Supplementary Materials for

The human splicing code reveals new insights into the genetic determinants of disease

Hui Y. Xiong[1,2,3,*], Babak Alipanahi[1,2,3,*], Leo J. Lee[1,2,3,*], Hannes Bretschneider[1,3,8], Daniele Merico[4,5,6], Ryan K.C. Yuen[4,5,6], Yimin Hua[7], Serge Gueroussov[2,6], Hamed S. Najafabadi[1,2,3], Timothy R. Hughes[2,3,6], Quaid Morris[1,2,3,6], Yoseph Barash[1,2,9], Adrian R. Krainer[7], Nebojsa Jojic[10], Stephen W. Scherer[3,4,5,6], Benjamin J. Blencowe[2,4,6], Brendan J. Frey[1,2,3,4,6,8,10,@]

[@]To whom correspondence should be addressed: frey@psi.toronto.edu

This PDF file includes:

Materials and Methods
Figs. S1 to S32
References

Other Supplementary Materials for this manuscript includes the following:

Tables S1 to S18

1. Department of Electrical and Computer Engineering, University of Toronto
2. Donnelly Centre for Cellular and Biomolecular Research, University of Toronto
3. Canadian Institute for Advanced Research, Program on Genetic Networks, Toronto, Canada
4. McLaughlin Centre, University of Toronto
5. The Centre for Applied Genomics, Hospital for Sick Children, Toronto
6. Department of Molecular Genetics, University of Toronto
7. Cold Spring Harbor Laboratory, New York
8. Department of Computer Science, University of Toronto
9. School of Medicine, University of Pennsylvania
10. eScience Group, Microsoft Research, Redmond

MATERIALS AND METHODS

1. RNA-seq data processing

We used RNA-seq data to build and test our regulatory model of splicing. This section describes the key datasets and their processing methods. To apply our regulatory model to disease-causing genetic variants, we also performed whole genome sequencing from autism spectrum disorder (ASD) patients and control samples, which is described in Sec. S10.

1.1 RNA-seq datasets

Four RNA-seq datasets were used in this study. The training of our regulatory model was based on 75bp single-end RNA-seq data from the Illumina Human BodyMap 2.0 project (NCBI GEO accession GSE30611), which was derived using poly-A selected mRNA from sixteen diverse human sources, including adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testis and thyroid tissues, plus white blood cells. This data set was also used extensively in testing the model's performance on wild type (WT) predictions. Further evaluation of model performance used human and mouse RNA-seq data from various normal tissues generated in Kaessmann's lab (*53*), RNA-seq data from a recent MBNL knockdown study (*14*), and lymphoblastoid RNA-seq data and their matching genotype data from a population study (*15*).

1.2 Aligning RNA-seq reads

Using Tophat (*55*), reads were mapped to hg19 and a comprehensive set of splice junctions (by combining RefSeq and Ensembl annotations with a high quality set of junctions mined from EST data, details in Sec. S2), while also allowing confident discovery of novel junctions. When needed, we used Cufflinks (*56*) to quantify transcript and gene expression levels, based on Ensembl 73 annotations. When counting uniquely mapped reads to a splice junction, no more than 3 mismatches were allowed and at least 8 bases were required to align to both sides of the junction. This resulted in three sets of reads for each exon and tissue, corresponding to the junctions in isoforms with the exon included and excluded. For an exon triplet C1-A-C2, let N_{inc} denote the number of reads mapped to the two inclusion junctions C1A and AC2 and N_{exc} denote the number of reads mapped to the exclusion junction C1C2. Then, the total junction read coverage is defined as $N_r = N_{inc}/2 + N_{exc}$ for each exon and tissue. We used two techniques to estimate the percent of transcripts with the exon spliced in, referred to as PSI or Ψ (*57*). The simple, standard estimate was computed by normalizing N_{inc} and N_{exc} separately by the number of mappable positions, including a pseudo-count of 1 in each term, and then computing the ratio of $\left(\frac{N_{inc}}{2}\right) \Big/ \left(\frac{N_{inc}}{2} + N_{exc}\right)$ for each exon and sample. As described below, we also computed a more useful estimate that accounts for uncertainties due to low counts, read stacks and other effects. For this estimate, the number of reads mapped to each mappable position in each junction was further recorded.

1.3 Quantification of PSI using the 'positional bootstrap'

Since average RNA-seq coverage is proportional to gene expression and the probabilities of reads at different positions within a transcript are not equal due to various position-dependent biases, the junction read coverage varies greatly across cassette exon sequences. Therefore, it is important to treat Ψ values probabilistically and take into account the variable amount of uncertainty for each exon and tissue based on RNA-seq data (*58*). We used a Beta model, which is derived under a Bayesian inference framework by treating Ψ as the parameter of a Bernoulli distribution. The prior for Ψ is set to the uniform distribution over [0 1], which is $Beta(1,1)$. By definition, the probability of sampling an inclusion isoform is Ψ while the probability of sampling an exclusion isoform is (1-Ψ). Thus, after obtaining *i* inclusion reads and *e* exclusion reads, the posterior of Ψ is $Beta(1+i,1+e)$.

Adapting this model to junction reads from RNA-seq data, we obtained the Beta model for Ψ estimation:

$$p(\psi) = Beta(1+\frac{N_{inc}}{2},1+N_{exc}),$$

$$p(\psi) \propto \psi^{\frac{N_{inc}}{2}}(1-\psi)^{N_{exc}}.$$

Note that as $N_r = N_{inc}/2 + N_{exc}$ increases, the value of Ψ becomes more certain and converges to a delta function centered at $\left(\frac{N_{inc}}{2}\right)\Big/\left(\frac{N_{inc}}{2}+N_{exc}\right)$. However, this Beta model only captures the uncertainty in Ψ due to insufficient junction read coverage after summing over all mappable positions, and does not account for position-dependent biases.

To account for uncertainty introduced by position-dependent RNA-seq biases, we developed a bootstrapping method that incorporates the Beta model described above. If a problematic position of an isoform (*e.g.*, inclusion) has a high sequence-dependent preference or has a read stack, the Ψ obtained by the Beta model will erroneously favor that isoform. To address this problem, we bootstrap the mappable positions, *i.e.*, we randomly sample positions, sum the reads for each sample, and obtain a distribution over the resulting Beta model estimates of Ψ. In the above example, the single problematic position will have a significant probability of being omitted in some bootstrap samples, so that the final distribution of Ψ will have a much larger variance than in the Beta model without performing bootstrapping. As a result, uncertainty associated with outlier junction positions is discovered and addressed. On the other hand, if the number of mapped reads is the same across all positions, bootstrapping does not alter the Ψ estimates obtained from the Beta model.

Specifically, in the bootstrapping model, N_{inc} and N_{exc} are random variables computed with the sum of n_p draws with replacement from the n_p position-specific number of mapped junction reads, where n_p is the number of mappable junction positions. It is described in the following generative process:

$$r_i \sim Uniform(1,2,\ldots,n_p)\,,$$

$$N_{inc} = \frac{1}{2}\sum_{i=1}^{n_p} n_{C1A}^{(r_i)} + n_{AC2}^{(r_i)}\,,$$

$$N_{exc} = \sum_{i=1}^{n_p} n_{C1C2}^{(r_i)}\,,$$

$$\psi \sim Beta(1+\frac{N_{inc}}{2}, 1+N_{exc})\,,$$

where $n_{C1A}^{(p)}$ denotes the number of reads mapped to the *p*-th mappable position of the C1A junction. Note that for each bootstrapped Ψ value, the same set of r_i is used for all three junctions, so that bootstrap samples always include triplets of matching positions that share RNA sequence. For example, matching C1A and C1C2 positions share the same C1 sequence. Doing this minimizes the impact of sequence dependent bias of C1 on the predicted Ψ, because the same bias is represented in both isoforms. For each exon and tissue, we simulated the above process 20,000 times and used the empirical distribution of the simulated samples as the distribution of Ψ. For example, the variance of this distribution was used to select low-noise cases for testing.

2. Exon identification

Various exon sets were constructed to train and test our regulatory model and to study genetic variations.

2.1 Cassette exons used in training the splicing regulatory model

Cassette exons were first mined from RefSeq annotations. We downloaded Human UniGene sequences (build 226) from NCBI and aligned them to hg19 using megablast (*59*) and Splign (*60*), with custom filtering strategies developed to keep only high quality spliced alignments that originated from normal tissue samples. Cassette exons were then mined based on EST evidence and combined with those annotated in RefSeq. In order to obtain a non-redundant and high quality set of training examples for our splicing regulatory model, we further filtered out cassette exons that overlap with other exons, highly similar exons, and exons that are very short (<10nt) or very long (>6000nt). This resulted in a total of 10,689 cassette exons for training, which will be referred to as the *AS-EST* exon set from now on. By combining RefSeq and EST evidence, we also built a set of 33,159 constitutive exons.

2.2 Verification and filtering of cassette exons by RNA-seq

Detecting alternatively spliced exons by RNA-seq is confounded by several factors including read coverage, erroneous reads and each exon having different expression levels and splicing patterns in different tissues. We performed various analyses to check the consistency between the mined AS-EST exon set and BodyMap RNA-seq data, while limiting the effect of these confounding factors.

Fig. S1 shows the fraction of AS-EST exons detected to be alternatively spliced by RNA-seq, based on having at least *k* junction reads from the minor isoform across all 16 tissues, where *k* ranges from 1 to 10. The 10,689 AS-EST exons were binned into 10 bins

of equal size based on total junction read coverage summed across tissues. The fraction of AS-EST exons detected to be alternative was computed for each bin and plotted against the medium junction read coverage in each bin. If sequencing error were the main reason behind detecting more alternative exons with higher coverage, the false positive rate would decrease exponentially with growing *k*. The consistent growing trend of alternative exon detection across all *k* values clearly shows that sequencing error didn't play a major role in the detection.

Next we restrict the standard estimate of Ψ of detected alternative exons to be within a certain range. As before, the 10,689 AS-EST exons were binned into 10 bins of equal size based on total junction read coverage. The detected fraction of alternative exons, which were required to have minimum junction read coverage of 10 and Ψ within a specified range (5-95% and 10-90% respectively) in at least one tissue, were plotted against the median junction coverage of each bin as shown in Fig. S2. The fraction of detected exons increases with junction read coverage initially, with a small drop towards the end. This is somewhat to be expected since for highly expressed genes, it is more likely that a rare splice variant is present in the EST library. However, the drop is relatively small and most of the exons with enough coverage are still detected to be alternative.

In summary, we found that given enough coverage, the majority of exons in the AS-EST set were indeed alternatively spliced in at least one of the 16 tissues profiled by BodyMap. To better use RNA-seq evidence to test the performance of our splicing regulatory model, we filtered the AS-EST exons based on Ψ estimated by the positional bootstrap and define the following four exon-tissue combinations, also termed *events* hereafter as in (*5*), to be used later: (1) *AS-All* events: exon-tissue combinations where the bootstrap standard deviation is less than 10%; (2) *AS-Detected* events: further restrict the AS-All events to only include exons whose bootstrapped Ψ is more than one standard deviation greater than 5% and more than one standard deviation less than 95% in at least one tissue (9,209 out of the original 10,689 exons were kept this way); (3) *AS-Strict* events: further restrict the AS-All events to only include exons whose bootstrapped Ψ is at least one standard deviation greater than 5% and at least one standard deviation less than 95% in at least one tissue where the bootstrap standard deviation is less than 10% (1,654 exons were kept); (4) *AS-Extreme* events: further restrict AS-All to only include the exon-tissue combinations whose bootstrapped Ψ is more than one standard deviation greater than 5% and more than one standard deviation less than 95% (6.8% of AS-All events). Fig. S3 shows the distribution of Ψ estimated by positional bootstrap for these four sets of AS events based on BodyMap RNA-seq data.

2.3 Other exon sets used

To facilitate genome-wide analysis in human, we also constructed unique exon triplets based on all internal exons in RefSeq transcripts and their two flanking exons, referred to as *Triplet-RefSeq*, exon triplets based on all internal exons in Ensembl 72 transcripts and their two flanking exons, referred to as *Triplet-Ensembl*, and exon triplets based on all internal exons in Ensembl 72 transcripts with the canonical flag and their two flanking exons, referred to as *Triplet-Ensembl-canonical*.

3. Description of features

The current RNA feature set is based on our previous one (*3*), but includes ~40% more features and some modifications of previously defined features, which were found to improve prediction performance significantly. The following notes describe these changes and additions:

- Four nucleosome positioning features were introduced to encode computationally predicted nucleosome positions around exon A. Nucleosome occupancy scores were computed using (*61*) for the alternative exon and flanking introns. Features were defined as the average and maximal occupancy scores in the first 100nt in each intron and the first or last 50nt of the alternative exon.
- Twelve Alu related features were introduced to account for Alu repeats around exon A. Alu motif searches were executed using AF-1 (*62*) over the intronic sequences 300nt up and downstream of the alternative exon, and in an extended range of up to 2000nt in those introns.
- ~350 new binding motif features were included, including motifs for general splicing related RNA binding proteins (RBPs), SR and SR-related proteins (SC35, SRp20, 9G8, ASF/SF2, SRp30c, SRp38, SRp40, SRp55, SRp75, Tra2α/β), and hnRNP proteins (hnRNPA1, hnRNPA2/B1, hnRNPF/H, hnRNPG). The added features were based on motif counts in each of the 7 intronic and exonic regions as in (*3*).
- PTC features were removed since they rely on knowing the splicing pattern of the full transcript and cannot be reliably computed based on local DNA sequences.
- To compensate for the loss of PTC features and test for the protein coding potentials of various combinations of exons, four binary 'translatability' features were introduced: Translatable.C1, Translatable.C1A, Translatable.C1C2 and Translatable.C1AC2. These features were computed locally without relying on full transcript information. Translatable.C1A equals 1 if the sequence comprised of exons C1 and A can be translated without a stop codon in at least one of the three possible translation frames and equals 0 otherwise. The translatability features for other exon combinations were computed similarly.
- Feature names were modified to be more succinct, but the correspondence to previous feature names should be clear.
- Some redundant features were combined, *e.g.*, 'Frame.shift.NaN' and 'Frame.noShift.NaN' features were complimentary to each other and they were combined into a single feature 'FrameShift.A'.

Overall, the number of RNA features grew from 1,014 to 1,393, and the complete list is in Table S1.

4. Bayesian inference of the splicing regulatory model

We assembled the human splicing regulatory model using a Bayesian machine learning framework. Each of the 10,698 cassette exons in AS-EST was treated as a training case. An ensemble of neural network models that relates the RNA features and

the observed Ψ values was fitted to the exons in the training dataset. Each model seeks to maximize an information-theoretic 'code quality' measure (*3*), which is defined as the amount of information provided by the predictions of the model beyond a naïve guesser:

$$CQ = \sum_e \sum_t D_{KL}(\mathbf{q}_{t,e} \mid \bar{\mathbf{q}}_t) - D_{KL}(\mathbf{q}_{t,e} \mid \mathbf{p}_{t,e}),$$

where $\mathbf{q}_{t,e}$ is the target splicing pattern for exon *e* in tissue *t*, $\bar{\mathbf{q}}_t$ is the prediction of the optimal guesser that ignores the RNA features, $\mathbf{p}_{t,e}$ is the prediction made by the regulatory model not trained on exon *e* and D_{KL} is the Kullback–Leibler divergence between two distributions. It can also be interpreted as a likelihood function of the predictions $\mathbf{p}_{t,e}$ based on partial counts. Refer to (*3*, *11*) for detailed explanations of this objective function.

The structure of a single model in the ensemble is a two-layer neural network with sigmoidal hidden units shared across tissues. It is capable of modeling complex non-linear and context-dependent interactions between the RNA features and the splicing patterns. In this model, features are combined to form the inputs to a maximum of 30 hidden variables. Each of these hidden variables applies a sigmoidal non-linearity to its input. Subsequently, these non-linear hidden variables are combined by a softmax function to produce the prediction. The tissues were trained jointly as separate output units and shared the same set of hidden variables, enabling information about RNA feature usage to be combined across tissues. In this model, there are in total 41,820 potential input-to-hidden parameters and 960 hidden-to-output parameters.

Because of the complexity of this model, fitting a single model using a standard maximum likelihood learning method suffers severely from overfitting. Therefore, we adopted a Bayesian Markov chain Monte Carlo (MCMC) approach to search over billions of models with different structure and parameter values, and the final combined model substantially outperforms other popular machine learning techniques including linear regression, nearest neighbors and support vector machines. The learning algorithm and comparison to other techniques in terms of prediction accuracy have been described in detail in (*11*), where the method was tested using mouse microarray data. Here, we applied the model to human RNA-seq data and found that it worked well. We also found that accuracy was improved by the addition of many of the new features, including frequencies and locations of Alu elements and nucleosome positioning features.

As described in detail below, we used two different training sets to train two types of computational models, for the purpose of (1) predicting low, medium and high Ψ in each tissue, and (2) predicting whether or not there is tissue-variable splicing. Both computational models were trained using some common parameter settings, as follows: (1) Each feature was normalized by dividing the feature value by its maximum absolute value across the training set. This results in feature values between -1 and 1, but with the value of 0 preserved, since it usually has special meaning, such as the absence of a motif. (2) A sparsity prior of 0.95 was used for all input-to-hidden connection. (3) No sparsity between hidden units and output units was used. (4) A zero-mean Gaussian prior with unit variance was used for the weights on all active connections. (5) All weights were discretized to be between -5 and 5 in steps of 0.1. (6) MCMC was run for about 3000 iterations (full passes through all the parameters) for each experiment while discarding the first 100 burn-in samples. (7) Five-fold cross-validation with 6 different random

partitions of exons was used to obtain 30 different splicing regulatory models so that code prediction variability could be evaluated.

Next, we describe the targets used in training these two computational models.

4.1 Targets for the low/medium/high (LMH) model

We defined the low-medium-high (LMH) splicing pattern using three numbers, corresponding to the probabilities that Ψ is low (0-33%), medium (34%-66%) and high (67%-100%). To produce these splicing patterns, the distribution of Ψ generated by the bootstrapped Beta model was discretized according to the above ranges, resulting in three probabilities that sum to 1. We only kept events where N_r is at least 10 reads for training, although we later tested the model on cases with less than 10 reads (see Section S5.1). In total, 38% of events were kept and 85% of exons had at least one tissue that was kept. The events that did not pass the read count threshold were treated as missing data during training. Overall, for the events that passed the threshold, 62% were labeled as 'high', 7% were labeled as 'medium' and 31% were labeled as 'low'.

4.2 Targets for the tissue-regulated-splicing (TRS) model

The tissue-regulated-splicing (TRS) model aims to identify exons that have tissue-dependent splicing but does not aim to predict the values of differences between tissues. The TRS splicing pattern was also defined as three categories: 'low-across-tissues', 'high-across-tissues' and 'tissue-regulated'. They were derived using the LMH splicing patterns described above. For tissues that passed the junction coverage threshold ($N_r \geq 10$), if the splicing category (low, medium, or high) with the highest assigned probability for all tissues is low or medium, the exon is classified as 'low-across-tissues'. In other words, if Ψ of the exon is more likely to be between 0%-33% or between 34%-66% than between 67%-100% for all tissues with confident measurements, that exon is classified as 'low-across-tissues'. Similarly, if the most probable splicing category for all tissues is either high or medium, the exon is classified as 'high-across-tissues'. Otherwise, the exon is classified as 'tissue-regulated'. For these tissue- regulated exons, there is at least one tissue whose Ψ estimate is most probable between 67%-100% and at least one tissue whose Ψ estimate is most probable between 0%-33%, representing at least one significant tissue-dependent splicing difference. In total, 3087 exons were confidently labeled as 'high-across-tissues', 1506 exons were confidently labeled as 'low-across-tissues' and 589 exons were confidently labeled as 'tissue-regulated'. Note that our TRS splicing pattern targets were based on the list of 16 tissues studied. If we incorporate more tissues, more exons may exhibit tissue-dependent splicing. In addition, if RNA-seq coverage is increased, more exons may exhibit tissue-dependent splicing within the 16 tissues because of the omission of tissues with low coverage in our current dataset. Furthermore, the set of 5.5% (589 out of 10,689) 'tissue-regulated' exons used in training was identified with a very stringent threshold to ensure high quality training. If the threshold were lowered, a lot more exons would be identified as 'tissue-regulated'. For example, 2,688 exons out of 10,689 alternative exons that we analyzed (25.2%) have at least one tissue pair whose expected Ψ difference is greater than 15% with a z-score greater than 2. Table S2 shows the number of tissue-regulated exons identified by varying the minimum threshold of ΔΨ and the z-score of this difference, based on Ψ quantified by positional bootstrap on BodyMap RNA-seq data.

4.3 Simple splicing regulatory model based on multinomial regression

To demonstrate the importance of capturing context dependence when predicting splicing, we trained a multinomial regression model to compare with our Bayesian neural network model. As a generalization of logistic regression, the multinomial regression model is linear in the log odds ratio domain and does not have hidden variables. This model was trained using the same objective function, RNA features, splicing patterns and dataset partitions as the Bayesian neural network model described above. The features were normalized in the same way so that the numerical values of each feature ranged between -1 and 1. The multinomial regression model was trained with gradient descent and early stopping. The initial parameter values were set to small random values (mean zero Gaussian with $\sigma^2 = 0.001$) for the feature-to-output connections. The biases for each output unit were set to the optimal values so that the regression predicts the average probabilities for all training examples initially. As a result, the code quality starts around 0. To avoid overfitting, 1/3 of training data points were randomly chosen and set aside as the validation dataset used for early stopping. Fig. S4 shows the training performance and validation performance for several runs of gradient descent using different data partitions. We observed that the validation performance peaks well before training performance saturates. The final model of the multinomial regression is defined by the set of parameter values with the best validation performance.

5. Validation of WT predictions

In order to successfully apply the human splicing regulatory model to analyze genetic variations, it needs to have high prediction accuracy and generalize well to unseen cases. Therefore, we have extensively evaluated our regulatory model with various types of experimental data, including different sources of RNA-seq data, RT-PCR data, RNA binding protein (RBP) binding affinity data, splicing factor knockdown data, and matching genotype/phenotype data. These are described in detail in this section.

5.1 Validation using RNA-seq Data

We evaluated the prediction performance of the regulatory model using test sets that differ from the training data in several different ways. These include the ability of the proposed model to generalize across exons, chromosomes, datasets and species in predicting absolute Ψ levels and tissue-differential splicing on a genome-wide scale.

Because the regulatory model can potentially overfit the training data, we ensured that prediction performance was always evaluated using held out exons, which are not seen by the training procedure. Based on this principle, we performed five-fold cross validation for each splicing prediction task, in order to increase the number of exons used for model training and to reduce the noise in the estimate of prediction performance. In this procedure, the AS-EST exons were randomly partitioned into 5 bins of equal size, each containing 20% of the exons. A regulatory model was trained using the exons in four of the five bins and then tested on the exons in the remaining bin, which were held out during training. This procedure was repeated five times to obtain one test prediction for each exon in the AS-EST exon set. These test predictions were then evaluated for each of the 6 random partitions to produce the final estimation of the model performance.

Under the assumption that exons come from a certain probability distribution, this procedure produces an unbiased estimate of the performance of the learned model on novel exons from that distribution. When performing cross-chromosome and cross-species evaluation, this procedure was modified as described below. We also tested with randomly permuted exons to further demonstrate that our model does not suffer from over-fitting.

Furthermore, since predictions obtained from the regulatory model are in the form of three probabilities: p_{low}, p_{medium} and p_{high}, we obtained real-valued predictions of Ψ by computing the expected value of Ψ under this predictive distribution, whose bins are centered at 1/6, 3/6 and 5/6:

$$\hat{\psi} = \frac{1}{6} p_{low} + \frac{3}{6} p_{medium} + \frac{5}{6} p_{high} .$$

To correct for the fact that the bins bias the predictions toward 1/6 and 5/6 because of the quantization, we subtract 1/6 and multiply by 3/2 to get values between 0 and 1. The variance in Ψ is predicted as follows:

$$\hat{\sigma}^2 = \left(\frac{1}{6} - \hat{\psi}\right)^2 p_{low} + \left(\frac{3}{6} - \hat{\psi}\right)^2 p_{medium} + \left(\frac{5}{6} - \hat{\psi}\right)^2 p_{high} .$$

High variance predictions correspond to cases where the model spreads probability out over the low, medium and high categories.

5.1.1 ROC curves for low versus high Ψ

We used the bootstrap model to define a simple labeling scheme for high inclusion and low inclusion events in order to test the splicing model across gene expression ranges, exon sets, chromosomes, datasets and species. In this scheme, the distributions of Ψ are computed by bootstrapping and events with an expected Ψ above 66% and standard deviation of Ψ less than 15% are labeled as 'high inclusion', while events with an expected Ψ below 33% and standard deviation of Ψ less than 15% are labeled as 'low inclusion'. All other events, such as those with a standard deviation of Ψ greater than 15%, are discarded. The performance of the code in predicting absolute inclusion is evaluated by its ability of distinguishing 'high-inclusion' events from 'low-inclusion' events. The ROC curve for each tissue is plotted in Fig. S5a, and the overall AUC is 95.5%. If we only test on the top 50% high confidence predictions, then the overall AUC becomes 99.1%.

In Fig. S6, we grouped events based on their junction read coverage N_r. We observe that our model generalizes well across different ranges of gene expression. In particular, our model generalizes well when $N_r < 10$ without being trained on any events in that range, which indicates that our model is able to predict splicing in genes with low expression for which direct measurement of Ψ is challenging.

5.1.2 ROC curves for high inclusion versus non-high inclusion and low inclusion versus non-low inclusion

We also evaluated performance of the LMH model in distinguishing low inclusion events from events that are likely medium or high, and high inclusion events from events that are either medium or low. For the 'high inclusion' ROC curve, the positive data

points are defined as the events whose high category in the LMH splicing pattern have probabilities above 0.5 ($q_{high} > 0.5$) while the negative data points are the events whose low or medium category in the LMH splicing pattern is above 0.5 ($q_{medium} > 0.5$ or $q_{low} > 0.5$). Similarly, for the 'low inclusion' ROC curve, positive data points are the events with $q_{low} > 0.5$ and negative data points are the events with $q_{medium} > 0.5$ or $q_{high} > 0.5$. Real-valued predictions of Ψ were obtained as described above and used to produce the ROC curves in Fig. S7, and the overall AUCs are 91.4% and 92.8% respectively.

5.1.3 ROC curves for tissue-dependent differences

To demonstrate that our new model can make accurate predictions for tissue-differential splicing comparing to our previously developed model (*3*), we evaluated its performance on the task of predicting differences in Ψ between pairs of tissues. For each pair of tissues and each direction (sign of difference), we identified a 'positive' set of exons that have significant evidence supporting tissue-dependent inclusion and a 'negative' set of exons that have significant evidence supporting no substantial difference. To produce these sets, we computed the distribution of the Ψ difference by randomly picking pairs of Ψ from the bootstrap-generated samples of the two tissues, which is based on the assumption that Ψ distributions of the two tissues generated by bootstrapping are independent given RNA-seq data. Exons with expected Ψ differences greater than 15% and z-scores greater than 2 are defined as inclusion. Exons with expected Ψ differences below 15% and combined standard deviations less than 25% are defined as no change. For each pair of tissues and each direction of change, we tested the ability of the model to distinguish between positive and negative exons. To remove biases caused by the mismatch between this task and the original task of predicting absolute inclusion levels, we constructed a simple logistic regression classifier that takes the output of the LMH model as input. The logistic classifier was trained by only using the data used to train the LMH model, *i.e.*, it was not trained with validation or test data. For each tissue pair and direction of change, the logistic classifier takes six inputs corresponding to the log-probabilities of low, medium and high inclusion for the two tissues. The ROC curves for all tissues pairs are shown in Fig. S8, with an overall AUC of 89.1%. At a false positive rate of 1%, our new method correctly identifies 29.9% of cases, which improves substantially upon the previously published accuracy of 7.8% (*3*).

5.1.4 ROC curves using different sets of test exons

Although we have verified that most of the exons in AS-EST are indeed alternatively spliced, we also evaluated our regulatory model on the three event sets with increasingly stronger AS evidence: 'AS-Detected', 'AS-Strict' and 'AS-Extreme', to avoid possible biases introduced by constitutive or nearly constitutive exons. Since the estimated Ψ values will be noisier when imposing more stringent AS criteria on exons, restricting the prediction in this way will lead to decreased performance. However, as shown in Fig. S9, where ROC curves for different tissues are combined, and summarized in Table S3, the resulting performances are still extraordinary good on the task of

distinguishing low versus high Ψ. Similarly high performances were also obtained for other prediction tasks (data not shown).

5.1.5 Testing with randomly permutated exons

To further demonstrate the validity of our training and testing scheme, we performed the following random shuffling experiment. Using six random partitions and five-fold cross validation as described previously, we trained models with a dataset where the mapping between RNA features and splicing patterns were randomly scrambled. The resulting models were then tested on held out datasets, which were not permuted. The ROC curves for predicting absolute inclusion is plotted in Fig. S10. As expected, the splicing regulatory model so trained has no generalization power, since their performance on held out dataset is nearly random. Similar results were also obtained for predicting tissue-regulated splicing differences (data not shown).

5.1.6 Generalization across chromosomes, assays and species

To examine the ability of the splicing regulatory model to generalize across entire human chromosomes, we trained and tested a version of the regulatory model using a chromosome based 5-fold cross-validation. In this cross-validation partitioning, entire chromosomes were randomly assigned to five bins such that each bin has roughly the same number of exons. As a result, the prediction ability of the regulatory model is evaluated using exons from chromosomes that are not seen by the training procedure. As shown in Fig. S6b, the performance of the model is only marginally affected by partitioning according to chromosomes with an overall AUC of 93.7%. This provides additional support that the inferred code generalizes well and accounts for splicing mechanisms that regulate all chromosomes.

