BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 15 2008, pages 1707–1714
doi:10.1093/bioinformatics/btn284

Gene expression

FIRMA: a method for detection of alternative splicing from exon array data

E. Purdom[1,*], K. M. Simpson[2], M. D. Robinson[2,3], J. G. Conboy[4], A. V. Lapuk[4] and T. P. Speed[1,2]

[1]Department of Statistics, University of California at Berkeley, 367 Evans Hall #3860, Berkeley, CA 94720–3860, USA, [2]The Walter and Eliza Hall Institute, 1G Royal Parade, Parkville, Victoria, 3050, [3]Department of Medical Biology, University of Melbourne, Parkville, Victoria 3010, Australia and [4]Life Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

Received on February 25, 2008; revised on May 18, 2008; accepted on June 6, 2008

Advance Access publication June 23, 2008

Associate Editor: David Rocke

ABSTRACT

Motivation: Analyses of EST data show that alternative splicing is much more widespread than once thought. The advent of exon and tiling microarrays means that researchers now have the capacity to experimentally measure alternative splicing on a genome wide level. New methods are needed to analyze the data from these arrays.

Results: We present a method, finding isoforms using robust multichip analysis (FIRMA), for detecting differential alternative splicing in exon array data. FIRMA has been developed for Affymetrix exon arrays, but could in principle be extended to other exon arrays, tiling arrays or splice junction arrays. We have evaluated the method using simulated data, and have also applied it to two datasets: a panel of 11 human tissues and a set of 10 pairs of matched normal and tumor colon tissue. FIRMA is able to detect exons in several genes confirmed by reverse transcriptase PCR.

Availability: R code implementing our methods is contributed to the package aroma.affymetrix.

Contact: epurdom@stat.berkeley.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Alternative splicing is thought to have several roles in complex organisms, primarily in increasing protein diversity (Maniatis and Tasic, 2002). It can affect the intracellular localization, binding properties or stability of a protein, or regulate its expression via nonsense-mediated decay (NMD) (Stamm *et al.*, 2005). These events usually occur in a regulated manner, but if an aberrant splicing event occurs, it can be causative for, or symptomatic of, disease. More than 15% of heritable human diseases are known to be associated with mutations in splice sites or in splicing regulatory elements (Matlin *et al.*, 2005). In particular, aberrant pre–mRNA splicing events are known to be implicated in several types of cancer (Brinkman, 2004; Venables, 2004).

Previously thought to be a relatively uncommon phenomenon, alternative splicing has recently been shown to be widespread throughout the genome. Analyses of data on human expressed sequence tags (ESTs) give estimated lower bounds between 35% and 59% for the proportion of genes which have at least one splice variant (Modrek and Lee, 2002). The frequency of *functional* alternative splicing events is probably lower than this. Several groups have searched for alternative splicing events conserved between human and mouse, and their results suggest that the proportion of functionally alternatively spliced genes is ~10% (Sorek *et al.*, 2004; Sugnet *et al.*, 2004; Yeo *et al.*, 2005). A weakness of all EST-based methods is that they are biased towards genes which have greater EST coverage (Modrek and Lee, 2002).

Several kinds of alternative splicing have been observed (see Black, 2003, for a recent review). The most common form is skipping or inclusion of one or more 'cassette' exons (roughly 40–50% of cases based on bioinformatic evidence (Clark and Thanaraj, 2002; Sugnet *et al.*, 2004), these being exons which are wholly present in some transcripts, and wholly absent in some others. Alternatively, mutually exclusive cassette exon usage can take place; e.g. exon A or exon B forms part of the transcript, but never A and B together (more generally, multiple exons can exhibit mutual exclusivity). Usage of alternative $3'$ or $5'$ splice sites can result in shortening or lengthening of an exon. Other types of alternative splicing that have been observed are alternative promoter usage, alternative polyadenylation sites and intron retention. Additionally, any combination of the above may occur in an alternatively spliced transcript (Black, 2003).

Skipping or inclusion of internal cassette exons is the most common kind of alternative splicing, and possibly the easiest to detect and verify. For this reason, we have focused on identifying specific exons showing patterns of differential alternative expression and have not approached the problem of reconstructing more complicated transcript patterns.

Our algorithm FIRMA has been developed for analyzing the Affymetrix exon array, Santa Clara, California, USA, which queries the expression level of well annotated and as well as predicted exons. In brief, FIRMA scores each exon as to whether its probes systematically deviate from the expected gene expression level. With a small number of probes per exon (four or less), this is a challenging microarray platform to analyze—such deviations can

[*] To whom correspondence should be addressed.

come from a myriad of biological and technical factors unrelated to alternative splicing. We show that FIRMA performs well in detecting exon-specific changes in expression and therefore can contribute substantially to the detection of regulated alternative splicing. Of course a single scoring method can only be one step in the analysis, and any results must be evaluated in the light of these other complications.

2 MATERIALS

The GeneChip® Human Exon 1.0 ST (sense target) array is a whole-genome array, containing over 1.4 million probesets of up to four perfect match (PM) probes each, spread across exons from all known genes, plus a number of additional regions based on other annotation sources, including GENSCAN predictions and ESTs from dbEST. In the design phase, sequences from all the annotation sources were mapped to the July 2003 version of the human genome (UCSC hg16, NCBI 34). Regions which had some evidence from one or more sources for being transcribed were divided into probe selection regions (PSR) according to the presence of canonical splice sites, CDS start and stop positions or polyadenylation sites. Probes were then selected from within PSRs >25 bp in length. Each PSR corresponds to a probeset, which generally contains four possibly overlapping probes (sometimes fewer). About a quarter of the probesets are based solely on EST evidence, while another quarter are based solely on GENSCAN predictions (GeneChip® Exon Array Design Technical Note, Affymetrix).

