Materials and methods

Compilation of exons and alternative splicing events

We built an alternative splicing database by aligning RefSeq, mRNA and EST sequences to the mouse genome (coverage >50% and identity >95%). Splicing graphs were employed to extract exons, introns and typical types of alternative splicing events as described previously (S*1*). To further improve the comprehensiveness of the collection of exons, we included all Refseq/UCSC Known Gene exons (S*2*), in addition to the GU/AG internal exons from our database, resulting in a non-redundant set of 274,309 exons in total. Alternative splicing events used in our Bayesian network analysis consist of 13,357 cassette exons (CASS), 3,262 events of tandom cassette exons (TACA, 7,531 exons), 682 events of mutually exclusive exons (MUTX, 1,458 exons), 4,728 alternative 5′ splice site events (ALT5), 6,355 alternative 3′ splice site events (ALT3), 4,838 alternative polyA/5′ splice site usage events (APA5), and 2,273 alternative polyA/3′ splice site usage events (APA3) (See table S5 for the diagram of each type). All these events use GU/AG splice sites. For each exon, we added annotations of alternative splicing in mouse, as well as the conservation of alternative splicing patterns in human and rat, which were mapped to the mouse genome by the program liftOver obtained from the UCSC Genome Browser (S*2*).

We also compiled candidate exons with a high level of cross-species sequence conservation, but not annotated in the mouse genome. More specifically, we combined all exons mapped from human and rat (compiled using the same method as for mouse), and exons predicted by NSCAN (S*3*) and exoniphy (S*4*). Redundancies were removed and only exons whose coordinates did not match known mouse exons were kept.

Genomic mapping of HITS-CLIP tags generated by Illumina 1G sequencer

HITS-CLIP experiments were performed as previously described and sequenced by the Illumina platform (See table S1 for a summary) (S*5*-*7*). Raw reads of 32 nt in length obtained from some earlier experiments (datasets 2, 3, 8 in table S1) were filtered to keep only those with an average quality score ≥20 due to a relatively high sequencing error rate of the platform, which was improved during this study. Reads (CLIP tags) were mapped to the mouse genome (mm9) by the ELAND program included in the Illumina Genome Analyzer pipeline. To increase sensitivity,

each read was trimmed iteratively at the 3′ end and aligned using different sizes from 25 to 32 nt, requiring ≤ 2 mismatches. A read was kept for analysis only if it was mapped to an unambiguous locus. If unambiguous mapping was possible with different sizes, the one with minimum mismatches and maximum size was kept. For each individual CLIP experiment, we further collapsed tags with the same starting genomic coordinates to remove potential RT-PCR duplicates, and identify unique tags for further analysis.

Clustering of HITS-CLIP tags

A two-state hidden Markov model (HMM) (S*8*) was used to define CLIP clusters. The algorithm first calculated the number of overlapping CLIP tags at each nucleotide position, and then sampled the resulting CLIP tag coverage profile at a 5-nt resolution. This sampled profile was used to segment the genome into CLIP clusters and non-cluster regions, as represented by the two states, "+" and "-", respectively (fig. S2A). To reduce computation, we first partitioned the genome into segments by grouping neighboring reads ≤ 200 nt apart. Only segments with ≥2 reads were kept for further analysis (termed g200_2 set, 386,723 segments, mean 224 nt, median 134 nt, std 328 nt). We then ran HMM for two rounds.

In the first round, a positive training set was generated by an initial clustering procedure to group together neighboring reads ≤ 20 nt apart. This resulted in 56,840 clusters with ≥ 10 reads (median 147 nt, std 149 nt), from which a nonparametric distribution of CLIP coverage for the "+" state was estimated. Since the g200_2 set was very loosely defined, and thus robust CLIP-positive regions were expected to be a small proportion of all the segments, we used the whole set as a proxy to estimate the distribution of the background CLIP tag coverage for the "-" state. The average length of the positive set, $\mu[+]$, was also estimated from the initial clusters (176 nt, corresponding to 35.2 sampled values at the 5-nt resolution); the average length of the negative set, $\mu[-]$, was empirically set to be $20 \times \mu[+]$ (3525 nt, 704.9 sampled values). The two averages were used to estimate the transition probabilities between the states. The HMM was then trained and a Viterbi algorithm was used to infer the hidden states.

In the second round, the positive and negative training sets were refined. The predicted CLIP clusters in the first round were used as the positive training set. Segments harboring predicted clusters were removed from the g200_2 set and the remaining segments were used as the negative

training set. The refined parameters were shown in fig. S2 A and B. The rest was the same as in the first round.

The two-round procedure is conceptually similar to the Baum-Welch algorithm (S*8*), an iterative method to decompose unlabeled data and to estimate model parameters. In our case, the results are very robust using reasonably chosen parameters in a wide range, and the predictions from the first and second rounds are also very similar, so the computationally intensive iterations can be avoided. The resulting clusters were ranked by peak height (PH), i.e., the number of tags in the position with the highest coverage.

Ab initio prediction of Nova-bound YCAY clusters

Many RNA-binding proteins, including Nova, bind clusters of short and frequently degenerate motifs, which are difficult to characterize and predict using consensuses or position weight matrices of individual motifs. These concerns were partially addressed in our previous algorithm to predict Nova-bound YCAY clusters, although it was based on a set of heuristic rules derived from a small subset of Nova target exons, and was not optimized for global prediction (S*9*). To improve the accuracy, we developed a hidden Markov model (HMM) (S*8*) for ab initio Nova-bound YCAY cluster prediction. This model took advantage of a large number of in vivo Nova-binding sites defined by HITS-CLIP data to optimize model parameters and evaluate the performance, but not for prediction. Details of model design, evaluation and comparison with the previous heuristic approach will be described elsewhere. Briefly, three types of essential features were characterized to improve signal-to-noise ratios. Clustering of YCAY elements were explicitly modeled by the distance between neighboring elements and also the number of elements in each cluster. The conservation of each YCAY element was measured by a branch length score (BLS) (S*10*), which took the divergence among species into consideration, and was effective in predicting Fox targets (S*11*); twenty mammalian species were used for this calculation (S*12*). Accessibility or single-strandedness is the probability of each YCAY element located in single-stranded regions (estimated by RNAplfold (S*13*)), which was also included in the model because of the distinct distribution observed around CLIP clusters (fig. S2F; also see descriptions below). The model took YCAY elements (represented by these features) in each sequence as input and predicted YCAY clusters ranked by a log-likelihood ratio:

Log [*P*(feature| cluster)/*P*(feature|non-cluster)].

To give an unbiased assessment of the prediction accuracy, we split the whole dataset randomly into two halves. Two models were then trained on one half of the data, and tested on the other half, and vice versa. The specificity and sensitivity of the pooled predictions were evaluated by comparing predicted YCAY clusters with the footprint regions (i.e., +/- 50 nt around CLIP cluster peak) of robust CLIP clusters, to get the standard receiver operating characteristic (ROC) curve (fig. S3). To do this, a subset of ~2000 non-repetitive CLIP clusters with PH≥ 15 and located in internal exons with 1kb extension on both sides was used as a surrogate of the true positive dataset. To define a true negative set, we randomly picked the same number of 100-nt sequences from exons with 1kb extension on both sides without any CLIP tags. We call a region predicted as positive if an overlapping YCAY cluster above certain thresholds exists, and negative otherwise. Note that ROC curves usually compare the whole range of specificity and sensitivity between 0 and 1. However, there is an intrinsic threshold in the Viterbi algorithm, so that the sensitivity of HMM-based predictions never reaches 1 in our case. Nevertheless, this does not affect the comparison in performance, because only the region on the left side is of interest to have a reasonably high specificity.

After ensuring the effectiveness of the method, a full model was trained on the complete dataset to predict the Nova-bound YCAY clusters described in this study.

Processing of Affymetrix exon and exon-junction array data

Splicing changes between wild-type (WT) and Nova knockout (KO) mouse brains (or spinal cords) were measured by Affymetrix Exon Array ST 1.0 (S*14*) or custom Affymetrix exon junction arrays, as described previously (S*5, 15*); either Nova2 KO mice or Nova1/2 double KO (dKO) mice were used (See table S2 for a summary). Exon-junction array data were analyzed by ASPIRE2, as described previously (S*5*). This algorithm reports a ΔI for each alternative exon, which represents the proportional change of exon inclusion (e.g., a change from 60% inclusion in WT brains to 20% inclusion in Nova KO brains gives ΔI=0.4).

For Affymetrix exon arrays, exon intensities and gene intensities were summarized using extended probe sets and core probe sets, respectively. This was done by the PLIER/IterPLIER

model implemented in Affymetrix power tools (APT). Normalized exon intensities were derived by subtracting gene intensities from exon intensities in the $\log_2$ scale (with a pseudo count of 8 for $\log_2$ transformation). The statistical significance of splicing changes between WT and KO mice was evaluated by the empirical Bayes method (a regularized t-test) (S*16*). To minimize noises, we also applied multiple filtering criteria, as recommended by the vendor (http://www.affymetrix.com/support/technical/technotes/id_altsplicingevents_technote.pdf) with minor modifications. Specifically, a probe set was excluded for analysis unless (i) the average $\log_2$ gene intensity was ≥7 in both WT and KO samples; (ii) the average $\log_2$ exon intensity was ≥7 in either WT or KO samples; (iii) the difference in the averages of absolute exon intensity between WT and KO samples exceeded half of the difference in the average gene intensity. The last filtering criterion removed probe sets with low variance, which frequently indicated either the absence or saturation of signal.

In addition to the quality-filtering criteria, we call an alternative splicing event (for exon-junction arrays) or probe set (for exon arrays) with Nova-dependent splicing if $|\Delta I|>0.15$ for exon-junction array data, or $P<0.005$ and fold change >1.5 for exon array data, for all the descriptive analyses presented in this study (figs. S4 and S11D). These thresholds were not as stringent as necessary to derive the highest possible confidence predictions of Nova-dependent exons, but were determined empirically to include ~60% of previously validated Nova target exons. However, Bayesian network analysis does not require this thresholding.

Probe sets from exon arrays and alternative splicing events from exon-junction arrays were then mapped to exons in our database. When multiple probe sets were mapped to the same exon, the probe set indicating the most significant splicing change was used. We also paid special attention to probe set mapping in Bayesian network analysis. For alternative splicing events of types ALT5, ALT3 and APA5, a probe set has to be mapped to the alternative part of each exon (fig. S7) to be considered.

Summary of datasets for Bayesian network analysis

The biochemical map of Nova-RNA interactions in the mouse brain was derived from a large set of 81.2 million CLIP tags compiled from 20 independent HITS-CLIP experiments, among which 51.7 million (64%) tags were mapped unambiguously to mm9. After removing potential RT-

PCR duplicates, we obtained 4.4 million unique tags that represented independent captures of Nova-RNA interaction sites in the mouse brain (table S1; dataset S1). The genomic distribution of these unique tags is similar to that observed from a smaller dataset (S*5*). Overall, 76% unique tags were mapped to genic regions, as defined by RefSeq and UCSC Known Genes (S*2*); additional 6% reads were mapped to the extended 3′ UTRs (downstream 10K nt) (fig. S1A). The background level, as estimated by the tag density in intergenic regions, was as low as ~0.05 RPKM (reads per Kb per million reads (S*17*)), compared to 0.3 RPKM in genic regions (excluding the peak region at 3′end) and 4.1 RPKM near polyA sites. There is also an enrichment of tags in introns flanking alternative exons compared to constitutive exons (fig. S1B), which is consistent with the major role of Nova as a splicing regulator.

To identify Nova-binding sites, we defined 279,631 CLIP clusters ranked by PH (fig. S2 A and B; dataset S2), which varies from 2 to 824 tags. In particular, 36,186 and 6,507 clusters had a PH ≥10 and 30 tags, respectively (fig. S2C). Although HITS-CLIP experiments were performed in different brain regions of varying ages, a set of highly reproducible CLIP clusters were obtained. For clusters with PH ≥ 10 and ≥ 30, 98% and 100% have a biological complexity ≥ 5 (i.e. independently detected in ≥ 5 different mouse brains, as defined in ref. (S*7*)), respectively (fig. S2D). The high-affinity Nova-binding tetramer YCAY (S*18, 19*), but not YAAY or YACY, is enriched in cluster regions compared with flanking sequences (fig. S2E). Interestingly, Nova-binding sites in the brain have a clear preference for single-stranded sequences, presumably due to the higher accessibility of YCAY elements (fig. S2F). This is consistent with in vitro selection experiments demonstrating high-affinity Nova-binding sites with YCAY clusters in the loop region of a hairpin structure (S*18, 19*). Taken together, these observations confirm the reliability and robustness of the CLIP data.

To complement the biochemical Nova-binding map, we applied our hidden Markov model (HMM)-based algorithm to make ab initio predictions of Nova-binding YCAY clusters in the whole mouse transcriptome, including full-length introns. This resulted in 841,501 predicted YCAY clusters ranked by a confidence score (YCAY cluster score), as described above. Overlay of the HMM-predicted YCAY clusters and robust CLIP clusters gave a substantial overlap between the two datasets, as observed from a clear shift of the ROC curve towards the top-left direction from the diagonal, which represents random predictions (fig. S3). With a threshold to include all HMM-predicted YCAY clusters, the model achieved a specificity of 85%, and

sensitivity of 74%. With a more stringent threshold corresponding to a specificity of 99%, the model still had a sensitivity of 20%, indicating the effectiveness of the bioinformatic predictions of Nova-binding sites. Note that a number of predicted YCAY clusters currently lacking support from CLIP data might still be real, partly because the current sequencing depth is unlikely saturated, and vice versa (see *Gabrg2* in Fig. 3A in the main text for an example and discussion below).

To assess the functional outcome of Nova-RNA interactions, we used custom Affymetrix exon-junction arrays (S*5, 15*) as well as standard Exon Arrays (S*14*). Previous microarray experiments compared WT and Nova1 or Nova2 single KO mouse brains (S*5, 15*), which might miss a subset of target exons due to redundant or complementary roles of Nova1 and Nova2. To address this concern, we made additional comparisons of WT and Nova double-KO (dKO) mice in E18.5 brains and spinal cords, respectively. Altogether, analysis of the four microarray datasets (table S2) detected 1,331 exons showing Nova-dependent splicing changes in at least one dataset, in addition to many probe sets without exon annotations, probably representing intronic splicing intermediates or novel exons (fig. S4A). A relatively small percentage of these exons (15%) showed robust splicing changes in at least two datasets, which are correlated with conserved alternative splicing patterns (fig. S4 B and C). In many other cases, significant changes were not observed or only observed in one dataset, while more moderate but consistent changes were observed in other datasets. These observations suggest a limited statistical power for microarrays alone to distinguish functional splicing targets from noise.

Taken together, these genome-wide surveys of Nova-RNA interactions and Nova-dependent splicing result in a large number of candidate Nova targets, providing an unprecedented and rich resource with which to define a high-quality, global Nova splicing-regulatory network.

Design of Bayesian networks for Nova target prediction

Bayesian network belongs to a class of graph models, because it can be visualized by a graph, with nodes representing variables and directed edges representing causal relationships between variables. In the past few years, variations of Bayesian networks have been used to predict transcription factor binding motifs, infer regulatory modules from genomic and microarray data, transcription factor networks, and protein-protein interactions (S*20-25*). We decided to use

Bayesian networks, among many possible machine learning approaches, to integrate multiple types of data because of its advantage in modeling the statistical dependency among variables, and its flexibility to handle heterogeneous data types (i.e., continuous and categorical variables), missing data (which is common in e.g., microarray data as a result of data filtering to remove unreliable measurements, described above), and hidden variables (S*26, 27*). Furthermore, the model learned from Bayesian network is interpretable if the network is designed carefully.

The key components in the design of Bayesian networks include the network structure (nodes and edges), and types of conditional probability distributions (CPDs) specifying the relationship of each node with its parent nodes. We designed a Bayesian network for each type of alternative splicing events—cassette exons (CASS), tandem cassette exons (TACA), mutually exclusive exons (MUTX), alternative 5′ (ALT5) and 3′ (ALT3) splice sites, and alternative polyA usage coupled with 5′ (APA5) or 3′ (APA3) splice site choices (table S5). For each alternative splicing event, the model took all observed data, including YCAY clusters and CLIP clusters in the alternative exon and flanking introns, Nova-dependent splicing changes and evolutionary signatures (with missing data tolerated), to predict the probability of Nova-regulated inclusion P(In) or exclusion P(Ex) through direct protein-RNA interactions, or absence of direct regulation P(0). Below we describe the network for cassette exons in detail and specify differences in other types of events.

Network structure. It is possible to learn network structure from the data, which is however limited to relatively simple networks. We used a pre-defined network structure to model each alternative splicing event, according to our prior knowledge that considers relationships between variables. Taking cassette exons as an example (Fig. 1A in the main text), a Nova target was defined by the presence of both Nova-dependent splicing and Nova binding in either the alternative exon, upstream intron (UI), or downstream intron (DI). These four variables were treated as discrete hidden variables, whose states were inferred from observed evidence, including YCAY clusters (which determine Nova binding), CLIP clusters (which are evidence of Nova binding) in each region, and exon or exon-junction microarray data, reading-frame preservation, conservation of alternative splicing pattern (which are direct measurements or evolutionary manifestations of Nova-dependent alternative splicing) (S*28*). Based on this network, the joint probability of an exon can be decomposed as follows:

P(Exon|Network)=P(YCAY clusters, Binding, CLIP evidence, Splicing change, Array evidence, Evolutionary signatures, Target|Network)
=P(Target|Binding, Splicing change) × P(Binding|YCAY cluster) × P(CLIP evidence|Binding) × P(Splicing change| Binding) × P(Array evidence|Splicing change) × P(Evolutionary signatures|Splicing change).

Here Binding={Binding UI, Binding exon, Binding DI}, with each component assumed to be conditionally independent. Similar notational simplifications were used for YCAY clusters, CLIP evidence, arrays, and evolutionary signatures. The overall likelihood function is the product of the probability of each individual exon, assuming all exons are independent.

Summarized CLIP and YCAY cluster score. Since each alternative splicing event frequently has multiple CLIP or YCAY clusters, we tried different ways to obtain a combined measure of CLIP clusters and YCAY clusters, respectively. The following empirical rules observed from previously validated Nova target exons should be reflected in the summarization method. First, Nova binding in different positions inside the alternative exon, UI or DI in general affects exon exclusion or inclusion consistently, but the strength is dependent on the distance to the splice sites (either 5′ or 3′ splice site). Nova binding close to the splice sites generally has a stronger positive or negative effect, but there is no clear cut, as assumed in previous studies (e.g. ref (S*9*, *11*)). Second, when a Nova-binding site is too close to the splice site, it generally represses exon inclusion even from the downstream intron, presumably by blocking the splice site.

Based on these considerations, we extended each exon for 30 nt on both sides (and truncated each intron for 30 nt on both sides, accordingly). Each YCAY cluster or CLIP cluster was then assigned to the extended exon, truncated UI, or DI, respectively, according to the position of the cluster center (for YCAY clusters) or peak (for CLIP clusters). The size of exon extension was based on the length of predicted YCAY clusters (median=56 nt, mean=72 nt, for all clusters with score≥10; we used approximately the half-size). The YCAY (CLIP) clusters assigned to each region were then weighted according to the distance to the 5′ and 3′ splice sites, respectively, and then summed together to get a regional score, with respect to each splice site. Specifically, we used a weighting function $e^{-d_i^2/\sigma^2}$ for exonic clusters and $e^{-(d_i-30)^2/\sigma^2}$ for intronic clusters, where d_i is the distance to the splice site, σ is a scaling constant and i is the index of the cluster. We used σ=300, as estimated from our previous RNA map (S*5*), so that a cluster was down-weighted

by ½ when the binding site was ~280 nt away from the splice site. Finally, the weighted sum $s = \sum_i s_i e^{-d_i^2/\sigma^2}$ (for exonic clusters; or $\sum_i s_i e^{-(d_i-30)^2/\sigma^2}$ for intronic clusters) was calculated with respective to 5′ and 3′ splice sites, respectively. Therefore, each cassette exon was measured by six regional YCAY (CLIP) cluster scores denoted as $s_{UI5'ss}$, $s_{UI3'ss}$, $s_{E3'ss}$, $s_{E5'ss}$, $s_{DI5'ss}$, and $s_{DI3'ss}$. For Bayesian network analysis, the two scores with respect to splice sites in each region (UI, exon, and DI) were further combined, and the maximum was used to represent the overall strength of YCAY (CLIP) clusters in each region.

Conditional probability distributions. The conditional probability distributions (CPDs) are summarized in table S3. The CPDs of Nova binding in different regions were chosen to be sigmoidal logistic functions. For the summarized CLIP cluster score, we tried different types of CPD, such as normal and Poisson distributions, and finally used a negative Binomial distribution because of the over-dispersion ("heavy tail") of the scores. Splicing array data were approximated by normal distributions. For exon-junction arrays, we directly used the ΔI scores from ASPIRE2 (Fig. 1E in the main text) (S*5*), and for exon arrays, we used the $\log_{10}$-transformed *P*-values, with a sign indicating the direction of splicing change (fig. S5 A-C). Other nodes representing discrete variables with discrete parents were modeled by tables. In particular, the "Target" node is deterministic and predicts a direct target by (inferred) presence of Nova binding and Nova-dependent splicing change. In summary, the network for cassette exons is represented by 17 nodes and 78 parameters.

Implementation of Bayesian networks. Our Bayesian networks were implemented based on the Bayes Net Toolbox (BNT) for MATLAB (http://bnt.googlecode.com) (S*29*). This package allows one to specify the network structure and CPDs. We implemented negative Binomial distribution and Poisson distribution as new types of CPDs not available in the package, while the other types of CPDs are included in the package. For model parameter estimation, we used an Expectation Maximization (EM) algorithm included in the package that maximized the joint likelihood function, due to the presence of missing values and hidden variables.

The training of Bayesian network is semi-supervised in nature, because exons without class labels, or more generally with missing data, can still contribute to the parameter estimation. This effectively avoided over-fitting, as verified by a 10-fold cross validation procedure (described below), although the number of known targets are relatively small. For cassette exons, the

training dataset consisted of 634 exons, including 50 exons with previous RT-PCR validations, 206 exons with significant Nova-dependent splicing detected in ≥1 microarray datasets and consistent changes (in the same direction) in all four datasets, and 378 exons with $|\Delta I|<0.05$ exon-junction array data and no significant changes in any of the exon array datasets. These exons were selected to roughly represent three populations of exons (Nova-dependent inclusion, exclusion, or no direct effect), and to save time in model training. During training, class labels were specified explicitly only for Nova target exons previously validated by RT-PCR (n=50); for all the remaining exons (n=584), labels were inferred by the model. The EM procedure used to estimate model parameters converged quickly after ~20 iterations. Different starting points generally gave the same results, suggesting that a good (local) maximum was obtained. For prediction, we used the junction tree algorithm, which provided exact inferences and a reasonable speed given the scale of our network. Class labels were not specified for all exons at this stage.

