



Technical Note

Identifying and Validating Alternative Splicing Events

An introduction to managing data provided by GeneChip® Exon Arrays

GeneChip® Exon Arrays are powerful tools for the discovery and study of mRNA transcript diversity. For the first time, researchers will obtain approximately 1.4 million data points from each sample in a single experiment. This increased data density also poses a number of data analysis challenges, including management of a higher number of potential false positives from the much larger data set.

In addition, exon arrays provide a new dimension of genomic information beyond classical gene expression results from microarrays—alternative splicing. For the analysis of alternative splicing, new algorithms will need to be developed and tested in various data sets. This is an active area of research and we anticipate that new developments and methods will continue to emerge with the increasing availability of sample data sets on exon arrays.

In this Technical Note, we present practical recommendations for managing some of these challenges based on our experience at Affymetrix. A sample workflow operating mainly within Affymetrix® Expression Console™ Software and other GeneChip®-compatible™ software programs for exon array analysis is described in detail. Our experience shows that systematic filtering of the raw array results, as detailed here, is critical to the improvement of validation rate in the subsequent RT-PCR validation experiments. Most of the analysis strategies discussed in this Technical Note will be applicable to any statistical method developed for identifying alternative splicing events.

INTRODUCTION

Alternative splicing is a major source of protein diversity for higher eukaryotic organisms, and is frequently regulated in a developmental stage-specific or tissue-specific manner. Current estimates suggest that 50 to 75 percent (or more) of human genes have multiple isoforms.

Splice variants from the same gene can produce proteins with distinct properties and different (even antagonistic) functions. In addition, a number of genetic mutations involved in human disease have been mapped to changes in splicing signals or sequences that regulate splicing. Thus, an understanding of changes in splicing patterns is critical to a comprehensive understanding of biological

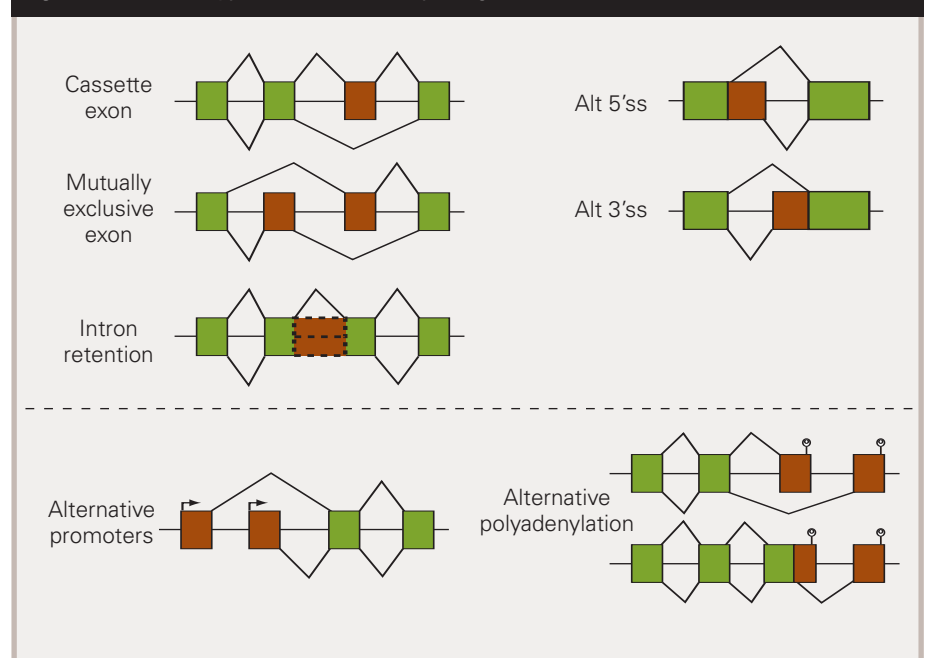
regulation and disease mechanisms.

This Technical Note provides detailed guidelines for those using exon arrays for alternative splicing analysis, to help researchers generate meaningful interpretation of exon array data more quickly. These guidelines include the following:

- An introduction to alternative splicing prediction algorithms when comparing changes that have occurred between two groups of samples
- Description of an analysis workflow
- Practical considerations in filtering data
- Experimental verification of alternative splicing events

A list of technical support materials is included for convenient reference. It is highly recommended that users review these reference

Figure 1: Different types of alternative splicing events.



documents to become familiar with the array design and basic algorithms associated with exon arrays prior to use of this Technical Note to walk through the actual analysis workflow. For detailed information about using this workflow to generate biologically significant results comparing colon cancer and normal samples, see Gardina, *et al.*

BACKGROUND

EXON ARRAY DESIGN

GeneChip® Exon 1.0 ST Arrays are an incredibly powerful tool for the study of alternative splicing. The ability to treat individual exons (or parts of exons) as independent objects makes it possible to observe differential skipping or inclusion of exons. This is not possible (or at least, certainly not optimal) on more classical expression array designs that focus on transcription activities at the 3' end of a gene.

The exon array was designed to be as inclusive as possible at the exon level, drawn from annotations ranging from empirically determined, highly curated mRNA sequences to *ab-initio* computational predictions (for more information, see Technical Note, "GeneChip® Exon Array Design"), enabling the *discovery* of new alternative splicing events. This is an advantage over exon-exon junction arrays, which are typically designed against only observed or annotated junctions.

The GeneChip Human Exon 1.0 ST Array contains approximately 5.4 million probes ("features") grouped into 1.4 million probe sets, interrogating over 1 million exon clusters, which are exon annotations from various sources that overlap by genomic location. A Probe Selection Region (PSR) represents a region of the genome that is predicted to act as an integral, coherent unit of transcriptional behavior. A PSR is the target sequence from which probes are designed. In many cases, each PSR is an exon; in other cases, due to potentially overlapping exon structures (or alternative splice site utilization), several PSRs may form contiguous, non-overlapping subsets of a true biological exon.

The median size of PSRs is 123 bp with a

minimum size of 25 bp. About 90 percent of the PSRs are represented by four Perfect Match (PM) probes (a "probe set"). Such redundancy allows robust statistical algorithms to be used in estimating presence of signal, relative expression and existence of alternative splicing.

The exon arrays do not include a paired Mismatch probe for each PM probe. Instead, surrogate background intensities are derived from approximately 1,000 pooled probes with the same GC content as each PM probe. One commonly used set of background probes is called the "antigenomic" background probe set, which contains sequences that are not present in the human genome (or a few other genomes) and are not expected to cross-hybridize with human transcripts.

In addition, exon arrays provide robust gene-level expression analysis. The median number of probes for each RefSeq gene is 30 to 40 distributed along the entire length of the transcript, as compared to probes selected only at the 3' end in classical gene expression microarrays.

PROBE SET ANNOTATIONS AND TRANSCRIPT CLUSTERS

The plethora of exon architectures (as shown schematically in Figure 1, e.g., cassette exons, mutually exclusive exons, alternative splice sites, alternative transcriptional starts and stops), the variations in quality of transcript annotations and the necessity of rapidly incorporating new genomic knowledge have led to a dynamic design for reconstituting exons into genes.

A set of rules was created for virtual assembly of the probe sets (exon-level) into transcript clusters (gene-level) based on the confidence level of the supporting evidence and the juxtapositions of the exon borders (White Paper, "Exon Probe set Annotations and Transcript Cluster Groupings v1.0"). The mapping between probe sets and transcript clusters is defined by *meta-probe set* lists as described below (in order of decreasing confidence). The number of clusters cited below reflects the version of mapping files provided by Affymetrix as of November 2006. Updated mapping files incorporating the latest information may be downloaded

directly from Affymetrix' web site.

- Core: RefSeq transcripts and full-length mRNAs (17,800 transcript clusters)
- Extended: Core + cDNA-based annotations (129,000 clusters)
- Full: Extended + *ab-initio* gene predictions (262,000 clusters)

SIGNAL ESTIMATION ALGORITHMS

Several statistical methods may be used to combine information from probes belonging to the same gene, or exon, to generate expression signal values of the gene or exon. For example, RMA and PLIER are two of the most commonly used algorithms. Although this Technical Note focuses only on the workflow with PLIER, the basic principles apply to RMA and other signal estimation analysis methods, as well.

Relative expression can be determined using the PLIER algorithm (White Paper, "Guide to Probe Logarithmic Intensity Error (PLIER) Estimation"), a robust M-estimator that uses a multi-chip analysis to fit a model for feature response and target response for each sample. The target response is the PLIER estimate of the signal of a probe set (exon-level).

Gene-level PLIER estimates are derived by combining all probe sets predicted to map into the same transcript cluster (according to the meta-probe set list). Since PLIER is a model-based algorithm, exons that are alternatively spliced in the samples, therefore exhibiting different expression patterns compared to the constitutive exons, will have down-weighted effect in overall gene-level target response values (White Paper, "Gene Signal Estimates from Exon Arrays").

IterPLIER is a variation that iteratively discards features (probes) that do not correlate well with the overall gene-level signal and then recalculates the signal estimate to derive a robust estimation of the gene expression value primarily based on the expression levels of the constitutive exons.

Presence/absence of exons is determined using "Detection Above Background" (DABG), as documented in the Detection Call section of the Technical Note, "Statistical Algorithms Reference Guide," using surrogate background intensities as described above.

