Transcriptome-wide splicing quantification in single cells

Yuanhua Huang [1], and Guido Sanguinetti [1,2,*]

[1]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK; [2]Centre for Synthetic and Systems Biology (SynthSys), University of Edinburgh, Edinburgh, EH9 3BF, UK

[*]To whom correspondence should be addressed. Email: G.Sanguinetti@ed.ac.uk

Abstract

Single cell RNA-seq (scRNA-seq) has revolutionised our understanding of transcriptome variability, with profound implications both fundamental and translational. While scRNA-seq provides a comprehensive measurement of stochasticity in transcription, the limitations of the technology have prevented its application to dissect variability in RNA processing events such as splicing. Here we present BRIE (Bayesian Regression for Isoform Estimation), a Bayesian hierarchical model which resolves these problems by learning an informative prior distribution from multiple single cells. BRIE combines the mixture modelling approach for isoform quantification with a regression approach to learn sequence features which are predictive of splicing events. We validate BRIE on several scRNA-seq data sets, showing that BRIE yields reproducible estimates of exon inclusion ratios in single cells and provides an effective tool for differential isoform quantification between scRNA-seq data sets. BRIE therefore expands the scope of scRNA-seq experiments to probe the stochasticity of RNA-processing.

1 Results and Discussion

Next generation sequencing (NGS) technologies have revolutionised our understanding of RNA biology, illustrating both the diversity of the transcriptome and the richness and complexity of the regulatory processes controlling transcription and RNA processing. Recent, efficient RNA amplification techniques have been coupled with NGS to yield transcriptome sequencing protocols to measure the abundance of transcripts within single cells, known as single-cell RNA-seq (scRNA-seq)[1]. scRNA-seq has provided unprecedented opportunities to investigate the stochasticity of transcription and its importance in cellular diversity. Groundbreaking applications of scRNA-seq include the ability to discover novel cell types[2], to study transcriptome stochasticity in response to external signals[3], to enhance cancer research by dissecting tumour heterogeneity[4], to mention but a few. However, such advances have been limited to explore variability between single cells at the gene level, and we know very little about the global variability of RNA splicing between individual cells. Bulk RNA-seq splicing quantification algorithms cannot be easily adapted to the single cell case due to the minute amounts of starting material, low cDNA conversion efficiency and uneven transcript coverage resulting in intrinsically low coverage and potentially high technical noise[5]. This considerably limits the usefulness of scRNA-seq to investigate questions about RNA processing and splicing at the single cell level.

Splicing analysis has been revolutionised by the advent of (bulk) RNA-seq techniques. Early studies[6] quantified splicing by considering junction reads that are uniquely assigned to an inclusion/ exclusion isoform, necessitating very high coverage depth to achieve confident predictions. The situation can be considerably improved by using probabilistic methods based on mixture modelling, an idea that is at the core of standard tools such as Cufflinks[7] and MISO[8]. Nevertheless, low coverage represents a challenge even for probabilistic methods. Recent work has shown that improved predictions at lower coverage can be achieved by incorporating informative prior distributions within probabilistic splicing quantification algorithms, leveraging either aspects of the experimental design, such as time series[9], or auxiliary data sets such as measurements of PolII localisation[10]. Such auxiliary data are not normally available for scRNA-seq data. Nevertheless, recent studies have also demonstrated that splicing (in bulk cells) can be accurately predicted from sequence-derived features[11]. This suggests that overall patterns of read distribution may be associated with specific sequence words, so that one may be able to construct informative prior distributions that may be learned directly from data.

Here we introduce the Bayesian Regression for Isoform Estimation (BRIE) method, a statistical model that achieves extremely high sensitivity at low coverage by the use of informative priors learned directly from data. Figure 1a presents a schematic illustration of BRIE (see Methods for precise definitions and details of the estimation procedure). The bottom part of the figure represents the standard mixture model approach to isoform estimation introduced in MISO[8] and Cufflinks[7]. This module takes as input the scRNA-seq data (aligned reads) and forms the likelihood of our Bayesian model. The standard mixture model likelihood is then coupled with an informative prior in the form of a regression model (top half of Figure 1a), where sequence derived features are used to explain a priori some of the variability in inclusion ratios. Crucially, the regression parameters can be learned across multiple single cells, thus regularising the task and enabling robust predictions in the face of very low coverage. In the Methods and Supplementary Material we give details of the features used and show that indeed they can be used to provide a highly accurate supervised learning predictor of splicing on bulk RNA-seq data sets (Suppl. Fig S1).

