

Gene Signal Estimates from Exon Arrays

I. Introduction:

With exon arrays like the GeneChip® Human Exon 1.0 ST Array, researchers can examine the transcriptional profile of an entire gene (Figure 1). Being able to gather data for each individual exon enables the investigation of phenomena such as alternative splicing, alternative promoter usage, and alternative termination while also providing more probe level data to determine the overall rate of expression a particular locus. The calculation of a gene level signal estimate is of particular interest for many analysis as it provides a familiar starting point for the analysis of expression data. Generating gene level signal estimates from exon array data is the focus of this whitepaper. Throughout the rest of this article, the phrase “transcriptional locus”, or “locus”, will describe a genome location that is transcribed rather than using the term “gene” which often implies protein-coding transcripts.

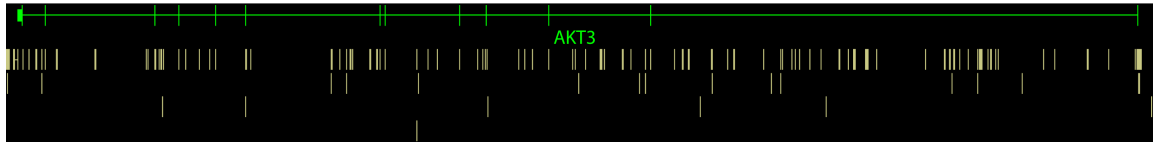


Figure 1: Akt3 RefSeq Transcripts (green) and Probe Sets (gold). Note that many of the probe sets are more exploratory in nature, being outside of regions that are annotated by RefSeq as being transcribed.

The GeneChip® Human Exon 1.0 ST Array is inclusive by design. In addition to containing probe sets that interrogate exons of RefSeq (Pruitt and Maglott 2001) genes, mRNAs and ESTs from GenBank (Benson et al. 1999), it also contains probe sets for exons predicted by *ab-initio* gene finders such as GENSCAN (Burge and Karlin 1997), TWINSKAN (Korf et al. 2001), geneid (Parra et al. 2000), and even GENSCAN Suboptimal exon predictions. Having so much content on the array allows the detection of novel alternative events, but presents a new challenge. This is because including alternative exons can negatively affect the estimate for the entire transcriptional unit. Additionally, much of the content is exploratory in nature, such as the GENSCAN suboptimal predictions, and by nature will have a lower probability of being actually transcribed compared with other well-annotated sequences. Having such a rich set of possible probe sets for each locus provides both the benefit of more data and the challenge of finding the best probe sets to use for gene-level signal estimation in a given experimental study.

The first step in estimates signal for a particular locus is to determine the boundaries for each locus, or gene, and which probe sets are contained within the boundaries. It is important to avoid mistakenly joining two loci, or genes, that are transcribed separately or splitting a single locus in two or more loci.

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

Affymetrix has used genomic annotations such as RefSeq, cDNAs, and gene predictions to create discrete locus definitions (Wheeler 2005). Briefly, annotations that overlap exons on the same genomic strand were iteratively merged into a larger locus annotation. The probe sets on the GeneChip Exon Array were then assigned to these locus annotations, thus partitioning them into distinct transcriptional units, corresponding in many cases to genes.

Different, sometimes nested, loci definitions were created for different levels of annotation confidence so that research can clearly identify more exploratory content from well annotated content.

- High confidence annotations such as RefSeq transcripts and full-length mRNAs make up the “core” loci.
- The core annotations in addition with all other cDNA based annotations make up the “extended” loci.
- The “full” loci are the extended loci plus the ab-initio gene predictions.
- There are many single exon annotations that do not overlap the exons of other multi-exon annotations, but are contained in the intron of a larger multi-exon annotation. In some cases, such as a suspected alternative exon, it may be desirable to associate these single exon annotations with the locus they are nested within. This is referred to as being “bounded” by an annotation.

Users may select what confidence level of loci definitions to use and also exclude probe sets associated with a particular annotation. For example, the exploratory GENSCAN Suboptimal probe sets will not be appropriate for studies examining transcription changes in well-annotated genes.

II. Uncertainty in Loci Definition and its Effect on PLIER signal estimates:

To explore the effect of uncertainty of which probe sets to use for estimating the signal of a particular locus using PLIER (Hubbell 2005) signal estimates, we performed the following experiments:

- Addition of decoy probe sets likely to have low signal to simulate the inclusion of speculative content that is rarely, if ever, transcribed.
- Addition of decoy probe sets likely to have high signals that do not correlate with the original locus to simulate the effect of mistakenly adding probe sets from a transcriptionally distinct locus.
- Effect of using probe sets from more speculative annotations.
- Effect of adding probe sets likely to have low signals on the signal estimates of spike-in data sets.

II.A. Addition of Decoy Probe Sets:

Rather than generating simulated data we will sample from the GENSCAN suboptimal only probe sets as a proxy for low signal probe sets (mean PLIER estimated signal across all GeneChips: 76 +/- 574) and the full length mRNA

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

probe sets for high signal probe sets (mean PLIER estimated signal across all GeneChips: 339 +/- 2,654). To facilitate the study of adding low- and high-signal decoy probe sets, we identified a set of probe sets that appeared to be constitutive for 1674 loci, with 4-11 constitutive probe sets in each locus, on chromosomes 1, 10, 11, 12, and 22. These constitutive probe sets were used as a gold standard to generate PLIER signal estimates for those loci. For this experiment, we used a set 48 GeneChip® Human Exon 1.0 ST Arrays hybridized to 16 different tissues with 3 technical (assay) replicates for each tissue. Tissue replicates were quantile normalized and then the different tissues were median normalized to each other. For both the low- and high-signal decoy probe sets, we then generated pair wise correlations for each PLIER estimate of the test case vs the PLIER estimate for the gold standard set (Figure 2, Figure 3).

