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

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

[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]Department of Quantitative Biomedicine, University of Zurich, Winterthurerstrasse 190, CH-8057, Zurich, Switzerland
[4]MRC Human Genetics Unit, Institute of Genetics & Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, 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 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

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

Introduction

Single-cell RNA-sequencing (scRNA-seq) enables the study of genome-wide transcriptional heterogeneity in cell populations that is not captured by bulk experiments [1, 2, 3]. On the broadest level, this heterogeneity can reflect the presence of distinct cell subtypes or states. Alternatively, it can be due to gradual changes along biological processes, such as development and differentiation. Several clustering and pseudotime inference methods have been developed to characterise these types of heterogeneity [4, 5]. However, there is a limited availability of computational tools tailored to study more subtle variability within seemingly homogeneous cell populations. This variability can reflect deterministic or stochastic events that regulate gene expression and, among others, has been reported to increase prior to cell fate decisions [6] as well as during ageing [7].

This article complements existing scRNA-seq workflows based on the Bioconductor ecosystem (e.g. [8, 9]), providing a detailed framework for transcriptional variability analyses. Firstly, we briefly discuss the sources of variability that arise in scRNA-seq data and the strategies that have been designed to control or attenuate technical noise in these assays. Subsequently, we describe a step-by-step workflow which uses *scater* [10] and *scran* [8] to perform quality control (QC) as well as initial exploratory analyses. To robustly quantify transcriptional variability we use *BASiCS* [11, 12, 13] — a Bayesian hierarchical framework that jointly performs data normalisation, technical noise quantification and downstream analyses, whilst propagating statistical uncertainty across these steps. Our analysis pipeline includes practical guidance to assess the convergence of the Markov Chain Monte Carlo (MCMC) algorithm that is used to infer model parameters as well as recommendations to interpret and post-process the model outputs. Finally, through a case study in the context of immune cells, we illustrate how *BASiCS* can be used to identify highly and lowly variable genes within a cell population, as well as to compare expression profiles between experimental conditions or cell types.

All source code used to generate the results presented in this article is available on Github. To ensure the reproducibility of this workflow, the analysis environment and all software dependencies are provided as a Docker image [14]. The image can be obtained from Docker Hub.

Sources of variability in scRNA-seq data

Stochastic variability within a seemingly homogeneous cell population — often referred to as transcriptional *noise* — can arise from intrinsic and extrinsic sources [15, 16]. Extrinsic noise refers to stochastic fluctuations induced by different dynamic cellular states (e.g. cell cycle, metabolism, intra/inter-cellular signalling) [17, 18, 19]. In contrast, intrinsic noise arises from stochastic effects on biochemical processes such as transcription and translation [15]. Intrinsic noise can be modulated by genetic and epigenetic modifications (such as mutations, histone modifications, CpG island length and nucleosome positioning) [20, 21, 22] and usually occurs at the gene level [15]. Cell-to-cell gene expression variability estimates derived from scRNA-seq data capture a combination of these effects, as well as deterministic regulatory mechanisms [16]. Moreover, these variability estimates can also be inflated by the technical noise that is typically observed in scRNA-seq data [23].

Different strategies have been incorporated into scRNA-seq protocols to control or attenuate technical noise. For example, external RNA spike-in molecules (such as the set introduced by the External RNA Controls Consortium, ERCC [24]) can be added to each cell's lysate in a (theoretically) known fixed quantity. Spike-ins can assist quality control steps [10], data normalisation [25] and can be used to infer technical noise [23]. Another strategy is to tag individual cDNA molecules using unique molecular identifiers (UMIs) before PCR amplification [26]. Reads that contain the same UMI can be collapsed into a single molecule count, attenuating technical variability associated to cell-to-cell differences in amplification and sequencing depth (these technical biases are not fully removed unless sequencing to saturation [25]). However, despite the benefits associated to the use of spike-ins and UMIs, these are not available for all scRNA-seq protocols [27].