To examine the ability of our regulatory model to generalize across assays and species, we downloaded the RNA-seq data from (*53*), referred to as Kaessmann's data hereafter, which was produced by a set of different biological samples and sequencing machines in a different lab. We mapped Kaessmann's mouse and human data consisting of five tissues (brain, heart, kidney, liver and testis), which is a subset of the sixteen tissues used to train our model. The generalization ability across the BodyMap dataset and Kaessmann's dataset was evaluated by testing the regulatory model trained on BodyMap data using Kaessmann's dataset as labels. ROCs for both absolute Ψ levels and tissue-differential Ψ values were produced using the same labeling methods as described above. As shown in Fig. S6c and summarized in Table S3, predictions generalize well across assays and biological samples.

We also evaluated the performance of the regulatory model inferred from human splicing data using mouse exons and mouse RNA-seq data for brain, heart, kidney, liver and testis generated in Kaessmann's study. To avoid making predictions for a mouse exon with a model that has been trained on the orthologous human exon, we first identified 1,967 orthologous exons in our training set, using a 90% sequence similarity threshold including indels (note that although the exon sequences of these orthologous exons are similar, their flanking intron sequences are generally different which result in different feature vectors). When making predictions for a mouse exon, only the models that were not trained on the orthologous human exon were used. The ROC curves for absolute inclusion prediction are shown in Fig. S6d and the overall AUC is 90.3%.

5.2 Evaluation using RT-PCR data

To further study the ability of our regulatory model to predict tissue-dependent splicing differences, we selected a set of exons for RT-PCR validation in the following 10 matching but independent tissue samples from BodyMap RNA-seq: brain, heart, skeleton muscle, adrenal, lung, colon, liver, breast, kidney and ovary. Exons were selected based on the following criteria after applying our method to the 10,689 AS-EST exons: (1) the top 200 exons predicted by the TRS model to exhibit tissue variable Ψ; (2) the top 10,000 exon-tissue pairs predicted by the LMH model to have tissue-dependent differential Ψ (out of the ~900,000 overall exon-tissue pairs in 10 tissues); (3) having adequate gene expression levels in most of the tissues, as determined by RNA-seq data, to increase chances of successful RT-PCR experiments; (4) maximization of the number of predicted tissue-dependent differential Ψ values for each exon, especially within tissue types that have fewer tissue-specific splicing patterns (as opposed to the dominant brain, heart and muscle specific inclusion patterns). When applying the regulatory model (TRS or LMH), care was taken to ensure that predictions for a specific exon were made by a version of the model that was trained without the test exon in the training set (one such fold per partition), and the ranking was averaged across the six partitions. This procedure was used to select 14 exons. Their RT-PCR measured Ψ values are summarized in Table S5, and the accompanying gel images with quantified Ψ values are shown in Fig. S11. Overall, among exons and tissue pairs where the RT-PCR-measured Ψ differs by more than 5% ($n = 232$), we examined cases with low predicted variance ($n = 193$) and found that the direction of tissue-dependent splicing change is correctly predicted in 89.6% of cases.

5.3 Comparison to recently published RBP binding data

To see if our model can effectively account for information obtained from independent measurement on binding affinities of RNA-binding proteins (RBPs), we performed the following analysis. For each exon A, we first define the following six regions: up to 300 bases of the 3' end of I1 upstream of A, ±6 nt around the splicing junction between I1 and A, up to 300 bases of the 5' end of A, up to 300 bases of the 3' end of A, ±6 nt around the splicing junction between A and the downstream intron I2, and up to 300 bases of the 5' end of I2. Please note that the regions for each exon may have different lengths due to the limited length of the exon and its two flanking introns. In addition, the two exonic regions overlap when the exon length is less than 600nt, and become identical if an exon is shorter than 300nt. For each of the 98 RBPs measured in (*13*), we then scanned through each of the six previously defined regions and summed the z-scores of all overlapping heptamers based on the z-score matrices published in (*13*). Because the two junction regions have length 12, the heptamers always cross the exon-intron boundary. Finally these summed z-scores were normalized by the length of the region so that we have one feature for each region and protein. The correlation between these features and the RNA-seq profiled PSIs for each event, before and after subtracting code-predicted PSIs, are plotted in Fig. 1c. We also used linear regression to identify code features that are associated with RBPs. Specifically, for each RBP feature vector at each of the six regions defined above, we used LASSO (*63*) to train a model with code features as covariates. The number of covariates included in the regression was chosen to

minimize the Akaike information criterion (AIC). Model performance was tested by 10-fold cross-validation. Fig. S12 shows code features that have obtained a regression coefficient ≥ 1 for at least one RBP in at least one region. For ~90% of RBPs, code features can explain >80% of variance in RBP affinities across regions 1, 3, 4, and 6. By converting the resulting Pearson correlations to z-scores with Fisher's transformation, this corresponds to P<1e-200 for all four regions under the normal distribution assumption.

5.4 Analysis of MBNL knockdown data

To test if our model can account for the regulatory effects of *trans*-elements, we used RNA-seq data generated from Hela cell lines in a recent MBNL knockdown study (*14*). Based on gene expression levels estimated by Cufflinks, we estimated that MBNL expression was reduced by ~50% in the knockdown sample versus the control sample. By profiling the 10,689 AS-EST exons and the similarly mined 33,159 constitutive exons, we identified 333 exons with increased PSI and 331 exons with decreased PSI as estimated by positional bootstrap (z-score≥1 and $\Delta\Psi$>5%). Besides these exons, 26,951 out of 43,848 exons had a combined knockdown and control PSI standard deviation less than 5%. These exons are labeled as no change.

To compute the model predictions for the knockdown, we set the 24 MBNL-related features to the average value found in the training dataset. By comparing the knockdown prediction to the original prediction, we computed a MBNL regulatory score for each exon, similar to the regulatory score for SNVs (see Sec. S7.1). We found that the exons affected by MBNL knockdown had significantly higher predicted MBNL regulatory scores as described in the main text. We also found that the MBNL-related features themselves were also predictive. For example, the single most predictive feature is the conservation weighted MBNL motif count in the I2_5' region, which corresponds to a p-value of 3.1e-11, and combining all MBNL features produce a p-value of 2.5e-11. To further test if our code is capable of making more accurate predictions based on not only MBNL feature differences but also the context of other RNA features, we computed the distribution of AUCs by bootstrapping the exons. We found that our regulatory model had an AUC of 70.7% with standard deviation of 1.0% and the combined feature differences had an AUC of 59.8% and a standard deviation of 1.1%. Using a normal approximation, this corresponds to a highly significant p-value of 1.4e-14, which supports the superior performance of our regulatory model compared to directly using feature differences.

5.5 Analysis of individual genotype/phenotype data

Although a comprehensive population study is beyond the scope of the current paper, we examined genotype and RNA-seq data from lymphoblastoid cell lines of four individuals from a recent population study (*15*). Lymphoblastoid RNA-seq data was processed with our pipeline described earlier to obtain positional bootstrapped Ψ's. We also mapped their haplotype-resolved genomes to the set of ~300k common SNPs (to be described in Sec. S7). For each exon with a common SNP nearby, we used our regulatory model to predict the overall change in PSI in white blood cells as an approximate to lymphoblastoid cell lines, which our model was not trained on. When more than one common SNPs are found within the exon triplet and flanking introns, predicted changes

in PSI are combined by summing. For four randomly selected individuals, we found 99 events with measured delta PSI between two samples greater than 15%, measured standard deviation on PSI less than 15% and predicted delta PSI above the 25th percentile of all common SNPs. We used the above thresholds to ensure that these 99 differences are significant both in model prediction and in RNA-seq measurement. In this set of events, 72 are correctly predicted and 27 are incorrectly predicted, resulting in an accuracy of 73%.

6. Feature analysis using the splicing regulatory model

The analysis on MBNL knockdown data is an example showing that the *cis*-regulatory code inferred by our model is consistent with some underlying biological mechanisms. Here we present more in-depth analysis on these *cis*-features: first by feature sensitivity then by feature relevance.

6.1 Analysis using feature sensitivity

To ascertain the quantitative effect that a single feature F is predicted to have on Ψ for a given *cis*- and *trans*-context, we define the exon-specific feature sensitivity, denoted by $\Delta\Psi/\Delta F$, which is an estimate of the partial derivative of predicted Ψ with respect to a feature F. It equals the difference in predicted Ψ that a small change in the RNA feature F makes, while holding all other feature values constant. For example, suppose F is the FOX motif count in region I1, then a positive $\Delta\Psi/\Delta F$ for brain indicates that introducing an extra FOX motif in I1 will increase the value of predicted Ψ, when all other RNA features are left unchanged. In general, the magnitude of this 'feature sensitivity' indicates how much the feature affects splicing in the given context, and its sign indicates whether the feature inhibits or promotes splicing. Because our regulatory model can make non-linear predictions with interactive features, both the magnitude and sign of $\Delta\Psi/\Delta F$ can be different for different exons depending on their *cis*-element contexts. The distributions of $\Delta\Psi/\Delta F$ for the top 16 features, as determined by the frequency with which the feature was selected during Bayesian inference, are plotted in Fig. S13. For each tissue and the 100 features most strongly selected during learning, we computed a histogram of feature sensitivity across *cis*-contexts defined by different exons, shown in Fig. S14. Most features either positively or negatively affect Ψ across cis-contexts, but interestingly, 40 of the top 100 features frequently switch the direction of their effect in at least one tissue, depending on *cis*-context. To explore the effects of *trans*-context, we separated the sensitivities by tissue and found that the same feature can have quite different sensitivity in different tissues, *e.g.*, while features have nearly identical sensitivities in breast and adipose tissue, they are often quite different in brain tissue. Fig. S15 plots the discrepancy in feature sensitivities for every pair of tissues, and illustrates rich compositional structure. In Fig. S16, for every pair of tissues, we plot the feature sensitivity in one tissue against the feature sensitivity in the other tissue for the top 20 features in all exons. The importance of context dependence is also evident when the effects of genetic variations are analyzed, as described in the main text and Sec S7.6.

6.2 Visualization by feature relevance

Here we aim to provide a 2-D visualization of pairwise feature relationships based on similarity of *cis*- and *trans*-context dependence. We derived an exon-specific relevance for RNA feature x_i, which is a score that captures the overall effect of the i-th feature on the predicted Ψ. It is defined as the difference between the predicted Ψ using the original feature set and the predicted Ψ after x_i is integrated away. Ideally, the conditional distribution of x_i should be used in the integration. However, since the conditional distribution is unknown, the marginal distribution of x_i is used as surrogate, and it is approximated by the empirical distribution of x_i on the training set.

Specifically, let $\hat{\psi}^{e,t}(\mathbf{x})$ denote the expected Ψ for exon e in tissue t predicted by the LMH model given features $\mathbf{x} = \{x_1, \ldots, x_N\}$. The predicted Ψ for exon e in tissue t with feature x_i marginalized out is then

$$\tilde{\psi}^{e,t}_{-x_i}(\mathbf{x}) = \sum_{x_i} \hat{\psi}^{e,t}(\mathbf{x}),$$

where the sum is over all values that x_i takes in the training set. $\tilde{\psi}^{e,t}_{-x_i}(\mathbf{x})$ was computed by numerical integration since most of our features take only few values. The *relevance* of feature x_i for exon e in tissue t is defined as

$$s^{e,t}_{x_i} = \hat{\psi}^{e,t}(\mathbf{x}) - \tilde{\psi}^{e,t}_{-x_i}(\mathbf{x}),$$

which is a measure of how much and in which direction the predicted Ψ changes when the feature is unknown or removed from the model. We further define the relevance of x_i for exon e by averaging over tissues

$$s^{e}_{x_i} = \frac{1}{T} \sum_{t} s^{e,t}_{x_i}$$

Putting all exon-specific feature relevance together results in the following matrix

$$\mathbf{S} = \begin{pmatrix} s^1_{x_1} & \cdots & s^E_{x_1} \\ \vdots & \ddots & \vdots \\ s^1_{x_F} & \cdots & s^E_{x_F} \end{pmatrix}$$

where each row is like the signature of a feature. We may view the rows as vectors in the high-dimension space spanned by the columns of the feature relevance matrix, using their pairwise Euclidean distances as a similarity measure for the features. Features that affect the same set of exons will have very small distances while features that influence different sets of exons will have large distances.

In order to visualize the rows of the feature relevance matrix in two-dimension, we applied a powerful non-linear dimensionality reduction technique called *t-distributed stochastic neighbor embedding* (t-SNE) (*64*). Because many features have little predictive power and their entries in the feature relevance matrix are very small, we only selected the 200 strongest features to visualize, where the strength of a feature x_i is defined as

$$\theta_{x_i} = \sum_{e} \left| s^{e}_{x_i} \right|.$$

Next we applied the version of t-SNE available as part of the *Matlab Toolbox for Dimensionality Reduction*, available from http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html and set the perplexity parameter to 10 to achieve a good tradeoff between preserving local versus global structures. The embedded 2D visualization is shown in Fig. S17.

7. Genome-wide analysis of SNVs

To analyze the effects of SNVs on a genome-wide scale, we used the following three SNV datasets: 1) the common SNPs track of dbSNP135 on hg19 downloaded from the UCSC genome browser (*65*); 2) the point mutations in HGMD Professional version 2011.3 (*17*); 3) GWAS catalog downloaded from http://www.genome.gov/gwastudies/ on Dec. 7, 2013. For each SNV dataset, we mapped them to various exon sets (AS-EST, Triplet-RefSeq, Triplet-Ensembl, Triplet-Ensembl-canonical etc.), which include an exon and up to 300nt of flanking introns, for different analyses. The total number of unique exon-SNV combinations analyzed is 658,420, and the total number of unique SNVs is 410,412. In addition, we also analyzed rare variants on chromosome 1 (Sec. S7.7), and thousands more SNVs were studied when applying our regulatory model to three specific diseases (Sec. S8-10).

7.1 Quantifying the effect of a SNV using our regulatory model

After pairing up a mutant and its wild type (WT) sequence corresponding to after and before introducing a SNV, we used two measures to quantify its effect on an exon. First we compute the difference between mutant and WT PSIs predicted by our code across all tissues and take the one with the largest magnitude. This will be simply denoted by $\Delta\Psi$ hereafter. We also derived a 'regulatory score' for each SNV based on the LMH model, which is computed by summing up the absolute differences of the three low/medium/high probabilities and averaging across all tissues. In order to display a smooth distribution of genome-wide regulatory scores as in Fig. 4, we created 56 bins from -5 to 0.5 with a bin width of 0.1, and applied a Gaussian smoothing filter with a standard deviation of 0.2.

7.2 Other Computational methods

When comparing to other methods that can detect aberrant splicing, since there already exists many in the literature that can predict the effect of splice site mutations accurately (see (*66*) for a review), we focused on two methods that are specifically designed to predict the effects of SNVs deep into exons and/or introns: Skippy (*48*), which can be applied only to exonic SNVs, and Spliceman (*47*, *67*), which can be applied to both exonic and intronic SNVs. For both methods, the web versions of these tools were used (http://research.nhgri.nih.gov/skippy/input.shtml for Skippy and http://fairbrother.biomed.brown.edu/spliceman/index.cgi for Spliceman) and run according to the instructions provided on their webpages. Fig. S18 shows the result of comparing these two methods to our splicing code for detecting disease SNVs.

To score variants according to their overlap with functional annotations, we annotated exonic and intronic sequences using the feature detectors that we previously developed (*3*), plus additional ones described in Sec. S3. Annotations include donor and

acceptor signal strengths, information about conservation in intronic and exonic regions, binding sites of splicing-associated proteins and exonic and intronic splicing enhancers and silencers. Also included are more complex, derived annotations, such as lengths of proximal introns and exons, information about whether splicing induces a frame shift or a nonsense codon, information about secondary structures in introns, exons and junctions, nucleosome positioning, retroviral sequences. For each wild type exon and mutant, the annotations were converted into a normalized feature vector (as described in Sec. S4) and a score was computed by summing the absolute feature differences.

7.3 Proximal HGMD analysis

To determine if the splicing regulatory effects of HGMD SNVs are simply due to their genomic locations, we compared the predicted splicing effects for pairs of nearby common and HGMD SNVs. Using a maximum distance of 10nt, we found 212 pairs of intronic SNVs that are more than 6nt away from the splice site. We performed the KS test for the regulatory scores on this position-matched set of common and HGMD mutations. HGMD SNVs are still significantly enriched for high regulatory scores with a KS-statistics of 14.6% (P=0.0096), which is similar to the result for all intronic and non-splice-site SNPs combined without controlling for genomic locations. Relaxing the threshold of distance to 50nt, we found 689 pairs of SNVs. The KS-stastics for this set is 12.1% with P=4e-5. The resulting CDFs are plotted in Fig. S19.

7.4 Analyzing GWAS SNVs and sQTLs

Out of a total of 15,204 SNPs in the GWAS catalogue, 618 SNPs can be mapped to *triplet-RefSeq* and also overlap with the UCSC common SNPs track. It is observed that the GWAS-implicated SNPs are significantly enriched for exonic variants: 25.9% of GWAS-implicated SNPs are in exons compared to 15.0% overall (P=1.4e-12, one sided hypergeometric test). The GWAS SNPs are also enriched for missense and nonsense mutations (P=1.6e-10 and P=4.0e-4 respectively, one sided hypergeometric test). However, a two sample Kolmogorov–Smirnov (KS) test performed on the distribution of the predicted regulatory scores of the intronic GWAS-implicated mutations found no significant difference compared to other intronic mutations. There are 457 GWAS-implicated and 26,2347 non-GWAS-implicated intronic SNPs used in this KS test. This is in striking contrast to the intronic HGMD SNVs whose regulatory scores are much higher than intronic common SNPs. To analyze the power of the two-sample KS test in this situation, we randomly replaced 5% of intronic GWAS-implicated SNPs with randomly chosen intronic HGMD SNVs. The KS test is able to reject the null hypothesis that disease-associated set and the common SNPs set have the same distribution of regulatory score 99.5% of the time with $p<0.05$.

We also analyzed the 1,763 splicing QTLs (sQTLs) reported in (*51*), of which 453 exon-SNV combinations (or 324 unique SNVs) can be mapped to 'triplet-RefSeq'. These sQTLs were discovered by correlation analysis based on RNA-seq data of blood samples from 922 individuals. Using our model, a regulatory score is computed for each mapped sQTL and compared to the distribution of all common SNPs. Using the KS test, we found that the splicing QTLs are significantly enriched for higher regulatory score, with a p-value of 4.2e-10 as stated in the main text. This is true even after breaking them down

into exonic and intronic SNVs: P=2.2e-3 for 165 exonic sQTLs and P=1.2e-6 for 297 intronic sQTLs (KS test).

7.5 Condel analysis

We used the Condel webserver (*20*) that combines predictions of several methods, including SIFT (*68*) and PolyPhen2 (*69*), to predict the implications of missense SNVs in HGMD in terms of protein function. For each missense SNV, Condel annotates it as either 'deleterious' or 'neutral'. A 'score' between 0 and 1 accompanies this label such that a score of 1 implies the highest likelihood of pathogenicity and a score of 0 implies neutrality. In order to pick the most reliably predicted SNVs from deleterious and neutral sets, we sorted SNVs in each set based on their Condel scores and picked the top and bottom 25% of SNVs from deleterious and neutral sets, respectively. We then compared their code-predicted $\Delta\Psi$'s. Fig. S20 depicts the CDFs of the $\Delta\Psi$'s of two sets for the 25% (P= 6.65e-20, KS test), and also the 10% (P=5.46e-19, KS test) most reliable SNVs. Further, for the 25% threshold, approximately five times more SNVs that are neutral to protein function disrupt splicing ($\Delta\Psi$ >5%), compared to those that are deleterious (see Fig. S20).

7.6 Analysis of context dependency

By taking into account *cis*-context dependence, our computational model can make more accurate predictions for Ψ, but we wondered whether the predicted effects of mutations also depend on context. Therefore, we sought to analyze pairs of SNVs that induce very similar changes of $\Delta\mathbf{F}$ in their wild type feature vectors, but where the wild type feature vectors are themselves different (note that whereas above F referred to a singled feature, here $\mathbf{F}$ refers to the entire feature vector). For such pairs of SNVs, linear methods, such as correlation analysis, would predict very similar changes in Ψ, since if $\Psi=\mathbf{AF}$, then we have $\Delta\Psi_1=\mathbf{A}\Delta\mathbf{F}_1$ and $\Delta\Psi_2=\mathbf{A}\Delta\mathbf{F}_2$, and $\Delta\Psi_1-\Delta\Psi_2 = \mathbf{A}(\Delta\mathbf{F}_1-\Delta\mathbf{F}_2) \approx 0$. In contrast, context-dependent models are capable of predicting different changes in Ψ.

For each SNV, we identified another SNV whose $\Delta\mathbf{F}$ was most similar, according to the normalized dot product (angle) between $\Delta\mathbf{F}$ for the two SNVs. If that angle was greater than 5 degrees, we discarded the pair, but otherwise we kept it. To ensure that the changes in the two feature vectors were identical, rather than just similar, we modified $\Delta\mathbf{F}$ for the two SNVs to be equal to $(\Delta\mathbf{F}_1-\Delta\mathbf{F}_2)/2$. This ensures that linear models would predict exactly equal changes in Ψ, *i.e.*, $\Delta\Psi_1-\Delta\Psi_2 = 0$. We applied this correction to the mutant feature vectors, re-applied the computational model, and computed $\Delta\Psi_1$ and $\Delta\Psi_2$ for the two SNVs. Fig. S21 plots $\Delta\Psi_1$ against $\Delta\Psi_2$, and we found that for 43% of the cases, $\Delta\Psi_1$ and $\Delta\Psi_2$ differed by more than 5%, indicating that the predicted effects of mutations are highly context dependent.

7.7 Analysis of rare variants

Since our method was derived (trained) using the reference genome, without any mutation data, its accuracy in predicting the effects of mutations should not directly depend on population frequency. To demonstrate this, we analyzed chromosome 1 variants of three types: disease annotated (based on HGMD), common with no disease annotation (minor allele frequency or MAF>1%) and rare with no disease annotation (0.1%<MAF<1%). We downloaded the rare variants via ANNOVAR (*54*) at

http://www.openbioinformatics.org/annovar/annovar_download.html, selecting ones whose maximum MAF across major populations is between 0.1% and 1%. To match the genomic distributions of the three sets of variants, we only kept variants that are located within the exonic and intronic regions used by our code for HGMD mutations, and among intronic mutations we only kept those that are at least 3nt from a splice site. Further, to focus on variants that don't change protein sequence, which may have a disease cause that is unrelated to splicing regulation, we only kept synonymous exonic mutations and intronic mutations. This procedure resulted in 860 intronic and 1584 synonymous exonic rare variants, 316 intronic and 2395 synonymous exonic common variants, and 140 intronic and 333 synonymous exonic disease-annotated variants. For each variant from the three sets (common, rare, disease-annotated), we applied our technique to predict the mutation-induced change in PSI for each tissue and then computed the maximum absolute change across tissues. We used the KS test to compare the distribution of the mutation-induced change in PSI for the three different sets.

We observed a significant difference in the regulatory scores generated by our method for disease variants and all other variants (both rare and common): P=3.2e-39 for intronic mutations and P=1.2e-10 for exonic mutations. However, no significant difference was observed in the regulatory scores for non-disease annotated rare and common variants, although the common and rare variant sets are larger than the disease-annotated set.

8. Analysis of SMA genes *SMN1/2*

To carry out mutagenesis experiments for validation, we constructed parental *SMN1/2* minigenes pCI-*SMN1* and pCI-*SMN2* as described in (*24*). The previous 200-nt shortened intron 6 was modified to 283nt long, comprising first 61nt of intron 6, 3nt linker (TCT) and last 219nt of intron 6. All nucleotide differences that occur naturally between endogenous *SMN1* and *SMN2* were carried over to the two minigenes. All mutants (*SMN1* G-44A, A100G, and A215G, and *SMN2* G-35T, A-133G and C-134A) were transfected into HEK293 cells by electroporation; total RNA was isolated with Trizol reagent (Invitrogen), and 1 μg of each RNA sample was used per 20-μl reaction for first-strand cDNA synthesis with Oligo-dT and ImProm II reverse transcriptase (Promega). Standard radioactive RT-PCR was performed, and splicing was analyzed in native polyacrylamide gels as described in (*70*).

Although SMA is a neural disease, we performed experiments on HEK293 cells instead of neuron cells. One reason is due to the easy transfection of HEK-293 cells, but more importantly, we have previously tested many splicing-modulatory antisense oligonucleotides in neonatal and adult transgenic mice (to follow *in vivo* effects on *SMN*2 splicing in spinal cord, brain, liver, muscle, heart, and kidney in murine models of SMA), as well as in HEK-293 cells and patient fibroblasts (*24*, *70–72*) and have not observed cell-type specific differences in the splicing regulation of the ubiquitously expressed *SMN*1 and *SMN*2 genes. In particular, ISIS-SMNRx is an investigational drug in phase-2 trials involving intrathecal injection to correct splicing in spinal-cord motor neurons, and we originally identified this oligonucleotide by screening in HEK-293 cells.

There are a large number of published mutagenesis studies aiming to understand various regulatory mechanisms of exon 7 splicing. After an extensive literature survey,

we identified over 300 variations, including substitutions, insertions and deletions. For all of these, we used our regulatory model to predict the mutation-induced ΔΨ, and compared our results with the corresponding studies in the literature (*23*, *24*, *73–77*). Since these studies were generally carried out in cell lines that our splicing model was not trained on, we took the average of ΔΨ across all 16 tissues as a surrogate. For each mutation, a value of ΔΨ was computed using either *SMN1* or *SMN2* as the wild type, depending on the experiment conducted in the original study. When both wild type sequences were tested in an experiment, a more appropriate one was chosen as the reference. To estimate the uncertainty in the code-predicted ΔΨ, we used all 30 LMH computational models trained from different partitions and folds to compute the sample variance of ΔΨ. To identify confident predictions of change, we computed a z-score and applied a threshold of ±1. Table S6 summarizes all *in silico* mutational analyses for *SMN1/2* mutations and lists the classification accuracies for the direction of regulation in each region or mechanism of interest. More details on the *in vivo* selection of exon 7 and representative experiments in each of the four regions of interest are plotted in Figs. S22-26. The same data was also used to generate Fig. 5d in the main text.

9. Analysis of nonpolyposis colorectal cancer genes *MLH*1/*MSH*2

We downloaded all *MLH1* and *MSH2* single nucleotide substitution variants from the international society for gastrointestinal hereditary tumors (InSIGHT) mutation database (*26*) as of September 18, 2012. We only considered mutations located in exons, plus mutations located at the 3' end of the upstream intron and 5' end of the downstream intron. We computed ΔΨ between the wild type and the mutant (variant). To assess the significance of the predictions for each variant, we determined the percentile rank of its ΔΨ among those of common SNPs.

A total of 977 mutations were analyzed (536 in *MLH1* and 441 in *MSH2*), 156 of which introduced a stop codon (63 and 93). From the 977 mutations, if nonsense mutations are considered, 421 of them (243 and 178) had ΔΨ larger than the 95% percentile of common SNPs, and if nonsense mutations are ignored, this number reduces to 265 (180 and 85). All of the code predictions for *MLH1* and *MSH2* variants are listed in Tables S7 and S8, respectively. Moreover, significant predictions excluding nonsense mutations are plotted in Fig. 6a in the main text. In that figure, coding sequence (CDS) numberings for *MLH1* and *MSH2* are based on GenBank NM_000249.3 and NM_000251.2 respectively, where the numbering starts at the A of the ATG-translation initiation codon.

To further evaluate the accuracy of the predictions, we compiled a set of positive (aberrant splicing in the form of varied exon inclusion) and negative (negligible/no change in splicing) test cases, as determined by RT-PCR experiments. Overall, we complied a set of 229 mutations (150 and 79). Validations of predictions for *MLH1* and *MSH2* are listed in Tables S9 and S10, respectively. It can be seen that the majority of the predictions are concordant with RT-PCR experiments. Ignoring the novel/cryptic splice site cases and cases for which different studies report inconsistent outcomes, the code achieves an AUC of 92.4% (P = 2.8e-23, one-sided permutation test, n=134) for *MLH1* and an AUC of 93.8% (P = 8.7e-15, n=73) for *MSH2*.

When analyzing the three cognate mutations c.1976G>[T|C|A] in exon 17 of *MLH*1, ESEFinder 3.0 (*32*) detected higher scoring SRSF5 (SRp40) ESEs for all three mutations (the WT score of 4.33 increased to 6.09, 5.15 and 7.79 respectively), while both C1976G>T and C1976G>C also gained a novel SRSF2 (SC35) ESE with scores of 2.60 and 2.83. Therefore, ESE evidence points toward increased exon inclusion for all three mutations. However, this contradicts experimental evidence of increased exon skipping for C1976G>T and C1976G>C as described in the main text. In contrast, our computational model confidently predicts increased exon skipping for all three mutations (Table S10).

10. Analysis of rare variants implicated in ASD

To study rare mutations in autism spectrum disorder (ASD) with our splicing regulatory model, we used brain tissue samples from five autism cases obtained from the Autism Tissue Program, detailed in Table S11. They are all Caucasians without any other known cytogenetic findings for autism (*e.g.*, chromosome 15q duplication). We also carefully selected the following three control groups: (1) *CEU Subjects* (Utah residents with Northern and Western European ancestry): four unrelated Caucasian samples ('NA06985', 'NA06994', 'NA07357', and 'NA12004'), with whole genome sequencing data publically available from Complete Genomics (ftp://ftp2.completegenomics.com/), two males and two females; (2) *TSI Subjects* (Tuscans in Italy): four unrelated Caucasian samples ('NA20502', 'NA20509', 'NA20510', and 'NA20511'), with whole genome sequencing data publically available from Complete Genomics, three males and one female; (3) *IMS Subjects* (In-house Male Subjects): four unrelated Caucasian samples sequenced by Complete Genomics (GS12066, GS12067, GS13808, GS13809). The whole genome sequencing data of five autism cases and four in-house controls have been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/) under accession number EGAS00001000928.