PLIER appeared relatively robust to the addition of limited amounts of low-signal decoy probe sets (Figure 2), but was more affected by the addition of higher-signal decoy probe sets (Figure 3). It should be noted that the addition of low-signal decoy probe sets will also reduce the absolute values of the PLIER estimates. As feature response of probes is constrained to have an average of 1, the low-signal probe sets make the other probe sets look like they are more sensitive to the presence of the target in comparison and thus reduce the signal estimate (see Hubbell 2005 for discussion of constraints (Hubbell 2005)).

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

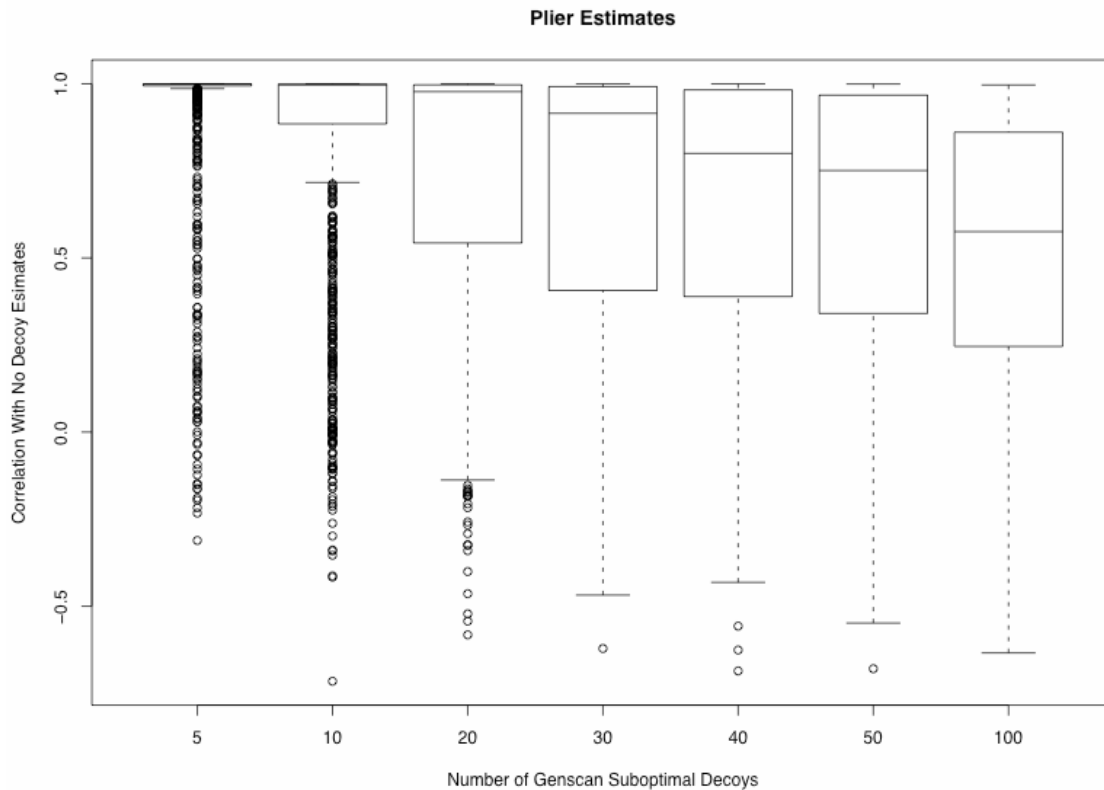


Figure 2: Correlation with original estimates as low-signal decoy probe sets were added. As more low-signal GENSCAN Suboptimal decoy probe sets were added to the gold standard probe sets, the correlation with the original estimates decreases. However, the robustness of PLIER is evident as even when adding 10 unrelated probesets the correlation remains high.