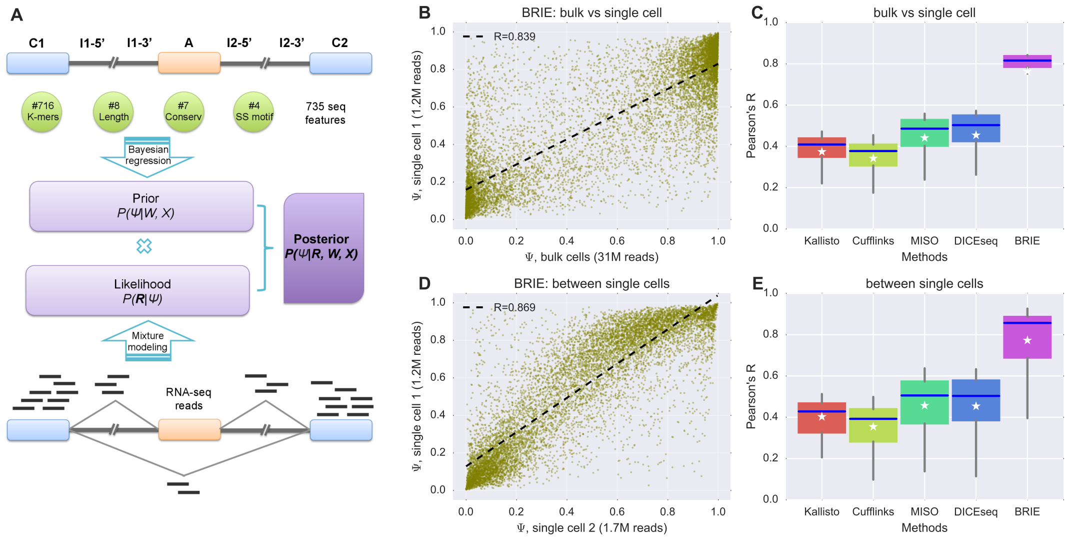

Figure 1: BRIE improves splicing estimates by using sequence features. (A) A cartoon of the BRIE method for isoform estimation, which combines a likelihood computed from RNA-seq data (bottom part) and an informative prior distribution learned from 735 sequence-derived features (top). (B) Scatter plot of exon inclusion ratio estimated by BRIE between bulk RNA-seq and single-cell RNA-seq data from HCT116 cells. (C) Comparing BRIE with other 4 methods on correlation between bulk and single-cell RNA-seq data. (D) Scatter plot of exon inclusion ratio estimated by BRIE between two different single cells. (E) Comparing BRIE with other 4 methods on correlation between pairs of single cells.

To assess the suitability of BRIE as a tool for scRNA-seq splicing quantification, we compared its operational characteristics with four methods for splicing quantification from bulk

RNA-seq: MISO[8] and Cufflinks[7], two of the first and still very widely used probabilistic methods, DICE-seq[9], a modification of MISO using informative priors (for multiple time points), and Kallisto[12], which was recently proposed as one of the most efficient and robust quantification tools. Simulation studies (see Methods and Supplementary Fig S2) show that BRIE achieves significantly higher accuracy in splicing quantification at extremely low coverage values. To assess its performace on real scRNA-seq data, we use 20 scRNA-seq libraries from individual HCT116 human cells from the benchmark scRNA-seq study of Wu et al[13] (see Methods for details). Importantly, a bulk RNA-seq data set in the same conditions was also obtained from one million cells. Figure 1b-e show the results: BRIE clearly outperforms all other methods by a large margin, both in terms of correlation between estimates from different single cells (Fig 1e), and in terms of correlations between estimates from individual single-cells and bulk (Fig 1c). Example scatterplots for both comparisons are given in Fig 1d and 1b, clearly showing very consistent predictions. The high correlation between bulk and scRNA-seq predictions is particularly remarkable, as the analysis of the two data sets is not done with a shared prior. Similarly high correlations were found between splicing estimates obtained by BRIE in single cells and estimates from bulk RNA-seq obtained by other methods (see Suppl. Fig S3). These statistical advantages are reflected in a more effective and confident quantification: considering genes with quantified uncertainty smaller than 0.3 (a threshold adopted e.g. in[14] to select for downstream analysis), Figure S4 shows that BRIE retained 10.9% out of 11,478 genes on average from each single cell (41.1% across all cells), as compared with 3.1% and 5.6% for MISO and DICE-seq, respectively.

