Open access • Posted Content • DOI:10.1101/328138

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

Stephanie M. Linker, Lara Urban, Stephen J. Clark, Mariya Chhatriwala ...+7 more authors

Institutions: European Bioinformatics Institute, Babraham Institute, Wellcome Trust Sanger Institute, Goethe University Frankfurt

Published on: 22 May 2018 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Alternative splicing, RNA splicing, Exon, DNA methylation and Exon skipping

Linker et al. Genome Biology (2019) 20:30
https://doi.org/10.1186/s13059-019-1644-0

Genome Biology

RESEARCH Open Access

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

Stephanie M. Linker[1,2†], Lara Urban[1,2†], Stephen J. Clark[3], Mariya Chhatriwala[4], Shradha Amatya[4], Davis J. McCarthy[1,2], Ingo Ebersberger[5,6], Ludovic Vallier[4,7,8], Wolf Reik[3,4,9], Oliver Stegle[1,2,10*] and Marc Jan Bonder[1,2*]

Abstract

Background: Alternative splicing is a key regulatory mechanism in eukaryotic cells and increases the effective number of functionally distinct gene products. Using bulk RNA sequencing, splicing variation has been studied across human tissues and in genetically diverse populations. This has identified disease-relevant splicing events, as well as associations between splicing and genomic features, including sequence composition and conservation. However, variability in splicing between single cells from the same tissue or cell type and its determinants remains 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 show that variation in single-cell splicing can be accurately predicted based on local sequence composition and genomic features. We observe moderate but consistent contributions from local DNA methylation profiles to splicing variation across cells. A combined model that is built based on genomic features as well as DNA methylation information accurately predicts different splicing modes of individual cassette exons. These categories include the 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 alternative splicing events have been implicated in human diseases [5]. Bulk RNA sequencing (RNA-seq) of human tissues and cell lines has been applied to identify and quantify different splicing events [6], where in particular exon skipping at cassette exons, the most prevalent form of alternative splicing [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 as well as the length of the exon and flanking introns [5, 8]. Although there is some evidence for a role of DNA methylation in splicing regulation, this relationship is not fully understood and alternative models have been proposed [9–11]. The transcriptional repressor CTCF has

* Correspondence: oliver.stegle@embl.de; bonder.m.j@gmail.com
†Stephanie M. Linker and Lara Urban contributed equally to this work.
[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK
Full list of author information is available at the end of the article

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 [9]. Alternatively, increased DNA methylation of the MeCP2 pathway has been associated with 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, thus leading to an increased exon inclusion rate [10]. 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 [11]. 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 in single-cell RNA-seq have enabled investigating splicing variation at a single-cell resolution [8, 12, 13]. We here leverage recent protocols for parallel sequencing of RNA and bisulfite-treated DNA from the same cell (single-cell methylation and transcriptome sequencing; scM&T-seq [14]) to study single-cell splicing while accounting for cell-specific DNA methylome profiles. We apply our approach to investigate the associations between single-cell splicing variation and DNA methylation at two states of human induced pluripotent stem (iPS) cell 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 one cell line (joxm_1) of the Human Induced Pluripotent Stem Cell Initiative (HipSci) [15, 16]. We profiled 93 cells from 2 different cell types, namely cells in the iPS state (iPS) and cells following 3 days of differentiation towards definitive endoderm (endoderm). After quality control, this resulted in 84 and 57 cells, respectively (the "Methods" section), which were used for analysis. In each cell, we quantified cassette exon inclusion rates (the "Methods" section, Additional file 1: Table S1, Additional file 2: Table S2). We quantified splicing rates for between 1386 and 4917 cassette exons in each cell (minimum coverage of 5 reads), estimating splicing rates (PSI) as the fraction of reads that include the alternative exon versus the total number of reads at the cassette exon (the "Methods" section). Differences in sequencing depth and cell type explained most of the differences in the number of quantified splicing events between cells (Additional file 3: Figure S1, Additional file 1: Table S1, Additional file 2: Table S2). DNA methylation profiles were imputed using DeepCpG [17], yielding on average 23.1 M CpG sites in iPS and 21.6 M CpG sites in endoderm cells. We considered 6265 iPS and 3873 endoderm cassette exons that were detected in at least 10 cells for further analysis.

Initially, we explored whether individual cells express only a single splice isoform ("cell model"; the "Methods" section), or whether multiple isoforms are present in a given cell ("gene model"; the "Methods" section, Fig. 1a), a question that has previously been investigated in bulk and single-cell data [18, 19]. Specifically, we compared the observed distribution of splicing rates PSI in our data to the expected values when assuming a binomial distribution according to the cell model [18], as well as the expected distribution according to the gene model (the "Methods" section, Fig. 1a). Globally, our data rule out the cell model; however, we also observed deviations from the gene model, in particular for exons with intermediate levels of splicing ($0.2 < PSI < 0.8$, Fig.1b).

Methylation heterogeneity across cells is associated with splicing variability

Next, to identify locus-specific correlations between DNA methylation heterogeneity and variation in splicing across cells, we tested for associations between differences in imputed DNA methylation levels across cells and splicing rates (Spearman correlation; the "Methods" section).

For each cassette exon, we tested for associations between the splicing rate (PSI) and variation in DNA methylation in each of 7 sequence contexts: the upstream, alternative, and downstream exons, and the 5′ and 3′ end of the 2 introns (the "Methods" section, Fig. 1c). Genome-wide, this identified 424 cassette exons with a methylation-splicing associations in iPS cells (out of 5564 tested cassette exons, $Q < 0.05$, Additional file 3: Figure S2a, Additional file 4: Table S3) and 245 associations in endoderm cells (out of 2811 tested, $Q < 0.05$, Additional file 3: Figure S2a, Additional file 4: Table S3). The majority of these associations were observed in the upstream alternative exon (~ 75%), with approximately equal numbers of positive (increased DNA methylation is linked to increased alternative exon inclusion) and negative (increased DNA methylation is linked to decreased alternative exon inclusion) associations. In iPSC, 58% of correlations are positive, and 55% of the correlations are positive in endoderm cells. Most associations could be detected significantly in more than 1 context for a given exon with consistent effect directions (Additional file 3: Figure S2b, c). Similarly, we observed largely concordant associations across the 2 cell types in our data. Among the exons that are expressed in both iPS and endoderm ($n = 3743$), 77% of the associations identified in iPS were nominally replicated in endoderm

