Transcriptional variability analysis of CD4$^+$ T cells using BASiCS

Alan O'Callaghan Nils Eling John C. Marioni Catalina A. Vallejos

Contents

1 1. Introduction

To date, three versions of the *BASiCS* model have been proposed (Catalina A. Vallejos, Marioni, and Richardson 2015a) (Catalina A. Vallejos, Richardson, and Marioni 2016) (Eling et al. 2018). They differ in how inference is performed (e.g. using different priors) and the type of downstream analysis that is enabled by the model.

- **Vallejos et al (2015)** (Catalina A. Vallejos, Marioni, and Richardson 2015b): the original model uses information from extrinsic spike-in molecules (e.g. those introduced by (External RNA Controls Consortium 2005)) as *control features* to quantify technical noise. This enables the estimation of two sets of cell-specific normalisation parameters (s_j and ϕ_j) capturing technical (e.g. amplification biases) and biological (e.g. mRNA content) systematic differences across cells (Catalina A. Vallejos et al. 2017). A probabilistic decision rule (based on δ_i) was proposed to identify *highly variable genes* (HVGs) that capture the major sources of heterogeneity within the analysed cells (Brennecke et al. 2013). HVG detection is often used to perform feature selection, choosing the input set of genes for subsequent analyses. A similar rule was developed to highlight *lowly variable genes* (LVGs) that exhibit stable expression across the population of cells. These may relate to essential cellular functions and can assist the development of new data normalisation or integration strategies (Lin et al. 2019).
- **Vallejos et al (2016)** (Catalina A. Vallejos, Richardson, and Marioni 2016): the model was extended to enable **differential expression** analyses between two pre-specified groups of cells (e.g. different experimental conditions or cell types). This is achieved by comparing the posterior distribution of gene-specific parameters (μ_i and δ_i). While several differential expression tools were previously proposed for scRNA-seq data (e.g. (Kharchenko, Silberstein, and Scadden 2014; Finak et al. 2015)), some evidence suggests that these do not generally outperform popular bulk RNA-seq tools (Soneson and Robinson 2018). Moreover, most of these methods are only designed to uncover changes in overall expression, ignoring the more complex patterns that can arise at the single cell level (Lähnemann et al. 2020). Instead, *BASiCS* embraces the high granularity of scRNA-seq data, uncovering changes in cell-to-cell transcriptional variability. As noted by Catalina A. Vallejos, Richardson, and Marioni (2016), the inverse relationship that is observed between mean expression and over-dispersion derived from (bulk and) scRNAseq can affect the interpretation of such analyses. In particular, genes that are differentially expressed between two groups of cells are likely to exhibit changes in both mean expression and variability, due to the inverse relationship between these two quantities. Thus, comparisons of variability between populations must be restricted to genes that do not exhibit changes in mean expression.
- **Eling et al (2018)** (Eling et al. 2018): the model was extended to account for the strong relationship that is typically observed between gene-specific mean expression and over-dispersion estimates. Eling

et al. (Eling et al. 2018) introduced a *joint prior* specification for these parameters. This joint prior assumes that genes with similar mean expression (μ_i) have similar over-dispersion parameters δ_i. Effectively, this shrinks over-dispersion estimates towards a global trend that captures the relationship between mean and over-dispersion (Figure XX). This improves posterior inference for over-dispersion parameters when the data is less informative (e.g. small sample size, lowly expressed genes) (Eling et al. 2018). This information-sharing approach is conceptually similar to that performed by Love, Huber, and Anders (2014) and others, where sparse data is pooled to obtain more reliable estimates. The global trend is then used to derive gene-specific *residual over-dispersion* parameters ϵ_i that are not confounded by mean expression. Similar to the DM values implemented in *scran*, these are defined as deviations with respect to the overall trend (Figure XX). *BASiCS* also provides a probabilistic decision rule to perform differential expression analyses between two pre-specified groups of cells (Catalina A. Vallejos, Richardson, and Marioni 2016; Eling et al. 2018). Furthermore, the model was extended using a horizontal integration framework to allow its use in the absence of spike-in genes. This is useful for droplet-based scRNAseq protocols, given that it is not possible to ensure that each droplet contains a specified quantity of spike-in molecules. In this horizontal integration framework, technical variation is quantified using replication (Carroll 2005). In the absence of true technical replicates, we assume that population-level characteristics of the cells are replicated using appropriate experimental design. This requires that cells from the same population have been randomly allocated to different batches. Given appropriate experimental design, *BASiCS* assumes that biological effects are shared across batches, while technical variation leads to spurious differences between cells in different batches. It is this version of the model that we focus on here, and that we recommend for most users. Previous versions of the model are available within the package, but are primarily useful for reproducibility purposes or for analysing datasets that contain spike-in genes.

