

## Characterization of the Rice Whole Genome Array

*Sherman Chang*

The customized rice whole genome array contains 50,187 probe sets including 50,119 rice genes, 45 Affymetrix control probe sets, 5 control probe sets and 17 rice control probe sets for cRNA QC purpose designed by TMRI. The rice genes selected in the rice whole genome array are based on gene prediction program against rice myriad V8 contigs database and there are 21,538 rice genes with high evidence, 12,407 genes with some evidence and 16,172 genes with low evidence. Each gene contains 10 perfect-match probes on average and mismatch probes are completely eliminated. The feature size of the rice genome array is 18X18  $\mu\text{m}$ , which is smaller than the first rice genome array (20X20  $\mu\text{m}$ ). In this study, we will determine the basic parameters of the rice whole genome array.

We have designed experiments to answer the following questions:

- How reproducible is the rice whole genome array?
- What the expression level that can be considered as meaningful measurement of the expression data?
- How can one assess the quality of the cRNA preparation?
- What's the detection sensitivity and dynamic range of the 50K rice whole genome array?
- How data generated from the 50K rice whole genome arrays correlate with the 21K rice genome arrays?

### Experiment design

#### *Rice genomic DNA labeling and cRNA synthesis*

3  $\mu\text{g}$  of Arabidopsis genomic DNA was labeled with biotin-ddNTP in the presence of random hexamers using DNA Klenow fragment at 37°C for 2 hours and the labeled DNA was hybridized to the array overnight, the array was scanned after post-hybridization wash.

Rice RNAs were isolated and subjected to cDNA and cRNA synthesis. cRNAs were hybridized to the array in duplicate from various samples and data were used to compute reproducibility of the 50K rice whole genome array.

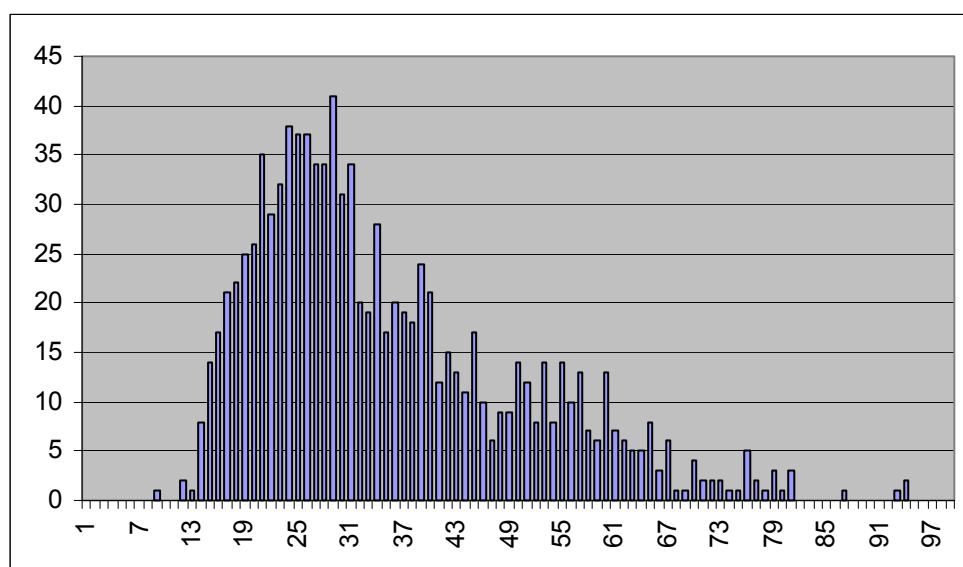
Equal molars of spike controls (including BioB, BioC, BioD and CreX) were mixed and serial dilution of the spike controls were hybridized to the 50K rice whole genome arrays with rice cRNAs. The range of spike controls used was shown below, and after hybridization to the 50K rice whole genome array, data were collected and used to compute the dynamic range of the 50K rice whole genome array.

spike	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Bio-B (pM)	0	0.01	0.02	0.04	0.1	0.1	0.2	0.4	0.8	1	2	4	8	10	20	80	100	200	400
Bio-C (pM)	0	0.01	0.02	0.04	0.1	0.1	0.2	0.4	0.8	1	2	4	8	10	20	80	100	200	400
Bio-D (pM)	0	0.01	0.02	0.04	0.1	0.1	0.2	0.4	0.8	1	2	4	8	10	20	80	100	200	400
Cre (pM)	0	0.01	0.02	0.04	0.1	0.1	0.2	0.4	0.8	1	2	4	8	10	20	80	100	200	400