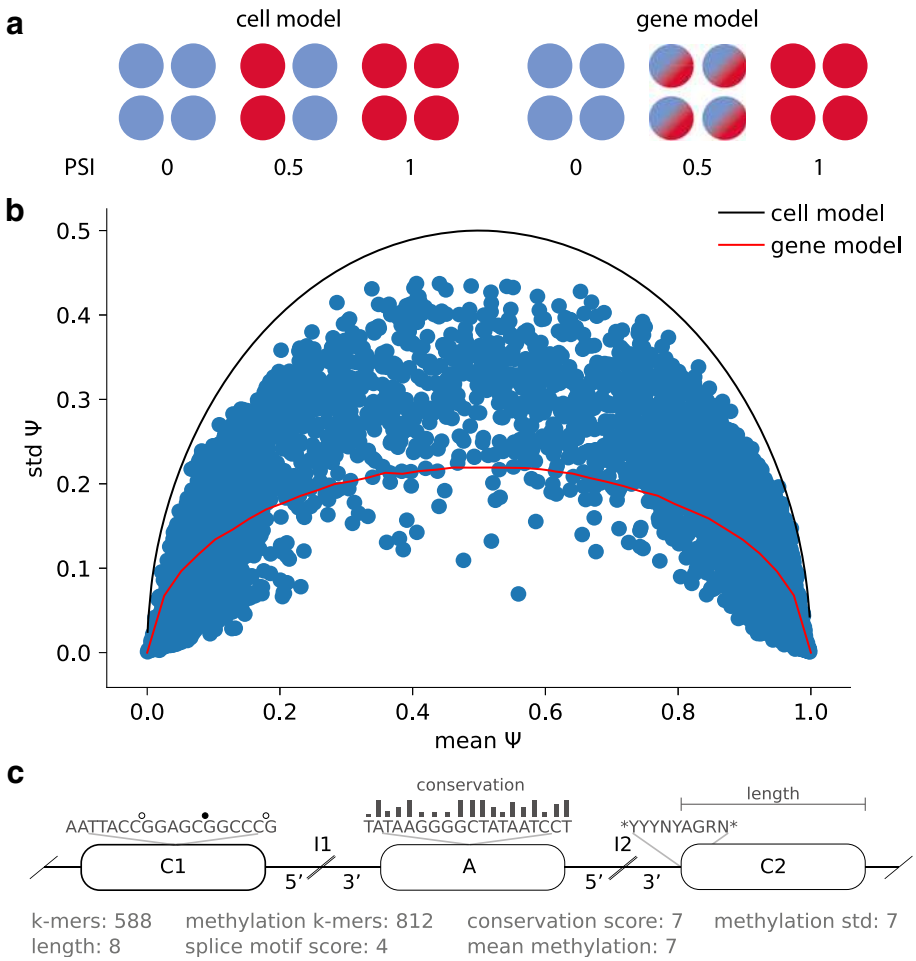

Fig. 1 Single-cell splicing and considered features for modeling splicing rates. **a** Two canonical splicing models. The "cell model" assumes that splicing variation is due to the differential splicing between cells, with each cell expressing one of two splice isoforms. The "gene model" corresponds to the assumption that both splice isoforms can be expressed in the same cells. **b** Mean-variance relationships of splicing rates in iPS cells. Shown is the standard deviation of splicing rates across cells for the same cassette exon (standard deviation of PSI) as a function of the average inclusion rate of the cassette exons across cells, considering 84 iPS cells. Solid lines correspond to the expected relationship when either assuming a "cell model" (black line) or when assuming the "gene model" (red line). **c** Illustration of the considered features and genomics contexts for predicting splicing variation. "A" denotes the alternative exon; "I1" and "I2" correspond to the upstream and downstream flanking introns, respectively; and "C1" and "C2" to the upstream and downstream flanking exons, respectively. The 5′ and 3′ ends (300 bp) of the flanking introns are considered separately

cells ($P < 0.05$, with a consistent effect direction), and 89% of the associations identified in endoderm were also observed in iPS cells ($P < 0.05$, with a consistent effect direction). Genes with negative associations between DNA methylation in the 3 upstream regions and PSI were enriched for HOXA2 transcription factor binding sites (iPS—78/118 query genes linked to HOXA2, adjusted $P = 6.02 \times 10^{-4}$; endoderm—60/90 query genes linked to HOXA2, adjusted $P = 9.03 \times 10^{-3}$; enrichment based on g:Profiler [20]).

Prediction of splicing at the single-cell level

To gain insights into the global determinants of splicing, we trained regression models to predict genome-wide cassette exon splicing rates in individual cells using local genomic and epigenetic features (Fig. 1c). Briefly, for each cell type, we combined splicing rates across all cassette exons and cells and trained global regression model using alternative sets of input features (assessed using tenfold cross-validation; the "Methods" section). Initially, we considered models based on a set of 607 "genomic" features derived from local sequence composition (based on *k*-mers), sequence conservation, and the length of the seven sequence contexts of each cassette exon ("genomic" features, the "Methods" section, Additional file 5: Table S4). Notably, the performance that was similar to previous approaches to predict splicing rates using bulk [5] and single-cell [8] RNA-seq ($r^2 = 0.704$, $r^2 = 0.668$;

assessed using tenfold cross-validation (CV); Fig. 2a, Additional file 3: Figure S3). To facilitate the comparison with previous studies using bulk RNA-seq, we also considered a model that was trained using aggregate splicing rates across cells ("pseudo-bulk PSI", bPSI), which resulted in similar prediction accuracies ($r^2 = 0.745$ and $r^2 = 0.733$ for iPS and endoderm cells, respectively, Additional file 3: Figure S4).

