VALIDATION OF RICE OLIGONUCLEOTIDE ARRAY

Jirapa Phetsom

A Thesis Submitted in Partial Fulfillment of the Requirements for

the Degree of Doctor of Philosophy in Biotechnology

Suranaree University of Technology

Academic Year 2006

การตรวจสอบโอลิโกนิวคลิโอไทด์อเรย์ของข้าว

นางสาวจิรภา เพชร์สม

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต

สาขาวิชาเทคโนโลยีชีวภาพ

มหาวิทยาลัยเทคโนโลยีสุรนารี

ปีการศึกษา 2549

VALIDATION OF RICE OLIGONUCLEOTIDE ARRAY

Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

Thesis Examining Committee

(Asst. Prof. Dr. Chokchai Wanapu)

Chairperson

(Asst. Prof. Dr. Mariena Ketudat-Cairns)

Member (Thesis Advisor)

(Assoc. Prof. Dr. James Ketudat-Cairns)

Member

(Dr. Rodjana Opassiri)

Member

(Dr. Sompid Samipak)

Member

(Assoc. Prof. Dr. Saowanee Rattanaphani)
Vice Rector for Academic Affairs

(Asst. Prof. Dr. Suwayd Ningsanond)
Dean of Institute of Agricultural Technology

จิรภา เพชร์สม : การตรวจสอบคุณภาพโอลิโกนิวคลิโอไทด์อเรย์ของข้าว (VALIDATION OF RICE OLIGONUCLEOTIDE ARRAY) อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.มารินา เกตุทัต-คาร์นส์, 158 หน้า

โอลิโกนิวคลิโอไทด์อเรย์ของข้าว 2 ขนาด ได้แก่ 20K และ 45K ที่นำมาศึกษา ได้ถูกสร้างขึ้นโดยการสนับสนุนจาก National Science Foundation (NSF) อเรย์ของข้าวขนาด 20K ถูกออกแบบด้วย TIGR gene model version 2 ซึ่งได้ทำขึ้นก่อนที่การศึกษายีโนมข้าวจะเสร็จสิ้น หลังจากที่ข้อมูลลำดับนิวคลิโอไทด์ของยีโนมสมบูรณ์ จึงได้มีการออกแบบอเรย์ขนาด 45K โดยใช้ TIGR gene model version 3 อเรย์ทั้งสองระบบได้รับการตรวจสอบด้วยเทคนิคการวิเคราะห์ทางสถิติและชีววิทยา ได้นำการวิเคราะห์ทางสถิติใช้ในการเปรียบเทียบและตรวจสอบความน่าเชื่อถือของข้อมูล 20K และ 45K ซึ่งประกอบไปด้วยค่าเฉลี่ยของความเข้มของแสง, ค่าลอการิทึม ($\log_2$) ของอัตราส่วน, กราฟการกระจายตัวของข้อมูลในแบบจุด (scatter) และแบบแท่ง (smoothed histogram) การวิเคราะห์ข้อมูลทางชีววิทยาได้จากการจัดกลุ่มของยีนที่สนใจในฐานข้อมูลการจำแนกยีน (Gene Ontology)

ได้ทำการตรวจสอบอุณหภูมิที่เหมาะสม ในการจับกันระหว่างตัวอย่างกับตัวติดตามของอเรย์ขนาด 20K ที่ 42, 46, และ 50 องศาเซลเซียส ซึ่งตัวอย่าง cDNA ที่นำมาทดสอบ ได้มาจากใบข้าวสายพันธุ์ nipponbare ที่ปลูกในที่มีแสงและที่มืด ผลการวิเคราะห์ทางสถิติพบว่าข้อมูลอเรย์ที่อุณหภูมิ 42 องศาเซลเซียส แสดงค่าความเข้มของสารตัวติดตาม รวมทั้งค่าแสดงความสัมพันธ์ของข้อมูลในชุดทดสอบได้สูงที่สุด จึงได้เลือกใช้ 42 องศาเซลเซียส ในการศึกษาอเรย์ขนาด 45K ตัวอย่างที่นำมาทดสอบกับอเรย์ 45K ประกอบไปด้วยข้าวสี่สายพันธุ์ได้แก่ Nipponbare, Kitaake, Taipei 309, และ IR24 ที่ปลูกในสภาวะที่มีแสงและที่มืดเช่นเดียวกับการทดสอบกับอเรย์ 20K จากการวิเคราะห์ผลของอเรย์ทางชีววิทยาโดยอาศัยฐานข้อมูลการจัดกลุ่มความสัมพันธ์ของยีน เพื่อการจัดความสัมพันธ์ของข้อมูลอเรย์ใน 20K และ 45K พบว่ายีนในกลุ่มที่เกี่ยวข้องกับการสังเคราะห์ด้วยแสง มีผลต่อสภาวะการทดสอบ นอกจากนั้นข้อมูลของอเรย์ 45K สามารถนำมาศึกษากลุ่มของยีนที่มีลำดับใกล้เคียงกันได้ เช่น glycosyl hydrolase family 1

สาขาวิชาเทคโนโลยีชีวภาพ ลายมือชื่อนักศึกษา_____

ปีการศึกษา 2549 ลายมือชื่ออาจารย์ที่ปรึกษา_____

ลายมือชื่ออาจารย์ที่ปรึกษาร่วม_____

JIRAPA PHETSOM : VALIDATION OF RICE OLIGONUCLEOTIDE ARRAY. THESIS ADVISOR : ASST. PROF. MARIENA KETUDAT-CAIRNS, Ph.D. 158 PP.

RICE ARRAY/VALIDATION/OLIGONUCLEOTIDE/MICROARRAY

Two rice oligonucleotide microarrays of 20K and 45K were constructed with support from the National Science Foundation (NSF). The rice 20K array was designed with TIGR gene model version 2 prior to the complete genome assembly. After the genome sequence was completed, the rice 45K array was designed with TIGR gene model version 3. These two rice array systems were validated with statistical and biological analysis. The statistical analyses were used to compare and test the reliability of the 20K and 45K arrays, which included average mean intensity, $\log_2$ratio, scatter plot, and smoothed histogram. The biological analysis was done by classifying candidate genes with Gene Ontology database. To obtain images with high quality, the annealing temperatures of 42, 46, and 50 ^{o}C were tested in the 20K array. The testing samples were derived from nipponbare leaves treated with light and dark growing conditions. The array data for 42 ^{o}C showed the highest signal intensity and correlation coefficient value, therefore, it was selected and used in further study of the 45K array. The tested samples for the 45K array, which consisted of four rice cultivars of Nipponbare, Kitaake, Taipei 309, and IR24, were treated with light and dark growing condition similar to the 20K array samples. The biological analysis with Gene Ontology classification, both the 20K and 45K arrays results showed that the photosynthesis related genes were affected by the treatment

condition. The 45K array data can be used to analyze multigene families which have similar homologous sequences. Glycosyl hydrolase family 1 was used as an example of multigene family for data analysis from the 45K array.

School of Biotechnology

Academic Year 2006

Student's Signature_____

Advisor's Signature_____

Co-advisor's Signature_____

ACKNOWLEDGEMENTS

My study is supported by the university lecturer development program from the Ministry of Education, Thailand.

I would like to express my sincere gratitude to all kind help during this work: my advisor, Asst. Prof. Dr. Mariena Ketudat-Cairns, I'm thankful for her kind attitude, valuable advice and many stimulating discussions throughout my PhD work. I am also profoundly indebted to my other co-advisor, Assoc. Prof. Dr. James R. Ketudat-Cairns, for sharing valuable knowledge and inspiration. I am grateful to Prof. Dr. Pamela C. Ronald, her kind attitude of accepting me to join the notable rice array project. I would like to thank Dr. Ki-hong Jung, and Dr. Geun Cheol Lee for sharing their knowledge and working experiences.

The also thank my friends and my colleagues in School of Biotechnology at SUT and all the people in the UCD laboratory for their warm wishes and encouragement.

Most importantly, I would like to express my deepest gratitude to my dear family for their love, support, understanding and encouragement.

Jirapa Phetsom

CONTENTS

Page

CONTENTS (Continued)

Page

CONTENTS (Continued)

Page

CONTENTS (Continued)

Page

CONTENTS (Continued)

Page

CONTENTS (Continued)

Page

LIST OF TABLES

Table **Page**

LIST OF TABLES (Continued)

Table **Page**

LIST OF TABLES (Continued)

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES (Continued)

CHAPTER I

INTRODUCTION

Rice is considered a model species for monocots. Drafts of the genomic sequence had been generated by Monsanto (japonica cultivar Nipponbare), Syngenta (japonica cultivar Nipponbare) and the Beijing Genomics Institute (indica cultivar). The small size of the rice genome has more benefits to be a model for molecular biology studying. The rice gene sequences are also relatively conserved with other economically important cereal crops such as maize, wheat, barley, sorghum, rye, and sugarcane. Moreover, EST and full length cDNA sequences are available publicly. Recently, a variety of techniques allow parallel assessment of gene expression on a global scale, including serial analysis of gene expression (SAGE) (Velculescu *et al*., 1995), differential display, oligonucleotide arrays (Lockhart *et al*., 1996), and cDNA microarray (Schena *et al*., 1995; Schena *et al*., 1996).

Microarray technique is powerful tool to monitor gene expression patterns of tens of thousands of genes in a short time (Lockhart *et al*., 1996). Several vendors provide rice arrays, such as Agilent, Affymetrix, and BGI. Currently, DNA microarrays are manufactured using either cDNA or oligonucleotides as gene probes. A cDNA microarray is created by spotting amplified cDNA fragments in a high density pattern onto a glass slide (Lockhart *et al*., 1996; Velculescu *et al*., 1995). Oligonucleotides array are either spotted or constructed by chemically synthesizing

approximately 25-mer oligonucleotide probes directly onto a glass or silicon surface using photolithographic technology (Schena *et al*., 1995).

The rice array provided by the NSF Rice Oligonucleotide project was used as model in this study. The array composed of long oligonucleotide of 50- to 70-mers covering most of the rice genome. Two array systems of 20K and 45K were used in this study. The 20K array was designed based on TIGR gene model version 2. This array consisted of 21,120 spots which corresponding to 20,229 unique oligos for unique genes and 891 spots for control elements. After the rice genome sequence was completed, the project reconstructed the oligos and printed the 45K array. The new 45K array was designed based on gene models from the TIGR Rice Annotation Database that has EST and/or full-length cDNA support. These sequences were cross referenced to the Kikuchi full-length cDNA dataset. The oligo design of the 45K array consists of 43,312 oligos corresponding to 44,974 TIGR V3 rice gene models. The rice whole genome sequences were computed with PICKY2 software to calculate the optimum annealing temperature and unique sequence for each specific gene.

The 20K and 45K arrays used the same platform, glass slides spotted with 50- to 70-mer synthetic oligonucleotides. The 45K array is composed of two slides, but the 20K array is presented on a single slide. To validate the array quality, the rice samples were prepared with light and dark treated conditions. The nipponbare rice seedlings were grown for 2 weeks before collecting the leaf tissue. The dark and light samples were labeled with different fluorescent dyes, cyanine 3 and cyanine 5. To optimize the annealing temperature on the 20K array, three different temperatures

were set up at 42, 46, and 50 ^{o}C. Hybridization, washing, and imaging were performed with automated machine and specific software. Each hybridization produced a pair of 16-bit images, which were processed by the GenePix software to measure the intensity values. The software analyzed the image with two specific channels for R and G (R = red for Cy5 and G = green for Cy3), which were characterized as light and dark samples. The array data were analyzed with the R program using LMGene package for both the 20K and the 45K array. The differentially expressed genes were selected to confirm the reliability of microarray technique among three different annealing temperatures.

The rice seedlings were prepared similarly for the 20K and 45K arrays. However, for the 45K array, four different rice cultivars were used to compare with the 20K array results. Several methods can be used to validate the array data for example real-time, quantitative RT-PCR, northern analysis or nuclease protection assay. In this study, the quantitative RT-PCR technique was selected to confirm the array result.

The microarray experiment has been performed by many laboratories with different platforms. For sharing and mining the data, standard procedures were set up for users, Minimum Information About a Microarray Experiment (MIAME) (Brazma *et al.*, 2001) was one of them. This tool can be used as the data repository for all users, and it is also easy for searching the prior array experiments which have been done.

In this research, the NSF rice oligonucleotide arrays of 20K and 45K were validated using light/dark rice treatment. Statistical analysis and RT-PCR were used to validate the array results. Gene ontology was also performed to classify the array results in order to interpret their biological meaning.

Objective of this research

1. Validation of 20K array with three different annealing temperatures (42, 46, and 50 ^{o}C), statistical and biological analysis.
2. Validation of 45K array with statistical and biological analysis.

CHAPTER II

LITERATURE REVIEW

2.1 Rice

Rice (*Oryza sativa* L.) is a model organism and one of the most important food crops in the world, providing a food staple for more than one quarter of the world population. Rice is also considered as a model cereal plant for genetic and molecular studies because its genome is smaller than other cereals (Devos and Gale, 2000). The progress study is emerged from the genome sequence project that facilitates the researcher to understand the molecular biology of rice. The draft genome sequence for *Oryza sativa* L. ssp. indica (Yu *et al*., 2002) and *Oryza sativa* L. ssp. japonica (Goff *et al*., 2002) provide a rich resource for understanding the biological process in rice.

2.1.1 Rice cultivars

Rice consists of many cultivars with various origins and genotypes. All of cultivars are derived from two different types, japonica and indica. For the collection of rice, the international agricultural research institute in Asia is established in the International Rice Research Institute (IRRI). The IRRI collects information about rice to assist research. Many rice cultivars are collected in IRRI, which al keeps theirinformation and gramne reference. The sample rice cultivars used in this research are listed in Table 2-1.

Table 2-1 The property of rice in various cultivars.

Name	Cultivar	Country	Reference
KITAAKE	Japonica	JAPAN	(Murray *et al*., 2002)
NIPPONBARE	Japonica	JAPAN	(Ohtsubo *et al*., 1991)
TAIPEI 309	Japonica	TAIWAN	(Buchholz *et al*., 1998)
IR24	Indica	PHILIPPINES	(Yasui *et al*., 1996)

2.2. Microarray

2.2.1 Principle of microarray

Microarray is an analytical tool to explore the genome with speed and precision. The microarray on a glass chip composes tens of thousands of gene probes. These gene probes are used to hybridize to fluorescent samples prepared by using the messenger RNA (mRNA) as a template to obtain labeled cDNA from cells, tissues, and other biological sources. The fluorescent molecules are hybridized to the DNA on the array via Watson-Crick duplex formation. These fluorescent molecules react with sequences on the chip that cause the spots to glow. The glowing spots represent the proportion of activity of expressed gene. The microarray tool allows the analysis of the entire genome in a single experiment (Schena, 2003). Microarrays were developed at Standford University by Schena and co-workers in early 1990s.

Other methodology to determine gene expression include semi-quantitative reverse transcription (RT-PCR), real-time PCR, northern blot, ribonuclease protection assay, and in situ hybridization or immuno histochemistry. Array results had been

compared with northern blots and showed good relation between methods (Taniguchi *et al*., 2001). The comparison between array results and real-time RT-PCR method, showed accuracy, but inconsistency for genes with lower than four-fold expression changes (Rajeevan *et al*., 2001).

2.2.2 Making the array

Two main methods widely used in making microarray are robotic spotting and *in situ* synthesis. The spotted microarrays provide the DNA probes spotted on a glass surface with a spotting robot. The *in situ* synthesized oligonucleotide arrays build the oligo probes base-by-base on the surface. This technique performs with creating the phosphodiester bond between the hydroxyl group of sugar molecule and phosphate group of the last nucleotide (Stekel, 2003).

Spotted arrays consist of three main types which can be classified in two ways, by type of DNA probe or by the attachment chemistry of the probe on glass slide (table 2.2). The DNA probes can be polymerase chain reaction (PCR) products or synthetic oligonucleotides.

Table 2-2 Type of attachment chemistry and DNA probes

DNA probes	Surface chemistry	
	Covalent	**Non-covalent**
Oligonucleotides	√	-
cDNAs	√	√

Note: derived from "Microarray Bioinformatics" (p. 4), Stekel, D., 2003, Cambridge : Cambridge University Press.

The chemical attachment can be classified as covalent and non-covalent bonding. In covalent bonding, a primary aliphatic amine (NH_2) group is added to the DNA probes, which are attached to the glass by making a covalent bond between this group and chemical linker on the glass slide. The amine group is usually added to the 5' end of oligonucleotide and PCR primer. In non-covalent bonding, the bonding is due to the electrostatic attraction between the phosphate backbone of the DNA probe and NH_2 group attached to the glass slide. The electrostatic bonding occurs at several locations along the DNA backbone making the probes tethered to the glass slide at many points.

2.2.3 Array platform

2.2.3.1 cDNA

The first cDNA array construction was performed from cDNA clones of *Arabidopsis* and human peripheral blood lymphocytes (Schena, 1996; Schena *et al*., 1995). A plasmid was used as a vector to carry the inserted DNA, which is specific to the gene sequence. The universal primers corresponding to the vector sequences were used to amplify the gene sequences. The amplified products were analyzed by gel electrophoresis and printed onto surface-modified glass slides by robotic printing. Amplified DNAs can be fixed onto the slides electrostatically, or through cross linking by heat or UV. Moreover, the PCR products can be covalently attached at their 5' ends to modified slides via an amine or other active group.

2.2.3.2 Short oligonucleotide

The first large-scale manufacturing of arrays is announced by the Affymetrix company that develope the photolithographic method to produce array (Lipshutz *et al*., 1999; Lockhart *et al*., 1996). The photolithographic method uses the synthetic linker with photolabile protecting group attaches to the glass substrate. The mask is applied to pre-determine areas on the substrate, which direct light to remove the exposed groups. Chemical coupling is followed by reaction between the de-protected groups and bi-functional deoxynucleosides. The step is repeated until the desired sequence and length of oligonuclotide are achieved by changing the new mask in

every round. The array is represented on a 1.28 x 1.28 cm chip containing up to 500,000 different oligonucleotide sequences of 25 bases in length. Each transcript or gene sequence is represented by eleven to twenty different oligonucleotides. These oligonucleotides are tiling or covering a portion of the 3' end of mRNA. The advantage of the platform is that it includes controls to detect background noise and cross-hybridization from unrelated probes by synthesizing oligonucleotides as perfect match and mismatch pairs (Lipshutz *et al*., 1999; Lockhart *et al*., 1996). The mismatch (MM) oligonucleotide has a one-base mismatch in the center position.

2.2.3.3 Long oligonucleotide

Long oligonucleotides are 40–80 bases in length. The oligonucleotides can be printed onto the same substrates (slides) as cDNAs, with the same printing device. The long oligonucleotides in the array should have very similar melting temperatures or G–C (guanosine–cytosine) content, have very little homology with other oligonucleotides, be entirely contained within an exon, and have no repetitive or hairpin sequences. Several commercial sources are available to print any size of oligonucleotides onto an appropriate substrate. The following companies provide this service: ClonTech, http://www.clontech.com; Compugen, http://www.cgen.com; MWG Biotech, http://www.mwg-biotech.com; and Operon, http://www.operon.com)

Long oligonucleotide arrays have more advantages than the other platforms. The longer length shows more specificity in hybridization than shorter length. The similar length of oligonucleotide has almost the same melting temperature and G-C content. Moreover, the concentration of oligonucleotides per spot is more consistent

than cDNA arrays. The single stranded oligonucleotides do not require a denaturation step and renaturation, which means that they hybridize efficiently (Barrett and Kawasaki, 2003).

2.2.4 NSF oligonucleotide array

Rice genome annotation release 4 was announced in an official website in January 12, 2006. This database supports several tools such as Flanking Sequence Tag searches, alignment with the Gene Indices, alignment with maize/wheat genetic markers, domain and motifs, and TIGR rice genome statistics. The NSF supported rice oligo 45K array was designed using gene models from the TIGR Rice Annotation Database that have EST and/or full-length cDNA support. And also, these oligos were cross referenced to the Kikuchi full-length cDNA dataset (www.ricearray.org).

The rice 20K array was designed based on TIGR V2 rice gene model. This array was provided on a single slide with 20,190 unique oligos. The spotted oligos array was printed on glass. The total number of spots including empty and controls is 21,120. The control elements consisted of 217 *Hph* spots and 674 empty oligonucleotide spots.

The NSF rice whole genome (45K) oligo array is composed of slide “A” and “B” which distribute spots in an equal. The two slides have 40,297 unique and 3,203 shared oligo probes. The 45K array was designed from 45,116 TIGR V3 rice gene models and matched to the TIGR Rice Annotation Project release 4, or flcDNA, or the

TIGR Plant Transcript Assembly Rice build 2 containing the oligo at 100% identity to databases. The oligos presented the annotation provided by blastx and UniGene. The 45K array included 163 oligos targeted for chloroplast and mitochondria, and 8 transgenes of *GUS*, *BAR* and *Hph*. All transgenes can be used as control depending on the experimental design (www.ricearray.org).

2.2.4.1 Array design (PICKY)

Picky is a program freeware for oligo microarray design that identifies probes with very unique and specific to input sequences. The software could calculate factors based on user including optimal probe length, ideal percentage of guanine and cytosine content, target-melting temperature, salt concentration and the maximum length to which a target sequence matches any non-target sequence (Chou *et al*., 2004).

Picky also considers more sophisticated design parameters. Picky accepts minimum and maximum oligo lengths instead of a fixed length. Within the specified range it can adjust the length of oligos to achieve greater specificity and uniformity among all oligos. Rather than requesting the melting temperature (T_m) as a parameter from users, Picky takes the minimum separation temperature as a parameter, and ranks best oligo candidates by comparing their target and non-target melting temperatures. Only oligos that provide at least the minimum separation temperature will be considered in Picky, and it is their joint temperature separations that finally

determine the optimal Tm suggested by Picky for microarray experiments. Picky also handles multiple target and non-target gene sets, where the non-target sets are used as a screening background while oligos are being designed for the target sets. This allows, for example, a small budget experiment to study a handful of genes from a large genome, but still guarantees the results will be as good as those obtained from a whole genome microarray (Chou. H. H. unpublished).

2.2.4.2 Designed diagram

The important step in experimental designed is to decide the number of technical replicates and how these will be paired together on array. Figure 2-1(a) uses two samples A and B and assigns as red and green dyes. The arrow indicates the reversed dyes between two samples and is allowed to repeat by using four (or six or more) arrays to compare the same biological samples, as shown in figure 2-1(b) (Churchill, 2002).

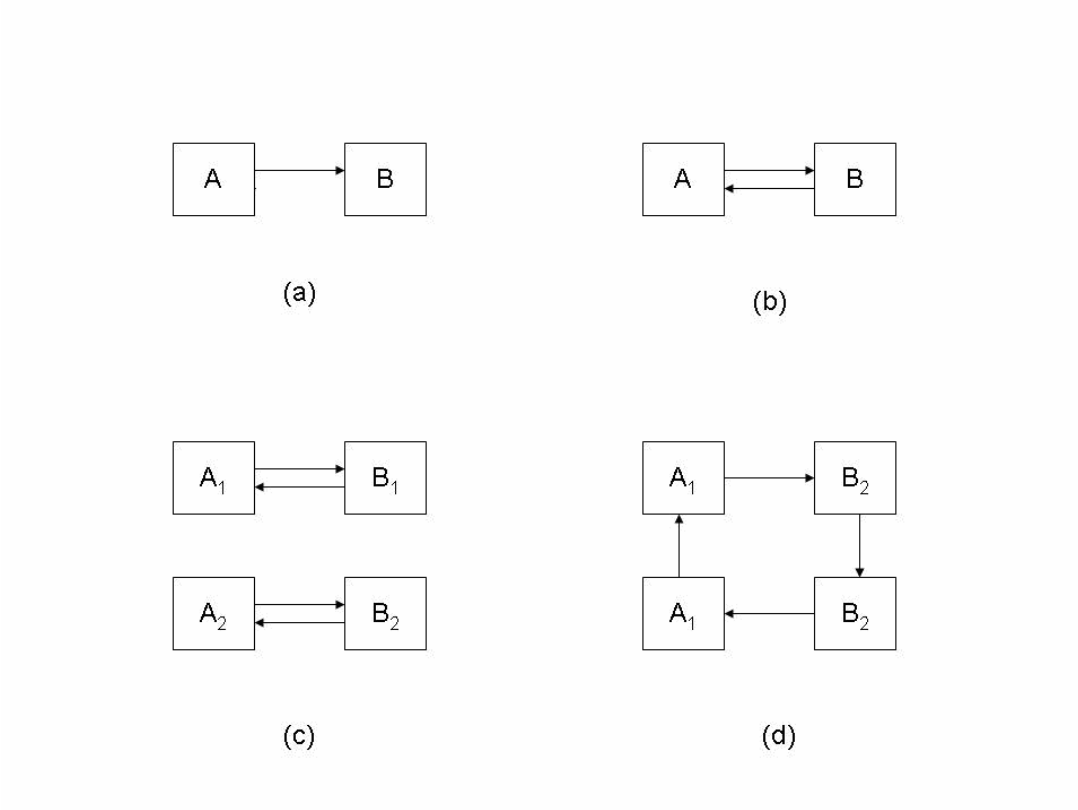


Figure 2-1 Experimental design diagram for direct comparison of two samples. Boxes; representing sample labeled as A or B. Subscript 1 and 2 indicate the number of biological replicates of the same treatment. Arrow; represent hybridization between samples. The sample at the tail of the arrow is labeled with red (cy5) dye, and the sample at the head of the arrow is labeled with green (cy3) dye. (a) dye swap, (b) a pair of dye swap, (c) a replicated pairs of dye swap, and (d) simple loop design. (modified from Churchilll, 2002)

2.3 Data processing

2.3.1 Image processing

The images of the microarray derive from the scanner which usually generates the raw data as TIFF (Tagged Information File Format) images. The computer software is the major analysis tool for array images and converting the images into the numerical information that quantifies gene expression. The image processing involves in feature extraction which impacts on the quality of array data (Stekel, 2003).

The different technology creating the microarray in different platforms, commercial array consists of *in situ* synthesis (Affymetrix), inkjet *in situ* synthesis (Rosetta, Agilent and Oxford Gene Technology), and pin-spotted microarray. Affymetrix has integrated image processing algorithms into the GeneChip experimental process. High quality inkjet arrays and image processing software are provided by the company. For pin-spotted arrays, various software programs are supported to quantify the array images and are provided by different sources. The computation analysis with image software converts the digital TIFF images of hybridization intensity into numerical measures of signal intensity. The quantification creates the numerical intensity for each channel on each feature. This process is called feature extraction and is composes of four steps; 1) identification of the position of the features on the microarray, 2) identification of the pixels on the image that are part of each feature, 3) identification nearby pixels that will be used for

background calculation for each feature, and 4) calculation of numerical values for the intensity of the feature, the intensity of the background and quality control information (Stekel, 2003).

The software provides the algorithms to automatically find the positions of the features. However, most of software still requires manual supervision of the feature extraction process to ensure that all features are found by the software. Identification of pixels comprising the features in the feature extraction procedure is called segmentation. The segmentation is processed with software that determines which pixels in the area of a feature are part of the feature, and their intensity will count towards a quantitative measurement of intensity. There are four common methods for segmentation which include fixed circle, variable circle, histogram, and adaptive shape.

The fixed circle segmentation places a circle of fixed size over the region of the feature and uses all the pixels in the circle. The problem with fixed circle segmentation is that it gives inaccurate results if the features are different sizes. Variable circle segmentation fits a circle of variable size onto the region containing the feature. This method is able to resolve the problem with different feature sizes, but performs poorly with irregularly shaped features. Histogram segmentation fits a circle over the region of the feature and background and then looks at a histogram of the intensities of the pixels in the feature. The brightest and dimmest pixels are not used in the quantification of feature intensity. Histogram segmentation produces reliable results for irregularly shaped features. However, histogram methods can be

unstable for small features if the circular mask is too large. The fourth method, adaptive shape segmentation, can solve the problem with irregular shapes. The algorithm requires a small number of seed pixels in the center of each feature to start. Then, it extends the region of each feature by adjoining pixels that are similar in intensity to their neighbors. Examples of software packages that implement the segmentation methods are listed as following;

Table 2-3 Segmentation algorithms of common image-processing software packages

Segmentation method	**Software**
Fixed circle	ScanAnalyze
	GenePix
	QuantArray
Variable circle	QuantArray
	GenePix
	Dapple
	Agilent feature extraction
Histogram	ImaGene
	QuantArray
Adaptive Shape	Spot

Note: derived from “Microarray Bioinformatics” (p. 66), Stekel, D., 2003, Cambridge : Cambridge University Press.

Errors in the features commonly include signal intensity of feature, non-specific hybridization, and other fluorescence from the glass. These errors in the fluorescent signals could be estimated by calculating the background signal from pixels that are near each feature but are not part of any feature. The background intensity is subtracted from the feature intensity to provide a more reliable estimate of hybridization intensity to each feature.

2.3.2 Data repository (MIAME) (Brazma *et al*., 2001)

Many commercial and freeware storage databases are available. The most widely used database when publishing microarray experiments is Minimum Information About a Microarray Experiment (MIAME). Some others databases are BioArray Software Environment (BASE), Microarray Gene Expression Data Society (MGED), ArrayExpress at EBI and GEO at NCBI.

The concept of MIAME is to establish a standard protocol for recording and reporting microarray-based gene expression data which verify the database and public repositories. In various laboratories, the microarray experiments have different platforms and designs which produce data in various formats. The MIAME tool is useful in defining the content and structure of the necessary information, rather than the technical format.