ALTERNATIVE SPLICING ALGORITHMS

Alternative splicing by definition is differential exon usage. It is not sufficient, however, to simply identify exons with differential expression patterns; we must also account for differential transcription of the gene itself.

For example, if a gene is expressed two-fold higher in sample “A” than in sample “B,”

then all of the constitutively expressed exons in that gene are also likely to have two-fold higher signal values. Thus, in order to focus on the differential inclusion of individual exons, we need to normalize to the transcription rate of the gene, as shown in Figure 2.

This concept led to the development of the “Splicing Index” (Srinivasan K., *et al.*) and

was the motivation for other methods (MiDAS, Robust PAC, etc.). Different analytical methods for performing splicing analysis may deal with this issue in different ways, but ultimately, the total transcription activity and splicing change must be included in the analysis for the method to be successful.

The Splicing Index is a conceptually simple algorithm that aims to identify exons (actually, PSRs) that have different inclusion rates (relative to the gene level) between two sample groups. The normalized exon intensity (NI), as described in Figure 3, is the ratio of the probe set intensity to the gene intensity as estimated by PLIER (or IterPLIER).

The Splicing Index Value (SI) is calculated by taking the log ratio (base 2) of the NI in Sample 1 and the NI in Sample 2. A hypothetical example of this is shown in Figure 4, where the inclusion rate of the measured exon is 10 times lower in Sample 2, despite the fact that the actual intensity is higher.

An SI of 0 indicates equal inclusion rates of the exon in both samples, positive values indicate enrichment of that exon in Sample 1, and negative values indicate repression or exon skipping in Sample 1 relative to Sample 2.

To identify exons that have statistically significant changes in inclusion rates between two groups, a Student’s t-test can be performed on the gene-level normalized exon intensities. The absolute value of the Splicing Index represents the *magnitude* of the difference for exon inclusion between the two samples (or groups of samples). It is important to consider both the p-value from the t-test and the magnitude of the change since the best candidates for validation are likely to have very small p-values and very large SIs (either positive or negative).

MiDAS, which is implemented as a command-line program of the Affymetrix Power Tools (APT), is conceptually similar to the Splicing Index but allows simultaneous comparisons between multiple sample groups. MiDAS employs the gene-level normalized exon intensities in an ANOVA model to test the hypothesis that no alternative splicing occurs for a particular exon. In the case of only two sample groups, ANOVA reduces to a t-test.

Figure 2: Exon-level signal consists of the gene-level expression and the inclusion of that exon as a part of the gene (splicing rate).

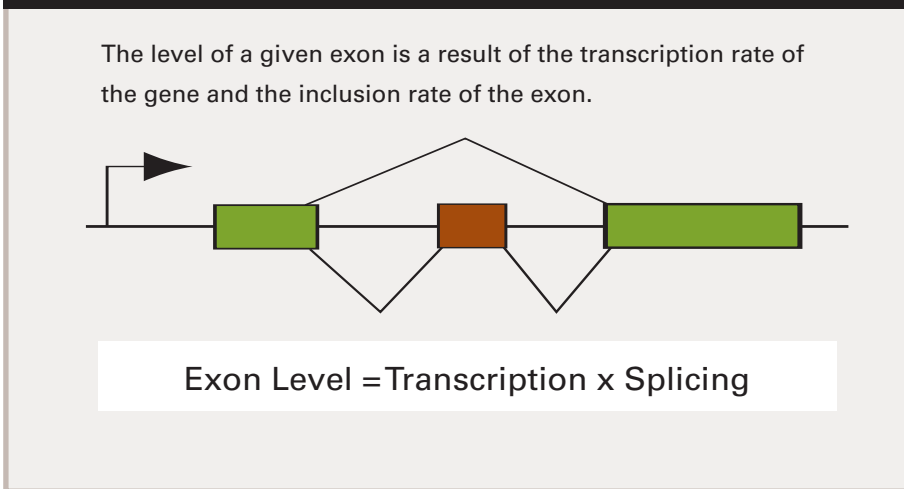


Figure 3: Gene-level normalized intensity (NI).

$$\text{Gene-level Normalized Intensity (NI)} = \frac{\text{Probe set intensity}}{\text{Expression level of the "gene"}}$$

Figure 4: Splicing Index Value (SI).

NI	Sample 1	Sample 2
$\frac{\text{Probe set intensity}}{\text{Gene level}}$	$\frac{500}{500} = 1.0$	$\frac{600}{6000} = 0.1$
Sample 1 has 10x higher inclusion level		
SI	$\text{Splicing Index} = \log_2 \frac{\text{Sample 1 NI}}{\text{Sample 2 NI}}$ $= \log_2 \frac{1.0}{0.1} = +3.32$	

For a more detailed description of algorithms from Affymetrix, refer to the White Paper “Alternative Transcript Analysis Methods for Exon Arrays,” available at www.affymetrix.com. Various GeneChip®-compatible™ software products for the exon application also introduce additional options and algorithms for alternative splicing analysis. A complete list can be found at http://www.affymetrix.com/products/software/compatible/exon_expression.affx.

CAVEATS OF MIDAS AND THE SPLICING INDEX

There are several caveats of the currently implemented, first-generation splicing algorithms that are important to consider when interpreting exon array results.

First, the alternative splicing results are highly dependent on the accurate annotation of the transcript clusters and precise quantitation of gene-level estimates. Mis-annotated transcript clusters or otherwise incorrect gene-level predictions may produce less reliable results.

For example, genes with many alternatively spliced exons (a large percentage relative to the total number of exons) or instances where only a sub-set of the gene is expressed (due to alternative transcriptional starts/stops) are likely to have some inaccuracy in gene-level estimates. One possible improvement that may increase the robustness of gene-level estimates is to only use probe sets interrogating constitutive exons. However, this can be extremely dependent on specific samples and difficult to predict. The Splicing Index works best when the gene has a large number of constitutive exons and a small number of alternative exons.

Furthermore, these algorithms assume that the splicing pattern is consistent among all of the samples within a group. This is probably not always the case, e.g., tissue-map experiments that combine multiple tissue types into a single group. Utilizing groups consisting of multiple tissue types makes it impossible to discover splicing differences between two tissues within the same group. Large variations of splicing pattern (and gene expression, to a lesser extent) within a group will reduce the significance of the t-test, resulting in a larger p-value. One

potential solution for a tissue-map experiment with many different tissue types is to treat each tissue as a different group and do many pair-wise comparisons of a single tissue to all other tissues. MiDAS is capable of including more than two groups using ANOVA.

The third limitation is the requirement for replicates in the statistical t-test (minimum of three samples per group). While it is actually possible to run the algorithms with fewer than three samples per group, the t-test needs the replicates to estimate the intra-group variability to calculate meaningful p-values. In cases where there are insufficient replicates within a sample group, it may be possible to create logical groups that might be expected to have similar splicing patterns. As with many algorithms, the result is likely to be more robust with increased numbers of replicates, larger group sizes and increased consistency within each group.

CONTEXT FOR INTERPRETING ALTERNATIVE SPLICING DATA

Alternative splicing introduces a new level of complexity relative to the analysis of gene expression. Several things are critical to consider for planning, evaluating and interpreting microarray data when screening for altered splicing events:

1. Alternative splicing is often a subtle event: there is rarely an all-or-none change in exon inclusion between one tissue and another. A more likely scenario may be a shift from 50 percent inclusion of an exon in one tissue to 80 percent inclusion in a different tissue. A small change in the ratio of isoforms may be biologically significant.

2. Alternative splicing is a very common phenomenon: ~ 75 percent (or more) of all genes appear to be alternatively spliced. Between any two tissue types there may be thousands of probe sets that vary in their inclusion into transcripts. The splicing pattern can also be inherited and therefore differ from individual to individual, resulting in heterogeneity in the population examined. This adds an additional factor in interpreting results or identifying disease-specific alternative splicing events in the mixed background.

3. We are typically dealing with a mixed population of related transcripts from a locus rather than the singular “gene” that we usually associate with other expression arrays. This is a messy concept, but it more closely reflects real biology. The exon array does not detect transcripts *per se*; it detects exons that can be virtually reassembled according to the meta-probe sets. It has no information about which exons are actually physically linked together in transcripts, but instead treats genes as a collection of all the exons associated with that locus. There may, in fact, be a number of combinations of transcripts that would account for the observed exon probe set signals; therefore, known transcripts from public databases may be a useful guide for interpreting the results.

4. Differential gene expression may further complicate differential alternative splicing analysis. In order to calculate differential exon inclusion into gene transcripts, we always have to account for relative differences in gene expression between two sample types. Therefore, predictors of alternative splicing must accurately estimate both gene-level signals and exon-level signals.

The combination of these factors means that typical considerations in designing a good microarray study such as sample size and sample quality become even more critical. For example, tumors may be a mixture of different tumor stages and probably also mixed with normal tissue. Even normal tissue may be composed of various cell types exhibiting different splicing patterns; for instance, colonic sections can consist of both smooth muscle and epithelial tissue.

Due to the inherent heterogeneity of the biology, the data needs filtering by multiple methods at both the exon and gene level to reduce false positives (described in detail in the **Workflow** and **Filtering Methods** sections below). Visually inspecting the probe set intensities in a genomic context is also a useful way to improve true discovery rate.