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

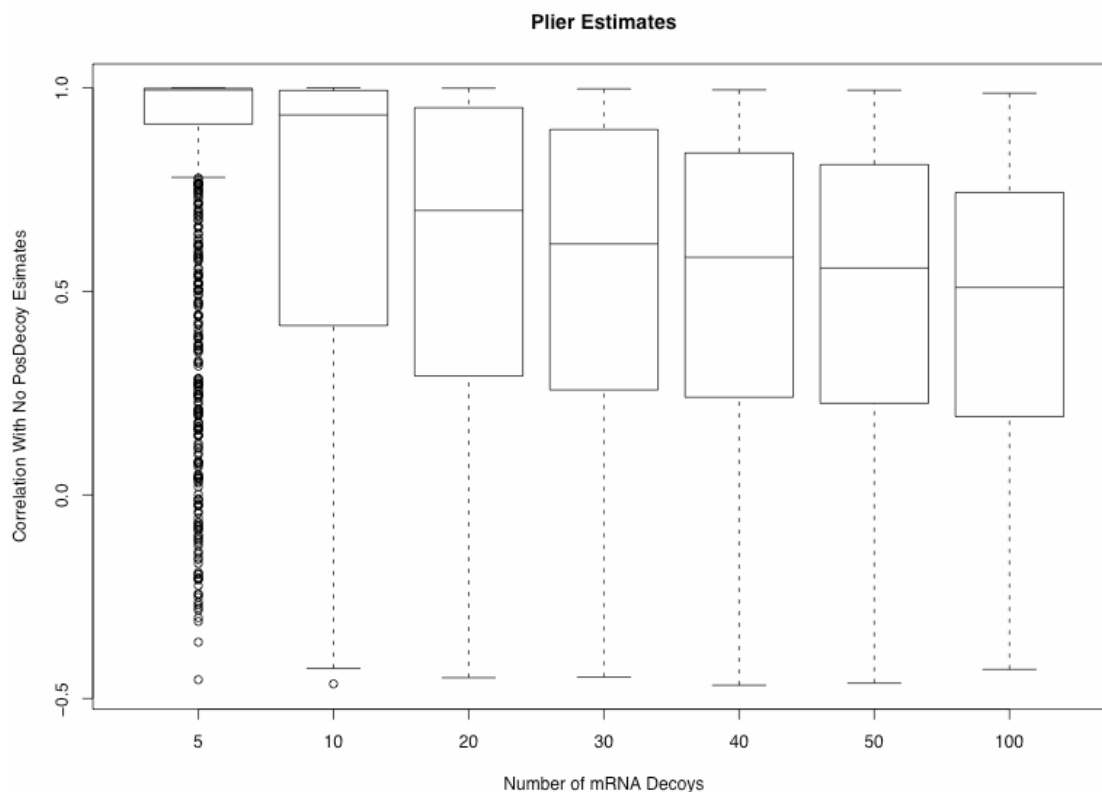


Figure 3: Correlation with original Plier estimates as high-signal decoy probe sets were added. Including decoy probe sets from full length mRNAs decreases the correlation with the original estimates. PLIER is not as resistant to higher-signal decoy probe sets as it is to low-signal decoy probe sets (Figure 2).

II.B. Using More Speculative Annotations:

In addition to adding artificial decoy sets to the gold standard set described above, we also observed the effect of adding probe sets resulting from annotations that are more exploratory in nature for a particular locus (Figure 4). The annotation file provided with Human Exon array details the annotations associated with each probe set. Using just the probe sets resulting from full-length mRNA and RefSeq exons results in loci expression estimates that are very highly correlated.

Expression estimates from extended annotations, those resulting from any cDNA, are also highly correlated with the gold standard set. Expression estimates from full annotations, those resulting from a gene prediction program, were also highly correlated, but not as much as the cDNA based annotations.

Including probe sets that were bounded by a gene (located in an intron, but not overlapping an exon) resulted in the lowest rate of correlation. The effect on a particular gene Akt3 (Figure 1) can be seen in the form of both lower PLIER

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

estimates and lower correlation with the constitutive set across 48 samples (Figure 5).

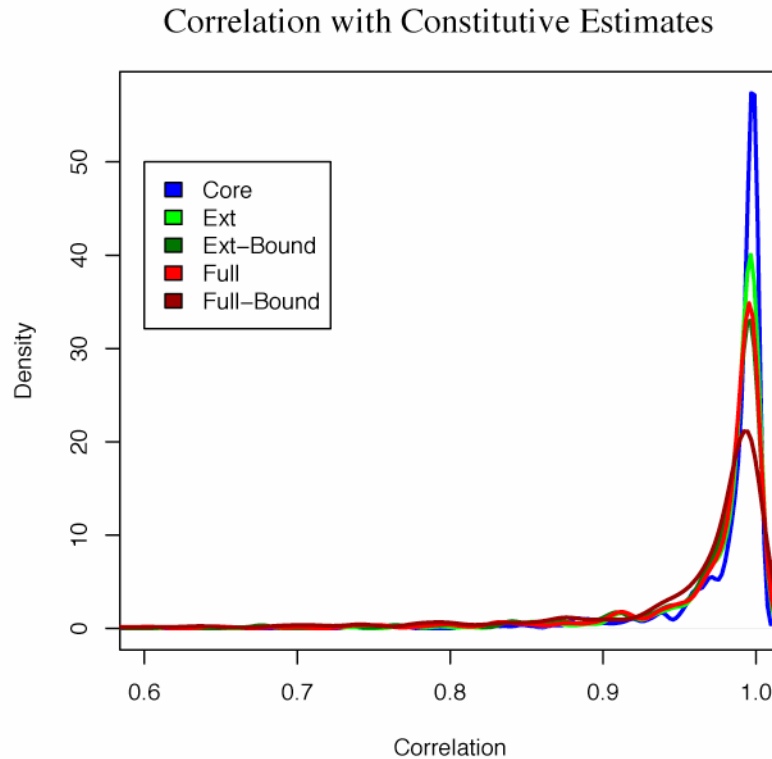


Figure 4: Effect of Adding Lower Confidence Probe Sets on Correlation with Constitutive Set. As probe sets with more speculative annotations are included in the loci expression estimates, the correlation with a bioinformatically enriched constitutive set becomes worse as illustrated above with the above empirically estimated probability density plot (similar to a smoothed-out histogram). Core probe sets are those based on RefSeqs and full-length mRNAs, extended probe sets use all cDNA transcripts and full probe sets combine all annotations including predicted exons. Ext-Bound and Full-Bound sets include probe sets that are contained within the intron of a locus on the same strand, but do not overlap an exon. For this plot 610 loci on chromosome 1 were used.

