked" into target samples have shown that the intensity of the signal does not reflect the absolute concentration of mRNAs in the original sample (Chudin et al., 2001; Irizarry et al., 2003a). Overall processing effects (target preparation, hybridization, washing and staining efficiencies, GeneChip image variation), non-specific binding, and

probe affinity for the substrate introduce bias in the signals from each array. Therefore, pre-processing steps such as background correction (within chip) and scaling/normalization (between chips) are necessary prior to statistical analysis of differential expression (Bolstad et al., 2003).

Different models have been developed to correct for non-biological sources of variation. The Microarray Analysis Suite 5.0 (MAS) statistical algorithm from Affymetrix (Hubbell et al., 2002; Liu et al., 2002), the Model Based Expression Index (a.k.a. dChip) of Li and Wong (Li and Wong, 2001a) and the Robust Multi-chip Average (RMA) of Irizarry *et al* (Irizarry et al., 2003b) stand out as the three most widely used methods for high-density oligoarray data pre-processing (Han et al., 2004; Choe et al., 2005; Fan et al., 2005; Li et al., 2005).

In MAS, background estimates are obtained from the lowest 2% probe intensities in each of 16 rectangular regions on the array. A weighted average of these values (relative to the distance of the probe from the centre of each region) is subtracted from each probe signal intensity. The signal for each PM probe is further processed by subtracting either the MM probe intensity from the corresponding PM probe intensity or an ideal mismatch (IM) value when MM $\geq$ PM. Following log transformation of the difference between the PM and MM (or IM), a single value for each probeset is obtained using a Tukey bi-weight average. A linear normalization (applying the same scaling factor to all probesets on one array) is then performed by scaling all the arrays to a particular target signal value (mean) or a particular baseline array (Hubbell et al., 2002; Liu et al., 2002).

For normalization, dChip uses the "invariant set" method wherein a set of "invariant" PM probes is selected by ranking all PM probes on a target and baseline array according to signal intensity and calculating a proportional rank difference (absolute rank difference over total number of PM probes) iteratively until there is no difference between sets. These "invariant" PM probes are assumed to represent probes from non-differentially expressed probesets. A

smooth (running median) curve is calculated and then used to generate new normalized values for each probe on the chip (Li and Wong, 2001a). After normalization, the dChip method models the PM probe signal intensities according to the following equation:

$$PM_{ij} = v_j + \theta_i \Phi_j ,$$

where $i = 1,\ldots,$ I is the number of arrays in a dataset, $j = 1, \ldots,$ J is the number of probes in a probeset, v_j represents non-specific hybridization, θ_i is a model based expression index (MBEI), and Φ_j is a probe sensitivity index (Li and Wong, 2001b). The probe sensitivity index (Φ_j) is included in the model to account for the larger variation in probe signal intensities within a probeset relative to the variation in probe signal intensities within a dataset composed of many replicate arrays (Li and Wong, 2001a). Thus, the MBEI is a weighted average of PM probe intensities within a probeset in which the weights are defined by the variability of each probe signal intensity within the probeset. Iterative least squares fitting is used to estimate the parameters.

The last normalization method, RMA, was modeled to reflect observed probe intensities in a spike-in/dilution series dataset where background noise was tightly controlled. In this model, signal intensities are adjusted according to a background correction method that is based on the distribution of PM probe intensities on the natural scale (Irizarry et al., 2003a). Observed PM intensities are decomposed into a true signal component (exponentially distributed) and a noise component (normally distributed). Background correction is followed by a form of intensity-based normalization whereby the PM probe intensities are ranked in each array, ranks are averaged and the original probe intensities are replaced by these rank averaged intensities (quantile normalization). Following $\log_2$ transformation of background adjusted and normalized PM intensities, a signal for each probeset is obtained using a Tukey median polish.

The performance of each of these pre-processing methods was compared using spike-in/dilution series datasets and experimental datasets. A common

assumption in gene expression profiling is that most genes in a control *vs.* treatment dataset are not differentially expressed and among the differentially expressed genes, there are just as many up-regulated as down-regulated genes (Draghici, 2003; Bolstad et al., 2004). The purpose of pre-processing is to adjust signal intensities so that the difference in expression ("fold change") on a log scale between most corresponding probesets in a dataset will reflect this assumption by being close to zero across all signal intensity levels.

One way to assess the performance of different normalization methods is to draw Minus vs. Average (MvA) plots, where the difference between probeset signal intensities or "fold-change" is plotted against the average signal intensity of corresponding probesets. Ideally, most points should fall along the M = 0 axis across average signal intensity levels, A. Typically, lower signal intensity probesets show more deviation from M = 0 axis using MAS and dChip, which implies that genes expressed at lower levels may be systematically included among lists of differentially expressed genes. The main advantage of RMA is a compression of these data points along the M = 0 axis, suggesting increased precision in estimating signal intensities (Irizarry et al., 2003a; Han et al., 2004). However, this compression of the data may come at a cost of accuracy when estimating fold change for low signal intensity probesets, especially in "noisy" experimental datasets (Seo et al., 2004).

In a spike-in/dilution series study, the concentration of spiked-in cRNAs was more accurately reflected in observed signal intensities for RMA than the other two methods, MAS and dChip (Irizarry et al., 2003b). However, an experiment in which a noisier experimental dataset was pre-processed using these three different methods followed by statistical analysis of differential gene expression and validation with qRT-PCR showed that RMA was less sensitive than the other two methods (Li et al., 2005). But, the lower sensitivity of RMA might have been a factor of the low sample number (N = 6) in this experimental dataset instead of "noisiness", as suggested by the authors. In another study, involving a total of 75

experimentally-derived samples, a list of expected differentially expressed genes served as a benchmark to assess the sensitivity of each pre-processing method (Galfalvy et al., 2003). The authors concluded that the sensitivity of RMA was superior to that of MAS and dChip.

2.5 Methods for selecting differentially expressed genes

Once the data from GeneChips is adjusted for background noise and normalized, with a single expression measure extracted for each probeset, these expression measures may be compared to obtain lists of differentially expressed genes. Here, microarray studies are confronted with issues of multiple testing of large numbers of probesets (> 20 000 on most GeneChips) in much smaller samples (typically 3 per condition). The first attempt at differential gene expression analysis consisted of ranking probesets according to observed fold changes (*i.e.* the ratio of signals intensities between corresponding probesets) and selecting an arbitrary criterion (usually two-fold) as a threshold for inclusion in the list of differentially expressed genes (Lee et al., 1999). This approach was found to be unreliable because it did not take random variability into account which resulted in a large number of falsely labeled differentially expressed genes (Miller et al., 2001; Yang et al., 2002; Draghici, 2003). Since then, a variety of statistical methods have been applied to find differentially expressed genes between two experimental conditions (control *vs.* treatment or, in our case, GM crop *vs.* conventional crop). In essence, these methods rely on classical hypothesis testing on a gene-per-gene basis and offer some measure of control over Type I and Type II error.

Again, there is no "gold standard" for statistical analysis of differential gene expression. The classical t-test (assuming unequal variances and number of samples in each group) is often applied on a probeset-specific basis according to the following model:

$$t_i = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{s_i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $\bar{x}_{i1}$ and $\bar{x}_{i2}$ are the group means for each probeset, n is the total number of samples in a dataset, n_1 and n_2 are the number of samples in each group, and s_i is the pooled standard deviation (Simon et al., 2003).

The advantage of a gene-specific t-test is that it is not affected by fluctuations in variance across genes. As mentioned in the previous section, variation in probe signal intensity decreases with increasing mean signal intensity. The main disadvantage of the t-test is the family-wise error-rate (FWER) associated with the generation of multiple p-values (Slonim, 2002). If, for example, the p-value is set at 0.05 and there are 40 000 probesets on a GeneChip, the expected number of false positives among the list of differentially expressed genes is 0.05*40 000 = 2000 differentially expressed genes (*i.e.* there's a 5% chance of committing a Type I error each time the null hypothesis is tested) (Leung and Cavalieri, 2003). A popular multiple testing adjustment, the Bonferroni procedure, guarantees a smaller FWER by setting the p-value so as to obtain only 5% false positives on the whole dataset (0.05/40 000 = 0.00000125 in the previous example). Such low p-value thresholds lower the power of the statistical test and are generally deemed too conservative when applied to microarray data (Miller et al., 2001; Draghici, 2003).

In recent years, a "penalized' t-test, the Significance Analysis of Microarrays (SAM), has become one of the most widely used alternative to the classical t-test (Wu and Irizarry, 2005). SAM relies on a classical gene-specific *t*-test with a small positive constant (s_0) added to the denominator in order to stabilize the variance across pre-processed probeset intensities

$$d_i = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{s_0 + s_i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

This constant (s_0) is computed according to the distribution of the d_i scores and the standard error of the difference between means. It aims to minimize the coefficient of variation of d_i as a function of s_i in moving windows across the data from all arrays in the two-class comparison. To find potential differentially expressed genes, the d_i scores are compared to average d_i scores obtained from permutated datasets. The average d_i scores across permutations are denoted as

$$\bar{d}_i = \sum_{b=1}^{B} \frac{d_i^b}{B},$$

where b=1,..., B is the total number of permutations. A user defined cutoff delta determines which genes are significant

$$\left| d_i - \bar{d}_i \right| \geq \Delta .$$

The typical SAM plot projects the rank ordered d_i scores against $\bar{d}_i$ scores resulting in a straight line that goes through the x and y intercepts at (0,0). Points to the left of the y intercept represent down-regulated genes (original d_i score negative) while the points to the right represent up-regulated genes (original d_i score positive). Correspondingly, significant probesets are located above or below the user defined cutoff delta.

This method also controls the number of falsely labeled differentially expressed genes by calculating the false discovery rate (FDR). Here, the FDR is the number of expected false positives among differentially expressed genes. To calculate the FDR, the smallest original d_i score among up-regulated probesets and the highest original d_i score among down-regulated probesets in the list of differentially expressed genes is noted

$$d_0 = \max_{d_i \leq \bar{d}_i - \Delta} d_i, d_1 = \min_{d_i \geq \bar{d}_i + \Delta} d_i .$$

For each permutation set, the number of probesets whose d_i scores are either higher than the lowest original positive score cutoff (d_0) or lower than the highest original negative score cutoff (d_1) is noted. The numbers are ranked and the

median is calculated to obtain the median number of false positives. The median number of false positives is multiplied by π_0, the proportion of true null probesets, to obtain $V(\Delta)$:

$$\hat{\pi}_0 = \frac{\#\{d_i(original) \in (q25, q75)\}}{(.5p)},$$

$$V(\Delta) = median\ number\ of\ false\ positives(\pi_0),$$

where p is the number of original d_i scores multiplied by the number of permutations. The median false discovery rate (FDR) is estimated as

$$\hat{FDR}(\Delta) = \frac{V(\Delta)}{R(\Delta)} * 100,$$

where

$$R(\Delta) = \sum_i \left\{ \left| d_i - \bar{d}_i \right| \geq \Delta \right\},$$

is the total number of significant genes.

The performance of a differential gene expression test statistic is intimately related to the methods employed during pre-processing steps. The combination of RMA and SAM was shown to have higher sensitivity and specificity in the detection of differentially expressed genes when applied to a spiked-in dataset compared to the combination of MAS and SAM (Wu and Irizarry, 2005; Yang et al., 2005). Also, the combination of RMA and SAM seemed to outperform the combination of RMA and the classical t-test (Yang et al., 2005). A recent study in which a list of potential differentially expressed genes was determined *a priori* also compared the ability of different combinations of pre-processing and statistical analysis methods to detect these differentially expressed genes in a relatively more "noisy" experimental dataset. The MAS and SAM combination identified 17/20 predetermined differentially expressed genes, the dChip and SAM combination identified 9/20 pre-determined differentially expressed genes, and the RMA and SAM combination identified 5/20 pre-determined differentially expressed genes (Li et al., 2005). The authors of the study attributed the lower sensitivity of the dChip and MAS as well as RMA and MAS combinations to the "noisiness" of their biologically-derived dataset compared to the spiked-

in/dilution series dataset. Since dChip and MAS require multiple samples to estimate signal intensity precisely, the lower sensitivity of the two combinations applied in the Li *et al.* (2005) study may have been the product of an insufficient number of samples (total of 6 GeneChips) in the dataset rather than the "noisiness" of the dataset.

2.6 The Affymetrix Soybean GeneChip

The soybean GeneChip was designed in collaboration with the Soybean Research Community as part of the Affymetrix GeneChip Consortia Program. Three organisms are represented on the chip: 37 594 probesets were designed from *Glycine max* publicly available EST and mRNA sequences, approximately 15 800 probesets belong to the water mold *Phytophtorae sojae* and approximately 7500 probesets belong to the cyst nematode *Heterodera glycines*. Each probeset is composed of 11 PM probes and 11 MM probes. Including the Affymetrix spike-in control probesets, there are 61170 probesets and 1 354 900 probes on the chip.

NCBI Soybean UniGene Build 13 (November 5, 2003) served as a template for the design of the soybean probesets. This Unigene Build assembled 184 912 publicly available 5' ESTs, 42 570 3' ESTs, 4334 ESTs of unknown orientation, and 837 mRNAs (P. Cooper, NCBI, personal communication). The soybean ESTs deposited in dbEST were generated through the large scale Soybean Public EST Project (Shoemaker et al., 2002). Members of the Soybean Public EST Project created the soybean cDNA libraries (approximately 80) from tissues representing different plant developmental stages, organs, genotypes (mostly Williams 82), and biotic and abiotic stresses (Shoemaker et al., 2002).