Next, we considered using an extended feature set in the linear models, using up to 826 DNA methylation features, including a *k*-mer alphabet that takes the methylation status of cytosines into account, as well as the DNA methylation rate and variance across CpG sites in each of 7 sequence contexts of a cassette exon (the "Methods" section). We considered 2 strategies to aggregate the methylation features, either (i) to capture patterns of methylation in individual cells ("genomic and cell methylation" features) or (ii) using averaged methylation features across all cells ("genomic and mean methylation" features; Additional file 5: Table S4, Fig. 1c). The inclusion of either type of methylation features into the model yielded an increased prediction accuracy, where larger gains in prediction accuracy were observed when accounting for DNA methylation information in individual cells ("genomic and cell methylation" versus "genomic and mean methylation"). Notably, the inclusion of DNA methylation features did not improve the accuracy of models for average splicing rates ("pseudo-bulk" models; Additional file 3: Figure S4). This observation in combination with the results from the association analysis between methylation and splicing indicates that DNA methylation can primarily explain the cell-to-cell variation in splicing at individual loci, whereas genomic features by design explain the variation across loci. These findings were consistent across iPS and endoderm cells, and we observed analogous benefits of accounting for DNA methylation when applying the same models to previous scMT-seq datasets from mouse embryonic stem (ES) cells [14] (the "Methods" section, Additional file 3: Figure S3 & S4).

Next, to assess the relevance of the considered features, we considered regression models based on individual features trained in individual cells. Consistent with previous bulk studies [5, 7], this identified features derived from the alternative exon and its neighboring contexts, namely the 3′ end of the upstream intron and

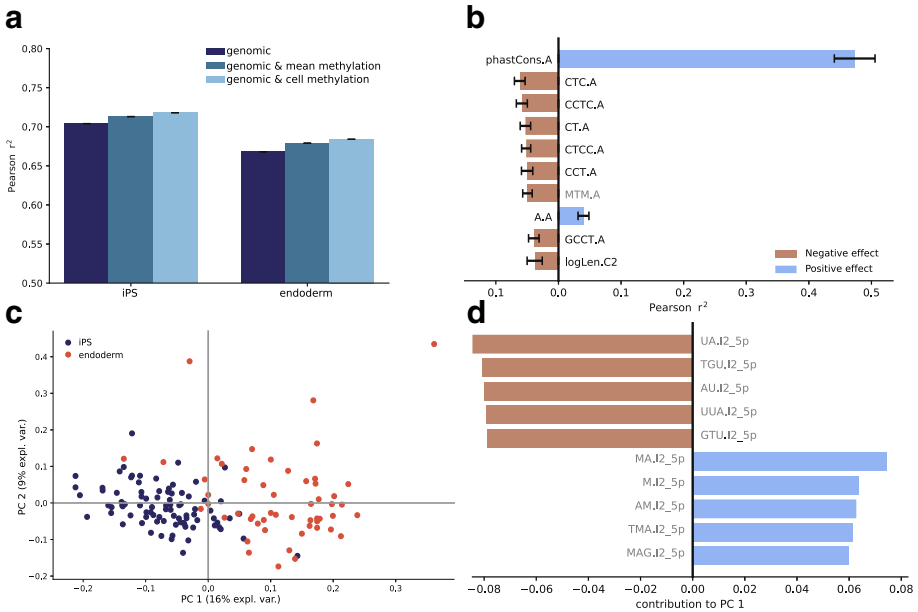

Fig. 2 Regression-based prediction of single-cell splicing variation. **a** Prediction accuracy of alternative regression models for predicting splicing rates in single cells. Shown are out of sample r^2 (based on tenfold cross-validation) in iPS cells (left) and endoderm cells (right). The genomic model (genomic, dark blue) was trained using sequence *k*-mers, conservation scores and the length of local contexts (size of the cassette exon, length of flanking introns) as input features. Other models consider additional features that capture average methylation features aggregated across cells (genomic and mean methylation, blue) or cell-specific methylation features (genomic and cell methylation, light blue). Error bars denote ± 1 standard deviation across four repeat experiments. **b** Relevance of individual features for predicting splicing rates, quantified using correlation coefficients between individual features and splicing rates. Shown are the average feature importance scores across all cells with error bars denoting ± 1 standard deviation across cells. Features are ranked according to absolute correlation coefficient with methylation features shown in gray. **c** Principal component analysis on the feature relevance profiles as in **b** across all cells. **d** Weights of the ten most important features that underpin the first principal component in **c** (shown are the five features with the largest positive and negative weight respectively), which include *k*-mers with methylation information of the downstream intron I2. Methylation features are shown in gray

the 5′ end of the downstream intron, as most informative (Additional file 6: Table S5). Within these contexts, sequence conservation of the alternative exon was the most relevant individual feature. Other relevant features included the *k*-mers CT, CTC, and CCT of the alternative exon (Fig. 2b), sequence patterns that show a close resemblance to CTCF-binding motifs. Although CTCF or CTCF-like motifs have previously been implicated splicing, these previous studies identified motifs upstream [9] or downstream [21] of the alternative exon as associated with increased splicing, whereas the *k*-mers in our model are located in the alternative exon and associated with decrease the inclusion rate [9, 21].

The relevance of the cell-specific features for splicing prediction as quantified by regression weights was markedly consistent across iPS and endoderm cells. This consistency extends to the mouse ES cell dataset, where again features of the alternative exon and sequence conservation scores were identified as the most relevant predictors for splicing (Additional file 6: Table S5, Additional file 3: Figure S5). Despite the overall consistency in feature relevance ($r^2 = 0.79$, average correlation between weights across all iPS and endoderm cells), principal component analysis (PCA) applied to the feature relevance matrix across all cells identified subtle coordinated axes of variation of the feature relevance (Fig. 2c). The first two principal components (PC) clearly separate iPS from endoderm cells, differences that are primarily attributed to *k*-mers of the downstream intron (I2) that contain methylated and unmethylated cytosine bases (Fig. 2d, Additional file 7: Table S6). Consistent with this, a single-cell methylation model trained on endoderm cells yielded only moderate prediction accuracy in iPS cells ($r^2 = 0.52$), highlighting the cell-type specificity of splicing models that account for DNA methylation information. 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.

