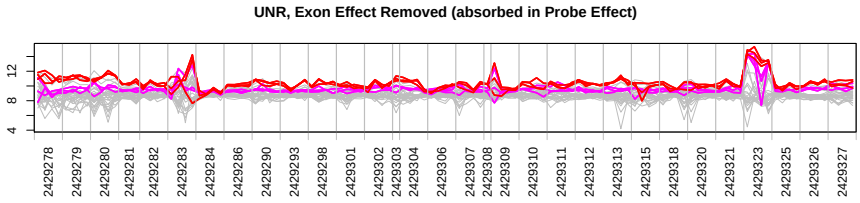
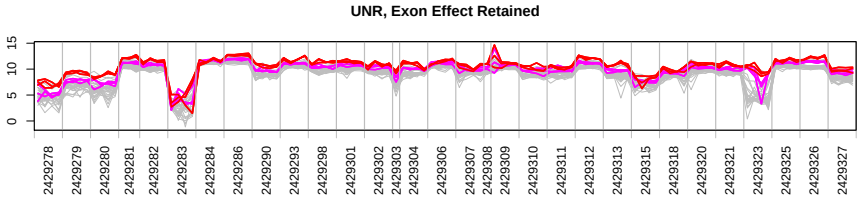


(a) No Exon Effect



(b) Exon Effect

Figure S1: Intensity plots of UNR demonstrating the visual effect of the estimate of probe affinity. Plotted are the intensities with (a) the estimated non-nested probe effect $\hat{p}_k$ removed and (b) the estimated nested probe effect $\hat{p}_{k(j)}$ removed.

Supplementary Text

S1 Algorithm: Exon Effect

As described in the main text, performing standard RMA will not estimate *nested* probe affinities (i.e. specific to the exon). The estimate $\hat{p}_k$ for probe k (with k ranging from one to the number of probes on the chip) from the standard RMA model will be an estimate of $e_j + p_{k(j)}$, where j is the exon containing probe k and $k(j)$ indexes the probes in exon j. This means that subtracting the probe effect, as is done in many of our plots, will erase any overall shift away from the overall gene expression level. In Figure S1 we visually demonstrate the difference in $\hat{p}_k$ and $\hat{p}_{k(j)}$; there we estimate $\hat{p}_{k(j)}$ by using the estimate $\hat{e}_j = \text{median } \hat{p}_k$ for k in exon j.

S2 Simulation Details

Data are simulated according to the following model:

$$y_{ij} = \log_2(B_j + I_{ij} \times 2^{(c_i + p_j)}) + \epsilon_{ij}, \tag{4}$$

with y_{ij} the $\log_2(PM)$ for chip i and probe j, and where

$$\begin{aligned} \log_2 B_j &\sim N(\mu_{\mathrm{BG}}, \sigma^2_{\mathrm{BG}}) \\ c_i &\sim N(\mu_c, \sigma^2_c) \\ p_j &\sim N(0, \sigma^2_p) \\ \epsilon_{ij} &\sim N(0, \sigma^2_{\mathrm{noise}}). \end{aligned}$$

The indicator variable I_{ij} is as follows:

$$I_{ij} = \begin{cases} 1 & \text{probe } j \text{ is present on chip } i \\ 0 & \text{probe } j \text{ is absent on chip } i. \end{cases} \tag{5}$$

This model features additive background, multiplicative noise, and probe–specific affinities. Additive background (as opposed to noise) is observed in most, if not all, microarray experiments, and can contain contributions from multiple sources (dark current in photon detectors, scattered light from the scanning laser, etc. (Bengtsson and Hössjer, 2006)).

We chose values for the simulation parameters by obtaining rough estimates of "typical" values from real data. We fitted a set of genes from the Affymetrix tissue panel data set to achieve this. The standard deviation of the residuals, σ_{noise}, generally lies between 0.5 and 1, with many values lying between 0.7 and 0.8, and so we set σ_{noise} equal to 0.7. To determine reasonable values for the additive background, we looked at the distribution of signals for the $\sim$ 16,000 genomic and $\sim$ 20,000 antigenomic control probes on the array. The distribution of background signals is not normal; in fact, it is skewed somewhat to the right. However, for the purposes of the simulation, this is probably not relevant as σ_{BG} is the standard deviation of the probe–wise background within a single gene, and not that for all genes on the chip. We set μ_{BG} equal to 5 (approximately the empirical mean of the antigenomic control probe signal) and σ_{BG} to 0.35 (slightly higher than the empirical σ, which is approximately 0.2).