## Results

### *Background determination for the rice whole genome array*

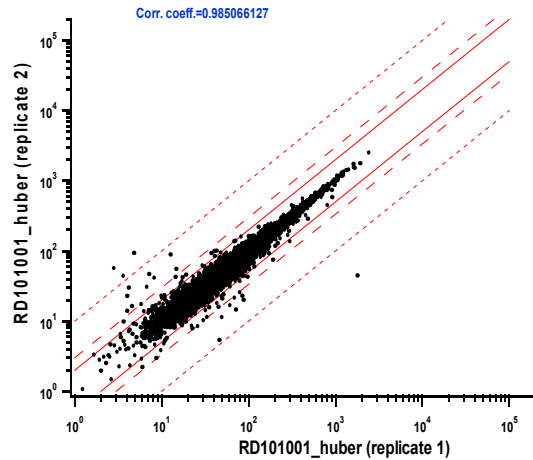
To determine background value for the rice whole genome array, expression levels of 36 Affy-negative control probes in total of 61 samples from RD101 and RD102 were computed. The expression values of Affy-negative control probes ranged from 9 to 94 with mean value of  $34.72 \pm 15.14$  and median value of 30.44. Histogram analysis indicates that expression level between 25 and 31 from Affy-negative controls can be used as background expression level for the rice whole genome array.



### *Reproducibility*

To determine reproducibility of the rice whole genome array, 3 ug of rice genomic DNA was labeled in duplicate and hybridized to the rice genome array, the hybridization signals of the rice genomic DNAs were computed and compared. A false positive is indicated if a probe is scored quantitatively as changing by at least two fold and the relative hybridized signal is greater than 30. There are only 46 out of 50,119 genes shown greater than two fold difference and the false positive rate is 0.092%. In addition, reproducibility was also measured using pool of rice cRNAs samples. 10 replicated hybridization results also indicated the false positive rate for the rice whole genome array was very low with a false positive rate of  $(0.19 \% \pm 0.03)$ . Scatter plots from duplicate samples also indicated that false positive rate was very low.

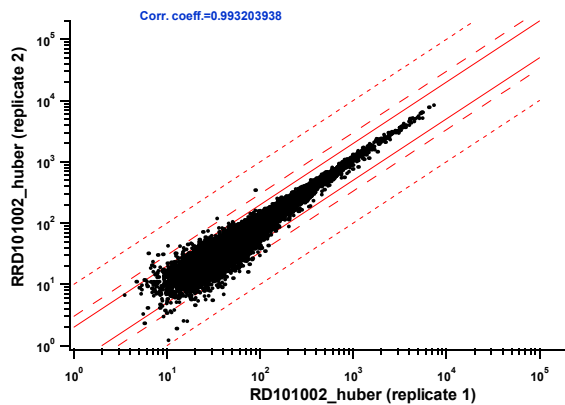
Replicate	1	2	3	4	5	6	7	8	9	10
false positive	107	103	78	86	107	80	102	118	66	108
total rice genes	50119	50119	50119	50119	50119	50119	50119	50119	50119	50119
false positive rate (%)	0.21	0.21	0.16	0.17	0.21	0.16	0.20	0.24	0.13	0.22



Scatter plot of hybridization signals from biotin-labeled rice genomic DNA in duplicate

gene name	probeset	mean	stdev
polyubiquitin (RUBQ1)	Ctrl_AF184279.1-3_at	1	
polyubiquitin (RUBQ1)	Ctrl_AF184279.1-M_s_at	0.88	0.22
polyubiquitin (RUBQ2)	Ctrl_AF184280.1-3_at	1	
polyubiquitin (RUBQ2)	Ctrl_AF184280.1-5_x_at	0.68	0.15
polyubiquitin (RUBQ2)	Ctrl_AF184280.1-M_at	1.72	0.25
cyclophilin 2	Ctrl_L29470.1-3_at	1	
cyclophilin 2	Ctrl_L29470.1-5_at	0.46	0.11
cyclophilin 2	Ctrl_L29470.1-M_at	0.43	0.11
GAPDH	Ctrl_U31676.1-3_at	1	
GAPDH	Ctrl_U31676.1-5_s_at	0.50	0.15
GAPDH	Ctrl_U31676.1-M_s_at	1.40	0.33
actin	Ctrl_X16280.1-3_s_at	1	
actin	Ctrl_X16280.1-5_s_at	0.27	0.07
actin	Ctrl_X16280.1-M_s_at	0.42	0.05

