Combined single cell profiling of expression and DNA methylation reveals splicing regulation and heterogeneity

Stephanie Linker[1,2,#], Lara Urban[1,2,#], Stephen Clark[3], Mariya Chhatriwala[4], Shradha Amatya[4], Davis McCarthy[1,2], Ingo Ebersberger[5,6], Ludovic Vallier[4,7,8], Wolf Reik[3,4,9], Oliver Stegle[1,2,*,@], Marc Jan Bonder[1,2,*,@]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK
[2]European Molecular Biology Laboratory, Genome Biology Heidelberg, Germany
[3]Epigenetics Programme, The Babraham Institute, Cambridge, UK.
[4]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
[5]Applied Bioinformatics Group, Institute of Cell Biology and Neuroscience, Goethe University Frankfurt, Max-von-Laue-Str. 13, 60438 Frankfurt am Main, Germany
[6]Senckenberg Biodiversity and Climate Research Centre (BiK-F), Frankfurt am Main, Germany.
[7]Wellcome Trust – MRC Cambridge Stem Cell Institute, Anne McLaren Laboratory, University of Cambridge, Cambridge CB2 0SZ, UK
[8]Department of Surgery, University of Cambridge, Cambridge CB2 0QQ, UK
[9]Centre for Trophoblast Research, University of Cambridge, Cambridge, UK

shared first authors
* shared last authors
@ corresponding authors

Abstract

Background

Alternative splicing is a key mechanism in eukaryotic cells to increase the effective number of functionally distinct gene products. Using bulk RNA sequencing, splicing variation has been studied both across human tissues and in genetically diverse individuals. This has identified disease-relevant splicing events, as well as associations between splicing and genomic variations, including sequence composition and conservation. However, variability in splicing between single cells from the same tissue and its determinants remain poorly understood.

Results

We applied parallel DNA methylation and transcriptome sequencing to differentiating human induced pluripotent stem cells to characterize splicing variation (exon skipping) and its determinants. Our results shows that splicing rates in single cells can be accurately predicted based on sequence composition and other genomic features. We also identified a moderate but significant contribution from DNA methylation to splicing variation across cells. By combining sequence information and DNA methylation, we derived an accurate model (AUC=0.85) for predicting different splicing modes of individual cassette exons. These explain conventional inclusion and exclusion patterns, but also more subtle modes of cell-to-cell variation in splicing. Finally, we identified and characterized associations between DNA methylation and splicing changes during cell differentiation.

Conclusions

Our study yields new insights into alternative splicing at the single-cell level and reveals a previously underappreciated link between DNA methylation variation and splicing.

Keywords

Single-cell analysis, Alternative splicing, DNA methylation, Splicing prediction, Cell differentiation, Multi-omics

Background

RNA splicing enables efficient gene encoding and contributes to gene expression variation by alternative exon usage [1]. Alternative splicing is pervasive and affects more than 95% of human genes [2]. Splicing is known to be regulated in a tissue-specific manner [3, 4] and individual splicing events have been implicated in human diseases [5]. Bulk RNA-sequencing of cell populations has been used to identify and quantify different splicing events [6], where in particular exon skipping at cassette exons, the most frequent alternative splicing event [1], has received considerable attention.

Different factors have been linked to splicing of cassette exons, including sequence conservation [7] and genomic features such as the local sequence composition and the length of the exon and flanking introns [8, 9]**.** Although there is some evidence for an epigenetic component of splicing regulation, this relationship is not fully understood and alternative models for the role of DNA methylation in splicing have been proposed [10–12]. The transcriptional repressor CTCF has been shown to slow down RNA polymerase II (Pol II), resulting in increased exon inclusion rates. By inhibiting CTCF binding, DNA methylation can cause reduced exon inclusion rate [10]. Alternatively, increased DNA methylation of the MeCP2 pathway has been associated to increased exon inclusion rates. MeCP2 recruits histone deacetylases in methylated contexts that wrap the DNA more tightly around the histones. This interplay between MeCP2 and DNA methylation slows down Pol II and lead to an increased exon inclusion rate [11]. Finally, HP1, which serves as an adapter between DNA methylation and transcription factors, increases the exon inclusion rate if it is bound upstream of the alternative exon. Binding of HP1 to the alternative exon leads to increased exon skipping [12]. These alternative mechanisms point to a complex regulation of splicing via an interplay between DNA sequence and DNA methylation, both in proximal as well as distal contexts of the alternative exon.

Technological advances to perform RNA-seq in single cells have most recently enabled studies that have started to investigate splicing variation at single-cell resolution [9, 13, 14]. Leveraging recent protocols for parallel sequencing of RNA and bisulfite treated DNA from the same cell (single-cell methylation and transcription sequencing; scM&T-seq [15]), we here extend such analysis by accounting the DNA methylome into account. For the first time, we study associations between single-cell splicing variation and DNA methylation at two stages of human iPS differentiation.

Results

Single-cell splicing variation during endoderm differentiation

We applied parallel single-cell methylation and transcriptome sequencing (scM&T-seq) to differentiating induced pluripotent stem (iPS) cells from a single donor ("joxm_1") of the Human Induced Pluripotent Stem Cell Initiative (HipSci) [16, 17]. We profiled 93 cells in the iPS state, as well as following three days of differentiation towards definitive endoderm (endoderm). After quality control this resulted in 86 and 59 cells, respectively (Methods), which were used for analysis. In each cell we quantified cassette exon inclusion rates (methods, **Fig. 1a**). We detected and quantified splicing for between 1,386 and 14,434 exons per cell (minimum coverage five reads), where splicing rates (PSI) were estimated as the fraction of reads that include the alternative exon versus the total number of reads at the cassette exon (Methods). Sequencing depth and differentiation stage were the most important determinants of differences between cells (**Figure S1**). We considered 6,282 (iPS) and 4,096 (endoderm) cassette exons that were detected in at least ten cells for further analysis.

Initially, we explored whether individual cells express only a single splice isoform ("cell model"), or whether multiple isoforms are present in a given cell ("gene model"; **Fig. 1b**), a question that has previously been investigated in bulk data [18, 19]. Globally, our data (**Fig. 1c**) rule out the cell model, however we also observed deviations from the gene model, in particular for exons with intermediate levels of splicing (**Fig. 1c**). We assessed relationships between cellular properties and the consistency with the two splicing models, focusing on the intermediate splicing ranges ($0.2 < PSI < 0.8$, **Fig. 1c**). This identified differences between cells with high and low splicing activity (defined as the abundance of splice factors; Methods, $P=5x10^{-5}$ and $P=0.001$ for iPS and endoderm, respectively, **Figures S2a-b**). Additionally, we observed that iPS cells have splicing patterns that are more consistent with the gene model as compared to differentiated cells ($P=8.x10^{-12}$, **Figure S2c**). Finally, we observed an enrichment for cells in G2/M cell cycle stage for splicing according to the gene model in iPS cells, and G1 cells in endoderm (**Figures S2d-e**).

