

Characterization of the *Phytophthora* GeneChip

By George Aux
September 15, 2002

Phytophthora infestans and *Phytophthora porii*, oomycete pathogens of economic importance, primarily affect members of the Solanaceae family. These organisms are each believed to contain ~18,000 genes. This characterization experiment was conducted to provide researchers a reference dataset and serve as a pilot study to obtain basic information about gene expression patterns and the ability of this microarray to measure transcription.

To measure the genome-wide expression levels this microarray was designed to integrate Affymetrix's basic chip design strategy, Syngenta's proprietary sequence collection and public datasets. Based on sequences derived from ESTs, gene predictions, and literature, 19,324 probe sets were designed for this chip, representing 15,645 *P. infestans* specific sets, 3,407 specific to *P. porii*, and 272 control probe sets.

The majority of the probes sets on this array consist of 13 individual perfect match (PM) probes and the same number of mismatch probes (MM). A single PM and its corresponding MM make a probe pair. Probe pairs are typically designed from a ~1500 bp sequence representing each gene and are biased towards the 3' end. The feature size on this chip size is 18 x 18 μm .

The Affymetrix controls are designed to qualify overall hybridization efficiency, monitor the quality of chip manufacture, and provide a built in positive and negative controls. Syngenta, along with the *Phytophthora* Consortium, generated a group of probe sets used specifically to monitor cDNA synthesis, overall sample and processing quality, and cross hybridization. Of these control probe sets, 38 sets are designed to measure expression of *Phytophthora* housekeeping genes. This group includes probes designed to measure gene expression from genes such as actin, tubulin, cyclophilin, etc. There are 181 sets designed to measure cross hybridization with host organisms, including tomato and potato, as well as other plants, namely *Arabidopsis*.

Methods:

The labeling of RNA followed standard Affymetrix SOP's but uses 5 μg total RNA to synthesize cDNA and biotin labeled cRNA. The cRNA is fragmented by heat and ion-mediated hydrolysis at 94°C for 35 minutes to produce ~100 bp fragments. 12.5 μg of labeled cRNA is used for the hybridization, which occurs at 45°C over 16 hours.

Genomic DNA is biotin labeled using random hexamers and the Klenow fragment incubated at 37°C for 2 hours to produce labeled ssDNA fragments between ~100-200 bp. Labeling uses 3 μg of genomic DNA and the subsequent hybridization to the gene chip uses this entire aliquot of labeled DNA. The staining, washing and scanning procedures applied to the RNA and DNA hybridized chips follow Affymetrix's SOPs.

Technical replicates of cRNA are prepared for the determination of dynamic range, sensitivity, reproducibility, and background noise using pooled total RNA from 2 samples as a template. Once all cRNAs for the replicates have been created they are subjected to QC analysis to determine the quality and quantity of preparations. The

replicates were synthesized from a pooled total RNA sample created by mixing equal masses of each of two biological replicates RNA samples, namely PB001005 and PB001006. The results from the hybridization experiments provide a quantitative measure of cRNA synthesis quality. Ratios and raw expression values indicative of cDNA synthesis efficiency are obtained by comparison of probe sets designed from the 3', middle or 5' regions of each control genes coding. The 14 technical replicates used in this experiment are individually spiked with a cocktail of labeled cRNAs from selected control genes (bioB, bioC, bioD, and cre). This cocktail consists of equimolar concentrations of each control gene's labeled cRNA fragments. A serial dilution of this cocktail across a range of 0 – 1600 pM for each control gene is used for spiking. A plot of the expression values versus spike concentration defines the saturation level and linear range of resolution.

The technical replicates will be examined to determine the reproducibility of chip data while the biological replicates will provide an estimate of the expected variation due to sampling and processing. Scatter plots of the expression values from both biological and technical replicates will be used in conjunction with correlation coefficients to obtain a measure of expected % of false positives, basal levels of cross hybridization signals to plant specific probe sets, and a basic expression profile for *P. infestans*. The built-in Affymetrix negative control probe sets are examined with a histogram analysis to characterize the range and distribution of values. Based on this analysis and calculation of the mean and median, the level of background noise can be estimated.