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

Locus Estimate: Akt3

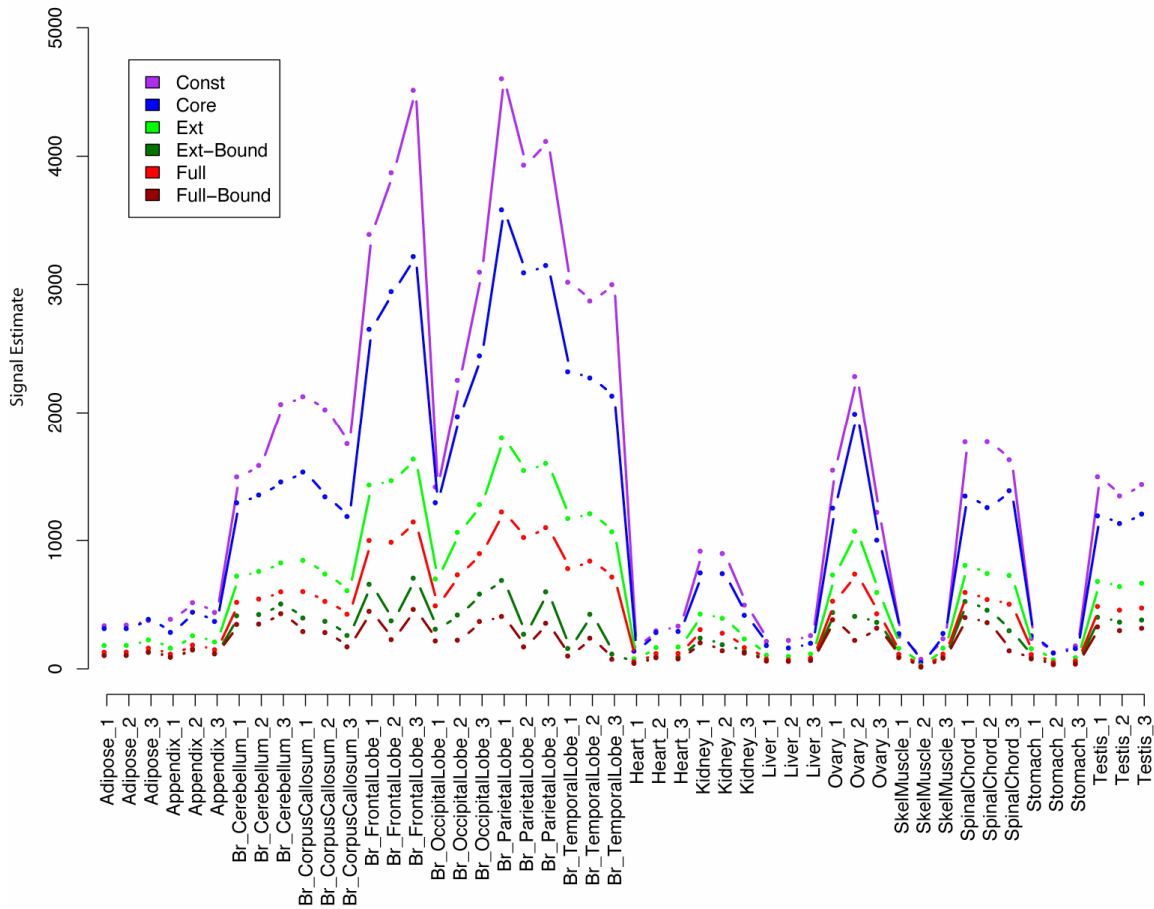


Figure 5: Intensity Estimates for Akt3 Different Annotation Qualities. As more exploratory probe sets are added to the transcriptional locus estimates, the overall correlation is still high, but the range of PLIER estimates becomes compressed.

II.C. Effect of Decoy Sets On Differentiating Known mRNA Concentrations:

We also investigated the effect of adding low-signal probe sets on the ability to differentiate between spiked-in concentrations of mRNAs. For this experiment, a latin square design was used with 4 groups of spike-in RNAs at 4 different concentrations of 0, 0.18, 0.36, and 0.71 pM with 3 replicates each. A receiver operator curve (ROC) was calculated with the ability to differentiate between spike-ins that are present in different concentrations and those present at the same concentrations (Figure 6). The addition of low-signal decoy probe sets decreases the ability of the PLIER expression estimates to differentiate between spike-in concentration pools (Figure 7).

