Computational identification of splicing phenotypes from single cell transcriptomic experiments

Yuanhua Huang[1,2,#] and Guido Sanguinetti[3,4,#]

[1]School of Biomedical Sciences, [2]Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong SAR; [3]School of Informatics, University of Edinburgh, UK; [4]SISSA, Trieste, Italy.
[#]Corresponding authors.

Abstract

RNA splicing is an important driver of heterogeneity in single cells, and a major determinant of the dynamical state of developing cells. However, the intrinsic coverage limitations of scRNA-seq technologies make it challenging to associate specific splicing events to cell-level phenotypes. Here, we present BRIE2, a scalable computational method that resolves these issues by regressing single-cell transcriptomic data against cell-level features. We show on different biological systems that BRIE2 effectively identifies differential splicing events that are associated with disease or developmental lineages, and detects differential momentum genes for improving RNA velocity analyses. BRIE2 therefore extends the scope of single-cell transcriptomic experiments towards the identification of splicing phenotypes associated with biological changes at the single-cell level.

1 Introduction

Single-cell RNA-sequencing (scRNA-seq) has rapidly become the key technology to disentangle transcriptional heterogeneity in cell populations. Over the last five years, scRNA-seq has been successfully applied both to identify discrete cell states or sub-populations in normal or diseased tissues, e.g. [1, 20], and to infer continuous stages in cellular processes, e.g., pseudo-time [19] and cell differentiation [2]. More recently, scRNA-seq has further been applied to multiple-sample designs with different donors, tissues, diseases or treatments. These experiments enable the discovery of cell type specific marker genes [11] or key pathways that are associated with the meta labels [20].

Beyond gene level information, RNA processing within a gene also holds rich information for both categorical cell states and continuous cell differentiation. A key RNA processing step is splicing, where a precursor mRNA (pre-mRNA or unspliced RNA) is spliced by removing intronic, non-coding regions, resulting in mature mRNA (or spliced RNA). Alternative splicing of exons further extends the molecular feature space, greatly contributing to cellular heterogeneity. A variety of studies have found that the abundance of splicing isoforms enables to identify cell states [22] or disease conditions [6]. Additionally, the intrinsic kinetics of splicing provides a footprint of cellular dynamics during cell differentiation, which has motivated the recent flourishing of RNA velocity studies [15, 4] and time-series scRNA-seq on metabolically labelled nascent RNAs [5, 7, 3, 17].

Despite the fundamental role of RNA splicing, stochasticity in splicing is much less understood than stochasticity in gene-level expression, primarily due to the technical difficulties in recovering splicing information from scRNA-seq data. First, scRNA-seq data is highly sparse, particularly for droplet based protocols including the popular 10x genomics. This high sparsity,

along with minute initial molecule counts, leads to very high technical noise in scRNA-seq data, hence requiring careful statistical modelling. Second, splicing adds new layers of complexity to scRNA-seq analyses, and the requirements to quantify relative abundances of isoforms from indirect observations of fragment counts creates considerable computational difficulties. For all these reasons, the level of heterogeneity in splicing between different cells has been difficult to quantify. Perhaps more importantly, the identification of single-cell level splicing phenotypes, splicing events associated with disease or genetic changes, has been largely unfeasible, hindering an understanding of the role of splicing changes and aberrations in cellular state.

In this work, we address these issues by directly incorporating the association of splicing phenotypes within the splicing quantification task itself. We introduce BRIE2, a Bayesian hierarchical model that predicts the splicing ratio (spliced vs unspliced or exon-inclusion vs exclusion) from a set features associated with cell-type/ state, as well as with the specific splicing event to be quantified. This enables us to robustly identify genes that are associated with each cell level feature, while controlling and quantifying in a Bayesian manner the uncertainty from the noise and sparsity of the data. Additionally, BRIE2 provides us with an efficient way to select biologically relevant features for RNA-velocity analyses, which leads to more consistent and interpretable visualisations of biological process dynamics.

2 Results and discussion

2.1 Model Description and Evaluation

An unavoidable difficulty in splicing quantification from short-read protocols derives from the fundamental ambiguity of the data, as the vast majority of reads cannot be unambiguously assigned to a single isoform. This problem is compounded in scRNA-seq by the generally low number of reads, which frequently results in no unambiguous reads being mapped to a specific isoform. Our earlier work, BRIE [9], resolved this issue by introducing latent variables conditioned on sequence features through a Bayesian regression approach, therefore using genomic sequence to regularise and inform splicing predictions. BRIE2 innovates over BRIE in two important ways: first of all, it augments the set of regressor features to include cell-specific features such as cell-type/ developmental stage (Fig. 1, Supplementary Fig. S1, and Methods). This enables us to statistically associate cell-level features with specific splicing events, thus defining quantitatively splicing as a single-cell level intermediate phenotype, but it considerably increases the complexity of the model (as data from all cells needs to be analyzed jointly). To cope with the added complexity, BRIE2 is formulated as a variational discriminative model, thus enabling the use of advanced software (Tensorflow) and hardware (GPUs), and leading to orders of magnitudes in computational seed-ups (>1,000 speed-ups; see Supplementary Fig. S2 and Methods).