Sample Name	Sample Description	Parameters Measured
PB001001	<i>P. infestans</i> Genomic DNA	reproducibility background
PB001002	<i>P. infestans</i> Genomic DNA	reproducibility background
PB001003	<i>P. infestans</i> RNA Biological Replicate 1	biological and sampling variation background
PB001004	<i>P. infestans</i> RNA Biological Replicate 2	biological and sampling variation background
PB001005	<i>P. infestans</i> RNA Biological Replicate 3	biological and sampling variation background
PB001006	<i>P. infestans</i> RNA Biological Replicate 4	biological and sampling variation background
PB001008	<i>P. infestans</i> RNA Technical Replicate 1 no spike	Dynamic range sensitivity reproducibility background
PB001009	<i>P. infestans</i> RNA Technical Replicate 2 0.04 pM spike	Dynamic range sensitivity reproducibility background
PB0010010	<i>P. infestans</i> RNA Technical Replicate 3 0.1 pM spike	Dynamic range sensitivity reproducibility background
PB0010011	<i>P. infestans</i> RNA Technical Replicate 4 0.2 pM spike	Dynamic range sensitivity reproducibility background
PB0010012	<i>P. infestans</i> RNA Technical Replicate 5 0.4 pM spike	Dynamic range sensitivity reproducibility background
PB0010013	<i>P. infestans</i> RNA Technical Replicate 6 1.0 pM spike	Dynamic range sensitivity reproducibility background
PB0010014	<i>P. infestans</i> RNA Technical Replicate 7 4.0 pM spike	Dynamic range sensitivity reproducibility background
PB0010015	<i>P. infestans</i> RNA Technical Replicate 8 10.0 pM spike	Dynamic range sensitivity reproducibility background
PB0010016	<i>P. infestans</i> RNA Technical Replicate 9 40.0 pM spike	Dynamic range sensitivity reproducibility background
PB0010017	<i>P. infestans</i> RNA Technical Replicate 10 100 pM spike	Dynamic range sensitivity reproducibility background
PB0010018	<i>P. infestans</i> RNA Technical Replicate 12 200 pM spike	Dynamic range sensitivity reproducibility background
PB0010019	<i>P. infestans</i> RNA Technical Replicate 13 400 pM spike	Dynamic range sensitivity reproducibility background
PB0010020	<i>P. infestans</i> RNA Technical Replicate 14 800 pM spike	Dynamic range sensitivity reproducibility background
PB0010021	<i>P. infestans</i> RNA Technical Replicate 15 1600 pM spike	Dynamic range sensitivity reproducibility background

Table 1

Results and Discussion:

One dataset critical for proper characterization of this array is obtained from the hybridization of biotin labeled genomic DNA. This information will give a researcher an idea of how well the probe sequences represent potential transcripts from genomic sequences unique to the variety or strain being studied. It is believed

that the comparison of these ‘genomic profiles’ may lead to the development techniques that will enable the prediction of sequence variation between samples.

Based on analysis of data from the genomic DNA hybridizations it is clear there is a much larger portion of marginal calls than expected. The overall intensity of signals from the array, analyzed at the probe level, is comparable to other arrays such as the Rice 50K and Arabidopsis 24K. The large number of marginal calls may be related to the fact that the probe sequences don’t accurately represent the genomic sequences from the strain used in the characterization. The variation among the probe sets may be excessive causing an increase in false absent calls. The mismatch (MM) probes can potentially bind labeled DNA fragments from the intended target gene because the MM may be the perfect match for that strain. Because MAS5.1 does not allow for data processing without inclusion of MM probe sets, Syngenta has developed an algorithm to perform this function. Preliminary results of analysis of data from Tomato suggest an improvement in the number of present calls and overall expression values. Ultimately, observation of the probe level figures and consideration of data from both MM and PM probes, will provide needed support for any conclusions drawn from these genomic DNA hybridizations and guide development of future analysis tools.

The biological replicates analyzed in this characterization came from RNA isolated from mycelium of individuals of the same genetic background, grown under the identical environmental conditions. These replicates add value to this study by providing an estimate of the variation in measurements of gene expression due to sampling and pre-synthesis processing. By studying the changes in expression values we can develop an understanding of potential patterns of variation and apply these with other selection criteria to generate the best possible candidate gene lists.

One of the first technical challenges addressed was the importance of the overall quality of RNA isolated from *Phytophthora* specimens. The majority of our initial total RNA samples contained significant amounts of DNA contamination. To maintain consistency in our analytical techniques high quality samples are required.

Two samples of RNA were obtained from tissue of the same genetic background. Ultraviolet spectrophotometric and agarose gel analysis were utilized to determine basic qualitative data and make comparisons. Based on the results of these measurements it is clear that one sample had a higher fraction of DNA, as evidenced by the distinct band on the gel. These two samples were then processed to create cDNA and biotinylated cRNA as per SBI Trait Expression Protocols.

