

# Characterization of the Tomato Genome Array

Wenyong Xu

*Lycopersicon esculentum* (tomato) is estimated to have 1,005 Mbp of genomic DNA, while Arabidopsis's genome is approximately 172 Mbp. The tomato GeneChip® genome array contains 22821 probe sets including 9 tomato house keeping genes with 19 probe sets, 22714 tomato EST assembly sequences, 43 public tomato sequence not in assembly, 45 Affymetrix control probe sets. In sum, there are 22776 probe sets for tomato genes. Each probe set contains 11 pairs of perfect-match and mismatch probes for cross-hybridization control. The feature size of tomato chip is 18 x 18 µm. Probe sequence selection is based toward the 3'-end of the ORF. Among 22714 tomato EST assembly sequences, 16800 probe sets with description using cutoff e-value as 1.00E-04, 5914 probe sets are with no description. It was estimated that there are 35,000 genes in tomato genome (Vander Hoeven, et al. 2002 Plant Cell). Therefore, it covers approximately 65% of the genes in the genome. The 45 Affymetrix controls are designed to qualify overall hybridization efficiency, monitor the quality of chip manufacture, and provide a built in negative control. 9 probe sets are designed to measure expression of *Lycopersicon esculentum* housekeeping genes such as actin, tubulin, cyclophilin, glyceraldehyde 3-phosphate dehydrogenase (GAPDH), etc. In this study, we will determine the basic parameters of the tomato genome array.

We have designed experiments to answer the following questions:

1. What is background threshold level and what is expression level that can be considered as meaningful measurement of the expression data?
2. How reproducible is the tomato genome array? Biological reproducibility and technique replicate.
3. How can one assess the quality of the cRNA preparation?
4. What's the detection sensitivity and dynamic range of the tomato genome array?

## Experiment design:

The labelling process was started with 1.5µg DNA with biotin-dNTP in the presence of random hexamers using DNA Klenow fragment at 37°C for 2 hours and pool tubes' product together as 3µg and 9µg. The labelled DNA was hybridized to the array overnight. The array was scanned after post-hybridization wash.

Tomato RNAs were isolated and subjected to cDNA and cRNA synthesis. cRNAs were hybridized to the arrays from various samples and data were used to compute reproducibility of the tomato genome array. In order to testing biological reproducibility, sample S1-S5 were pooling replicates from young leaf tissue of 4 leaf-stage tomato Bonnie Best. Each pool consists of 3 leaves from 3 individual plants (fresh weight ~0.27g). Total RNA was extracted using RNAwiz and purified by RNeasy kit. The tomato RNA was subjected to cDNA and cRNA synthesis for RNA hybridization.

Technical replicates of cRNA are prepared for the determination of dynamic range, sensitivity, reproducibility, and background noise using pooled total. Once all cRNAs for technical replicates have been created they are subjected to QC analysis to determine the quality and quantity of preparations. These individual preparations are then pooled and from this pool 16 aliquots of 12.5 µg cRNA are produced for use with the spike dilution experiment. These 16 technical replicates are individually spiked with a cocktail of labelled cRNAs from selected control genes (bioB, bioC, bioD, and cre). This cocktail consists of equimolar concentrations of each control gene's labelled cRNA fragments. A serial dilution of this cocktail across a range of 0 – 1000 pM is used for spiking the technical replicates. A plot of the expression values versus spike concentration will define the saturation level and linear range of resolution. Equal molar of spike controls (including BioB, BioC, BioD and CreX) were mixed and serial dilution of the spike controls were hybridized to the tomato genome arrays with pool of tomato cRNAs. The range of spike controls used was shown below, and after hybridization to the tomato genome array, data were collected and used to compute the dynamic range of the tomato genome array.

Spike	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q
BioB (pM)	1000	500	250	100	80	40	20	8	4	2	1	0.8	0.4	0.2	0.1	0
BioC (pM)	1000	500	250	100	80	40	20	8	4	2	1	0.8	0.4	0.2	0.1	0
BioD (pM)	1000	500	250	100	80	40	20	8	4	2	1	0.8	0.4	0.2	0.1	0
Cre (pM)	1000	500	250	100	80	40	20	8	4	2	1	0.8	0.4	0.2	0.1	0

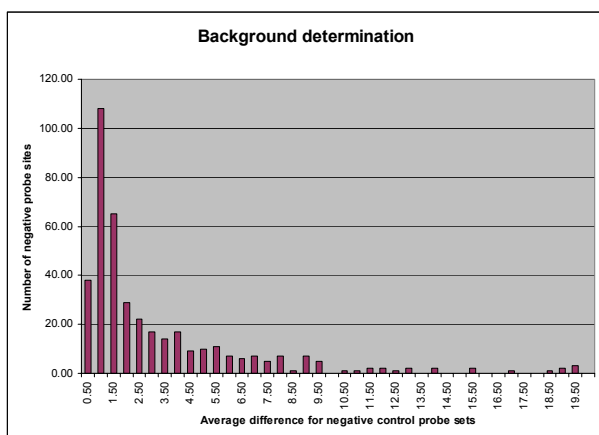
## Results:

