Supplementary Results

Besides linear models we employed convolutional neural networks (CNN) to model the relationship between DNA methylation and splicing. This might help us to get more insights into the complex relationship between DNA methylation and the increased or decreased splicing rates.

For each exon, a sequence window of ±400 bp around the centre of the alternative exon was used to train a CNN that predicts inclusion (PSI ≥ 1/3) and exclusion (PSI < 1/3) of the alternative exon. We chose this sequence region since the linear models revealed the relevance of the alternative exon when predicting splicing rates. In addition, the alternative exon has been used in previous analyses [1, 2]. As most exons are shorter than 200 bp [3], the chosen sequence region mostly covers the whole exon as well as the most relevant regions from the adjunctive introns. Besides using a four-letter code for the four DNA bases, we also integrated methylation data in a five-letter code that includes methylated ('M') and un-methylated ('U') cytosines (Methods). The sequences of both four-letter and five-letter code were one-hot encoded and multiplied with the conservation score of the region, as derived from BRIE.

We obtained scores for each position of the sequence by a sliding window approach of 58 filters of length 10 bp. A bias of 1 was added. The rectified linear unit (relu) was used as activation function. Subsequent layers include a maximum pooling step with pooling length of 8 bp and a hidden layer with 50 nodes.

The cassette exons were split into training, validation and test set (60%, 20%, 20%). The model was trained until weight convergence (patience: 3 iterations, maximum: 10 iterations).

The five-letter code model performed slightly better than the four-letter code model (AUC=0.912 and 0.908, respectively) (**Figure S8**). Similar to the linear models that predict inclusion and exclusion of the alternative exon, the methylation information does not improve predictability of splicing much. The CNN performances are comparable to state-of-the-art models that were applied to bulk data [1, 2]. To our knowledge this is the first CNN trained on single-cell splicing data, and that allows inclusion of DNA methylation information.

Here we show that a deep neural network can be used to predict single-cell splicing with good performance. The inclusion of methylation information by designing a five-letter alphabet that includes methylated and unmethylated cytosines led to a minimal improved prediction of splicing levels. To make this model, which is to our knowledge the first deep model to predict splicing from DNA sequence in a single-cell setting, accessible to the scientific community we uploaded it to the kipoi.org model zoo, so it can be employed to make regulatory gene predictions using genomic and epigenomic variation. Our dataset did not contain enough samples to try and predict splicing variability using a deep neural network.

References

1. Wainberg M, Alipanahi B, Frey B. Does conservation account for splicing patterns? BMC Genomics. 2016;17:787. doi:10.1186/s12864-016-3121-4.
2. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347:1254806. doi:10.1126/science.1254806.
3. Sakharkar MK, Chow VTK, Kangueane P. Distributions of exons and introns in the human genome. In Silico Biol. 2004;4:387–93. doi:2004040032 [pii].

Supplementary Tables

Supplementary Table 1. R^2, effect and sign of features predicting sPSI using linear ridge regressions, for differentiation both time points. (6 sheets). **a.** effect sizes, sign and correlation value of each genomic feature to prediction of PSI in single iPS cells. **b.** effect sizes, sign and correlation value of each genomic feature to prediction of PSI in single endoderm cells. **c.** effect sizes, sign and correlation value of each genomic & mean methylation feature to prediction of PSI in single iPS cells. **d.** effect sizes, sign and correlation value of each genomic & mean methylation feature to prediction of PSI in single endoderm cells. **e.** effect sizes, sign and correlation value of each genomic & single cell methylation feature to prediction of PSI in single iPS cells. **f.** effect sizes, sign and correlation value of each genomic & single cell methylation feature to prediction of PSI in single endoderm cells.