Ultimately, experimental confirmation by a different technology—RT-PCR, for example—will provide confidence in the results obtained from statistical analysis. However, experimental validation is laborious, so

approaches to narrow down the candidate list computationally may save much effort in the lab. It should also be noted that in some cases filtering the data in an attempt to reduce false positives may involve a trade-off with the loss of true positives. The suggested filtering methods are intended to highly enrich the results with true positives, thereby maximizing the likelihood of positive validation. Depending on the specific aims of an individual experiment, a different approach may be more appropriate.

In addition, some users may be interested in using exon arrays to detect unique individuals or samples within the experimental group that exhibit different splicing patterns compared to the rest of the samples. This analysis can be used effectively to detect individual-to-individual splicing variation or exon-skipping mutations, frequently observed in cancer samples. For this type of analysis, the methods described here are certainly relevant but the details will need to be modified to meet the specific research objectives.

WORKFLOW

This workflow (as schematically shown in Figure 5) begins with CEL files and concludes with empirical validation of the results. Some introductory concepts are described here:

- Some of the workflow occurs outside of Expression Console with external applications for scripting/filtering (Perl) and statistics (like Partek and other GeneChip®-compatible™ software packages).
- There are actually two parallel analyses (gene-level and exon-level) carried out in Expression Console that ultimately converge at MiDAS (or other alternative algorithms) for alternative splicing prediction.
- The assumption for this entire approach is that there are two or more sample groups, and we intend to find differential alternative splicing between the groups. Alternative splicing can occur within one tissue, but in this scenario, we are only concerned with the way it varies in different tissues or conditions.
- There are two points at which meta-probe set options are invoked:

1. In signal estimation, the meta-probe set defines the exons to be used to calculate gene-level signal (PLIER, IterPLIER, RMA, etc.), and therefore defines the set of genes to include in the analysis.

2. In MiDAS, the meta-probe set defines the exons that are mapped to the input genes.

The result is that you could, for instance, specify only Core (highly confident) genes, but use MiDAS to predict alternative splicing for the Full (speculative) set of exons associated with those Core genes.

- Command-line formats are provided (gray box in next column) for those using UNIX-based systems. The analogous input options for the Windows GUI version of Expression Console are explained in detail in the user's manual.

PHASE I: SIGNAL ESTIMATION

Listed below is a suggested example to conduct exon array analysis. Although there are other possible ways to perform these same functions, we describe one of them here as a starting point for new users.

A) Gene level

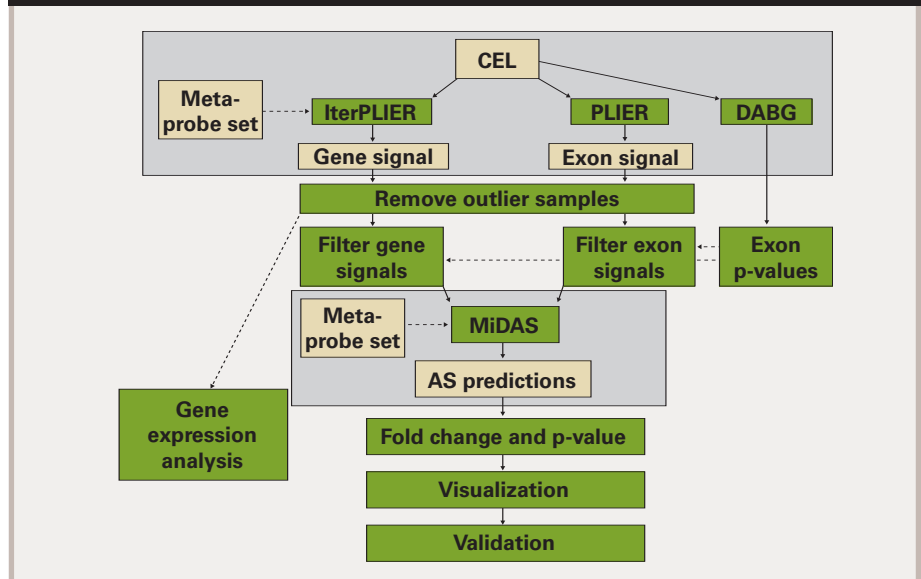
- Use Expression Console
- IterPLIER (PM-GCBG)
- Antigenomic.bgp as the GC background pool. Antigenomic sequences are not found in the human genome or several other genomes.
- Sketch normalization at 50,000 data points
- Meta-probe set = Core (this is the most conservative set of genes with the highest confidence)
- Set of CEL files

Do not use DABG for gene-level estimation of “Present/Absent.” Detection calls for genes will be estimated separately with another method (as described below during the filtering steps).

For those using the command-line APT programs on UNIX:

```
probe set-summarize -a quant-
norm.sketch=50000.bioc=false,pm-
gcbg,iter-plier -p HuEx-1_0-st-
v2.pgf -c HuEx-1_0-st-v2.clf -b
antigenomic.bgp -s meta-probe
set.core.txt MyCelDirectory/*.CEL
```

Figure 5: Workflow for gene expression and alternative splicing analysis with exon arrays. Tan boxes represent files or data sets and green boxes represent processes or programs. Functions within the gray boxes occur within Expression Console, or Affymetrix APT. Solid arrows are the main data flow and dashed lines are accessory flows.



B) Exon level

- Use Expression Console
- PLIER (PM-GCBG). Most probe sets only have four probes, which is too limited to be useful with IterPLIER at the individual exon level.
- Antigenomic.bgp
- Sketch normalization at 50,000 data points
- DABG (PM-only)-produces p-values for detection above background
- It is possible to limit the analysis to a subset of exons by providing a list file. However, in general, it may be best to begin with an inclusive set like *probe set-list.main.txt* to minimize any bias in analysis.
- Set of CEL files

```

For those using the command-line APT
programs on UNIX:
probe set-summarize -a plier-gcbg-
sketch -p HuEx-1_0-st-v2.pgfc -c
HuEx-1_0-st-v2.clf -b
antigenomic.bgp -s probe set-
list.main.txt -a pm-only,dabg -x 2
MyCelDirectory/*.CEL
    
```

PHASE II: REMOVING OUTLIERS

Outlier samples should be identified and eliminated, or at a minimum, accounted for. This is particularly true in analysis of exon array data since low sample quality is likely to be highly influential in generating noise, therefore leading to high false positive rates.

A typical approach is the Principle Component Analysis (PCA) for identification of possible outliers, followed by ANOVA to test their effect (using a standard statistical package, such as Partek's Genomics Suite). Examples are shown in Figure 6. Samples found as extreme outliers should be removed from the analysis.

PHASE III: FILTERING SIGNAL DATA

In order to obtain meaningful splicing information and to decrease the chances of false positives (thereby increasing the verification rate), a number of filtering steps can be performed.

Within the analysis described in this Technical Note, validation of splicing events in

Figure 6: Principle Component Analysis (PCA) to identify outlier samples. The examples shown here are the colon and normal paired samples run on exon arrays. The CEL files are available at www.affymetrix.com. Exon-level signals from each sample are mapped by PCA, and paired tumor-normal samples are joined by lines. The circled sample data (back left) appears to be an outlier in two dimensions (1 and 3) of the PCA mapping. However, it is Patient #3 (black arrows) that behaves contrary to the majority of the samples; while most of the tumor samples tend to have a higher value on the chief PCA component (i.e., rightward on the X-axis) than the normal tissue, Patient #3 tends to the opposite direction. The aberrant behavior of Patient #3 is confirmed by ANOVA (shown in Figure 7 below). This also illustrates one of the advantages of having paired samples.

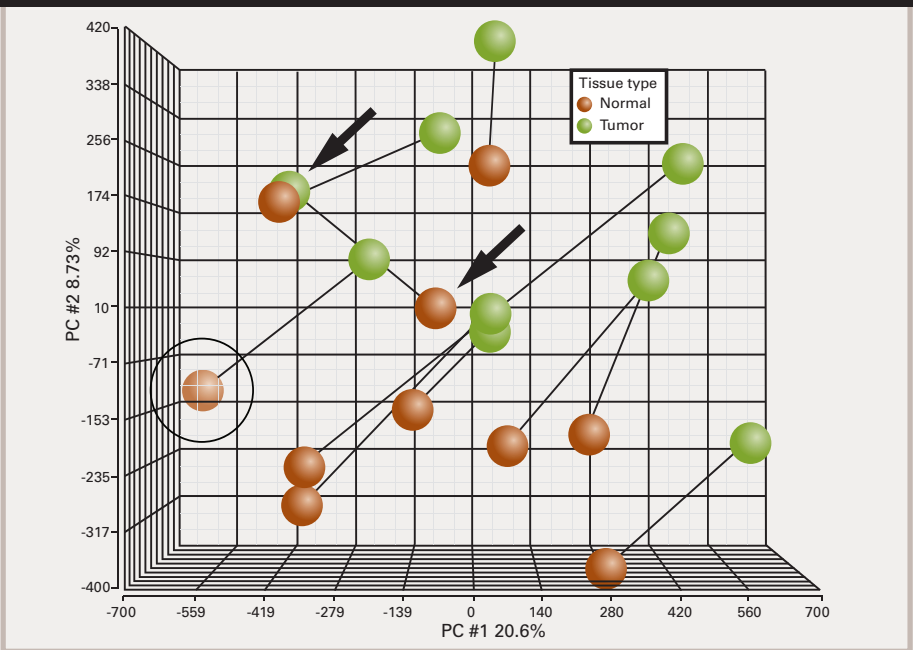


Figure 7: ANOVA analysis of signal and noise. Exclusion of the sample pair from Patient #3 ("Del_3") in ANOVA analysis greatly improves the signal-to-noise ratio for discrimination by tissue type (normal vs. tumor). Therefore, Patient #3 was removed from the remainder of the analysis described below. No other sample pair had this magnitude of effect. Be aware, however, that removing outliers reduces sample size; therefore removal of outlier samples should be done cautiously. Furthermore, noise ratios (normalized against error) from ANOVA also showed that the gender, patient and tumor-stage categories were relevant (F Ratio > 1), and thus should be included as factors in later analysis.

