csDEX: Condition-specific differential exon expression (Supplementary information)

Martin Stražar, Jernej Ule and Tomaž Curk

June 7, 2017

Contents

1 csDEX implementation and results on simulated data **1**
1.1 Wald test based model approximation 1
1.2 Comparison of dispersion fitting methods 7

2 Experiments on ENCODE RNA-seq datasets **8**
2.1 Number of unique exonic parts per gene 8
2.2 Comparison of differential exon usage methods 8
2.3 Batch effects / sample clustering 12
2.4 Binding and motif enrichment 17
2.4.1 Binding enrichment signal (eCLIP) 17
2.4.2 Motif enrichment signal 17
2.5 Retrieving TARDBP-regulated cryptic exons 20

3 Supplementary tables with experimental details **22**

1 csDEX implementation and results on simulated data

In this section, we give details on model approximation based on Wald Test, with investigation its effects on timing and accuracy in comparison to the full-model.

1.1 Wald test based model approximation

In the following section, we describe a low-rank model approximation to speed up re-computation of alternative models. The strategy is based on the observation,

that the distance between parameter vectors for the null and alternative model is inversely proportional to the number of conditions. The setting is depicted on Supplementary Fig. 1.

Let n_e be the number of exonic parts of a gene and n_c the number of experimental conditions. Let $\mathbf{X}_0 \in \mathbb{R}^{n_e n_c \times (n_e + n_c)}$ denote a binary design matrix, with one row per observation, and columns encoding each pair of exonic part e and condition c. This is an established description of general linear models that is given by the Eq. 1 below. The low-rank approximation is performed in the same manner for both the count- and PSI-based models, hence we make no distinction between the two in this description. The null models for both distributions have a similar form

$$\text{link}(\mu_{ec}) = \beta_e + \beta_c \tag{1}$$

where the link function equals *log* (the count model) or the *logit* (the PSI model). Let the parameter vector $\boldsymbol{\beta}_0 \in \mathbb{R}^{n_e + n_c}$ be a local maximizer of the null model likelihood. For all $n_e n_c$ alternative models, the design matrices $\mathbf{X}_a \in \mathbb{R}^{n_e n_c \times (n_e + n_c + 1)}$ and the parameter vectors $\boldsymbol{\beta}_a \in \mathbb{R}^{n_e + n_c + 1}$, define an additional factor for the candidate interaction e', c' being tested:

$$\text{link}(\mu_{ec}) = \beta_e + \beta_c + \delta_{ee'}\delta_{cc'}\beta_{e'c'}. \tag{2}$$

Note that the matrix $\mathbf{X}_0$ is equal to each matrix $\mathbf{X}_a$ in all but the last column. The number of model parameters is of order $n_e + n_c$, while the number of fitted equals $n_e \times n_c$, respectively. Thus, the *number of fitted values per parameter* equals $\frac{n_e n_c}{n_e + n_c}$ and rapidly saturates as n_c grows, with the derivative of the order n_c^{-2} (Supplementary Fig. 2a). Hence, as n_c grows, the mean squared-error (the distance) between corresponding components of vectors $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_a$ is expected to decrease. This is due to single additional factor having a diminishing effect on all other parameters (Supplementary Fig. 2b). Hence, re-using the information from the null model fit $\boldsymbol{\beta}_0$ can be used to reduce the time required to find each $\boldsymbol{\beta}_a$.

We propose a model approximation algorithm based on the Wald test of parameter significance. Let $\beta_{0,i}$ be the i-th component of the null model parameter vector. The corresponding estimated variance σ_i^2 is defined by the Iterative reweighted least-squared (IRLS) algorithm as the i-th diagonal entry in the Fisher information matrix (**?**). For each $\boldsymbol{\beta}_{0,i}$, the Wald test statistic is defined as

$$w_i = \frac{\beta_{0,i}^2}{\sigma_i^2} \tag{3}$$

which follows a χ^2 distribution with one degree of freedom. Thus, the statistical significance of $\beta_{0,i}$ being nonzero is quantified, subject to an user-defined significance level α.

Let $\mathcal{Z}$ be a subset of indices $\{1, 2, ...n_e + n_c\}$ for which w_i is not statistically significant subject to α. The parameters at indices in $\mathcal{Z}$ are not significantly different from zero and can be related to exonic parts and/or conditions with a majority of corresponding observations equal to zero (for both count and PSI data). These parameters are unlikely to change significantly when the interaction factor is added in Eq. 2, and do not significantly affect the null model likelihood.

We construct reduced design matrices for the null and alternative models by merging the non significant parameters into an *residual* factor ϵ_{ec}. The residual for each observation is computed as a dot product

$$\epsilon_{ec} = \mathbf{X}_0[ec, \mathcal{Z}]\boldsymbol{\beta}_0[\mathcal{Z}], \tag{4}$$

yielding new formulations for the *reduced* null and *reduced* alternative models as follows:

$$\begin{aligned} \text{link}(\mu_{ec}) &= \beta_e + \beta_c + \epsilon_{ec} \\ \text{link}(\mu_{ec}) &= \beta_e + \beta_c + \epsilon_{ec} + \delta_{ee'}\delta_{cc'}\beta_{e'c'} \end{aligned} \tag{5}$$

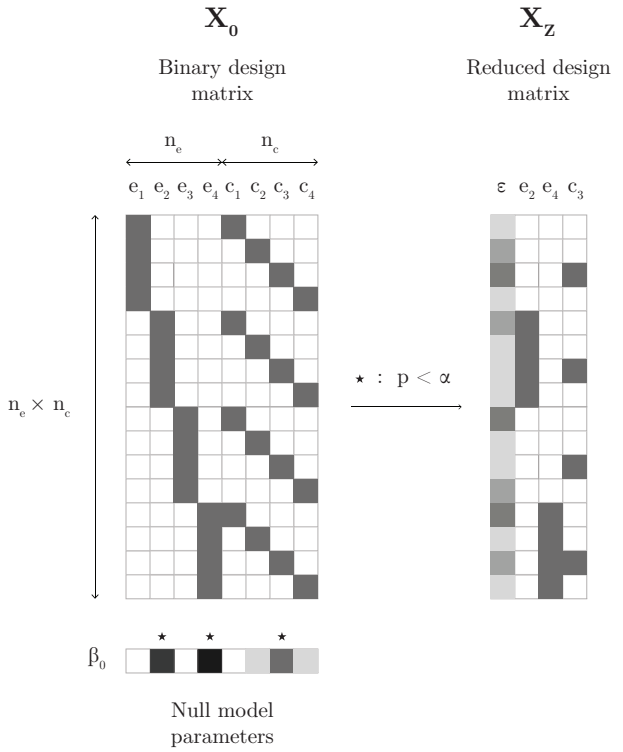
where β_e or β_c are nonzero if and only if $e \notin \mathcal{Z}$ or $c \notin \mathcal{Z}$, respectively. The reduced null and reduced alternative models have $n_e + n_c - |\mathcal{Z}| + 1$ and $n_e + n_c - |\mathcal{Z}| + 2$ parameters, respectively, where $|\cdot|$ denotes cardinality of the set, and an intercept factor is assumed implicitly.