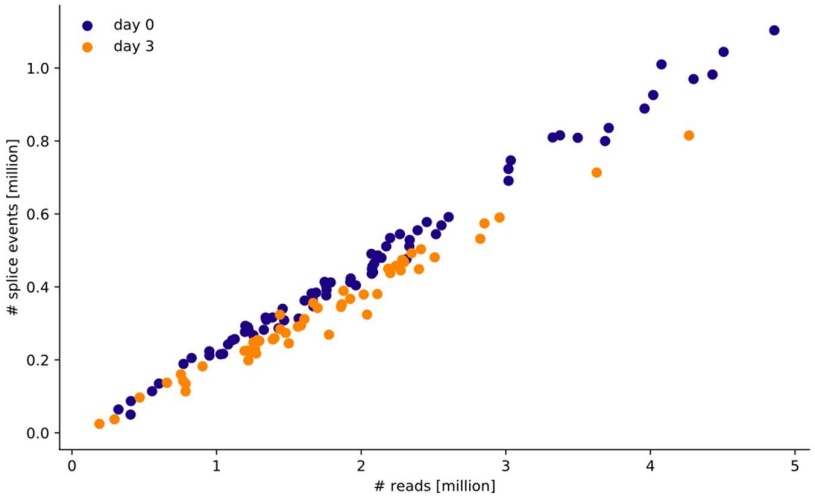
Supplementary Table 2. Contributions of features to PC1 in the PCA on all cell-specific feature weights. (2 sheets) **a**. contribution of features to PC1. **b**. contribution of features to PC2.

Supplementary Table 3. Assignment of splicing categories to cassette exons for both differentiation time points. (2 sheets). **a**. splicing category in iPS (day-0). **b**. splicing category in endoderm (day-3).

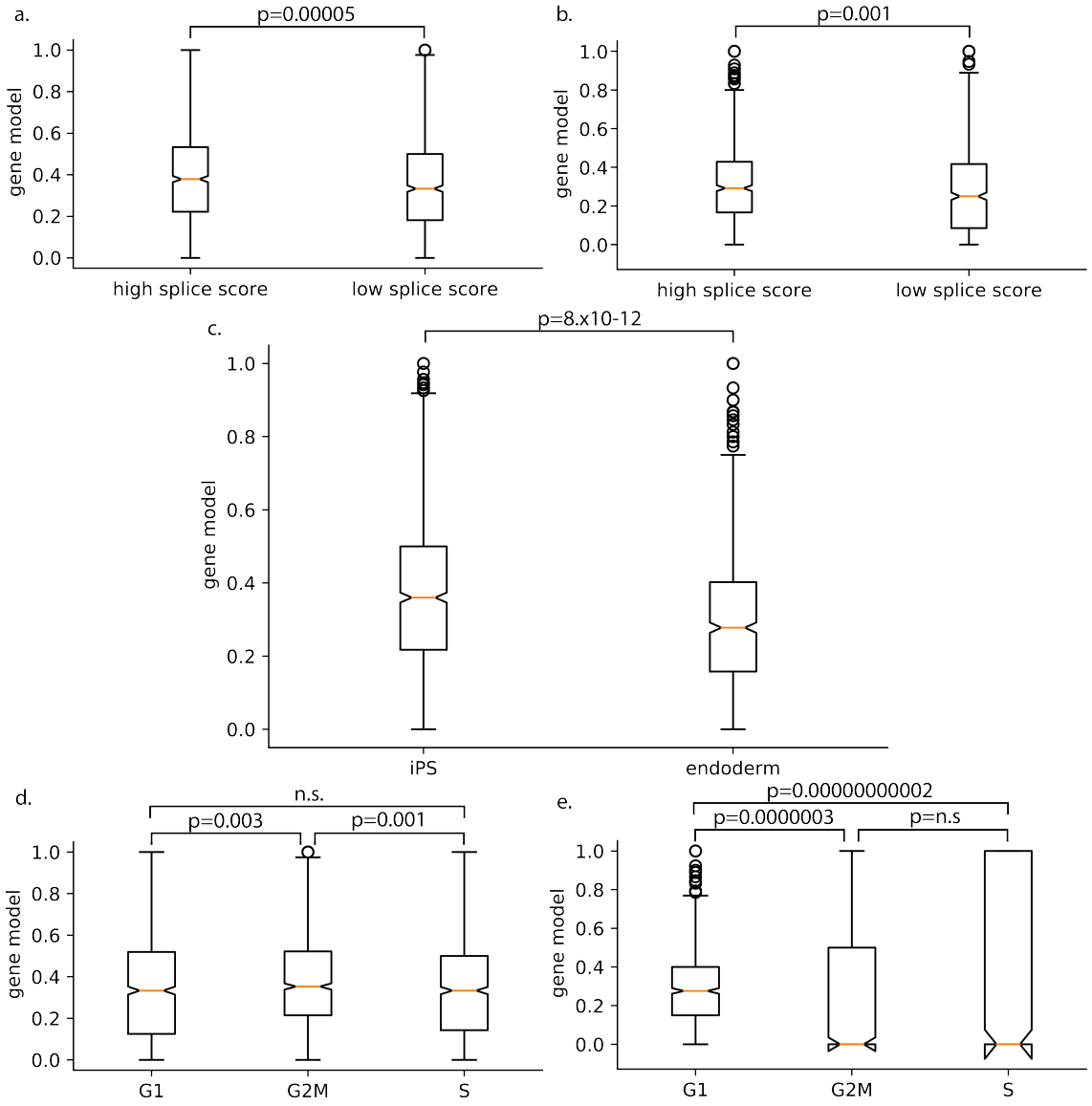
Supplementary Table 4. AUC, effect and sign of all features when predicting splicing categories in iPS cells (10-fold cross validation). The effect is defined as the absolute effect size weighted by the standard deviation of the feature and the sign is the sign of the effect size. **a-c.** Prediction of each splicing category based on genomic (**a**), genomic and methylation (**b**) and just mean methylation (**c**) features. **d-m.** Prediction of presence/absence of category switching between the two differentiation time points, for each splicing category, with and without methylation features. (13 sheets)