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

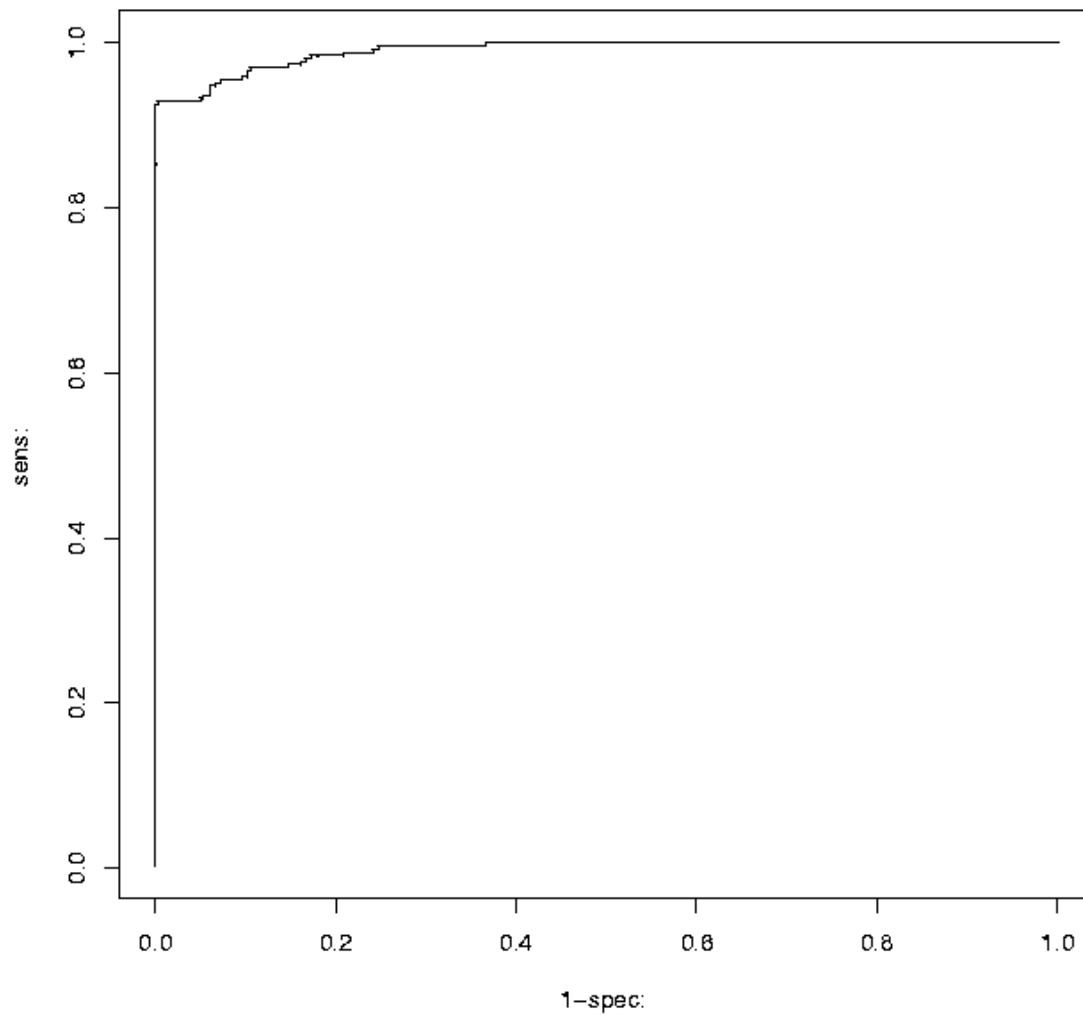


Figure 6: Signal ROC Plot for Differentiating Between Spike-in Concentrations of .71pM and .36pM. With no decoy probe sets, the GeneChip Exon Array differentiates well between spike-ins at different concentrations and those added at the same concentration.

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

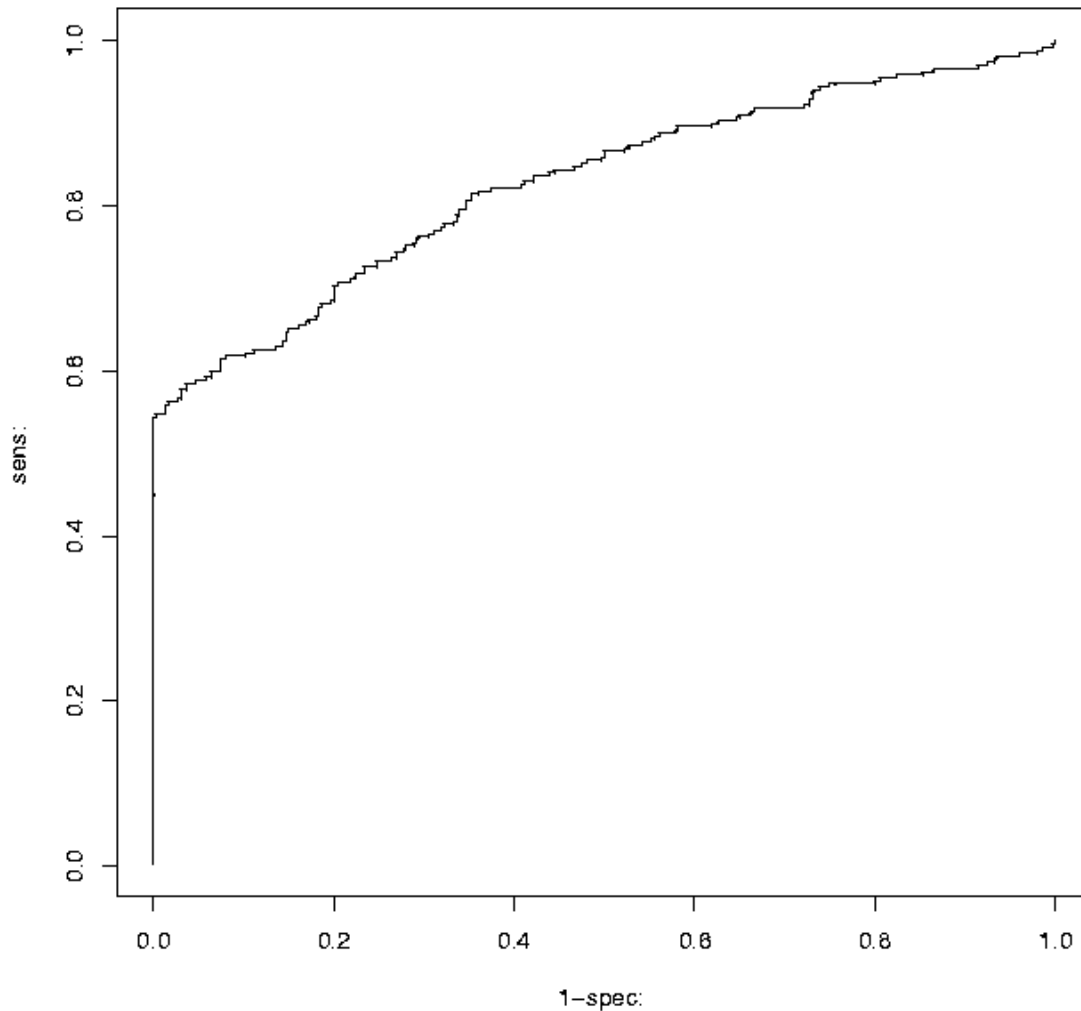


Figure 7: Signal ROC Plot of Spike-in Concentration (.71pM vs .36pM) Differentiation with 20 Decoy Probe Sets. The addition of low-signal probe sets decreases the ability of the PLIER expression estimates to differentiation between spike-in concentration pools.

