

Characterization of the Wheat GeneChip® Array

John McElver

Triticum aestivum (bread wheat) is estimated to have 16,979 Mbp of genomic DNA (Bennett & Leitch), while rice's genome is approximately 420 – 490 Mbp. The gene content of rice is estimated to be 40,000 to 60,000 genes. Since rice can effectively be considered a "1x" cereal genome and bread wheat is known to be hexaploid (three 1x genomes, so called A, B & D parental genomes), the gene number in wheat may be as high as 180,000, although that number is unlikely.

The Wheat GeneChip® has 38,577 probe sets. Therefore, it is not a genome-wide expression chip, rather it is like the early 8k Arabidopsis chip in that it is not based on a sequenced genome and represents only a portion of the transcriptome. There are 13 probes per probe set for all the experimental probe sets (Ta*) and control probe sets.

EST clustering for the Wheat GeneChip® design was done by Alan Evans at Jealott's Hill in late 2002. EST's of all *Triticum* species from Genbank and the legacy Incyte sequences of *Triticum aestivum* cv Savannah were clustered and only contig sequences with at least 2 contributing sequences were submitted for use on the Wheat GeneChip®. According to Evans, the probe sets should be viewed as coming from clusters rather than genes.

I am presently uncertain how wheat full-length cDNA/mRNA sequences from Genbank or EMBL were used in the chip design.

The control gene set itself is somewhat problematic, in that it contains only two (2) genes from wheat with polyadenylated RNAs. Since the labeling of eukaryotic RNA for the Affymetrix GeneChip system utilizes reverse transcription from an oligo-dT primer, the problems presented in selection of probe set sequences from non-polyadenylated transcripts should be obvious.

The other control genes include the 5.8S, 18S, and 28S rRNAs, which, in addition to their non-polyadenylated status, are actually cleaved from a single transcript, thus these six (6) control gene probe sets are only querying one promoter in any case. Four other control probe sets are from snRNAs, which are also non-polyadenylated.

Further, the control probe set mRNA selected for 5'/3' signal comparison is "leaf nonphosphorylating glyceraldehyde-3-phosphate dehydrogenase mRNA". I do not know its level of expression in tissue other than leaves, thus its value as a control is uncertain.

In the main, however, these difficulties will not prevent use of the Wheat GeneChip®, as the signal differences for a given probe set across a set of samples are not dependent on the controls.

Each cluster's probe set contains 13 perfect-match probes and mismatch probes are completely eliminated. The feature size of the Wheat GeneChip® array is 18X18 µm, which is smaller than the first rice genome array (20X20 µm). In this study, we will determine the basic parameters of the Wheat GeneChip®.

We have designed experiments to answer the following questions:

- How reproducible is the wheat 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 Wheat GeneChip® array?

Experiment design

Wheat genomic DNA labeling and cRNA synthesis

Wheat genomic DNA (cv Fielder) was labeled with biotin-dNTP in the presence of random octomers 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. Because the LIMS wasn't functioning, some scans were performed before sample entry in the LIMS. This information is summarized in the chart below:

Please note that the cv Bolero sample isn't in the LIMS.

Sample Number	Sample Name (keep)	time/treatment/genotype other (tissues etc.)	Sample ID	SCAN ID

	<i>short)</i>				
renamed as #1 below in LIMS	fielder1	Genomic DNA	3 ug std protocol	NA	
	bolero1	Genomic DNA	3 ug std protocol	NA	
1	WB001001	Wheat genomic DNA	reproducibility/background (3 ug Fielder DNA)	WB001024	WB001024SYNG 03174
2	WB001002	Wheat genomic DNA	reproducibility/background (3 ug Fielder DNA)	WB001002	WB001002SYNG 03170
3	WB001003	Wheat genomic DNA	Labeling Test (fielder double cut 6 ug)	WB001003	WB001003SYNG 03171
4	WB001004	Wheat genomic DNA	Labeling Test (Fielder genomic DNA, 6 ug labeled in 1.5 ug lots and pooled)	WB001004	WB001004SYNG 03172

Wheat 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 Wheat GeneChip®.

Equal molar amounts of spike controls (including BioB, BioC, BioD and CreX) were mixed and serial dilution of the spike controls were hybridized to the Wheat GeneChip® arrays with wheat cRNAs. The range of spike controls used was shown below, and after hybridization to the Wheat GeneChip®, data were collected and used to compute the dynamic range of the Wheat GeneChip®.