BRIE can also be used for differential splicing detection across different data sets. To do so, we compute the evidence ratio (Bayes factor, BF) between a model where the two data sets are treated as replicates (null hypothesis) and an alternative model where the two data sets are treated as separate. We use the Savage-Dickey density-ratio approach and relax it in order to obtain more robust estimates (see Methods). To estimate a background level of differential splicing between identical cells, we considered again the 20 single cell HCT116 libraries from Wu et al[13], and compared all possible pairs of cells. Figure 2a shows the fraction of genes called as differentially spliced at different BF thresholds in this control experiment; as we can see, this number is always very small, and around 1% at the normally recommended threshold of BF=10. This level of background calling could be partly attributed to intrinsic stochasticity or to residual physiological variability that was not controlled for in the experiment, such as cell cycle phase. As an additional comparison, we considered two bulk RNA-seq methods for differential splicing, MISO and the recently proposed rMATS[15]. Both methods could only call a negligible number of events, far fewer than the expected number of false positives, confirming that bulk methods are not suitable for scRNA-seq splicing analysis.

We then considered a mouse early development scRNA-seq data set[16], and compared the single cell transcriptomic profiles from cells from mouse embryos at 6.5 and 7.75 days. We compared both the profiles of individual cells at the same and different time points; the results are summarised in Figure 2b. Comparing individual cells at 6.5 days yielded approximately 1% of events called as significantly differential (BF$\geq$ 10) at 6.5 days. Comparing this result with our investigation of HCT116 cells suggests that murine cells at 6.5 days are still similar to a homogeneous population, from the splicing point of view. The percentage nearly doubled at 7.75 days, suggesting that differential splicing becomes more widespread at this later stage of differentiation. A similar fraction of exon skipping events were differentially called between cells at 7.75 days and cells at 6.5 days. To define a group of differentiation-associated skipping events, we considered events that we called as differential in at least 10% of 7.75 vs 6.5 comparisons. The resulting 159 events were highly enriched for organelle and intracellular part GO terms ($p < 0.01$) (see Supplementary Table S1 and S2). Figure 2c shows the example of DNMT3B,