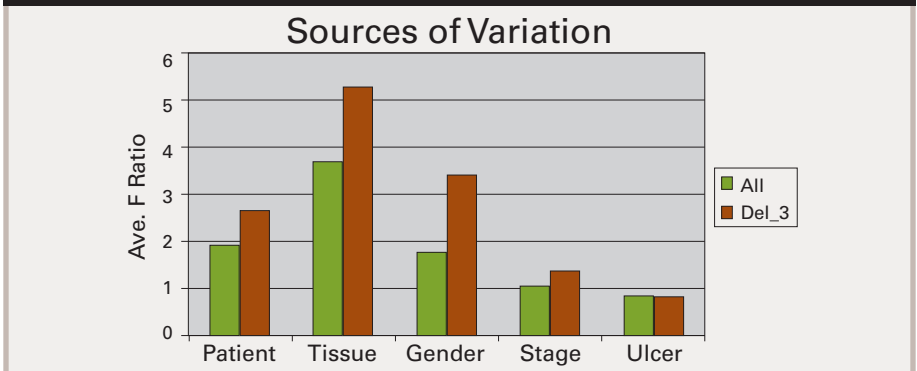
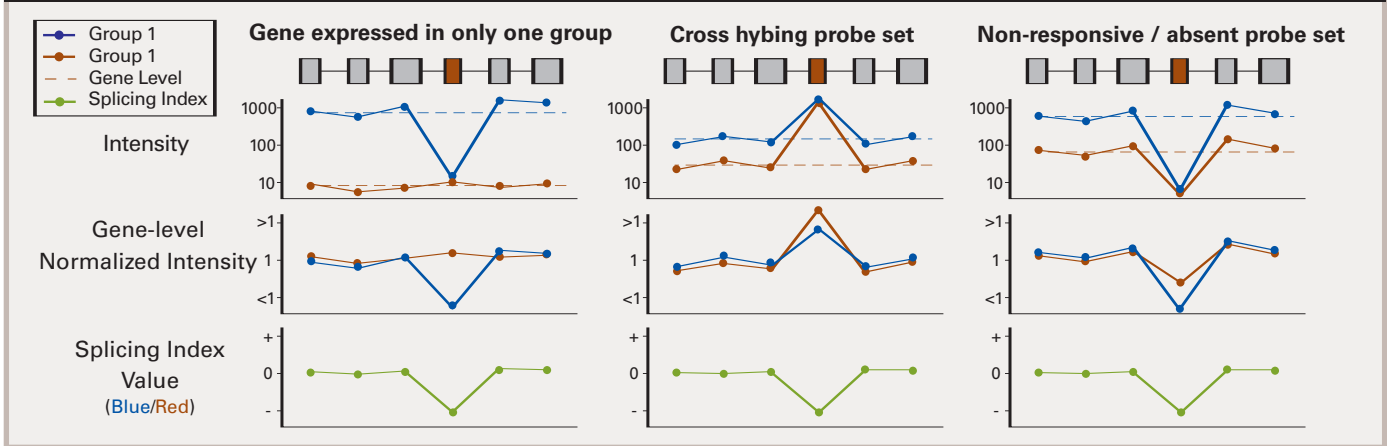


Figure 8: Several scenarios that may lead to artifactual predictions of alternative splicing events. In each case, it is the combination of misleading probe set results and differential gene expression that creates a false prediction of alternative splicing.



the laboratory (e.g., by RT-PCR) is by far the most laborious step. Therefore, additional effort made during computational analysis to reduce false positive rate will ultimately decrease time spent in the lab. In this section, we present some suggestions for minimizing the impact of artifactual signals. In addition, several more stringent optional filtering steps are possible that may further lessen false positives and make it easier to find “low-hanging fruit” for validation. Figure 8 illustrates several examples of possible scenarios that may lead to misidentification of alternative splicing events.

Several filtering methods applied to signal data are suggested here and are described in detail in the **Filtering Methods** section. Some of these methods have been implemented in third-party software, but otherwise can be performed using simple scripts, e.g., in Perl.

PRIMARY (MANDATORY) FILTERING

1. Remove any gene (transcript cluster) that is not expressed in both sample groups—to eliminate the scenario illustrated in the left column of Figure 8.
2. Remove any exon (probe set) that is not expressed in at least one sample group—to eliminate the scenario illustrated in the most right column of Figure 8.

SECONDARY (SUGGESTED) FILTERING

1. Remove probe sets with high potential for cross-hybridization—to eliminate the

scenario illustrated in the middle column of Figure 8.

2. Require a minimum gene signal level.
3. Remove probe sets with very large exon/gene intensity ratio—which may also implicate cross-hybridization to other gene sequences.
4. Remove probe sets with very low exon/gene intensity ratio in the group that is expected to have the higher rate of inclusion—which may also implicate non-linearity of the probe set.

TERTIARY (OPTIONAL) FILTERING

1. Remove genes that have very large differential expression.
2. Limit search to high-confidence exons.
3. Restrict the search to only highly expressed genes.
4. Focus on exons that have gene-level normalized intensities near 1.0 in the group predicted to have a higher inclusion rate and near 0 in the group predicted to have a lower inclusion rate.
5. Filter probes with unusually low variance.
6. Limit search to known alternative splicing events.

PHASE IV: MIDAS

Subsequent to filtering, both gene-level and exon-level signal intensities are input into MiDAS.

One option to carry out alternative splicing analysis through MiDAS is detailed here:

- Gene signals—Core (filtered); the Core gene-level meta-probe set is chosen at the signal-estimation stage
- Exon signals—all exons (filtered)
- Meta-probe set—Full; at this point the meta-probe set tells MiDAS which set of exons to evaluate. The combination described here means that MiDAS will look at all the exons (“Full”) associated with the input Core genes. A more conservative approach may be to look only at Core exons by inputting the Core meta-set probe file.
- `cel_ids.txt`—This file tells MiDAS how samples are partitioned into Groups (e.g., “brain,” “lung,” “kidney”) that should be compared for differential splicing. The ANOVA in MiDAS can compare multiple groups but does not incorporate additional factors like gender, tumor stage, etc.

For those using the command-line APT programs on UNIX:

```
midas -c cel_ids.txt -g CC.genes-core.i-plier.sum.txt -e CC.exons-main.plier.sum.txt -m meta-probe-set.full.txt
```

MiDAS outputs predictions for alternative splicing events as p-values. These should not necessarily be treated as true p-values but rather as scores that reflect relative ranking. The false positive rate is likely to be much higher than

indicated by p-values and the high-ranking candidates should be further screened/filtered and ultimately verified empirically.

PHASE V: POST-MIDAS FILTERING

1. Keep only probe sets with p-values less than a particular cutoff (i.e., p-value $< 1 \times 10^{-3}$).

Probe sets with the smallest p-values are the most likely to have significant differences in inclusion rate. The cut-off value used can easily be lowered to increase stringency. The optimal cut-off value is really dependent on the number of targets you wish to see on your final list. In addition, you can do a multiple testing correction (such as Bonferroni or the Benjamini-Hochberg False Discovery Rate) to determine the p-value at which the differences are considered statistically significant.

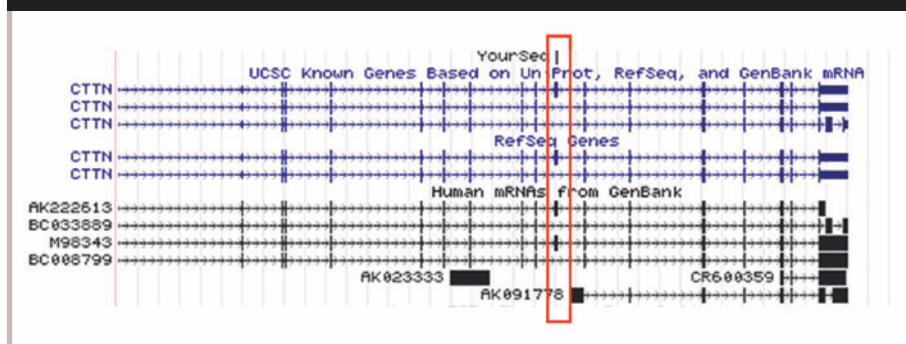
2. Sort results by magnitude (absolute value of Splicing Index) and keep only probe sets that have a minimum difference of 0.5.