After inference, the model output three probabilities for each exon that represented Nova-dependent exon inclusion $P(\text{In})$, exclusion $P(\text{Ex})$, or absence of direct effect $P(0)$. Given these probabilities, the estimation of false discovery rate (FDR) is straightforward. All exons were sorted in the ascending order by $P(0)$, giving a ranked list $P_i(0)$, $i=1,2,\ldots, N$, where $N=13{,}357$ is the total number of cassette exons. The FDR of the top K predictions is $\sum_{i=1}^{K} P_i(0) \Big/ K$ by definition.

Evaluation of over-fitting by 10-fold cross validation. The model described above was trained using a selected subset of exons including all validated Nova target exons, and thus denoted as the "full model". One should be cautious about potential model over-fitting for typical machine learning methods, especially when the dimension of the parameter space is high compared to the sample size. We expect that the learned Bayesian networks in this study should not suffer from over-fitting because both unlabeled and labeled exons contributed to parameter estimation. As a demonstration, we performed an iterative 10-fold cross validation. Specifically, in each iteration, 90% exons (including 45/50 validated exons) in the original training set were used to train a new Bayesian network, which was then used to test the remaining 10% of exons independent from those used to train the model (among all other exons, to calculate FDR). The prediction scores of 50 previously validated exons by 10-fold cross validation models trained on independent datasets were compared to those predicted by the the full Bayesian network model, from which a high

correlation was observed (r^2=0.97, fig. S6). Similar results were obtained from prediction scores of 12,723 exons not included in training of any of the models (R^2=0.98, average), indicating no apparent over-fitting.

Other types of alternative splicing events. For other types of alternative splicing events, the networks were designed with the same or very similar structures as that of cassette exons (fig. S7). The slight difference in network structures was due to difference in regions important for alternative splicing regulation. For instance, for alternative 5′ (or 3′) splice sites, we considered Nova binding, YCAY clusters, and CLIP clusters in alternative and constitutive parts of the alternative exon, and the downstream (or upstream) intron. For alternative polyadenylation coupled with splice site choices (APA5 and APA3), the regions to consider Nova binding and its evidence were adjusted accordingly, and reading-frame preservation did not apply. For mutually exclusive exons, tandom cassette exons, and APA3 exons that involve multiple non-overlapping alternative exons, each alternative exon was modeled and predicted separately. In addition, we used the parameters learned from cassette exons to other types of events, when possible, because cassette exons had the largest sample size and presumably gave the most precise estimates. To be more specific, the CPDs of nodes representing Nova binding, CLIP data and splicing array data learned from cassette exons were directly used and fixed during training.

Estimating the accuracy, specificity and sensitivity of the Bayesian networks

We first estimated the accuracy of Nova targets predicted by the Bayesian networks. For this purpose, we compared the predicted targets with AEDB (S*30*), a collection of manually curated alternative exons with functional characterizations in the literature but blind to our prediction. Specifically, we extracted the exon sequences in the database (ftp://ftp.ebi.ac.uk/pub/databases/astd/aedb/aedb-sequence_data.txt), which were aligned to the genome (mm9) by BLAT (S*31*), to match exons in our database. As a result, we obtained 31 exons, including 9 exons that were validated to be Nova targets in our previous studies (S*5, 6, 9, 15, 32, 33*). The remaining 22 exons were novel predictions, most of which had nothing known about splicing regulation despite the functional importance (table S6). We tested these 22 exons by comparing E18.5 WT and Nova1/2 dKO mouse brains using semi-quantitative RT-PCR (fig. S9; also see below). The FDRs of the 31 exons vary in a wide range, with a median rank of 288

(out of 588 predicted events in total). Therefore, these exons represent an unbiased subset of predicted Nova target exons in terms of confidence scores.

The validation above suggested an accuracy (validation rate) of 90.3% (28/31). By extrapolating this accuracy to the complete list of predictions, we estimated the number of false positive (FP) predictions to be around 57 (588×9.7%). Therefore, we can estimate the specificity of prediction by TN/(TN+FP)=1-FP/(TN+FP), where TN is the number of true negative predictions (non-target exons). Since TN was difficult to estimate, we used a lower bound to get a conserved estimate of the specificity. Our Bayesian network analysis included ~40,000 alternative splicing events. Using a conservative estimate of TN=20,000, we obtained a lower bound of the specificity to be 99.7%.

Lastly, we estimated the sensitivity of prediction using Nova targets validated in previous studies. We used either all types of alternative exons (58/77=75.3%), or the most prevalent type of alternative exon (i.e., cassette exons, 39/50=78%, based on the full model or 10-fold cross validation).

Searching for Nova-regulated exons not annotated in the mouse genome

We also searched additional Nova targets in novel exons not annotated in the mouse genome and thus not included in our Bayesian network analysis. This search was guided by Affymetrix exon array data and CLIP/YCAY clusters, as summarized in fig. S8.

Affymetrix Exon Array ST 1.0 was designed to include predicted exons. We assigned each probe set to annotated exons and introns according to our database described above. We then focused on "intronic" probe sets and identified 42 probe sets with robust changes in ≥2 datasets (fig. S8A). Manual examination of these probe sets identified 14 novel exons, including 11 cases with no or incomplete mRNA/EST evidence in the mouse genome, and 3 cases for which exon annotation was missing in our exon database (table S4). A majority of these novel exons were observed in human or rat, in mRNA-Seq data (S*17*), or predicted bioinformatically (S*3*, *4*), suggesting that they are bona fide exons. Most of these exons also have CLIP and/or YCAY clusters at positions consistent with the direction of Nova-dependent splicing change (table S4).

Alternatively, we started from 1,805 CLIP clusters with peak height ≥ 10 and overlapping YCAY clusters with a score ≥10 (FDR <12%, estimated from control YAAY clusters). Among these, 966 clusters were relatively far from known exons (≥400 nt) and had a bi-modal distribution of the sequence conservation level. The majority (66%) had an average 30-way vertebrate phastCons score ≤0.2 (S*34*) , whereas a smaller subset (34%) had phastCons scores > 0.2 (fig. S8B). We focused on 50 of 966 clusters within 400 nt from predicted exons (describe above), and with a phastCons score ≥ 0.2, because these are most likely to be functional Nova binding sites to regulate the nearby exons. We further excluded 9 clusters near predicted terminal exons and 2 clusters in genes embedded in the intronic regions of other genes, yielding 39 clusters. These 39 clusters corresponded to 38 unique exons, including 18 exons also found in human, 18 exons predicted by NSCAN and 3 exons predicted by exoniphy. Two thirds of the exons have a size that is multiple of three, and their alternative splicing is thus expected to preserve the reading frame. This is consistent with the observations from well annotated Nova targets.

In total, these two approaches identified 49 novel targets, and one additional exon (*Grik2*) validated in previous studies. Nine exons were tested by RT-PCR using P10 WT and Nova2 KO brains (fig. S10; also see below).

Nova target prediction using only CLIP clusters, YCAY clusters, or microarrays, or by other machine learning methods

Among the 698 Nova target events in the final network, only 189 (27.1%) events have robust Nova-dependent splicing changes detected in ≥ 2 microarray datasets, while 304 events (43.6%) have no significant change in any microarray dataset. To further evaluate the performance of CLIP clusters, bioinformatic YCAY clusters, or microarrays alone for Nova target prediction, we focused on cassette exons. We first built reduced Bayesian networks that used only CLIP clusters or YCAY clusters alone (fig. S11 A and B), and predicted the same number of Nova target cassette exons as we did using the full Bayesian network model (i.e., top 363 events). The top 363 candidates predicted from each dataset included 44-46% target cassette exons predicted by the full Bayesian network model (16-fold enrichment as expected by chance, $P<10^{-162}$, Fisher's exact test; fig. S11C), and 29/59 (49%) exons validated previously (S*5, 6, 9, 15, 32, 33*) or in this study (Fig. 1G in the main text and fig. S11E), indicating that each dataset made important and

comparable contributions for Nova target prediction despite the lower predictive power. In addition, predictions based on CLIP clusters and bioinformatic YCAY clusters are also largely complementary to each other, with 119 (33%) common events predicted independently by both methods (12-fold enrichment, $P<10^{-99}$, Fisher's exact test), and 244 events (67%) predicted by only one method but not the other. As an additional comparison, 29-31% of the top 363 events predicted by CLIP or YCAY clusters alone have significant Nova-dependent splicing observed in ≥1 microarray datasets (8-fold enrichment, $P<10^{-66}$, Fisher's exact test; fig. S11D). As expected, in both cases more stringent predictions with higher CLIP or YCAY cluster scores gave a larger overlap with the microarray data, confirming the effectiveness of the scoring methods used for CLIP and YCAY clusters. Similarly, compared to exons with splicing changes detected only in one microarray dataset, those with robust splicing changes detected in ≥2 microarray datasets gave a larger overlap with the top 363 predictions using CLIP clusters (53.3% vs. 14.6%, $P=4.2\times10^{-15}$, Fisher's exact test) or YCAY clusters (46.7% vs. 14.8%, $P=5\times10^{-11}$, Fisher's exact test) alone, suggesting that robust detection of splicing changes in multiple microarray datasets also contribute to more reliable predictions of bona fide and direct Nova targets. Exons with splicing changes observed in a single microarray dataset are nevertheless very informative, as the overlap expected by chance is much lower (2.7%, $P<10^{-28}$, Fisher's exact test). While the correlation between CLIP clusters, YCAY clusters and microarray datasets clearly demonstrated their importance for Nova target predictions, these observations again suggest much limited sets of Nova target exons that can be confidently predicted by analysis of individual datasets and the benefit of data integration.

To demonstrate the effectiveness of our integrative approach using Bayesian networks for accurate Nova target prediction, we then made a direct comparison with two other widely used machine learning algorithms, i.e., naïve Bayes and logistic regression (S*35*), implemented in the software R (S*36*). Binary classification models were built using each method to predict cassette exons with Nova-dependent inclusion or exclusion (in a 10-fold cross validation procedure to avoid over-fitting), taking microarray data, CLIP cluster scores and YCAY cluster scores, in the same format provided to the Bayesian network (except that missing values in microarray data were replaced by zeros for logistic regression because it does not allow missing data). Predictions were ranked according to the confidence scores. Among the top 363 target exons predicted by each of these two methods, 51.7-54.5% exons were also predicted by Bayesian network, suggesting a substantial discrepancy of these methods compared to the Bayesian

network integrating similar sets of data (fig. S11C). At the same stringency (i.e. top 363 targets), both methods predicted 36/59 targets validated previously or in this study, giving a sensitivity of 61%, as compared to 75-78% (39/50 target cassette exons, or 58/77 target events overall) by Bayesian networks (Fig. 1G in the main text). In addition, Bayesian networks also outperformed in predicting "weaker" targets beyond the stringent threshold we use. Logistic regression appeared to perform better than naïve Bayes in predicting the very top candidates (i.e., rank<350), but less well in candidates with moderate or relatively low ranks, as judged from the overlap with Baysian network predictions and the validated targets.

Although accurate estimation of the validation rate for each compared method requires experimental testing of an unbiased set of candidate targets, as we did for predictions by Bayesian networks, we derived approximate estimates (fig. S11E). To illustrate the method, here we use Naïve Bayes as an example. From the overall validation rate (90%) and sensitivity (75%) of the Bayesian networks, we estimate that 363*0.9=327 bona fide Nova target cassette exons were predicted by the Bayesian network analysis, and that Nova has 327/0.75=436 bona fide target cassette exons in total. Therefore, ~109 Nova targets are likely missed by Bayesian network analysis. According to the sensitivity of naïve Bayes (61%), ~109×0.61=66 of these 109 target exons will be predicted by this method. Similarly, among the 198 exons predicted by both naïve Bayes and Bayesian networks, ~198×0.9=178 are bona fide targets. Thus, the validation rate of naïve Bayes in the top 363 predictions is estimated to be ~67% (244/363). A similar or lower validation rate can be estimated for predictions by logistic regression, or using only microarray data, CLIP clusters, or YCAY clusters. Taken together, our Bayesian network approach is very effective in data integration and gave much more favorable results in terms of accuracy and sensitivity than the other methods we tested.

Analysis of combinatorial regulation

We focused on Nova target cassette exons for detailed analysis of combinatorial splicing regulation. From a total of 363 exons, a non-redundant set of 325 exons was defined (dataset S3). For example, if two cassette exons overlapped and were spliced to the same flanking exons, only the one with the most supporting transcripts was kept. An average-linkage hierarchical clustering (S*37*) was performed for these non-redundant exons using the six regional YCAY scores $s_{UI5'ss}$,

$s_{UI3'ss}$, $s_{E3'ss}$, $s_{E5'ss}$, $s_{DI5'ss}$, and $s_{DI3'ss}$ described above. These scores reflected the strength of Nova binding in upstream intron, exon, or downstream intron, and Nova regulation through either 5′ or 3′ splice site. Seven clusters with distinct combinatorial binding patterns were identified (Fig. 2A, clusters I-VII). To generate sequence conservation profiles, 30-way vertebrate phastCons scores (S*12*) were extracted for the 30 nt sequences from 5′ and 3′ splice site of the cassette exon, and for the 200 nt sequences near 5′ or 3′ splice site of the upstream and downstream introns, respectively. The average conservation level was calculated. As a control for this analysis, we used all cassette exons in the mouse genome (Fig. 2B).

To search for specific cofactors regulating Nova targets, we used 49 hexamer motifs enriched in or near brain- or cerebellum-specific alternative exons (S*38*). For this analysis, we examined 200 non-redundant target cassette exons activated by Nova and 125 exons repressed by Nova, separately. For each hexamer, its relative frequency in Nova target exons in comparison to all cassette exons was calculated in two ways for each exonic or intronic region and used to generate 2D scatter plots shown in fig. S12. The x-axis shows the relative frequency using the observed frequency in all cassette exons as the denominator. In contrast, the y-axis shows the relative frequency using the frequency expected from the base composition of all cassette exons as the denominator. As expected, the two measures generally correlate well. A more stringent filtering can be achieved by requiring the relative frequency estimated by both methods above certain thresholds (e.g. 1.5 fold).

Putative Fox targets were predicted based on the conserved UGCAUG motif specifically recognized by Fox proteins, as described previously (S*11*). Since the original analysis was performed in human, the exons were mapped to the mouse genome by the program liftOver obtained from UCSC Genome Browser (S*2*).

To test Nova and Fox combinatorial regulation in 293T cells, we selected a subset of Nova target exons, in addition to *Gabrg2* exon9, that were also putative Fox targets. The human orthologs of these exons were identified and further filtered to remove exons without detectable expression in 293T cells. This was done using a published Affymetrix gene expression microarray dataset (intensity >100 in untreated 293T cells; GSE2451 (S*39*)). Finally, 18 exons which passed this filtering were used for RT-PCR analysis (described below); two exons lacking bands of expected sizes were excluded.

Among the 17 tested exons (including *Gabrg2*), seven exons responded to both Nova and Fox, one exon responded to Nova only, and two exons responded to Fox only. The splicing outcomes of these exons and the position of Nova and Fox binding sites are summarized in Fig. 3 in the main text and fig. S13. The relatively moderate validation rate is likely due to our specific experimental settings and the biological difference between cell culture systems and brain. First, we performed moderate overexpression of Nova and Fox (0.5 μg individually and 0.25+0.25 μg in combination) optimized to best demonstrate the synergistic regulation of *Gabrg2* exon 9 by Nova and Fox, by avoiding saturation. It is clear from our previous experiments that a more dramatic effect can be achieved by a higher expression level of Nova/Fox (e.g. 2μg in transfection, ref. (S*33*) and data not shown). Some exons appeared to be less sensitive to moderate Nova/Fox expression, although they are bona fide targets. Second, the cell culture system did not completely recapitulate the physiological cellular environment in the brain, and some additional factors required for combinatorial regulation might not be expressed in 293T cells used for this study. Supporting either of these two interpretations, we found that among the 6 exons that were validated Nova targets in the brain, only two showed Nova-dependent splicing in 293T cells (roughly comparable to the overall validation rate of Nova regulation, i.e. 8/17, in 293T cells). Lastly, some bioinformatically predicted Fox targets might be false positives.

Gene ontology (GO) and pathway analysis

GO analysis was performed using the online tool DAVID (S*40*) (table S7). This tool can also be used to analyze KEGG pathways (table S8) and keywords extracted from protein databases (PIR and Uniprot) (Fig. 4 in the main text). A total of 13,054 genes with detectable expression in at least one of the three exon array datasets (average $\log_2$ intensity ≥ 7 in WT samples) were used as background genes for comparison with Nova target genes.

Kinases and phosphatases

Genes encoding kinases and phosphatases were obtained from the database PhsphoregDB (S*41*). Among the 684 genes in the database, 662 genes with Entrez Gene IDs were used for this study.

The enrichment of kinases and phosphatases in Nova targets were tested against all genes expressed in the brain, as described above.

Phosphorylation sites

We downloaded 16,123 mouse protein sequences and their annotations from the Uniprot/Swissprot protein database (http://www.uniprot.org)(S*42*). Sequences without Entrez gene annotations were excluded. Among the remaining 15,944 sequences, 13,881 were aligned to the mouse genome by exonerate (S*43*), using the protein2genome mode (coverage>0.5, identity>0.9). We then extracted the coordinates (relative to protein sequences) and annotation of 21,499 experimentally determined phosphorylation sites from the database, among which 18,749 sites were successfully mapped back to the mouse genome. In particular, these included a subset of 1,710 sites that were experimentally determined to be phosphorylated in the mouse brain in vivo, based on five published large-scale mass spectrometry studies (S*44-48*).

With the genomic coordinates of each phosphorylation site, we can calculate the number and frequency of phosporylation sites encoded by each exon. To compare the frequency of phosphorylation sites among different groups of exons, i.e., constitutive exons, overall alternative exons, brain-specific alternative exons that are not Nova targets, and Nova target exons, we took into consideration the fact that the current protein database is likely incomplete, at least at the exon level. Therefore, exons in our database were matched with exons defined by protein sequences; only exons with matched protein sequences were used to calculate the frequency of phosphorylation sites per amino acid (Fig. 4B in the main text and fig. S14).

Nova target genes implicated in genetic diseases

To search for Nova target genes implicated in human disease genes (table S9), we used two databases, OMIM (S*49*) and HGMD (S*50*). Autism susceptibility genes were obtained from Simons Foundation Autism Research Initiative (SFARI) gene database (https://sfari.org/sfari-gene) (S*51*). Mouse homologs of these genes were identified using the Homologene database (http://www.ncbi.nlm.nih.gov/homologene). The enrichment of disease genes in Nova targets

were compared with all genes expressed in the brain, the same set of background genes for GO term analysis, as described above.

Biochemical assays

Transfection. $1x10^6$ HEK293T cells were plated onto 6 cm dishes in 3ml 10% DMEM the day before transfection. Prior to transfection, culture media was replaced with 3ml antibiotic-free 10% DMEM. Transfection mixes containing 150ul opti-mem, 6ul lipofectamine 2000™ (Invitrogen) and DNA (2.25 μg total) were prepared according to manufacturer's instructions and added to the culture dishes. 24H post transfection cells were scrapped in ice-cold PBS and spun down (1.500 r.p.m. 5min). One half was resuspended in 1ml Trizol for RNA extraction and the other half was resuspended in 150 μl lysis buffer (Hepes pH7.4, 150 mM KCl, 5 mM $MgCl_2$, 0.5 mM DTT, 1% NP-40) for protein analysis. Wild type and mutant $GABA_A$ receptor minigenes were generated in a previous study (S*33*).

Immunoblot. 1μg/μl protein samples were prepared with laemmeli loading dye and 15 μl were loaded into 8% SDS-PAGE Novex Tris-Glycine gels (Invitrogen). After protein transfer onto PVDF Hybond (Millipore) 0.4um membranes, the following antibodies were used to immunoblot Nova1, Flag-Fox2 and γ-tubulin, respectively: rabbit anti-Nova1 serum, mouse HRP-conjugated anti-Flag (A8592 Sigma) and mouse anti-γ-tubulin (T6557 Sigma).

RT-PCR. Radio-labeled semi quantitative RT-PCR was performed as described previously (S*32, 33, 52*), using primers located in flanking exons of the alternative spliced region (or in the alternative exon in some cases) to amplify specific isoforms. Bayesian network predicted target exons were tested using three pairs of E18.5 WT and Nova dKO brains. Novel exons predicted from Affymetrix Exon Arrays and CLIP/YCAY clusters were tested using three pairs of P10 WT and Nova2 KO cortex. Targets under the potential regulation by Nova and Fox were tested in 293T cells, as described above. Biological replicates were evaluated to generate the results in Fig. 3B in the main text. Four replicates (two biological replicates, each with two technical replicates from independent RT-PCR analyses) were used to generate the results presented in Fig. 3C in the main text and fig. S13. For each candidate, different cycle numbers were tested to ensure linear amplification of PCR products. Radiolabeled ^{32}P-dCTP was added to the PCR reactions for the

last two cycles, except for *Gabrg2*, in which case the forward primer was radio labeled. RT-PCR primer sequences used in this study are listed below.

