bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

bayNorm: Bayesian gene expression recovery, imputation and normalisation for single cell RNA-sequencing data

Wenhao Tang[1], François Bertaux[1,2,3], Philipp Thomas[1] ,Claire Stefanelli[1], Malika Saint[2,3], Samuel Marguerat[2,3,4] and Vahid Shahrezaei[1,4]

[1] Department of Mathematics, Faculty of Natural Sciences, Imperial College, London SW7 2AZ, UK

[2] MRC London Institute of Medical Sciences (LMS), Du Cane Road, London W12 0NN, UK

[3] Institute of Clinical Sciences (ICS), Faculty of Medicine, Imperial College London, Du Cane Road, London W12 0NN, UK

[4] Correspondence to samuel.marguerat@imperial.ac.uk or v.shahrezaei@imperial.ac.uk

Normalisation of single cell RNA sequencing (scRNA-seq) data is a prerequisite to their interpretation. The marked technical variability and high amounts of missing observations typical of scRNA-seq datasets make this task particularly challenging. Here, we introduce bayNorm, a novel Bayesian approach for scaling and inference of scRNA-seq counts. The method's likelihood function follows a binomial model of mRNA capture, while priors are estimated from expression values across cells using an empirical Bayes approach. We demonstrate using publicly-available scRNA-seq datasets and simulated expression data that bayNorm allows robust imputation of missing values generating realistic transcript distributions that match single molecule FISH measurements. Moreover, by using priors informed by dataset structures, bayNorm improves accuracy and sensitivity of differential expression analysis and reduces batch effect compared to other existing methods. Altogether, bayNorm provides an efficient, integrated solution for global scaling normalisation, imputation and true count recovery of gene expression measurements from scRNA-seq data.

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Introduction

scRNA-seq is a method of choice for profiling gene expression heterogeneity genome-wide across tissues in health and disease[1, 2]. Because it relies on the detection of minute amounts of biological material, namely the RNA content of one single cell, scRNA-seq is characterised by unique and strong technical biases. These arise mainly because scRNA-seq library preparation protocols recover only a small fraction of the total RNA molecules present in each cell. As a result, scRNA-seq data are usually very sparse with many genes showing missing values (i.e. zero values, also called dropouts). The fraction of all transcripts recovered from a cell is called capture efficiency and varies from cell to cell, resulting in strong technical variability in transcripts expression levels and dropouts rates. Moreover, capture efficiencies tend to vary between experimental batches resulting in confounding "batch effects". Correcting for these biases in order to recover scRNA-seq counts reflecting accurately the original numbers of transcripts present in a cell remains a major challenge in the field [3-5]

A common approach to scRNA-seq normalisation is the use of cell-specific global scaling factors. These methods are based on principles developed for normalisation of bulk RNA-seq experiments and assume that gene specific biases are small[3]. Typically, read counts per cell are divided by a cell specific scaling factor estimated either from spike-in controls[6], or directly from the transcriptome data using methods developed initially for bulk RNA-seq[7-9] or specifically for scRNA-seq[10, 11]. A recent method called SCnorm extended the global scaling approach by introducing different scaling factors for different expression groups[12].

Importantly, scaling methods do not correct for cell-to-cell variations in dropout rates, as genes with zero counts remain zero after division by a scaling factor. Several approaches have been designed to tackle this problem. A series of methods use zero-inflated distribution functions, to explicitly model the dropout charachteristics[13-15]. Alternatively, other studies have proposed to infer dropouts based on expression values pooled across cells or genes[16-19]. For instance, scImpute pools expression values across similar cell subpopulations in each dataset and imputes dropouts using a Gamma-Normal mixture model and population specific thresholds[18].

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Similarly, the MAGIC package is based on pooling gene expression values across cells using a network-based similarity metric[19]. Another method is based on K-nearest neighbour smoothing, which uses Poisson distribution and aggregate information from similar cells[20]. Conversely, the SAVER approach pools expression values across genes within each cell using a Gamma-Poisson Bayesian model[17]. The Gamma-Poisson model is also used in two other packages called Splatter and scVI for simulating and normalising scRNA-seq data respectively[21, 22]. scVI belongs to new class of approaches which implement deep learning variational autoencoder or autoencoder methods[16, 21, 23-25]. For instance, DCA, an autoencoder method, utilises a zero-inflated negative binomial noise model[16]. Experimental batch-to-batch variations are another common source of technical variability in scRNA-seq data. The origin of batch effects is not fully understood but results at least in part from differences in average capture efficiencies across experiments[26]. Several methods have been developed to specifically remove batch effect in scRNA-seq data[27-29].

The methods discussed above, treat normalisation, imputation, and batch effect correction as separate tasks. Moreover, they rely on strong assumptions such as the zero-inflation model. Here we provide a detailed account of a novel integrated approach called bayNorm which performs all the processing steps discussed above at the same time using minimal assumptions. We compared its performance with a series of available packages focusing on true count recovery, differential expression analysis and batch effect correction.

The bayNorm rationale

bayNorm is a Bayesian implementation of global scaling normalisation that simultaneously imputes missing values in scRNA-seq data. bayNorm generates for each gene (*i*) in each cell (*j*) a posterior distribution of original expression counts (x_{ij}^0), given the observed scRNA-seq read count for that gene (x_{ij}) (**Fig. 1a**). Using the Bayes rule we have:

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

$$P(x_{ij}^0|x_{ij}) = \frac{P_{\beta_j}(x_{ij}|x_{ij}^0)P(x_{ij}^0)}{P(x_{ij})}$$

Where $P(x_{ij}^0|x_{ij})$ is the posterior distribution of true gene expression counts of a given gene in a given cell. $P_{\beta_j}(x_{ij}|x_{ij}^0)$ is a likelihood function that depends on the cell specific capture efficiency (β_j). Specific capture efficiencies can be estimated using spike-in controls or directly from the data using scaling factors provided by different methods[3] and normalised to the dataset's mean capture efficiency $< \beta >$ (see Methods). $P(x_{ij}^0)$ is a gene specific prior expression distribution and $P(x_{ij})$ is the marginal likelihood. The outputs of bayNorm are either samples (3D array) or point estimates (2D array) from the posterior distributions (**Fig. S1**).

The binomial model is an appropriate choice for the bayNorm likelihood function

The bayNorm likelihood function $P_{\beta_j}(x_{ij}|x_{ij}^0)$ is at the core of the approach and describes the empirical distribution of the raw experimental scRNA-seq counts. The binomial model describes the random sampling of a fraction of a cell transcriptome with constant probability. This is a simple model of transcript capture in scRNA-seq[30] and we therefore hypothesised that it would be a good choice for bayNorm likelihood function. For the prior $P(x_{ij}^0)$, we assume a negative binomial model, which describes the bursty distribution of mRNAs in simple models of gene expression [31, 32]. Gene specific prior parameters are estimated using an empirical Bayes approach by pooling gene expression values across multiple cells of the dataset (see Methods for details).

To validate our choice of binomial likelihood model and prior estimates, we generated simulated scRNA-seq data based on these assumptions and investigated how closely they captured statistics of several published scRNA-seq datasets (**Fig. 1 b-e, Fig. S2-7**)[12, 30, 33, 34]. The simulations assumed mRNA counts per cell that followed negative binomial distributions and used gene specific priors obtained with bayNorm (**Fig. 1**, Binomial_bayNorm), or sampled from estimates obtained with a modified version of the Splatter package (**Fig. 1**,

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

'Binomial_Splatter', Supplementary Notes 1)[22]. These were compared with simulations generated with the original Splatter package which is based on the Gamma-Poisson distribution[22]. We note that in Splatter, scaling factors are multiplicative to the Gamma distribution's mean. In bayNorm, however, the cell specific capture efficiencies, which act as scaling factors, are set as the probability parameter of the binomial model. We found that the binomial model captures the variance-mean relationship of experimental scRNA-seq data well (**Fig. 1b**).

Another important feature of scRNA-seq data is their large amount of missing values, or dropouts, and several models have been proposed to explain this phenomenon[14, 15, 26, 35, 36]. We therefore investigated how well the binomial model would capture dropout rates in experimental data. Our simulated dataset generated using the 'Binomial_bayNorm' function reproduced accurately the dependence of dropout fractions on gene expression means performing better than Splatter (**Fig. 1c-e**). Moreover, a parameter free approximation based on the binomial model predicted the dropout fraction to depend on an exponential of the negative mean expression ($e^{-\bar{x}}$, see Methods). This functions produced a very close fit to the experimental data providing additional support for our choice of the binomial model (**Fig. 1c**). Notably, the Binomial_bayNorm simulation protocol using inferred gene-specific priors together with cell specific parameters (β_j) was the only one that recovered the distribution of dropout rates per gene observed in experimental data (**Fig. 1d**). Finally, the results presented on **Fig. 1b-e** could be replicated consistently using several additional experimental scRNA-seq datasets (**Fig. S2-7**).

The datasets discussed so far include sequencing scores corrected for PCR amplification biases using unique molecular identifiers (UMIs)[37]. Some popular protocols, however, do not include UMIs, and are therefore likely to be less well described by the binomial distribution due to technical variability arising from PCR amplification bias. Accordingly, their dependence of dropout fractions on the mean expression has been reported to be more complex than in UMI-based datasets [36]. We investigated this issue further and found that a simple scaling of non-UMI raw data by a constant factor produced a reasonable match to the binomial model (**Fig. S9**; see Methods). This scaling factor can be interpreted as the average number of times

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

original mRNA molecules were sequenced after PCR amplification. This indicates that, provided appropriate scaling, non-UMI datasets are also compatible with the bayNorm model. Importantly, as bayNorm recovered dropouts rates successfully in both UMI-based and non-UMI protocols without the need of specific assumptions, we conclude that invoking zero-inflation models is not required to describe scRNA-seq data. Consistent with this, the differences in mean expression levels of lowly expressed genes observed between bulk and scRNA-seq data, which were suggested to be indicative of zero-inflation, were recovered by our simulated data using the binomial model only (**Fig. S10**)[38].

We note that the ability of simulation protocols to recover the statistics of experimental data depended intimately on the value of cell-specific capture efficiencies (β_j). We used different ways to estimate β (spike-in, Scran scaling factors, trimmed means, or housekeeping genes; Supplementary Note) together with different $<\beta>$ in the Binomial_Splatter simulation protocol. We found that changes in β_j values affected recovery of the distribution of dropout rates per cell. (**Fig. S8**). In particular, we found that the use of spike-in controls or of housekeeping reference gene expression levels did not improve estimates of capture efficiencies (**Fig. S8c-f**). Altogether, this analysis demonstrates that accurate statistics of experimental scRNA-seq data can be consistently retrieved using the binomial model and empirical Bayes estimation of gene expression parameters implemented in bayNorm along with accurate estimates of cell-specific capture efficiencies.

bayNorm enables recovery of true gene expression distributions from scRNA-seq data

Single-cell RNA-seq provides a unique opportunity to study stochastic cell-to-cell variability in gene expression at a near genome-wide scale. However, doing this requires normalisation approaches able to retrieve from scRNA-seq data transcripts levels matching quantitatively *in vivo* mRNA numbers[33]. With this in mind, we evaluated bayNorm performance in reconstructing true gene expression levels from a series of experimental scRNA-seq datasets that contained matched single molecule fluorescence *in situ* hybridisation (smFISH) measurements for a series of genes. We used global mean capture efficiencies $<\beta>$

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

estimated directly from smFISH together with gene specific priors informed by the sequencing data (**Fig S11**). After bayNorm normalisation, scRNA-seq counts reproduced accurately count distributions obtained by smFISH for several mRNAs (**Fig 2a-b**). We then compared bayNorm performance with a series of published normalisation methods (Supplementary note 4, **Fig 2**). All methods captured mean smFISH counts across different genes well (**Fig. 2c-d, Fig S11**). However, noise in gene expression (coefficient of variation, CV) and expression dispersion (Gini coefficient) measured by smFISH were better captured by bayNorm compared to normalisation by scaling or by several recent normalisation and imputation methods (**Fig. 2e-f, Fig. 2g-h**) [12, 16-19]. bayNorm's good performance could also be confirmed in a series of simulation studies (**Fig S12**, Supplementary note 1). In summary, bayNorm combined with gene specific priors inferred directly from the scRNA-seq data, retrieves gene expression variability matching closely smFISH data.

bayNorm enables accurate and sensitive differential expression analysis

Differential genes expression analysis (DE) in scRNA-seq studies is challenging as several factors including variability in capture efficiencies, dropout rates, sequencing depth, and experimental batch effects can introduce significant, yet spurious, differential expression signal. Normalisation and imputation approaches have, therefore, a significant impact on the sensitivity and accuracy of DE analysis protocols. Two features of the bayNorm approach have the potential to improve the performance of DE analysis. Firstly, bayNorm posterior distribution of original counts maintains the uncertainty resulting from small capture efficiencies and could therefore reduce false positive DE discovery rates[39]. Secondly, the use of priors specific to each group of cells compared in the DE analysis could increase true positive discovery rates. With this in mind, we have assessed bayNorm performance in DE analysis using several experimental scRNA-seq datasets and compared it to other existing methods. To identify DE genes we use MAST[13], which performs well in terms of false positives rates, precision and recall[40]. MAST was first applied to individual sample from the bayNorm posterior distribution (3D array, **Fig. S1**). Differentially expressed genes were then called based on the median of Benjamini-Hochberg adjusted P-values of the individual samples[28].

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

As mentioned above, differences in capture efficiencies between cells is a source of technical variability that could affect DE analysis. To test bayNorm's ability to correct for this bias, we selected the 100 cells with the highest and lowest capture efficiencies based on total counts in a recent UMI-based scRNA-seq study[30]. We then applied bayNorm to the 200 cells using global prior estimation based on the combination of the two groups (see Methods). In this design, the two groups of cells differ based only on their capture efficiencies, and significant differential expression is therefore not expected. **Fig. 3a** shows the number of genes called differentially expressed as a function of increasing average expression levels using a series of normalisation and imputation methods[12]. bayNorm normalised data show almost no differentially expressed genes, outperforming all the other methods. Moreover, $\log_2$ gene expression ratios between cells of the two groups, were consistently close to zero, confirming bayNorm ability to correct for biases inherent to different capture efficiencies in UMI-based datasets (**Fig. 3b**).

Sequencing depth is another parameter affecting DE analysis especially because it impacts on the dropout rates of lowly expressed genes. Moreover, differences in sequencing depth are likely to affect levels of capture efficiencies, especially for non-UMI datasets where PCR biases are not accounted for. To assess bayNorm's ability to correct for this source of bias, we used a benchmark dataset published by Bacher and colleagues[12] that consists of non-UMI based scRNA-seq data for two groups of cells isolated from a single culture and sequenced to a depth of either 1 million or 4 million reads per cell. bayNorm and other imputation methods performed well in this setting (**Fig. S13**). However, a global scaling approach on its own led to poor results, unless performed independently on groups of genes with similar mRNA expression levels as in SCnorm. Finally, bayNorm corrected robustly for variability in sequencing depth when applied to a series of simulated datasets (**Fig. S14-15**)[12].

We have shown that bayNorm is efficient at removing spurious differential expression from scRNA-seq data caused by variability in capture efficiencies and sequencing depth. We next explored bayNorm performance in supporting sensitive and robust detection of genes truly regulated between samples. To do this, we used two experimental scRNA-seq datasets[41, 42] and lists of benchmark DE genes derived from matched bulk RNA-seq data[40, 43]. To maximise

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

sensitivity, we used priors specific to each groups of cells in the comparison (we call this design "local priors"). With the first dataset, bayNorm normalised data generated an AUC value as high as other normalisation methods demonstrating that the approach supports sensitive DE detection (**Fig. 3c**). Analysis of the second dataset (UMI-based)[42] confirmed this observation with bayNorm performing better than all other methods (**Fig. 3d**). Importantly, bayNorm performance did not depend on the number of cells in each group, except for groups with very low numbers of cells (**Fig. 3d**, **Fig. S16**). Finally, using a series of simulated datasets, we explored situations where the compared groups have different mean capture efficiencies and found that bayNorm supported robust DE detection in all cases (**Fig. S17**).

Three important parameters should be considered before bayNorm normalisation: i) the choice of priors, ii) the choice of average capture efficiencies $< \beta >$, iii) the choice of bayNorm output format. Prior parameters can be either estimated for all cells across groups (global) or within each group (local). Since priors are gene specific, applying bayNorm across homogeneous cells (*i.e.* using global prior) allows for mitigating technical variations (**Fig S18a-b**). On the other hand, using priors estimated "locally" within each group amplifies differences in signals between heterogeneous groups of cells increasing sensitivity (**Fig S18c-d**). Average capture efficiencies $< \beta >$ are specific to each scRNA-seq protocol and reflect their overall sensitivity. This value represents the ratio of the average number of mRNA molecules sequenced per cell to the total number of mRNA molecules present in an average cell. It is not always easy to determine as quantitative calibration methods such as smFISH are not widely used, and approaches based on spike-in controls have important shortcomings[3]. We investigated the impact of inaccurate estimation of $< \beta >$ on biases in DE detection. Critically we found that DE results based on bayNorm normalised data are not affected significantly by a 2 fold change of $< \beta >$ (**Fig. S20-S21**). Finally, bayNorm output consists of either samples from its posterior distributions (3D array) or the modes of these distributions as point estimates (2D arrays). For DE analysis using MAST, 3D arrays reduces false positive rates but 2D arrays perform slightly better in terms of AUC (**Fig S18c-d**). **Fig S19** shows DE results for two other non-parametric methods: ROTS[44] and Wilcoxon test[40]. Both approaches perform equally well

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

with 3D arrays but show variable results when applied to 2D arrays with the Wilcoxon test performing less well.

