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

Single-cell RNA-seq (scRNA-seq), a recent technology that combines efficient RNA amplification with high throughput sequencing, has revolutionised our understanding of transcriptome variability among cell population. It has profound implications both fundamental and translational, for example, in dissecting tumour heterogeneity (1). However, intrinsic limitations of scRNA-seq, stemming from the minute quantity of initial RNA retrieved from single cells, have prevented its application to dissect variability in RNA splicing, as methods from bulk RNA-seq cannot handle the low coverage and high drop-out rates of scRNA-seq.

Here we present BRIE (Bayesian Regression for Isoform Estimation), a Bayesian hierarchical model which pools genetic and expression information to perform robust splicing quantification from scRNA-seq data. BRIE consists of two modules: a likelihood part (bottom part of Fig 1) which uses the scRNA-seq data (aligned reads) within a mixture model approach to isoform estimation (as used in standard methods such as MISO (2) and Cufflinks (3)). The likelihood module is coupled with an informative prior distribution in the form of a Bayesian regression model, where the prior probability of inclusion ratios is regressed against sequence-derived features (upper part of Fig 1). This exploits the fact that splicing events are highly predictable from sequence (4) to help quantification when data is lacking. Importantly, the prior distribution can be learned across multiple single cells, thus transferring information across the whole experimental design.

BRIE model has been implemented as a standard Python package, which is freely available at `http://github.com/huangyh09/brie`. The full manuscript is available on bioRxiv (5).

This architecture effectively enables BRIE to simultaneously trade-off two tasks: in the absence of data (drop-out genes), the informative prior provides a way of imputing missing data, while for highly covered genes the likelihood term dominates, returning a mixture-model quantification. For intermediate levels of coverage, BRIE uses Bayess theorem to trade off imputation and quantification.

We validate BRIE on both simulated and real scRNA-seq data sets, showing that BRIE yields reproducible estimates of exon inclusion ratios in single cells. With the simulated RNA-seq data, we first assessed the performance of BRIE with different coverage levels, and see that the use of an informative prior in BRIE can bring very substantial performance improvements at low coverage, with a gain of almost 20% in correlation between estimates and ground truth in RPK=25. We also mimicked the drop-out events with the parameters estimated from real single cell experiments. Again with the simulated data, we see that if drop-out happens (namely no reads sequenced for expressed genes), BRIE can produce a good imputation of the isoform usage simply by taking the mean of the informative prior learned from sequence features (Pearsons R: 0.6∼0.7).

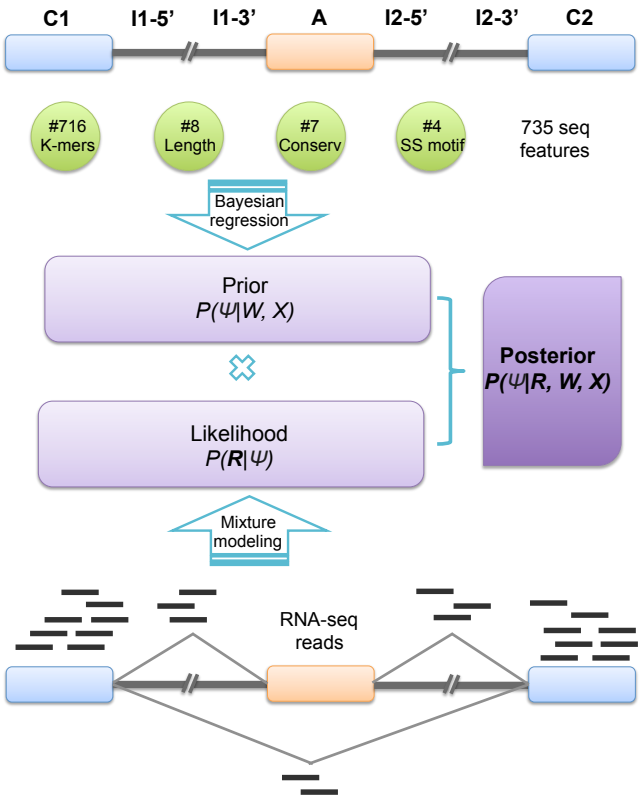

Figure 1. A cartoon of the BRIE method for isoform estimation. BRIE combines a likelihood computed from RNA-seq data (bottom part) and an informative prior distribution learned from 735 sequence-derived features (top).

To further assess BRIE's performance on real scRNA-seq data, we used 96 scRNA-seq libraries from individual HCT116 human cells from the benchmark scRNA-seq study of Wu et al (6) (see Methods for details). Importantly, a bulk RNA-seq data set in the same conditions was also obtained from one million cells. Figure 2 shows that BRIE clearly outperforms all other methods by a large margin, both in terms of correlation between estimates from different single cells (Fig 2f), and in terms of correlations between estimates from individual single-cells and bulk (Fig 2c).