Methods

This step-by-step scRNA-seq workflow is primarily based on the Bioconductor package ecosystem [28]. A graphical overview is provided in Figure 1 and its main components are described below.

Input data

```
library("SingleCellExperiment")
```

We use *SingleCellExperiment* to convert an input matrix of raw read-counts (molecule counts for UMI-based protocols) into a `SingleCellExperiment` object which can also store its associated metadata, such as gene- and cell-specific information. Moreover, when available, the same object can also store read-counts for spike-in molecules (see `altExp()`). A major advantage of using a `SingleCellExperiment` object as the input for scRNA-seq analyses is the interoperability across a large number of Bioconductor packages [28].

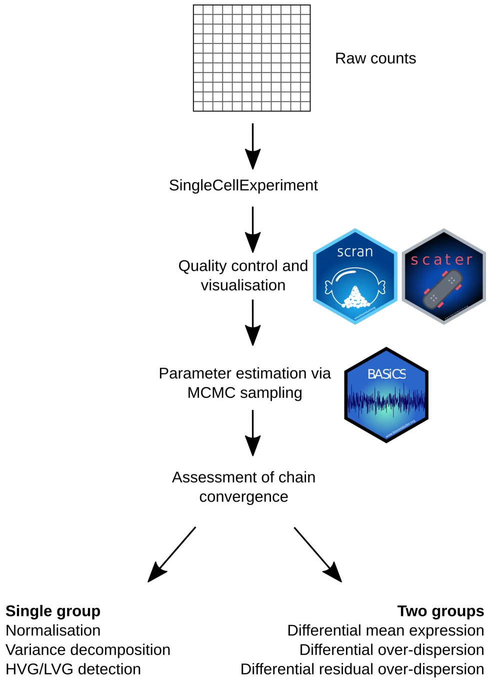
Raw counts
SingleCellExperiment
Quality control and visualisation
scran
scater
Parameter estimation via MCMC sampling
BASiCS
Assessment of chain convergence
Single group
Normalisation
Variance decomposition
HVG/LVG detection
Two groups
Differential mean expression
Differential over-dispersion
Differential residual over-dispersion

Figure 1. Graphical overview for the scRNA-seq analysis workflow described in this manuscript. Starting from a matrix of expression counts, we use the scater and scran Bioconductor packages to perform QC and initial exploratory analyses. To robustly quantify transcriptional heterogeneity within seemingly homogeneous cell populations, we apply the BASiCS Bioconductor package and illustrate how BASiCS can be used to analyse a single or multiple pre-specified groups of cells.

Quality control and exploratory analysis

```
library("scater")
library("scran")
```

An critical step in scRNA-seq analyses is QC, removing low quality samples that may distort downstream analyses. Among others, QC diagnostics can help to identify samples that contain broken cells, that are empty or that contain multiple cells [29]. Moreover, lowly expressed genes for which less reliable information is available are typically also removed. The OSCA online book provides an extensive overview on important aspects of how to perform QC of scRNA-seq data, including exploratory analyses [28].

Here, we use the *scater* package [10] to calculate QC metrics for each cell (e.g. total read-count) and gene (e.g. percentage of zeroes across all cells), respectively. Moreover, we use the visualisation tools implemented in *scater* to explore the input dataset and its associated QC diagnostic metrics. For further data exploration we use the *scran* package [8]. *scran* can perform *global scaling* normalisation, calculating cell-specific scaling factors that capture global differences in read-counts across cells (e.g. due to sequencing depth and PCR amplification) [30]. Moreover, *scran* enables exploratory analyses of transcriptional variability. For example, it can be used to infer an overall trend between mean expression and the squared coefficent of variation (CV^2) for each gene. To derive variability estimates that are not confounded by this overall trend, *scran* also defines gene-specific DM (distance to the mean) estimates as the distance between CV^2 and a rolling median along the range of mean expression values [31]. DM estimates enable exploratory analyses of cell-to-cell heterogeneity, but a measure of uncertainty is not readily available. As such, gene-specific downstream inference (e.g. differential variability testing) is precluded.

```
library("ggplot2")
```

References

[1] Oliver Stegle, Sarah a. Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, jan 2015. ISSN 1471-0056. doi: 10.1038/nrg3833. URL http://www.nature.com/doifinder/10.1038/nrg3833.

