GLOBAL CORRELATION ANALYSIS BETWEEN REDUNDANT PROBE SETS USING A LARGE COLLECTION OF ARABIDOPSIS ATH1 EXPRESSION PROFILING DATA

Xiangqin Cui[1,3] and Ann Loraine [1,2,4]
[1]*Section on Statistical Genetics, Department of Biostatistics*
[2]*Department of Genetics*
[3]*Department of Medicine*
[4]*corresponding author*
University of Alabama at Birmingham
Birmingham, AL 35294
{aloraine,xcui}@uab.edu

Oligo-based expression microarrays from Affymetrix typically contain thousands of redundant probe sets that match different regions of the same gene. We used linear regression and correlation to survey redundant probe set behavior across nearly 500 quality-screened experiments from the Arabidopsis ATH1 array manufactured by Affymetrix. We found that expression values from redundant probe set pairs were often poorly correlated. Pre-filtering expression results using MAS5.0 "present-absent" calls increased the overall percentage of well-correlated probe sets. However, poor correlation was still observed for a substantial number of probe set pairs. Visual inspection of non-correlated probe sets' target genes suggests that some may be inappropriately merged gene models and represent independently expressed, but neighboring loci. Others may reflect differential regulation of alternative 3-prime end processing. Results are on-line at http://www.transvar.org/exp_cor/analysis.

1. INTRODUCTION

Affymetrix microarrays contain thousands of probes which are grouped into probe sets, collections of probes that (typically) hybridize to 300-500 bp sequence segments near the three prime ends of target transcripts. Due to the high frequency of alternative mRNA processing (splicing and polyadenylation) in many eukaryotic genomes, Affymetrix arrays typically include multiple probe sets that match predicted or known mRNA variants produced by the same gene. Because these probe sets measure different regions (or transcripts) of the same gene, we designate these as "redundant probe sets."

Thanks to new public resources that archive and distribute expression data from hundreds, sometimes thousands, of microarray experiments, it is now possible to survey the behavior of individual probe sets across many different experimental conditions and laboratory settings. The Nottingham Arabidopsis Stock Centre's NASCArrays is perhaps the acme of such services[1]. For a nominal fee, users can subscribe to the NASC AffyWatch service, which delivers quarterly DVDs bearing expression data in the form of array 'CEL' files, which contain numeric, probe intensity data from microarray scans. These CEL files, the majority of which are from the ATH1 microarray[2], are contributed by users who enjoy discounted array processing service from NASC in exchange for donating their data for public use.

Because the ATH1 array is based on a solved genome, data from NASC provide an unprecedented opportunity to investigate the long-range behavior of redundant probe sets. Toward this end, we analyzed co-expression patterns among redundant probe sets using a database that contained data from nearly 500 quality-screened ATH1 array hybridizations.

2. METHODS

We obtained probe set to gene mappings and gene structure annotations (version 6) from the Arabidopsis Information Resource (TAIR)[3]. To simplify the analysis, we purged all probe sets that mapped to multiple genes. Using methods described previously[4], we created an expression database containing quality-screened data from 486 array hybridizations compiled from AffyWatch Release 1.0. Array data were processed using RMA[5], followed by quantile-quantile normalization. We also processed the CEL files using MAS5.0[6] and generated Present, Absent, and Marginal "calls" for each probe set. All array processing was done using the BioConductor software under default settings[7].

We used R to perform linear regression and compute Pearson's correlation coefficient for each pair of redundant probe sets that measure the same gene. Results from these analyses, including scatter plots showing regression results, are posted as Supplementary Files at our Web site http://www.transvar.org/exp_cor/analysis.

We manually inspected gene models using the Integrated Genome Browser (http://genoviz.sourceforge.net) and IGB Quickload site http://www.transvar.org/data/quickload, which serves probe set-to-genome alignments generated by Affymetrix and Arabidopsis gene annotations (versions 5 and 6). To assess cDNA evidence, we used the Sequence Viewer tool at the TAIR Web site.