In summary, our analysis demonstrates that in addition to correcting for technical biases, bayNorm also supports robust and accurate DE analysis of a wide range of experimental and simulated scRNA-seq datasets.

bayNorm correction of experimental batch effects

scRNA-seq protocols are subject to significant experimental batch effects[34]. In cases where the study design does not take this problem into account by distributing cases and controls across batches for instance, batch effects can lead to artefactual differences in gene expression of single cells, resulting in inaccurate biological conclusions. bayNorm can mitigate batch effects in two ways. First, as described above, bayNorm efficiently corrects for differences in capture efficiencies which is a pervasive source of batch-to-batch variability[38]. Second, the use of bayNorm data-informed priors is an efficient way to mitigate batch variation by estimating prior parameters across different batches but within the same biological condition. To investigate bayNorm's performance for batch effect correction we use data from the Tung study [34] where scRNA-seq data were obtained in triplicates for three induced pluripotent stem cell lines (iPSC) derived from three individuals. Sequencing libraries were prepared in three experimental batches, each containing one repeat of each line[34]. We first used priors calculated within each individual, but across batches (bayNorm local (individual)). This strategy allows for maintaining differences between individuals while minimising batch effects as illustrated by PCA analysis (**Fig. 4a-b, Fig S22**). To assess the normalisation performance quantitatively, we extracted the number of genes differentially expressed between each pair of batches within the same individual (**Fig S23**). We defined the ratio of the number of DE genes (adjusted $P_{MAST} < 0.05$) and the total number of genes (13058) to be the false positive rates (FPR). In theory, batch effects should be the main source of differential expression between these samples[34]. In parallel, we tested whether bayNorm also maintained differences between individuals using the same settings. To do this, we defined DE genes between the iPSC lines NA19101 and NA19239 and compared it to a benchmark list of 498 DE genes[43]. Efficient batch effect correction is expected to minimise FPR while maximizing

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Area Under the Curve (AUC) values of DE detection between individuals. We find that using bayNorm with "within individual" local priors (estimated across different batches within the same line) outperformed other methods in terms of correcting batch effects while maintaining meaningful biological information. As expected, using bayNorm and global priors (estimated across batches and individuals, bayNorm global) preserves low FPR, but reduces AUC significantly. Finally, using bayNorm with "within batch" local priors (bayNorm local (batch)) result in higher false positive rates, which is also expected.

Overall we have showed that the flexibility of priors selection afforded by bayNorm Bayesian approach enables robust correction of batch effects, while maintaining sensitive detection of differentially expressed genes.

Conclusions

We introduced bayNorm, a versatile Bayesian approach for implementing global scaling that simultaneously provides imputation of missing values and true counts recovery of scRNA-seq data. Bayesian methods have been applied to different aspects of RNA-seq data analysis before[11, 15, 17, 45]. The approach most related to bayNorm is taken by SAVER which uses a Poisson-Gamma model and pooling information across genes for true-count recovery[17]. In contrast, bayNorm uses a binomial model of mRNA capture as likelihood and achieves similar or improved performance relative to SAVER on real and simulated data (Fig 2-4). We showed that using the binomial model and an empirical Bayes approach to estimating gene expression priors across cells results in simulated data almost identical to experimental scRNA-seq measurements. Importantly, this suggests that zero-inflated models are not required to explain the frequency of dropout observed in scRNA-seq. Although designed initially for UMI-containing scRNA-seq protocols, a simple scaling factor makes bayNorm applicable to non-UMI data as well. This flexibility will allow using this approach with most present and future scRNA-seq datasets. We showed using datasets that combine smFISH and scRNA-seq, that bayNorm is accurately recovering true gene expression across a wide range of expression levels. This approach could therefore be particularly useful for quantitative analysis of more difficult scRNA-seq datasets, such as those generated from small quiescent cells or microbes, for instance. In fact, we have recently used bayNorm successfully in the first scRNA-seq study

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

of fission yeast[46]. One of the most powerful features of bayNorm is its use of gene expression priors directly calculated from gene expression values across cells. We showed that by grouping cells according to experiment design or phenotypic features increased significantly the robustness and sensitivity of differential expression analysis. This allows almost complete removal of sequencing depth and capture efficiency biases, and reduced batch effects. Critically, this approach preserved accurate and sensitive detection of benchmark DE genes.

Accurate estimation of cell capture efficiencies (or scaling factors) is central to most scRNA-seq normalisation methods including bayNorm. Interestingly, we observed that the choice of cell specific capture efficiencies affect how closely simulated data recovers statistics of real data**.** We therefore propose that comparison of drop-out rates per cell in simulated datasets and experimental data could be used as a tool to inform appropriate choice of global scaling factors and mean capture efficiency estimates. The option to tailor bayNorm priors based on phenotypic information about cell subpopulations will be a powerful asset for discovery of gene expression programmes associated with specific phenotypic features of single cells such as cell size[46]. Finally, the concepts and mathematical framework behind bayNorm will be useful if combined with other emerging theoretical approaches such as deep learning, for instance [16, 21, 23-25]. Overall, bayNorm provides a simple and integrated solution to remove the technical biases typical of scRNA-seq approaches, while enabling robust and accurate detection of cell-specific changes in gene expression. bayNorm has been made freely available as an R package (see Methods), and will be submitted to Bioconductor.

Acknowledgements

We are grateful to Dan Hebenstreit for critical reading of the manuscript. The benchmark lists used in the Islam and Tung studies were kindly provided by Maria K. Jaakkola and Chengzhong Ye respectively. We would like to thank Rhonda Bacher for providing R code for running MAST and producing figures 3a and 3b. We would like to thank Mo Huang for the code for preprocessing data from the Torre Study. We would also like to thank Lennart Kester for providing smFISH data used in the Grün study. This research was supported by the UK Medical Research Council, and a Leverhulme Research Project Grant (RPG-2014-408). WT is

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

supported by a Roth Scholarship from the Department of Mathematics at Imperial College. PT acknowledges a fellowship from The Royal Commission for the Exhibition of 1851. The authors used the computing resources of the UK Medical Bioinformatics partnership (UK MED-BIO; aggregation, integration, visualisation and analysis of large, complex data), which is supported by the UK Medical Research Council (grant no. MR/L01632X/1) and the Imperial College Research Computing Service (DOI: 10.14469/hpc/2232) for access to their HPC facilities (CX1 cluster).

References

[1] Baslan, T., and Hicks, J. (2017) Unravelling biology and shifting paradigms in cancer with single-cell sequencing, *Nature Reviews Cancer 17*, 557.

[2] Chen, X., Teichmann, S. A., and Meyer, K. B. (2018) From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture.

[3] Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities, *Nature methods 14*, 565.

[4] Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I., and Enard, W. (2018) Quantitative single-cell transcriptomics, *Briefings in functional genomics*.

[5] Bacher, R., and Kendziorski, C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments, *Genome biology 17*, 63.

[6] Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., and Marioni, J. C. (2013) Accounting for technical noise in single-cell RNA-seq experiments, *Nature methods 10*, 1093-1095.

[7] Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome biology 15*, 550.

[8] Robinson, M. D., and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data, *Genome biology 11*, R25.

[9] Robinson, M. D., and Smyth, G. K. (2007) Moderated statistical tests for assessing differences in tag abundance, *Bioinformatics 23*, 2881-2887.

[10] Lun, A. T., Bach, K., and Marioni, J. C. (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts, *Genome biology 17*, 75.

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

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[12] Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., Newton, M., and Kendziorski, C. (2017) SCnorm: robust normalization of single-cell RNA-seq data, *Nature methods 14*, 584.

[13] Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., and Prlic, M. (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data, *Genome biology 16*, 278.

[14] Pierson, E., and Yau, C. (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis, *Genome biology 16*, 241.

[15] Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014) Bayesian approach to single-cell differential expression analysis, *Nature methods 11*, 740-742.

[16] Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2018) Single cell RNA-seq denoising using a deep count autoencoder, *bioRxiv*, 300681.

[17] Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., and Zhang, N. R. (2018) SAVER: gene expression recovery for single-cell RNA sequencing, *Nature Methods*, 1.

[18] Li, W. V., and Li, J. J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data, *Nature communications 9*, 997.

[19] van Dijk, D., Nainys, J., Sharma, R., Kathail, P., Carr, A. J., Moon, K. R., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe'er, D. (2017) MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data, *BioRxiv*, 111591.

[20] Wagner, F., Yan, Y., and Yanai, I. (2018) K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data, *bioRxiv*, 217737.

[21] Lopez, R., Regier, J., Cole, M. B., Jordan, M., and Yosef, N. (2018) Bayesian Inference for a Generative Model of Transcriptome Profiles from Single-cell RNA Sequencing, *bioRxiv*, 292037.

[22] Zappia, L., Phipson, B., and Oshlack, A. (2017) Splatter: simulation of single-cell RNA sequencing data, *Genome biology 18*, 174.

[23] Ding, J., Condon, A., and Shah, S. P. (2018) Interpretable dimensionality reduction of single cell transcriptome data with deep generative models, *Nature communications 9*, 2002.

[24] Wang, D., and Gu, J. (2017) VASC: dimension reduction and visualization of single cell RNA sequencing data by deep variational autoencoder, *bioRxiv*, 199315.

[25] Grønbech, C. H., Vording, M. F., Timshel, P. N., Sønderby, C. K., Pers, T. H., and Winther, O. (2018) scVAE: Variational auto-encoders for single-cell gene expression data, *bioRxiv*, 318295.

[26] Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2017) Missing data and technical variability in single-cell RNA-sequencing experiments, *Biostatistics*.

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[27] Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018) scmap: projection of single-cell RNA-seq data across data sets, *Nature methods 15*, 359.

[28] Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors, *Nature biotechnology 36*, 421.

[29] Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species, *Nature biotechnology 36*, 411.

[30] Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells, *Cell 161*, 1187-1201.

[31] Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006) Stochastic mRNA synthesis in mammalian cells, *PLoS biology 4*, e309.

[32] Shahrezaei, V., and Swain, P. S. (2008) Analytical distributions for stochastic gene expression, *Proceedings of the National Academy of Sciences*.

[33] Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., Kim, J., Murray, J., and Raj, A. (2018) Rare cell detection by single-Cell RNA sequencing as guided by single-molecule RNA FISH, *Cell systems 6*, 171-179. e175.

[34] 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, *Scientific reports 7*, 39921.

[35] Anders, S., and Huber, W. (2012) Differential expression of RNA-Seq data at the gene level–the DESeq package, *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)*.

[36] Andrews, T. S., and Hemberg, M. (2018) Dropout-based feature selection for scRNASeq, *bioRxiv*.

[37] Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014) Quantitative single-cell RNA-seq with unique molecular identifiers, *Nature methods 11*, 163.

[38] Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2017) Missing data and technical variability in single-cell rna-sequencing experiments. bioRxiv, May.

[39] Pimentel, H., Bray, N. L., Puente, S., Melsted, P., and Pachter, L. (2017) Differential analysis of RNA-seq incorporating quantification uncertainty, *Nature methods 14*, 687.

[40] Jaakkola, M. K., Seyednasrollah, F., Mehmood, A., and Elo, L. L. (2016) Comparison of methods to detect differentially expressed genes between single-cell populations, *Briefings in bioinformatics 18*, 735-743.

[41] Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq, *Genome research 21*, 1160-1167.

bioRxiv preprint first posted online Aug. 3, 2018; doi: http://dx.doi.org/10.1101/384586. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[42] Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A., and Mikkelsen, T. S. (2014) Characterization of directed differentiation by high-throughput single-cell RNA-Seq, *BioRxiv*, 003236.

[43] Ye, C., Speed, T. P., and Salim, A. (2017) DECENT: Differential Expression with Capture Efficiency AdjustmeNT for Single-Cell RNA-seq Data, *bioRxiv*, 225177.

[44] Elo, L. L., Filén, S., Lahesmaa, R., and Aittokallio, T. (2008) Reproducibility-optimized test statistic for ranking genes in microarray studies, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 5*, 423-431.

[45] Hardcastle, T. J., and Kelly, K. A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data, *BMC bioinformatics 11*, 422.

[46] Saint, M., Bertaux, F., Tang, W., Sun, X.-M., Game, L., Köferle, A., Bähler, J., Shahrezaei, V., and Marguerat, S. (2018) Single-cell phenotyping and RNA sequencing reveal novel patterns of gene expression heterogeneity and regulation during growth and stress adaptation in a unicellular eukaryote, *bioRxiv*, 306795.

Main figures

Figure 1

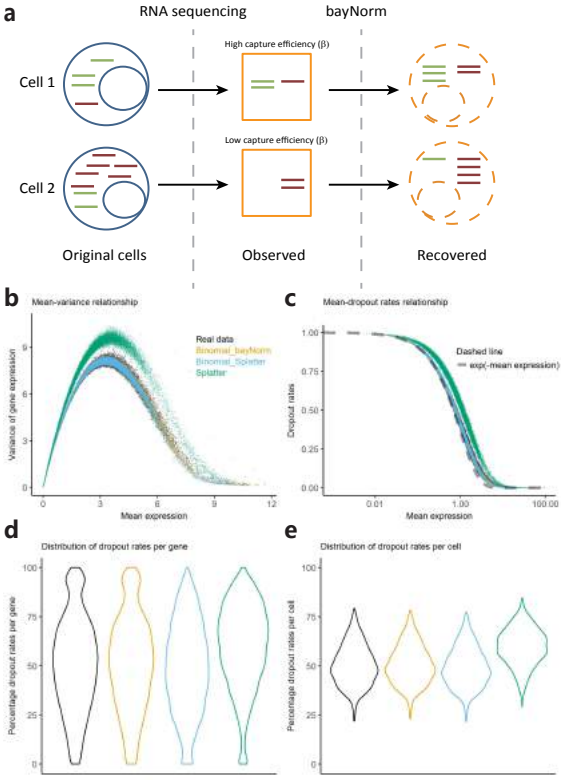


Figure 1: A binomial model of mRNA capture is consistent with the statistics of raw experimental scRNA-seq data. (a) Cartoon illustration of the bayNorm approach. Only a fraction of the total number of mRNAs present in the cell is captured during scRNA-seq library preparation. This occurs with a global probability called capture efficiency (β). Using cell-specific estimates of β, bayNorm aims at recovering the original number of mRNA of each gene present in each cell. (b)-(e) Comparisons between raw experimental scRNA-seq data from the Klein study[1] and synthetic data obtained using the Binomial_bayNorm (orange), Binomial_Splatter (blue), or Splatter[2] (green) simulation protocols (Supplementary Note 1). (b) Variance vs mean expression relationship. (c) Dropout rates vs mean expression relationship (note that Binomial_Splatter and Binomial_bayNorm are on top of each other in this panel). The dotted line shows the $e^{(-\text{Mean expression})}$ function. (d) Distribution of dropout values per gene. (e) Distribution of dropout values per cell.