We then validated the BRIE2 model against realistic simulated data in order to assess its ability to accurately reconstruct splicing ratios and to detect splicing phenotypes. By comparing to ground truth in simulations, we found BRIE2 retains high accuracy (Pearson's R>0.95) on splicing ratio quantification when there are more than 4 unambiguous reads, no matter if informative prior is learned from features (Supplementary Fig. S3a-c). Importantly, the inclusion of cell-level features enables a significant increase in accuracy for lowly covered genes (Peason's R increasing from 0.88 to 0.98, Supplementary Fig. S3d). Additionally, the estimated coefficients of cell or gene features are highly correlated with the ground truth values used to generate the data (Pearson's R=0.93 for coefficients with FDR<0.1, Supplementary Fig. S3e-f).

To assess the quality of the variational approximation, we compared results to MCMC based estimation on a real data set with 130 mesoderm cells [21] with 19 gene level sequence features,

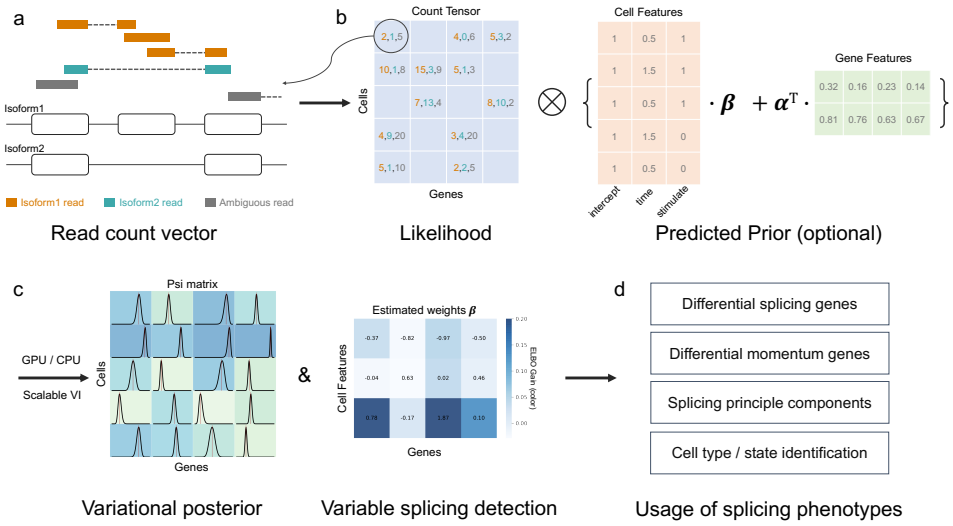

Figure 1: Illustration of BRIE2. (a) Reads are counted into isoform 1, isoform 2 or ambiguous groups according to its alignment identity, which constitutes a cell-by-gene-by-3 tensor. (b) The posterior distribution of isoform proportion Psi is defined by combining the likelihood from read counts and informative prior predicted by cell level covariates and/or gene sequence features. (c) A `logit-normal` variational posterior and coefficients on covariates are optimised to approximate the exact posterior, where the evidence lower bound gain (EBLO) between including and excluding a certain cell feature set can be leveraged to select splicing phenotypes. (d) The selected differential splicing events or differential momentum genes on RNA velocity can be used as markers for downstream analysis, and the estimated Psi can be used for dimension reduction to enhance cell type/ state identification.

observing that the estimates by variational inference in BRIE2 are highly concordant with the MCMC estimates both for Psi (Pearson's R>0.99 for confident genes, Supplementary Fig. S4) and feature coefficients (Pearson's R=0.87).

BRIE2 can also detect genes with differential splicing ratios associated with cell level covariates by performing Bayesian model selection. To do so, we run BRIE2 twice for both with and without the candidate feature(s), and calculate the difference on the evidence lower bounds, termed `ELBO_gain`. The `ELBO_gain` approximates the empirical Bayes factor and mimics the log likelihood ratio in hypothesis tests, hence in practice it can be transformed to a p value through a chi-square distribution (Methods). We found the transformed p values are well calibrated in the null model (Supplementary Fig. S5). Further, in detecting splicing events that are significantly associated with cell level features, BRIE2 returns excellent performance in both sensitivity and specificity (Supplementary Fig. S6): on cell feature with moderate correlation to splicing ratio (Pearson's R=0.47), BRIE2 achieves AUROC>0.986 in detecting 400 significant events out of 2,248 splicing events. Therefore, in practice, both `ELBO_gain` and its transformed false discovery rate (FDR) after multiple testing correction can be used as a significance cut off. In the rest of the manuscript, we use $\texttt{FDR} < 0.05$ as a general threshold for significance.

2.2 BRIE2 discovers hundreds of differential splicing events associated with multiple sclerosis

Next, we applied BRIE2 to analyse alternative splicing in multiple sclerosis, a neurological autoimmune disease. Falcão *et al* have generated 2,208 mouse cells using the SMART-seq2 protocol, with equal number of cases and controls [6]. Here, we analysed 3780 exon-skipping events that satisfied the quality control, e.g., more than 30 cells with unique reads (Methods), resulting in 1,876 cells that have more than 3,000 total reads on the above genes.

