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]. On the broadest level, this heterogeneity can reflect the presence of distinct cell subtypes or states. Alternatively, cell-to-cell expression heterogeneity can be due to gradual changes along processes that evolve over time, such as development and differentiation. Several clustering and pseudotime inference methods have been developed to characterise these sources of heterogeneity [4, 5]. However, there is a limited availability of computational tools tailored to study more subtle variability within a seemingly homogeneous cell population. This variability can be due to deterministic or stochastic events that regulate gene expression and, among others, has been reported to increase prior to cell phate decisions [6] as well as throughout ageing [7].

This article complements existing scRNA-seq workflows based on the Bioconductor package ecosystem (e.g. [8, 9]). We describe a step-by-step analysis which uses *scater* and *scran* to perform quality control as well as initial exploratory analyses [10, 8]. To robustly quantify transcriptional variability within and between pre-specified cell populations (such as experimental conditions or cell types) we use *BASiCS* [11, 12, 13] — 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. Among others, *BASiCS*, has led to new insights about the heterogeneity of immune cells [7].

Within a population of cells, *BASiCS* decomposes the total observed variability in expression measurements into technical and biological components [11]. This enables the identification of *highly variable genes* (HVGs) that capture the major sources of heterogeneity within the analysed cells [14]. HVG detection is often used as feature selection, to identify the input set of genes for subsequent analyses. *BASiCS* can also highlight *lowly variable genes* (LVGs) that exhibit stable expression across the population of cells. These may relate to essential cellular functions and can assist the development of new data normalisation or integration strategies [15].

In order to compare expression patterns between two or more pre-specified groups of cells, *BASiCS* provides a probabilistic decision rule to perform differential expression analyses [12, 16]. Whilst several differential expression tools have been proposed for scRNA-seq data (e.g. [17**?**]), some evidence suggests that these do not generally outperform popular bulk RNA-seq tools [**?**]. Moreover, most of these methods are only designed to uncover changes in overall expression, ignoring the more complex patterns that can arise at the single cell level [**?**]. Instead, *BASiCS* embraces the high granularity of scRNA-seq data, uncovering changes in cell-to-cell expression variability that are not confounded by differences in technical noise or in overall expression.

In this manuscript, we briefly discuss the sources of variability that arise in scRNA-seq data and some of the strategies that have been designed to control or attenuate technical noise in these assays. We also summarise the main features of the Bioconductor packages that are used throughout this workflow and provide a description for the underlying statistical model implemented in *BASiCS*. This 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 to post-process the model outputs. Finally, we provide two step-by-step case studies using naive $CD4^+$ T cells [7] and samples collected during embryonic somitogenesis [18]. These examples were chosen to illustate the usage of *BASiCS* in the context of different scRNA-seq protocols.

All source code used to generate the results presented in this article is available at `https://github.com/VallejosGroup/BASiCSWorkflow`. To ensure the reproducibility of this workflow, the analysis environment and all software dependencies are provided as a Docker [**?**] image at: [ADD LINK].

Sources of variability in scRNA-seq datasets

An overarching goal of scRNA-seq experiments is to characterise gene expression heterogeneity with cellular resolution. The focus of this article is to quantify the strength of this heterogeneity within seemingly homogeneous cell populations. Here, we briefly describe the underlying sources of heterogeneity that can be captured by cell-to-cell variability estimates derived from scRNA-seq data.

Stochastic variability within a population of cells is often referred to as transcriptional *noise* and can arise from intrinsic and extrinsic sources [19, 20]. Classically, extrinsic noise is defined as stochastic fluctuations in cellular components induced by cells residing in different dynamic states (e.g. cell size, cell cycle, metabolism, intra- and inter-cellular signalling) [21, 22, 23]. In contrast, intrinsic noise arises from stochastic effects on biochemical processes such as transcription and translation [19]. Intrinsic noise can be modulated by genetic and epigenetic modifications (such as mutations, histone modifications, CpG island length and nucleosome positioning) [24, 25, 26] and is usually measured at the level of individual genes [19]. Cell-to-cell gene expression variability estimates derived from scRNA-seq data capture a combination of these effects, as well as deterministic regulatory mechanisms [20]. Moreover, these variability estimates can also be inflated by the technical noise that is typically observed in scRNA-seq assays [14].

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 [27]) can be added to each cell's lysate in a (theoretically) known fixed quantity. Spike-ins can assist quality control steps [10], data normalisation [28] and can be used to infer technical noise [14]. Another strategy is to tag individual cDNA molecules using unique molecular identifiers (UMIs) before PCR amplification [29]. 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 [28]). However, despite the benefits associated to the use of spike-ins and UMIs, these are not available for all scRNA-seq protocols [30].