The mean chip effect, μ_c, is arbitrary, as it is the mean expression level of the simulated gene. Similarly, σ_c is also arbitrary (we set it equal to 1.5). This is true to a lesser extent for the standard deviation of the probe effects, σ_p. The probe effects from fitting the example genes in the tissue panel data had a mean of 0 and a standard deviation of 3, and so we chose σ_p equal to 3.

The simulated alternatively spliced gene used for evaluation purposes is illustrated in Figure S2. We chose to simulate a gene with ten 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 dropped exons.

500 simulations were performed for each of two different values of μ_c (7 and 10) and four different probabilities of including a splice isoform (prob=0.1, 0.3, 0.5, 0.8), *i.e.*, eight cases in total. The two values of μ_c were chosen to mimic differing scenarios – one where the expression is close to background, and one where it is far above background.

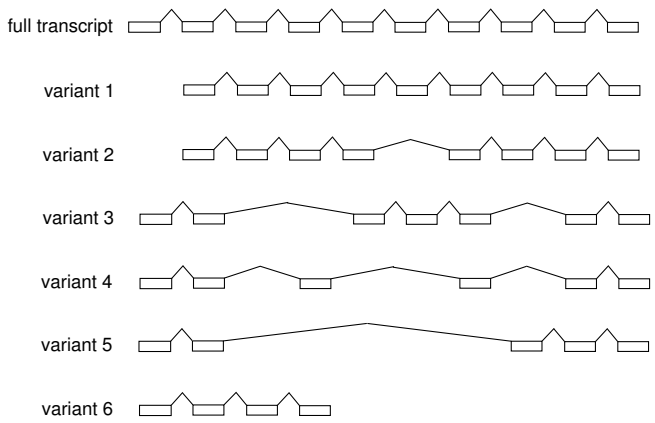


Figure S2: Structure of the simulated gene, and its isoforms, used to evaluate the detection algorithm. Boxes denote cassette exons, while lines connect the adjacent exons in a particular "transcript".

For each run, the simulated data set contained one gene with ten exons (as in Figure S2), in 40 chips (samples). We sampled random numbers from $U(0, 1)$ to determine which of the 40 chips would contain either the full transcript, or one of the variants 1 through 6 (the variant to be included was also chosen randomly). We did not simulate the case where a mixture of transcripts was included in a given sample.

Receiver–Operator Characteristic (ROC) curves are shown in Figures S3 and S4.

S3 Implementation of Algorithm in R

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. Furthermore, designation of probes belonging both to different exons as well as genes is not supported. Instead we implemented the FIRMA algorithm in the `aroma.affymetrix` package for large datasets which makes use of persistent memory (Bengtsson *et al.*, 2008).

The package relies on the standard chip definition file (cdf) that contains information regarding the location and grouping of probes on the chip; it also supports cdfs with grouping of probes into exons as well as genes. Affymetrix provides a cdf for the Human Exon 1.0 ST chip without any clustering of probesets to genes but merely of probes into probesets. We created custom cdfs appropriate for analysis in `aroma.affymetrix` that give the gene clusterings of Affymetrix ("transcript clusters") as well as a custom cdf file that creates gene clusters based on the human Ensembl gene build 47 (Hubbard *et al.*, 2007). The cdfs used in this paper retain the probeset definitions of Affymetrix. These cdfs

