Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data

[illegible]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome 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]Cambridge Institute for Medical Research, University of Cambridge, Cambridge Biomedical Campus, Hills Road, Cambridge CB2 0XY, UK
[4]MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK
[5]The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK
[6]Department of Statistical Science, University College London, 1-19 Torrington Place, London WC1E 7HB, UK
[7]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XY, UK
[8]Lead Contact
*Correspondence: marioni@ebi.ac.uk (J.C.M.), catalina.vallejos@igmm.ed.ac.uk (C.A.V.)

https://doi.org/10.1016/j.cels.2018.06.011

SUMMARY

unique in its composition, covering a range of different cell types and experimental protocols (see STAR Methods and Table S1).

BASiCS model appears to underestimate d_i for lowly expressed genes when the sample size is small (with respect to the parameter estimates obtained based on the full dataset of 939 cells). In contrast, the shrinkage introduced by our regression BASiCS model aids parameter estimation, leading to robust estimates even for the smallest sample size. This is particularly important for rare cell populations where large sample sizes are difficult to obtain. A similar effect is observed for genes with medium and high expression levels, where the non-regression BASiCS model appears to overestimate d_i. We also observe that estimates of residual over-dispersion parameters e integration approach is based on experimental designs where cells from a population are randomly allocated to multiple independent experimental replicates (here referred to as ‘‘batches’’). In such an experimental design, the no-spikes implementation of BASiCS assumes that biological effects are shared across batches and that technical variation will be reflected by spurious differences. As shown in Figures 4C and 4D, posterior inference under the no-spikes BASiCS model closely matches the original implementation for datasets where spike-ins and batches are available. Technical details about the no-spikes implementation of BASiCS are discussed in STAR Methods and Figure S4.

Expression Variability Dynamics during Immune Activation and Differentiation

focus on samples collected 2, 4, and 7 days post malaria infection, for which more than 50 cells are available.

To study global changes in over-dispersion along the differentiation time course, we first compared posterior estimates for the gene-specific parameter d_i, focusing on genes for which mean expression does not change (see Figure 6A and STAR Methods). This analysis suggests that the expression of these genes is most tightly regulated at day 4, when cells are in a highly proliferative state. Moreover, between days 4 and 7, the a broader applicability of the BASiCS software and allow statistical testing of changes in variability that are not confounded by technical noise or mean expression.

In general, stable gene-specific variability estimates ideally divergence of Th1 and Tfh differentiation was previously identified (Lönnberg et al., 2017). This decrease in variability on day 4 is potentially due to the induction of a strong pan-lineage proliferation program. However, we observe that not all genes follow this trend and uncover four different patterns of variability changes. Second, we observe that several Tfh and Th1 lineage-associated genes change in expression variability

ribosome biogenesis, and proliferation of mouse CD8 T cells, In Proceedings of the National Academy of Sciences.

Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science *329*, 533–538.

Tung, P.Y., Blischak, J.D., Hsiao, C.J., Knowles, D.A., Burnett, J.E., Pritchard, J.K., and Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. Sci. Rep. *7*, 39921.

Vallejos, C.A., Marioni, J.C., and Richardson, S. (2015). BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Comput. Biol. *11*, e1004333.

Vallejos, C.A., Richardson, S., and Marioni, J.C. (2016). Beyond comparisons of means: understanding changes in gene expression at the single-cell level. Genome Biol. *17*, 70.

Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J.C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat. Methods *14*, 565–571.

Vallejos, C.A., and Steel, M.F.J. (2015). Objective Bayesian survival analysis using shape mixtures of log-normal distributions. J. Am. Stat. Assoc. *110*, 697–710.

West, M., and Harrison, J. (1989). Bayesian Forecasting and Dynamic Models (Berlin: Springer).

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873–881.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

Zeng, C., Mulas, F., Sui, Y., Guan, T., Miller, N., Tan, Y., Liu, F., Jin, W., Carrano, A.C., Huising, M.O., et al. (2017). Pseudotemporal ordering of single cells reveals metabolic control of postnatal b cell proliferation. Cell. Metab. *25*, 1160–1175.e11.

STAR★METHODS

KEY RESOURCES TABLE

CONTACT FOR REAGENT AND RESOURCE SHARING

level. However, the well known confounding effect between mean and variability that typically arises in scRNA-seq datasets (Brennecke et al., 2013) can preclude a meaningful interpretation of these results.

Modeling the Confounding between Mean and Dispersion

Here, we extend BASiCS to account for the confounding effect described above. For this purpose, we estimate the relationship between mean and over-dispersion parameters by introducing the following joint prior distribution for ðm,,

In both cases, the posterior probability threshold a_R is chosen to control the expected false discovery rate (EFDR) (Newton et al., 2004 from the same population to multiple independent experimental replicates (hereafter these are referred to as

Equation 39

$$
\begin{aligned}
\Sigma &\propto (1 □ 1/\ ) □ \; □ 1'_{□1} \; I_{□1} □ 1_{□1}1'_{□1} \; (1_{□1} \ ) \\
&= 1 □ \frac{1}{} □ \frac{1}{2} 1'_{□1}(I_{□1} + 1_{□1}1'_{□1})1_{□1} \; (\text{see Miller, 1981}) \\
&= 1 □ \frac{1}{} □ \frac{1}{2} 1'_{□1}(1_{□1} + (□ 1)1_{□1}) \\
&= 1 □ \frac{1}{} □ \frac{1}{2} \; 1'_{□1}1_{□1} \equiv 0
\end{aligned}
\qquad \text{(Equation 49)}
$$

Proposition 3. Under the same assumptions as in Proposition 1. Let $m_{□ι,}$ be the vector obtained after removing elements i and r from $(m_1, \ldots, m\ )'$. It can be shown that

$$
\log(m_ι)|\log\left(m_{□ι,}\right) □ N\left(\frac{1}{2}\left[\log(m_0) □ 1'_{□2}\log\left(m_{□ι,}\right)\right], \frac{1}{2}a_m^2\right), \qquad \text{(Equation 50)}
$$

where $1_{q□2}$ denotes a $(q□2)$ -dimensional vector of ones.