2.7 Soybean as a model crop plant

Soybean ranks as one of the most important grains and oilseed crops in Canada with an estimated seed production of 3 million tonnes in 2004 (Statistics Canada,

2004). According to Chris Beckham, oilseeds analyst at Agriculture and Agri-Food Canada, approximately 50 to 65% of soybean grown in 2004 was GM (personal communication). In 2002, a Statistics Canada survey estimated GM soybean at around 30% of the total production in Ontario and Quebec (Hategekimana and Beaulieu, 2002). In comparison, in 2004, 85% of the US production estimated at 85 million tonnes was GM (NASS, 2004).

All commercial GM soybean have been transformed to resist the broad-spectrum herbicide glyphosate, an amino acid analogue (Duke, 2005). A unique transformation event, soybean breeding line 40-3-2 (Asgrow A5403, Monsanto, St-Louis, MO), was approved by Canadian and US regulatory agencies for environmental release and use as food and feed (Canada 1995, USA 1994). By law, this glyphosate-resistant variety became the only source of glyphosate resistance in subsequent breeding programmes (Raymer and Grey, 2003; Sneller, 2003). Worldwide adoption rates of these glyphosate-resistant soybean varieties have been rising steadily over the past 10 years (James, 2004; Duke, 2005).

Glyphosate targets a key step in the aromatic amino acids biosynthesis pathway which is present in all plants and some bacteria, including *Bradyrhizobium japonicum* (Zablotowicz and Reddy, 2004). Instead of catalyzing the transfer of the enolpyruvyl moiety of phosphoenolpyruvate (PEP) to shikimate-3-phosphate, the enzyme 5-enolpyruvyl-shikimate-3-phosphate synthase (EPSPS) binds glyphosate, forming a stable compound and effectively blocking the aromatic amino acid pathway (Barry et al., 1992). Since it is not inhibited by glyphosate, constitutively overexpressed *Agrobacterium* spp. CP4-EPSPS confers resistance to the herbicide by supplementing endogenous EPSPS. The CP4-EPSPS gene was introduced by particle bombardment of a gene cassette containing a portion of the 35S cauliflower mosaic virus promoter (CaMV), the *Petunia hybrida* EPSPS chloroplast transit peptide, the CP4-EPSPS coding sequence and a portion of the 3' untranslated region of the nopaline synthase gene (*nos*) terminator (Padgette et al., 1995). A more detailed characterization of border sequences

revealed a 254 nucleotide portion of CP4-EPSPS and a 534 nucleotide sequence of unknown origin flanking the 3' *nos* terminator (Windels et al., 2001).

Proximate analysis, amino acid and isoflavone analyses were performed on the original glyphosate-resistant line (40-3-2) and its parent (Asgrow A5403) and no significant differences were found between the two (CFIA, 1995a; Health Canada, 1996; Padgette et al., 1996; Taylor et al., 1999). A higher susceptibility to *Fusarium solani* and water stress was reported for glyphosate treated soybean (Sanogo et al., 2000; King et al., 2001; Sanogo et al., 2001). Withholding applications of glyphosate, Elmore et al. reported a 5% yield reduction in glyphosate-resistant soybean lines compared to untransformed soybean sister lines (Elmore et al., 2001).

3. MATERIALS AND METHODS

3.1 Selection of soybean varieties

To isolate the effect of a particular genetic transformation, the most efficient experimental design consists of comparing the GM and parental lines simultaneously grown under identical conditions (Kok and Kuiper, 2003). However, in the particular context of safety assessment of GM crops, the conventional "counterpart" is expanded to include a variety of genotypes to determine if the magnitude of observed effects is within the range of naturally observed variation among crops that have a history of safe use. This strategy was adopted by the CFIA in numerous safety assessments, including those of glyphosate-resistant soybean, canola, and corn (CFIA, 1995a, 1995b, 1999). Therefore a sampling of soybean varieties were selected for this study according to the following criteria: regulatory approval of the GM varieties for unconfined release, similar performance of the GM and conventional varieties in field trials, and history of safe use.

Four mid to late (2550-2750 crop heat units) soybean varieties were selected for this experiment on the basis of the similarity of their performance in the CRAAQ and OOPSC field trials from 2001-2004 in terms of yield and days to maturity (CRAAQ, 2002-2005; OOPSCC, 2002-2005). Two varieties, OAC Bayfield and S03-W4, are the products of conventional breeding while the other two varieties, 2601R and PS46R, are descendants of the glyphosate-resistant soybean 40-3-2 line. The fifth variety, Mandarin (Ottawa), was released in 1934 and is a major ancestor of North American varieties, having contributed 18-55% to the genomes of present-day varieties (Lohnes and Bernard, 1991; Kisha et al., 1998). Sharing the same maturity group, Mandarin (Ottawa) and OAC Bayfield have already been compared in two genetic improvement studies (Kumudini et al., 2002; Cober et al., 2005).

Soybean variety OAC Bayfield was developed by the University of Guelph and registered in 1993 (Tanner et al., 1998). Soybean variety S03-W4 was developed by Syngenta Seeds Inc. (Minneapolis, MN) and registered in 1998. Soybean variety 2601R is a descendant of the glyphosate-resistant 40-3-2 line, and was registered by First Line Seeds Ltd. (Guelph, ON) in 1998 (CFIA, 2005). Soybean variety PS46R is also a descendent of 40-3-2 and was registered by First Line Seeds Ltd. (Guelph, ON) in 2000 (CFIA, 2005). Finally, Mandarin (Ottawa) was obtained from the Eastern Cereal and Oilseed Research Centre, Agriculture and Agri-Food Canada.

3.2 Growth conditions

OAC Bayfield, 2601R, PS46R, S03-W4 and Mandarin soybean were sown in 20 cm plastic pots in Premier Horticulture Promix BX (Promix, Dorval, Canada) previously autoclaved at 80°C for 1 hour. For each variety, 10 seeds were sown in 2 pots, and the pots were placed on either side of an E15 plant growth chamber (Conviron, Winnipeg, Canada). Conditions were set as follows: ambient humidity, 16 hour photoperiod, and 25/19°C day/night temperatures. Peak light intensity (540 $\mu mol \cdot m^{-2} \cdot sec^{-1}$) was measured with an LI-1800 portable spectroradiometer (LI-COR, Lincoln, NE). Plants were watered with distilled water as required.

3.3 Plant growth monitoring and harvest of plant material

Growth monitoring was done on a daily basis by noting the number of plants at each growth stage as defined by Fehr *et al.* (1971). At the V1 stage, the unifoliate leaves are completely unrolled. At the V2 stage, the first trifoliate leaf is completely unrolled. Leaves are considered completely unrolled when the outer leaflets of the leaf at the node directly above are no longer touching (Fehr et al., 1971). When more than 50% of the plants in one variety had reached the V2 stage, completely unrolled first trifoliate leaves were harvested by cutting the

petiole a few millimeters below the leaflets. The leaves were immediately frozen in liquid nitrogen and stored at -80°C.

3.4 PCR screening

Second trifoliate leaves were collected from each plant and PCR was performed to confirm/infirm the presence of CP4-EPSPS using primers (*sttmf3a* and *sttmr2a*) designed by Padgette *et al.* (1995). Equal amounts of leaf material (approximately 3 mg) were mixed and ground for the pooled CP4-EPSPS detection experiments. DNA was extracted from 100 mg of leaf material using the DNeasy Plant Mini kit (Qiagen, Valencia, CA). PCR reactions were prepared by mixing 5 µl of DNA to the following Invitrogen PCR reagents (Invitrogen, Carslbad, CA): 2.5 µl of 10x buffer [200 mM Tris HCl (pH 8.4), 500 mM KCl], 2.5 µl of 2mM dNTP, 0.5 µl of 50mM MgCl2, 1.25 units of Platinum TAQ, 2.5 µl each of 5 uM *sttmf3a* and *sttmr2a*, and 9.25 µl of H_2O. The PCR program was as follows: 4 min. at 94°C, 30 cycles of 30 s. at 94°C, 30 s. at 55°C, and 30 s. at 72°C, followed by 5 min. at 72°C.

3.5 RNA extraction and RNA quality assessment

The samples were randomly assigned to six RNA extraction groups. Total RNA was prepared from trifoliate leaves, previously ground in liquid nitrogen with a mortar and pestle, using the RNeasy Plant Mini Kit (Qiagen). RNA integrity was tested for each sample using the Agilent 2100 bioanalyzer (Palo Alto, CA). Electropherograms were generated using the Agilent 2100 bioanalyzer Expert 2100 Software (Agilent).

3.6 GeneChip expression profiling

Five samples of total RNA from each soybean variety were selected for hybridization to Affymetrix Soybean GeneChips (total 25 chips) following total

RNA integrity assessment. Target preparation, hybridization and scanning were carried out at the McGill University and Genome Quebec Innovation Centre Microarray platform using the protocol recommended by Affymetrix (Affymetrix, 2004). Briefly, 5 μg of total RNA was used to generate double-stranded cDNA using a T7-linked oligo(dT) primer and SuperScript II reverse transcriptase (Invitrogen) following the instructions for the One cycle-cDNA synthesis kit (Affymetrix). cRNA were synthesized using the IVT labelling kit from Affymetrix, resulting in biotinylated cRNA. Labelled cRNA were cleaned and fragmented using the Sample Cleanup Module reagents (Qiagen). Spike controls B2, bio-B, bio-C, bio-D, and Cre-x were added to the hybridization cocktail before overnight hybridization at 45°C for 16 h. Arrays were washed and stained in an Affymetrix Fluidics Station prior to scanning on the GeneChip Scanner 3000 (Affymetrix). Image acquisition and processing was done with the Microarray Analysis Suite 5.0 (Affymetrix).

3.7 Data pre-processing and statistical analysis

Data pre-processing and statistical analysis was done using packages in R-Bioconductor (Gentleman et al., 2004). All computations were carried out on a desktop PC (P4) running the Debian Linux operating system and equipped with 2 gigabytes of random access memory (RAM).

To remove *P. sojae* and *H. glycines* probes, an alternative chip environment was created in R-Bioconductor using the *altcdfenvs* package (Gautier, 2004). The dataset was pre-processed using functions available in the *affy* package (Gautier et al., 2004). For MAS, dChip and RMA, the parameters of the *expresso* function were applied as described in the section "Pre-processing steps for GeneChip data". Since MAS and dChip source code are not publicly available, the implementation of those algorithms in the *affy* package was not expected to yield exactly the same results as the published algorithms (Li and Wong, 2001a; Hubbell et al., 2002; Liu et al., 2002).

Statistical analysis was performed using the package Significance Analysis of Microarrays for R (SAMr, version 1.01) available at http://www-stat.stanford.edu/~tibs/SAM/Rdist/index.html. The parameters of the *samr* function were set to "resp.type = Two class unpaired" and "nperms = 500", the maximum number of permutations possible using available computer software and hardware.

3.8 Annotation of the Affymetrix Soybean GeneChip probesets

The following steps were carried out to complete the annotations provided by Affymetrix for the Soybean GeneChip. The initial file contained a description of the consensus sequences used as a template for probe design in text format. A new file was created with *Glycine max* consensus sequences only (prefixed Gma) and definition lines were created and edited to conform to fasta and tabular format standards, respectively. The algorithm BLASTN (Basic Local Alignment Search Tool Nucleotides) was used to compare these sequences to The Institute for Genomic Research (TIGR) Soybean (*Glycine max*) Gene Index (GmGI) Release 12.0 (Altschul et al., 1990; TIGR, 2004). The results were parsed using XMLBlast::Report (D. Benz and J. Crow, Centre for Computational Genomics and Bioinformatics, University of Minnesota), a Perl script modified by Christophe Henquin (programmer) to allow for parsing of multiple blast results. These results were loaded into a MySQL database.

4. RESULTS AND DISCUSSION

4.1 Plant growth monitoring

Five days after planting (DAP), 90% of the seeds had germinated. Mandarin was harvested 15 DAP, when 1 plant was in the cotyledon stage, 3 plants were V1 and 6 plants were V2. OAC Bayfield, 2601R, PS46R, and S03-W4 were harvested 16 DAP, when 2 plants were in the cotyledon stage, 12 plants were V1 and 21 plants were V2. The plants did not show any symptoms of pathogen infection or abiotic stress. Figure 1 shows a soybean plant in the V2 stage.

4.2 Detection of CP4-EPSPS using PCR

To verify whether each plant in the varieties selected for the GM group was transformed with the CP4-EPSPS insert, PCR was performed on DNA extracted from second trifoliate leaves using primers designed to amplify a portion of the CP4-EPSPS insert (Padgette et al., 1995). To exclude the possibility of a mix-up, PCR was also performed on DNA extracted from pooled leaf samples collected from conventional soybean plants. The sensitivity of this assay was verified by adding one GM second trifoliate leaf to 29 second trifoliate leaves collected from conventional plants and by adding one GM second trifoliate leaf to 49 second trifoliate leaves collected from conventional plants. The CP4-EPSPS insert present in the GM second trifoliate leaf could be detected among the 29 second trifoliate leaves harvested from conventional plants and among the 49 second trifoliate leaves harvested from conventional plants. Thus, PCR was performed on DNA extracted from a pool of second trifoliate leaves in the case of conventional plants (a maximum of 20 in each group) and on each individual second trifoliate leaf for the GM varieties. The results are shown in Figure 2. This screening experiment showed that each of the GM plants used for the microarray experiment was transformed with the CP4-EPSPS insert, while none of the conventional plants was transformed with the CP4-EPSPS insert.