### *Gene detection in tomato genome array*

Using Affymetrix Algorithm (Version 5), data from total of 2 genomic hybridization experiments were collected and there were ~6700 out of 22776 probe sets were not detected in these samples and these results indicated that greater than 67% of genes were detected at least once. Using custom algorithm and ~50 as background cutoff, there was 85.5% to 87.6% of genes detected in individual experiment. For the RNA hybridization results, about 70% of genes were detected using Affymetrix Algorithm and about 81% of genes were detected using custom algorithm.

### *Background determination for the tomato genome array*

The question of how to determine background level for the tomato genome array becomes critical in order to determine whether the detection level of the tomato genes is meaningful. To determine background noise, the expression values of the built-in Affymetrix negative control probe sets are examined. A histogram analysis is performed to study the range and distribution of values for the negative control probes. Based on this analysis and calculation of mean and median, the approximate level of background noise can be determined. There are 27 negative control probes present in the tomato chip. Mean expression levels for these negative control genes based on 16 RNA hybridization chips are  $2.90 \pm 3.53$ .



### *Monitor the quality and integrity of cRNAs*

The cRNA yield was compared using the total RNA that was isolated by RNAwiz with or without purification by RNeasy column. With purification by RNeasy column, more than twice of cRNA yield with better OD<sub>260/280</sub> ratio was reached.

Positive control probes: there are 9 tomato housekeeping genes are designed to measure expression of *Lycopersicon esculentum* housekeeping genes such as actin, tubulin, cyclophilin, glyceraldehyde 3-phosphate dehydrogenase (GAPDH), etc. Among them, there are four genes containing 3', middle and 5' probes that can be used to monitor the quality and integrity of cRNAs. Based on 16 RNA hybridization chips, the ratios of 5', middle probe to 3' probe of these four genes were calculated and shown in the table. We studied the relationship between cDNA length and intensity decay. The intensity decayed 30% when the cDNA length from 3'-end is about 600bp, and the intensity decayed 53% when the cDNA length from 3'-end is 1220bp. The longer the cDNA length from 3'-end for each probe is, the less intensity of the probe is.

Gene	Probe set	mean	stdev
actin1	3'	1.00	
	5'	0.41	0.04
	M	0.57	0.06
actin2	3'	1.00	
	5'	0.57	0.04
	M	0.79	0.05
GAPDH (U93208)	3'	1.00	
	5'	0.37	0.04
	M	0.55	0.05
GAPDH (U97257)	3'	1.00	
	5'	0.42	0.05
	M	0.92	0.14

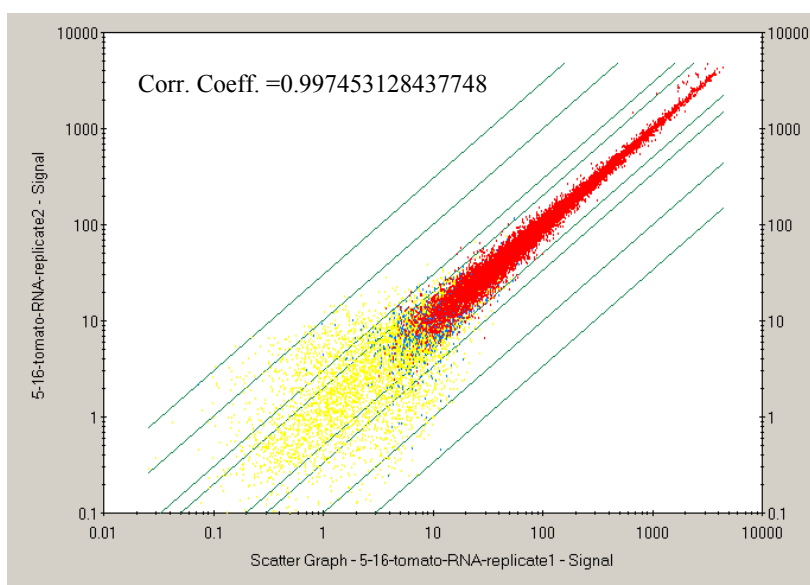
## Reproducibility

Reproducibility is measured by comparison of the replicates. Both technical and biological replicates were examined. The technical replicates allow for the determination of the reproducibility of chip data while the biological replicates provide an estimate of the expected variation due to sampling and processing. Scatter plots of the logarithm of expression values from both biological and technical replicates are used to determine the correlation coefficient, a measure of reproducibility, and the rate of false positives by identification and enumeration of outliers.

To measure reproducibility of expression profiling data, tomato leaf cRNA samples were pooled and 16 technical replicates showed the good correlation coefficient. The average of Corr. Coeff. for 120 pairs' comparisons is 0.993 and C.V. is 0.303%.