A. Primers for Bayesian network predicted exons:

Dcc_F	TCTCATTATGTAATCTCCTTAAAAGC
Dcc_R	TCACAGCCTCATGGGTAAGAG
Dclk1_F	ATGAGCATCAGCTGTCAGTAGC
Dclk1_R	GGAAAACCTGCCTCTCCTTATC
Epb4.1l2_F	GATTGGTGTTGTGGACCAAAG
Epb4.1l2_R1	GGATCTAAGACCGAATCCAATG
Epb4.1l2_R2	TTATCCCCGTCTACTCTCAAGG
Gnas_F	AAAAGCACCATTGTGAAGCAG
Gnas_R	TCAGGTTGTTTTTGATGTCCTG
Magi1_F	GTGCTCCCTGAGTACCTACCTG
Magi1_R	TTTTCTCCAGAGGAAGATGTCC
Kcnma1_F	ATGACGTCACAGATCCCAAAAG
Kcnma1_R	AGTTCCTCATGCCCCCATTAC
Ktn1_F	GCATTTGGAAATGGAGCTAGAG
Ktn1_R	CCTTCTTTCTCTCCGTTTGTTC
Numb_F	AGAGCATCAGCTCCTTGTGTTC
Numb_R	CAGAAGACTGACCCCACTCAG
Pak3_F	CATACGATTCATGTGGGTTTTG
Pak3_R	GACCGTTTCTTTGGAGTCGTAG
Ikzf1_F	GAGGACCTGTCCACTACCTCTG
Ikzf1_R	ATAGTTGCAAAGATGGCATTTG
Npr2_F	AAGCTGATGCTGGAGAAGGAG
Npr2_R	TGTTTGATGGCAACAACATTTC
Cyb5r4_F	CTGCCTCCAAGTACTCACCTTC
Cyb5r4_R	GACTGGACATGTGAGACAAAGC
Epb4.1_F	AAGGAAGCTGTGAAGGTTGAAG
Epb4.1_R	TTGATGTTAAGAGTCCGGAAGG
Smtn_F	TTCCCTGAGGCTTTTGACTATG
Smtn_R	CACATAGGTGAAGACGCACTTG
Bin1_F	CCCTGAGAAAGGGAACAAGAG
Bin1_R	ATTCACAGTTGCGGAGAAGG
Mpzl1_F	AACATTCGAAGCGGGATTATAC
Mpzl1_R	TATGGACATTTTCTGCACAAGC
Cacna1c_F2	AGTGATCCCTGGAATGTTTTTG
Cacna1c_F3	GATATAGCAATCACCGAGGTACAC

Cacna1c_R	AGGACTTGATGAAGGTCCACAG
Snap25_F2	AACAACTCGATCGTGTCGAAG
Snap25_F3	TGACGGACCTAGGAAAATTCTG
Snap25_R	CATCTGCTCCCGTTCATCC
Tpm2_F1	ATGAAGGATGAGGAAAAGATGG
Tpm2_R1	CTGAGGCTATCAGCGACTTGAG
Tpm2_R2	TTTTCAGCTCCTCCTCTAGGTC
Itsn1_F	ACTGGTGGAAAGGAGAAGTCAG
Itsn1_R1	AGACATTTCGATCACATGCAAC
Itsn1_R2	CTGCAAGTCGTTCACGTAGTTC
Tpm1_F	GCTGAGTTTGCAGAGAGATCAG
Tpm1_R2	TTATTTTACACTGGGCGAATTG
Hs6st2_F	GCTCTTCTCCAGGTTCTCCAC
Hs6st2_R	CCATTCACTCAAGTACCGTGAC
Nav2_F	GAAGACTCCTTGATGCCTTTTG
Nav2_R	GGAATTGGTAGCAGTGTCCTG

B. Primers for novel exons regulated by Nova:

Sfrs9-F	GGACCTCGAGGACTTGTTCTAC
Sfrs9-R	GGAAATCTGACCGTCTTGTAGG
Apc-F	ATCTGTCCTGCTGTGTGTGTTC
Apc-R	GCAACATCTCCAAAGGTCAAG
Mical3-F	GTCCCTGTGGAAATCTGTCTTC
Mical3-R	TCCTCCTCTGTGTAGGTTCTGG
Map3K9-F	AGCTCAGTTGTTCCAAAAGAGG
Map3K9-R	TATCTCCATAAGGCTGGTGAAC
Larch1-F	ATACGAGAGAACTCCCCTTCAG
Larch1-R	CAGCTCCTTCTCCTCTCTCATC
Myo9a-F	ACATCAGTAGCCTGGAATTTGC
Myo9a-R	GTCAGTACTTTTTCCTCCTGCTG
Nrxn3-F	TCACCCTTTCCTTTAAGACCTG
Nrxn3-R	GAGTGTAGCCCGTTGTGGTTAG
Ank3-F	AAGAGACATAAACTGGCCAACC
Ank3-R	ACTGATGTTCCTTCCAGGTCTC
Myh10-F	AGTGGCTGATGAAGAACATGG
Myh10-R	CTTCTCGTGATTTGGAATGATG
Mapk9-F	TGAGTGACAGTAAAAGCGATGG
Mapk9-R	GTTTGGTTCTGAAAAGGACGAC

C. Primers for exons to test for Nova and Fox combinatorial regulation in human 293T cells:

pGABA_F	AGAGCATCAGCTCCTTGTGTTC
pGABA_R	CAGAAGACTGACCCCACTCAG
hAlcam_F	TTAACTTGCACAGCAGAAAACC
hAlcam_R	TGTACAGCCAGTAGACGACACC
hCamk2g_F	TATTGAACAAGAAGTCGGATGG
hCamk2g_R	CTCATCTTCTGTGGTGGTGTTG
hEfna5_F	CGGAAGAAGGTCCTGTCTAAAG
hEfna5_R	GAACAGTAGGATTGCCAAAAGG
hNumb_F	ACAGATCACCAATGCCTTCAG
hNumb_R	GGACGCTCTTAGACACCTCTTC
hPpp3cb_F	TTCTGAGTATTTGCTCTGATGATG
hPpp3cb_R	GAGAAGACTCTTGCCATCTTGC
hSpna2_F	AGAATCTCCTGGAGGAGCAAG
hSpna2_R	ACTTCTTCTGGAGCACCTCAAC
hFam49b_F	GAAGGACAAGGACAGAATCACC
hFam49b_R	CCAGCTCCTCTGTATGACTGC
hHisppd1_F	AGATGAAGTTGATCGAGCTGTG
hHisppd1_R	TGCTTCTGTTCCACAAGAGTTC
hSipa1l2_F	GTTCTCCGATGGGTCCTTATC
hSipa1l2_R	CATATCTGTCAGGGTGCAGAAG
hSyne2_F	CTCACGAAGAGGACGAGGAG
hSyne2_R	TTCCATTTGCTTGTAGTGATGC
hCadm1_F	AAGCTCACTCGGATTATATGCTG
hCadm1_R	CAGAATGATGAGCAAGCACAG
hSmarce1_F	ACCATCTTATGCCCCACCTC
hSmarce1_R	CATCAGCGGCTTATCTGGTG
hPbrm1_F	CAACACCCAGACTACTCTTTCG
hPbrm1_R	CCACTCCTTGGTTCATCACAC
hAtp2b1_F	GCCAAATCTTGTGGTTTAGAGG
hAtp2b1_R	ATTCCGGTTTTTCTAACCCTTC
hArhgef12_F	CCGAGAGTCACCAACAGATAAG
hArhgef12_R	ACGAAGACTGGATTGTCTCCAC
h5730419I09Rik_F	TAATTCCCCATCCAAAGAAATG
h5730419I09Rik_R	AACGCTGATTGAGTCTGTTGTC

Supporting figures

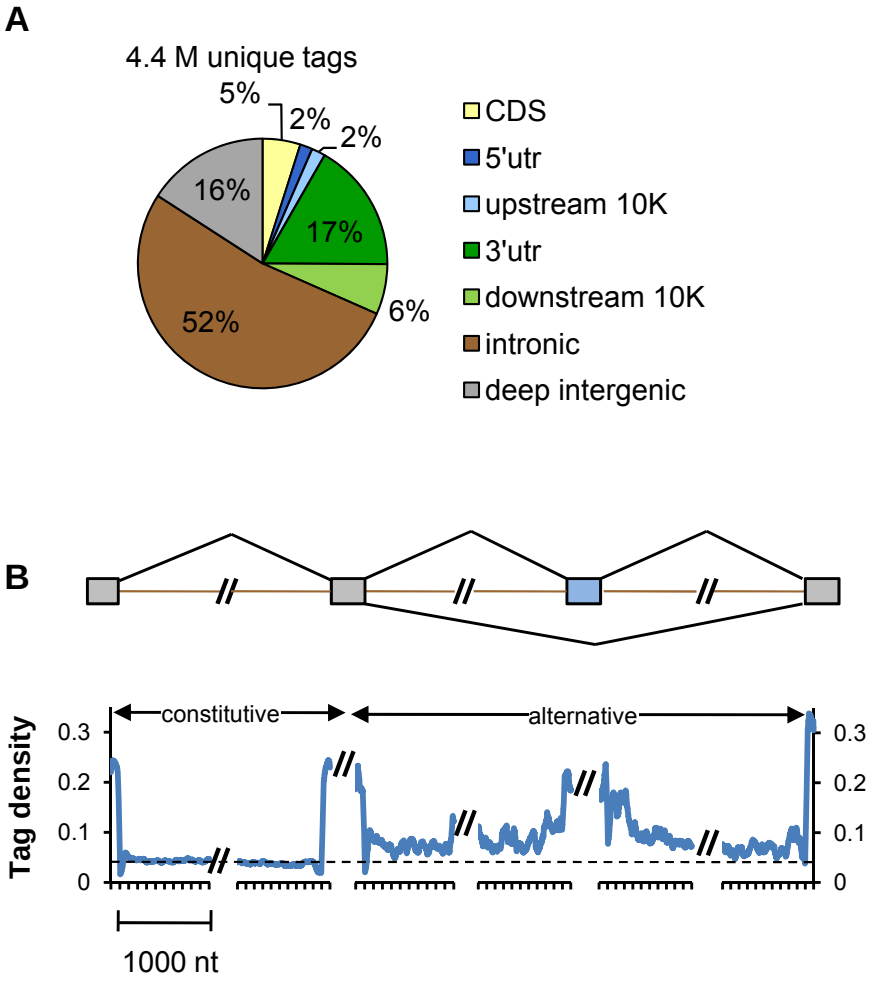


Figure S1. Distribution of HITS-CLIP tags.
(**A**) The distribution of 4,401,528 unique CLIP tags over the genome is shown. Coding sequences (CDS) , 5′ and 3′ untranslated (UTRs) were defined according to RefSeq and UCSC Known Gene transcripts. Genes were also extended for 10 kb on both sides to distinguish tags in extended regions and those in deep intergenic regions. (**B**) A composite map of tag density in exons and introns with constitutive or alternative splicing. For constitutively spliced regions, 90,770 exons, and 58,357 introns ≥2kb in length derived from RefSeq transcripts are shown. For alternatively spliced regions, the map was derived from 3,651 introns ≥2kb in length upstream of cassette exons, and 3,826 introns ≥2kb in length downstream of cassette exons.

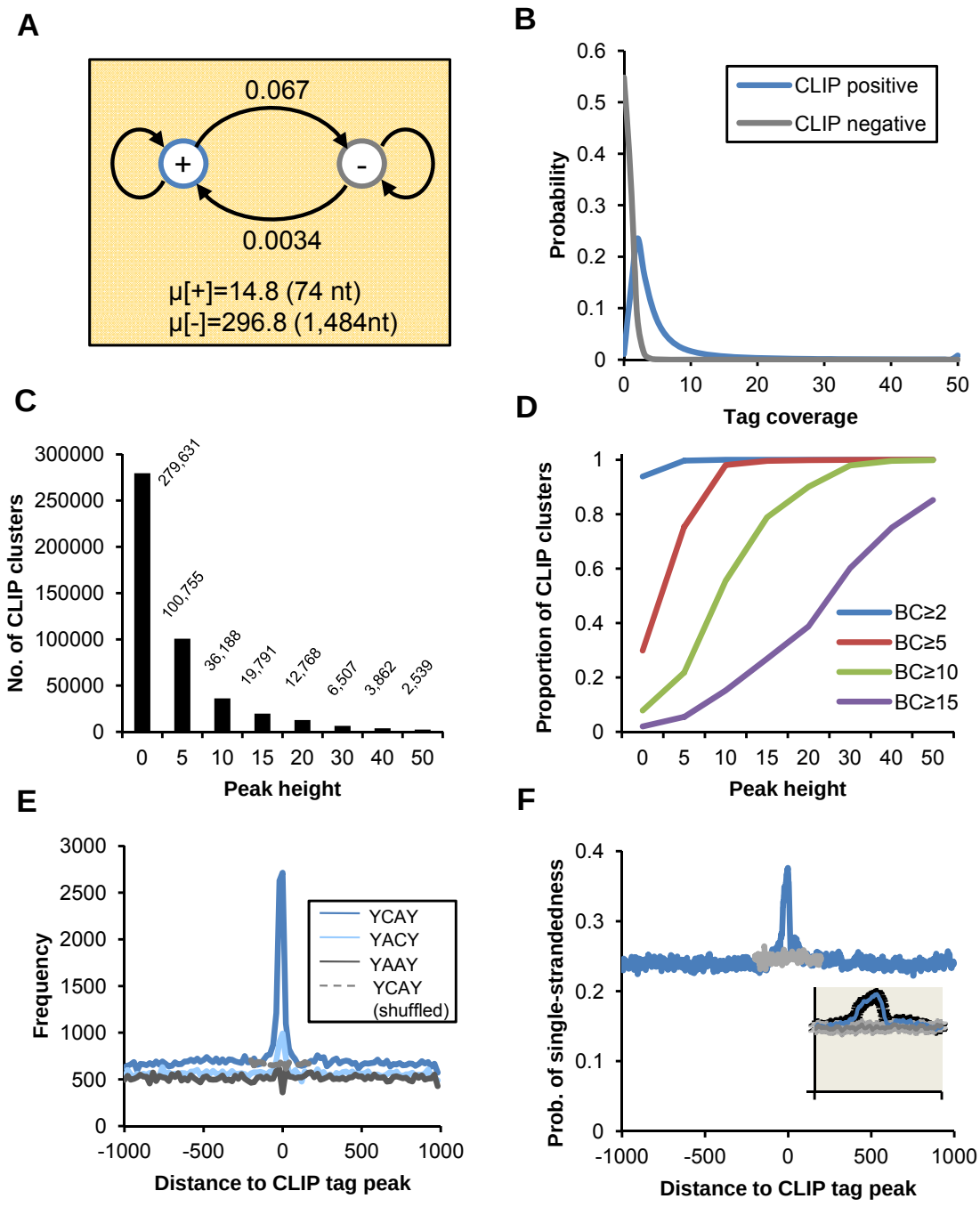


Figure S2. Identification and characterization of CLIP clusters.
(**A**) Schematic representation of the two-state HMM to define CLIP clusters. The transition probabilities between CLIP-positive regions (clusters) and CLIP-negative (background) regions indicated at the edges were estimated from the average length of initial CLIP clusters and background sequences. (**B**) Emission probability distribution of tag coverage for CLIP-positive and -negative regions. (**C** and **D**) Breakdown (**C**) and biological complexity (BC) (**D**) of CLIP clusters. CLIP clusters are defined by different thresholds of peak height (PH). (**E**) YCAY frequency around CLIP clusters, as estimated from 1,938 non-repetitive CLIP clusters with a size ≤200 nt and PH≥30 tags. The frequency of control motifs (YAAY and YACY), as well as YCAY in shuffled sequences (+/- 200 nt) , is also plotted for comparison. (**F**) The average probability of single-strandedness for each tetranucleotide around CLIP clusters as in (**E**). The gray curve +/- 200 nt around the CLIP tag peak is the probability calculated using permuted sequences in the region. The inset shows a zoom-in view +/- 200 nt around the CLIP tag peak, with error bars representing two standard errors.

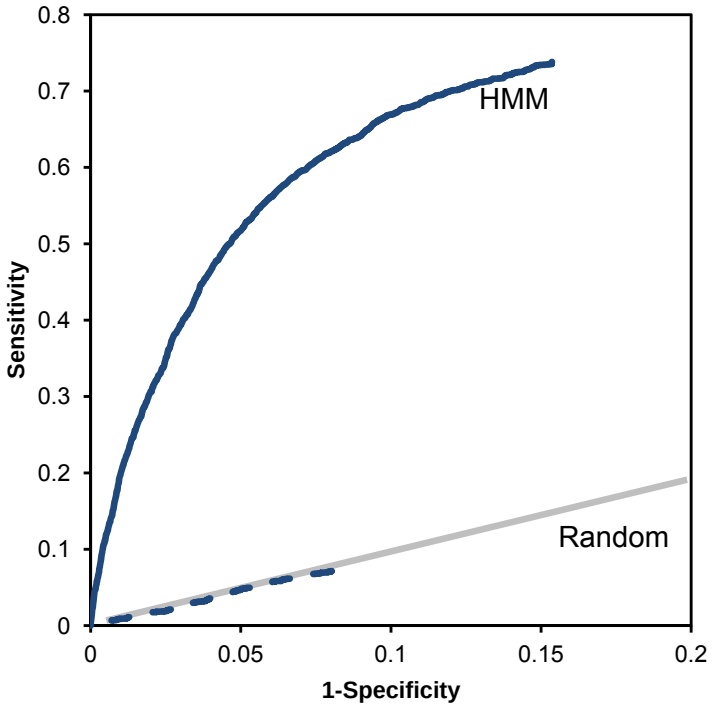


Figure S3. Evaluation of ab initio YCAY cluster prediction.
YCAY clusters were predicted in half of the dataset using the HMM trained on the other independent half-size dataset, and vice versa. Specificity and sensitivity were estimated from the overlap of predicted YCAY clusters with the footprint region (+/-50 nt) of CLIP clusters or random genic sequences of the same size, to obtain the receiving operating characteristic (ROC) curve (solid blue). The diagonal shows the performance of random guess derived from theoretical estimation (gray line) or from permutation experiments in which the labels of CLIP footprint and background regions were shuffled (dotted curve).

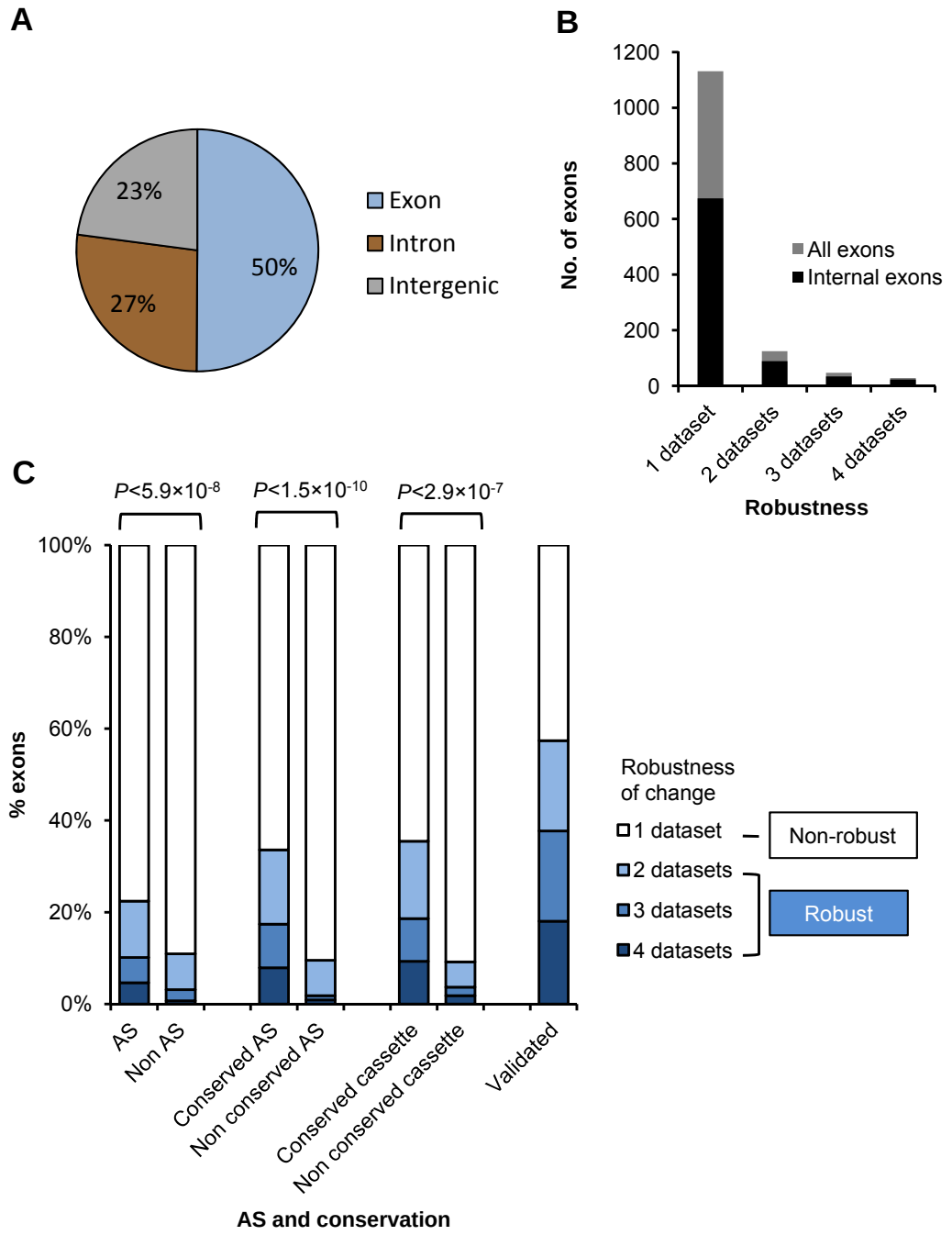


Figure S4. Robustness of Nova-dependent splicing changes.
(**A**) Breakdown of probe sets with significant Nova-dependent splicing in at least one of the three Affymetrix Exon Array datasets. (**B**) Probe sets from exon arrays or exon junction arrays were mapped to exons (or more strictly internal exons). Exons were grouped according to the number of datasets, in which significant Nova-dependent splicing was detected. (**C**) Comparison of different groups of exons in robustness of splicing changes, as represented by the number of datasets in which significant splicing changes were detected. To perform a statistical test, exons with splicing changes in ≥ 2 datasets were defined to have a robust splicing change. *P*-values indicating the difference of robustness derived from a Fisher's exact test are shown at the top.