Figure 2

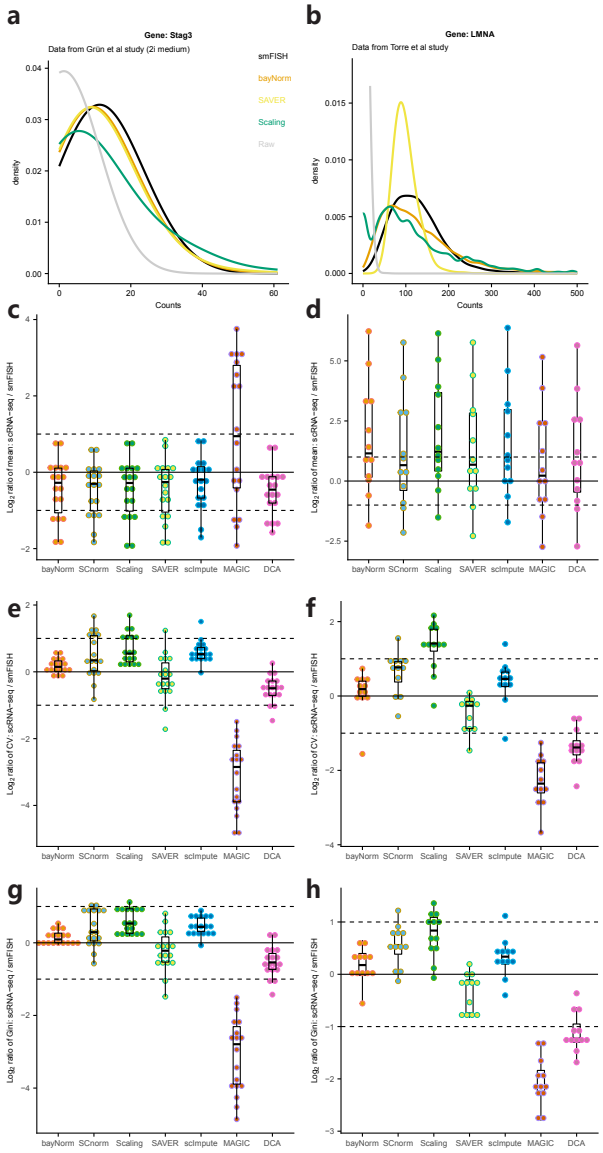


Figure 2: bayNorm recovers distributions of gene expression observed by smFISH. (a) Stag3 mRNA distribution for cells grown in 2i measured by smFISH or by scRNA-seq and normalised with different methods (from Grn study). Raw denotes unnormalised scRNA-seq data. (b) As in (a) for the LMNA gene (from Torre study). Legend as in (a). (c) Log_2 ratio between the means of scRNA-seq measurements for 18 genes normalised by different methods and their matched smFISH measurements (from Grn study). (d) As in (c) using 12 genes (Torre study). (e) Log_2 ratio between the CV of scRNA-seq measurements for 18 genes normalised by different methods and their matched smFISH measurements (from Grn study). (f) As in (e) using 12 genes (from Torre study). (g) Log_2 ratio between the Gini coefficients of scRNA-seq measurements for 18 genes normalised by different methods and their matched smFISH measurements (from Grn study). (h) As in (c) using 12 genes (from Torre study). For the bayNorm and SAVER normalised datasets, 20 or 5 samples were generated from posterior distributions for the Grn and the Torre studies, respectively. For bayNorm and SAVER, normalized counts across cells and samples are used. All normalised datasets except bayNorm and the Scaling method have been divided by the $< \beta >$ value used in bayNorm procedure. For this analysis smFish data were normalised for variation in total transcript numbers using either cell size measurements (Grn study) or expression levels of a house keeping gene (Torre study) as detailed in Supplementary note 3.

Figure 3

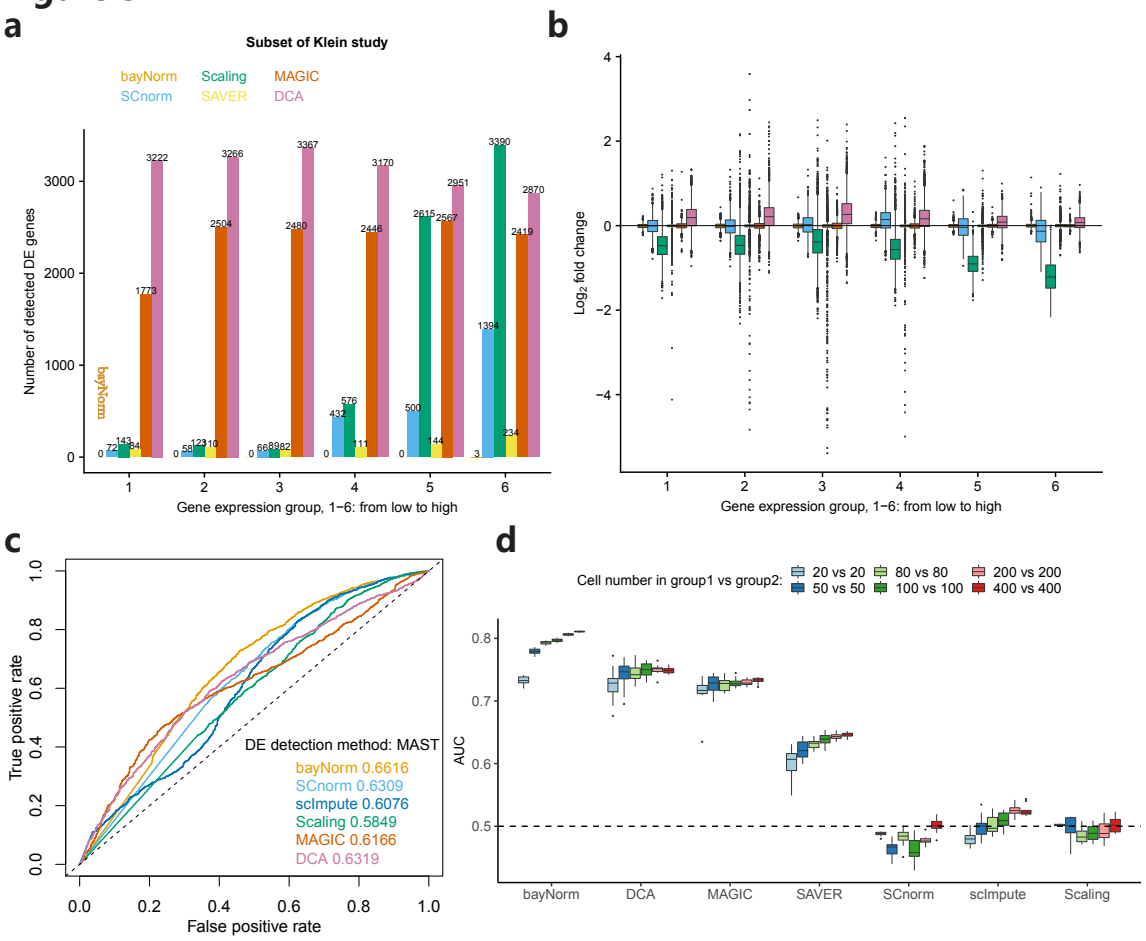


Figure 3: bayNorm enables robust and sensitive differential expression analysis. (a) Number of differentially expressed genes between the 100 cells with the highest and the 100 cells with the lowest total counts in (Klein study). DE genes were called using the MAST package ($P_{\mathrm{MAST}} < 0.05$) and plotted for 6 groups of genes with increasing mean expression (1-low 6-high). (b) Log_2 fold change from (a). (c) Differential expression analysis using MAST for different normalization methods (Islam study) using a benchmark list of DE genes obtained from matched bulk RNA-seq data[3]. (d) Differential expression analysis using data from Soumillon study[4]. Ten samples of 20, 50, 80, 100, 200 or 400 cells were selected randomly from two groups of stage-3 differentiated cells at day 0 (D3T0) or day 7 (D3T7). DE detection was performed between groups as described at the top of the figure using a list of DE genes obtained from matched bulk RNA-seq data as a benchmark (1000 genes with the largest magnitude of log fold-change between the D3T0 and D3T7 samples)[3].

Figure 4

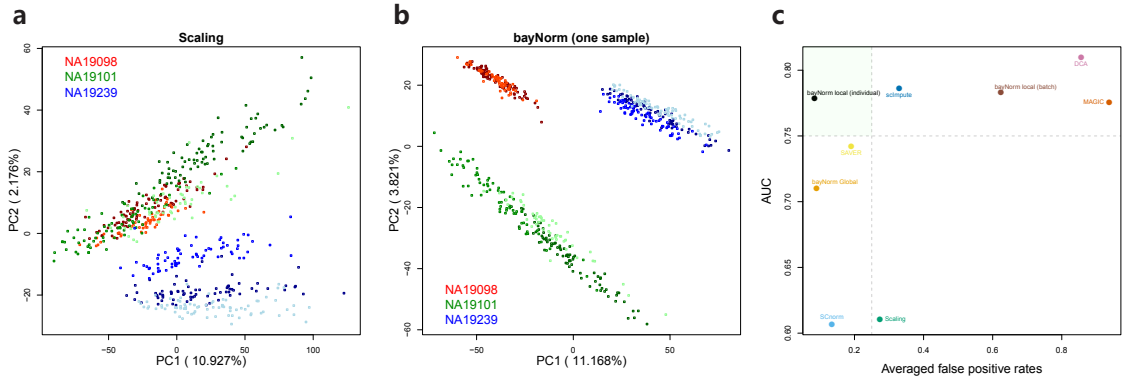


Figure 4: bayNorm normalisation reduces experimental batch effects. (a) PCA plots of data from the Tung study normalised using global scaling. Each colour represent a different cell line derived from a different individual. Colour shades represent different batches within a line/individual. (b) As in (a) using bayNorm normalization. (c) Differentially expressed genes were called between lines NA19101 and NA19239. DE genes from matched bulk RNA sequencing data were used as a benchmark set and AUC values were calculated. In parallel, DE genes were called between different batches within each line (7 pair of comparisons in total). The ratio of the number of DE genes per line and the total number of genes (13058) between batched is defined as the DE false positive rate (FDR). FDRs were averaged across the 7 pairs. Each normalisation method results in a pair of averaged FDR and AUC values that is displayed on the figure. Normalisation methods are colour-coded. The vertical and horizontal dashed lines represent 0.25 and 0.75 indicative cutoffs respectively. bayNorm was applied either across batches but within lines ("bayNorm local (individual)") or across all cells ("bayNorm global") or within each batch ("bayNorm local (batch)").

Methods

1 The Bayesian model used in bayNorm

A scRNAseq dataset is typically represented in a matrix of dimension $P \times Q$, where P denotes the total number of genes observed and Q denotes the total number of cells studied. The element x_{ij} ($i \in \{1, 2, \ldots, P\}$ and $j \in \{1, 2, \ldots, Q\}$) in the matrix represents the number of transcripts reported for the i^{th} gene in the j^{th} cell. This is equal to the total number of sequencing reads mapping to that gene in that cell for a non-UMI protocol. For UMI based protocols this is equal to the number of individual UMIs mapping to each gene[5, 6]. The matrix can include data from different groups or batches of cells, representing different biological conditions. This can be represented as a vector of labels for the cell groups or conditions (C_j).

A common approach for normalizing scRNAseq data is based on the use of a global scaling factor (s_j), ignoring any gene specific biases (for a recent review see[7]). The normalized data $\tilde{x}_{ij}$ is obtained by dividing the raw data for each cell j by the its global scaling factor s_j:

$$\tilde{x}_{ij} = \frac{x_{ij}}{s_j} \tag{1}$$

In bayNorm, we implement global scaling using a Bayesian approach. We assume given the original number of transcripts in the cell (x^0_{ij}), the number of transcripts observed (x_{ij}) follows a Binomial model with probability β_j [1], which we refer to as capture effeiciency and it represents the probability of original transcripts in the cell to be observed. In addition, we assume that the original number or true count of the i^{th} gene in the j^{th} cell (x^0_{ij}) follows Negative Binomial distribution with parameters mean (μ), size (or dispersion parameter, ϕ), such that:

$$\Pr(x^0_{ij} = n|\phi_i, \mu_i) = \frac{\Gamma(n + \phi_i)}{\Gamma(\phi_i)n!}(\frac{\phi_i}{\mu_i + \phi_i})^{\phi_i}(\frac{\mu_i}{\mu_i + \phi_i})^n$$

So, overall we have the following model:

$$\begin{aligned} x_{ij} &\sim \text{Binom}(x^0_{ij}, \text{prob} = \beta_j) \\ x^0_{ij} &\sim \text{NB}(\text{mean} = \mu_i, \text{size} = \phi_i) \end{aligned} \tag{2}$$

Using the Bayes rule, we have the following posterior distribution of original number of mRNAs for each gene in each cell:

$$\underbrace{\Pr(x^0_{ij}|x_{ij}, \beta_j, \mu_i, \phi_i)}_{\text{Posterior}} = \frac{\overbrace{\Pr(x_{ij}|x^0_{ij}, \beta_j)}^{\text{Likelihood}} \times \overbrace{\Pr(x^0_{ij}|\mu_i, \phi_i)}^{\text{Prior}}}{\underbrace{\Pr(x_{ij}|\mu_i, \phi_i, \beta_j)}_{\text{Marginal likelihood}}} \tag{3}$$

The prior parameters μ and ϕ of each gene were estimated using an empirical Bayesian method as discussed in detail in Section 4 below.

The marginal distribution for gene i in cell j is

$$\Pr(x_{ij}|\mu_i,\phi_i,\beta_j) = \sum_{n=0}^{+\infty} \underbrace{\binom{n}{x_{ij}}\beta_j^{x_{ij}}(1-\beta_j)^{n-x_{ij}}}_{\text{Binomial}} \underbrace{\binom{n+\phi_i-1}{\phi_i-1}(\frac{\phi_i}{\mu_i+\phi_i})^{\phi_i}(\frac{\mu_i}{\mu_i+\phi_i})^{n}}_{\text{Negative Binomial}} \tag{4}$$

$$= \binom{x_{ij}+\phi_i-1}{\phi_i-1}\left(\frac{\phi_i}{\mu_i\beta_j+\phi_i}\right)^{\phi_i}\left(\frac{\mu_i\beta_j}{\mu_i\beta_j+\phi_i}\right)^{x_{ij}}, \tag{5}$$

which follows from using

$$\binom{n+\phi_i-1}{\phi_i-1}\binom{n}{x_{ij}} = \binom{x_{ij}+\phi_i-1}{\phi_i-1}\binom{n+\phi_i-1}{n-x_{ij}}, \tag{6}$$

and

$$\sum_{n=x_{ij}}^{+\infty} z^n\binom{\phi_i+n-1}{n-x_{ij}} = \sum_{m=0}^{+\infty} z^{m+x_{ij}}\binom{\phi_i+m+x_{ij}-1}{m} = \frac{z^{x_{ij}}}{(1-z)^{\phi_i+x_{ij}}}, \tag{7}$$

with $z = \frac{\mu_i}{\mu_i+\phi_i}(1-\beta_j)$ in Eq. (4). Hence we have that the number of transcripts reported for the i^{th} gene in the j^{th} cell

$$x_{ij} \sim \text{NB}(\text{mean} = \mu_i\beta_j, \text{size} = \phi_i), \tag{8}$$

has a Negative Binomial distribution with mean $\mu_i\beta_j$ and size ϕ_i.

It can also be shown that the posterior distribution of x_{ij}^0 is a shifted Negative Binomial distribution. To sample from the posterior distribution, we note that the original count can be expressed as

$$x_{ij}^0 = x_{ij} + \zeta_{ij}, \tag{9}$$

where ζ_{ij} is the *lost* count satisfying

$$\zeta_{ij} \sim \text{NB}(\text{mean} = \frac{\mu_i(1-\beta_j)(x_{ij}+\phi_i)}{\mu_i\beta_j+\phi_i}, \text{size} = x_{ij}+\phi_i). \tag{10}$$

The posterior mean and variance then evaluate to

$$\text{E}[x_{ij}^0] = x_{ij}\frac{\mu_i+\phi_i}{\mu_i\beta_j+\phi_i} + \mu_i\frac{\phi_i-\phi_i\beta_j}{\mu_i\beta+\phi_i} \tag{11}$$

$$\text{Var}[x_{ij}^0] = \frac{(x_{ij}+\phi_i)\mu_i(1-\beta_j)(\mu_i+\phi_i)}{(\phi_i+\mu_i\beta_j)^2} \tag{12}$$

Note that when ϕ_i is small, the mean of posterior tends to $\frac{x_{ij}}{\beta_j}$. After estimating the posterior distribution for each gene in each cell, we can either sample a certain number of draws from it (3D array output, see Supplementary Figure S1) or extract the mean or mode of posterior [8] as $\tilde{x}$ (2D array output, see Supplementary Figure S1).

2 Binomial distribution and dropout probability

The binomial model of capture in scRNA-seq predicts the dropout rate for a particular gene:

$$\Pr(x_{ij}=0|x_{ij}^0,\beta_j) = (1-\beta_j)^{x_{ij}^0},$$

in a given cell j. Across a group of non-homogeneous cells, we may approximate this expression by

$$(1-\bar{\beta})^{(\bar{x}/\bar{\beta})}$$

For small $\bar{\beta}$ this expression tends to $\Pr(x = 0) = \exp(-\bar{x})$. In dropout vs mean expression (dropout-mean) (Figure 1c, Sup Figures S2c, S3c, S4c, S5c, S6c and S7c), the line "$\exp(-\bar{x})$" follows the lower limit of the trend. We note that a Poisson model of RNA-seq that is used by several authors also predicts dropout rates to be $\Pr(x = 0) = \lambda^0/0! \exp(-\lambda) = \exp(-\lambda)$, where $\lambda = \bar{x}$[9, 2, 10].

To further show that Binomial distribution can capture the relationship between dropout rates and mean expression, we simulated data based on real experimental data[1, 11, 12] by adapting simulation protocols proposed in the R package Splatter[2]. The details about the simulation procedure can be found in the supplementary note 1. The resulting dropout-mean plot of simulated data based on Binomial model is very close to that of the real scRNA-seq data for UMI-based protocols. As shown in the Supplementary Figures S2c, S3c, S4c, S5c and S6c, the dropout-mean trend of UMI data is close to the asymptotic line "$\exp(-\bar{x})$" (Binomial_Splatter and Binomial_bayNorm simulated data perform similar to each other and the real experimental data). data based on real experimental data[1, 11, 12] as discussed in the results and supplementary note 1. The resulting dropout-mean plot of simulated data based on Binomial model is very close to that of the real scRNA-seq data for UMI-based protocols.