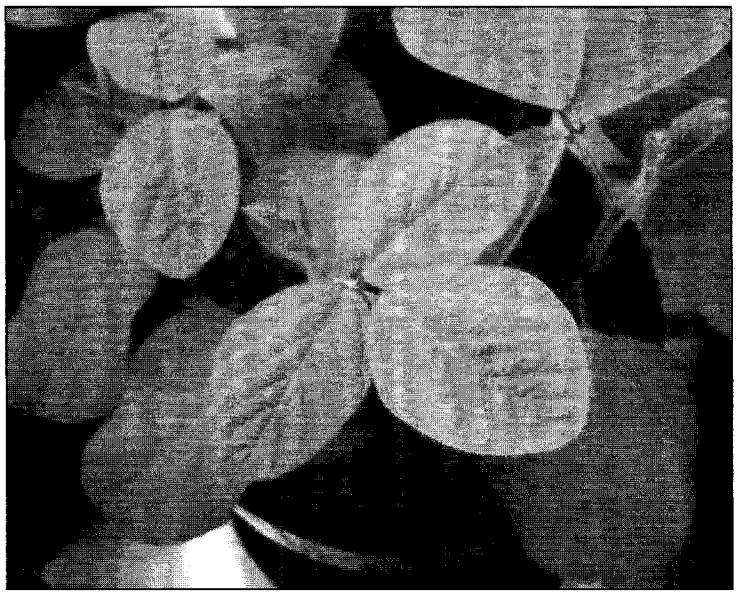

Figure 1. Soybean V2 stage. First trifoliate leaf is completely unrolled and outer leaflets of second trifoliate leaf are no longer touching.

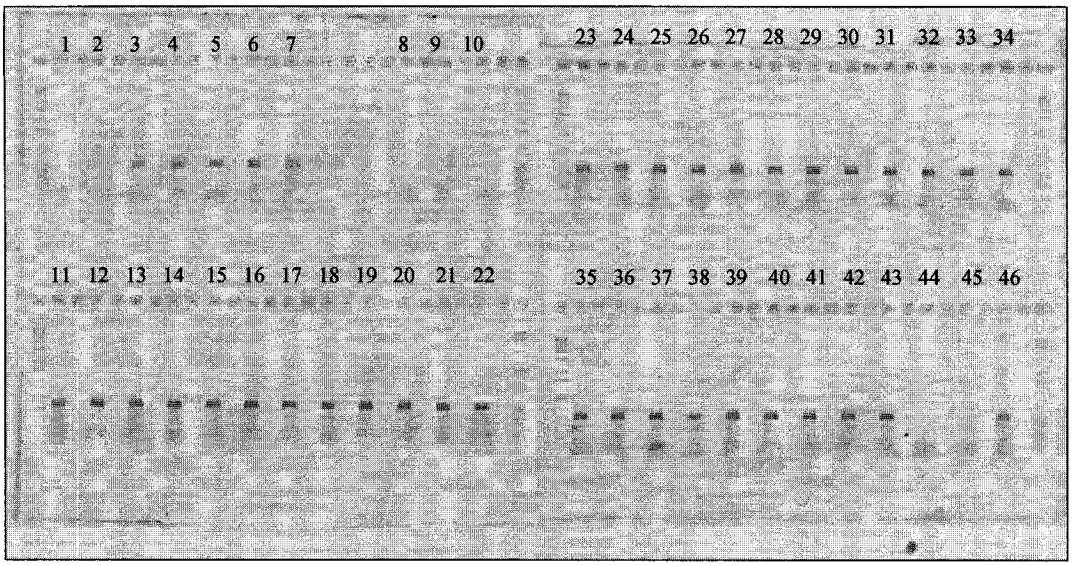

Figure 2. Agarose gel showing use of the PCR primer combination *sttmf3a* and *sttmr2a* to screen for plants transformed with the CP4-EPSPS insert (1 kb DNA ladder, Invitrogen). Lane 1: negative control with water instead of DNA, lane 2: template DNA was isolated from OAC Bayfield, lanes 3-4: template DNA was isolated from a pool of 1 GM and 49 conventional second trifoliate leaves, lanes 5-6: template DNA was isolated from a pool of 1 GM and 29 conventional second trifoliate leaves, lane 7: template DNA isolated from one 2601R plant, lane 8: template DNA isolated from Mandarin (Ottawa), lane 9: template DNA isolated from OAC Bayfield, lane 10: template DNA isolated from S03-W4, lanes 11-18: X, lanes 19-27: template DNA isolated from 9 diferent PS46R plants, lanes 28-36: template DNA isolated from 9 different 2601R plants, lanes 37-46: X. A single PCR product (~600 bp) corresponding to an amplified segment of the CP4-EPSPS insert was obtained from samples containing DNA from GM soybean second trifoliate leaves.

X: template DNA isolated from soybean varieties not discussed in this thesis.

4.3 Quality assessment of total RNA

To evaluate the quality of the RNA samples prior to hybridization to high-density oligoarrays, an electropherogram of each total RNA sample extracted from first trifoliate leaves was obtained using an Agilent 2100 bioanalyzer. Good quality total RNA is defined by the appearance of well spaced, prominent, narrow peaks for ribosomal RNAs (Krupp, 2005). A relatively low number of smaller peaks, indicative of low molecular weight degradation products, may also be present in electropherograms of good quality total RNA. Although total RNA profiles are expected to vary substantially between plant species and plant tissues (Krupp, 2005), they were not expected to vary substantially between samples from soybean first trifoliate leaves. Electropherograms of 10 randomly selected total RNA samples are shown in Figure 3. The other 15 samples selected for hybridization to GeneChips had similar total RNA profiles (data not shown). The well defined peaks on these electropherograms, and the high reproducibility of the RNA profiles, were indicative of good quality total RNA.

4.4 Quality assessment of cRNA

A typical preparation of "target" samples involves *in vitro* transcription of double-stranded cDNA to generate complementary biotin-labeled RNA for hybridization to GeneChips. The quality of the resulting biotinylated cRNA was also assessed using the Agilent 2100 bioanalyzer. Good quality cRNA is defined by a single broad peak devoid of smaller peaks. The electropherogram in Figure 4a is provided as an example of good quality cRNA obtained with the One-cycle cDNA synthesis kit (Affymetrix). Electropherograms of 10 randomly selected cRNA samples synthesized from soybean total RNA are shown superimposed in Figure 4b. The 15 other cRNA samples selected for hybridization to GeneChips had similar profiles (data not shown). Figure 4b shows a well-defined broad peak for each cRNA sample. Although the height of the peaks seems to vary from one sample to another, this observed variability was attributed to the varying concentration of cRNA between each sample, a situation that can be corrected by

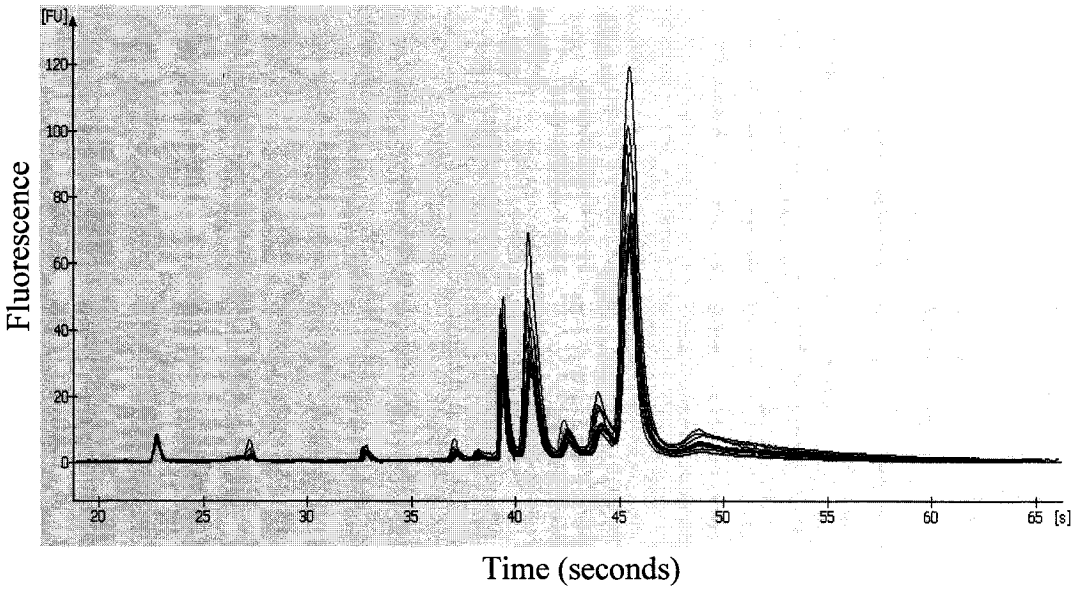

Figure 3. Overlaid electropherograms of 10 randomly selected total RNA samples from 2601R, OAC Bayfield, PS46R, S03-W4, and Mandarin (Ottawa) first trifoliate leaves (two electropherograms per soybean variety). Total RNA was extracted from soybean first trifoliate leaves using the RNeasy Plant Mini Kit (Qiagen) and the electropherograms were obtained using the Agilent 2100 bioanalyzer.

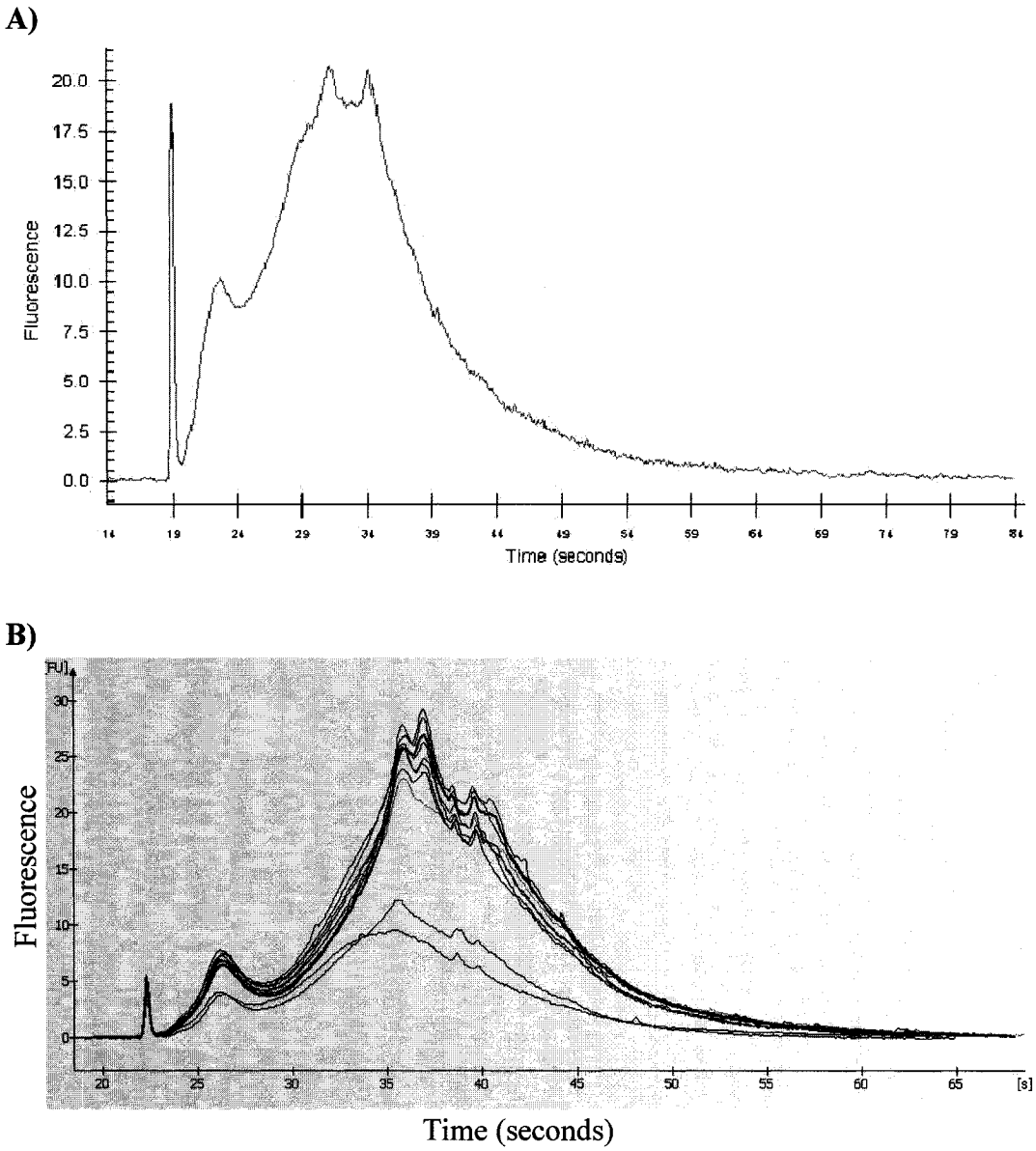

Figure 4. Electropherograms of cRNA samples synthesized from total RNA using the One-cycle cDNA synthesis kit (Affymetrix). (A) Good quality biotin-labeled cRNA from human cultured cells (from Affymetrix GeneChip Expression Analysis Technical Manual, 2004). (B) Overlaid electropherograms of 10 randomly selected cRNA samples from 2601R, OAC Bayfield, PS46R, S03-W4, and Mandarin (Ottawa) first trifoliate leaves (two electropherograms per soybean variety). All electropherograms were obtained using an Agilent 2100 bioanalyzer.

applying normalization techniques after the cRNA samples have been hybridized to GeneChips. Therefore the 25 samples were retained for hybridization to GeneChips.