Microarray experiments produce massive quantities of gene expression and other functional genomics data. The standard formats for presentation, analysis tools,

and databases are still inaccessible to broader users. The different microarray platforms and experimental designs produce data in various formats and with various normalization methods. It is difficult to integrate array data with different standardization. The widely acknowledged public repository for microarray data was developed by the National Center for Biotechnology Information (NCBI).

MIAME was developed by the Microarray Gene Expression Database group (MGE) (http://www.mged.org/), a grass-roots movement to develop standards for microarray data. Two guidelines were prepared by the Microarray Gene Expression Data Society (MGED) that lay out the minimum standards (Knudsen and Daston, 2005). The two guidelines are the minimum information about a microarray experiment (MIAME) and MIAME/Tox, which extends these guidelines to cover toxicogenomics experiments. These guidelines are available at http://www.mged.org/index.html.

The MIAME/Plant describes which biological details need to be captured for describing microarray experiments involving plants (Zimmermann *et al*., 2006). The objective of MIAME/Plant is to facilitate and normalize experiment and sample annotations.

Each section contains information that can be provided by using controlled vocabulary. The minimum information about a published microarray-based gene expression experiment includes the details and description in six sections as described below.

2.3.2.1 Experimental design: the set of hybridization experiment as a whole

The experimental design is all including a set of hybridizations that are inter-related and address a common biological question. Each experiment should have an author (submitter), contact information, and experiment title. The required information consists of type of experiment (such as normal-versus-disease comparison, time course, dose response, and so on), and experimental variables, which include parameters or conditions tested (such as time, dose, genetic variation or response to a treatment or compound). This section also provides general quality-related indicators, such as usage and types of replicates and quality control steps.

2.3.2.2 Array design: each array used and each element (spot, feature) on the array

This section provides a systematic definition of all arrays used in the experiment which includes the genes represented and their physical layout on the array. This section is composed of two important parts, a list of the physical arrays and their design. The array-type definition includes a description of the array as a whole, a description of each type of element or spot used, and a description of the specific properties of each element.

2.3.2.3 Samples: samples used, extract preparation and labeling

The MIAME "sample" concept represents the biological material or biomaterial used in gene expression profile study. This section includes the description of source of the original sample and any biological in vivo or in vitro treatments applied, the technical extraction of the nucleic acids, and their subsequent labeling.

2.3.2.4 Hybridizations: procedures and parameters

This section defines the laboratory conditions under which hybridizations were carried out. Other than a free-text description of the hybridization protocol, MIAME requires the specific hybridization parameters, choice of hybridization solution, nature of the blocking agent, wash procedure, quantity of labeled target used, hybridization time, volume, temperature and descriptions of the hybridization instruments.

2.3.2.5 Measurement: images, quantification and specifications

The actual experimental results are defined in this section. It consists of the three parts; (a) the original scans of the array (images), (b) the microarray quantification matrices based on image analysis, and (c) the final gene expression matrix after normalization and consolidation of possible replicates.

Image data should be provided as raw scanner image files (such as TIFF), accompanied by scanning information, such as scan parameters and laboratory protocols. For each experimental image, a microarray quantification matrix contains the complete image analysis output as directly generated by the image analysis software. The summarized information, gene expression matrix consists of sets of gene expression levels for each sample. At this point, the expression values may have been normalized, consolidated and transformed in order to present the data.

2.3.2.6 Normalization controls: types, values and specifications

A typical microarray experiment involves a number of hybridization assays which derive from multiple samples. The data from multiple samples are analyzed to identify relative changes in expression levels and identify differentially expressed genes. A typical experiment follows “reference design”, in which many samples are compared to a common reference sample that facilitates inferences about relative expression changes between samples.

For these comparisons, the reported hybridization intensities derived from image processing must first be normalized. The normalization adjusts for the number of technical variations between and within each hybridization. This process could balance the starting RNA and labeling, and detection efficiencies for each sample. The section 6 of the MIAME standard provides an opportunity for the specification of parameters relevant to normalization and control elements. The proposed standard includes (i) the normalization strategy (spiking, housekeeping genes, total array, or

other approach) (ii) the normalization and quality control algorithms used, (iii) the identities and location of the array elements serving as controls, their type (spiking, normalization, negative or positive hybridization controls), and (iv) hybridization extract preparation, detailing how the control samples are included in sample targets prior to hybridization.

2.3.3 Data analysis

The microarray technique is widely used to determine gene expression, and many models and algorithms have been developed for data analysis. Knowledge of mathematics, statistics and computer skills is necessary to utilize the analysis tools. An open source project on genomic data analysis is Bioconductor (Sanmiya et al., 1997). A wide range of packages available for microarray data analysis, however, command line interface programming skills are essential for use of Bioconductor and R computer language (Xia *et al*., 2005). These skills are complex for biologist to acquire and utilize. Therefore, many tools have been developed as user-friendly programs such as WebArray, which includes some packages from Bioconductor and others.

Before starting the data analysis, raw data must be normalized to adjust the variations which occurred in the quantity of starting RNA or efficiency of detection and fluorescent labeling of the sample. Then, the candidate genes can be determined with the false discovery rate (FDR), which is better than p-value because of specification of the confidence of microarray data (Pounds and Cheng, 2004).

2.3.3.1 Normalization

Normalization is important part for further data analysis. What normalization strategy is chosen depends on the experimental design. The array data could be affected by systematic (non-biological) differences between samples on a slide. The common methods of normalization can be divided into three types, including total intensity normalization, normalization using regression technique, and normalization using ratio statistics (Causton *et al*., 2003).

2.3.3.1.1 Total intensity normalization

Total intensity normalization relies on two assumptions. First, the total quantity of starting mRNA is the same for both samples. Second, the selected genes for normalization are not biased by significant differential expression between the samples. Then, comparison of the intensity for both samples, summed over all of the features in the normalization set should be equal.

2.3.3.1.2 Mean log centering

Total intensity normalization by mean log-centering normalization relies on the assumption that mean $\log_2$(ratio) should be equal to 0. The normalization constant can be calculated for all features in the normalization set. However, this approach has a pitfall in outlying to correct the significant up-regulated gene by the expression ratio.

2.3.3.1.3 Linear regression

For closely related samples, many genes have expression at nearly constant levels. In a scatter plot of intensities (or their logarithms) from the two samples, these genes would cluster along a straight line and fit slope 1, if the labeling and detection efficiencies are the same for both samples. Normalization of these data is equivalent to calculating the best fit slope using regression techniques (Chatterjee and Price, 1991) and adjusting the intensities to give a calculated slope of one.

2.3.3.1.4 Chen's ratio statistic

Another approach is the ratio statistics method, which was developed by Chen and his colleges in 1997 (Chen *et al*., 1997). In this method, housekeeping genes are used to calibrate the ratio of cy3 to cy5 which has to be one. The constant coefficient is calculated as in the first method for adjusting the entire array data.

However, the expression ratio has the disadvantage of representing the up- and down-regulated genes differently. The genes up-regulated by a factor of 2 have an expression ratio of 2, whereas genes down-regulated by the same factor have an expression ratio of -0.5. The alternative transformation is representing the array data as its base 2 logarithm, which has the advantage of giving symmetrically distributed array data. The base 2 logarithm calculates the ratio between dye1 and dye2 intensities and represents the up- and down-regulated gene in a similar fashion. For instance, genes up-regulated by a factor of 2 have a $\log_2$(ratio) of 1, genes down-

regulated by a factor of 2 have a $\log_2$(ratio) of -1, and a gene expressed at a constant level (ratio of 1) has a $\log_2$(ratio) equal to zero (Quackenbush, 2002).

2.3.3.1.5 Lowess normalization

Normalizing data is equivalent to calculating the best-fit slope using regression techniques (Hedenfalk, 2001) and adjusting the intensities to give a calculated slope of one. The LOWESS (LOcally WEighted Scatterplot Smoothing) regression technique has been proposed as a normalization method for microarray data (Cleveland, 1979). Most experiments present the nonlinear intensities, for which local regression techniques, such as LOWESS regression are more suitable. (Cleveland and Devlin, 1988).

The lowess method can be applied either locally (physical subset of the data) or globally (entire data set) (Yang *et al*., 2002). Local normalization is used to reduce artifacts in specific location on the array, or specific (housekeeping) genes. The microarray data in each experiment can use the global normalization to diminish the systematic errors, such as techniques or samples.

2.3.3.1.6 Global vs. local normalization

Most normalization algorithms can be applied either globally to the entire dataset or locally to subsets of the data, such as subgrids or pen groups. Applying these algorithms to a single subgrid can correct the local systematic

variation due to spotting pens, slide surface, and slight local differences in hybridization conditions across the array. If the exogenous spiked-in controls or set of selected housekeeping gene are selected, all of these control features must be present in each subgrid.

2.3.3.2 Scatter plots : (Diagonal and MA-plot)

A scatter plot is the most common graphical display for microarray data that represents intensity for both channels transformed data as $\log_2(R_{intensity})$ versus $\log_2(G_{intensity})$ in x- and y-axes. The spot distribution is typically diagonal (Yang and Dudoit, 2002). The spots distributed away from the diagonal represent differentially expressed genes. The cloud of data represents the distribution of spot intensities (Bowtell and Sambrook, 2002).

An alternative type of scatter plot is an MA plot. An M vs. A plot is representation of the R and G data in terms of the log intensity ratios M. The MA-plot is more revealing than $\log_2$ R vs. $\log_2$ G counterparts in terms of identifying spot artifacts and detecting intensity dependent patterns in the log ratios. "M" is presented the $\log_2$ (R/G) and "A" is $\frac{1}{2}\log_2(RxG)$.

The loess method performed a local scatter plot smoothing to the MA plot. This plot is constructed with the $\log_2$ ratio of average intensity and linear normalization. The data is normalized to the regression line by subtracting the fitted value on the line from the log ratio of each feature. The regression line is transformed

to a horizontal line through zero. The spots with the highest intensities lie above the line. Typically, the spots at points away from the diagonal, represent the differentially expressed genes. The data are rotated by 45^{o} along the axes. An MA-plot serves to increase the room available to represent the range of differential expression. The MA-plot is also useful to identify spot artifacts and detect intensity-dependent patterns in the log ratios M.

2.4 Gene Ontology

The goal of the Gene Ontology Consortium is to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes, even as knowledge of gene and protein roles in cells is accumulating and changing. The gene ontology produced a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism. Three independent ontologies accessible on the World-Wide Web (http://www.geneontology.org) are being constructed: biological process, molecular function and cellular component (Ashburner *et al*., 2000). The GO concept is intended to allow annotation of homologous gene and protein sequences in multiple organisms using a common vocabulary. The GO ontologies produce a controlled vocabulary that can be used for dynamic maintenance and interoperability between genome databases.

In 2000, the genomic sequences of three model systems were already available (budding yeast, *Saccharomyces cerevisiae*, completed in 1996 (Goffeau *et al*., 1996); the nematode worm *Caenorhabditis elegans*, completed in 1998 (1998); and the fruit

fly *Drosophila melanogaster* (Adams *et al*., 2000). The first comparison was processed on two complete eukaryotic genomes (budding yeast and worm) that revealed a large fraction of the genes displayed evidence of orthology. About 12% of the worm genes (~18,000) encode proteins whose biological roles could be inferred from their similarity to their putative orthologues in yeast, comprising about 27% of the yeast genes (~5,700). Most of these proteins have been found to have a role in the 'core biological processes' common to all eukaryotic cells, such as DNA replication, transcription, and metabolism. Each node in the GO was linked to other kinds of information, including the many gene and protein keyword databases such as SwissPROT, GenBank, EMBL, DDBJ, PIR, MIPS, YPD & WormPD, Pfam, SCOP, and ENZYME.

GO is classified into three classes as described below (Ashburner *et al.*, 2000);

2.4.1 Biological process

Biological process refers to a biological objective to which the gene or gene product contributes. A process is accomplished in one or more ordered assemblies of molecular functions, and the processes often involve a chemical or physical transformation. The biological processes include several terms such as "cell growth and maintenance", "signal transduction", "translation", "pyrimidine metabolism", and "cAMP biosynthesis".

2.4.2 Cellular component

Cellular component refers to the place in the cell where a gene product is active. Cellular component includes such terms as ‘ribosome’ or ‘proteosome’, specifying where multiple gene products would be found. It also includes terms such as ‘nuclear membrane’ or ‘Golgi apparatus’.

2.4.3 Molecular function

Molecular function is defined as the biochemical activity (including specific binding to ligands or structures) of a gene product. This definition is also applied to the capability of a gene product (or gene product complex) which carries as a potential. The samples of this class are represented as ‘enzyme’, ‘transporter’ or ‘ligand’, ‘adenylate cyclase’, and ‘Toll receptor ligand’.

In addition, mapping function to gene products in the genome consists of two parts which include ontology building and ontology annotation (Thomas *et al*., 2007). Ontology building is the formal representation of a domain of knowledge. Ontology annotation is association of specific genomic regions (genes, their regulatory elements and products) to parts of the ontology. Two complementary representations of gene function: the Gene Ontology (GO) and pathway ontologies are considered. Pathway ontologies provide detailed descriptions of the biochemical relationships between molecular types, and are connected to GO terms.

In 2006, the GO project (http://www.geneontology.org) developed and used a set of structured, controlled vocabularies for community use in annotating genes, gene products and sequences (also see http://song.sourceforge.net/). The GO Consortium GOC continued to improve to the vocabulary content, reflecting the impact of several novel mechanisms of incorporating community input. The Plant-Associated Microbe Gene Ontology (PAMGO) Interest Group collaborated with the GO Consortium to produce a new set of terms representing pathogenic and symbiotic processes from BioCyc databases.

Another database integration system is represented in term of "OntoFusion" (Alonso-Calvo *et al*., 2007). OntoFusion is a system for integrating databases that are either publicly available on the Internet or are directly accessible through DBMS. The system integrated seven significant and widely used public biomedical databases, including OMIM, PubMed, Enzyme, Prosite and Prosite documentation, PDB, SNP, and InterPro.

2.5 Photosynthesis

Photosynthesis is the primary process in plant, which stores the chemical energy from sunlight. The organic compounds are derived from dioxide and water with several processes occurred in plant cell. The photosynthesis process is located in chloroplast, which composes of double-membrane envelope and complex internal membranes. There are two types of internal membranes that are called "grana" and "stroma lamellae". These two membranes establish the thylakoids in the chloroplast.

Also, these two membranes contain the chlorophylls and other pigments for trapping the light energy. Plant pigments are classified in several types due to the absorption wavelengths, such as chlorophyll *a* and *b* absorb red, blue, indigo, and violet light. In particular blue light, it can enhance gene expression of farnesyl diphosphate synthase (*FPPS*) in etiolated rice (*Oryza sativa* L. variety Nipponbare) seedlings after three hours of illumination by a subtraction method (Sanmiya *et al*., 1997). The chlorophylls are also found in the thylakoids, and placed in the photosystem that is composed of photosystem I and II. Each photosystem contains several protein molecules complexed with pigment molecules, chlorophylls *a* and *b*. In addition, each photosystem has a complex of electron acceptor and donor molecules (Rost *et al*., 2006). It has been reported that chlorophyll-binding 22 kDa protein of Photosystem II can be induced with brief exposure to red light and increased of transcript level (Iwasaki *et al*., 1997). In higher plants, ferredoxin (Fd) and Fd : $NADP^+$ oxidoreductase (FNR, EC 1.18.1.2) are encoded by small multigene families, and the individuals transfer electrons to the dependent enzymes in the photosynthetic and the non-photosynthetic plastids (Sakakibara, 2003).

The light-regulated expression of photosynthetic genes is also adversely affected by tagetitoxin, a specific inhibitor of plastid-encoded RNA polymerase (Dhingra *et al*., 2004). The plastid *psbK-psbI-psbD-psbC* gene cluster of overlapping transcripts (RNAs *a-g*), have been reported that class *d* RNAs is induced by light. In dark growing condition, the accumulation of RNAs *a-g* has not found (Chena *et al*., 1994). Light can also modulate the development and differentiation of the photosynthetic organelle, chloroplast, by photomorphogenetic mechanisms, which are

involved in regulating transcription of various photosynthetic genes encoded by nuclear genome (Tyagi and Gaur, 2003).

The photosynthesis synthesizes the adenosine triphosphate (ATP) with the difference of proton concentration across the thylakoid membrane. The production of ATP in chloroplasts is called photophosphorylation, which is classified into two types of cyclic and noncyclic. Also, the nicotinamide adenine dinucleotide phosphate is produced in light reaction. The process occurs with driven the electron from water molecules to $NADP^+$ and forms NADPH. Within process, photosystem I and II support the electron by driving the energy supported each other.

The reduction of carbon dioxide to sugar is processed by enzymes catalysis. All the enzymes that participate in photosynthesis are occurred in the chloroplasts and found in stroma. The important enzyme with highest concentration in many leaf cells is ribulose bisphosphate carboxylase/oxygenase (rubisco). The process is called Calvin cycle that synthesizes the first product of three carbons. The product of phosphoglyceraldehyde can be converted to dihydroxyacetone phosphate and sugar phosphate, which involves in several enzymes (Rost *et al*., 2006).

In addition, light is not only used to activate the photosynthesis, but also stimulate the metabolic pathways related to energy transforming in the cell. Several enzymes, compartments, and processes are involved in light, which transform light to storage energy. Light can activate several activities in plant with related to plant development.

CHAPTER III

MATERIALS AND METHODS

Rice genes with differential expression in light and dark growing conditions were determined using microarray and RT-PCR techniques. All data were also analyzed to validate the arrays platform including the hybridization step.

3.1 Sample preparation

3.1.1 Plant materials

3.1.1.1 Light vs Dark condition for 20K array

Rice seedlings (Nipponbare cultivar) were used in this experiment. They were germinated without light for 3 days and then transferred to normal light in greenhouse and continuous dark in a dark growth chamber. The greenhouse and dark growth chamber temperature and humidity were set at 30 ^{o}C and 75%, respectively. After 2 weeks the leaves were collected and their RNA extracted.

3.1.1.2 Light vs Dark condition for 45K array

Four different rice cultivars (Nipponbare, Kitaake, IR24, and TP309) were germinated without light for 3 days and then transferred to a greenhouse for normal light and a growth chamber for continuous dark conditions. The temperature and humidity were set similar to the 20K array condition. After 2 weeks the leaves were collected and their RNA extracted.

3.1.2 RNA extraction

Rice leaves were collected and total RNA isolated with TRIZOL (Invitrogen, USA). The total RNA was treated with DNaseI for 15 minutes and purified with the RNeasy Midi Kit (Qiagene, USA). The total RNA was then enriched for poly-A RNA with the Oligotex mRNA Kit (Qiagen). All steps were done according to the manufacturer's instructions.

3.1.3 Labeling system

In this study, the samples were labeled with two different protocols derived from the CyScribe Post-labeling Kit (GE Healthcare, Bioscience, USA) and SuperScript Indirect cDNA labeling System (Invitrogen, USA).

3.1.3.1 CyScribe Post-labeling Kit

The cDNAs were labeled with either Cy3 or Cy5 fluorescent dyes (GE Healthcare, Bioscience, USA) using a two step procedure. The first step involved the incorporation of the amino allyl-dUTP (AA-dUTP) during cDNA sythesis. The *in vitro* reverse transcription was performed using 500 ng mRNA, combined with random hexamers and anchored oligo(dT) primers, then the reactions were incubated at 70°C for 5 min. After rapid cooling at 4 °C for 10 min, the reaction mixtures were brought to a final volume of 20 µl by adding Cyscript buffer, DTT, nucleotides, AA-dUTP and Cyscript reverse transcriptase. The mixtures were kept at 42 °C for 2 hours. After the first strand synthesis was completed, the RNA templates were hydrolyzed with 2 µl of 2.5 N NaOH and incubated at 37°C for 15 min. Neutralization was performed by adding 10 µl of 2 M HEPES to each reaction. The unincorporated short oligomers were removed with a filter column from Zymo research kit, according to the manufacturer's protocol (Genetix, UK). The allyl-modified cDNA were then resuspended in 32 µl of 50 mM sodium bicarbonate (pH 9.0). The mixtures were then ready to use to resuspend the lyophilized Cy3 or Cy5 in the coupling step.

3.1.3.2 SuperScript Indirect cDNA labeling System

For the SuperScript Indirect cDNA labeling system, the same concept as the CyScribe Post-Labeling Kit was used, except that the amount of reverse transcriptase enzyme was double and the incubation temperature was raise up to 46 ^{o}C. The other

components were similar to the previous kit. This protocol was recommended by the Invitrogen company.

Coupling amino allyl-modified cDNA with cyanine dye (CyDye) was prepared by the following protocol. The first step started with coupling fluorochrome to the cDNA by resuspending allyl-modified cDNA in lyophilized Cy3 or Cy5 and incubating for 1 hr at room temperature. After that, the reaction were quenched by adding 15 μl of 4 M hydroxylamine and incubatingr 15 min at room temperature. The dye coupled cDNA were then purified with the Zymo research kit. The coupling and quenching steps were done in the dark to prevent dye bleaching.

3.1.4 Hybridization

The hybridizations were performed with both Cy3 and Cy5 labeled cDNAs together. The Cy3 and Cy5 cDNAs were mixed and resuspended in hybridization solution by gently vortexing until the labeled cDNAs were completely resuspended. The mixture were heated at 95 ^{o}C for 30 sec and rapidly cooled on ice. The hybridization solution with the labeled cDNA were then pipetted on to the slides and covered with glass cover slips. The slides were placed in an automatic TECAN station for the hybridizing and automatic washing steps. The hybridization was done at 42 ^{o}C for 16 hours. After washing, the slides were removed and dried with a gentle stream of nitrogen. The slides were placed in a slide box and kept for scanning. The hybridization process followed the protocol from the UCD Array Core facility (www.array.ucdavis.edu).

For the 20K array, the hybridization step was also performed at 46 and 50 ^{o}C to test the annealing temperature and to validate the results. However, for the 45K array only the 42 ^{o}C hybridization was performed.

3.2 Data collection

The microarrays images were generated with a GenePix Pro 4000B scanner (Axon Instrument). The fluorescent scanner was equipped with lasers at the specific excitation wavelengths of 532 and 635 nm. The fluorescent signals from each spot were measured with 532 and 635 nm for two fluorescent dyes and captured by using GenePix Pro software.

3.2.1 Reading data from fluorescent signal

The Axon scanner with confocal laser system was used to scan the slide with the specific excitation wavelength for specific fluorescent dyes. Then, the photomultiplier tube (PMT) converts the resulting fluorescence into electrons which are counted and transformed into a 16 bit value for storage in a tiff image. The confocal imaging system had advantages in high specificity of the lasers, which gave the precise measurements of fluorescence values. However, both PMT sensitivity and laser power are involved in optimization of the dynamic range of the resultant image for each array.

In this study, the axon scanner model 4000B was used to scan the slides. The slide was fully scanned at 10 microns resolution with the laser power at 100. This scanner used dual laser scanning system to scan the slide at two wavelengths simultaneously. These two channels have specific wavelength for laser activated cyanine dyes. The cy3 and cy5 had the excited laser at 532 nm (green) and 635 nm (red), respectively. The emission filters were the 575DF35 (~557-592 nm) and 670DF40 (~650-690 nm) for cy3 and cy5, respectively. In each fluorescent channels, the image were stored in different files to analyze and calculate the relative expression levels of each gene.

3.2.2 Image processing

After scanning, the spot intensity was measured by GenePix software. This software was used to operate the array grid to address and adjust each spot for the foreground and background intensity measurements. The industry-standard formats GenePix Array List (GAL) file was used to identify the spot and oligo ID in separate array blocks. The measurements were reported in GenePix Results (GPR) files which represent each spot intensity for both cy3 and cy5 channels. The intensity values were calculated to average the intensity in each pixel for each spot. The GPR files were used to analyze in further steps, such as normalization and calculated logarithm base two. In addition, image processing was done for spot identification and local background determination. The spot intensity was measured in cy3 and cy5 at different channels. The average background intensity was derived from area

surrounding the spot. The background was subtracted from each spot and the foreground intensities with averaged mean intensity.

3.2.3 Statistical analysis

Following the image processing, the data was first normalized. The normalization is necessary to adjust the variation such as the different labeling, detection efficiency for the fluorescent dye, and quantity of starting mRNA. The average ratio of cy3 to cy5 and intensities were rescaled for minimal variation. In this study, the lowess normalization method was used with the concept of global normalization and linear regression analysis for adjusting the array data.

3.2.4 Significant genes identification

Following the normalization, the data were analyzed to identify genes which were differentially expressed. The post-normalization method normally used the cut off value to distinguish the differentially expressed genes. These values include the two-fold changes in up- and down- regulation, p-value at the highest confidence level, or Log_2 ratio of cy3 and cy5 intensities.

3.3 Data repository

The minimum information about a microarray experiment (MIAME) includes six sections as described in the following paragraphs.

3.3.1 Experimental design:

The hybridization was designed for a direct comparison between light and dark-treated tissues. The samples were prepared for two experimental comparisons: (i) two-week-old rice plant grown in continuous light; and (ii) two-week-old rice plant grown in continuous dark. In each comparison, the samples derived from continuous light and dark tissues were considered as the reference and test sample, respectively. The mRNAs were used for labeling in this experiment. After preparing the labeled samples, four biological replicates were used for hybridization. All mRNA samples were labeled with both Cy3 and Cy5 deoxy UTP. The reversed labeling with different dyes was to avoid the potential of different labeling efficiency and dye bias. The 20K array was designed with two biological and two technical replicates. The 45K array was designed with eight hybridizations which included four biological (rice strains) and four technical replicates (fluoro-flip). The sample paring is shown in Table 3-1 for the 20K array and Table 3-2 for the 45K array.

3.3.2 Array design:

The platform type used in these experiments are glass-slides spotted with 50 – 70mers synthetic oligonucleotides (table 3-3).

Table 3-1 Experimental design for 20K array in the light/dark experiment

Annealing temperatures	**20K array**	
	Cy5	**Cy3**
42 ^{o}C		
slide 1	dark	light
slide 2	light	dark
slide 3	dark	light
slide 4	light	dark
46 ^{o}C		
slide 1	dark	light
slide 2	light	dark
slide 3	dark	light
slide 4	light	dark
50 ^{o}C		
slide 1	dark	light
slide 2	light	dark
slide 3	dark	light
slide 4	light	dark

Table 3-2 Experimental design for 45K array in the light/dark experiment

	45K array	
	Cy5	**Cy3**
slide 1	Dark (Np)	Light (Np)
slide 2	Light (Np)	Dark (Np)
slide 3	Dark (Ki)	Light (Ki)
slide 4	Light (Ki)	Dark (Ki)
slide 5	Dark (TP)	Light (TP)
slide 6	Light (TP)	Dark (TP)
slide 7	Dark (IR)	Light (IR)
slide 8	Light (IR)	Dark (IR)

Note; Np ; nipponbare, Ki ; kitaake, TP ; taipei 309, IR ; IR24

All hybridization were done at 42 ^{o}C

Table 3-3 Array properties for 20K and 45K

Rice array	**20K**	**45K**	
Rice gene model	TIGR V2	TIGR V3	
Unique-oligo probes	20,190	40,098	
Shared-oligo probes	-	3,214	
Length of oligo	50-70 mer	50-70 mer	
Number of slides	1	2	
		A	B
Control elements (hygromycin gene)	217	240	216
Empty spots (without probe)	674	192	2,505

3.3.3 Samples: rice seedling leaves were used in this study.

Total RNA was extracted from tissue by TRIZOL reagent. RT was performed followed by the labeling with Cyscribe-post labeling kit for the 20K array, and SuperScript Indirect cDNA labeling System for the 45K array.

3.3.4 Hybridizations:

This section includes procedures and parameters in this step followed the hybridization's protocol as mention earlier in section 1.4.

3.3.5 Measurements:

The slide was scanned by an Axon scanner and spot intensities quantified with GenePix software. Each image was collected as a 16 bit (TIFF) file, as mentioned earlier. The images were quantified and specified for measurement of the signal intensity.

3.3.6 Normalization controls:

The Lowess normalization method was used in this study for adjusting the array data.

3.4 Array validation by RT-PCR

One method of array validation is quantitative RT-PCR. Thirty-two selected candidate genes were used to confirm the reliability of the array data. The primers for each gene were designed using tool Primer 3.0, with the size set at 20mer. The PCR technique was used to amplify the selected genes with specific primers with the averaged product of 300 base pairs. The quantification of transcripts was based on the expression relative to actin and ubiquitin genes. In this step, the mRNA samples were the same as those previously described.

3.4.1 Reverse Transcriptase-Polymerase Chain Reaction

The reactions were performed following the protocol from Invitrogen. The first strand synthesis reaction comprised of 200 ng of mRNA, 50 ng of oligo(dT), 10 nmole of dNTPs, 2 µl of 10X RT buffer, 4 µl of 25 mM $MgCl_2$, 2 µl of 0.1 M DTT, 40 units of RNaseOUT, and 200 units of SuperScript III.

The procedure started with mixing the mRNA, oligo(dT) primer, and dNTPs in total volume of 10 µl. The mixture were centrifuged briefly and incubated at 56^{o}C for 5 min, then place on ice for at least 1 min. After cooling, 10 µl of master mix of buffer, DTT, $MgCl_2$, RNaseOUT and SuperScript III were added. All reagents were briefly mixed and incubated at 50^{o}C for 2 hours. The reaction was terminated at 85^{o}C for 5 min and chilled on ice.