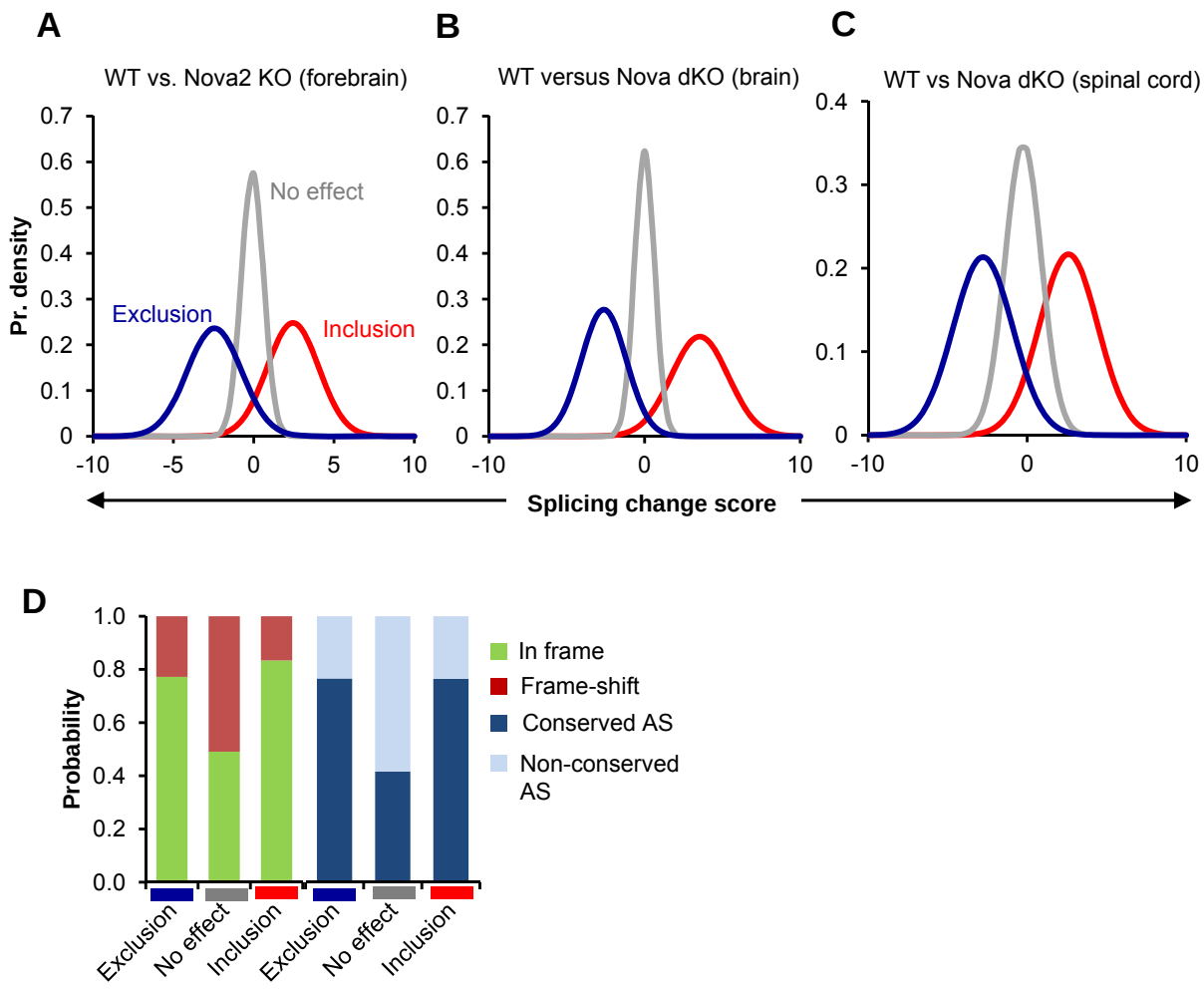


Figure S5. Conditional probability distributions (CPDs) of the Bayesian network to predict Nova target cassette exons.

The CPDs not included in Fig. 1 of the main text are shown. (**A**-**C**) CPDs of the three exon array datasets. For each panel, *P*-values are $\log_{10}$ transformed with a sign representing the direction of the splicing change, so that the distribution is approximately normal. The distributions for three populations of exons with Nova-dependent inclusion, exclusion and exons without Nova-dependent splicing (No effect) are shown in red, blue and gray, respectively. (**A**) P10 WT versus Nova2 KO forebrains. (**B**) E18.5 WT versus Nova1/2 double KO (dKO) whole brains. (**C**) E18.5 WT versus Nova1/2 dKO spinal cord. (**D**) Reading-frame preservation and conservation of alternative splicing in human or rat for exons with Nova-regulated inclusion, exclusion, and exons lack of Nova-dependent splicing (No effect).

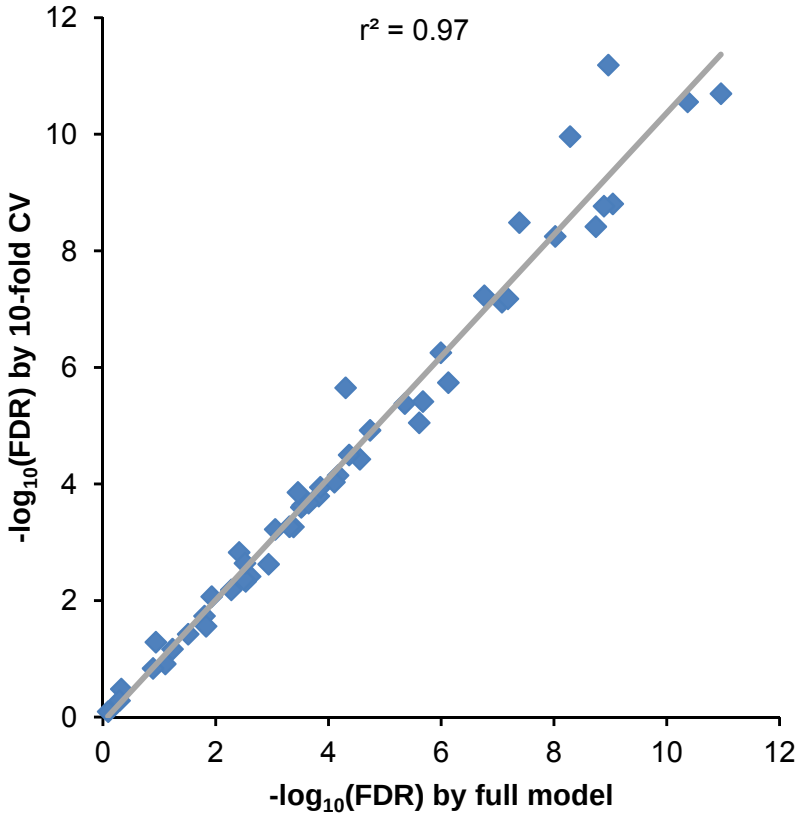


Figure S6. Evaluating over-fitting of the Bayesian network to predict Nova target cassette exons. Fifty previously validated target exons independent of this study were used for this evaluation. X-axis shows the FDR of each exon predicted by the full model, which was trained using the complete training dataset (including all validated exons). Y-axis shows the FDR of each validated exon predicted in 10-fold cross validation (CV). In this cross validation procedure, models were trained using 90% of the training data and used to predict the independent set of 10% exons left out.

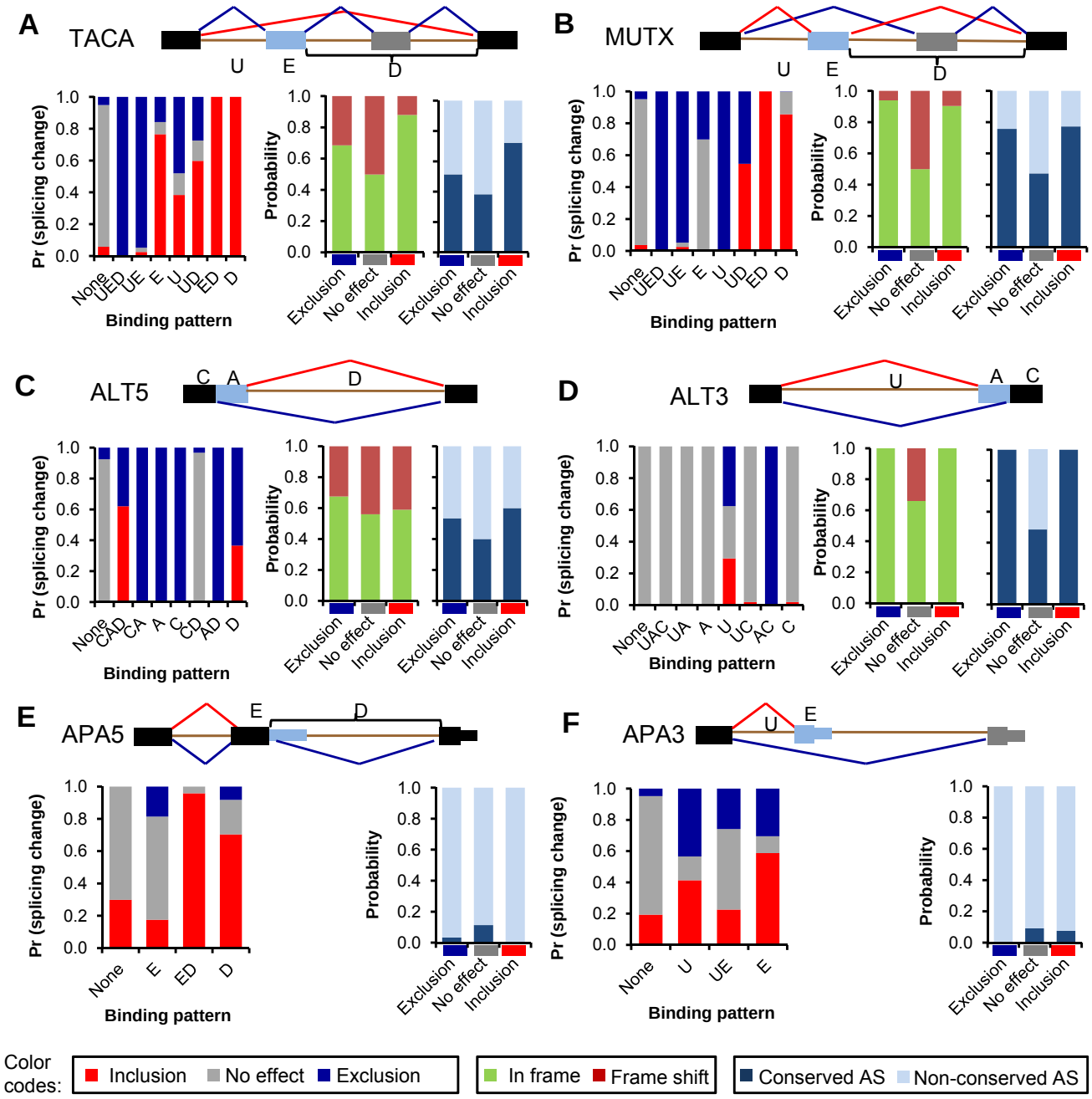


Figure S7. CPDs of Bayesian networks for other types of alternative splicing events.
(**A**) Tandem cassette exons (TACA); (**B**) Mutually exclusive exons (MUTX); (**C**) Alternative 5′ splice sites (ALT5); (**D**) Alternative 3′ splice sites (ALT3); (**E**) Alternative polyA sites coupled with 5′ splice site choices (APA5); (**F**) Alternative polyA sites coupled with 3′ splice site choices (APA3). For each panel, regions where Nova binding was considered in the model are indicated in the schematic diagram (U: upstream intron, E: exon, D: downstream intron, C: constitutive region, A: alternative region). Only CPDs learned for each specific type of alternative splicing are shown. Reading-frame preservation does not apply to APA5 and APA3.

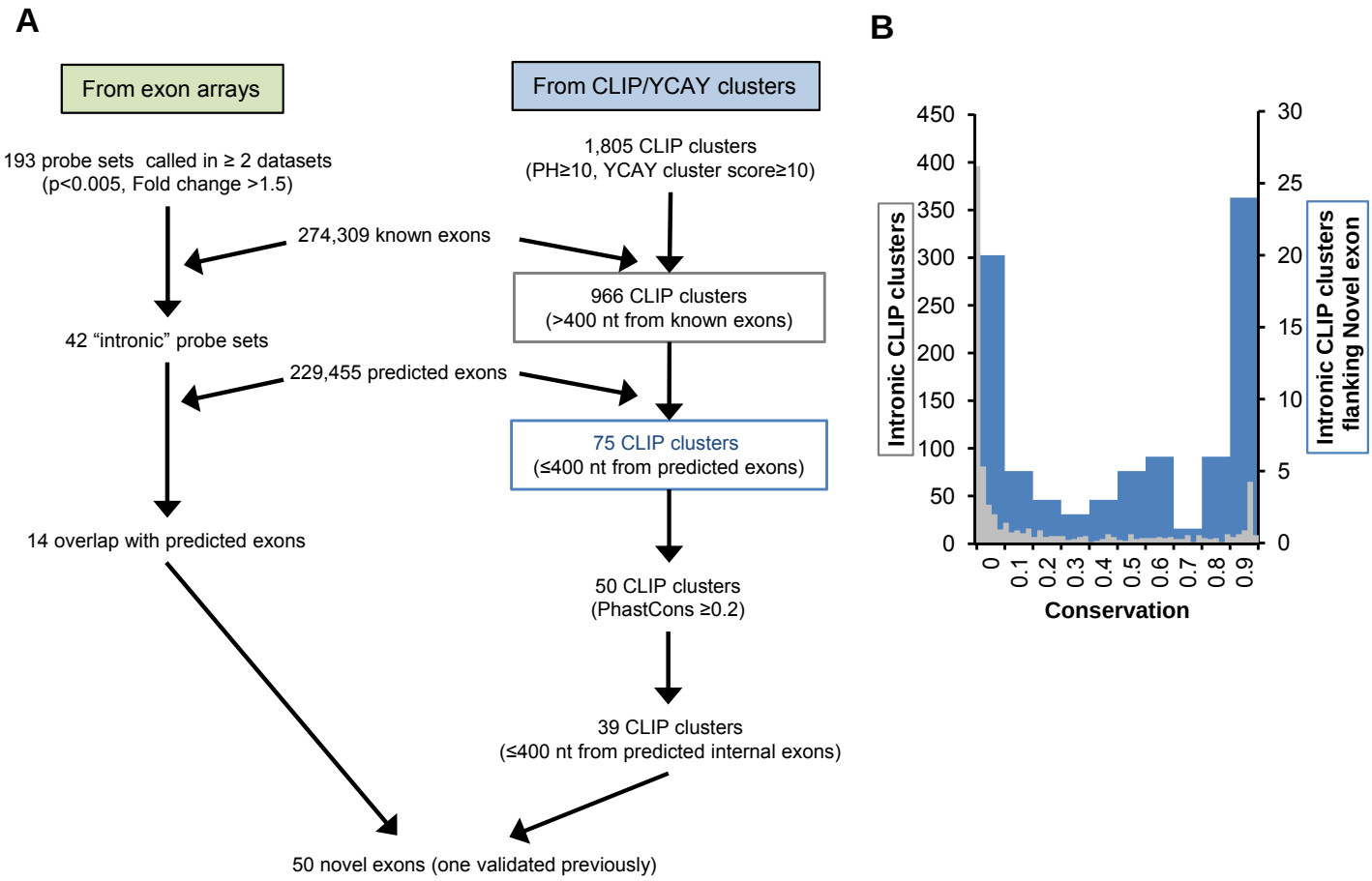


Figure S8. Prediction of novel exons regulated by Nova.
(**A**) Predictions were based on exon array probe sets interrogating predicted exons (left) or based on CLIP/YCAY clusters (right). Multiple filtering steps were performed to obtain a stringent subset of candidate targets. (**B**) Distribution of sequence conservation (phastCons scores) of CLIP clusters. 966 CLIP clusters far from known exons (≥ 400 nt), marked by a gray box in (**A**), are represented by gray bars (left axis); 75 CLIP clusters far from known exons and close to predicted exons (≤ 400 nt), marked by a blue box in (**A**), are represented by blue bars (right axis).

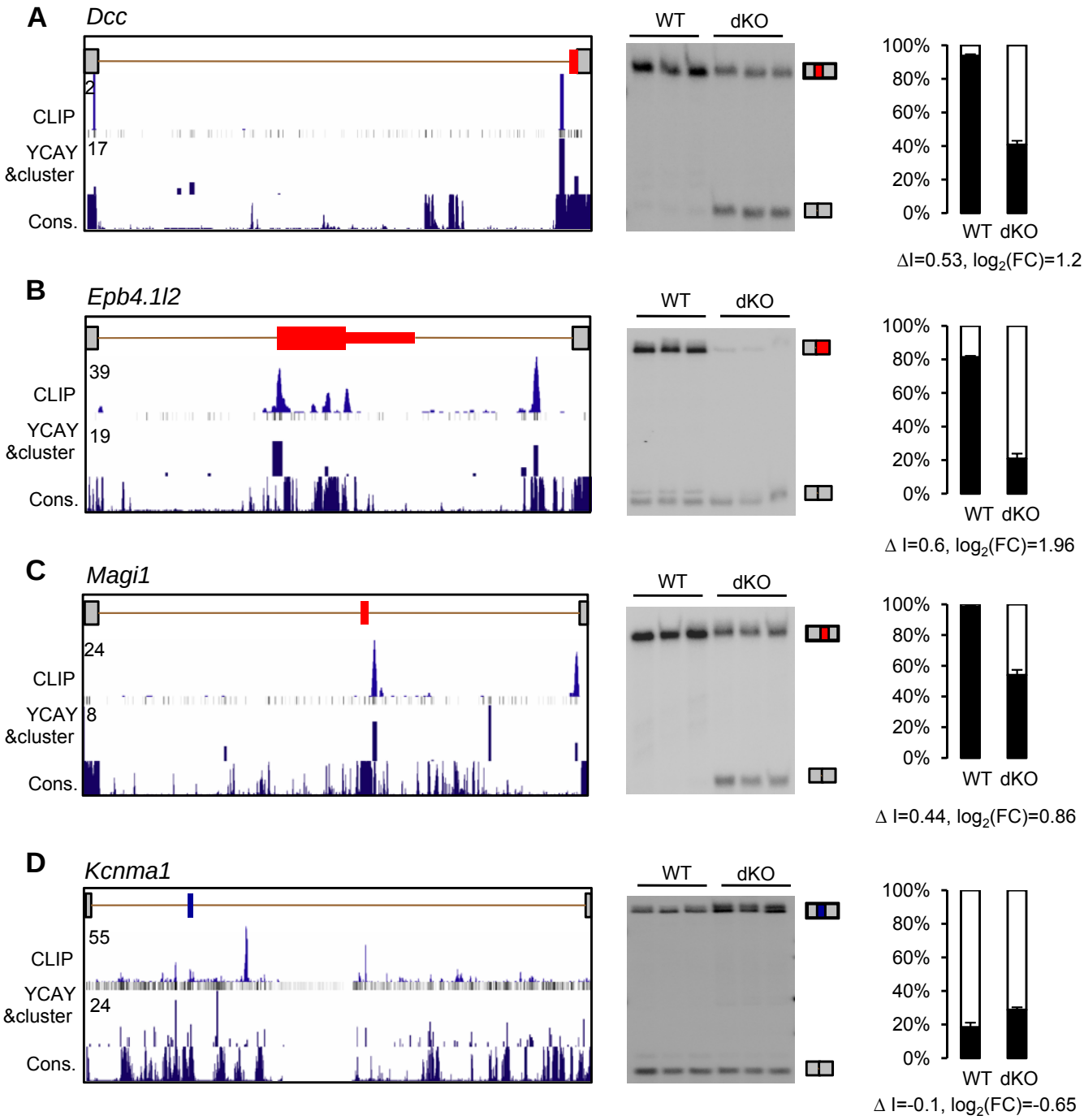


Figure S9. RT-PCR validation of Nova target exons predicted by Bayesian networks. Each panel shows a predicted Nova target alternative splicing event. Left: the diagram at the top shows the alternative exon(s) and flanking exons and introns. The alternative exon(s) with Nova-regulated inclusion or exclusion are highlighted in red or blue, respectively. Below the diagram show four tracks in order: CLIP tag coverage, positions of YCAY elements in vertical bars, positions and scores of predicted YCAY clusters, and phastCons conservation. Middle: The result of RT-PCR analysis comparing E18.5 WT and Nova dKO brains, each with three biological replicates. The alternative isoforms are indicated. Right: The quantitated results of RT-PCR. The black bars represent the longer isoform and the unfilled bars represent the shorter isoform. The error bars represent the standard error. When more than two isoforms are present, the two isoforms used for quantitation are indicated. The proportional splicing change ΔI and $\log_2$ fold change (long vs. short isoforms) are also indicated.

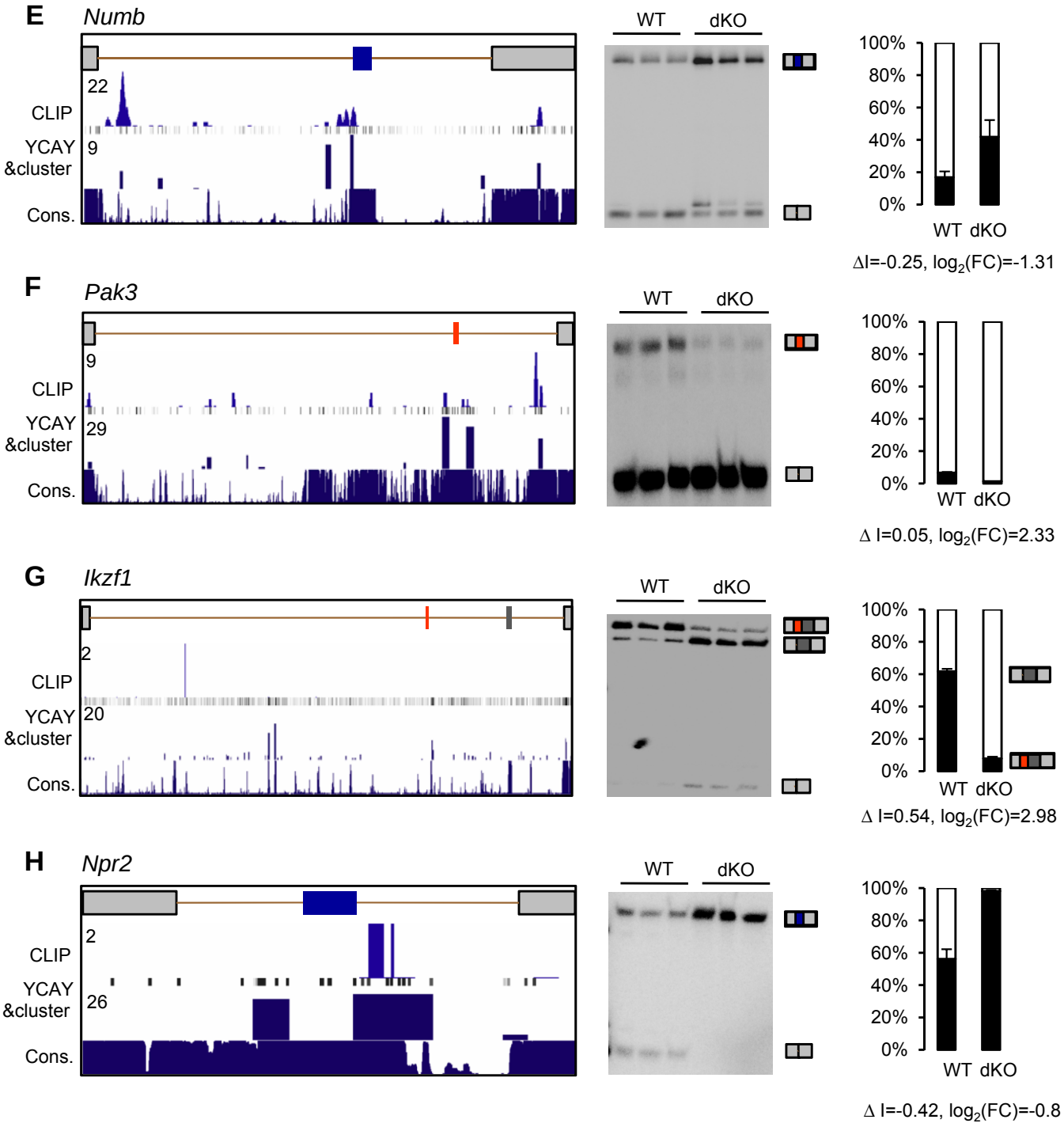


Figure S9. RT-PCR validation of Nova target exons predicted by Bayesian networks. (continued).