Probe sets with larger magnitudes of predicted changes are more likely to have more dramatic splicing changes. Small magnitude changes may still be biologically relevant; however, larger magnitude changes typically make better candidates for validation. As discussed above, the cut-off value can easily be raised to increase stringency of the filter depending on the study design and needs. This step is analogous to filtering by fold change for gene expression. Whereas p-values measure the ability to statistically separate two groups, the SI gives the magnitude of the difference. Unfortunately, MiDAS does not directly output the SI and it must be calculated externally to the program (e.g., with Perl or R).

PHASE VI: VISUAL FILTERING OF RESULTS

Despite all attempts to filter the data, several classes of false positives are not possible to be identified purely based on values from the statistical analysis itself. A simple manual inspection of the data in genomic context can catch many of these potential pitfalls. Previous experience has shown the benefits of doing this in two ways.

Figure 9: A candidate splicing event mapped onto the UCSC Genome Browser. The highlighted region that corresponds to the targeted probe set ("YourSeq") appears to be a cassette exon (one that is included or skipped in different known transcripts).



First, BLAT the sequence of the PSR identified to be alternatively spliced (can be obtained using the probe set ID on the NetAffx Data Analysis Center at www.affymetrix.com) to the UCSC Genome Browser (<http://genome.ucsc.edu>). BLATing the sequence can also provide information about potential cross-hybridization or if the probe set maps to multiple genomic locations.

View the probe set in the browser and zoom out slightly to get a broader view of the region around the exon or the entire gene that contains the exon of interest. The genome browser can provide important information about location of the probe set within the gene and if the exon is known to be alternatively spliced (based on EST and mRNA sequences). It may also be possible to observe overlapping, intervening or independent transcripts (on the same or opposite strand) and potential artifacts of transcript cluster annotations, such as multiple genes being combined into a single cluster.

While not an absolute requirement for validation, candidates consistent with known examples of alternative splicing are more likely to prove to be true positives. However, this filtering approach will also tend to suppress the discovery of novel splicing events.

Second, it is worthwhile to examine the probe set intensity data of your exon of interest and surrounding exons in an integrated browser such as the Affymetrix Integrated Genome Browser (IGB) or BLIS (Biotique Systems). This is a helpful step to ensure that the data is consistent with alternative splic-

ing. For example, if the exon of interest is predicted to be higher in group A (increased inclusion), but surrounding probe sets also appear higher, it may suggest that the region of the gene has an alternative start or stop that affects multiple exons. Also, if there are multiple probe sets for the exon, it is reassuring that the data for all of them are in agreement.

Visualization and careful observation of the data can often predict the likelihood of a positive result prior to verification. For example, consider an experiment where we are comparing tumor samples to normal samples to look for aberrant splicing patterns in cancer. We find an exon that is predicted by the Splicing Index (or similar algorithm) to be higher (increased inclusion) in the tumor samples. However, when we locate the position of the exon within the gene, we discover by looking at the genome browser that the targeted exon appears to be constitutive (always included) in numerous EST and mRNA sequences. This exon is very *unlikely* to be a simple case of alternative splicing (cassette exon), and thus is very unlikely to result in a positive validation. It is entirely possible that this seemingly constitutive exon is, in fact, alternatively spliced in cancer. However, in this case, the exon is predicted to be *higher* in the tumor samples, but the exon is already included in all of the transcripts of this gene. It is not possible to have an increase in inclusion when the exon is already present in 100 percent of transcripts. While the data may represent an interesting biological event, it is possible that the result is a false positive or may involve a

Figure 10: A view of two mutually exclusive exons (19a and 19b) in the BLIS viewer (Gardina, *et al*). This candidate splicing event was confirmed by RT-PCR. (BLIS normalizes the exon signal for each sample to the median exon signal for that sample across all the exons in the view.)

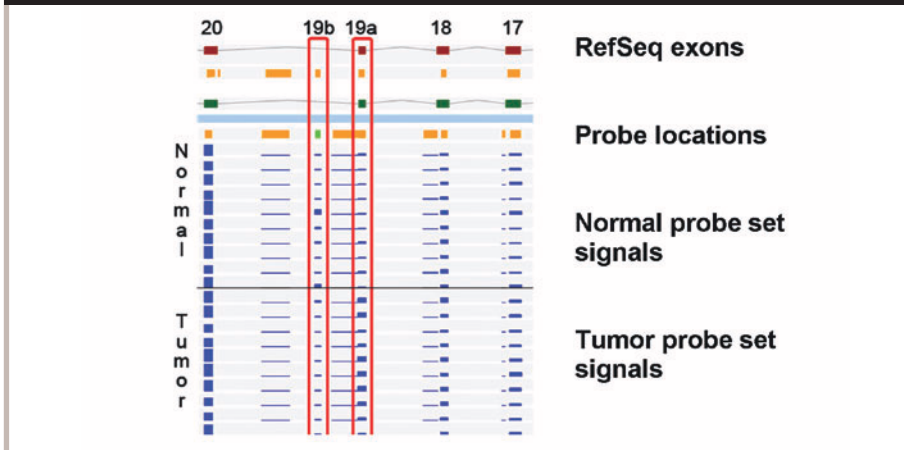
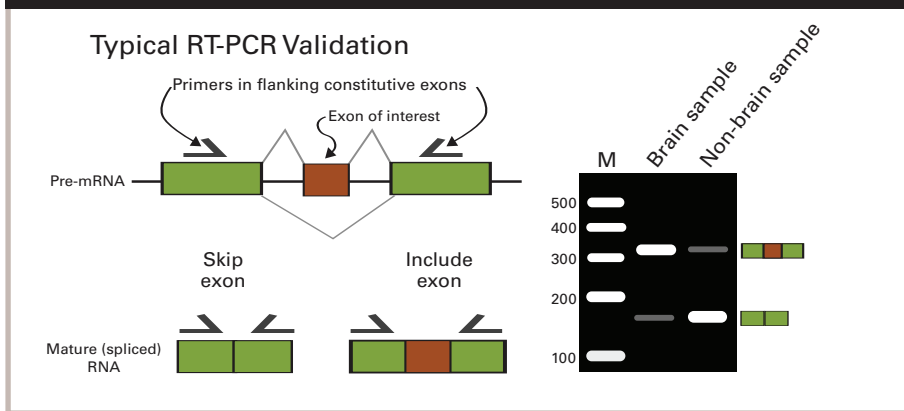


Figure 11: Primer design for RT-PCR validation of alternative splicing events.



more complex regulation (possibly involving alternative starts/stops) and as such may not be easily verified by the follow-up RT-PCR verification.

While the process of manual filtering can be time-consuming, a well-trained eye can view 100 or more potential hits in a reasonable amount of time. Our experience suggests that it is well worth the time and it can be coupled with the design of primer sequences for validation. Several commercially available GeneChip®-compatible™ Software Packages, including Partek's Genomics Suite and Biotique's X-Ray, provide easy click-through visualization of the raw signal data on an exon-by-exon basis that belong to the same gene, and can easily be used to conduct this step of visual inspection.

PHASE VII: RT-PCR VALIDATION

Validations are easily carried out via RT-PCR using primers in exons that flank the exon of interest. This works well for simple cassette exons and alternative 5' and 3' splice sites. As illustrated in Figure 11, when the PCR products are run on an agarose gel, there will be separation between the "include" product and the "skip" product based on size ("include" product is larger by the size of the alternative exon and therefore runs slower through the gel). Primers are best designed in flanking constitutive exons and it is usually possible to calculate the expected sizes of the PCR products ahead of time.

While carefully designed quantitative PCR (i.e., TaqMan) could provide more accu-

rate estimates of the absolute levels of each isoform, depending on the individual requirements of the experiment, it may not be necessary in the verification of results from the exon array. In most cases the two different bands representing two alternatively spliced isoforms are produced by the same primer pair, and it is easy to observe changes in the relative intensity of the two bands between samples. Amplification efficiencies may differ between the two products, but this bias will be the same for all samples. The key observation is a change in ratio (relative intensity) of the two products. By starting with equal amounts of input cDNA, simple RT-PCR can be considered semi-quantitative.

Since each validation case is unique and different types of alternative splicing events require different strategies, for best results, automated selection of primer sequences is not recommended. For example, detection of alternative starts and stops requires placement of a primer within the exon of interest since the target exon does not connect to flanking exons on *both* sides. In this case, two separate PCR reactions should be run. One determines presence/absence of the exon (as described above), and the other is a reference primer pair designed to measure the alternative pattern or expression level of the gene.

A similar design strategy can be used for mutually exclusive exons of similar size that cannot be resolved by separation on a gel. In addition, in order to discriminate inclusion versus skipping events on a gel, the overall size of the amplicon should be designed relative to the size of the target exon. Smaller exons should have smaller overall amplicon sizes so that the skip/include products will clearly resolve (separate) on the gel. Clearly, design of primers for validation is not a "one size fits all" situation.

There are several classes of validation targets that pose additional challenges to RT-PCR. Alternative starts/stops and mutually exclusive exons were mentioned above. Alternative transcriptional starts/stops that include multiple exons are also tricky and require validation strategies different from the standard flanking exon approach.

For cassette exons that are very large in size, it may be difficult to efficiently amplify

the “include” product using flanking exons due to the large size of the amplicon. This may also be the case for genes with multiple consecutive cassette exons.