1.0.1 Testing for changes in mean expression and over-dispersion

Differential mean and differential over-dispersion testing is done by computing the tail posterior propabilities of the difference in mean expression or over-dispersion between two conditions (A and B) being larger than an evidence threshold τ_0 or ω_0 Catalina A. Vallejos, Richardson, and Marioni (2016):

$$\mathrm{P}\left(\log\left(\frac{\mu_i^{(A)}}{\mu_i^{(B)}}\right) > \tau_0 | \mathrm{Data}\right) > \alpha_m$$

$$\mathrm{P}\left(\log\left(\frac{\delta_i^{(A)}}{\delta_i^{(B)}}\right) > \omega_0 | \mathrm{Data}\right) > \alpha_d$$

If the tail posterior probability is larger than a given propability threshold α_m or α_d, the gene is considered to be differentially expressed or differentially over-dispersed (Catalina A. Vallejos, Richardson, and Marioni 2016). The evidence threshold is usually fixed *a priori* and the probability threshold is defined to control the expected false discovery rate (EFDR) to (e.g. 10%) (Newton 2004, Vallejos2016).

As described by Catalina A. Vallejos, Richardson, and Marioni (2016), estimates of the over-dispersion parameters δ_i are negatively correlated to mean expression μ_i. This indicates that in homogeneous populations of cells, highly expressed genes tend to be less noisy than lowly expressed genes. Differential over-dispersion testing is therefore confounded by mean expression changes. When assessing changes in over-dispersion δ_i, only genes with no changes in mean expression are considered (see Catalina A. Vallejos, Richardson, and Marioni (2016)).

1.0.2 Correcting the mean-variability confounding effect

Eling et al. (2018) extended BASiCS to account for the confounding effect between mean expression and expression variability. For this purpose, we capture the relationship between mean and over-dispersion

parameters by introducing the following joint prior distribution for $(\mu_i, \delta_i)'$:

$$\mu_i \sim \text{log-Normal}\left(0, s_\mu^2\right), \qquad \delta_i|\mu_i \sim \text{log-t}_\eta\left(\text{f}(\mu_i), \sigma^2\right).$$

The latter is equivalent to the non-linear regression model:

$$\log(\delta_i) = \text{f}(\mu_i) + \epsilon_i, \qquad \epsilon_i \sim \text{t}_\eta(0, \sigma^2),$$

where $\text{f}(\mu_i)$ represents the over-dispersion (on the log-scale) that is predicted by the global trend (across all genes) for a given mean expression μ_i. Therefore, ϵ_i can be interpreted as a gene-specific *residual over-dispersion* parameter. Positive values for ϵ_i indicate more variation than expected for genes with similar expression level. Similarly, negative values of ϵ_i suggest less variation than expected (Eling et al. 2018).

