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

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 population of cells can relate to different underlying structures. On the broadest level, this heterogeneity can relate to the presence of distinct expression profiles associated to cell subtypes or discrete states, which could be characterised through clustering [12]. Alternatively, cell-to-cell expression heterogeneity can reflect gradual changes along processes that evolve over time and that can be characterised using pseudotime inference methods [13]. The focus of this article is on more subtle expression variability that can occur within a seemingly homogeneous cell population. This variability can be due to deterministic or stochastic events that regulate gene expression and has been reported to increase prior to cell phate decisions [**?**] as well as throughout ageing [14].

This article complements existing workflows that use the Bioconductor package ecosystem to analyse scRNA-seq datasets [15, 16], including the use of *scater* and *scran* to perform quality control steps and low-level preliminary analysis [17, 15]. We present a step-by-step computational workflow to robustly quantify transcriptional variability using the *BASiCS* package [18, 19, 20]. *BASiCS* implements a Bayesian hierarchical framework that simultaneously performs data normalisation (global scaling), technical noise quantification and selected downstream analyses whilst propagating statistical uncertainty across these steps. Within a population of cells, *BASiCS* decomposes the total observed variability in gene expression measurements into technical and biological components. This enables the identification of highly variable genes that [TBC]. Moreover, this variance decomposition enables detection of lowly variable genes with stable expression [TBC - CITE GIGASCIENCE PAPER]. When two or more groups of cells are available (e.g. experimental conditions or cell types), *BASiCS* uses differential expression analysis to identify genes whose expression patterns change [19].

Since the era of RNA sequencing, methods for differential expression testing of transcript counts across conditions have been developed [21, 22]. Due to high technical variability and sparsity in scRNA-seq data, new approaches were developed for differential expression testing for scRNA-seq data [23, 24, 25]. In contrast to bulk samples, scRNA-seq measures variations in gene expression across a population of cells, and can therefore be used to test for changes in expression variability between two conditions. To do this, `BASiCS` compares the gene-specific over-dispersion parameters between two conditions. These parameters are independent of technical noise and can be used as proxy for biological variability [19]. Similar to the mean-variability trend observed for normalised scRNA-seq data [26], the estimates for over-dispersion parameters decrease with mean expression [19]. To correct for this, BASiCS has been extended to model the mean-variability relationship and capture residual over-dispersion estimates that show no association to mean expression. Therefore, this extension allows to test changes in mean expression in parallel to changes in variability [27].

Two case studies exemplify the use of `BASiCS` for non-UMI and UMI scRNA-seq data. In the first case, `BASiCS` can be used to detect highly and lowly variable genes and to obtain robust, gene-specific estimates to assess biological variability in naive CD4$^+$ T cells [14]; for a similar workflow see [16]. Furthermore, we compare naive to activated CD4$^+$ T cells to highlight the use of `BASiCS` to test for changes in mean expression and expression variability. In the second case, we use droplet-based scRNA-seq data to detect more subtle transcriptional changes during embryonic somitogenesis [7]

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] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010. ISSN 1465-6906. doi: 10.1186/gb-2010-11-10-r106. URL http://genomebiology.com/2010/11/10/R106.

[22] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp616.

[23] Shintaro Katayama, Virpi Töhönen, Sten Linnarsson, and Juha Kere. SAMstrt: Statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics*, 29(22):2943–2945, 2013. ISSN 13674803. doi: 10.1093/bioinformatics/btt511.

[24] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–2, jul 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2967. URL http://www.ncbi.nlm.nih.gov/pubmed/24836921.

[25] Mihails Delmans and Martin Hemberg. Discrete distributional differential expression (D(3)E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC bioinformatics*, 17(1):110, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-0944-6. URL http://dx.doi.org/10.1186/s12859-016-0944-6{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26927822{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4772470.

[26] 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.

[27] 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.