3.4.2 Polymerase chain reaction with gene specific primers

The reaction mixture from the RT step was used as template for gene specific amplification with gene specific primers. The control genes for quantification were actin and ubiquitin. The PCR reaction for actin and ubiquitin were carried on 23 cycles of denaturing at 94^{o}C for 1 min, annealing at 50^{o}C for 30 sec, and polymerization at 72^{o}C for 30 sec. For the other gene specific primers the PCR reactions were run for 30 cycles of the same condition as the actin and ubiquitin genes. The running cycle for gene specific primers was increased to 30 cycles due to weak signal of the PCR product in gel electrophoresis at 23 cycles. The specific primer sets are listed in Table 3-4.

Table 3-4 The 32 specific primer sets for validating the array result were listed in forward and reverse nucleotide sequence.

Locus ID	Annotation	Forward primer (5'-> 3')	Reverse primer (5'-> 3')
LOC_Os08g33820.1	chlorophyll a/b-binding protein precursor	CCTGAGGTGCTGACGAAGAT	GGTTTCACCAGCCTGAACTC
LOC_Os01g45280.2	Carbonic anhydrase, putative	ACATGGTCCCAGCTTACTGC	GTACTCCACGCGTTCACATC
LOC_Os03g38950.1	expressed protein	CATGAACGGTCTTCCTCCAT	CGAGTGGATCAAACATGTGC
LOC_Os05g41640.3	phosphoglycerate kinase	CCCTGGAAACAACACAGACA	TCTGGGGGAAACAACGTAAG
LOC_Os03g03720.1	glyceraldehyde-3-phosphate dehydrogenase, type I, putative	GTGGCCAACATTATCAGCAA	GACCACCTTGACCATGTCG
LOC_Os10g37180.1	putative glycine decarboxylase subunit	ATGCCCTCAAGATCTCATGC	AGGTCGGAGTGACAAAATGG
LOC_Os01g55570.1	hypothetical protein	TGTCCAGAGTTCGTTCAGCA	GGCGAAAACAAGCTTCTCAC
LOC_Os04g53230.1	glycine cleavage system T protein	GTTTAGGTGCCCGTGACAGT	GCGACAAGAGAGTTGCATGA
LOC_Os08g10020.1	Photosystem II 10 kDa polypeptide PsbR, putative	GAAGCTGAGGATGGCAACAT	CATCACATCGCATTCAGAGG
LOC_Os03g16050.1	fructose-1,6-bisphosphatase	TGTTCTCCAACTGCCTCAAA	CTCCTTGAGGCTGTCCATGT

LOC_Os01g51410.1	glycine dehydrogenase	GGGCAGGTATACATGGATGG	AACGCTTTGCAACATCCTCT
LOC_Os02g01340.1	Oxidoreductase NAD-binding domain, putative	ACACCAACGATCAGGGAGAG	GATCATGATGTCGTCGATGC
LOC_Os03g36750.1	HAD-superfamily hydrolase, subfamily IA, variant 3, putative	CGTCCAGGTGTTCAAAGGTT	TTTCCATAATCCTGGGCAAG
LOC_Os03g61220.1	Transposable element protein, putative	GGGTACCTGACTGGAGCAAA	AAGAGCGAAGGGAACAGTGA
LOC_Os12g08770.1	photosystem i reaction centre subunit n, chloroplast precursor (psi-n)	TCGACGAGTACCTCGAGAAGA	CTTTCCCTTGTTGGTCTGGA
LOC_Os08g44810.1	malate dehydrogenase, NADP-dependent	TTGCCTCTGGTGAGGTTTTT	TTGAGTGGTTCCCCCAAATA
LOC_Os02g01150.1	D-isomer specific 2-hydroxyacid dehydrogenase, NAD binding domain, putative	TTACGGGCAGTTCCTGAAAG	ATTAACAATGCTGGGGCAAG
LOC_Os10g35370.2	putative dehydrogenase	ACATCACCAAGGGCTACGT	TTTTCGATCACCCAAGCTCT
LOC_Os03g51930.1	expressed protein	GTGGGGAACGCTGTAGTTGT	AAATGATTGGGCCATGCTAC
LOC_Os04g55710.1	transposon protein, putative, unclassified	TAGAGCGGCCATTAGCTTGT	GGCAGAAAGAAACAGGAGCA
LOC_Os04g51300.1	Peroxidase, putative	TTTGGGTCTAGGGACAGCAG	CTAATTGGCGCGAACAGTTT

LOC_Os03g52460.1	putative ADP-glucose pyrophosphorylase	GGCTTACCACAGGATGAAGC	CAGCCAGCAGTTCTCCTCTT
LOC_Os02g34810.1	Peroxidase, putative	GCCGAAAAATATGCAGAGGA	AGGAGGTCATCAGACCATCG
LOC_Os08g44320.1	Low molecular weight phosphotyrosine protein phosphatase, putative	CCGTCCTCTTCGTCTGTCTC	AAGCAGAAACGCTTGCATTC
LOC_Os08g02210.2	expressed protein	CAAAATTCCGACGCAAAACT	ATTTTCCCAGGCAAGGAGAT
LOC_Os08g14440.3	expressed protein	ACATCAACATCACCGTCCAA	CCAGAACAGTTGATGCATGG
LOC_Os07g08970.1	expressed protein	GGAACATCATTTTGCCAAGG	GCTTCCTCTTTGGAGCATCA
LOC_Os01g73500.1	expressed protein	GCCGCATCTCCTCAACATA	TACGGGTAGCGAGATTGTCC
LOC_Os12g17910.1	rubisco subunit binding-protein alpha subunit, chloroplast precursor (60 Kda chaperonin alpha subunit) (cpn-60 alpha) (fragment)	GAAGAGCAACTTGGGACAGC	TTGTACCCGACTTCCCACTC
LOC_Os08g43560.1	Peroxidase, putative	TGACAAGGCATTGTTGGAAG	TCGAAGGACACAAACCACTG
LOC_Os10g40030.1	putative WW-domain oxidoreductase	TTGCCTCCGAGTTTGACTCT	GCCACATAACATGTCGTTGC
LOC_Os03g48040.1	putative ferredoxin	AGAGGTGTCGTCCACGAGTT	TGCCAAACAATCGCTAACAG

Os03g0234200	ubiquitin1	CACGGTTCAACAACATCCAG	TGAAGACCCTGACTGGGAAG
OSIGCRA124K06	Actin1	ACGGCGATAACAGCTCCTCTT	CCTCTTCCAGCCTTCCTTCAT

CHAPTER IV

RESULTS

4.1 Light/dark 20K array

4.1.1 Statistical analysis

4.1.1.1 Average mean intensity

Analysis of microarray data was done with R software using the LMGene package. The basic statistical analysis was performed by observing the mean intensity of red and green dyes. The 20K arrays were tested with different annealing temperatures (42, 46, and 50 ^{o}C) to find the optimum temperature which obtained the highest signal intensity and reliable results. The array data for each annealing temperature came from 4 slides, except for the 46 ^{o}C experiment, in which data for slide number 4 was lost during data upload into the R software. Therefore, further analysis displayed only three slides for 46 ^{o}C.

The average intensity among three different annealing temperatures was calculated to compare the mean intensity of red and green dye. Table 4-1 represents the results for average mean for red and green dye at 42, 46 and 50 ^{o}C. The mean intensities of red and green in the 42 ^{o}C experiment were 451.63 and

455.18, respectively. The average mean intensity at 42 ^{o}C was 453.39, which was the highest value when compare to the other two temperatures. Among three annealing temperatures, the 42 ^{o}C obtained the highest signal intensity and was chosen for the annealing temperature in the next experiment. In addition, the standard deviation for three array data sets was not calculated because of the wide range of signal intensities in both cy3 and cy5.

4.1.1.2 Control elements

The 20K array also contains control elements of the hygromycin resistance gene and empty spots. These elements were used as positive and negative controls. When transgenic rice samples were tested, the hygromycin gene was used as a positive control. In contrast, when wild type rice was used, the hygromycin spots were used as the negative control.

The ratio of red and green intensity were checked from the three different annealing temperatures to achieve the error rate over 2 fold and 4 fold changes (Table 4-2). The control elements on the 20K array were composed of 217 hygromycin and 674 empty spots. The $\log_2$ ratio of red and green intensities were calculated and reported in error percentages. At 42 ^{o}C, the average percentage error was 0.69 for empty spots and 0.037 for hygromycin spots. This error was also found at 46 ^{o}C, which showed an average percentage error of 0.30 for empty spots but no error was found in hygromycin spots. The calculation was also done for 50 ^{o}C, however, no

error was shown. This result agrees with the theory that higher annealing temperature could reduce the background and cross hybridization.

Table 4-1 The basic statistical analysis of average means intensity for red and green colors.

	Mean	**Rmean**	**Gmean**	**Gap|R-G|**
42 ^{o}C				
Slide1	565.67	563.11	568.25	5.14
Slide2	467.12	402.34	531.91	129.56
Slide3	205.05	196.64	213.46	16.82
Slide4	575.72	644.45	507.10	137.24
average	**453.39**	**451.63**	**455.18**	**74.69**
46 ^{o}C				
Slide1	397.99	373.60	422.40	48.80
Slide2	456.75	464.30	449.18	15.12
Slide3	465.92	503.96	427.88	76.07
average	**440.22**	**447.28**	**433.15**	**46.66**
50 ^{o}C				
Slide1	357.15	334.05	380.26	46.20
Slide2	399.64	431.82	367.46	64.35
Slide3	406.53	371.41	441.64	70.22
Slide4	401.76	433.69	369.84	63.85
average	**391.27**	**392.74**	**389.80**	**61.15**

Note; Rmean is referred to mean intensity of red color

Gmean is referred to mean intensity of green color

Gap|R-G| is referred to the different intensity value between red and green colors

Table 4-2 The control elements on the 20K array; empty and hygromycin spots for 2- and 4-fold changes with three different annealing temperatures.

	Empty spot		**Hygromycin spot**	
	Error > 2 fold	**Error > 4 fold**	**Error > 2 fold**	**Error > 4 fold**
42 ^{o}C				
Slide1	0	0	0	0
Slide2	1	0	0	0
Slide3	5	0	0	0
Slide4	0	0	1	0
Percentage error (%)	**0.69**	**0**	**0.037**	**0**
46 ^{o}C				
Slide1	1	0	0	0
Slide2	1	0	0	0
Slide3	0	0	0	0
Percentage error (%)	**0.30**	**0**	**0**	**0**
50 ^{o}C				
Slide1	0	0	0	0
Slide2	0	0	0	0
Slide3	0	0	0	0
Slide4	0	0	0	0
Percentage error (%)	**0**	**0**	**0**	**0**

4.1.1.3 MM plot

The MM plot was used as one of the analytical methods to visualize the data. This plot was generated from the $\log_2$ ratio of cy3 and cy5 intensity. The normalization method in this study was computed in R software with locally weighted linear regression (loess) method.

This experiment was composed of 4 slides, which were designed to give biological replication (slides 1, and 3) and technical replication (slides 2, and 4). Slides 1 and 2 presented the pair of technical replication (dye-swap). The MM plots were executed on array data of all three tested annealing temperature (Table 4-3). Figures 4-1 to 4-5, represent the MM plots for the array data derived from 42, 46, and 50 ^{o}C annealing temperatures before and after normalization. All array data sets showed similar patterns of spot distribution. The correlation coefficient reached 0.8 for all data sets which explained the high correlation on the technical replication (Table 4-3). At 42 ^{o}C, array data showed 0.88 for the average correlation coefficient, which was higher than the other two annealing temperatures.

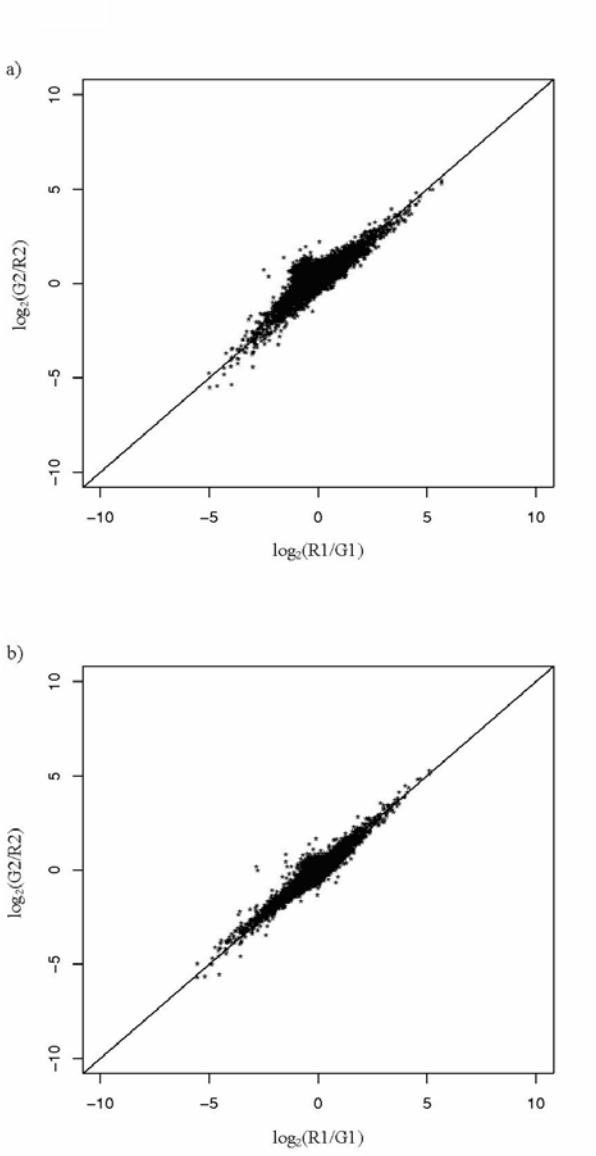


Figure 4-1 MM plot of the signal intensities obtained from the 42 ^{o}C annealing temperature array data (slides 1 and 2). The x-axis represents the ratio of cy3 (light) over cy5 (dark) intensity (slide 1). The y-axis represents the ratio of cy3 (dark) over cy5 (light) intensity (slide 2). (a) The array data transformed to the $\log_2$(ratio) before normalization, (b) The array data were normalized and then transformed to the $\log_2$(ratio).

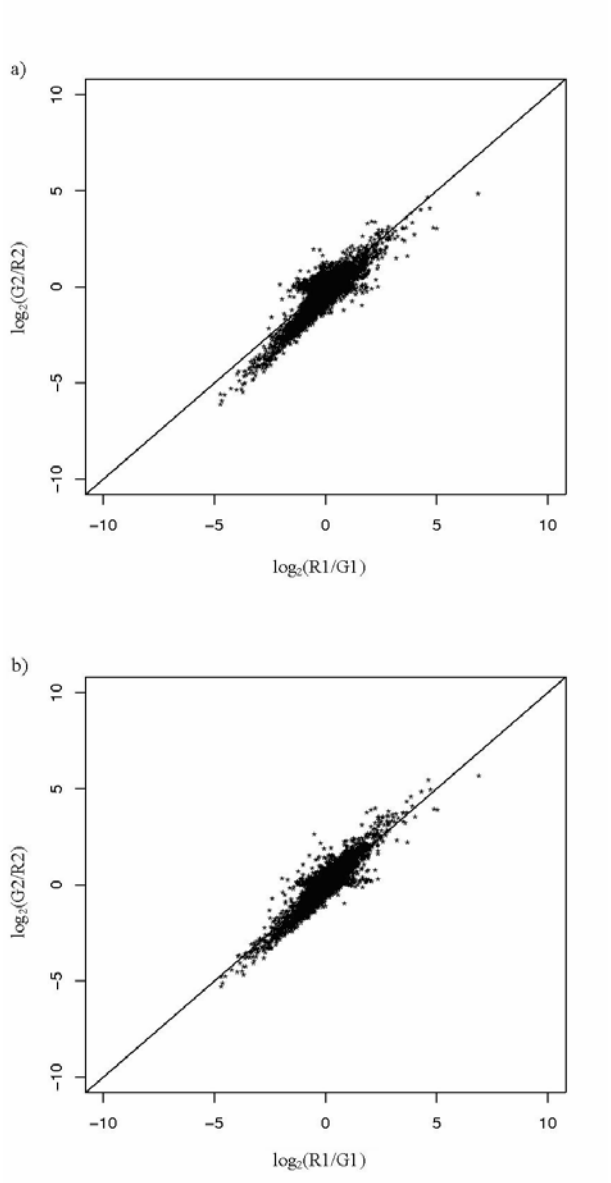


Figure 4-2 MM plot of the signal intensities obtained from the 42 ^{o}C annealing temperature array data (slides 3 and 4). The x-axis represents the ratio of cy3 (light) over cy5 (dark) intensity (slide 3). The y-axis represents the ratio of cy3 (dark) over cy5 (light) intensity (slide 4). (a) The array data transformed to the $\log_2$(ratio) before normalization, (b) The array data were normalized and then transformed to the $\log_2$(ratio).

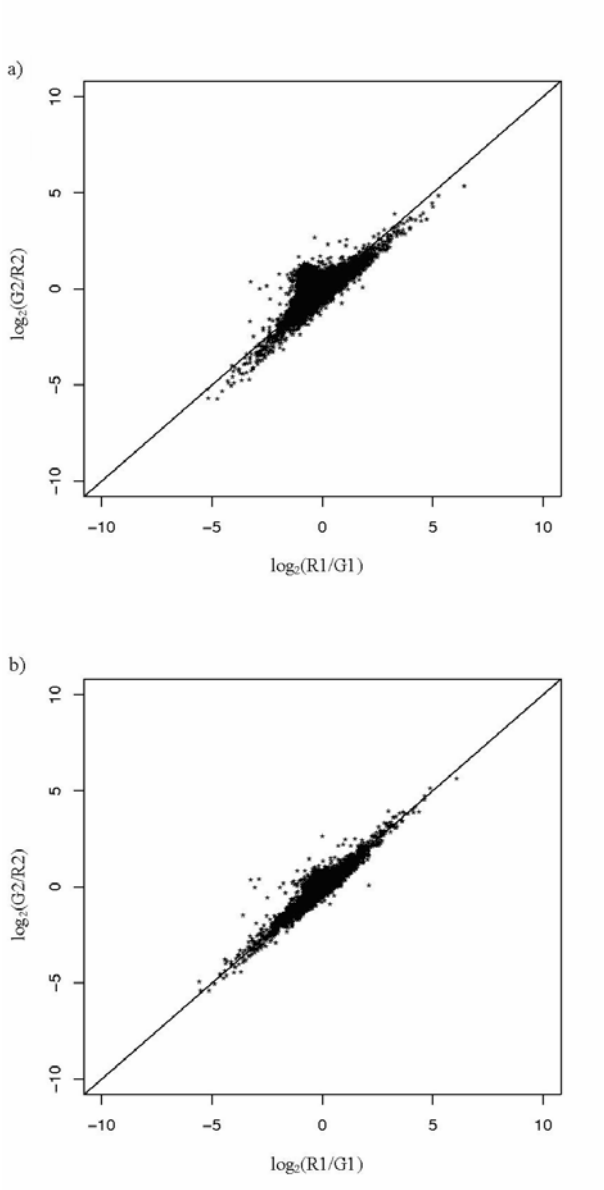


Figure 4-3 MM plot of the signal intensities obtained from the 46 ^{o}C annealing temperature array data (slides 1 and 2). The x-axis represents the ratio of cy3 (light) over cy5 (dark) intensity (slide 1). The y-axis represents the ratio of cy3 (dark) over cy5 (light) intensity (slide 2). (a) The array data transformed to the $\log_2$(ratio) before normalization, (b) The array data were normalized and then transformed to the $\log_2$(ratio).

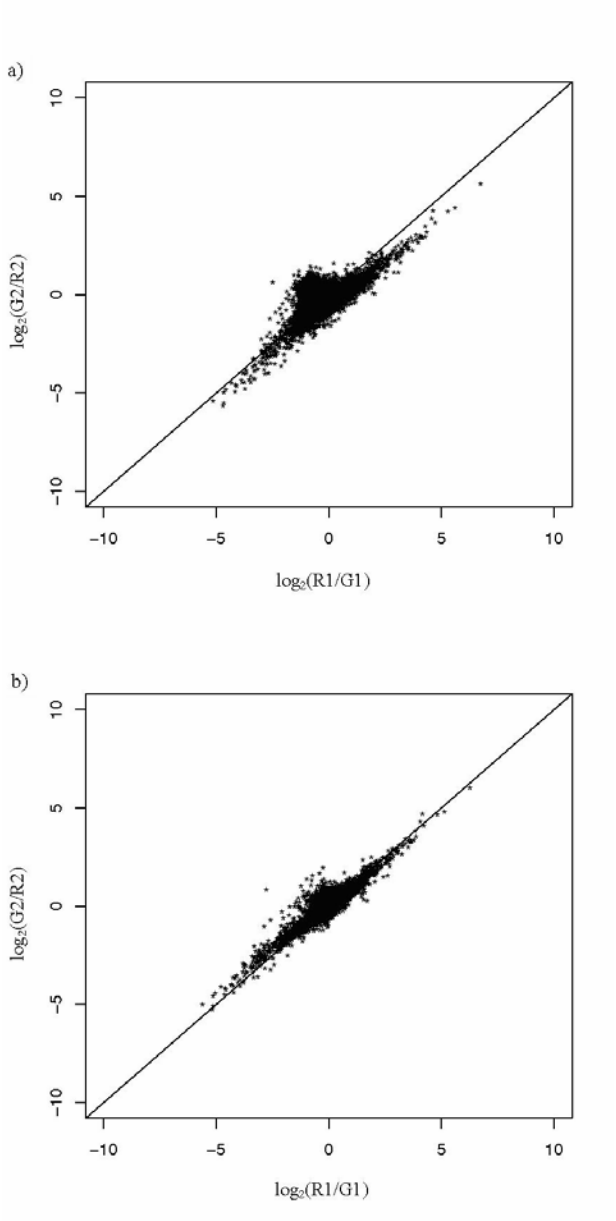


Figure 4-4 MM plot of the signal intensities obtained from the 50 ^{o}C annealing temperature array data (slides 1 and 2). The x-axis represents the ratio of cy3 (light) over cy5 (dark) intensity (slide 1). The y-axis represents the ratio of cy3 (dark) over cy5 (light) intensity (slide 2). (a) The array data transformed to the $\log_2$(ratio) before normalization, (b) The array data were normalized and then transformed to the $\log_2$(ratio).

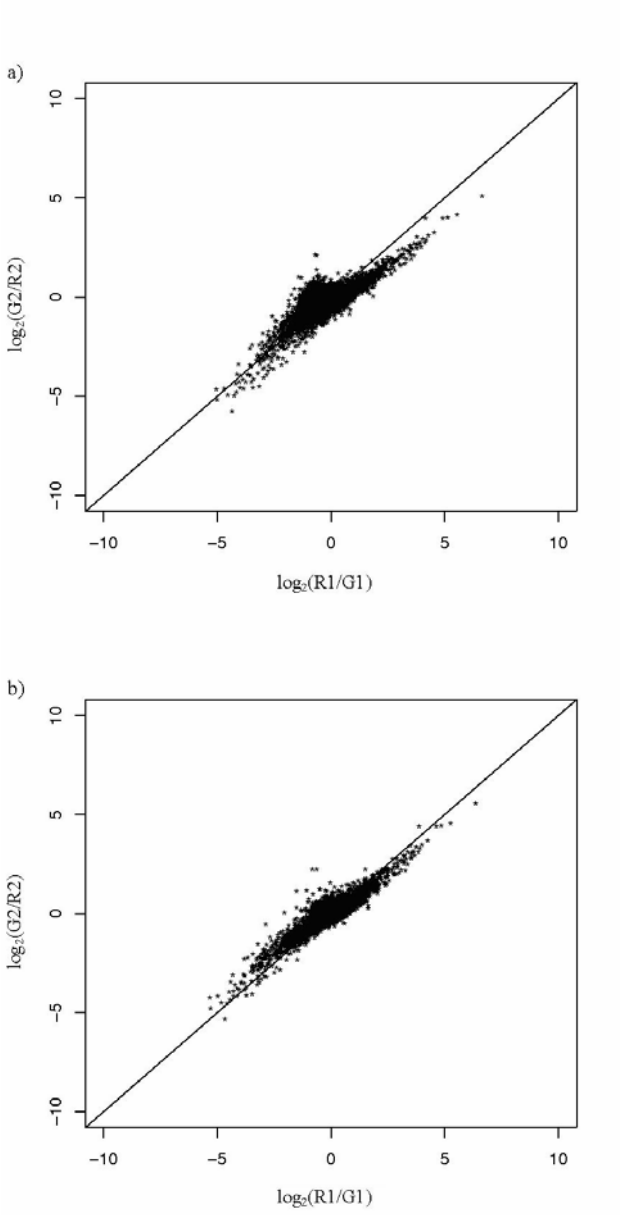


Figure 4-5 MM plot of the signal intensities obtained from the 50 ^{o}C annealing temperature array data (slides 3 and 4). The x-axis represents the ratio of cy3 (light) over cy5 (dark) intensity (slide 3). The y-axis represents the ratio of cy3 (dark) over cy5 (light) intensity (slide 4). (a) The array data transformed to the $\log_2$(ratio) before normalization, (b) The array data were normalized and then transformed to the $\log_2$(ratio).

Table 4-3 The MM plot of array data between light and dark samples

Annealing temperature (^{o}C)	Comparison		Correlation coefficient
	X	Y	
42	Slide 1	Slide 2	0.92
	Slide 3	Slide 4	0.82
	Slide 1	Slide 4	0.92
	Slide 2	Slide 3	0.88
average			**0.88**
46	Slide 1	Slide 2	0.88
	Slide 2	Slide 3	0.80
average			**0.84**
50	Slide 1	Slide 2	0.85
	Slide 3	Slide 4	0.83
	Slide 1	Slide 4	0.81
	Slide 2	Slide 3	0.87
average			**0.84**

Note; X referred to x-axis, Y-referred to y-axis

4.1.1.4 MA plot

MA plot is an alternative type of scatter plot that was constructed by combining the array data. The $\log_2$ ratio (M) are plot on the y-axis against the log of the geometric mean of the signal strengths (A) for each spot on the slide. M = $\log_2$(cy5/cy3) represents a normalized $\log_2$ ratio of the two dyes. The positive "M" values indicated higher normalized signal intensity in the leaf RNA sample which was treated with the dark condition. The negative M values indicate the higher signal

intensity in the leaf RNA sample which was treated with the light condition, and M values of zero indicated equal intensity in the two samples. A = $[\log_2(cy5)+\log_2(cy3)]/2$ is the average logarithmic signal intensity. The higher "A" values indicate brighter signal, lower "A" values indicate dimmer signals. In addition, M and A were calculated for each element on each array.

The MA plot represents the grey spots of normalized data and black spots of raw data. Figure 4-6, the 42^{o}C annealing temperature array data of all 4 slides showed the MA plot with most of the spots distributed in range of M from -2 to 2. The 46 ^{o}C and 50 ^{o}C annealing temperature array data were plotted and displayed in Figures 4-7 and 4-8, respectively. The spot distribution showed similar patterns to the 42 ^{o}C array data. For the "A" value, array data from the 42, 46, and 50 ^{o}C annealing temperatures showed a small distribution at low signal intensity area (A<8.0) in all slides. However, some of the array data with annealing at 42 ^{o}C (slides 3 and 4) and 50 ^{o}C (slide 4) showed the different patterns of spot distribution at low intensity when compared to other slides. These data showed spot distributions that overlapped the x-axis or had M values equal to zero. These patterns, referred to as Rmean and Gmean intensities were close before normalization. The normalized and unnormalized patterns were similar in these data sets.

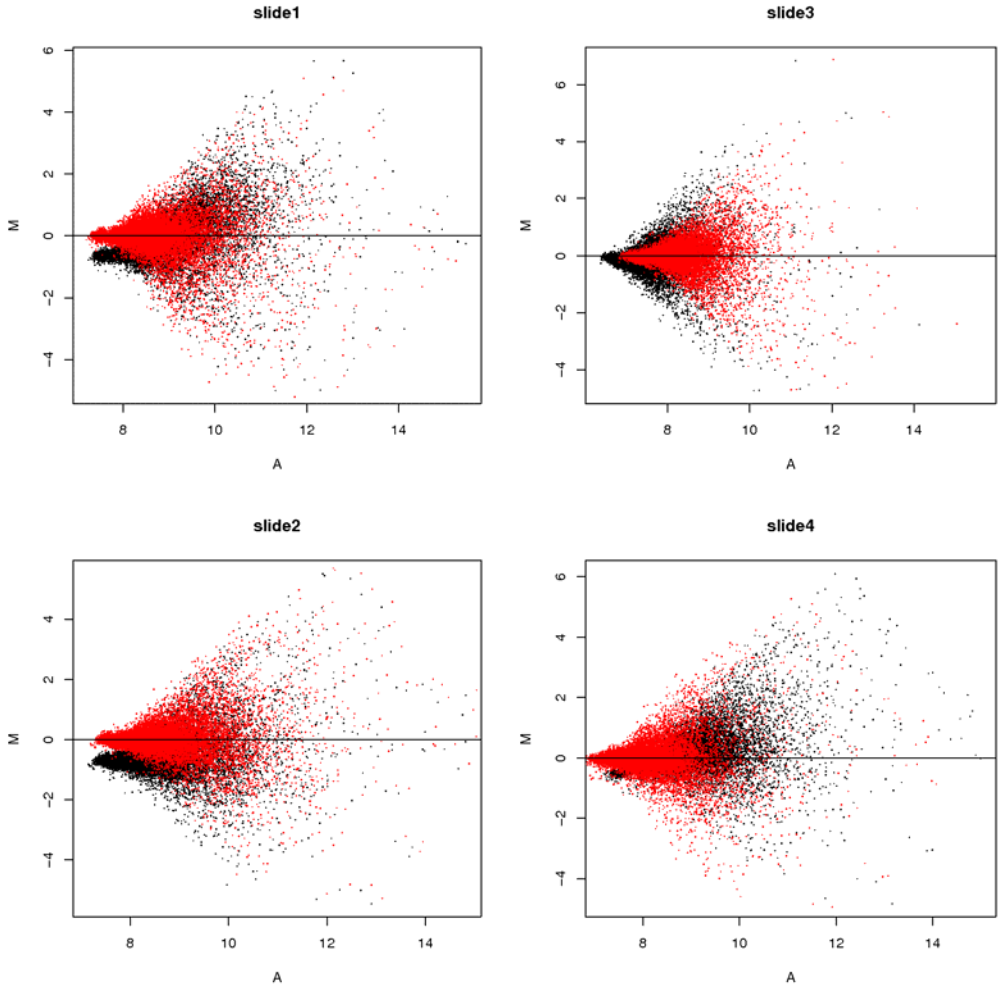


Figure 4-6 MA-plot demonstrating the normalized array data set with annealing at 42 ^{o}C.