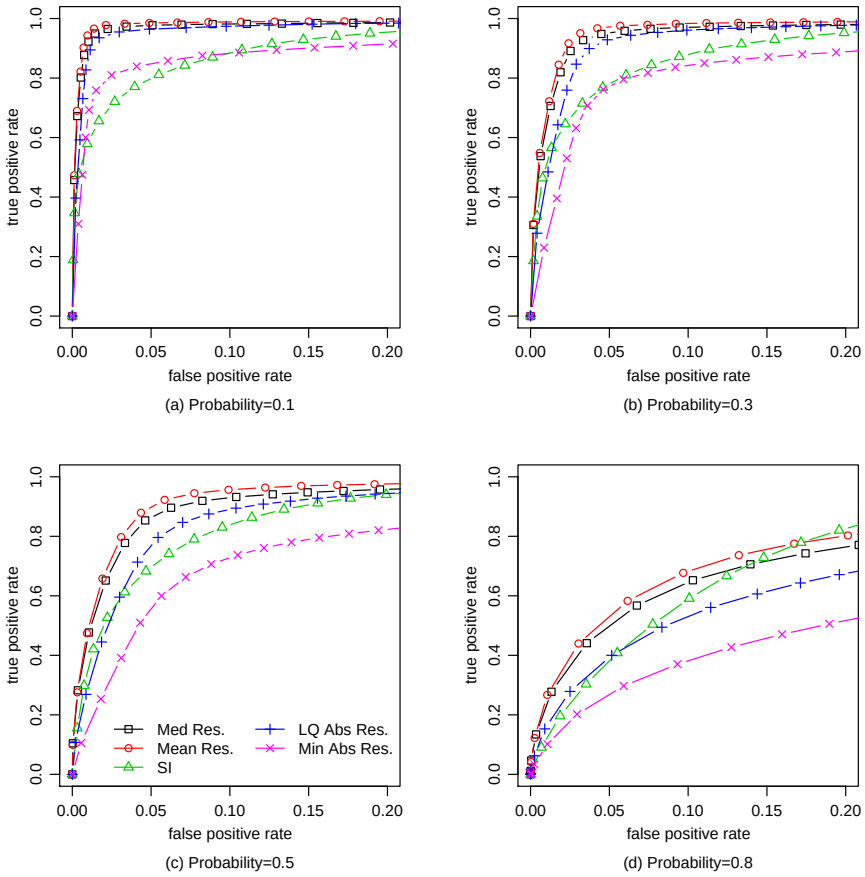


Figure S3: ROC curves for Single-Sample statistics for varying probabilities of a chip containing splicing. Note that the false positive rate (x-axis) only ranges from 0 to 0.2 to better show the main features of the plots.

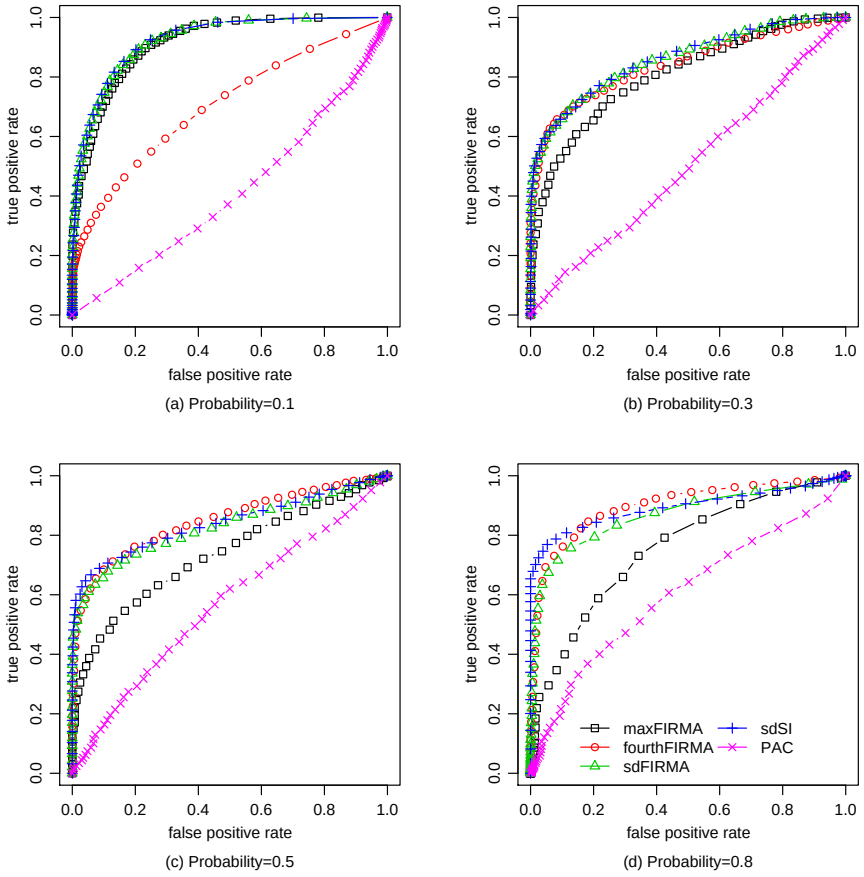


Figure S4: ROC curves for All-Sample statistics for varying probabilities of a chip containing splicing.

are publicly available in the support pages of `aroma.affymetrix`.