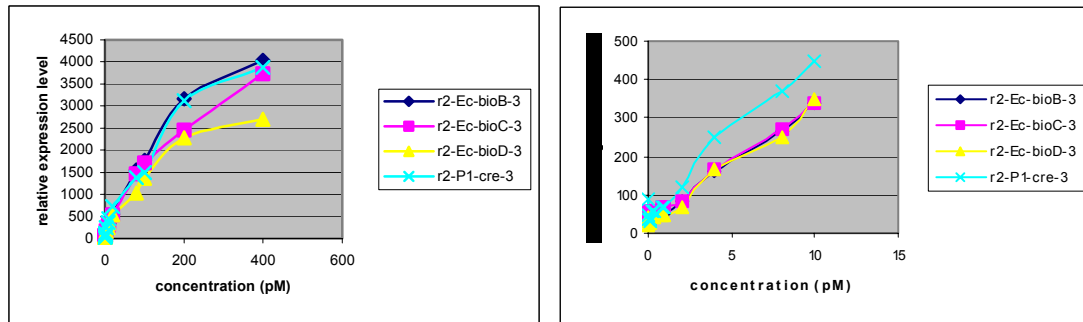
Scatter plot of expression levels from biotin-labeled rice cRNA in duplicate



Positive control probes: There are five positive probes in the rice whole genome array including glyceraldehyde-3-phosphate dehydrogenase (Gpc), actin, cyclophilin 2, polyubiquitin (RUBQ1) and polyubiquitin (RUBQ2) gene. Each gene contains 3', middle and 5' probes that can be used to monitor the quality and integrity of cRNAs. The ratio of 5', middle probe to 3' probe for these positive controls are shown in the table based on data from 36 rice whole genome arrays.

### Dynamic Range and Sensitivity

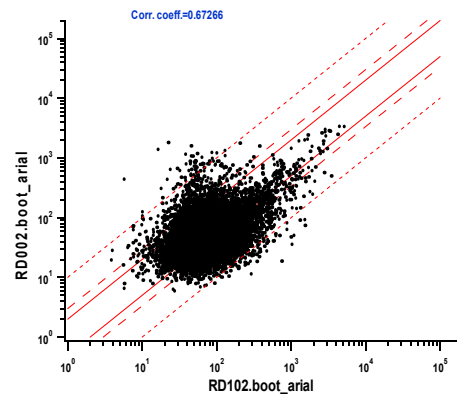
To determine the dynamic range of the 50K rice whole genome array, an equal molar of spike control mixture including BioB, BioC, BioD and CreX was prepared and serial dilution of spike control mixture was mixed with 10 µg rice cRNAs and hybridized to the 50K rice whole genome array. As the concentration of spike increases, so does the relative expression level of the spike and the linear dynamic range is between 0.4 pM and 200 pM (see attached file) and there is 500 fold linearity for the rice 50K whole genome array.



### Correlation between the Two Generations of Rice GeneChip Microarrays

The rice 50K whole genome array is highly reproducible and greater dynamic range; however, how the data generated from the rice 50K whole genome array is comparable to the first rice 21K genome array? 11989 common genes on both the rice 21K and 50K genome arrays were selected and correlation coefficient was calculated. If data from these genome arrays are highly correlated, the correlation coefficient will be close to 1. If they are not correlated then the correlation coefficient will be close to 0. The mean correlation coefficient was  $0.64 \pm 0.08$  among 36 rice samples (see attached file). The correlation coefficient between these two rice genome arrays is much lower than that between the 8K and 26K arabidopsis genome arrays which is 0.85. Rice genes on the rice whole genome array are primarily based on gene prediction software and the accuracy of predicted genes is around 40 to 50%. In addition, selection of probe sets from the same gene for these two rice genome arrays may be different and these might account for low correlation of expression profiles among 11989 common genes between these two rice genome arrays. The rice whole genome array contains genes with high evidence, genes with some evidence and genes with low evidence, so if genes with high evidence are selected, the correlation coefficient between these two rice genome arrays may be improved. Among 11989 common genes on these two rice genome arrays, there are 7334 genes with high evidence of rice genes and correlation coefficient was re-computed based on these rice genes with high evidence. The mean correlation coefficient was  $(0.71 \pm 0.08)$  which is still lower than as expected. These results suggest that data generated from the 21k rice genome array may not be compared directly to those from the 50k rice whole genome array.

	common genes	genes with high evidence
number of genes	11989	7334
corr coeff.	0.64	0.71
stdev	0.08	0.08



### Conclusions

- Data generated from the 50K rice whole genome array is highly reproducible and false positive rate for the array is less than 0.2%.
- Background level for the rice whole genome array is around 30 based on Affy-negative controls.
- The dynamic linear range for the rice whole genome array is approximately 500 fold and the sensitivity of detection for the array is around 0.4 pM based on 4 bacterial spike control genes.
- Correlation coefficient among 11,989 genes between the 21K and 50K rice genome array is 0.64.