We first applied BRIE2 to quantify `Psi` by regressing on an intercept term, i.e., a constant

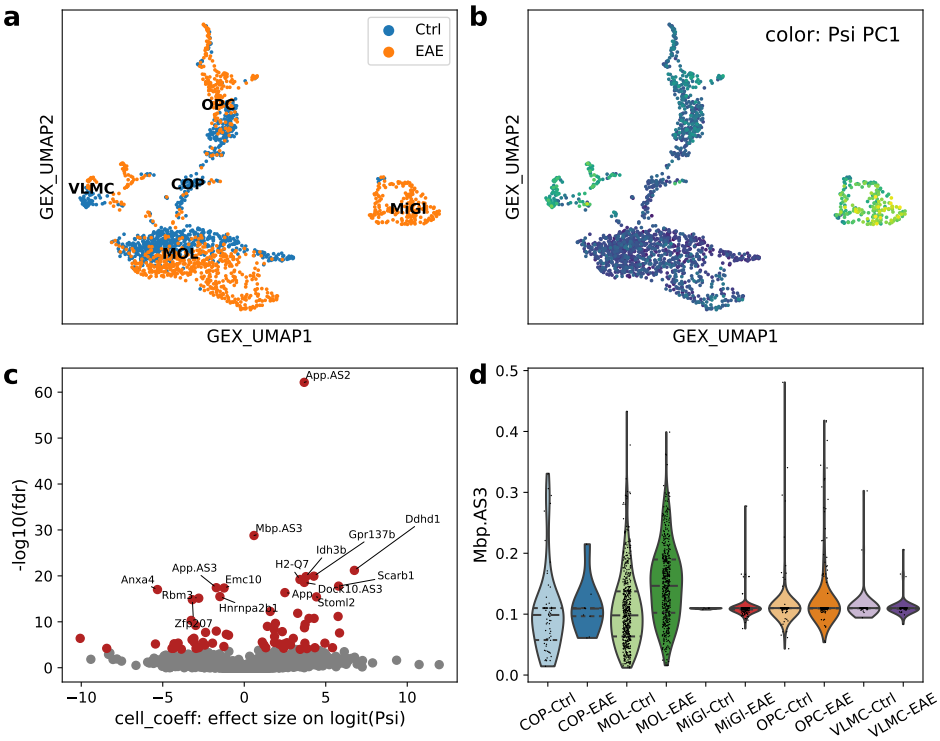

Figure 2: Differential splicing events on multiple sclerosis. (a) Umap visualization of gene level expression, annotated with cell types and EAE state. (b) Umap coloured by the first principle component based on Psi matrix, which suggests that Psi PC has a global impact on cell types. (c) Volcano plot between -log10(FDR) and effect size on logit(Psi) for detecting differential splicing between EAE and control cells by BRIE2. (d) Violin plot on example gene Mbp (the exon3) for estimated Psi between EAE and control in each cell type. Psi values in panel (b) and (d) are quantified by only using unity cell feature for aggregation, but not EAE state label. EAE: Experimental Autoimmune Encephalomyelitis.

cell feature. In other words, a prior is learned by aggregating all cells, which reflects an average-based imputation. Based on the `Psi` matrix, we performed a principal component analysis and found that the `Psi` principal component has strong cell type specificity (see Fig. 2a-b for the first PCs). By comparing clusters identified with gene expression and `Psi` PCs, we found that top 20 `Psi` PCs can accurately predict the cell clusters (overall AUC = 0.97, Supplementary Fig. S7).

Furthermore, by incorporating disease state and strain labels as cell covariates, BRIE2 detects 368 differential splicing events across 348 genes with FDR<0.05 that are associated with disease condition (Fig. 2c, Supplementary Fig. S8). Particularly, the myelin genes Mbp (FDR=1e-28; Fig. 2d) and Pdgfa (FDR=1e-5) are both identified as differential splicing events, which was highlighted in the original study [6] by using BRIE1. In addition, when leveraging these 368 MS-related splicing events, we found their `Psi` principal components can predict the disease state on MOL, the largest cell type with well balance, at moderate level (AUC=0.76), and can enhance disease state predictions, as compared with using gene expression alone (AUC from 0.954 to 0.968, Supplementary Fig. S9).

2.3 Differential momentum genes improve RNA velocity analyses

Global RNA-processing efficiency has recently been used to define the concept of RNA velocity associated with an individual cell [15, 4], which is rapidly becoming a major tool to study the