The parameter vector for the reduced null model is computed directly from $\boldsymbol{\beta}_0$, i.e., does not require model refitting. Specifically, the parameter corresponding to the residual factor, $\boldsymbol{\beta}_{0\mathcal{Z},\epsilon} = 1$, while the parameters for β_e and β_c, $e, c \notin \mathcal{Z}$ do not change. Note that the reduced null model will have the same likelihood as the original null model. For the reduced alternative model, the value of $\boldsymbol{\beta}_{a\mathcal{Z},\epsilon}$ is obtained through IRLS likewise to all other parameters.

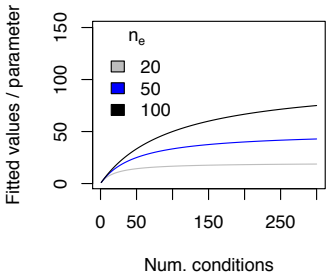
This definition still allows application of likelihood ratio or difference in deviance model selection tests, since the reduced null is a special case of the reduced alternative model. The reduced alternative model will have a likelihood that is *less than or equal* to the original alternative model. This implies lower statistical power, but does not affect the probability of Type I error (false positive rate) under the same significance threshold.

Depending on cardinality of $\mathcal{Z}$ (number of zero observations), the savings in the dimension of the parameter space can be substantial, greatly reducing the time required for testing alternative models with large number of exonic parts and/or conditions.

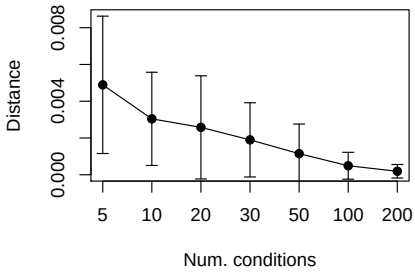
The effect of low-rank approximation, dependence of timing and accuracy on parameter α is shown on Suppl. Figs. 2-3.



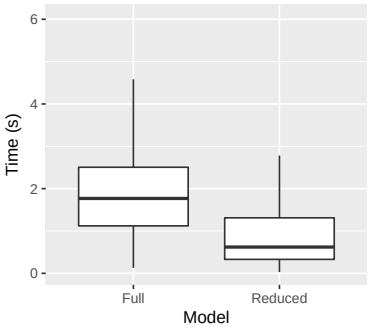
Supplementary Figure 1: Overview of the reduced model algorithm, showing a hypothetical dataset with $n_e = 4$ exonic parts and $n_c = 4$ conditions. The $\boldsymbol{\beta}_0$ is a vector of parameters corresponding to the null model, and statistically significant values (Wald test, $p < \alpha$) are marked with a star ($\star$). The statistically non-significant parameters and the corresponding columns of the original design matrix (left) are merged the residual factor ϵ, obtaining the reduced design matrix (right). An intercept (bias) factor is assumed, but not shown.



(a)



(b)

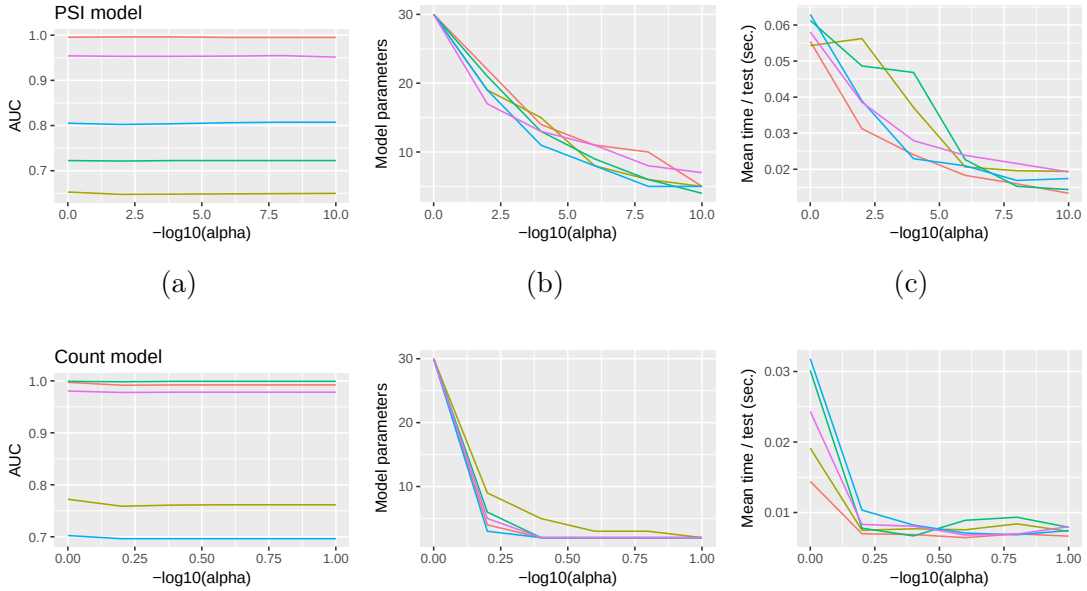


(c)

Quantile	Full (s)	Reduced (s)
99%	6.04	4.64
75%	2.53	1.33
50%	1.78	0.63
25%	1.13	0.33
1%	0.30	0.10

(d)

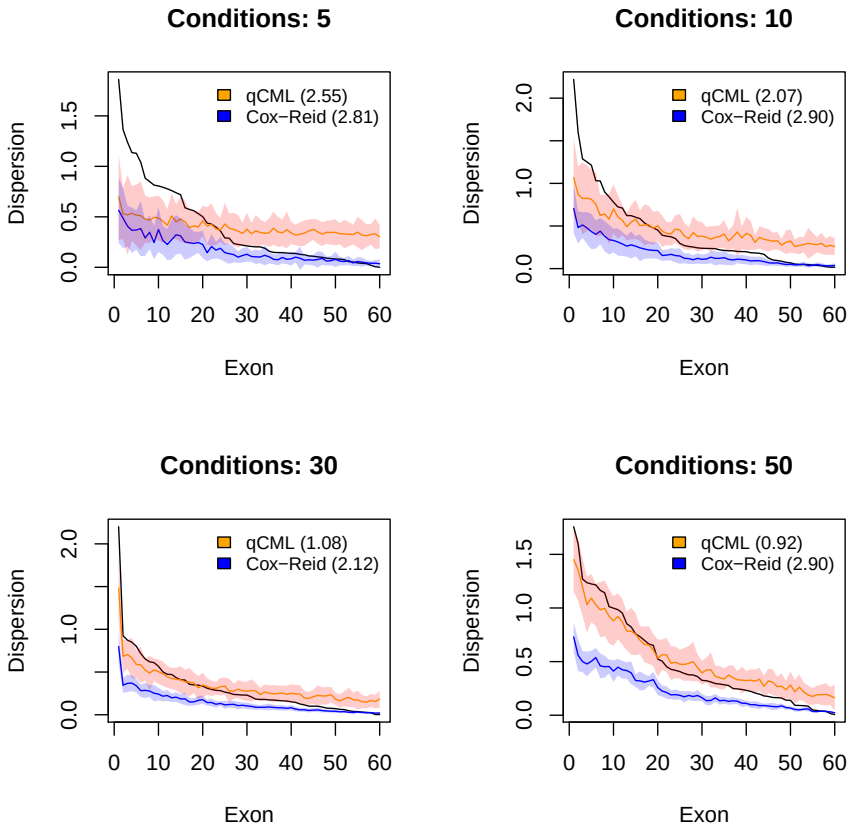
Supplementary Figure 2: Empirical evaluation of model reduction on a real biological dataset, which is constructed as described in Section 2.2.1 (main text). **a)** The relationship between the number of conditions (n_c) to the ratio of fitted values per parameter ($\frac{n_e n_c}{n_e+n_c}$). **b)** Model fits for a randomly selected gene. The distance between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_a$ is inversely proportional to n_c. Standard deviations are computed over all possible interacting pairs e, c. **c)** The distribution of times (in seconds) per one model fit in the whole dataset when using the full and reduced models. **d)** Quantiles of the time distributions in c).