SampleID	Spike Concentration
WB001010	0.00 pM
WB001011	0.01 pM
WB001012	0.02 pM
WB001013	0.04 pM
WB001014	0.1 pM
WB001015	0.2 pM
WB001016	0.4 pM
WB001017	1.0 pM
WB001018	4.0 pM
WB001019	10.0 pM
WB001020	40.0 pM
WB001021	100.0 pM
WB001022	200.0 pM
WB001023	400.0 pM

Results:
Gene detection in Wheat GeneChip® array

Data from total of 4 genomic hybridization experiments were collected and analyzed with the custom algorithm, with an additional hybridization not processed (cv Bolero, see above). The first two hybridizations were performed using the SOPs developed at TMRI; 3 ug of genomic DNA labeled with the BioPrime labeling kit (Invitrogen 18094-011). As these hybridizations were weak, relative to the standard protocol with Rice and Arabidopsis, we decided to test 2 additional labeling protocols.

The net result of these genomic DNA labeling tests indicates the need for optimization of wheat labeling, as it is such a large genome. An additional GeneChip Project will be conducted specifically for this purpose. Present/absent call results from these experiments (for the Ta* probe sets only):

	3 ug Expt 1	3 ug Expt 2	6 ug Cut	6 ug Separate
Present	7638	11332	27812	27407
Marginal	28994	25299	8821	9227
Absent	1932	1933	1931	1930

Additionally, 830 probe sets were called absent in each of the above experiments.

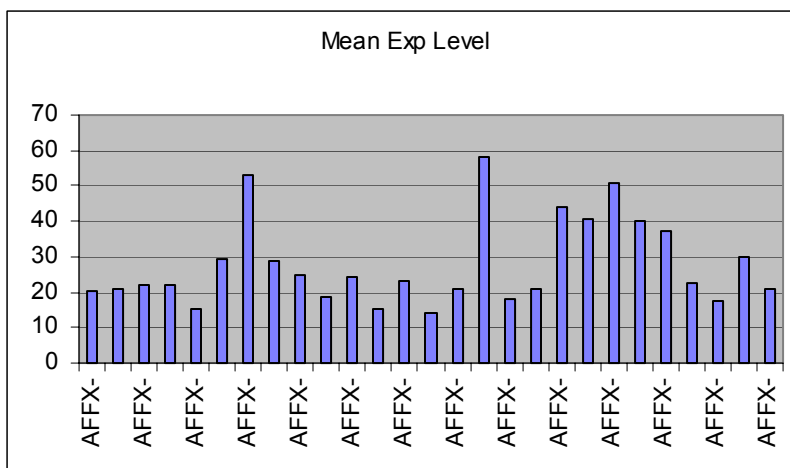
Correlation coefficients among the genomic hybridizations are listed below:

	3 ug Expt 1	3 ug Expt 2	6 ug Cut	6 ug Separate
3 ug Expt 1	1.00	0.95	0.82	0.82
3 ug Expt 2		1.00	0.81	0.78
6 ug Cut			1.00	0.97
6 ug Separate				1.00

As mentioned, these results indicate a new project is required to optimize labeling and hybridization for this large genome, so that the Wheat GeneChip will be useful for SFP and deletion detection.

Background determination for the Wheat GeneChip array

The question of how to determine the background level for the Wheat GeneChip® becomes critical in order to determine whether the detection level of the wheat 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 wheat chip. Mean expression levels for these negative control genes based on 19 RNA hybridization chips are 27.90 with a standard deviation of 13.4.



Monitor the quality and integrity of cRNAs

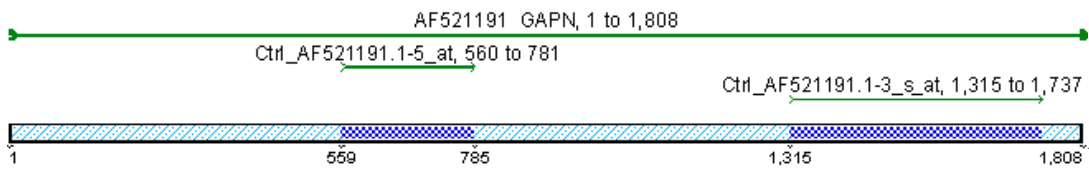
Positive control probes: there are two wheat housekeeping genes are designed to measure expression of *Triticum aestivum* housekeeping. They are AF262983 wheat cyclophilin A2 and AF521191 GAPN. Of these two, GAPN contains 3' and 5' probes that can be used to monitor the quality and integrity of cRNA synthesis. Based on 19 RNA hybridization chips, the ratios of 5' and 3' probe of this gene was calculated and shown below. 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.