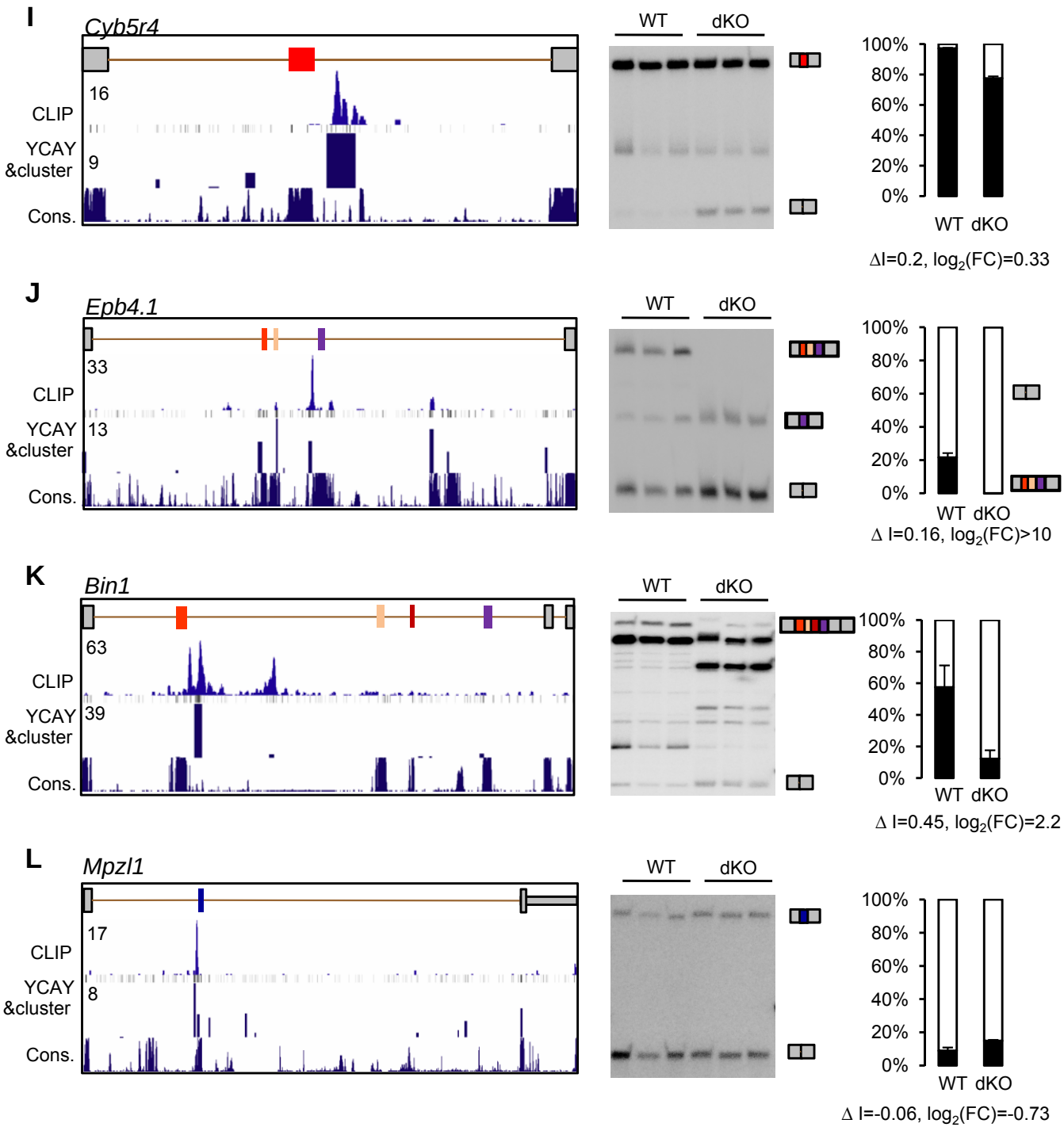


Figure S9. RT-PCR validation of Nova target exons predicted by Bayesian networks. (continued).

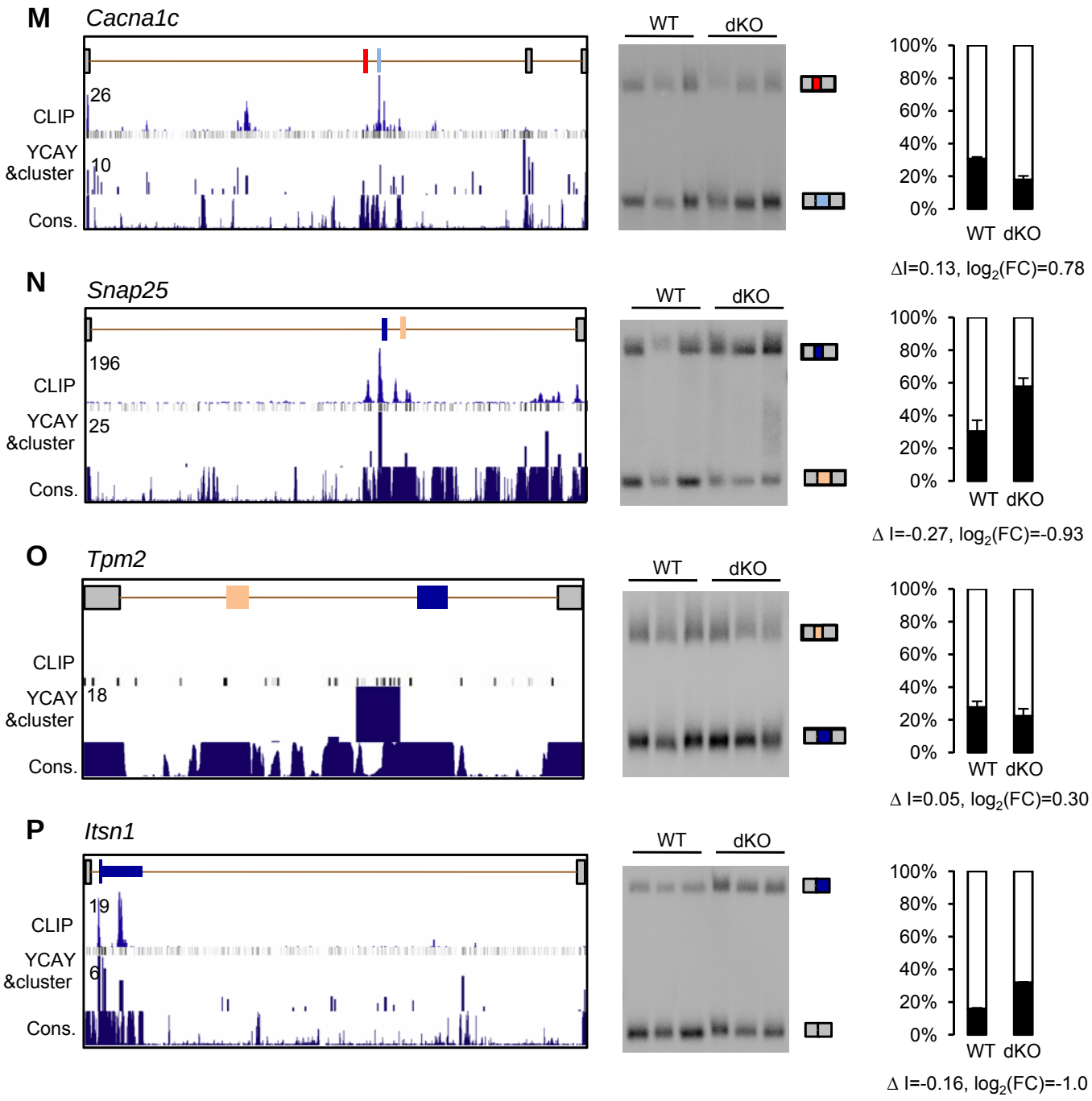


Figure S9. RT-PCR validation of Nova target exons predicted by Bayesian networks. (continued).

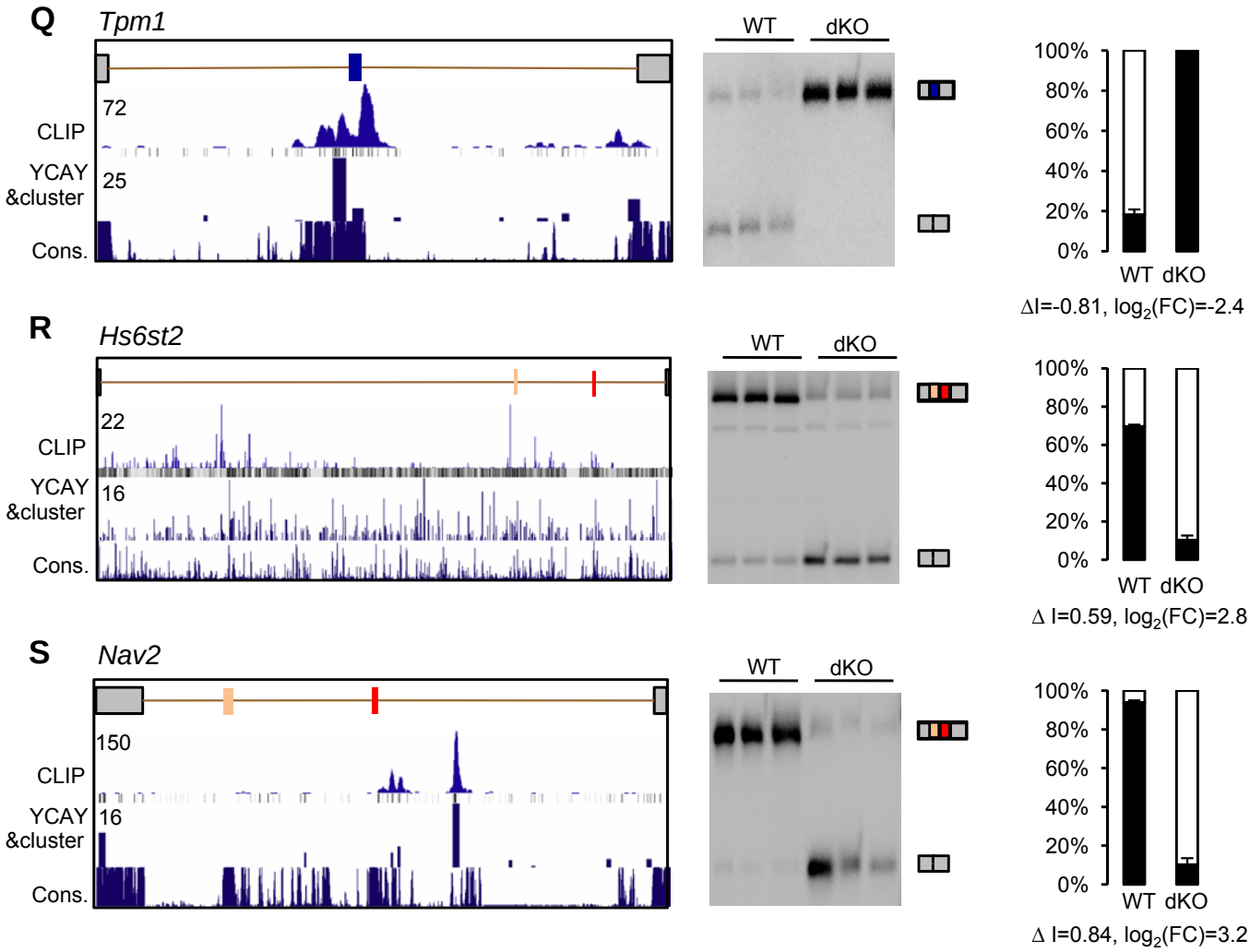


Figure S9. RT-PCR validation of Nova target exons predicted by Bayesian networks. (continued).

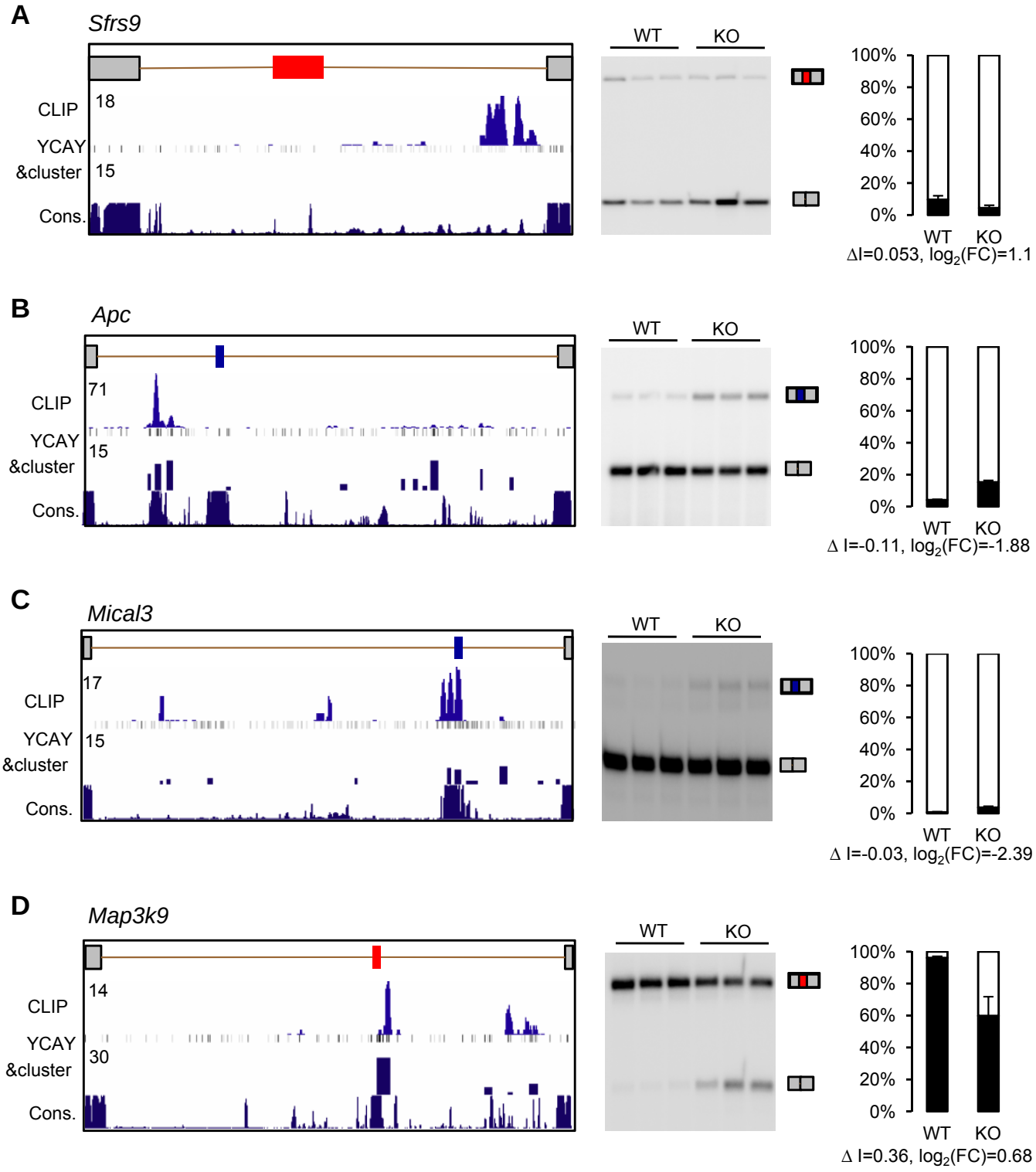


Figure S10. RT-PCR validation of novel exons regulated by Nova.
RT-PCR analysis was performed to compare P10 WT and Nova2 KO cortex. Three biologcal replicates were used for each group. See Fig. S9 legends for more details.

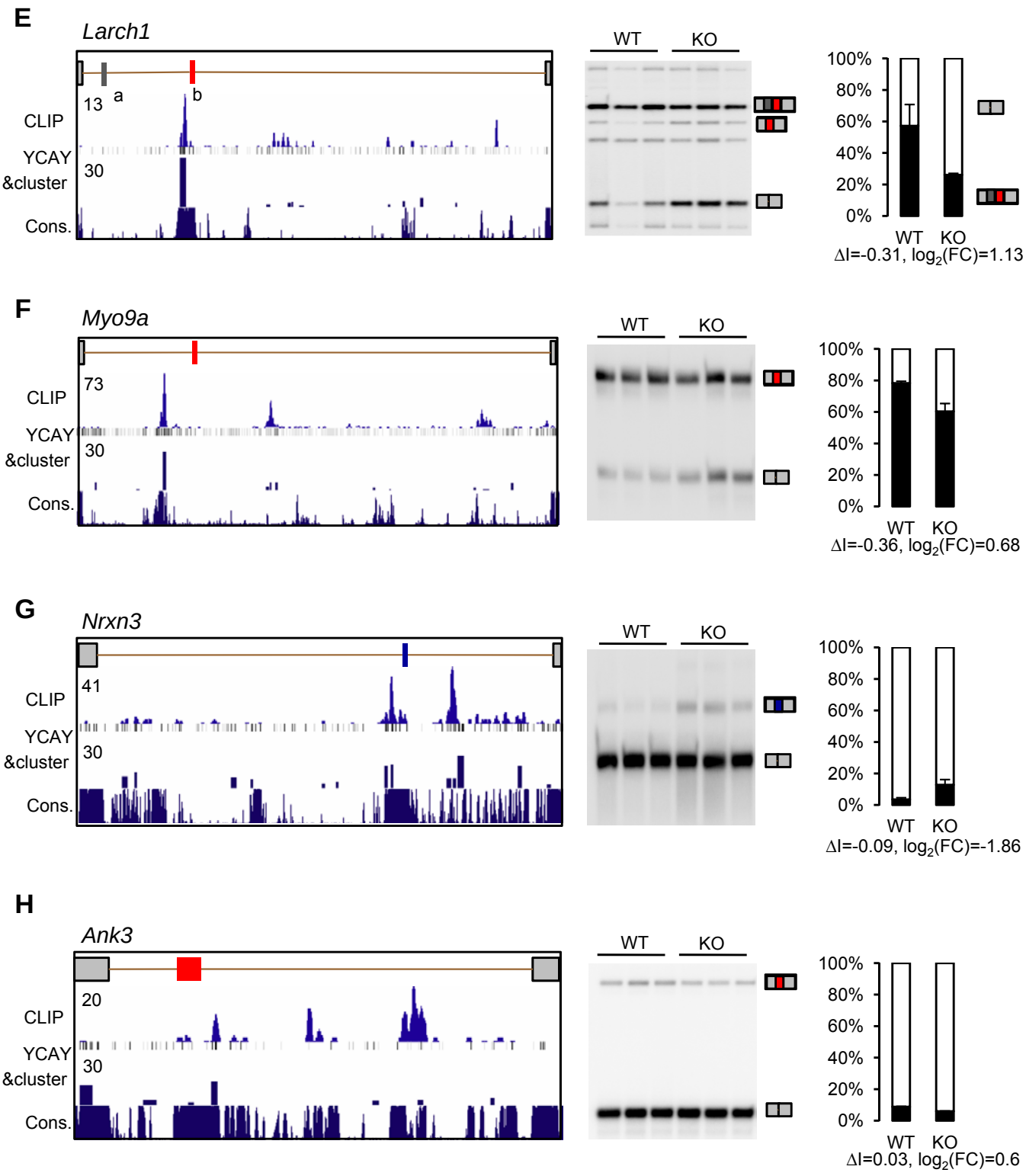
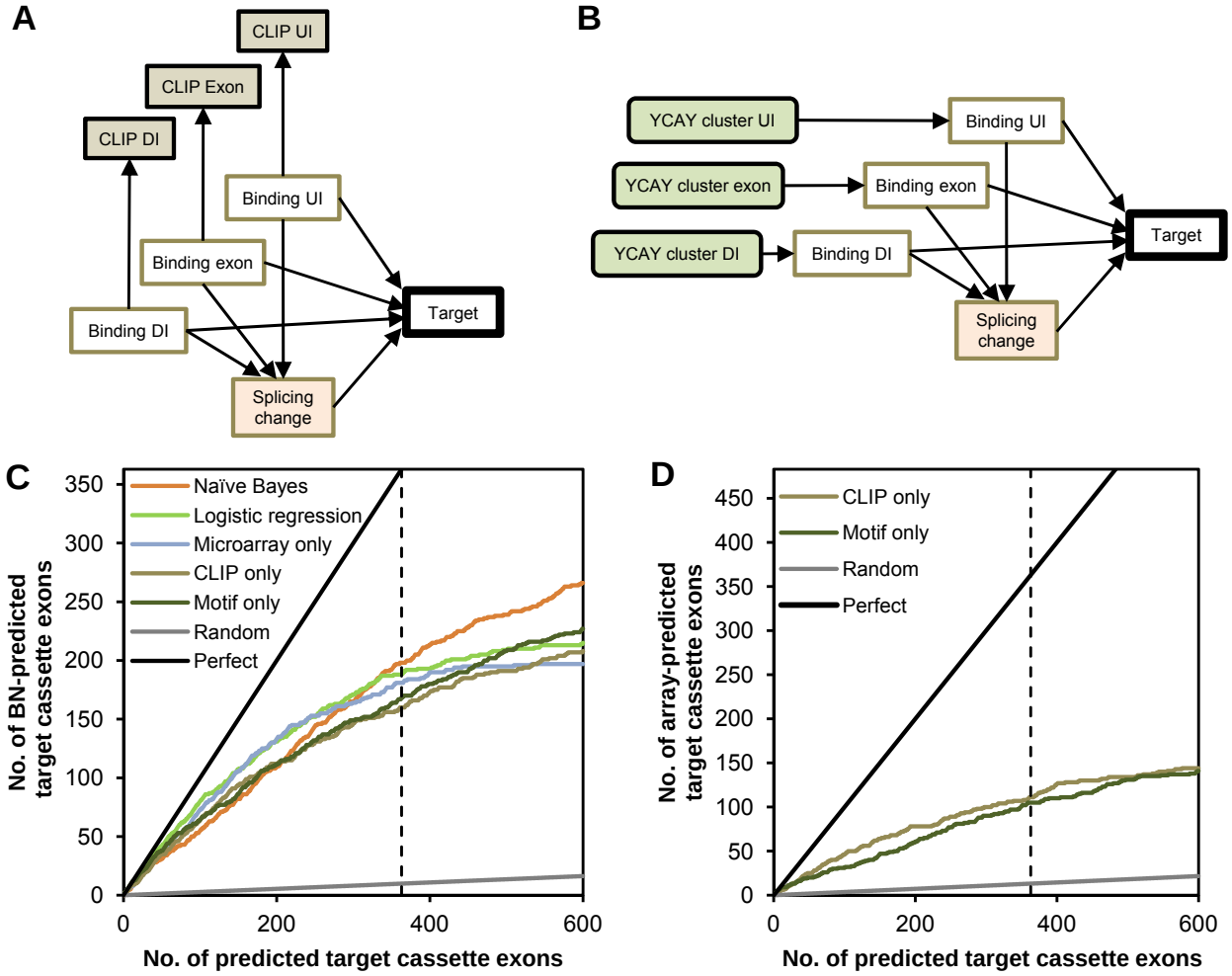


Figure S10. RT-PCR validation of novel exons regulated by Nova. (continued).