Supplementary Figure 3: Effect of significance level α on synthetic count (a-c) and PSI (d-f) datasets described in Section 2.1 (main text). Five replicate datasets are generated and shown in different colors. **a, c)** Change in retrieval accuracy (Area under receiver-operating characteristic, AUC). **b, d)** Change in the number of significant non-zero model parameters. **c, f)** Change in mean time per model fit.

1.2 Comparison of dispersion fitting methods

We compare the qCML and Cox-Reid dispersion estimation methods depending on different number of conditions on Suppl. Fig. 4.



Supplementary Figure 4: Dispersion fitting with qCML (orange) and Cox-Reid dispersion estimate (blue) for increasing number of conditions: 5, 10, 30, 50 (left to right). Root mean square error is shown in parentheses. True dispersion parameter used for data sampling is shown in black.

2 Experiments on ENCODE RNA-seq datasets

In the following, we present details on data preparation and supplementary results for ENCODE RNA-seq datasets.

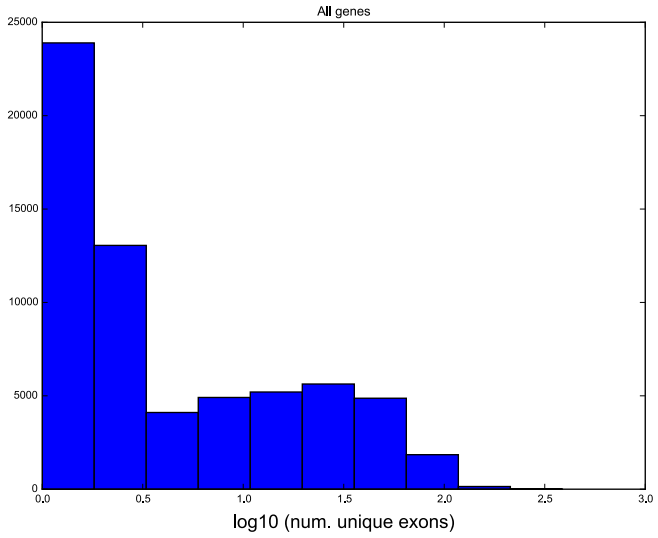
2.1 Number of unique exonic parts per gene

The distribution of the number of unique exonic parts per gene, used to construct a subset of the original annotation is shown on 5.

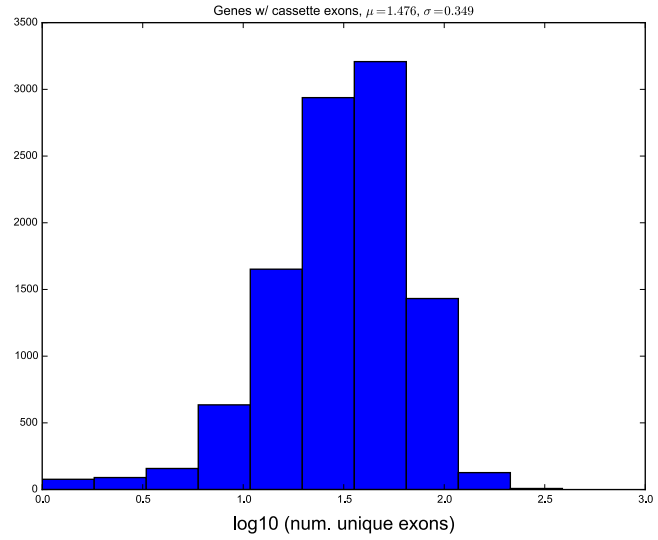
2.2 Comparison of differential exon usage methods

We compare the methods csDEX-PSI, csDEX-count, DEXSeq and rMATS using

- Uniqueness scores (Suppl. Fig. 6), and
- Precision of retrieving known alternative splicing events (Suppl. Fig. 7)

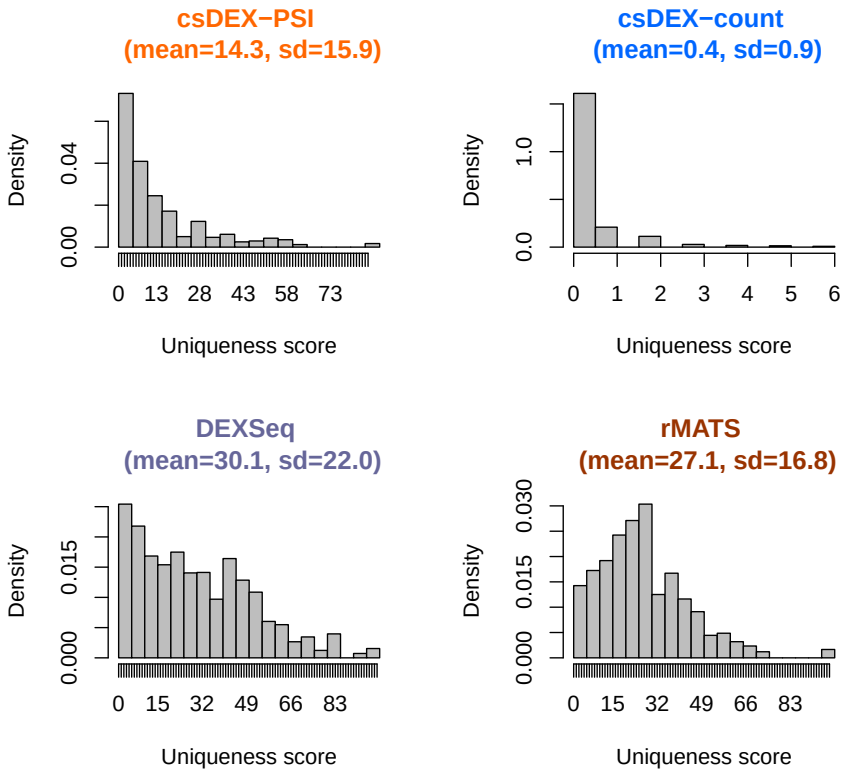


(a)