Probe Set	Mean Expression Level	Percent of 3'
CTRL_AF521191.1-3_S_AT	1108.96	100.00%
CTRL_AF521191.1-5_AT	40.15	3.62%

These data indicate an appreciable amount of signal decay. However, as the 3' set is a "_s" set; there is probable cross hybridization with other sets on the chip and precise conclusions are impossible.

(NB: As discussed above, there are other control gene sets, specifically the rRNA and snRNA genes. Of these, the 18S and 28S genes do have 5', Middle, and 3' probe sets, however, as any synthesis of cDNA is not primed from a 3' poly-A tail, rather, from the oligo-dT primer's random binding to the rRNA, utilization of these probe sets for determining 3' to 5' synthesis efficacy would be of little value.

Further, the selection of the 5' probe region for the GAPN would be better called a "middle" probe region, as indicted on the contig map below:



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, wheat leaf cRNA samples were pooled and 12 technical replicates showed a good correlation coefficient. The average coefficient of correlation for 91 pairs' comparisons is 0.9718 and C.V. is 0.0298%. Four of these samples were run several days later from the majority of this set. These had coefficients of correlation as low as 0.916 compared to the first day's set. The lowest coefficient of correlation among the first day's set was 0.9885. This represents strong evidence for the group SOP of running all samples on the same day, if possible.

To measure the false positive rate, 10 of the samples run on the same day were analyzed using the settings of the LIMS. A false positive would be a fold change greater than 2-fold up or down compared to the reference sample.

Comparison	Number >2 or <-2 Fold Change	False Positive Rate
WB001010 vs. WB001011	87	0.23%
WB001010 vs. WB001012	74	0.19%
WB001010 vs. WB001013	78	0.20%
WB001010 vs. WB001014	104	0.27%
WB001010 vs. WB001015	96	0.25%
WB001010 vs. WB001016	134	0.35%
WB001010 vs. WB001017	149	0.39%
WB001010 vs. WB001018	118	0.31%
WB001010 vs. WB001019	97	0.25%

Number of Probe sets used 38580

Biological Replicates:

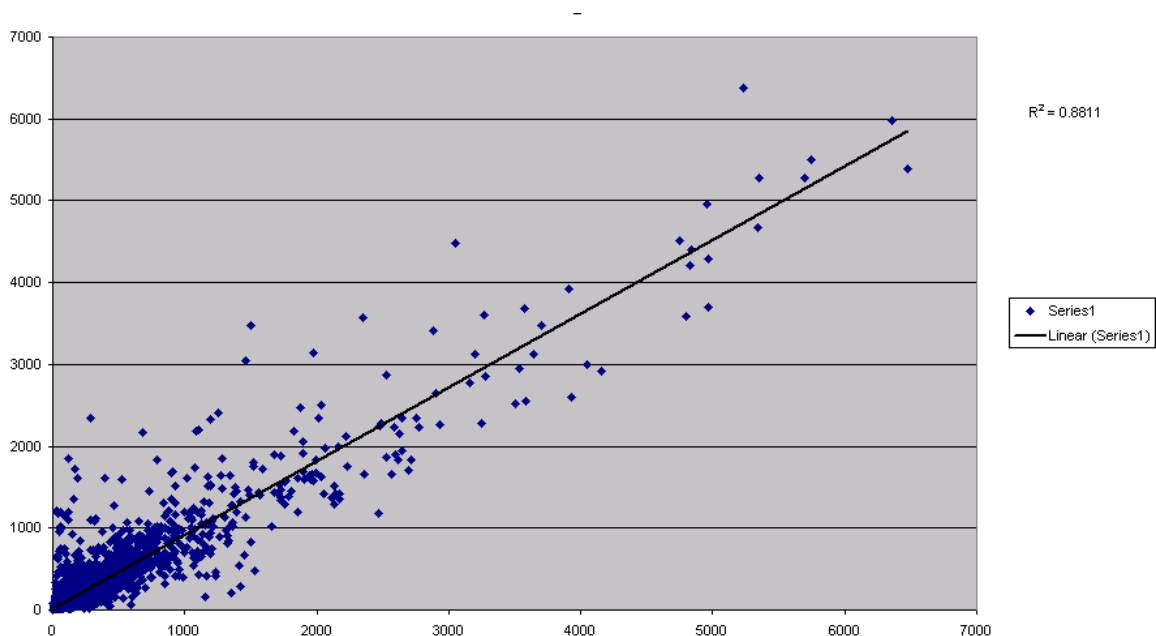
In order to finish the chip characterization as quickly as possible, the biological replicates were not true replicates in the strictest sense. Three cultivars were used, and the some of the samples were taken on separate days. They all are, however, leaf samples. Further, they are indicative of the variation that arises in experimental samples if the strictest control of sampling time and place are not adhered to.