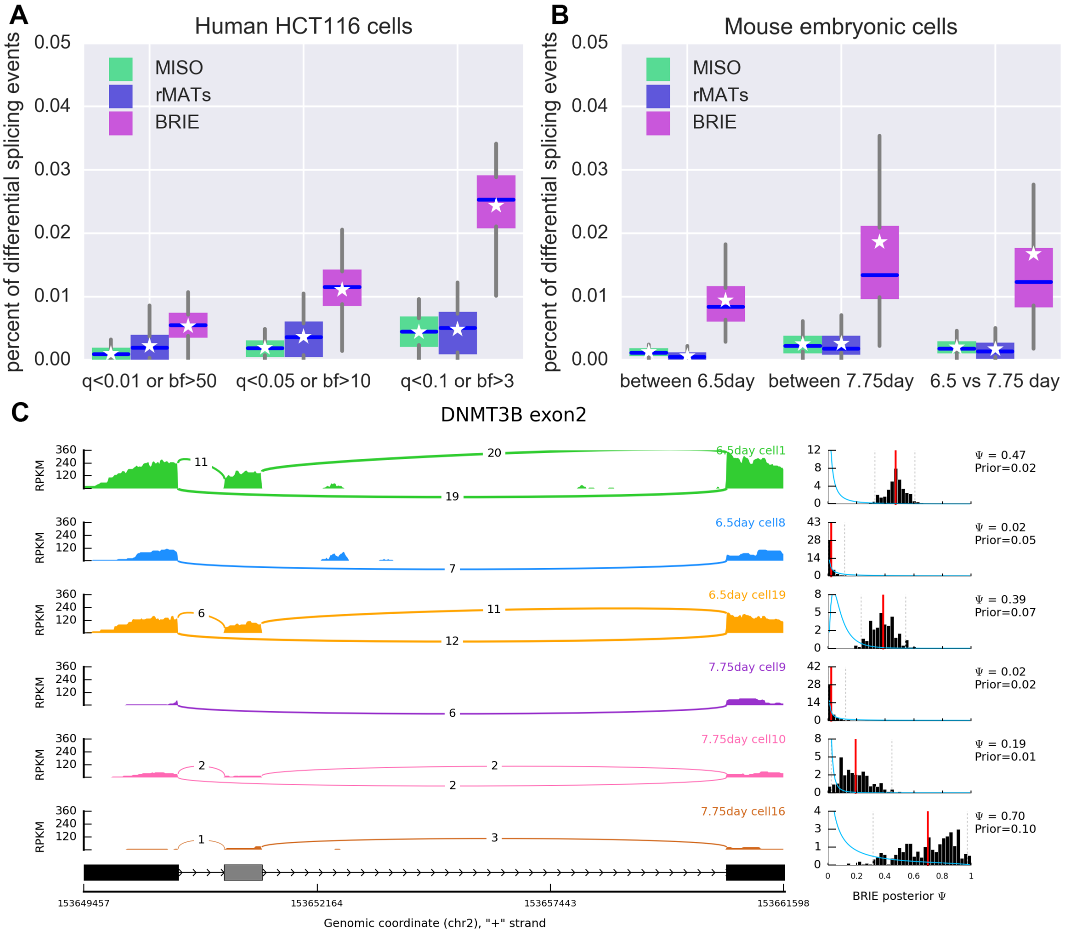

Figure 2: Detection of differential splicing between cells. (A) Percentage of differential splicing events between human HCT116 cells, detected by MISO, rMATS and BRIE with different thresholds. MISO and BRIE use Bayes factor (bf) and rMATS uses q value, i.e., false discovery rate. (B) Percentage of differential splicing events between mouse early embryonic cells at 6.5 day or 7.75 day. The threshold is $bf > 10$ for MISO and BRIE, and $q < 0.05$ for rMATS. (C) An example exon-skipping event in DNMT3B in 3 mouse cells at 6.5s days and 3 cells at 7.75days. The left panel is sashimi plot of the reads density and the number of junction reads. The right panel is the prior distribution in blue curve and a histogram of the posterior distribution in black, both learned by BRIE. For the histogram, the red line is the mean and the dash lines are the 95% confidence interval.

a regulator of DNA methylation maintenance, which is known to undergo functionally relevant alternative splicing [17]. DNMT3B exhibited differential splicing between 7.75 days and 6.5 days in 153 out of 400 comparisons between individual single cells, clearly highlighting the strong differential inclusion effect. Four more example events, all of which have shown differential splicing in more than 100 pairs of comparisons, are presented in Supplementary Figure S5.

Our results demonstrate that BRIE can provide a reliable and reproducible method to quantify splicing levels within single cells. Alternative splicing is a major mechanism of regulation of the transcriptome, and splicing analyses within bulk studies have revealed important associations of splicing with disease. Therefore, the ability to quantify alternative splicing in individual cells would considerably expand the relevance of scRNA-seq technology to investigate variations in RNA processing, and its relevance to diseases. We believe the usage of a data-driven informative prior is essential for this task: directly using bulk RNA-seq methods on scRNA-seq is not a viable route due to the limitations of the technology, an observation that was made earlier[18] that our results confirm. Recent work[19] has addressed the issue of *detection* of alternative splicing across a population of single cells, but as far as we are aware BRIE is the first method to be able to *quantify* splicing in individual single cells, and to detect differential splicing between individual cells from scRNA-seq data. BRIE provides a flexible framework for modelling and, while sequence features are particularly appealing due to their ease of usage and availability, additional side information, such as DNA methylation and chromatin accessibility, could easily be incorporated. BRIE cannot be deployed on all scRNA-seq protocols, as it assumes that sequenced reads can be distributed along whole transcripts. Naturally, protocols such as CEL-seq or STRT-seq that bias reads towards the ends of the transcript cannot provide information about exon skipping events that may be very far from the ends of a transcript. We believe that the availability of splicing quantification approaches such as BRIE can therefore be an important consideration in experimental design, particularly at a time when single-cell omic technologies are about to start being more routinely employed.