10.1 Sequencing and variant calling

Genomic DNA of the autism cases was extracted from brain tissues, and sequenced by Complete Genomics (CG) (*78*), with average depth of coverage >40x. Around 10ug of non-degraded DNA was provided for whole genome sequencing as required by CG. Sequencing reads were aligned to the human reference genome GRCh37. Genetic variations that occur in known genes, including a description of the variation's putative effect on the protein (for example, frameshift, nonsynonymous, etc.) were annotated in-house by CG. Other information, such as small insertions/deletions (indels), CNVs, structural variants (SVs) were also provided. Variations in the genome that correspond to known polymorphisms in dbSNP were annotated for each variant detected. All the low-quality variant calls were filtered (varScore less than 20 for homozygous calls and less than 40 for other calls). Only SNVs were kept; all heterozygote SNVs in which both alleles mismatched the reference allele were also removed. The median number of high quality SNVs per subject is ~3.2M. Lastly, we removed all SNVs already observed in dbSNP138, leaving a median of 42K rare SNVs per subject. More detailed information is provided in Table S12.

10.2 Checking systematic biases among subjects

In order to ascertain that there are no systematic biases between case and different control sets, we added subjects from African ancestry in Southwest USA (ASW; 'NA19700', 'NA19701', 'NA19703', 'NA19704'), subjects from Japanese in Tokyo, Japan (JPT; 'NA18940', 'NA18942', 'NA18947', 'NA18956'), and three in house subjects: two from Middle East (MEA) and one Indonesian-Canadian (INC). We computed a genomic distance between any two subjects by: 1) Taking the union of all high quality SNVs from 29 subjects (12 used in the study and 17 for checking the biases) in chromosomes 1 to 22 (~12.6M SNVs in total); 2) If a subject's SNVs did not include a SNV in the union, we assumed its alleles match the reference allele; 3) For any two subjects, we summed the Manhattan distance across all 12.6M SNVs in the union, such that the distance between matching homozygotes is zero, between heterozygote and homozygote is one, and between two different homozygotes is two. As shown in Fig. S27, a dendrogram formed by genomic distances between any two subjects shows clear separation between Caucasians and non-Caucasians, while no discernable bias is observed between ASD subjects and other Caucasian subjects.

10.3 Analysis using our splicing regulatory model

For the 5 cases and 12 controls, we mapped their high-quality, rare SNVs onto *Triplet-Ensembl-canonical* and used our regulatory model to predict the variant-induced $\Delta\Psi$. Since the majority of SNVs are either intergenic or very deep into introns (more than 300nt away from the closest splice site), only a fraction of variants were analyzed; the median number of variants per subject is 1,035. Tables S13.1-13.4 list the code-predicted $\Delta\Psi$ for the subjects in each of the four sets (one case and three control sets) as well as other relevant information.

10.4 Targeted functional enrichment analysis to assess different $\Delta\Psi$ thresholds

Biologically relevant splicing alterations in ASD subjects are supposed to specifically affect genes with a known role in neurodevelopmental pathways, or at least expressed at higher levels in brain. To evaluate different $\Delta\Psi$ thresholds, we tested the enrichment of ASD subjects compared to control subjects in genes with predicted splicing alterations, using the following curated and neurally-related gene-sets. This analysis was intended to select $\Delta\Psi$ thresholds to achieve higher biological specificity.

1) **DevCNS**: Central nervous system development (GO:0007417)
2) **Synaptic**: Synaptic components and regulators, set union of genes in (*79*) and (*80*).
3) **NeuroPh**: Genes associated with neurodevelopmental or neurobehavioral phenotype in human or mouse.
4) **BrainExpr_HI**: Genes in the top 25% expression tier, requires support from at least 5 data points (including replicates) based on BrainSpan RNA-seq RPKM values as available from the Allen Brain Atlas resource.
5) **BrainExpr_AL**: Genes in the top 75% expression tier, requires support from at least 5 data points based on BrainSpan RNA-seq RPKM values.

First of all, we found that only reduction in exon inclusion produced a significant enrichment in neurally-related gene-sets; see Fig. S28. As expected, we typically found the highest enrichment at the most stringent $\Delta\Psi$ thresholds (1st, 2nd and 3rd percentiles),

with a decline at the less stringent thresholds (4th and 5th percentiles). However, the 1% threshold tended to have inconsistent results using different control subsets, probably because of stochastic variability caused by the small number of genes with predicted alterations (Fig. S29). Therefore, we used the 2% and 3% thresholds for all other ASD analyses. We additionally show that enrichment tests using all controls, or only a subset of controls (based on ethnic groups), all display enrichment of neurally-related gene-sets in ASD subjects, suggesting that ethnic differences are not driving results (Fig. S29).

Enrichment was tested using Fisher's Exact Test: the contingency matrix for the test was composed by counting the number of genes with predicted splicing alterations in ASD subjects or controls, within or outside the tested gene-set; genes with predicted splicing alterations in both ASD subjects and controls were discarded prior to the contingency matrix construction, as they would introduce double counts. Notably, this approach is robust to biases such as gene length or variability (propensity to have genetic variants at the locus), because the same biases are expected to be present in ASD subjects as well as control subjects and thus cancel out.

These five gene sets have some overlaps and their overlap percentages are listed in Table S14. Fig. S28 & S29 depict the enrichment ratios that are calculated as:

$$\frac{(\text{ASD\_predicted\_gene-set})/(\text{ASD\_predicted\_total})}{(\text{Control\_predicted\_gene-set})\,/\,(\text{Control\_predicred\_total})}.$$

10.5 Complete functional enrichment with ΔΨ thresholds of 2nd and 3rd percentiles

We tested enrichment for all gene-sets derived from Gene Ontology annotations and pathway databases using predictions of reduced exon inclusion at 2nd and 3rd percentiles. With $P < 0.01$ (Fisher's exact test), there are 8 gene sets for the 2nd percentile and 16 gene-sets for the 3rd percentile (Tables S15.1-2). To further assess the significance of our results, we performed two more experiments. First we swapped cases and controls and redid the enrichment analysis (also listed in Tables S15.1-2 as the 'Inverse p-value' column), which resulted in 2 and 1 GO terms with $P < 0.01$ for the 3rd and 2nd percentiles, none of which are neuro-related. Second we performed enrichment analyses by comparing one control group against another and found no significant GO terms, with the detailed results listed below:

a) CEU vs IMS, 3rd percentile: 0 significant GO terms, smallest $P > 0.05$.
b) CEU vs IMS, 2nd percentile: 0 significant GO terms, smallest $P > 0.1$.
c) TSI vs IMS, 3rd percentile: 0 significant GO terms smallest $P > 0.01$, only GO:0005975 (carbohydrate metabolic process) has a $P < 0.05$.
d) TSI vs IMS, 2nd percentile: 0 significant GO terms, smallest $P > 0.05$.

10.6 Computing empirical FDRs

To compute empirical false discovery rates (FDRs), we performed the following random permutation test. For the case group and each control group, *Ensembl* gene IDs were randomly permuted and the genes with code-predicted ΔΨ below 2nd and 3rd percentile thresholds were selected. This permutation and sampling procedure ensures that genes over-represented in affected exons are adequately represented, thus modeling selection biases that would be missed by a random uniform gene sampling. Each Gene Ontology and pathway derived gene-set was also tested for enrichment using Fisher's Exact Test as described before. A panel of brain and neuron-related Gene Ontology terms

(including all their offspring terms) was defined to assess the specificity of enrichment results to ASD:

- GO:0097458 neuron part
- GO:0043005 neuron projection
- GO:0045202 synapse
- GO:0007399 nervous system development
- GO:0007417 central nervous system development
- GO:0030182 neuron differentiation
- GO:0050808 synapse organization
- GO:0019226 transmission of nerve impulse

Offspring terms were extracted from the Bioconductor package GO.db 2.9.0. The permutation and enrichment procedure was iterated 500 times and the following statistics were recorded:

1) The mean and median number of gene-sets passing the 0.01 nominal *p*-value for the 500 iterations
2) The mean and median number of brain and neuron-related gene-sets passing the 0.01 nominal *p*-value for the 500 iterations
3) The empirical FDR, defined as mean or median number of 0.01-significant gene-sets observed for the permuted data divided by the number of 0.01-significant gene-sets for the original data
4) The empirical FDR when restricting to brain and neuron-related gene-sets
5) The fraction of iterations with the same or more 0.01-significant gene-sets than original results (interpretable as a 'global' enrichment significance *p*-value)
6) The fraction of iterations with the same or more 0.01-significant brain and neuron-related gene-sets than the original results

We found permuted data produced a smaller but consistent number of significant gene-sets, but they did not produce a significant enrichment in brain and neuron-related gene-sets. When using all gene sets to assess the original (not permuted) gene enrichment results, the mean-based empirical FDRs are 0.4605 and 0.4175 for the 2rd and 3nd percentile thresholds respectively, or median-based empirical FDRs of 0.25 and 0.25, and iteration fractions of 0.138 and 0.1, which are not significant. When restricting to brain and neuron-related gene-sets, the mean-based empirical FDRs become 0.0316 and 0.0384 for the 2rd and 3nd percentile thresholds, median-based empirical FDRs are less than 0.002 for both, and iteration fractions are 0.01 and 0.006, which are all significant.

10.7 Enrichment analysis for brain-expressed genes

To test the enrichment for Brain-related and Brain-agnostic genes, we defined the following BrainSpan derived gene-sets, of roughly equal size, based on brain expression levels:

- BSpan_VH_thr4.86: genes with at least 5 BrainSpan (available from http://developinghumanbrain.org) data points for which log2(rpkm) >= 4.86, thus deemed expressed at (very) high levels in brain.
- BSpan_HM_thr3.32: genes with at least 5 BrainSpan data points for which 4.86 > log2(rpkm) >= 3.32, thus deemed expressed at high/medium levels in brain.
- BSpan_ML_thr0.84: genes with at least 5 BrainSpan data points for which 3.32 > log2(rpkm) >= 0.84, thus deemed expressed at medium/low levels in brain.

- BSpan_Ab_thr.MIN: genes with BrainSpan data points failing all previous criteria, thus deemed expressed at very low level or not expressed in brain.

The enrichment analysis is done as described before and the results are listed in Table S18.

11. Development of the mutation analysis web tool

To make our human splicing code available to all medical researchers and practitioners, we have developed a web tool that allows anybody to access and run variant analyses on our servers. The web tool is comprised of two parts: a webserver that hosts the web application and database, and a computational server that runs the resource intensive splicing regulatory model. Both servers run Ubuntu Linux. The web application is written in Python using the Flask web framework (http://flask.pocoo.org/) and stores application data in a MongoDB database (http://www.mongodb.org/). The computational model of splicing is implemented in Matlab while the feature extraction pipeline is mostly written in Perl, with the use of third party tools to compute features based on RNA secondary structures, nucleosome positioning and Alu repeats. The webserver communicates with the computational server using the Celery distributed task queue (http://celery.readthedocs.org/). The predictions generated by the computational server are stored in a MongoDB database on the webserver so that predictions on the same variants, even submitted by different users, are not computed twice.

The mutation analysis web tool, available at <http://tools.genes.toronto.edu>, provides an easy-to-use interface for users to input one or more SNVs in standard VCF format, where each entry specifies the genetic locus, wild type nucleotide, variant nucleotide, and optionally an identifier chosen by the user to reference the SNV. Fig. S30 shows a screenshot of the job submission page. Upon submission, the tool automatically determines if there are any exons affected by the submitted SNVs. If yes, it further runs our splicing regulatory model and displays the prediction results for them. We have carefully designed the result pages to clearly show all relevant information and statistics. On top of the result page for a completed job, it shows the mapping of SNVs to internal RefSeq exons, explaining any cases in which a SNV cannot be analyzed, *e.g.*, because it is too deep into the intron or not inside an annotated gene. Below, predictions from our splicing code are shown in a table with the following information:

- $\Delta\Psi$ between the wild type and variant
- The percentile of the mutant $\Delta\Psi$ among all common SNP $\Delta\Psi$'s
- The log-regulatory score
- The percentile of the regulatory among all common SNP regulatory scores
- The wild type Ψ

Fig. S31 shows a screenshot for part of this global result page. Clicking on any exon in the result table will further bring up a detailed result page with information such as the RefSeq transcript ID, exon number, the coordinates of the cassette exon and the flanking exons, and a list of RNA features changed by the mutation. A screenshot is shown in Fig. S32. Please note that we are continuingly improving our web tool, including both the user interface and the underlying computational engine.

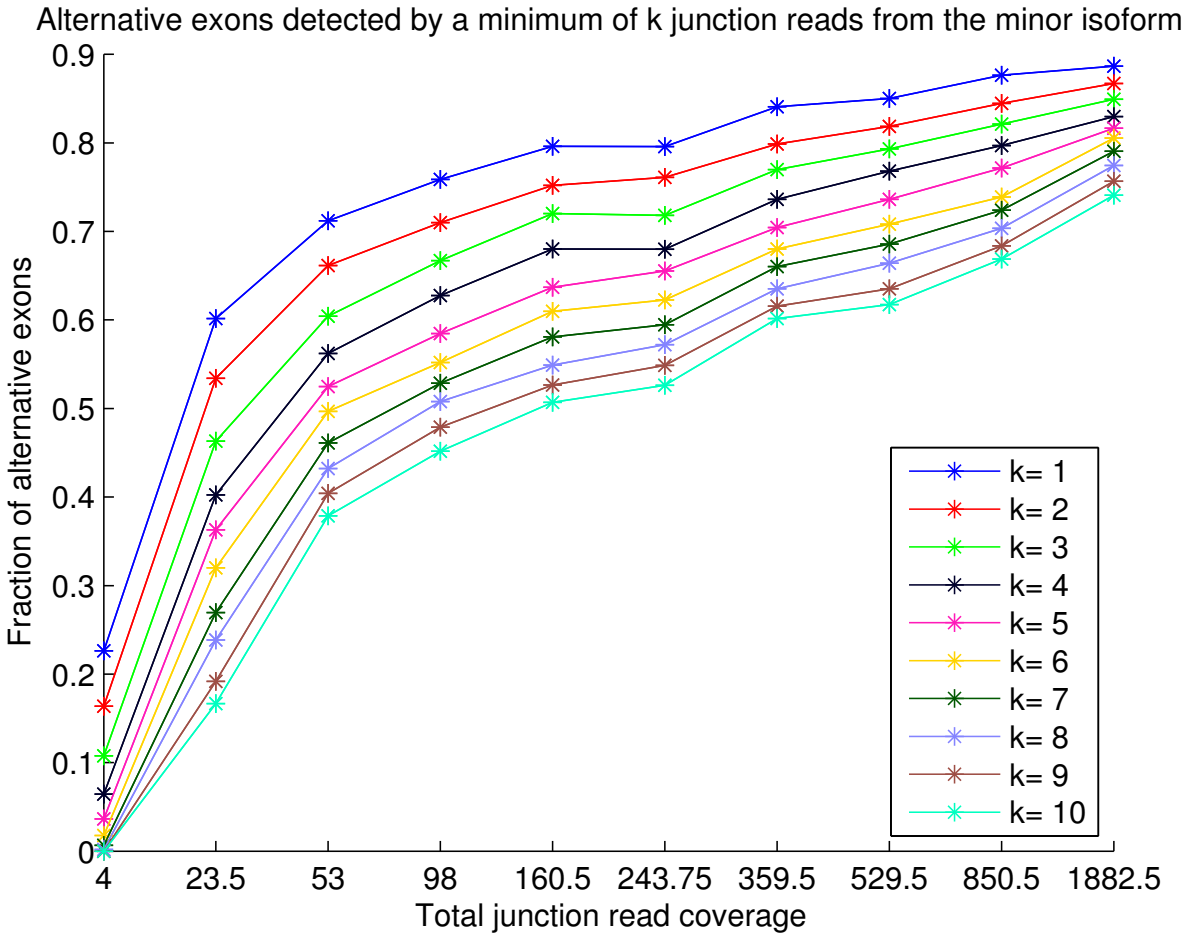

Fig. S1.

Detecting alternative exons by the number of mapped junction reads *k* from the minor isoform across different total junction read coverage over 16 tissues.

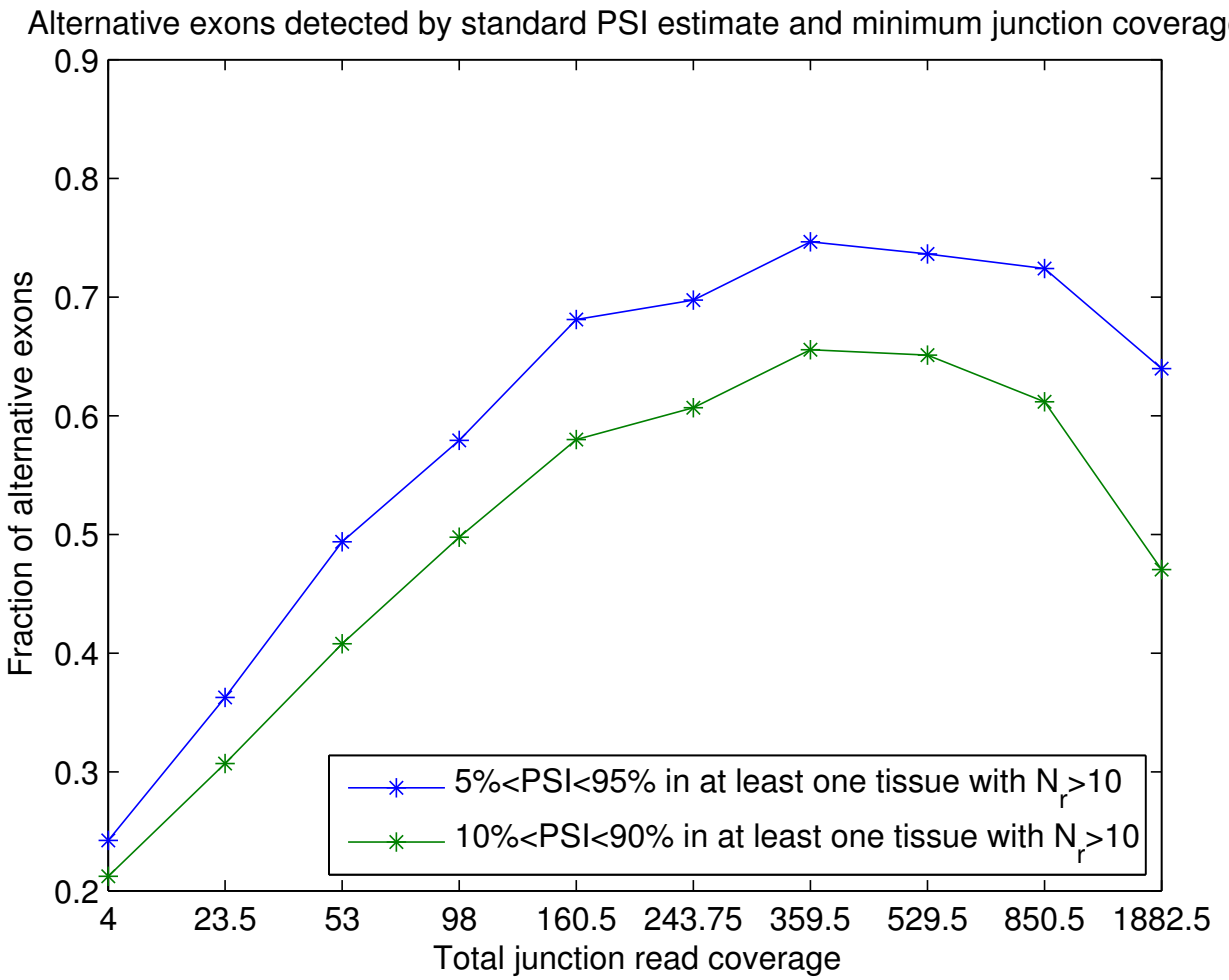

Fig. S2

Detecting alternative exons by restricting PSI and minimum junction coverage.

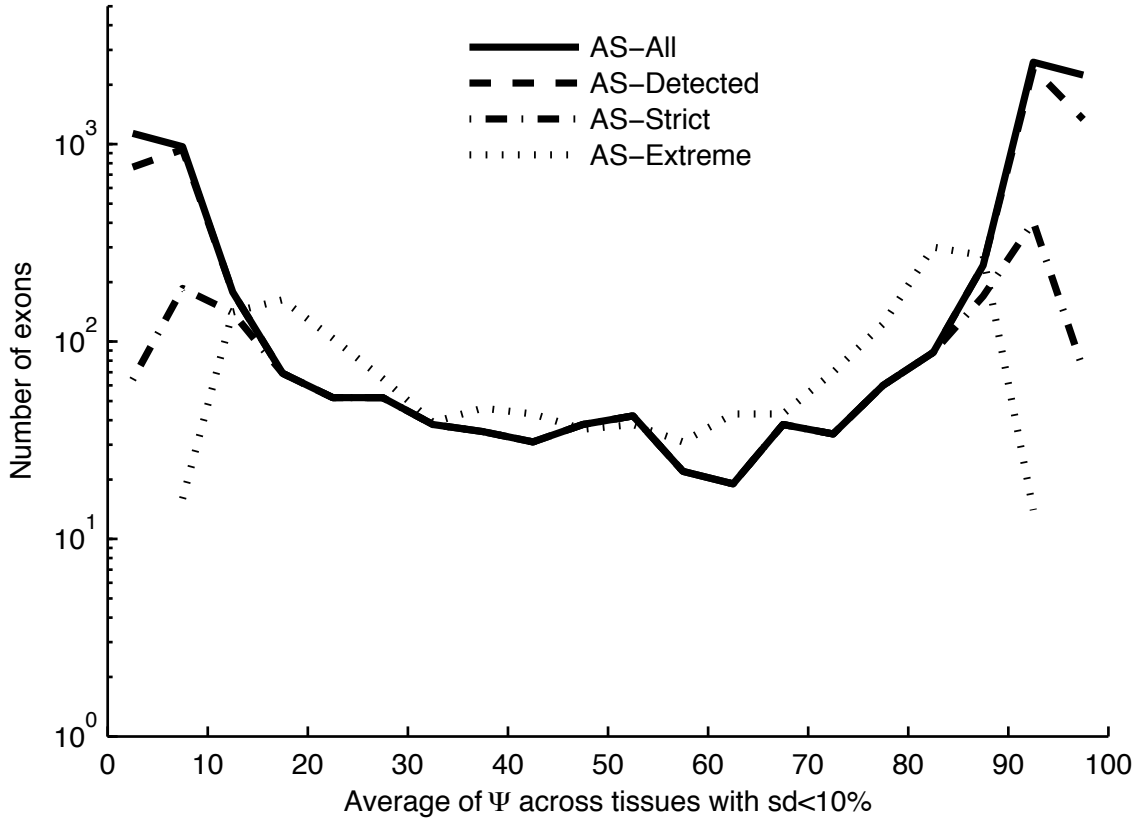

Fig. S3

The distribution of expected Ψ, estimated by positional bootstrap from BodyMap data, over all 16 tissues for different sets of AS events. For AS-All, AS-Detected and AS-Strict, the number of exons are plotted; for AS-Extreme, the number of events or exon-tissue combinations are plotted.

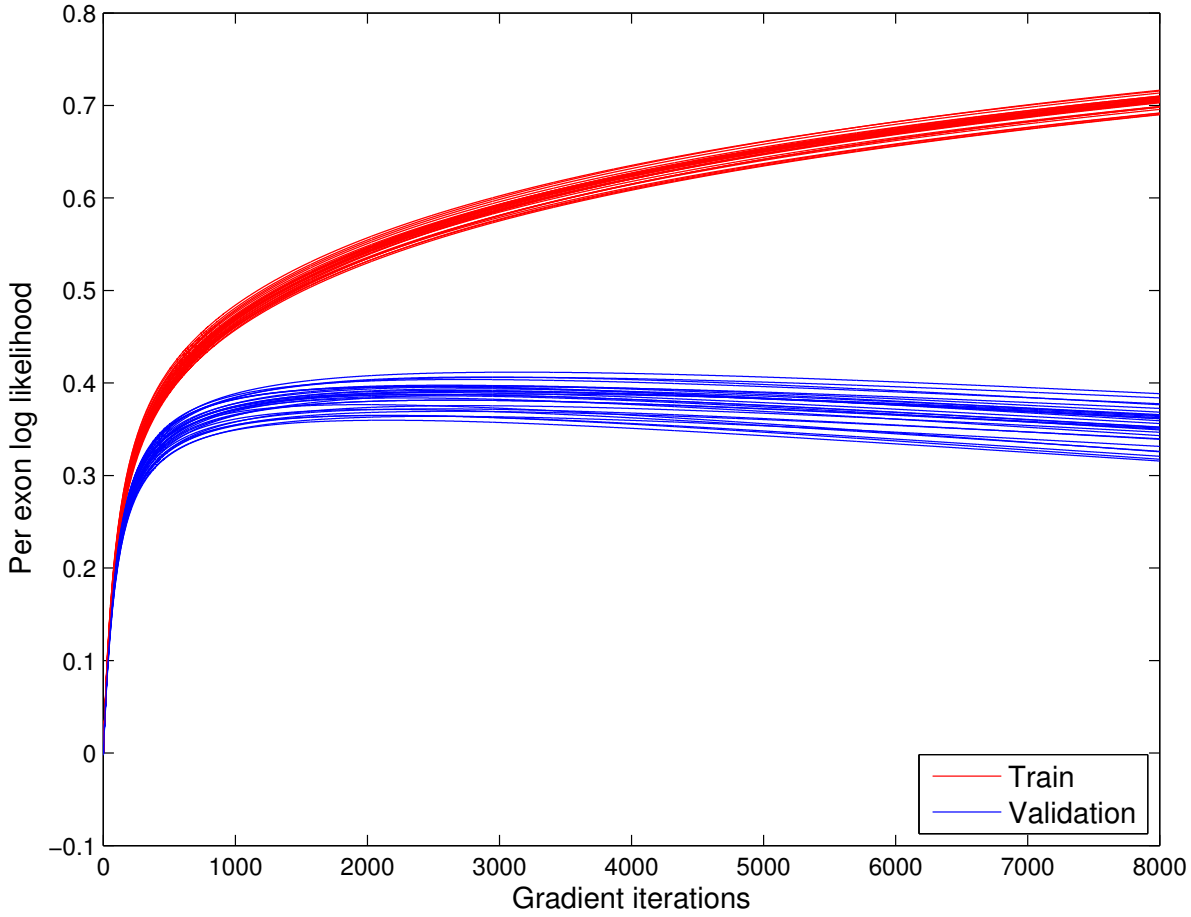

Fig. S4

Training and validation set log-likelihoods for multinomial regression models versus number of gradient iterations, for the low/medium/high dataset (five folds, six partitions).

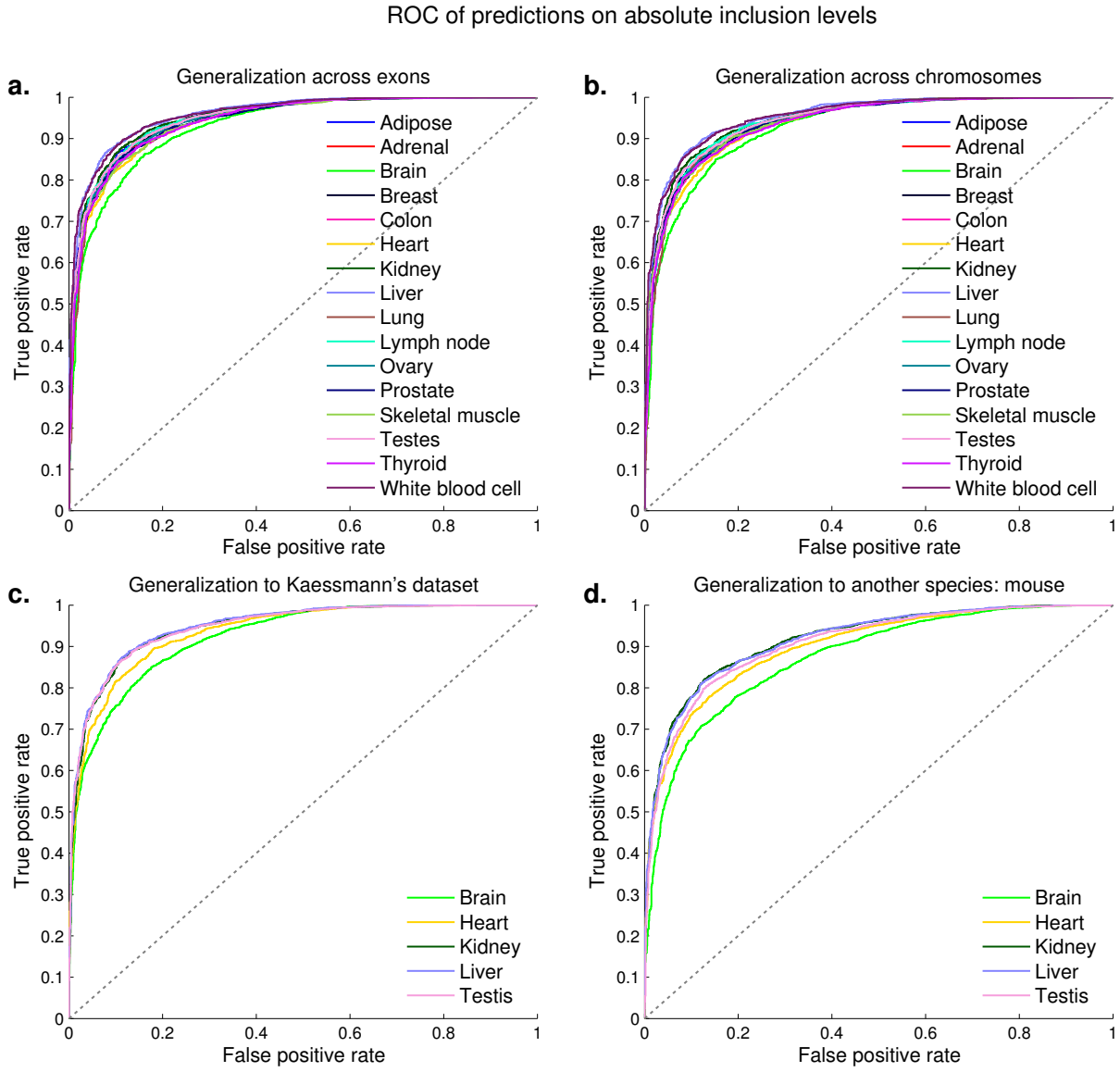

Fig. S5

ROC curves of the proposed model for different prediction tasks. (a) Generalization to held-out exons. (b) Generalization to held out chromosomes. (c) Generalization to an independently prepared dataset. d) Generalization to mouse.