The synthesis reactions were successful and the overall cRNA yields were almost identical for each sample. The results from the hybridizations reflected the apparent lack of impact DNA contamination has on the measurement of expression. The two samples correlate between 95 and 96%, when considering only the probes designed for *P. infestans*. The outliers are quite few, an average of 21 probe sets varied 5 fold or more and had expression values greater than 50 when comparing sample 1 and 2. Syngenta has chosen these two to be used as biological replicates and serve a reference samples for other researchers preparing RNA for hybridization to this array. To generate the highest quality data the samples should be as similar as possible in terms of their overall integrity.

DNA REPLICATES

(Perfect and Mismatch Probes)

Whole Probe Set

	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001001	7951	886	10487	19324
PB001002	7610	914	10800	19324
average	7780.5	900	10643.5	19324
%	40%	5%	55%	100

P.infestans probe sets only

	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001001	7677	811	7157	15645
PB001002	7333	853	7459	15645
average	7505	832	7308	15645
%	48%	5%	47%	100

P.porri probe sets only

	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001001	236	73	3098	3407
PB001002	238	61	3108	3407
average	237	67	3103	3407
%	7%	2%	91%	100

(Perfect Match Probes)

Whole Probe Set

	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001001	16613	1743	968	19324
PB001002	16561	1795	968	19324
average	16587	1769	968	19324
%	86%	9%	5%	100

P.infestans probe sets only

	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001001	14299	1046	300	15645
PB001002	14214	1103	328	15645
average	14256.5	1074.5	314	15645
%	91%	7%	2%	100

P.porri probe sets only

	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001001	2228	665	514	3407
PB001002	2254	665	488	3407
average	2241	665	501	3407
%	66%	20%	15%	100

RNA REPLICATES

(Perfect and Mismatch Probes)

Whole Probe Set

	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001008	16613	1743	968	19324
PB001021	16561	1795	968	19324
average	16587	1769	968	19324
%	86%	9%	5%	19324

P.infestans probe sets only

	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001008	14299	1046	300	15645
PB001021	14214	1103	328	15645
average	14256.5	1074.5	314	15645
%	91%	7%	2%	15645

P.porri probe sets only

	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001008	2228	665	514	3407
PB001021	2254	665	488	3407
average	2241	665	501	3407
%	66%	20%	15%	3407

(Perfect Match Probes)

Whole Probe Set

	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001008	14903	3448	973	19324
PB001021	14474	3881	969	19324
average	14688.5	3664.5	971	19324
%	76%	19%	5%	19324

P.infestans probe sets only

	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001008	12971	2125	549	15645
PB001021	12659	2427	559	15645
average	12815	2276	554	15645
%	82%	15%	4%	15645

P.porri probe sets only

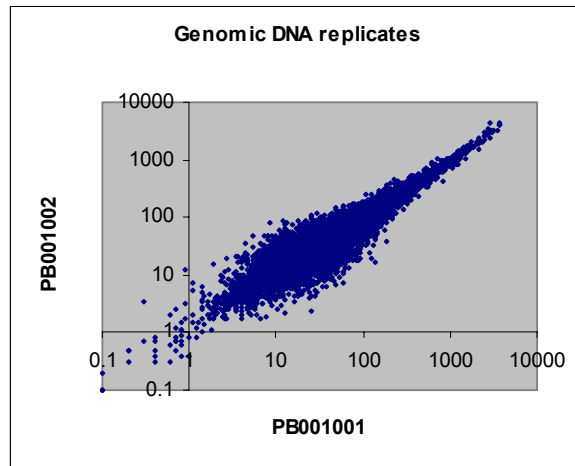
	<i>present</i>	<i>marginal</i>	<i>absent</i>	<i>total</i>
PB001008	1863	1222	322	3407
PB001021	1712	1378	317	3407
average	1787.5	1300	319.5	3407
%	52%	38%	9%	3407

Charts 1 (top) 2 (bottom)