The array contains only PM probes, with a small number of generic mismatch probes for the purposes of background correction. There are no probes which span exon–exon junctions.

Association of probesets with genes is not made at design time. Instead, these ‘main-design’ probesets are annotated afterwards, using their alignment to the genome (Exon Probeset Annotations Whitepaper, Affymetrix). This process has been undertaken by Affymetrix, first for NCBI Build 34 of the genome, and more recently for Build 35. The result is that each probeset is assigned to a ‘transcript cluster’, and also has an annotation quality indicator associated with it.

3 ALGORITHM

Our method is developed to evaluate levels of alternative splicing for the situation where there are no replicates nor pre-defined groups in the samples or alternative splicing does not consistently follow the groupings that exist. The second situation can be quite common, for example in disease versus normal where alternative splicing may exist in only a proportion of the diseased samples for a given gene. Alternatively there may be groupings, such as tissue type, where patterns of alternative splicing are shared amongst several tissue types, but the tissue types that share a similar splicing pattern may be different in different genes. For this reason, our algorithm is sample-by-exon specific: each exon and sample pairing is given a score that is comparable across either samples, genes or exons.

3.1 Alternative splicing detection method

The two major steps in our method are estimation of the expression levels of each gene, using the robust multichip analysis (RMA) approach (Irizarry *et al.*, 2003), and detection of alternative splicing using a suitably defined score from auxiliary information from the estimation step. We call the combined approach finding isoforms using robust multichip analysis (FIRMA).

RMA itself involves three steps: background correction, normalization and summarization of the probe-level data (Irizarry *et al.*, 2003). The following discussion assumes that the first two steps have already been performed.

We extract the normalized probe-level data for the probes belonging to a transcript cluster. The exact set of probesets which are used may depend on the aim of the experiment. If this is detection of novel exons, then all probesets might be used (though note the problems with non-expressed regions mentioned in the Discussion). If alternative splicing of well-annotated exons is of more interest, then the analysis might be restricted to well-annotated probesets. In the development below, we will refer to ‘exons’, though in fact the analysis is done on probesets, which usually coincide with an exon.

The final step in RMA is to estimate the gene expression level of each sample by fitting the following additive model for each gene:

$$\log_2(PM_{ik}) = c_i + p_k + \epsilon_{ik}, \qquad (1)$$

where c_i is the chip effect (expression level) for chip i, p_k is the probe effect for probe k, (and which can be interpreted as a relative probe affinity if we use the constraint $\sum_k p_k = 0$), and $\log_2(PM)_{ik}$ is the log (base 2) of the background-corrected, normalized PM signal for probe k on chip i (Irizarry *et al.*, 2003). The model is fitted using iteratively reweighted least squares (IRLS) (Marazzi, 1993).

For the exon array, we can consider a more general additive model which includes the possibility of alternative splicing or different levels of expression per exon,

$$\log_2(PM_{ijk(j)}) = c_i + e_j + d_{ij} + p_{k(j)} + \epsilon_{ijk(j)}, \qquad (2)$$

where (again assuming a zero-sum constraint for these parameters) e_j is the relative change in exon expression for exon j, d_{ij} is the interaction between chip and exon giving the relative change for sample i in exon j, and $p_{k(j)}$ is the *nested* relative probe effect for the k-th probe in exon j.

The parameter d_{ij} indicates the discrepancy of a given sample in exon j from the expected expression for that exon. It is large values of this parameter that indicate differential alternative splicing. Rather than fit this extended model and estimate d_{ij} explicitly, we propose to fit the standard RMA model in (1) for the exon array. If there is a large discrepancy in some samples (a large d_{ij}) then we will see this as large residuals for the probes for that sample in that exon.

In this way, we frame the problem of detecting alternative splicing as a problem of outlier detection, rather than estimation of an interaction effect. By robustly fitting without the term d_{ij} we avoid the additional noise that would be added to all of our parameter estimates, since there are at most four observations to estimate this term. We do assume, however, limited levels of alternative splicing so that the other terms in the model are still well estimated with our robust estimation procedure even though d_{ij} is excluded from the model.

Based on this logic, let

$$r_{ijk} = y_{ijk} - \hat{c}_i - \hat{p}_k,$$

be the residuals from fitting the standard model in Equation (1). Then for each exon j and sample i, a summary score based on the four residuals from exon j and sample i gives a measure of the discrepancy d_{ij} in the expression of the exon in that sample.

Any number of scoring functions could be used. The most obvious choice is the mean. More robust alternatives would be the median residual, the lower quartile or even the smallest of the absolute residuals. We considered these various options in scoring. Ultimately, we determined that the median of the residuals in an exon gave the best tradeoff between sensitivity to the size and sign of the residuals and robustness to the small number of probes (see Section 4, below, for simulation results that compare the scores). This gives us a final score statistic,

$$F_{ij} = \underset{k \in \text{exon } j}{\text{median}} \quad r_{ijk}/s.$$

The estimate of standard error, s, is estimated by the median absolute deviation (MAD) of the residuals and helps to make the scores comparable between different genes.

Note that the term e_j is not estimated separately if we fit the standard RMA model—it will be absorbed into the probe estimates. Large values of the parameter e_j can also indicate alternative splicing, but only where the overall shift in level of expression of the exon is shared amongst all of the samples. The FIRMA scores could be adjusted to include this information but so far we are not convinced that the data is clean enough for the benefits to outweigh the additional noise. Often a shift in expression level will come from the many possible complications in the annotation or the probe definitions in the array, and with at most four probes it will be difficult to isolate the biological signal from these technical problems (see Section 5, below). Since most experiments will be designed to find differential splicing amongst the samples included, rather than shared effects, this should not be a problem. See Supplementary Materials for an illustration of this effect on the UNR gene discussed subsequently.