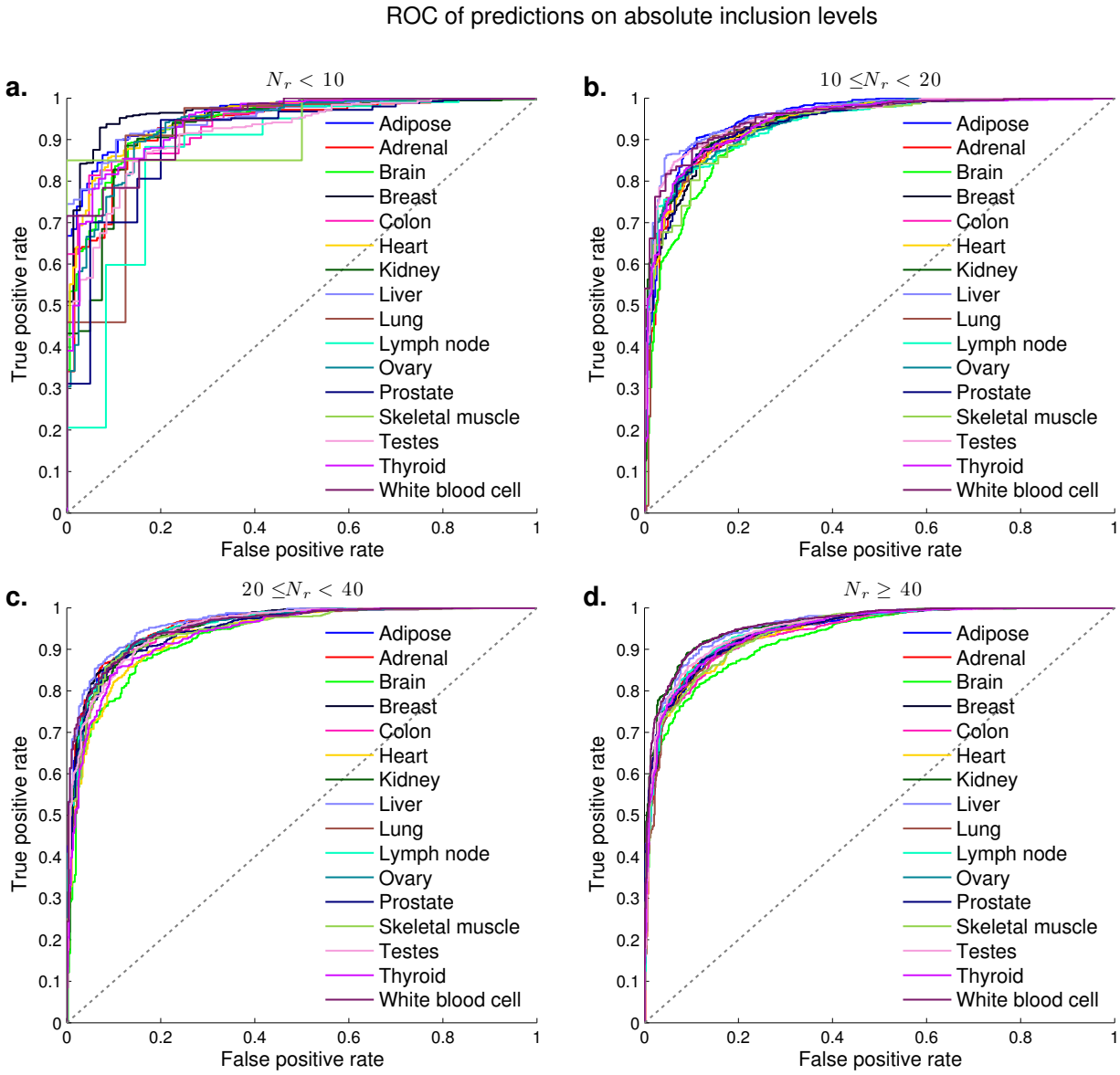

Fig. S6

ROC curves of the inferred splicing code for exons in diverse gene expression ranges. The exons are binned according to the junction coverage indicated at the top of each panel.

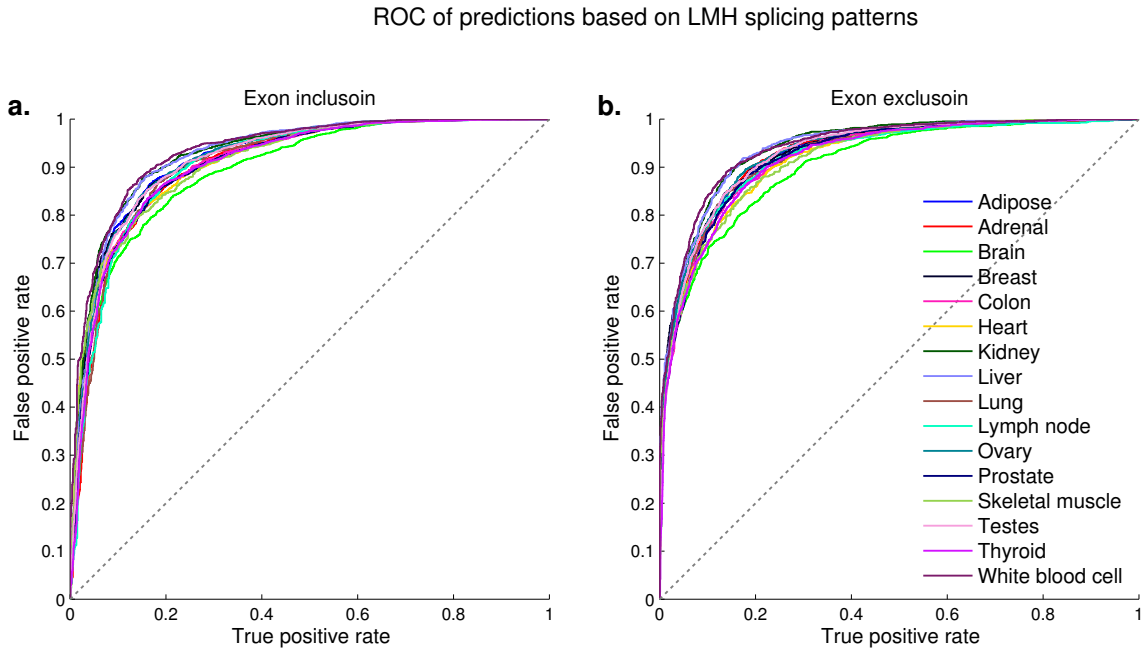

Fig. S7

ROC curves of predictions using low/medium/high labels. Exons are sorted by the predicted probability of being in the high inclusion category and low inclusion category respectively.

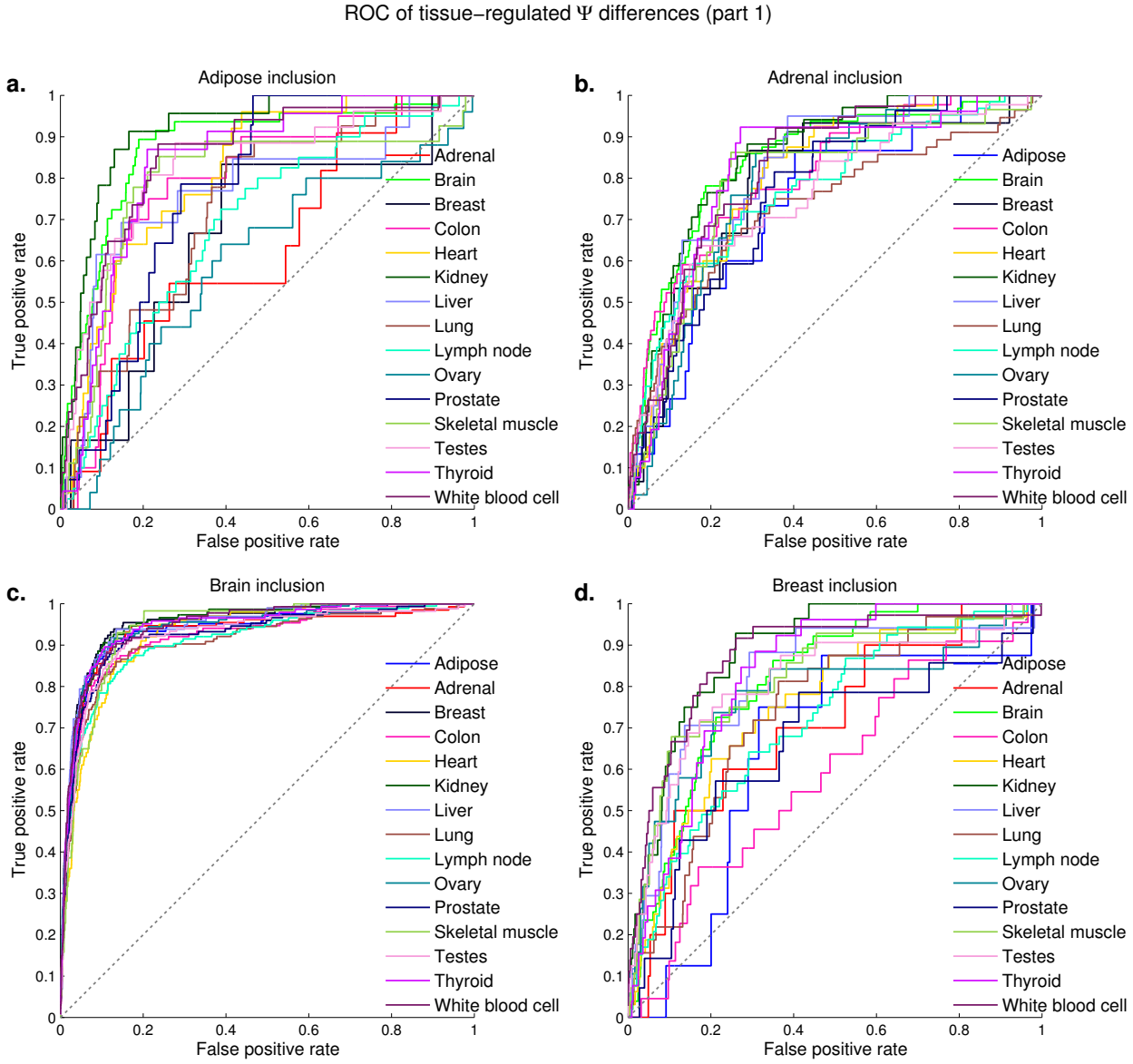

Fig. S8

Part 1 of 4. ROC curves of tissue-regulated Ψ differences: Each panel is for the task of identifying increased inclusion in a particular tissue compared to all other tissues (color coded). For example, comparing panel (c) to other panels, we observe that brain-specific inclusion events were predicted better than many other types of tissue-specific inclusion events

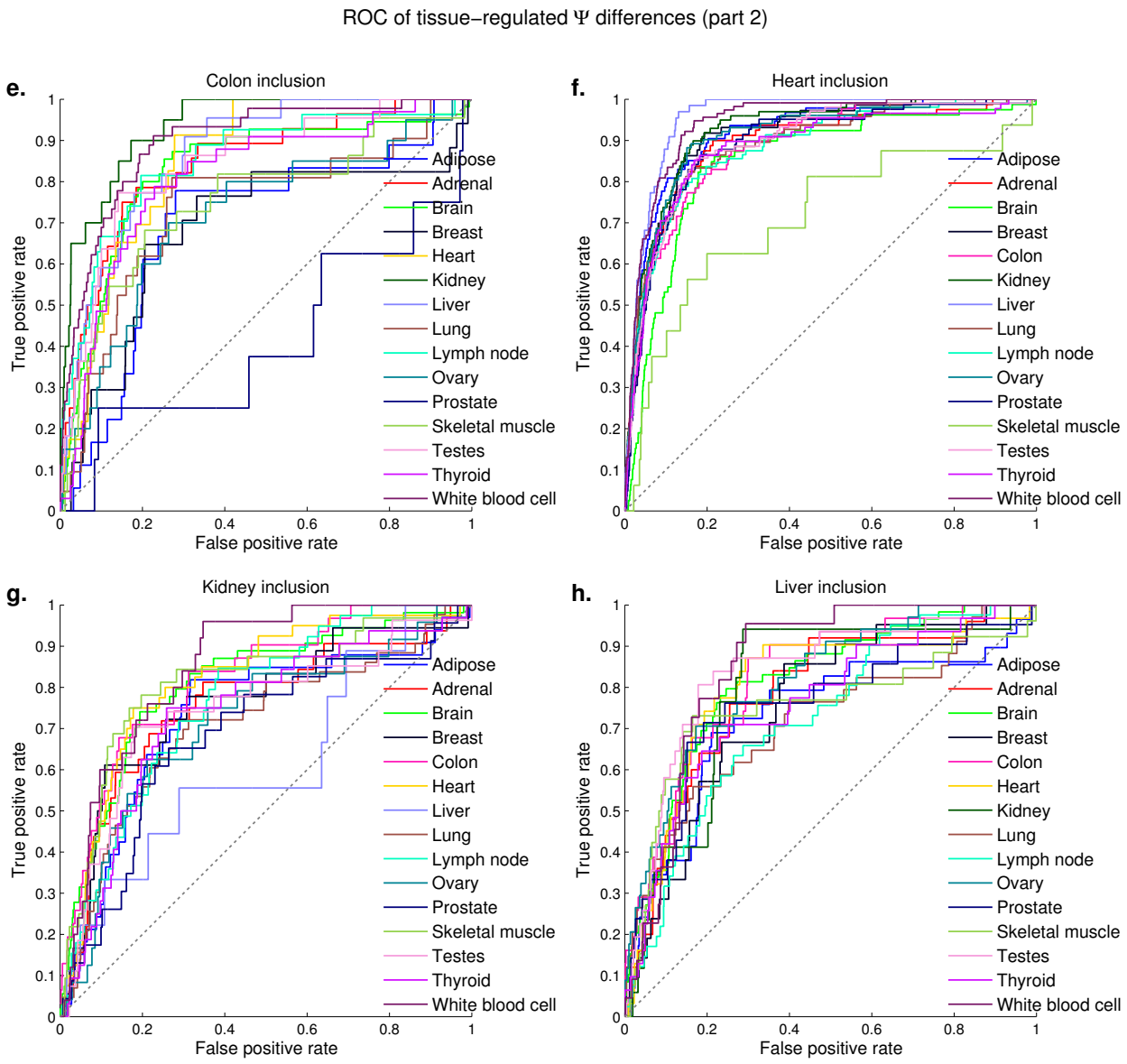

Fig. S8

Part 2 of 4. ROC curves of tissue-regulated Ψ differences: Each panel is for the task of identifying increased inclusion in a particular tissue compared to all other tissues (color coded). For example, comparing panel (c) to other panels, we observe that brain-specific inclusion events were predicted better than many other types of tissue-specific inclusion events

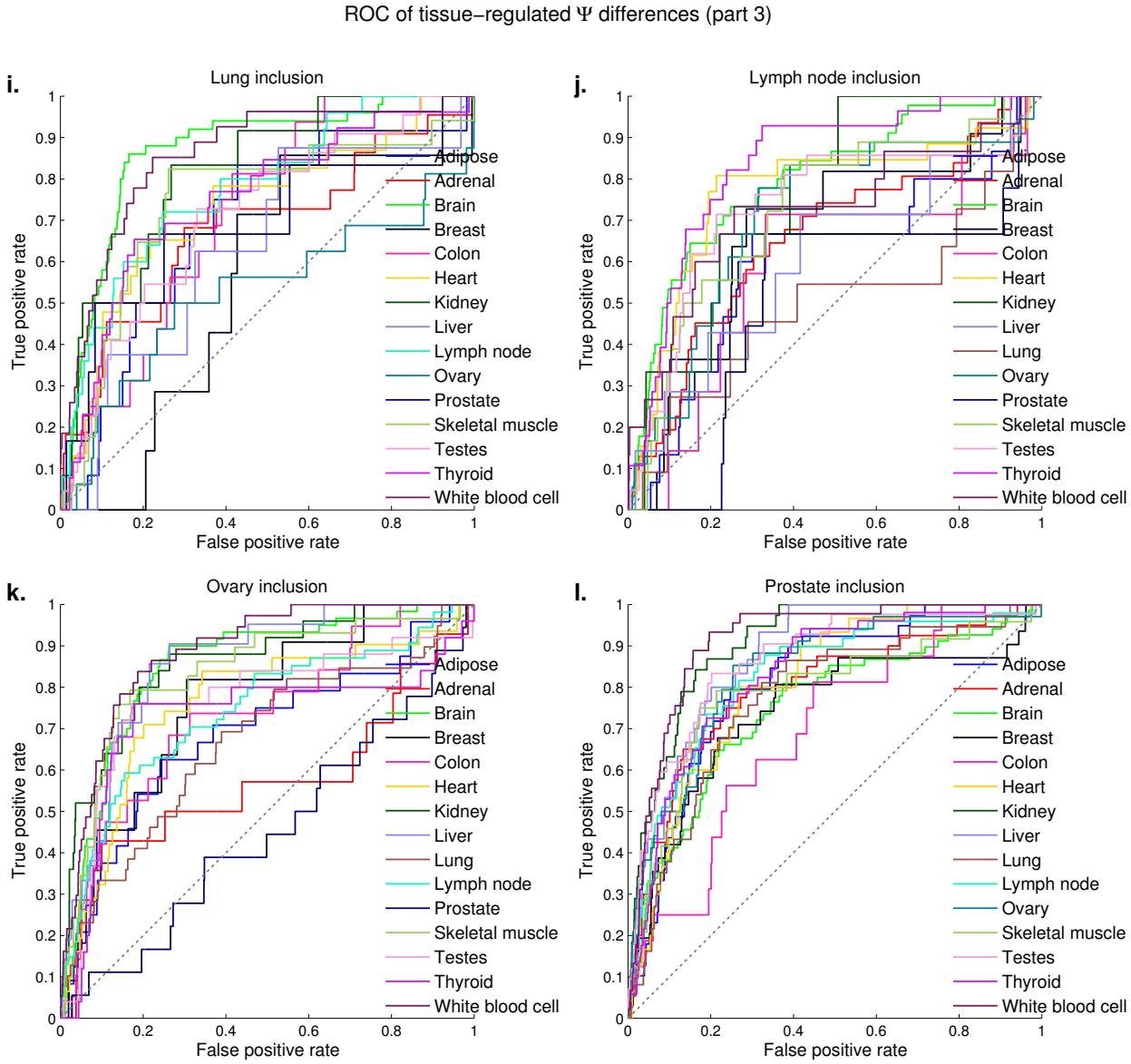

Fig. S8

Part 3 of 4.

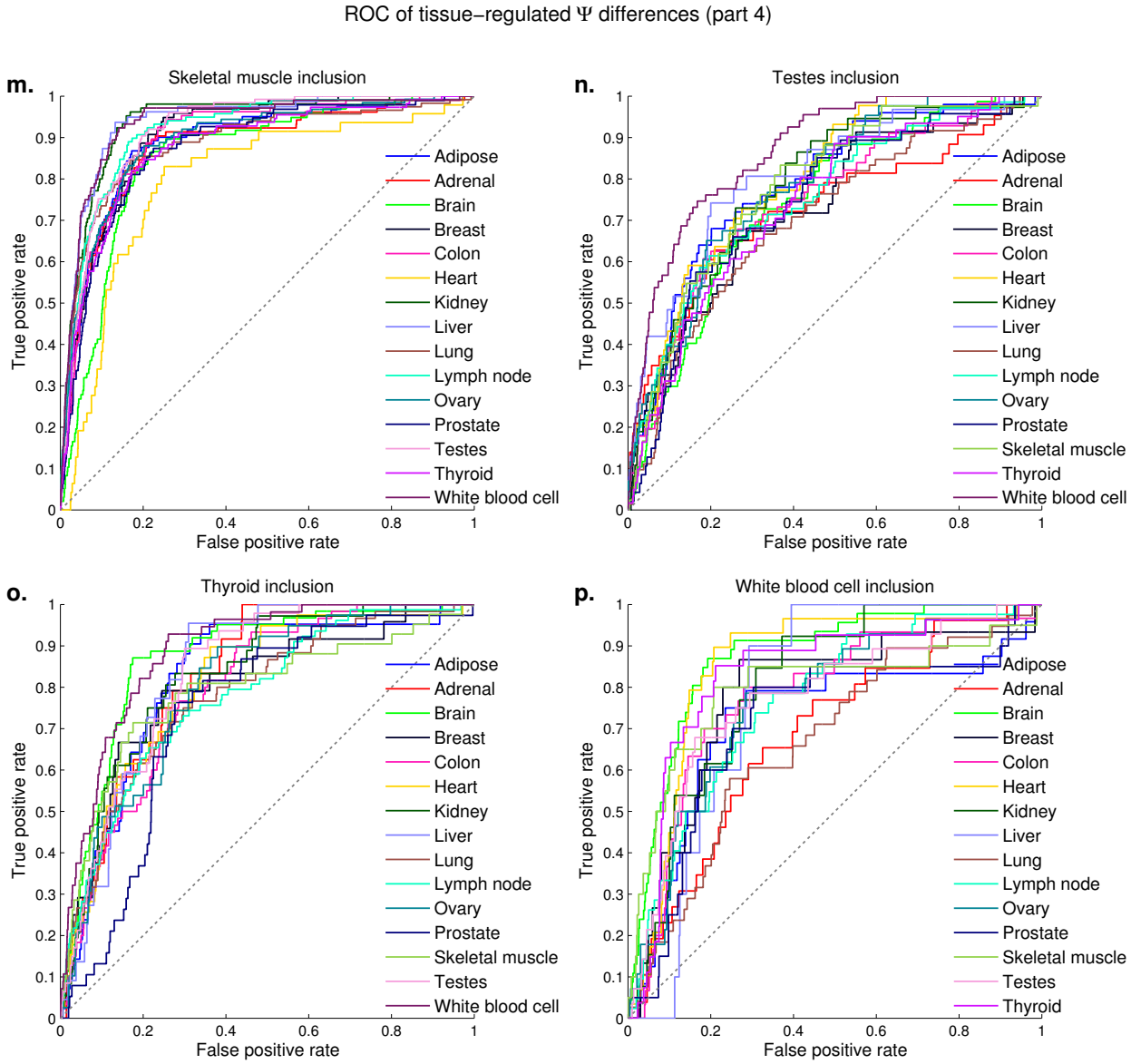

Fig. S8

Part 4 of 4.

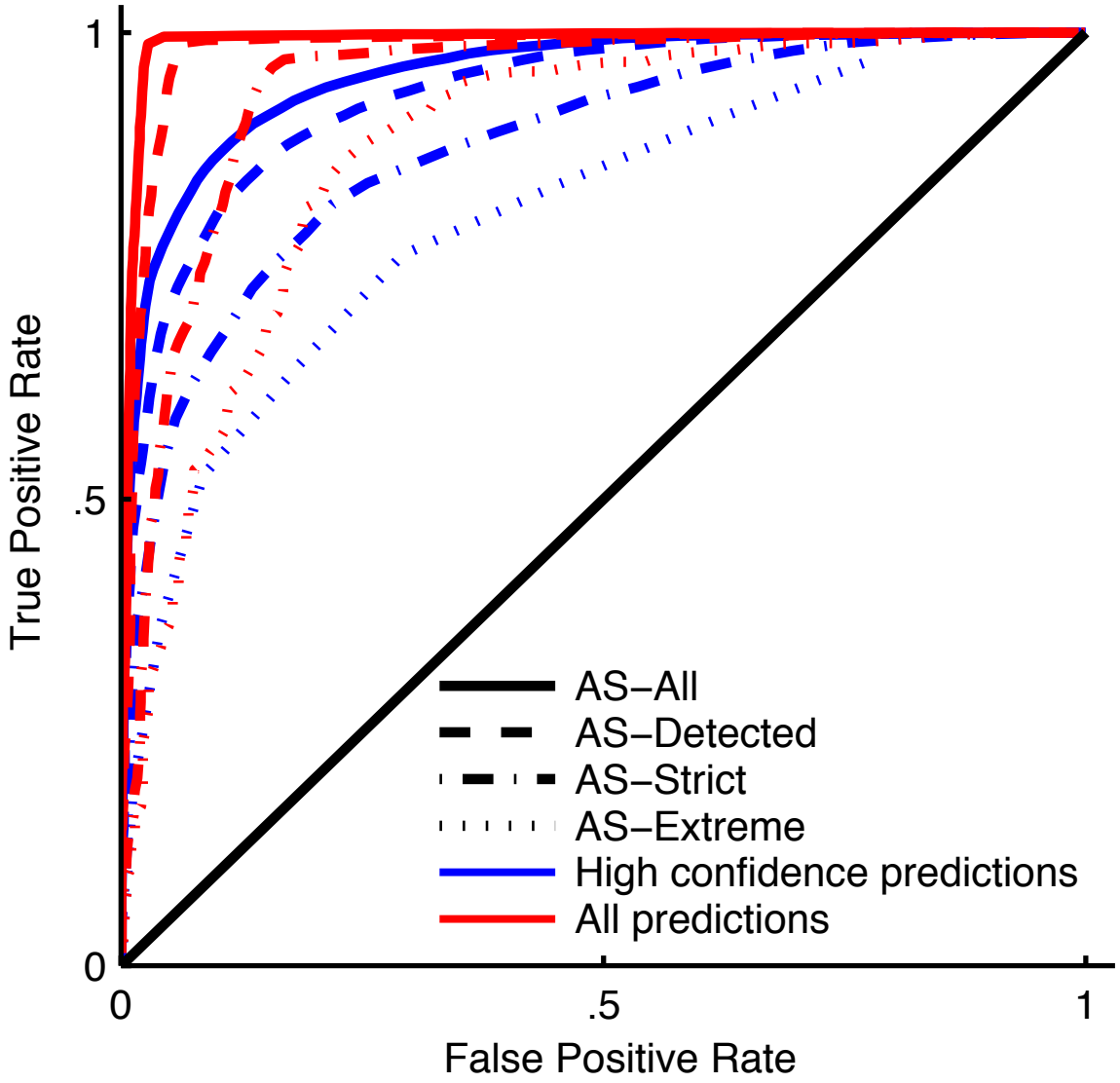

Fig. S9

ROC curves of the splicing regulatory model for different sets of AS events.

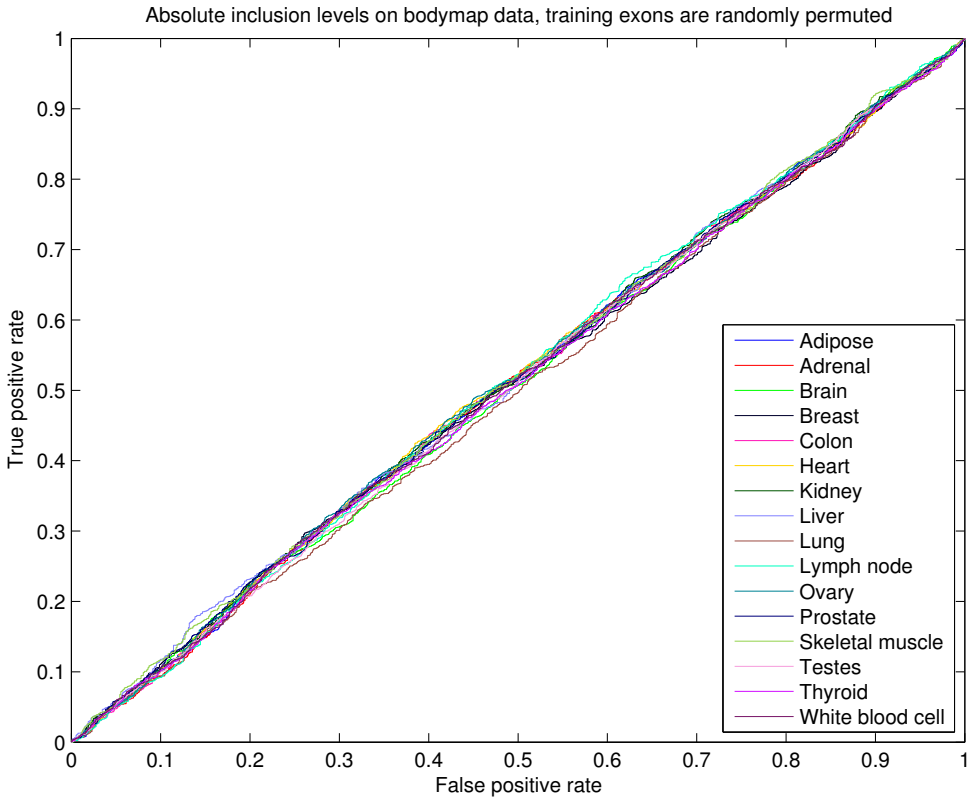

Fig. S10

ROC curves of predicting absolute inclusion with randomly permuted exons.