3. RESULTS

The ATH1 array contains 21,148 probe sets that uniquely map to 20,987 protein-coding genes in the Arabidopsis genome as determined by extensive sequence analysis performed at TAIR. Of these 21,148 probe sets, 309 are redundant probe sets measuring 148 genes (Table 1.) To simplify the analysis, we focused on the 142 genes interrogated by two probe sets each.

We hypothesized that if redundant probe sets measure related molecular entities, i.e., transcripts whose synthesis is driven by the same promoter, they should exhibit a high degree of correlation across a broad range of conditions. To test this, we computed Pearson's correlation coefficient (r) and performed linear regression between each pair of redundant probe sets. Interestingly, we found that many redundant probe sets are not well-correlated (Figure 1A).

Table 1. Breakdown of redundant probe sets per gene on the ATH1 expression microarray

Probe sets per gene	Genes
1	20,839
2	142
3	4
>3	2

It is commonly believed that less than half of the genes in a genome are simultaneously expressed[8,9]. If true, the low degree of correlation between some redundant probe sets may be the result of including low expression readings whose target gene was not actually expressed. The readings from these probe sets might represent random noise and, therefore, exhibit low correlation. To reduce the influence of the probe set readings not derived from bona fide expression of their respective target genes, we ran the "present-absent" call procedure in MAS5.0 for each probe set on each array and eliminated probe set readings called as "Absent" from the analysis. We then re-computed linear regression and Pearson's correlation coefficient for the redundant probe set pairs in which both partner was called as "present" in at least 20 chips. This filtering step removed 55 genes, leaving 87 for further correlation analysis.

This PA filtering followed by correlation analysis generated two notable results. First, we found surprisingly small correspondence in P versus A calls between redundant probe sets (Supplementary File 4). Second, we found that eliminating probe set readings called as "Absent" by MAS5.0 removed many of the genes that were found to be poorly-correlated according in the first (no PA filtering) analysis (Figure 1B).

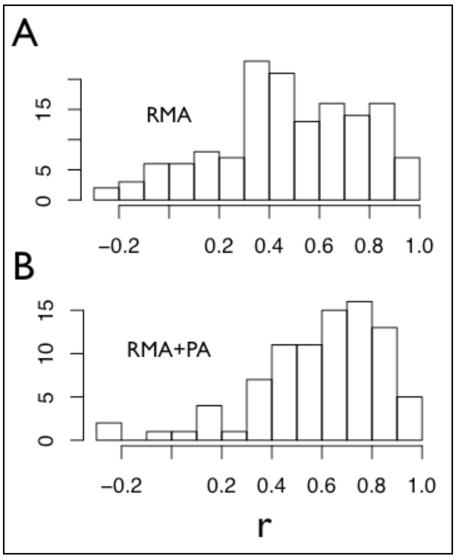

Figure 1. Correlation (r) computed using RMA expression values before (**A**) and after (**B**) PA filtering.

We next explored the possibility that for some of the poorly correlated probe sets, the annotated structure of the target gene may have inappropriately merged adjacent or overlapping genes into a single gene model. If this were true, then we might observe a negative correlation between putative transcript size and expression correspondence between redundant probe sets since inappropriately merged gene models would likely be unusually large.

To test this, we computed Pearson's correlation coefficient comparing average transcript size per gene (log scale) and R-squared from the linear regression, which is the percentage of variation in one probe set that can be explained by variation in the other. (Note that transcript sizes are approximately log-normally distributed; see Supplemental File 2.) We found that there was indeed a weak negative correlation (r = -0.28) between average transcript size per gene and R-squared, suggesting that some fraction of the genes with non-correlated, redundant probe sets might represent gene models that should be split.

Many genes currently included in the Arabidopsis version 6 annotations are based originally on the results of computational analysis and manual curation. For many of these gene models, some additional evidence is needed before they can be accepted as accurate. Currently, the gold standard for assessing the correctness of a gene model is the existence of one or more full-length or partial cDNA sequences that cover the gene region in question. Using the Integrated Genome Browser to visualize probe sets and the TAIR Web site Sequence Viewer to visualize gene structures and cDNA alignments, we investigated cDNA support for genes whose redundant probe sets generated non-correlated expression values.