S4 Details of the Genome-Wide Scan for Muscle Enriched Genes

Because the pre-commercial array was designed against the November 2002 genome (hg13), we re-annotated their probesets based on the most recent assembly at the time the analysis was performed (hg18) in order to identify the probesets on the commercial array that corresponded to the confirmed region of alternative splicing

We scored each probeset by finding the minimum FIRMA score in each of the two tissue groups (heart and muscle) and then take the maximum,

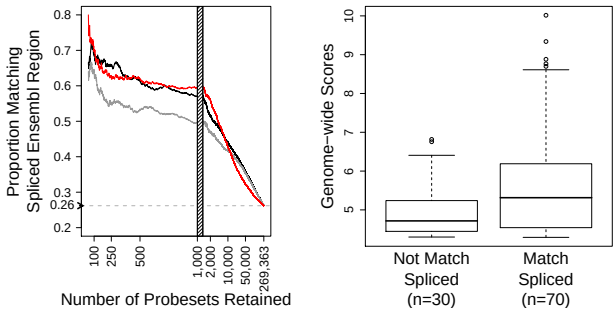
$$S_j = max\{min_{i \in \text{heart}} F_{ij}, min_{i \in \text{muscle}} F_{ij}\},$$

where F_{ij} is the FIRMA score in the ith sample and jth probeset. This created an all-sample score, to use our terminology from the main paper. Furthermore, we excluded probesets without at least 3 probes (roughly 20,000 probesets). To find the muscle-enriched candidates, we took the probesets with the largest positive value for this score. For searching for muscle-enrichment, we filtered low expressing probesets where the probeset summary expression in both the muscle or heart arrays was below three on the log-scale. This was based at looking at the distribution of scores and resulted in removing roughly another 100,000 probesets.

In Figure S5a we plot the proportion of splicing calls that match splicing events as a function of the number of probesets kept for the muscle-specific score and two general (non-tissue specific) scores. For the comparison with the Ensembl database, however, we used the largest absolute value of in either heart or muscle as the muscle-specific score. We also changed our filtering criteria to filter low-expressing probesets in which all of the tissues, not just heart or muscle, had an replicate below our threshold, which resulted in removing only around 45,000 probesets. For searching for general splicing amongst any tissues, we calculated the range of FIRMA scores per probeset as well as the largest FIRMA score in any tissue. We expect the range to be more sensitive, since it will detect situations when there are a large number of extreme residuals (like with ABLIM discussed above). And indeed the range does better in terms of matching splicing events in the Ensembl database. We note that the baseline comparison is 0.26 – the proportion of all the 269,363 non-filtered probesets that match a splicing event in Ensembl.

S5 Details of Colon Cancer Analysis

We tried to use the criteria of Gardina *et al.* (2006) in filtering the probesets to allow for better comparison. Gardina *et al.* had the following criteria for removing probesets: 1) probesets with a DABG p-value < 0.05 (an algorithm of



(a) Proportion of Splice Calls Supported by Ensembl

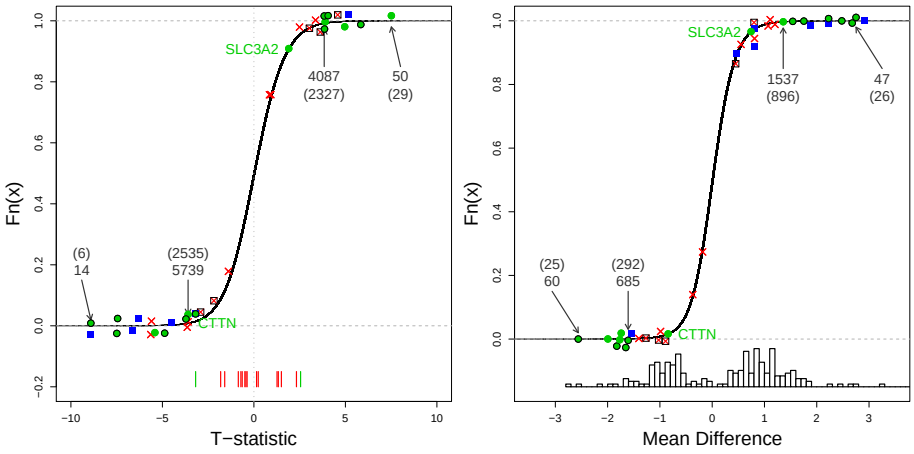
(b) Genome-wide Scores of Top 100 Probesets