[2] Sanjay M. Prakadan, Alex K. Shalek, and David A. Weitz. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nature Reviews Genetics*, 18(6):345–361, 2017. ISSN 1471-0056. doi: 10.1038/nrg.2017.15. URL http://www.nature.com/doifinder/10.1038/nrg.2017.15.

[3] Simona Patange, Michelle Girvan, and Daniel R. Larson. Single-cell systems biology: Probing the basic unit of information flow. *Current Opinion in Systems Biology*, 8:7–15, 2018. ISSN 24523100. doi: 10.1016/j.coisb.2017.11.011. URL https://doi.org/10.1016/j.coisb.2017.11.011.

[4] Vladimir Yu Kiselev, Tallulah S. Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics 2018*, 2019. ISSN 1471-0064. doi: 10.1038/s41576-018-0088-9. URL https://www.nature.com/articles/s41576-018-0088-9?utm{_}source=feedburner{&}utm{_}medium=feed{&}utm{_}campaign=Feed{%}3A+nrg{%}2Frss{%}2Fcurrent+{%}28Nature+Reviews+Genetics+-+Issue{%}29.

[5] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, May 2019. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-019-0071-9.

[6] Mitra Mojtahedi, Alexander Skupin, Joseph Zhou, Ivan G. Castaño, Rebecca Y. Y. Leong-Quong, Hannah Chang, Kalliopi Trachana, Alessandro Giuliani, and Sui Huang. Cell Fate Decision as High-Dimensional Critical State Transition. *PLOS Biology*, 14(12):e2000640, December 2016. ISSN 1545-7885. doi: 10.1371/journal.pbio.2000640.

[7] 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.

[8] 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.

[9] Beomseok Kim, Eunmin Lee, and Jong Kyoung Kim. Analysis of Technical and Biological Variabilityin Single-Cell RNA Sequencing. In *Computational Methods for Single-Cell Data Analysis*, volume 1935, pages 25–43. 2019. ISBN 978-1-4939-9056-6. doi: 10.1007/978-1-4939-9057-3. URL http://www.ncbi.nlm.nih.gov/pubmed/30758827{%}0Ahttp://link.springer.com/10.1007/978-1-4939-9057-3{_}12{%}0Ahttp://link.springer.com/10.1007/978-1-4939-9057-3.

[10] Davis J. McCarthy, Kieran R. Campbell, Aaron T.L. Lun, and Quin F. Wills. Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186, 2017. ISSN 14602059. doi: 10.1093/bioinformatics/btw777.

[11] Catalina A. Vallejos, John C. Marioni, and Sylvia Richardson. BASiCS: Bayesian analysis of single-cell sequencing data. *PLOS Computational Biology*, 11:e1004333, 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004333. URL http://dx.plos.org/10.1371/journal.pcbi.1004333.

[12] 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.

[13] Nils Eling, Arianne C. Richard, Sylvia Richardson, John C. Marioni, and Catalina A. Vallejos. Robust expression variability testing reveals heterogeneous T cell responses. *bioRxiv*, page 237214, 2017. doi: 10.1101/237214. URL https://www.biorxiv.org/content/early/2017/12/21/237214.full.pdf+html.

[14] Carl Boettiger. An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1):71–79, January 2015. ISSN 0163-5980. doi: 10.1145/2723872.2723882.

[15] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002. ISSN 1095-9203. doi: 10.1126/science.1070919. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{&}db=PubMed{&}dopt=Citation{&}list{_}uids=12183631.