A false positive is based on the comparison analysis using the Affetrix Change Algorithm. A false positive is indicated if a probe Change call is "Increase" or "Decrease". 10 randomly picked replicated hybridization results indicated that the false positive rate for tomato arrays was low with a false positive rate of  $(0.3 \pm 0.067)\%$ . Scatter plots from duplicate samples also indicated that false positive rate was very lower.

sample ID	Number of FP	Number of probes	False positive rate (%)
TB001032 vs. TB001033	46	22776	0.2
TB001020 vs. TB001021	82	22776	0.36
TB001020 vs. TB001022	69	22776	0.3
TB001022 vs. TB001023	52	22776	0.23
TB001024 vs. TB001025	86	22776	0.34
TB001020 vs. TB001024	84	22776	0.37
TB001020 vs. TB001025	79	22776	0.35
TB001026 vs. TB001027	53	22776	0.23
TB001030 vs. TB001031	83	22776	0.36

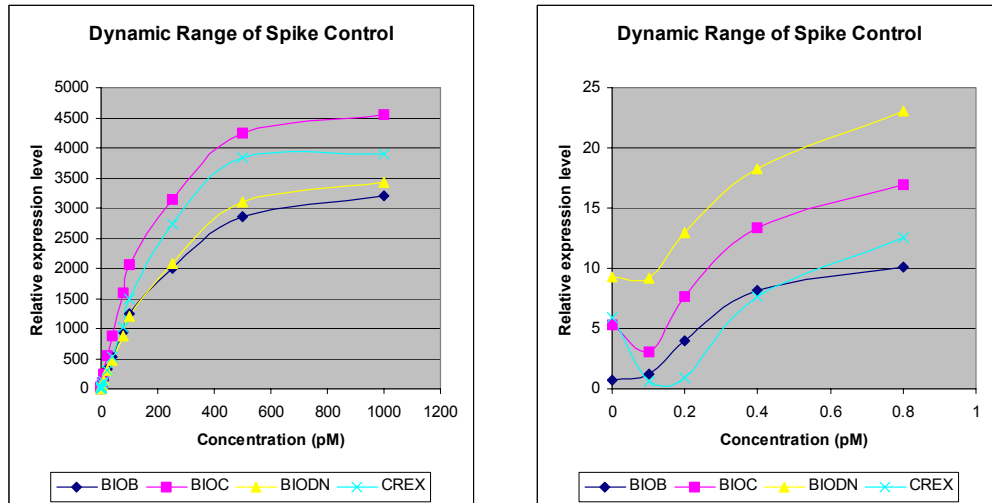


The results of biological replicates indicated that there was very low variation due to sampling and processing. Five samples' hybridization results were scattered and the average of 10 Corr. Coeff. is  $(0.99 \pm 0.0029)$

Corr. Coeff.	TB001015-S1	TB001016-S2	TB001017-S3	TB001018-S4	TB001019-S5
TB001015-S1	1.000	0.990	0.992	0.986	0.993
TB001016-S2	0.990	1.000	0.989	0.984	0.992
TB001017-S3	0.992	0.989	1.000	0.989	0.992
TB001018-S4	0.986	0.984	0.989	1.000	0.986
TB001019-S5	0.993	0.992	0.992	0.986	1.000

### Dynamic Range and Sensitivity

To determine the dynamic range of the tomato 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 12.5 µg tomato cRNAs and hybridized to the tomato genome array.



As shown in the figures, the linear dynamic range is between 0.4 pM and 500 pM (see attached file) and there is greater than 500 fold linearity for the tomato genome array. Using Affymetrix algorithm, the sensitivity of detection for the array is around 0.8 pM based on the 4 bacterial spike control genes.

### In Conclusions

1. Background level for the tomato genome array is less than 20 based on Affy-negative controls on the tomato array.
2. Data generated from tomato genome array is highly reproducible and false positive rate among technical replicates for the array is less than 0.4%.
3. The dynamic linear range for the array is greater than 500 fold and sensitivity of detection for the array is around 0.8 pM based on the 4 bacterial spike control genes using Affymetrix algorithm.
4. Using affymetrix algorithm, there were about 6700 out of 22776 tomato genes not detected by genomic DNA hybridization. This may be related to high intensity of the mismatch probes. However, based on custom algorithm that only analyzes the perfect match probes and cutoff background is about 50, there were more than 85% of genes detected in individual DNA hybridization experiment.
5. Based on the cRNA yield and quality, we recommend the following protocol for total RNA isolation and purification: isolated the RNA using RNAwiz with purification by RNeasy column.
6. Based on the ratio of 5'/3' and 5'/M for four positive control genes, the intensity decreased to 70% when the cDNA length from 3'-end is about 600bp and to 47% when the cDNA length is 1220bp.