Tables are available at:
https://drive.google.com/open?id=1S13cstWYu6KdcdIr-bUcDWLTO0JkaWR9

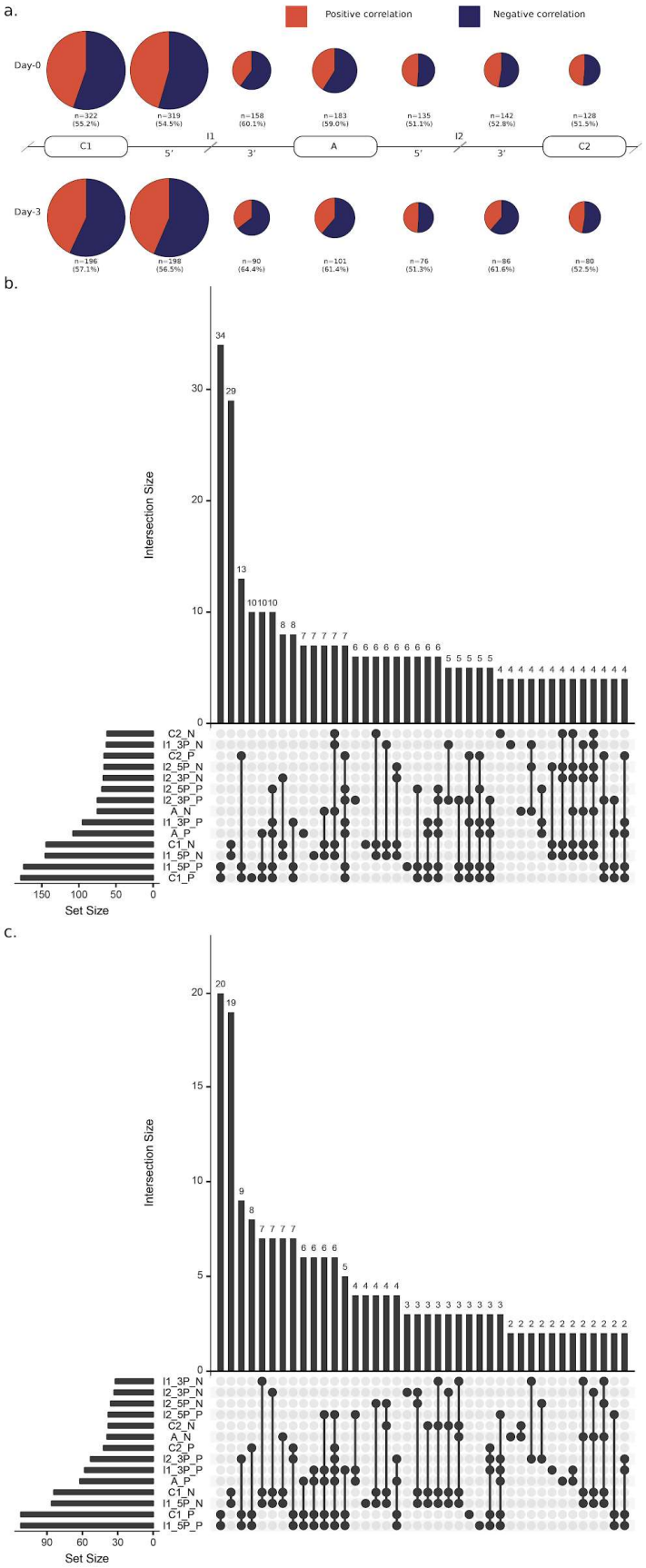
Supplementary figures



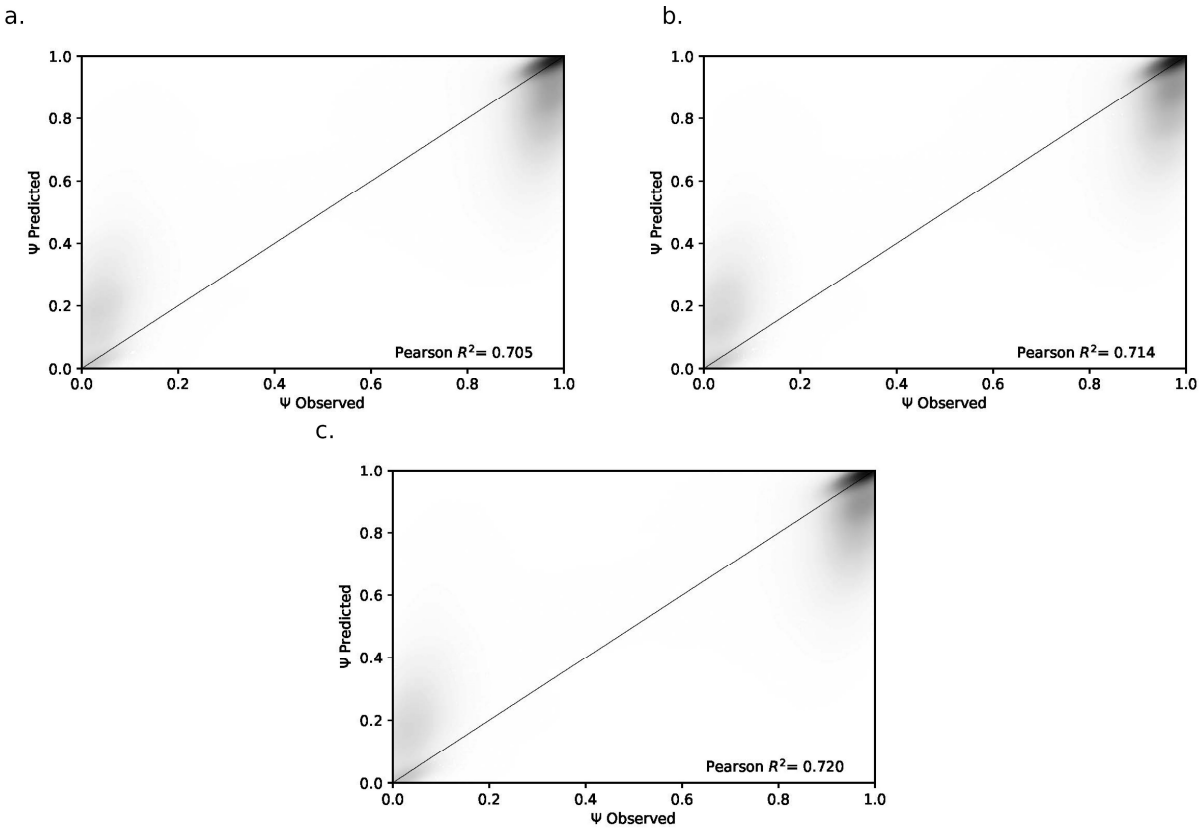
Supplementary Figure 1. Number of detected exon skipping events (minimum coverage 5 reads) versus the number of reads per single cell, for iPS (day-0, blue) and endoderm (day-3, orange) cells, respectively.