3 Estimation of capture efficiencies

Cell specific capture efficiency β_j and global scaling factor (s_j) are closely related. We can transform scaling factors estimated by different methods (see below) into β_j values with the following formula:

$$\beta_j = (s_j/\bar{s})\bar{\beta} \tag{13}$$

$\bar{\beta}$, a scalar, is an estimate of global mean capture efficiency across all cells, which ranges between 0 and 1.

There are two different methods for estimating $\bar{\beta}$ and β_j:

1. If spike-ins or smFISH data are available they can be used to estimate capture efficiencies. We can either divide the total number of observed spik-ins in each cell by the total number of input spike-ins, or we can fit a linear regression[1] to estimate the cell specific β_j. If smFISH data is available, we can fit a linear regression between the mean expression of raw data (response variable) and the mean expression of the smFISH data (explanatory variable). The coefficient of the explanatory variable can be used as $\bar{\beta}$[13].

2. The raw data itself can be directly used for estimation of cell specific global scaling factors (s_j). Then equation 13 and an estimate of $\bar{\beta}$ can be used to estimate β_j. There are different methods available for estimation of global scaling factors. Some were developed for bulk RNA-seq data[14, 15] and some are specific to scRNA-seq data[16, 17]. The value of $\bar{\beta}$ depends on the protocol used and can be batch dependent. For example, for Droplet based protocol, it is about 0.06[1] or 0.12[18]. $\bar{\beta}$ can also be estimated by spike-ins or smFISH data as explained above.

We finally note, that estimates of capture efficiency discussed above will assume cells have simular original transcript content. Therefore, the bayNorm outputs estimates of original transcript counts for a typical cell, which is corrected for variation in cell size and transcript content. This is usually desirable for down-stream analysis such as DE detection. However, if one is interested in absolute origianl count and has additional information such as cell size or total transcirpt content per cell, the capture efficiencies can be approporiatly rescaled for this purpose.

4 Estimation of prior parameters

4.1 Maximisation of marginal distribution

Using an emperical bayes approach, one can use the maximisation of marginal likelihood distribution of the observed counts across cells to estimate prior parameters [19]. Let M_i denotes the marginal likelihood function for the i^{th} gene across cells. Assuming independence between cells, the log-marginal distribution for the i^{th} gene is

$$\log M_i = \sum_{j=1}^{Q} \log \Pr(x_{ij}|\mu_i, \phi_i, \beta_j), \tag{14}$$

where $\Pr(x_{ij}|\mu_i, \phi_i, \beta_j)$ is the Negative Binomial in Eq. (5). Maximizing of Eq. (14) yields the pair (μ_i, ϕ_i).

The above optimization needs to be done for each of the P genes. We refer to the ϕ and/or μ estimated by maximizing marginal distribution as BB estimates for convenience, because bayNorm utilizes spectral projected gradient method (spg) from the R package named "BB". Optimizing the marginal distribution with respect to both μ and ϕ (2D optimization) is computationally intensive. If we had a good estimate μ, then we could optimize the marginal distribution with respect to ϕ alone, which would be much more efficient.

4.2 Method of Moments

A heuristic way of estimating μ_i and ϕ_i is through a variant of the Method of Moments. The first step is to do a simple normalization of the raw data, to scale expressions given the cell specific capture efficiencies (β_j). The simple normalized count x^s_{ij} is calculated as following:

$$x^s_{ij} = x_{ij} \frac{\langle \sum_{i=1}^{P} x_{ij}/\beta_j \rangle_j}{\sum_{i=1}^{P} x_{ij}}, \tag{15}$$

where the numerator of the scaling factor of x_{ij} is obtained by taking the average of scaled total counts across cells.

Based on simple normalized data, we are able to estimate prior parameters μ and ϕ of the Negative Binomial distribution using the Method of Moments Estimation (MME), which simply equates the theoretical and empirical moments. This estimation method is fast and simulations suggests it provides good estimates of μ but the drawback is that the estimation of ϕ show a systematic bias (see Supplementary Figure S24 a-b).

4.3 The combined method

Based on simulation studies (Supplementary Figure S24), the most robust and efficient estimation of μ and ϕ can be obtained using the following combined approach, which is the default setting in bayNorm:

1. Based on simple normalized data, we use the MME method for each gene to obtain MME estimated μ and ϕ.

2. Although the BB estimated ϕ is much closer to the true ϕ, many estimates are at the upper boundary of the search space (Supplementary Figures S24 c-d). So, we find adjusting the MME estimated ϕ by a factor which can be estimated by fitting a linear regression between MME estimated ϕ and BB estimated ϕ works best (Supplementary Figures S24 c-d). This adjusted MME estimated ϕ together with the MME estimated μ and estimates of β_j can be used in approximating posterior distribution for each gene in each cell.

Cells are grouped together for prior estimation, based on cell-specific attributes (C_j). Prior estimation can be done over all cells irrespective of the experimental condition. We refer to this procedure as "global". Alternatively, suppose that there are multiple groups of cells in the datasets and we have reasons to believe each group could behave differently. Then we can estimate the prior parameters "μ and ϕ" within each group respectively (within groups with the same C_j value). We refer to this procedure as "local". Estimating prior parameters across a certain group of cells based on "global" procedure allow for removing potential batch effects. Multiple groups normalization based on "local" procedure allows for amplifying the inter-groups' differences while mitigating the intra-group's variability, which is suitable for DE detection.

5 Code availability

The R package bayNorm is available at `https://github.com/WT215/bayNorm`.

The codes for producing figures in the paper are provided at `https://github.com/WT215/bayNorm_papercode`.

In the Bacher study, the code for running MAST and log fold change calculation was kindly provided by Rhonda Bacher, the author of SCnorm[20].

In the Torre study, the code for transforming counts per million normalized data to UMI data was kindly provided by Mo Huang, the author of SAVER[9].

References

[1] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[2] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

[3] Chengzhong Ye, Terence P Speed, and Agus Salim. Decent: Differential expression with capture efficiency adjustment for single-cell rna-seq data. *bioRxiv*, page 225177, 2017.

[4] Magali Soumillon, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, and Tarjei S Mikkelsen. Characterization of directed differentiation by high-throughput single-cell rna-seq. *BioRxiv*, page 003236, 2014.

[5] Tom Smith, Andreas Heger, and Ian Sudbery. Umi-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome research*, 27(3):491–499, 2017.

[6] Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann. zumis: A fast and flexible pipeline to process rna sequencing data with umis. *bioRxiv*, page 153940, 2017.

[7] 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, 2017.

[8] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[9] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature Methods*, page 1, 2018.

[10] Florian Wagner, Yun Yan, and Itai Yanai. K-nearest neighbor smoothing for high-throughput single-cell rna-seq data. *bioRxiv*, page 217737, 2018.

[11] Eduardo Torre, Hannah Dueck, Sydney Shaffer, Janko Gospocic, Rohit Gupte, Roberto Bonasio, Junhyong Kim, John Murray, and Arjun Raj. Rare cell detection by single-cell rna sequencing as guided by single-molecule rna fish. *Cell systems*, 6(2):171–179, 2018.

[12] Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7:39921, 2017.

[13] Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637–640, 2014.

[14] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol.*, 15(12):550, 2014.

[15] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):R25, 2010.

[16] Aaron TL 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.

[17] Catalina A Vallejos, John C Marioni, and Sylvia Richardson. Basics: Bayesian analysis of single-cell sequencing data. *PLoS computational biology*, 11(6):e1004333, 2015.

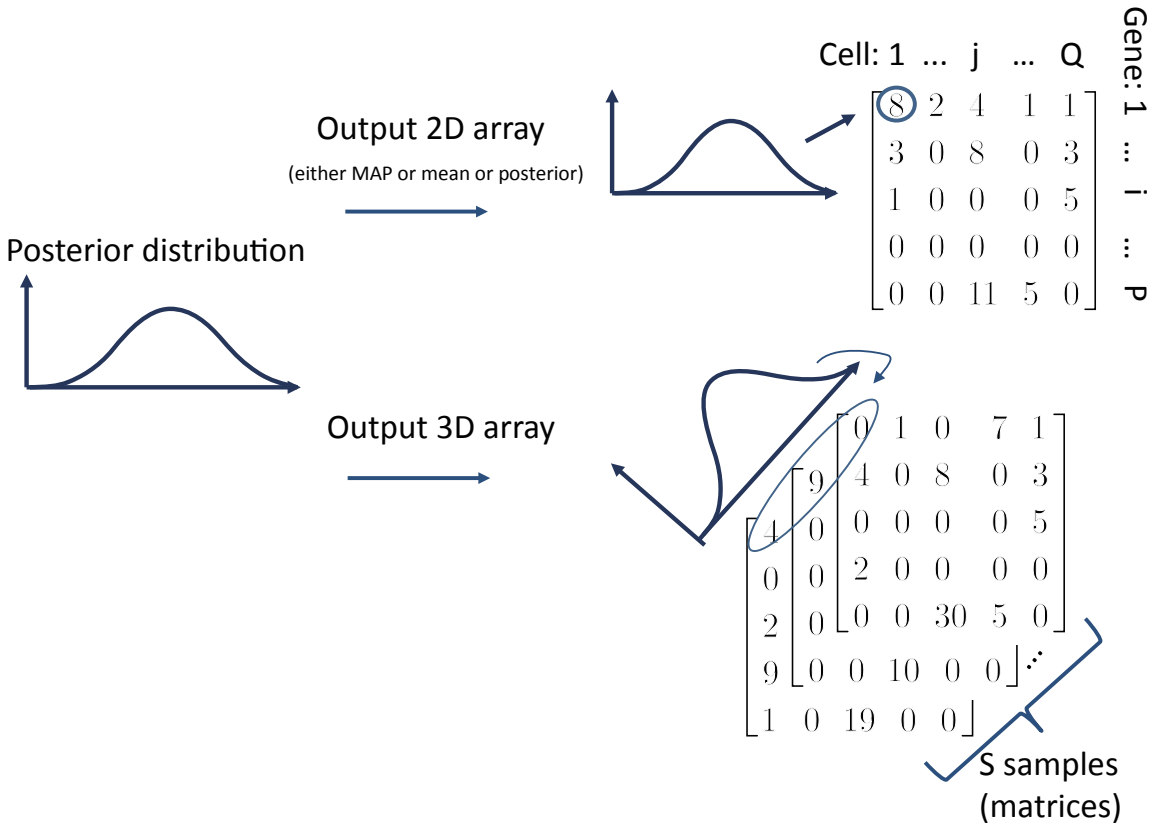
[18] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161 (5):1202–1214, 2015.

[19] Bradley P Carlin and Thomas A Louis. *Bayesian methods for data analysis*. CRC Press, 2008.

[20] Rhonda Bacher, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski. Scnorm: robust normalization of single-cell rna-seq data. *Nature methods*, 14(6):584, 2017.

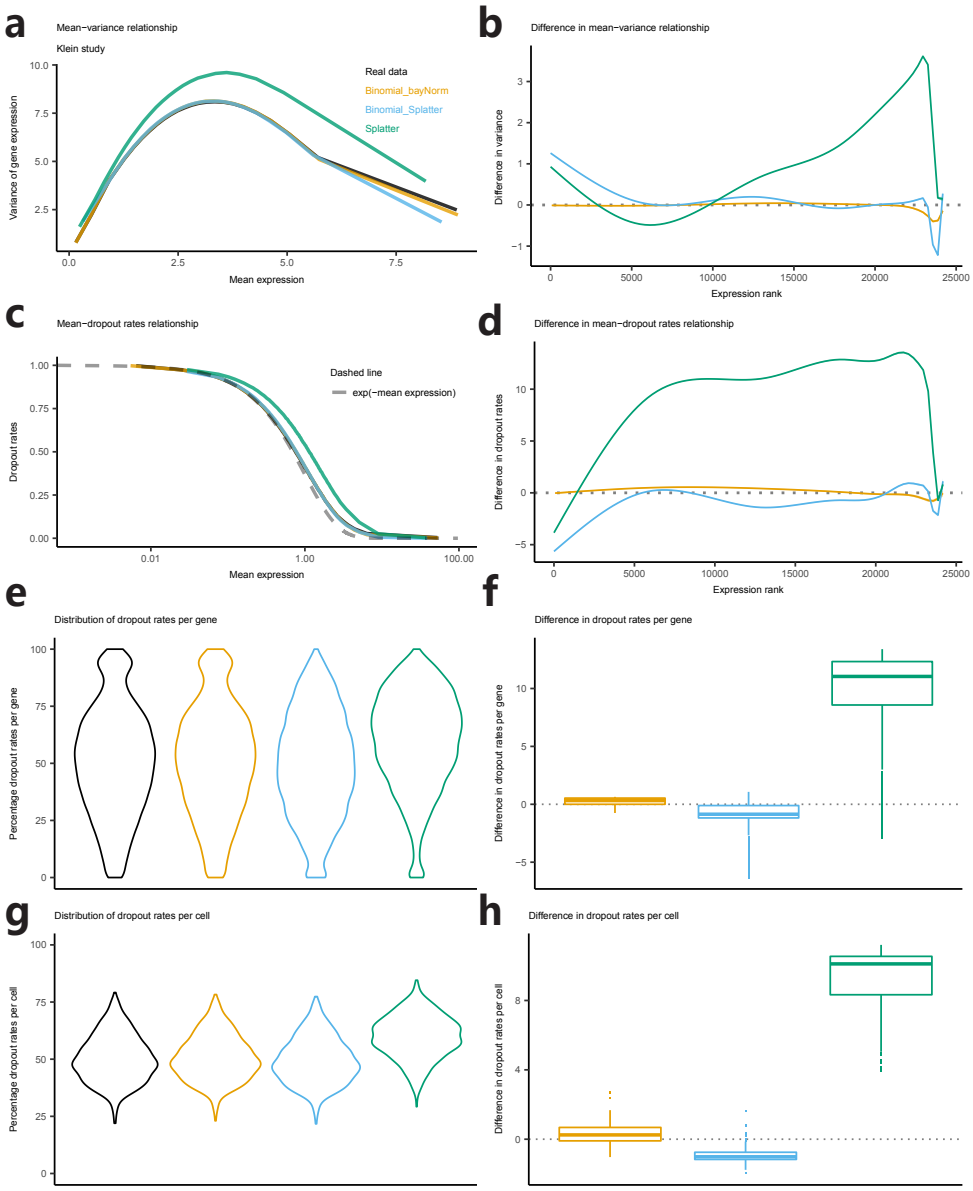
Supplementary figures

Figure S1



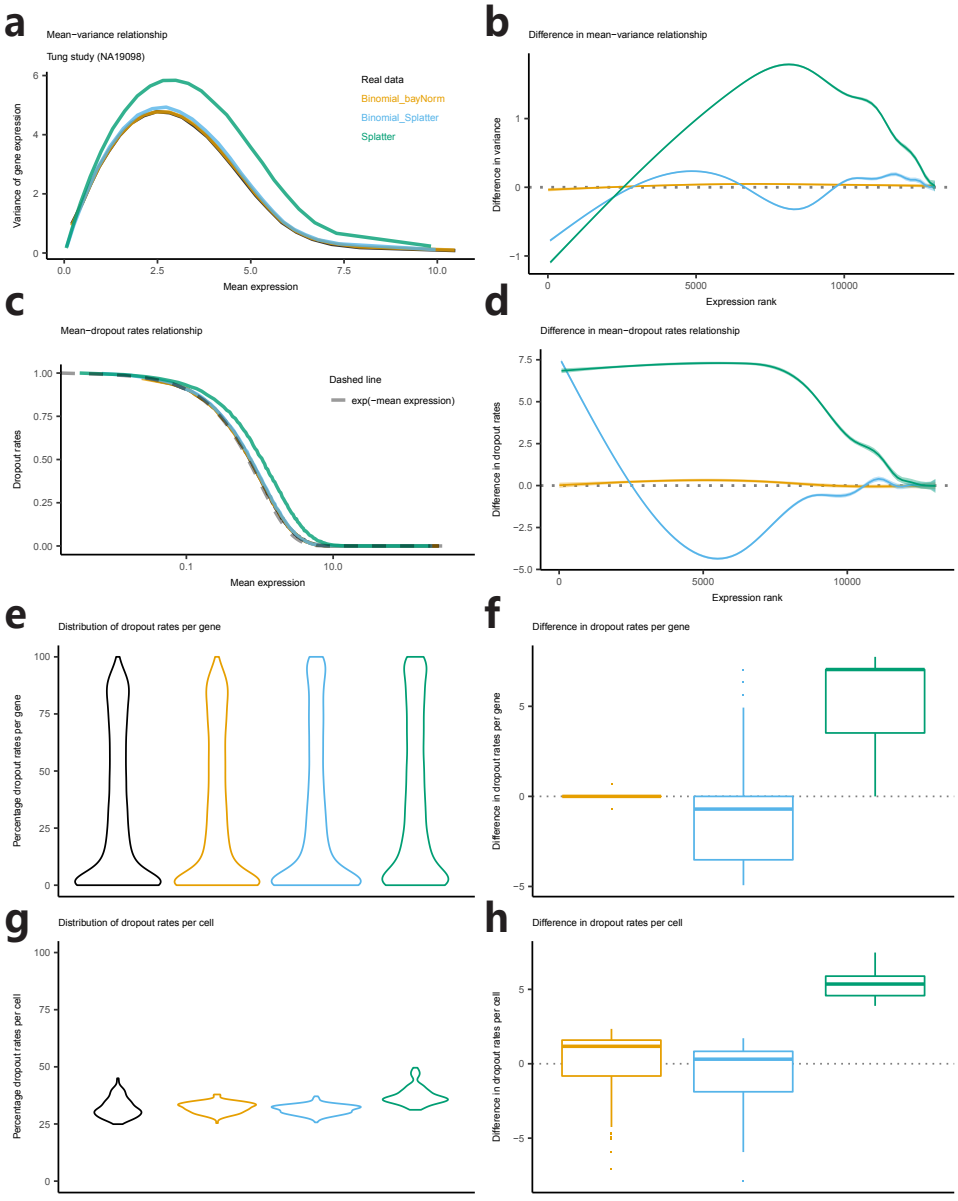
Supplementary Figure 1: Output of bayNorm. For each gene in each cell, we have a posterior distribution as bayNorm is a Baysian method (See methods). Final bayNorm output is either S samples randomly sampled from the posterior distributions (3D arrays), or the mode or mean of the posterior used as point estimates (2D arrays).