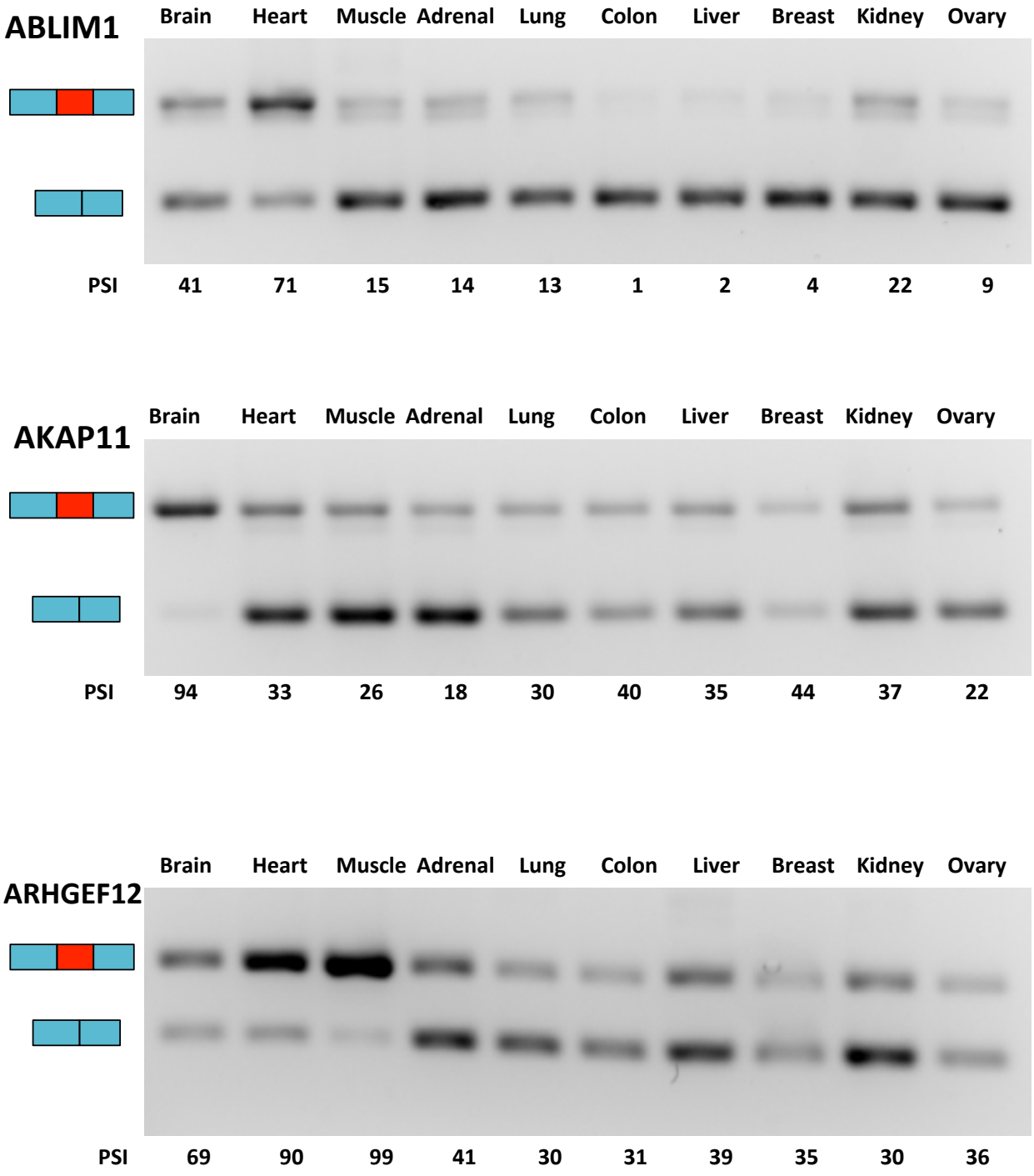

Fig. S11
Part 1 of 5: RT-PCR gel images.

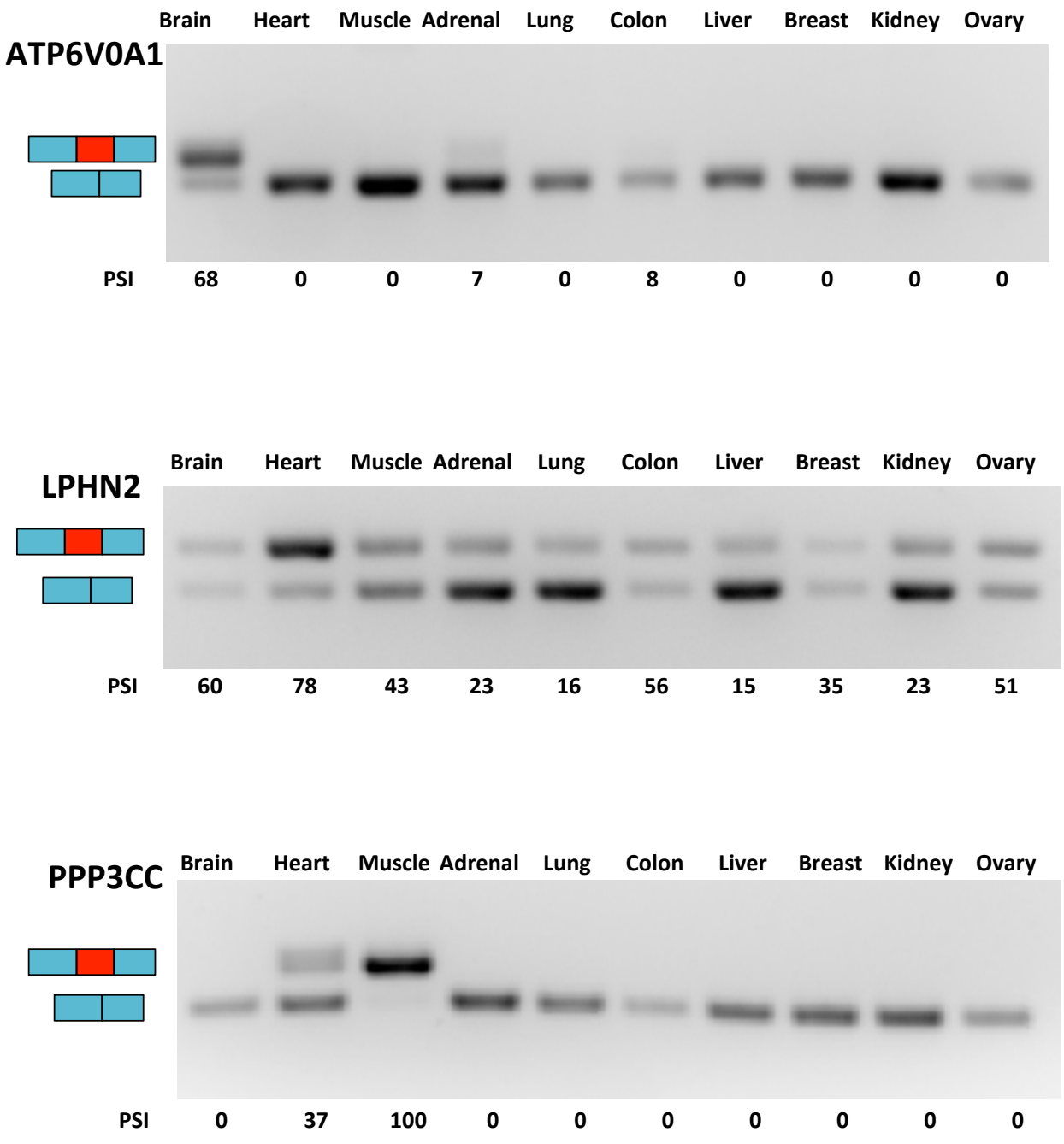

Fig. S11

Part 2 of 5: RT-PCR gel images.

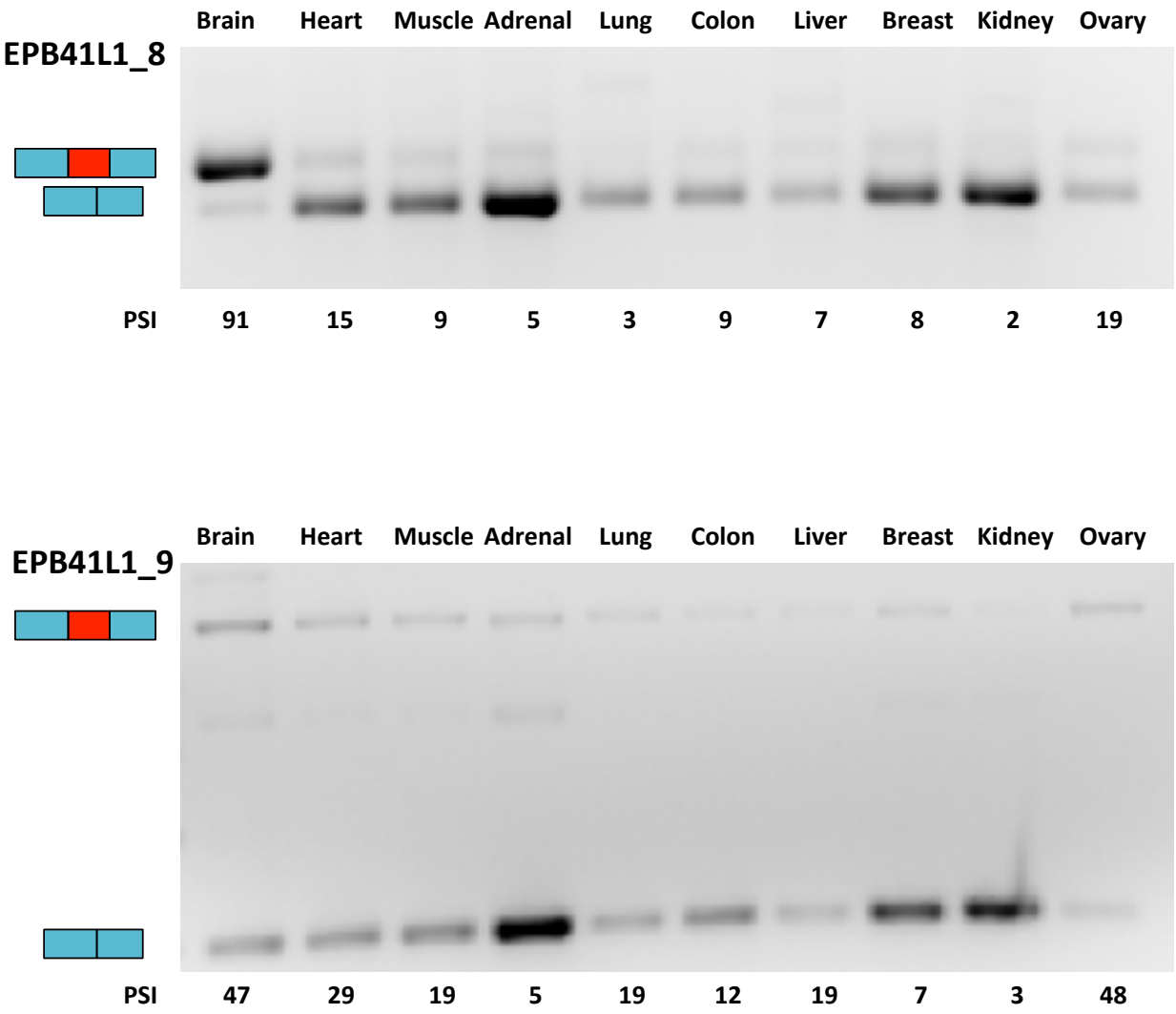

Fig. S11

Part 3 of 5: RT-PCR gel images.

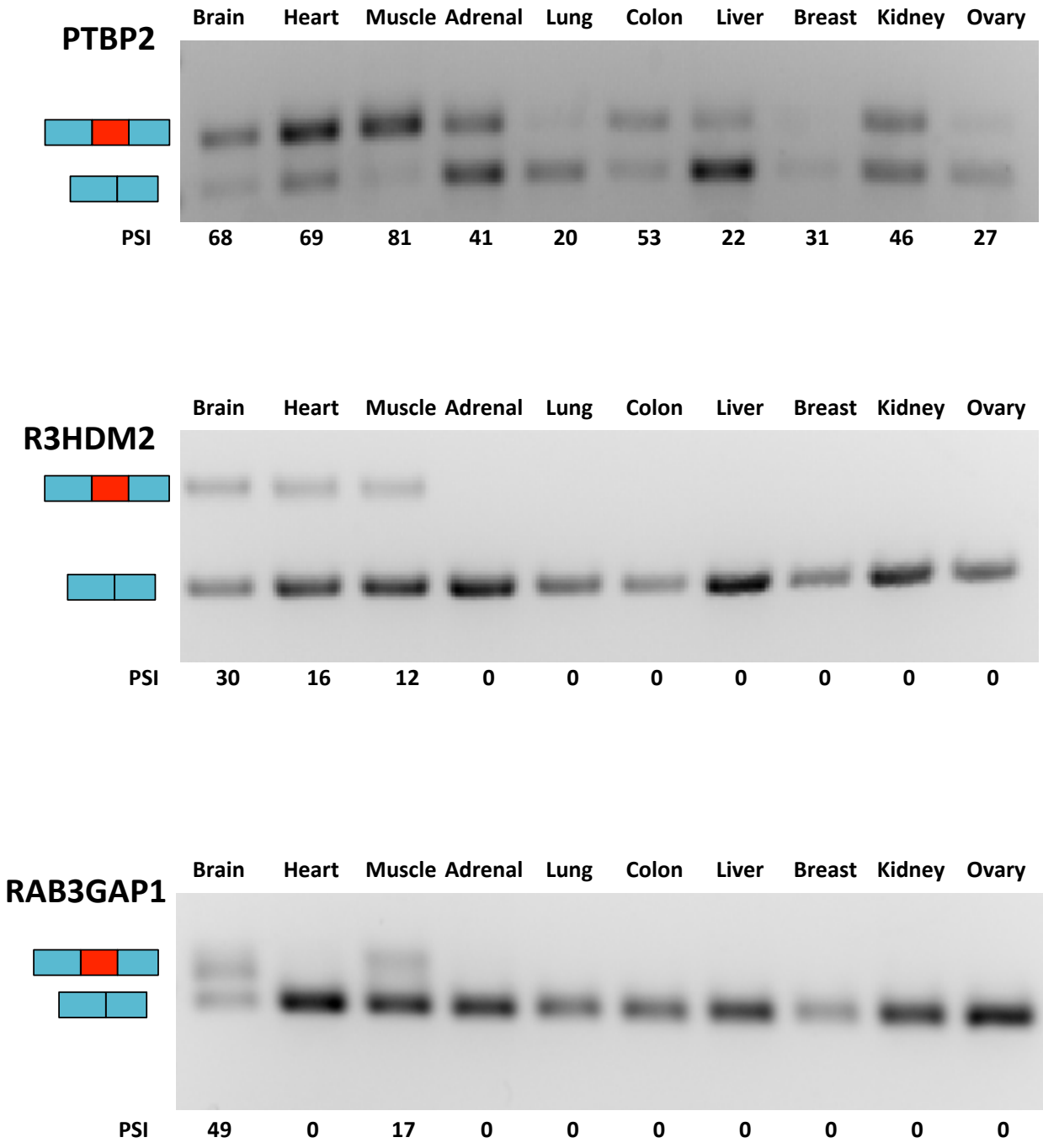

Fig. S11
Part 4 of 5: RT-PCR gel images.

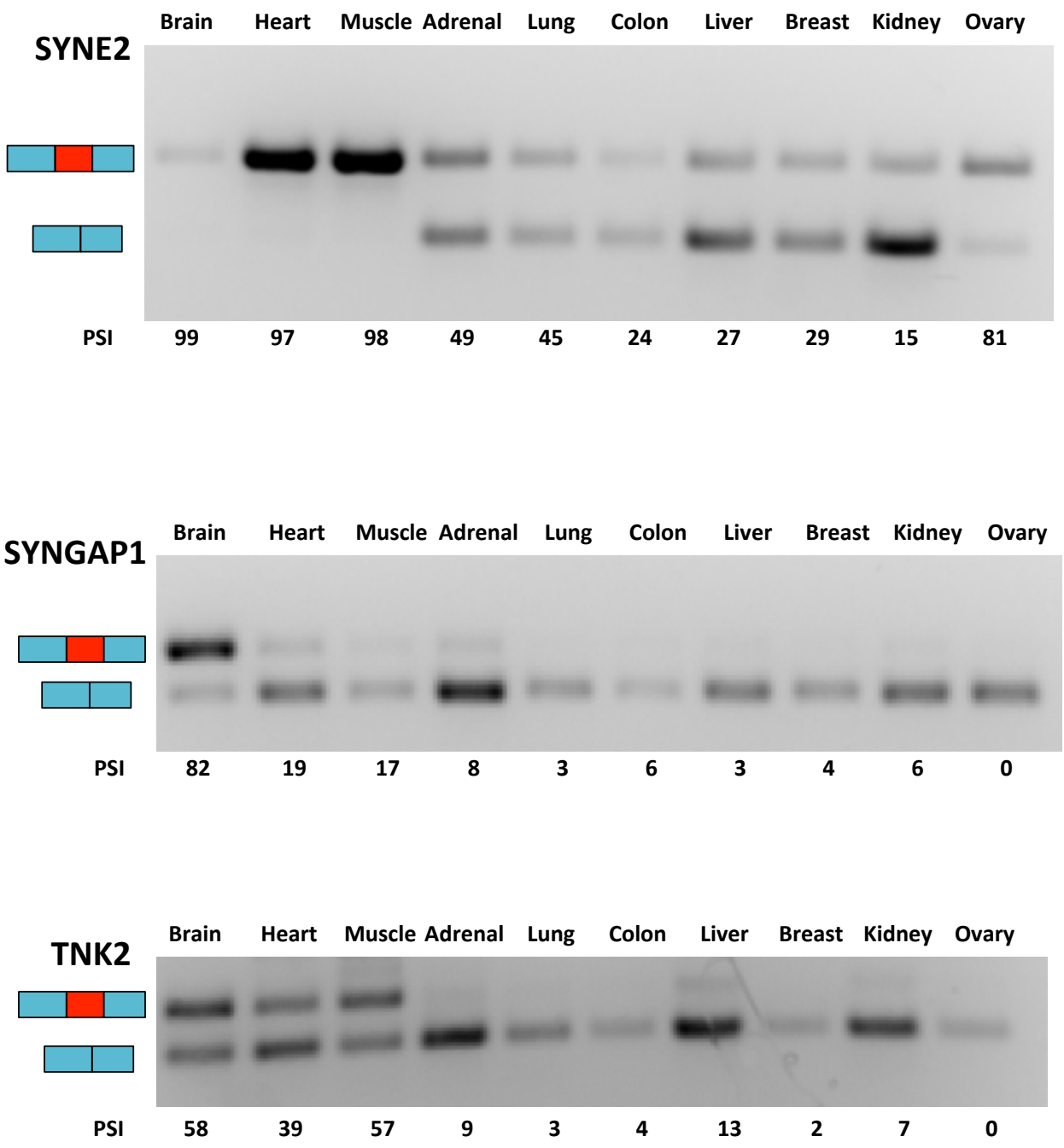

Fig. S11

Part 5 of 5: RT-PCR gel images.

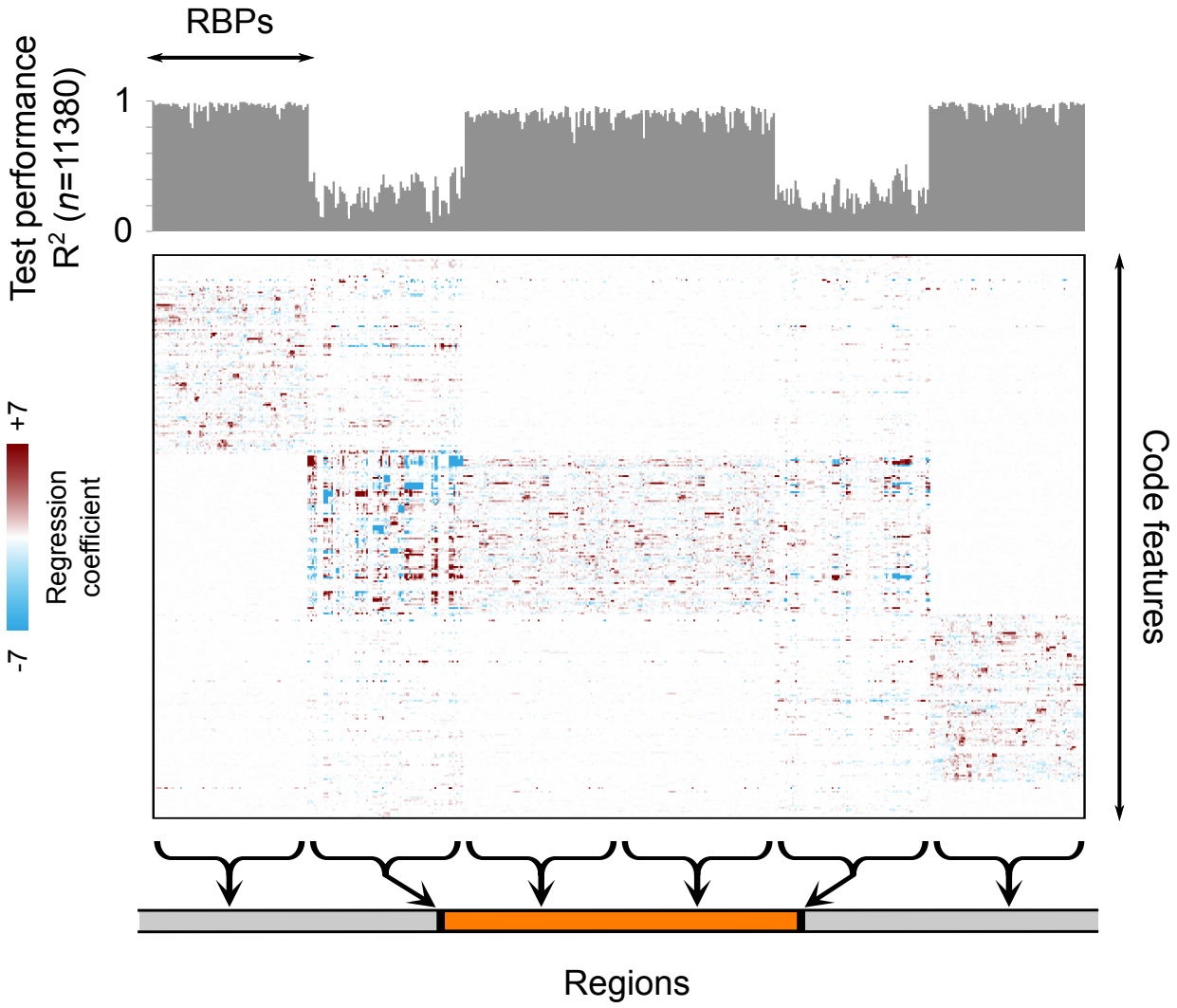

Fig. S12

Test performance and LASSO coefficients in different regions for a linear regression model that associates RBP binding affinities with our splicing code features.

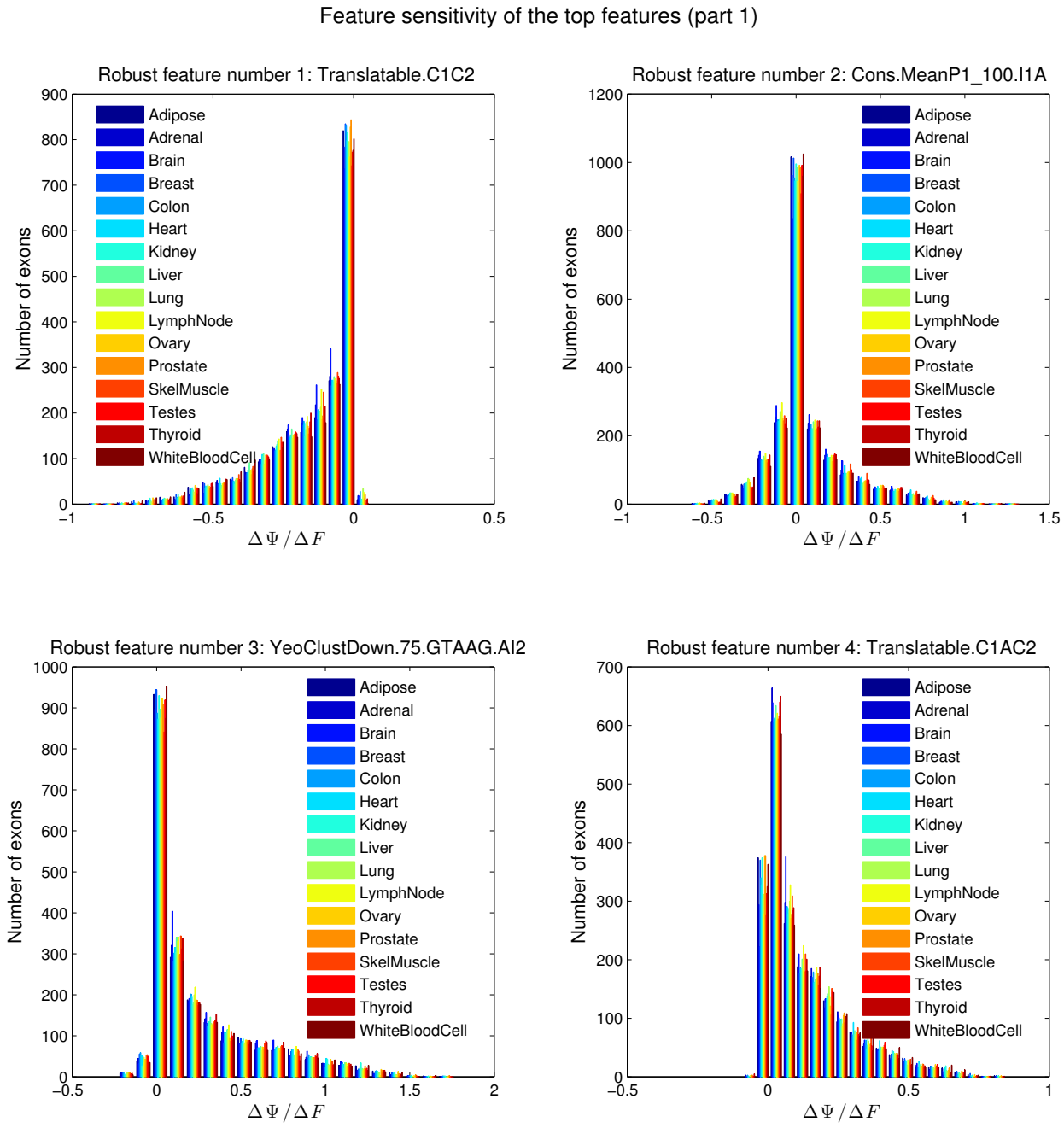

Fig. S13

Part 1 of 4: Histogram of feature sensitivity ($\Delta\Psi/\Delta F$) for the top 16 features. Top features were chosen by sorting the average activity of the feature-to-hidden connections in the Bayesian neural network ensemble. Each feature was scaled to be between -1 and 1 before the sensitivity analysis. We observe that the sensitivities of many features (40 out of the top 100) are highly context-dependent and their sign switch between exons.

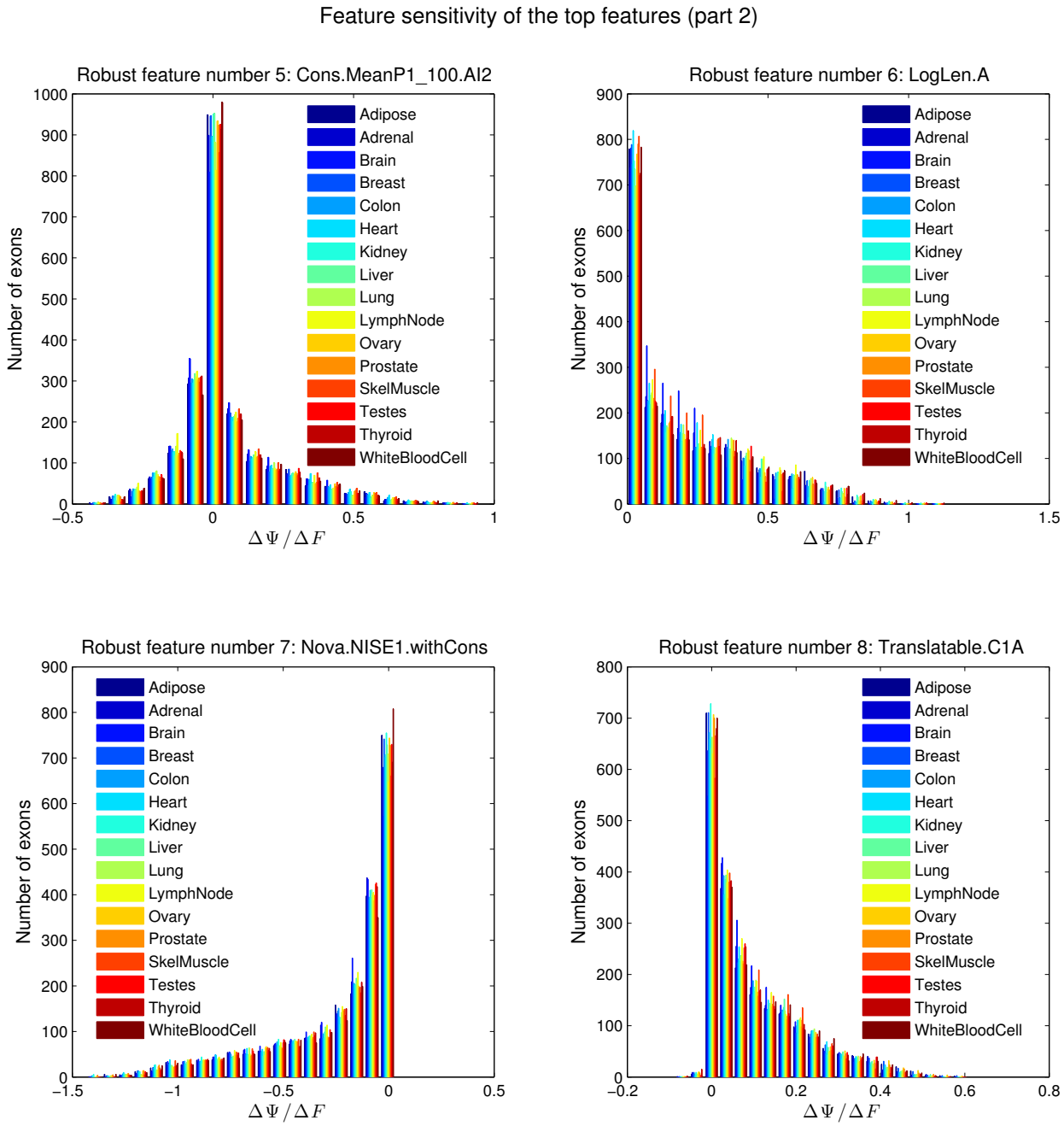

Fig. S13

Part 2 of 4.