Figure S5: (a) The proportion of (exon) splice calls that match spliced regions in Ensembl versus the number of top scoring probesets kept. After the first 1000 probesets, the x-axis switches to a log-scale (in the number of probesets), designated by the shaded rectangle. Different genome-wide scores for the probesets are shown: largest absolute FIRMA scores in muscle or heart (red); range of FIRMA scores across all tissues (black); and largest absolute FIRMA score in any tissue (grey). (b) Boxplots of top 100 probeset-summarized genome-wide scores for muscle specific alternative splicing, divided as to whether the probeset matched spliced regions in Ensembl.

Affymetrix for scoring probesets) 2) all probesets from genes with a signal > 70 3) probesets with a SI > 0.50 4) Core Exons. However, we did not find any of these values to be easily replicable. For instance, we could not find the DABG algorithm documented in any Affymetrix white-paper. And the criterion based on absolute measures, such as gene signal and SI, are highly dependent on the normalization method – implementing their values cut out a large proportion of the 200 probesets on their list that passed their filters. There was also no indication of how many probesets actually passed their filter to use as a guide to pick roughly similar values for the scale of our normalization. Ultimately, we decided to do minimal filtering that was similar to theirs but would keep both their list of 200 and (practically) all of their validated probesets; thus with a single filtering we could still evaluate all of their results. Thus the filtering was much more conservative than their final recommendation (though their validations were a result of experimenting with different filters). This resulted in a simple filtering based on probeset expression level cutoff of 0.70 (removing probesets with 50% or more of the samples below that level to be like the DABG step: 14,223 probesets) and a filtering of probesets from low expressed genes (signal of < 11 on the intensity scale: 27,444 probesets). Note that we did not remove probesets with less than three probes because Gardina *et al.* clearly did not do so – in fact a couple of the probesets they chose to validate had only one probe.

We also looked at using the T-statistic rather than the mean difference

(which we used in the main paper). In Figure S6a, we show the same kind of plot as Figure 5 but based on a t-statistic based on the individual FIRMA scores. We see that there is not as clear a division of the confirmed and non-confirmed probesets, as well as being of much wider range of ranks. While we chose the mean based on theoretical properties, this gives mild support to this choice, though it is still difficult to conclude based on the 53 probesets.

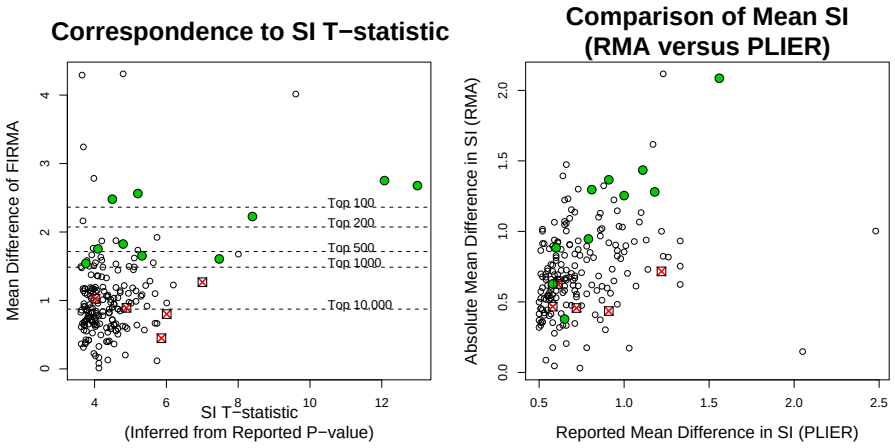


(a) Ecdf of T-statistic

(b) Ecdf of Mean Difference with Histogram of top 200

Figure S6: (a) Empirical Cdf of the t-statistic based on the paired differences in FIRMA scores. (b) Empirical cdf of the Mean Difference (as in main paper) but with a histogram of the values for the 200 top probesets of Gardina *et al.* (2006). Plotting symbols of the probesets are the same as in Figure 5.