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 [22], the "Methods" section). We observed only limited benefits when including DNA methylation information (Additional file 3: Supplementary Results and Figure S6). These results line up with the locus-specific DNA methylation and the linear regression results, supporting the hypothesis that global splicing information is primarily encoded by DNA sequence and conservation, and DNA methylation is linked to splicing in a locus-specific manner.

Prediction of splicing modes for individual exons

Next, we set out to study the differences between different exons and their splicing patterns. We classified cassette exons into five categories, using a scheme similar to that of Song et al. [12]: (1) excluded, (2) included, and three intermediate splicing categories: (3) overdispersed, (4) underdispersed, and (5) multimodal (Fig. 3a, b, Additional file 8: Table S7, the "Methods" section). We trained multinomial regression models (the "Methods" section) and assessed their classification performance (using four tenfold cross-validations) using analogous feature sets as considered for the regression models on single-cell splicing (Additional file 5: Table S4). A model based on genomic features yielded a macro-average AUC of 0.85 in iPS (Fig. 3c) and 0.84 in endoderm cells (Additional file 3: Figure S7), where again sequence conservation in different contexts was the most informative feature (Additional file 9: Table S8). 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 for the included and excluded categories 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, Additional file 3: Figure S7a). These prediction accuracies exceed previously reported results in bulk data [5]. Even higher accuracies were achieved when training a model to discriminate between included and excluded exons only (AUC = 0.99 in iPS), whereas the ability to discriminate intermediate splicing categories only was generally lower (AUC = 0.7–0.9, Additional file 9: Table S8). Notably, in contrast to the prediction of splicing rates, the inclusion of the DNA methylation features did not improve the prediction performance of these categorical models (Fig. 3d, Additional file 3: Figure S8a).

We 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, Additional file 3: 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, Additional file 3: Figure S9a-b).

We assessed the consistency of these results across iPS and endoderm cells, as well as in mouse ES cells. To do this, we trained the genomic model on endoderm cassette exons and assessed this model's predictions on iPS-specific cassette exons, which resulted in a prediction accuracy that was similar to the within cell-type prediction

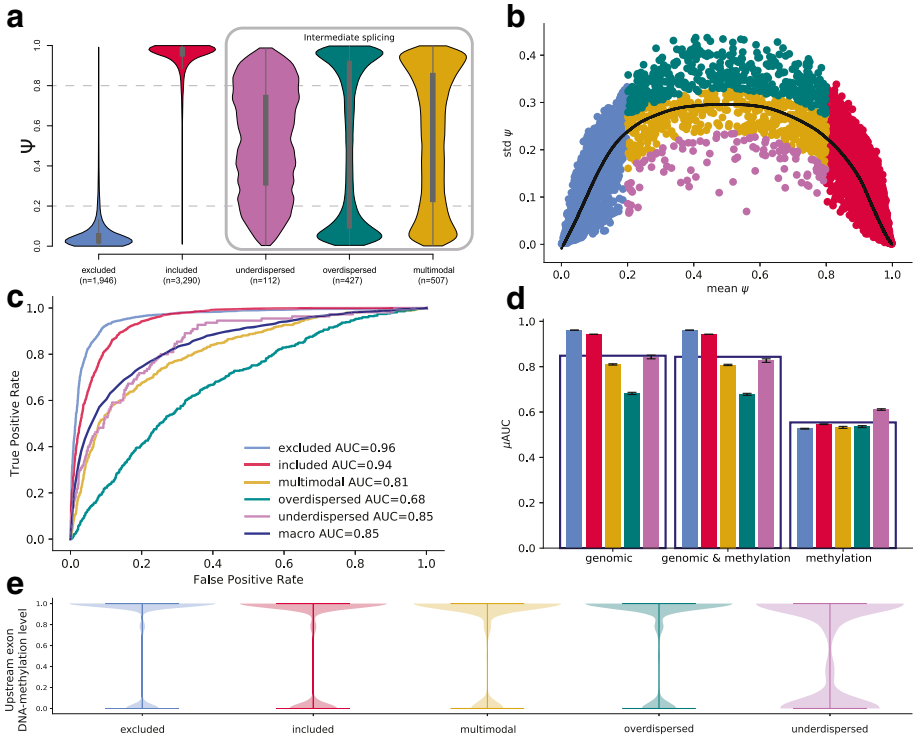

Fig. 3 Classification of cassette exons based on single-cell splicing patterns in iPS cells. **a** Single-cell splicing rate (PSI) distributions of the 5 splicing categories (inspired by Song et al. [12]) in 84 iPS cells. Intermediate splicing categories that can only be defined based on single-cell information are framed by a gray box. **b** Variation of PSI (standard deviation) across cells as a function of the average inclusion rate of cassette exons across 84 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** Performance of logistic regression models 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 and methylation," center) as well as only DNA methylation features ("methylation," right). The genomic model includes *k*-mers, conservation scores, and region lengths (see Fig. 1c). 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 ± 1 standard deviation across 4 repeat experiments. **e** Distribution of DNA methylation levels in the upstream exon (C1) per splicing category. Methylation is decreased in underdispersed exons

performance (macro-AUC = 0.82, Additional file 3: Figure S10a). However, the inclusion of the DNA methylation features into the model resulted in a decline in the cross-prediction performance (macro-AUC = 0.54, Additional file 3: Figure S10b). As in the linear model cross-replication analysis, this finding emphasizes the importance of cell type-specific DNA methylation for accurately predicting splicing. Next, we observed that the performance for splicing category prediction in mouse ES cells was very similar to the performance in the endoderm and iPS cells (macro-AUC = 0.82, in the genomic and the genomic and methylation model). We observed the same distinct distributions of DNA methylation in the upstream exon of underdispersed cassette exons (Additional file 3: Figure S9c). However, the relationship between the DNA methylation levels and underdispersed cassette exons category could not be replicated in the mouse ES cells (Additional file 3: Figure S7b).

Splicing category switches across cell differentiation

Finally, we assessed the changes in the splicing category switches between cell types. Similar to previous observations in the context of neuronal iPS differentiation [12], 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.