Figure S2



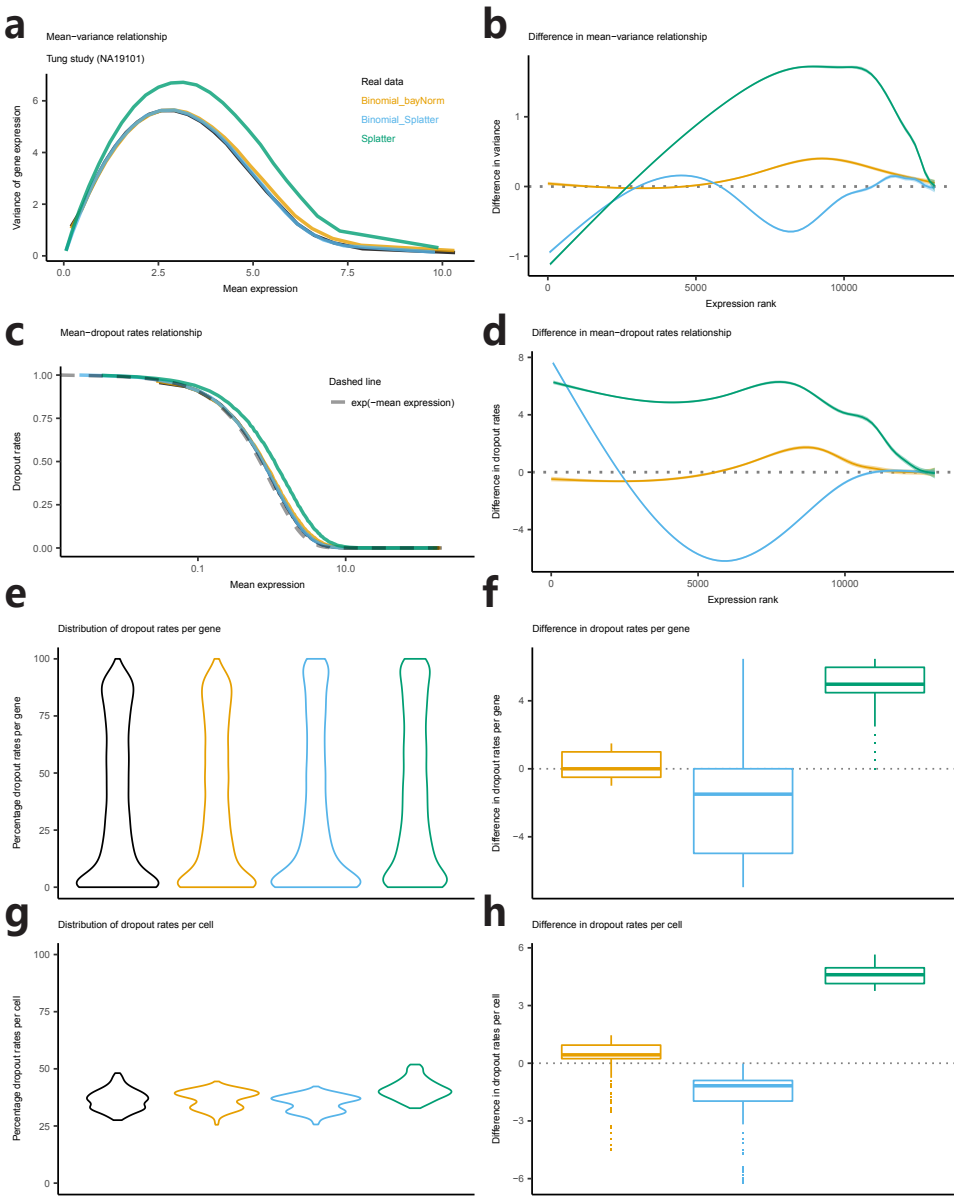
Supplementary Figure 2: Simulation analysis based on the Klein study. Comparison between simulated data and experimental data in terms of: (a-b) variance-mean relationship, (c-d) dropout rates-mean relationship, (e-f) distribution of proportions of zeros per gene, (g-h) distribution of proportions zeros per cell. (b), (d), (f) and (h): ranked difference between statistics of experimental data and simulated data for (a), (c), (e) and (g) respectively. The smoothed lines in (a) and (c) were obtained by binning x values and calculating the mean of y values in each bin[1]. The smoothed lines in (b) and (d) were generated by ggplot2 (geom_smooth with method set to "auto").

Figure S3



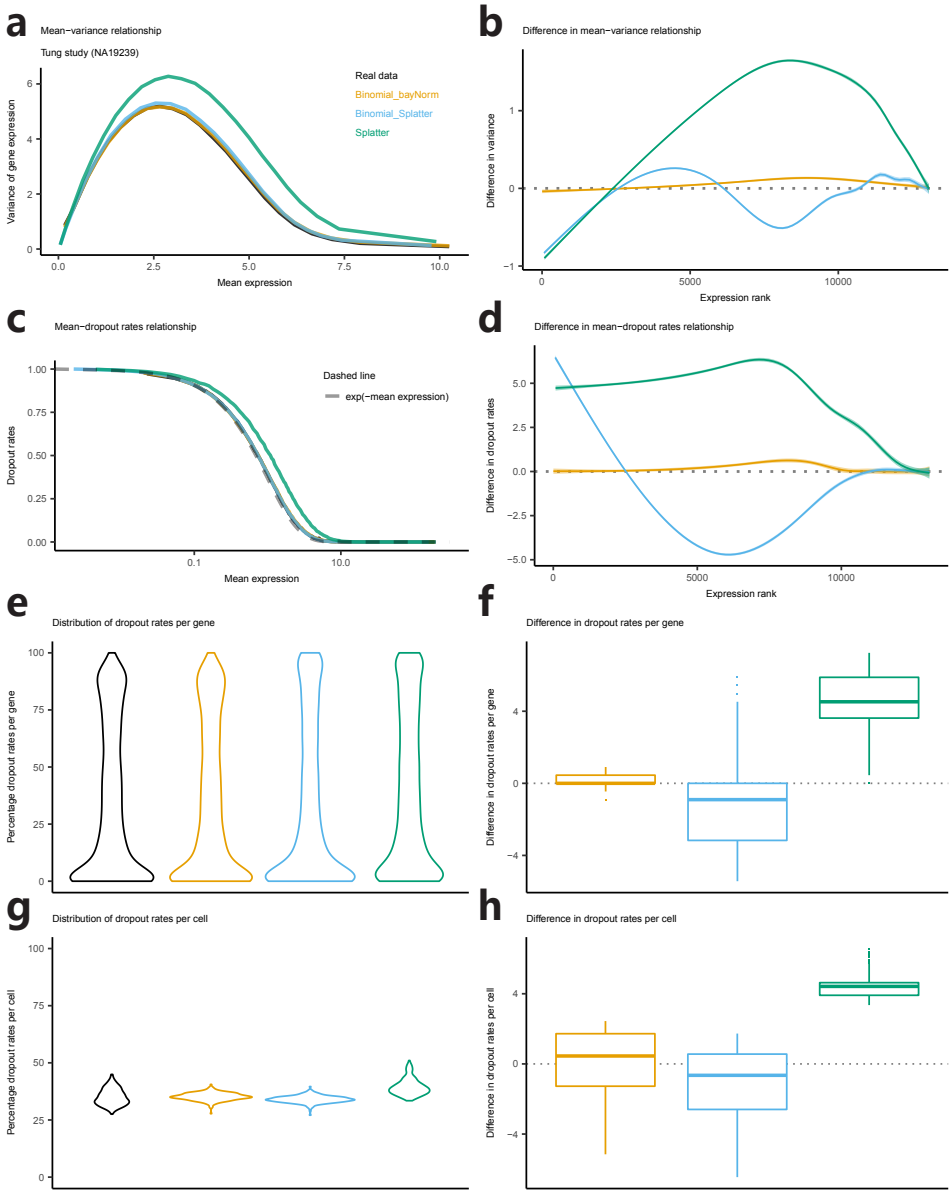
Supplementary Figure 3: Simulation analysis based on the Tung study (Individual NA19098). Comparison between simulated data and real data in terms of: (a-b) variance-mean relationship, (c-d) dropout rates-mean relationship, (e-f) distribution of proportions of zeros per gene, (g-h) distribution of proportions zeros per cell. (b), (d), (f) and (h): ranked difference between statistics of experimental data and simulated data for (a), (c), (e) and (g) respectively. The smoothed lines in (a) and (c) were obtained by binning x values and calculating the mean of y values in each bin[1]. The smoothed lines in (b) and (d) were generated by ggplot2 (geom_smooth with method set to "auto").

Figure S4



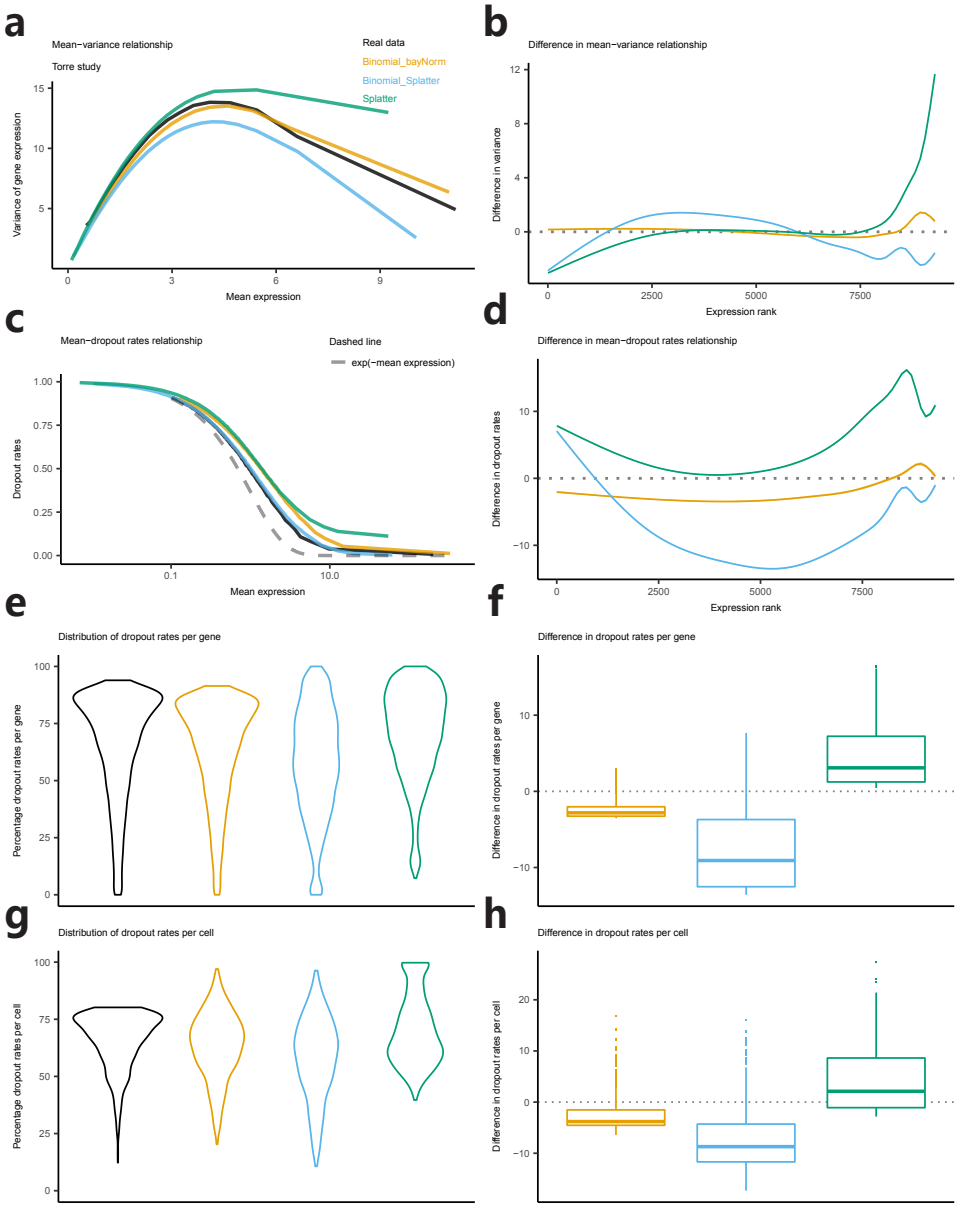
Supplementary Figure 4: Simulation analysis based on the Tung study (Individual NA19101). Comparison between simulated data and real data in terms of: (a-b) variance-mean relationship, (c-d) dropout rates-mean relationship, (e-f) distribution of proportions of zeros per gene, (g-h) distribution of proportions zeros per cell. (b), (d), (f) and (h): ranked difference between statistics of experimental data and simulated data for (a), (c), (e) and (g) respectively. The smoothed lines in (a) and (c) were obtained by binning x values and calculating the mean of y values in each bin[1]. The smoothed lines in (b) and (d) were generated by ggplot2 (geom_smooth with method set to "auto").

Figure S5



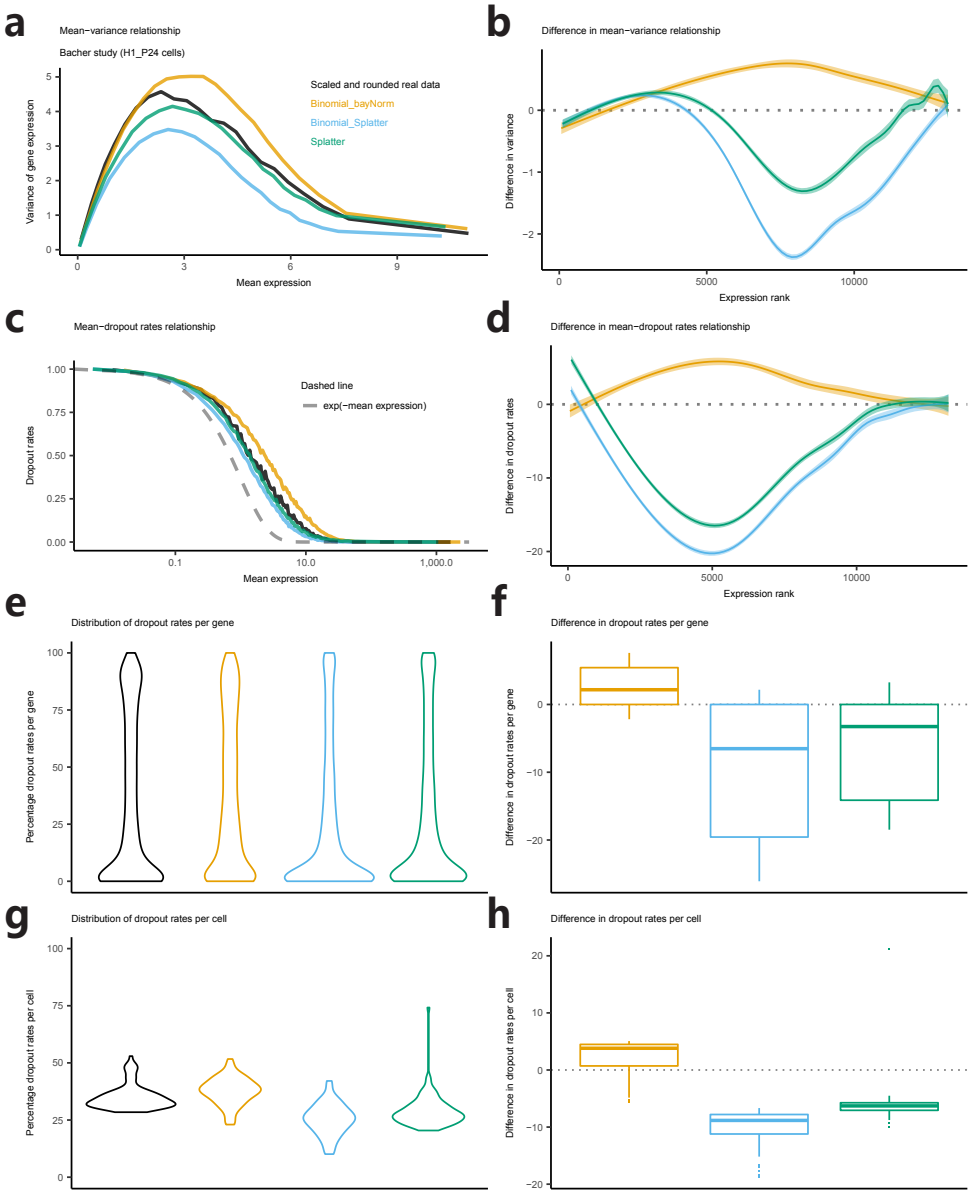
Supplementary Figure 5: Simulation analysis based on the Tung study (Individual NA19239). Comparison between simulated data and real data in terms of: (a-b) variance-mean relationship, (c-d) dropout rates-Mean relationship, (e-f) distribution of proportions of zeros per gene, (g-h) distribution of proportions zeros per cell. (b), (d), (f) and (h): ranked difference between statistics of experimental data and simulated data for (a), (c), (e) and (g) respectively. The smoothed lines in (a) and (c) were obtained by binning x values and calculating the mean of y values in each bin[1]. The smoothed lines in (b) and (d) were generated by ggplot2 (geom_smooth with method set to "auto").