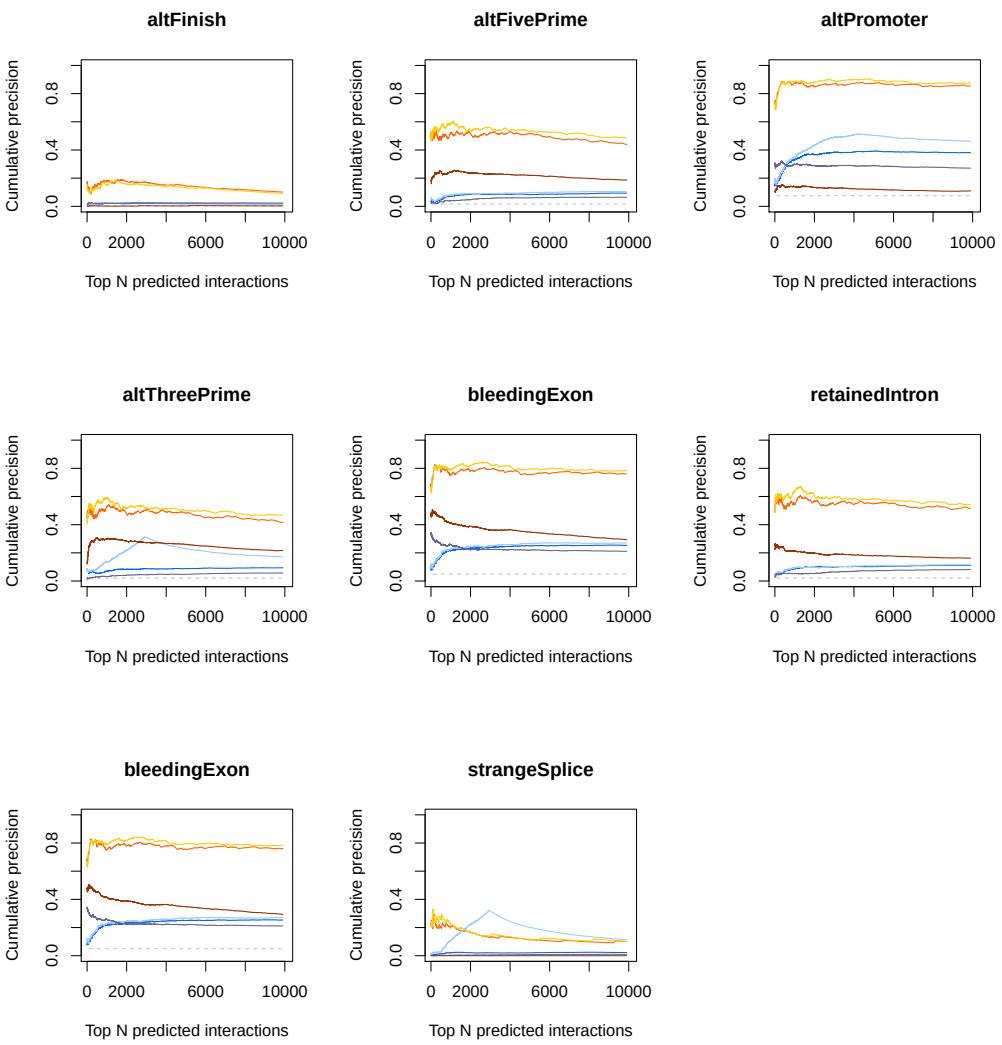


(b)

Supplementary Figure 5: Distribution of the number of unique exons in $\log_{10}$ scale for genes in the hg19 genome. a) All genes. b) A subset of all genes; the genes containing at least one annotated cassette exon in the *knownAlt* track (maximum likelihod parameters, mean and std. deviation 1.476 ± 0.349 (absolute counts $10^{1.476} = 29.88$, $10^{1.476-0.349} = 13.39$, $10^{1.476+0.349} = 66.68$).



Supplementary Figure 6: Distributions of uniqueness score. The distributions are computed for each feature, condition pair (e, c) on interaction list retrieved by each model. The uniqueness score is defined for (e, c) as the number of times e appears in the interaction list for different c' - the lower uniqueness score implies that the feature is differentially used in a smaller number of conditions.



Supplementary Figure 7: Evaluation of differential-exon usage detection methods for alternative splicing events as annotated in UCSC altEvent track. The list of candidate interactions provided by each of the methods is ranked by statistical significance (p-value). The plots show cumulative probabilities of known event of corresponding types within lists of size 30 to 10,000. The gray line show the fraction of each splicing event in the dataset.

2.3 Batch effects / sample clustering

The complete RNA-seq dataset in the ENCODE project at the time of writing consisted of 190 shRNA knockdowns. The dataset can be divided into 19 batches - a set of knockdown corresponding to the same control experiments, performed at different dates. We performed two clustering analyses to investigate for possible presence of batch effects that can be responsible for a non-negligible portion of technical variability in sequencing studies (**?**).

We create a design matrix consisting of 420 RNA-seq samples and 139,759 exonic parts (dataset for annotation with genes from 13 to 66 exonic parts, Suppl. Table 3, online) with corresponding Percent-spliced In (PSI) value. For each sample, additional metadata on experiment date and control sample is included. We use the Principal component analysis (PCA) and Multi-dimensional scaling (MDS) clustering methods to assess the similarities of samples within- and between- batches, with the results summarized in Suppl. Table 1.

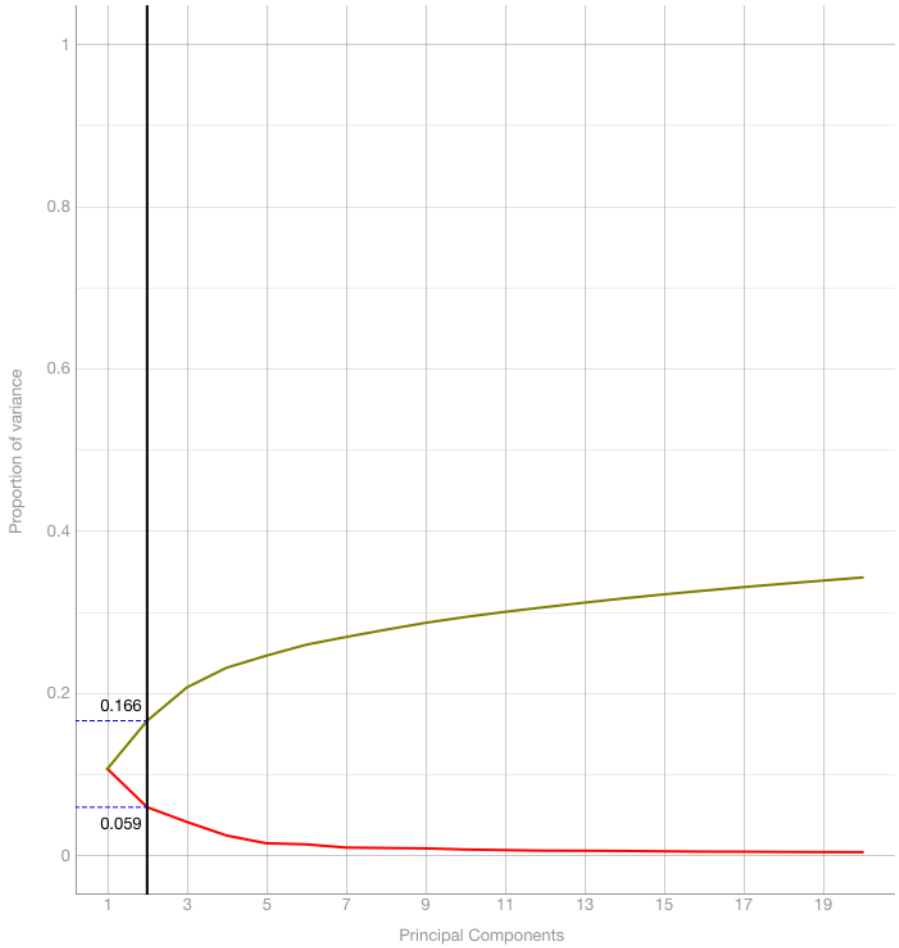
The detailed results of PCA on the design matrix with PSI values are available in(Suppl. Table 10, online). The histogram of explained variance per principal components displays a heavy-tailed distribution. Nevertheless, the two principal components (PC1 and PC2) with largest explained variance account for a non-negligible 16.6 % of total variance (Suppl. Fig. 8). A scatter plot of principal component coordinates per each sample displays clustering of experiments in the same batch (Suppl. Fig. 9). The samples corresponding to the same batch tend to be closer to each other in the coordinate system spanned by PC1 and PC2 (average within-batch Euclidean distance 60.885 ± 43.118; average between-batch distance: 198.794 ± 108.319).