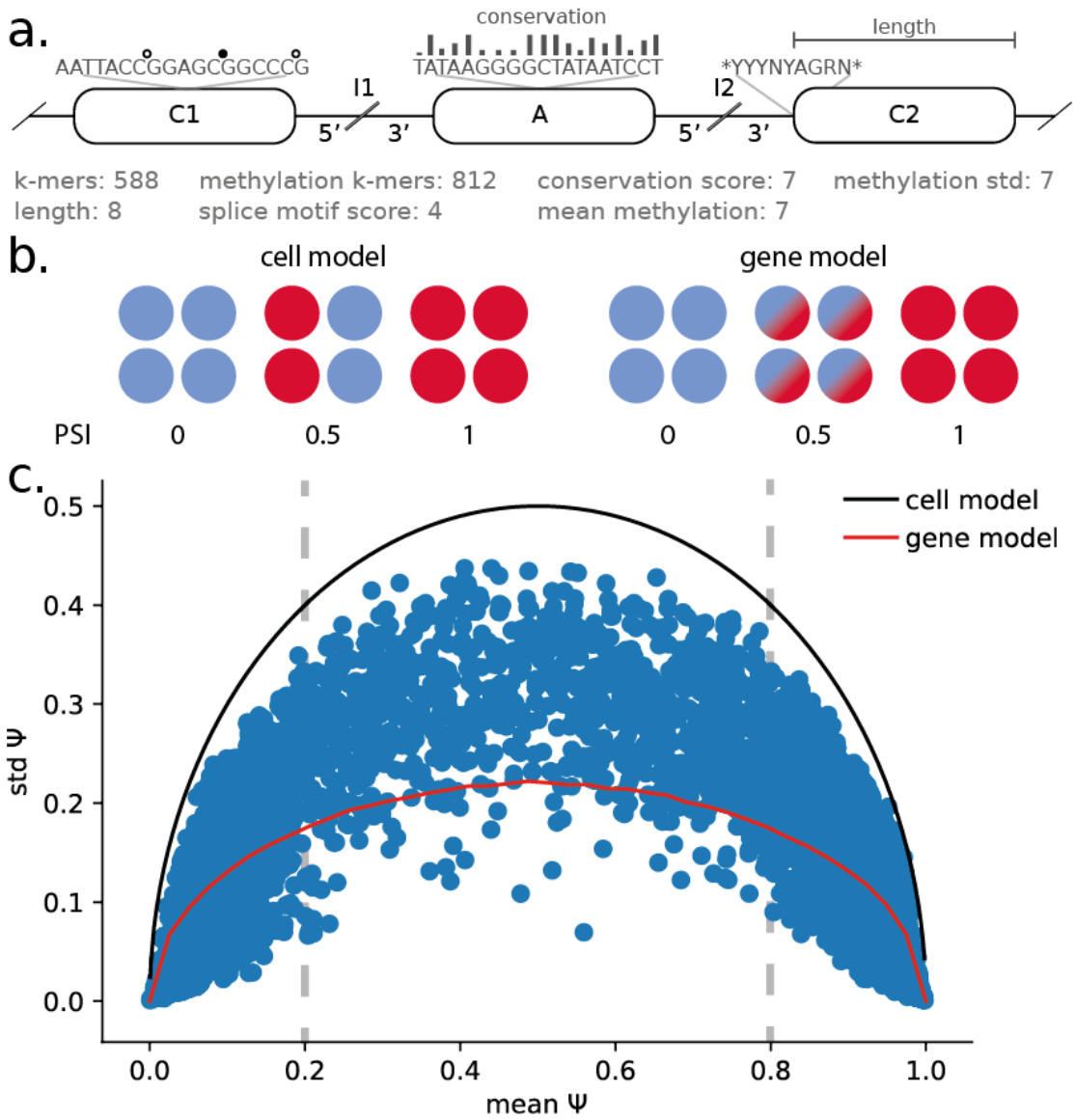

Figure 1 | **Single-cell splicing and considered features for modeling splicing rates. a.** Illustration of the considered sequence contexts (top) and the total number of extracted features (bottom). Sequence contexts that were considered to extract genomic and epigenetic splicing features. “A” denotes the alternative exon, “I1” and “I2” correspond to the upstream and downstream flanking introns and “C1” and “C2” to the upstream and downstream flanking exons, respectively. The 5’ and 3’ ends (300bp) of flanking introns are considered separately. **b.** Illustration of two canonical splicing models. The "cell model" assumes that splicing variation is due to differential splicing between cells, where each cell expresses one of two splice isoforms. The "gene model" corresponds to the assumption that both splice isoforms can be expressed in the same cells. **c.** Cell-to-cell variation in splicing rate across cells (standard deviation of PSI) as a function of the average inclusion rate of cassette exons, considering 86 iPS cells. Solid lines correspond to the expected trends either assuming a "cell model" (black line) or the "gene model" (red line). Grey vertical dashed lines mark intermediate ranges of splicing (0.2 < PSI < 0.8).

Methylation heterogeneity across cells is associated with splicing variability

Next to try and identify the locus specific link between DNA methylation heterogeneity and splicing, we tested for associations between differences in DNA methylation levels across cells and splicing rates (Spearman correlation; Methods). For each cassette exon, we tested for associations between variation in DNA methylation and splicing in each of seven sequence contexts: the three exons, and the 5' and 3' parts of the two introns (methods, **Fig. 1a**). Genome-wide, this analysis identified 424 cassette exons with a methylation-splicing associations in iPS cells (out of 5,280 tested cassette exons, $Q < 0.05$, **Figure S3a, Table S1**) and 253 associations in endoderm cells (out of 2,622 tested, $Q < 0.05$, **Figure S3a, Table S1**). DNA methylation variation in the upstream alternative exon was most frequently associated with splicing variation (~60%), with approximately equal numbers of negative and positive effects (55% negative in iPS, 57% negative in endoderm). Most associations could be detected in more than one context for a given exon with consistent effect directions (**Figures S3b, S3c**). Similarly, we observed largely concordant effects across differentiation stages, with 87% of the associations detected in endoderm cells also being significant in iPS cells. Our associations point to selected set of genes with a relationship between DNA methylation and splicing of specific genes. Genes with negative associations between DNA methylation and splicing were enriched for HOXA2 transcription factor binding sites (adjusted $P=2.3x10^{-2}$ and adjusted $P=1.2x10^{-3}$ in iPS and endoderm cells; using G:Profiler). Genes with a positive methylation-splicing associations were enriched for LHX3 transcription factor binding at the iPS state (adjusted $P=3.3e^{-2}$), while no enrichments were observed in endoderm cells.

Prediction of splicing at single-cell level

To gain insights into the global determinants of splicing, we trained regression models that related genomics and epigenetic features (see **Fig. 1a**) to splicing rates in single cells. Briefly, we pooled splicing information from cassette exons across cells and trained separate regression models for iPS and endoderm cells. Initially, we considered 607 features that explain sequence composition (based on k-mers) and sequence conservation ("genomic" features, Methods). We considered an additional set of up to 826 features derived from DNA methylation, including an extended k-mer alphabet that takes the methylation status into account, as well as DNA methylation average and variance (across CpG sites) in each of the seven sequence contexts of the exon per cell or across cells for mean methylation models (Methods). Methylation features were either incorporated on bulk average ("genomic & mean methylation" features) or individual cell level ("genomic & cell methylation" features).

Notably, the model to predict single-cell splicing based on genomic features yielded comparable performance to previous attempts to predict splicing using bulk [8] and single-cell [9] RNA-seq (R^2=0.706, R^2=0.670; assessed using 10-fold cross validation; **Fig. 2a**, **Figures S4, S5**). To facilitate the comparison with previous results using bulk RNA-seq, we also considered predicting aggregate methylation rates across cells ("pseudo bulk PSI" (bPSI)), which resulted in similar prediction accuracies (R^2=0.747 and R^2=0.732 for iPS and endoderm cells, **Figure S6**). The inclusion of DNA methylation features increased the prediction accuracy, where larger gains were observed when including cell-matched DNA methylation information ("genomic & cell methylation" versus "genomic & mean methylation"). This in combination with our previous results suggests that DNA methylation is most predictive of cell-to-cell variation in splicing at the same locus, whereas genomic features capture variation across different loci.