E

Methods	No. of predicted target cassette exons	Overlap with BN- predicted target cassette exons	Sensitivity	Validation rate
Byesian network (BN)	363	363	75-78%	90%
Naïve Bayes	363	198	61%	67%
Logistic regression	363	188	61%	65%
Microarray only	363	181	54%	61%
CLIP only	363	160	49%	54%
Motif only	363	168	49%	56%
Random	363	10	3%	3%

Figure S11. Prediction of Nova target exons using different datasets or methods
(**A**,**B**) Design of the reduced Bayesian network (BN) model used to predict Nova target cassette exons using only CLIP clusters (**A**) or YCAY clusters (**B**), respectively. (**C**) The overlap of 363 Nova target cassette exons predicted by the integrative Bayesian network with target cassette exons predicted by other machine learning methods (i.e. naïve Bayes and logistic regression, which combine microarray data, CLIP clusters, and YCAY clusters), or predictions from microarrays, CLIP clusters or YCAY clusters (motif) alone. For microarray-only predictions, one exon array dataset (E18.5 WT versus Nova dKO brains) was used to simplify exon ranking. The perfect overlap and that expected by chance are also shown for comparison. The dotted line indicates the overlap when 363 targets were predicted by each method. (**D**) Similar to (**C**) but the overlap between 483 cassette exons with Nova-dependent splicing in ≥ 1 microarray datasets and target cassette exons predicted using only CLIP clusters or YCAY clusters (motif) is shown. (**E**) Summary of sensitivity and validation rate of each method when 363 targets were predicted.

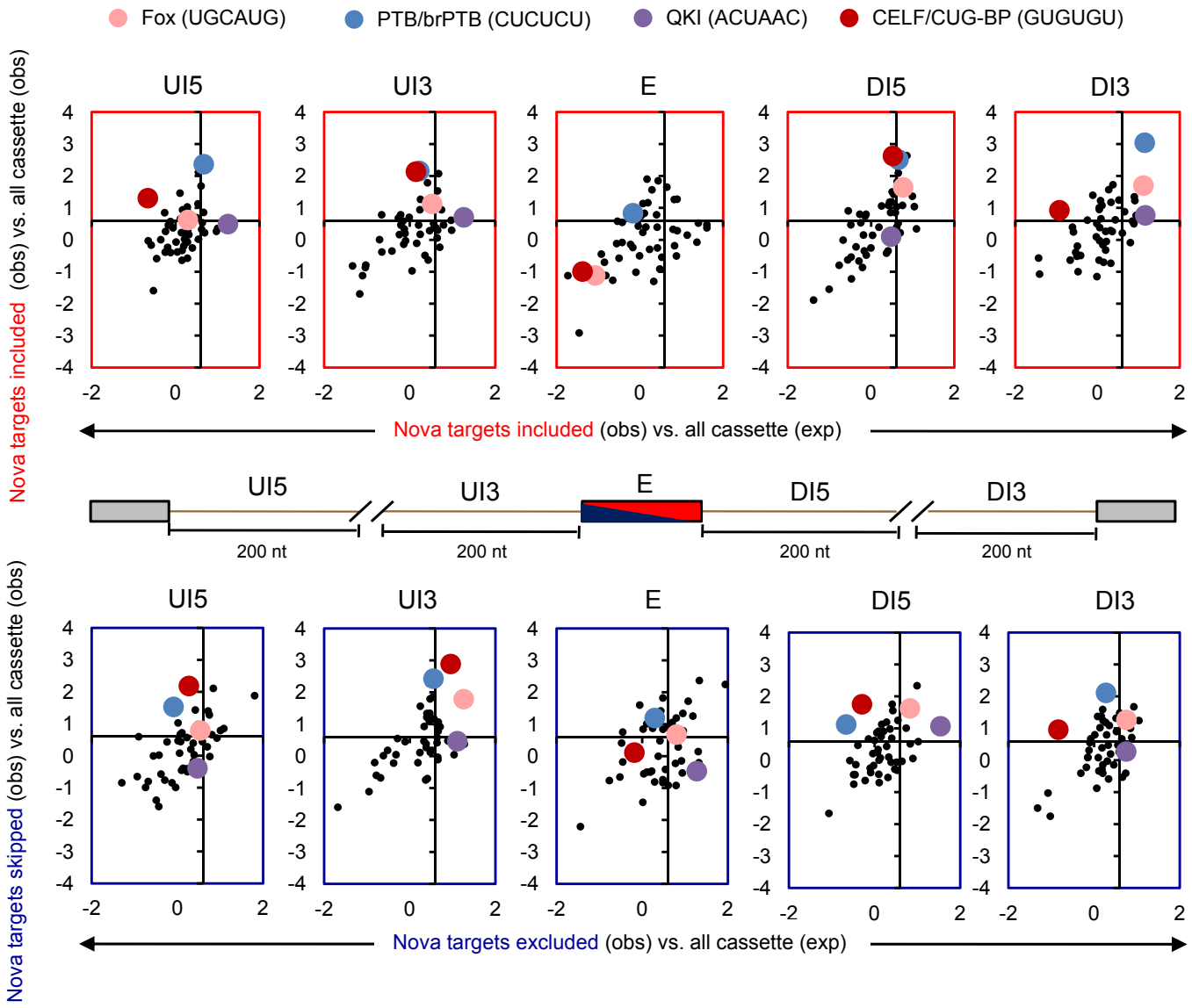


Figure S12. Enrichment of additional splicing-regulatory motifs in Nova-regulated exons. The hexamer motifs were derived from brain-specific exons previously by the Burge lab (S*38*). The scatter plots in the top and bottom panels represent cassette exons with Nova-regulated inclusion and exclusion, respectively. X-axis compares the frequency of each motif in Nova target cassette exons and flanking intronic sequences (200 nt from 5′ or 3′ splice site) with background frequency observed in corresponding regions of all cassette exons. Y-axis compares the frequency of each motif in Nova target exons and flanking intronic sequences with the background frequency in corresponding regions of all cassette exons expected from the base composition. The relative frequency is shown in $\log_2$ scale. An arbitrary threshold corresponding to 1.5 fold change is indicated. Motifs of several known splicing factors are highlighted.

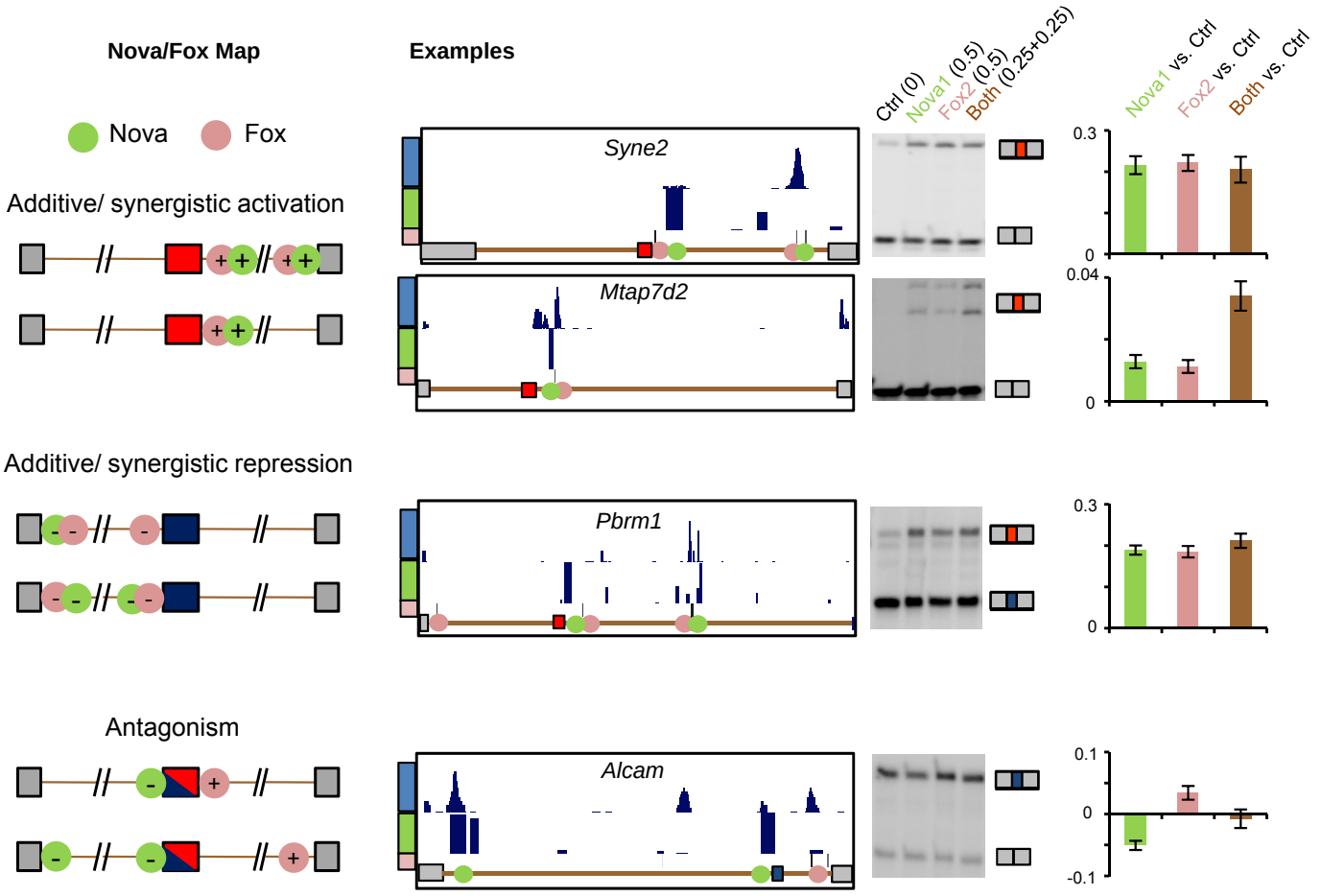


Figure S13. A combinatorial RNA-regulatory map of Nova and Fox.
Typical types of Nova and Fox binding, and the resulting splicing patterns are shown on the left of each panel. Nova and Fox binding in downstream introns activate exon inclusion (red) through an additive or synergistic action (top panel), whereas Nova and Fox binding in upstream introns or exons repress exon inclusion (blue, shown in the middle panel). Nova binding in the upstream intron and Fox binding in the downstream intron, or vice versa, have an antagonistic effect on exon inclusion (red and blue, shown in the bottom panel). Additional examples of each category not included in Fig. 3 of the main text are shown on the right (see Fig. 3 legend in the main text for more details).

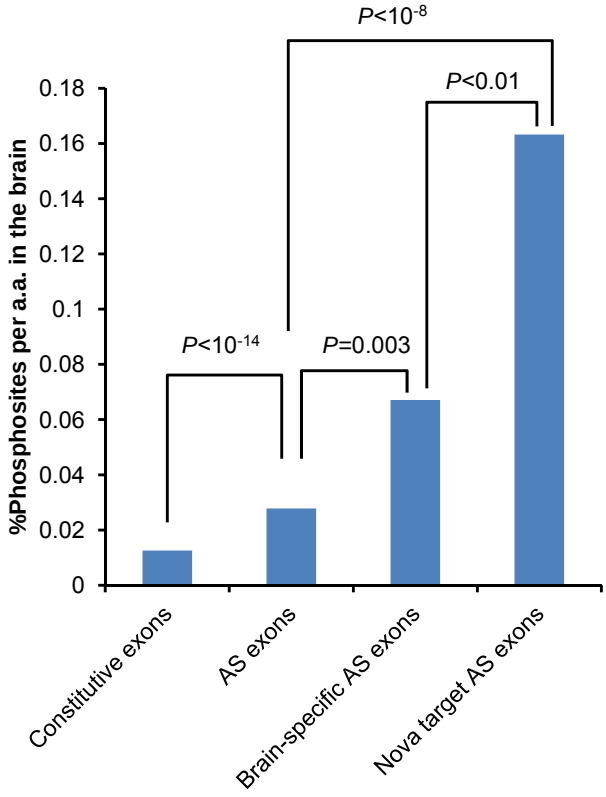


Figure S14. Phosphorylation sites experimentally determined in the brain are enriched in Nova target exons.

Percentage of phosphorylation sites per amino acid is shown for different groups of exons, and compared by a Fisher's exact test.

Supporting tables

Table S1. Summary of Nova CLIP experiments.

Dataset	Sample	Age	Antigen	Biological replicates	No. reads	Unique reads
1	Cortex*	P10	Nova1, Nova2	3	412,686	168,593
2	Cortex	P10	Nova2	1	1,831,677	411,336
3	Whole brain	P16	Nova1, Nova2	3	14,134,444	1,546,198
4	Cortex (+pilocarpine)	adult	Nova2	3	14,214,607	583,549
5	Cortex (+sham)	adult	Nova2	3	15,458,694	697,825
6	Hippocampus (+pilocarpine)	adult	Nova2	3	16,711,534	292,948
7	Hippocampus (+sham)	adult	Nova2	3	16,540,852	271,552
8	Cortex (nuclear)	P10	Nova2	1	1,927,248	429,527
			Total	20	81,231,742	4,401,528

* Sequenced by 454 (data from ref. (S5)).

Table S2. Summary of microarray datasets.

Data set	Comparison	Tissue	Age	Replicates per group	Array
1	WT vs. Nova2 KO	cortex	P10	4	Exon-junction Array
2	WT vs. Nova2 KO	forebrain	P10	4	Exon Array
3	WT vs. dKO	brain	E18.5	4	Exon Array
4	WT vs. dKO	spinal cord	E18.5	4	Exon Array

Table S3. Summary of nodes in the Bayesian network for cassette exons.

Node ID	Node Name	Possible values	Parent Node	CPD	No. parameters
1	YCAY cluster UI	Continuous [0,+∞)	N/A	Root	0
2	YCAY cluster exon	Continuous [0,+∞)	N/A	Root	0
3	YCAY cluster DI	Continuous [0,+∞)	N/A	Root	0
4	Binding UI	0=No binding 1=Binding	YCAY cluster UI	Logistic	2
5	Binding exon	0=No binding 1=Binding	YCAY cluster exon	Logistic	2
6	Binding DI	0=No binding 1=Binding	YCAY cluster DI	Logistic	2
7	CLIP UI	Discrete [0,+∞)	Binding UI	Negative binomial	2×2=4
8	CLIP exon	Discrete [0,+∞)	Binding exon	Negative binomial	2×2=4
9	CLIP DI	Discrete [0,+∞)	Binding DI	Negative binomial	2×2=4
10	Splicing change	-1=Exclusion 0=No effect 1=Inclusion	Binding UI Binding exon Binding DI	Tabular	(2×2×2)×3=24
11	Junction array dataset	Continuous [-1,1]	Splicing change	Normal	2×3=6
12	Exon array dataset 1	Continuous [-∞,+∞]	Splicing change	Normal	2×3=6
13	Exon array dataset 2	Continuous [-∞,+∞]	Splicing change	Normal	2×3=6
14	Exon array dataset 3	Continuous [-∞,+∞]	Splicing change	Normal	2×3=6
15	Reading frame	0=Frame-shift 1=Frame-preserving	Splicing change	Tabular	3×2=6
16	AS conservation	0=Non-conserved 1=Conserved	Splicing change	Tabular	3×2=6
17	Target	-1=Exclusion 0=No effect 1=Inclusion	Binding UI Binding exon Binding DI Splicing change	Deterministic*	0
				Total No. of parameters:	78

UI and DI represent upstream and downstream introns, respectively.

* Deterministic CPD for the node “Target”:

$$P(\text{Target} \mid \text{Binding}, \text{Splicing change}) = \begin{cases} 0 & \text{if Splicing change}=0 \\ & \text{or } \sum \text{Binding}=0 \\ \text{Splicing change} & \text{other wise} \end{cases}$$

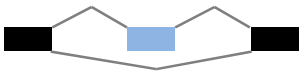
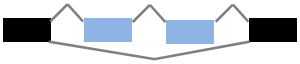
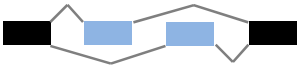
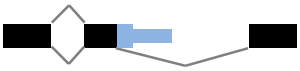
Table S4. Novel exons regulated by Nova.

Symbol	Exon/Intron coordinates (mm9)	Strand	Reading frame preservation	CLIP	YCAY	Region*	Method
Grik2 (known)	chr10:48823320-48823454	-	1	13	33.50	Intron	Exon Array, CLIP/ YCAY clusters
Slc4a3	chr1:75543974-75544518(?)	+	?	35	30.30	Exon	Exon Array
Agrn	chr4:155551931-155552115(?)	-	?	10	7.10	UI 5´SS	Exon Array
Sfrs9	chr5:115,778,355-115,778,758(?)	+	?	18		DI 3´SS	Exon Array
Apc	chr18:34460916-34460969	+	1	71	8.40	UI 3´SS	Exon Array
Mical3	chr6:120903652-120903702	-	1	17	4.40	UI 3´SS /Exon	Exon Array
Hrb	chr1:82877322-82877393	+	1	18	14.50	UI 3´SS of UE	Exon Array
Epb4.1l2	chr10:25210516-25212115(?)	+	NA	31	19.07	Exon	Exon Array
Sorbs1	chr19:40373953-40374143	-	1	10	34.20	Exon ALT	Exon Array
Centg3	chr5:23985941-23986166	+	0	79	5.62	DI 5´SS	Exon Array
Ccdc136	chr6:29371991-29372097	+	0	23	7.34	UI 3´SS	Exon Array
Wnk1	chr6:119879844-119879885	-	1	6	8.37	DI 5´SS	Exon Array
Rlbp1l1	chr4:9196499(?)-9196881	+	0	1	33.50	Exon ALT	Exon Array
Map3k9	chr12:82828318-82828386	-	0	3	8.33	UI 3´SS	Exon Array
Cadm3	chr1:175285131-175285358	-	1	21	46.87	Exon	CLIP/YCAY clusters
Stxbp1	chr2:32645519-32645597	-	0	44	30.59	UI 3´SS	CLIP/YCAY clusters
Lrch1	chr14:75192934-75193038	-	1	13	27.76	UI 3´SS	CLIP/YCAY clusters
Ppfia2	chr10:106272233-106272319	+	1	15	26.08	UI 3´SS	CLIP/YCAY clusters
Mtap2	chr1:66457148-66457393	+	1	145	25.88	UI 3´SS	CLIP/YCAY clusters
Myo9	chr9:59760344-59760397	+	1	73	21.21	Exon	CLIP/YCAY clusters
Matr3	chr18:35732678-35732734	+	1	25	20.50	DI 5´SS	CLIP/YCAY clusters
Ank3	chr10:69448837-69456510	+	1	75	20.36	Exon	CLIP/YCAY clusters
Pbx1	chr1:170287875-170288065	-	0	14	19.16	Exon	CLIP/YCAY clusters
Map2k5	chr9:63077898-63078011	-	1	14	19.02	Exon	CLIP/YCAY clusters
Cabin1	chr10:75172193-75172328	-	0	59	18.60	UI 3´SS	CLIP/YCAY clusters
Slc23a2	chr2:131926100-131926323	-	0	11	18.05	UI 3´SS	CLIP/YCAY clusters

Nrxn3	chr12:90430225-90430237	+	0	41	17.34	Exon	CLIP/YCAY clusters
Stim1	chr7:109581711-109581747	+	0	16	17.13	UI 3´SS	CLIP/YCAY clusters
5730419I09Rik	chr6:143017081-143017266	-	1	106	16.30	Exon	CLIP/YCAY clusters
Prdm2	chr4:142735293-142735349	-	1	24	15.25	Exon	CLIP/YCAY clusters
Ptprd	chr4:75746351-75746656	-	1	28	15.03	UI 3´SS	CLIP/YCAY clusters
Foxp1	chr6:98890989-98891158	-	0	56	14.98	DI 5´SS	CLIP/YCAY clusters
Wnk1	chr6:119882687-119882779	-	1	13	14.70	DI 5´SS	CLIP/YCAY clusters
Ryr2	chr13:11708828-11708845	-	1	27	14.62	Exon	CLIP/YCAY clusters
Gtl2	chr12:110805800-110805806	+	0	620	14.23	Exon	CLIP/YCAY clusters
Sorbs1	chr19:40384071-40384823	-	1	28	13.92	Exon	CLIP/YCAY clusters
Rims2	chr15:39386611-39386643	+	1	11	13.39	Exon	CLIP/YCAY clusters
Slc25a36	chr9:96989910-96990012	-	0	11	13.20	Exon	CLIP/YCAY clusters
Memo1	chr17:74619828-74619981	-	0	21	12.74	Exon	CLIP/YCAY clusters
C230096C10Rik, Ubr	chr4:138958424-138958583	+	0	10	12.71	Exon	CLIP/YCAY clusters
Sbf2	chr7:117484804-117484878	-	1	10	12.60	DI 5´SS	CLIP/YCAY clusters
Aof2	chr4:136119353-136119364	-	1	12	12.60	DI 5´SS	CLIP/YCAY clusters
Nrxn3	chr12:90429597-90429620	+	1	34	12.39	UI 3´SS	CLIP/YCAY clusters
Ank3	chr10:69419416-69419514	+	1	10	12.28	DI 5´SS	CLIP/YCAY clusters
Slc4a10	chr2:62077683-62077744	+	0	49	12.03	Exon	CLIP/YCAY clusters
Clip1	chr5:124063077-124065344	-	1	10	11.39	UI 3´SS	CLIP/YCAY clusters
Fnbp1	chr2:30903725-30903808	-	1	16	10.75	Exon	CLIP/YCAY clusters
Ppp1r12a	chr10:107686012-107686125	+	1	52	10.69	Exon	CLIP/YCAY clusters
Mtap2	chr1:66454160-66454396	+	1	40	10.54	Exon	CLIP/YCAY clusters
Tanc2	chr11:105620236-105620346	+	1	14	10.20	UI 3´SS	CLIP/YCAY clusters

* UI: upstream intron; DI: downstream intron; Exon ALT: exonic region between the two alternative splice sites; UE: upstream exon; SS: splice site (5´ or 3´ as indicated).

Table S5. Summary of Nova-regulated alternative splicing events.