4.5 Quality assessment of target preparation

R-Bioconductor functions available in the *affy* package were used to assess the quality of the "target" preparations for each of the 25 soybean GeneChips. In particular, the *AffyRNAdeg* function orders the 11 PM probes per probeset relative to the 5' end of the template consensus sequence. Following this, probe intensities are averaged by location across probesets on each GeneChip in a dataset. For example, the sum of the intensity signal from all probes located at position 7 (out of 11) relative to the 5' end of the template consensus sequence is divided by the number of probesets on the GeneChip. Since RNA exonucleases typically attack RNA molecules from the 5' or 3' end, probe signal intensities systematically lowered at one or both ends may indicate degradation of the cRNA (Alberts et al., 2002). Figure 5 shows the mean signal intensity of probes ordered from the 5' to the 3' ends of the template consensus sequence for each of the 25 GeneChips. The overall "flatness" of the 25 lines indicates that the integrity of the 25 cRNA samples was preserved throughout the steps leading to hybridization.

Finally, scanner images of the GeneChips were reconstructed using the *image* function available in the *affy* package. Visual inspection of each GeneChip did not reveal any obvious experimental artifacts such as air bubbles or salt stains (data not shown).

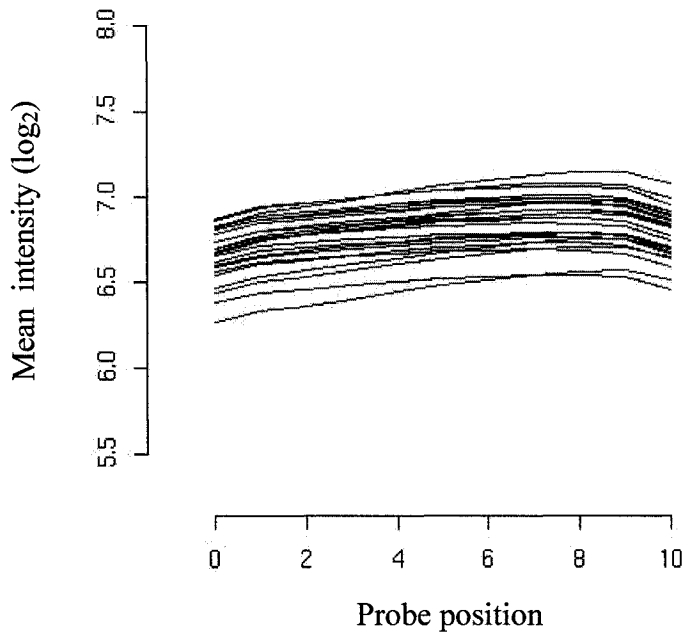

Figure 5. Quality assessment of target preparations using the *AffyRNAdeg* function in R-Bioconductor. Each of the 25 GeneChips is represented by a line. The average fluorescence intensity (log_2) is given for each of the 11 probes. The probes are ordered from 0 to 10 according to their physical location relative to the 5' end of the template consensus sequence (probe position 0 is nearest the 5' end).

4.6 Comparison of pre-processing methods

The goal of the present study was to explore the use of high-density oligoarrays as a tool to assess substantial equivalence of GM soybean. Since high-density oligoarrays are a relatively recent technology, there is still some uncertainty concerning the interpretation of signal intensities which is reflected in the number and variety of methods that have been developed to obtain the most accurate and sensitive measure of gene expression (Galfalvy et al., 2003). Since there is no "gold standard" for pre-processing GeneChip data, three pre-processing methods were applied to the dataset (MAS, dChip and RMA). The main advantage of such a combinatorial approach is to reduce the number of false positives by compensating for the biases that affect each strategy. In the particularly sensitive context of assessing substantial equivalence, such false leads could discredit a useful tool that is still in development. However, the main disadvantage of the approach is that the least sensitive and accurate method also determines the outcome of the statistical analysis. Nevertheless, the following graphs illustrate the effects of each pre-processing method on the dataset.

Figure 6 shows the overall effect of each pre-processing method on probeset intensities. Each sample is represented in the form of a boxplot. The thick black line in each box is the median probeset signal intensity, the top and bottom edges of the box are the third and first quartiles, respectively, and the points at the extreme top and bottom range of the plot are the maximum and minimum probeset signal intensities, respectively. As seen in Figure 6a, the distribution of probeset intensities varies widely between arrays when no pre-processing method is applied to the dataset. Hence, any comparison of the GM group (2601R and PS46R) to the Conventional group (OAC Bayfield, S03-W4 and Mandarin) would lead to misleading conclusions as per which genes are differentially expressed. Normalization of the all the samples in the dataset was achieved through MAS, dChip and RMA pre-processing (Figures 6b, c, and d).

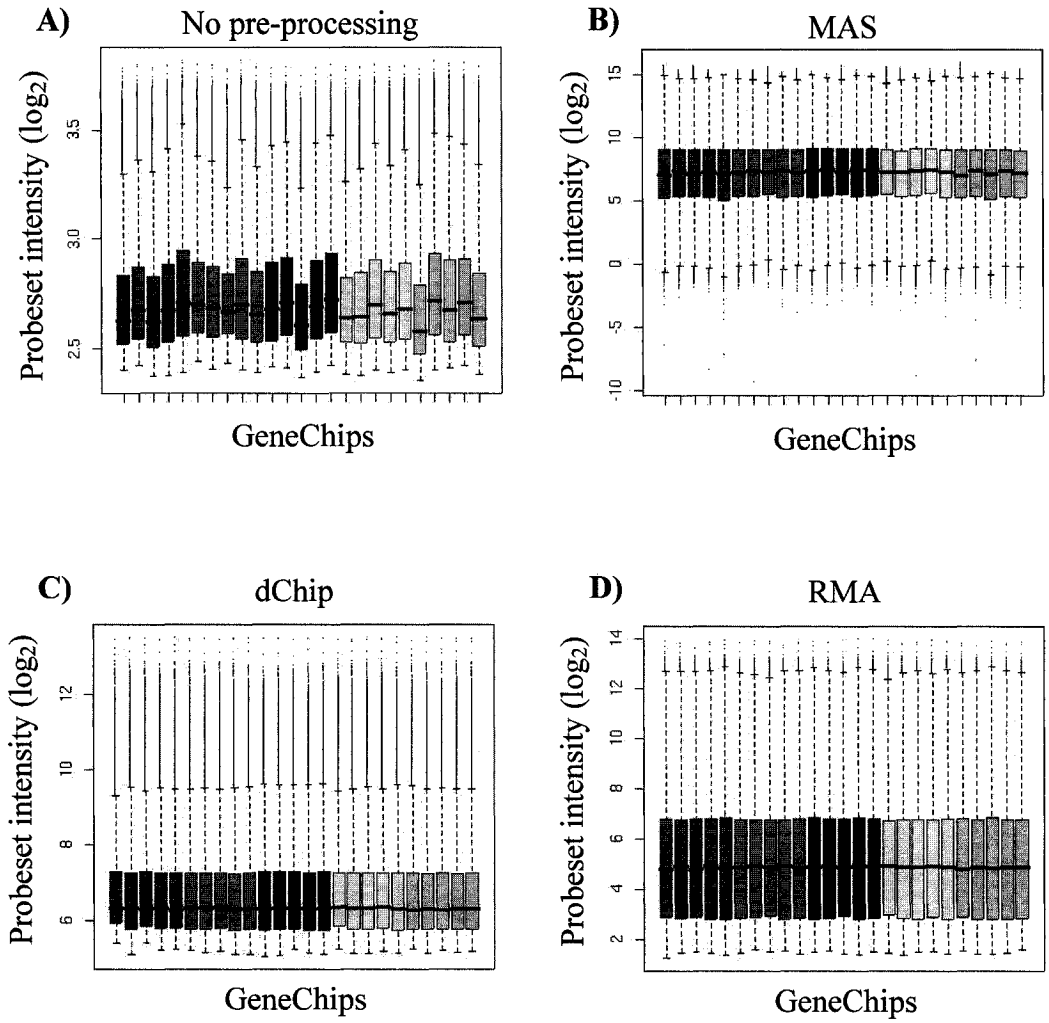

Figure 6. Box plots of log-scale probeset intensities for the 25 arrays in the soybean GeneChips dataset following application of three different pre-processing methods. (A) No pre-processing method was applied to the soybean GeneChips dataset. PM probe intensities in each probeset were averaged and the resulting probeset intensities were $\log_2$ transformed. (B) The soybean GeneChips dataset was pre-processed with MAS (probeset intensities were $\log_2$ transformed). (C) The soybean GeneChips dataset was pre-processed with dChip (probeset intensities were $\log_2$ transformed). (D) The soybean GeneChips dataset was pre-processed with RMA. The first group of five arrays from the left is variety 2601R (red), the second group is variety Bayfield (green), the third group is variety Mandarin (blue), the fourth group is variety PS46R (yellow) and the fifth group is variety S03-W4 (grey).

The effects of pre-processing on probeset intensities were also assessed using scatterplots, MvA plots and "SD vs. Mean" plots. The representative plots, selected to illustrate the effects of MAS, dChip and RMA pre-processing, are shown in Figures 7, 8, and 9, respectively.

In Figure 7a, sample 2601R-2 (the second 2601R first trifoliate leaf that was collected) is plotted against 2601R-6 (the sixth 2601R first trifoliate leaf that was collected), Bayfield-5 and Mandarin-7. These comparisons show that the points (corresponding probeset intensities) are less dispersed in the first scatterplot when compared to the points on the two other plots, because both samples come from the same variety. In Figure 7b, the difference between probesets intensities is plotted against the average intensity of the paired probesets leading to the creation of an MvA plot. Since the data are log transformed, the difference between corresponding probeset intensities ($\log_a x_{i2} - \log_a x_{i1}$) is the same as fold change expressed as $\log_a x_{i1} / \log_a x_{i2}$, where x is the intensity of probeset i from sample 1 or 2. M values of zero indicate that the probesets are not differentially expressed. Since we expect only a small fraction of the probesets to be differentially expressed, the points should scatter around 0 on the y-axis. With the MAS pre-processing method, most probesets expressed at low intensity appear differentially expressed. Figure 7c, in which the standard deviation (SD) of each probeset in the entire dataset is plotted against the mean intensity of each probeset in the entire dataset, shows a strong relationship between variance and intensity, especially for low intensity probesets. Ideally, to make comparisons between probesets relevant, this function should have a small constant value across mean probeset intensities.

The effects of dChip pre-processing are shown in Figure 8. Figure 8a shows that the points scatter more tightly in the 2601R-2 *vs.* 2601R-6 plot than in the other two. The bias seen in Figure 7b is less present in Figure 8b, where the difference between probeset intensities is more tightly scattered around M = 0 across probeset intensities. Figure 8c shows that the dChip pre-processing method has a

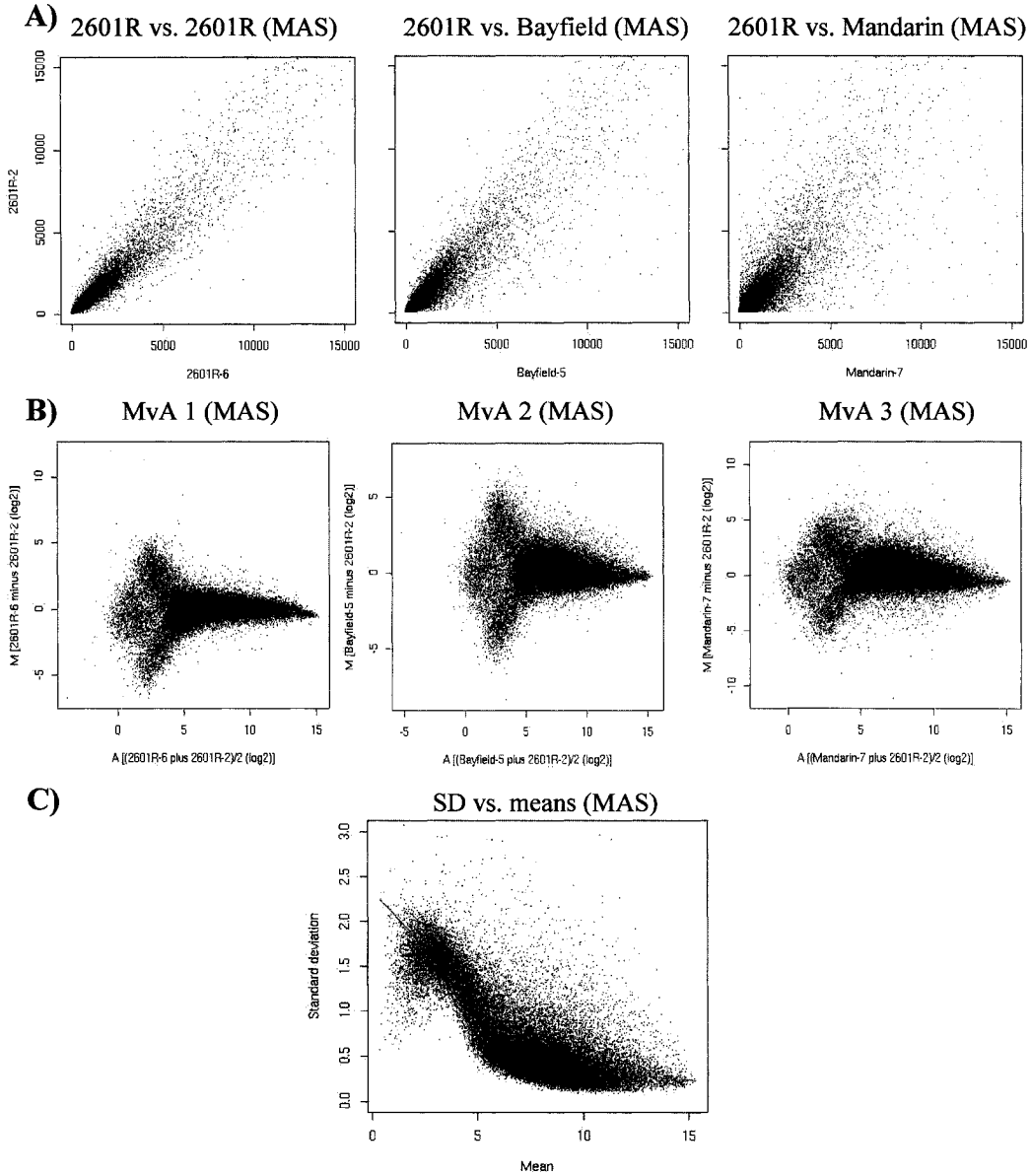

Figure 7. Effect of MAS pre-processing on probeset intensities. (A) Scatterplots of first trifoliate leaf 2601R-2 *vs.* first trifoliate leaf 2601R-6, first trifoliate leaf 2601R-2 *vs.* first trifoliate leaf Bayfield-5, first trifoliate leaf 2601R-2 *vs.* first trifoliate leaf Mandarin-7. (B) MvA (Minus vs. Average) plots in which the difference between corresponding probesets intensities ($\log_a x_{i2} - \log_a x_{i1}$) is plotted against their average intensities ($\log_a x_{i1} + \log_a x_{i2}$)/2. MvA1 compares first trifoliate leaf 2601R-2 and first trifoliate leaf 2601R-6, MvA2 compares first trifoliate leaf 2601R-2 and first trifoliate leaf Bayfield-5, MvA3 compares first trifoliate leaf 2601R-2 and first trifoliate leaf Mandarin-7. (C) Standard deviations *vs.* means for all arrays with Loess smoother (red).