Because existing bioconductor functions used to analyze GeneChip® data require all the data to be in memory at once, we could not use them because of the size of the Human Exon array. Instead we implemented the FIRMA algorithm in the `aroma.affymetrix` package for large datasets which makes use of persistent memory `aroma.affymetrix`. See Supplementary Materials for more information.

3.2 Other splicing detection methods

Numerous other alternative splicing detection algorithms have been proposed. Many of the techniques extended linear models similar to our motivating model in Equation (2).

Affymetrix proposes several techniques based on their proposal for fitting gene and exon expression levels (Alternative Transcript Detection Whitepaper, Affymetrix). Their general approach differs from our RMA approach in two ways. First they use their algorithm PLIER (rather than IRLS) to robustly fit the standard linear model in (1). Second they propose two separate estimations using the standard linear model in Equation (1): first fit the standard model using all of the probes in the gene to get an estimate of the gene intensity for each sample ($\hat{G}_i$) and then fit the same model but only on the probes in a particular exon to get an estimate of the exon intensity in each sample ($\hat{E}_{ij}$). We can interpret the estimates in terms of our general model (2). We can see that $\log(\hat{G}_i)$ is an estimate of c_i and $\log(\hat{E}_{ij})$ is an estimate of $c_i+e_j+d_{ij}$. Affymetrix defines the normalized exon intensity as $NI_{ij} = \hat{E}_{ij}/\hat{G}_i$, which implies that the log of the normalized exon intensity ($\log NI$) estimates e_j+d_{ij}.

While the normalized exon intensity gives a sample-by-exon score, Affymetrix does not propose the normalized exon intensity as a score for alternative splicing. Instead they propose statistics that give a score *per exon*: the pattern-based correlation (PAC) statistic and microarray detection of alternative splicing (MIDAS). PAC is the correlation across samples of $\hat{E}_{ij}$ and $\hat{G}_i$, $PAC_j = \text{cor}_i(\hat{E}_{ij}, \hat{G}_i)$. MIDAS is an ANOVA test for differences in the group means of the $\log NI$. MIDAS obviously requires predefined groupings or replicates of the samples.

From the definition, it is clear that PAC works best when there are enough differentially spliced samples to significantly weaken the correlation between gene and exon expression levels. In general, we found the PAC to be a weak measure of alternative splicing for an exon, both in simulations and in datasets we examined, probably because only a small proportion of samples were differentially spliced in any gene.

Other methods are related to alternative splicing but address different questions or platforms. DECONV (Wang *et al.*, 2003) is an algorithm which attempts to estimate the relative concentrations of known isoforms by maximum likelihood methods. ANOSVA (Cline *et al.*, 2004) assumes replicated samples and fits all of the parameters in the exon model in Equation (2) and then tests for d_{ij} that are non-zero. Other methods developed recently are the correlation-based method of Le *et al.* (2004), and the Bayesian method (GenASAP) of Shai *et al.* (2006), who attempt to estimate relative expression of two isoforms in the same sample, a difficult problem. Their algorithm was developed for a custom array containing splice junction probes, and gives good agreement with RT-PCR assays undertaken for validation purposes.

4 IMPLEMENTATION

4.1 Simulation study

To test the scoring methods discussed in Section 3.1, we devised a simulation model. All analyses were done in R (R Development Core Team, 2006). Synthetic data were generated and our FIRMA scores, as well as Affymetrix's PAC and *NI* scores, were calculated from these data. We did not test MIDAS or ANOSVA, as both require replicated samples, and our aim here is to detect alternative splicing without replication.

Data are simulated according to the following model:

$$y_{ij} = \log_2(B_j + I_{ij} \times 2^{(c_i+p_j)}) + \epsilon_{ij}, \tag{3}$$

with y_{ij} the $\log_2(PM)$ for chip i and probe j. The additive background, B_j, is modeled as a log-normal variable, the chip effect c_i as a normal variable, and the probe affinity p_j and the residuals ϵ_{ij} as mean-zero normal variables. The indicator variable I_{ij} is 1 when probe j is expressed on chip i, and 0 otherwise.

This model features additive background, multiplicative noise, and probe-specific affinities. We chose values for the simulation parameters by obtaining rough estimates of 'typical' values from real data We simulated a gene with 10 exons (four probes per exon) with six variants, each one with one fewer exon than the preceding one. When a variant was included in the data, we set $I_{ij} = 0$ for all four probes belonging to the dropped exon. See Supplementary Materials for more details regarding the implementation.

500 simulations were performed for each of two different values of the mean chip effect (7 and 10) and four different probabilities of including a splice isoform ($P = 0.1$, 0.3, 0.5, 0.8), i.e. eight cases in total. The two values of the mean chip effect were chosen to mimic differing scenarios—one where the expression is close to background, and one where it is far above background. Forty chips were simulated for each run, and each was randomly selected to

have either the full transcript or one of the six variants (the variant to be included was then also chosen randomly).

Four different summaries of the residuals mentioned earlier were tested on the simulated data: the mean and the median of the residuals and the minimum and the lower quartile of the absolute residuals belonging to a probe in the exon. We also calculated a version of Affymetrix's log*NI*, with estimates $\log(\hat{E}_{ij})$ and $\log(\hat{G}_i)$ given by the IRLS estimation used by RMA rather than PLIER for ease of comparison. These are all single-sample methods, i.e. they aim to detect alternative splicing in each exon for each sample.