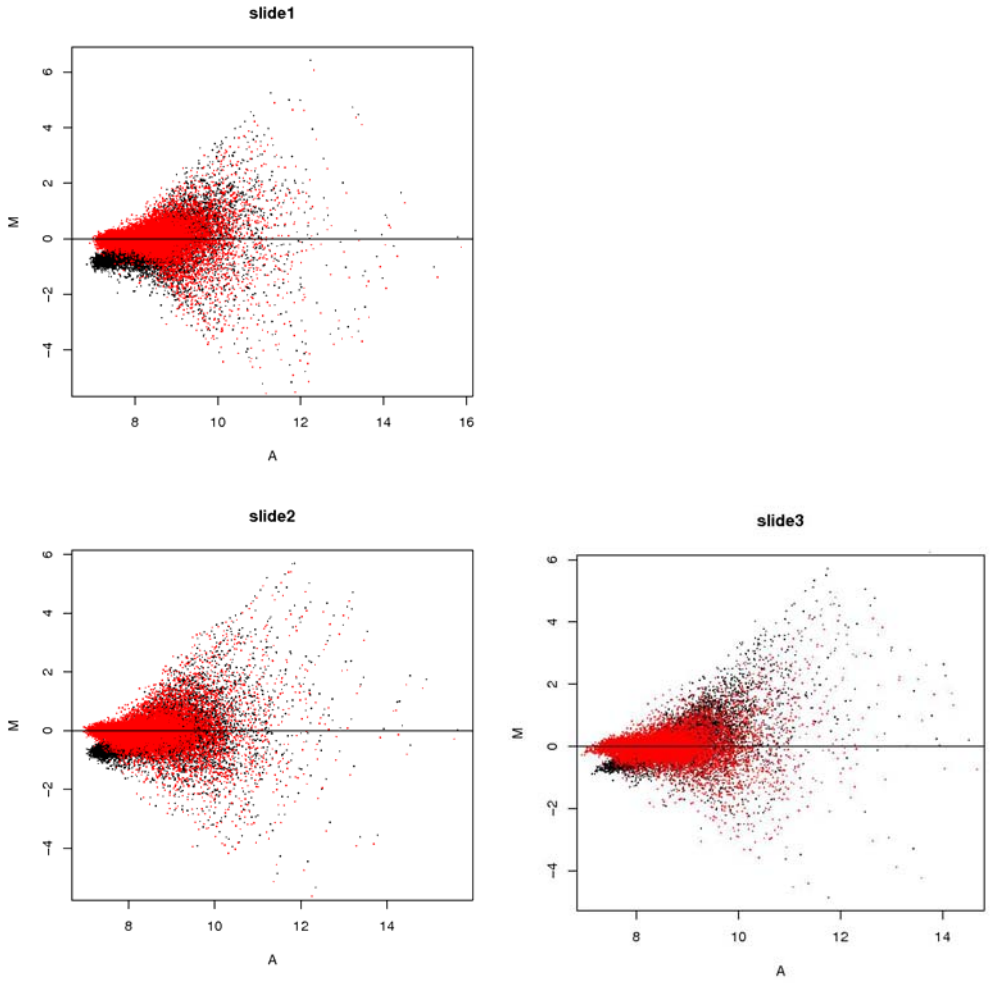


Figure 4-7 MA-plot demonstrating the normalized array data set with annealing at 46 ^{o}C.

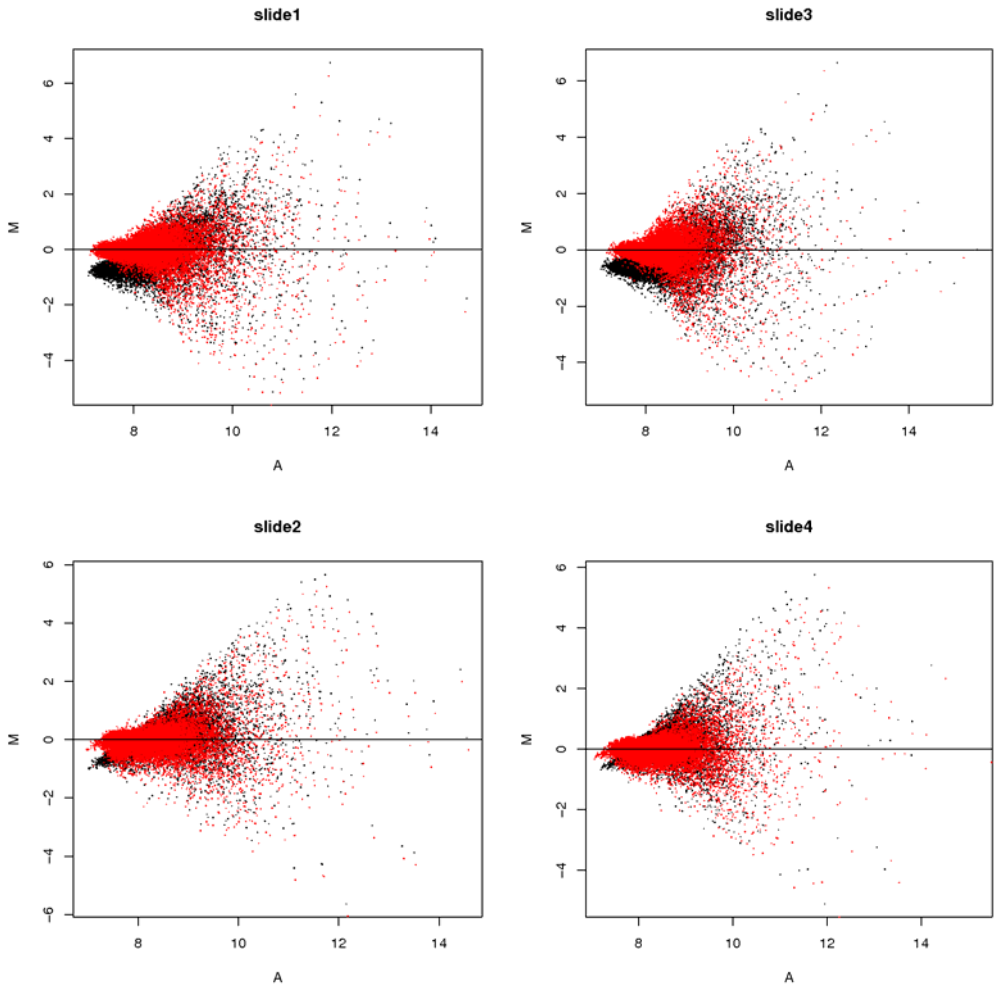


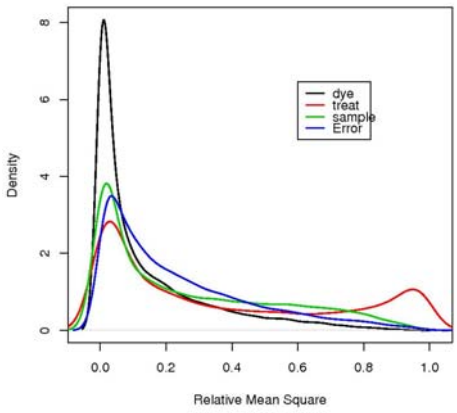
Figure 4-8 MA-plot demonstrating the normalized array data set with annealing at 50 ^{o}C.

4.1.1.5 Smoothed histogram analysis of sources of variation

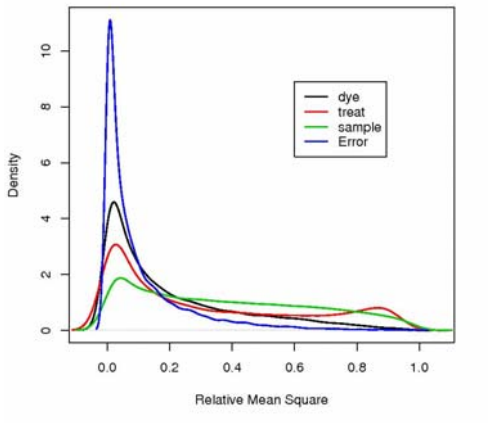
The array data was also analyzed by applying statistical methods to explore the variable factors. The systematic error had been classified as dye, slide, treatment, sample, and error factors. The ANOVA was used for each gene to find the relative mean square and a graph was plotted as smoothed histogram. The histogram illustrates the properties of distribution in a large data set. The density or frequency of data displayed on the y-axis was plotted against the relative mean square of the intensity in the array data set on the x-axis. The sources of variation were plotted to compare and determine variable factors which give the highest effect on the array data set.

Figure 4-9 represents the smoothed histogram of the tested annealing temperatures of 42, 46, and 50 ^{o}C. The variable factors, sample, dye, treatment, and error were observed. The effects of variable factors were monitored from the high relative mean squares and high density of array data. In all three annealing temperatures, the treatment factor showed the highest effect on the array data.

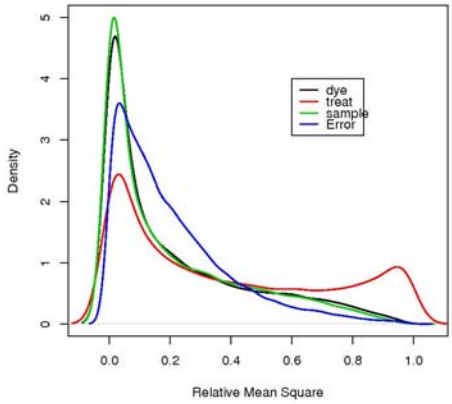
To analyze the smoothed histogram results, only the values of relative mean square higher than 0.6 are worth considering (or may be significant). The relative mean square values lower than 0.6 are not significant.



(a)



(b)



(c)

Figure 4-9 The smoothed histograms representing the variable factors on array results at three different annealing temperatures. a) 42 ^{o}C; b) 46 ^{o}C; c) 50 ^{o}C The dye factor represents in black color, treatment factor represents in red color, sample factor represents in green color, and error factor represents in blue color.

4.1.1.6 Significantly differentially expressed gene identification

To find the significantly differentially expressed genes, the FDR (p-value) cut offs of 0.01 and 0.05 were used. The array data confidence depends on the cut off value, which is critical for further analysis. The significant gene lists were compared between the three different annealing temperatures to determine the reliability of the array results. Table 4-4 represents the number of significant genes on the 42 ^{o}C and 50 ^{o}C annealing temperature lists. The FDR cut off (p-value) at 0.01, gave 3,089 and 2,530 differentially expressed genes at 42 ^{o}C and 50 ^{o}C, respectively. The FDR (p-value) at 0.05 gave the number of genes list up to 4,260 and 3,677 at 42 ^{o}C and 50 ^{o}C, respectively. In addition, when only changes over two-fold were considered, the list of differentially expressed genes was down to 1,205 for the 42 ^{o}C array data. When the cut-off was increased up to four-fold changes, only 307 differentially expressed genes were found for the 42 ^{o}C array data. The number of genes on the list with 50 ^{o}C annealing temperature was lower than 42 ^{o}C, with 2,530 genes at FDR (p-value) 0.01 and 3,677 genes at FDR 0.05. The candidate number at 50 ^{o}C was also lower than 42 ^{o}C when the cut-offs by fold change at over two- and four-fold changes were used. The higher annealing temperature can reduce the background and increase the specificity of the oligonucleotide probe to the target labeled gene, therefore, lower numbers of significant genes were found. The matched candidate genes were counted between 42 ^{o}C and 50 ^{o}C. The numbers matched at the cut-off values of 0.01 and 0.05 were 2,184 genes and 3,132 genes, respectively. The numbers matched was also counted in the list genes with greater than two- and four-fold changes. After matching the candidate genes of 50 ^{o}C array to the 42 ^{o}C array, the matched number is

close to the candidate gene number for 50°C within the two- and four-fold changes groups. This number implied that the cut-off value affected on candidate gene number and annealing temperature reduced the redundant candidate gene number as well.

Table 4-4 The significantly differentially expressed gene number of the 42 °C and 50 °C annealing array results. The FDR (p-value) cut-off values of 0.01 and 0.05 and fold change cut-off values were applied to generate the gene list.

Number of candidate gene (Annealing temperature)	**FDR (p-value)**		**Fold change**	
	≤ 0.01	**≤ 0.05**	**≥ 2**	**≥ 4**
42 °C	3,089	4,260	1,205	307
50 °C	2,530	3,677	940	220
Matched number	**2,184**	**3,132**	**911**	**215**

4.1.2 Biological analysis

4.1.2.1 Gene ontology

Gene ontology (GO) has been categorized into three main classes of cellular component, biological process and molecular function (Ashburner et al., 2000). Each class is composed of several subclasses. The classification of a gene product might be associated with, or located in the one or more cellular components. The GO database can aid the interpretation of the candidate gene list or annotation.

4.1.2.1.1 Gene ontology of 20K array

GO was applied on the microarray data to annotate the functions and gene products. GO was also used as an analysis tool to categorized the oligonucleotide probes of the 20K array. The analysis manipulated the array data by matching the oligonucleotide probes with the GO database and categorized them into subclasses. The candidate genes were also compared to the GO database and the matched candidate genes in each GO subclass counted. From Table 4-5, the 20K array contained only 7,967 (38%) oligo probes that corresponded to the GO database. The molecular function was categorized for 7,227 oligos (34%). The biological process and cellular component were identified for 5,630 oligos (27%) and 3,015 oligos (14%), respectively. This result correlated with the GO classification that one gene could be active in one or more biological processes, or perform one or more biological functions. The GO subclasses are represented in Tables 4-6 to 4-8, that show the corresponding GO subclass for each GO class. A high proportion of candidate genes could be placed in a biological process class, followed by molecular function and cellular component classes, respectively.

One oligonucleotide probe corresponded to several GO subclasses which contained a large number of oligonucleotide probes in each subclasses. The molecular function contained 7,227 oligonucleotide probes, which were classified into many subclasses, each containing a number of oligonuclotide probes. The percentage of classified oligonucleotides in the GO subclasses are represented in Tables 4-6 to 4-8 for molecular function, cellular component, and biological process, respectively.

"All GO" terms refer to total number of classified oligonucleotide probes within the GO database.

Table 4-5 The GO classification of oligonucleotide probes found on the 20K array

GO class	oligo matched GO (number)	% oligo on array
Biological process	5,630	26.65
Cellular component	3,015	14.27
Molecular function	7,227	34.21
All GO-terms	7,967	37.72
micro-array	21,120	

Table 4-6 Oligo numbers in molecular function categories on the 20K array

Molecular function subclass	Total	Percentage (%)
transferase activity	1,888	12.57
hydrolase activity	1,563	10.41
protein binding	1,552	10.33
catalytic activity	1,507	10.03
nucleotide binding	1,413	9.41
kinase activity	1,282	8.53
binding	891	5.93
DNA binding	755	5.02
transporter activity	721	4.80
carbohydrate binding	635	4.23
transcription factor activity	508	3.38
nucleic acid binding	337	2.24
RNA binding	272	1.81
transcription regulator activity	250	1.66
signal transducer activity	233	1.55
receptor activity	223	1.48
oxygen binding	180	1.19
structural molecule activity	161	1.07
chaperone activity	109	0.72
lipid binding	109	0.72
enzyme regulator activity	93	0.61
translation factor activity, nucleic acid binding	79	0.53
nuclease activity	72	0.48
motor activity	64	0.42
All GO	15,014	

Table 4-7 Oligo numbers in cellular component categories on the 20K array

Cellular component subclass	Total	Percentage (%)
membrane	1,523	32.12
plasma membrane	623	13.14
nucleus	413	8.71
plastid	350	7.38
cell wall	297	6.26
endoplasmic reticulum	228	4.80
mitochondrion	158	3.33
intracellular	143	3.01
cytosol	138	2.91
cell	135	2.84
thylakoid	120	2.53
ribosome	119	2.51
cytoplasm	114	2.40
extracellular region	85	1.79
unlocalized protein complex	55	1.16
nuclear membrane	54	1.13
cytoskeleton	38	0.80
peroxisome	30	0.63
nucleolus	26	0.54
nucleoplasm	26	0.54
Golgi apparatus	23	0.48
extracellular matrix (sensu Metazoa)	14	0.29
All GO	4,741	

Table 4-8 Oligo numbers in biological process categories on the 20K array

Biological process subclasses	Total	Percentage (%)
response to biotic stimulus	1,346	8.48
signal transduction	1,103	6.95
protein modification	1,095	6.90
response to endogenous stimulus	1,084	6.83
response to stress	948	5.97
biosynthesis	804	5.06
response to abiotic stimulus	670	4.22
transport	640	4.03
response to external stimulus	587	3.69
lipid metabolism	575	3.62
protein metabolism	566	3.56
catabolism	546	3.44
amino acid and derivative metabolism	521	3.28
metabolism	483	3.04
transcription	480	3.02
secondary metabolism	381	2.40
cell differentiation	359	2.26

Table 4-8 (Continued)

Biological process subclasses	Total	Percentage (%)
carbohydrate metabolism	342	2.15
electron transport	300	1.89
growth	286	1.80
protein biosynthesis	279	1.75
development	264	1.66
cell organization and biogenesis	232	1.46
morphogenesis	204	1.28
behavior	187	1.17
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	184	1.16
flower development	183	1.15
DNA metabolism	142	0.89
death	107	0.67
reproduction	106	0.67
cell death	105	0.66
cell cycle	103	0.65
physiological process	82	0.52
cell growth	81	0.51
post-embryonic development	70	0.44
response to extracellular stimulus	62	0.39
cell growth and/or maintenance	61	0.39
generation of precursor metabolites and energy	59	0.37
cellular process	50	0.31
pollination	42	0.26
cell homeostasis	27	0.17
ripening	22	0.14
photosynthesis	19	0.12
embryonic development	12	0.08
regulation of gene expression, epigenetic	11	0.07
All GO	15,868	

4.1.2.1.2 Gene ontology of significantly expressed genes

The GO analysis was done on the significant gene lists for 42 ^{o}C and 50 ^{o}C annealing temperatures with the cut-off values at 0.001 and 0.05 FDR. The numbers were counted for calculating the relative ratios between GO classes and total candidate genes.

Table 4-9 represents the candidate gene numbers for 42 and 50 ^{o}C annealing temperature array data. It was found that 1,868 candidate genes were counted for the 0.001 FDR cut-off and 4,260 were counted for the 0.05 FDR cut-off value. The 50 ^{o}C showed 1,461 and 3,677 candidate genes at 0.001 and 0.05 cut-off values, respectively. The relative ratio represented as the percentages by matching GO over the total number of candidate genes was shown in Table 4-10. The biological process and cellular component showed similar percentages at 42 ^{o}C and 50 ^{o}C annealing temperatures. The FDR 0.001 and 0.05 demonstrated the same percentage for 42 ^{o}C and 50 ^{o}C. The highest percentage was found in molecular functions in both 42 ^{o}C and 50 ^{o}C. The array result at 42 ^{o}C and 50 ^{o}C has the correlation on the candidate gene list.

Table 4-9 The GO classification of candidate genes at FDR cut-off values of 0.001 and 0.05.

Annealing temperature (^{o}C)/FDR	**Number of match candidate gene**			**Total number of candidate gene**
	Biological process	**Cellular component**	**Molecular function**	
42 /(0.001)	696	383	926	1,868
42 /(0.05)	1,500	832	2,044	4,260
50 /(0.001)	546	300	738	1,461
50 /(0.05)	1,320	722	1,763	3,677

Table 4-10 The relative percentage of candidate genes on three GO classes.

Annealing temperature (^{o}C)/FDR	Relative percentage (%)		
	Biological process	Cellular component	Molecular function
42 /(0.001)	37	20	49
42 /(0.05)	35	19	47
50 /(0.001)	37	20	50
50 /(0.05)	35	19	47

For the 20K array light/dark experiment, the optimum annealing temperature at 42 ^{o}C was chosen based on statistical and GO analysis. Further results represented only array data for annealing at 42 ^{o}C. The result showed the cellular component matching genes comprised several subclasses, which corresponded to significantly expressed genes. The cellular component class represented over 30% in the ribosome, peroxisome, and thylakoid subclasses. The mitochondrion and plastid subclasses also showed over 20%. The cellular component subclasses is demonstrated in Table 4-11, these subclasses belong to the important component in photosynthesis that relates to array results affected in light/dark treatment. The result clearly illustrated that in the biological processes, the photosynthesis subclass showed the highest percent of affected genes (68%). The protein synthesis, generated precursor in metabolism or nucleic acid metabolism subclasses displayed around 20% of genes changing expression levels in light/dark treatment. The molecular function, categorized as structural molecule activity and RNA binding subclasses also showed over 20% on this treatment.

The oligonucleotide probes for the photosynthesis subclass on the 20K array were identified and shown in Table 4-12. The photosynthesis subclass classified by GO term showed 19 oligos. Only 13, and 14 oligos matched significantly expressed genes with FDR cut-off value at 0.001 and 0.05, respectively. In addition, nine more oligos that were not identified by the GO term, but were identified while looking at the significant differential expression. All of these 9 oligos showed significant differential expression at FDR less than 0.05. The classified oligonucleotide probes for photosynthesis are composed of the essential elements for photosynthesis as chlorophyll *a/b* binding protein and ferredoxin (table 4-13). Ten oligo probes denoted chlorophyll *a/b* binding protein were represented on the 20K array. All ten oligos showed significant differential expression with a 0.01 FDR cut off. Fourteen ferredoxin oligos were identified, but only 9 oligos showed significant differential expression.

Table 4-11 GO categories on significant gene list for 42 ^{o}C by cut off at ≤ 0.1% FDR

GO categories	Total oligo (20K)	Significant gene (≤ 0.1%FDR)	Percentage (%)
cellular component subclass			
ribosome	119	47	39.49
peroxisome	30	10	33.33
thylakoid	120	37	30.83
plastid	350	97	27.71
mitochondrion	158	39	24.68
Biological process subclass			
photosynthesis	19	13	68.42
protein biosynthesis	279	75	26.88
generation of precursor metabolites and energy	59	15	25.42
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	184	36	19.56
Molecular function subclass			
structural molecule activity	161	52	32.29
RNA binding	272	69	25.36

Table 4-12 Biological process subclass photosynthesis oligonucleotide probes on the 20K array. The FDR and Log_2Ratio results from the 42 ^oC experiment.

Locus ID	Photosynthesis (biological process)	FDR	Log_2Ratio
GO classification			
LOC_Os01g05490.1	triosephosphate isomerase	2.36E-02	0.71
LOC_Os01g41710.1	Chlorophyll A-B binding protein, putative	1.47E-06	4.85
LOC_Os01g52240.1	Chlorophyll A-B binding protein, putative	7.65E-05	1.66
LOC_Os01g62420.4	triosephosphate isomerase	1.47E-01	0.70
LOC_Os01g71190.1	Photosystem II reaction centre W protein, PsbW, putative	1.22E-04	1.78
LOC_Os02g52650.1	Chlorophyll A-B binding protein, putative	7.31E-06	2.18
LOC_Os03g45710.1	putative ferredoxin	2.24E-01	0.18
LOC_Os03g48040.1	putative ferredoxin	5.19E-05	1.83
LOC_Os03g56670.1	photosystem-1 F subunit precursor	4.84E-07	2.47
LOC_Os03g61960.1	putative ferredoxin	3.56E-01	-0.18
LOC_Os04g33630.1	ferredoxin [2Fe-2S], putative	9.91E-01	0.02
LOC_Os05g37140.1	ferredoxin [2Fe-2S], root - rice	7.73E-01	0.07
LOC_Os07g05360.1	Photosystem II 10 kDa polypeptide PsbR, putative	4.58E-05	-1.66
LOC_Os07g05480.1	Photosystem I psaG / psaK, putative	1.52E-07	4.17
LOC_Os07g37550.1	Chlorophyll A-B binding protein, putative	4.62E-07	4.50
LOC_Os07g38960.1	Chlorophyll A-B binding protein, putative	6.23E-07	3.69
LOC_Os08g10020.1	Photosystem II 10 kDa polypeptide PsbR, putative	4.47E-07	3.51
LOC_Os08g33820.1	chlorophyll a/b-binding protein precursor	8.60E-08	5.33
LOC_Os09g17740.1	chlorophyll a/b binding protein 1.	9.71E-07	5.17
Additional oligo probes (20K array)	*Not in the GO classification of photosynthesis gene*		
LOC_Os01g31690.1	Manganese-stabilizing protein / photosystem II polypeptide	9.81E-04	1.21

Table 4-12 (Continued)

Locus ID	Photosynthesis (biological process)	FDR	Log_2Ratio
LOC_Os07g25430.1	Photosystem I reaction centre subunit IV / PsaE, putative	5.83E-07	2.87
LOC_Os12g08770.1	photosystem I reaction centre subunit n, chloroplast precursor (psi-n)	1.63E-06	3.08
NA	photosystem I reaction center subunit XI, putative	1.52E-07	3.15
NA	psbD, photosystem II 44 kDa protein, chloroplast	1.95E-05	1.82
NA	psbC, photosystem II protein D2, chloroplast	5.80E-05	1.85
NA	photosystem I subunit IX, chloroplast	2.94E-03	1.05
NA	psbB, photosystem II 47 kDa protein, chloroplast	1.25E-02	0.52
NA	psaC, photosystem I subunit VII, chloroplast	3.89E-02	-0.78

NA; not available

Table 4-13 Chlorophyll and ferredoxin identified in the 20K array with the FDR and Log_2Ratio of the 42 ^oC experiment.

Locus ID	Annotation	FDR	Log_2Ratio
	Chlorophyll		
LOC_Os01g41710.1	Chlorophyll A-B binding protein, putative	6.56E-05	4.91
LOC_Os01g52240.1	Chlorophyll A-B binding protein, putative	1.13E-04	1.76
LOC_Os02g52650.1	Chlorophyll A-B binding protein, putative	2.28E-04	1.78
LOC_Os07g08150.1	Chlorophyll A-B binding protein, putative	3.31E-05	1.28
LOC_Os07g08160.1	Chlorophyll A-B binding protein, putative	4.73E-05	3.45
LOC_Os07g37550.1	Chlorophyll A-B binding protein, putative	7.36E-07	4.31
LOC_Os07g38960.1	Chlorophyll A-B binding protein, putative	1.54E-05	3.43
LOC_Os08g33820.1	chlorophyll a/b-binding protein precursor	3.36E-07	5.18

Table 4-13 (Continued)

Locus ID	Annotation	FDR	Log_2Ratio
LOC_Os09g17740.1	Chlorophyll a/b binding protein 1.	4.48E-06	5.07
N/A	probable chlorophyll a/b-binding protein - rice	9.49E-04	1.48
	Ferredoxin		
LOC_Os01g25490.1	Similar to ferredoxin-nitrite reductase (EC 1.7.7.1) precursor - rice	2.46E-02	0.43
LOC_Os01g72090.1	phytochromobilin:ferredoxin oxidoreductase, chloroplast precursor(ec 1.3.7.4) (phytochromobilin synthase) (pfb synthase)	1.07E-02	0.35
LOC_Os02g17700.1	NADP adrenodoxin-like ferredoxin reductase	1.45E-02	0.36
LOC_Os02g52730.1	ferredoxin--nitrite reductase	9.61E-01	0.06
LOC_Os03g45710.1	putative ferredoxin	2.24E-01	0.18
LOC_Os03g48040.1	putative ferredoxin	5.19E-05	1.83
LOC_Os03g57120.1	ferredoxin-NADP+ reductase	5.04E-03	0.38
LOC_Os03g61960.1	putative ferredoxin	3.56E-01	-0.18
LOC_Os04g33630.1	ferredoxin [2Fe-2S], putative	9.90E-01	0.02
LOC_Os05g37140.1	ferredoxin [2Fe-2S], root - rice	7.72E-01	0.07
LOC_Os06g01850.1	ferredoxin-nadp reductase, leaf isozyme, chloroplast precursor (ec 1.18.1.2) (fnr).	2.39E-07	5.03
LOC_Os07g05400.1	ferredoxin-nadp reductase, embryo isozyme, chloroplast precursor(ec 1.18.1.2) (fnr)	7.63E-05	-1.23
LOC_Os07g46460.1	ferredoxin-dependent glutamate synthase (fragments)	9.06E-07	3.49
NA	ferredoxin-thioredoxin reductase, variable chain (ftr-v) (ferredoxin-thioredoxin reductase subunit a) (ftr-a)	1.51E-05	2.16

NA; not available

4.2. Light/dark 45K array

4.2.1 Statistical analysis

4.2.1.1 Average mean intensity

The 45K array data was derived from the light/dark experiment with various rice cultivars. Labeled cDNAs from four different rice cultivars labeled cDNAs were hybridized to the 45K array at 42 ^{o}C. These rice cultivars were composed of nipponbare, kitaake, TP309 and IR24. This experiment was also intended to verify the 20K array result of the light/dark experiment. The 45K array is composed of two slides called 45K_A and 45K_B. R software using the LMGene package was used as the analysis tool for the 45K array data, similar to the previous experiment with the 20K array. The average mean intensity was calculated to observe the array quality in both slides.