We also can compare the reported scores of Gardina *et al.* for the top 200 probesets. In figure S6b we superimpose the histogram of mean difference of the FIRMA scores for those 200 probesets on top of the ecdf shown in the main paper. We see that the bulk of the (unvalidated) probesets on their list do not score well with the mean FIRMA difference (this does not change if we use the t-statistic based on the FIRMA scores, not shown). In figure S7a we show a scatter plot of the FIRMA mean difference versus the inferred t-statistic of Gardina *et al.* (inferred from reported p-values). As noted in the text, the confirmed and non-confirmed probesets in the top 200 do not separate well based on the t-statistic of the SI. While it would be useful to also use these results to compare the SI and the FIRMA algorithm more directly, the different choices in filtering and the different techniques of normalization and summarization also account for a great deal of the difference. For example, in figure S7b we show a scatter plot of the mean difference in SI, using estimates based on the IRLS (RMA) versus the PLIER algorithm; there is only a very weak correlation and thus much of the differences we see in the two results may be due to this.

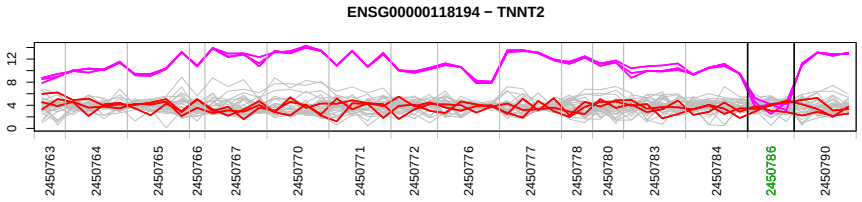


(a) (b)

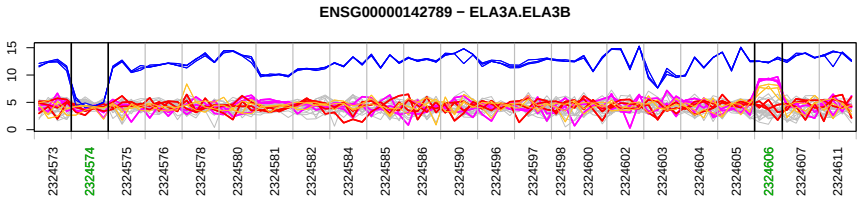
Figure S7: S7a A scatter plot of the FIRMA based t-statistics versus the (inferred) SI based t-statistics reported by Gardina *et al.* (2006) for the top 200 probesets reported by Gardina *et al.* (2006). S7b A scatter plot of the mean log difference of the Splicing Index when gene and exon effects are estimated by RMA versus the reported log difference of the Splicing Index (thus estimated with PLIER) for the top 200 probesets of Gardina *et al.*. Superimposed on both plots are the probesets which correspond to a probeset chosen for validation. The plotting symbols for these probesets are as in Figure 5

S6 Effect of Non-responsive Probesets

As discussed in the main text, if all of the probes in a probeset are non-responsive, the model for differential alternative splicing will often detect this as alternative splicing. We show some expression patterns in Figure S8 that demonstrate some problematic or confusing patterns that were found in various genome-wide searches that we performed. These are all examples of when a few arrays are expressed for the gene while the remaining arrays are not expressed. In such cases drops in expression to background level could be a splicing event in those few expressed arrays; however since there is no signal in that probeset from the other arrays, this could be a non-transcribed region of the genome or a case of differential expression. This is not a problem in the algorithm or model, per se, but can confound results.



(a) TNNT2



(b) ELA3A/ELA3B

Figure S8: Examples of situations with patterns of differential expression that are difficult to interpret.