	WB001005	WB001006	WB001007	WB001008	WB001009
WB001005	1	0.93951	0.90771	0.82337	0.85607
WB001006		1	0.97798	0.89334	0.89417
WB001007			1	0.92515	0.91333
WB001008				1	0.92577
WB001009					1

Coefficients of correlation of "biological replicates".

Samples are:

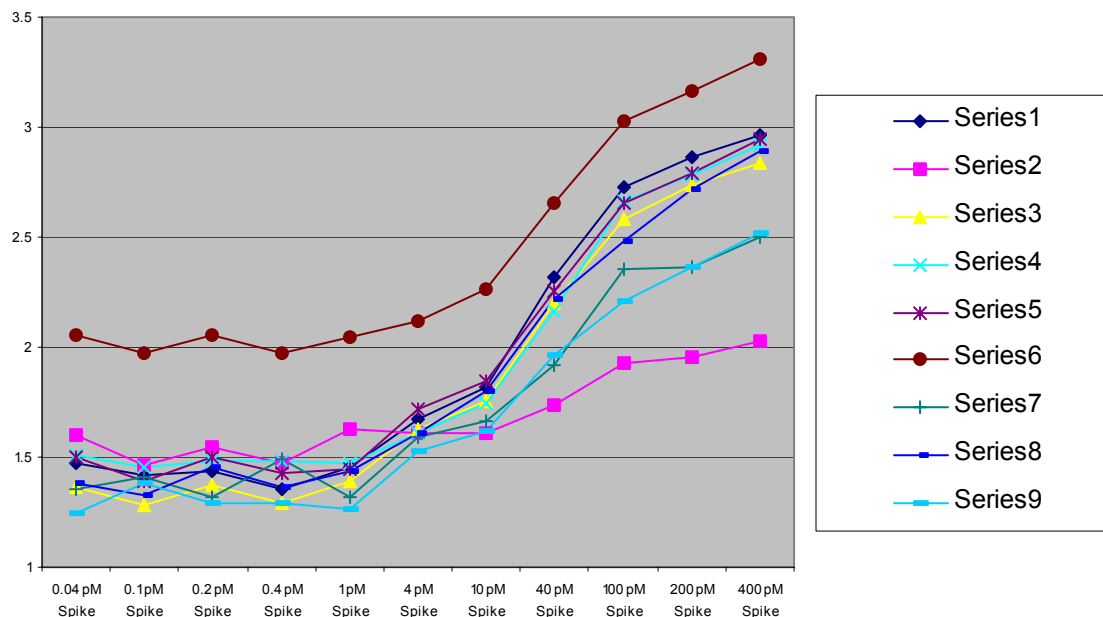
- WB001005 Fielder leaf, 6 weeks old
- WB001006 Savannah, St C5 from Jacky Pallas Day 5
- WB001007 Savannah, St C7 from Jacky Pallas Day 7
- WB001008 Savannah, S C2 from Jacky Pallas Day 2
- WB001009 Hereward, HC1 from Jacky Pallas Day 1



Scatterplot of sample 1 vs Sample 2.

Dynamic Range and Sensitivity

To determine the dynamic range of the Wheat GeneChip 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 μ g wheat cRNAs and hybridized to the wheat array.



As shown in the figures, the linear dynamic range is at least between 0.4 pM and 400 pM and there is greater than 500 fold linearity for the wheat array. Using the MAS algorithm, the sensitivity of detection for the array is around 0.8 pM based on the 4 bacterial spike control genes. These data compare with the Tomato Characterization by Wenying Xu, where the spikes went to at least 1000 pM. Please refer to that characterization document for further discussion of spike.

In Conclusion

1. Background level for the Wheat GeneChip array is less than 40 based on Affy-negative controls on the wheat array.
2. Data generated from Wheat GeneChip 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 the MAS algorithm for perfect match only probe sets, at least 27812 genes are detected when labeled genomic DNA is hybridized to the chip. This should increase when genomic DNA labeling is optimized.
5. Based on the cRNA yield and quality, we recommend the following protocol for total RNA isolation and purification: isolate the RNA using RNAwiz with purification by RNeasy column.

References:

Citation: Bennett MD, Leitch IJ. 2003. Angiosperm DNA C-values database (release 4.0, Jan. 2003) <http://www.rbgekew.org.uk/cval/homepage.html>