After observing the category switches between the cell types, 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

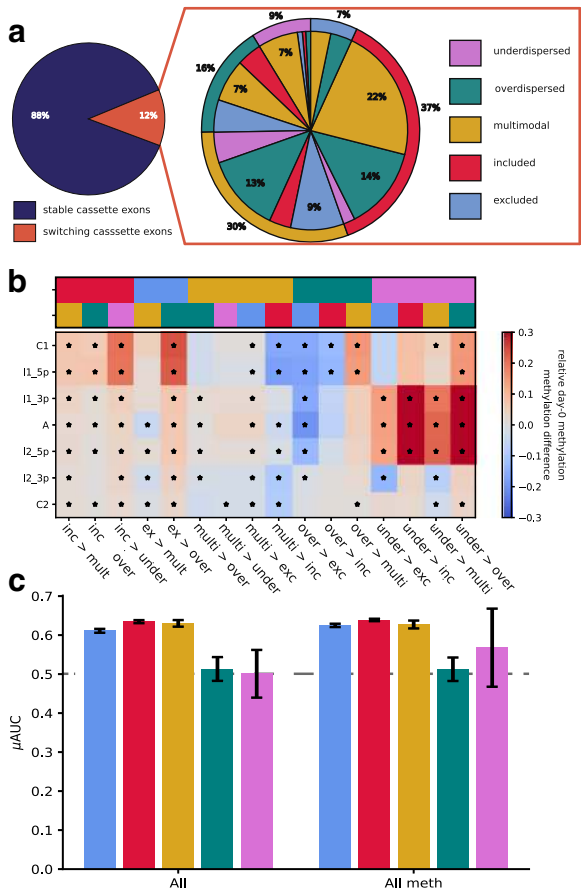

Fig. 4 Comparison of splicing category distributions between iPS and endoderm cells. **a** Pie chart showing the number of category switches between iPS and endoderm cells (left panel). The zoom-in (right panel) shows details of different category switches. The outer pie chart shows the splicing category 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** 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). **c** Performance of logistic ridge regression models that predict the absence/presence of switching splicing categories between iPS and endoderm states. DNA methylation information improves the prediction of the under- and overdispersed cassette exons. The categories are colored according to **a**. Error bars denote ± 1 standard deviation across four repeat experiments

exons during differentiation (Fig. 4c for prediction performance, Additional file 10: Table S9). This model had limited power to predict category switches (AUCs 0.51–0.64), and DNA methylation did not significantly improve the prediction of any category 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 cell types. The DNA methylation levels of cassette exons that switched category only changed minimally (Additional file 3: Figure S11). However, we observed that DNA methylation of the alternative exon of switching cassette exons differed from non-switching cassette exons at the iPS state (Fig. 4b). 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.

Discussion

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

We show that single-cell splicing of cassette exons is influenced by genomic features as previously assessed in bulk data, but also by DNA methylation differences. We observe that DNA methylation is related to splicing phenotypes, with the strongest link to single-cell splicing ratios. When assessing splicing variation in bulk populations (pseudo-bulk), most of the information encoded in DNA methylation is lost. A reason for this might be the strong correlation between genomic and methylation features, in particular between DNA methylation and cytosine-related features. Additionally, our results indicate that the relationship between splicing and DNA methylation is locus-specific (Additional file 3: Figure S2). This may also explain why DNA methylation has limited benefits when prediction average splicing rates across cells or splicing quantified using bulk RNA-seq.

Besides the 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 (Additional file 3: Figure S2b). 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 [9]. Interestingly, there is a known link between DNA methylation and CTCF motifs [9]. 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 less predictive of splicing, we suggest a more complex involvement of DNA methylation in alternative splicing, potentially by locus-specific effects, which our current models are not able to capture.

In addition to modeling splicing ratios, we also considered categorical models of splicing to gain insights into the variability of splicing across cells (Fig. 3). The categories considered in our model reflect both the overall splicing rate and splicing variability across cells. Exons with included versus excluded splicing states could be accurately predicted. In contrast, the intermediate splicing categories which are reflective of single-cell variability could only be predicted with a lower accuracy. This might be due to the lower number of cassette exons assigned to these categories (multimodal $n = 506$, overdispersed $n = 427$, underdispersed $n = 110$, versus included $n = 3278$ and excluded $n = 1944$ in iPS cells) or reflect increased vulnerability to 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 (Additional file 6: Table S5). Interestingly, for intermediate categories, 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, most significantly in the upstream exon. We hypothesize that the lower DNA methylation levels of underdispersed cassette exons give the sequence motifs more power to control splicing levels, i.e., increased DNA methylation levels lead to more stochasticity in splicing. This hypothesis is supported by the effect direction of methylation features, which are opposite between overdispersed and underdispersed cassette exons. We finally observe that the methylation *k*-mers are on average less informative of splicing than non-methylation features, potentially further supporting our hypothesis.

By leveraging data from two cell types, we were able to assess the consistency of splicing prediction across cell types and the relevant genomic and methylation features (Fig. 2c). The differences between features predictive of splicing between iPS and endoderm cells were primarily observed within the (methylated) *k*-mers, which are consistent with the known alteration of transcription factor activity and DNA methylation differences between cell types. Next, we were able to confirm the findings from Song et al. [12] that only a limited number of cassette exons switch splicing categories between cell types (Fig. 4a). Additionally, as previously described in the context of neural differentiation [12], switches between included and excluded categories were not observed. Most of the category switches were observed within the three intermediate splicing categories. Hereby, 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 cell types. 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 lack methylation information cannot be recovered.

The intrinsic properties of single-cell data also affect the accuracy of the estimated splicing ratios per cassette exon. We opted for a lenient threshold on read depth to determine 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 cassette exon, an effect 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. On the contrary, sequencing multiple cells at once would decrease the fit of the cell model. Given that our results are robust across cassette exons, cell types, and species, the overall findings we report are however not likely to be affected.

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 in single cells. We assessed the impact of DNA methylation and cellular features on cassette exon splicing and were able to replicate our findings in two human cell types and mouse ES cells. We investigated the stability and variance of splicing between the two cell types, and importantly, we showed that DNA methylation primes splicing switches during differentiation.