Next, to assess the relevance of individual features, we built equivalent single-feature models for individual cells. Consistent with previous bulk studies [7, 8], this identified features derived from the alternative exon and its neighboring contexts, i.e. the 3' end of the upstream intron and the 5' end of the downstream intron, as most informative (**Table S2**). Within these contexts, sequence conservation of the alternative exon was found to be most relevant individually. Other high-ranking features included the k-mers CT, CTC and CCT of the alternative exon (**Fig. 2b)**, sequence patterns that show close resemblance to CTCF binding motifs, which has previously been linked to alternative splicing. However, unlike the previously described CTCF motifs that are found upstream of the alternative exon, these k-mers are located in the alternative exon and have an opposite effect direction [10, 20].

Although the most important features were consistent between iPS and endoderm cells (R^2=0.79, average correlation between weights across all cells), principal component analysis (PCA) applied to the feature relevance information from all cells identified more subtle coordinated changes of the feature relevance (**Fig. 2c**). The first two principal components (PC) clearly separate iPS from endoderm cells, differences that were mainly driven by k-mers of the downstream intron (I2) that contain methylated and unmethylated cytosine bases (**Fig. 2d**). This points towards a combination of differences in sequence composition, potentially transcription factor activity, and DNA methylation as the main determinants of cell-type specific splicing regulation (**Table S3**).

Finally, we considered more complex regression models based on convolutional neural networks to predict single-cell splicing based on DNA sequence and an extended genomics alphabet including base-level DNA methylation information (deposited at kipoi.org; methods). We observed only a limited increase in performance when including DNA methylation information (**Supplementary Results**, **Figure S7**). These results line up with our locus specific DNA methylation results and the linear regression results. Strengthening our idea that global splicing information is mainly encoded by DNA sequence and conservation, and DNA methylation is linked to splicing a locus specific manner.

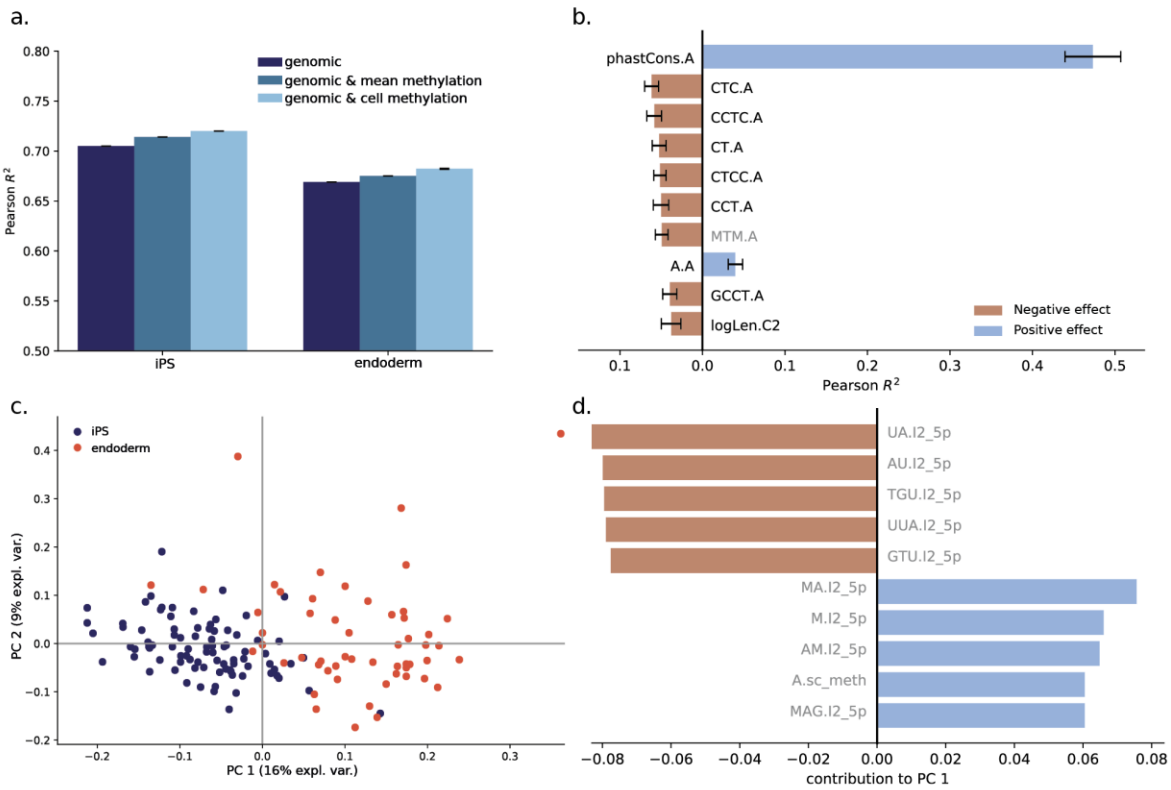

Figure 2. **Regression-based prediction of single-cell splicing variation. a.** Prediction accuracy (Pearson R^2 based on 10-fold cross validation) of alternative regression models for single-cell splicing rates in iPS cells (day-0) and endoderm cells (day-3). The genomic model is based on sequence k-mers, conservation scores and lengths of contexts (size of the cassette exon, length of flanking introns) as features (genomic features, dark blue). Other models account for average methylation rates across cells (genomic & mean methylation, blue), or cell-specific methylation rates (genomic & cell methylation, light blue). Error bars denote plus or minus one standard deviation across four repeat experiments. **b.** Features ranked by relevance for predicting splicing in iPS cells as determined by single-feature regression models trained on single cells. The most important features are features of the alternative exon, and include a methylated k-mer. Error bars denote plus or minus one standard deviation of the feature relevance across cells. Methylation features are indicated in grey. **c.** Principal component analysis on all cell-specific feature weights as shown in **b**. The first principal component (PC) primarily captures differences between differentiation states. **d.** The ten features with the largest contribution to the first PC (five positive and five negative features), which include k-mers with methylation information of the downstream intron I2. Methylation features are shown in grey.

Prediction of splicing modes of individual exons

Next, we set out to study differences between different exons and their splicing patterns. We classified cassette exons into five distinct categories, using scheme that is similar to Song *et al.* [13]: 1) excluded, 2) included, and three intermediate splicing categories: 3) overdispersed, 4) underdispersed and 5) multimodal (**Fig. 3a, 3b, Table S4**, Methods). We trained multinomial regression models (Methods) to classify individual exons using analogous features sets as considered for the regression models on single-cell splicing. A model based on genomic features yielded a macro-average AUC of 0.85 (**Fig. 3c**), where again sequence conservation in different contexts was the most informative feature (**Table S5**). Interestingly, we observed differences in the feature relevance across splicing categories: i) included and excluded exons, where the most relevant features were located in the alternative exon, and ii) the intermediate splicing categories, where features of the flanking exons were most informative. In general, predictions of included and excluded exons were most accurate (AUC=0.96 for both in iPS, AUC=0.94 for included in endoderm, AUC=0.96 for excluded in endoderm cells, **Fig. 3d**). These prediction accuracies exceed previously reported results in bulk data [8]. Even higher accuracies were achieved when training a model to differentiate between included and excluded exons only (AUC=0.99), whereas lower prediction accuracies were achieved for differentiating just the intermediate splicing categories (AUC=0.7 to 0.9, **Table S5**). The inclusion of methylation features did not improve the prediction performance (**Fig. 3d**, **Figure S8a**). Consistent with this, we also found that a model based on DNA methylation alone did not yield accurate predictions, although methylation contained some information for identifying underdispersed cassette exons (**Fig. 3d**, **Figure S8b**). Given this, we investigated the distribution of DNA methylation patterns across splicing categories, observing distinct distributions of DNA methylation in the upstream exon of underdispersed cassette exons (**Fig. 3e**). This effect was consistent, although less pronounced, in other sequence contexts (decreasing from the upstream to the downstream exon, **Figure S9**).