Table 4-14 shows the average mean intensity of both red and green colors. The data showed different mean intensities when compared among the eight slides in each set for 45K_A and 45K_B. The experiment contained eight arrays for four different rice cultivars. The 45K_A data set, contained slides 1 and 2, and represented array data for one rice cultivar with one biological (rice cultivar) and one technical (dye-swap) replicates. The calculation showed different mean intensities between slides 1 and 2, in 45K_A, and also showed the same pattern in 45K_B. The huge difference in slides 1 and 2 of 45K_A and 45K_B might have occurred due to dye

bias caused by different incorporation rates during the labeling step. Both red and green colors mean intensity showed obvious difference in all data sets, depending on the source of the sample. However, mean intensity could be observed at similar levels of Rmean and Gmean in both data sets.

Table 4-14 The basic statistical analysis of average mean intensities for red and green colors.

45K array	Mean	Rmean	Gmean	Gap|R-G|
45K_A				
Slide 1	616.63	665.07	568.20	96.87
Slide 2	1027.15	1438.12	616.18	821.93
Slide 3	562.80	692.63	432.98	259.65
Slide 4	494.21	696.41	292.02	404.39
Slide 5	469.95	503.65	436.25	67.395
Slide 6	669.24	1041.04	297.44	743.60
Slide 7	903.64	560.64	1246.64	685.99
Slide 8	943.58	873.44	1013.73	140.28
45K_B				
Slide 1	759.65	791.04	728.27	62.77
Slide 2	1152.40	1558.47	746.32	812.15
Slide 3	719.62	829.82	609.42	220.40
Slide 4	551.11	804.68	297.53	507.16
Slide 5	544.15	544.72	543.59	1.135
Slide 6	728.71	1123.75	333.68	790.07
Slide 7	811.06	522.70	1099.43	576.72
Slide 8	973.79	901.71	1045.87	144.17

Note; Rmean is referred to mean intensity of red color

Gmean is referred to mean intensity of green color

Gap|R-G| is referred to the different intensity value between red and green colors

4.2.1.2 Control elements

The 45K array contains control elements of hygromycin resistance gene oligos and empty spots similar to the 20K array. The control elements on 45K_A (slide A) include 240 hygromycin resistance gene oligos and 192 empty spots. On the 45K_B (slide B), hygromycin resistance gene oligos and empty elements are represented as 216 and 2,505 spots, respectively.

Table 4-15 presents the ratio of red and green intensities of the control elements in each slide. The error rates were calculated for the ratios over 2-fold and 4-fold. The percentage errors were displayed for average error of the control elements on the eight slides. For slide B, a large number of empty spots showed intensity ratios over 2-fold and 4-fold differences. The error rate in slide B was also higher than slide A. However, the proportion of empty spots on slide B was also higher than slide A. When determined slide by slide, the error spot showed the different number that referred to variation of the array.

The results explained the set of array data for slides A and B. To specify array data set of slide A, the symbol “slide A” was used. Similarly, to indicate the array data set of slide B, the symbol ‘slide B’ was used. Over two-fold error was found in slide A and B, which showed high percentage errors for empty spots of 3.10, and 5.96, respectively. The error rate of hygromycin resistance gene spots was also found in slide A as 1.25 and B as 0.29 at over two-fold differences. For over-four fold difference, the percentage error of hygromycin resistance gene oligos in slide A was

0.36, while no error was found in slide B. The error of empty spots was 0.97 in slide A and 3.44 in slide B with the 4-fold cut-off.

Table 4-15 The control elements on the 45K array calculated for 2 fold and 4 fold changes at 42 ^{o}C annealing temperatures.

	Empty spots		**Hygromycin resistance gene spots**	
	Error > 2 fold	**Error > 4 fold**	**Error > 2 fold**	**Error > 4 fold**
45K_A				
Slide 1	0	0	5	2
Slide 2	0	0	2	0
Slide 3	0	0	2	0
Slide 4	17	7	3	1
Slide 5	0	0	4	1
Slide 6	0	0	3	1
Slide 7	29	8	3	2
Slide 8	2	0	2	0
Percentage error (%)	**3.10**	**0.97**	**1.25**	**0.36**
45K_B				
Slide 1	307	217	1	0
Slide 2	8	0	1	0
Slide 3	214	212	0	0
Slide 4	0	0	0	0
Slide 5	0	0	0	0
Slide 6	60	0	0	0
Slide 7	222	21	1	0
Slide 8	385	240	2	0
Percentage error (%)	**5.96**	**3.44**	**0.29**	**0**

4.2.1.3 MM plot

Two slides of dye-swap data were used to calculate the $\log_2$ ratio of cy3 and cy5 intensities in the MM scatter plot. The ratio from each slide was represented on the x and y axes. This experiment was composed of dye-swap, paired as slides 1-2, 3-4, 5-6, and 7-8. The calculation of correlation coefficient was performed on alternating slides in different pairs as 1-4, 1-6, or 1-8. The other slides were also switched as displayed in Table 4-16. The data had been normalized with lowess method (by R software) before graph plotting. Table 4-16 shows similar correlation coefficients for pairs of slides A and B. The average correlation coefficients were 0.84 for slide A, and 0.80 for slide B. The similar correlation coefficients between array data from the two slides indicate consistent of biological and technical replication. Example of MM plots for the 45K array data are displayed in Figures 14-10 to 14-13. In Figure 4-10, shows MM plot of slides 1 and 2, most of the spot distributed linearly. All array data displayed similar patterns of spot distribution.

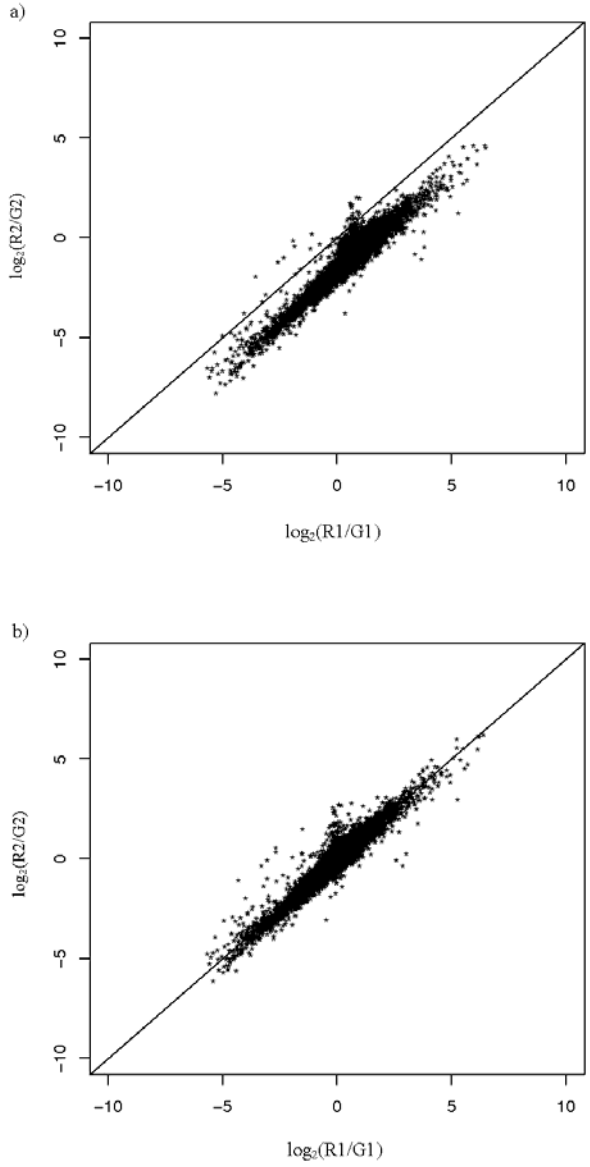


Figure 4-10 MM plot of the signal intensities obtained from 45K_A array data (slides 1 and 2). The x-axis represents the ratio of cy3 (light) over cy5 (dark) intensity (slide 1). The y-axis represents the ratio of cy3 (dark) over cy5 (light) intensity (slide 2). (a) The array data transformed to the $\log_2$(ratio) before normalization, (b) The array data were normalized and then transformed to the $\log_2$(ratio).

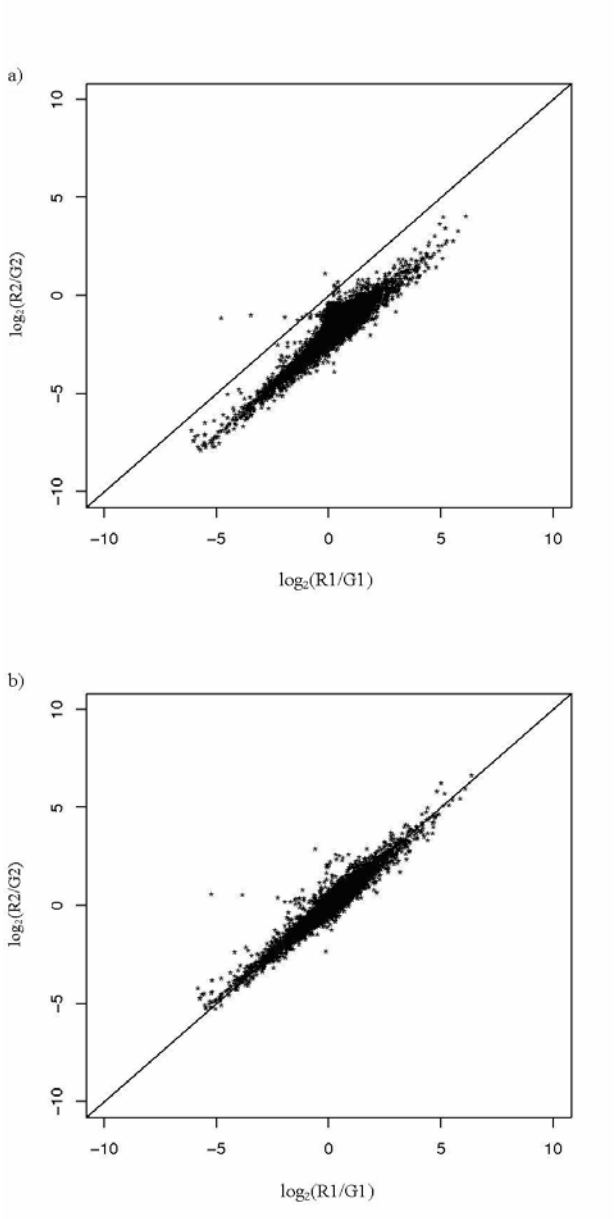


Figure 4-11 MM plot of the signal intensities obtained from 45K_A array data (slides 3 and 4). The x-axis represents the ratio of cy3 (light) over cy5 (dark) intensity (slide 3). The y-axis represents the ratio of cy3 (dark) over cy5 (light) intensity (slide 4). (a) The array data transformed to the $\log_2$(ratio) before normalization, (b) The array data were normalized and then transformed to the $\log_2$(ratio).

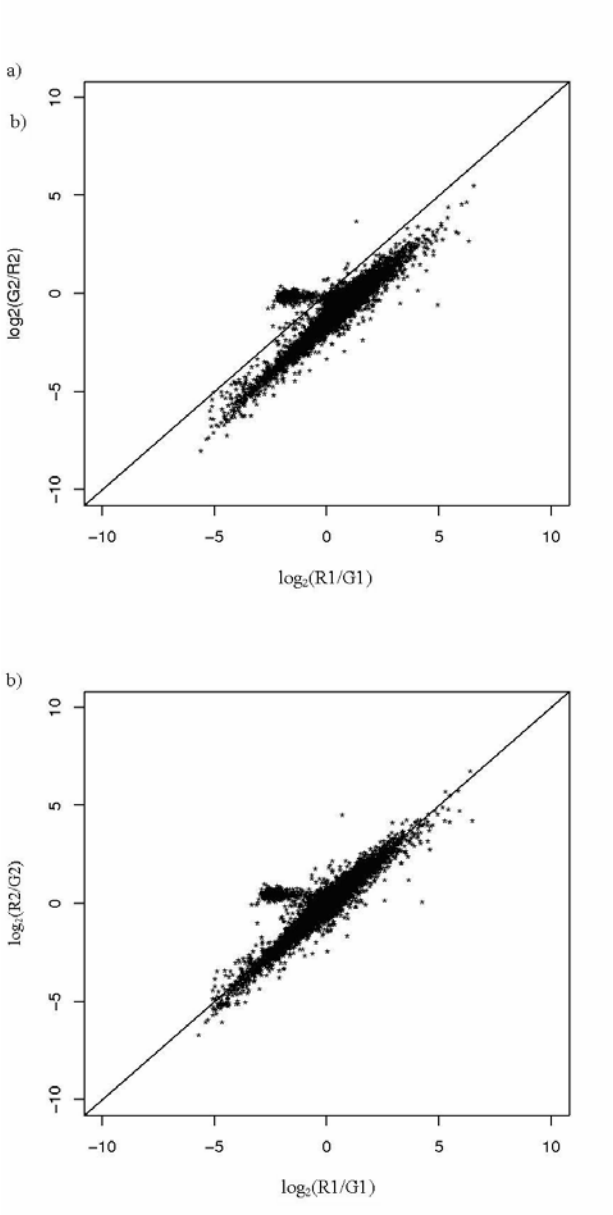


Figure 4-12 MM plot of the signal intensities obtained from 45K_B array data (slides 1 and 2). The x-axis represents the ratio of cy3 (light) over cy5 (dark) intensity (slide 1). The y-axis represents the ratio of cy3 (dark) over cy5 (light) intensity (slide 2). (a) The array data transformed to the $\log_2$(ratio) before normalization, (b) The array data were normalized and then transformed to the $\log_2$(ratio).

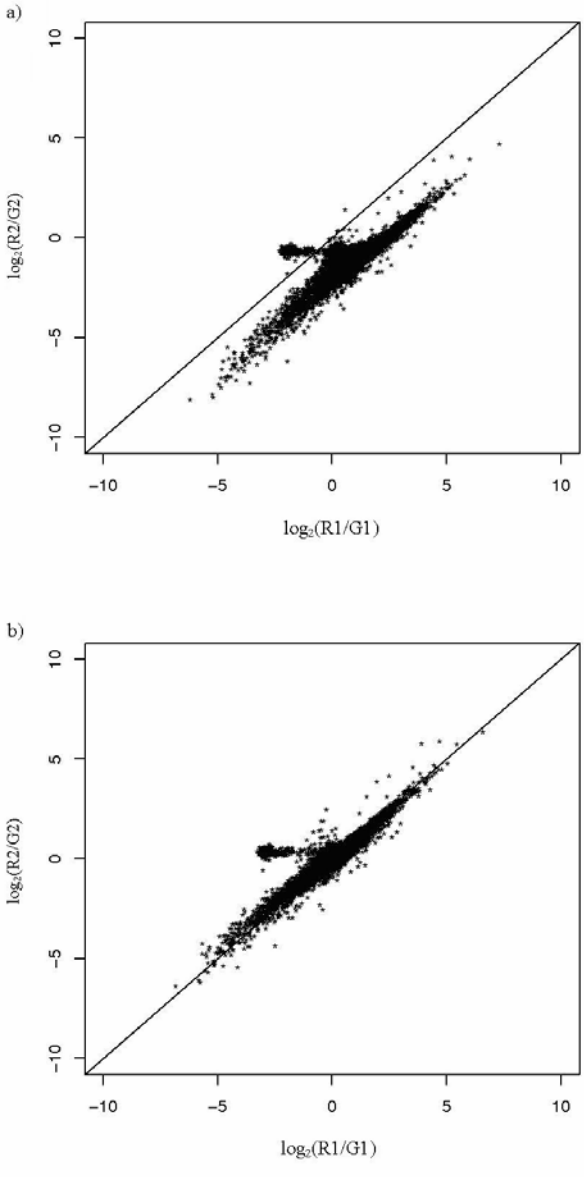


Figure 4-13 MM plot of the signal intensities obtained from 45K_B array data (slides 3 and 4). The x-axis represents the ratio of cy3 (light) over cy5 (dark) intensity (slide 3). The y-axis represents the ratio of cy3 (dark) over cy5 (light) intensity (slide 4). (a) The array data transformed to the $\log_2$(ratio) before normalization, (b) The array data were normalized and then transformed to the $\log_2$(ratio).

Table 4-16 The MM plot of array data between light and dark sample

Temperature (^{o}C)	Comparison		Correlation coefficient
	X	**Y**	
45K_A			
	Slide 1	Slide 2	0.94
	Slide 3	Slide 4	0.93
	Slide 5	Slide 6	0.90
	Slide 7	Slide 8	0.85
	Slide 1	Slide 4	0.81
	Slide 1	Slide 6	0.89
	Slide 1	Slide 8	0.85
	Slide 3	Slide 2	0.80
	Slide 3	Slide 6	0.77
	Slide 3	Slide 8	0.77
	Slide 5	Slide 2	0.84
	Slide 5	Slide 4	0.77
	Slide 5	Slide 8	0.83
	Slide 7	Slide 2	0.83
	Slide 7	Slide 4	0.76
	Slide 7	Slide 6	0.83
average			**0.84**
45K_B			
	Slide 1	Slide 2	0.78
	Slide 3	Slide 4	0.94
	Slide 5	Slide 6	0.88
	Slide 7	Slide 8	0.86
	Slide 1	Slide 4	0.87
	Slide 1	Slide 6	0.66
	Slide 1	Slide 8	0.55
	Slide 3	Slide 2	0.88
	Slide 3	Slide 6	0.84

Table 4-16 (Continued)

Temperature (^{o}C)	Comparison		Correlation coefficient
	X	Y	
	Slide 3	Slide 8	0.74
	Slide 5	Slide 2	0.84
	Slide 5	Slide 4	0.88
	Slide 5	Slide 8	0.75
	Slide 7	Slide 2	0.78
	Slide 7	Slide 4	0.83
	Slide 7	Slide 6	0.75
average			**0.80**

Note; X referred to x-axis, Y-referred to y-axis

4.2.1.4 MA plot

An MA plot was generated with 45K array data from one slide. The $\log_2$ ratios (M) are plotted on the y-axis against the mean intensities (A) on the x-axis. Figures 4-14 and 4-15 represent the MA plot for 45K_A slides 1-8. The MA plot represents the grey spots of normalized data and black spots of raw data. The spots distributed in the range of M from (-2) to 2 for all slides. At low intensity (A<8.0), spots located close to zero 'M' value on the y-axis were found in all slides. The spot distribution was observed in the 45K_B array data which showed a similar pattern as 45K_A. The MA plots for 45K_B slides 1-8 are displayed in Figures 4-16 and 4-17.

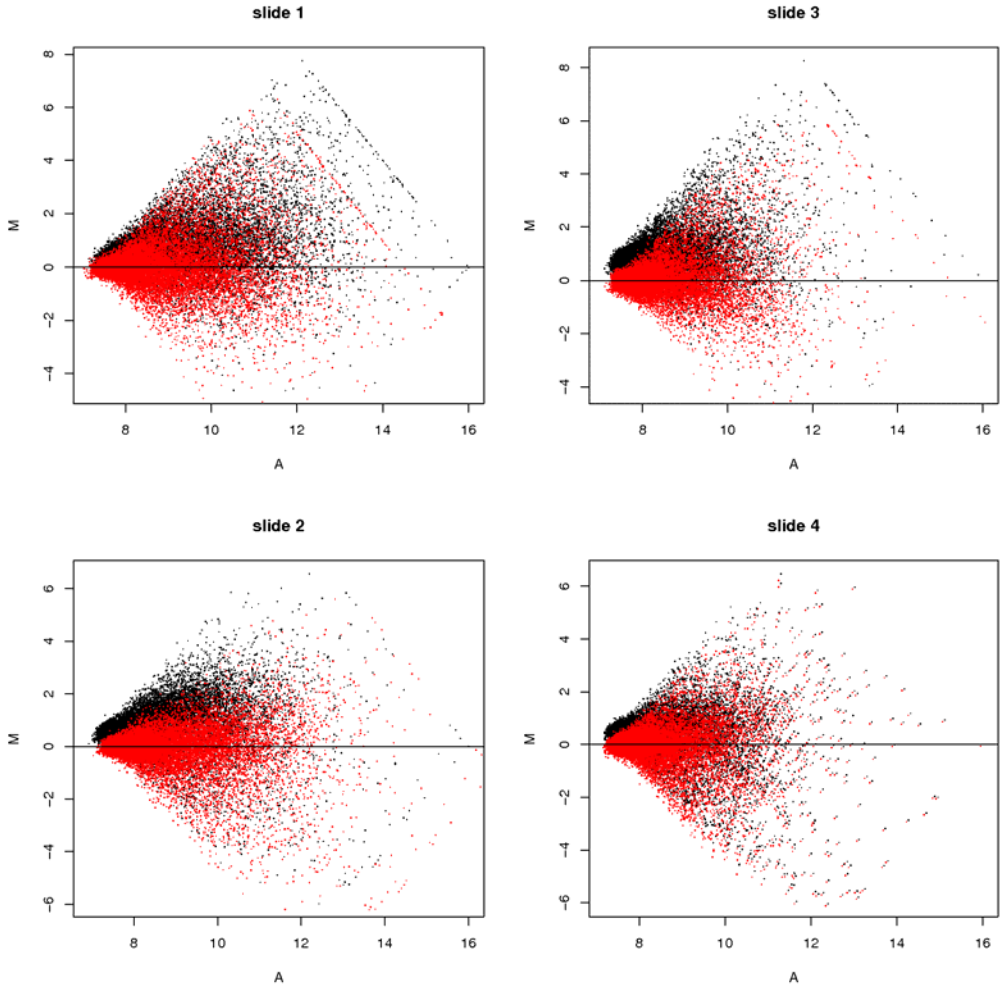


Figure 4-14 MA-plot demonstrating the normalized 45K_A array data.

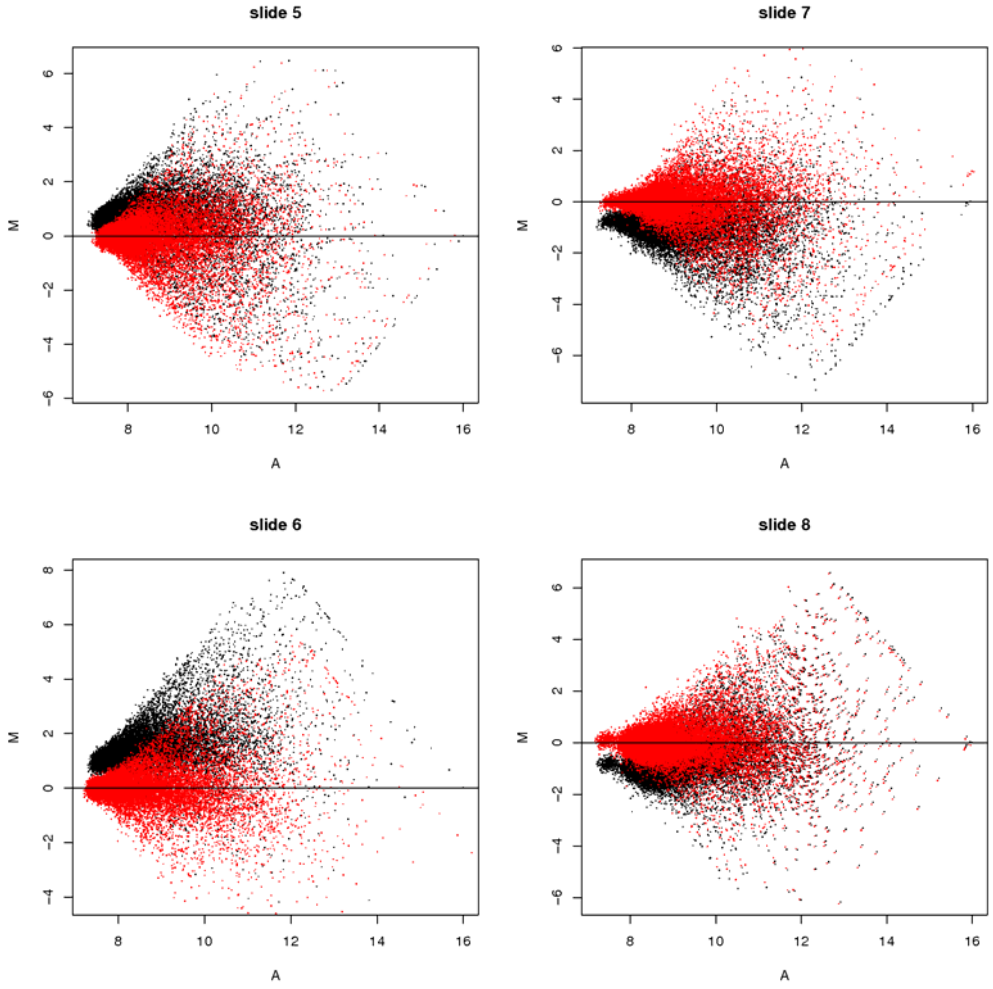


Figure 4-15 MA-plot demonstrating the normalized 45K_A array data.

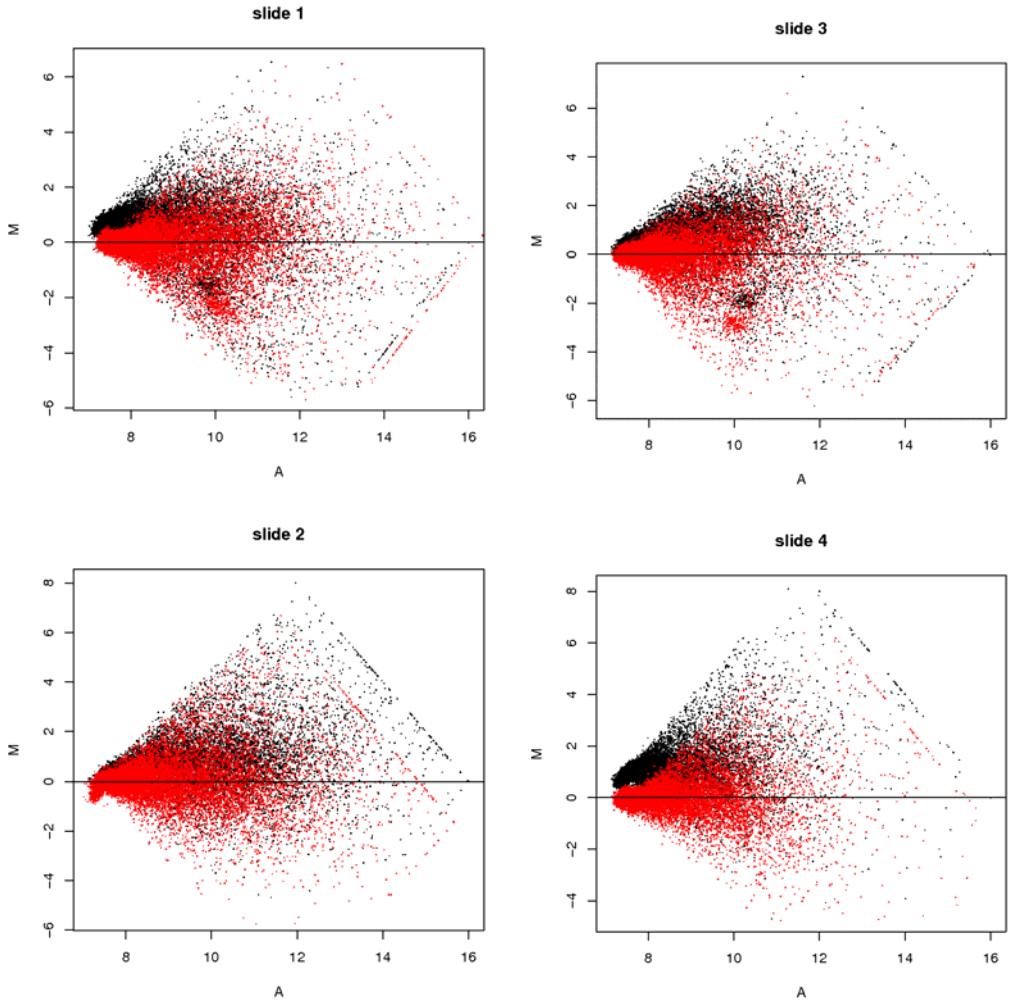


Figure 4-16 MA-plot demonstrating the normalized 45K_B array data.