Methods

Single-cell transcription and methylation data was generated from a single donor from the Human Induced Pluripotent Stem Cells Initiative (HipSci) [15, 16], using

the previously described protocol for single-cell methylation and transcriptome sequencing in the same cells (scM&T-seq) (see [14] for details). 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 1 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 × 1 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 × 1 with MEF medium [23] and pelleted gently by centrifuging at 285×*g* 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 h post-plating. Differentiation into definitive endoderm commenced 72 h post-plating as previously described [23].

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 × 1 with MEF medium as described above. Approximately 1×10^6 cells were resuspended in 0.5 mL of differentiation state-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 the cells with 5 mL PBS + 2% BSA + 2 mM EDTA (FACS buffer); BSA and PBS were nuclease-free. For the staining of cell surface markers Tra-1-60 (BD560380) and CXCR4 (eBioscience 12-9999-42), cells were resuspended in 100 μL of FACS buffer with enough antibodies to stain 1×10^6 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 the 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. [14], scM&T-seq library preparation was performed following the published protocols for G&T-seq [24] and scBS-seq [25], 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 [26], and Nextera XT libraries were generated from 100 to 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 7.4 M raw reads after trimming.

Gene expression quantification

For single-cell RNA-seq data, adapters were trimmed from reads using Trim Galore! [27–29], using default settings. Trimmed reads were mapped to the human reference genome build 37 using STAR [30] (version: 020201) in two-pass alignment mode, using the defaults proposed by the ENCODE consortium (STAR manual). Expression quantification was performed separately using Salmon [31] (version: 0.8.2), using the "--seqBias," "--gcBias," and "VBOpt" options on transcripts derived from ENSEMBL 75. Transcript-level expression values were summarized at the gene level (estimated counts) and quality control of scRNA-seq data was performed using scater [32]. Cells with the following features were retained for analysis: (i) at least 50,000 counts from endogenous genes, (ii) at least 5000 genes with non-zero expression, (iii) less than 90% of counts are assigned to the top 100 expressed genes per cell, (iv) less than 20% of counts are assigned to ERCC spike-in sequences, and (v) a Salmon mapping rate of at least 40%. These filters jointly removed 9 iPS cells and 36 endoderm cells from our analysis.

Splicing quantification

Of the 186 cells, 84 (iPS) and 57 (endoderm) cells passed QC on gene expression data as described above. Exon splicing rates in individual cells were quantified using the data-dependent module of BRIE [8]. 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 modeling framework that is similar to MISO [33]. 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 1386 and 4917 exons per cell (minimum coverage 5 reads, in total considered 6265 (iPS) and 3873 (endoderm) cassette exons that were detected in at least 10 cells for further analysis.

The following settings were used to quantify splicing with BRIE: exons have to be located on autosomes and input chromosomes and should not be overlapped by any other alternatively spliced exon. The surrounding introns have to be longer than 100 bp, the length of the alternative exon regions has to be between 50 and 450 bp with a minimum distance of 500 bp from the next TSS or TTS, and the exon has to be surrounded by AG-GT. The default annotation file gencode.v19.annotation.gtf and the reference genome GRCh37.p13.genome.fa were downloaded from https://www.gencodegenes.org/human/release_19.html (May 2018) and used for subsequent analyses.

We used three different measurements to quantify splicing ratios (PSI), namely single-cell splicing ratios, pseudo-bulk splicing ratios, and variance of splicing ratios. To calculate single-cell PSI 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 pseudo-bulk PSI per cassette exon, we aggregated the single-cell PSI values per cassette exon. The variance of PSI per cassette exon was defined as the standard deviation of PSI across single cells.

DNA methylation pre-processing and quantification

For DNA methylation data, single-cell bisulfite sequencing (scBS-seq) data was processed as previously described [25]. Reads were trimmed with Trim Galore! [27–29], using default settings for DNA methylation data and additionally removing the first 6 bp. Subsequently, Bismark [34] (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. We removed cells with low alignment rates (alignment rate < 15%) and cells with a library size of less than 1 M reads, resulting in 84 iPS cells and 53 endoderm cells with RNA and DNA methylation information.

To mitigate typically low coverage of scBS-seq profiles (20–40%; [17]), we applied DeepCpG [17] to impute unobserved methylation states of individual CpG sites. DNA methylation profiles in iPS and endoderm cells were imputed separately. The cell type-specific models were built using CpG and genomic information according to DeepCpG's setup of a joint model (see [17] for details and default values; see Additional file 1: Table S1 for imputation accuracy as measured on a validation set per sample).

Predicted methylation states were binarized according to DeepCpG probability outputs as follows: sites with a probability of equal to or lower than 0.3 were set to 0 (un-methylated base), all methylation sites with a probability of greater than 0.7 were set to 1 (methylated base). Intermediate methylation levels were handled as missing. After imputation the methylation data was aligned back to human genome version 37 to match the expression data, using the UCSC lift-over tool [35].

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.

Cell and gene model assumptions

To assess if our PSI variation patterns follow the gene or the cell model [18], we compared the distribution of splicing rates to a binomial distribution that is expected according to the cell model and to the expected distribution according to the gene model.

The cell model assumes that each individual cell expresses only a single splice isoform, and hence models PSI variation as a bimodal distribution at the single cell level. The alternative gene model assumes splicing regulation on the gene level. The mean PSI of a gene is determined by the sequence. Each time a gene is transcribed, the probability of exon inclusion equals mean PSI. However, the limited number of transcripts leads to fluctuation in the observed PSI, and the binomial distribution is restrained by the upper boundary of the standard deviation. To obtain this upper boundary, we simulated the PSI of each cell as a binomial distribution and calculated the standard deviation across the cells. We only considered genes that were covered by at least 5 reads per cell in least 10 cells. To obtain the mean standard deviation, we repeated this simulation 400 times.

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. [5, 8]. As these features were specifically designed to study exon skipping events at cassette exons, they capture sequence variation around the alternatively spliced exon. This region is first split in five genomic contexts: the alternative exon itself, the two neighboring exons and the introns between the exons. Logarithmic length, relative length, and the strength of the splice site motifs at the exon-intron boundaries were

calculated per genomic context. 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, the three exons and the 5′ and 3′ boundaries of the two introns. Only the two boundary contexts of the introns (300 bp length) were used since intron length is highly variable and the boundaries are found to be the most relevant contexts for splicing.