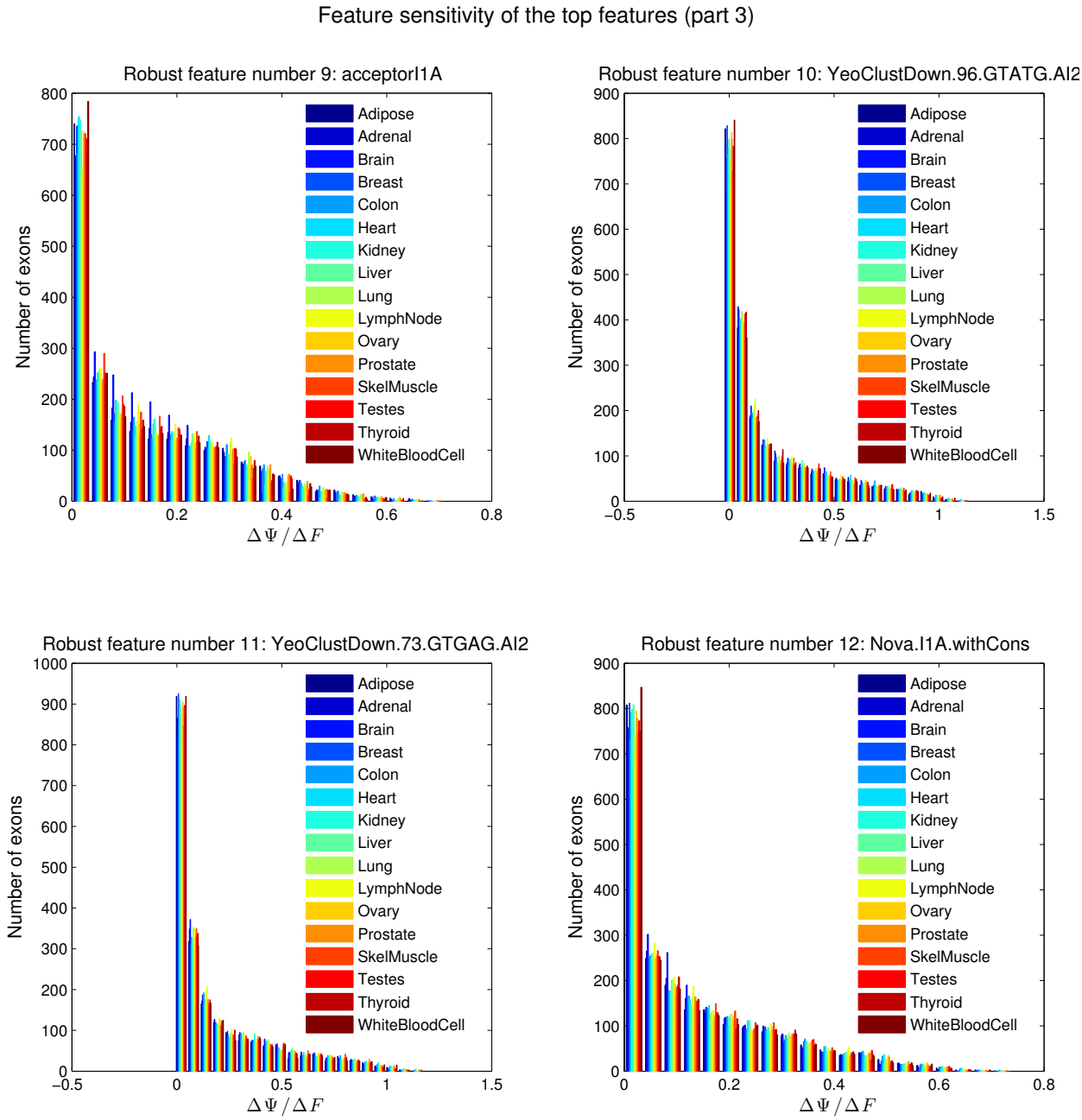

Fig. S13
Part 3 of 4.

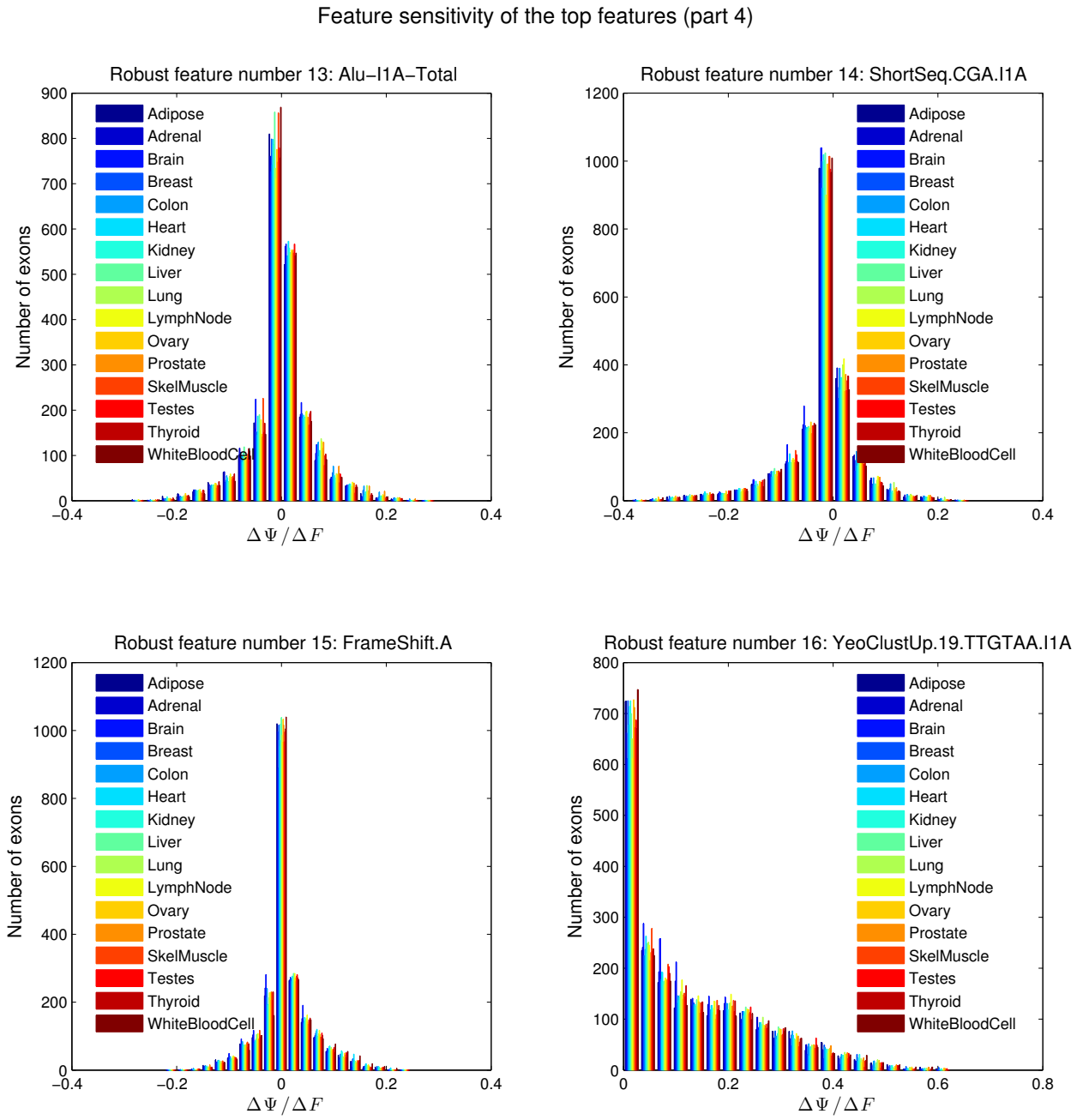

Fig. S13
Part 4 of 4.

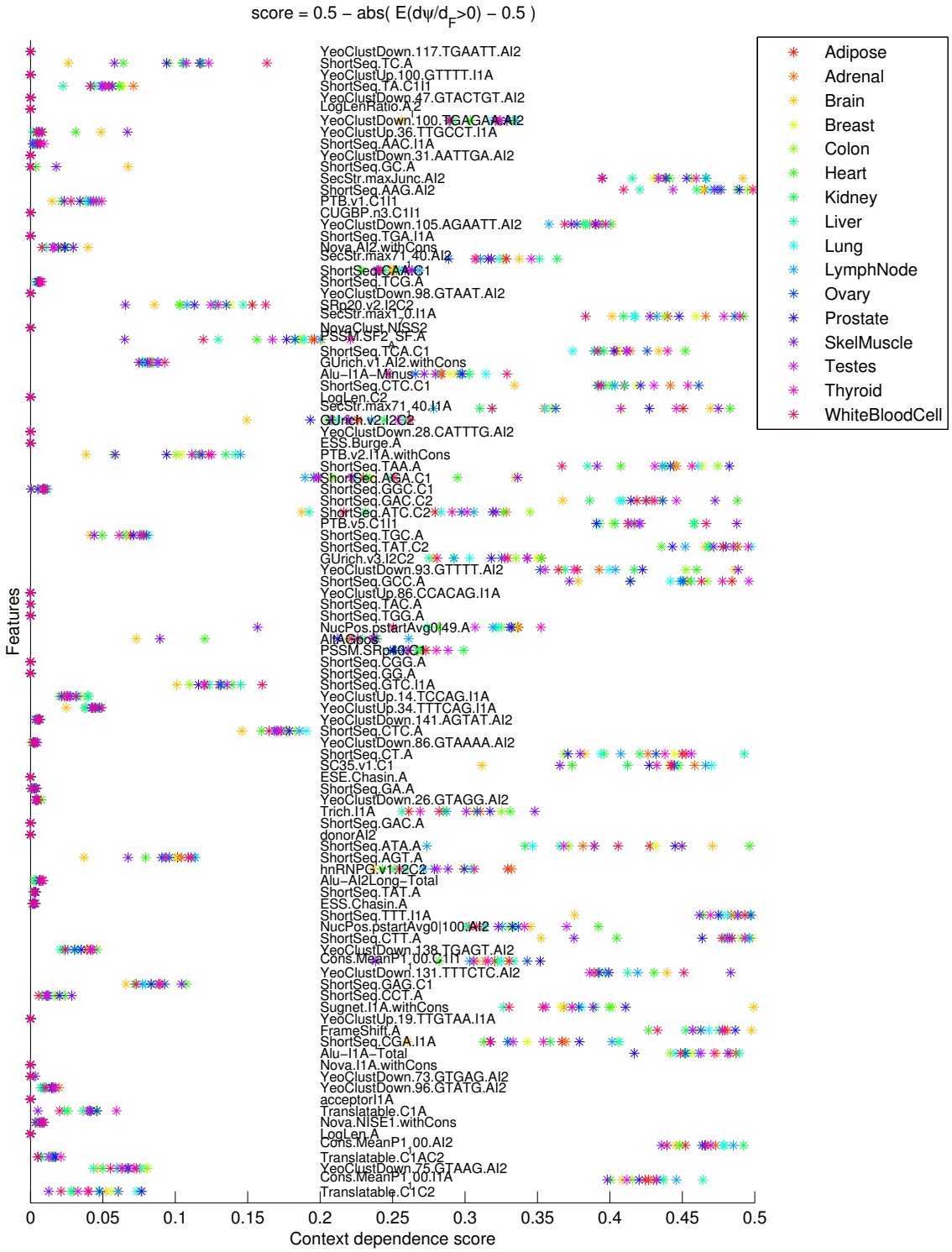

Fig. S14

For each of the top 100 features, the degree to which the effect of the feature switches sign across different exons was evaluated (horizontal axis), where a score of 0.5 means the features effect had the same sign for 50% of exons. 40 features exhibit strong sign switching.

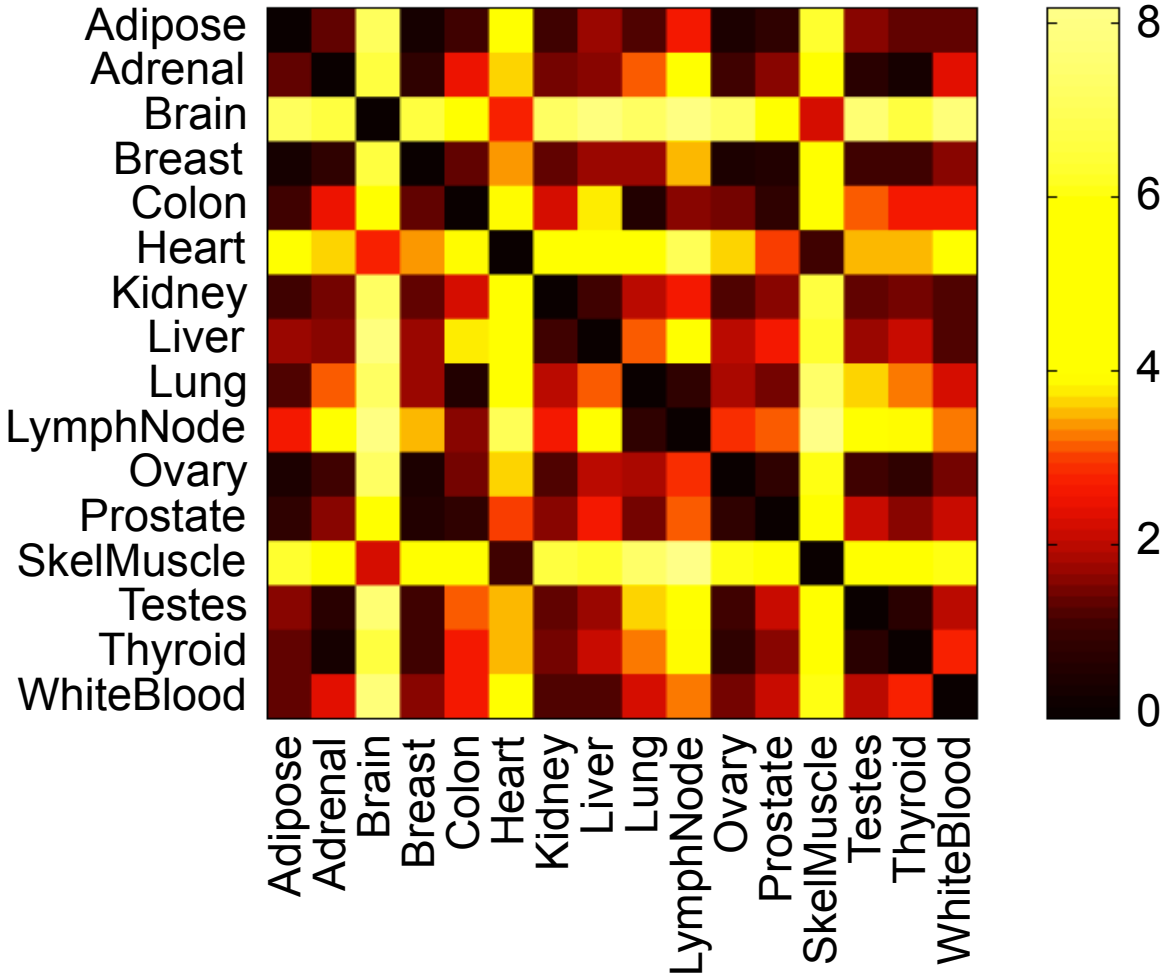

Fig. S15

For each pair of tissues, the fraction of variability in the sensitivities of all features that is different in the two tissues is plotted as $(1-r^2)x100$

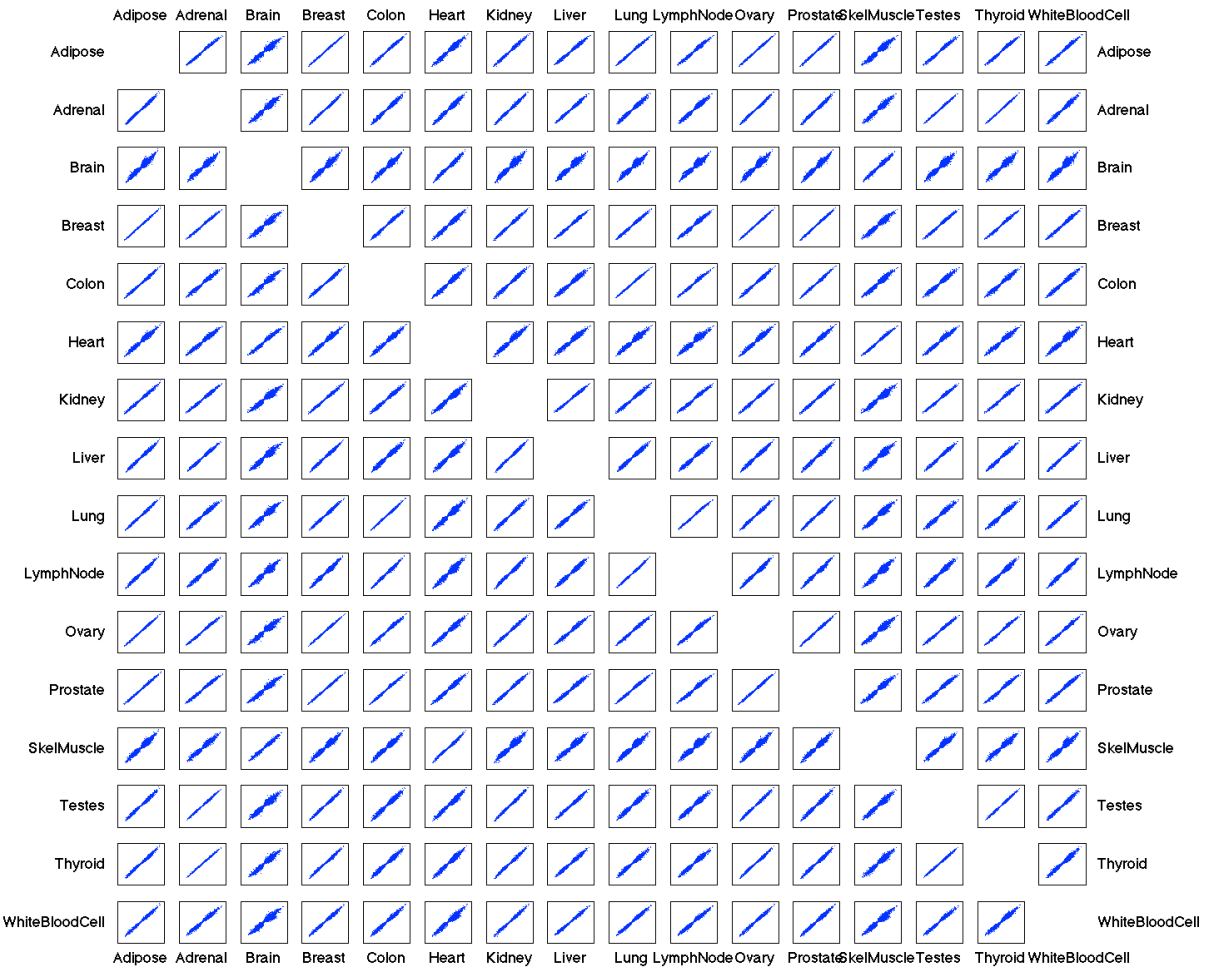

Fig. S16

Scatter plots of feature sensitivity ($\Delta\Psi/\Delta F$) across all held-out exons and across the top 20 features, for every pair of tissues. Each plot is a tissue pair and each point is an exon-feature combination. High correlations between most tissue pairs were observed, indicating that the sensitivity of Ψ to small changes of top features is similar for these tissue pairs. However, significant differences exist for some pairs. See Fig. S15 for a heat-map that summarizes these plots.

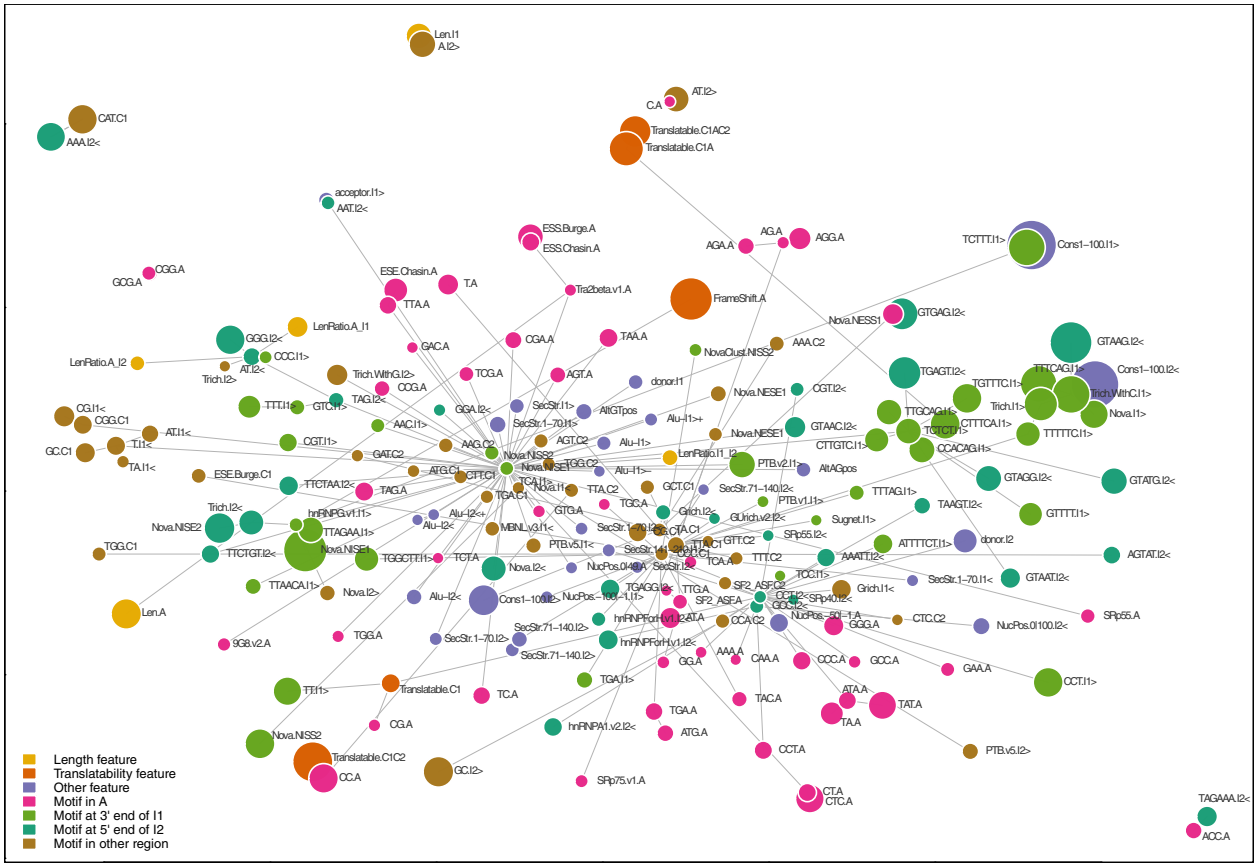

Fig. S17

Visualization of co-regulatory effects of features. The t-SNE visualization computes an embedding for the features such that their distance in the figure is small when their effect on the predicted Ψ is similar, and the size of a feature node is proportional to its strength. Because dimensionality reduction introduces errors, an edge is further drawn from a feature to its nearest neighbour when they are not so close by.

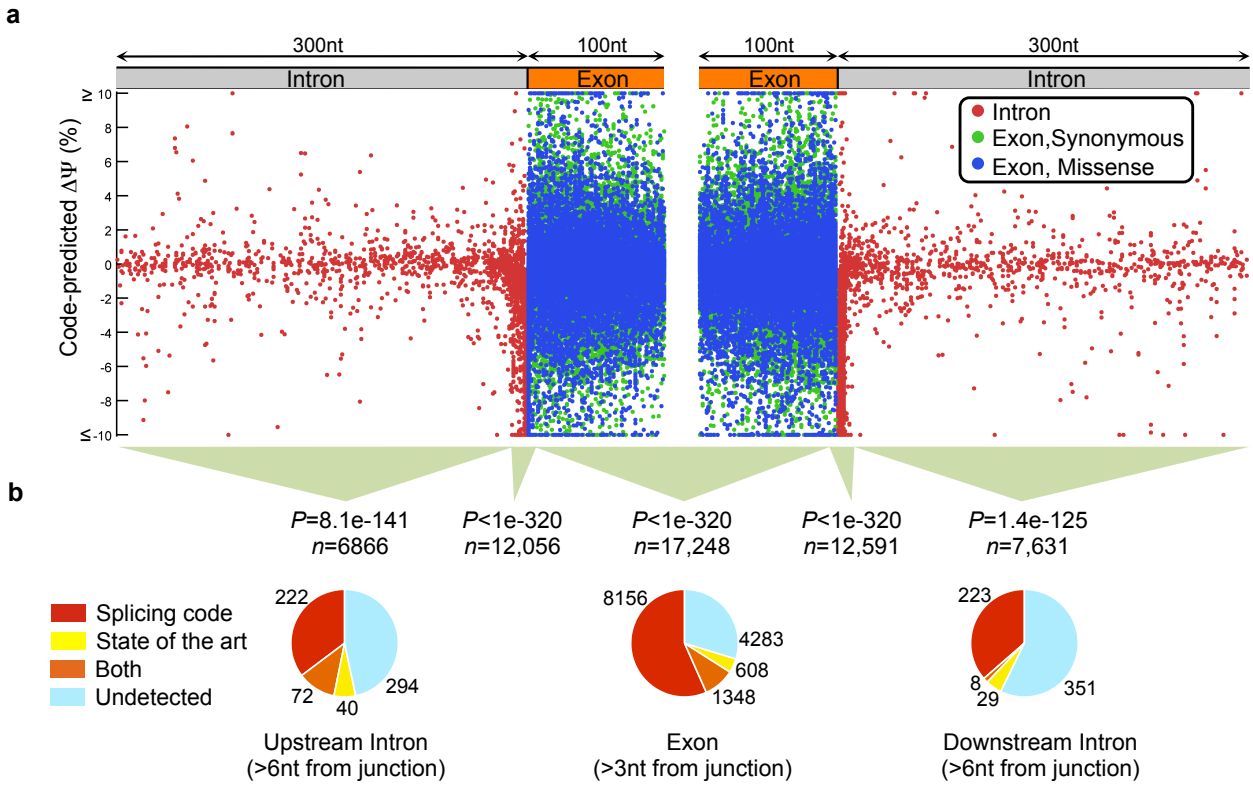

Fig. S18

Comparing the splicing code to other methods for detecting disease variants. (a) Locations and predicted ΔΨ of 100,022 disease annotated intronic SNVs and synonymous or missense exonic SNVs. (b) In every sequence region, the scores of disease SNVs tend to be larger than those of SNPs (Ansari-Bradley test for equal dispersion, n includes both types). Pie charts compare the fraction of disease SNVs detected by our splicing code to the state of the art (Spliceman and Skippy), using thresholds that detect 10% of SNPs.

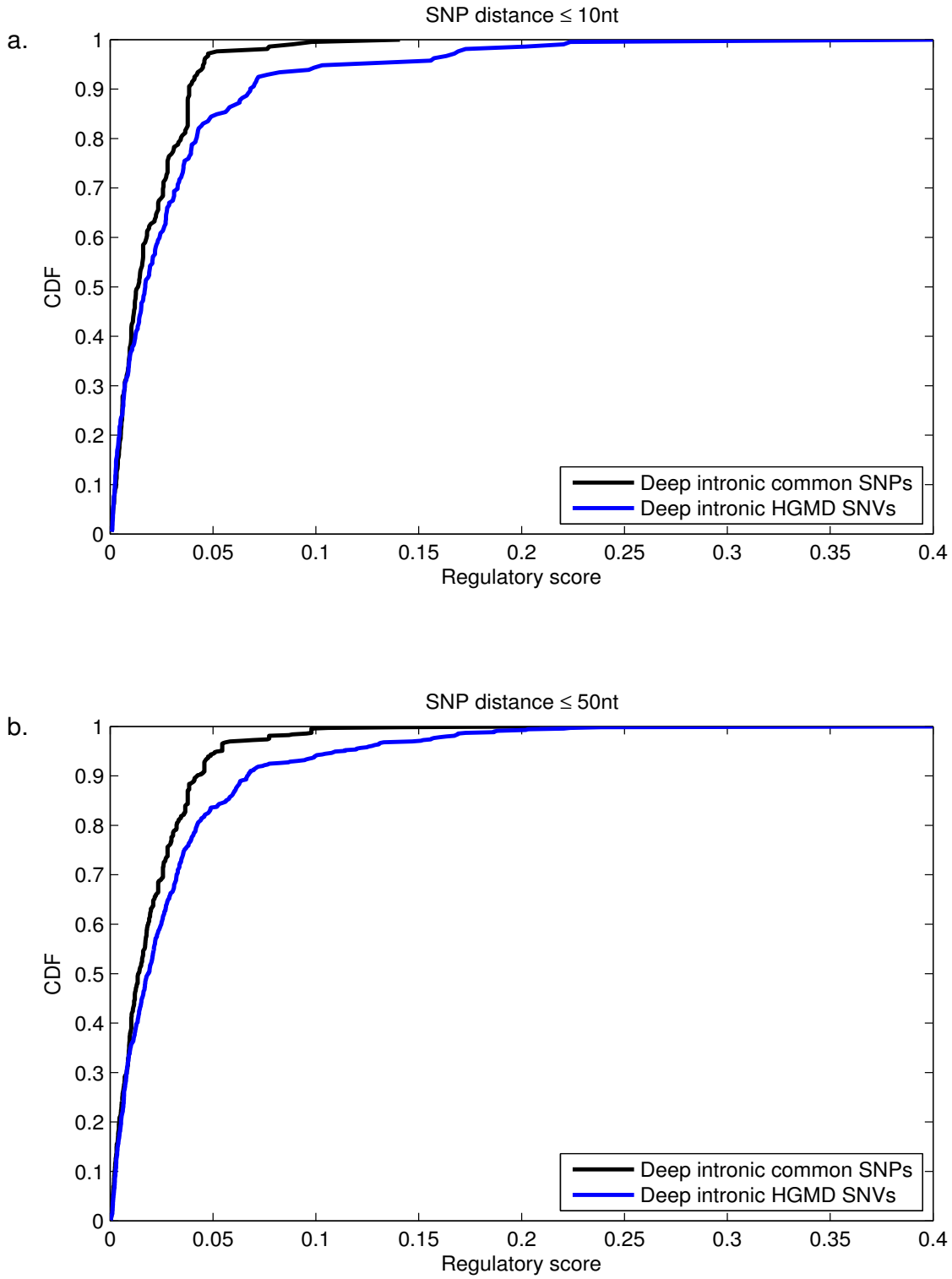

Fig. S19

CDF plots of regulatory scores derived from pairs of common and HGMD SNVs that are close to each other.

