Alternative Transcript Analysis Methods for 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. 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 level of expression a particular locus. Detecting relative changes in alternative transcript forms is the subject of this whitepaper.

We conclude that an ANOVA based method applied to a Splicing Index equal to the ratio of exon signal to gene signal works well in two test data sets: one a tissue panel and the other a set of cancerous and normal samples from the same tissue. For the tissue panel data set, a correlation based method, Robust PAC, performs well.

One of the most well studied alternative splicing events is the alternative utilization of “cassette exons”. Here, we describe the evaluation of the performance of different mathematical methods designed to detect alternatively spliced cassette exons from the exon array data.

To do this, we first tested the method’s robustness in an artificial data set by constructing “genes” comprised of a group of probe selection regions that are best suggested by experimentally-derived annotation evidence as biologically real and are always jointly expressed in a sample data set. Each such group of probe selection regions is referred to as the “core constitutive exons” of that “gene.” We then simulated alternative splice events in each such “gene” by substituting exons from alternative regions of the genome and constructed ROC curves as a way to measure the performance of the different methods.

II. Data set

II.A. Core Constitutive Exons

To generate a set of exons that appear to be constitutively included in all transcripts splicing graphs were generated according to the method of Sugnet, *et al*, 2004[1], using all human RefSeq, mRNA and EST sequences. The graphs were pruned by requiring that each exon be either also present in mouse cDNAs or present multiple times in human cDNAs. The graphs were then searched to identify exons that were constitutively included in the pruned graph. To ensure that an exon is likely to be present in all transcripts a minimum number of 5 mRNAs or 10 ESTs (or a combination of the two) had to contain the exon.

Exons making up the core constitutive genes tend to be well-expressed; their median signal is greater than 10 times the median signal of background probesets in both data sets.

II.B. Tissue panel

The tissue panel data set consists of 11 human tissue samples. The hybridization cocktail for each sample was scanned in triplicate, using three arrays per sample for a total of 33 scans. In this study we quantile normalized the triplicates together, then multiplicatively scaled the resulting scans so that they all had the same median (“median scaling”). All features, whether genomic or non-genomic are included in the both normalizations.

II.B.1 Alternative splice set generation

The alternative splice set was generated by taking the probes of one exon in each gene (the first in each set) and substituting them into another gene. The probes of the real exon in the substituted gene were moved to yet another gene, and after repeating over all 5,800 genes, each gene had an “exon” that did not belong to it.

This approach works as variation within replicate groups is not due to biological variation and hence, if the core constitutive genes have detectable expression in one or more of the tissues, then they will likely have different expression in the different tissues. Hence each gene in the alternative splice set will each contain an exon that will behave differently across tissues. The major caveat in creating this alternative splice set is that it is unknown how well this procedure mimics true biological alternative splicing behavior.

II.C. Colon Cancer

The Colon Cancer data set consists of 18 biologically distinct samples arranged in 9 pairs. Each pair is a normal colon tissue sample and a colon cancer sample from each of 9 different individuals. There are no replicate scans. All scans were normalized using median scaling.

II.C.1 Alternative Splice set generation

For the colon cancer set, biological variation in signal is large enough within normal and tumor groups so that around 98% of the genes did not show significant differences in means of signal between the group of cancer and normal samples (p-value > .05). Accordingly, the approach in creating an Alternative splice set as in the tissue panel data will not work well; replacing an exon by an exon from another gene will not induce detectable signal changes from one group to the other and hence methods of alternative splice detection based on changes in signal of exon normalized by gene will not detect anything.

Instead, we took a different approach: a “background” set of probesets was generated by randomly selecting 5,800 probesets from probesets with mean signal across all 18 samples less than the 45th percentile of mean signal of each probeset across all 18 samples.

An alternative splice data set was simulated by replacing probe intensities in the first exon in each gene with probe intensities from a “background” probeset *for the normal scans only*. Tumor scans were left untouched.