Additionally, incorrect selection of constitutive exons can lead to misinterpretation of the RT-PCR results. Independent intervening transcripts may be impossible to validate using this method since the exon of interest is never included in transcripts for that gene. Some of these problematic cases are illustrated in Figure 12.

Nevertheless, thoughtful and careful primer design can usually find a solution for validation. This also supports the need for visualization of the data in genomic context. Many of these problematic situations can be avoided by looking at EST/mRNA sequence data for evidence of these odd types of transcript structures. In many cases, if you didn't know the factors beforehand, RT-PCR may result in a single band and you may incorrectly identify the exon as a false positive.

FILTERING METHODS

In this section, we will discuss some of the inherently challenging cases associated with this type of array-based alternative splicing analysis, and propose a number of primary and optional methods to increase the true discovery rate.

AN IDEAL SPLICING EVENT

Prior to describing the specifics of filtering, it is helpful for us to conceptualize the ideal scenario for gene and exon expression that will maximize our ability to identify differential splicing. Factors that deviate from this ideal situation tend to lower the probability of correctly identifying splicing events mathematically. Characteristics of the ideal alternative splicing event include:

1. High gene expression (and consistent across all samples)
2. Equal gene expression in both sample groups
3. Most exons (probe sets) parallel the overall expression of the gene (in other words, most of the exons belonging to that gene are constitutive so that the

estimation of the gene-level signal is more reliable)

4. A single exon is alternatively spliced
5. The alternatively spliced exon is always included in one sample group and never included in the other sample group.

Figure 13 shows an example of a correctly identified alternative splicing event (later validated by RT-PCR).

ANOMALOUS PROBE SIGNALS

In some cases, technical anomalies may

give a false signal for probe set intensity due to cross-hybridization, probe set saturation or inherently weak and non-linear response. Although this is not a frequent event, the exon array represents 1.4 million probe sets, giving a larger number of potential false positives. In most cases, this signal will be about the same across all samples in all groups. However, if the associated gene-level signals are different between the two groups, the false probe set

Figure 12: Examples of challenging cases for RT-PCR validation where the primer pairs (marked by flat arrows) may not give anticipated validation results due to the complication of the biology. Therefore, careful selection of the primers and use of an alternative primer design strategy should be considered if negative results are obtained.

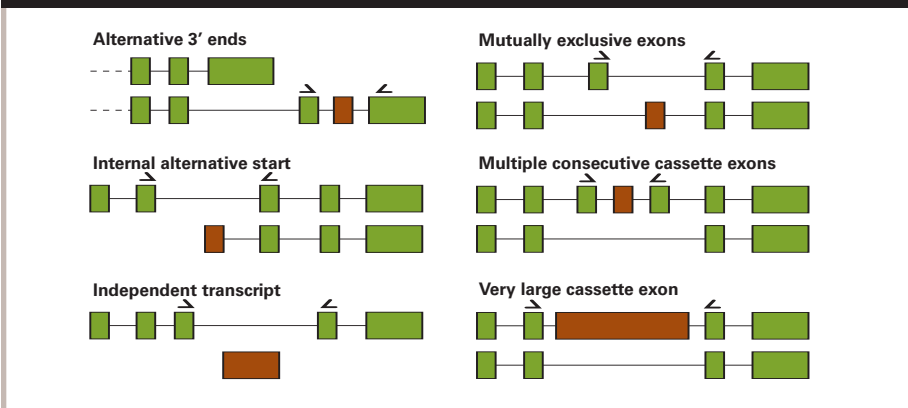
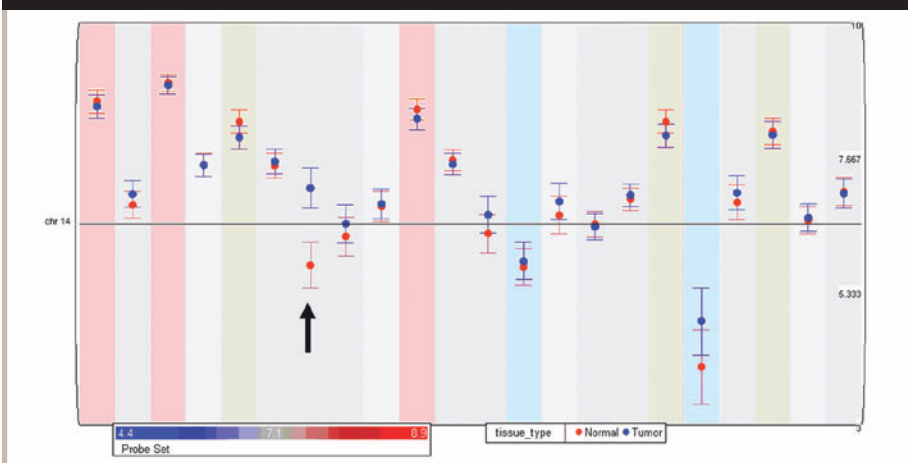


Figure 13: Exon-level signals from a single gene with a validated differential splicing event. The average signal and standard error for each probe set is shown for tumor samples (blue) and normal samples (red). The log₂ intensity scale is shown on the right-hand axis. In the validated splicing event (indicated by the arrow), the exon shows low expression relative to gene expression in normal tissues. Note that there are many characteristics of an ideal scenario: the gene is highly and equally expressed in both groups, and most of the exons behave uniformly. (The data here is visualized with the Partek Genomics Suite.)



signal will be interpreted as differential inclusion of the exon because the probe set signals are normalized against the gene signals from each group.

1. Cross-hybridizing or constitutively high probe signals. In this scenario, an individual

probe set signal may be stuck on “high” regardless of the actual expression of that exon. Therefore, this exon will appear to be relatively more included in the sample group that has lower expression for the cognate gene. A probable example of this

phenomenon is shown in Figure 14.

2. Poorly hybridizing or non-linear probes. This is the converse of the above problem, but the probe sets will appear to be weakly expressed in all samples (see Figure 15).

Characteristics of these anomalous probe set signals are:

1. Approximately equal intensity in both sample groups
2. Low variance
3. Possibly extreme intensities relative to gene expression level.

Implementation of a systematic filtering method will be possible as more data becomes available. Fortunately, such anomalous signals are also generally easy to identify by visual inspection.

PRIMARY (MANDATORY) FILTERING OF SPLICING INDEX RESULTS

There are two major criteria that must be met for genes and probe sets to be included in further downstream analysis after microarrays. Both relate to a common theme, where the lack of expression will very often be mistaken for alternative splicing when compared to actual expression (or even low expression) in another group, since the signals are so low that the arrays are simply measuring the noise of the experiments. Thus, the first two *mandatory* filtering steps of all splicing analyses should be:

1. **Remove exons (probe sets) that are not expressed in at least one group.** (The reason for only requiring detection in one group is that if the exon is alternatively spliced it is completely reasonable that it is absent in all of the samples of one group.)

The following is a good default set of rules as starting points to remove exons that are expressed at low level:

1. If the DABG p-value of the probe set is less than 0.05, the exon is considered as “Present” in a sample.
2. If the exon is called as Present in more than 50 percent samples of a group, the exon is considered to be expressed in that group.
3. If the exon is expressed in either group, accept it; otherwise, filter

Figure 14: A probable artifact from probe sets with constitutively high signals (indicated by the arrow). Even though this probe set gives the same signal for both tissue types, it appears to have a relatively high inclusion in normal samples because the gene-level signal is lower in normal than in tumor. This probe set displays three trademark symptoms of this artifact: it is higher than the neighboring probe set signals, it is approximately the same in both tissues and it shows an extremely low variance across the replicates.

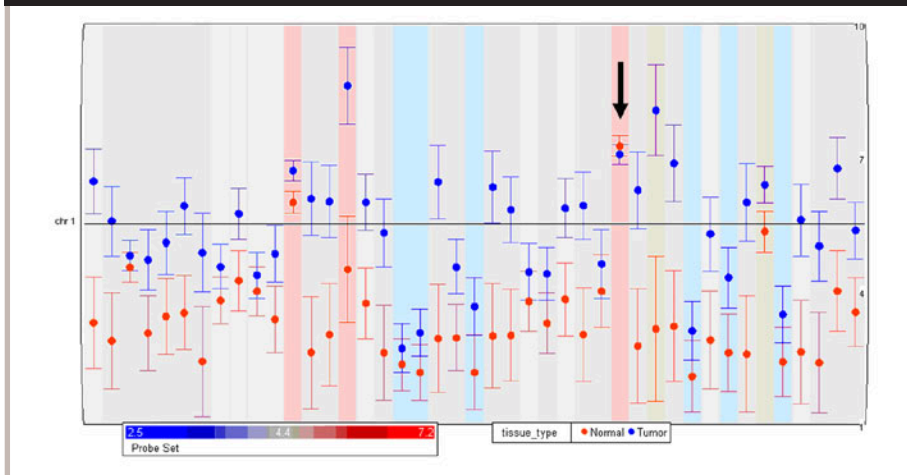
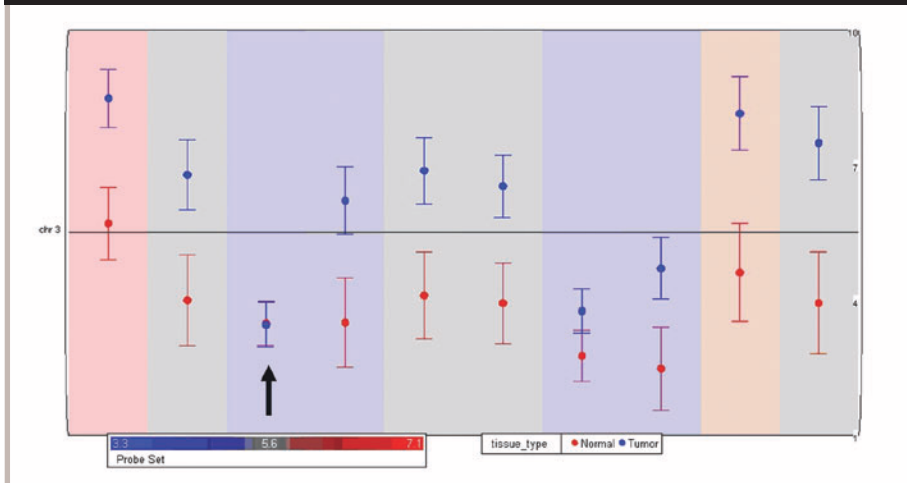