A similar result is obtained when using the MDS transformation, where the data is mapped to a two-dimensional space spanned by coordinates M1 and M2 (Suppl. Table 11, online). The Euclidean distances between samples in the M1-M2 space approximate the Spearman correlation distance in the original (139,759-dimensional) space. A similar clustering effect is seen after the MDS transformation (Suppl. Fig 10); the samples within the same batch then to be substantially closer (average within-batch Euclidean distance 0.047 ± 0.032; average between-batch distance: 0.105 ± 0.045).

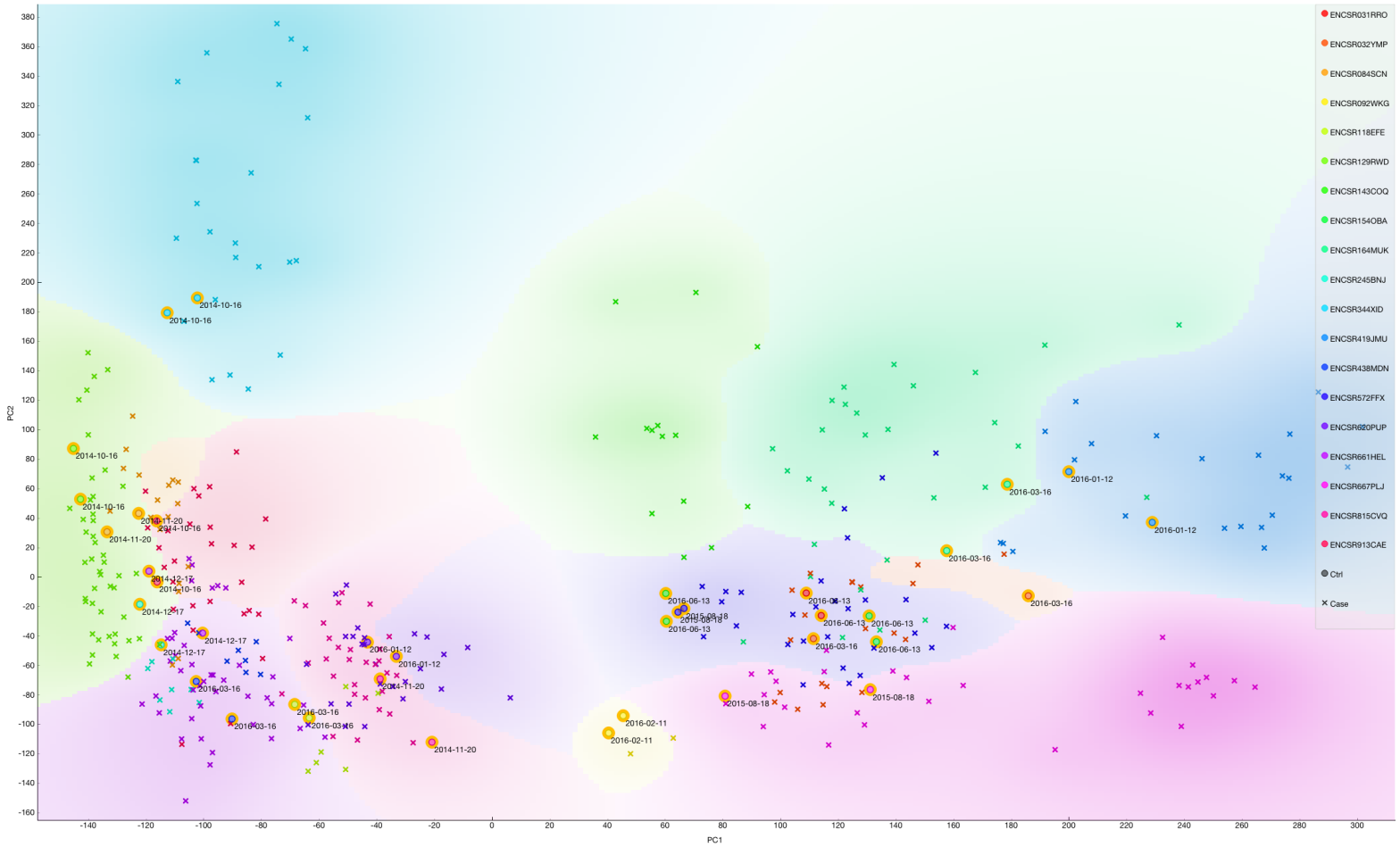
Together, the results of this analysis confirm a possible presence of batch effects, where a substantial proportion of variance could be related to non-biological variation, such as differences in laboratory conditions or personnel, and more.

	Control	N	date	dw.PCA	db.PCA	dw.MDS	db.MDS
1	ENCSR129RWD	48	2014-10-16	**66.38 ± 48.40**	196.61 ± 111.18	**0.05 ± 0.02**	0.11 ± 0.05
2	ENCSR661HEL	44	2014-12-17	**49.99 ± 30.28**	176.48 ± 112.12	**0.06 ± 0.04**	0.11 ± 0.05
3	ENCSR164MUK	32	2016-03-16	**86.73 ± 47.66**	222.25 ± 86.69	**0.05 ± 0.03**	0.10 ± 0.04
4	ENCSR667PLJ	32	2014-12-17	**81.03 ± 50.03**	235.08 ± 106.35	**0.04 ± 0.02**	0.11 ± 0.05
5	ENCSR815CVQ	32	2014-11-20	**37.86 ± 22.54**	156.06 ± 94.78	**0.04 ± 0.02**	0.08 ± 0.03
6	ENCSR572FFX	30	2015-08-18	**55.87 ± 32.28**	186.90 ± 83.63	0.06 ± 0.05	0.11 ± 0.04
7	ENCSR913CAE	30	2014-10-16	**55.14 ± 37.25**	167.72 ± 102.56	**0.05 ± 0.03**	0.10 ± 0.04
8	ENCSR620PUP	28	2016-01-12	**37.45 ± 19.95**	153.39 ± 91.19	**0.03 ± 0.02**	0.10 ± 0.04
9	ENCSR344XID	26	2014-10-16	**93.71 ± 59.50**	306.69 ± 95.53	**0.03 ± 0.01**	0.12 ± 0.05
10	ENCSR419JMU	24	2016-01-12	**66.30 ± 31.53**	286.55 ± 106.56	**0.04 ± 0.03**	0.12 ± 0.05
11	ENCSR032YMP	22	2016-03-16	**52.54 ± 28.28**	192.26 ± 89.68	0.06 ± 0.05	0.11 ± 0.05
12	ENCSR084SCN	20	2014-11-20	**50.36 ± 36.85**	178.82 ± 103.35	**0.04 ± 0.02**	0.11 ± 0.05
13	ENCSR143COQ	16	2016-06-13	**79.61 ± 50.83**	188.73 ± 62.33	0.06 ± 0.04	0.09 ± 0.04
14	ENCSR245BNJ	10	2014-12-17	**27.49 ± 17.37**	169.48 ± 117.10	**0.05 ± 0.03**	0.10 ± 0.05
15	ENCSR438MDN	10	2016-03-16	**25.60 ± 14.22**	156.53 ± 111.16	**0.04 ± 0.02**	0.10 ± 0.05
16	ENCSR118EFE	8	2016-03-16	**32.46 ± 16.81**	169.34 ± 104.85	**0.05 ± 0.04**	0.11 ± 0.05
17	ENCSR092WKG	4	2016-02-11	**19.78 ± 4.52**	178.04 ± 82.37	**0.03 ± 0.02**	0.11 ± 0.05
18	ENCSR031RRO	2	2016-06-13	**16.10 ± 0.00**	172.34 ± 85.71	**0.03 ± 0.00**	0.08 ± 0.04
19	ENCSR154OBA	2	2016-06-13	**17.76 ± 0.00**	184.56 ± 94.42	**0.04 ± 0.00**	0.10 ± 0.05
20	all	420		**60.89 ± 43.12**	198.79 ± 108.32	**0.05 ± 0.03**	0.10 ± 0.05