We note the log*NI* as a single-sample technique which has some possible advantages in our simulations. In particular, if the proportion of samples showing alternative splicing is high within an exon (say in the majority of samples), the high residuals will be found not in those samples classified by the simulation as spliced out, but rather the complementary set of samples; in such cases, the FIRMA values will call the wrong set of samples spliced. The log*NI* index, as we mentioned earlier, will also estimate which samples are different from the overall gene expression; thus for high proportions of splicing in the simulation, it may have an advantage in calling individual samples spliced. In reality, which samples are considered 'spliced' in such a situation would be generally a question of definition.

We also did some brief comparison of all-sample methods—methods that seek to pinpoint exons with alternative splicing, but do not determine in which samples the exon is alternatively spliced. The PAC is an example of such a method, as is the standard deviation (SD) of the log *NI*. As a comparison, we tried different summaries of our FIRMA score (based on the median residual): the maximum FIRMA score among the samples, the fourth largest among samples, and the SD across samples.

Table 1 summarizes the area under the Receiver Operating Characteristic (ROC) curve created using the single-sample and all-sample methods; note that the ROC curves are not comparable between the two different approaches because the events being classified are entirely different (with different underlying probabilities). There are several observations to make about Table 1. In the single-sample methods, the variants of the FIRMA scores perform better than the log*NI*, despite its possible advantages in the simulation study when there are high levels of splicing. For all-sample methods, the SD of either the FIRMA scores or the log*NI* scores seems to be competitive and both are consistently well behaved across different probabilities of splicing and expression level.

4.2 Application to real data

We used for our analyses two sets of biological samples evaluated on the Human Exon 1.0 ST chip publicly available from Affymetrix http://www.affymetrix.com/. The first is a set of tissues consisting of 11 different tissues (breast, cerebellum, heart, kidney, liver, muscle, pancreas, prostate, spleen, testes and thyroid) with three technical replicates of each. The second is a set of 10 matched normal-tumor colon cancer pairs analyzed by Gardina *et al.* (2006) (20 arrays).

We evaluated the performance of the algorithm on genes validated to have alternative splicing. This allowed us to determine, case-by-case, whether the algorithm gives sensible answers. We performed genome-wide searches for high scoring probesets to give perspective on how the scores for these validated genes compared to other genes on the array. In addition, our genome-wide search gave us candidate probesets—examples of probesets that would be found *de novo* by our algorithm—to further evaluate the behavior of the algorithm. However, such a genome-wide search creates a large number of additional questions; for example, what criteria should be used, and how should we exclude poorly behaving genes or exons, and how to evaluate the results with only a very limited knowledge of the truth. For the purposes of this article, we chose to focus on the question of defining a good initial metric to identify alternative splicing within a gene.

Table 1. Area under the ROC curve for the cases described in the text

			Mean expr = 7				Mean expr = 10			
		Method	0.1	0.3	0.5	0.8	0.1	0.3	0.5	0.8
Single-Sample		Median	0.94	0.92	0.89	0.78	**0.99**	0.98	0.96	0.84
		Mean	**0.97**	**0.95**	**0.93**	**0.83**	**0.99**	**0.99**	**0.97**	0.86
		LQ	0.93	0.90	0.86	0.74	0.98	0.97	0.94	0.80
		Min	0.84	0.81	0.77	0.67	0.94	0.92	0.87	0.71
		NI	0.86	0.85	0.83	0.78	0.97	0.96	0.95	**0.88**
All-Sample	FIRMA	Max	**0.87**	0.80	0.73	0.72	0.92	0.82	0.73	0.75
		4th	0.66	0.82	**0.84**	**0.86**	0.70	0.84	**0.86**	**0.90**
		SD	0.85	**0.84**	0.80	0.82	0.92	**0.87**	0.83	0.87
	Affymetrix	SD(NI)	0.86	**0.84**	0.81	0.82	**0.93**	**0.87**	0.85	0.89
		PAC	0.39	0.50	0.54	0.55	0.41	0.50	0.56	0.60

The numbers in the table header indicate the probability of including a variant isoform in any given sample. The largest area under the curve value for each probability (column) noted in bold font for both the single-sample and all-sample methods. Bold values highlight the largest (best) AUC in each column.

Our analysis used the procedure outlined in Section 3.1 and implemented in `aroma.affymetrix`: the chips were individually background corrected and then jointly quantile normalized (using all of the main-design probes) and then the standard RMA model in (1) was fit (per gene) using the gene definitions from Ensembl (Hubbard *et al.*, 2007). This means only the 332 532 probesets that mapped to the Ensembl annotation were retained, which constituted 23 385 gene definitions. FIRMA scores were calculated for each probeset and sample from the residuals from these fits. Our background correction is the standard RMA and does not make use of the control probes, but rather the main-design PM probes. To aid in evaluation of the algorithm we implemented simple filtering of low-expressed probesets (see Supplementary Materials for details). As has been noted in previous work (e.g. Gardina *et al.*, 2006), non-expressed probesets induce false positives on the array: their expression is merely background noise and thus no longer tracks the expression of the gene.