We matched the number of probes in the replacement exon, as PLIER[2] signal is calculated across all samples by using probe intensities. Replacement probes need not have the same GC content, and replacement probe intensities in each sample were adjusted by the ratio of surrogate intensity mismatch of original to replacement based on GC content.

Exons making up the core constitutive genes have much higher overall signal than the background probesets on the chip: the median signal (2.5) of background probesets is around the 20th percentile of all probeset signals and the median signal (35) of the first exon is around the 70th percentile of all probeset signals.

III. Splice Detection Methods

Several different methods are assessed; Pattern-Based Correlation (PAC) and two ANOVA methods: Microarray Detection of Alternative Splicing (MIDAS) presented here, and ANOSVA (Cline *et al*, 2005)[3].

Their suitability to any particular data set will depend on the structure of the data set and the goals of the experiment. Data set exploration using multiple methods, including variations on these methods, is recommended.

In general, many of the splice detection methods have a similar structure:

- Under the null hypothesis, exons or probes comprising a gene are assumed to be proportional to each other across different samples.
- A model is fit that predicts probe or exon response under the null hypothesis.
- A statistic is constructed that measures how deviant the data is from the model.
- This statistic is used to construct a p-value.

III.A. Splicing Index

The Splicing Index captures the basic metric for the analysis of alternative splicing. Specifically it is a measure of how much exon specific expression (with gene induction factored out) differs between two samples. The first step is to normalize the exon level signals to the gene level signals.

Equation 1: $$n_{i,j,k} = \frac{e_{i,j,k}}{g_{j,k}}$$

where $e_{i,j,k}$ is the exon signal estimate of the i exon, j experiment, and k gene. $g_{j,k}$ is the gene level signal estimate of the j experiment and k gene. The splicing index is then the ratio of normalized exon signal estimates from one sample or set of samples relative to another. For example, in the colon cancer data set a splicing index could be established for each gene by taking the median $n_{i,j,k}$ for the cancer samples and dividing by the median $n_{i,j,k}$ for the normal samples. Alternatively one might want to calculate a splicing ratio for each paired normal/cancer sample and then report the median ratio. Use of such an index can be found in Clark, et al.[4]

III.B. PAC

PAC assumes that in the absence of splicing, exon expression follows gene expressions across samples using the following model:

$$\text{Equation 1: } e_{i,j,k} = n_{i,k} g_{j,k}$$

where $e_{i,j,k}$ is the signal of the i-th exon of the j-th sample of the k-th gene, $g_{j,k}$ is the signal of k-th gene in the j-th sample, and $n_{i,k}$ is the ratio of exon i signal to its gene signal.

We use a robust measure of gene signal and correlate signal of each exon with this signal. Low correlations are indications of alternative splicing. The robust measure of gene signal allows multiple exons to not track the overall gene, so long as they remain in a “small enough” minority. In this paper we use PLIER.

An important class of experiments asks the following question: Is there an alternative splice variant present in one group out of two groups of samples. PAC has the problem that it will fail in two-sample cases, as correlation will always be +1 (exon response agrees in direction from one sample type to the other) or -1 (exon response disagrees in direction); this always happens no matter how small the change actually is; in the rare event that there is no numerical change correlation would be zero. Hence PAC is better suited to experiments with a relatively large number of sample types.

III.C. MIDAS

We introduce an alternative ANOVA based method based on measuring differences between exon level signal and aggregate gene level signal that we call Microarray Detection of Alternative Splicing using (MIDAS) that has good performance.

The basic idea is the following:

- We use PLIER to generate a robust estimate of gene-level signal by using data from all features in all exons in the gene. This signal has the virtue that it will be robust against exons that exhibit anomalous signal across samples, whether they be non-expressing probe sets, probe sets that are incorrectly assigned to a gene, or are true alternative splice exons. Since this array has no mismatch probes we use a surrogate estimate of mismatch by using the median intensity of all antigenomic probes on the same array that have the same GC content as the probe.