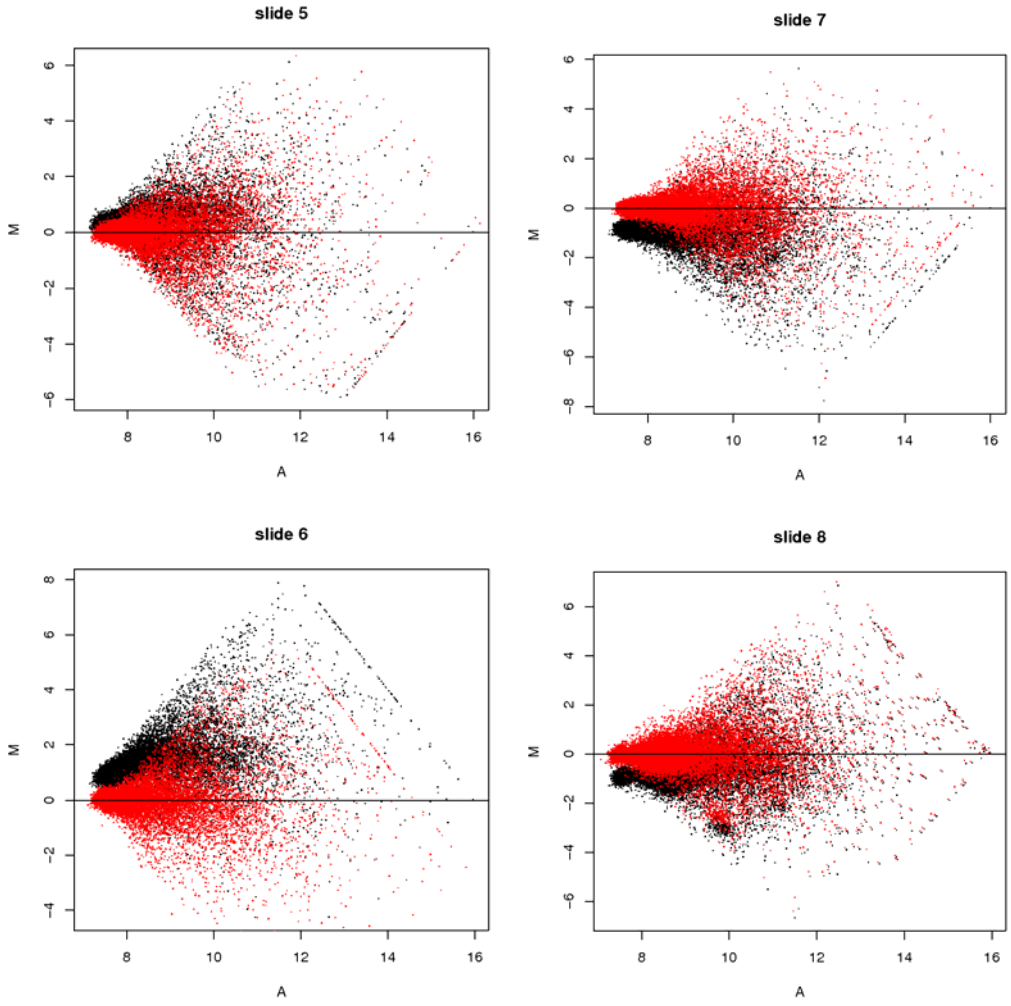


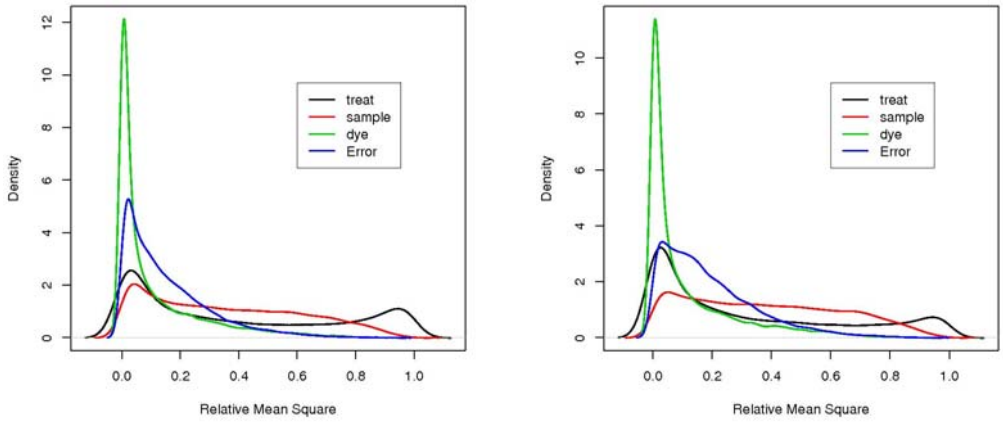
Figure 4-17 MA-plot demonstrating the normalized 45K_B array data.

4.2.1.5 Smoothed histogram analysis of sources of variation

The variable factors in the 45K array data were analyzed similar to the 20K array data. The analysis was divided into two sets of slides for 45K_A and 45_B. Data from all eight slides in each group were combined and smoothed histograms constructed. The systematic error had been classified as dye, slide, treatment, sample, and error factors. The density value (on the y-axis) was plotted against relative mean square on the x-axis. The effect of variable factors was determined from high density and relative mean square. Figure 4-18 shows the variable factors on the 45K_A and 45K_B data sets. The highest effect was observed from the treatment factor for both slides. This result was related to treatment conditions, which are reflected in the treatment factor.

4.2.1.6 Significantly differentially expressed gene identification

The significantly differentially expressed genes in the 45K array data were identified with the FDR cut off values of 0.001, 0.01, and 0.05. This criterion was applied in the gene list to categorize the candidate genes. The number of candidate genes identified with given FDR and fold change cut off values are represented in Table 4-17. The 45K_A array data showed 4,204 candidate genes at 0.001 FDR and this number dramatically increased as the FDR increased. Slide B (45K_B) showed a low candidate gene number when the low FDR value cut off was used similar to the 45K_A array data (slide A). The FDR (p-value) cut off at 0.001, gave 2,223 differentially expressed genes for slide B. At a higher cut off (0.01 FDR), 5,725,



(a) (b)

Figure 4-18 The smoothed histogram representing the effects variable factors on array results for the 45K array. a) 45K_A; b) 45K_B The dye factor represents in green color, treatment factor represents in black color, sample factor represents in red color, and error factor represents in blue color.

and 3,205 differential expressed genes were found in slides A and B, respectively. When the 2- and 4-fold change criteria were applied, the differentially expressed genes were found to be slightly different between 0.001, 0.01, and 0.05 FDR in slide A. The fold change was also divided as up- and down- regulated genes. At the 4-fold change (slide A), a similar number of up- and down-regulated genes were found at 0.01 and 0.05 FDR. This criterion was also applied in slide B for classified candidate gene. The total candidate genes on slide B were fewer than slide A because the total number of oligonucleotide probes on slide A was higher than slide B. The total slide

B candidate gene list was sorted by-2 and 4-fold change criteria and 659 and 181 up-regulated genes were found at FDR 0.001. At 0.01 FDR, the candidate genes reached up to 756 and 199 for up-regulated genes with the cut off at 2 and 4 fold, respectively. The result showed a slightly different candidate gene number when compared to FDR at 0.01 and 0.05 at 4-fold cut off. The up-regulated gene list at FDR of 0.01 was 381 and at 0.05 was 389, when considered at a 4-fold change cut off.

Table 4-17 Significantly differentially expressed genes derived from the 45K array data of slides A and B. The cut off value at FDR (p-value ≤ 0.001, ≤ 0.01 and ≤ 0.05 and fold change were applied in the gene list.

Cut off (FDR)	**Total** (Candidate gene)	**≥2 fold** (up)	**≥2 fold** (down)	**≥4 fold** (up)	**≥4 fold** (down)
	45K_A (23,040 spots)				
0.001	4,024	818	1214	212	471
0.01	5,725	871	1246	220	474
0.05	7,135	878	1252	220	474
	45K_B (20,727 spots)				
0.001	2,223	659	840	181	349
0.01	3,205	756	953	199	381
0.05	4,167	822	1013	207	389

4.2.2 Biological analysis

4.2.2.1 Gene ontology of 45K array

45K array was categorized using the Gene ontology database that was more updated than the previous one used to analyze the 20K array. This database was composed of three main classes of GO: cellular component, biological process and molecular function. The oligonucleotide probes from slide A (45K_A) and B (45K_B) were combined together and categorized with the GO database. The analysis performed the array data by matching oligonucleotide probes with the GO database, followed by subclass categorization. Table 4-18 shows the number of oligonucleotide probes which corresponded to genes in the GO database. The total oligonucleotide probes that matched the GO database, was 17,946 probes or 41%. When classified by GO class, molecular function was found to categorize the highest portion of oligonucleotides on the array with 36%. The biological process and cellular component categories represented lower percentages with 28 % and 18% of the oligos matched, respectively.

An example of GO subclasses were selected, which subclass composed more than 200 oligos corresponded to genes in the GO database. The GO subclasses are represented in Tables 4-19 to 4-21. The molecular function category comprises several subclasses, which show mainly portion in protein or DNA binding, transcription factor activity, and kinase activity. The cellular component showed the large portion in nucleus, mitochondrion, membrane, and cytoplasm. The biological

process showed the regulation of transcription and amino acid phosphorylation as the high portion of genes represented on the array.

Table 4-18 The GO classification of genes represented by oligonucleotide probes on the 45K array

GO class	number of oligo matched GO term	% oligo on array
Biological process	12,075	27.59
Cellular component	7,872	17.99
Molecular function	15,817	36.14
All GO-terms	17,946	36.14
micro-array	43,767	

Table 4-19 Oligo numbers in molecular function categories on the 45K array

Molecular function subclass	Total	Percentage (%)
protein binding	2,463	6.97
transcription factor activity	2,232	6.31
kinase activity	2,051	5.80
DNA binding	1,993	5.64
ATP binding	1,128	3.19
protein serine/threonine kinase activity	862	2.44
zinc ion binding	802	2.27
protein kinase activity	718	2.03
transferase activity, transferring glycosyl groups	632	1.79
RNA binding	601	1.70
nucleotide binding	544	1.54
nucleic acid binding	542	1.53
transporter activity	489	1.38
catalytic activity	440	1.24
oxygen binding	375	1.06
structural constituent of ribosome	374	1.06
calmodulin binding	325	0.92
hydrolase activity	316	0.89
oxidoreductase activity	296	0.84
hydrolase activity, hydrolyzing O-glycosyl compounds	283	0.80
carbohydrate binding	263	0.74
ATPase activity	237	0.67
calcium ion binding	236	0.67
transcriptional activator activity	225	0.64
binding	224	0.63
transcription regulator activity	224	0.63
UDP-glycosyltransferase activity	220	0.62
All GO	35,346	

Table 4-20 Oligo numbers in cellular component categories on the 45K array

Cellular component subclass	Total	Percentage (%)
nucleus	2,452	15.37
mitochondrion	1,591	9.98
membrane	1,516	9.51
cytoplasm	1,358	8.51
chloroplast	685	4.29
plasma membrane	623	3.91
anchored to membrane	374	2.35
cytosol	368	2.31
ribosome	253	1.59
extrinsic to plasma membrane	232	1.45
endoplasmic reticulum	228	1.43
intracellular	223	1.40
All GO	15,948	

Table 4-21 Oligo numbers in biological process categories on the 45K array

Biological process subclasses	Total	Percentage (%)
regulation of transcription, DNA-dependent	970	2.88
protein amino acid phosphorylation	940	2.79
defense response	923	2.74
regulation of transcription	920	2.73
proteolysis	574	1.70
transmembrane receptor protein tyrosine kinase signaling pathway	479	1.42
ubiquitin-dependent protein catabolism	450	1.33
electron transport	431	1.28
response to auxin stimulus	408	1.21
response to abscisic acid stimulus	400	1.18
protein biosynthesis	388	1.15
metabolism	373	1.11
unidimensional cell growth	339	1.01
hypersensitive response	338	1.00
signal transduction	330	0.98
embryonic development (sensu Magnoliophyta)	306	0.91
response to light stimulus	301	0.89
protein folding	293	0.87
response to heat	292	0.87
response to wounding	286	0.85
transport	274	0.81
development	262	0.78
response to water deprivation	247	0.73
response to cold	219	0.65
response to ethylene stimulus	201	0.60
All GO	33,683	

4.2.2.2 Gene ontology of significantly expressed genes

Data of the combination of slides A and B were used for the 45K GO analysis. After data from slide A and B was pooled, the gene list showed a very high number of candidate genes. To find the highly significant differentially expression, the candidate gene list needed to be compressed by use of a lower cut off value. The additional cut off at 0.0001 FDR was applied to the candidate gene list. The resulting

candidate genes were sorted in the three main classes of the GO database. The number of candidate genes is shown in Table 4-22. These three GO classes showed a large number of candidate genes when the FDR cut off of $p \leq 0.0001$, ≤ 0.01, and ≤ 0.05 were applied. The molecular function classified the highest number of candidate genes which corresponded to the large portion of this GO class represented on the array. The other GO classes displayed a large number of candidate genes as well.

Table 4-22 The GO classification of candidate genes by FDR cut off for 45K array data.

	Number (candidate gene)/FDR cut off			**Total number of classified oligos in GO database**
	≤ 0.01%	**≤ 0.1%**	**≤ 5%**	
Biological process	1,873	2,725	4,681	12,075
Cellular component	1,326	1,925	3,241	7,872
Molecular function	2,392	3,529	6,116	15,817
Classified oligo probe	2,726	4,015	6,951	17,946
Total candidate gene	4,192	6,247	11,302	43,767

The data for the 45K array light/dark experiment were categorized into GO subclasses to determine the candidate genes which showed the effect of the experiment. Tables 4-23 to 4-25 represent the GO subclasses identified by 0.001 and 0.05 FDR. The displayed GO subclasses in these tables were selected when over 50% of candidate genes over the total set of oligo probes in these subclasses were affected, and they contained over 10 oligo probes on array. The percentage of the candidate genes was calculated by dividing the number of significant genes in each subclass by the total oligo number that belonged to that subclass. The results in this part were

based on a cut-off of 0.001 FDR. The result for the cellular component subclass demonstrated that the photosystem complex, chloroplast, mitochondrion, and thylakoid were highly affected in the light/dark experiment for over 50% of the probes. The photosystem I and II antenna complexes showed 58% (7 oligos) and 100% (15 oligos), affected by the treatment, respectively. However, these antenna complexes showed 100% when a cut-off value of 0.05 FDR was used. Photosystems I and II reaction centers also showed 78% and 83% when a cut-off value of 0.05 FDR was used.

A shown inTable 4-24, for biological processes 85% of the probe genes in photosynthesis are 82% in electron transport in photosystem I were affected by light or dark. Interestingly, the photosynthesis in light harvesting and light harvesting in photosystem II subclasses show 100% (4 oligos) and 68% (26 oligos) of probes affected, respectively. In photosynthesis, separated light and dark reactions, showed 66% (16 oligos) and 100% (11 oligos) of the represented gene affected, respectively. The transport system showed over 70% as candidate genes in the anion, hexose phosphate and triose phosphate subclasses. The important metabolism, starch and fructose showed over 60% candidate genes, while the tricarboxylic acid cycle with 53%. The biosynthesis subclass showed over 50% candidate genes for porphyrin, chlorophyll, and amino acid synthesis. In addition, gluconeogenesis showed 62% in the significant gene list effect for the experiment.

The significant gene list for the molecular function subclasses is shown in Table 4-25. The highest percentage was represented in sterol regulatory element-

binding protein site 2 protease activity, and sigma factor activity. Both show 100% of the represented genes (12 total oligos) affected by the treatment. The molecular function subclass of glyceraldehyde-3-phosphate dehydrogenase activity classified as NADP(+) and non NADP(+) activity showed 75% and 68% in those subclasses. The phosphoenolpyruvate carboxylase, and pyruvate dehydrogenase (acetyl-transferring) activity, showed 70% and 55%, respectively. The transporter activity of glucose-6-phosphate and triose-phosphate showed over 70%, which was a higher percentage than glucose transporter activity, which showed 56%.

An example for annotated chlorophyll oligo probes on 45K array is shown in Table 4-26. The array results indicate the significant differential expression of these genes with a high confidence of FDR (p-value ≤ 0.05). All of these genes over-express in the light sample compared to the dark sample.

Table 4-23 GO categories (cellular component) in the significant gene list for the 45K array data with ≤ 0.001 and ≤ 0.05 FDR cut-offs.

Cellular component subclass	Total	Number (candidate gene) / (percentage)	
		≤ 5% FDR	≤ 0.1% FDR
photosystem II antenna complex	15	15 (100)	15(100)
thylakoid membrane	25	25 (100)	22(88)
light-harvesting complex	36	36 (100)	31(86)
chloroplast stromal thylakoid	23	21 (91)	19(83)
photosystem I (sensu Viridiplantae)	16	16 (100)	13(81)
microbody	16	13 (81)	12(75)
photosystem II reaction center	18	15 (83)	12(66)
photosystem I reaction center	14	11 (78)	9(64)
thylakoid (sensu Viridiplantae)	32	25 (78)	20(62)
photosystem II (sensu Viridiplantae)	36	28 (77)	22(61)
photosystem I antenna complex	12	12 (100)	7(58)
peroxisomal membrane	21	14 (66)	12(57)
oxygen evolving complex	23	16 (69)	13(56)
thylakoid lumen (sensu Viridiplantae)	122	83 (68)	67(55)
chloroplast inner membrane	106	74 (70)	58(55)
mitochondrial outer membrane	21	12 (57)	11(52)
chloroplast stroma	158	103 (65)	81(51)
photosystem II	12	6 (50)	6(50)

Note : The classified oligonucleotide probes showed in this table, were selected by using criteria of classified candidate genes in each subclasses showed over 50% of the represented genes.

Table 4-24 GO categories (biological process) in the significant gene list for the 45K array data with ≤ 0.001 and ≤ 0.05 FDR cut-off values.

Biological process subclass	Number (candidate gene) / (percentage)		
	Total	5% FDR	0.1% FDR
photosynthesis, dark reaction	11	11 (100)	11 (100)
photosynthesis, light harvesting	4	4 (100)	4 (100)
photosynthesis	53	50 (94)	45 (85)
photosynthetic electron transport in photosystem I	17	17 (100)	14 (82)
photorespiration	24	20 (83)	19 (79)
anion transport	15	11 (73)	11 (73)
thylakoid membrane organization and biogenesis	33	27 (82)	24 (72)
hexose phosphate transport	22	22 (100)	16 (72)
triose phosphate transport	22	22 (100)	16 (72)
photosynthesis, light harvesting in photosystem II	38	29 (76)	26 (68)
induction of apoptosis	19	15 (79)	13 (68)
photosynthesis, light reaction	24	21 (87)	16 (66)
porphyrin biosynthesis	15	11 (73)	10 (66)
nitrate assimilation	15	13 (86)	10 (66)
short-chain fatty acid metabolism	12	9 (75)	8 (66)
fructose metabolism	12	9 (75)	8 (66)
photosynthesis, light harvesting in photosystem I	23	20 (87)	15 (65)
starch metabolism	11	7 (63)	7 (63)
gluconeogenesis	16	10 (62)	10 (62)
photosystem II assembly	13	13 (100)	8 (61)
protein import into chloroplast stroma	20	16 (80)	12 (60)
chlorophyll biosynthesis	40	31 (77)	23 (57)
tricarboxylic acid cycle	15	10 (66)	8 (53)
photosynthetic electron transport	21	13 (62)	11 (52)
amino acid biosynthesis	23	18 (78)	12 (52)
systemic acquired resistance, salicylic acid mediated signaling pathway	31	21 (68)	16 (52)

Note : The classified oligonucleotide probes showed in this table, were selected by using criteria of classified candidate genes in each subclasses showed over 50% of the represented genes.

Table 4-25 GO categories (molecular function) in the significant gene list for the 45K array data with ≤ 0.001 and ≤ 0.05 FDR cut-off values.

Molecular function subclass	Total	Number (candidate gene) / (percentage)	
		5% FDR	5% FDR
sterol regulatory element-binding protein site 2 protease activity	12	12 (100)	12 (100)
sigma factor activity	12	12 (100)	12 (100)
glyceraldehyde-3-phosphate dehydrogenase (NADP+) activity	16	12 (75)	12 (75)
membrane alanyl aminopeptidase activity	15	11 (73)	11 (73)
triose-phosphate transporter activity	22	22 (100)	16 (72)
glucose-6-phosphate transporter activity	22	22 (100)	16 (72)
tryptophan synthase activity	11	11 (100)	8 (72)
phosphoenolpyruvate carboxylase activity	10	7 (70)	7 (70)
galactokinase activity	13	10 (77)	9 (69)
protochlorophyllide reductase activity	13	10 (77)	9 (69)
glyceraldehyde-3-phosphate dehydrogenase activity	19	15 (79)	13 (68)
chlorophyll binding	70	55 (78)	47 (67)
acyl-CoA oxidase activity	17	12 (71)	11 (65)
phosphatidylcholine-sterol O-acyltransferase activity	17	13 (76)	11 (65)
glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) activity	17	13 (76)	11 (65)
protein phosphorylated amino acid binding	21	14 (67)	13 (62)
protein tyrosine/serine/threonine phosphatase activity	31	20 (65)	19 (61)
pigment binding	12	12 (100)	7 (58)
sucrose synthase activity	12	7 (58)	7 (58)
tyrosine aminopeptidase activity	21	13 (62)	12 (57)
FMN adenylyltransferase activity	14	8 (57)	8 (57)
riboflavin kinase activity	14	8 (57)	8 (57)
alpha,alpha-trehalose-phosphate synthase (UDP-forming) activity	25	16 (64)	14 (56)
glucose transporter activity	45	35 (78)	25 (56)
pyruvate dehydrogenase (acetyl-transferring) activity	20	15 (75)	11 (55)

Note : The classified oligonucleotide probes showed in this table, were selected by using criteria of classified candidate genes in each subclasses showed over 50% of the represented genes.

Table 4-26 The identification of 45K array genes with chlorophyll annotated function. The FDR and Log_2Ratio for each oligo probe are presented.

Locus ID	Annotation	FDR	Log_2Ratio
LOC_Os01g41710.1; LOC_Os01g41710.2	Chlorophyll a-b binding protein 2, chloroplast precursor, putative, expressed	1.02 E-04	3.24
LOC_Os01g52240.1	Chlorophyll a-b binding protein 2, chloroplast precursor, putative, expressed	3.78E-07	5.95
LOC_Os02g18500.1	Chlorophyllase-2, chloroplast precursor, putative	4.43E-01	0.13
LOC_Os02g52650.1	Chlorophyll A-B binding protein, expressed	3.03E-07	4.20
LOC_Os03g39610.1; LOC_Os03g39610.3	Chlorophyll a-b binding protein, chloroplast precursor, putative, expressed	4.34E-06	4.73
LOC_Os04g16750.1	Photosystem I P700 chlorophyll a apoprotein A2, putative, expressed	2.88E-04	1.92
LOC_Os04g38410.1	Chlorophyll a-b binding protein CP24 10B, chloroplast precursor, putative, expressed	1.12E-05	4.32
LOC_Os07g37550.1	Chlorophyll a-b binding protein of LHCII type III, chloroplast precursor, putative, expressed	1.98E-06	5.00
LOC_Os07g38960.2	Chlorophyll a-b binding protein 7, chloroplast precursor, putative, expressed	2.24E-06	3.23
LOC_Os08g15260.1; LOC_Os10g21310.1	Photosystem II P680 chlorophyll A apoprotein, putative, expressed;Photosystem II P680 chlorophyll A apoprotein, putative, expressed	1.07E-03	1.46
LOC_Os08g33820.1	Chlorophyll a-b binding protein 4, chloroplast precursor, putative, expressed	2.66E-04	4.01

4.3 Confirmation of the significant genes by RT-PCR

4.3.1 Significant genes of 20K and 45K array

The 20K array data were selected for validation of the array results. The candidate genes were derived from the experiment with 42 ^{o}C annealing temperature, which was the optimum annealing temperature. Random selection was done in the range of 850 candidate genes with 0.0001 FDR. The criterion for selecting was set at over 2 fold up-regulation in light compared to dark treated leaf tissue. The list of 32 selected genes for the 20K and 45K array data were represented with their $\log_2$ ratio (table 4-27). The genes selected from the 20K array are also contained in the 45K array, except for three genes. The comparison of the $\log_2$ ratio from the 20K and 45K array data of the selected genes found that most selected genes in the 45K array data showed higher $\log_2$ values. All of the selected genes did not have similar $\log_2$ value in both 20K and 45K arrays.

4.3.2 RT-PCR

RT-PCR technique was used to verify the array data with 32 selected genes. The specific primers for specific genes were designed with the Primer 3.0 interactive tool. The reverse transcription was performed with mRNA template derived from four different rice cultivars, which were kitaake, nipponbare, TP309, and IR24. These mRNAs were derived from light/dark experiment in the 45K array. The control actin and ubiquitin genes were used to equilibrate the amount of transcript in each of the

mRNA samples. Figure 4-19 shows a gel electrophoresis image of all 32 selected genes. The samples were arranged by light treated leaf (L) followed by the dark treated leaf (D) that for each of the four rice cultivars.

Table 4-27 Comparison of the $\log_2$ ratio for selected genes for the 20K and 45K arrays

		Log_2Ratio	
Locus ID	**Annotation (45K array)**	**20K array**	**45K array**
LOC_Os08g33820.1	Chlorophyll a-b binding protein 4, chloroplast precursor, putative, expressed	5.33	4.01
LOC_Os01g45280.2	Carbonic anhydrase, putative	4.66	NA
LOC_Os03g38950.1	Expressed protein	4.44	3.27
LOC_Os05g41640.3	phosphoglycerate kinase	3.88	NA
LOC_Os03g03720.1	Glyceraldehyde-3-phosphate dehydrogenase B, chloroplast precursor, putative, expressed	3.81	5.43
LOC_Os10g37180.1	Glycine cleavage system H protein, mitochondrial precursor, putative, expressed	3.76	2.92
LOC_Os01g55570.1	Expressed protein	3.70	4.96
LOC_Os04g53230.1	Aminomethyltransferase, mitochondrial precursor, putative, expressed	3.53	5.24
LOC_Os08g10020.1	Photosystem II 10 kDa polypeptide, chloroplast precursor, putative, expressed	3.51	4.55
LOC_Os03g16050.1	Fructose-1,6-bisphosphatase, chloroplast precursor, putative, expressed	3.45	4.81
LOC_Os01g51410.1	Glycine dehydrogenase, mitochondrial precursor, putative, expressed;Glycine dehydrogenase, mitochondrial precursor, putative, expressed	3.30	4.39
LOC_Os02g01340.1	Ferredoxin-NADP reductase, leaf isozyme, chloroplast precursor, putative, expressed	3.25	4.65
LOC_Os03g36750.1	HAD-superfamily hydrolase, subfamily IA, variant 3 containing protein, expressed	3.24	4.47
LOC_Os03g61220.1	DEAD/DEAH box helicase family protein, expressed	3.12	3.59

Table 4-27 (Continued)

Locus ID	Annotation (45K array)	Log_2Ratio 20K array	Log_2Ratio 45K array
LOC_Os12g08770.1	Photosystem I reaction centre subunit N, chloroplast precursor, putative, expressed	3.08	3.62
LOC_Os08g44810.1	Malate dehydrogenase 1, chloroplast precursor, putative, expressed	3.04	3.23
LOC_Os02g01150.1	D-isomer specific 2-hydroxyacid dehydrogenase, NAD binding domain containing protein, expressed	2.94	3.61
LOC_Os10g35370.2	Protochlorophyllide reductase B, chloroplast precursor, putative, expressed	2.93	4.19
LOC_Os03g51930.1	expressed protein	2.66	NA
LOC_Os04g55710.1	Transposon protein, putative, unclassified, expressed	2.64	1.90
LOC_Os04g51300.1	L-ascorbate peroxidase, chloroplast precursor, putative, expressed	2.60	3.24
LOC_Os03g52460.1	Glucose-1-phosphate adenylyltransferase large subunit 3, chloroplast precursor, putative, expressed	2.38	4.22
LOC_Os02g34810.1	L-ascorbate peroxidase 8, chloroplast precursor, putative, expressed	2.31	3.52
LOC_Os08g44320.1	Low molecular weight phosphotyrosine protein phosphatase containing protein, expressed	2.24	2.60
LOC_Os08g02210.2	Expressed protein	2.20	4.19
LOC_Os08g14440.3	expressed protein	2.17	NA
LOC_Os07g08970.1	Expressed protein	2.06	2.87
LOC_Os01g73500.1	Expressed protein	1.90	2.36
LOC_Os12g17910.1	RuBisCO subunit binding-protein alpha subunit, chloroplast precursor, putative, expressed	1.89	1.56
LOC_Os08g43560.1	L-ascorbate peroxidase 4, putative, expressed	1.85	2.20
LOC_Os10g40030.1	Oxidoreductase, short chain dehydrogenase/reductase family protein, expressed	1.83	2.30
LOC_Os03g48040.1	Ferredoxin family protein, expressed	1.83	3.26

NA; not available

Figure 4-19 presents the gel images for RT-PCR of all 32 selected genes selected from 20K array data. The control genes were actin and ubiquitin, which showed equal signal intensity for all samples. Different numbers of PCR cycles were used to optimize the amplified product. Twenty-three cycles were used for the control genes and thirty cycles were used for the selected genes. All sample patterns were arranged in alternating light and dark treatments, which is obvious in some specific primer sets with up- and down-regulation in the light/dark pattern. However, some patterns showed signal from dark sample and equal signal in both light and dark samples. This error might be due to the different source of sample (rice cultivar), which show the different expression levels in each genes.