Methods

This step-by-step scRNA-seq analysis workflow is primarily based on the Bioconductor package ecosystem [31]. A graphical overview for the workflow is provided in Figure 1 and its main components are described below.

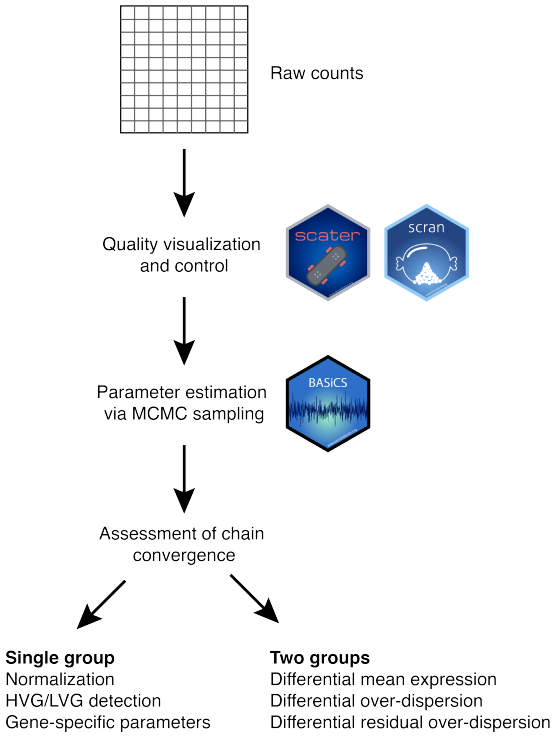

Figure 1. Graphical overview for the scRNA-seq analysis workflow described in this manuscript. Starting from a SingleCellExperiment object, we use the scater and scran Bioconductor packages to perform quality control and initial exploratory analyses. The primary focus of this workflow is to robustly quantify transcriptional heterogeneity within seemingly homogeneous cell populations. For this purpose, we apply the BASiCS Bioconductor package, illustrating how BASiCS can be used to analyse a single or multiple pre-specified groups of cells.

Input data - `SingleCellExperiment`

Here, we use the *SingleCellExperiment* package to convert an input matrix of raw read-counts (molecule counts for UMI-based protocols) into a `SingleCellExperiment` object. Such object can be used to store scRNA-seq data and its associated metadata, such as gene- and cell-specific information. Moreover, when available, the same object can be used to store read-counts for technical spike-in molecules (these can be accessed via the `altExp()` method). A major advantage of using a `SingleCellExperiment` object as the input for scRNA-seq analyses is that it enables interoperability across a large number of Bioconductor packages [31].

Quality control and exploratory analyses - `scater` and `scran`

An critical step in scRNA-seq analyses is to apply quality control (QC) diagnostics, removing low quality samples (wells or droplets, depending on the protocol) that may distort downstream analyses. Among others, QC can help to identify samples that contain broken cells, that are empty or that contain multiple cells [**?**]. Moreover, lowly expressed genes for which less reliable information is available are typically also removed. To perform QC of scRNA-seq data, we use the *scater* Bioconductor package [10]. In *scater*, the `calculateQCMetrics` function can be used to calculate standard QC metrics for each cell (e.g. percentage of mitochondrial reads) and gene (e.g. percentage of zeroes across all cells). The package also provides a suite of visualisation tools that can be used to explore the data under study and its associated QC diagnostic metrics.

The *scran* package offers additional tools for QC diagnostics and a variety of functions scRNA-seq data analysis [8]. In particular, it can perform *global scaling* data normalisation, calculating cell-specific scaling factors that capture global differences in read-counts across cells (e.g. due to variability in sequencing depth and PCR amplification) [32]. In *scran*, users can also explore how the observed variability in gene expression counts can be decomposed into technical and biological components (see the `decomposeVar` function). Moreover, the `trendVar` function can be used to infer an overall trend between gene-specific mean and variance estimates. To derive gene-specific variability estimates that are not confounded by this overall trend, the `DM` function calculates the distance between gene-specific squared coefficients of variation (CV^2) and a rolling median along the range of mean expression values [33]. DM variability estimates enable exploratory analyses of cell-to-cell heterogeneity, but a measure of uncertainty is not readily available. As such, gene-specific downstream inference (such as differential variability testing) is precluded.