In line with the probabilistic approach described above, we identified statistically significant differences in residual over-dispersion for those genes where the tail posterior probability of observing a large difference between ϵ_i^A and ϵ_i^B exceeds a certain threshold:

$$\text{P}(\mid \epsilon_i^A - \epsilon_i^B \mid > \psi_0 \mid \text{Data}) > \alpha_R,$$

where $\psi_0 > 0$ defines the minimum tolerance threshold. As a default choice, we assume $\psi_0 = \log_2(1.5)/\log_2(e) \approx 0.41$, which translates into a 50% increase in over-dispersion. The posterior probability threshold α_R is chosen to control the EFDR (e.g. 10%) (Newton 2004). To support interpretability of the results, we exclude genes that are not expressed in at least 2 cells per condition from differential variability testing.

Bochkina, N., and S. Richardson. 2007. "Tail Posterior Probability for Inference in Pairwise and Multiclass Gene Expression Data." *Biometrics* 63 (4): 1117–25. https://doi.org/10.1111/j.1541-0420.2007.00807.x.

Brennecke, Philip, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, et al. 2013. "Accounting for technical noise in single-cell RNA-seq experiments." https://doi.org/10.1038/nmeth.2645.

Carroll, Raymond J. 2005. "Measurement Error in Epidemiologic Studies," 38.

Eling, Nils, Arianne C. Richard, Sylvia Richardson, John C. Marioni, and Catalina A. Vallejos. 2018. "Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data." *Cell Systems* 7 (3): 1–11. https://doi.org/10.1016/j.cels.2018.06.011.

External RNA Controls Consortium. 2005. "Proposed methods for testing and selecting the ERCC external RNA controls." *BMC Genomics* 6 (June 2004): 150. https://doi.org/10.1186/1471-2164-6-150.

Finak, Greg, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, et al. 2015. "MAST: A Flexible Statistical Framework for Assessing Transcriptional Changes and Characterizing Heterogeneity in Single-Cell RNA Sequencing Data." *Genome Biology* 16 (1): 278. https://doi.org/10.1186/s13059-015-0844-5.

Kharchenko, Peter V, Lev Silberstein, and David T Scadden. 2014. "Bayesian approach to single-cell differential expression analysis." *Nature Methods* 11 (7): 740–42. https://doi.org/10.1038/nmeth.2967.

Lähnemann, David, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, et al. 2020. "Eleven Grand Challenges in Single-Cell Data Science." *Genome Biology* 21 (1): 31. https://doi.org/10.1186/s13059-020-1926-6.

Lin, Yingxin, Shila Ghazanfar, Dario Strbenac, Andy Wang, Ellis Patrick, David M Lin, Terence Speed, Jean Y H Yang, and Pengyi Yang. 2019. "Evaluating Stably Expressed Genes in Single Cells." *GigaScience* 8 (9): giz106. https://doi.org/10.1093/gigascience/giz106.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Newton, M. A. 2004. "Detecting Differential Gene Expression with a Semiparametric Hierarchical Mixture Method." *Biostatistics* 5 (2): 155–76. https://doi.org/10.1093/biostatistics/5.2.155.

Soneson, Charlotte, and Mark D Robinson. 2018. "Bias, Robustness and Scalability in Single-Cell Differential Expression Analysis." *Nature Methods* 15 (4): 255–61. https://doi.org/10.1038/nmeth.4612.

Vallejos, Catalina A., John C. Marioni, and Sylvia Richardson. 2015b. "BASiCS: Bayesian analysis of single-cell sequencing data." *PLOS Computational Biology* 11: e1004333. https://doi.org/10.1371/journal.pcbi.1004333.

———. 2015a. "BASiCS: Bayesian analysis of single-cell sequencing data." *PLOS Computational Biology* 11 (6): e1004333. https://doi.org/10.1371/journal.pcbi.1004333.

Vallejos, Catalina A, Sylvia Richardson, and John C Marioni. 2016. "Beyond comparisons of means: understanding changes in gene expression at the single-cell level." *Genome Biology* 17 (70). https://doi.org/10.1101/035949.

Vallejos, Catalina A, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. 2017. "Normalizing single-cell RNA sequencing data: challenges and opportunities." *Nature Methods* 14 (6): 565–71. https://doi.org/10.1038/nmeth.4292.