Of nine genes with non-correlated redundant probe sets (r < 0.3), only one was supported by cDNA evidence covering the entire gene model. However, visualization with the Integrated Genome Browser revealed that the probe sets associated with this gene (AT5G04440) appear to interrogate opposite strands of the chromosome, which explains why expression readings from these two probe sets were uncorrelated. No similarly trivial explanation could explain lack of correspondence (r = 0.09) between the two redundant probe sets interrogating AT4G12640, however. This lack of correspondence suggests that gene model AT4G12640 represents two independent transcriptional units.

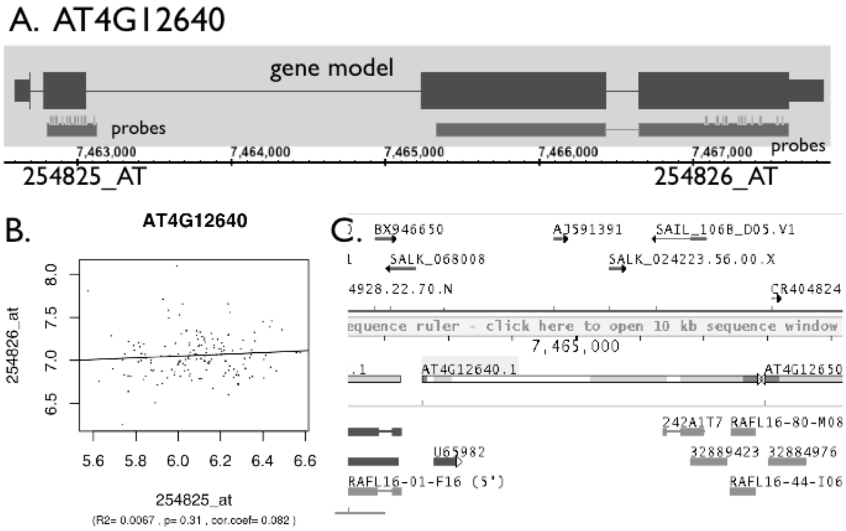

Figure 2. A. Alignment of ATH1 redundant probe sets to Arabidopsis chromosome 4, alongside gene model AT5G04440. Probes appear as vertical bars above rectangles representing the genomic alignment of probe set design sequences, from which the probe sequences were selected. **B.** Scatter diagram showing expression readings from the probe sets in A. **C.** TAIR Sequence Viewer showing lack of full cDNA support for this gene model.

4. DISCUSSION & CONCLUSIONS

We found that large-scale analysis of redundant probe sets reveals a surprising lack of correspondence of expression values between probe sets annotated as interrogating the same gene. Some discordance between redundant probe sets may arise from differential regulation of alternative splicing or polyadenylation. In many cases, however, it is more likely to result from incorrect gene models. We suggest that this lack of correspondence can be used to improve annotation, first as a means of checking probe set to gene mappings (as with AT5G04440) and second as a way to flag gene models that require further validation through cDNA sequencing or other means.

Acknowledgments

We thank Tapan Mehta and Vinodh Srinivasasainagendra for superb programming support. Funds from NSF grant 0217651 (PI David Allison) supported this work.

References

1. Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S. NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res* 2004; **32**: D575-577.
2. Redman JC, Haas BJ, Tanimoto G, Town CD. Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. *Plant J* 2004; **38**: 545-561.
3. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res* 2003; **31**: 224-228.
4. Persson S, Wei H, Milne J, Page GP, Somerville CR. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc Natl Acad Sci U S A* 2005; **102**: 8633-8638.
5. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003; **4**: 249-264.
6. Hubbell E, Liu WM, Mei R. Robust estimators for expression analysis. *Bioinformatics* 2002; **18**: 1585-1592.
7. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004; **5**: R80.
8. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 2004; **101**: 6062-6067.
9. Jongeneel CV, Delorenzi M, Iseli C, Zhou D, Haudenschild CD, Khrebtukova I, Kuznetsov D, Stevenson BJ, Strausberg RL, Simpson AJ, Vasicek TJ. An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res* 2005; **15**: 1007-1014.