Figure 15: A probable artifact due to non-responsive or non-linear probes. The indicated probe set appears to be alternatively spliced since its signal in tumor is relatively low compared to gene expression in tumor. Consequently, the exon appears to be 100 percent included in normal but only about 15 percent in tumor (the scale is log2). This probe set displays two trademark symptoms of this artifact: it has the same intensity in both tissues, and it shows low variance across replicate samples.



it out.

In a recent study of colon cancer splicing, this step removed almost two-thirds of the exons (540,000 remained out of 1.4 million probe sets). More stringent filtering options may be implemented to require that the exon is called as Present in more than 75 percent of the samples or have a DABG p-value of less than 0.01.

2. Remove genes (transcript clusters) that are not expressed in both groups. (Alternative splicing is meaningless unless the gene is expressed in both sample groups.)

This is more complicated, because we don't presently have a direct way of making Present/Absent call at the gene level. Instead, an approach utilizes the exon-level DABG results to indirectly estimate Present/Absent for the gene. In this case, we use the Core exons as a surrogate for gene expression (but this is only valid for Core genes).

The following is a good default set of rules to remove genes expressed at low level:

1. If the DABG p-value of the probe set is less than 0.05, the exon is considered as Present in a gene.
2. If more than 50 percent of the Core exons are called as Present, the gene is considered Present.
3. If the gene is called as Present in more than 50 percent of the samples in a group, the gene is considered expressed in that group.
4. If the gene is expressed in *both* groups, accept it; otherwise, filter it out.

This filtering step is easily made more stringent by increasing the percentage of samples (e.g., more than 75 percent) in which the gene is expressed, and/or by lowering the cut-off DABG p-value for detection. Assuming that the gene is in fact alternatively spliced, you may not want to increase the "50 percent of Core probe sets" requirement since it is likely that some of the exons will be skipped.

In a recent study of colon cancer splicing, this step removed almost half of the Core

genes (9,000 remained out of 17,800 Core transcript clusters). If Full gene sets are used for analysis, it is anticipated that a much larger proportion of the speculative content will be removed.

SECONDARY (SUGGESTED) FILTERING

1. Discard probe sets with high potential for cross-hybridization.

Probe sets that have the potential for cross-hybridization are more likely to result in false positives. This is because the signal from the probe set may originate from an entirely different gene that may have tissue-specific expression. On the exon array, each probe set has a value of potential to cross-hybridize in the annotation files. Probe sets with cross-hybridization values other than 1 might be candidates for removal from the analysis.

2. Require a minimum gene expression intensity of greater than 15.

Results from genes with low expression values close to the noise level may not be as reliable. Genes with higher expression values tend to have less noise. The actual optimal cut-off value is likely to be dependent on the experiment (and how the gene-level values are calculated). This value can be increased if higher stringency is desired.

In general, one can estimate the background gene signal and set an absolute threshold for calling the gene Present. In this case:

1. If the gene signal is greater than THRESHOLD, the gene is called as Present.
2. If the gene is called as Present in more than 50 percent of the samples in a group, the gene is expressed in that group.
3. If the gene is expressed in *both* groups, accept it; otherwise, filter it out.

3. Discard probe sets with very large gene-level normalized intensities (exon/gene > 5.0).

This filter relates to probes with potential cross-hybridization, but also includes probes with high background levels. The theory behind this filter is that exons having intensities much higher than the

median intensity across the gene are likely to have a signal that originates from somewhere outside of the gene. While there should be some consideration for probes with varying affinities (probe response), intensities that are dramatically different from the rest of the exons belonging to the same gene are certainly questionable. Even given the potential variability in probe response, accepting a value much beyond 5.0 (e.g., the exon is five times higher than the gene intensity in an absolute comparison sense) is not recommended. Large values for the gene-level normalized intensities can also result in somewhat inflated magnitudes of change.

4. Discard probe sets with very low gene-level normalized intensities in the group that is predicted to have the higher rate of inclusion (exon/gene < 0.20).

The theory behind this filtering step is somewhat related to the argument above, that exon intensities that are much different from the median intensity are unusual. However, it is a bit more complicated on the low end. It is entirely feasible that an exon could be included in only, say, 10 percent of transcripts in group A, but in only 1 percent of transcripts in group B. This difference still represents a 10-fold higher inclusion rate in group A, and the exon could still be considered enriched in group A. However, if our interest is in reducing the false positive rate and finding suitable validation targets, this class is better off being filtered out. In most of these cases it is more likely that the exon is simply not expressed at all.

TERTIARY (OPTIONAL) FILTERING

1. Remove genes that have very large expression differences between the two groups.

Despite the fact that the Splicing Index algorithm corrects for gene expression level, genes with very large differentials in gene expression between the two groups have a tendency to produce false positives. Large disparities in transcription rate make it more likely that probe set intensities in the two groups are disproportionately affected by background noise or saturation. A 10-fold difference is an arbitrary

cut-off that can be decreased if a less stringent filtering is desired:

$$|\log_2(\text{gene} / \text{gene})| > 3.32$$

At the extreme, we might begin by looking only at genes with approximately equal expression.

2. Limit search to “Ensembl/RefSeq Supported” or “Core” Exons.

Narrowing the set of analyzed exons to only well-annotated exons (“Core” exons) is a way of dramatically reducing the amount of speculative content. Because speculative exons have much lower levels of annotation/sequence support, they are more likely to result in false positives. The trade-off with this filter is that you will lose the ability to discover new splicing events involving novel exons. It is possible, however, to uncover new examples of alternative splicing that involve well-annotated exons (novel skipping of a known exon, for example.)

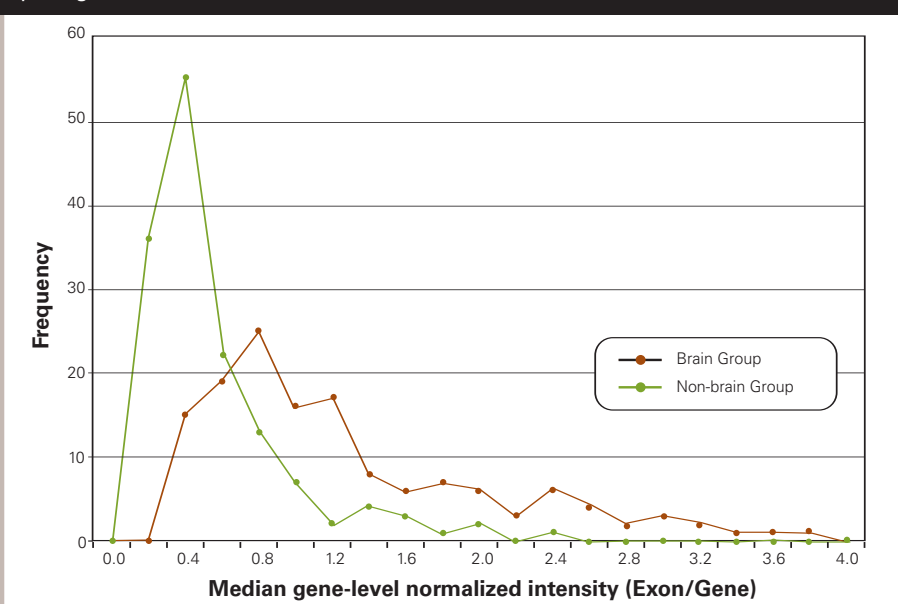
3. Increase stringency on gene expression level filter to restrict the search to only highly expressed genes.

Genes with higher expression levels have exon intensities that are further away from noise and thus are more reliable. A reasonable, but very conservative, threshold would be greater than 100. This filter can be used to find the lowest of the “low-hanging fruit,” but at the expense of potentially interesting alternative splicing events in low or moderately expressed genes.

4. Focus on exons that have gene-level normalized intensities near 1.0 in the group predicted to have a higher inclusion rate and gene-level normalized intensities close to 0 in the group predicted to have a lower inclusion rate.