4.2.1 *Panel of tissue samples*

FIRMA was applied to the tissue sample arrays from Affymetrix. Specifically, we asked whether FIRMA could predict muscle specificity for a group of exons previously validated by RT-PCR as being enriched in heart and skeletal muscle (Das *et al.*, 2007). Here we focused on the 11 exons which were contained in our Ensembl mapping. As shown in Figure 1, FIRMA successfully predicted most of the relevant probesets to be muscle-enriched. The observed variability in muscle-enrichment indicated by the color scale is to be expected, since the original RT-PCR validations showed considerable variation in muscle-specificity. Moreover, the predictions of muscle-specificity made in that study utilized a different set of tissue data that was heavily weighted towards brain. In particular based on the

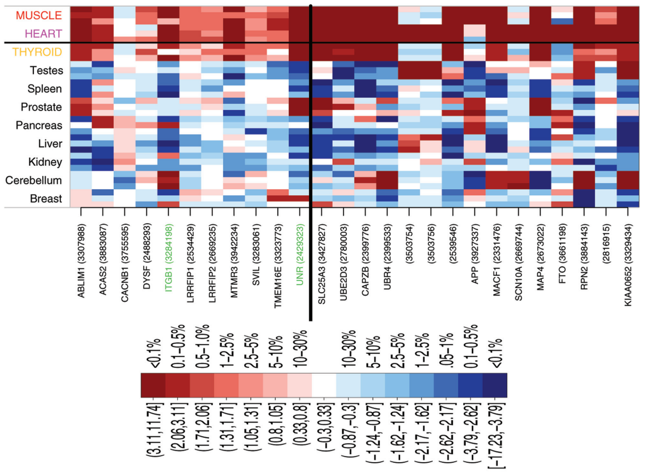


Fig. 1. FIRMA scores represented by color scale for all 11 of the validated probesets (left of dividing line) as well as 15 additional top scoring probesets (right of dividing line). Two validated genes (UNR and ITGB1) also ranked high enough to be included in the top 15 candidate genes and their labels are colored in green to note this. Note that the color scale is not evenly spaced but rather based on percentiles of all FIRMA scores in all non-filtered probesets and samples.

PCR data (Das *et al.*, 2007), CACNB1—which demonstrated no obvious evidence of splicing in the FIRMA analysis—might better have been classified as brain-depleted rather than muscle-enriched. We conclude that FIRMA was effective as a sample-specific measure of splicing in this control study.

We also performed a genome-wide search for high-scoring probesets as a comparison to this set of probesets with confirmed splicing. To provide better comparison to the validated genes, we searched for probesets with enriched expression value in the muscle or heart tissues, as compared to other tissues. We scored each probeset by finding the minimum FIRMA score in each of the two tissue groups and then took the maximum (see Supplementary Materials for details); this created an all-sample score, to use our terminology from above. This score ensured that at least one of the two tissues had uniformly high FIRMA scores in all three replicates. As we mentioned above, there are many issues in a genome-wide search which we do not treat; this technique was a simple way to get additional candidates for comparison.

In Figure 1 we show the FIRMA scores in each of the tissue samples for the 11 confirmed probesets plus 15 additional top candidate genes found *de novo* by FIRMA. Our FIRMA scores did reasonably well in determining alternative splicing, given the evidence available on the exon array. In all cases, except CACNB1, the FIRMA scores of the muscle tissues are decidedly positive, indicating enrichment. UNR, in particular, shows extremely high scores in the muscle and heart and much smaller scores in the other tissues; indeed UNR also has the highest score in our genome-wide scan. We show the example of UNR in more detail in Figure 3. Other tissues, in particular thyroid, also demonstrate high FIRMA scores in many of the probesets, demonstrating the flexibility of not pre-defining the groups in advance. The genes LRRFIP1, LRRFIP2, SVIL and DYSF have only modest FIRMA scores in their muscle and heart tissues relative to those of the other genes. CACNB1 shows no marked FIRMA scores in any of its tissues.

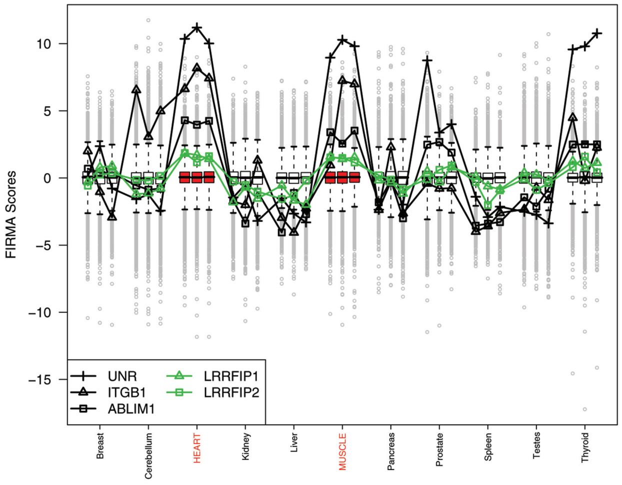
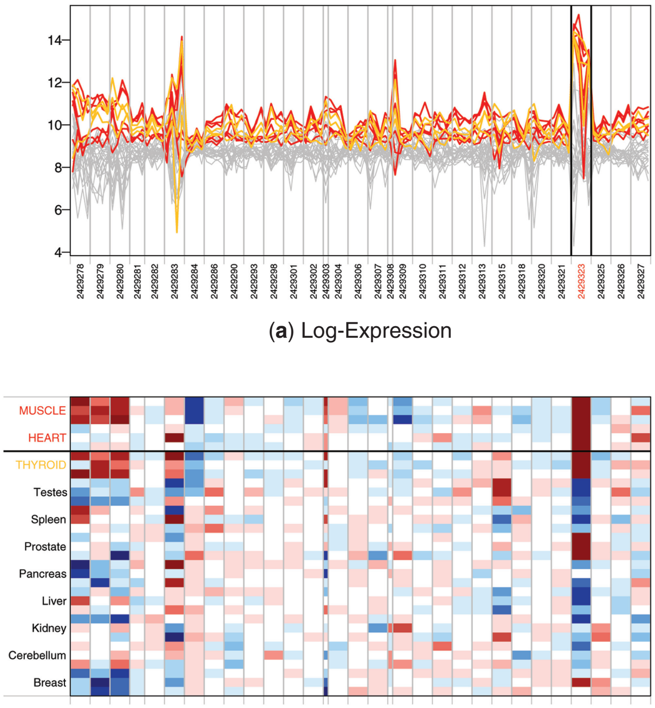


Fig. 2. Boxplots of the FIRMA scores for all probesets in the Ensembl annotation mapping (332 532 probesets) for each of the 33 arrays. The boxplots of muscle and heart tissues are colored red. The scores for the probesets of a few of the validated genes are mapped on top of the boxplots (see legend in graph).