Quantifying cell-to-cell transcriptional variability - `BASiCS`

The *BASiCS* package implements a Bayesian hierarchical framework which borrows information across all genes and cells to robustly quantify transcriptional variability [34]. Similar to the approach adopted in *scran*, *BASiCS* infers cell-specific global scaling normalisation parameters. However, instead of inferring these as a pre-processing step, *BASiCS* uses an integrated approach in which data normalisation and downstream analyses are performed simultaneously — whilst propagating statistical uncertainty. To quantify technical noise, the original implementation of *BASiCS* uses information from extrinsic spike-in molecules as control features, but the model has been extended to address situations in which spike-ins are not available [16].

BASiCS summarises the distribution of gene expression through gene-specific *mean* (μ_i) and *over-dispersion* (δ_i) parameters. Mean parameters μ_i quantify the overall expression for each gene i across the population of cells under study. Instead, δ_i captures the excess of variability that is observed with respect to what would be expected in a homogeneous cell population, after taking into account technical noise. This is used as a proxy to quantify transcriptional variability. Moreover, to account for the strong association that is typically observed between mean expression and over-dispersion estimates, we recently introduced gene-specific *residual over-dispersion* parameters ϵ_i [16]. Similar to DM values implemented in *scran*, these are defined as deviations with respect to an overall regression trend that captures the relationship between mean and over-dispersion values.

Parameter estimation is implemented in the `BASiCS_MCMC` function, which can be run using four different major settings (Table 1). If spike-in molecules are available and `WithSpikes = TRUE` (default), the model uses this information in order to infer technical noise. Alternatively, if `WithSpikes = FALSE`, the approach introduced in [16] is applied. We recommend to use `Regression = TRUE` (default) as the underlying informative prior improves the stability of posterior estimates for small sample sizes and lowly expressed genes [16]. Moreover, it provides additional flexibility for downstream analysis as it also infers gene-specific residual over-dispersion parameters ϵ_i.

The algorithm implemented in the `BASiCS_MCMC` function is an adaptive Metropolis within Gibbs sampler [**?**]. The `BASiCS_MCMC` function returns a `BASiCS_Chain` object, which can be used for further downstream analyses, many of which are detailed in this workflow. These objects contain draws from Markov chain Monte Carlo (MCMC) samplers, which are used to infer the posterior distribution over the model parameters [35]. Briefly, the posterior distribution quantifies how probable different parameter values are given the observed data. However, before assessing the posterior distribution, we must first ensure that the MCMC sampler has

Table 1. Four settings available for the the `BASiCS_MCMC` function.

	No regression	Regression
Using spike-in reads	`WithSpikes = TRUE`	`WithSpikes = TRUE`
	`Regression = FALSE`	`Regression = TRUE`
No spike-ins available	`WithSpikes = FALSE`	`WithSpikes = FALSE`
	`Regression = FALSE`	`Regression = TRUE`

converged to its stationary distribution, and has sampled efficiently from this distribution [36]. If these conditions are not met, then the estimated parameters may be inaccurate. The *coda* CRAN package contains a variety of functions to assess the convergence of a sampled MCMC chain. To use `coda` functions, the individual chains returned by `BASiCS` need to be transformed into a MCMC object that `coda` recognises using the `coda::mcmc` function. `BASiCS` also offers a number of functions to visualise and assess the convergence of MCMC chains. In particular, we will use `BASiCS_EffectiveSize` and `BASiCS_DiagPlot` to calculate and visualise the effective sample size generated by the MCMC samplers.

[talk about downstream analyses; hvg/lvg; differential testing then mention the ability to perform differential testing; finally the extention to account for mean/over-dispersion]

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] 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.

[15] Yingxin Lin, Shila Ghazanfar, Dario Strbenac, Andy Wang, Ellis Patrick, David M Lin, Terence Speed, Jean Y H Yang, and Pengyi Yang. Evaluating stably expressed genes in single cells. *GigaScience*, 8(9):giz106, September 2019. ISSN 2047-217X. doi: 10.1093/gigascience/giz106.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Catalina A. Vallejos, John C. Marioni, and Sylvia Richardson. BASiCS: Bayesian analysis of single-cell sequencing data. *PLOS Computational Biology*, 11(6):e1004333, 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004333. URL http://dx.plos.org/10.1371/journal.pcbi.1004333.

[35] A. F. M. Smith and G. O. Roberts. Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(17):3–23, September 1993. ISSN 00359246. doi: 10.1111/j.2517-6161.1993.tb01466.x.

[36] Mary Kathryn Cowles and Bradley P. Carlin. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91(434):883–904, June 1996. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1996.10476956.