Proof. Standard multivariate normal theory leads to

$$
\log(m_ι)|\log\left(m_{□ι,}\right) □ N_1(m, \Sigma), \qquad \text{(Equation 51)}
$$

with

$$
\begin{aligned}
m &= \log(m_0) + \; □ 1'_{□2} \; I_{□2} □ 1_{□2}1'_{□2} \; \log\left(m_{□ι,}\right) □ \log(m_0)1_{□2} \\
&= \log(m_0) □ 1'_{□2} \; I_{□2} + \frac{1}{2}1_{□2}1'_{□2} \; \log\left(m_{□ι,}\right) □ \log(m_0)1_{□2} \\
&\qquad (\text{see Miller, 1981}), \\
&= \log(m_0) □ \frac{1}{2} \; 1'_{□2}\log\left(m_{□ι,}\right) □ (□ 2)\log(m_0) \\
&= \frac{}{2}\log(m_0) □ \frac{1}{2}1'_{□2}\log\left(m_{□ι,}\right)
\end{aligned}
\qquad \text{(Equation 52)}
$$

and

$$
\begin{aligned}
\Sigma &= a_m^2 \; (1 □ 1/\ ) □ \; □ 1'_{□2} \; I_{□2} □ 1_{□2}1'_{□2} \; (1_{□2} \ ) \\
&= a_m^2 \; 1 □ \frac{1}{} □ \frac{1}{2}1'_{□2} \; I_{□2} + \frac{1}{2}1_{□2}1'_{□2} \; 1_{□2} \; (\text{see Miller, 1981}) \\
&= a_m^2 \; 1 □ \frac{1}{} □ \frac{1}{2}1'_{□2} \; 1_{□2} + \frac{□ 2}{2}1_{□2} \\
&= \frac{1}{2}a_m^2
\end{aligned}
\qquad \text{(Equation 53)}
$$

Implementation

Bayesian inference is implemented using an adaptive Metropolis within Gibbs algorithm (Roberts and Rosenthal, 2009). After integrating out the random effects $r_{ι,}$ the full conditionals required for this implementation are based on the following likelihood function:

$$
\prod_{ι=1}\prod_{=1}\prod_{ι=1} \frac{\Gamma\left(ι + \frac{1}{d_ι}\right)}{\Gamma\left(\frac{1}{d_ι}\right) \; ι!} \; \frac{\frac{1}{d_ι}}{n_ι m_ι +}
$$

Dictyostelium Cells

Antolović et al. (2017) studied changes in expression variability between 0 hours (undifferentiated), 3 hours and 6 hours of *Dictyostelium* differentiation. Raw data is available by direct download (see Data S1 in Antolović et al., 2017). Across all time-points, 5 cells were removed due to low quality. Technical spike-in genes that were not detected and biological genes with an average expression (across all cells) smaller than 1 count were removed. In total, 433 cells (131 cells and 3 batches at 0h, 157 cells and 3

Functional Annotation Analysis

We performed functional annotation analysis using DAVID version 6.8 (Dennis et al., 2003). All genes considered for differential testing were used as background. The functional annotation clustering function in DAVID was used to cluster annotation categories based on similarity and to sort them according to their enrichment score.

Stabilization of Posterior Inference for Small Sample Sizes

To compare parameter estimates of the regression and non-regression model across different sample sizes, we used the CA1 pyramidal neuron population from Zeisel et al. (2015). The regression BASiCS model was first run on the full population of 939 cells to generate se d ground truth parameter estimates. Subsequently, 50, 100, 150, 200, 250, 300 and 500 cells were randomly sub-sampled from the full population prior to parameter estimation. This procedure was repeated 10 times for each sample size. Based on parameter estimates using the non-regression model, we split the genes into three sets: lowly expressed ð$\mu_i<1.89$þ, medium expressed ð$1.89<\mu_i<5.37$þ and highly expressed ð$\mu_i>5.37$þ. These cut-off values were chosen such that a third of genes classifies into each category. We dissected the results of this experiment in three ways. First, we visualize boxplots showing all estimates of gene-specific parameters for a single sub-sampling experiment (Figure 3). Second, we computed the $\log_2$ fold change for estimates of gene-specific over-dispersion parameters δ_i between the regression and non-regression BASiCS models (Figures S3A–S3C). Third, for each sub-sampling experiment, sample size and gene set, we computed the median $\log_2$ fold change in μ_i and δ_i and the median difference for ϵ_i between estimates and the se d ground truth. The median and the range of these values across 10 sub-sampling experiment is used for visualization purposes (see Figure S3D–S3F).

External validation for posterior estimates of gene-specific model parameters was obtained using matched scRNA-seq and smFISH data of mouse embryonic stem cells grown in 2i and serum media (see Table S1 and Grün et al., 2014). As in Brennecke et al. (2013)