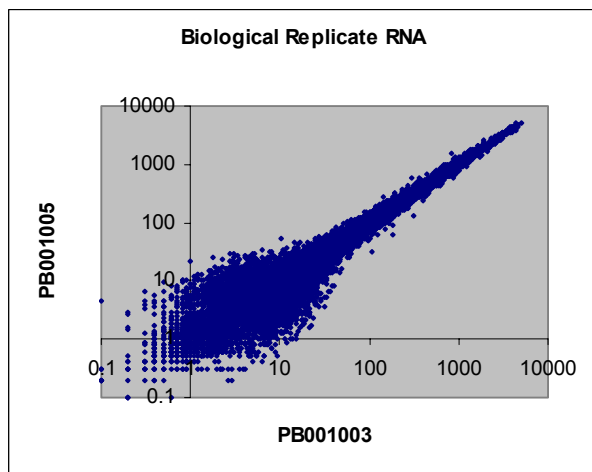
The data from the biological and technical replicates is shown in a series of graphs and charts. The correlation statistics demonstrate that the variation between two chips hybridized using the same sample vary $0.5\% \pm 0.1\%$ (Chart 3) for *P.infestans* probe sets. The scatter plots show the overall distribution of expression values measured and were be used to identify outliers. Although *P.porri* RNA was not used for these experiments the data is included for those probe sets.

The results from the technical replicate experiment show the background level, after processing with Microarray Suite 5.1, was quite low, 2.61 ± 1.02 (Chart 5). This compares with well other GeneChips that contain Mismatch Probes, such as the Tomato GeneChip. Two techniques were utilized to generate these figures. By

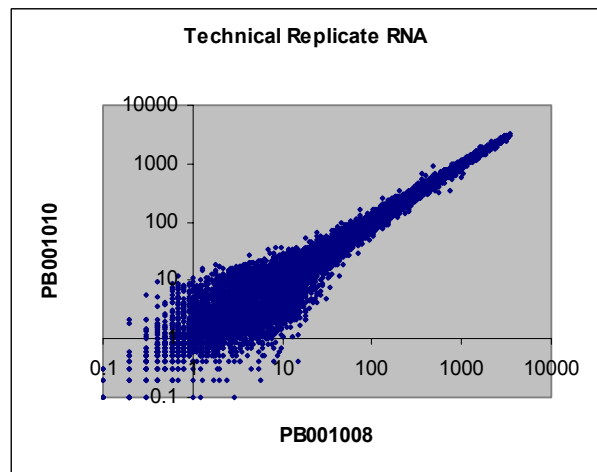
examination of individual probe sets (Graph 4) and the summary data provided through MAS 5.1 we get an accurate measure of the background and variation thereof. A report file was generated with MAS 5.1 for each of the replicates. This report file not only provides excellent qualitative control data also yields a cumulative measurement for each control gene displayed in Graph 5.



Graph 1



Graph 2



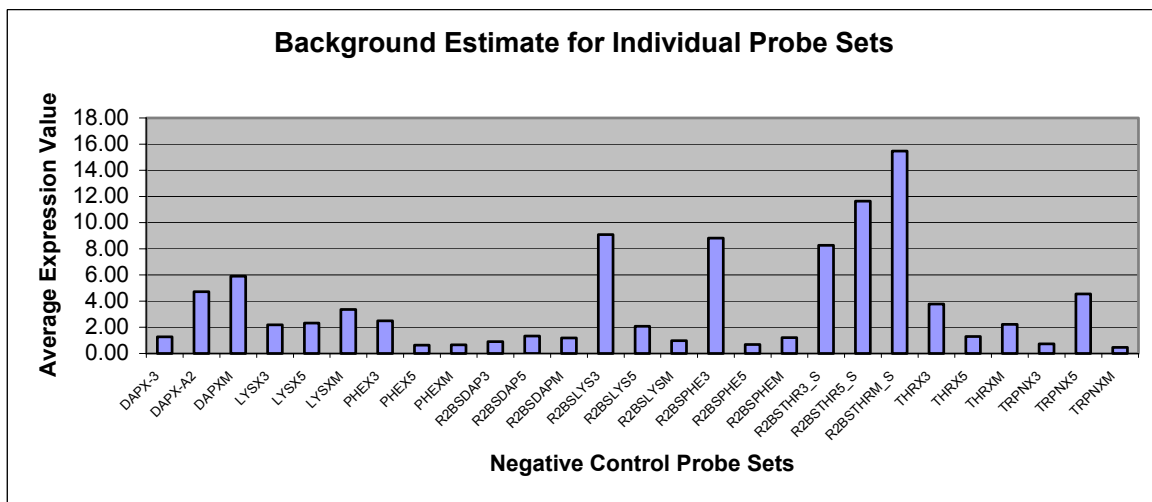
Graph 3

<i>P. infestans</i> Probes	All RNA samples	Technical replicates	Biological replicates
Mean	0.9773	0.9951	0.9855
Std Dev	0.0228	0.0018	0.0072
CV	0.0233	0.0018	0.0073
	PB001004	PB001005	PB001006
PB001003	0.9787	0.9957	0.9837
PB001004	1	0.9782	0.9925
PB001005	X	1	0.9842
PB001006	X	X	1