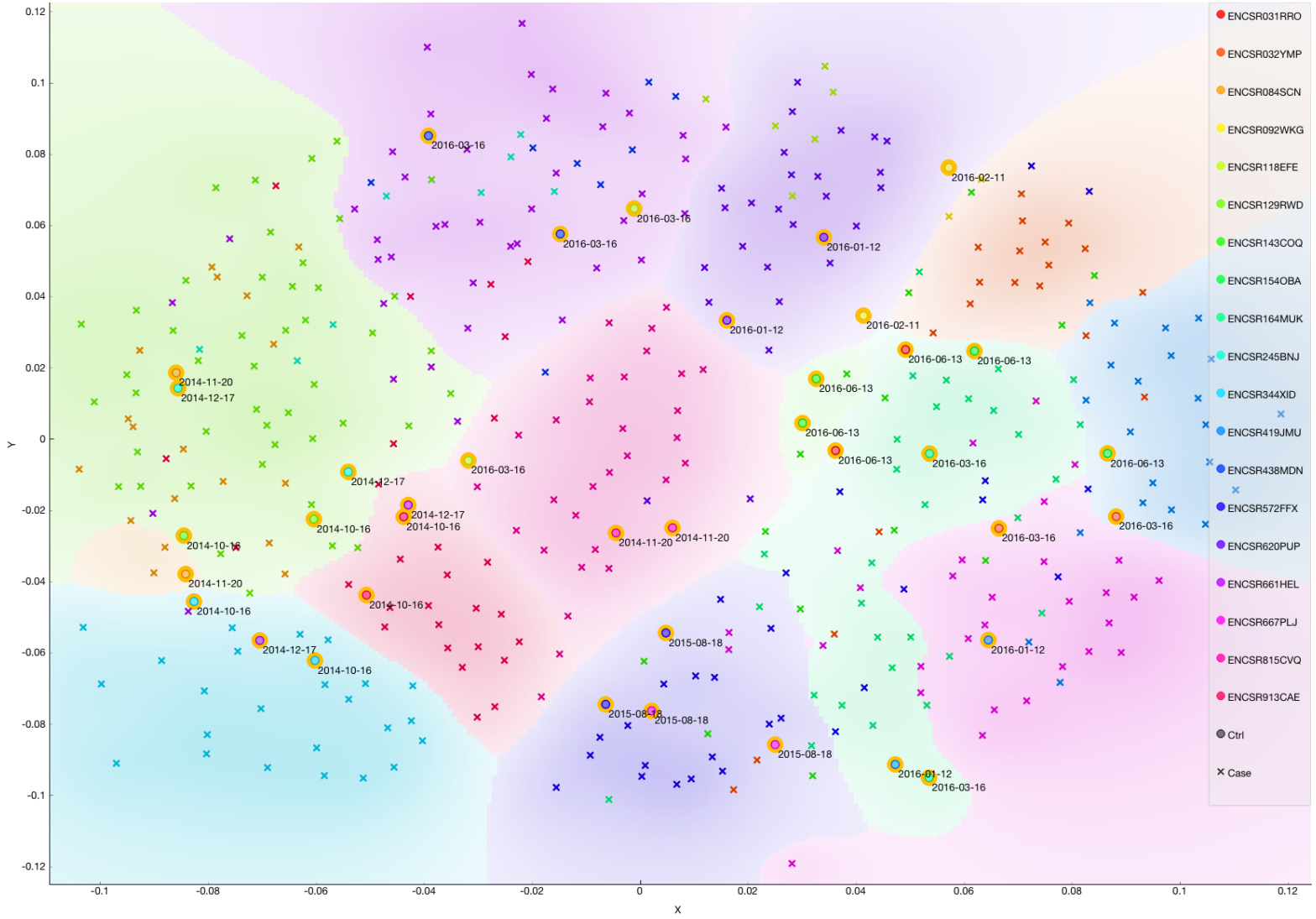
Supplementary Table 1: Summary of average distances for i) samples with the same control (distance within; dw) and ii) distances between samples of different controls (distance within; db) for PCA and MDS results. N, number of samples corresponding to each control.



Supplementary Figure 8: Results of the PCA. Fraction of explained variance per principal component (red), and the corresponding cumulative distribution (green).



Supplementary Figure 9: Results of the PCA. Scatter plot of the two components with highest explained variance. Each color represents samples controlled by a particular control experiment. RBP knockdowns are marked with a cross, while controls are marked with a circle.



Supplementary Figure 10: Scatter plot of the two-dimensional MDS analysis. Each color represents samples controlled by a particular control experiment. RBP knockdowns are marked with a cross, while controls are marked with a circle.

2.4 Binding and motif enrichment

In the following, we give a thorough definition of calculation of the binding and motif score enrichments. The results are presented for the *batch design* (Suppl. Fig. 11, Suppl. Table 12) and *case vs. control* design (Suppl. Fig. 12, Suppl. Table 13).

2.4.1 Binding enrichment signal (eCLIP)

We seek overlapping eCLIP tags at these regions and define the *binding signal* as enrichment in eCLIP tags on regions associated to regulated exonic parts (reg.), comparing to the background set (backg.). Specifically, we define the odds (fold enrichment) as ratio of the two probabilities

$$\text{(binding) odds}_i = \frac{P_i^{\text{reg.}}}{P_i^{\text{backg.}}} \tag{6}$$

where i is a sequence position in the interval [-150, +150] nt proximal to the 3' SS and $P_i^{\text{reg.}}$ is the probability of an eCLIP tag mapped to position i for the set of selected regulated (reg.) sequences, with $P_i^{\text{backg.}}$ defined analogously.