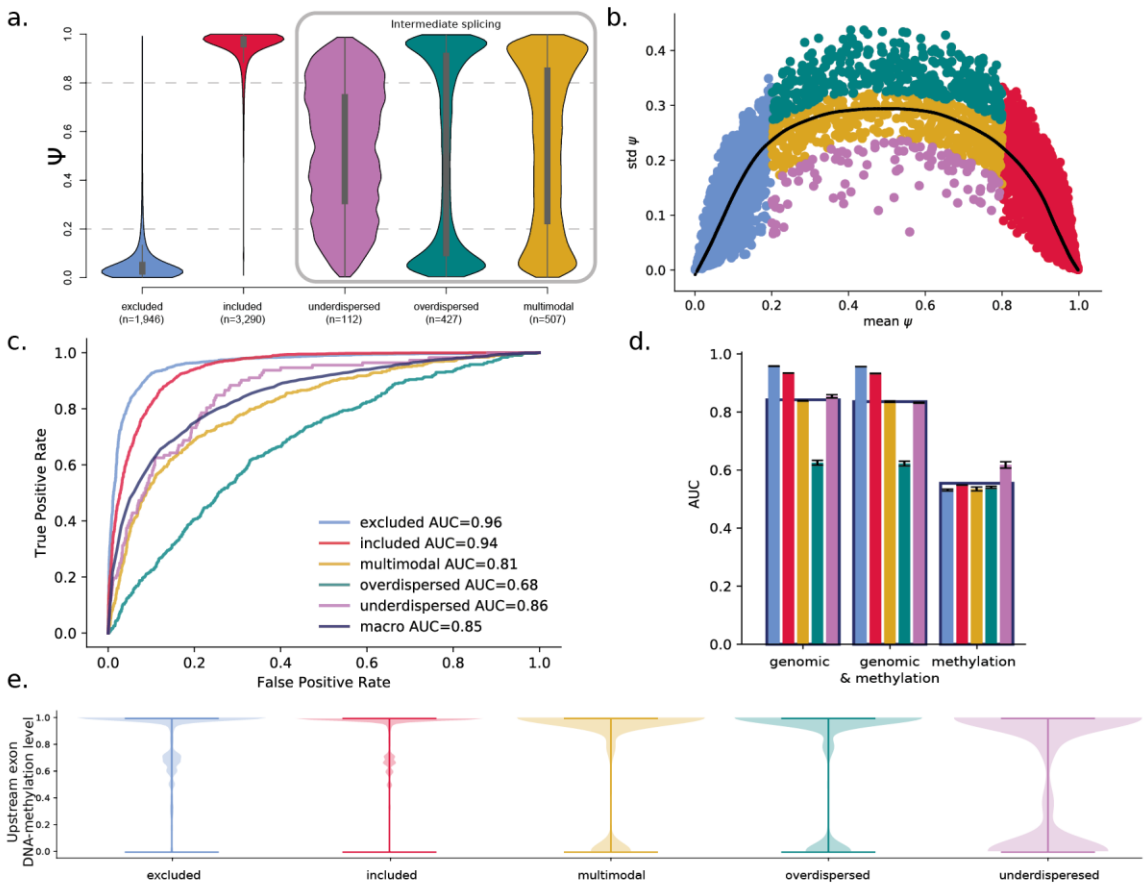

Figure 3. **Classification of cassette exons based on their single-cell splicing patterns. a.** Splicing rate (PSI) distributions for five splicing categories, inspired by Song *et al.* [13]. The intermediate splicing categories can only be defined based on single cell information are framed using a grey box. **b.** Variation of PSI (standard deviation) across cells as a function of the average inclusion rate of cassette exons across 86 iPS cells, colored according to their respective splicing category as defined in **a**. The solid black line denotes the LOESS fit across all cassette exons. **c.** Prediction performance of logistic regression for predicting splicing categories based on genomic features. Shown is the receiver operating characteristics for each splicing category and the macro average (area under the curve, AUC)**. d.** Prediction performance of alternative regression models for each splicing category, either considering a model trained using genomic features ('genomic', left), genomic and all DNA methylation features ('genomic & methylation', center) as well as only DNA methylation features ('methylation', right). The genomic model includes k-mers, conservation scores and region lengths (see Fig 1a). The genomic and methylation model additionally includes DNA methylation features. The methylation model includes average DNA methylation features per sequence context. Splicing categories are coded in color as in **a**. Error bars denote plus or minus one standard deviation across four repeat experiments. **e**. Distribution of DNA methylation levels in the upstream exon (C1) per splicing category. Methylation is decreased in underdispersed exons.

Splicing category switches across cell differentiation

Finally, we assessed changes in the splicing category switched between differentiation stages. Similar to previous observations in the context of neuronal iPS differentiation [13], we observed that a majority (88%) of the cassette exons retained their category during differentiation (**Fig. 4a**). We also observed no cassette exon that switched from included to excluded or *vice versa.* Instead, most (55%) of the switching events were observed within the three intermediate splicing categories. The most prevalent switch events were changes to the multimodal category; 51% of the underdispersed and nearly 45% of the overdispersed cassette exons in iPS cells switched to multimodal at the endoderm state.

Observing the category switches between the differentiation stages we set out to build a final set of logistic ridge regression models based on genomic and methylation features to predict category switching ability of cassette exons during differentiation (**Fig. 4b** for prediction performance, **Table S5**). This model had limited power to predict category switches, and DNA methylation did not significantly improve the prediction of any class although moderately higher predictions can be seen for the switching behavior of over- and underdispersed cassette exons.

Lastly we assessed if DNA methylation changed within the cassette exons switching between the differentiation time points. The DNA methylation levels of cassette exons that switched category only changed minimally between the differentiation time points (**Figure S10**). However, we observed that DNA methylation of the alternative exon of switching cassette exons differed from non-switching cassette exons at the iPS stage (**Fig. 4c**). DNA methylation of both, switching included and switching excluded cassette exons, was increased around C1 in comparison to their relevant non-switching counterparts. In the case of switching overdispersed cassette exons, we observed higher DNA methylation levels within and in the vicinity of the alternative exon.