2 Methods

2.1 Exon-skipping events annotation

Gene annotations were downloaded from GENCODE human release H22 and mouse release M6. 24,957 and 9,343 exon-skipping events were extracted from protein coding genes on human and mouse, respectively. In order to ensure high quality of the splicing events, we applied 6 constraints following two recent studies[20, 11] for filtering:

1) located on chromosome 1-22 (1-19 for mouse) and X
2) not overlapped by any other AS-exon
3) surrounding introns are no shorter than 100bp
4) length of alternative exon regions between 50 and 450bp
5) with a minimum distance of 500bp from TSS or TTS
6) surrounded by AG-GT, i.e., AG-AS.exon-GT

Consequently, 11,478 and 4,549 exon-skipping events from human and mouse respectively were finally used for this study.

2.2 Feature extraction for Bayesian regression

Following Xiong et al [11], we extract predictive sequence features from the following 7 genomic regions for each exon-skipping event (see cartoon in Figure 1a): C1 (constitutive exon 1), I1-5ss

(300nt downstream from the 5' splice site of intron 1), I1-3ss (300nt upstream from the 3' splice site of intron1), A (alternative exon), I2-5ss (300nt downstream from the 5' splice site of intron 2), I2-3ss (300nt upstream from the 3' splice site of intron 2), C2 (constitutive exon 2).

From these 7 regions, four types of splicing regulatory features are defined. First, 8 length related features are included, i.e., log length of C1, A, C2, I1, I2, and the ratio of the log length of A/I1, A/I2 and I1/I2. Second, the motif strengths of the 4 splice sites, i.e., I1-5'ss, I1-3'ss, I2-5'ss and I2-3'ss, were calculated from mapping each sequences to its averaged position weight matrix. Here, we considered -4nt upstream to +6nt downstream around 5'ss (11nt in total), and from -16nt to 4nt for 3'ss. Third, we also include evolutionary conservation scores for each of the 7 genomic regions, which were calculated by phastCons[21], and are available at the UCSC genome browser. We used the phastCons files in bigWig format with version hg38 for human and mm10 for mouse, where 99 and 59 vertebrate genomes were mapped to the human and mouse genome, respectively. Then the mean conservation scores for the above 7 regions were extracted by using `bigWigSummary` command-line utility. Lastly, 716 short sequences were extracted from the 7 regions, including 1-2mers for I1-5ss and I2-3ss (20 sequences each), and 1-3mers for C1, I1-3ss, I2-5ss and C2 (84 sequences each), and 1-4mers for A (340 sequences). In total, 735 splicing regulatory features were used to predict the exon inclusion ratio in Bayesian regression.

2.3 RNA-seq data and preprocessing

Bulk RNA-seq libraries for K562 cell line were produced by the ENCODE project[22], downloaded from Gene Expression Omnibus (accession number GSE26284); these were used to validate the prediction performance of the splicing regulatory features on bulk RNA-seq (Supplementary Figure S1).

Two single cell RNA-seq data sets were used to validate BRIE model. The first data set is from a benchmark study[13], consisting of 20 single cell RNA-seq libraries from the HCT116 cell line (GEO: GSE51254). These single-cell RNA-seq libraries were prepared with SMART-seq protocol, and have paired-end reads with read length of 125bp. By using a barcode, 48 cells were sequenced per lane, resulting in an average 2.2 million reads per cell. From the same study, two bulk RNA-seq libraries, each with 31.2M reads generated from 1 million HCT116 cells, were also used for comparison.

In order to study differential splicing across different cell types, scRNA-seq data from mouse embryo at embryonic day 6.5 and day 7.75[16] were used. From each of the two groups, 20 individual cells were used, which can be accessed at Array Express (E-MTAB-4079).

All above RNA-seq reads were aligned to the relevant genome reference by HISAT 0.1.6-beta with known splicing junctions.

2.4 Assessing BRIE via a simulation study

In order to assess the performance of BRIE, synthetic reads were generated for 11,478 human exon-skipping events by using Spanki [23]. We assume that the exon inclusion ratio follows a `logitNormal` distribution with mean $\mu = 0.5$ and $\sigma = 3$, which is similar as the distribution of exon inclusion in ENCODE K562 cell line. Then we set all splicing events at the same sequencing coverage, by fixing its RPK, i.e., reads per killo-base in each experiment. Finally, five different coverage levels are used, including $RPK = 25$ (very low, but comparable to a highly covered gene in a scRNA-seq experiment), $RPK = 50, RPK = 100, RPK = 200$ and $RPK = 400$.