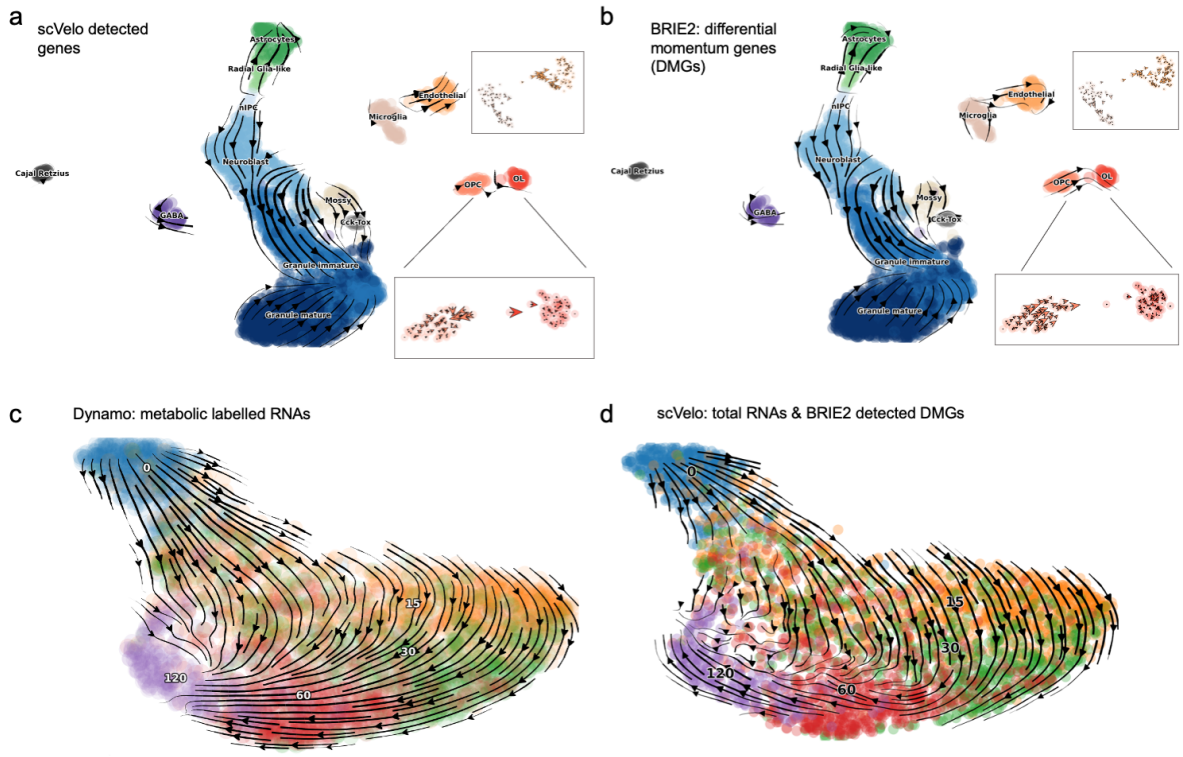

Figure 3: Differential momentum genes for RNA velocity. (a-b) Cell differentiation in neurogenesis inferred from RNA velocity by scVelo with different gene sets: (a) scVelo detected gene set requiring positive correlation between unlicensed and spliced RNAs; and (b) BRIE2 detected gene set that have differential spliced ratio in one cell type vs all others, which are termed as differential momentum genes. (c-d) State transitions of excitatory neurons inferred from RNA velocity with different methods: (c) Dynamo using metabolic labelling information measured by scNT-seq. (d) scVelo using total RNAs on 20 differential momentum genes detected by BRIE2. The colour denotes the time since stimulations: 0 (blue), 15 (orange), 30 (green), 60 (red) and 120 (purple) minutes.

dynamics of cellular processes at the single-cell level. The concept of RNA velocity is based on treating spliced and unspliced RNAs as two different isoforms, associating to each gene in each cell an RNA-processing speed which is then combined (and projected using any visualisation tool) to quantitate the dynamics of cellular processes at the molecular level.

Standard RNA velocity analyses are fully unsupervised, thus discarding available annotations during the (frequently crucial) step of selecting genes for velocity estimates. Instead, we propose to us BRIE2 to detect genes that have differential splicing ratios associated with cell-level covariates, thus providing a biologically informed approach to selecting features to compute RNA velocities associated with cell transitions. We term these genes as *differential momentum genes* (DMG), as the differential splicing ratio implies a departure from equilibrium between splicing and degradation rates, likely due to changes on synthesis rate associated with changes in cell type.

To see the impact of using DMGs in RNA-velocity analyses, we re-analyzed the neurogenesis data set in [4], which well illustrates the impact of gene selection on cell transition inference. We used BRIE2 to detect cell type specific DMGs by using each cell type as the testing covariate, and accounted for differences in coverage between cells by using gene detection rate as an additional cell-level covariate. We therefore examine the effect of using BRIE2 as a pre-selection step in velocity analyses, applying the same downstream modelling to DMGs and default genes selected by the package scVelo [4]. The stochastic model is used here for illustrating that the differentiation direction can be corrected by using informative genes. In Fig. 3, we compare

the cell differentiation paths inferred from RNA velocity based on the 634 genes selected by the package scVelo [4] and the 335 DMGs selected by BRIE2 (`FDR` < 0.05 in any cell type), both selected out of the initial 3,000 quality-pass genes (see read counts for top genes in Supplementary Fig. S10).

While the overall picture is broadly in agreement, DMGs obtained from BRIE2 highlighted a more obvious direction from oligodendrocyte precursor cells (OPCs) to myelinating oligodendrocytes (OLs) compared to scVelo, both using the stochastic model (Fig. 3a-b) and dynamical model (Fig. 2 in [4]) approaches of scVelo. Observing more in detail this biological transition (Supp. Fig. S11-12), we see that scVelo directions are inconsistent on a subgroup of OPCs, while DMGs at different FDR consistently estimate the correct transition direction from OPC to OL. These observations highlight the importance of feature selection when visualising cell transitions: in this light, DMGs detected by BRIE2 are likely to return more biologically informative angles, thanks to the use of annotations.