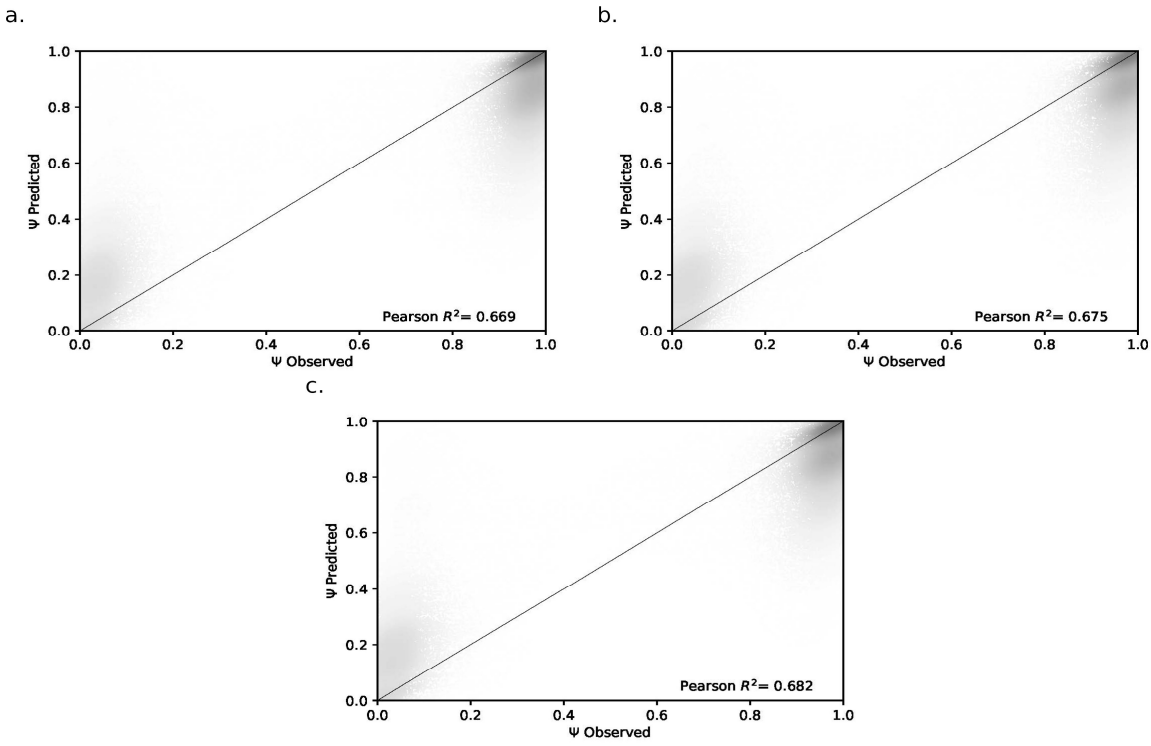
Supplementary Figure 2. Relationship between cell properties at intermediate splicing categories and splicing models (gene versus cell model) at the iPS (**a**,**c**,**e**) and endoderm (**b**,**d**,**e**) stage. Shown is the proportion of cells that follow the gene model as opposed to the cell model for alternative cell stratifications. **a.** Stratification of iPS cells by splicing activity (Methods), showing a reduction in the proportion of cells that follow the gene model for low splicing activity ($P=5x10^{-6}$, Wilcoxon test). **b.** Stratification of endoderm cells by splicing activity (Methods). A decreased proportion of single cells follows the gene model in the case of low splicing activity (P=0.001, Wilcoxon test). **c.** Stratification of iPS cells by cell cycle state (FACS measurements, Methods, G1 vs. G2M: P=0.003, S vs. G2M: P=0.001, Wilcoxon tests). **d.** Stratification of endoderm cells by cell cycle state (FACS measurements, Methods, G2M vs. G1: $P=3x10^{-7}$, S vs. G1: $P=2x10^{-11}$). **e.** Comparison between iPS and endoderm cells. An increased proportion of iPS cells follows the gene model in comparison to the endoderm cells ($P=8.x10^{-12}$).