Based on the ground truth, we add some noise to generate an informative prior, which has a Pearson's correlation coefficient of 0.8 with the truth. This correlation is similar as that achieved by supervised learning in human and mouse data sets. In addition, random features are also used to give a Null prior, which is named as BRIE.Null. Besides, BRIE and BRIE.Null, we also compare DICE-seq, MISO and Kallisto, in estimating the inclusion ratio from the simulated reads. Supplementary Fig. S2 clearly shows that the use of an informative prior can bring very substantial performance improvements at low coverage, with BRIE essentially maintaining its accuracy levels at all coverage values.

2.5 BRIE model for isoform estimate

Here, we define formally the BRIE statistical model. We consider exon inclusion / exclusion as two different isoforms and adopt the mixture modelling framework for isoform quantification, introduced in MISO[8]. The likelihood of isoform proportions Ψ_i for observing N_i reads $R_{i,1:N_i}$ in sample (single cell) i, can be defined as follows

$$P(R_{i,1:N_i}|\Psi_i) = \prod_{n=1}^{N_i} \sum_{I_{in}=1}^{2} P(R_{in}|I_{in})P(I_{in}|\Psi_i) \tag{1}$$

where the latent variable I_{in} denotes read identity, i.e., the isoform read n in cell i came from. For bulk RNA-seq methods like MISO[8] or Cufflinks[7], the conditional distribution of the read identity $I_{in}|\Psi_i$ is assumed to be a Multinomial distribution, and the prior distribution over Ψ_i is taken to be an uninformative uniform distribution (suitably adjusted to reflect the potentially different isoform lengths). The pre-computed term $P(R_{in}|I_{in})$ encodes the probability of observing a certain read coming from a specific isoform I_{in}.

BRIE enhances the mixture model approach by combining it with a Bayesian regression module to automatically learn an informative prior distribution by considering sequence features. First, we use a `logit` transformation of Ψ_i, i..e, $y_i = \texttt{logit}(\Psi_i)$. We then model the transformed exon inclusion ratio y_i as a linear function of a set of m covariates $X \in \mathbb{R}^m$ (here the covariates are the sequence features described previously): $y_i = W^\top X + \epsilon_i$, where W is a vector of weights shared by all samples and ϵ_i follows zero-mean Gaussian distribution. All exon skipping events are independently modelled with shared W parameters.

Here, we use a conjugate Gaussian prior for the weights, i.e., $W \sim \mathcal{N}(0, \Lambda^{-1})$, with a common choice of $\Lambda = \lambda\mathbf{I}$, for a positive scalar parameter λ. Thus, the graphical representation of the full model is shown in Supplementary Figure S6, and the full posterior is as follows (omitting the cell index for simplicity),

$$P(W, \sigma, \mathbf{\Psi}|\mathbf{X}, \mathbf{R}) \propto P(W|\lambda) \prod_{k=1}^{K} \{P(\Psi_k|X_k, W, \sigma) \prod_{n=1}^{N_k} \sum_{I_n^k=1}^{2} P(R_n^k|I_n^k)P(I_n^k|\Psi_k)\} \tag{2}$$