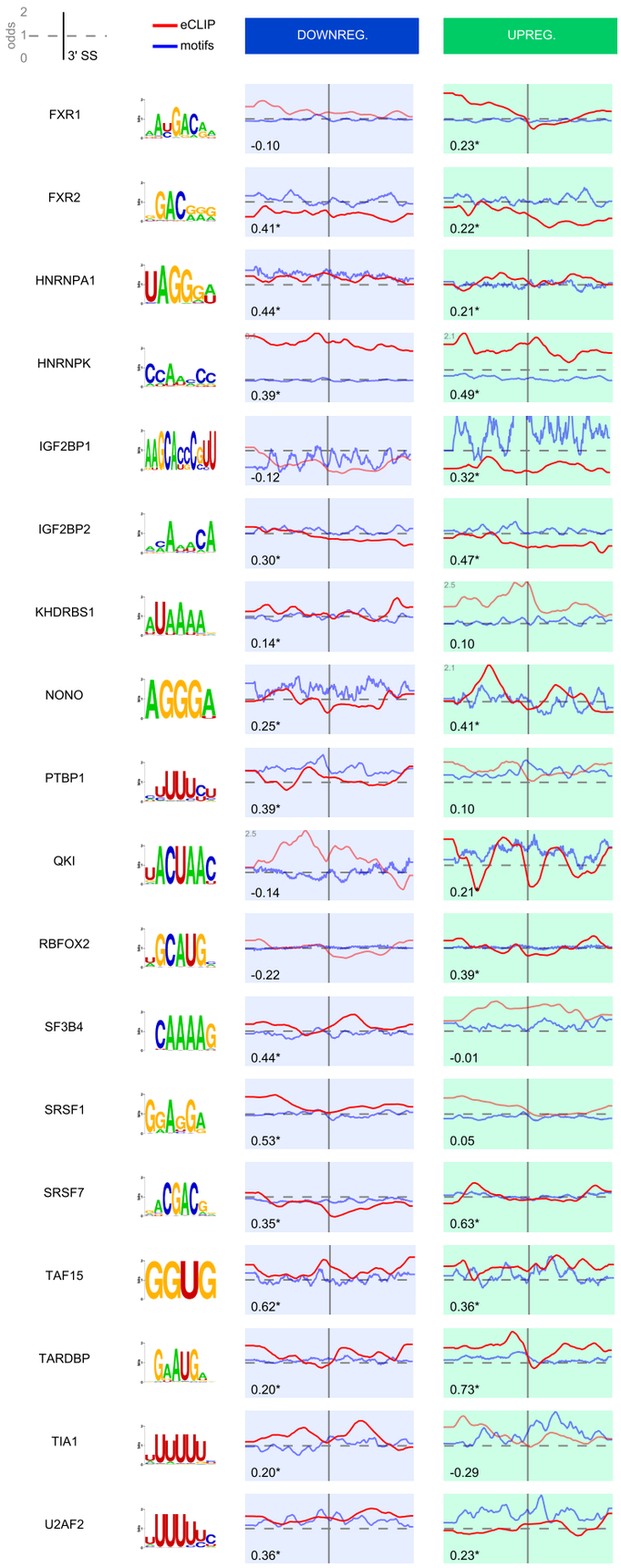
2.4.2 Motif enrichment signal

Similarly, we define a motif signal as follows. Let $\mathbf{W}$ be a $\ell \times 4$ motif probability matrix and $\mathbf{S}$ be a sequence binary matrix of size $L \times 4$ with a value of $\mathbf{S}_{i,a} = 1$ iff the nucleotide at position i is a. The odds of motif given by $\mathbf{W}$ at sequence position i are defined as the ratio of expected motif scores:

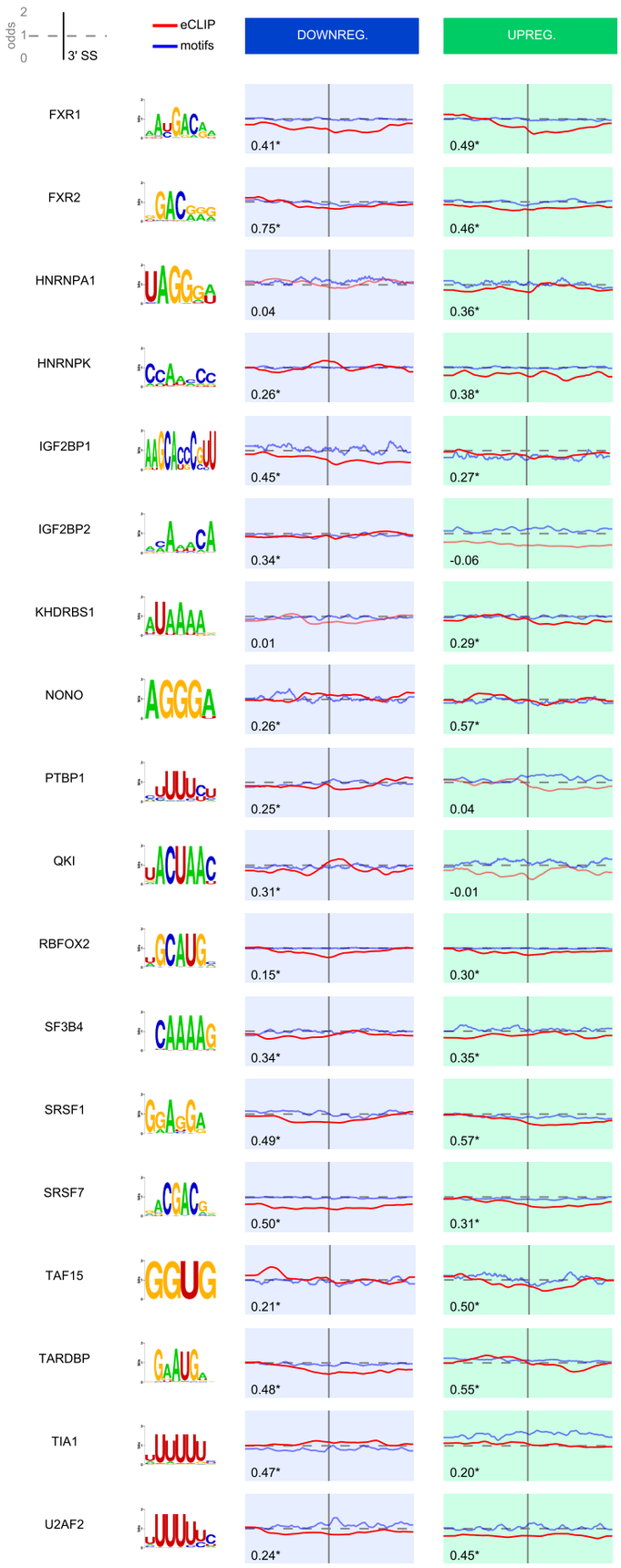
$$\text{(motif) odds}_i = \frac{\mathbb{E}_{\mathbf{S}\in\text{reg.}}\left[\sum_{a\in\{A,C,G,U\}}\sum_{j=1}^{\ell}\mathbf{S}_{i+j,a}\mathbf{W}_{j,a}\right]}{\mathbb{E}_{\mathbf{S}\in\text{backg.}}\left[\sum_{a\in\{A,C,G,U\}}\sum_{j=1}^{\ell}\mathbf{S}_{i+j,a}\mathbf{W}_{j,a}\right]} \tag{7}$$

where the expectations are taken over regulated (reg.) and background (backg.) sequences. Defined in this way, the odds give the ratio between the expected motif scores at sequence position i (**?**).