Furthermore, we examined how selection of DMGs improve the inference of cell state transition in time-series of neuronal scRNA-seq data generated by scNT-seq [17]. scNT-seq is a recently proposed technique where nascent RNAs are metabolically labelled, effectively providing a measurement of the age of a transcript. Using the information of metabolic labelling provides an effective ground truth and enables a consistent visualisation where cell transitions are strongly aligned with the time direction [18] (Fig. 3c). In the original paper, it was observed that such transitions are difficult to obtain only using the total RNAs; our own experimentations confirm that scVelo struggles to identify the right direction in the early stage of stimulation (i.e., 0 to 15 or 30min; Supplementary Fig. S13). Applying BRIE2 to detect DMGs by using the stimulation time as testing covariate, we found 280 DMGs significantly associated with time (FDR$<$0.01; Supplementary Fig. S14-S15), with 141 genes overlapped with the top 2,000 highly variable genes selected by scVelo. By projecting the RNA velocity on these 141 DMGs, the cell transitions are largely corrected to the expected direction along the time (Fig. 3d). This pattern remains even if varying the cut-off at FDR$<$0.001 for 89 more stringent DMGs or FDR$<$0.05 for 165 more lenient DMGs (Supplementary Fig. S16). Taken together, these results demonstrate that BRIE2 is an effective tool to select informative genes underlying dynamical processes.

3 Discussion

Splicing is a fundamental step in gene expression in higher eukaryotes, and has the potential to represent an important intermediate phenotype in single cell experiments. BRIE2 provides an effective and computationally efficient approach to link such intermediate phenotypes to cell-level covariates. Our results showed that BRIE2 identifies hundreds of splicing events linked to multiple sclerosis, and that inclusion of splicing events leads to improved cell-type classification on this translationally relevant data set.

While quantification of splicing events is certainly biologically important, it is likely to only be possible using technologies that sample evenly the transcriptome. Recent years, instead, have seen an increasing popularity of technologies which can upscale the number of cells assayed by sequencing only parts of the transcriptome (typically, the regions immediately preceding the polyA tail). Despite this enrichment, many such data sets still present a substantial number of intronic reads (presumably due to the abundance of repetitive A sequences within introns) which can be used to measure changes in RNA kinetics (so called RNA velocity) and therefore provide a more accurate description of transitional cell states in large data sets. Our results showed that, in the presence of cell annotations, BRIE2 can be a useful tool to select relevant genes (differential momentum genes) which provide a smoother and more interpretable description of cell transitions within RNA velocity studies.

4 Methods

4.1 Modelling of splicing isoform abundance

In this study, we jointly analyse N splicing genes (i.e., segments) across M cells, and we focus on two-isoform splicing events, for example exon-skipping (SE) and intron retention (IR). For a splicing gene i in a cell j, we use $\psi_{i,j}$ to denote the fraction of a certain isoform; for conventional reason, it refers to isoform with exon-inclusion in SE event. Without losing generality, we define the BRIE2 model on SE event here but it is applicable to any other two-isoform event.

In order to scale up the analysis across a large number of cells, reads aligned to a splicing gene are not modelled individually but rather aggregated into three groups depending on their isoform identity:

- group1: reads from isoform1 explicitly, e.g., on the junction between exon1 and exon2;
- group2: reads from isoform2 explicitly, e.g., on the junction between exon1 and exon3;
- group3: reads with ambiguous identity e.g., within exon3.

Thus, from the aligned reads file we could extract the count vector $\boldsymbol{s}_{i,j} = [s_{i,j,1}, s_{i,j,2}, s_{i,j,3}]$ for these three groups, with $n_{i,j} = \sum_{k\in 1,2,3} s_{i,j,k}$ as the total count. In addition, for each gene i we can pre-define the effective length $l_{i,h,k}$, i.e., the (effective or corrected) number of positions in isoform h that can generate read being located in the region of read group k. This gene specific 2-by-3 length matrix L_i can be defined from the exon structures encoded in the gene annotation, and the read counts are proportional to the effective lengths.

Given the total read counts $n_{i,j}$ and its according effective length matrix L_i, we could have the base likelihood of $\psi_{i,j}$ (or equivalently its transformation $z_{i,j} := \texttt{logit}(\psi_{i,j})$) for observing the three-group reads counts $\boldsymbol{s}_{i,j}$ by a multinomial distribution, whose proportion vector $\boldsymbol{\rho}_{i,j}$ is coded by $\psi_{i,j}$ and the effective length matrix L_i as follows,

$$\begin{aligned} p(\boldsymbol{s}_{i,j}|z_{i,j}) = p(\boldsymbol{s}_{i,j}|n_{i,j}, \psi_{i,j}, L_i) = \texttt{Multinomial}(\boldsymbol{s}_{i,j}|n_{i,j}, \boldsymbol{\rho}_{i,j}) \\ \rho_{i,j,k} = \frac{\psi_{i,j} l_{i,1,k} + (1-\psi_{i,j}) l_{i,2,k}}{\sum_{t\in 1,2,3} \psi_{i,j} l_{i,1,t} + (1-\psi_{i,j}) l_{i,2,t}}, k \in 1,2,3. \end{aligned} \tag{1}$$

