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. Single-cell RNA sequencing can be used to quantify this heterogeneity, but it is prone to strong technical noise. Here, we describe a step-by-step computational workflow which uses the BASiCS Bioconductor package to robustly quantify expression variability within and between known cell populations (such as experimental conditions or cell types). BASiCS provides 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 be used identify highly variable genes that drive the heterogeneity within the population as well as lowly variable genes that might exhibit housekeeping-like behavior. BASiCS also provides a probabilistic rule to identify changes in expression variability between cell populations, while 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 and 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 in order to quantify technical noise. The second dataset was generated using a droplet-based system, for which spike-in RNA is not available. The latter analysis provides an example, in which differential variability testing reveals insights regarding a possible early cell fate commitment process.

Keywords

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

*eling@ebi.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 remains otherwise undetected in bulk experiments [1, 2, 3]. Applications of scRNA-seq range from characterising cell types in immunity [4, 5, 6] and development [7, 8, 9] to dissecting the mechanisms for cell fate commitment [10, 11]. Transcriptional heterogeneity within a cell population can relate to different underlying sources. On the broadest level, this heterogeneity can reflect the presence of distinct expression profiles associated to cell subtypes or discrete states which could be identified through clustering [12]. Alternatively, cell-to-cell expression heterogeneity can be due to gradual changes along biological processes that evolve over time (such as development and differentiation) — these can be characterised using pseudotime inference methods [**?**]. More subtle expression variability within a seemingly homogeneous cell population can be due to deterministic or stochastic events and is the focus of this article. The stochastic component of this variability is referred to as transcriptional *noise* [13**?**].

Transcriptional noise can arise from intrinsic and extrinsic sources of variability. Classically, extrinsic noise is defined as stochastic fluctuations in cellular components, which is induced by cells residing in different dynamic states (e.g. cell size, cell cycle, metabolism, intra- and inter-cellular signalling) [14, 15, 16]. Instead, intrinsic noise arises from stochastic effects on biochemical processes such as transcription and translation [13]. Intrinsic noise can be modulated by genetic and epigenetic modifications (such as mutations, histone modifications, CpG island length and nucleosome positioning) [17, 18, 19] and is usually measured at the level of individual genes [13]. Cell-to-cell gene expression variability estimates derived from scRNA-seq data capture a combination of these effects, as well as deterministic regulatory mechanisms [**?**]. These variability estimates can also be inflated by the technical noise that is typically observed in scRNA-seq assays [20].

Different strategies have been implemented to quantify or attenuate technical noise in scRNA-seq experiments. For example, external RNA spike-in molecules (such as the ones introduced by the External RNA Controls Consortium, ERCCs [21]) can be added to each cell's lysate. Spike-ins can be used to inform quality control steps [22], data normalisation [23] as well as to infer technical background noise [20].

Some computational methods aim to denoise the data prior to downstream analysis (e.g. via imputation). Alternatively, computational approaches can be designed to simultaneously quantify technical variabli

was to quantify or attenuate technical noise in scRNA-seq assays. Some are experimental: eg spike-ins or UMIs. Some others are computational.

Moreover, technical noise inflates the observed cell-to-cell variability in gene expression [20]. To account for high amounts of technical noise that affects scRNA-seq data,

Fitting a regression trend between the variability and the mean abundance of the ERCC molecules allows the statistical detection of genes the show larger variability than the technical background .

Genes that show larger variability compared to spike-in molecules or the average variability are often referred to as 'highly variable genes' (HVG) and are used in computational scRNA-Seq analysis to select biologically informative genes for down-stream analysis [24]. Furthermore, spike-in molecules can be used to normalize gene expression for cells with differences in total mRNA content.

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] Tapio Lönnberg, Valentine Svensson, Kylie R. James, Daniel Fernandez-Ruiz, Ismail Sebina, Ruddy Montandon, Megan S. F. Soon, Lily G. Fogg, Arya Sheela Nair, Urijah N. Liligeto, Michael J. T. Stubbington, Lam-Ha Ly, Frederik Otzen Bagger, Max Zwiessele, Neil D. Lawrence, Fernando Souza-Fonseca-Guimaraes, Patrick T. Bunn, Christian R. Engwerda, William R. Heath, Oliver Billker, Oliver Stegle, Ashraful Haque, and Sarah A. Teichmann. Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves Th1/Tfh fate bifurcation in malaria. *Science Immunology*, 2(9):eaal2192, 2017. ISSN 2470-9468. doi: 10.1126/sciimmunol.aal2192. URL http://immunology.sciencemag.org/lookup/doi/10.1126/sciimmunol.aal2192.

[5] Alexandra Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, Laura Jardine, David Dixon, Emily Stephenson, Emil Nilsson, Ida Grundberg, David McDonald, Andrew Filby, Weibo Li, Philip L. De Jager, Orit Rozenblatt-Rosen, Andrew A. Lane, Muzlifah Haniffa, Aviv Regev, and Nir Hacohen. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335), 2017. ISSN 10959203. doi: 10.1126/science.aah4573.

[6] Grace X.Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8: 1–12, 2017. ISSN 20411723. doi: 10.1038/ncomms14049. URL http://dx.doi.org/10.1038/ncomms14049.

[7] Ximena Ibarra-Soria, Wajid Jawaid, Blanca Pijuan-Sala, Vasileios Ladopoulos, Antonio Scialdone, David J. Jörg, Richard C.V. Tyser, Fernando J. Calero-Nieto, Carla Mulas, Jennifer Nichols, Ludovic Vallier, Shankar Srinivas, Benjamin D. Simons, Berthold Göttgens, and John C. Marioni. Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nature Cell Biology*, 20(2):127–134, 2018. ISSN 14764679. doi: 10.1038/s41556-017-0013-z.

[8] Daniel E. Wagner, Caleb Weinreb, Zach M. Collins, James A. Briggs, Sean G. Megason, and Allon M. Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 4362:1–12, 2018. ISSN 10959203. doi: 10.1126/science.aar4362.

[9] Blanca Pijuan-Sala, Jonathan A. Griffiths, Carolina Guibentif, Tom W. Hiscock, Wajid Jawaid, Fernando J. Calero-Nieto, Carla Mulas, Ximena Ibarra-Soria, Richard C.V. Tyser, Debbie Lee Lian Ho, Wolf Reik, Shankar Srinivas, Benjamin D. Simons, Jennifer Nichols, John C. Marioni, and Berthold Göttgens. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495, 2019. ISSN 14764687. doi: 10.1038/s41586-019-0933-9.

[10] Mubeen Goolam, Antonio Scialdone, Sarah J L Graham, Iain C. MacAulay, Agnieszka Jedrusik, Anna Hupalowska, Thierry Voet, John C. Marioni, and Magdalena Zernicka-Goetz. Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell*, 165(1):61–74, 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.01.047. URL http://dx.doi.org/10.1016/j.cell.2016.01.047.

[11] Yusuke Ohnishi, Wolfgang Huber, Akiko Tsumura, Minjung Kang, Panagiotis Xenopoulos, Kazuki Kurimoto, Andrzej K Oleś, Marcos J Araúzo-Bravo, Mitinori Saitou, Anna-Katerina Hadjantonakis, and Takashi Hiiragi. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nature Cell Biology*, 16(1):27–37, 2014. ISSN 1476-4679. doi: 10.1038/ncb2881. URL http://dx.doi.org/10.1038/ncb2881.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.