# 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

```
library(SingleCellExperiment)
library(scater)
library(scran)
library(BASiCS)
```

### QC and exploratory analysis

As an initial step, we remove lowly expressed genes whose expression was largely not detected through sequencing. More specifically, we exclude genes whose expression was detected in less than 2 cells and those for which their average read count is below 1.

These thresholds need to be set specifically for each dataset, and careful gene-specific quality metrics need to be closely examined as suggested by the *SimpleSingleCell* Bioconductor workflow [1].

## References

[1] Aaron T. L. Lun, Davis J. McCarthy, and John C. Marioni. A step-by-step workflow for basic analyses of single-cell RNA-seq data. *F1000Research*, 5(2122), 2016. ISSN 2046-1402. doi: 10.12688/f1000research.9501.1.