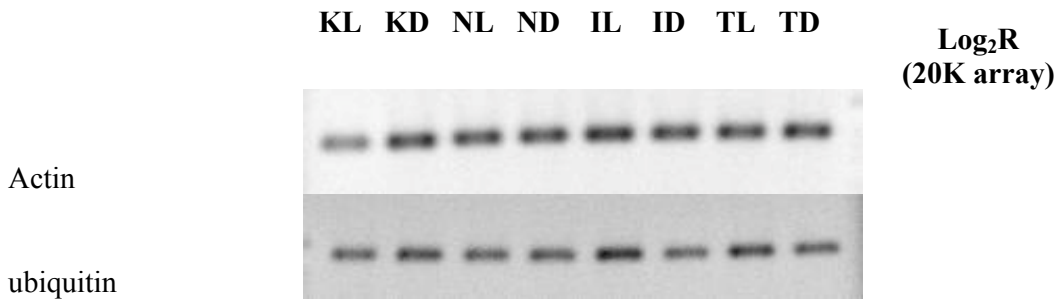


Figure 4-19 RT-PCR image for 32 genes selected from the 20K and 45K arrays. The image contains 8 specific lanes for samples. 1) KL; kitaake leaves treated with light, 2) KD; kitaake leaves treated with dark, 3) NL; nipponbare leaves treated with light, 4) ND; nipponbare leave treated with light, 5) IL; IR24 leaves treated with light, 6) ID; IR24 leaves treated with dark, 7) TL; TP309 leaves treated with light, 8) TD; TP309 leaves treated with dark.

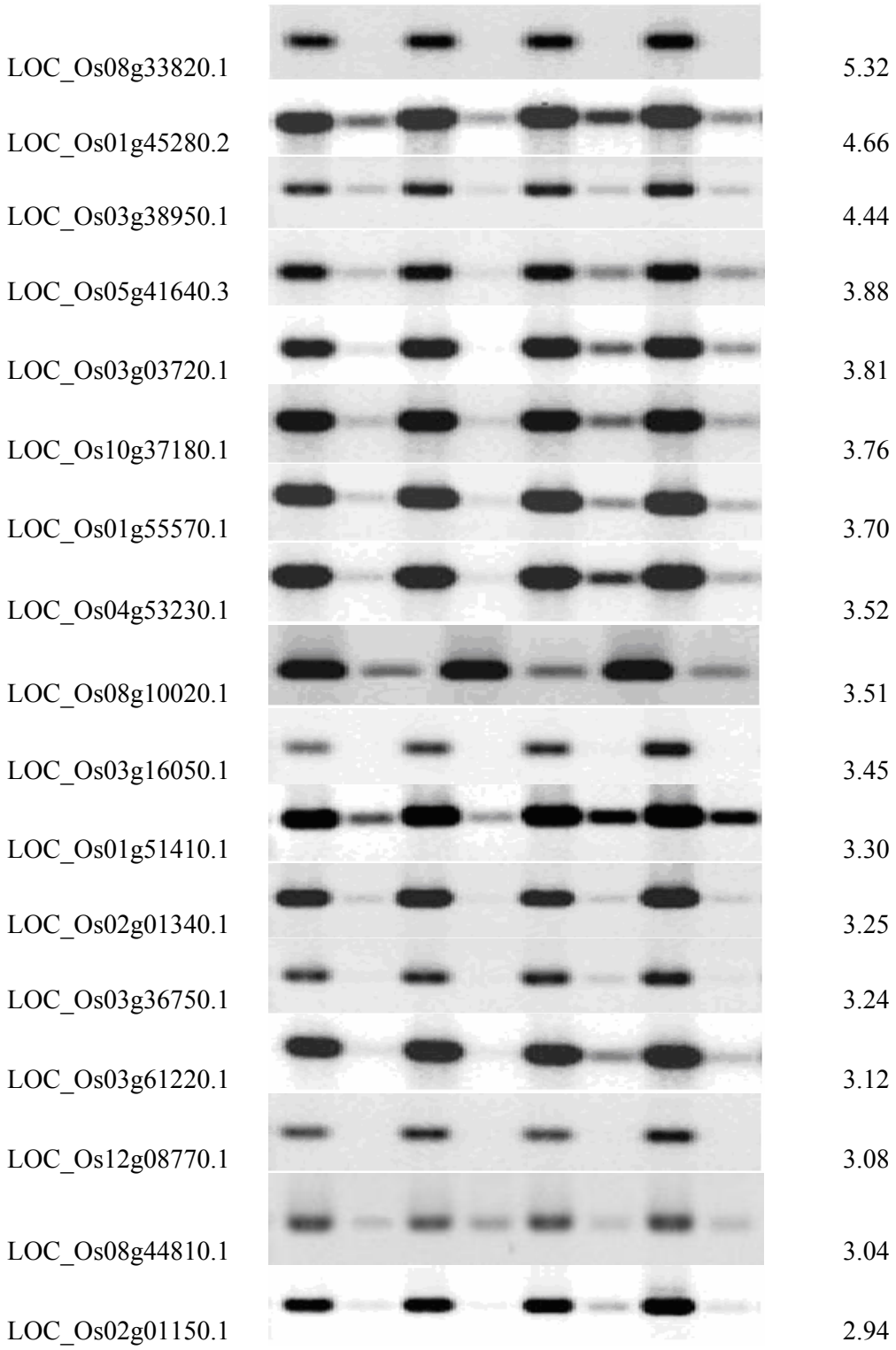


Figure 4-19 (Continued)

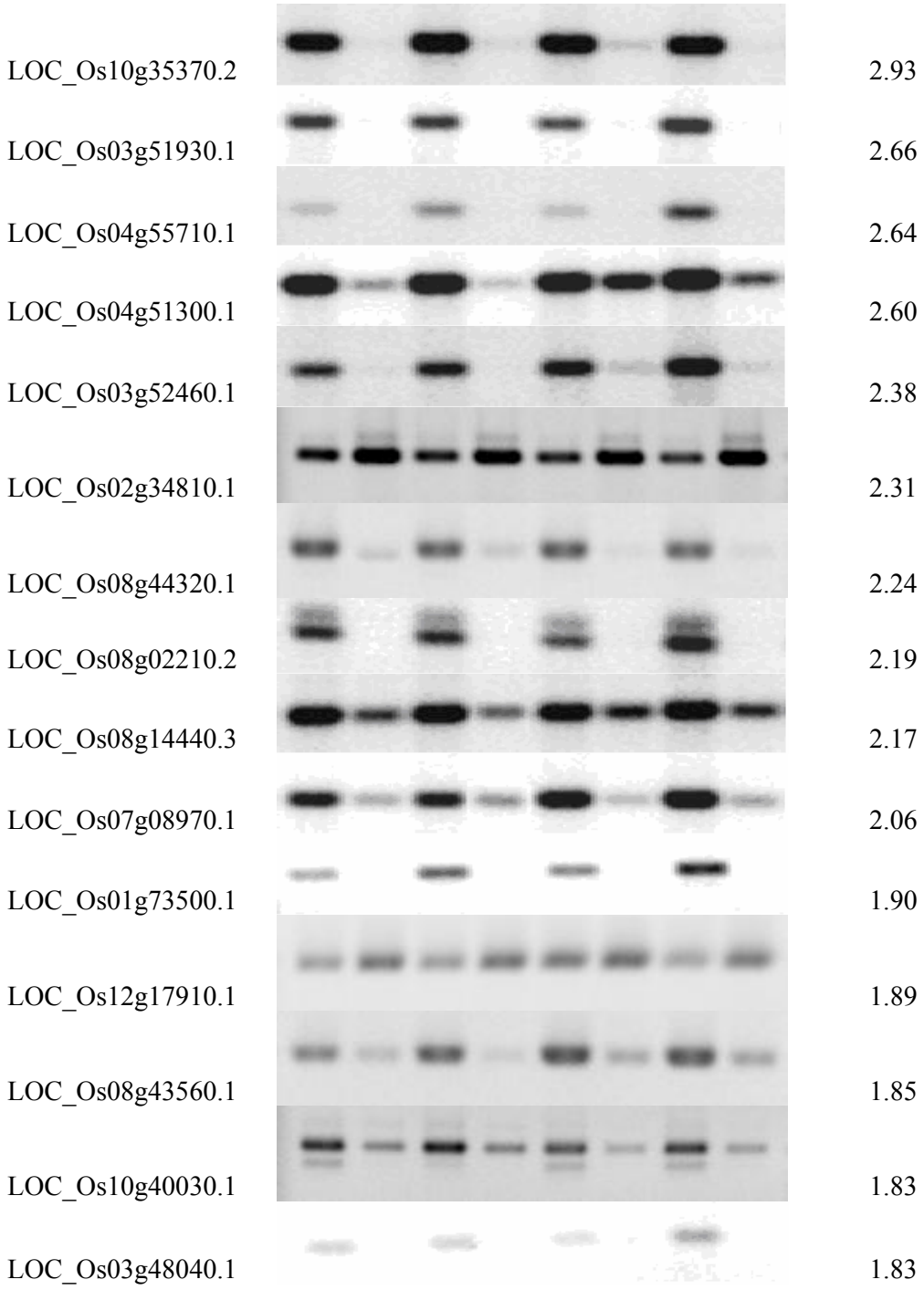


Figure 4-19 (Continued)

4.4 Identification of Glycosyl hydrolase family 1

The annotation file was determined to annotate the function of glycosyl hydrolase family 1 (GH 1). Only 26 oligonucleotides were found corresponding to this term (table 4-28). The 45K array results showed down-regulation in most of the oligos. The FDR cut-off value at 0.05 was used to identify the significantly differential expressed genes in GH 1 groups. Out of 26 genes only 13 genes were found to be significantly differential expressed (table 4-29). The significantly differential expressed genes showed down-regulation with $\log_2$ of -0.28 to -1.19, and up-regulation with $\log_2$ of 0.23 to 2.37, in the light treated leaf.

Table 4-28 The identified GH 1 oligonucleotides of the 45K array data

Locus ID	Annotation	FDR (p-value)	Log_2(Ratio)
LOC_Os01g70520.1; LOC_Os01g70520.2	Glycosyl hydrolase family 1 protein, expressed	6.79 E-08	2.16
LOC_Os11g45710.1; LOC_Os11g45710.3	Glycosyl hydrolase family 1 protein, putative, expressed	1.93 E-07	1.62
LOC_Os11g45710.2	Glycosyl hydrolase family 1 protein, putative, expressed	1.97 E-07	2.28
LOC_Os07g46280.1; LOC_Os07g46280.2; LOC_Os07g46280.3	Glycosyl hydrolase family 1 protein, expressed	3.73 E-07	-2.37
LOC_Os09g31430.1; LOC_Os09g31430.2	Glycosyl hydrolase family 1 protein, expressed	6.95 E-06	-1.31
LOC_Os09g33680.1; LOC_Os09g33680.2	Glycosyl hydrolase family 1 protein, expressed	2.53 E-05	1.10
LOC_Os04g43360.1	Glycosyl hydrolase family 1 protein, expressed	1.22 E-04	0.58
LOC_Os01g67220.1; LOC_Os01g67220.2; LOC_Os01g67220.3	Glycosyl hydrolase family 1 protein, expressed	1.40 E-04	1.02
LOC_Os05g30250.1	Glycosyl hydrolase family 1 protein, expressed	3.69 E-04	1.98

Table 4-28 (Continued)

Locus ID	Annotation	FDR (p-value)	Log_2(Ratio)
LOC_Os03g11420.1	Glycosyl hydrolase family 1 protein, expressed	6.84 E-04	-1.19
LOC_Os04g43410.1	Glycosyl hydrolase family 1 protein, expressed	1.23 E-03	-1.01
LOC_Os01g32364.1	Glycosyl hydrolase family 1 protein, expressed	1.82 E-02	0.23
LOC_Os03g49610.1	Glycosyl hydrolase family 1 protein, expressed	3.31 E-02	-0.28
LOC_Os06g21570.1	Glycosyl hydrolase family 1 protein, expressed	5.41 E-02	0.33
LOC_Os05g30350.1; LOC_Os05g30350.2	Glycosyl hydrolase family 1 protein, expressed	5.94 E-02	1.29
LOC_Os03g49600.1; LOC_Os03g49600.2; LOC_Os03g49600.4	Glycosyl hydrolase family 1 protein, expressed	8.30 E-02	0.40
LOC_Os09g33710.1	Glycosyl hydrolase family 1 protein, expressed	1.4 E-01	0.28
LOC_Os04g43390.1	Glycosyl hydrolase family 1 protein, expressed	1.69 E-01	-0.61
LOC_Os01g59819.1	Glycosyl hydrolase family 1 protein, expressed	4.28 E-01	0.08
LOC_Os01g59819.1	Glycosyl hydrolase family 1 protein, expressed	5.07 E-01	0.09
LOC_Os09g31410.1	Glycosyl hydrolase family 1 protein, expressed	6.59 E-01	-0.14
LOC_Os03g20710.1	Glycosyl hydrolase family 1 protein	9.4 E-01	0.01
LOC_Os10g17650.1	Glycosyl hydrolase family 1 protein, expressed	9.99 E-01	0.05
LOC_Os05g30300.1	Glycosyl hydrolase family 1 protein, expressed	9.99 E-01	0.05
LOC_Os05g30390.1	Glycosyl hydrolase family 1 protein	9.99 E-01	0.04
LOC_Os12g23170.1	Glycosyl hydrolase family 1 protein, expressed	9.99 E-01	-0.04

Table 4-29 The identified glycosyl hydrolase family 1 with Opassiri, 2006.

Locus ID	Annotation (Opassiri *et al*., 2006)	FDR (p-value)	Log_2(Ratio)
LOC_Os03g49610.1	Os3*bglu*8	3.31E-02	-0.28
LOC_Os01g32364.1	Os1*bglu*1	1.82E-02	0.23
LOC_Os04g43410.1	Os4*bglu*18	1.23E-03	-1.01
LOC_Os03g11420.1	Os3*bglu*6	6.84E-04	-1.19
LOC_Os05g30250.1	Os5*bglu*19	3.69E-04	1.98
LOC_Os01g67220.1; LOC_Os01g67220.2; LOC_Os01g67220.3	NA	1.41E-04	1.02
LOC_Os04g43360.1	Os4*bglu*14	1.22E-04	0.58
LOC_Os09g33680.1; LOC_Os09g33680.2	Os9*bglu*31	2.53E-05	1.10
LOC_Os09g31430.1; LOC_Os09g31430.2	Os9*bglu*30	6.95E-06	1.31
LOC_Os07g46280.1; LOC_Os07g46280.2; LOC_Os07g46280.3	Os7*bglu*26	3.73E-07	2.37
LOC_Os11g45710.2	NA	1.97E-07	2.28
LOC_Os11g45710.1; LOC_Os11g45710.3	Os11*bglu*36	1.93E-07	1.62
LOC_Os01g70520.1; LOC_Os01g70520.2	Os1*bglu*5	6.79E-08	2.16

NA; not available

CHAPTER V

DISCUSSIONS

5.1 Statistical analysis

5.1.1 Statistical analysis of 20K array light/dark experiment

5.1.1.1 Mean intensity

In this study, R program installed LMGene package was used for array data analysis. To optimize the annealing temperature, the 20K arrays were hybridized at annealing 42, 46, and 50 ^{o}C. The spot intensity was determined with two channels for both cy3 (G) and cy5 (R). The basic statistical analysis represented the mean, Rmean, and Gmean for cy3 and cy5 intensity. The average mean intensity showed the highest value at 42 ^{o}C and decreased the higher temperature. The high signal intensity also increased the image quality. High annealing temperature improves the hybridization specificity between probes and targets, and decreases the background intensity. However, the signal intensity is the most critical parameter that influences the reproducibility, specificity and sensibility of microarray experiment (Yang *et al*., 2001).

5.1.1.2 Normalization

Normalization is an important part of data analysis in order to adjust the balancing of individual hybridization intensities for improving the biological meaning (Quackenbush, 2002). The normalization can be used to reduce the variations which occur from unequal quantities of starting RNA, different labeling set or slides, and systematic biases in the measured expression levels (Yang *et al*., 2002). In this experiment, the non-linear regression method was applied to the array data set. The common method for non-linear regression is call "loess" or "lowess". The loess approach was previously used by Yang and his colleage (Yang *et al*., 2002). The loess normalization method was installed in LMGene package and used for this array data. The conceptual idea for normalization is similar to equilibrate the expression level measured by northern analysis or quantitative reverse transcription PCR (RT-PCR) that related to the reference genes. These reference genes are assumed to have constant expression level between samples. The normalized data were further transformed to $\log_2$(ratio) of cy3 over cy5 to create reliable significant differential expression gene list.

5.1.1.3 MA plot

The range of signal intensity was also observed using the MA plot which was classified as scatter plot. A scatter plot is the most common graphical display following the microarray data (Bowtell and Sambrook, 2003). The MA-plot was described by Dudoit *et al*. (Dudoit *et al*., 2002). This plot represents the array data in

both red and green intensities. "M" is the $\log_2$ (R/G) and "A" is the ½$\log_2$(RxG). The array data are transformed to "M" and plot on the y-axis against the "A" for each spot on the slide. The array data were normalized with loess method before transformed to "M" and "A". The normalization can remove the intensity-dependent effects in the $\log_2$(ratio) value. The 20K array data derived from two different samples, therefore, these data created different spot distribution patterns. Light and dark array data sets displayed the spot distributions that are likely to merge and partly cover the patterns. A straight line had been fitted to the data points which demonstrated a clear trend in the cy3 and cy5 responses. At low intensities, the cy3 channel was responding more strongly with un-normalized data, while at high intensities, the cy5 channel was responding more strongly with normalized data. The separated spots were signified as different signal intensities and expressed genes. Most of array data set showed similar MA-plot pattern, except some data sets showed spots over fusion in the x-axis which cause by normalized and un-normalized with an equal Rmean and Gmean intensities.

5.1.1.4 MM plot

The alternative scatter plot called "MM plot" which is the transformed array data to $\log_2$(ratio) of cy3 over cy5. The scatter plot constructed by the $\log_2$(ratio) from two array data sets which plotted on the x- and y-axes. The 'cloud' of data represented the distribution of spot intensities. The correlation of array data between replicates could be determined by MM plot as well. The correlation coefficient is used to quantify the level of relation between two data sets. All array data sets were

used to create the MM plots and all of them showed high correlation coefficient among each other with an average of 0.84 to 0.88. The correlation coefficient represents the value between (-1) to +1. A value of -1 implies the strong negative correlation. A value of +1 implies the strong positive correlation. Most array experiments performed with different sources of sample were difficult to form the perfect correlation coefficient value. Therefore, the correlation coefficient of 0.84 - 0.88 are high enough confident of worthy relation among sample set.

5.1.1.5 Smoothed histogram

The systematic variation could be determined using the smoothed histogram of source of variations. These variation factors consisted of slide, error, sample, dye, and treatment. The array data of the relative mean square of intensity were plotted against the frequency of spot. The highest mean square intensity and spot frequency represent the highest variation factor. In our experiment, all array data sets showed that the treatment factor was the highest effect on the array data set.

To reduce the data variability, all arrays should be ordered from the same manufacturing batch. Some array data usually show that the dye factor is the highest effect on array data due to the dye properties and incorporation rate during labeling step (data not shown) which is not a good sign. In this case, the results will be due to the dye rather than the treatment.

5.1.1.6 Array design for 20K experiment

The 20K array was designed using direct comparison between two samples on the same slide. This method is a unique and powerful feature of the two-color microarray system (Churchill, 2002) The dye-swap technique was also performed in this experiment and proposed as an effective design for direct comparison of two samples (Kerr and Churchill, 2001). The technical replicate with dye-swap technique is also useful for reducing technical variation in which independent biological samples are used. Especially, the dye-swap technique could apply for the different sources of sample which plausible causes the variation of the sample and reduction of the correlation coefficient between data sets. The bias with occurred in technical variation can be minimized by balancing dyes and samples (Kerr and Churchill, 2001). To create the balancing, design of even number of technical replicates from each biological sample and assign equal numbers of these to each dye label (Churchill, 2002) can be performed. Moreover, the dye-swap experiment can control the problem which associated with different corporation rate of fluorescent dyes during labeling (Wei *et al*., 2001).

In each annealing temperature tested, the light/dark samples consisted of four pairs with two biological and two technical replicates. Excluding tested of 46 ^{o}C annealing temperature that contained only two biological and one technical replicates. The scale of replication also required to yield significant data and improved the reliability of array data. However, optimizing the annealing temperature, the 46 ^{o}C

array data set was observed only for basic statistical analysis of signal intensity and influent factor in variation sources to determine the quality of array data.

5.1.2 Statistical analysis of 45K array light/dark experiment

To validate the 45K array, the experiment was performed with light/dark samples similar to the 20K array but hybridized at only 42 ^{o}C was tested. However, four rice seedling cultivars which included nipponbare, kitaake, TP309, and IR24 were used. These samples carried out eight arrays which composed of four biological and four technical replicates. In each rice cultivar designated as one biological and one technical replicates. All these array data sets were combined and analyzed using R program. The basic statistical analysis with mean intensity of cy3 and cy5 represented the variable values among slides that plausible caused by background intensity and incorporation rate of fluorescent dye in the labeling step. The signal intensities were pictured by MA plot. The MA plot showed similar pattern for all 45K array data sets which were similar to the 20K array data as well. The correlation level between two data sets was determined using MM plot which showed an average correlation coefficient in the range of 0.84 to 0.88. At these values, the array data showed high correlation between data sets and reliable to analyze in the further steps. The array data for 45K were divided into two slides due to the high number of oligonucleotide probes (43,767) which were unable to contain in a single slide. These array data sets from two slides were determined and the source of variation showed treatment factor as the highest effect on both array data sets.

The meaningful interpretation of results depends on the appropriateness of data analysis technique. The statistical method provides an improved approach for identifying gene exhibiting changes in expression between conditions. Initial methods for interpretation of changes in expression microarray data used a simple fold-change statistic (DeRisi *et al*., 1996). However, the criterion cut off value at 2-fold changes is not stringent enough at low intensity (Mariani *et al*., 2003). In this study, the FDR (p-value) was also used as criteria to cut off the significant differentially expressed gene list. The FDR cut off value at 0.05 and 0.01 brought up the large number of candidate gene. The condensable candidate gene list was required for further analysis. Therefore, the FDR of 0.001 was used to exclude not high significant candidate genes.

5.2 Biological analysis (20K and 45K array light/dark experiment)

5.2.1 Gene Ontology (GO)

20K array data were analyzed to improve the biological meaning by classifying the candidate gene with GO database. The GO database for 20K and 45K array contained mainly three classes of biological processes, molecular function, and cellular component. Most oligonucleotide probes on the 20K and 45K arrays were classified into GO database for determining the sensitivity of array for specific biological experiment. The applications of GO database are widely used in GO terms and gene product annotation. The GO database facilitated the interpretation of the

results of microarrays data (2006). Classification of GO were applied in array results with the tested annealing temperature of 42^{o}C for both the 20K and 45K.

The GO classification of three main classes showed that the molecular function obtained the highest significant expressed gene portion. The 20K and 45K array data represented similar GO results that consequence by the large portion of oligonucleotide probes belonged to molecular function and less portion in biological process and cellular component. The GO database for 45K was a bit different from the 20K GO database. The 45K GO database showed more details in GO subclass and annotated function of specific oligonucleotide probe with different GO terms.

Both 20K and 45K array data showed highly relative percentage of significant expressed genes in photosynthesis subclass which corresponded to experiment with light as treatment factor. The 20K array result categorized by biological process with cut off 0.1% FDR showed the significant changed in photosynthesis as 68% related to total number of photosynthesis represented oligonucleotide probe on array. And over 20% in protein biosynthesis, generation of precursor metabolites and energy, and nucleobase, nucleoside, nucleotide and nucleic acid metabolism, were represented as differentially expressed genes. As previous knowledge, light can affect genes with products involved in photosynthesis pathway. The synthesis of the photosynthetic machinery includes the structures and enzymes that essential in photosynthesis system. These products enhanced foliage growth and metabolism in the light, as well as the reduced extension growth of the stem.

The GO term corresponded to photosynthesis subclass included chlorophyll *a/b*-binding protein and ferredoxin which compatible with oligonucleotide probes on 20K and 45K arrays. These oligonucleotide probes were denoted as candidate genes with highly confident on FDR value showed up-regulation in light treated condition. The 45K array data belonged to several subclasses of biological process which referred to photosynthesis. Most of subclass showed highly relative percentage of candidate gene with 0.1%FDR. The 20K and 45K array data showed chlorophyll *a/b* binding protein as average $\log_2$ of 4.2. These 20K and 45K array data were also found that the relative $\log_2$ of some ferredoxin genes are down-regulation and some are up-regulation with the ranged of (-1.5) to 5.0, respectively. The general photosynthetic pigments include phytochromes, cryptochromes, phototropins, and chlorophylls. These pigments receive the signal transmitted from light and change to active form in photosynthesis. Light signals are identified mainly by phytochromes and cryptochromes that regulate plant growth and development. The 45K array found that cryptochromes showed up- and down-regulation with $\log_2$ 1.6 and (-0.22) but 20K array do not contain the cryptochromes oligonucleotide probe. Early light-responsive genes included a large proportion of transcription factors of different DNA binding motifs (Casal and Yanovsky, 2005). The 20K and 45K arrays contained 139 and 289 probes corresponded to transcription factors. All of these factors influenced DNA, RNA and protein synthesis even with small or large amount of these products. The products derived from biosynthesis pathways and enzymes, show the significant differential expression between light and dark growing conditions as expected.

Chlorophyll is a complex molecule essential in all photosynthetic organisms. The role of pigments is to absorb light energy, converting it into chemical energy. They are found on the chloroplast membranes and the chloroplasts are arranged in the cells. Their membranes are at right-angles to the light source to maximize light absorption. In most plants, the photosynthetic pigments can be grouped into two classes; the chlorophylls and the carotenoids. In green plants, the primary photosynthetic pigments are chlorophylls *a* and *b*. The chlorophyll is also identified as *a, b, c, d,* and *e* (algae and protistans). The higher land plants and green algae have chlorophyll *a/b*-binding proteins (CAB) as a light-harvesting complex (LHC). The light-harvesting complex proteins chlorophyll *a/c*-binding protein (CAC) and peridinin chlorophyll *a*-binding protein (PCR) from the dinoflagellate (*Alexandrium tamarense*) have been reported the significant changes during light/dark cycle. The expression analysis found that the mRNA levels of CAC were 7-fold higher during the light than during the dark period (Kobiyama *et al*., 2005). Chlorophyll *a/b* (*CAB*) genes of rice (*Oryza sativa*) were reported to be known as circadian clock-controlled genes by light regulation (Sugiyama *et al*., 2001).

In 1996, Millar and Kay reported the *CAB* genes and promoter in *Arabidopsis* were regulated by light. There have been several reported on the mechanism of transcriptional regulation of the CAB gene. The transcription factor binding elements has been identified in several plant species. The transcription factors binding sites were reported as G-box in *Arabidopsis* (Anderson and Kay, 1995), GT-box in *Dunaliella* (Escoubas *et al.*, 1995.), and GATA motif in maize (Shiina *et al*., 1997). The light intensity is also important factor for regulation of LHC gene expression.

The report found that CAB genes transcription in *Dunaliella* were increased after shifting from high light to low-light conditions (Laroche *et al*., 1991). The chlorophylls absorb energy from violet-blue and reddish orange-red wavelengths, and little from the intermediate (green-yellow-orange) wavelengths. All these wavelengths were included in the light treated leaf samples in the array experiment.

Even though, the 20K and 45K array lacked oligonucleotide probe for phototropin. There were reported that these genes are related to light growing condition. As previous reported, a blue light photoreceptor, phototropin 1 and 2 genes (*OsPHOT1*, *OsPHOT2*) in *Oryza sativa* had been characterized as light regulated gene. The report showed both *OsPHOT1* and *OsPHOT2* were significant highly expressed in leaves of etiolated rice seedlings. The transcript levels of *OsPHOT1* and *OsPHOT2* were also regulated by different wavelengths of light. The transcript levels of *OsPHOT2* increased both in coleoptiles and leaves after illumination. Additionally, red and far-red light irradiation also caused down-regulation of *OsPHOT1* mRNA in rice coleoptiles and up-regulation of *OsPHOT2* mRNA both in coleoptiles and leaves (Jain et al., 2007). The multiple phytochromes in red light-induced enhancement of phototropic response were reported to be involved in subsequent irradiation with unilateral blue light (Janoudi *et al*., 1997). In Arabidopsis, blue light irradiation caused not only the down-regulation of transcript levels of *PHOT1* transcript but also the accumulation of its protein product (Sakamoto and Briggs, 2002). In rice, white light caused down-regulation of *OsNPH1a*/*OsPHOT1* and up-regulation of *OsNPH1b*/*OsPHOT2* (Kanegae *et al*., 2000).

The major part of photosynthesis compose of photosystem which identified as photosystem I and photosystem II. Photosystems are arrangements of chlorophyll and other pigments packed into thylakoids. Photosystem I, uses chlorophyll *a*, in the form referred to P700. Photosystem II uses a form of chlorophyll *a* known as P680. Both "active" forms of chlorophyll *a* function in photosynthesis due to their association with proteins in the thylakoid membrane. The 20K and 45K contained 17 and 35 oligonucleotide probes of the photosystem in which almost all showed up-regulation up to $\log_2$ 4.1. For instance, photosystem I reaction centre subunit N showed up-regulation in both 20K and 45K array data with $\log_2$ 3.1 and 3.6, respectively.