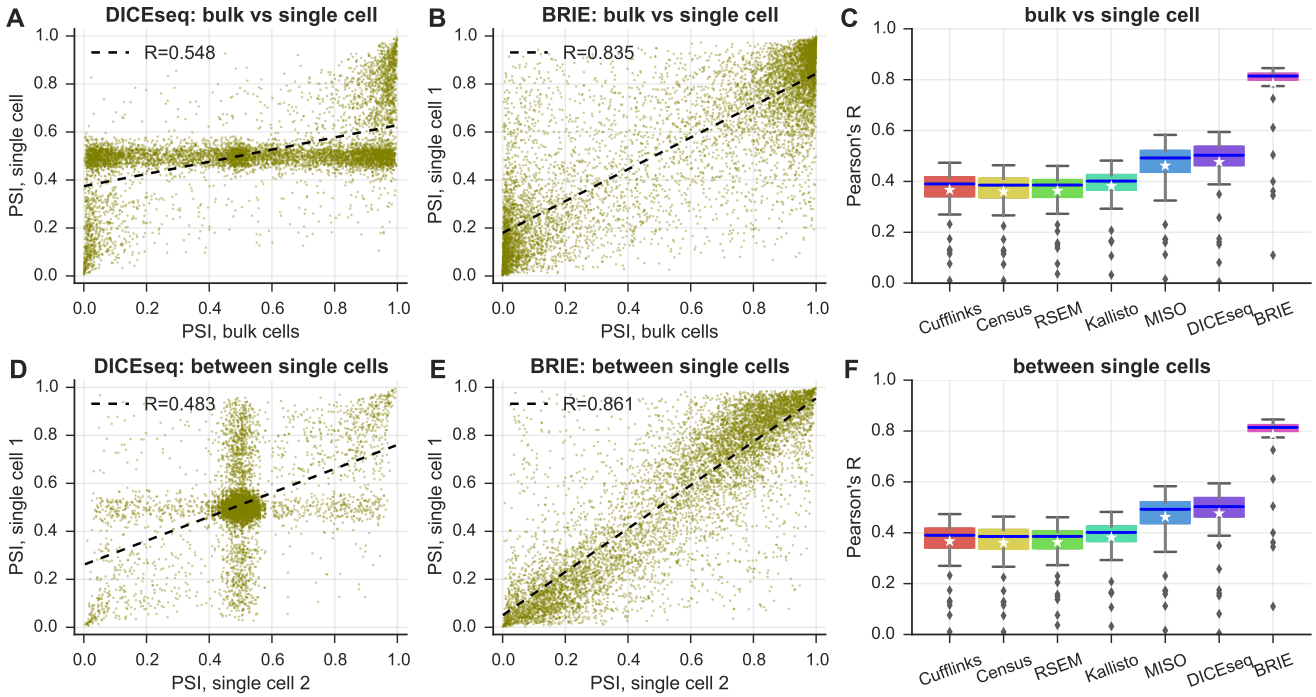

Figure 2. BRIE improves splicing estimates by using sequence features. (A-C) Pearson's correlation between between bulk and single cells on exon inclusion ratio ψ in HCT116 cells. Scatter plot of ψ estimates by DICEseq (A), or estimated by BRIE (B). Box-plot for all methods (C) in 96 cells. (D-F) Pearson's correlation between single cell pairs. Scatter plot of ψ estimates by DICEseq (D), or estimated by BRIE (E). Box-plot for all methods (F) in 4,608 cell pairs.

Example scatter plots for both comparisons are given in Fig 2e and 2b, clearly showing very consistent predictions. Notably, the performance of other methods was strongly degraded by the inability to handle the large drop-out rates (see Fig 2a and 2d for DICE-seq, where many estimates of splicing are centred around the uninformative prior value of 0.5). 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.

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 original preprint (5)).

To estimate a background level of differential splicing between identical cells, we considered again the 20 single cell HCT116 libraries from Wu et al (6), and compared all possible pairs of cells. Figure 3a 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 (7). 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 (8), 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 3b. 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. Figure 3c shows the example of DNMT3B, a regulator of DNA methylation maintenance, which is known to undergo functionally relevant alternative splicing (9). 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.

Overall, our results demonstrate 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 splicing. As splicing is implicated in a number of disease and developmental processes, BRIE can considerably enhance the usefulness of scRNA-seq technologies in both fundamental and translational biology.