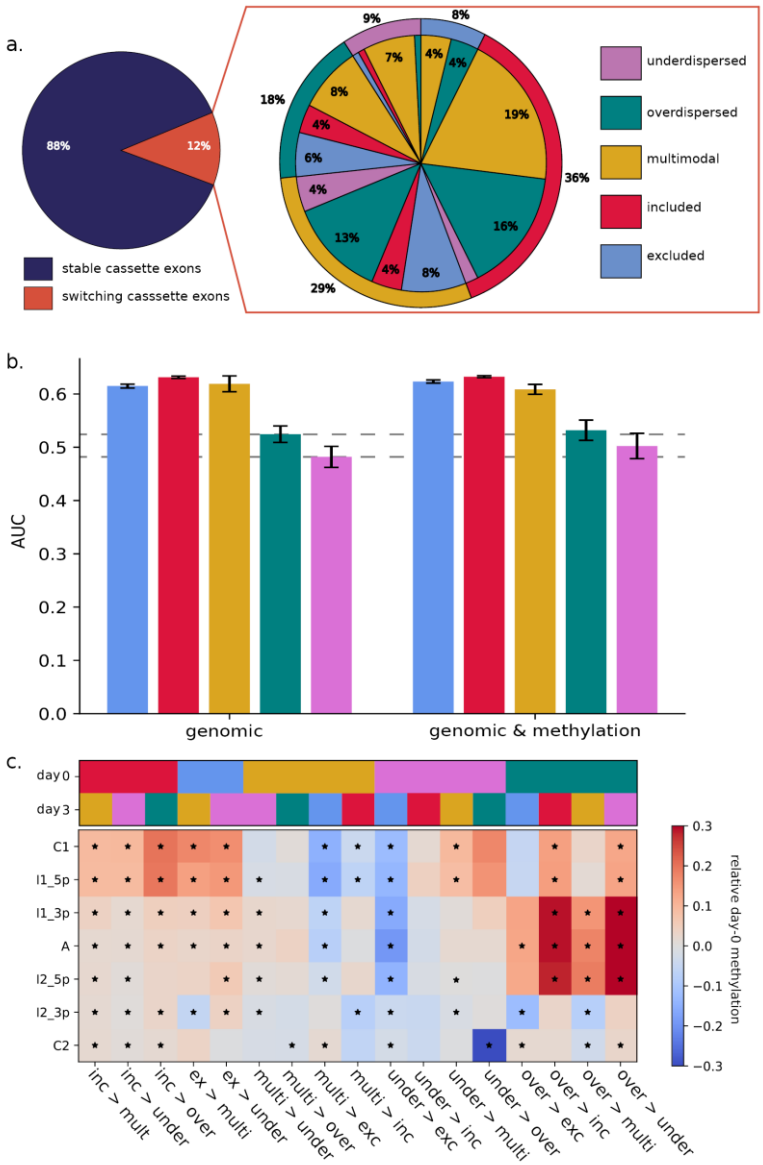

Figure 4. Comparison of splicing category distributions between iPS and endoderm cells. **a.** Pie chart showing the number of category switches between iPS and endoderm cells (left panel). The zoom-in (right panel) shows details of different category switches. The outer pie chart shows the splicing category of each cassette exon at the iPS state and the internal pie chart shows the respective category at endoderm state. Non-annotated slices in the pie chart reflect ~1% of the data. **b.** Performance of logistic ridge regression models that predict absence/presence of switching splicing categories between iPS and endoderm states. DNA methylation information improves prediction of the under- and overdispersed cassette exons. The categories are colored according to **a**. Error bars denote plus or minus one standard deviation across four repeat experiments. For comparison dashed lines are added to show the differences in prediction accuracies using the two feature sets. **c.** DNA methylation changes associated with the observed category switches. The top panel shows the iPS and endoderm splicing categories colored according to **a**. The bottom panel shows DNA methylation levels within the seven sequence contexts of a cassette exon as compared to the DNA methylation levels of the cassette exons that do not switch in their splicing category. Significant changes ($Q < 0.05$) are marked with a star. DNA methylation of the alternative exon and its vicinity is increased in cassette exons that switch from the underdispersed category. Cassette exons that switch from either included or excluded to any other splicing category show increased DNA methylation of the upstream exon (C1).

Discussion

Here, we present the first analysis of alternative splicing in single cells that considers both genomic and epigenetic alterations factors. Our study is focused on variation of splicing in cassette exons at two different stages of iPS differentiation. We show that splicing events do not strictly follow the previously described cell or gene model of splicing patterns, but instead we find a substantial proportion of exons that are better described by an intermediate model (**Fig. 1c**).

We show that single-cell splicing of cassette exons is influenced by genomic features previously assessed in bulk data, but also by cellular features and by DNA methylation differences. We observe that DNA methylation is related to the assessed splicing phenotypes, with the strongest link being observed in single-cell splicing ratios. When assessing splicing variation in bulk populations ("pseudo bulk") most of the information encoded in DNA methylation was lost (**Fig. 2a**). One reason for this might be the strong correlation between our genomic and methylation features, in particular between DNA methylation and cytosine-related features. Additionally, our results suggest that the relationship between splicing and DNA methylation is locus specific as observed when linking DNA methylation in a loci specific manner (**Figure S3**). This might also explain why we don't see an increase in prediction accuracy when we move the across cell features.

Next to sequence conservation, a feature that has previously been described in bulk studies [7], the most relevant features to predict splicing were the k-mers CTC, CT and CCT within the alternative exon (**Fig. 2b**). These k-mers point towards involvement of CTCF. Previous work has shown that CTCF motifs within introns are linked to splicing by slowing down RNA Polymerase II, thereby leading to a higher chance of exon inclusion. Interestingly, there is a known link between DNA methylation and CTCF motifs [21]. Methylation of CTCF binding sites can block CTCF and thereby result in decreased inclusion rates of an exon. As the methylated k-mer equivalents were in general less predictive of splicing, we suggest a more complex involvement of DNA methylation in alternative splicing, potentially a locus specific effect, which our current features and models are not able to capture.

In addition to modelling splicing ratios, we also modelled splicing categories to gain insights into variability of splicing across single cells (**Fig. 3**). The categories considered in our model reflect both the overall splicing rate and characteristics of splicing variability across cells. Exons with included versus excluded splice states, which is similar to splicing states previously

considered in bulk studies, could be accurately predicted. In contrast, the intermediate splicing categories, which are reflective of single-cell variability, were less accurate. This could be due to the lower number of cassette exons assigned to these categories (multimodal n=507, overdispersed n=427, underdispersed n=112, versus included n=3,290 and excluded n=1,946 in iPS cells), or reflect increased vulnerability top assay noise or more complex regulatory dependencies. As in the linear regression models, we observed that DNA sequence conservation scores were the most informative features for predicting splicing categories (**Table S3**). Interestingly, for intermediate classes the genomic information in the vicinity of the alternative exon rather than of the exon itself seemed to be predictive of splicing variability. Whereas DNA methylation did not contribute to improving the splicing prediction, we observe that DNA methylation levels of underdispersed cassette exons were significantly reduced in all genomic contexts, and most significantly in the upstream exon.

By leveraging data from two differentiation time points, we were able to show the stability of splicing prediction and of the relevant genomic and methylation features, and could simultaneously assess splicing category maintenance during cell differentiation (**Fig. 2c**). The differences between features being predictive of splicing at the two time points were mainly observed within the (methylated) k-mers, which is consistent with the known alteration of transcription factor activity and DNA methylation differences between two differentiation states. Next, we were able to confirm the findings from Song *et al.* [13] within a different differentiation set-up that only a limited number of cassette exons switch splicing categories between differentiation states (**Fig. 4a**). Additionally, also as described in context of a neural differentiation before, switches between included and excluded category were not observed. Most of the category switches were observed within the three intermediate splicing categories. Furthermore, DNA methylation differences seemed to predate the switching ability. Using ridge regression, we were able to predict if a cassette exon would switch its splicing category between the differentiation time points. Again, DNA methylation seemed to be particularly informative of intermediate splicing. It improved the predictability of switching in over- and underdispersed categories.