a.
Positive correlation
Negative correlation
Day-0
n=322 (55.2%)
n=319 (54.5%)
n=158 (60.1%)
n=183 (59.0%)
n=135 (51.1%)
n=142 (52.8%)
n=128 (51.5%)
C1
5′
I1
3′
A
5′
I2
3′
C2
Day-3
n=196 (57.1%)
n=198 (56.5%)
n=90 (64.4%)
n=101 (61.4%)
n=76 (51.3%)
n=86 (61.6%)
n=80 (52.5%)
b.
Intersection Size
34
29
13
10 10 10
8 8
7 7 7 7 7
6 6 6 6 6 6 6 6 6
5 5 5 5 5
4 4 4 4 4 4 4 4 4 4 4 4 4
C2_N
I1_3P_N
C2_P
I2_5P_N
I2_3P_N
I2_5P_P
I2_3P_P
A_N
I1_3P_P
A_P
C1_N
I1_5P_N
I1_5P_P
C1_P
150 100 50 0
Set Size
c.
Intersection Size
20
19
9
8
7 7 7 7
6 6 6 6
5
4 4 4 4 4
3 3 3 3 3 3 3 3 3 3
2 2 2 2 2 2 2 2 2 2 2
I1_3P_N
I2_3P_N
I2_5P_N
I2_5P_P
C2_N
A_N
C2_P
I2_3P_P
I1_3P_P
A_P
C1_N
I1_5P_N
C1_P
I1_5P_P
90 60 30 0
Set Size

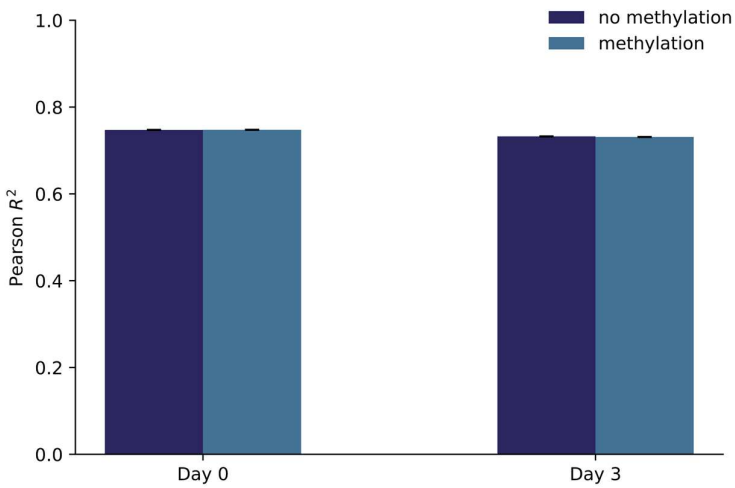
Supplementary Figure 3. Additional results for the single-exon association analysis between DNA methylation and splicing. **a.** Distributions of positive and negative associations between DNA methylation and splicing per sequence context for iPS (day-0) and endoderm (day-3) cassette exons, respectively. **b.** Upset plot showing the overlap between associations found in multiple sequence contexts including effect directions (in iPS cells). **c.** Upset plot showing the overlap between associations found in multiple sequence contexts including effect directions (in endoderm cells).