The alignment between binding and enrichment signals (vectors of fold enrichment within 300 nt regions) is scored using *cross-correlation*, defined as a maximal Pearson correlation when one of the signals is allowed to be shifted by at most 50 nt.



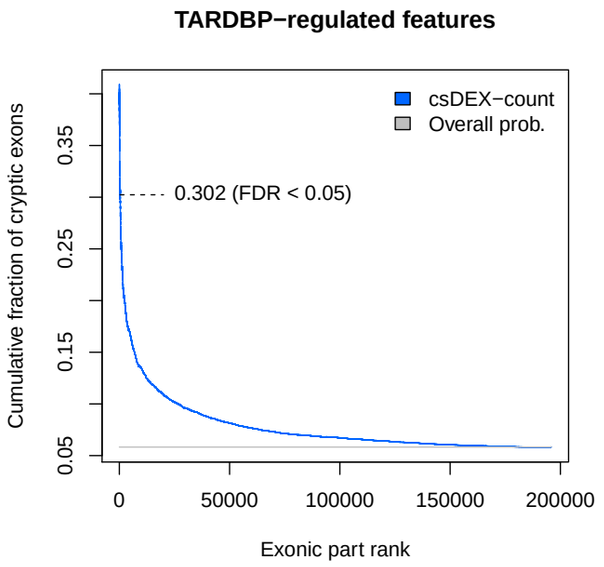
Supplementary Figure 11: *Batch design*; Fold enrichment (odds) of binding and motif score probabilities when comparing up- and down- regulated exonic parts for each RBP. A background (reference) set is composed of 20,000 non-regulated exonic parts. The plots show [-150, 150] nt regions centered at 3' splice sites (3' SS), represented by a black vertical line. The gray dashed line represents fold enrichment of 1 (i.e. no enrichment). The numbers represent values of cross-correlation with maximum allowed displacement of 50 nt. Line plots are shown in darker color when the cross-correlation > 0.15.



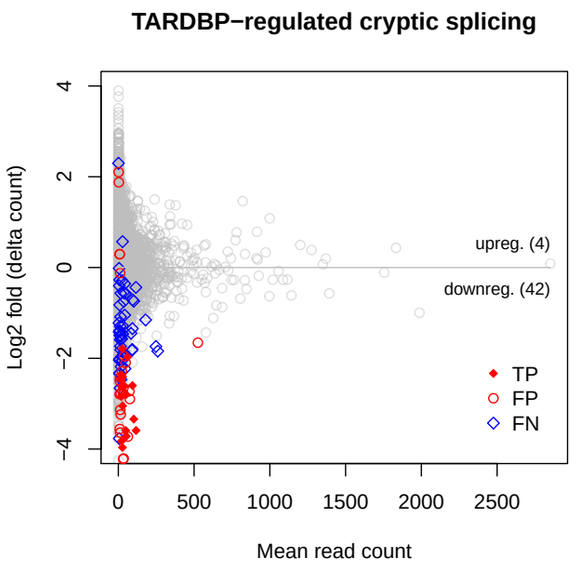
Supplementary Figure 12: *Case vs. control design*; Fold enrichment (odds) of binding and motif score probabilities when comparing up- and down- regulated exonic parts for each RBP. A background (reference) set is composed of 20,000 non-regulated exonic parts. The plots show [-150, 150] nt regions centered at 3' splice sites (3' SS), represented by a black vertical line. The gray dashed line represents fold enrichment of 1 (i.e. no enrichment). The numbers represent values of cross-correlation with maximum allowed displacement of 50 nt. Line plots are shown in darker color when the cross-correlation > 0.15.

2.5 Retrieving TARDBP-regulated cryptic exons

The dataset containing TARDBP regulated exons (see main text, Section 3.2.5). The cumulative probability of cryptic exons dependent on the exonic part rank (p-value returned by csDEX-count) is shown on Suppl. Fig. 13. The relationship between number of reads mapping to and exonic part and $\log_2$ fold change against expected expression is shown on Suppl. Fig. 14.



Supplementary Figure 13: Cumulative probability of cryptic exons among all TARDBP-specific differentially regulated exonic parts. The gray line shows the overall density of cryptic exons among all regulated exons (5.6%). The density of cryptic exons within the subset of exons predicted by csDEX-count is 30.2 %.



prediction	cryptic	non-cryptic	total
YES	20	26	46
NO	51	11319	11370
total	71	11345	11416

Supplementary Figure 14: The dependence of mean read count to model-predicted $\log_2$-fold change in expression (MA plot). All cryptic exonic parts are marked with a diamond () and others with a circle. The exonic parts downregulated by TARDBP, predicted by the csDEX-count model at FDR<5% are shown in red color. Not-retrieved true cryptic exons are shown in blue. The confusion matrix is shown in the table below.

3 Supplementary tables with experimental details

Supplementary Table 2: The genes with 5 to 15 unique exons and corresponding non-overlapping exonic parts. Merged with the *knownAlt* annotation. See the online file `Homo_sapiens.GRCh37.75.cassetteExon.5.15.tab`

Supplementary Table 3: The genes with 13 to 66 unique exons and corresponding non-overlapping exonic parts. Merged with the *knownAlt* annotation. See the online file `Homo_sapiens.GRCh37.75.cassetteExon.13.66.tab`

Supplementary Table 4: The list of BAM files aligned to the hg19 genome, used to evaluate the count-based models. See the online file `metadata-count.tsv`

Supplementary Table 5: The list of transcript quantification files aligned to the hg19 genome, used to evaluate the PSI-based models. See the online file `metadata-PSI.tsv`

Supplementary Table 6: The list of BAM files aligned to the hg38 genome, used to evaluate the retrieval of TARDBP cryptic exons. See the online file `metadata-TARDBP-hg38.tsv`

Supplementary Table 7: The list of BED files from eCLIP experiments in the K562 samples. See the online file `metadata-eCLIP.tsv`

Supplementary Table 8: The GFF annotation file including cryptic exonic parts provided by **?**. See the online file `Homo_sapiens.GRCh38.TARDBP.cryptic.gff.gz`

Supplementary Table 9: The high-confidence cryptic exons provided by **?**, retrieved on Mar 28, 2017. See the online file `hg38-cryptic-exons.tsv`.

Supplementary Table 10: Results of the Principal component analysis (PCA). The coordinates of two components with higest explained variance corresponding to each exonic part. See the online file `all_samples_rows_pca.csv`.

Supplementary Table 11: Results of Multidimensional scaling (MDS). The coorinates of the two-dimensional space retaining the Spearman correlation distance. See the online file `all_samples_rows_mds.csv`.

Supplementary Table 12: The binding enrichment at proximal regions for the upregulated and downregulated exonic parts in the dataset of genes with 13-66 exons. The RBP knockdowns are grouped in 19 batches, corresponding to the control experiment and the knockdown date. See the online file `rnamaps_batch.csv`.

Supplementary Table 13: The binding enrichement at proximal regions for the downregulated exonic parts in the dataset of genes with 13-66 exons. Each RBP knockdown is grouped with the corresponding control. See the online file `rnamaps_case.csv`.