(**a**) Log-Expression

(**b**) Firma Scores

Fig. 3. Analysis for UNR: (**a**) Log expression values for UNR, with the estimated probe effect subtracted off ($\log(PM)_{ijk} - \hat{p}_k$). Probes are plotted along the x-axis, with grey vertical lines separating the probesets; each plotted line corresponds to an array. (**b**) UNR FIRMA scores for all probesets of UNR, aligned to match (a); the color scale is the same as in Figure 1. In both plots, the lines and labels of muscle and heart tissues are colored red and those of thyroid colored yellow.

These results are consistent with their performance on the array based on visual inspection. In Figure 2 we show the scores of a few of these probesets as compared to the scores of all probesets used in our analysis.

On the other hand some probesets show a general spread of FIRMA scores across samples with lower relative scores, despite showing evidence of splicing on visual inspection of the data. Looking closely at the data and residuals for all of the genes (supplied in the Supplementary Materials) we see that this is most common in two related types of cases: (1) when there is a wide spectrum of the level at which the tissues show alternative splicing and (2) when there is a relatively separated level of expression for alternatively spliced arrays, but there are more tissues than just heart and muscle which share in the alternative splicing. In both cases the residuals for many (if not most) of the samples will be large in that probeset, not just the enriched samples, though obviously the scores will have different signs. One such example is the gene ABLIM1. Here, there is an increase in expression of heart and muscle tissue, but also a relatively similar level of increase in thyroid and prostate. The remaining tissues show some variable change in expression in this probeset, all of which is lower than the enriched tissues and some of which is lower than their overall gene expression. As a result, both the high- and low-expressed tissues for this probeset have large residuals, but none have extremely large residuals despite the large increase in expression (see Supplementary Materials for figures demonstrating this behavior). The scores could be adjusted to be more informative as to which arrays are different by either explicitly comparing to the overall estimated gene expression or arbitrarily defining a different point of comparison than the middle (zero-level) residual; however, this could also add a great deal of noise and remains an area of further exploration. Despite these drawbacks, ABLIM1 still ranks in the top 100 probesets in our genome-wide scan (rank = 63).

To evaluate the new muscle-specific probesets predicted by FIRMA, we used the UCSC genome browser to identify and examine the exons in which the probesets reside. The majority of these candidates have properties similar to the known muscle-enriched exons defined in Das *et al.* (2007): they are alternatively spliced, relatively short exons that are evolutionarily conserved among vertebrates, and they possess some of the putative splicing regulatory elements, located in flanking introns, known to be over-represented near other muscle-enriched exons. Two candidate probesets from SLC25A3 and CAPZB also were identified in the muscle-enrichment screen of Das *et al.*, though not selected for their RT-PCR screen. These properties strongly suggest that FIRMA has successfully identified a number of new muscle-specific alternative exons.

Using the package biomaRt in R (Durinck *et al.*, 2005), we classified all of the probesets as to whether they were contained in all the transcripts for that gene in the Ensembl database and compared this classification to their FIRMA scores. This merely gave an indication of the performance of the algorithm and clearly is not equivalent to a true positive rate—just because the region is known to have a splicing event does not mean that it was spliced in our samples. Similarly, we could have identified new splicing regions not in the database. Moreover, the ‘splicing events’ detected by our automatic scan may be quirks of Ensembl’s clustering of transcripts into genes and not actual splicing events.

Despite these many caveats, the proportion of splicing calls that match splicing events in Ensembl is roughly increasing as a function of our genome-wide score, indicating that the FIRMA scores are tracking real splicing events (Figure S5a in Supplementary Materials). If we look at just the top 100 probesets from our genome-wide scan, 70 probesets matched a splicing event in Ensembl, and the genome-wide scores for those that did is somewhat higher than for those that did not (Figure S5b in Supplementary Materials). If the Ensembl database was complete, we hope for far more successes in the top hundred probesets. Manually inspecting the data showed that many of the probesets that did not match an Ensembl splicing event still have interesting expression patterns, suggesting that the algorithm has good prospects of utilizing the Human Exon array to find new splicing events.

4.2.2 Colon cancer data For the colon cancer dataset, Gardina *et al.* (2006) performed RT-PCR on 41 genes chosen based on *t*-tests of the *NI*. Roughly, a third were confirmed as cancer-specific splicing in the samples. Thirty-six of these probesets could be uniquely identified and also matched the probesets we analyzed, 16 of which corresponded to confirmed cancer-specific splicing based on the RT-PCR results (i.e. found by RT-PCR to consistently have a different isoform in the cancer from the normal). Gardina *et al.* also performed RT-PCR on nine additional genes reported in the literature as related to colon cancer. Based on their descriptions, we identified 16 probesets from these previously reported genes, which brought the total number of probesets to 52. This gives a larger number of validations (and wider range of outcomes) than the tissue data. Moreover most of the probesets were selected for further validation based on the same data that we are analyzing.

Because the data consists of paired normal and tumor samples, we took the pairwise difference of FIRMA scores then calculated the mean difference per probeset for our all-sample score. We used the mean rather than the *t*-statistic because the FIRMA score is already on a comparable scale across genes—and this scale is more informative as to the level of splicing relative to the noise of the gene. See Supplementary Materials for more details of the implementation.

