BASiCS workflow: a step-by-step analysis of expression variability using single cell RNA sequencing data

Alan O'Callaghan[*1], **Nils Eling**[2], **John C. Marioni**[3,4], **and Catalina A. Vallejos**[†1,5]

[1]**MRC Human Genetics Unit, Institute of Genetics & Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK**
[2]**Department of Quantitative Biomedicine, University of Zurich, Winterthurerstrasse 190, CH-8057, Zurich, Switzerland**
[3]**European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK**
[4]**Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, CB2 0RE, UK**
[5]**The Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, UK**

Abstract Cell-to-cell gene expression variability is an inherent feature of complex biological systems, such as immunity and development. Single-cell RNA sequencing is a powerful tool to quantify this heterogeneity, but it is prone to strong technical noise. In this article, we describe a step-by-step computational workflow that uses the BASiCS Bioconductor package to robustly quantify expression variability within and between known groups of cells (such as experimental conditions or cell types). BASiCS uses an integrated framework for data normalisation, technical noise quantification and downstream analyses, whilst propagating statistical uncertainty across these steps. Within a single seemingly homogeneous cell population, BASiCS can identify highly variable genes that exhibit strong heterogeneity as well as lowly variable genes with stable expression. BASiCS also uses a probabilistic decision rule to identify changes in expression variability between cell populations, whilst avoiding confounding effects related to differences in technical noise or in overall abundance. Using a publicly available dataset, we guide users through a complete pipeline that includes preliminary steps for quality control, as well as data exploration using the scater and scran Bioconductor packages. The workflow is accompanied by a Docker image that ensures the reproducibility of our results.

Keywords

Single-cell RNA sequencing, expression variability, transcriptional noise, differential expression testing

[*]a.b.o'callaghan@sms.ed.ac.uk
[†]catalina.vallejos@igmm.ed.ac.uk

Introduction

Parameter estimation is implemented in the `BASiCS_MCMC` function using an adaptive Metropolis within Gibbs algorithm [1]. The primary inputs for `BASiCS_MCMC` correspond to:

- `Data`: a `SingleCellExperiment` object created as described in the previous sections.
- `N`: the total number of MCMC iterations.
- `Thin`: thining period for output storage (only the `Thin`-th MCMC draw is stored).
- `Burn`: the initial number of MCMC iterations to be discarded.
- `Regression`: if `TRUE` a join prior is assigned to μ_i and δ_i [2], and residual over-dispersion values ϵ_i are inferred. Alternatively, independent log-normal priors are assigned to μ_i and δ_i [3].
- `WithSpikes`: if `TRUE` information from spike-in molecules is used to aid data normalisation and to quantify technical noise.
- `PriorParam`: Defines the prior hyper-parameters to be used by *BASiCS*. We recommend to use the `BASiCS_PriorParam` function for this purpose. If `PriorMu = "EmpiricalBayes"` in `BASiCS_PriorParam`, μ_i's are assigned a log-normal prior with gene-specific location hyper-parameters defined via an empirical Bayes framework. Alternatively, if `PriorMu = "default"`, location hyper-parameter are set to be equal 0 for all genes.

References

[1] Gareth O. Roberts and Jeffrey S. Rosenthal. Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, January 2009. ISSN 1061-8600, 1537-2715. doi: 10.1198/jcgs.2009.06134.

[2] Nils Eling, Arianne C. Richard, Sylvia Richardson, John C. Marioni, and Catalina A. Vallejos. Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell Systems*, 7(3): 1–11, 2018. ISSN 24054712. doi: 10.1016/j.cels.2018.06.011. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471218302783.

[3] Catalina A Vallejos, Sylvia Richardson, and John C Marioni. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biology*, 17(70), 2016. doi: 10.1101/035949. URL http://biorxiv.org/content/early/2016/01/05/035949.abstract.