AS type	AS diagram	No. AS events
Bayesian network predictions:		
CASS		363
TACA		141
MUTX		37
ALT5		9
ALT3		9
APA5		13
APA3		16
	Bayesian network total	**588**
Others:		
	Validated	29
	Robust splicing change	27
	Novel exons	49
	Total	**698**

In each diagram, alternatively spliced exons are shown in blue and constitutive exons are shown in black. CASS: cassette exons; TACA: tandem cassette exons; MUTX: mutually exclusive exons; ALT5 and ALT3: alternative 5′ and 3′ splice sites; APA5 and APA3: alternative polyA usage coupled with 5′ or 3′ splice site choices.

Table S6. Nova target exons with functional characterizations in the literature.

Gene symbol[†]	Coordinates (mm9)	In/Ex[‡]	Phospho-rylation [¶]	Description	Ref.
Agrn (known)	chr4:155,541,391-155,542,929	In	indirect	Exon Z is critical for aggregation of postsynaptic proteins at neuromuscular junction (NMJ), which requires the receptor-tyrosine kinase MuSK. The conserved tripeptide asparagineglutamate-isoleucine in the exon is necessary and sufficient for full MuSK phosphorylation activity.	(S*53-55*)
Ank3 (known)	chr10:69,461,615-69,468,404	Ex		Alternative 3´ splice sites change a region of 588 nt (insert C), encoding Ser/Thr-rich amino acids in the regulatory domain. This insert has the highest expression in skin and reduced level in kidney, brain and testis.	(S*56, 57*)
Atp2b1 (known)	chr10:98,481,349-98,485,478	In	indirect	This exon (154 nt, near 3´ end) overlaps with the calmodulin-binding domain and thus might modulate calmodulin binding by alternative splicing. Inclusion of the exon also shifts the site of cAMP-dependent phosphorylation and likely alters the regulatory behavior of the isoform. Cryptic donor sites inside the exon also produce additional isoforms.	(S*58*)
Atp2b2 (known)	chr6:113,745,490-113,756,289	Ex		The three tandem cassette exons, located in the intracellular loop separating membrane-spanning segments 2 and 3, can be inserted in different combinations that vary in different tissues. Different isoforms differ in ATPase activity.	(S*59-62*)
Bin1 (validated)	chr18:32,584,474-32,591,747	In		Bin1 encodes a tumor suppressor. Exons12A-D are brain-specific. Exon 12A is not included in normal nonneruronal tissues, but is included in tumor cell lines. Aberrant inclusion of the exon in melanoma cells abolishes Bin1 to inhibit malignant transformation by c-Myc or adenovirus E1A and to induce programmed cell death. The isoform IIb that lacks exon 12C (24 nt) and D (108 nt) affects the interaction with clathrin, which is potentially important for synaptic vesicle endocytosis.	(S*63-65*)
Cacna1c (validated)	chr6:118,580,417-118,602,076	In		The two mutually exclusive exons overlap with critical transmembrane segments of the Ca^{2+} channels, with differential expression in CNS. Both isoforms are equally expressed in newborn and fetal rat heart and only a single isoform (the 3´ exon) is predominant in adult rat heart.	(S*66-68*)
*Camk2b**	chr11:5,879,676-5,882,575	Ex	kinase	The isoforms differ in the autonomy generated by Ca^{2+} oscillations.	(S*69*)
Cyb5r4 (validated)	chr9:86,950,600-86,953,910	In		Exon 12 codes for the flavin-adenine dinucleotide binding domain of cb5/cb5r and the protein lacking this exon may function in a dominant negative fashion in limiting the amount of damage caused by the production of reactive oxygen species by cb5/cb5r.	(S*70*)
Dcc (validated)	chr18:71,538,309-71,543,877	In		Dcc is a receptor for netrin-1. It is implicated as a tumor suppressor, with frequent allelic losses in colorectal cancer. Alternative splicing of the exon is thought to introduce a hinge in the extraceullular domain.	(S*71*)
Dclk1 (known)	chr3:55,320,795-55,337,673	In	indirect; kinase	Inclusion of exon 19 causes a frameshift of exon 20 and results in a truncated protein in the C terminus, which has reduced autophosphorylation activity. This exon is also regulated during development, with more inclusion in adult brain.	(S*72, 73*)

*Dlg1**	chr16:31,853,929-31,856,577	In		This exon (insert I4), one of the four alternatively spliced exons in the region, is specifically included in brain and liver. Alternative splicing in this region is involved in the localization of Dlg1 to cell-cell contact although the specific role of the insert I4 was not assessed.	(S*74*)
Epb4.1 (validated)	chr4:131,513,558-131,523,890	In	direct	Three consecutive alternative exons in this region overlapping with the spectrin-actin binding (SAB) domain show complex tissue-specific splicing patterns and may mediate red cell membrane mechanical stability and deformability. In rat brain, the predominant isoform includes all the three exons, and is required for fodrin-actin-4.1R ternary complex formation, which might be essential for the shape and membrane integrity of neural cells. Phosphorylation of a tyrosine on one of the exons by EGFR reduced the ability of protein 4.1 to promote the assembly of the complex.	(S*75-77*)
Epb4.1l2 (validated)	chr10:25,208,757-25,222,193	In	direct	The four consecutive alternative exons are paralogous to the alternatively spliced region of *Epb4.1* (4.1R) encoding the spectrin-actin binding (SAB) domain, which can induce fodrin-actin complex formation. An additional isoform uses a novel 3´UTR, resulting in a truncated product without the SAB domain.	(S*76, 78*)
Epb4.1l3 (known)	chr17:69,611,466-69,624,235	In	direct	The three consecutive alternative exons are paralogous to the alternatively spliced region of Epb4.1 (4.1R) encoding the spectrin-actin binding (SAB) domain. However, the encoded proteins cannot induce fodrin-actin complex formation.	(S*76, 78*)
*Fnbp1**	chr2:30,895,925-30,909,738	In	direct	The fully included isoform (RapostlinL) is predominantly expressed in brain whereas isoforms with complete or partial exclusion of the exon (RapostlinM and RapostlinS) are ubiquitous. The insert region is important for neurite branching.	(S*79*)
Gabrg2 (known)	chr11:41,725,284-41,729,991	In	direct	Inclusion of exon 9 adds an additional eight amino acids to an intracellular loop of the protein, and generates a site which can be phosphorylated by PKC. Mice engineered to only express the short form of the $GABA_AR\gamma2$ subunit lacking the exon display a higher level of anxiety and increased sensitivity to benzodiazapines than control.	(S*80*)
*Glra2** (known)	chrX:161,719,755-161,762,566	In		The switch of the two exons are under strong regulation in developing neurons.	(S*52, 81*)
Gnas	chr2:174,159,713-174,167,208	-	probably direct	This region includes potential phosphorylation sites for protein kinase A.	(S*82*)
*Grik1**	chr16:87,896,142-87,914,649	In	direct	The exon encodes an ER retention signal, which controls receptor trafficking in both heterologous cells and neurons. The ER retention motif consists of a critical arginine (Arg-896) and surrounding amino acids, which, if disrupted, promote ER exit and surface expression of the receptors, as well as alter their physiological properties. The Arg-896-mediated ER retention is regulated by a mutation that mimics phosphorylation of Thr-898, but not by PDZ interactions.	(S*83*)
Grin1 (known)	chr2:25,165,895-25,169,124	Ex	indirect	The exon encodes amino acids in the extracellular N-terminal domain. Use of the exon lowers the affinity of NR1 receptors to NMDA and increases potentiation by protein kinase C. The splicing of the exon is also regulated developmentally, with the lowest inclusion at embryonic day E19. It reaches peak expression at P14 in hippocampus and cerebellum.	(S*84-86*)
*Grin1** (known)	chr2:25,146,705-25,151,457	In	direct	Alternative splicing of exon 21 deletes the original stop codon and causes a switch of the 3´ ends, which is activity-dependent. This switch will change the recruitment of nascent NMDARs to ER exit sites mediated through the divaline motif encoded in the 3´ end of the transcript.	(S*87-89*)

Hs6st2 (validated)	chrX:48,742,897-49,033,585	In		The two exons are specifically included in brain, which inserts 40 amino acids and results in a long isoform that differs from the shorter isoform in preferences for sulphation sites in HS substrates.	(S*90*)
Ikzf1 (validated)	chr11:11,607,788-11,654,177	In		This gene encodes a B- and T-cell transcription factor implicated in leukemia. Alternative splicing of exons 3-6 in this region generates isoforms with deletion in zinc finger domains, which dramatically affects DNA-binding specificity of the protein and plays a dominant negative role in the lymphoid pathway. The dominant negative isoform IK6 is also expressed in pituitary tumors with a predominant cytoplasmic localization which can reverse the effect of the DNA-binding nuclear form IK1 on the FGFR4 promoter.	(S*91-93*)
Itsn1 (validated)	chr16:91,869,192-91,888,915	Ex		This gene encodes a protein involved in clathrin-mediated encocytosis. Alternative splicing of the region generates a short isoform and a long isoform with different domain architectures. The ratio of the two forms is developmentally regulated, with the short form ubiquitously expressed and the long form mainly in neurons. The long form functions as a guanine nucleotide exchange factor for Cdc42 and modulates actin cytoskeleton.	(S*94, 95*)
Kcnma1 (validated)	chr14:24,205,482-24,255,248	Ex	direct	BK channel subunits are derived from a single gene with extensive alternative splicing. This exon is included in the STREX variant, and phosphorylation of a PKA consensus site encoded by the exon mediates the inhibition of the channel, whereas a downstream conserved C-terminal PKA consensus motif mediates activation.	(S*96*)
Ktn1	chr14:48,344,852-48,351,148	-		Change a coil-coiled region with multiple heptad repeats.	(S*97*)
Magi1 (validated)	chr6:93,651,284-93,658,214	In	phospha-tase	The exon encodes the alpha segment between PDZ2 and PDZ3, which is included prevalently in brain.	(S*98*)
Map4k4 (known)	chr1:40,067,402-40,071,068	Ex	kinase	Map4k4 is overexpressed in many tumor cell lines. The exon (M7) encodes a serine-rich sequence.	(S*99*)
Mpzl1 (validated)	chr1:167,522,373-167,534,873	Ex	direct	The PZR1b variant lacking this exon does not have the immunoreceptor tyrosine-based inhibitory motifs and the ability to recruit tyrosine phosphatase SH2. In human HT-1080 cells, it also shows a dominant negative effect by blocking ConA induced tyrosine phophorylation of full length PZR and recruitment of tyrosine phosphatase SHP-2.	(S*100*)
Nav2 (validated)	chr7:56,806,967-56,812,440	In	direct	Exons 16 and 17; *Nav2* is a homolog of the *C. elegans unc-53* and a member of neuron navigators involved in axon guidance. Nav2 hypomophic mice show sensory defects.	(S*101, 102*)
*Nova1** (known)	chr12:47,801,467-47,821,843	Ex	direct	This exon is autoregulated by Nova1 through a negative feedback loop. The exon encodes 24 amino acids between KH1 and KH2 domains, and can be phosphorylated in vivo, by kinases including GSK3.	(S*32*)

Npr2 (validated)	chr4:43,655,119-43,657,368	Ex	direct	Exon 9 overlaps three out of six putative phosphorylation sites (Ser 523, Ser 526, Thr 529) and the putative ATP-binding motif. The region encoded by exon 9 is indispensable in C-type natriuretic peptide (CNP) induced activation of GC-B (Guanyl cyclase-B), although this region is not critical for either basal GC activity or CNP binding. Because of the dominant negative phenotype of the isoform lacking exon 9 (forming complexes with isoform containing exon 9) it is possible that CNP/GC-B signalling can be regulated by relative levels of various isoforms, adjusting the magnitude of a CNP signal in brain. Mutagenesis and comigration studies in 293 cells using synthetic phosphopeptides identified phosporylation in five residues (Ser-513, Thr-516, Ser-518, Ser-523, and Ser-526) within the kinase homology domain. Elimination of all of the phosphorylation sites resulted in a completely dephosphorylated receptor whose CNP-dependent cyclase activity was decreased by >90%, indicating that phosphorylation of the kinase homology domain is a critical event in the regulation of NPR-B.	(S*103, 104*) (S*104*).
Numb (validated)	chr12:85,136,389-85,140,607	Ex		This exon overlaps with a proline rich region (PRR). Inclusion of the exon keeps low throughout early rat neuronal development, peaks at E10 and decreases thereafter. In P19 cells, Inclusion of the exon promotes cell proliferation whereas skipping of the exon promotes differentiation during mammalian neurogenesis. Only the exon-skipped isoform mediates neuronal cell fate choice in *Drosophila*.	(S*105*)
Pak3 (validated)	chrX:140,144,191-140,149,832	In	kinase	p21-activated kinases (PAK) are involved in the control of cyotoskeleton dynamics and cell cycle progression. Inclusion of the exon overlapping with the autoinhibitory domain generates a variant termed PAK3b that displays a high kinase activity in starved cells that is not further stimulated by active GTPases. The 15-amino-acid insertion by the inclusion of the exon within the autoinhibitory domain impedes the ability of PAK3b to bind to the GTPases Rac and Cdc42 and changes its specificity toward the GTPases.	(S*106*)
*Sh2b1** (known)	chr7:133,610,508-133,612,155	Ex	indirect?	Alternative splicing of exons 7-9 introduces frame-shifts in four isoforms with distinct C-terminal domains. All variants are phosphorylated on tyrosine specifically in response to IGF-1 and PDGF stimulation. cDNA expression of the four variants caused variant-dependent levels of stimulation of IGF-I and PDGF-induced mitogenesis.	(S*107*)
Smtn	chr11:3,417,530-3,421,971	-		The exon overlaps with smoothlin CH-domain.	(S*108*)
Snap25 (validated)	chr2:136,589,309-136,599,756	Ex		SNAP25 is a SNARE protein contributing to the formation of the exocytotic fusion complex in neurons. Mutually exclusive splicing of the two exons is regulated during brain development and modifies the organization and sequence context of the central Cys quartet. These exons provide sites for fatty acid acetylation. Developmental switch of the two exons is essential for postnatal viability. Rescue experiments using specific isoforms in SNAP25 knockout mice suggested that the two isoforms differ in their ability to stabilize synaptic vesicles in the primed state.	(S*109-111*)
*Sorbs1**	chr19:40,373,953-40,386,353	Ex	indirect	Exon 31 is between the second and the third SH3 domains. Yeast two-hybrid experiments suggest that inclusion of this exon together with exon 32a (extended version of exon 32 by alternative 3´ splice sites) is important for Sorbs1 to function as an A-kinase anchoring protein (AKAP) that serves as scaffolds for multimolecular protein complexes containing PKA, and targets PKA to the insulin receptor, lipid rafts, actin stress fibers, or focal adhesions.	(S*112*)
*Stau2****	chr1:16,221,084-16,336,092	Ex		Stau2 is mainly expressed in the brain and involved in mRNA transport in neurons. Alternative 3´UTR usages generate distinct isforms Stau2 (62/59) vs. Stau2 (52) that alters the dsRBD5.	(S*113*)

Tpm1 (validated)	chr9:66875782-66878905	Ex	The cassette exon is near the 3′ end of the striated muscle form. The exon encodes amino acids that are likely to specify the troponin T site A binding domain.	(S*114*)
Tpm2 (validated)	chr4:43531274-43532169	Ex	Exons 6 and 7 are mutually exclusive. Exon 7 is specifically included in muscle, due at least in part to the repression of PTB. PTB interacts with critical cis-regulatory sequences upstream of exon 7 and blocks the usage of the exon in non-muscle cells. The predicted YCAY cluster is in a region that is important for repression of exon 7 as demonstrated by mutagenesis.	(S*115, 116*)
*Tpm3** (known)	chr3:89,891,596-89,893,781	Ex	Mutually exclusive splicing generates two isoform Tm5NM-1 and Tm5NM-2. Tm5NM-2 is sorted specifically to the Golgi complex, whereas Tm5NM-1, which differs by a single alternatively spliced internal exon, is incorporated into stress fibers.	(S*117*)

† Exons without asterisk were obtained from AEDB (*S30*). Additional exons reported in the literature identified by manual searches are indicated by an asterisk. Previously validated Nova target exons are labeled as “known” in the first column, whereas exons validated in this study are labeled as “validated”.

‡ In: Nova-dependent exon inclusion; Ex: Nova-dependent exon exclusion

¶ direct: there are phosphorylation sites encoded by the alternative exon so that alternative splicing can directly affect the availability of the phosphorylation sites; indirect: there are no known phosphorylation sites encoded by the alternative exon, but alternative splicing affects the regulation of phosphorylation indirectly according to the literature. Kinase or phosphatase indicates whether the gene encodes a kinase or phosphatase, respectively.

Table S7. Gene ontology analysis of Nova target genes.

GO Term	Gene count	%	P-Value	Fold Enrichment	Benjamini FDR
Biological process					
GO:0016043~cellular component organization	93	26.05	4.02E-12	2.02	6.93E-09
GO:0007399~nervous system development	53	14.85	1.01E-09	2.48	8.72E-07
GO:0032989~cellular component morphogenesis	31	8.68	2.32E-09	3.56	1.33E-06
GO:0030030~cell projection organization	30	8.40	3.55E-09	3.60	1.53E-06
GO:0007268~synaptic transmission	22	6.16	8.39E-09	4.64	2.90E-06
GO:0048667~cell morphogenesis involved in neuron differentiation	22	6.16	1.18E-08	4.55	3.40E-06
GO:0051179~localization	104	29.13	1.63E-08	1.66	4.03E-06
GO:0000902~cell morphogenesis	28	7.84	1.64E-08	3.57	3.53E-06
GO:0019226~transmission of nerve impulse	24	6.72	2.47E-08	4.01	4.73E-06
GO:0007154~cell communication	34	9.52	4.41E-08	2.93	7.60E-06
Cellular component					
GO:0045202~synapse	38	10.64	1.64E-15	4.85	5.00E-13
GO:0044459~plasma membrane part	82	22.97	2.06E-15	2.51	3.16E-13
GO:0042995~cell projection	47	13.17	3.24E-14	3.62	3.24E-12
GO:0030054~cell junction	42	11.76	3.53E-14	4.00	2.65E-12
GO:0005856~cytoskeleton	61	17.09	6.36E-12	2.60	3.81E-10
GO:0005886~plasma membrane	104	29.13	6.82E-11	1.84	3.41E-09
GO:0042734~presynaptic membrane	11	3.08	1.07E-09	14.80	4.61E-08
GO:0016323~basolateral plasma membrane	19	5.32	3.11E-09	5.83	1.17E-07
GO:0044456~synapse part	23	6.44	5.65E-09	4.55	1.88E-07
GO:0043005~neuron projection	25	7.00	8.93E-09	4.09	2.68E-07
Molecular function					
GO:0005515~protein binding	198	55.46	9.43E-15	1.49	4.50E-12
GO:0008092~cytoskeletal protein binding	35	9.80	2.57E-10	3.51	6.14E-08
GO:0003779~actin binding	23	6.44	9.83E-07	3.40	1.56E-04
GO:0030695~GTPase regulator activity	26	7.28	1.17E-06	3.06	1.39E-04
GO:0060589~nucleoside-triphosphatase regulator activity	26	7.28	1.61E-06	3.01	1.54E-04
GO:0005488~binding	270	75.63	2.82E-05	1.10	0.002242
GO:0005516~calmodulin binding	13	3.64	2.91E-05	4.48	0.001983
GO:0005083~small GTPase regulator activity	17	4.76	6.58E-05	3.25	0.003916
GO:0030234~enzyme regulator activity	31	8.68	2.04E-04	2.07	0.010742
GO:0008017~microtubule binding	9	2.52	2.04E-04	5.47	0.009681

The top 10 terms in each of the three gene ontology (GO) term categories are shown.

Table S8. KEGG pathways enriched in Nova target genes.

KEGG pathway	Gene count	Fold Enrichment	Benjamini FDR	Genes
mmu04020 Calcium signaling pathway	17	3.5	0.001	*Atp2b1, Atp2b2, Cacna1c, Cacna1d, Cacna1b, Cacna1g, Camk2a, Camk2g, Camk2b, Grin1, Gnas, Plcb4, Ppp3cb, Ppp3cc, Ryr2, Slc8a1, Erbb4*
mmu04720 Long-term potentiation	10	5.3	0.003	*Cacna1c, Camk2a, Camk2g, Camk2b, Gria2, Grin1, Plcb4, Ppp1r12a, Ppp3cb, Ppp3cc*
mmu04514 Cell adhesion molecules (CAMs)	12	4.1	0.003	*Alcam, Cadm1, Cadm3, Mpzl1, Neo1, Nrxn3, Nfasc, Nfasc, Ptprf, Ptprm, Nlgn1, Nrcam,Nrxn1*
mmu04520 Adherens junction	10	4.4	0.006	*Actn4, Baiap2, Ctnna2, Ctnnd1, Pard3, Smad2, Smad4, Ptprf, Ptprm,Sorbs1*
mmu04360 Axon guidance	13	3.3	0.006	*Ablim1, Cxcl12, Dcc, Epha5, Efna5, Ablim2, Ntng1, Pak3, Ppp3cb, Ppp3cc, Arhgef12, Robo2, Unc5c*
mmu04912 GnRH signaling pathway	10	3.6	0.017	*Cacna1c, Cacna1d, Camk2a, Camk2g, Camk2b, Gnas, Mapk8, Mapk9, Map2k4, Plcb4*
mmu04310 Wnt signaling pathway	12	2.8	0.032	*Apc, Camk2a, Camk2g, Camk2b, Smad2, Smad4, Mapk8, Mapk9, Plcb4, Porcn, Ppp3cb, Ppp3cc*
mmu04930 Type II diabetes mellitus	6	5.2	0.048	*Cacna1c, Cacna1d, Cacna1b, Cacna1g, Mapk8, Mapk9*
mmu04260 Cardiac muscle contraction	7	4.2	0.049	*Tpm2*, *Cacna1d*, *Cacna1c*, *Tpm1*, *Ryr2*, *Slc8a1*, *Tpm3*
mmu04012 ErbB signaling pathway	8	3.3	0.074	*Camk2a, Camk2g, Camk2b, Mapk8, Mapk9, Map2k4, Pak3, Erbb4*
mmu04530 Tight junction	9	2.7	0.01	*Actn4, Cask, Ctnna2, Pard3, Epb4.1, Epb4.1l1, Epb4.1l2, Epb4.1l3, Magi1*

Table S9. Nova target genes implicated in genetic diseases.