The novelties of our analyses also represent their main limitations. Single-cell sequencing intrinsically delivers fewer reads to assess gene expression and DNA methylation levels. Especially the genome coverage of the bisulfite-treated DNA sequencing remains low due to the low quantities of starting material. Using computational imputation, we were able to mitigate this effect to some extent, however imputation strategies have limitations and in particular loci that completely lack methylation information cannot not be recovered.

The intrinsic properties of single-cell data similarly affect the accuracy of estimated splicing ratios per cassette exon. We opted for a lenient threshold on read depth to determine the splicing ratio, which delivered more cassette exons to train our models, but also rendered splicing ratios less accurate in comparison to deep-sequenced bulk data. The low read depth increases the chance of missing an isoform or a complete cassette exon, what is known as a dropout. Dropouts in single-cell RNA-seq data can have a strong impact on the fit of the cell- or gene-model. If one of the isoforms was completely unobserved, this would decrease the fit of the gene model. Opposed to that, sequencing multiple cells at once would decrease the fit of the cell model. Given that our results seem stable across cassette exons and differentiation time points, the overall trends we report are not likely to be affected. Together with the constant improvement of single-cell techniques, these results make us hopeful that we can give a more definitive answer on the fit of the cell- or gene-model of splicing variability in the future.

Conclusions

In summary, we showed for the first time that alternative splicing and splicing variability across cells can be predicted with genomic and DNA methylation information within single cells. We assessed the impact of DNA methylation and cellular features on cassette exon splicing, and we were able to replicate our findings using two differentiation time points. We investigated stability and variance of splicing between the two differentiation time points and, importantly, we showed that DNA methylation primes splicing switches during the differentiation process.

Methods

Data generation

Single cell transcription and methylation data was generated from a single donor from the Human Induced Pluripotent Stem Cells Initiative (HipSci) [16, 17], using the previously described protocol for single-cell methylation and transcriptome sequencing in the same cells (scM&T-seq) [15]. Line joxm_1, an induced pluripotent stem cell (iPSC) line derived from fibroblasts cells from HipSci project, was cultured and triggered into differentiation towards endoderm. scM&T-seq data was generated for 93 cells (together with one empty well as negative control and two 15-cell and 50-cell positive controls) at the undifferentiated time point (iPS) and the definitive endoderm time point (endoderm), yielding 186 cells for analysis.

Cell handling and differentiation

The joxm_1 IPSC line was cultured in Essential 8 (E8) media (LifeTech) according to the manufacturer's instructions. For dissociation and plating, cells were washed 1x with DPBS and dissociated using StemPro Accutase (Life Technologies, A1110501) at 37°C for 3 - 5 min. Colonies were fully dissociated through gentle pipetting. Cells were washed 1x with MEF medium [22] and pelleted gently by centrifuging at 285xg for 5 min. Cells were re-suspended in E8 media, passed through a 40 µm cell strainer, and plated at a density of 60,000 cells per well of a gelatin/MEF coated 12 well plate in the presence of 10 µM Rock inhibitor – Y27632 [10 mM] (Sigma, Cat # Y0503 - 5 mg). Media was replaced with fresh E8 free of Rock inhibitor every 24 hours post plating. Differentiation into definitive endoderm commenced 72 hours post plating as previously described [22].

FACS preparation and analysis of cells

During all staining steps, cells were protected from light. Cells were dissociated into single-cells using accutase and washed 1x with MEF medium as described above. Approximately 1 x 10^6 cells were re-suspended in 0.5 mL of differentiation stage specific medium containing 5 µL of 1 mg/mL Hoechst 33342 (Thermo Scientific). Staining with Hoechst was carried out at 37°C for 30 min. Unbound Hoechst dye was removed by washing cells with 5 mL PBS + 2% BSA + 2 mM EDTA (FACS buffer); BSA and PBS were nuclease-free. For staining of cell surface markers Tra-1-60 (BD560380) and CXCR4 (eBioscience 12-9999-42), cells were re-suspended in 100 µL of FACS buffer with enough antibodies to stain 1 x 10^6 cells according

to the manufacturer's instructions, and were placed on ice for 30 min. Cells were washed with 5 mL of FACS buffer, passed through a 35 µM filter to remove clumps, and re-suspended in 250 µL of FACS buffer for live cell sorting on the BD Influx Cell Sorter (BD Biosciences). Live/dead marker 7AAD (eBioscience 00-6993) was added just prior to analysis according to the manufacturer's instructions and only living cells were considered when determining differentiation capacities. Living cells stained with Hoechst but not Tra-1-60 or CXCR4 were used as gating controls.

scM&T-seq

As previously described in Angermeuller *et al.* [15], scM&T-seq library preparation was performed following the published protocols for G&T-seq [23] and scBS-seq [24], with minor modifications as follows. G&T-seq washes were performed with 20 µl volumes, reverse transcription and cDNA amplification were performed using the original Smart-seq2 volumes [25] and Nextera XT libraries were generated from 100-400 pg of cDNA, using 1/5 of the published volumes. RNA-seq libraries were sequenced as 96-plexes on a HiSeq 2000 using v4 chemistry and 125 bp paired end reads. BS-seq libraries were sequenced as 24-plexes using the same machine and settings, which yielded a mean of 7.4M raw reads after trimming.

DNA methylation pre-processing and quantification

For DNA methylation data, single-cell bisulfite sequencing (scBS-seq) data were processed as previously described [24]. Briefly, reads were trimmed with Trim Galore! [26–28], using default settings for DNA methylation data and additionally removing the first 6 bp. Subsequently, Bismark [29] (v0.16.3) was used to map the bisulfite data to the human reference genome (build 38), in single-end, non-directional mode, which was followed by de-duplication and DNA methylation calling using default settings. All but two single-cell libraries (alignment rate <15%) yielded good alignment rates (mean = 43%), with negative control wells having very low mappability (mean = 2%). Additionally, we removed seven samples with a library size of less than 1M reads.

To mitigate typically low coverage of scBS-seq profiles (20-40%; [30]), we applied DeepCpG [30] to impute unobserved methylation states of individual CpG sites. DNA methylation profiles in iPS and endoderm cells were imputed separately, using the default settings of the method. Predicted methylation states were binarized according the DeepCpG probabilities as follows: sites with a probability of equal or lower than 0.3 were set to 0 (un-methylated base), all methylation sites with a probability of great than 0.7 were set to 1 (methylated base).

Intermediate methylation levels were handled as missing. After imputation the methylation data was lifted back to human genome version 37 to match the expression data, using the UCSC lift-over tool [31].

We integrated the imputed methylation information into the DNA sequence by distinguishing methylated ('M') and un-methylated ('U') cytosines. Cytosines without methylation information after imputation were assigned the value of the closest cytosine with methylation information. If there was no methylation information within 900 bp around the cytosine, its state was set to un-methylated.

Gene expression quantification

For single-cell RNA-seq data, adapters were trimmed from reads using Trim Galore! [26–28], using default settings. Trimmed reads were mapped to the human reference genome build 37 using STAR [32] (version: 020201) in two-pass alignment mode, using the defaults proposed by the ENCODE consortium (STAR manual). Expression quantification was performed separately using Salmon [33] (version: 0.8.2), using the "--seqBias", "--gcBias" and "VBOpt" options on transcripts derived from ENSEMBL 75 [23]. Transcript-level expression values were summarized at gene level (estimated counts) and quality control of scRNA-seq data was performed using scater [34]. Cells with at least 50,000 counts from endogenous genes, at least 5,000 genes with non-zero expression, less than 90% of counts came from the 100 most-expressed genes in the cell, less than 20% of counts from ERCC spike-in sequences and a Salmon mapping rate of at least 40% were retained for analysis.