III. Strategies for Coping with Exploratory Content:

One natural solution to determine which probe sets to use for summarizing a particular locus is to only use probe sets contained within high quality annotations such as RefSeq transcripts. Leveraging years of work cloning and sequencing genes combined with expert manual annotation assures that the probe sets chosen belong to that locus. This approach does not address the issue of including probe sets to alternative events, but if most exons are constitutive for a given gene, they should dominate the PLIER expression

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

estimate rather than the alternative probe sets. Robust methods such as PLIER and RMA (Irizarry et al. 2003) should be minimally affected by a limited amount of alternative splicing at a particular locus.

III.A. Selecting Features to Use in a Given Data Set:

Another strategy is to use feature sets from more exploratory annotations and iteratively discard those that appear to be performing poorly. This approach takes advantage of PLIER's ability to identify some of the signal at a particular locus and iteratively exclude features that are not correlated with that signal. As a proof-of-principle study, we implemented a procedure, IterPLIER, to use all features to generate an initial PLIER estimate and found the 22 features best correlated with that estimate, re-estimated the expression estimate with those 22 features using PLIER and then selected 11 features that were best correlated with the new expression estimates to generate a final PLIER signal estimate.

The correlation with the signal estimates without decoy probe sets compared to the result of IterPLIER on the low-intensity decoy set (Figure 8) and high-intensity decoy set (Figure 9) are improved to those resulting from using the initial PLIER estimates (Figure 1, Figure 2). The results for IterPLIER on the Akt3 locus (Figure 10) illustrate the improved correlation for all but the full-bounded probe set groups. This method also improves the ability of PLIER expression estimates to differentiate between spike-in concentration pools in the latin square experiments.

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

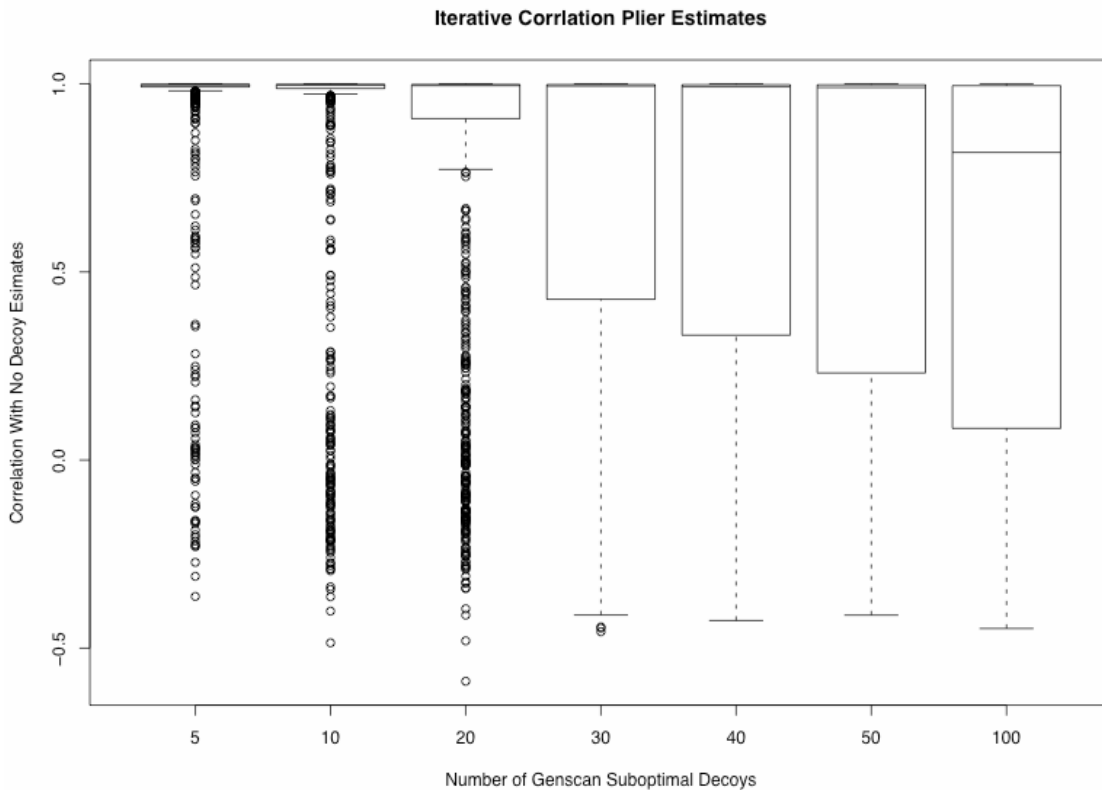


Figure 8: Correlation with original set as low-intensity decoys are added using an IterPLIER. The IterPLIER expression estimates are better correlated to the gold standard PLIER expression estimates than the raw PLIER expression estimates (Figure 2).