Figure S6



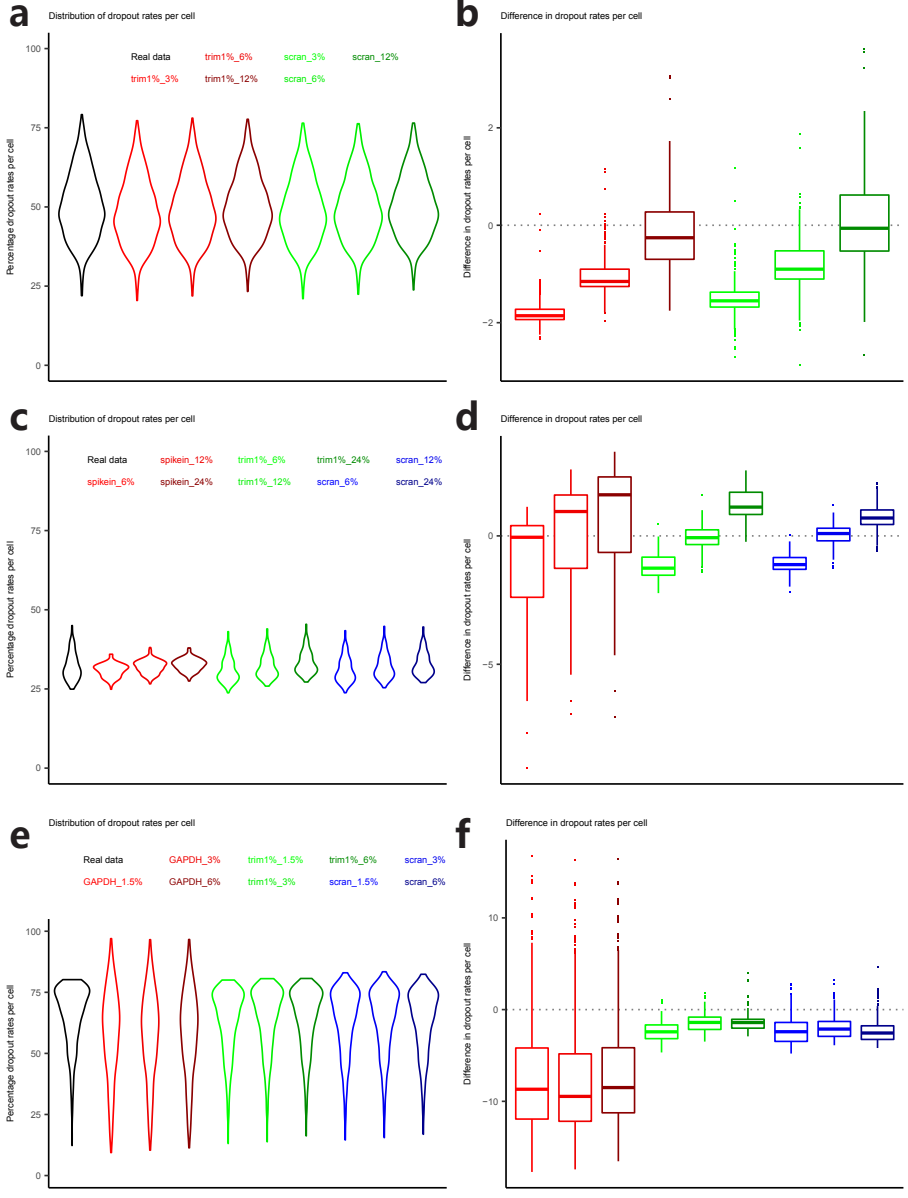
Supplementary Figure 6: Simulation analysis based on the Torre study. Comparison between simulated data and real data in terms of: (a-b) variance-mean relationship, (c-d) dropout rates-mean relationship, (e-f) distribution of proportions of zeros per gene, (g-h) distribution of proportions zeros per cell. (b), (d), (f) and (h): ranked difference between statistics of experimental data and simulated data for (a), (c), (e) and (g) respectively. The smoothed lines in (a) and (c) were obtained by binning x values and calculating the mean of y values in each bin[1]. The smoothed lines in (b) and (d) were generated by ggplot2 (geom_smooth with method set to "auto").

Figure S7



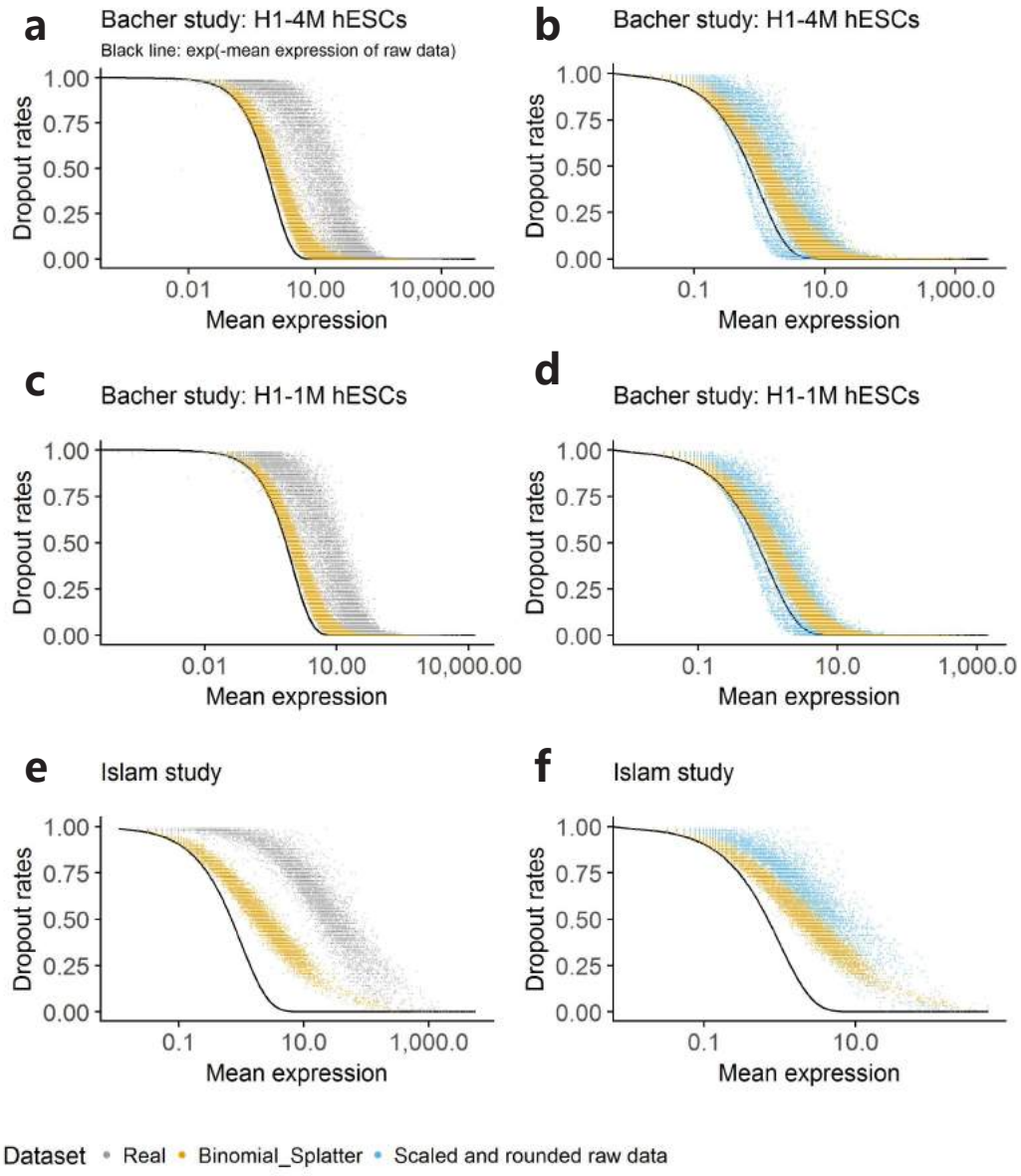
Supplementary Figure 7: SSimulation analysis based on H1_P24 cells in the Bacher study. Comparison between simulated data and real data in terms of: (a-b) variance-mean relationship, (c-d) dropout rates-mean relationship, (e-f) distribution of proportions of zeros per gene, (g-h) distribution of proportions zeros per cell. (b), (d), (f) and (h): ranked difference between statistics of experimental data and simulated data for (a), (c), (e) and (g) respectively. The smoothed lines in (a) and (c) were obtained by binning x values and calculating the mean of y values in each bin33. The smoothed lines in (b) and (d) were generated by ggplot2 (geom_smooth with method set to "auto"). Experimental data were scaled by 20 and rounded before being used as input of the three simulation protocols.

Figure S8



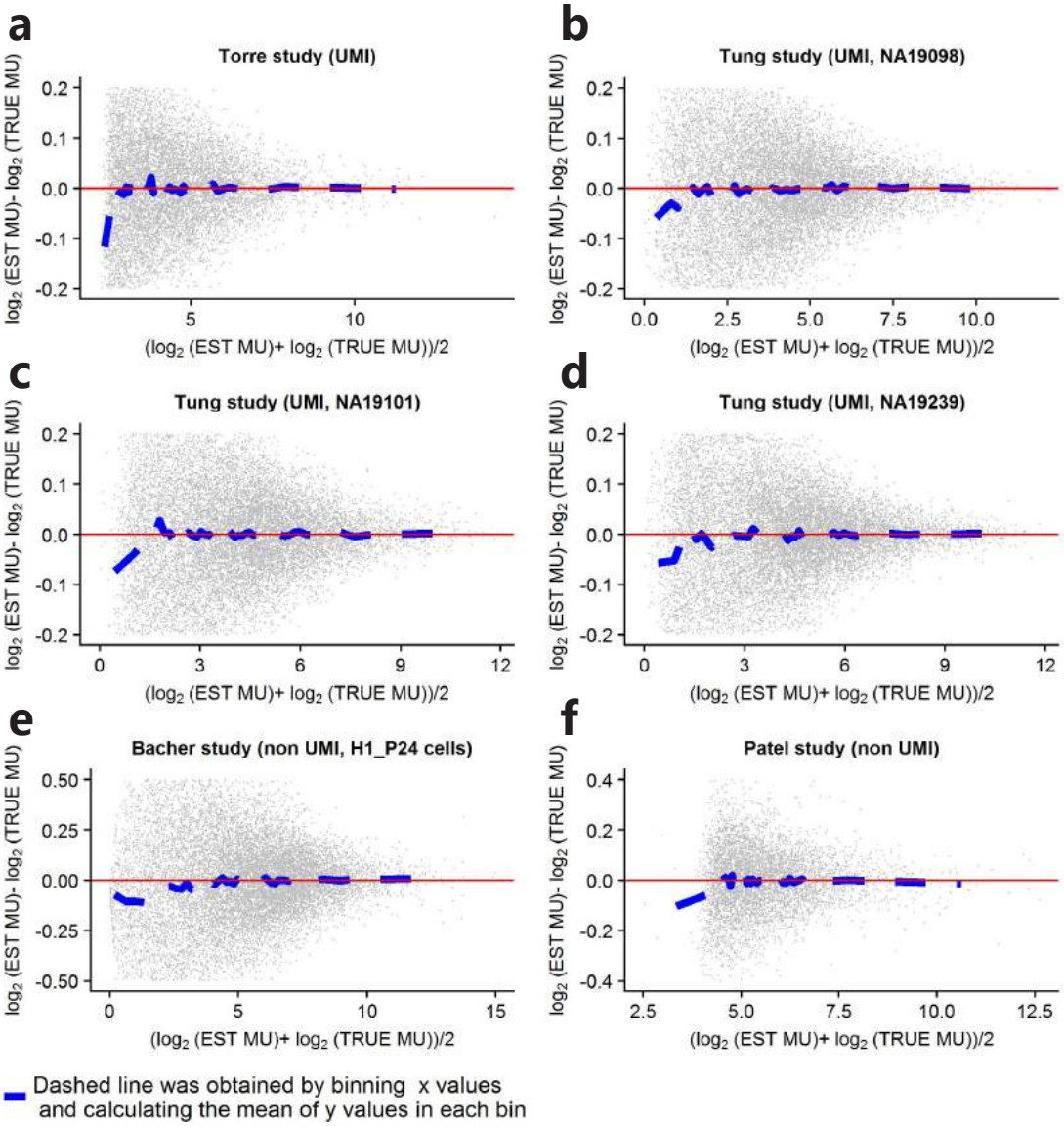
Supplementary Figure 8: Impact of different mean capture efficiencies and different size factor estimates on the Binomial_Splatter simulation protocol. Results are based on (a-b) the Klein study. (c-d) the NA19098 sample from the Tung study. (e-f) the Torre study. Scaling factors were estimated using different methods: (1) trim1%: 1% of counts were trimmed from each end of the counts in a specific cell before computing the mean. (2): "scran": scaling factors were estimated with the R package scran[2]. (3) "spikein": total counts of observed spike-ins in each cell were used as scaling factors. (4) "GAPDH": the expression of the housekeeping gene GAPDH was used as scaling factors. The percentage at the end of each label indicates the mean capture efficiency $< \beta >$ (see Methods).