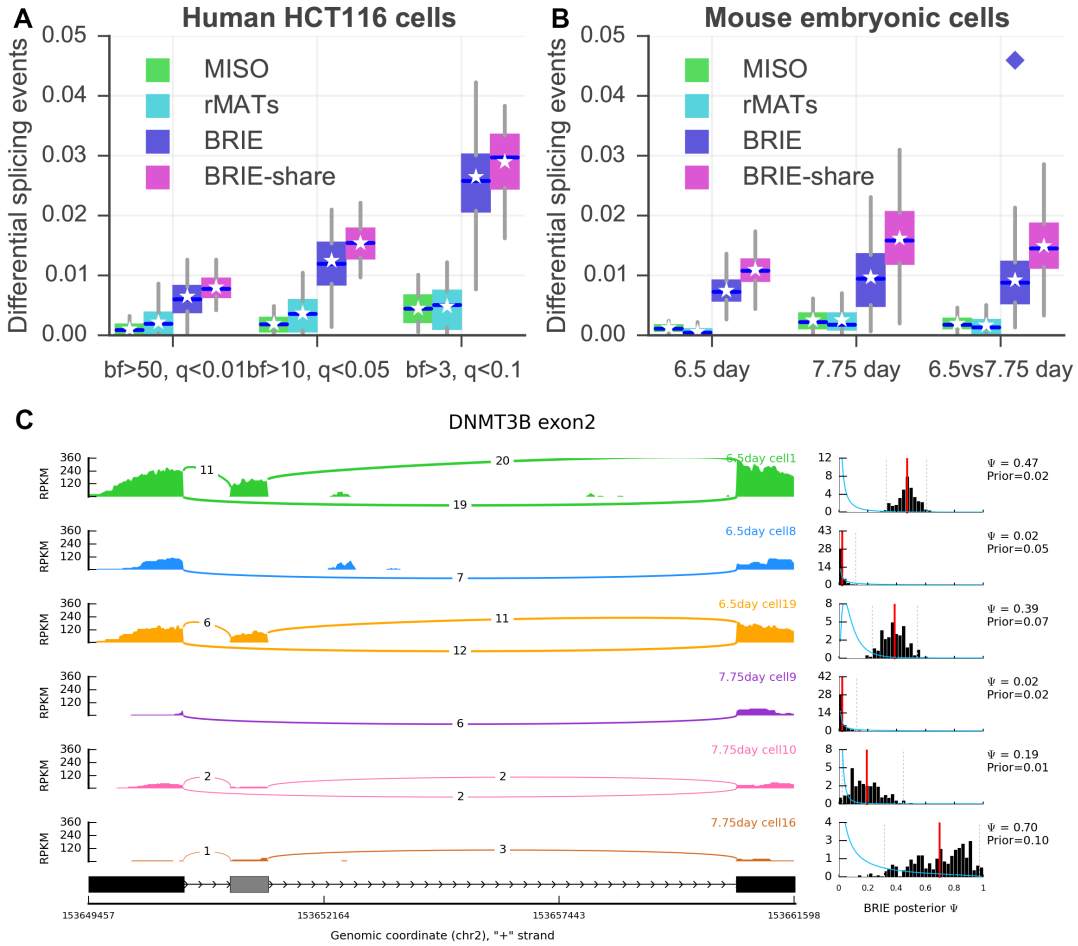

Figure 3. Detection of differential splicing between cells. (A) Percentage of differential splicing events between human HCT116 cells, detected by MISO, rMATS, BRIE and its mode with shared weights (i.e., BRIE.share) with different thresholds. MISO and BRIE use Bayes factor (bf) and rMATS uses false discovery rate (q value). (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. Diamond indicates pooling reads of 20 cells in each group. (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.

Current work to extend this study of splicing in single cells includes two directions. First, we are trying to apply BRIE into a wider range of real scRNA-seq data sets to show the power of splicing events as markers for identifying cell types or cell states. Second, we are interested in adding a Gaussian process into the Bayesian hierarchical model to account for the pseudo-time trajectory of the splicing among cell population.

REFERENCES

1. Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., et al. (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science,* **344**(6190), 1396–1401.
2. Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods,* **7**(12), 1009–1015.
3. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology,* **28**(5), 511–515.
4. Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J., and Frey, B. J. (2010) Deciphering the splicing code. *Nature,* **465**(7294), 53–59.
5. Huang, Y. and Sanguinetti, G. (2017) Transcriptome-wide splicing quantification in single cells. *bioRxiv,* p. 098517.
6. Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., Mburu, F. M., Mantalas, G. L., Sim, S., Clarke, M. F., et al. (2014) Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods,* **11**(1), 41–46.
7. Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., and Xing, Y. (2014) 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.
8. Scialdone, A., Tanaka, Y., Jawaid, W., Moignard, V., Wilson, N. K., Macaulay, I. C., Marioni, J. C., and Göttgens, B. (2016) Resolving early mesoderm diversification through single-cell expression profiling. *Nature,* **535**(7611), 284–293.
9. Duymich, C. E., Charlet, J., Yang, X., Jones, P. A., and Liang, G. (2016) DNMT3B isoforms without catalytic activity stimulate gene body methylation as accessory proteins in somatic cells. *Nature Communications,* **7**.