2.6 Inference in the BRIE model

As shown above, BRIE model involves the whole set of exon-skipping events, thus there are thousands of parameters to infer jointly, which can lead to very high computational costs which are not easily distributed. Therefore, we introduce an approximate method to alternately learn Ψ and W. Also, to alleviate computational burdens, there is an option to merge reads from all cells to learn parameters. For simplicity, we set λ empirically, using the value $\lambda = 0.1$ which gave the best predictive performance on tests on ENCODE data. Then, we collapse W and σ by taking their expected value in Bayesian regression given a set of Ψ, i.e., $W = (\mathbf{X}^\top\mathbf{X} + \sigma^2\Lambda)^{-1}\mathbf{X}^\top\mathbf{Y}$ and $\sigma = \texttt{std}(\mathbf{Y} - W^\top\mathbf{X})$. At a single exon-skipping event level, we used an adaptive Metropolis-Hastings sampler to sample Ψ, where a univariate Gaussian distribution is used for proposal with adaptive variance, i.e., $\eta = 2.38*\texttt{std}(y^{(1:m)})$. At this step, we could run short parallel MCMC chains on multiple events to alleviate computational costs, for example $h = 50$ steps if the total iteration is $n*h = 1000$. Pseudocode to sample from the (approximate) posterior distribution of Ψ is given in Algorithm 1. Also, this model supports fixed W and σ, which can be learned from other data sets, e.g. bulk RNA-seq; then the line 3 and 5 will be turned off in Algorithm 1. The convergence of the sampling is diagnosed by using the Geweke diagnostic Z score; in our experiments 1000 burn-in steps appeared to be sufficient in all cases.

```
Algorithm 1: Approximation of $\mathbf{\Psi}, W, \sigma$
   Data: $\mathbf{X}, \mathbf{R}, \Lambda$; optional: $W$ and $\sigma$
   Result: $\mathbf{\Psi}, W, \sigma$
1  initialization $\mathbf{Y}^{(0)} = \mathbf{0}; \sigma = 1.0; \eta = 1.0$
2  for $i \leftarrow 0$ to $n$ do
3      $W^{(i)} = (\mathbf{X}^\top \mathbf{X} + \sigma^2 \Lambda)^{-1} \mathbf{X}^\top \mathbf{Y}^{(i*h)}$
4      $\bar{\mathbf{Y}} = W^{(i)\top} \mathbf{X}$
5      $\sigma = \texttt{std}(\mathbf{Y}^{(i*h)} - \bar{\mathbf{Y}})$
6      for $k \leftarrow 1$ to $K$ do
7          if $i * h > 10$ then
8              $\eta = 2.38 * \texttt{std}(y_k^{(0:i*h)})$
9          for $j \leftarrow i * h$ to $(i+1) * h$ do
10             Sample: $\mu \sim U(0,1)$; $y_k^* \sim Q_y(y_k^* | y_k^{(j)}, \eta)$
11             Calculate: $P(y_k^*|R) = \mathcal{N}(y_k^* | \bar{y_k}, \sigma) P(R|y_k^*)$
12             if $\mu < \texttt{min}\left\{ \frac{P(y_k^*|R) \times Q_y(y_k^{(j)}|y_k^*, \eta)}{P(y_k^{(j)}|R) \times Q_y(y_k^*|y_k^{(j)}, \eta)}, 1 \right\}$ then
13                 $y_k^{(j+1)} \leftarrow y_k^*; \Psi_k^{(j+1)} \leftarrow \texttt{logistic}(y_k^*)$
14             else
15                 $y_k^{(j+1)} \leftarrow y_k^{(j)}; \Psi_k^{(j+1)} \leftarrow \texttt{logistic}(y_k^{(j)})$
16 return $W^{(0:n)}, \mathbf{\Psi}^{(0:n*h)}$
```

2.7 Detection of differential splicing using Bayes factors

The Bayes factor[24] is a posterior odds in favor of a hypothesis relative to another, and is also able to detect whether splicing in two cells or conditions are different or not.

To detect differential splicing between two cells (or conditions), A and B, $\delta = \Psi_A - \Psi_B$, we introduce a null hypothesis (H_0) as $\delta \approx 0$, and the alternative hypothesis (H_1) as $\delta \not\approx 0$. Here, D is the data used to sample the posterior of Ψ in two cells. Then, the Bayes factor in favor of the alternative hypothesis on observing data D is defined as follows,

$$\texttt{BF} = \frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_0)P(H_0)} \tag{3}$$

As usual, we assume that both hypotheses have the same prior, i.e., $P(H_1) = P(H_0)$, and we can clearly see that $P(D|H_0) = P(D|\delta \approx 0, H_1)$. Therefore, by taking the Savage-Dickey density ratio [25], we could simplify the calculation of BF as follows,

$$\texttt{BF} = \frac{P(D|H_1)}{P(D|\delta \approx 0, H_1)} = \frac{P(\delta \approx 0|H_1)}{P(\delta \approx 0|D, H_1)} = \frac{P(-\epsilon < \delta < \epsilon|H_1)}{P(-\epsilon < \delta < \epsilon|D, H_1)} \tag{4}$$

where ϵ can be set as 0.05.