By definition, we have $l_{i,1,2} = l_{i,2,1} = 0$ for all genes. Taking the assumption of conditional independence, we could have the joint likelihood for all N splicing genes in M cells by taking their product as follows,

$$p(S|Z) = \prod_{i=1}^{N} \prod_{j=1}^{M} p(\boldsymbol{s}_{i,j}|z_{i,j}).$$

4.2 Bayesian regression on splicing

In BRIE2 model (see graphical representation in Supplementary Fig. S2), we aim to identify the regulatory factors on splicing from both gene level features $\boldsymbol{x}$ (e.g., splice site motif) and / or cell level features $\boldsymbol{y}$ (e.g., cell type) via a generalised linear model. Specifically, we assume that the `logit` of the fraction of isoform1 $z_{i,j}$ is linear prediction of $\boldsymbol{x}$ and $\boldsymbol{y}$ as follows,

$$z_{i,j} = \boldsymbol{x_i}^\top \boldsymbol{\alpha_j} + \boldsymbol{\beta_i}^\top \boldsymbol{y_j} + \epsilon_{i,j}, \tag{2}$$

where we could use a deterministic way by taking $\epsilon_{i,j} := 0$, and assume the uncertainty only comes from the regression weights. On the other hand, we could introduce $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma_i)$ to

account for gene specific over dispersion, which is particularly important for the phenomenon of mono-isoform in single cells.

With considering the over dispersion by adding a gene specific σ_i, we could have a predicted distribution with $z_{i,j}$

$$p(z_{i,j}|\boldsymbol{\alpha_j}, \boldsymbol{\beta_i}, \sigma_i) = \mathcal{N}(z_{i,j}|\boldsymbol{x_i}^\top \boldsymbol{\alpha_j} + \boldsymbol{\beta_i}^\top \boldsymbol{y_j}, \sigma_i^2), \tag{3}$$

which can be treated as an informative prior on z (and according logit-normal distribution for ψ).

4.3 Bayesian Inference in BRIE2

Besides estimating the parameters for regression model in Eq.(3), it is often of high interest to approximate the posterior distribution of the isoform abundance Ψ or its logit transformation Z. Therefore, it is crucial to keep Z as auxiliary variable instead of marginalizing out. By taking the product of the base distribution defined in Eq.(1), and the predicted prior in Eq.(3), we could have the joint distribution to which the posterior distribution $p(Z|S, \mathrm{A}, \mathrm{B}, \boldsymbol{\sigma})$ is proportional as follows,

$$\begin{aligned} p(Z|S, \mathrm{A}, \mathrm{B}, \boldsymbol{\sigma}) &\propto p(Z, S, \mathrm{A}, \mathrm{B}, \boldsymbol{\sigma}) \\ &= p(S|Z)p(Z|\mathrm{A}, \mathrm{B}, \sigma). \end{aligned} \tag{4}$$

This posterior is intractable and it also has parameters to optimize. In the BRIE v1, we used an approximate algorithm to alternately optimize the parameters and sampling the posterior with Metropolis-Hastings algorithm [9]. Here, instead we are using a variational inference to approximate the posterior. Namely, we introduce a fully factorized distribution (mean-field) as a variational posterior, and we assume it is Gaussian, the same form as the predicted prior distribution in Eq.(3):

$$q(Z|\boldsymbol{\mu}, \boldsymbol{\delta}) = \prod_{i=1}^{N}\prod_{j=1}^{M} \mathcal{N}(z_{i,j}|\mu_{i,j}, \delta_{i,j}^2) \tag{5}$$

Then the inference becomes an optimisation problem for minimising the Kullback–Leibler (KL) divergence between the exact Eq.(4) and variational posteriors Eq.(5),

$$\begin{aligned} \mathtt{KL}(q(Z|\boldsymbol{\mu}, \boldsymbol{\delta})||p(Z|S, \mathrm{A}, \mathrm{B}, \boldsymbol{\sigma})) &= \mathbb{E}[\log q(Z|\boldsymbol{\mu}, \boldsymbol{\delta})] - \mathbb{E}[\log p(Z|S, \mathrm{A}, \mathrm{B}, \boldsymbol{\sigma})] \\ &= \mathbb{E}[\log q(Z|\boldsymbol{\mu}, \boldsymbol{\delta})] - \mathbb{E}[\log p(Z|\mathrm{A}, \mathrm{B}, \boldsymbol{\sigma})] - \mathbb{E}[\log p(S|Z)] + \log p(S). \end{aligned} \tag{6}$$

As the $\log p(S)$ is a constant term, minimizing the KL divergence is equivalent to maximizing the evidence lower bounder (ELBO)

$$\begin{aligned} \mathtt{ELBO}(q) &= -\mathbb{E}[\log q(Z|\boldsymbol{\mu}, \boldsymbol{\delta})] + \mathbb{E}[\log p(Z|\mathrm{A}, \mathrm{B}, \boldsymbol{\sigma})] + \mathbb{E}[\log p(S|Z)] \\ &= -\mathtt{KL}(q(Z|\boldsymbol{\mu}, \boldsymbol{\delta})||p(Z|\mathrm{A}, \mathrm{B}, \boldsymbol{\sigma})) + \mathbb{E}[\log p(S|Z)], \end{aligned} \tag{7}$$

where $\mathbb{E}[\cdot]$ denotes expectation over variational distribution $q(Z|\boldsymbol{\mu}, \boldsymbol{\delta})$ as a shortcut. The first part in ELBO is the KL divergence between the posterior and prior distribution on Z, which could be calculated analytically. The second term $\mathbb{E}[\log p(S|Z)]$ in ELBO (Eq. (7)) is difficulty to calculate due to the intrinsic mixture of two isoforms in the base likelihood Eq (1). Therefore, a cheap Monte Carlo exception [14] is introduced by sampling R samples on Z following its posterior distribution $q(Z)$:

$$\mathbb{E}_{q(Z)}[\log p(S|Z)] = \frac{1}{R}\sum_{r=1}^{R} \log p(S|Z^{(r)}) = \frac{1}{R}\sum_{r=1}^{R}\sum_{i=1}^{N}\sum_{j=1}^{M} \log p(\boldsymbol{s}_{i,j}|z_{i,j}^{(r)}). \tag{8}$$