In Figure 4 we show the paired difference in FIRMA scores for the set of 52 validation probesets. In bulk, the mean value of the difference in FIRMA scores separated those probesets that confirmed to have cancer-specific splicing from those that were not confirmed. In Figure 5, we plot these probesets on top of the empirical cumulative distribution function (cdf) of the mean difference in FIRMA. Almost all of the probesets are in the extremes of the distribution, which is not surprising given that they were chosen for further validation. Again, this plot shows that the confirmed and the non-confirmed probesets are generally well separated by the mean difference of FIRMA scores (of roughly 1.5 in absolute value), with two notable exceptions of the probesets from SLC3A2 and CTTN.

A more precise measure of the success rate of our algorithm is not really feasible based on just these probesets. The mean difference in FIRMA values nicely separated the confirmed from the non-confirmed probesets, but the ranks of the confirmed probesets in a genome-wide comparison ranged widely (from 47 to 1537 excepting SLC3A2 and CTTN). Among the probesets that ranked higher we find promising candidates not selected by Gardina *et al.* (2006). Filtering choices dramatically change rankings, making rankings difficult to evaluate or compare. Our filtering was fairly conservative, as we tried to not remove the validated probesets; more aggressive filtering will bring down the ranks of the top confirmed probesets but at the cost of also removing some of them

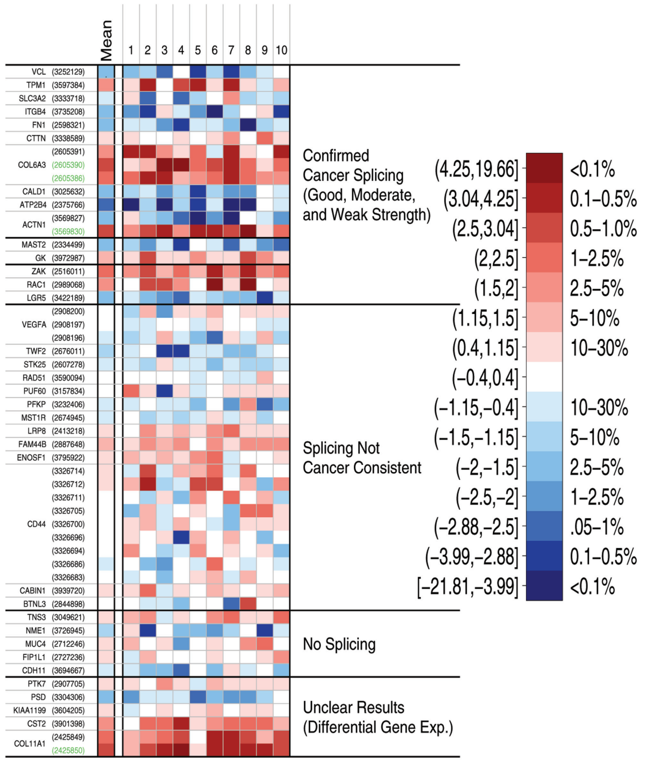


Fig. 4. Paired differences in FIRMA scores (represented by color scale) for all 52 of the probesets from the RT-PCR experiments of Gardina *et al.* (2006). The probesets (rows) are grouped according to their characterizations of the RT-PCR results. Again, the color scale is not equally spaced intervals but based on the percentiles.

from consideration. This highlights the delicate role of filtering in reducing the false positive rate at the cost of losing viable candidates. In fact, Gardina *et al.* chose the probesets to subject to further RT-PCR experiments from many different filtering runs, in addition to a manual review of the data. Ultimately more precise evaluation of the algorithm and of filtering techniques will require more validations of high ranking probesets based on FIRMA scores (or a more complete understanding of the transcriptome for colon cancer).

As a final comparison, we examine the final list of top 200 candidate probesets of Gardina *et al.* (2006) (based on their preferred filtering). Only 10 of the 16 confirmed probesets were contained in this final list. Those 10 probesets ranged from rank 1 to 166 in their own list (compared to rank 47 to 861 with the mean difference of FIRMA). Presumably they chose candidates across a wide range of scores to better evaluate their algorithm; this implies that a wide range of rankings for these probesets is not surprising. Five non-confirmed probesets were also intermingled in their list of top 200 (with ranks ranging from 7 to 121). In comparison, the mean difference of FIRMA scores widely separated the confirmed and non-confirmed probesets (the best rank of the five non-confirmed probesets was 2112). However, with only 15 total probesets in which we can strictly compare the results (all of which were chosen on the basis of their algorithm and very differing filtering), we cannot draw any general conclusion regarding the relative behavior of the algorithms. In fact, much of the difference we see may be due to different normalization and summarization choices (see Supplementary Materials for more details).

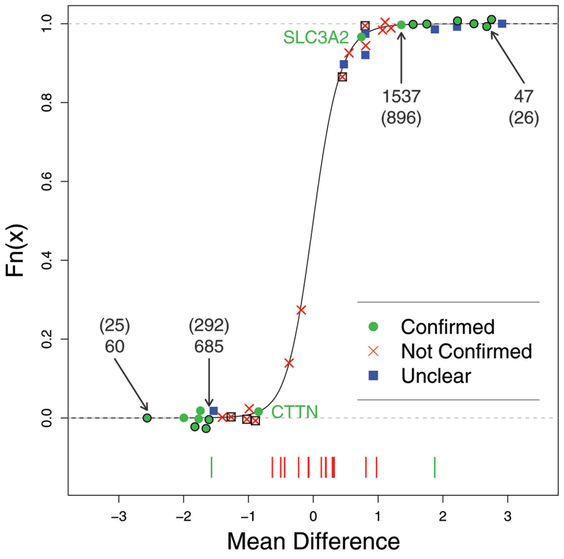


Fig. 5. Empirical cdf of the mean difference of the FIRMA scores for the colon cancer dataset. Superimposed on the plot are the 36 probesets chosen for further validation by Gardina *et al.* (2006); these points are color coded according to the reported RT-PCR results. Points outlined in black are probesets that are also among their final list of top 200 probesets. Below the plot, noted with line segments, are the scores of the probesets previously reported in the literature and also tested by RT-PCR by Gardina *et al.* The rank of a few of the confirmed probesets are marked; in parenthesis is the rank amongst only positive (enriched) or negative (reduced) values.