A similar phenomena can happen when all of the arrays show expression for a gene, except for a particular probeset where all of the arrays drop to background level. If this results in a simple linear shift of the expression of the arrays, then such a shift will be included in the probe affinity estimate (see above) and will not result in a high FIRMA score. However, often when all of the arrays are at background level, the level of the expression of the arrays will not remain the same but be generally random and in addition the standard deviation of the signal often changes dramatically. We can see this by looking at all of the probesets, not just those corresponding to our Ensembl cdf, for CACNB1 (Figure S9) a gene that shows some levels of differential gene expression in the different arrays. We can see that most, though not all, the probesets not

supported by Ensembl annotation (labeled in green) express at background level and do not show the levels of variability in expression otherwise apparent for this gene. We also note that the same behavior can also be seen in the first two probesets, which are supported by the Ensembl annotation and probably represent alternative 3'/5' endings. Such heteroscedasticity will result in larger residuals for these probesets and thus larger FIRMA scores.

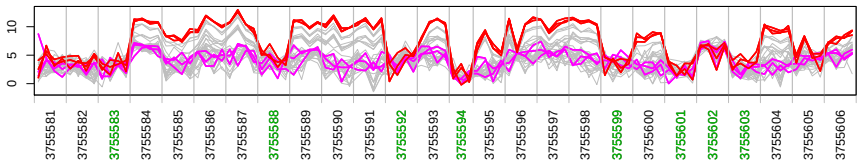


Figure S9: The log intensity expression of CACNB1 including all probesets associated with Affymetrix's transcript cluster. Probesets not corresponding to the Ensembl mapping that we used in the main text have labels colored in green. The nested probe effects $\hat{p}_{k(j)}$ are subtracted off here as in Figure S1 to emphasize the shift in expression level of these probesets.

The influence of poorly expressing probesets will particularly be a problem when additional probesets with less support from known annotation are analyzed. This heteroscedasticity will cause a large number of false positives, a problem that explodes in magnitude once more putative regions are included. The large increase in such probesets will also affect the chip effect estimates by overwhelming the ability of the robust fitting to estimate the true gene expression level. This is not specific to the FIRMA model but is a similar problem for summaries based on the SI, for example.

S7 Legends for Additional Figures

We provide additional material in the Supplementary Materials in the form of files that contain the same plots for many different genes; legends for these figures are given here.

Figure S10: Validated_FirmaScores.pdf: Each page in these files gives the FIRMA scores (top), residuals (middle) and normalized intensity values (bottom) for the eleven validated muscle-enriched genes alphabetically ordered. The x-axis for all of the plots indexes the probes, grouped by probeset; probesets are ordered along the chromosome. The label of the probeset of interest is in red. Per gene, the x-axis of each of the plots is aligned. For the residual and intensity plots, each chip is a separate line. Intensity values are on the log-scale. For the FIRMA scores, the value of the score for each sample is indicated by the same color scale as Figure 1 in the main text. The lines and labels of muscle and heart tissues are colored red and those of thyroid colored yellow

Figure S11: Enrich15Candidates_FirmaScores.pdf: Each page in this file plots gives the FIRMA scores (top), residuals (middle) and normalized intensity values (bottom) (as described in legend for Figure S10) for the fifteen candidate probesets highlighted in the paper. Below the intensity plots are corresponding transcripts from Ensembl. The green gene model represents the coordinates of the Affymetrix probesets corresponding to the intensity plots. The yellow gene model represents the concatenation of all exons in Ensembl corresponding to that gene. The purple gene models correspond to actual transcripts reported in Ensembl identified to that gene.

References

Bengtsson, H. and Hössjer, O. (2006). Methodological study of affine transformations of gene expression data with proposed robust non–parametric multi–dimensional normalization method. *BMC Bioinformatics*, **7**, 100.

Bengtsson, H., *et al.* (2008). aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Technical Report 745, Department of Statistics, University of California, Berkeley.

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

Brinkman, B. (2004). Splice variants as cancer biomarkers. *Clin. Biochem.*, **37**, 584–94.

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 Research*, **35**.

Durinck, S., *et al.* (2005). Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**(16), 3439–3440.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *JASA*, **99**(465), 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 Research*, **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 Research*, **17**, 1210–1218.

Le, K., *et al.* (2004). Detecting tissue–specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Research*, **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., Clark, F., and Smith, C. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, **6**, 386–398.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature Genetics*, **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., Shamir, R., and Ast, G. (2004). How prevalent is functional alternative splicing in the human genome? *Trends in Genetics*, **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. on Biocomputing*, **9**, 66–77.

Venables, J. (2004). Aberrant and alternative splicing in cancer. *Cancer Research*, **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. *Proceedings of the National Academy of Science*, **102**, 2850–55.