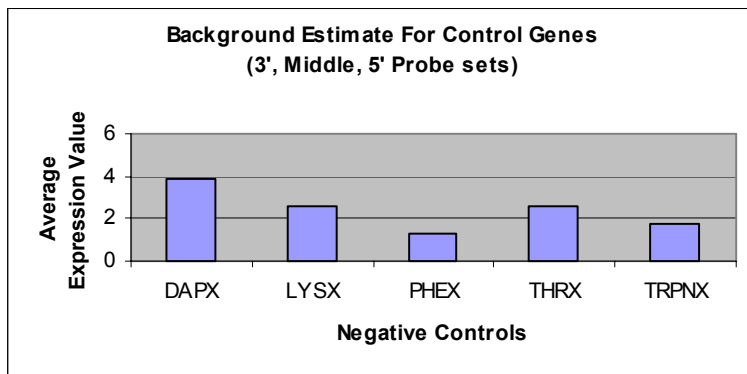
Chart 3

<i>P. porri</i> Probes	All RNA samples	Technical replicates	Biological replicates
Mean	0.9797	0.9884	0.9738
Std Dev	0.0134	0.0069	0.0153
CV	0.0137	0.0070	0.0157
	PB001004	PB001005	PB001006
PB001003	0.9716	0.9849	0.9904
PB001004	1	0.9782	0.9925
PB001005	X	1	0.9894
PB001006	X	X	1

Chart 4



Graph 4



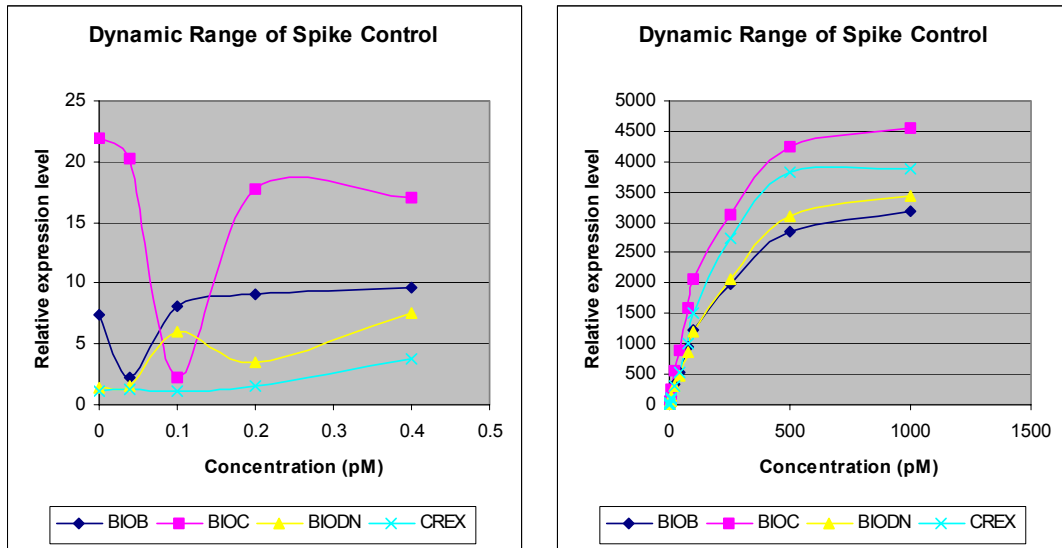
Graph 5

Median Expression Level	2.61
Mean Expression Level	2.43
Standard Deviation	1.02

Chart 5

As shown in Graph 6 the linear dynamic range runs from near 0.4 pM to 400 pM and there is greater than 500 fold linearity for this microarray. Using the

MAS5.1 algorithm, the sensitivity of detection for the array is around 0.4 pM based on the 4 bacterial spike control genes. These data compare well with the Tomato and Wheat Characterizations by Wenying Xu and John McElver respectively. Those documents may be considered as further references.