In practice, $R = 3$ samples are sufficient to give good approximate and are used by default. Given the expression of ELBO, we could use a (stochastic) gradient descent algorithm, e.g., Adam by default [13], to achieve the maximum of ELBO. Here, we use TensorFlow platform to obtain an automated derivation of the gradient. Also, the re-parametrization trick [14] for gradient is fully supported for Gaussian distribution in TensorFlow.

4.4 Detecting differential splicing events

BRIE2 allows to detect genes (i.e., splicing events) that are significantly associated with one or multiple cell level covariates. This is equivalently to select Model 1 ($\mathcal{M}_1$) with none-zero coefficient versus Model 0 ($\mathcal{M}_0$) with zero coefficient for given cell feature(s) on a per gene basis. Therefore, BRIE2 will be run twice for both $\mathcal{M}_1$ with all provided cell features and $\mathcal{M}_0$ with leaving the candidate feature(s) out.

Then we compare the relevant evidence lower bounds $\texttt{EBLO}_1$ and $\texttt{EBLO}_0$, and obtain an $\texttt{ELBO\_gain} = \texttt{ELBO}_1 - \texttt{ELBO}_0$, which approximates the empirical Bayes factor. As the weights of cell features are fitted as a point estimate by maximizing the ELBO between the exact and variational posteriors, the `ELBO_gain` also mimics the log likelihood ratio in hypothesis test way, hence in practice one can transform the `ELBO_gain` to a p value through chi-square distribution with the degree of freedom equal to the number of testing features.

When testing events with alternative splicing associated with multiple sclerosis, we used the mouse strain, EAE state and intercept as covariates in $\mathcal{M}_1$ and left EAE state out in $\mathcal{M}_0$. When testing genes with spliced ratio associated cell type, each time we include proportion of detected genes, intercept, and one of 14 cell types as covariates in $\mathcal{M}_1$ and left the cell type out in $\mathcal{M}_0$. This tests have been repeated 14 times for all cell types.

4.5 Simulations

Simulations were performed to evaluate the quantification of ψ, feature coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}$ (Supplementary Fig. S3), and detection of genes significantly associated with cell level features (Supplementary Fig. S6). In both situations, we used an experimental data set with 130 cells and 2,248 splicing events from [21] as seed data. Here, we kept the same total read count of each event and cell as the seed data, and generate the isoform-specific count vectors with three read categories via a multinomial distribution parametrised by pre-defined ψ and fixed effective lengths L as in Eq.(1).

The core simulation is to sample $\psi_{i,j}$ from a logit-normal distribution $\texttt{LN}(\mu_{i,j}, \sigma^2)$, where $\mu_{i,j}$ is determined by cell features with noises and σ is set to 3 by default. In Supplementary Fig. S3, we took the mean $\mu_{i,j}$ as a product of five principle components calculated from gene expression and their according coefficients that are estimated from the seed data. In Supplementary Fig. S6, we independently sampled all $\mu_{i,j}$, and a cell feature vector $\boldsymbol{x} = \{x_1, ...x_{130}\}$ across 130 cells from a Gaussian distribution $\mathcal{N}(0, 3^2)$. We then randomly picked 400 genes as significant genes by replacing the mean vector $\boldsymbol{\mu}_j$ by this cell features $\boldsymbol{x}$. Here, we varied the σ among 1, 3 and 5 to mimic different levels of correlation between cell feature and splicing ratio ψ for systematically evaluating BRIE2's performance in detecting differential splicing events.

4.6 Data processing and gene filtering

For benchmarking BRIE2, we used 130 mouse embryonic cells at day 6.5 (80 cells) and day 7.75 (50 cells) that were generated by [21] with SMART-seq2 protocol [16]. This data set has also been used as an illustration data set in BRIE1 [10]. Here, we used HISAT v2.2.0 [12] to align the reads to mouse genome GRCm38.p6, combined with ERCC92 spike RNAs. Then *brie-count*

command line in BRIE v2.0.3 with all default parameters was used to count the reads aligned to each of the 8,253 alternative splicing events, which was extracted from GENCODE vM24 by using *briekit* at lenient thresholds.

The same processing except removing ERCC92 reference was applied to another SMART-seq2 data set on 2,208 mouse cells in the topic of multiple sclerosis [6], where BRIE2 was used to detect differential alternative splicing between disease and control cells. Here and in general, where detecting differential splicing and only cell level features are applicable, we filtered out clearly less informative genes. By default in *brie-quant*, we filter out events with 1) less than 50 total reads or 10 unique reads across all cells, or 2) less than 30 cells with unique reads, or 3) the fraction of unique reads on minor isoform less than 0.001.