- We use PLIER to similarly generate an estimate of signal for each exon.
- Under the null hypothesis of no alternative splicing for an exon, we would expect the difference between the logged signal for the exon and its gene to be a constant across all samples.

In other words, we expect that observed signal from each exon will have a constant ratio with observed signal from its gene.

A detection metric or statistic will be based on log of the Splicing Index, i.e, the difference between logged signal of each exon and its gene. We add a small constant before logging to stabilize variance (see discussion in III.A above).

III.C.1 MIDAS relation to ANOVA

Statistics based on the Splicing Index can be calculated either for each gene or for each exon. A model for possible splicing is:

Equation 2: $e_{i,j,k} = \alpha_{i,k} p_{i,j,k} g_{j,k}$

where $e_{i,j,k}$ is the signal of the *i*-th exon of the *j*-th sample of the *k*-th gene, $g_{j,k}$ is the signal of *k*-th gene in the *j*-th sample, $\alpha_{i,k}$ is the ratio of exon *i* signal to its gene signal in the sample where it is maximally expressed, and $0 \le p_{i,j,k} \le 1$ (with $p_{i,j,k} = 1$ in the sample where exon I is maximally expressed) is the proportionate expression of this exon of this gene in tissue *j*.

Dividing both sides $g_{j,k}$ obtains the Splicing Index and taking logs reduces this to an additive model (ignoring the possibility of zero signal):

Equation 3: $\log(e_{i,j,k} / g_{j,k}) = \log(e_{i,j,k}) - \log(g_{j,k}) = \log(\alpha_{i,k}) + \log(p_{i,j,k})$

III.C.2 Exon-level detection vs. Gene Level detection

Gene-level MIDAS is a 2-way ANOVA that includes an error term and possible interactions comparing:

Equation 4: $\log(e_{i,j,k}) - \log(g_{j,k}) = \log(\alpha_{i,k}) + \log(p_{i,j,k}) + \gamma_{i,j,k} + \varepsilon_{i,j,k}$

to the reduced model:

Equation 5: $\log(e_{i,j,k}) - \log(g_{j,k}) = \log(\alpha_{i,k}) + \varepsilon_{i,j,k}$

So we ask the question: are effects other than exon effects present; i.e., test $\log(p_{i,j,k}) = \gamma_{i,j,k}$ =0 across samples and exons.

Exon-level MIDAS considers the situation an exon at a time. In this narrow view $\log(\alpha_{i,k})$ is constant and hence it is appropriate to use classical 1-way ANOVA that compares the full model:

Equation 4: $$\log(e_{i,j,k})-\log(g_{j,k})=\log(\alpha_{i,k})+\log(p_{i,j,k})+\varepsilon_{i,j,k}$$

to:

Equation 5: $$\log(e_{i,j,k})-\log(g_{j,k})=\log(\alpha_{i,k})+\varepsilon_{i,j,k}$$

to test the hypothesis of no alternative splicing by testing for the constant effects model $\log(p_{i,j,k})=0$ for all *J* samples.

As discussed in IV, the log model is inappropriate when exons or gene are not expressed. In practice, stabilizing variance by adding a constant (as we do here and further discussed in IV) will go a long way to reducing the false positive rate.

Both PAC and MIDAS show considerable improvement in the ROC curves when using exon-level detection over gene-level detection. See IV.B for further discussion.

III.D. ANOSVA

A detailed description of the ANOSVA can be found in (Cline *et al*, 2005).In short, ANOSVA assumes all probe responses can be modeled as proportional to each other across different samples in the absence of alternative splicing; this reduces to an additive ANOVA model after taking logs. The ANOSVA method then tests for an alternative hypothesis of non-zero interactions between samples and exons using an F-statistic. Preliminary evaluation of ANOSVA on exon array data did not yield good performance for exon array data.