Chloroplast composes of thylakoid membrane which is essential organell in photosynthesis. The differential expressed genes classified using GO term that showed the high relative percentage on thylakoid membrane also. The 20K array results showed the cellular component class displayed over 30% of ribosome, peroxisome, and thylakoid groups seized an affected on light/dark growing condition. And also, over 20% of the plastid and mitochondrion groups were affected by the growing condition. *Arabidopsis* also exhibited the global change in gene expression in response to high light (HL) condition. The heat shock protein genes, dehydroascorbate reductase, and ascorbate peroxidase were also found to be increased in expression after exposure to HL condition. The repressed genes included the flowering genes BEL1 and FHA and the chlorophyll synthesis enzyme protochlorophyllide oxidoreductase (Rossel *et al*., 2002).

The GO classification of 20K array data found that over 20% of protein biosynthesis, generation of precursor metabolites and energy, and nucleobase, nucleoside, nucleotide and nucleic acid metabolism, were represented as differential expressed genes. As reported that UDP-glucose pyrophosphorylase is an important cytosolic enzyme producing/utilizing UDP-glucose (UDPG). Though, 20K and 45K array data showed slightly different in expression level of this gene when compared light over dark sample. However, the previous report in *Arabidopsis*, the *Ugp* gene was investigated by taken inorganic phosphate (Pi) status, light/dark and sucrose effects. The reported indicated that the *Ugp* was highly expressed in darkened leaves of *pho1* (P-deficiency mutant). And the daily light exposure enhanced *Ugp* expression both in wild type (wt) and *pho* mutants. (Ciereszko *et al*., 2005). In addition, the light stimulated *Ugp* expression in *Araabidopsis* had been reported as up-regulated regardless the Pi-deficiency (Ciereszko *et al*., 2001a). In leaves, photosynthesis is a potent generator of sugars and osmotica, and the light/dark effects may have direct relevance to the signal transduction pathways. Light generally lead to an increase in starch and soluble sugar contents in leaves. The biosynthesis of metabolites or precursors is essential in plant growing and adapting under stress condition that reasonable in array results with comparing light/dark growing condition. Moreover, the *UGP* expression was found as up-regulated by sucrose and low temperature stress in potato tubers (Spychalla *et al*., 1994) and *Arabidopsis* leaves (Ciereszko *et al*., 2001b). In the light, photosynthetic metabolism leads to accumulation of sugars and starch, while dark conditions lead to starch degradation and glycolytic metabolism of sugars (Kleczkowski, 1994). Interestingly, brasinosteroids had been reported as requirement to maintain the skotomorphogenic

pattern in darkness (Li *et al*., 1996) and brassinosteroid biosynthesis enzymes are down regulated in 6-day-old light-grown compared to dark-grown seedlings.

In addition, histone acetylation is an important component of chromatin structure that affects gene transcription. Histone acetylation effects on chromatin and light-regulated gene expression. Hyperacetylation of histones relaxes chromatin structure and is associated with transcriptional activation. The report showed the role of Arabidopsis GCN5 and HD1 in controlling histone acetylation levels over several light-responsive genes. It was previously found that the histone acetyltransferase (TAF1) is required for light regulation of gene expression. Histone acetyltransferase (GCN5) and histone deacetylase (HD1) are also involved and play roles balance the light regulation of gene expression (Benhamed et al., 2006). In 45K array, 20 corresponded probes for histone deacetylase showed slight changes in expression level similar to the data from 20K (8 histone deacetylase probes). The 20K and 45K array data showed small changes in gene expression for histone deacetylase, the changes in histone acetyltransferase gene were found. These results were supported by the roles of their functions with affected gene transcription which indirectly related to amount of protein working in process.

The GO classification of molecular function on the 20K array found that structural molecule activity and RNA binding represented over 20% of the candidate gene list. The communication between chloroplast and nucleus was reported as important process because proteins in the chloroplast were encoded by both

chloroplast and nuclear genomes, and some important molecules have been identified in the signaling pathway (Strand *et al*., 2003; Surpin *et al*., 2002).

5.3 Confirmation of candidate gene by RT-PCR

The expression level 32 genes selected were confirmed within light and dark samples. The sample was divided into four rice cultivars of nipponbare, kitaake, TP309, and IR24. The mRNAs were used as template for reverse transcription. After preparing cDNAs, the specific primers for 32 genes were used to amplify in both light and dark cDNA samples. The control genes in this step composed of actin and ubiquitin. These two control genes were used to equilibrate the amount of starting transcript in mRNAs samples. The amplified condition was optimized to obtain an equal amount of actin and ubiquitin product from all samples. The amplified products were loaded in agarose gel and performed the electrophoresis for visualization. The samples from light and dark, were placed as light alternating dark for all rice cultivars. From the gel images, most of specific primer sets gave products well matched to the array data with up-regulation in light samples. However, dark sample gave small signal from gel image, even if, expected no signal from dark sample. And also, at low level of up-regulation in light sample displayed similar signal of gel image in both light and dark samples which might cause by crossed hybridization among probes and targets on array. This error could happen with low level expression and scanning problem with artifact area on slide and unequal focuses of cy3 and cy5 laser.

The validate array results with RT-PCR might be error with efficiency of PCR reaction and primers. The preferable primers could achieve the superior result, even though; the intricate primers may give an undesirable result. Therefore, important results obtained with microarray must be confirmed with other techniques such as real-time quantitative PCR or northern blotting.

5.4 Identification of Glycosyl hydrolase family 1 in the 45K array

The 45K array data were analyzed the differential expressed gene of glycosyl hydrolase family 1 (GH1). All significant GH 1 genes are classified in GO database and report that genes contain hydrolase activity and hydrolyze O-glycosyl compounds. These genes also show the cellular component classification in cell wall (http://www.ricearray.org/rice_digital_northern_search.shtml). The rice EST database reports that all significant GH 1 genes derive from shoot library. However, all significantly differential expressed GH 1 genes were found in other tissue libraries, such as 2 week-old leaf, root, callus, panicle, flower, suspension culture, or seedling. Two corresponded of GH 1 oligonucleotides, Os9*bglu*30 showed up-regulation with $\log_2$ of 1.31 and Os7*bglu*26 with $\log_2$ of 2.37. These genes have been reported in rice EST database that could be found in rice 2 week-old leaf. The up-regulated genes are not only found in shoot library but also found in leaf with abiotic and biotic stress (Opassiri *et al*., 2006).

The functional characterization of glycosyll hydrolase family 1 is classified in supported GO database. The Glycosyl hydrolase family 1 (GH1) corresponding the

terms in the GO database (TIGR's Eukaryotic of Gene Ontology Classification), represents the GO subclasses in separated seven GO terms. The rice array database (www.ricearray.org) showed GO terms for glycosyl hydrolase family 1, including GO:0005618 cell wall, GO:0006950 response to stress, GO:0007275 multicellular organismal development, GO:0007582 physiological process, GO:0009607 response to biotic stimulus, GO:0009991 response to extracellular stimulus, and GO:0016787 hydrolase activity.

CHAPTER VI

CONCLUSIONS

6.1 Light/dark experiment with 20K array

The quality assessment of 20K array data from three different annealing temperatures by statistical analysis found that 42 ^{o}C was the optimum annealing temperature. At 42 ^{o}C, the results showed the strongest signal intensity, the highest correlation coefficient value, and treatment factor shows the highest effect on the array data. The treatment factor was also affected in the array data of 46 and 50 ^{o}C annealing temperatures. Signal intensity is the most crucial factor to determine the array quality, reproducibility, and sensitivity, therefore, 42 ^{o}C annealing temperature which had the highest signal intensity was chosen for further works.

To determine the candidate genes, FDR was used for cut off with the p-value less than 0.01 and 0.05. The candidate genes were classified into GO classes and their subclasses. The results showed biological process represented photosynthesis subclass as the highest relative percentage when compared with the other subclasses in this GO class.

6.2 Light/dark experiment with 45K array

Array validation with a 45K array light/dark experiment was performed with four different rice cultivars. The annealing temperature was set up at 42 ^{o}C. The quality assessment with statistical analysis showed similar mean intensities and high correlation coefficient values between data sets. The treatment factor shows the highest effect on the array data from both slides A and B.

The 45K array data were classified with a GO database with a different format from the 20K array. The GO subclasses contained more details on GO terms with various annotations. The identified candidate genes with the FDR cut off less than 0.05 and 0.001 showed high relative percentages on photosynthesis and related photosynthesis subclasses.

6.3 Confirmation of the array result with RT-PCR

The array results from the 20K and 45K arrays were well matched with the RT-PCR results. The control genes in the RT-PCR step consisted of actin and ubiquitin. The pattern of up-regulation in light and down-regulation in dark samples was clear in the gel images, except for some genes. The unclear pattern of gel image was found some genes, which had low expression levels.

The microarray technique is a new and powerful tool to monitor global gene expression in a single experiment. Its applications are widespread across many fields

of plant and animal biological and biomedical research. The array result could suggest further experiments and established the proper gene expression network. Nevertheless, mRNA is only one intermediate between DNA and protein. Post-transcriptional and post-translational controls also play a major role in modulating protein expression. The important array results have to be validated by other methods such as real-time PCR, northern blot, western blot, and SDS-PAGE analysis before designing further experimental steps.

6.4 Identification of glycosyl hydrolase family 1 in the 45K array

The 45K array can be used to study multigene family with example of glycosyl hydrolase family 1. The significant differentially expressed GH 1 genes showed up- and down-regulation in light treated leaf. With 2-fold cut off, only four corresponded GH 1 genes showed down-regulation and other four corresponded GH 1 genes showed up-regulation, in light treated leaf. Also, two corresponded GH 1 genes showed the predicted location in chloroplast with PSORT. The prediction of localization and putative function can relate the GH 1 genes with converting the enzyme and hormone into active forms under light and dark growing condition.

The corresponded GO subclass or term of glycosyl hydrolase family 1 could be used to be the reference of functional identification. However, the GO database is still developed and updated for more benefits not only in functional classification of animals but also in plants.

In conclusion, the 20K and 45K rice UCD rice array was validate in this research. This research work was also able to identify some genes differentially expressed in light and dark conditions.

REFERENCES

1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. **Science**. 282(5396): 2012-8.

2006. The Gene Ontology (GO) project in 2006. **Nucleic Acids Res**. 34(Database issue): D322-6.

Adams, M.D. et al. (2000). The genome sequence of *Drosophila melanogaster*. **Science**. 287(5461): 2185-95.

Anderson, S.L., and Kay, S.A. (1995). Functional dissection of circadian clock- and phytochrome-regulated transcription of the Arabidopsis *CAB*2 gene. Proc Natl **Acad. Sci. USA.** 92(5): 1500-4.

Alonso-Calvo, R., Maojo, V., Billhardt, H., Martin-Sanchez, F.,Garcı´a-Remesal, M., and Pe´rez-Rey,D. (2007). An agent- and ontology-based system for integrating public gene, protein, and disease databases. **J**. **Biomed Inform**. 40(1): 17-29.

Ashburner, M., Ball, C.A., Blake, A.J., Botstein, D., Butler, H., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nat. Genet.** 25(1): 25-9.

Barrett, J.C., and Kawasaki, E.S. (2003). Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. **Drug Discov. Today.** 8(3): 134-41.

Benhamed, M., Bertrand, C., Servet, C., and Zhou, D.X. (2006). Arabidopsis GCN5, HD1, and TAF1/HAF2 interact to regulate histone acetylation required for light-responsive gene expression. **Plant Cell.** 18(11): 2893-903.

Bowtell, D. and Sambrook, J. (2003). **DNA microarrays.** Scatter plots: Diagonal and MA plots. Cold Spring Harbor Laboratory Press, New York, 712 pp.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. **Nat. Genet.** 29(4): 365-71.

Buchholz, W.G., Teng, W., Wallace, D., Ambler, J.R., and Hall, T.C. (1998). Production of transgenic rice (*Oryza sativa* subspecies japonica cv. Taipei 309). **Methods in molecular biology**. 81: 383-396.

Casal, J.J., and Yanovsky, M.J. (2005). Regulation of gene expression by light. **Int. J. De.v Biol.** 49(5-6): 501-11.

Causton, H.C., Quackenbush, J., and Brazma, A. (2003). **Microarray gene expression data analysis.** Blackwell Publishing.

Chatterjee, S., and Price, B. (1991). **Regression analysis by example.** John Wiley & Sons, New York.

Chen, Y., Dougherty, E.R., and Bittner, M.L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. **J. Biomed. Optics.** 2: 364-374.

Chena, G.S., Lub, J.H., Chenga, M.C., Chena, L.F.O., and Loc, P.K. (1994). Four promoters in the rice plastid psbK-psbI-psbD-psbC operon. **Plant Science**. 99(2): 171-182.

Chou, H.H., Hsia, A.P., Mooney, D.L., and Schnable, P.S. (2004). Picky: oligo microarray design for large genomes. **Bioinformatics.** 20(17): 2893-902.

Churchill, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. **Nat. Genet.** 32 Suppl: 490-5.

Ciereszko, I., Johansson, H. and Kleczkowski, L.A. (2001). Sucrose and light regulation of a cold-inducible UDP-glucose pyrophosphorylase gene via a hexokinase-independent and abscisic acid-insensitive pathway in Arabidopsis. **Biochem. J.** 354(Pt 1): 67-72.

Ciereszko, I., Johansson, H. and Kleczkowski, L.A. (2005). Interactive effects of phosphate deficiency, sucrose and light/dark conditions on gene expression of UDP-glucose pyrophosphorylase in Arabidopsis. **J. Plant Physiol.** 162(3): 343-53.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatter plots. **J. Amer. Stat. Asso.** 74: 829-836.

Cleveland, W.S., and Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. **J. Am. Stat. Assoc.** 83: 596–610.

Ciereszko, I., Johansson, H., Hurry, V., and Kleczkowski, L.A. (2001a). Phosphate status affects the gene expression, protein content and enzymatic activity of UDP-glucose pyrophosphorylase in wild-type and pho mutants of Arabidopsis. **Planta.** 212(4): 598-605.

Ciereszko, I., Johansson, H., and Kleczkowski, L.A. (2001b). Sucrose and light regulation of a cold-inducible UDP-glucose pyrophosphorylase gene via a hexokinase-independent and abscisic acid-insensitive pathway in Arabidopsis. **Biochem J.** 354(Pt 1): 67-72.

Ciereszko, I., Johansson, H., and Kleczkowski, L.A. (2005). Interactive effects of phosphate deficiency, sucrose and light/dark conditions on gene expression of UDP-glucose pyrophosphorylase in Arabidopsis. **J. Plant Physiol**. 162(3): 343-53.

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., et al. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. **Nat. Genet.** 14(4): 457-60.

Devos, K.M., and Gale, M.D. (2000). Genome relationships: the grass model in current research. **Plant Cell.** 12(5): 637-46.

Dhingra, A., Khurana, J.P., and Tyagi, A.K. (2004). Involvement of G-proteins, calmodulin and tagetitoxin-sensitive RNA polymerase in light-regulated expression of plastid genes (*psbA, psaA and rbcL*) in rice (*Oryza sativa* L.) **Plant Science.** 166(1): 163-168.

Dudoit, S., Yang, Y.H., Callow, M.J., and Speed, T.P. (2002). Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. **Statistical Sinica.** 12: 111–139.

Escoubas, J.M., Lomas, M., LaRoche, J., and Falkowski, P.G. (1995). Light intensity regulation of cab gene transcription is signaled by the redox state of the plastoquinone pool. **Proc. Natl. Acad. Sci. USA.** 92(22): 10237–10241.

Goff, S.A., Ricke, D., Lan, T. H., Presting, G., Wang, R., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). **Science.** 296(5565): 92-100.

Goffeau, A. et al. (1996). Life with 6000 genes. **Science.** 274(5287): 546, 563-7.

Hedenfalk, I.e.a. (2001). Gene-expression profiles in hereditary breast cancer. **N. Engl. J. Med.** 344: 539–548.

Iwasaki, T., Saito, Y., Harada, E., Kasai, M., and Shoji, K. (1997). Cloning of cDNA encoding the rice 22 kDa protein of Photosystem II (PSII-S) and analysis of light-induced expression of the gene. **Gene**. 185(2): 223-9.

Jain, M., Sharma, P., Tyagi, S.B., Tyagi, A.K., and Khurana, J.P. (2007). Light regulation and differential tissue-specific expression of phototropin homologues from rice (*Oryza sativa* ssp. indica). **Plant Science.** 172: 164-171.

Janoudi, A.K., Gordon, W.R., Wagner, D., Quail, P., and Poff, K.L. (1997). Multiple phytochromes are involved in red-light-induced enhancement of first-positive phototropism in Arabidopsis thaliana. **Plant Physiol.** 113(3): 975–979.

Kanegae, H., Tahir, M., Savazzini, F., Yamamoto, K., Yano, M., et al. (2000). Rice NPH1 homologues, OsNPH1a and OsNPH1b, are differently photoregulated. **Plant Cell Physiol.** 41(4): 415-23.

Kerr, M.K., and Churchill, G.A. (2001). Experimental design for gene expression microarrays. **Biostatistics.** 2(2): 183-201.

Kleczkowski, L.A. (1994). Inhibitors of photosynthetic enzymes/carriers and metabolism. **Ann. Rev. Plant. Mol. Biol.** 45: 339-367.

Knudsen, T.B., and Daston, G.P. (2005). MIAME guidelines. **Reprod. Toxicol.** 19(3): 263.

Kobiyama, A., Yoshida, N., Suzuki, S., Koike, K., and Ogata, T. (2005). Differences in expression patterns of photosynthetic genes in the dinoflagellate *Alexandrium tamarense*. **European Journal of Protistology.** 41: 277–285.

Laroche, J., Mortain-Bertrand, A., and Falkowski, P.G. (1991). Light Intensity-Induced Changes in cab mRNA and Light Harvesting Complex II Apoprotein Levels in the Unicellular Chlorophyte Dunaliella tertiolecta. **Plant Physiol.** 97(1): 147-153.

Li, J., Nagpal, P., Vitart, V., McMorris, T.C., and Chory, J. (1996). A role for brassinosteroids in light-dependent development of Arabidopsis. **Science.** 272(5260): 398-401.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. (1999). High density synthetic oligonucleotide arrays. **Nat. Genet.** 21(Suppl 1): 20-4.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. **Nat. Biotechnol.** 14(13): 1675-80.

Mariani, T.J., Budhraja, V., Mecham, B.H., Gu, C.C., Watson, M.A., and Sadovsky, Y. (2003). A variable fold change threshold determines significance for expression microarrays. **Faseb J.** 17(2): 321-3.

Millar, A.J., and Kay, S.A. (1996). Integration of circadian and phototransduction pathways in the network controlling CAB gene transcription in Arabidopsis. **Proc. Natl. Acad. Sci. USA.** 93(26): 15491-6.

Murray, L.E., akaiwa, F., Goto,F., Yoshihara, T., Theil, E.C., and Beard, J.L. (2002). Transgenic rice is a source of iron for iron-depleted rats. **The Journal of nutrition.** 132: 957-960.

Ohtsubo, H., Umeda, M., and Ohtsubo, E. (1991). Organization of DNA sequences highly repeated in tandem in rice genomes. **Japanese journal of genetics.** 66: 241-254.

Opassiri, R., Pomthong, B., Onkoksoong, T., Akiyama, T., Esen, A., and Ketudat Cairns, J.R. (2006). Analysis of rice glycosyl hydrolase family 1 and expression of Os4*bglu*12 beta-glucosidase. **BMC Plant Biol.** 29 : 6-33.

Pounds, S., and Cheng, C. (2004). Improving false discovery rate estimation. **Bioinformatics.** 20: 1735-1754.

Quackenbush, J. (2002). Microarray data normalization and transformation. **Nat Genet.** 32 Suppl: 496-501.

Rajeevan, M.S., Vernon, S.D., Taysavang, N., and Unger, E.R. (2001). Validationof array-based gene expression profiles by real-time (kinetic) RT-PCR. **J. Mol. Diagnosis.** 3: 26-31.

Rossel, J.B., Wilson, I.W., and Pogson, B.J. (2002). Global changes in gene expression in response to high light in Arabidopsis. **Plant Physiol.** 130(3): 1109-20.

Rost, T.L., Barbour, M.G., Stocking, C.R. and Murphy, T.M. (2006). **Plant Biology**. Thomson Nelson, Toronto.

Sakakibara, H. (2003). Differential response of genes for ferredoxin and ferredoxin:NADP+ oxidoreductase to nitrate and light in maize leaves. **J Plant Physiol**. 160(1): 65-70.

Sakamoto, K., and Briggs, W.R. (2002). Cellular and subcellular localization of phototropin 1. **Plant Cell.** 14(8): 1723-35.

Sanmiya, K., Iwasaki, T., Matsuoka, M., Miyao, M., and Yamamoto, N. (1997). Cloning of a cDNA that encodes farnesyl diphosphate synthase and the blue-light-induced expression of the corresponding gene in the leaves of rice plants. **Biochim Biophys Acta.** 1350(3): 240-6.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. **Science.** 270(5235): 467-70.

Schena, M. (1996). Genome analysis with gene expression microarrays. **Bioessays**. 18(5): 427-31.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., and Davis, R.W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A, 93(20): 10614-9.

Schena, M. (2003). **Microarray analysis.** John Wiley & Sons, Inc., New Jersey.

Shiina, T., Nishii, A., Toyoshima, Y., and Bogorad, L. (1997). Identification of promoter elements involved in the cytosolic Ca(2+)-mediated photoregulation of maize cab-m1 expression. **Plant Physiol.** 115(2): 477-83.

Spychalla, J.P., Scheffler, B.E., Sowokinos, J.R., and Bevan, M.W. (1994). Cloning, antisense RNA inhibition, and the coordinated expression of UDP-glucose pyrophosphorylase with starch biosynthetic genes in potato tubers. **J. Plant Physiol.** 144: 444-453.

Stekel, D. (2003). **Microarray Bioinformatics.** Cambridge university press.

Strand, A., Asami, T., Alonso, J., Ecker, J.R., and Chory, J. (2003). Chloroplast to nucleus communication triggered by accumulation of Mg-protoporphyrinIX. **Nature.** 421(6918): 79-83.

Sugiyama, N., Izawa, T., Oikawa, T., and Shimamoto, K. (2001). Light regulation of circadian clock-controlled gene expression in rice. **Plant J.** 26(6): 607-15.

Surpin, M., Larkin, R.M., and Chory, J. (2002). Signal transduction between the chloroplast and the nucleus. **Plant Cell.** 14 Suppl: S327-38.

Taniguchi, M., Miura, K., Iwao, H., and Yamanaka, S. (2001). Quantitative assessment of DNA microarrays; comparison with Northern blot analyses. **Genomics.** 71: 34-39.

Thomas, P.D., Mi, H., and Lewis, S. (2007). Ontology annotation: mapping genomic regions to biological function. **Curr. Opin. Chem. Biol.** 11: 4-11.

Tyagi, A.K., and Gaur, T. (2003). Light Regulation of Nuclear Photosynthetic Genes in Higher Plants. **Critical Reviews in Plant Sciences**. 22(5): 417-452.

Wei, Y., Lee, J.M., Richmond, C., Blattner, F.R., Rafalski, J.A., and LaRossa, R.A. (2001). High-density microarray-mediated gene expression profiling of Escherichia coli. **J. Bacteriol.** 183(2): 545-56.

Xia, X., McClellan, M., and Wang, Y. (2005). WebArray: an online platform for microarray data analysis. **BMC Bioinformatics.** 6: 306-312.

Yang, Y.H., and Dudoit, S. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. **Nucleic Acids Res**. 30(4): e15.

Yang, M.C., Ruan, Q.G., Yang, J.J., Eckenrode, S., Wu, S., McIndoe, R.A., and She, J.X. (2001). A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. **Physiol Genomics.** 7(1): 45-53.

Yasui, H., Yazawa, S., Yoshimura, A., and Iwata, N. (1996). RFLP mapping of genes for resistance to Japanese green rice leafhopper (Nephotettix cincticeps Uhler) in rice cultivars DV 85 and IR 24. **Breed Sci.** 46(Suppl. 2): 174-174.

Yu, J., Hu, S.,Wang, J., Wong, G.K., Li, S., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). **Science.** 296(5565): 79-92.

Zimmermann, P., Schildknecht, B., Craigon, D., Hernandez, G.M., Gruissem, W., et al. (2006). MIAME/Plant - adding value to plant microarrray experiments. **Plant Methods.** 2: 1-3.

APPENDICES

Appendix A

I Labeling system

1.1 CyScribe Post-Labeling Kit system

The preparation of Cy3 and Cy5 labeled cDNA for microarray hybridization was done with the CyScribe Post-Labeling Kit (GE Healthcare, Bioscience, USA)
All reagents were molecular biology grade and free of contaminating nucleases. Nuclease free water was used for the preparation of solutions.

Components are listed as follows:

- CyScript™ reverse transcriptase
- nucleotide mix
- amino allyl-dUTP
- anchored oligo (dT)
- random nonamers
- 5 ´CyScript reaction buffer
- 0.1 M Dithiothreitol (DTT)
- Cy3-NHS ester and Cy5-NHS ester

Additional solutions;

- 2.5 M NaOH (10 ml)

NaOH	1 g
Water to	10 ml

Sterilize by filtration with a 0.45 micron filter. Store up to 3 months.

- 2 M HEPES free acid (10 ml)

HEPES free acid	4.77 g
Water to	10 ml

Sterilize by filtration with a 0.45 micron filter. Store up to 3 months.

- 0.1 M Sodium Biocarbonate (Coupling)

Sodium Biocarbonate	0.84 g
Water to	100 ml

(Adjust pH to 9 with 1 M NaOH and sterilize by filtration with 0.45 micron filter. Dispense into aliquots and store at -15 ^{o}C to 30 ^{o}C for up to 3 months.)

- 4 M Hydroxylamine hydrochloride (100 ml)

Hydroxylamine hydrochloride (Sigma, H2391)	27.8 g
Water to	100 ml

1.2 SuperScript™ Indirect cDNA Labeling System

Components are listed as following;

- SuperScript™ III Reverse Transcriptase
- 5X First-Strand Buffer
    - 250 mM Tris-HCl (pH 8.3, room temp)
    - 375 mM KCl
    - 15 mM $MgCl_2$,
- 0.1 M Dithiothreitol (DTT)
- dNTP Mix
    - dATP, dGTP, dCTP, and dTTP
    - aminoallyl-modified nucleotide
    - aminohexyl-modified nucleotide
- Anchored Oligo(dT)20 primer (2.5 µg/µl in DEPC-treated water)
- Random hexamer primers (0.5 µg/µl in DEPC-treated water)
- 20 mg/ml Glycogen
- DMSO
- RNaseOUT™
- DEPC-treated Water
- 3 M Sodium Acetate pH 5.2

Additional solutions;

- 1 N NaOH (100 ml)

NaOH	4.0 g
Water to	100 ml

- 1 N HCl (100 ml)

HCl	3.65 g
Water to	100 ml

Appendix B

I Bioconductor (http://bioconductor.org/)

Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data. Bioconductor is primarily based on the R programming language and compatible with other programming languages.

1.1 R software

The main features of Bioconductor project propose that "R and the R package system are the main vehicles for designing and releasing software. R (www.r-project.org) is a widely used open source language and environment for statistical computing and graphics."

1.2 Install BioConductor (http://bioconductor.org/download/)

The construction steps are listed as following

1.2.1 Install R

"Download the most recent version of R from CRAN. The R FAQ and the R Installation and Administration Manual contain detailed instructions for installing R on various platforms (Linux, OS X, and Windows being the main ones). Start the R program. On Windows and OS X, this will usually mean double-clicking on the R application. On Unix-like systems, type "R" at a shell prompt."

1.1.2 Install BioConductor packages

"Right now, the easiest way to install BioConductor packages is using the biocLite.R installation script. Here's how to use it:

In R command window, type the following:

```
>source("http://bioconductor.org/biocLite.R")

>biocLite()
```

With these commands, the packages are installed in R software. List of package is including;

affy	affydata	affyPLM
annaffy	annotate	Biobase
Biostrings	DynDoc	gcrma
genefilter	geneplotter	hgu95av2
limma	marray	matchprobes
multtest	reposTools	ROC
vsn	xtable	

1.1.3 Install LMGene package

LMGene package was designed for analysis the array data of the NSF rice oligonucleotide array platform. The LMGene package is defined as "LMGene Software for Date Transformation and Identification of Differentially Expressed Genes in Gene Expression Arrays." This package is also defined as "LMGene package for analysis of microarray data using a linear model and glog data transformation in the R statistical package."

Source of package locates at ;

"http://bioconductor.org/packages/1.9/bioc/html/LMGene.html"

Author	David Rocke and Geun Cheol Lee.
Maintainer	Geun Cheol Lee

BIBLIOGRAPHY

Jirapa Phetsom was born on March 23rd, 1976 in Suphaburi, Thailand. She graduated a Bachelor degree from the Department of Biotechnology, Faculty of Technology, Khon kaen University in 1997. She received the Master degree from Department of Biotechnology, Khon kaen University, in 1999. She started the Doctoral degree in school of Biotechnology, Institute of Agricultural at Suranaree University of Technology, in 2003. During the doctoral degree studying, she was supported by the university lecturer development program from the Ministry of Education, Thailand.