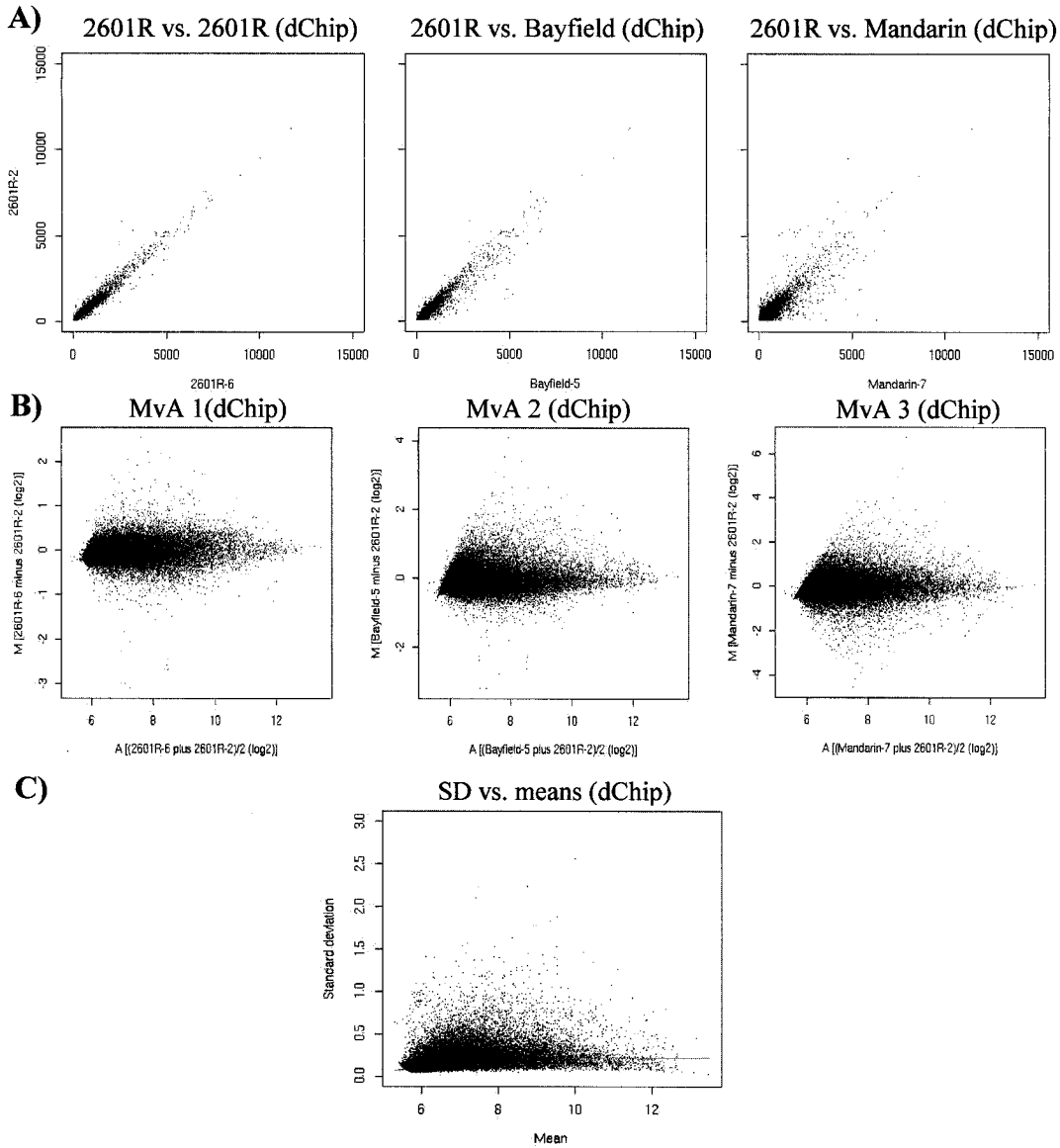

Figure 8. Effect of dChip pre-processing on probeset intensities. (A) Scatterplots of first trifoliate leaf 2601R-2 *vs.* first trifoliate leaf 2601R-6, first trifoliate leaf 2601R-2 *vs.* first trifoliate leaf Bayfield-5, first trifoliate leaf 2601R-2 *vs.* first trifoliate leaf Mandarin-7. (B) MvA (Minus vs. Average) plots in which the difference between corresponding probesets intensities ($\log_a x_{i2} - \log_a x_{i1}$) is plotted against their average intensities ($\log_a x_{i1} + \log_a x_{i2}$)/2. MvA1 compares first trifoliate leaf 2601R-2 and first trifoliate leaf 2601R-6, MvA2 compares first trifoliate leaf 2601R-2 and first trifoliate leaf Bayfield-5, MvA3 compares first trifoliate leaf 2601R-2 and first trifoliate leaf Mandarin-7. (C) Standard deviations *vs.* means for all arrays with Loess smoother (red).

variance stabilizing effect, *i.e.* the relationship between variance and intensity is less present than in the MAS pre-processed dataset. However, there seems to be an overcorrection of probeset intensities (*i.e.* standard deviation very small) for genes with low expression levels.

Finally, Figure 9 shows the effects of the RMA pre-processing method. Again, Figure 9a shows that the points are more tightly scattered in the 2601R-2 *vs.* 2601R-6 plot than in the other two. Figure 9b shows that M, the intensity difference between corresponding probesets, is less affected by A, the mean intensity of corresponding probesets, when compared to the other two pre-processing methods. Figure 9c shows that the RMA pre-processing method also has a variance stabilizing effect, perhaps with overcorrection of probeset signals at the low mean intensity end.

In general, variability in MAS probeset intensity measurements was high for genes with low expression levels, and decreased as the measurements increased, reaching levels comparable to dChip and RMA for higher intensity probesets. The RMA and dChip pre-processing methods detected probeset intensities with low and relatively constant variability, perhaps underestimating variability for probesets with low intensities. Stabilization of variance may have been achieved in dChip and RMA through the use of PM probe signal intensities only and by combining signal intensities for probes across all the arrays in the dataset to determine the levels of background noise and cross-hybridization.

4.7 Statistical analysis of differential gene expression

Significance Analysis of Microarrays (SAM) was applied to the MAS, dChip and RMA pre-processed datasets to compare gene expression in first trifoliate leaves in the GM group (2601R and PS46R) to gene expression in first trifoliate leaves in the Conventional group (OAC Bayfield, Mandarin and S03-W4), on a gene-per-gene basis. Two factors had a major impact on the outcome of SAM: the

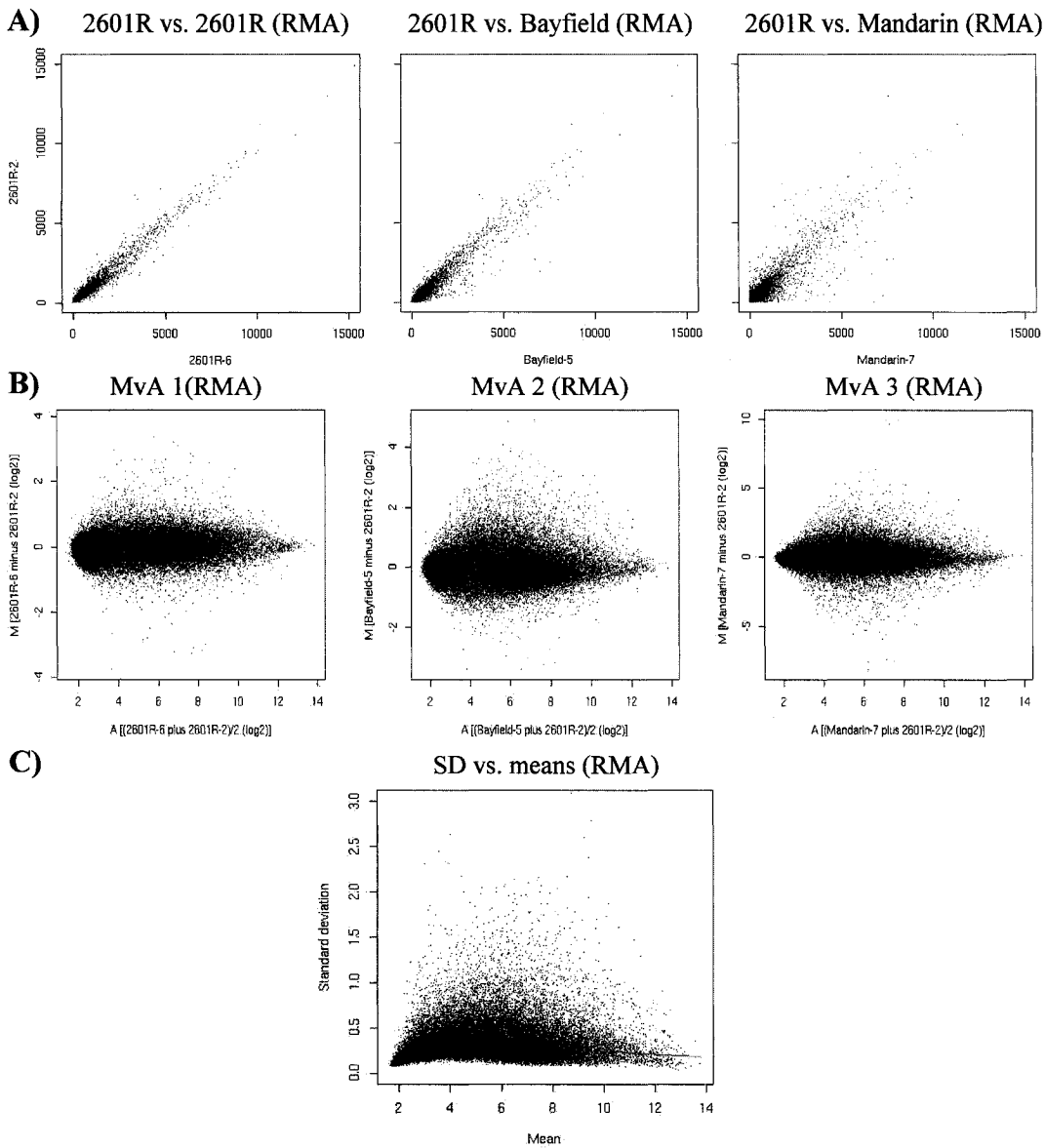

Figure 9. Effect of RMA pre-processing on probeset intensities. (A) Scatterplots of first trifoliate leaf 2601R-2 *vs.* first trifoliate leaf 2601R-6, first trifoliate leaf 2601R-2 *vs.* first trifoliate leaf Bayfield-5, first trifoliate leaf 2601R-2 *vs.* first trifoliate leaf Mandarin-7. (B) MvA (Minus vs. Average) plots in which the difference between corresponding probesets intensities ($\log_a x_{i2} - \log_a x_{i1}$) is plotted against their average intensities ($\log_a x_{i1} + \log_a x_{i2}$)/2. MvA1 compares first trifoliate leaf 2601R-2 and first trifoliate leaf 2601R-6, MvA2 compares first trifoliate leaf 2601R-2 and first trifoliate leaf Bayfield-5, MvA3 compares first trifoliate leaf 2601R-2 and first trifoliate leaf Mandarin-7. (C) Standard deviations *vs.* means for all arrays with Loess smoother (red).

permutation matrix and the delta value selected as a basis for "significance". The permutation matrix is the number of permutations performed (rows) and the sample composition of each group (columns). Since the available computing power limited the number of permutations to 500, the composition of the permutation matrix had an effect on the number of significant probesets, regardless of the selected delta threshold (data not shown). The second factor, delta, the difference between the original d_i score and permutated $\bar{d}_i$ scores, also heavily influenced the outcome of SAM. To compare the different pre-processing methods, we selected a delta value of 1. Thus, a gene was included among the significant genes list when the absolute difference between d_i and $\bar{d}_i$ was greater than 1. Figure 10 shows that selecting a slightly lower delta threshold (for example delta = 0.5) would have resulted in a much higher number of significant probesets as well as a higher FDR, especially for the dChip pre-processing method.