Altogether, 607 features were calculated for these genomic contexts per cassette exon: PhastCons scores [36] that describe sequence conservation, length of the sequence contexts, and sequence composition-based *k*-mer frequencies (with $k \leq 3$) ("genomic" features, the "Methods" section, Additional file 5: Table S4). The *k*-mers reflect the percentage of nucleotides in the context that match the respective specific motif. The PhastCons scores were retrieved for alignments of 99 vertebrate genomes with the human genome from hg19.100way.phastCons.bw from UCSC (May 2018) [35].

In addition to the genomic features, we defined up to 826 DNA methylation features derived from the imputed DNA methylation information, 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 7 sequence contexts of a cassette exon. Methylation features describe the methylation patterns of either individual cells ("genomic and cell methylation" features) or averaged across cells ("genomic and mean methylation" features; Additional file 5: Table S4). More specifically, for the single-cell PSI model, we considered cell-specific methylation levels; the *k*-mer features were extended by including un-methylated (U) and methylated (M) cytosine into the alphabet as follows: 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 averaged methylation values as described above.

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 and splicing variability across cells (inspired by Song et al. [12]): (1) excluded (mean PSI < 0.2), (2) included (mean PSI > 0.8), (3) overdispersed, (4) underdispersed, and (5) multimodal (Fig. 3a). The latter three categories categorize the extent of splicing variation across cells, since cassette exons with intermediate average splicing rates (here $0.2 \leq$ mean PSI ≤ 0.8, Fig. 1) exhibit substantial differences in splicing variance. To characterize cells into these three categories, we calculated the distribution of the distance between the observed and the expected variation per cell type. The expected variation was calculated by a scaled binomial standard deviation, where the scaling factor and the mean splice rate of the alternative exon [18] are 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 third quartile plus 1.5x interquartile range (IQR) (corresponding to > 0.016 in iPS and > 0.022 in endoderm). Likewise, for the definition of the underdispersed cassette exons, we used the first quartile minus 1.5x IQR as the threshold (corresponding to less than − 0.032 in iPS and less than − 0.039 in endoderm cells). The remaining cassette exons were assigned to the multimodal category.

Relating DNA methylation heterogeneity and splicing

We applied Spearman correlation to link splicing at a single locus to variation in DNA methylation observed between cells. The test was performed per sequence context of the cassette exon (Fig. 1c). We only considered cassette exons where variation in splicing and variation of DNA methylation of the relevant context were observed. In total, 5280 iPS and 2622 endoderm cassette exons were tested. The *P* values were adjusted for multiple testing using the *Q* value [37, 38] package in R. The gene enrichment across the cassette exons was performed using g:Profiler [20] (version: 2017-10-25, g:Profiler Ensembl 90), using all observed cassette exons per cell type as background. Multiple testing correction for the enrichments was performed within g:Profiler.

Prediction of PSI and categories

We applied linear ridge regression to model single-cell and pseudo-bulk PSI and (multi-class) logistic ridge regression to model 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-versus-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 based on each single feature. Per feature, we report, in the case of the linear models, 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 the performance and parameters of the models by using a tenfold cross validation (CV) with fixed

training-validation splits. To assess the variability of prediction performances, we repeated the CV procedure four times with different CV splits. Error bars indicate ± 1 standard deviation of the respective statistic (AUC, r^2).

Replication cohort

To replicate our results, we processed the mouse ES single-cell scM&T-seq data ($n = 80$) presented in Angermueller et al. [14]. We reprocessed the aligned RNA and DNA methylation data to quantify splicing following the same protocols that were applied to the human data, with the following changes: GRCm38 was used as a reference for imputation, genome and transcriptome annotations were based on gencode v18 ("GRCm38.p6.genome.fa" as genomic, "gencode.vM18.annotation.gff3" as transcriptomic reference, available at ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M18/ [August 2018]), and conservation scores were taken from "mm10.60way.phastCons.bw" downloaded from UCSC [35] (August 2018).

Out of the 80 cells, in total, 12 cells did not pass quality control on the transcriptome data, Cells with less than 500,000 sequenced reads or had less than 80% of the reads aligned to the genome were removed. Additionally, 4 cells did not pass quality on the DNA methylome data. Cells with less than 1 million reads aligned and bismark mapping efficiency below 7% were discarded. The filters yielded 68 cells that were used for the splicing analysis and 64 that are used for the analyses including DNA-methylation data. In these cells, we quantified between 649 and 1433 cassette exons per mouse ES cell (minimum coverage of 5 reads); in the replication analysis, we considered 2194 exons that were supported by at least 1 cells.

Availability of source code

Python and R were used for data processing, modeling, and visualization of the results. All regression models are based on implementations available in the package scikit-learn [39]. Software and scripts are available as jupyter notebooks at https://github.com/PMBio/scmt_splicing [40].

Additional files

Additional file 1: Table S1. Sample meta-data and quality control characteristics per cell. a. iPS cells and endoderm, c. mouse ES cells. (XLSX 68 kb)

Additional file 2: Table S2. Splicing rates per cell and tissue type. a. iPS cells, b. endoderm, c. mouse ES cells. (XLSX 6381 kb)

Additional file 3: Supplementary Results & **Figure S1-S11.** Supplementary results on the deep modeling of the splicing states in single cells and the Figures S1 to S11. (PDF 2190 kb)

Additional file 4: Table S3. Associations between DNA methylation and splicing as observed in the iPS (a-g) and endoderm (h-n) single cells. Each table contains significant associations per sequence context per tissue. The cassette exon, *P* value (nominal and after multiple testing correction), effect size and effect direction are shown. a. C1 iPS. b. I1 '5 iPS. c. I1 '3 iPS. d. A iPS. e. I2 '5 iPS. f. I2 '3 iPS. g. C2 iPS. h. C1 endoderm. i. I1 '5 endoderm. j. I1 '3 endoderm. k. A endoderm. l. I2 '5 endoderm S. m. I2 '3 iPS. n. C2 endoderm. (XLSX 163 kb)