Splicing quantification

Of the 192 cells, 86 (iPS) and 59 (endoderm) cells passed QC on both DNA methylation and gene expression data as described above. Exon splicing rates in individual cells were quantified using the data-dependent module of BRIE [9]. BRIE calls splicing at predefined cassette exons and quantifies splicing using exon reads in single-cell data. By default BRIE combines informative prior learned from sequence features and a likelihood calculated from RNA-seq reads by a mixture modelling framework that is similar to MISO [35]. As our aim is to model the local and global determinants of splicing, we used splicing rate estimates based on the observed data at individual exons only. We detected and quantified splicing for between 1,386 and 14,434 exons per cell (minimum coverage five reads, in total considered 6,282 (iPS) and 4,096 (endoderm) cassette exons that were detected in at least ten cells for further analysis.

We used three different measurements to quantify splicing ratios (PSI), namely single-cell splicing ratios (sPSI), pseudo bulk splicing ratios (bPSI) and variance of splicing ratios (vPSI). To calculate sPSI values per cassette exon per cell, we only considered splicing events that were supported by at least five reads and limited the analysis to cassette exons which were observed in at least ten cells. To derive bPSI values per cassette exon, we aggregated the sPSI values per cassette exon. The vPSI per cassette exon were defined as the standard deviation of the sPSI across cells.

Sequence features

The genomic features used to predict the splicing ratios and its variance were based on the features described by BRIE and Xiong et al. [8, 9]. As these features were specifically designed to study exon skipping events at cassette exons, they are designed to capture sequence variation in the following five genomic contexts: the alternative exon itself, the two neighboring exons and the introns between the exons. The following features were calculated per genomic context: the (log) length, the relative length and the strength of the splice site motifs at the exon-intron boundaries. The strength of the splice site was defined as the similarity between this splice site and known splice motives. Additional features were calculated on seven genomic contexts, namely the alternative exon itself, the two neighboring exons and the 5' and 3' boundaries of the introns. Only the two boundary contexts of the introns (300 bp length) were used since intron length is highly variable and the boundaries are the most relevant contexts for splicing. The following features were calculated for these genomic contexts: PhastCons scores [36] which reflect sequence conservation and k-mer frequencies (with $k \leq 3$). The k-mers reflect the percentage of nucleotides in the context that match the respective specific motif.

In addition to the genomic features, we defined DNA methylation features for each of the seven genomic contexts. For the sPSI model, we considered cell-specific methylation levels per context (methylation rate within the context) and extended the k-mer features by including un-methylated ('U') and methylated ('M') cytosine into the alphabet. Cytosines without methylation information after imputation were assigned the value of the closest cytosine with methylation information. If there was no methylation information within 900 bp around the cytosine, its state was set to un-methylated. For the bPSI model, we included the mean frequencies of the k-mers that contained 'M' or 'U' across cells and the mean and standard deviation of the seven contexts across cells per time point.

Splicing categories

In bulk RNA-seq data, splicing events can be broadly categorized into two major categories: included and excluded. Leveraging the single-cell information, we defined more fine-grained splicing categories that reflect both, splicing rates (sPSI) and splicing variability (vPSI) across cells (inspired by Song et al., 2017 [13]): 1) excluded (mean PSI < 0.2), 2) included (mean PSI > 0.8), 3) overdispersed, 4) underdispersed and 5) multimodal (Fig. 3). The later three categories categorize the extent of splicing variation across cells, since cassette exons with intermediate average splicing rates (here 0.2≤ mean PSI ≤0.8) exhibit substantial differences in splicing variance (**Fig. 1**). To characterize cells into these three categories, we calculated the distribution of the distance between the observed and the expected variation per day. The expected variation was calculated by a scaled binomial standard deviation, using: where is the scaling factor and the mean splice rate of the alternative exon [18], fit to all data points. We then defined the overdispersed cassette exons as those for which the deviation from the expected PSI was higher than the 3rd quartile plus 1.5x interquartile range (IQR, corresponding to > 0.016 in iPS and > 0.022 in endoderm). Likewise, for definition of the underdispersed cassette exons, we used the 1st quartile minus 1.5x IQR as threshold (corresponding to < -0.032 in iPS and < -0.039 in endoderm cells). The remaining cassette exons were assigned to the multimodal category.

Relating intermediate splicing levels and cell features

We investigated cell features in cassette exons with intermediate splicing levels (0.2 < PSI < 0.8). This was done by relating 1) cell-cycle state measured by FACS to these values, 2) Differentiation state by comparing the two tissue types and 3) cell-level splicing activity. To do so we used a Wilcoxon test to compare categorical values to the PSI values, the splicing activity was split based on mean splicing activity.

Prediction of PSI and categories

We applied linear ridge regression to model sPSI and bPSI and (multi-class) logistic ridge regression to model the PSI categories. The models are based on only the genomic features or on both genomic and DNA methylation features. The performance of linear models was evaluated using Pearson R^2 between predicted and observed splicing rates. For the multi-class prediction models, we applied a one-vs-rest scheme and report the per-category and the macro-average area under the receiver operating curves (AUC). To determine the most relevant individual features, we additionally trained regression models with a single feature at a time. Per feature we report, in the case of the linear models, using Pearson correlation (R, R^2) and, in the case of the logistic models, the absolute weight multiplied by the standard deviation of the feature, and the AUC. We assessed performance and parameters of models by using a 10-fold cross validation (CV) with fixed training-validation splits. To assess variability of prediction performances, we repeated the CV procedure four times with different cross validation splits. Error bars indicate plus or minus one standard deviation of the statistics in question (AUC, R^2).

Relating DNA methylation heterogeneity and splicing

Next to linear models across cells and sites we applied spearman correlation to link splicing at a single locus to variation in DNA methylation observed between cells, by spearman correlation. The test was performed and corrected per context within the cassette-exon (seven context considered, **Fig. 1a**). We only considered cassette exons where variation in splicing and variation of DNA methylation of the relevant context was observed. In total 5,280 cassette-exons are tested cassette exons for iPS and 2,622 for endoderm. The P-values obtained from the test are adjusted for multiple testing using the Q-value [37, 38] package in R. The gene enrichment across the cassette exons which are significantly related to DNA-methylation was performed using G:Profiler [39] using all observed cassette exons per tissue as a background. Multiple testing correction for the enrichments was performed within G:Profiler.

Cell features based on expression

We defined a splicing score for each of the cells to reflect the activity of the splicing machinery in each of the individual cells. This was done by taking the first PC of a PCA on the splicing ratios of genes that are known to be associated with alternative splicing according to Gene Ontology (GO, GO:0008380, [34]). The sign of the score was determined based on the genes in extremes of the rotation information as returned by the PCA.

List of abbreviations

splicing ratio	PSI
single-cell splicing ratio	sPSI
pseudo bulk splicing ratio	bPSI
variance of splicing ratio	vPSI
induced pluripotent stem cell	iPS cell

Declarations

Ethics approval and consent to participate

All samples for the HipSci resource were collected from consented research volunteers recruited from the NIHR Cambridge BioResource (http://www.cambridgebioresource.org.uk). Samples were collected initially under ethics for iPSC derivation (REC Ref: 09/H0304/77, V2 04/01/2013), with later samples collected under a revised consent (REC Ref: 09/H0304/77, V3 15/03/2013).