III.E.DECONV

Another published algorithm for estimating relative concentrations of different splice variants is described in (Wang, Hubbell, et al)[5], where it is applied to data on an experimental microarray.

If we define a set of splice variants each consisting of all exons but one, this algorithm can be recast into a model for probe-level intensities $x_{h,i,j,k}$ similar in form to Equation 2 above, but with an extra multiplicative term denoting the probe affinity $a_{h,i,k}$ of probe *h* of exon *i* of gene *k* in tissue *j*:

Equation 6: $$x_{h,i,j,k}=a_{h,i,k}\alpha_{i,k}p_{i,j,k}g_{j,k}+\varepsilon_{h,i,j,k}$$

Parameters are estimated using an iterative maximum likelihood estimation method, including $p_{i,j,k}$. For any one gene *k*, the relative ratios of the $p_{i,j,k}$ across tissues *j* represent relative concentration estimates of exons. Alternative splicing would be present if these relative ratios deviate from each other in a statistically significant way. This model differs from MIDAS in that error is treated additively rather than multiplicatively, and gene signal is estimated directly from the MLE rather than from a separate PLIER fit.

IV. Non-expressing probesets and genes

The Splicing Index finds probesets whose response data disagrees with a model where splicing is assumed absent. However, this can occur in cases with no alternative splicing:

- On the GeneChip® Human Exon 1.0 ST Array many probesets are based on gene prediction algorithms and EST singletons; as such, many may be interrogating regions that are not actually transcribed in a particular sample. Such probesets will typically have low probe intensities corresponding to noise and the Splicing Index will not generally track the gene signal.
- When the gene and probeset do correspond to biological reality, the gene may be poorly expressed or unexpressed in the biological samples. Both the gene and the probe set will have low probe intensities corresponding to noise and the Splicing Index will have no meaning.

Hence it is important that any statistical method based on the Splicing Index should be able to handle low intensity probe sets.

Handling of probesets or genes not expressed in sample is an active area of research and while there is no specific assessment of the methods in the context of the model breakdown above, we took the following steps to guard against model breakdown:

IV.A. Stabilize Variance

Most of the models discussed use the log of the Splicing Index and for these we stabilized variance by adding a constant before taking logs. This trades bias for variance. We chose a constant approximately equal to the 20%-ile of probeset signals. This appears to be well within the background level of probeset signals and since the core constitutive genes tend to be well-expressed (see discussion in II.C.1 above), the bias will be relatively small.

IV.B. Use exon-level models

Probe sets that show only noise against a backdrop of true variation in gene signal will tend to generate higher values using MIDAS F-statistics. If the detection statistic is calculated for the gene as a whole, then such exons must be removed before this calculation otherwise the gene will be flagged as a candidate for alternative splice events. We exploit the robust nature of PLIER estimates of gene signal to automatically filter out such probesets from the gene signal; this naturally leads to an exon-level approach.

V. Performance Evaluation

A ROC curve measures how well a statistic differentiates true alternatives from false positives. To do this exactly requires a known set that does not exhibit alternative splicing (the null set) to be compared with a known set that does exhibit alternative splicing (the alternative set). However, in our situation, we do not have complete knowledge, and hence the null set is likely contaminated by some real alternative splicing.

The effect of mixing some true alternative splice data into the null set will tend to cause the ROC curves to be constrained to a diamond (Bourgon)[6] around the diagonal of ROC plot and hence reduce the differences between statistics for large p-values. Mixing null data into the alternative set also affects the shape of the diamond; in this case the differences between statistics for small p-values will be affected.

It is our belief that small-pvalues are much more important than large p-values in assessing performance of different statistics that detect alternative splicing, and hence the error in construction of the null set is unimportant in our conclusion.