Additional file 5: Table S4. Description of features that are used per sequence context to predict splicing, and assignment of these features to the models that include them. (XLSX 59 kb)

Additional file 6: Table S5. Pearson r^2, effect and sign of features predicting sPSI using linear ridge regressions, for both tissue types. (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. (XLSX 572 kb)

Additional file 7: Table S6. Contributions of features to the PCs in the PCA on all cell-specific feature weights that are used for splicing prediction in iPS and endoderm cells. a. PC1. b. PC2. (XLSX 83 kb)

Additional file 8: Table S7. Assignment of splicing categories to cassette exons for both tissue types. a. Splicing categories in iPS cells. b. Splicing categories in endoderm cells. c. Splicing category changes between iPS and endoderm cells. d. Splicing categories in mouse serum cells. (XLSX 297 kb)

Additional file 9: Table S8. AUC, effect and sign of all features when predicting splicing categories in iPS, endoderm and mouse cells (tenfold 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. Prediction of each splicing category based on genomic (a-c), genomic and methylation (d-f) and just mean methylation (g-i) features, for iPS (a,d,g), endoderm (b,e,h) and mouse serum cells (c,f,i). (XLSX 676 kb)

Additional file 10: Table S9. AUC, effect and sign of all features when predicting splicing categories switching from iPS to endoderm (tenfold 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-j. Prediction of presence/absence of category switching between the two tissue types, for each splicing category, with and without methylation features. (XLSX 314 kb)

Abbreviations

ES cell: Embryonic stem cell; iPS cell: Induced pluripotent stem cell; PSI: Splicing ratio

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 & R Argelaguet for helpful discussions and comments on the manuscript.

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.M.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.

Availability of data and materials

All sample metadata 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 corresponding to the main analysis of this manuscript are available via the SRA under bioproject: PRJEB15062 [41], sample metadata is available in Additional File 1: Table S1. The replication dataset, originally generated for the Angermueller et al. [14] paper , is available via the SRA under bioproject: PRJNA300666.
The source code used to produce the linear and categorical models, and data analysis including plots are available as jupyter notebooks via https://github.com/PMBio/scmt_splicing [40].

Authors' contributions

SML contributed to the data processing, splicing ratio modeling, deep modeling, and writing of the manuscript. LU contributed to the splicing variation modeling, supervision of splicing modeling and deep modeling, and writing of the manuscript. DJM contributed to the gene expression data processing and cell-level feature calculations and contributed to the editing. MC, SC, SA, LV, and WR contributed to the data generation and editing. IE supervised the master thesis project this study is based on and contributed to the editing. OS contributed to the supervision of modeling, writing of the manuscript, and study design. MJB contributed to the methylation data processing, supervision of splicing modeling, and writing of the manuscript. All authors read and approved the final manuscript.

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).

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK. [2]European Molecular Biology Laboratory, Genome Biology Unit, 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, Germany. [6]Senckenberg Biodiversity and Climate Research Centre (BiK-F), Frankfurt, 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. [10]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany.

Received: 5 November 2018 Accepted: 29 January 2019
Published online: 11 February 2019

References

1. Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem. 2003;72:291–336. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. 2015;347:1254806. https://doi.org/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. https://doi.org/10.1371/journal.pcbi.1000147.
7. Wainberg M, Alipanahi B, Frey B. Does conservation account for splicing patterns? BMC Genomics. 2016;17:787. https://doi.org/10.1186/s12864-016-3121-4.
8. Huang Y, Sanguinetti G. BRIE: transcriptome-wide splicing quantification in single cells. Genome Biol. 2017;18:123. https://doi.org/10.1186/s13059-017-1248-5.
9. 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.
10. 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. https://doi.org/10.1038/cr.2013.110.
11. 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. https://doi.org/10.1016/j.celrep.2015.01.038.
12. 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. https://doi.org/10.1016/j.molcel.2017.06.003.
13. Welch JD, Hu Y, Prins JF. Robust detection of alternative splicing in a population of single cells. Nucleic Acids Res. 2016;44:e73.
14. 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. https://doi.org/10.1038/nmeth.3728.
15. 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. https://doi.org/10.1093/nar/gkw928.
16. Kilpinen H, Goncalves A, Leha A, Afzal V, Ashford S, Bala S, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. 2016.
17. 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. https://doi.org/10.1186/s13059-017-1189-z.
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. https://doi.org/10.1038/nature12172.
20. 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.
21. 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. https://doi.org/10.1261/rna.2696311.

22. Avsec Z, Kreuzhuber R, Israeli J , Xu N, Cheng J, Shrikumar A, et al. Kipoi: accelerating the community exchange andreuse of predictive models for genomics. BioRxiv. 2018:375345. https://doi.org/10.1101/375345.
23. 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.
24. 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&amp. T-seq. Nat Protoc. 2016;11:2081–103. https://doi.org/10.1038/nprot.2016.138.
25. 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. https://doi.org/10.1038/nprot.2016.187.
26. 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. https://doi.org/10.1038/nprot.2014.006.
27. Krueger F. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. 2012. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
28. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 2011;17:10–12
29. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
30. 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.
31. 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. https://doi.org/10.1038/nmeth.4197.
32. 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.
33. 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. https://doi.org/10.1038/nmeth.1528.
34. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2. https://doi.org/10.1093/bioinformatics/btr167.
35. 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. https://doi.org/10.1101/gr.229102.
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. https://doi.org/10.1101/gr.3715005.
37. Storey JD. False discovery rates. USA: Princet Univ Princeton; 2010.
38. Bass Jds, Dabney A, Robinson D. qvalue: Q-value estimation for false discovery rate control. 2015. http://github.com/jdstorey/qvalue.
39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011 12;2825–30.
40. Linker S, Urban L, Clark SJ, Chhatiwala M, Amatya S,McCarthy DJ et al. Source code available from: https://doi.org/10.5281/zendo.2542560.
41. Linker S, Urban L, Clark SJ, Chhatiwala M, Amatya S,McCarthy DJ et al. [Dataset] Available from: https://www.ebi.ac.uk/ena/data/view/PRJEB15062