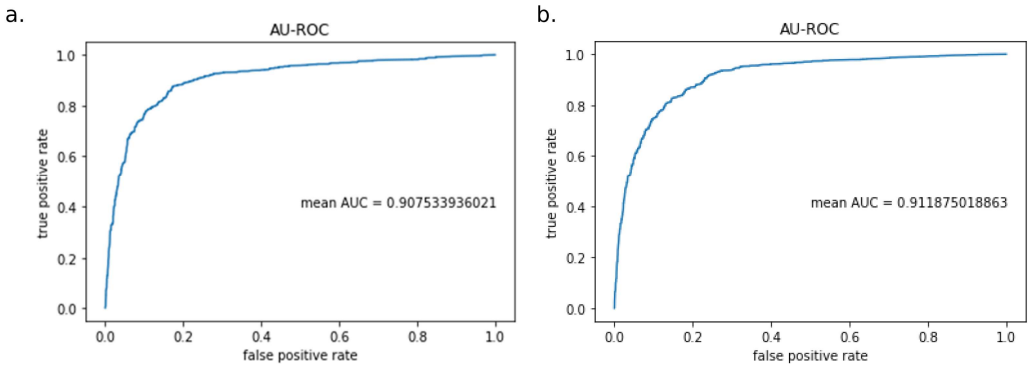
Supplementary Figure 4. Scatter plots between observed and predicted single-cell splicing rates and Pearson R^2 of splicing predictions for iPS cells (day-0). The three alternative regression models are based on different predictive features. **a.** Only genomic features. **b.** Genomic features and mean methylation information across cells. **c.** Genomic features and cell-matched methylation information.



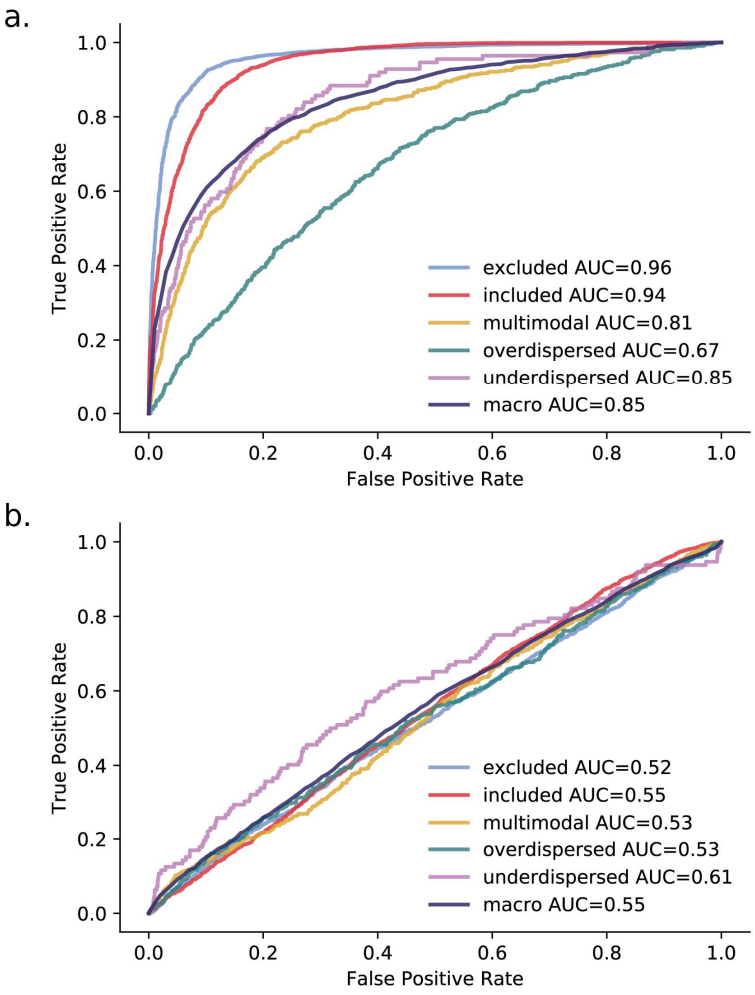
Supplementary Figure 5. Scatter plots between observed and predicted single-cell splicing rates and Pearson R^2 of splicing predictions for endoderm cells (day-3). The three alternative regression models are based on different predictive features. **a.** Only genomic features. **b.** Genomic features and mean methylation information across cells. **c.** Genomic features and cell-matched methylation information.