In any case, qualitatively we can see the empirical ROC curves seem to be well-behaved; errors in construction of the null set probably do not affect p-values less than 0.1

V.A. Tissue Data Set Performance

Results vary considerably depending on the data set. With no filtering for the tissue panel both PAC and MIDAS perform equivalently:

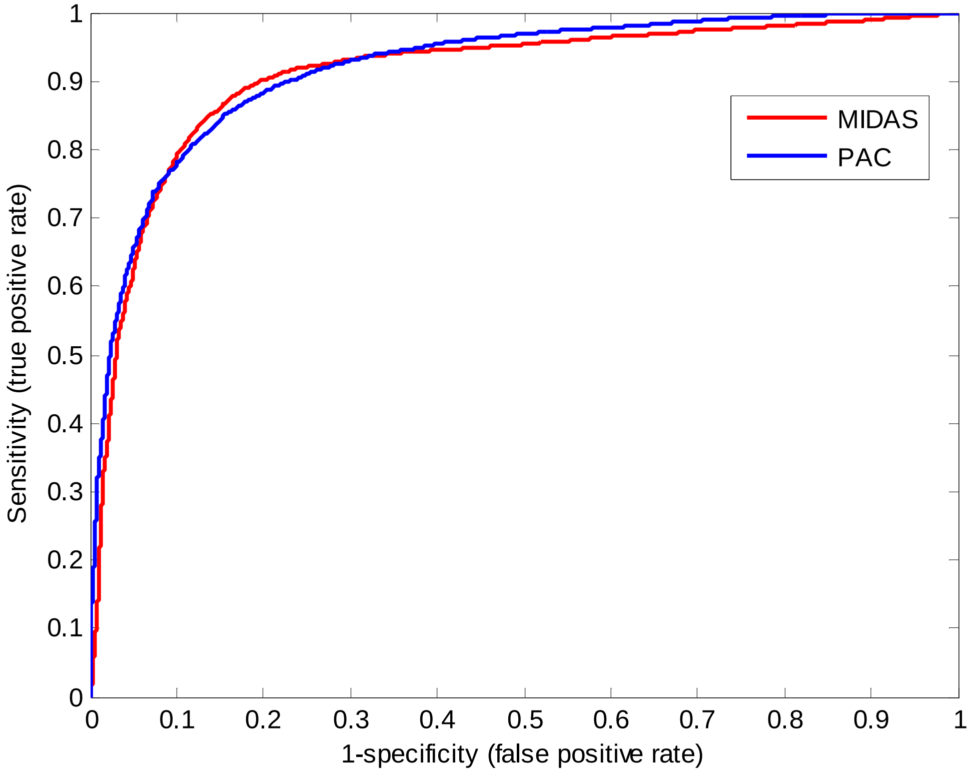

Figure 1: ROC curves for Alternative Splice detection in the Tissue Panel

V.B. Colon Cancer/Normal Data Set Performance

On the colon cancer data set, PAC is not applicable, and MIDAS reduces to a 2-sample t-test.