This filter is a bit theoretical, but does have some basis in fact. For the sake of simplicity in this example, let’s assume that all probes have equivalent affinities (probe response). Let’s also assume that you are looking for a dramatic change in splicing pattern such that an exon is fully included in all transcripts in one group and fully skipped in all transcripts in the other group. In this perfect case,

Figure 16: Histogram of gene-level normalized intensities for validated alternative splicing events.



the exon would be expected to have a gene-level normalized intensity of 1.0 (exon intensity equivalent to gene expression level) in group A, and a gene-level normalized intensity of 0 (exon not expressed) in group B.

Real data is, of course, not this idealized. However, there is some empirical support for this theory. Figure 16 shows a histogram of gene-level normalized intensities for validated brain-enriched exons in both the brain samples (red) and non-brain samples (green). The peak for the brain samples (exon included) is between 0.8 and 1.0, and the peak for the non-brain samples (exon skipped) is between 0.2 and 0.4.

5. Filter probes with unusually low variance.

This filter seems a bit counterintuitive at first, but the idea is that probes with very low variance (relative to others within the gene) across the samples are very likely to be either absent in all samples (not expressed) or saturated in all samples. If the signal is above background it can escape the DABG filter even though the exon is “Absent.” By definition, an alterna-

tively spliced exon will have probe set intensities that change between samples (Present in one sample, Absent in another). This will also tend to eliminate probes that are cross-hybridizing or non-responsive.

6. Limit search to known alternative splicing events.

One way of quickly screening for candidates that are most likely to result in positive validations is to limit the analysis set to exons that have prior evidence of being involved in an alternative splicing event. This filter is more difficult to implement since it relies on bioinformatic predictions of alternative splicing based on EST/mRNA sequences or annotations. The UCSC Genome Browser is one good source of transcript and alternative splicing predictions. In addition, there are a number of publicly accessible databases of alternative splicing events available on the Internet. It may be possible for this prediction to be run once to create a list of probe sets that could be used to filter all subsequent analyses. It should be pointed out that most exon-exon junction arrays are filtered in this way by default since they are typically designed to observe

junctions. Thus, it might be expected that they would have higher validation rates on the whole compared to an exon array design. Using this filter does, however, surrender the ability to discover new alternative splicing events.

OVERALL STRATEGIES

It is important to remember that after the CEL files are produced, the most laborious step in the analysis is validation in the laboratory, e.g., by RT-PCR. Therefore, it may be wise to begin with an extremely conservative analysis that minimizes false positives (i.e., using all or most of the recommended filtering steps), then relax the search for candidate splicing events until the false positive rate becomes unacceptable.

This strategy may be modified depending on time constraints since it postulates iterative cycles of analysis/validation. The final success rate will likely be heavily dependent on sample quality, sample size and the inherent biological differences between the sample groups.

As an example, a study comprising a panel of relatively pure normal tissue samples produced a validation rate of 85 percent, while a much noisier comparison of colon tumor versus normal tissue demonstrated a validation rate of 35 percent. Researchers might also have particular interests that guide their strategies, e.g., searching for targets of a particular splicing factor or alternative splicing events in a particular pathway.

SPlicing EVENTS THAT MAY BE MISSED BY THE ANALYSIS

Here is a partial list of situations where a real splicing event may be missed (false negative) by current splicing algorithms along with brief discussions of each:

▪ Alternative splicing outside of transcript clusters

The MiDAS uses transcript cluster information to generate the gene-level call, thus any probe set outside of a transcript cluster is excluded from the

analysis. Depending on how you define “transcript cluster,” the excluded probe sets represent a significant portion of the total number of probe sets on the exon array.

▪ No probe set for the alternatively spliced exon

There are several reasons why an exon may not be represented by a probe set on the exon array. Some of the reasons for lack of probe set include very small exons (less than 25 bp), over-fragmented exons resulting in multiple PSRs that are all below the minimum size and lack of evidence (or prediction) for existence of the exon. It is also possible that the sequence of the exon made it impossible to build probes, the designed probes give very weak signal or the sequence of the exon is repeated elsewhere in the genome such that data from the probes for that exon were discarded or filtered out.

▪ Alternatively spliced product is rapidly degraded

It is possible that mRNA, including an alternatively spliced exon, is turned over at a high rate so that the signal is not detectable by the array. In many cases, inclusion of an exon (or mis-splicing) alters the protein coding reading frame or incorporates a premature termination codon (PTC). Cells have a mechanism called Nonsense Mediated Decay (NMD) for detecting and destroying these messages. It has also been shown that in some cases this mechanism is exploited as a means of regulating gene expression: purposeful inclusion of a PTC so that the mRNAs are degraded by NMD to silence expression of the gene.

▪ Only a fraction of annotated exons from a gene are expressed

Many genes have alternative transcriptional starts or alternative 3' ends. In some cases, these alternative starts and stops may result in only a fraction of the well-annotated exons in a transcript cluster being expressed. Our filtering approaches require that more

than 50 percent of the well-annotated exons within a transcript cluster be detected above background for the gene to be considered as expressed. Thus, if expression involves fewer than half of the exons, the algorithm will incorrectly call the gene as absent.

SUMMARY

Exon arrays are powerful tools that enable researchers to monitor genome-wide gene expression and alternative splicing beyond classical microarrays. By following the basic guidelines in this Technical Note, novel splicing events may be uncovered that are critical to biology and disease studies, adding a new dimension to genome research.

FAQs

1. Why should DABG not be used for gene-level Present/Absent calls?

There is a strong assumption in DABG that all the probes are measuring the same thing (i.e., the same transcript). This is not the case at the gene level due to alternative splicing. For example, probes for a cassette exon that is skipped will contribute to a misleadingly insignificant p-value.

2. Can we determine frameshifts or the introduction of nonsense codons by alternative splicing events?

No. The resolution of the Human Exon 1.0 ST Array is not nearly sufficient to determine single nucleotide changes. This would require junction-type arrays, which focus on specific events that are known a priori. The Human Exon 1.0 ST Array is designed more for genome-wide discovery of large-scale alterations in transcript structure.

3. What validation rate should be expected from the analysis?

A validation rate of 80 percent would be excellent, but rates down to 30 percent might be acceptable in discovery-based research. Generally, the predicted splicing events that consistently survive different types of filtering are more likely to be true positives. The acceptable validation rate will be a balance between the researcher's available time and

resources to the validation versus the desire to extend the limits of the search. As a first discovery tool, exon arrays will generate information not previously possible with other traditional technologies. In some cases, the discovery of even 10 new splicing events in an exon study might be highly significant. As more data sets become available on exon arrays, it is anticipated that further algorithm development will become more sophisticated and the validation rate will improve.

4. Can MiDAS handle multiple sample groups?

The ANOVA in MiDAS can compare multiple groups, e.g., brain, kidney and heart. It does not handle additional factors like gender, tumor stage, etc. More sophisticated ANOVA methods have been implemented in third-party packages.

REFERENCES

- Technical Note, GeneChip® Exon Array Design
White Paper: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation
White Paper: Exon Probe Set Annotations and Transcript Cluster Groupings v1.0
White Paper: Gene-Signal Estimates from Exon Arrays
White Paper: Alternative Transcript Analysis Methods for Exon Arrays
Technical Note, Statistical Algorithms Reference Guide
Gardina, P. J., *et al.* Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics* 7:325 (2006).
Srinivasan K., *et al.* Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods* 37(4):345-59 (2005).

A complete listing of GeneChip®-compatible™ software products for exon applications can be found at http://www.affymetrix.com/products/software/compatible/exon_expression.affx.

AFFYMETRIX, INC.

3420 Central Expressway
Santa Clara, CA 95051 USA
Tel: 1-888-DNA-CHIP (1-888-362-2447)
Fax: 1-408-731-5441
sales@affymetrix.com
support@affymetrix.com

AFFYMETRIX, UK Ltd.

Voyager, Mercury Park,
Wycombe Lane, Wooburn Green,
High Wycombe HP10 0HH
United Kingdom
UK and Others Tel: +44 (0) 1628
552550
France Tel: 0800919505
Germany Tel: 01803001334
Fax: +44 (0) 1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com


AFFYMETRIX JAPAN K.K.

Mita NN Bldg., 16 F
4-1-23 Shiba, Minato-ku,
Tokyo 108-0014 Japan
Tel: +81-(0)3-5730-8200
Fax: +81-(0)3-5730-8201
salesjapan@affymetrix.com
supportjapan@affymetrix.com

www.affymetrix.com Please visit our web site for international distributor contact information.

For research use only. Not for use in diagnostic procedures.

Part No. 702422 Rev. 1

©2006 Affymetrix, Inc. All rights reserved. Affymetrix®,  GeneChip®, HuSNP®, GenFlex®, Flying Objective™, CustomExpress®, CustomSeq®, NetAffx™, Tools To Take You As Far As Your Vision®, The Way Ahead™, Powered by Affymetrix™, GeneChip-compatible™, and Command Console™ are trademarks of Affymetrix, Inc. Products may be covered by one or more of the following patents and/or sold under license from Oxford Gene Technology: U.S. Patent Nos. 5,445,934; 5,700,637; 5,744,305; 5,945,334; 6,054,270; 6,140,044; 6,261,776; 6,291,183; 6,346,413; 6,399,365; 6,420,169; 6,551,817; 6,610,482; 6,733,977; and EP 619 321; 373 203 and other U.S. or foreign patents.