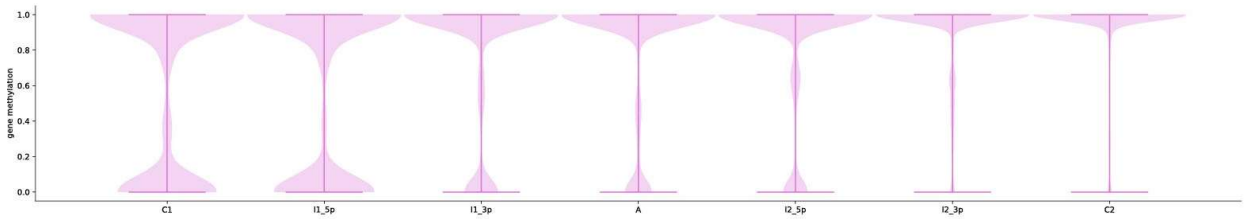
Supplementary Figure 6. Pearson R^2 of the linear models predicting pseudo-bulk splicing by aggregating read counts across cells (left: iPS (day-0) cells, right: endoderm (day-3) cells). A model that includes DNA methylation information (R^2=0.747, R^2=0.732 at day-0 and day-3), yields comparable prediction accuracy to a model without DNA methylation information (R^2=0.748, R^2=0.731, respectively).



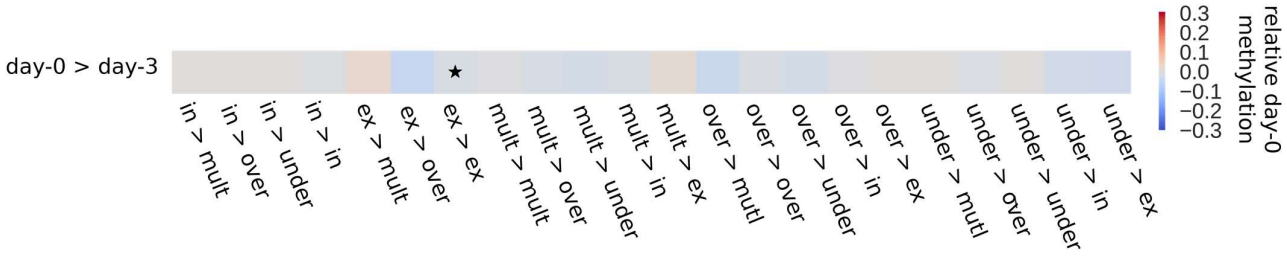
Supplementary Figure 7. AU-ROC curves and AUCs of a CNN when predicting splicing from genomic sequences in iPS (day-0) cells. **a.** Performance of the CNN based on only genomic information (four-letter code model). **b.** Performance of the CNN based on genomic and methylation information (five-letter code model).



Supplementary Figure 8. AU-ROC curves and AUCs when predicting splicing categories with one-vs.-rest logistic ridge regression models and different sets of predictive features. The performances of each individual splicing category and the macro performance across all categories are shown. **a.** Genomic and DNA methylation features. **b.** Mean methylation features.



Supplementary Figure 9. The distribution of DNA methylation rates per sequence context across overdispersed cassette exons of iPS (day-0) cells.



Supplementary Figure 10. iPS (day-0) vs. endoderm (day-3) DNA methylation rate differences of the category-switching cassette exons. A significant decrease in DNA methylation rates is observed between the excluded day-0 and excluded day-3 categories. Labels shown at the bottom: included (in), excluded (ex), multimodal (mult), overdispersed (over), underdispersed (under).