Figure S9



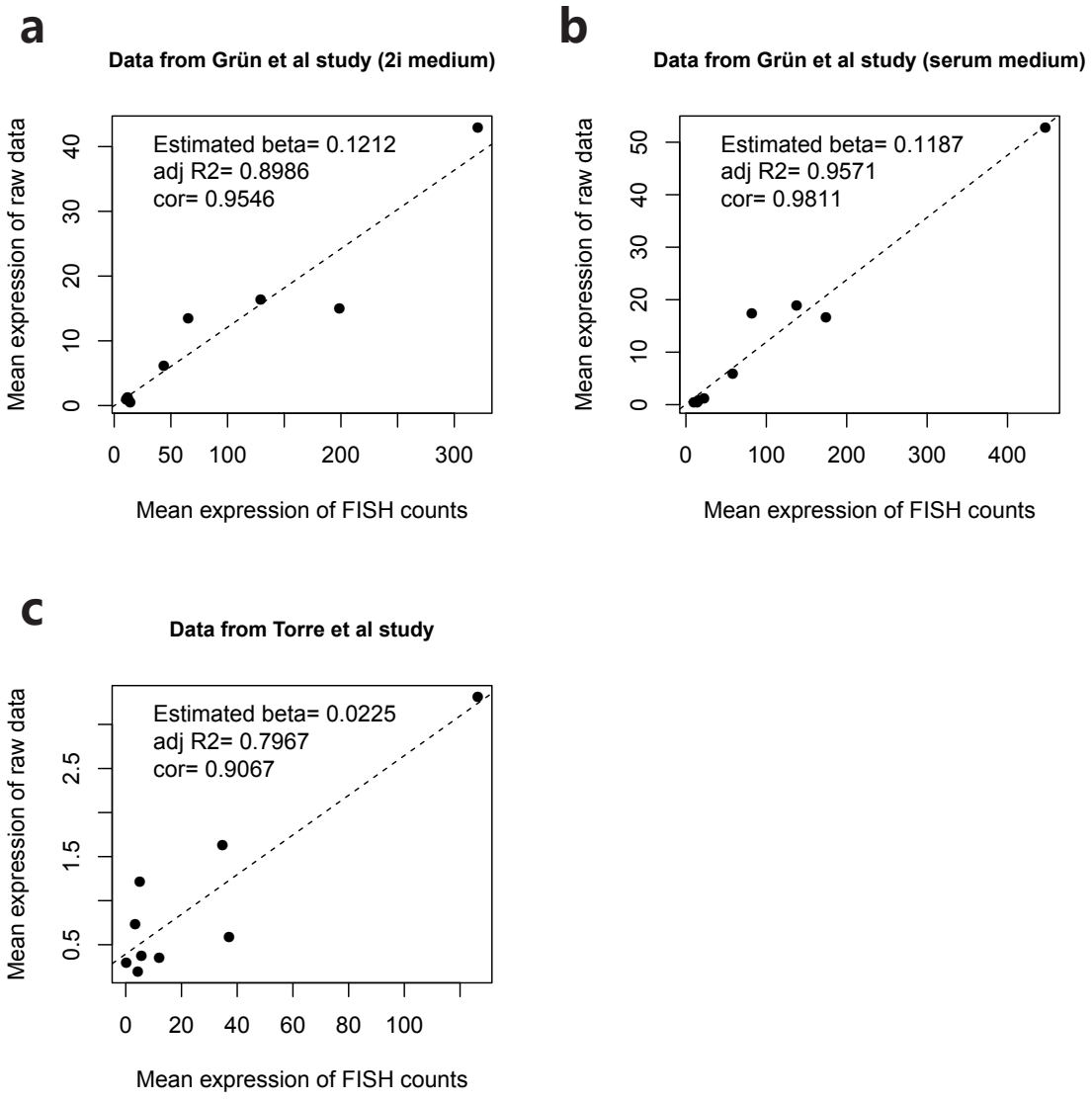
Supplementary Figure 9: Comparison between simulated data and raw experimental data in terms of the relationship between dropout rates and mean expression. (a-b) non-UMI data from the Bacher study (H1 hESCs from the 4 million mapped reads group). In (b) raw experimental data were divided by 20 and rounded. (c-d) non-UMI data from the Bacher study (H1 hESCs from the 1 million mapped reads group). In (d) raw experimental data were divided by 10 and rounded. (e-f) non-UMI data from the Islam study. In (f) raw experimental data were divided by 10 and rounded. (b), (d) and (f) are comparisons between Binomial_Splatter simulated data and scaled and rounded real experimental data. Parameters in Binomial_Splatter simulations were generated from scaled raw data as illustrated in the Supplementary Note 4.

Figure S10



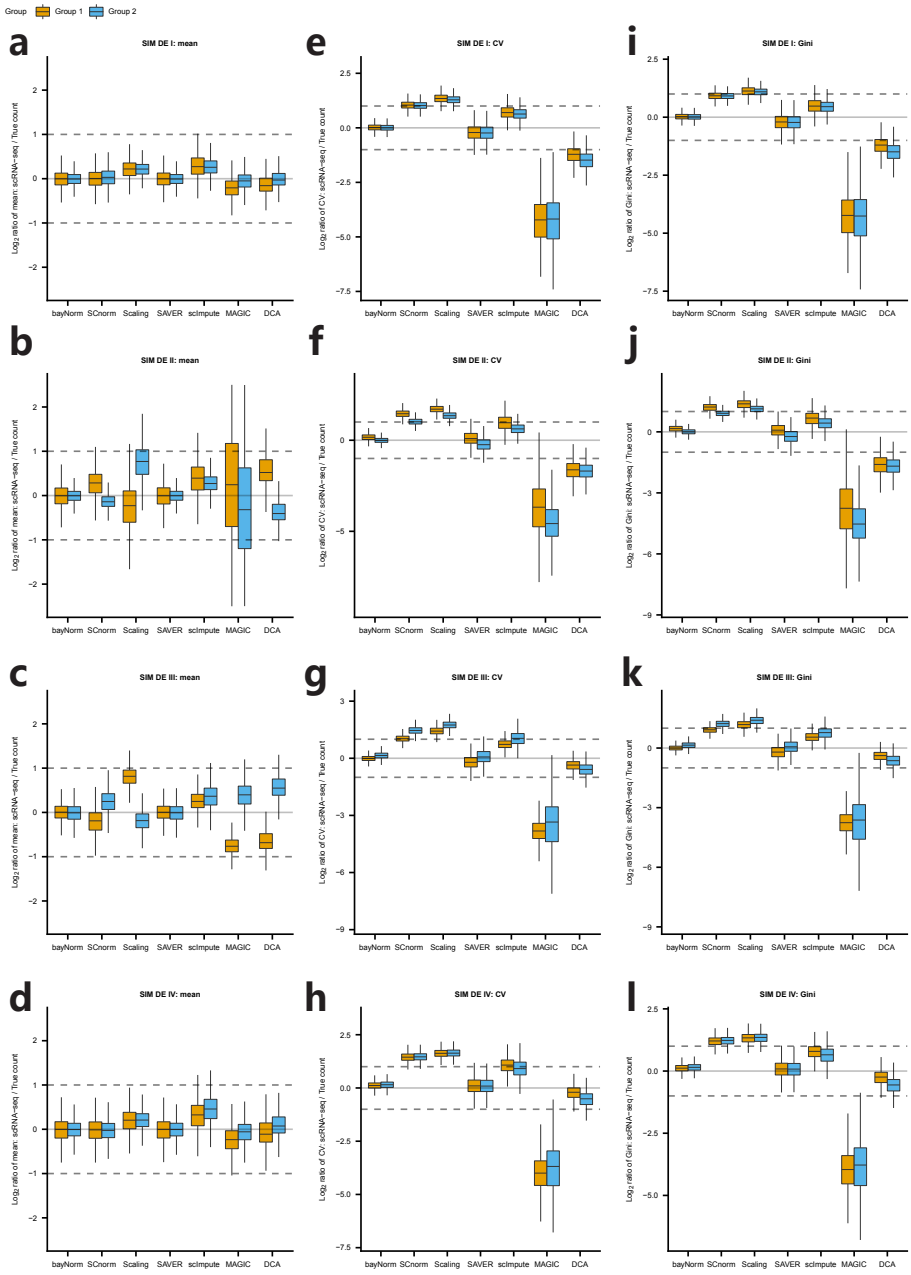
Supplementary Figure 10: MA plots based on simulated data using the Binomial_bayNorm protocol. TRUE MU stands for the MME estimated μ output from bayNorm and "EST MU" stands for the mean expression of Binomial_bayNorm simulated data scaled by the β used in bayNorm. Binomial_bayNorm simulation protocol was applied to (a) the Torre study, (b) individual NA19098 from the Tung study, (c) individual NA19101, (d) individual NA19239, (e) the Bacher study and (f) the Patel study. The dashed line was obtained by binning x values and calculating the mean of y values in each bin[1].

Figure S11



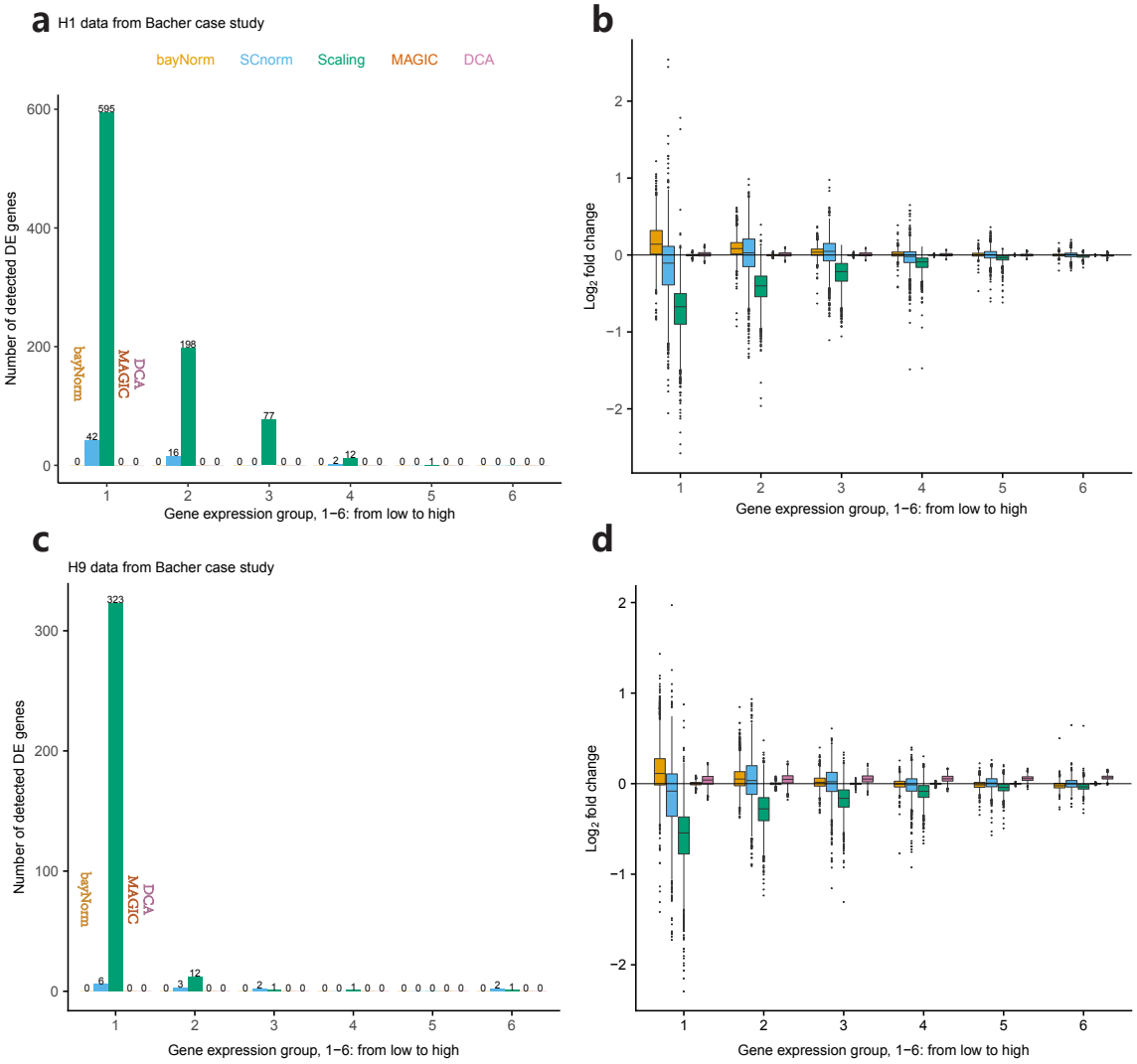
Supplementary Figure 11: Linear regression of mean expression of scRNA-seq experimental raw data vs smFISH data. (a) 2i medium single cell data from the Grn study. (b) Serum medium single cell data from the Grn study. (c) Data from the Torre study. The coefficient of explanatory variable of linear regression is used as mean beta.

Figure S12



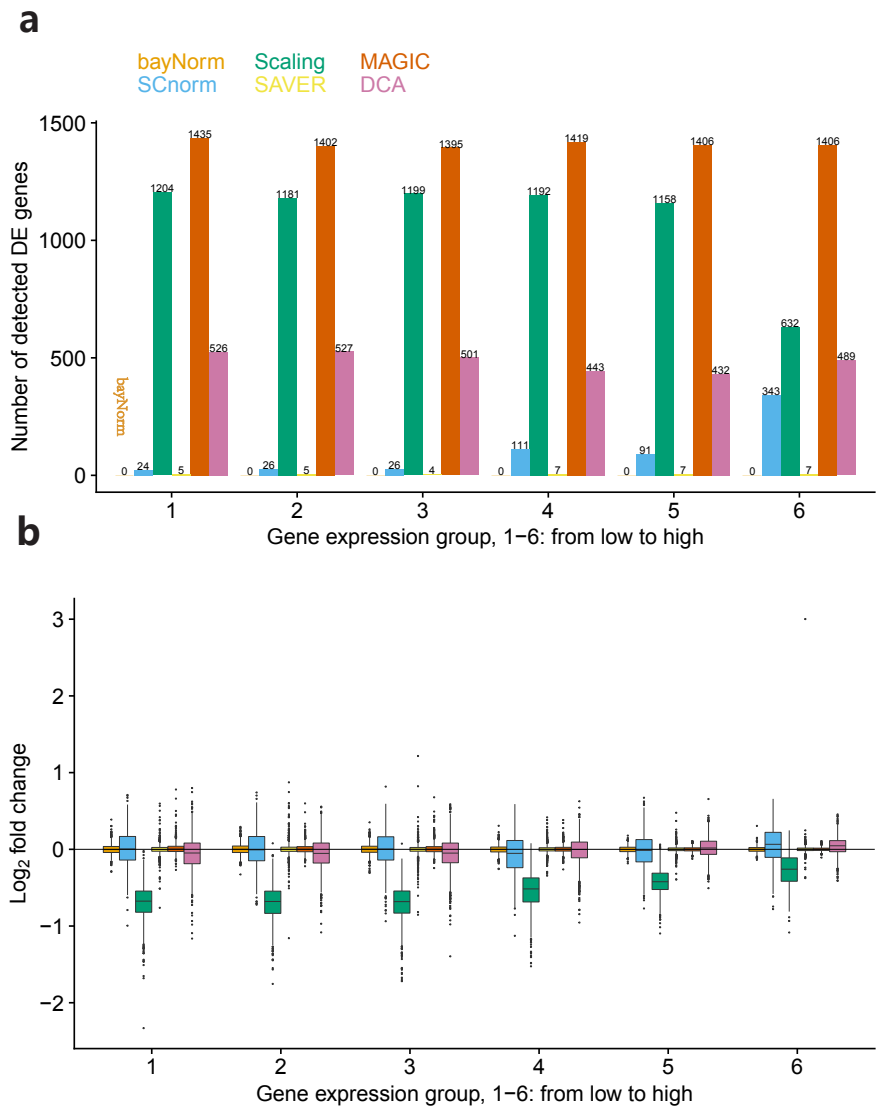
Supplementary Figure 12: Recovering the mean, CV and Gini of gene expression using simulated scRNA-seq data. For the four simulation studies: SIM DE I-IV (See Supplementary note 1 and 2 for details about simulation studies), Log_2 ratio between true simulated data (dataset before binomial downsampling) and normalized simulated scRNA-seq data for mean gene expression (a-d), CV (e-h) and Gini coefficients (i-l). (a), (e) and (i) are based on SIM DE I simulated data. (b), (f) and (j) are based on SIM DE II simulated data. (c), (g) and (k) are based on SIM DE III simulated data. (d), (h) and (l) are based on SIM DE IV simulated data. Except for the bayNorm and scaling methods, the normalized datasets have been divided by their corresponding mean capture efficiencies (either 0.1 or 0.05) for a fair comparison.

Figure S13



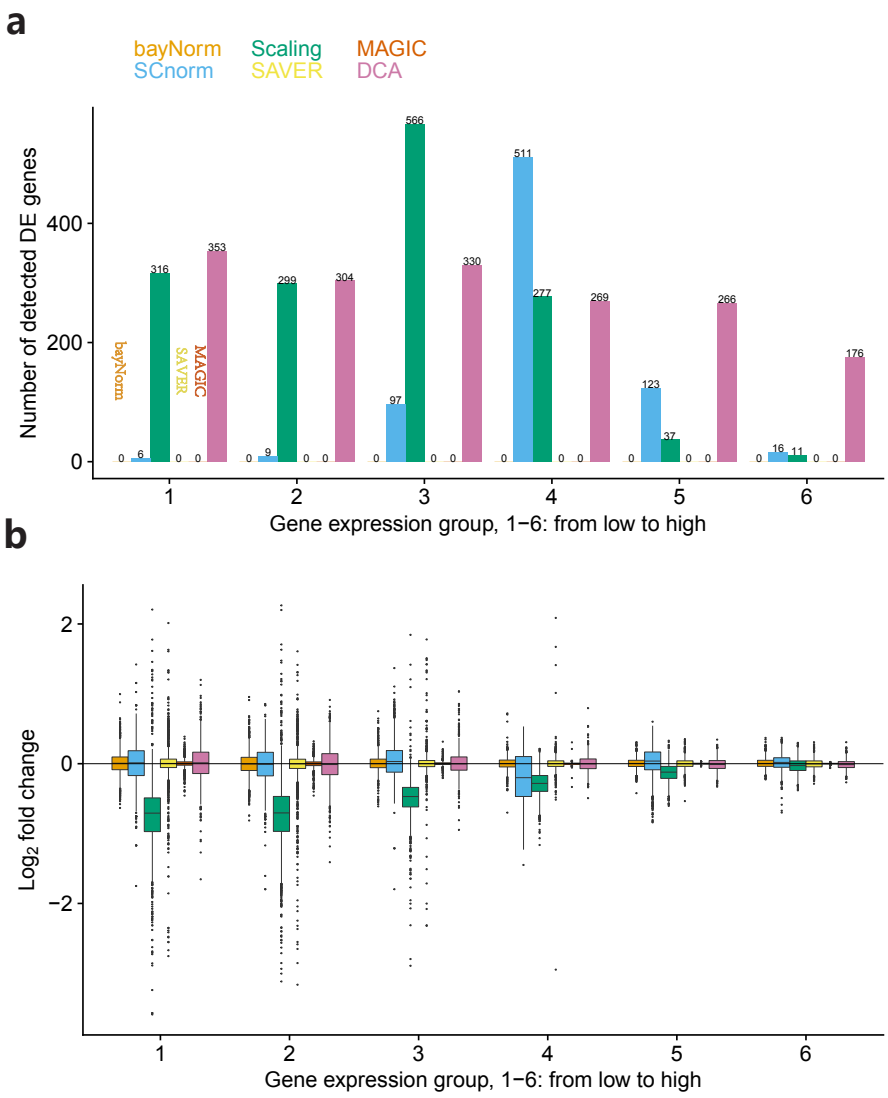
Supplementary Figure 13: bayNorm correction for differences in sequencing depths. Data from H1 (a-b, 13181 genes in total) and H9 (c-d, 13195 genes in total) hESC cells are shown. (a) Number of DE genes called by MAST as a function of gene expression groups ($P_{\text{MAST}} < 0.05$). (b) Log_2 fold change as a function of gene expression group. In (a) bayNorm is based on 20 posterior samples (3D array). In (b), bayNorm is based on the mean of posteriors (2D array).