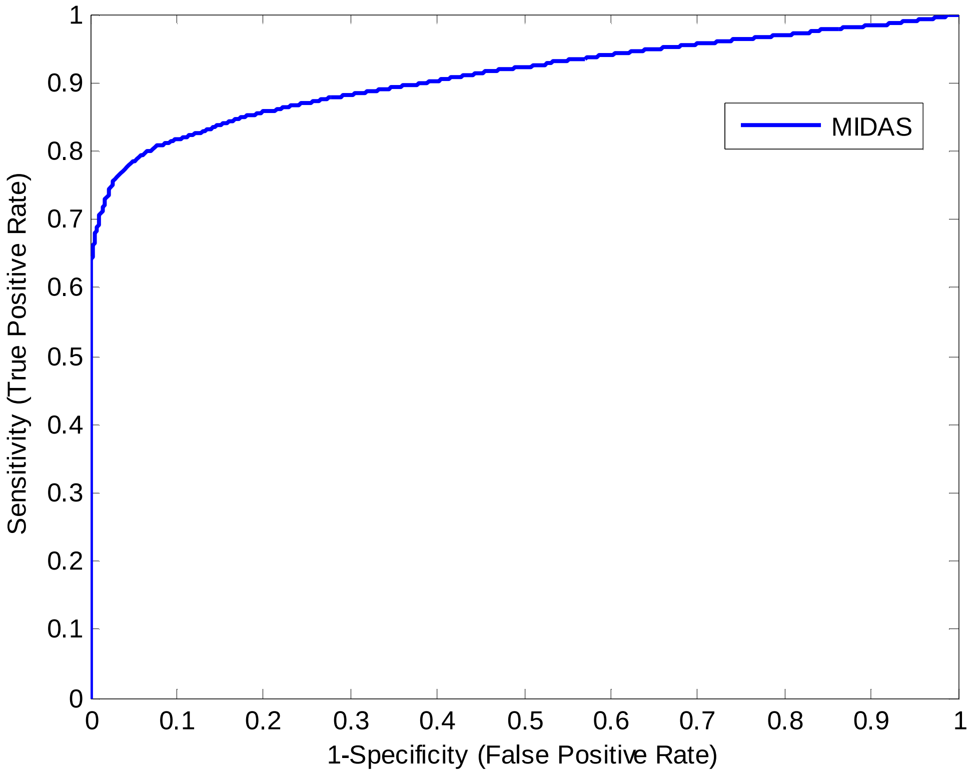

Figure 2: ROC curve for Colon samples (Normal vs. Tumor) using MIDAS

V.C. Methods stratified by number of exons per gene

A histogram of the distribution of exons is shown below in Figure 3:

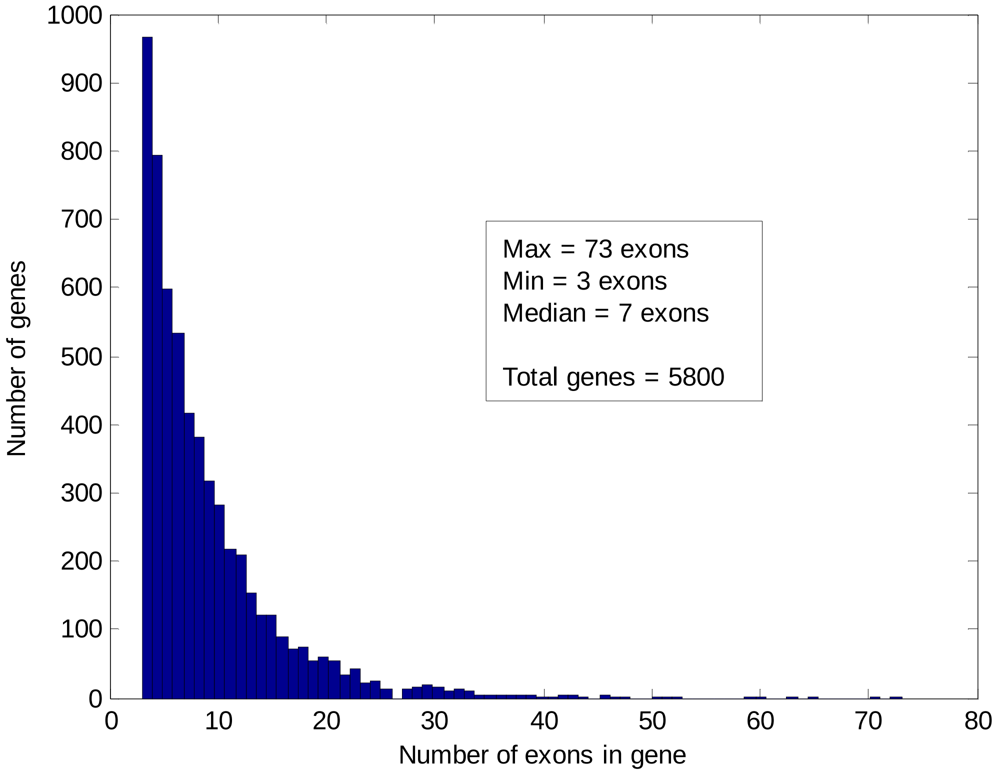

Figure 3: Histogram of exons per Constitutive Core Gene

We split exons into approximately 3 equal groups (Table 1):