Figure 11 shows the results of each SAM analysis. To obtain these plots, the distribution of d_i scores was plotted against $\bar{d}_i$ scores. Significant probesets can be found above (up-regulated genes) and below (down-regulated genes) the dotted line corresponding to a delta threshold of 1. Green colored points represent significant up-regulated probesets that are differentially expressed by more than two-fold (natural scale). Red colored points represent significant down-regulated probesets that are also differentially expressed by more than two-fold (natural scale). The two-fold limit is a standard cut-off in microarray studies (Deavours and Dixon, 2005; Gomez-Mena et al., 2005; Kristensen et al., 2005) and was set in accordance with findings reported in a recent spike-in study (Choe et al., 2005). In this study, it was shown that, at low concentrations, where signals are typically noisier, a minimum two-fold change in target concentrations could be detected. Therefore, significant probesets also had to meet this condition to be included

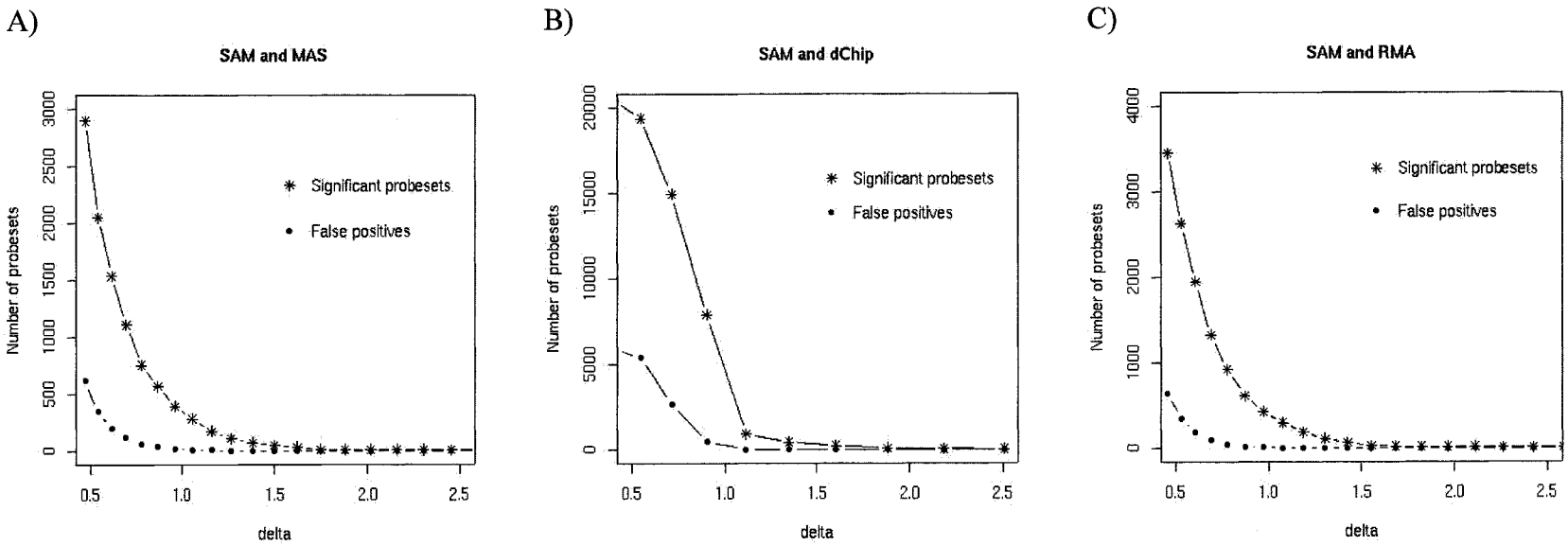

Figure 10. Effect of delta threshold on the number of significant probesets and the number of false positives. Delta is the difference between the original d_i score and permutated $\overline{d}_i$ scores. SAM statistical analysis was applied to the dataset pre-processed with (A) MAS method, (B) dChip method and (C) RMA method. Each plot represents the number of significant probesets and false positives at each delta threshold (from ~0.5 to ~2.5).

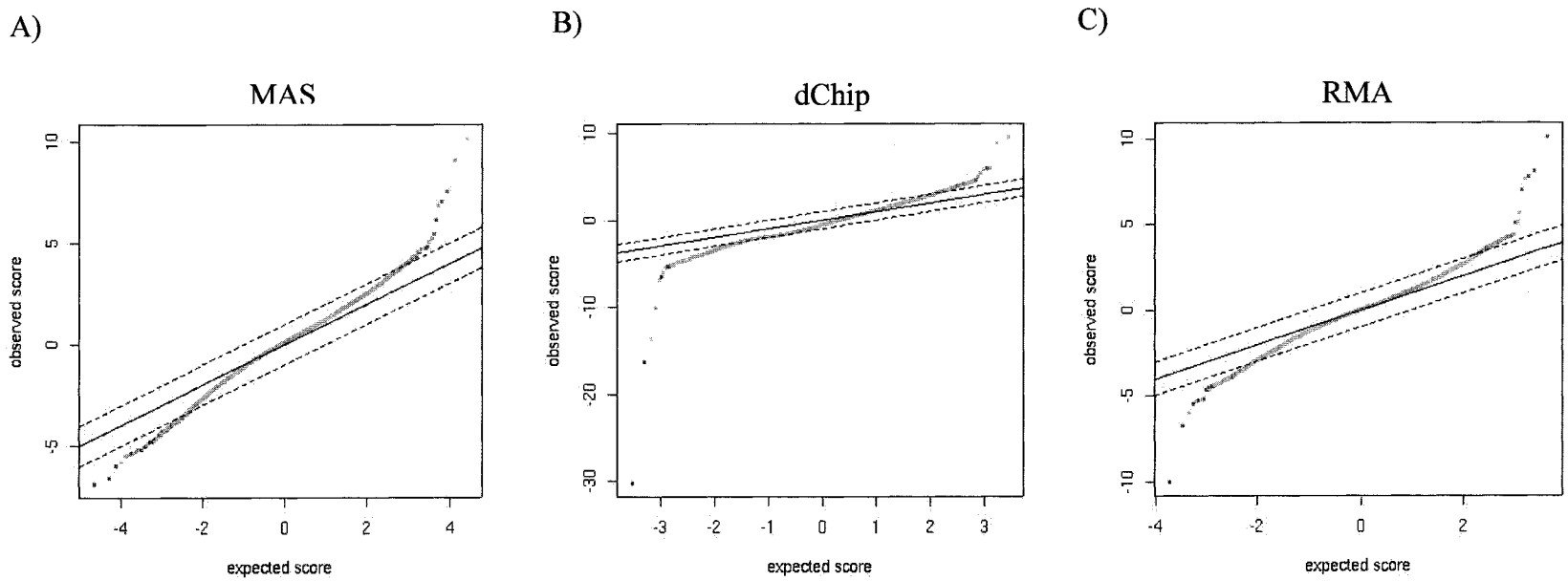

Figure 11. SAM Q-Q plots for (A) the soybean GeneChips dataset pre-processed with MAS, (B) the soybean GeneChips dataset pre-processed with dChip, and (C) the soybean GeneChips dataset pre-processed with RMA. In each plot, the distribution of observed d_i scores is plotted against the distribution of permutated $\overline{d}_i$ scores. Significant probesets are plotted above and below the dotted line defined by a delta threshold equal to 1. Delta is the difference between the original d_i score and permutated $\overline{d}_i$ scores. In green, significant probesets up-regulated in the GM soybean varieties by two-fold (natural scale). In red, significant probesets down-regulated in the GM soybean varieties by two-fold (natural scale).

among the list of differentially expressed genes. Table 2 summarizes the outcome of the different SAM analyses.

Of the three pre-processing methods, dChip had the lowest FDR (1.3%) for a given delta threshold (delta = 1). The FDR procedure applied in SAM may be defined as the estimated proportion of false positives among significant probesets (Tusher et al., 2001). Thus, a lower FDR for a given delta threshold indicates that dChip was the most sensitive pre-processing method. However, the majority of significant probesets in the dChip pre-processed dataset did not meet the minimum two-fold change criterion. For a relatively similar FDR, SAM identified less significant probesets in the RMA pre-processed dataset, more of which were differentially regulated by more than two-fold. The number of significant probesets was found to be similar for the MAS and RMA pre-processed dataset.

In accordance with our strategy to reduce bias by applying SAM to a dataset pre-processed with three different methods, we combined the gene lists and focused on the genes present in more than one gene list. As shown in Figure 12, when applying the two-fold condition, 11 out of 113 (10%) differentially expressed genes called by SAM were common to all three pre-processed datasets, 35 out of 108 (32%) DEGs identified by SAM were in agreement between the MAS and RMA pre-processed dataset, 14 out of 67 (21%) were in agreement between the dChip and RMA pre-processed dataset, and 11 out of 89 (12%) were in agreement between the MAS and dChip pre-processed datasets. Since we did not know which (if any) of the genes were differentially expressed, it is not possible to assess which pre-processing method offered the most accurate estimates of gene expression. However, if SAM identified all the differentially expressed genes in the dChip pre-processed dataset, without omitting other truly differentially expressed genes, the dChip pre-processing method coupled to SAM would yield the most sensitive and specific results. If, on the contrary, many differentially

Table 2. Summary of SAM analyses comparing GM soybean varieties to conventional soybean varieties. SAM was performed on the MAS, dChip and RMA pre-processed dataset and significant probesets were obtained by selecting a delta threshold equal to 1. The last two columns contain the number of significant probesets differentially expressed by more than two-fold.

Pre-processing method	Total number of significant probesets	Median FDR (%)	Total number of up-regulated probesets	Total number of down-regulated probesets	Number of up-regulated significant probesets (FC>2)	Number of down-regulated significant probesets (FC<0.5)
MAS	368	4.0	67	301	22	59
dChip	1259	1.3	133	1126	8	11
RMA	377	1.6	107	270	30	32

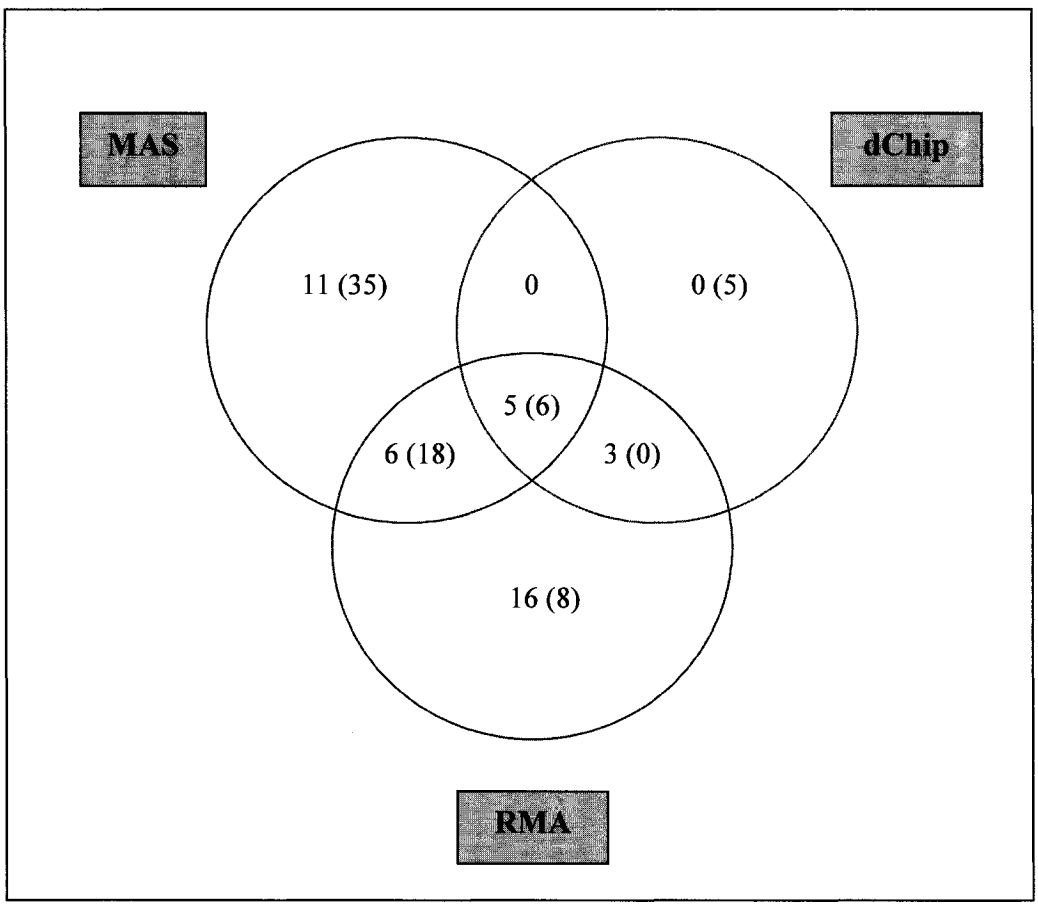

Figure 12. Number of significant probesets (delta = 1) up-regulated by more than two-fold using each pre-processing method with SAM. The number of down-regulated probesets is shown in parentheses.

expressed genes were missing from the SAM analysis performed on the dChip dataset but included in the other two, the dChip pre-processing method would still be sensitive but not as specific as the other two. Tables 3 and 4 offer a closer look at the probesets selected by all three pre-processing methods.

