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

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

[1]**European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK**
[2]**Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, CB2 0RE, UK**
[3]**MRC Human Genetics Unit, Institute of Genetics & Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK**
[4]**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 which 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 two publicly available datasets, we guide users through a complete pipeline which includes preliminary steps for quality control as well as data exploration using the scater and scran Bioconductor packages. Data for the first case study was generated using the Fluidigm@ C1 system, in which extrinsic spike-in RNA molecules were added as a control. The second dataset was generated using a droplet-based system, for which spike-in RNA is not available. This analysis provides an example, in which differential variability testing reveals insights regarding a possible early cell fate commitment process. 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

[*]eling@ebi.ac.uk
[†]catalina.vallejos@igmm.ed.ac.uk

Reproducibility

The following R, Bioconductor and package version were used for this workflow:

R version: R Under development (unstable) (2020-01-28 r77738)

Bioconductor version: 3.11

Packages: BASiCS 1.99.1, scran 1.15.29, scater 1.15.32, coda 0.19.3, goseq 1.39.0

For the full list of packages used, please see the Session Info.

C1 Fluidigm data: Analysis of naive CD4$^+$ T cells

For the first case study, we will use scRNA-seq data of CD4$^+$ T cells, which were processed using the C1 Single-Cell Auto Prep System (Fluidigm®) using 10–17 μm integrated fluidic circuits (IFCs). Martinez-Jimenez *et al.* profiled naive and activated CD4$^+$ T cells from young and old animals across two mouse strains to test for changes in expression variability that occur during organismal ageing [1]. They extracted naive or effector memory CD4$^+$ T cells from spleens of young or old animals and filtered using either magnetic-activated cell sorting (MACS) or fluorescence activated cell sorting (FACS) (labelled as MACS-purified Naive, FACS-purified Naive or FACS-purified Effector Memory). For clarification, naive CD4$^+$ T cells are also referred to as 'unstimulated' CD4$^+$ T cells.

In addition to profiling naive CD4$^+$ T cells, the authors stimulated half of the naive cells for 3 hours using *in vitro* antibody stimulation (labelled as Active). They processed naive as well as activated CD4$^+$ T cells using the C1 Fluidigm® system to capture and lyse cells, and to reverse transcribe and amplify mRNA prior to sequencing. The authors isolated cells from B6 (C57BL/6J, *Mus musculus domesticus*) and CAST (*Mus musculus castaneus*) animals to profile the evolutionary conservation of transcriptional dynamics during ageing. Additionally, the authors added external spike-in RNA to aid in quantifying technical variability across all cells. They performed all experiments in replicates (also referred to as batches) to control for batch effects.

We will begin the workflow with obtaining the data before quality control, running the BASiCS model, and performing further downstream analysis.

Obtaining the data

The raw counts of the full dataset can be obtained from ArrayExpress under the accession number E-MTAB-4888. In this dataset, the column names contain the library identifier of the original experiment, while the row names of the matrix store gene names. The dataset contains reads mapped to ERCC spike-in genes [2], which `BASiCS` uses to estimate and remove technical noise.

References

[1] Celia P. Martinez-Jimenez, Nils Eling, Hung-Chang Chen, Catalina A Vallejos, Aleksandra A Kolodziejczyk, Frances Connor, Lovorka Stojic, Timothy F Rayner, Michael J T Stubbington, Sarah A Teichmann, Maike de la Roche, John C Marioni, and Duncan T Odom. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science*, 1436:1433–1436, 2017. doi: 10.1126/science.aah4115.

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