Min exons/gene	3+	5+	10+
Max exons/gene	5	10	15
Number genes	1762	2252	1786

Table 1: Genes binned into groups by number of exons per gene

We see fairly similar ROC curves (Figure 4), with discrimination improving with larger number of exons per gene, possibly because the Plier estimate of gene signal is more stable with a larger number of exons.

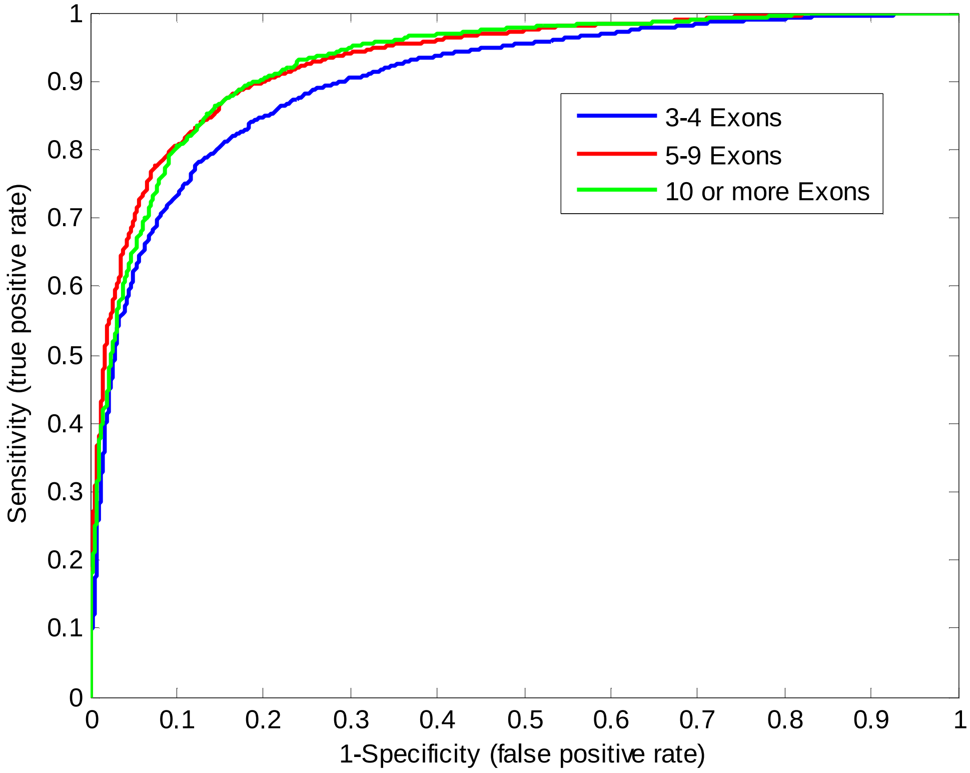

Figure 4: ROC curves for genes binned by size using MIDAS in the Tissue Panel

V.D. Application in Practice

In practice, using exon-level detection instead of gene-level detection forces the issue of multiple comparisons, where the larger the number of exons in the gene, the greater the chance of high statistical significance by chance alone.

V.D.1 P-values

In practice it is also desirable to use p-values. Using MIDAS on the tissue panel data, the ANOVA F-statistic gives p-values that are about 3 times too low (Figure 5), so in practice one would not want to take p-values too seriously. Different structure in replicates and sampling might not give the same relation.