Graph 6

False positive rate, examined in Chart 6, depicts the expected number of probe sets likely to be found with a significant difference between two identically prepared samples. This experimental data set reflects variation in synthesis as well as hybridization. The significant difference here is determined by MAS5.1 software. Two samples are examined simultaneously, comparing probe sets and determining if a sample has increased or decreased in relation to an arbitrary baseline dataset. Because the samples used in this analysis are all technical replicates, 10 random comparisons are examined in Chart 6 and are used to calculate the false positive rate. The average number of probe sets, which vary significantly from sample to sample, is 127, or $0.67\% \pm 0.47$ of the 19052 *Phytophthora* probe sets.

Baseline Dataset	Experimental Dataset	# of Probe Sets With Significant Difference	False Positive Rate
PB001008	PB001009	122	0.6%
PB001010	PB001011	122	0.6%
PB001012	PB001013	56	0.3%
PB001014	PB001015	282	1.5%
PB001016	PB001017	283	1.5%
PB001018	PB001019	38	0.2%
PB001020	PB001021	58	0.3%
PB001008	PB001014	122	0.6%
PB001018	PB001014	45	0.2%
PB001009	PB001021	142	0.7%

Chart 6

The quality of cDNA synthesis can be measured by examination of data supplied in Chart 7. Shown is the relative expression value as measured and interpreted by MAS5.1 of each probe set designed to measure 5 'housekeeping genes'. Each gene is measured using 3 probe sets designed from the 3', middle or 5' portions of the control genes sequences. The ratio of the values of these probe sets will provide an indication of how well mRNA of given length are utilized to produce cDNA templates during cDNA synthesis. The cDNA synthesis reaction is designed to prime from the 3' end of a transcript. Because this reaction is rarely optimized to produce full length products for each transcript in a total RNA sample, one would expect a degree of preference toward truncated cDNA products for some transcripts. The data from the measurements of the control genes demonstrates that the cDNA product length is indeed correlated positively with lack of ability to produce full length cDNAs. Typically probes for each gene are designed from the 3' end of the representative gene or EST sequence to minimize error or variation in measurement associated with incomplete cDNA synthesis.

Gene Name	Probe Set Name	Mean Expression Level	Standard Deviation	Length of cDNA (bp)	% of 3' value
peptidylprolyl isomerase (cyclophilin)	CTRL_MY-08-C-01-3_S_AT	2855.07	283.38	516	100%
peptidylprolyl isomerase (cyclophilin)	CTRL_MY-08-C-01-M_S_AT	2661.53	244.52	516	89%
peptidylprolyl isomerase (cyclophilin)	CTRL_MY-08-C-01-5_S_AT	2702.12	495.04	516	95%
elongation factor 1	CTRL_E7.928.C3-3_S_AT	1983.71	146.62	955	100%
elongation factor 1	CTRL_E7.928.C3-M_S_AT	1606.29	147.10	955	81%
elongation factor 1	CTRL_E7.928.C3-5_S_AT	1203.30	100.32	955	61%
ribosomal L3	CTRL_E7.579.C1-3_S_AT	3223.44	451.30	1167	100%
ribosomal L3	CTRL_E7.579.C1-M_S_AT	2899.87	261.28	1167	90%
ribosomal L3	CTRL_E7.579.C1-5_S_AT	85.42	13.19	1167	3%
actin A	CTRL_M59715-3_S_AT	3076.38	874.56	1128	100%
actin A	CTRL_M59715-M_S_AT	3036.98	684.96	1128	99%
actin A	CTRL_M59715-5_S_AT	3143.17	492.11	1128	102%
beta tubulin	CTRL_E7.5523.C1-3_S_AT	1890.24	97.13	1551	100%
beta tubulin	CTRL_E7.5523.C1-M_S_AT	939.34	75.02	1551	50%
beta tubulin	CTRL_E7.5523.C1-5_S_AT	859.84	60.34	1551	45%

Chart 7

Conclusions:

In conclusion this study provided a thorough analysis of the potential of this particular microarray. It should be noted that the results were obtained only with hybridizations from one of the two species of *Phytophthora* the GeneChip was designed for. The manufacture of the microarray and integration of control sequences should provide adequate quality assurance to future experiments. The interpretation of data from these experiments, while limited to a mere global characterization, demonstrate that the system is capable a measuring a diverse population of transcripts with a reasonable expectation of reproducibility and accuracy. With continued use, a better understanding of this microarray's utility will unfold.