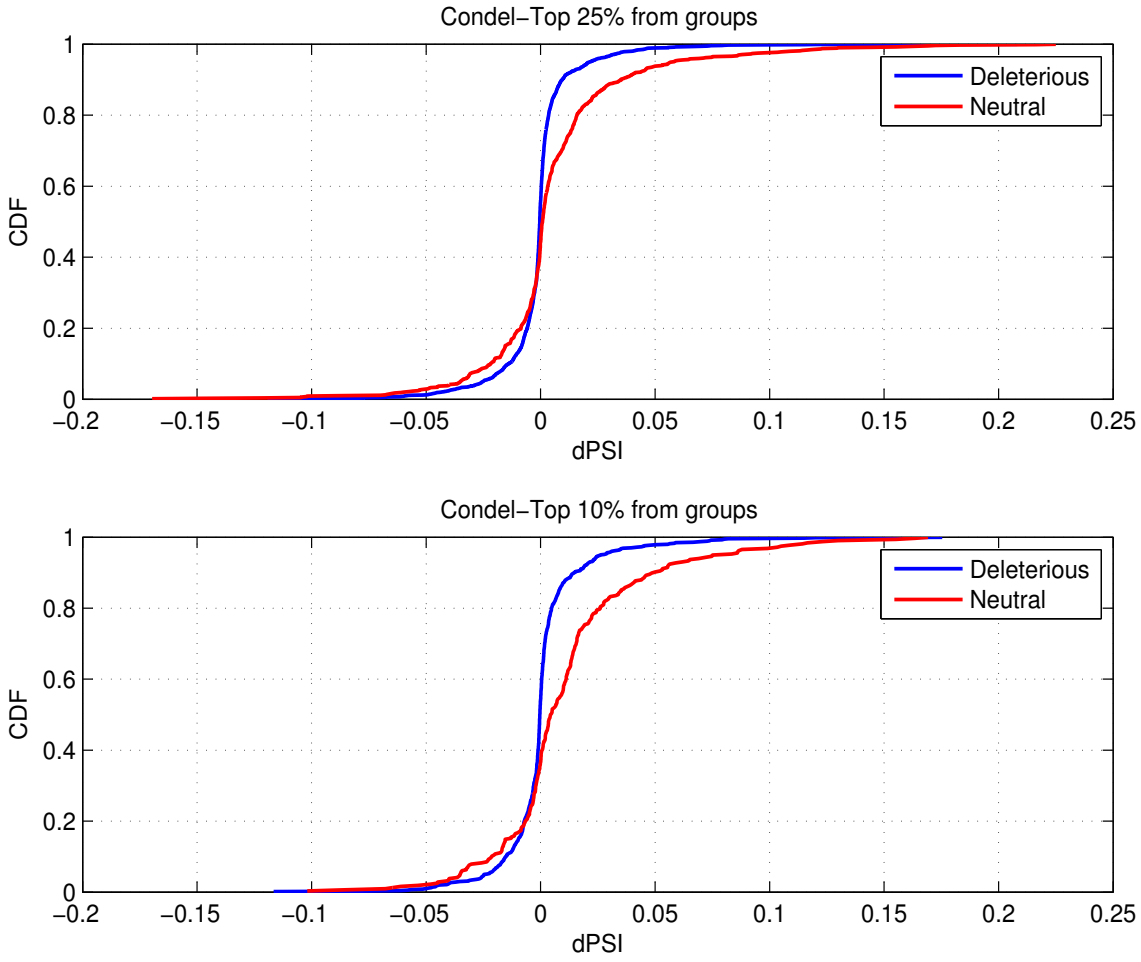

Fig. S20

CDF plots of ΔΨ's for two sets with the 25% and 10% most reliable SNVs predicted by Condel.

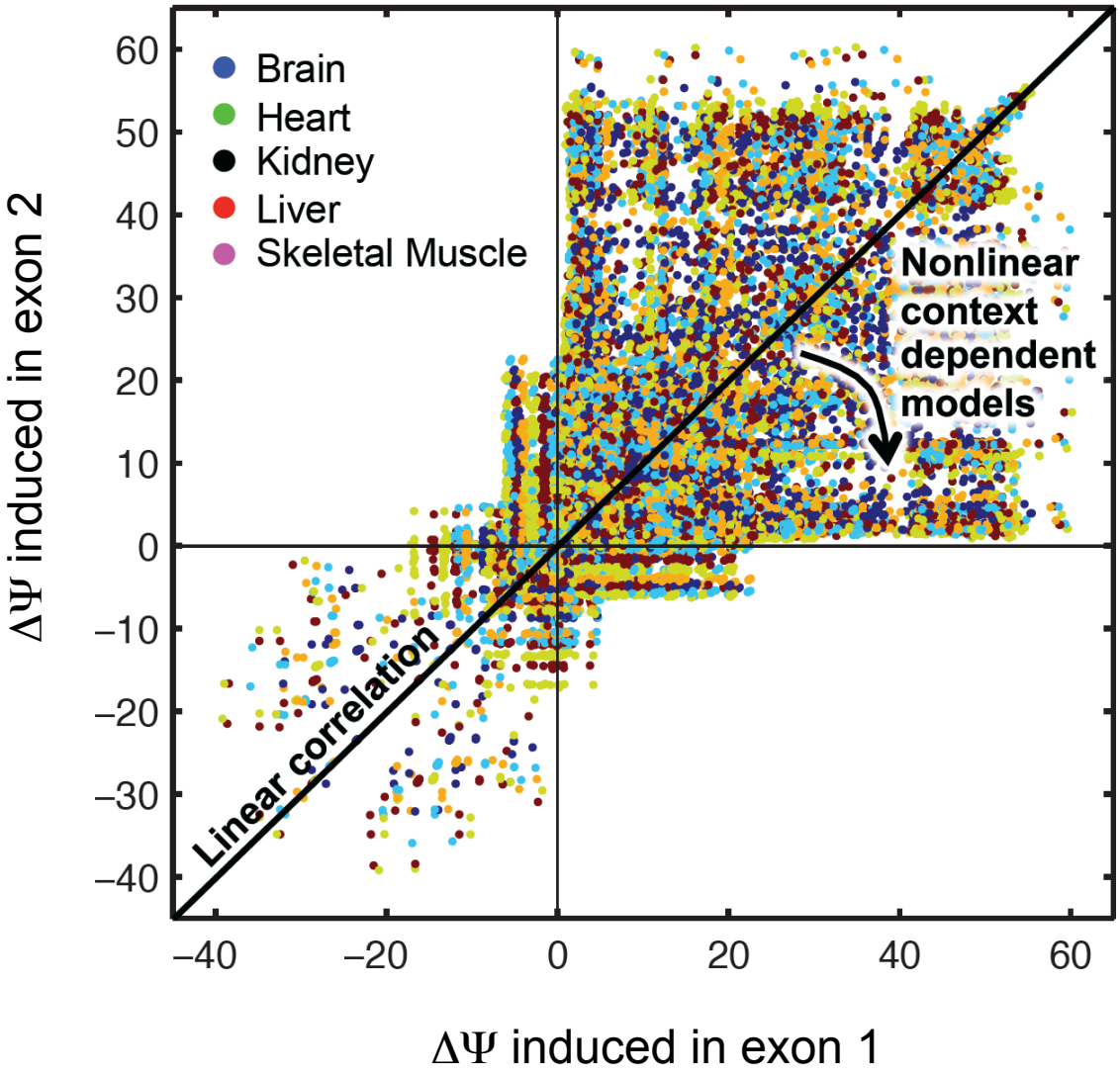

Fig. S21

Context dependence can be further examined by plotting the change in Ψ induced by point mutations in pairs of exons that have different wild type feature vectors but where the mutation-induced changes in the feature vectors are identical. For correlation analysis and linear models, all points would be on the diagonal line. Off-diagonal points are caused by context-dependent effects of SNVs.

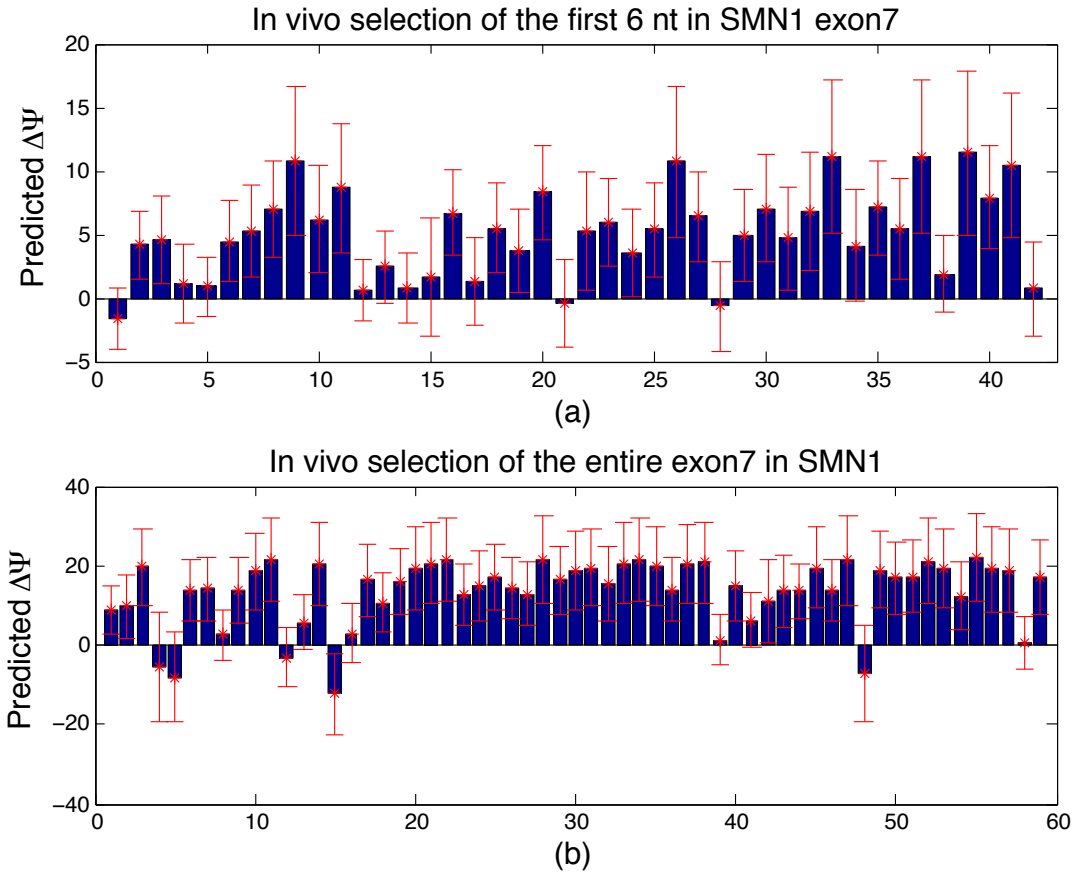

Fig. S22

Predicted ΔΨ for *SMN1* exon 7 *in vivo* selection. (a) *in vivo* selection of the first 6nt of exon 7 as performed in (*73*), with the same order as listed in Fig. 2(B) of the original paper. (b) *in vivo* selection of the entire exon 7 as performed in (*74*), with the same order as listed in Fig. 5 of the original paper. Blue bars indicate the magnitude of predicted ΔΨ's while red error bars show their sample standard deviation. For all mutations, the correct sign of ΔΨ should be positive.

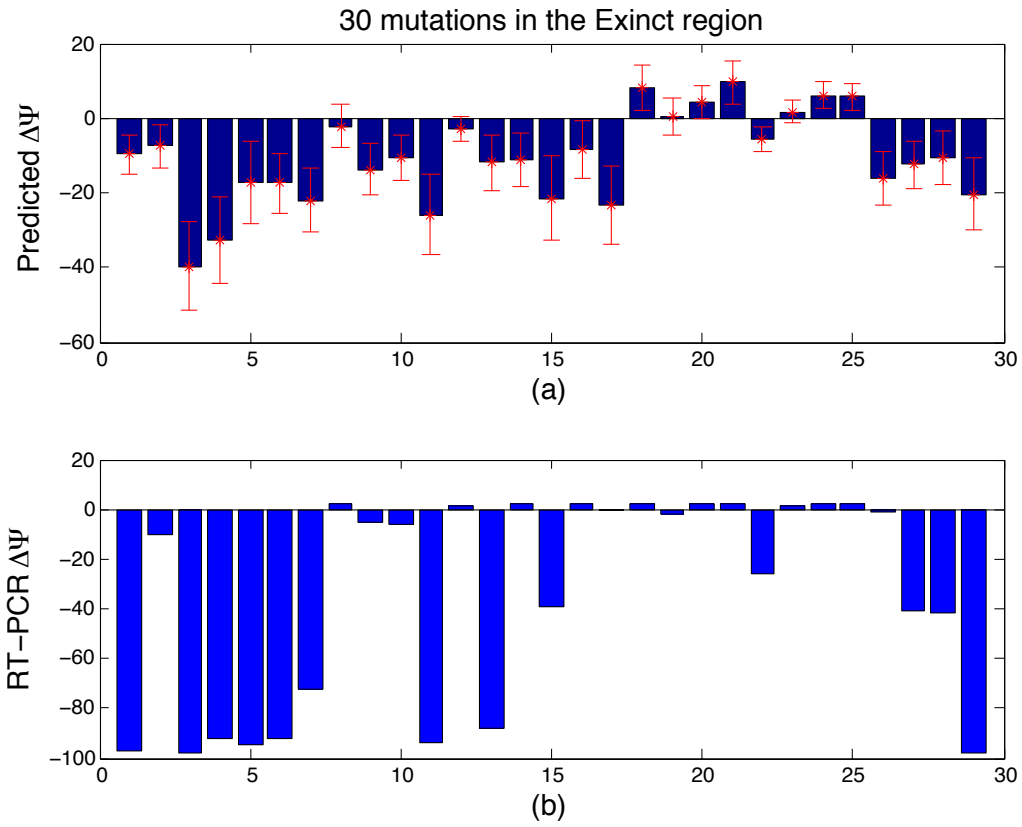

Fig. S23

Mutations in the 5' Exinct region of exon 7 as listed in Table 1 of (*75*). (a) Code-predicted ΔΨ (blue bars) and sample standard deviations (red error bars). (b) RT-PCR measured ΔΨ in the original study.

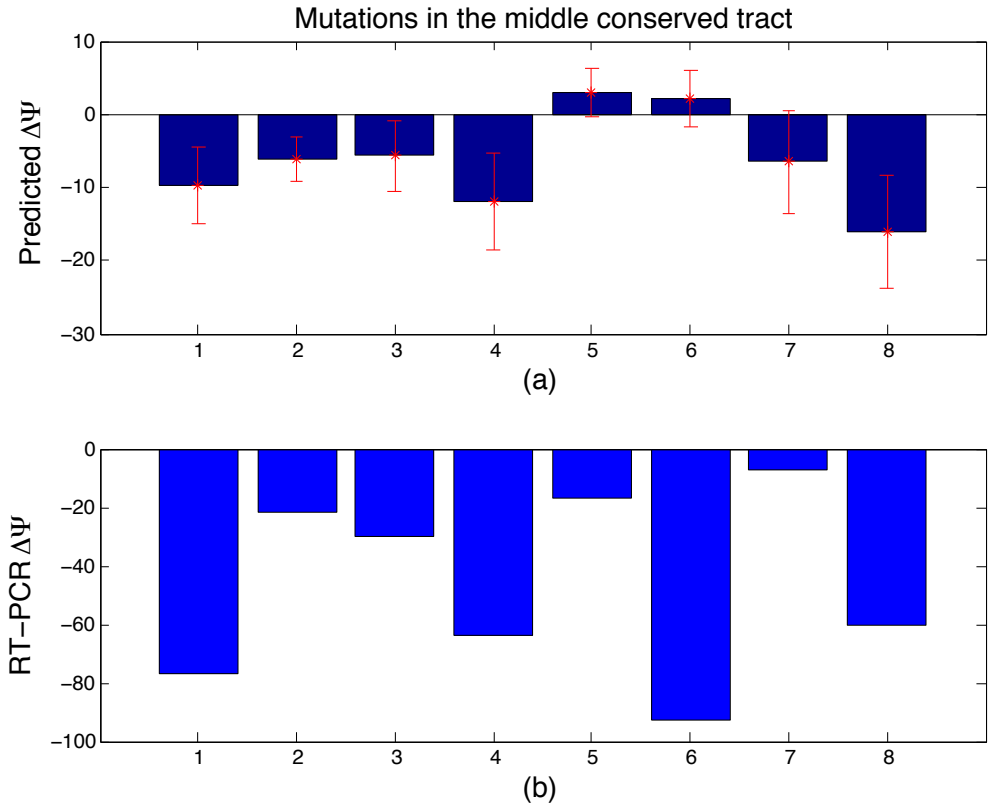

Fig. S24

Mutations in the middle conserved tract of exon 7 as shown in Fig. 8 of (*74*). (a) Code-predicted ΔΨ (blue bars) and sample standard deviations (red error bars). (b) RT-PCR measured ΔΨ in the original study.

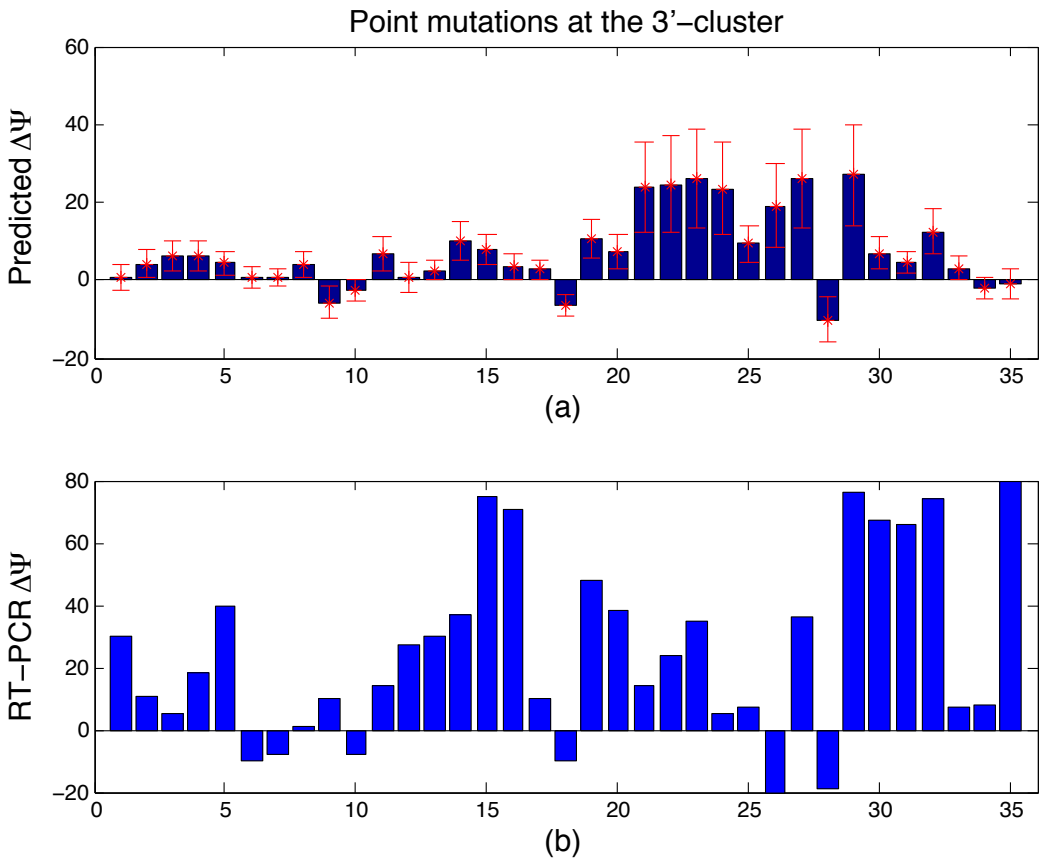

Fig. S25

Mutations in the 3'-cluster of exon 7 as listed in Table 1 of (*76*). (a) Code-predicted ΔΨ (blue bars) and sample standard deviations (red error bars). (b) RT-PCR measured ΔΨ in the original study.

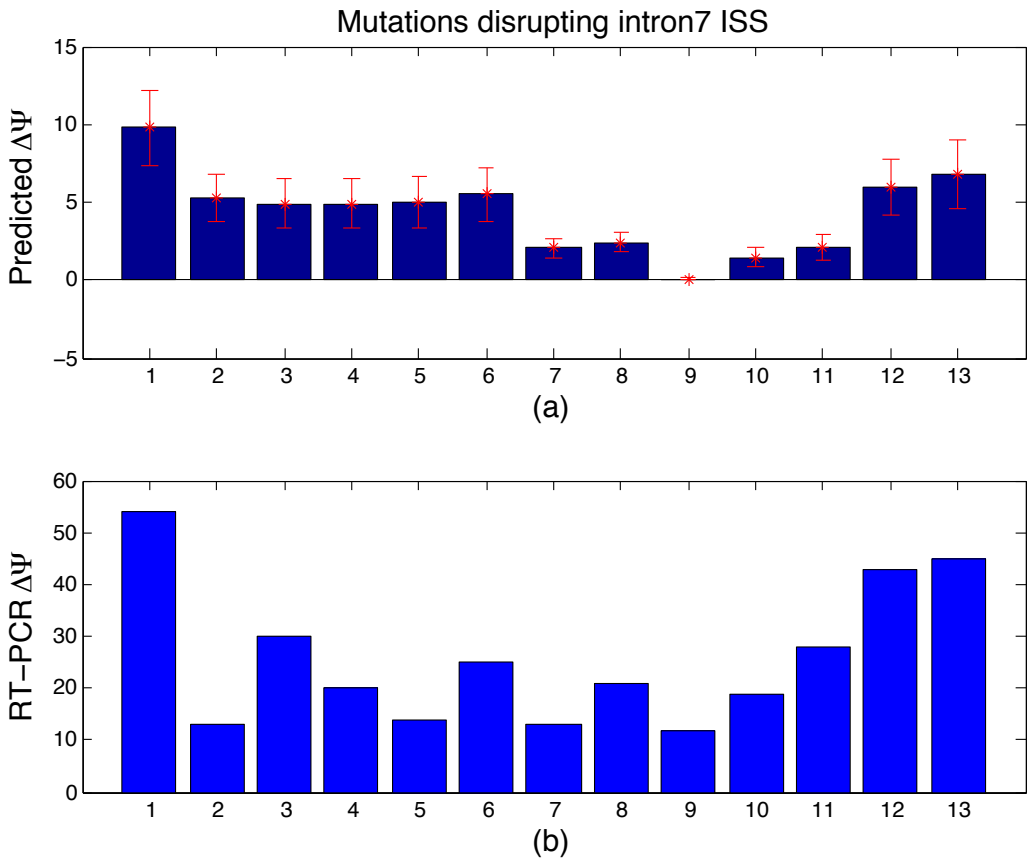

Fig. S26

Mutations that disrupt the intron 7 ISS at +10 to +24 as shown in Fig. 3 of (*24*). (a) Code-predicted ΔΨ (blue bars) and sample standard deviations (red error bars). (b) RT-PCR measured ΔΨ in the original study.

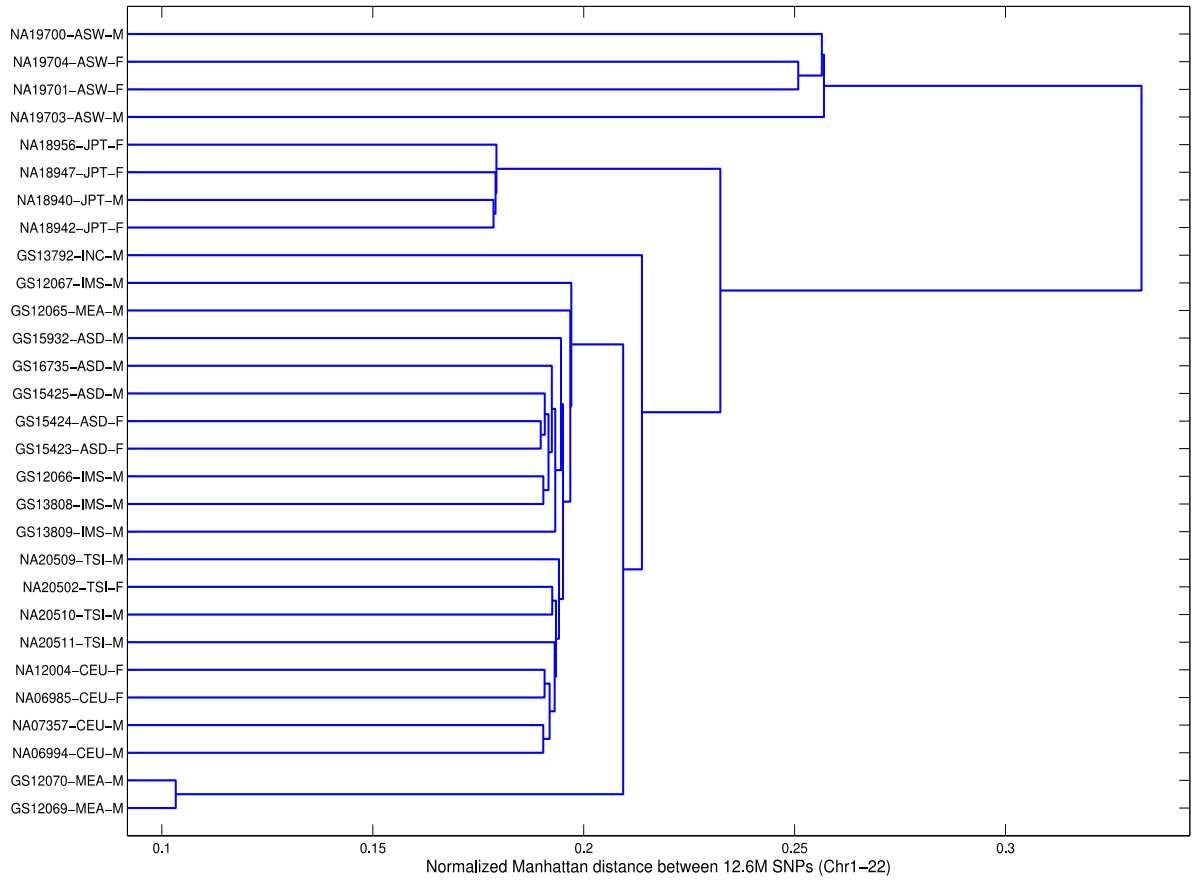

Fig. S27

A dendrogram formed from genomic distances between subjects by running the UPGMA algorithm.

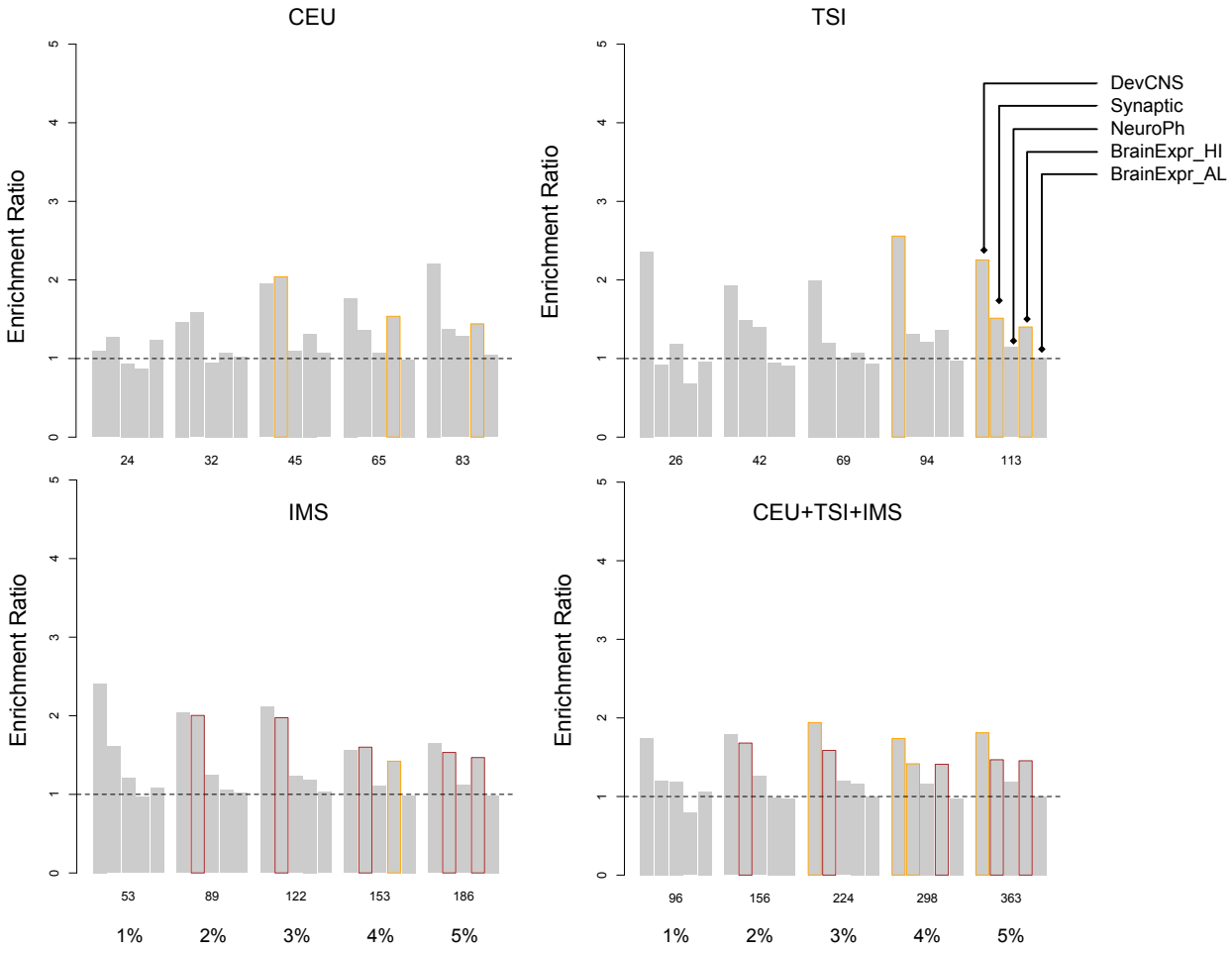

Fig. S28

Enrichment ratios for enrichment analysis in genes with absolute value of $\Delta\Psi$ larger than first to fifth percentiles of common SNPs in ASD vs. control groups. The five bars from left to right correspond to: DevCNS, Synaptic, NeuroPh, BrainExpr_HI and BrainExpr_AL. A dark red border indicates p-value < 0.05 and a yellow border indicates 0.05 < p-value < 0.1. The numbers below bars denote the number of control genes used in the enrichment analysis.