We have also observed that when using the method on RefSeq genes, highly significant p-values can be associated with exons and genes that consistently exhibit low expression across sample types (model failures as discussed in Section IV). P-values can also be high when an exon from one gene is incorrectly assigned to another gene. While the ratio of exon signal to gene is variable, observing a very high ratio in one or more of the sample types is an indication that an error of this type may have occurred.

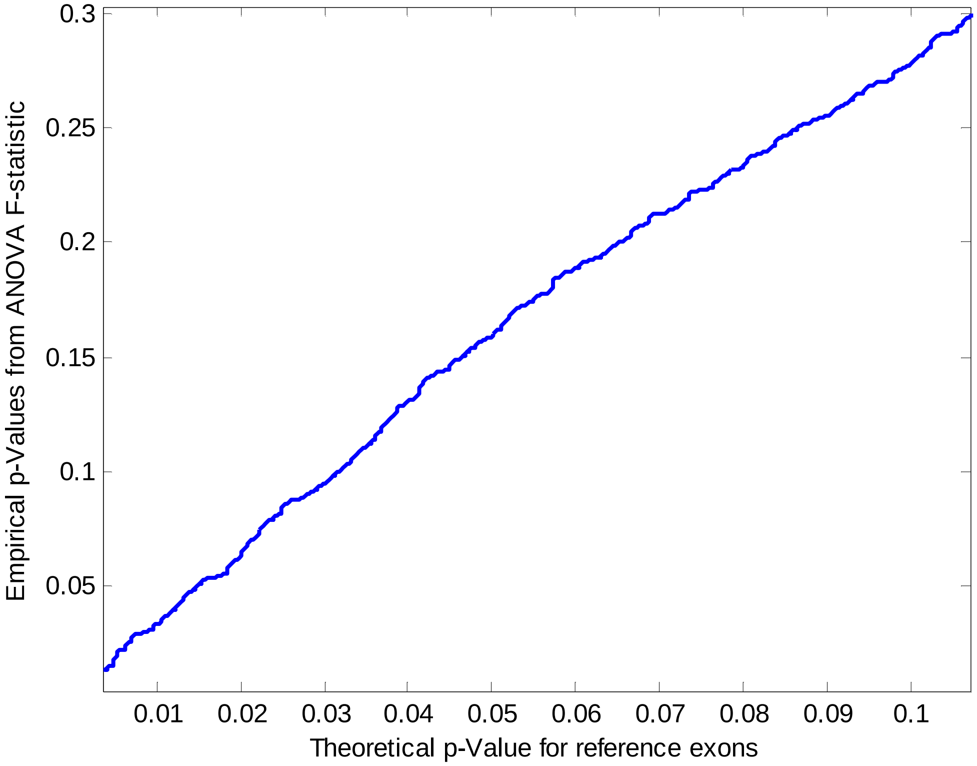

Figure 5: Comparison of Empirical p-Values with Theoretical p-Values

VI. References

[1] Transcriptome and genome conservation of alternative splicing events in humans and mice, **Sugnet CW, Kent WJ, Ares M Jr, Haussler D. Pac Symp Biocomput. 2004;:66-77**.

[2] PLIER: An M-Estimator for Expression Array, **Hubbell, E. Affymetrix White Paper**, 2005

[3] ANOSVA: a statistical method for detecting splice variation from expression data, **Cline M, Blume J, Cawley S, Clark T, Hu JS, Lu G, Salomonis N, Wang H, Williams A. Bioinformatics 2005 21(suppl_1):i107-i115**

[4] Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays, **Clark, T, Sugnet, C, Ares, M. Science 2002 May 3;296(5569):907-10**

[5] Gene structure-based splice variant deconvolution using a microarray platform. **Wang H, Hubbell E, Mei G, Cline M, Lu G, Clark T, Siani-Rose MA, Ares, M, Kulp D, Haussler. Bioinformatics 2003 19(suppl 1):i315-i322**

[6] ROC curves with misclassification, **Bourgon, Richard**. (in preparation)