For RNA velocity analysis, a data set on dentate gyrus development was used, which was generated by [8] with droplet protocol with 10x Genomics platform. The cell type annotation, UMAP visualization coordinates, and processed count matrices for both spliced and unspliced RNAs across 2,930 cells and 13,913 genes were downloaded from the tutorial in scVelo [4]. Only the top 3,000 highly variable genes with minimum 30 shared counts were used as suggested by scVelo. For detecting differential momentum genes, we only kept genes that were detected with at least one read in >15% of the cells. ScVelo v0.2.1 downloaded from PyPI is in use.

Additionally, an scNT-seq data set on excitatory neurons were obtained from original paper [17]. This processed data set has 3,066 quality controlled cells and 44,021 genes. It also has UMAP visualization coordinates, time annotation and layers of spliced, unspliced, and new RNAs. Therefore, no any pre-processing is needed on this data set. The RNA velocity inference by Dynamo (v0.95.2.dev142+9c30240) is based the same scripts provided on the original paper [17]. When running scVelo dynamical model, we select genes with at least 30 shared counts and either top 2,000 highly variable genes (Fig. 3b, Supplementary Fig. S13a, S14) or top 8,000 highly variable genes though with only 4880 genes pass the requirement (Supplementary Fig. S13b).

4.7 Code availability

BRIE2 is an open-source Python package available at `https://pypi.org/project/brie/`. All the analysis notebooks and the processed data sets can be found at `https://github.com/huangyh09/brie-tutorials`.

References

[1] Nadim Aizarani, Antonio Saviano, Laurent Mailly, Sarah Durand, Josip S Herman, Patrick Pessaux, Thomas F Baumert, Dominic Grün, et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature*, 572(7768):199–204, 2019.

[2] Ricard Argelaguet, Stephen J Clark, Hisham Mohammed, L Carine Stapel, Christel Krueger, Chantriolnt-Andreas Kapourani, Ivan Imaz-Rosshandler, Tim Lohoff, Yunlong Xiang, Courtney W Hanna, et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, 576(7787):487–491, 2019.

[3] Nico Battich, Joep Beumer, Buys de Barbanson, Lenno Krenning, Chloé S Baron, Marvin E Tanenbaum, Hans Clevers, and Alexander van Oudenaarden. Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. *Science*, 367(6482):1151–1156, 2020.

[4] Volker Bergen, Marius Lange, Stefan Peidli, F Alexander Wolf, and Fabian J Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, pages 1–7, 2020.

[5] Florian Erhard, Marisa AP Baptista, Tobias Krammer, Thomas Hennig, Marius Lange, Panagiota Arampatzi, Christopher S Jürges, Fabian J Theis, Antoine-Emmanuel Saliba, and Lars Dölken. scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature*, 571(7765):419–423, 2019.

[6] Ana Mendanha Falcão, David van Bruggen, Sueli Marques, Mandy Meijer, Sarah Jäkel, Eneritz Agirre, Elisa M Floriddia, Darya P Vanichkina, Anna Williams, André Ortlieb Guerreiro-Cacais, et al. Disease-specific oligodendrocyte lineage cells arise in multiple sclerosis. *Nature medicine*, 24(12):1837–1844, 2018.

[7] Gert-Jan Hendriks, Lisa A Jung, Anton JM Larsson, Michael Lidschreiber, Oscar Andersson Forsman, Katja Lidschreiber, Patrick Cramer, and Rickard Sandberg. NASC-seq monitors RNA synthesis in single cells. *Nature communications*, 10(1):1–9, 2019.

[8] Hannah Hochgerner, Amit Zeisel, Peter Lönnerberg, and Sten Linnarsson. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nature neuroscience*, 21(2):290–299, 2018.

[9] Yuanhua Huang and Guido Sanguinetti. BRIE: transcriptome-wide splicing quantification in single cells. *Genome biology*, 18(1):123, 2017.

[10] Yuanhua Huang and Guido Sanguinetti. Using BRIE to Detect and Analyze Splicing Isoforms in scRNA-Seq Data. In *Computational Methods for Single-Cell Data Analysis*, pages 175–185. Springer, 2019.

[11] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89, 2018.

[12] Daehwan Kim, Ben Langmead, and Steven L Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4):357–360, 2015.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR 2015*, 2015.

[14] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR 2014*, 2014.

[15] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018.

[16] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols*, 9(1):171–181, 2014.

[17] Qi Qiu, Peng Hu, Xiaojie Qiu, Kiya W Govek, Pablo G Cámara, and Hao Wu. Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nature Methods*, pages 1–11, 2020.

[18] Xiaojie Qiu, Yan Zhang, Dian Yang, Shayan Hosseinzadeh, Li Wang, Ruoshi Yuan, Song Xu, Yian Ma, Joseph Replogle, Spyros Darmanis, et al. Mapping vector field of single cells. *Biorxiv*, page 696724, 2019.

[19] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[20] David Schafflick, Chenling A Xu, Maike Hartlehnert, Michael Cole, Andreas Schulte-Mecklenbeck, Tobias Lautwein, Jolien Wolbert, Michael Heming, Sven G Meuth, Tanja Kuhlmann, et al. Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nature communications*, 11(1):1–14, 2020.

[21] 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):289–293, 2016.

[22] Yan Song, Olga B Botvinnik, Michael T Lovci, Boyko Kakaradov, Patrick Liu, Jia L Xu, and Gene W Yeo. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Molecular cell*, 67(1):148–161, 2017.