[16] Nils Eling, Michael D. Morgan, and John C. Marioni. Challenges in measuring and understanding biological noise. *Nature Reviews Genetics*, 20(9):536–548, September 2019. ISSN 1471-0056, 1471-0064. doi: 10.1038/s41576-019-0130-6.

[17] Cristopher J. Zopf, Katie Quinn, Joshua Zeidman, and Narendra Maheshri. Cell-Cycle Dependence of Transcription Dominates Noise in Gene Expression. *PLoS Computational Biology*, 9(7):1–12, 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1003161.

[18] Kazunari Iwamoto, Yuki Shindo, and Koichi Takahashi. Modeling Cellular Noise Underlying Heterogeneous Cell Responses in the Epidermal Growth Factor Signaling Pathway. *PLoS Computational Biology*, 12(11):1–18, 2016. ISSN 15537358. doi: 10.1371/journal.pcbi.1005222.

[19] Daniel J. Kiviet, Philippe Nghe, Noreen Walker, Sarah Boulineau, Vanda Sunderlikova, and Sander J. Tans. Stochasticity of metabolism and growth at the single-cell level. *Nature*, 514(7522):376–379, 2014. ISSN 0028-0836. doi: 10.1038/nature13582. URL http://www.nature.com/doifinder/10.1038/nature13582.

[20] James Eberwine and Junhyong Kim. Cellular Deconstruction: Finding Meaning in Individual Cell Variation. *Trends in Cell Biology*, 25(10):569–578, 2015. ISSN 18793088. doi: 10.1016/j.tcb.2015.07.004. URL http://dx.doi.org/10.1016/j.tcb.2015.07.004.

[21] Andre J. Faure, Jörn M. Schmiedel, and Ben Lehner. Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. *Cell Systems*, 5(5):471–484, 2017. ISSN 24054720. doi: 10.1016/j.cels.2017.10.003.

[22] Michael D. Morgan and John C. Marioni. CpG island composition differences are a source of gene expression noise indicative of promoter responsiveness. *Genome Biology*, 19(1):1–13, 2018. ISSN 1474760X. doi: 10.1186/s13059-018-1461-x.

[23] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah a Teichmann, John C Marioni, and Marcus G Heisler. Accounting for technical noise in single-cell RNA-seq experiments., 2013. ISSN 1548-7105. URL http://www.ncbi.nlm.nih.gov/pubmed/24056876.

[24] 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.

[25] Catalina A Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods*, 14(6):565–571, 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4292. URL http://www.nature.com/doifinder/10.1038/nmeth.4292.

[26] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, February 2014. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.2772.

[27] Ashraful Haque, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1):75, December 2017. ISSN 1756-994X. doi: 10.1186/s13073-017-0467-4.

[28] Robert A. Amezquita, Vince J. Carey, Lindsay N. Carpp, Ludwig Geistlinger, Aaron T. L. Lun, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, Levi Waldron, Hervé Pagès, Mike Smith, Wolfgang Huber, Martin Morgan, Raphael Gottardo, and Stephanie C. Hicks. Orchestrating Single-Cell Analysis with Bioconductor. Preprint, Genomics, March 2019.

[29] Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Frederik Otzen Bagger, Davis James Mccarthy, John C Marioni, and Sarah A Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17(29):1–15, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0888-1. URL http://dx.doi.org/10.1186/s13059-016-0888-1.

[30] Aaron T. L. Lun, Karsten Bach, and John C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):75, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0947-7. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0947-7.

[31] Aleksandra A. Kolodziejczyk, Jong Kyoung Kim, Jason C.H. Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N. Natarajan, Alex C. Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, John C. Marioni, and Sarah A. Teichmann. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17:471–485, 2015. ISSN 19345909. doi: 10.1016/j.stem.2015.09.011. URL http://linkinghub.elsevier.com/retrieve/pii/S193459091500418X.