Symbol	Disease	Source	Neuronal Phenotype	Fox targets*
Actn4	Glomerulosclerosis, focal and segmental	HGMD, OMIM		E
Aff3	Mesomelic dysplasia	HGMD		G
Ank1	Spherocytosis	HGMD, OMIM		G
Ank2	Cardiac arrhythmia / Long QT syndrome / Long QT syndrome ?	HGMD, OMIM		G
Ap1s2	Mental retardation	HGMD, OMIM	Y	
Apc	Adenomatous polyposis coli / Adenomatous polyposis coli ? / Adenomatous polyposis coli and CHRPE / Adenomatous polyposis coli, association with ? / Adenomatous polyposis coli, attenuated / APC with desmoid tumour / Colorectal adenoma / Colorectal cancer / Colorectal cancer, predisposition to, association / Desmoid tumours / Hepatoblastoma / Juvenile polyposis coli / Multiple adenomas / Thyroid cancer	HGMD, OMIM		
Arhgap26	JUVENILE MYELOMONOCYTIC LEUKEMIA; JMML	OMIM		
Arhgef12	Increased insulin sensitivity, association with	HGMD, OMIM		E
Arhgef6	Mental retardation, X-linked	HGMD, OMIM	Y	E
Arhgef9	Hyperekplexia and epilepsy	HGMD, OMIM	Y	
Arl13b	Joubert syndrome	HGMD, OMIM	Y	
Atp2b2	Deafness, autosomal recessive 12, modifier of	HGMD, OMIM	Y	E
Auts2	Autism / Mental retardation	HGMD	Y	G
Bin1	Myopathy, centronuclear, autosomal recessive	HGMD, OMIM		
Cabin1	Neurofibromatosis 2 / Schwannomatosis	HGMD	Y	
Cacna1c	Brugada syndrome (shorter-than-normal QT interval) / Timothy syndrome	HGMD, OMIM	Y	E
Cacna1g	Myoclonic epilepsy, juvenile	HGMD	Y	G

Cadm1	Autism spectrum disorder	HGMD	Y	E
Camta1	Impaired episodic memory performance, assoc. with	HGMD	Y	
Cask	Mental retardation, X-linked ? / Microcephaly, mental retard., brainstem & cerebellar hypoplasia	HGMD, OMIM	Y	G
Cdh23	Hearing loss, non-syndromic / Non-syndromic autosomal recessive deafness / Usher syndrome 1 / Usher syndrome 1d	HGMD, OMIM	Y	
Chl1	Mental retardation / Schizophrenia, association with	HGMD	Y	
Cxcl12	HIV 1, resistance to, association with	HGMD, OMIM		
Dcc	Colorectal cancer, risk, assoc. with ?	HGMD		
Ddr1	Schizophrenia, association with	HGMD	Y	
Dlg3	Mental retardation	HGMD	Y	E
Dst	Oesophageal atresia and psychomotor retardation	HGMD	Y	G
Ensa	Reduced insulin secretion, association with	HGMD		
Epb4.1	Elliptocytosis	HGMD, OMIM		E
Epha5	Mental retardation	HGMD	Y	G
Erbb4	Breast and colorectal cancer, association with / Increased promoter activity, association with	HGMD		E
Ercc2	Basal cell carcinoma, reduced risk, association with / Increased response to UV, association with / Lung adenocarcinoma, increased risk, association with / Reduced risk oligodendroglioma development, assoc. / Trichothiodystrophy / Xeroderma pigmentosum	HGMD, OMIM		
Fam126a	Hypomyelination & congenital cataract / Hypomyelination & congenital cataract ?	HGMD, OMIM	Y	E
Flnb	Atelosteogenesis / Boomerang dysplasia / Larsen syndrome, autosomal dominant / Spondylocarpotarsal syndrome	HGMD, OMIM	Y	
Gabrg2	Epilepsy, childhood absence with febrile seizures / Febrile seizures / Generalized epilepsy with febrile seizures plus	HGMD, OMIM	Y	E
Gad1	Cerebral palsy, spastic, symmetric, autosomal recessive / Schizophrenia, familial, association with	HGMD, OMIM	Y	
Gdi1	Mental retardation, non-specific, X-linked	HGMD, OMIM	Y	G
Gnas	Albright hereditary osteodystrophy / Albright hereditary osteodystrophy & pseudohypopar / Albright hereditary osteodystrophy, assoc. with ? / Essential hypertension, association with / Growth retardation, facial dysmorphism, hypotonia / McCune-Albright syndrome / Progressive osseous heteroplasia / Pseudohypoparathyroidism 1a / Pseudohypoparathyroidism 1a, with testotoxicosis / Pseudohypoparathyroidism 1b / Psychiatric disorder, risk, assoc. with ? / Reduced expression, association with / Trauma-related	HGMD, OMIM		

	bleeding, association with			
Gphn	Hyperekplexia / Molybdenum cofactor deficiency	HGMD, OMIM	Y	G
Grik1	Juvenile absence epilepsy, association with ? / Lung cancer, susceptibility to, association with	HGMD	Y	G
Grik2	Mental retardation	HGMD, OMIM	Y	G
Gtf2i	WILLIAMS-BEUREN SYNDROME; WBS	OMIM	Y	
Gyk	Glycerol kinase deficiency	HGMD, OMIM		
Hk1	Haemolytic anaemia	HGMD, OMIM		
Idh3b	Retinitis pigmentosa	HGMD	Y	
Inpp4a	Atopic asthma, reduced risk, association with	HGMD		G
Itgb4	Epidermolysis bullosa / Epidermolysis bullosa simplex / Epidermolysis bullosa with pyloric atresia / Epidermolysis bullosa, junctional / Epidermolysis bullosa, without pyloric atresia	HGMD, OMIM		
Kcnma1	Generalized epilepsy and paroxysmal dyskinesia	HGMD, OMIM	Y	E
Kcnq2	Epilepsy, benign neonatal / Epilepsy, rolandic & benign neonatal convulsions / Epilepsy, rolandic without neonatal seizures / Infantile seizures / Peripheral nerve hyperexcitability	HGMD, OMIM	Y	E
Kel	Kell blood group variation / Kell blood group variation ? / Null allele	HGMD		
Kif21a	CFEOM1 & Marcus Gunn jaw-winking syndrome / Congenital fibrosis of the extraocular muscles 1	HGMD, OMIM		G
Lcorl	STATURE QUANTITATIVE TRAIT LOCUS 13; STQTL13	OMIM		
Lrch1	Knee osteoarthritis, association with	HGMD		
Magi2	Infantile spasms / Infantile spasms in Williams-Beuren syndrome / Myoclonic epilepsy / Seizures	HGMD	Y	
Nav2	Colorectal cancer, increased risk, assoc. with	HGMD		G
Nedd4l	Altered splicing, association with / Epilepsy, photosensitive generalised / Epilepsy, photosensitive generalised ? / Impaired ENaC regulation, association with	HGMD	Y	
Npr2	Acromesomelic dysplasia, Maroteaux type / Short stature	HGMD, OMIM		
Nptn	Increased transcriptional activity, association with / Schizophrenia, reduced risk, association with ?	HGMD	Y	
Nrxn1	Autism spectrum disorder / Autism spectrum disorder ? / Mental retardation, autism, vertebral malformations / Schizophrenia	HGMD, OMIM	Y	G

Nrxn3	Alcohol dependence, association with	HGMD	Y	G
Ntng1	Rett syndrome	HGMD	Y	
Pak3	Mental retardation syndrome, X-linked / Mental retardation with neuropsychiatric features	HGMD, OMIM	Y	
Pde8b	Adrenal hyperplasia	HGMD		
Pnpla6	Motor neuron disease	HGMD, OMIM	Y	
Porcn	Focal dermal hypoplasia	HGMD, OMIM		
Prdm2	Bone mineral density, association with	HGMD		
Ptprf	Obesity, reduced risk, association with	HGMD	Y	G
Ptprz1	HELICOBACTER PYLORI INFECTION, SUSCEPTIBILITY TO	OMIM		
Rims1	Cone-rod dystrophy	HGMD, OMIM	Y	G
Robo2	Urinary tract anomalies / Vesicoureteral reflux	HGMD, OMIM		G
Ryr2	Arrhythmogenic right ventricular cardiomyopathy, assoc. with / Arrhythmogenic right ventricular dysplasia type 2 / Catecholaminergic polymorphic ventricular tachycardia / Ventricular tachycardia, polymorphic	HGMD, OMIM		
Sbf2	Charcot-Marie-Tooth disease 4b2	HGMD, OMIM		
Slc1a2	Progressing stroke, incr. risk, association with	HGMD	Y	E
Slc4a10	Autism	HGMD	Y	G
Smad2	Congenital heart defects ?	HGMD		G
Smad4	Haemorrhagic telangiectasia / Juvenile polyposis and haemorrhagic telangiectasia / Juvenile polyposis coli / Juvenile polyposis syndrome	HGMD, OMIM		
Sorbs1	Obesity and diabetes, reduced risk, association	HGMD		E
Srr	Schizophrenia, association with	HGMD	Y	
Stim1	IMMUNE DYSFUNCTION WITH T-CELL INACTIVATION DUE TO CALCIUM ENTRY DEFECT	OMIM		
Stxbp1	Epileptic encephalopathy, early infantile	HGMD, OMIM	Y	
Syne2	Increased spine bone mineral density in men, association / Muscular dystrophy, Emery-Dreifuss	HGMD		E
Tpm1	Cardiomyopathy, dilated / Cardiomyopathy, hypertrophic	HGMD, OMIM		G

Tpm2	Cap disease / Distal arthrogryposis syndrome 1 / Muscle weakness and distal limb deformity / Nemaline myopathy	HGMD, OMIM		
Tpm3	Fibre-type disproportion myopathy, congenital / Nemaline myopathy / Nemaline myopathy ?	HGMD, OMIM		
Trp53bp1	Lung cancer, susceptibility to, association with	HGMD		
Ttn	Cardiomyopathy, dilated / Cardiomyopathy, hypertrophic / Myopathy / Tibial muscular dystrophy / Titin deficiency	HGMD, OMIM		G
Whsc1	WOLF-HIRSCHHORN SYNDROME; WHS	OMIM	Y	
Wnk1	Ambulatory blood pressure variation, association with / Blood pressure, assoc. with ? / Colorectal cancer, increased risk, assoc. with / Neuropathy, hereditary sensory, type II / Protein kinase deficiency / Pseudohypoaldosteronism 2	HGMD, OMIM	Y	G

* G indicates a gene putatively regulated by both Nova and Fox, but not necessarily on the same exon; E represents a gene with at least one exon under predicted combinatorial regulation by Nova and Fox.

Supporting references

S1. C. Zhang, M. L. Hastings, A. R. Krainer, M. Q. Zhang, *Proc. Natl. Acad. Sci. USA* **104**, 15028 (2007).

S2. R. M. Kuhn *et al.*, *Nucl. Acids Res.* **37**, D755 (2009).

S3. S. S. Gross, M. R. Brent, *J Comput Biol* **13**, 379 (2006).

S4. A. Siepel, D. Haussler, paper presented at the Proc. 8th Annual Int'l Conf. on Res Comput Biol, New York, 2004.

S5. D. D. Licatalosi *et al.*, *Nature* **456**, 464 (2008).

S6. J. Ule *et al.*, *Science* **302**, 1212 (2003).

S7. S. W. Chi, J. B. Zang, A. Mele, R. B. Darnell, *Nature* **460**, 479 (2009).

S8. L. R. Rabiner, *Proc. IEEE* **77**, 257 (1990).

S9. J. Ule *et al.*, *Nature* **444**, 580 (2006).

S10. A. Stark *et al.*, *Nature* **450**, 219 (2007).

S11. C. Zhang *et al.*, *Genes Dev* **22**, 2250 (2008).

S12. W. Miller *et al.*, *Genome Res.* **17**, 1797 (2007).

S13. S. H. Bernhart, I. L. Hofacker, P. F. Stadler, *Bioinformatics* **22**, 614 (2006).

S14. T. Clark *et al.*, *Genome Biol.* **8**, R64 (2007).

S15. J. Ule *et al.*, *Nature Genet.* **37**, 844 (2005).

S16. G. K. Smyth, *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).

S17. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, *Nat Meth* **5**, 621 (2008).

S18. R. J. Buckanovich, R. B. Darnell, *Mol. Cell. Biol.* **17**, 3194 (1997).

S19. Y. Y. L. Yang, G. L. Yin, R. B. Darnell, *Proc. Natl. Acad. Sci. USA* **95**, 13254 (1998).

S20. N. Friedman, *Science* **303**, 799 (2004).

S21. E. Segal *et al.*, *Nat Genet* **34**, 166 (2003).

S22. N. Friedman, M. Linial, I. Nachman, D. Pe'er, *J Comput Biol* **7**, 601 (2000).

S23. K. Murphy, S. Mian. (Technical Report, Computer Science Division, University of California, Berkeley, 1999).

S24. Y. Wang, X.-S. Zhang, Y. Xia, *Nucl. Acids Res.* **37**, 5943 (2009).

S25. R. Jansen *et al.*, *Science* **302**, 449 (2003).

S26. D. Heckerman, in *Learning in Graphical Models,* M. Jordan, Ed. (MIT Press, Cambridge, MA, 1999).

S27. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, ed. 1, 1988).

S28. Y. Xing, C. Lee, *Nature Rev. Genet.* **7**, 499 (2006).

S29. K. P. Murphy, *Comput Sci Stat* **33**, (2001).

S30. S. Stamm *et al.*, *Nucleic Acids Res.* **34**, D46 (2006).

S31. W. J. Kent, *Genome Res.* **12**, 656 (2002).
S32. B. K. Dredge, G. Stefani, C. C. Engelhard, R. B. Darnell, *EMBO J.* **24**, 1608 (2005).
S33. B. K. Dredge, R. B. Darnell, *Mol. Cell. Biol.* **23**, 4687 (2003).
S34. A. Siepel *et al.*, *Genome Res.* **15**, 1034 (2005).
S35. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Springer, New York, ed. 2, 2009).
S36. R. Ihaka, R. Gentleman, *J. Comput. Graph. Statist.* **5**, 299 (1996).
S37. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc Natl Acad Sci U S A* **95**, 14863 (1998).
S38. E. T. Wang *et al.*, *Nature* **456**, 470 (2008).
S39. B. Groschel, F. Bushman, *J. Virol.* **79**, 5695 (2005).
S40. G. Dennis *et al.*, *Genome Biol* **4**, R60 (2003).
S41. A. Forrest *et al.*, *BMC Bioinformatics* **7**, 82 (2006).
S42. The UniProt Consortium, *Nucl. Acids Res.* **37**, D169 (2009).
S43. G. Slater, E. Birney, *BMC Bioinformatics* **6**, 31 (2005).
S44. R. P. Munton *et al.*, *Mol Cell Proteomics* **6**, 283 (2007).
S45. B. A. Ballif, G. R. Carey, S. R. Sunyaev, S. P. Gygi, *J Proteome Res* **7**, 311 (2007).
S46. J. C. Trinidad, C. G. Specht, A. Thalhammer, R. Schoepfer, A. L. Burlingame, *Mol Cell Proteomics* **5**, 914 (2006).
S47. B. A. Ballif, J. Villen, S. A. Beausoleil, D. Schwartz, S. P. Gygi, *Mol Cell Proteomics* **3**, 1093 (2004).
S48. M. O. Collins *et al.*, *J Biol Chem* **280**, 5972 (2005).
S49. V. A. McKusick, *Mendelian inheritance in man. A catalog of human genes and genetic disorders*. (Johns Hopkins University Press, Baltimore, ed. 12, 1998), pp. 3972.
S50. P. Stenson *et al.*, *Hum. Mutat.* **21**, 577 (2003).
S51. S. Banerjee-Basu, A. Packer, *Dis Model Mech* **3**, 133.
S52. K. B. Jensen *et al.*, *Neuron* **25**, 359 (2000).
S53. W. Hoch, M. Ferns, J. T. Campanelli, Z. W. Hall, R. H. Scheller, *Neuron* **11**, 479 (1993).
S54. P. Scotton *et al.*, *J. Biol. Chem.* **281**, 36835 (2006).
S55. M. Ruggiu *et al.*, *Proc Natl Acad Sci U S A* **106**, 3513 (2009).
S56. L. L. Peters *et al.*, *J. Cell Biol.* **130**, 313 (1995).
S57. A. A. Hopitzan, A. J. Baines, M.-A. Ludosky, M. Recouvreur, E. Kordeli, *Exp Cell Res* **309**, 86 (2005).
S58. E. E. Strehler, M. A. Strehler-Page, G. Vogel, E. Carafoli, *Proc Natl Acad Sci U S A* **86**, 6908 (1989).
S59. M. C. Chicka, E. E. Strehler, *J. Biol. Chem.* **278**, 18464 (2003).
S60. H. P. Adamo, J. T. Penniston, *Biochem. J.* **283**, 355 (1992).
S61. H. Roger, H. U. G. Martin, I. Tomoko, E. S. Emanuel, C. Ernesto, *Eur. J. Biochem.* **205**, 333 (1992).
S62. H. Hilfiker, D. Guerini, E. Carafoli, *J Biol Chem.* **269**, 26178 (1994).

S63. R. Wechsler-Reya, D. Sakamuro, J. Zhang, J. Duhadaway, G. C. Prendergast, *J. Biol. Chem.* **272**, 31453 (1997).
S64. K. Ge *et al.*, *Proc Natl Acad Sci U S A* **96**, 9689 (1999).
S65. R. R. Antoine, *J Neurochem* **70**, 2369 (1998).
S66. N. M. Soldatov, *Proc Natl Acad Sci U S A* **89**, 4628 (1992).
S67. R. J. Diebold *et al.*, *Proc Natl Acad Sci U S A* **89**, 1497 (1992).
S68. T. P. Snutch, W. J. Tomlinson, J. P. Leonard, M. M. Gilbert, *Neuron* **7**, 45 (1991).
S69. K. U. Bayer, P. D. Koninck, H. Schulman, *EMBO J* **21**, 3590 (2002).
S70. B. J. Curry *et al.*, *Genomics* **83**, 425 (2004).
S71. M. A. Reale *et al.*, *Cancer Res* **54**, 4493 (1994).
S72. H. A. Burgess, O. Reiner, *J. Biol. Chem.* **277**, 17696 (2002).
S73. B. M. Engels, T. G. Schouten, J. van Dullemen, I. Gosens, E. Vreugdenhil, *Mol Brain Res* **120**, 103 (2004).
S74. M. McLaughlin *et al.*, *J Biol Chem* **277**, 6406 (2002).
S75. J. P. Huang *et al.*, *J. Biol. Chem.* **268**, 3758 (1993).
S76. A. Kontrogianni-Konstantopoulos, C. Frye, E. J. Benz, S. Huang, *J Biol Chem* **276**, 20679 (2001).
S77. G. Subrahmanyam, P. J. Bertics, R. A. Anderson, *Proc Natl Acad Sci U S A* **88**, 5222 (1991).
S78. M. Parra *et al.*, *Genomics* **84**, 637 (2004).
S79. T. Kakimoto, H. Katoh, M. Negishi, *J Biol Chem* **279**, 14104 (2004).
S80. B. J. Krishek *et al.*, *Neuron* **12**, 1081 (1994).
S81. V. K. David, N. Alan, A. S. J. Paul, *J Neurobiol* **52**, 156 (2002).
S82. P. Bray *et al.*, *Proc Natl Acad Sci U S A* **83**, 8893 (1986).
S83. Z. Ren *et al.*, *J Biol Chem.* **278**, 52700 (2003).
S84. G. M. Durand, M. V. Bennett, R. S. Zukin, *Proc Natl Acad Sci U S A* **90**, 6731 (1993).
S85. R. S. Zukin, M. V. L. Bennett, *Trends Neurosci* **18**, 306 (1995).
S86. W. G. Tingley, K. W. Roche, A. K. Thompson, R. L. Huganir, *Nature* **364**, 70 (1993).
S87. M. Hollmann *et al.*, *Neuron* **10**, 943 (1993).
S88. Y. Mu, T. Otsuka, A. C. Horton, D. B. Scott, M. D. Ehlers, *Neuron* **40**, 581 (2003).
S89. S. M. Clinton, V. Haroutunian, K. L. Davis, J. H. Meador-Woodruff, *Am J Psychiatry* **160**, 1100 (2003).
S90. H. Habuchi *et al.*, *Biochem. J.* **371**, 131 (2003).
S91. K. Hahm *et al.*, *Mol. Cell. Biol.* **14**, 7111 (1994).
S92. L. Sun, A. Liu, K. Georgopoulos, *EMBO J* **15**, 5358 (1996).
S93. S. Ezzat, S. Yu, S. L. Asa, *Am J Pathol* **163**, 1177 (2003).
S94. C. Pucharcos, C. Casas, M. Nadal, X. Estivill, S. de la Luna, *Biochim. Biophys. Acta.* **1521**, 1 (2001).
S95. N. K. Hussain *et al.*, *Nat Cell Biol* **3**, 927 (2001).

S96. L. Tian *et al.*, *Proc Natl Acad Sci U S A* **101**, 11897 (2004).

S97. E. Leung *et al.*, *Immunol Cell Biol* **74**, 421 (1996).

S98. R. P. Laura, S. Ross, H. Koeppen, L. A. Lasky, *Exp Cell Res* **275**, 155 (2002).

S99. J. H. Wright *et al.*, *Mol. Cell. Biol.* **23**, 2068 (2003).

S100. R. Zhao, Z. J. Zhao, *Biochem Biophys Res Commun* **303**, 1028 (2003).

S101. T. Maes, A. Barcel, C. Buesa, *Genomics* **80**, 21 (2002).

S102. P. J. Peeters *et al.*, *Dev Brain Res* **150**, 89 (2004).

S103. N. Tamura, D. L. Garbers, *J. Biol. Chem.* **278**, 48880 (2003).

S104. L. R. Potter, T. Hunter, *J Biol Chem* **273**, 15533 (1998).

S105. J. M. Verdi *et al.*, *Proc Natl Acad Sci U S A* **96**, 10472 (1999).

S106. V. Rousseau, O. Goupille, N. Morin, J.-V. Barnier, *J. Biol. Chem.* **278**, 3912 (2003).

S107. N. Yousaf, Y. Deng, Y. Kang, H. Riedel, *J Biol Chem* **276**, 40940 (2001).

S108. S. S. M. Rensen *et al.*, *Cardiovasc Res* **55**, 850 (2002).

S109. I. C. Bark, M. C. Wilson, *Gene* **139**, 291 (1994).

S110. I. C. Bark, K. M. Hahn, A. E. Ryabinin, M. C. Wilson, *Proc Natl Acad Sci U S A* **92**, 1510 (1995).

S111. J. B. Sorensen *et al.*, *Cell* **114**, 75 (2003).

S112. S. A. Matson, G. C. Pare, M. S. Kapiloff, *Biochim Biophys Acta* **1727**, 145 (2005).

S113. T. F. Duchaine *et al.*, *J Cell Sci* **115**, 3285 (2002).

S114. N. Ruiz-Opazo, B. Nadal-Ginard, *J. Biol. Chem.* **262**, 4755 (1987).

S115. J. S. Grossman *et al.*, *RNA* **4**, 613 (1998).

S116. D. M. Helfman, R. F. Roscigno, G. J. Mulligan, L. A. Finn, K. S. Weber, *Genes Dev* **4**, 98 (1990).

S117. J. M. Percival *et al.*, *Mol. Biol. Cell* **15**, 268 (2004).