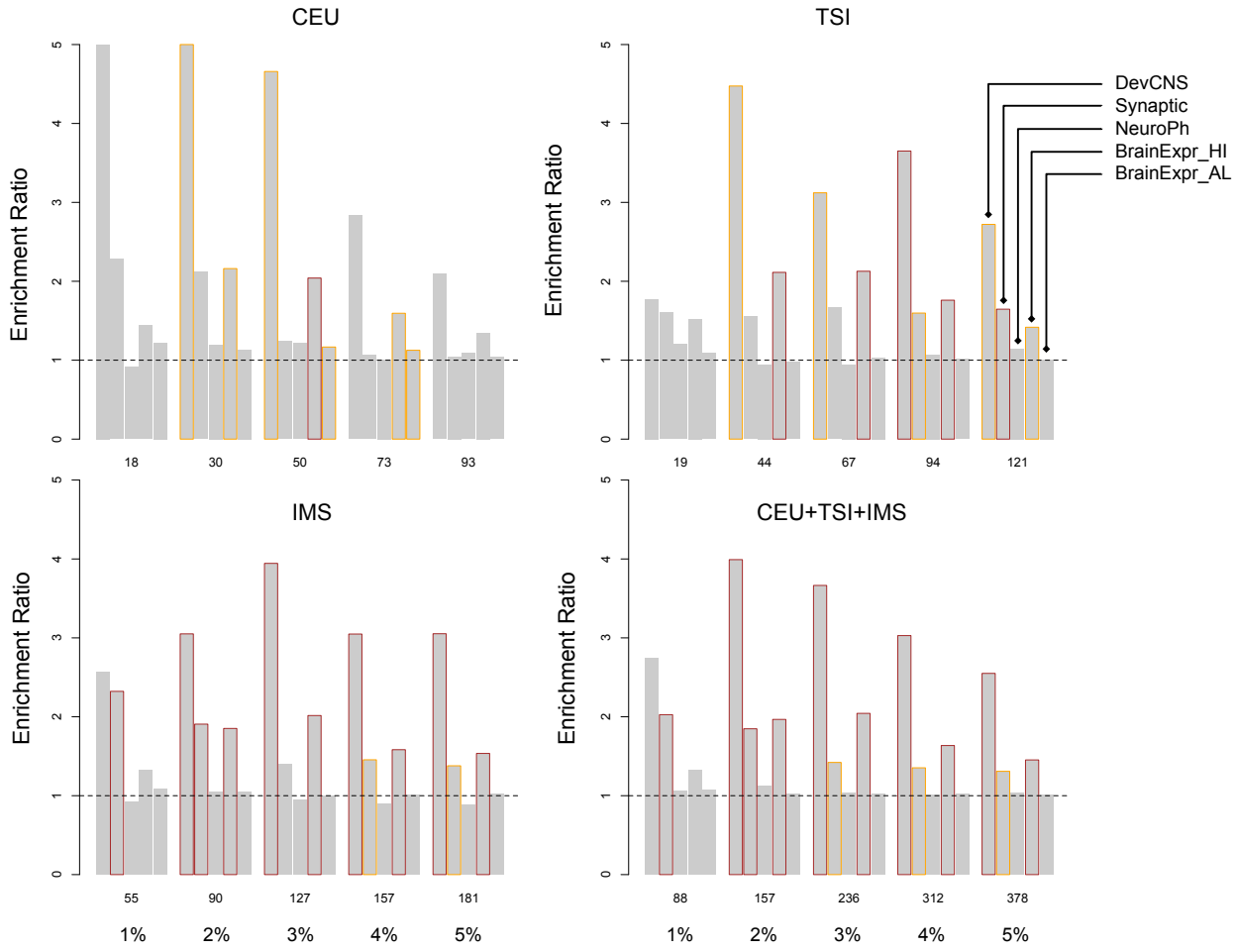

Fig. S29

Enrichment ratios for enrichment analysis in genes with $\Delta\Psi$ less than first to fifth percentiles of common SNPs in ASD vs. control groups. The five bars from left to right correspond to: DevCNS, Synaptic, NeuroPh, BrainExpr_HI and BrainExpr_AL. A dark red border indicates p-value < 0.05 and a yellow border indicates 0.05 < p-value < 0.1. The numbers below bars denote the number of control genes used in the enrichment analysis.

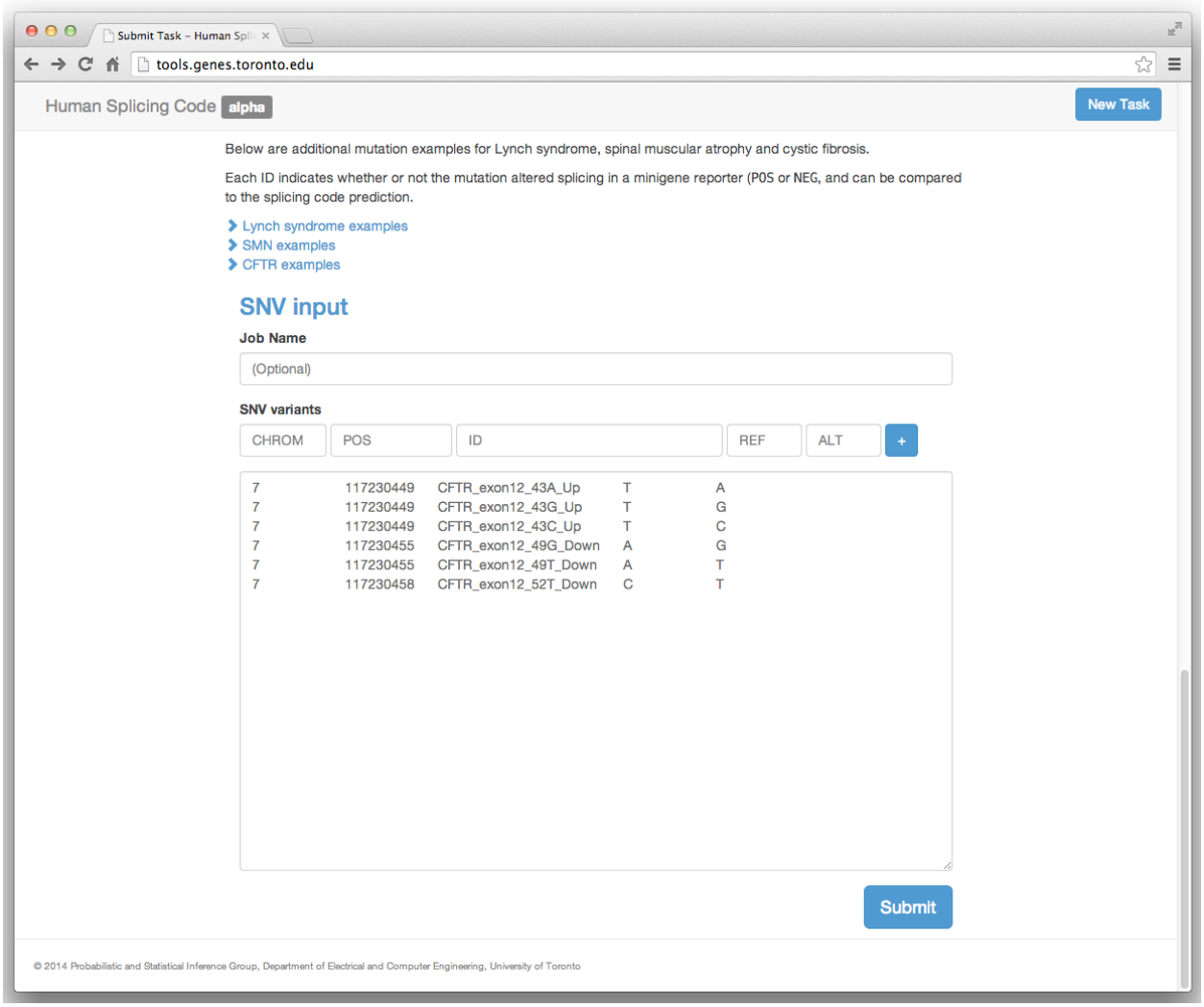

Fig. S30

Screenshot for part of the job submission page of the mutation analysis web tool.

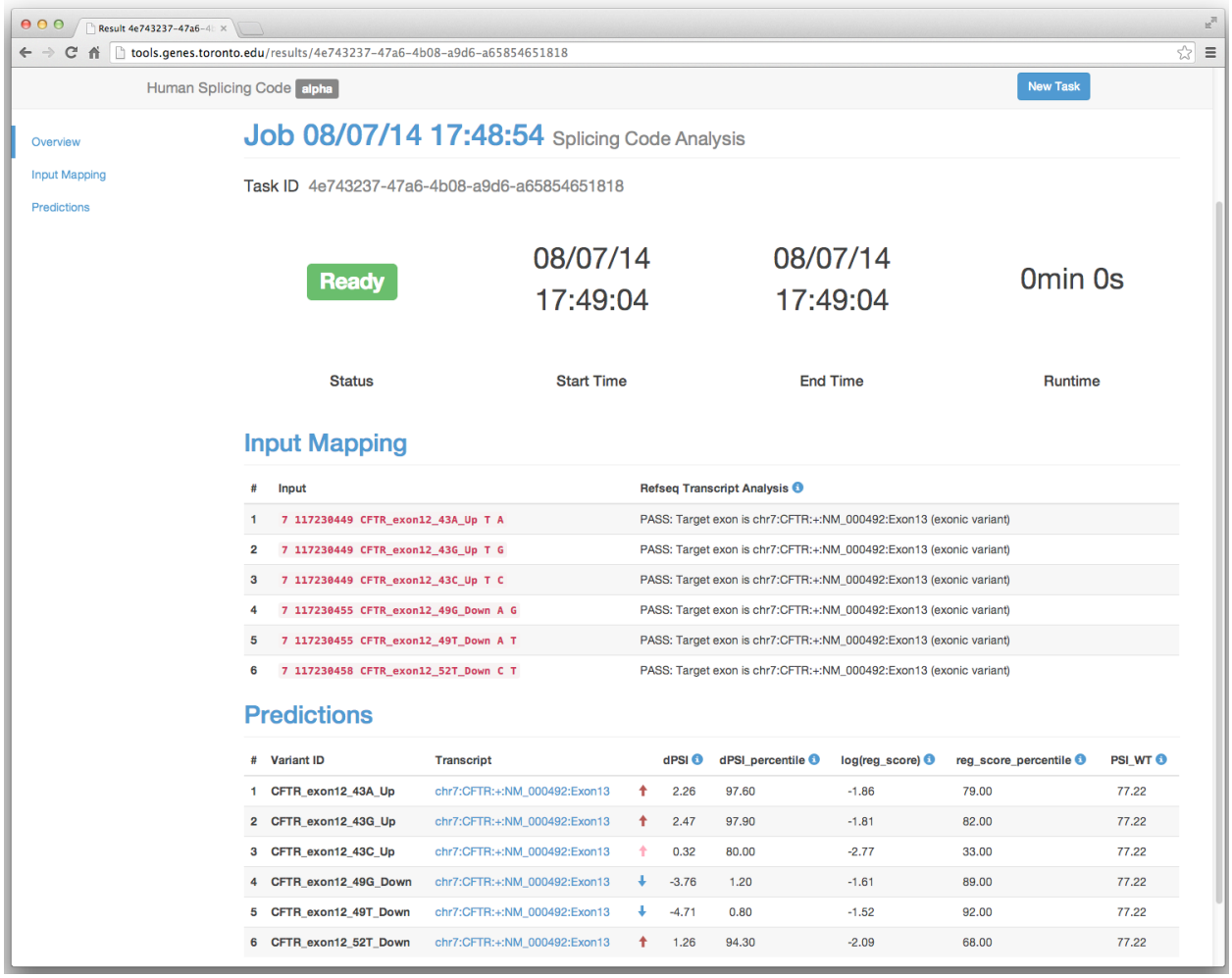

Fig. S31

Screenshot for the global result page of the mutation analysis web tool.

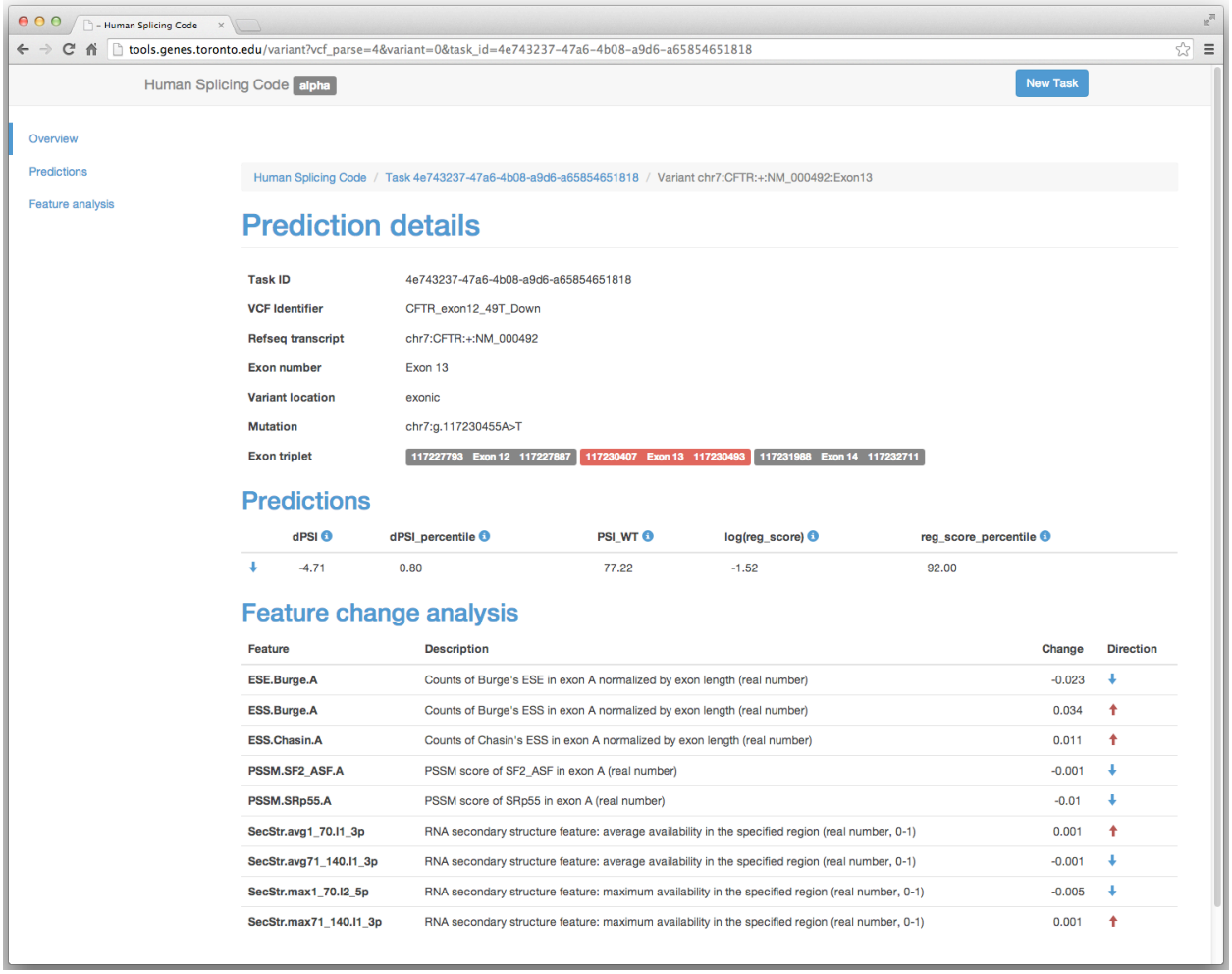

Fig. S32

Screenshot for the detailed result page of the mutation analysis web tool.

REFERENCES

1. K. Lindblad-Toh *et al.*, A high-resolution map of human evolutionary constraint using 29 mammals, *Nature* **478**, 476–82 (2011).
2. B. E. Bernstein *et al.*, An integrated encyclopedia of DNA elements in the human genome, *Nature* **489**, 57–74 (2012).
3. Y. Barash *et al.*, Deciphering the splicing code, *Nature* **465**, 53–9 (2010).
4. C. Zhang *et al.*, Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls, *Science* **329**, 439–43 (2010).
5. N. L. Barbosa-Morais *et al.*, The Evolutionary Landscape of Alternative Splicing in Vertebrate Species, *Science* **338**, 1587–1593 (2012).
6. E. Segal, J. Widom, From DNA sequence to transcriptional behaviour: a quantitative approach, *Nat Rev Genet* **10**, 443–56 (2009).
7. F. Gnad, A. Baucom, K. Mukhyala, G. Manning, Z. Zhang, Assessment of computational methods for predicting the effects of missense mutations in human cancers., *BMC Genomics* **14 Suppl 3**, S7 (2013).
8. M. Kircher *et al.*, A general framework for estimating the relative pathogenicity of human genetic variants, *Nat Genet* **46**, 310–315 (2014).
9. M. Mort *et al.*, MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing, *Genome Biol* **15**, R19 (2014).
10. T. Sterne-Weiler, J. R. Sanford, Exon identity crisis: disease-causing mutations that disrupt the splicing code, *Genome Biol* **15**, 201 (2014).
11. H. Y. Xiong, Y. Barash, B. J. Frey, Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context, *Bioinformatics* **27**, 2554–62 (2011).
12. Y. Barash *et al.*, AVISPA: a web tool for the prediction and analysis of alternative splicing, *Genome Biol* **14**, R114 (2013).
13. D. Ray *et al.*, A compendium of RNA-binding motifs for decoding gene regulation, *Nature* **499**, 172–7 (2013).
14. H. Han *et al.*, MBNL proteins repress ES-cell-specific alternative splicing and reprogramming, *Nature* **498**, 241–5 (2013).
15. T. Lappalainen *et al.*, Transcriptome and genome sequencing uncovers functional variation in humans, *Nature* **501**, 506–11 (2013).
16. S. T. Sherry *et al.*, dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res* **29**, 308–11 (2001).
17. P. D. Stenson *et al.*, The Human Gene Mutation Database: 2008 update, *Genome Med* **1**, 13 (2009).
18. F. Supek, B. Miñana, J. Valcárcel, T. Gabaldón, B. Lehner, Synonymous mutations frequently act as driver mutations in human cancers, *Cell* **156**, 1324–35 (2014).
19. M. Kimura, *The Neutral Theory of Molecular Evolution* (1983; http://www.worldcat.org/isbn/0521317932).

20. A. González-Pérez, N. López-Bigas, Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel, *Am J Hum Genet* **88**, 440–9 (2011).
21. L. A. Hindorff *et al.*, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *PNAS* **106**, 9362–7 (2009).
22. R. N. Singh, Evolving concepts on human SMN pre-mRNA splicing, *RNA Biol* **4**, 7–10 (2007).
23. T. Kashima, N. Rao, J. L. Manley, An intronic element contributes to splicing repression in spinal muscular atrophy, *PNAS* **104**, 3426–31 (2007).
24. Y. Hua, T. A. Vickers, H. L. Okunola, C. F. Bennett, A. R. Krainer, Antisense Masking of an hnRNP A1/A2 Intronic Splicing Silencer Corrects SMN2 Splicing in Transgenic Mice, *Am J Hum Genet* **82**, 834–848 (2008).
25. R. A. Barnetson *et al.*, Classification of ambiguous mutations in DNA mismatch repair genes identified in a population-based study of colorectal cancer, *Hum Mutat* **29**, 367–74 (2008).
26. P. Peltomäki, H. Vasen, Mutations associated with HNPCC predisposition -- Update of ICG-HNPCC/INSiGHT mutation database, *Dis Markers* **20**, 269–276 (2004).
27. S. Arnold *et al.*, Classifying MLH1 and MSH2 variants using bioinformatic prediction, splicing assays, segregation, and tumor characteristics, *Hum Mutat* **30**, 757–70 (2009).
28. B. Betz *et al.*, Comparative in silico analyses and experimental validation of novel splice site and missense mutations in the genes MLH1 and MSH2, *J Cancer Res Clin Oncol* **136**, 123–34 (2010).
29. M. Nyström-Lahti *et al.*, Missense and nonsense mutations in codon 659 of MLH1 cause aberrant splicing of messenger RNA in HNPCC kindreds, *Genes Chromosom Cancer* **26**, 372–375 (1999).
30. P. Lastella, N. C. Surdo, N. Resta, G. Guanti, A. Stella, In silico and in vivo splicing analysis of MLH1 and MSH2 missense mutations shows exon- and tissue-specific effects, *BMC Genomics* **7**, 243 (2006).
31. J. Kosinski, I. Hinrichsen, J. M. Bujnicki, P. Friedhoff, G. Plotz, Identification of Lynch syndrome mutations in the MLH1-PMS2 interface that disturb dimerization and mismatch repair, *Hum Mutat* **31**, 975–82 (2010).
32. P. J. Smith *et al.*, An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers, *Hum Mol Genet* **15**, 2490–508 (2006).
33. J. D. Buxbaum *et al.*, The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders, *Neuron* **76**, 1052–6 (2012).
34. C. Betancur, Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting, *Brain Res* **1380**, 42–77 (2011).
35. B. Devlin, S. W. Scherer, Genetic architecture in autism spectrum disorder, *Curr Opin Genet Dev* **22**, 229–37 (2012).
36. I. Iossifov *et al.*, De novo gene disruptions in children on the autistic spectrum, *Neuron* **74**, 285–99 (2012).

37. Y. Jiang *et al.*, Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing, *Am J Hum Genet* **93**, 249–63 (2013).
38. R. Anney *et al.*, A genome-wide scan for common alleles affecting risk for autism, *Hum Mol Genet* **19**, 4072–82 (2010).
39. T. C. Südhof, Neuroligins and neurexins link synaptic function to cognitive disease, *Nature* **455**, 903–11 (2008).
40. I. Voineagu *et al.*, Transcriptomic analysis of autistic brain reveals convergent molecular pathology, *Nature* **474**, 380–4 (2011).
41. R. F. Wintle *et al.*, A genotype resource for postmortem brain samples from the Autism Tissue Program, *Autism Res* **4**, 89–97 (2011).
42. D. Pinto *et al.*, Functional impact of global rare copy number variation in autism spectrum disorders, *Nature* **466**, 368–72 (2010).
43. M. Uddin *et al.*, Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder, *Nat Genet* **46**, 742–7 (2014).
44. E. Skafidas *et al.*, Predicting the diagnosis of autism spectrum disorder using gene pathway analysis, *Mol Psychiatry* **19**, 504–10 (2014).
45. E. B. Robinson *et al.*, Response to "Predicting the diagnosis of autism spectrum disorder using gene pathway analysis," *Mol Psychiatry* **19**, 859–861 (2014).
46. E. Khurana *et al.*, Integrative annotation of variants from 1092 humans: application to cancer genomics, *Science* **342**, 1235587 (2013).
47. K. H. Lim, L. Ferraris, M. E. Filloux, B. J. Raphael, W. G. Fairbrother, Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes, *PNAS* **108**, 11093–8 (2011).
48. A. Woolfe, J. C. Mullikin, L. Elnitski, Genomic features defining exonic variants that modulate splicing, *Genome Biol* **11**, R20 (2010).
49. B. E. Stranger *et al.*, Relative impact of nucleotide and copy number variation on gene expression phenotypes, *Science* **315**, 848–53 (2007).
50. J. A. Tennessen *et al.*, Evolution and functional impact of rare coding variation from deep sequencing of human exomes, *Science* **337**, 64–9 (2012).
51. A. Battle *et al.*, Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals, *Genome Res* **24**, 14–24 (2014).
52. U. Braunschweig, S. Gueroussov, A. M. Plocik, B. R. Graveley, B. J. Blencowe, Dynamic integration of splicing within gene regulatory pathways, *Cell* **152**, 1252–69 (2013).
53. D. Brawand *et al.*, The evolution of gene expression levels in mammalian organs, *Nature* **478**, 343–8 (2011).
54. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, *Nucleic Acids Res* **38**, e164 (2010).
55. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics* **25**, 1105–11 (2009).
56. A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, L. Pachter, Improving RNA-Seq expression estimates by correcting for fragment bias, *Genome Biol* **12**, R22 (2011).
57. Y. Katz, E. T. Wang, E. M. Airoldi, C. B. Burge, Analysis and design of RNA sequencing experiments for identifying isoform regulation, *Nat Methods* **7**, 1009–15 (2010).

58. B. Kakaradov, H. Y. Xiong, L. J. Lee, N. Jojic, B. J. Frey, Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data, *BMC Bioinformatics* **13 Suppl 6**, S11 (2012).
59. Z. Zhang, S. Schwartz, L. Wagner, W. Miller, A greedy algorithm for aligning DNA sequences, *J Comput Biol* **7**, 203–14 (2000).
60. Y. Kapustin, A. Souvorov, T. Tatusova, D. Lipman, Splign: algorithms for computing spliced alignments with identification of paralogs, *Biol Direct* **3**, 20 (2008).
61. L. Xi *et al.*, Predicting nucleosome positioning using a duration Hidden Markov Model, *BMC Bioinformatics* **11**, 346 (2010).
62. R. Shankar, B. Kataria, M. Mukerji, Finding Alu in primate genomes with AF-1, *Bioinformation* **3**, 287–288 (2009).
63. B. Efron, T. Hastie, I. Johnson, R. Tibshirani, Least angle regression, *Ann Stat* **32**, 407–499 (2004).
64. L. Van Der Maaten, G. Hinton, Visualizing data using t-SNE, *J Mach Learn Res* **9**, 2579–2605 (2008).
65. L. R. Meyer *et al.*, The UCSC Genome Browser database: extensions and updates 2013, *Nucleic Acids Res* **41**, D64–9 (2013).
66. F. Desmet, D. Hamroun, Bioinformatics identification of splice site signals and prediction of mutation effects, *Res Adv Nucleic Acid Res Kerala Glob Res Netw* , 1–14 (2010).
67. K. H. Lim, W. G. Fairbrother, Spliceman--a computational web server that predicts sequence variations in pre-mRNA splicing, *Bioinformatics* **28**, 1031–2 (2012).
68. P. Kumar, S. Henikoff, P. C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, *Nat Protoc* **4**, 1073–81 (2009).
69. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations, *Nat Methods* **7**, 248–9 (2010).
70. Y. Hua, T. a Vickers, B. F. Baker, C. F. Bennett, A. R. Krainer, Enhancement of SMN2 exon 7 inclusion by antisense oligonucleotides targeting the exon, *PLoS Biol* **5**, e73 (2007).
71. Y. Hua *et al.*, Antisense correction of SMN2 splicing in the CNS rescues necrosis in a type III SMA mouse model, *Genes Dev* **24**, 1634–1644 (2010).
72. K. Sahashi *et al.*, TSUNAMI: an antisense method to phenocopy splicing-associated diseases in animals, *Genes Dev* **26**, 1874–84 (2012).
73. N. N. Singh, E. J. Androphy, R. N. Singh, An extended inhibitory context causes skipping of exon 7 of SMN2 in spinal muscular atrophy, *Biochem Biophys Res Commun* **315**, 381–8 (2004).
74. N. N. Singh, E. J. Androphy, R. N. Singh, In vivo selection reveals combinatorial controls that define a critical exon in the spinal muscular atrophy genes, *RNA* **10**, 1291–305 (2004).
75. L. Cartegni, M. L. Hastings, J. a Calarco, E. de Stanchina, A. R. Krainer, Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2, *Am J Hum Genet* **78**, 63–77 (2006).

76. N. N. Singh, R. N. Singh, E. J. Androphy, Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes, *Nucleic Acids Res* **35**, 371–89 (2007).
77. N. K. Singh, N. N. Singh, E. J. Androphy, R. N. Singh, Splicing of a Critical Exon of Human Survival Motor Neuron Is Regulated by a Unique Silencer Element Located in the Last Intron, *Mol Cell Biol* **26**, 1333–1346 (2006).
78. R. Drmanac *et al.*, Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays, *Science* **327**, 78–81 (2010).
79. J. C. Darnell *et al.*, FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism, *Cell* **146**, 247–61 (2011).
80. A. Bayés *et al.*, Characterization of the proteome, diseases and evolution of the human postsynaptic density, *Nat Neurosci* **14**, 19–21 (2011).