As BRIE samples Ψ_A and Ψ_B following their posteriors, the distribution of $P(\delta|D, H_1)$ is readily to approximate by empirically re-sampling $\Psi_A - \Psi_B$. With a set of re-sampled $\delta_{1:M}$, we take the proportion of $|\delta_i| < \epsilon$ as the posterior probability $P(-\epsilon < \delta < \epsilon|D, H_1)$. Similarly, we could sample a set of $\hat{\Psi}_A$ and $\hat{\Psi}_B$ following their prior distributions, and use the same procedure to approximate the prior probability $P(-\epsilon < \delta < \epsilon|H_1)$.

2.8 Software

BRIE model has been implemented as a standard Python package, which is freely available in the following repository: http://github.com/huangyh09/brie.

Acknowledgements

We thank Chris Ponting and Jean Beggs for useful feedback on an early draft of the manuscript. G.S. acknowledges support from the European Research Council under grant MLCS306999. Y.H. is supported by the University of Edinburgh through a Principal Career Development scholarship.

References

[1] Dominic Grün and Alexander van Oudenaarden. Design and analysis of single-cell sequencing experiments. *Cell*, 163(4):799–810, 2015.

[2] Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, 2015.

[3] Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, et al. Single cell RNA Seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363, 2014.

[4] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[5] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095, 2013.

[6] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[7] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010.

[8] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, 2010.

[9] Yuanhua Huang and Guido Sanguinetti. Statistical modeling of isoform splicing dynamics from RNA-seq time series data. *Bioinformatics*, 32(19):2965–2972, 2016.

[10] Peng Liu, Rajendran Sanalkumar, Emery H Bresnick, Sündüz Keleş, and Colin N Dewey. Integrative analysis with ChIP-seq advances the limits of transcript quantification from RNA-seq. *Genome Research*, 26(8):1124–1133, 2016.

[11] Hui Y Xiong, Babak Alipanahi, Leo J Lee, Hannes Bretschneider, Daniele Merico, Ryan KC Yuen, Yimin Hua, Serge Gueroussov, Hamed S Najafabadi, Timothy R Hughes, et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218):1254806, 2015.

[12] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.

[13] Angela R Wu, Norma F Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E Rothenberg, Francis M Mburu, Gary L Mantalas, Sopheak Sim, Michael F Clarke, et al. Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods*, 11(1):41–46, 2014.

[14] J David Barrass, Jane EA Reid, Yuanhua Huang, Ralph D Hector, Guido Sanguinetti, Jean D Beggs, and Sander Granneman. Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. *Genome Biology*, 16(1):1, 2015.

[15] Shihao Shen, Juw Won Park, Zhi-xiang Lu, Lan Lin, Michael D Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, 111(51):E5593–E5601, 2014.

[16] Antonio Scialdone, Yosuke Tanaka, Wajid Jawaid, Victoria Moignard, Nicola K Wilson, Iain C Macaulay, John C Marioni, and Berthold Göttgens. Resolving early mesoderm diversification through single-cell expression profiling. *Nature*, 535(7611):284–293, 2016.

[17] Christopher E Duymich, Jessica Charlet, Xiaojing Yang, Peter A Jones, and Gangning Liang. DNMT3B isoforms without catalytic activity stimulate gene body methylation as accessory proteins in somatic cells. *Nature Communications*, 7, 2016.

[18] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

[19] Joshua D Welch, Yin Hu, and Jan F Prins. Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Research*, 44(8):e73–e73, 2016.

[20] Joao Curado, Camilla Iannone, Hagen Tilgner, Juan Valcárcel, and Roderic Guigó. Promoter-like epigenetic signatures in exons displaying cell type-specific splicing. *Genome biology*, 16(1):1–16, 2015.

[21] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

[22] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[23] David Sturgill, John H Malone, Xia Sun, Harold E Smith, Leonard Rabinow, Marie-Laure Samson, and Brian Oliver. Design of RNA splicing analysis null models for post hoc filtering of Drosophila head RNA-Seq data with the splicing analysis kit (Spanki). *BMC Bioinformatics*, 14(1):1, 2013.

[24] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[25] Isabella Verdinelli and Larry Wasserman. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618, 1995.