Availability of data and material

All sample meta-data can be accessed via the HipSci data portal (http://www.hipsci.org), which references to EMBL-EBI archives that are used to store the HipSci data. DNA methylation and RNA expression data will be shared via the European Nucleotide Archive (accession number pending).

Competing interests

The authors declare that they have no competing interests

Funding

This work was funded with a strategic award from the Wellcome Trust and UK Medical Research Council (WT098503). Work at the Wellcome Trust Sanger Institute was further supported by Wellcome Trust grant WT090851. S.L. received support from the Stiftung Familie Klee. L.U. received support from core funding of the European Molecular Biology Laboratory and the European Union's Horizon 2020 research and innovation programme (grant agreement number N635290). M.J.B. was supported by a fellowship from the EMBL Interdisciplinary Postdoc (EI3POD) program under Marie Skłodowska-Curie Actions COFUND (grant number 664726). Research from L.V. laboratories is supported by the European Research Council advanced grant New-Chol, the Cambridge University Hospitals National Institute for Health Research Biomedical Research Center, a core support grant from the Wellcome Trust and Medical Research Council to the Wellcome Trust – Medical Research Council Cambridge Stem Cell Institute (PSAG028). WR was supported by grants from BBSRC (BB/K010867/1), Wellcome Trust (095645/Z/11/Z), EU BLUEPRINT, and EpiGeneSys.

Authors' contributions

S.L.: data processing, splicing ratio modeling, deep modeling

L.U.: splicing variation modeling, supervision of splicing modeling and deep modeling, writing of the manuscript

D.M.: gene expression data processing, cell-level feature calculations

M.C.: data generation, editing

S.C.: data generation, editing

S.A.: data generation, editing

I.E.: supervision of original master project, editing

L.V.: data generation, editing

W.R.: data generation, editing

O.S.: supervision of modelling, writing of the manuscript, study design

M.J.B.: methylation data processing, supervision of splicing modelling, writing of the manuscript

Acknowledgements

We thank the staff in the Cellular Genetics and Phenotyping and Sequencing core facilities at the Wellcome Trust Sanger Institute. We acknowledge the participation of all NIHR Cambridge BioResource volunteers, and thank the NIHR Cambridge BioResource centre staff for their contribution. We thank the National Institute for Health Research and NHS Blood and Transplant. The NIHR/Wellcome Trust Cambridge Clinical Research Facility supported the volunteer recruitment. We thank Y. Huang for helpful discussions and comments on the manuscript.

References

1. Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem. 2003;72:291–336. doi:10.1146/annurev.biochem.72.121801.161720.
2. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338:1587–93. doi:10.1126/science.1230612.
3. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, et al. Function of alternative splicing. Gene. 2013;514:1–30. doi:10.1016/j.gene.2012.07.083.
4. Revil T, Gaffney D, Dias C, Majewski J, Jerome-Majewska LA. Alternative splicing is frequent during early embryonic development in mouse. BMC Genomics. 2010;11:399. doi:10.1186/1471-2164-11-399.
5. 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 (80-). 2015;347:1254806. doi:10.1126/science.1254806.
6. Sammeth M, Foissac S, Guigó R. A general definition and nomenclature for alternative splicing events. PLoS Comput Biol. 2008;4:e1000147. doi:10.1371/journal.pcbi.1000147.
7. Wainberg M, Alipanahi B, Frey B. Does conservation account for splicing patterns? BMC Genomics. 2016;17:787. doi:10.1186/s12864-016-3121-4.
8. 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.
9. Huang Y, Sanguinetti G. BRIE: transcriptome-wide splicing quantification in single cells. Genome Biol. 2017;18:123. doi:10.1186/s13059-017-1248-5.
10. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature. 2011;479:74–9.
11. Maunakea AK, Chepelev I, Cui K, Zhao K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. Cell Res. 2013;23:1256–69. doi:10.1038/cr.2013.110.
12. Yearim A, Gelfman S, Shayevitch R, Melcer S, Glaich O, Mallm J-P, et al. HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. Cell Rep. 2015;10:1122–34. doi:10.1016/j.celrep.2015.01.038.
13. Song Y, Botvinnik OB, Lovci MT, Kakaradov B, Liu P, Xu JL, et al. Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron

Differentiation. Mol Cell. 2017;67:148–161.e5. doi:10.1016/j.molcel.2017.06.003.
14. Welch JD, Hu Y, Prins JF. Robust detection of alternative splicing in a population of single cells. Nucleic Acids Res. 2016;44:e73.
15. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods. 2016;13:229–32. doi:10.1038/nmeth.3728.
16. Streeter I, Harrison PW, Faulconbridge A, The HipSci Consortium, Flicek P, Parkinson H, et al. The human-induced pluripotent stem cell initiative-data resources for cellular genetics. Nucleic Acids Res. 2017;45:D691–7. doi:10.1093/nar/gkw928.
17. Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature. 2017;546:370–5. doi:10.1038/nature22403.
18. Faigenbloom L, Rubinstein ND, Kloog Y, Mayrose I, Pupko T, Stein R. Regulation of alternative splicing at the single-cell level. Mol Syst Biol. 2015;11:845. http://www.ncbi.nlm.nih.gov/pubmed/26712315.
19. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013;498:236–40. doi:10.1038/nature12172.
20. Brooks AN, Aspden JL, Podgornaia AI, Rio DC, Brenner SE. Identification and experimental validation of splicing regulatory elements in Drosophila melanogaster reveals functionally conserved splicing enhancers in metazoans. RNA. 2011;17:1884–94. doi:10.1261/rna.2696311.
21. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature. 2011;479:74–9. doi:10.1038/nature10442.
22. Hannan NRF, Segeritz C-P, Touboul T, Vallier L. Production of hepatocyte-like cells from human pluripotent stem cells. Nat Protoc. 2013;8:430–7. http://www.ncbi.nlm.nih.gov/pubmed/23424751.
23. Macaulay IC, Teng MJ, Haerty W, Kumar P, Ponting CP, Voet T. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. Nat Protoc. 2016;11:2081–103. doi:10.1038/nprot.2016.138.
24. Clark SJ, Smallwood SA, Lee HJ, Krueger F, Reik W, Kelsey G. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). Nat Protoc. 2017;12:534–47. doi:10.1038/nprot.2016.187.
25. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9:171–81. doi:10.1038/nprot.2014.006.

26. Krueger F. Trim Galore!
27. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10. doi:10.14806/ej.17.1.200.
28. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
29. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2. doi:10.1093/bioinformatics/btr167.
30. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18:67. doi:10.1186/s13059-017-1189-z.
31. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12:996–1006. doi:10.1101/gr.229102.
32. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
33. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9. doi:10.1038/nmeth.4197.
34. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33:1179–86.
35. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010;7:1009–15. doi:10.1038/nmeth.1528.
36. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50. doi:10.1101/gr.3715005.
37. Storey JD. False Discovery Rates. Princet Univ Princeton, USA. 2010; January:1–7.
38. Bass Jds, Dabney A, Robinson D. qvalue: Q-value estimation for false discovery rate control. 2015. http://github.com/jdstorey/qvalue.
39. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016;44:W83–9.