Table 3. Summary statistics and annotations for up-regulated significant probesets (FC>2) common to all three pre-processing methods.

Probeset ID	Top TIGR TC score	Description
Gma.3314.2.S1_x_at	TC226073	similar to UP|Q71SU8 (Q71SU8) Protease inhibitor, partial (75%)
Gma.1327.2.S1_s_at	TC203332	weakly similar to UP|Q8H9E9 (Q8H9E9) Resistant specific protein-2, partial (60%)
GmaAffx.3560.1.S1_at	TC225360	homologue to UP|MSK1_MEDSA (P51137) Glycogen synthase kinase-3 homolog MsK-1, partial (97%)
Gma.3921.2.S1_x_at	TC226191	Unknown
Gma.17814.1.S1_at	TC225831	weakly similar to UP|Q9ATM2 (Q9ATM2) Small basic membrane integral protein ZmSIP1-2, partial (78%)

Table 4. Summary statistics and annotations for down-regulated significant probesets (FC<0.5) common to all three pre-processing methods.

Probeset ID	Top TIGR TC score	Description
Gma.1379.2.A1_at	TC214227	homologue to UP|RL44_GOSHI (Q96499) 60S ribosomal protein L44, partial (65%)
Gma.4071.1.S1_at	TC205912	weakly similar to UP|O24040 (O24040) LTCOR11, partial (61%)
Gma.11115.1.S1_s_at	TC204156	homologue to UP|Q40151 (Q40151) Hsc70 protein, partial (40%)
GmaAffx.22419.1.S1_at	TC211582	Unknown
Gma.1043.1.S1_at	TC205422	Unknown
Gma.4564.1.A1_at	TC210170	Unknown

5. CONCLUSION

The objective of this thesis was to explore the use of high-density oligoarrays as tools for the substantial equivalence assessment of GM crops using soybean as the model plant. Since the methods currently employed to assess substantial equivalence are targeted toward the detection of pre-determined compounds, and, therefore, geared toward the assesment of "known" risks, they may lack true power to detect unintended effects of genetic modification. The gene expression profiling method presented in this thesis was designed with the objective of complementing current safety assessments by providing broader coverage of analytes, in this case mRNA, and a more in-depth analysis of the potential impacts of genetic modification on plant composition.

While applying this method to assess substantial equivalence of GM crops, we focused on assessing the quality of the samples used to generate the dataset, comparing pre-processing methods used to extract summary measures for each probeset, and finally, applying a statistical analysis (SAM) to obtain a list of genes that were differentially expressed between the GM soybean varieties and the conventional soybean varieties. In the absence of prior studies on differential gene expression in CP4-EPSPS transformed crops or validation techniques such as qRT-PCR, it is not possible to determine if the proposed lists (Tables 3 and 4) reflected the "true" state of gene expression in CP4-EPSPS transformed soybean. However, qRT-PCR analyses have tended to confirm the results of microarray data analyses (Larkin et al., 2005; Zhang et al., 2005). Finally, our analysis was limited to one plant organ (first trifoliate leaf) collected from five varieties of soybean grown in a controlled environment, and to the approximately 37 000 genes that were represented on the Affymetrix Soybean GeneChip. Experiments involving different plant organs or different environmental conditions would provide more information on differential expression in CP4-EPSPS transformed soybean.

Sampling from five different soybean varieties might also have had an impact on the results of the statistical analysis of gene expression. On GeneChips, the relative signal intensity of 25-mer probes reflects relative cRNA abundance in addition to mRNA sequence variation and alternative splicing mimicking relative cRNA abundance. Affymetrix GeneChips have recently been used to conduct reverse genetics experiments, more precisely to identify single-nucleotide polymorphisms in the form of differential probe signal intensities between cRNA samples from different barley genotypes (Rostoks et al., 2005). Thus, differential expression identified using the three pre-processing methods and SAM may not have happened as a result of genetic modification but rather as a product of the soybean varieties different affinities for the DNA probe sequences on the Affymetrix Soybean GeneChip. However, this result would be unlikely given the relatively low frequency of single-nucleotide polymorphisms in commercial soybean (Zhu et al., 2003).

In summary, although there are no standardized methods to generate, pre-process and analyze GeneChip datasets, results from this thesis indicate that high-density oligoarrays provide a highly sensitive exploratory tool for the substantial equivalence assessment of GM crops.

6. REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno R (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science **252:** 1651-1656.

Affymetrix (2004) GeneChip expression analysis - technical manual. Affymetrix. http://www.affymetrix.com/support/technical/manual/expression_manual.affx, 194 pp.

Alba R, Fei Z, Payton P, Liu Y, Moore SL, Debbie P, Cohn J, Ascenzo M, Gordon JS, Rose JKC, Martin G, Tanksley SD, Bouzayen M, Jahn MM, Giovannoni J (2004) ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. The Plant Journal **39:** 697-714.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) Molecular Biology of the Cell, Ed 4. Garland Science, New York, 1616 pp.

Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ (1990) Basic Local Alignment Search Tool (BLAST). Journal of Molecular Biology **215:** 403-410.

Atkinson RG, Schroder R, Hallett IC, Cohen D, MacRae EA (2002) Overexpression of polygalacturonase in transgenic apple trees leads to a range of novel phenotypes involving changes in cell adhesion. Plant Physiology **129:** 122-133.

Barry G, Kishore G, Padgette M, Kolacz K, Weldon M, Re D, Eichholtz D, Fincher K, Hallas L (1992) Inhibitors of amino acid biosynthesis: strategies for imparting glyphosate tolerance to crop plants. *In* BK Singh, E Flores, JC Shannon, Biosynthesis and molecular regulation of amino acids in plants. American Society of Plant Physiologists, Rockville, MD, pp 139-145.

Beachy R, Bennetzen J, Chassy B, Chrispeels M, Chory J, Ecker J, Noel J, Kay S, Dean C, Lamb C, Jones J, Santerre C, Schroeder J, Umen J, Yanofsky M, Wessler S, Zhao Y, Parrott W (2002) Divergent perspectives on GM food. Nature Biotechnology **20:** 1195-1196.

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19:** 185-193.

Bolstad BM, Collin F, Simpson KM, Irizarry RA, Speed TP (2004) Experimental design and low-level analysis of microarray data. International Review of Neurobiology **60:** 25-58.

Bresnahan GA, Manthey FA, Howatt KA, Chakraborty M (2003) Glyphosate applied preharvest induces shikimic acid accumulation in hard red spring wheat (*triticum aestivum*). Journal of Agriculture and Food Chemistry **51:** 4004-4007.

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

CFIA (1995a) Decision document DD95-05: Determination of environmental safety of Monsanto Canada Inc.'s glyphosate tolerant soybean (*Glycine max L.*) line GTS 40-3-2. Canadian Food Inspection Agency, Plant Biosafety Office, Ottawa, Canada. http://www.inspection.gc.ca/english/plaveg/bio/dd/dd9505e.shtml.

CFIA (1995b) Decision document DD95-02: Determination of environmental safety of Monsanto Canada Inc.'s Roundup® herbicide-tolerant *Brassica napus* canola line GT73. Canadian Food Inspection Agency, Plant Biosafety Office, Ottawa, Canada. http://www.inspection.gc.ca/english/plaveg/bio/dd/dd9502e.shtml.

CFIA (1999) Decision Document 1999-33: Determination of the safety of Monsanto Canada Inc.'s Roundup Ready® Corn (*Zea mays L.*) line GA21. Canadian Food Inspection Agency, Plant Biosafety Office, Ottawa, Canada. http://www.inspection.gc.ca/english/plaveg/bio/dd/dd9933e.shtml.

CFIA (2004) Directive 94-08 (Dir94-08): Assessment criteria for determining environmental safety of plants with novel traits. Canadian Food Inspection Agency, Plant Biosafety Office, Ottawa, Canada. http://www.inspection.gc.ca/english/plaveg/bio/dir/dir9408e.shtml.

CFIA (2005) List of varieties with novel traits and their progeny registered under the Canada Seeds Act and Regulations. Canadian Food Inspection Agency, Ottawa, Canada. http://www.inspection.gc.ca/english/plaveg/variet/pntvcne.shtml.

Charlton A, Allnutt T, Holmes S, Chisholm J, Bean S, Elis N, Mullineaux P, Oehlschlager S (2004) NMR profiling of transgenic peas. Plant Biotechnology Journal **2:** 27-35.

Choe S, Boutros M, Michelson A, Church G, Halfon M (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biology **6:** 1-16.

Chudin E, Walker R, Kosaka A, Wu S, Rabert D, Chang T, Kreder D (2001) Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip® arrays. Genome Biology **3:** 1-10.

Cober ER, Morrison MJ, Ma B, Butler G (2005) Genetic improvement rates of short-season soybean increase with plant population. Crop Science **45:** 1029-1034.

Committee on Identifying and Assessing Unintended Effects of Genetically Engineered Foods on Human Health (2004) Safety of genetically engineered foods: Approaches to assessing the unintended health effects. The National Academies Press, Washington, DC, 235 pp.

Conner AJ, Jacobs JME (1999) Genetic engineering of crops as potential source of genetic hazard in the human diet. Mutation Research/Genetic Toxicology and Environmental Mutagenesis **443:** 223-234.

Corpillo D, Gardini G, Vaira AM, Basso M, Aime S, Paolo G, Fasano AM (2004) Proteomics as a tool to improve investigation of substantial equivalence in genetically modified organisms: The case of a virus-resistant tomato. Proteomics. **4:** 193-200.

CRAAQ (2002-2005) Résultats des essais de mais-grain et de cultivars de plantes oléoprotéagineuses et recommandations de cultivars de céréales. Centre de référence en agriculture et agroalimentaire du Québec, Sainte-Foy, Quebec, Canada. www.craaq.qc.ca.

Deavours BE, Dixon RA (2005) Metabolic engineering of isoflavonoid biosynthesis in alfalfa. Plant Physiology **138:** 2245-2259.

Draghici S (2003) Data analysis tools for DNA microarrays. Chapman and Hall/CRC, London, UK, 477 pp.

Duke SO (2005) Taking stock of herbicide-resistant crops ten years after introduction. Pest Management Science **61:** 211-218.

Dunwell JM (2005) Transgenic crops: The current and next generation. *In* L. Pena (ed.), Methods in Molecular Biology, Vol 286. Humana Press Inc., Totowa, NJ, pp 377-397.

EFSA (2004) Guidance document of the scientific panel on genetically modified organisms for the risk assessment of genetically modified plants and derived food and feed. European Food Safety Authority Journal **99:** 1-94.

Elmore RW, Roeth FW, Nelson LA, Shapiro CA, Klein RN, Knezevic SZ, Martin A (2001) Glyphosate-resistant soybean cultivar yields compared with sister lines. Agronomy Journal **93:** 408-412.

Fan W, Pritchard J, Olson J, Khalid N, Zhao L (2005) A class of models for analyzing GeneChip® gene expression analysis array data. BMC Genomics **6:** 1-10.

Fare TL, Coffey EM, Dai H, He YD, Kessler DA, Kilian KA, Koch JE, LeProust E, Marton MJ, Meyer MR, Stoughton RB, Tokiwa GY, Wang Y (2003) Effects of atmospheric ozone on microarray data quality. Annals of Chemistry **75:** 4672-4675.

Fehr WR, Caviness CE, Burmood DT, Pennington JS (1971) Stage of development descriptions for soybean, *Glycine max* (L.) Merrill. Crop Science **11:** 929-931.

Fiehn O, Kloska S, Altmann T (2001) Integrated studies on plant biology using multiparallel techniques. Current Opinion in Biotechnology **12:** 82-86.

Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D (1991) Light-directed, spatially addressable parallel chemical synthesis. Science **251:** 767-773.

Galfalvy HC, Erraji-Benchekroun L, Smyrniotopoulos P, Pavlidis P, Ellis SP, Mann J, Sibille E, Arango V (2003) Sex genes for genomic analysis in human brain: internal controls for comparison of probe level data extraction. BMC Bioinformatics **4:** 38-52.

Gall GL, Colquhoun IJ, Davis AL, Collins GJ, Verhoeyen ME (2003) Metabolite profiling of tomato (*Lycopersicon esculentum*) using 1H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. Journal of Agriculture and Food Chemistry **51:** 2447-2456.

Gautier L (2004) Alternative CDF environment for 2(or more)-genomes chip. R-Bioconductor. http://www.bioconductor.org/docs/vignettes.html, 3 pp.

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) Affy - analysis of Affymetrix GeneChip data at the probe level. Bioinformatics **20:** 307-315.

Gentleman R, Carey V, Bates D, 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 A, Sawitzki G, Smith C, Smyth G, Tierney L, Yang J, Zhang J (2004) Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology **5:** 1-16.

Gerrits N, Turk SCHJ, van Dun KPM, Hulleman SHD, Visser RGF, Weisbeek PJ, Smeekens SCM (2001) Sucrose metabolism in plastids. Plant Physiology **125:** 926-934.

Gomez-Mena C, de Folter S, Costa MMR, Angenent GC, Sablowski R (2005) Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. Development **132:** 429-438.

Han E-S, Wu Y, McCarter R, Nelson JF, Richardson A, Hilsenbeck SG (2004) Reproducibility, sources of variability, pooling, and sample size: Important considerations for the design of high-density oligonucleotide array experiments. The Journals of Gerontology Series A Biological Sciences and Medical Sciences **59:** B306-315.