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

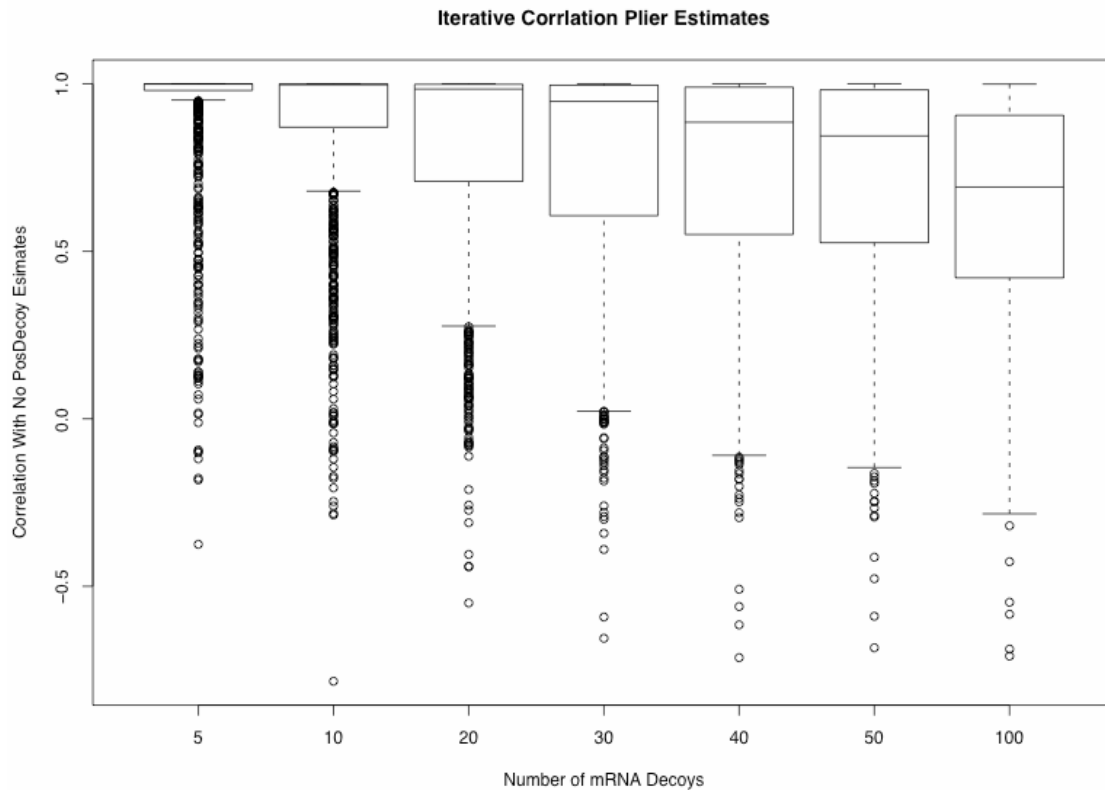


Figure 9: Correlation with original set as high-intensity decoys are added using IterPLIER. The IterPLIER expression estimates are better correlated to the gold standard PLIER expression estimates than the raw PLIER expression estimates (Figure 3).

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

Locus Estimate: Akt3

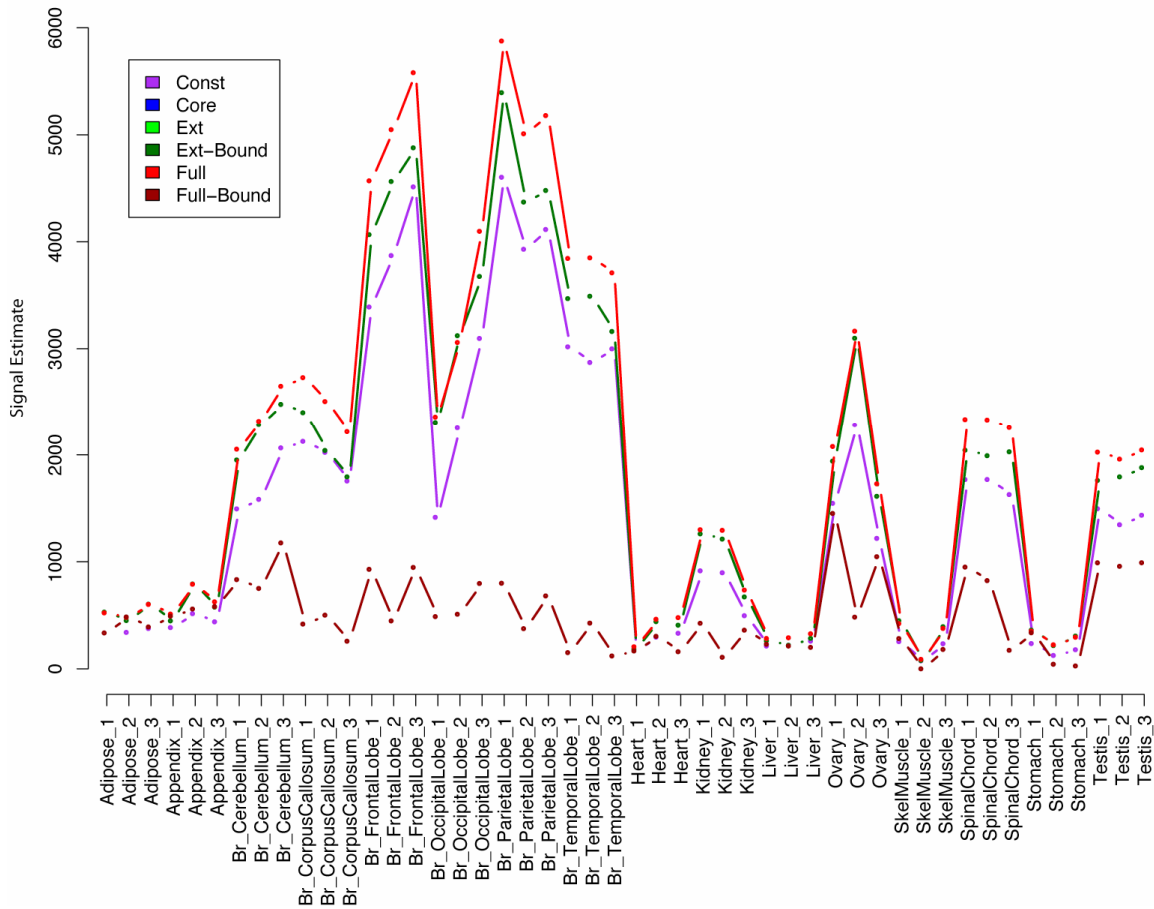


Figure 10: Locus estimates using Iter PLIER. The final signal estimates are plotted above illustrate that signal estimates consistent with the original bioinformatically enriched constitutive gold standard set were obtained for all cases but the full-bounded set. These results compare favorably to the raw PLIER expression estimates (Figure 5).