5 DISCUSSION

It is anticipated that many researchers will be interested in genome-wide studies of alternative splicing under particular conditions. In such cases, a method of ranking genes by evidence of splice variation is needed. We have reported a method, FIRMA, for detecting alternatively spliced exons in individual samples, without replication, from GeneChip® Human Exon 1.0 ST data. We have demonstrated that our method also shows promise in application to real data. It is able to detect known alternatively spliced exons in single samples in complex datasets. We also provide software in an R package to allow general analysis of the array as well as calculate our FIRMA scores.

Our studies so far have not taken into account the fact that in some probesets, some or all probes overlap in sequence. This will occur any time a probeset is designed against a small exon. Overlapping probes introduce additional correlation which may bias alternative splicing detection. The problem is confounded if the overlapping region contains a SNP, for then variability due to the SNP can be identified as alternative splicing (Kwan *et al.*, 2007). Probes containing SNPs could be manually filtered out before fitting the RMA model, but this indicates the problem that overlapping probes create. Additionally, a probe may have a high residual not because it is in an alternatively spliced exon, but because it is poorly performing (either unresponsive, or hyper–responsive due to strong cross-hybridization).

As noted above the problem becomes much more complicated for all of the algorithms when all the probes in a probeset are unresponsive. This is often a problem of annotation—the probeset queries either an intronic region or a non-transcribed region of the genome. From the point of view of the models of alternative splicing, there is little to separate such behavior from legitimate alternative splicing. This can be a problem even when the analysis is limited to probesets with strong annotation support. If the analysis expands to find truly novel patterns of expression and therefore includes many more speculative probesets, this problem can be overwhelming and post-filtering of probesets would be insufficient (see Supplementary Materials for discussion). We are currently investigating ways to filter such completely non-responding probes to give a more coherent framework for analysis.

So far, we have not addressed the issue of determining a threshold for alternative splicing discovery in real data. The threshold for calling an exon alternatively spliced is also not obvious. If we assume the residuals from fitting the standard model in (1) have a normal distribution, the null distribution (in the absence of splicing) of the median of the residuals could be calculated explicitly and form the basis of a cutoff. However, the distribution of the median of less than four observations will be sensitive to the assumption of normality. In a genome-wide study, it will probably be necessary to derive a threshold empirically, for example by estimating the null hypothesis as in Efron (2004) or by using control genes. Our future work will address these issues.

ACKNOWLEDGEMENTS

We would like to thank Henrik Bengtsson for helpful discussions regarding implementation of FIRMA in `aroma.affymetrix`, Steffen Durinck for providing R code for plotting the Ensembl gene transcripts, and Alan Williams for useful discussions regarding the design of the array.

Funding: This work was supported by National Science Foundation, Postdoctoral Research Fellowships in Biological Informatics (to E.P.). T.S. was supported by grants from National Cancer Institute (U24 CA126551-01) and National Institutes of Health (NIH) 5R01GM083084.

Conflict of Interest: none declared.

REFERENCES

Black,D. (2003) Mechanisms of alternative pre–messenger RNA splicing. *Ann. Rev. Biochem.*, **72**, 291–336.

Brinkman,B. (2004) Splice variants as cancer biomarkers. *Clin. Biochem.*, **37**, 584–594.

Clark,F. and Thanaraj,T. (2002) Categorization and characterization of transcript–confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, **11**, 451–464.

Cline,M.S. *et al*. (2004) ANOSVA: a statistical method for detecting splice variation from expression data. *Bioinformatics*, **21** (suppl. 1), i107–i115.

Das,D. *et al*. (2007) A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. *Nucleic Acids Res.*, **35**, 4845–4857.

Durinck,S. *et al*. (2005) Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.

Efron,B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *JASA*, **99**, 96–104.

Gardina,P.J. *et al*. (2006) Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, **7**, 325.

Hubbard,T.J.P. *et al*. (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610—-D617.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kwan,T. *et al*. (2007) Heritability of alternative splicing in the human genome. *Genome Res.*, **17**, 1210–1218.

Le,K. *et al*. (2004) Detecting tissue–specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res.*, **32**, e180.

Maniatis,T. and Tasic,B. (2002) Alternative pre–mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.

Marazzi,A. (1993) *Algorithms, Routines and S Functions for Robust Statistics*. Wadsworth & Brooks/Cole, Pacific Grove, California.

Matlin,A. *et al*. (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, **6**, 386–398.

Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.

R Development Core Team (2006) *R: a Language and Environment for Statistical Computing*. Vienna, Austria.

Shai,O. *et al*. (2006) Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics*, **22**, 606–613.

Sorek,R. *et al*. (2004) How prevalent is functional alternative splicing in the human genome? *Trends Genet.*, **20**, 68–71.

Stamm,S. *et al*. (2005) Function of alternative splicing. *Gene*, **344**, 1–20.

Sugnet,C.W. *et al*. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.*, **9**, 66–77.

Venables,J. (2004) Aberrant and alternative splicing in cancer. *Cancer Res.*, **64**, 7647–7654.

Wang,H. *et al*. (2003) Gene structure–based splice variant deconvolution using a microarray platform. *Bioinformatics*, **19** (suppl. 1), i315–i322.

Yeo,G.W. *et al*. (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl Acad. Sci. USA*, **102**, 2850–2855.