Hategekimana B, Beaulieu M (2002) Genetically modified crops: Steady growth in Ontario and Quebec. Statistics Canada, Ottawa, Canada. www.statcan.ca/english/freepub/21-004-XIE/21-004-XIE2002112.pdf, 11 pp.

Health Canada (1994) Guidelines for the safety assesment of novel foods, Volume II, Genetically modified microorganisms and plants. Health Canada, Food Directorate, Ottawa, Canada. http://www.hc-sc.gc.ca/fn-an/legislation/guide-ld/nvvlii01_e.html, 27 pp.

Health Canada (1996) Glyphosate tolerant soybean 40-3-2. Health Canada, Food Directorate, Ottawa, Canada. http://www.hc-sc.gc.ca/fn-an/gmf-agm/appro/ofb-096-100-d-rev_e.html.

Health Canada (2004) Imidazolinone tolerant Clearfield^{TM} wheat (Teal 11A). Health Canada, Food Directorate, Ottawa, Canada. http://www.hc-sc.gc.ca/fn-an/gmf-agm/appro/nf-an102decdoc_e.html.

Hubbell E, Liu W-M, Mei R (2002) Robust estimators for expression analysis. Bioinformatics **18:** 1585-1592.

Hughes TR, Shoemaker DD (2001) DNA microarrays for expression profiling. Current Opinion in Chemical Biology **5:** 21-25.

ILSI (2005) Crop composition database Version 2.0. International Life Science Institute. www.cropcomposition.org.

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

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003b) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research **31:** 1-8.

James C (2004) ISSA briefs: Global status of biotech/GM crops: 2004 (Preview). The International Service for the Acquisition of Agri-biotech Applications (ISAAA), Ithaca, NY, http://www.isaaa.org/, 12 pp.

Jefferson RA, Kavanagh TA, Bevan MW (1987) GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. The EMBO Journal **6:** 3901-3907.

King CA, Purcell LC, Vories ED (2001) Plant growth and nitrogenase activity of glyphosate-tolerant soybean in response to foliar glyphosate applications. Agronomy Journal **93:** 179-186.

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

Kohli A, Twyman RM, Abranches R, Wegel E, Stoger E, Christou P (2003) Transgene integration, organization and interaction in plants. Plant Molecular Biology **52:** 247-258.

Kok EJ, Kuiper HA (2003) Comparative safety assessment for biotech crops. Trends in Biotechnology **21:** 439-444.

Konig A, Cockburn A, Crevel RWR, Debruyne E, Grafstroem R, Hammerling U, Kimber I, Knudsen I, Kuiper HA, Peijnenburg AACM (2004a) Assessment of the safety of foods derived from genetically modified (GM) crops. Food and Chemical Toxicology **42:** 1047-1088.

Konig A, Kleter G, Hammes W, Knudson I, Kuiper H (2004b) Genetically modified crops in the EU: food safety assessment, regulation, and public concerns - Overarching report. European Network on Safety Assessment of Genetically Modified Food Crops (ENTRANSFOOD), Wageningen, UR, The Netherlands. www.entransfood.com, 99 pp.

Kristensen C, Morant M, Olsen CE, Ekstrom CT, Galbraith DW, Lindberg Moller B, Bak S (2005) Metabolic engineering of dhurrin in transgenic *Arabidopsis* plants with marginal inadvertent effects on the metabolome

and transcriptome. Proceedings of the National Academy of Sciences of the United States of America **102:** 1779-1784.

Krupp G (2005) Stringent RNA quality control using the Agilent 2100 bioanalyzer. Agilent Technologies, Palo Alto, CA, www.agilent.com/ch/labonachip, 10 pp.

Kuiper HA, Kok EJ, Engel K-H (2003) Exploitation of molecular profiling techniques for GM food safety assessment. Current Opinion in Biotechnology **14:** 238-243.

Kumudini S, Hume DJ, Chu G (2002) Genetic improvement in short-season soybeans: II. Nitrogen accumulation, remobilization, and partitioning. Crop Science **42:** 141-145.

Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J (2005) Independence and reproducibility across microarray platforms. Nature Methods **2:** 337-344.

Lee CK, Klopp RG, Weindruch R, Prolla TA (1999) Gene expression profile of aging and its retardation by caloric restriction. Science **285:** 1390-1393.

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Research **33:** D71-74.

Lehesranta SJ, Davies HV, Shepherd LVT, Nunan N, McNicol JW, Auriola S, Koistinen KM, Suomalainen S, Kokko HI, Karenlampi SO (2005) Comparison of tuber proteomes of potato varieties, landraces, and genetically modified lines. Plant Physiology **138:** 1690-1699.

Leung YF, Cavalieri D (2003) Fundamentals of cDNA microarray data analysis. Trends in Genetics **19:** 649-659.

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

Li C, Wong WH (2001b) Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. Genome Biology **2:** 11-19.

Li J, Pankratz M, Johnson JA (2002) Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. Toxicological Sciences **69:** 383-390.

Li J, Spletter ML, Johnson JA (2005) Dissecting tBHQ induced ARE-driven gene expresson through long and short oligonucleotide arrays. Physiological Genomics **21:** 43-58.

Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DJ (1999) High density synthetic oligonucleotide arrays. Nature Genetics **21 (1 Suppl):** 20-24.

Liu Wm, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho Mh, Baid J, Smeekens SP (2002) Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics **18:** 1593-1599.

Lockhart DJ, Winzeler EA (2000) Genomics, gene expression and DNA arrays. Nature **405:** 827-836.

Lohnes DG, Bernard RL (1991) Ancestry of U.S./Canadian commercial cultivars developed by public institutions. Soybean Genetics Newsletter **18:** 243-255.

Macdonald P, Yarrow S (2003) Regulation of Bt crops in Canada. Journal of Invertebrate Pathology **83:** 93-99.

Mah N, Thelin A, Lu T, Nikolaus S, Kuhbacher T, Gurbuz Y, Eickhoff H, Kloppel G, Lehrach H, Mellgard B, Costello CM, Schreiber S (2004) A comparison of oligonucleotide and cDNA-based microarray systems. Physiological Genomics **16:** 361-370.

Manetti C, Bianchetti C, Bizzarri M, Casciani L, Castro C, D'Ascenzo G, Delfini M, Di Cocco ME, Lagana A, Miccheli A (2004) NMR-based metabonomic study of transgenic maize. Phytochemistry **65:** 3187-3198.

Meyers BC, Galbraith DW, Nelson T, Agrawal V (2004) Methods for transcriptional profiling in plants. Be fruitful and replicate. Plant Physiology **135:** 637-652.

Miller RA, Galecki A, Shmookler-Reis RJ (2001) Interpretation, design, and analysis of gene array expression experiments. The Journals of Gerontology Series A Biological Sciences and Medical Sciences **56:** B52-57.

Momma K, Hashimoto W, Ozawa S, Kawai S, Katsube T, Takaiwa F, Kito M, Utsumi S, Murata K (1999) Quality and safety evaluation of genetically engineered rice with soybean glycinin: Analyses of the grain composition and digestibility of glycinin in transgenic rice. Bioscience, Biotechnology, and Biochemistry **63:** 314-318.

NASS (2004) Acreage. USDA, National Agriculture Statistics Service, Agricultural Statistics Board, Washington, DC. http://usda.mannlib.cornell.edu/reports/nassr/field/pcp-bba/.

OECD (1993) Safety evaluation of foods derived by modern biotechnology, concepts and principles. Organisation for Economic Cooperation and Development. http://www.oecd.org/document/63/0,2340,en_2649_201185_1905919_1_1_1_1,00.html, 74 pp.

OOPSCC (2002-2005) Ontario Soybean Variety Trials. Ontario Oil and Protein Seed Crop Committee, Harrow, Ontario, Canada. www.oopscc.org.

Padgette SR, Kolacz KH, Delannay X, Re DB, LaVallee BJ, C.N. Tinius, Rhodes WK, Otero YI, Barry GF, Eichholz DA, Peschke VM, Nida DL, Taylor NB, Kishore GM (1995) Development, identification, and characterization of a glyphosate-tolerant soybean line. Crop Science **35:** 1451-1561.

Padgette SR, Taylor NB, Nida DL, Bailey MR, MacDonald J, Holden LR, Fuchs RL (1996) The composition of glyphosate-tolerant soybean seeds is equivalent to that of conventional soybeans. Journal of Nutrition **126:** 702-716.

Page A, Angell S (2002) Transient expression of reporter proteins can alter plant gene expression. Plant Science **163:** 431-437.

Park PJ, Cao YA, Lee SY, Kim J-W, Chang MS, Hart R, Choi S (2004) Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. Journal of Biotechnology **112:** 225-245.

Raymer PL, Grey TL (2003) Challenges in comparing transgenic and nontransgenic soybean cultivars. Crop Science **43:** 1584-1589.

Rostoks N, Borevitz J, Hedley P, Russell J, Mudie S, Morris J, Cardle L, Marshall D, Waugh R (2005) Single-feature polymorphism discovery in the barley transcriptome. Genome Biology **6:** 1-10.

Sanogo S, Yang XB, Scherm H (2000) Effects of herbicides on *Fusarium solani* sp. *glycines* and development of sudden death syndrome in glyphosate-resistant soybean. Phytopathology **90:** 57-66.

Sanogo S, Yang XB, Lundeen P (2001) Field response of glyphosate-tolerant soybean to herbicides and sudden death syndrome. Plant Disease **85:** 773-779.

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467-470.

Schulze A, Downward J (2001) Navigating gene expression using microarrays: A technology review. Nature Cell Biology **3:** E190-E195.

Seki M, Satou M, Sakurai T, Akiyama K, Iida K, Ishida J, Nakajima M, Enju A, Narusaka M, Fujita M, Oono Y, Kamei A, Yamaguchi-Shinozaki K, Shinozaki K (2004) RIKEN *Arabidopsis* full-length (RAFL) cDNA and its applications for expression profiling under abiotic stress conditions. Journal of Experimental Botany **55:** 213-223.

Seo J, Bakay M, Chen Y-W, Hilmer S, Shneiderman B, Hoffman EP (2004) Interactively optimizing signal-to-noise ratios in expression profiling: project-specific algorithm selection and detection p-value weighting in Affymetrix microarrays. Bioinformatics **20:** 2534-2544.

Shewmaker CK, Sheehy JA, Daley M, Colburn S, Ke DY (1999) Seed-specific overexpression of phytoene synthase: increase in carotenoids and other metabolic effects. The Plant Journal **20:** 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, Granger C (2002) A compilation of soybean ESTs: generation and analysis. Genome **45:** 329-338.

Simon RM, Korn EL, McShane LL, Wright GW, Zhao Y (2003) Design and analysis of microarray investigations. Springer-Verlag, New York, 199 pp.

Slonim DK (2002) From patterns to pathways: gene expression data analysis comes of age. Nature Genetics **32:** 502-508.

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

Statistics Canada (2004) Field Crop Reporting Series: November Estimate of Production of Principal Field Crops, Vol. 83, no. 8. Statistics Canada, Ottawa, Ontario, Canada. http://www.statcan.ca/Daily/English/041208/d041208a.htm, 6 pp.

Stewart CN (2001) The utility of green fluorescent protein in transgenic plants. Plant Cell Reports **20:** 376-382.

Tanner JW, Luzzi BM, Gostovic P, Montminy W, Hume DJ (1998) OAC Bayfield soybean. Canadian Journal of Plant Science **78:** 625-626.

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

TIGR (2004) Soybean (*Glycine max*) Gene Index (GmGI). The Institute for Genomic Research. http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=soybean.

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America **98:** 5116-5121.

Wang H-Y, Malek R, Kwitek A, Greene A, Luu T, Behbahani B, Frank B, Quackenbush J, Lee N (2003) Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. Genome Biology **4:** 1-13.

Whitham SA, Quan S, Chang H-S, Cooper B, Estes B, Zhu T, Wang X, Hou Y-M (2003) Diverse RNA viruses elicit the expression of common sets of genes in susceptible *Arabidopsis thaliana* plants. The Plant Journal **33:** 271-283.

Williams EJB, Bowles DJ (2004) Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Research **14:** 1060-1067.

Windels P, Taverniers I, Depicker A, Van Bockstaele E, De Loose M (2001) Characterization of the Roundup Ready® soybean insert. European Food Research Technology **213:** 107-112.

Wu Z, Irizarry R (2005) A statistical framework for the analysis of microarray probe-level data. The Berkeley Electronic Press. http://www.bepress.com/jhubiostat/paper73/, 33 pp.

Yang I, Chen E, Hasseman J, Liang W, Frank B, Wang S, Sharov V, Saeed A, White J, Li J, Lee N, Yeatman T, Quackenbush J (2002) Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biology **3:** 1-12.

Yang YH, Xiao Y, Segal MR (2005) Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics **21:** 1084-1093.

Zablotowicz RM, Reddy KN (2004) Impact of glyphosate on the *Bradyrhizobium japonicum* symbiosis with glyphosate-resistant transgenic soybean: a minireview. Journal Environmental Quality **33:** 825-831.

Zhang X, Feng B, Zhang Q, Zhang D, Altman N, Ma H (2005) Genome-wide expression profiling and identification of gene activities during early flower development in *Arabidopsis*. Plant Molecular Biology **58:** 401-419.

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics **163:** 1123-1134.

Zik M, Irish VF (2003) Global identification of target genes regulated by APETALA3 and PISTILLATA floral homeotic gene action. Plant Cell **15:** 207-222.