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

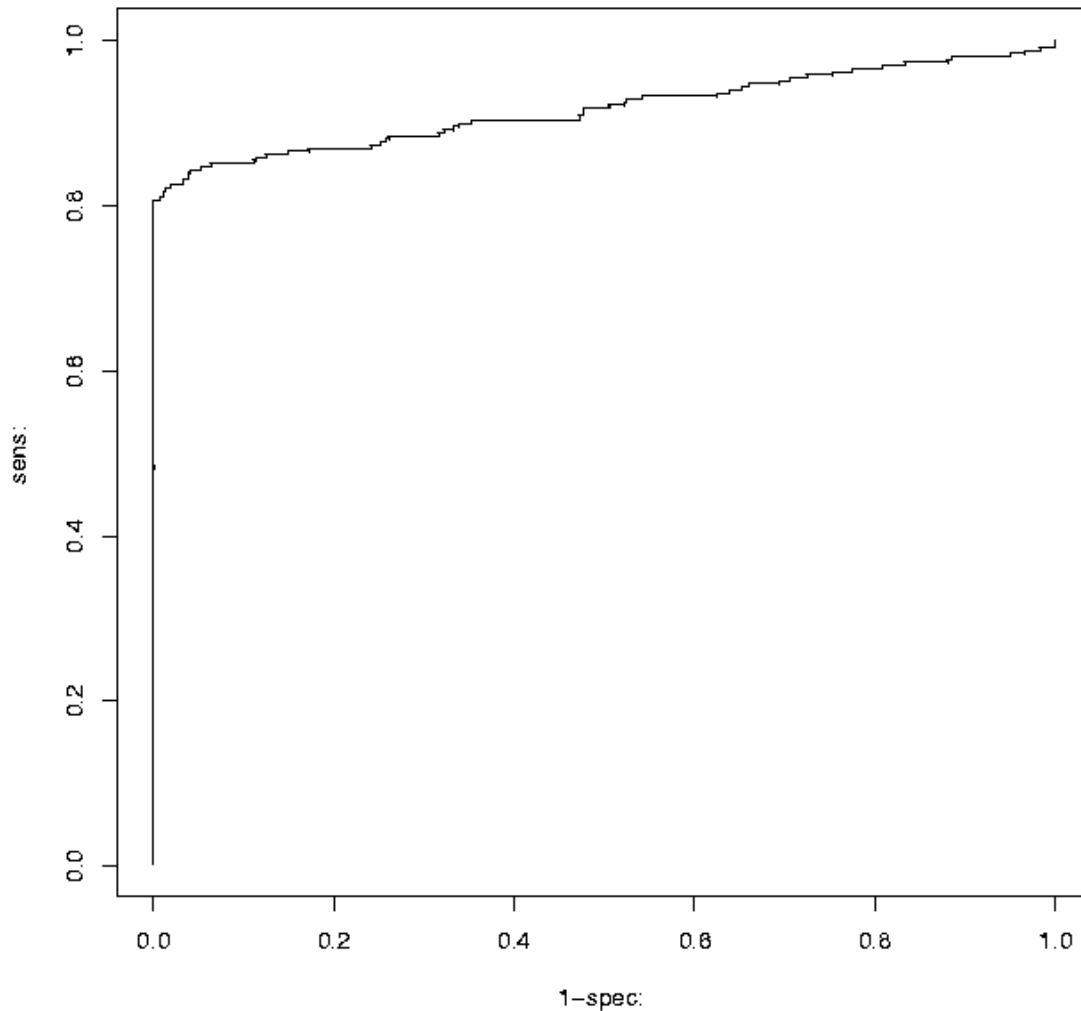


Figure 11: IterPLIER Improves Signal ROC performance in Presence of 20 Decoy Probe Sets (.71pM vs .36pM). The ability of IterPLIER expression estimates to differentiate between spike-in concentrations is improved compared to raw PLIER expression estimates (Figure 7).

IV. Summary:

Determining the best way to estimate locus intensity, and hence the underlying relative target response, is still an area of active research. Here we have shown that PLIER is more robust with respect to additional low-signal probe sets than to additional conflicting high-intensity probe sets. These results suggest that loci which may be transcribed separately should be joined with caution, while low confidence annotations may be included in an analysis with less risk. As the new Human Exon array has more data for each locus we now have an opportunity to

Gene Signal Estimates from Exon Arrays

Revision Date: 2005-09-27

Revision Version: 1.0

improve our estimates, but also face more choices about the type of estimates that are most appropriate for the study at hand.

V. References:

- Benson, D.A., M.S. Boguski, D.J. Lipman, J. Ostell, B.F. Ouellette, B.A. Rapp, and D.L. Wheeler. 1999. GenBank. *Nucleic Acids Res* **27**: 12-17.
- Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**: 78-94.
- Hubbell, E. 2005. Plier White Paper. In *Affymetrix White Paper*.
- Irizarry, R.A., B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**: e15.
- Korf, I., P. Flicek, D. Duan, and M.R. Brent. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1**: S140-148.
- Parra, G., E. Blanco, and R. Guigo. 2000. GeneID in Drosophila. *Genome Res* **10**: 511-515.
- Pruitt, K.D. and D.R. Maglott. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* **29**: 137-140.
- Wheeler, R. 2005. Gene Bounds: Probe Set Grouping, Annotation, and Evidence. In *Affymetrix White Papers*.