Figure S14



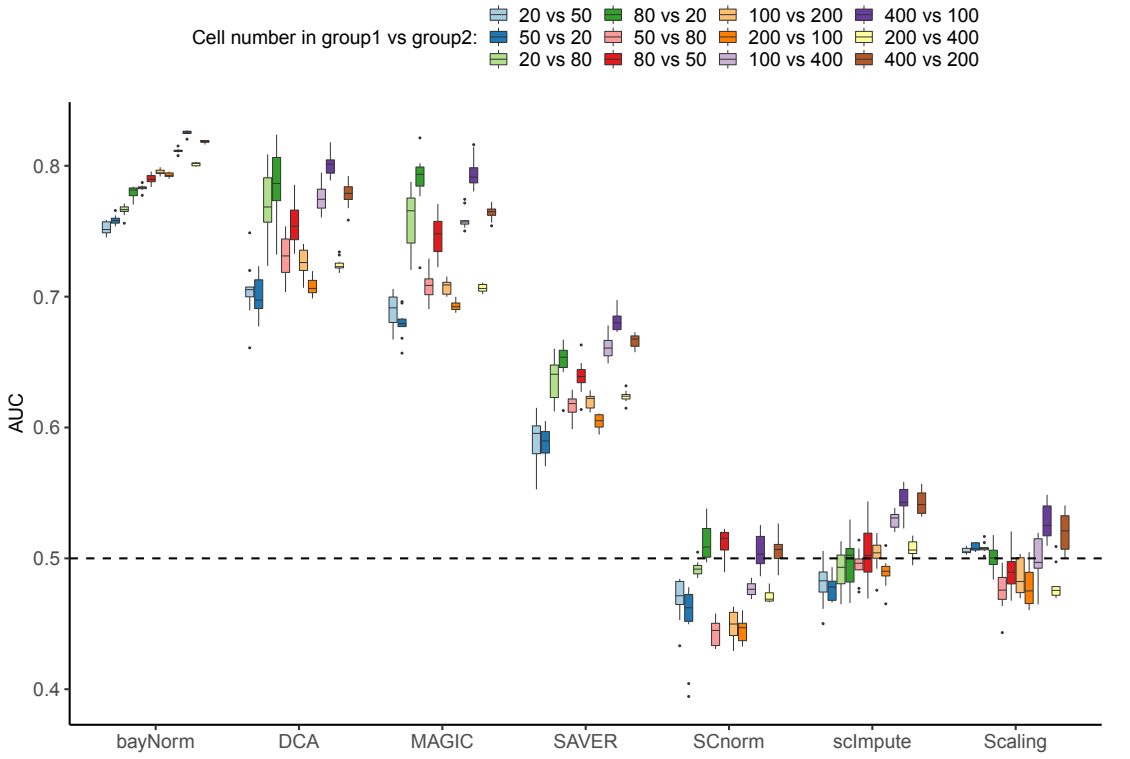
Supplementary Figure 14: Simulation analysis, SIM noDE study I (See Supplementary note 1 and 2 for details about simulation studies). (a) Number of detected DE genes (MAST) as a function of expression group ($P_{\text{MAST}} < 0.05$, 9999 genes in total). (b) Log_2 fold change of mean expression between two groups for different expression groups. For bayNorm and SAVER, 10 samples were generated and the median of p-values across the 10 samples was used in (a). In (b), bayNorm and SAVER are based on mean of posterior distributions.

Figure S15



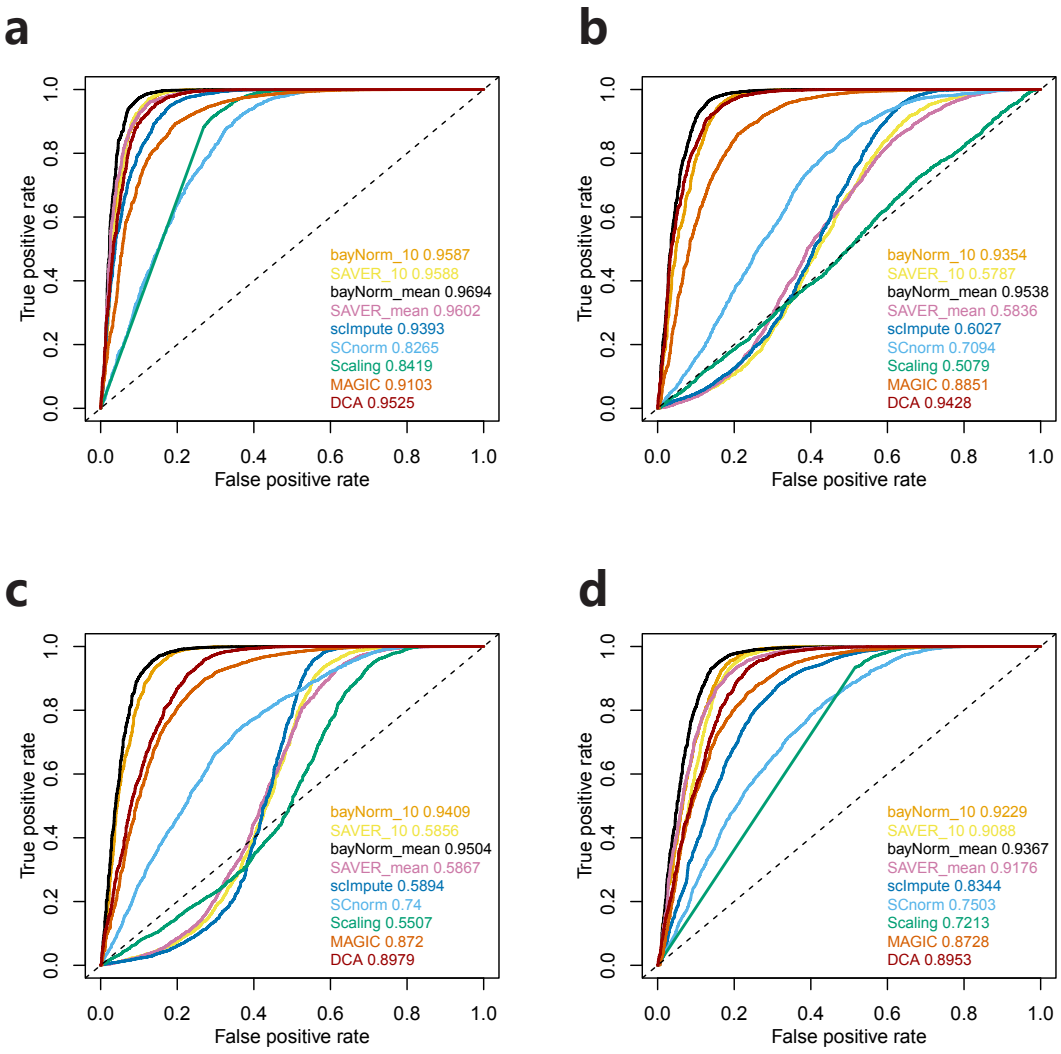
Supplementary Figure 15: Simulation analysis, SIM noDE study II (See Supplementary note 1 and 2 for details about simulation studies). (a) Number of detected DE genes (MAST) as a function of expression group ($P_{\text{MAST}} < 0.05$, 9598 genes in total). (b) Log_2 fold change of mean expression between two groups for different expression groups. For bayNorm and SAVER, 10 samples were generated and the median of p-values across the 10 samples was used in (a). In (b), bayNorm and SAVER are based on mean of posterior distributions.

Figure S16



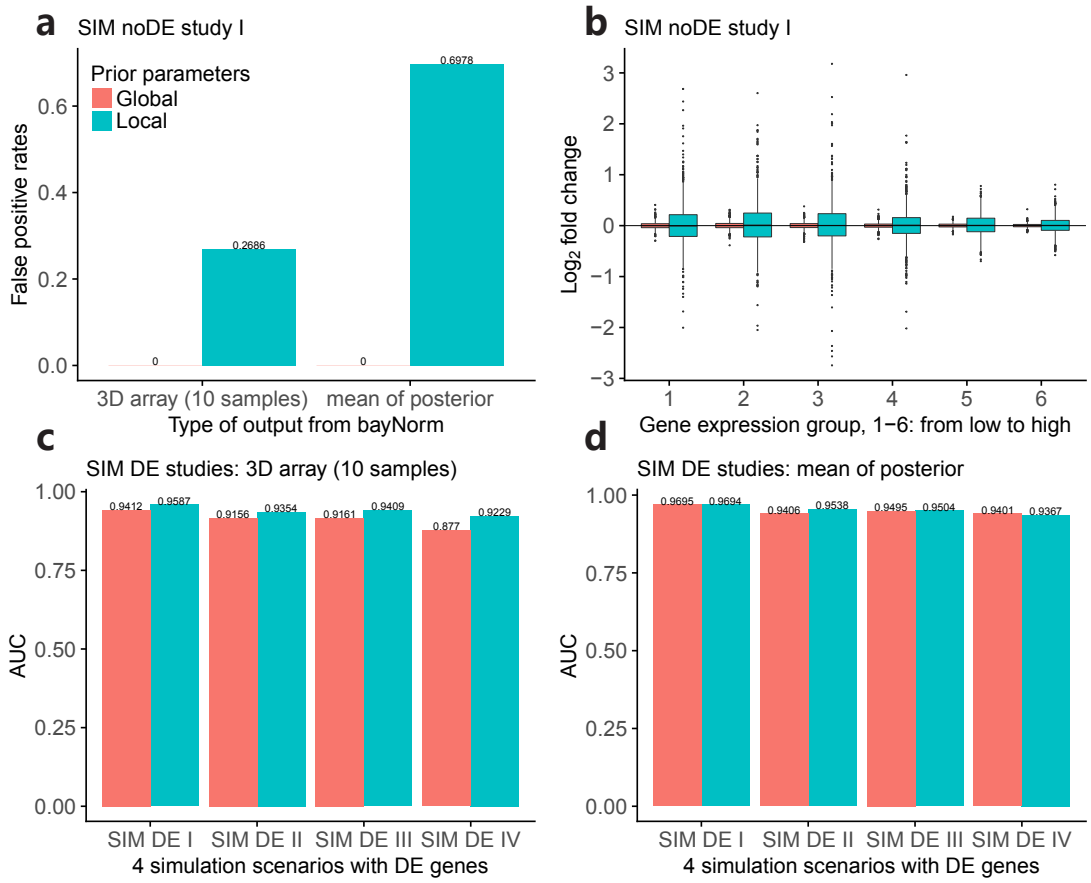
Supplementary Figure 16: DE detection for unbalanced groups of cells (UMI data from the Soumillon study). Ten samples of 20, 50, 80, 100, 200 and 400 cells were randomly selected from each group. DE detection was performed using MAST between groups as described at the top of the figure using a list of DE genes obtained from matched bulk RNA-seq data as a benchmark (1000 genes with the largest magnitude of log fold-change between the D3T0 and D3T7 samples)[3].

Figure S17



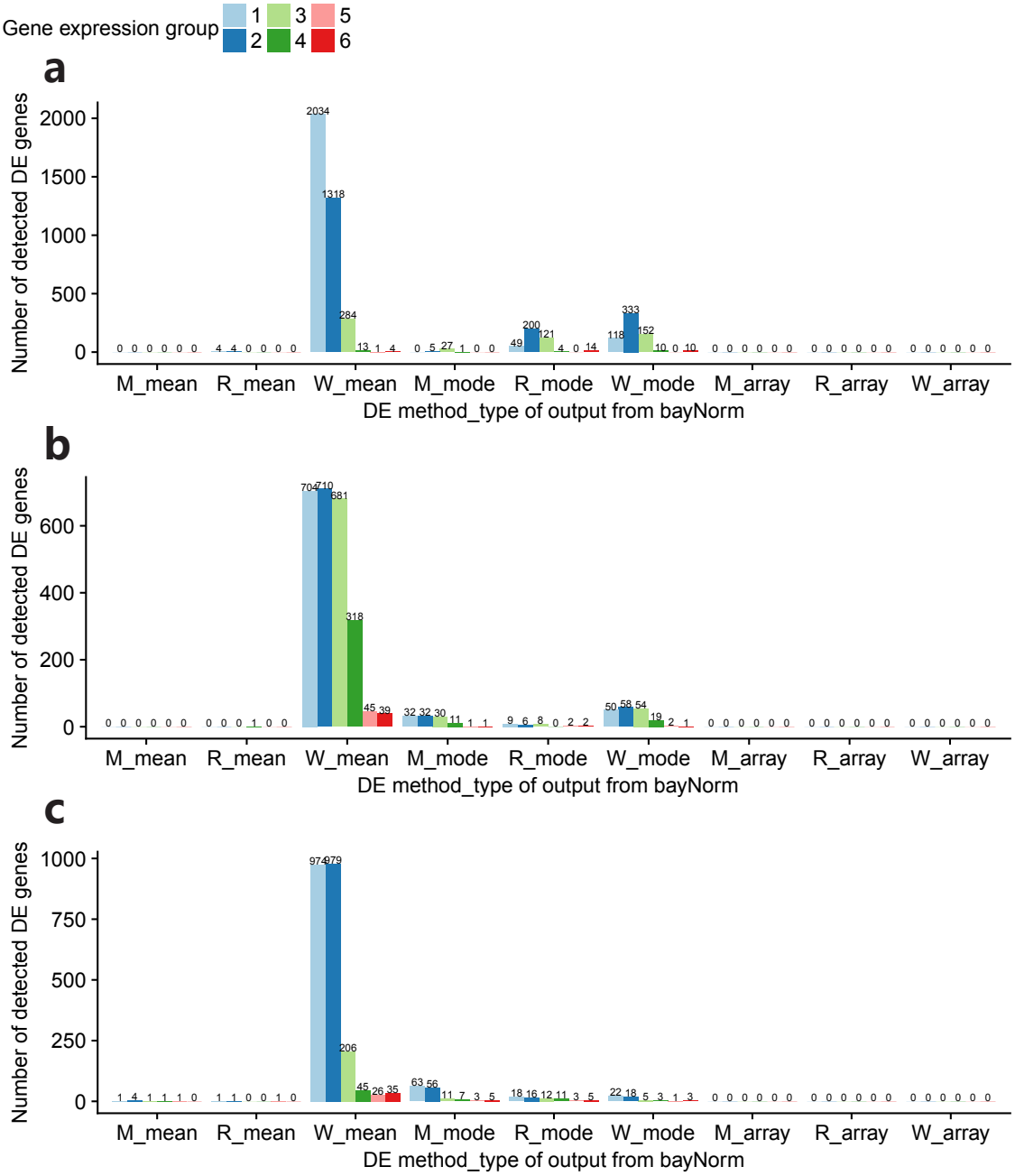
Supplementary Figure 17: DE analysis on simulated scRNA-seq data, SIM DE study (See Supplementary note 1 and 2 for details about simulation studies). (a-d) represent four simulation scenarios and DE detection is based on MAST. (a) SIM I: mean capture efficiencies are set to 0.1 for the two groups. (b) SIM II: mean capture efficiencies are set to 0.05 and 0.1 in group 1 and group 2 respectively. (c) SIM III: mean capture efficiencies are set to 0.1 and 0.05 in group 1 and group 2 respectively. (d) SIM IV: mean capture efficiencies are set to 0.05 in both groups. 2000 out of 10000 genes were simulated to be DE genes in group 1. bayNorm_10 and SAVER_10 are based on 10 samples from posterior distributions (3D arrays). DE detection was performed on each sample and the median of adjusted MAST P-values were used.

Figure S18



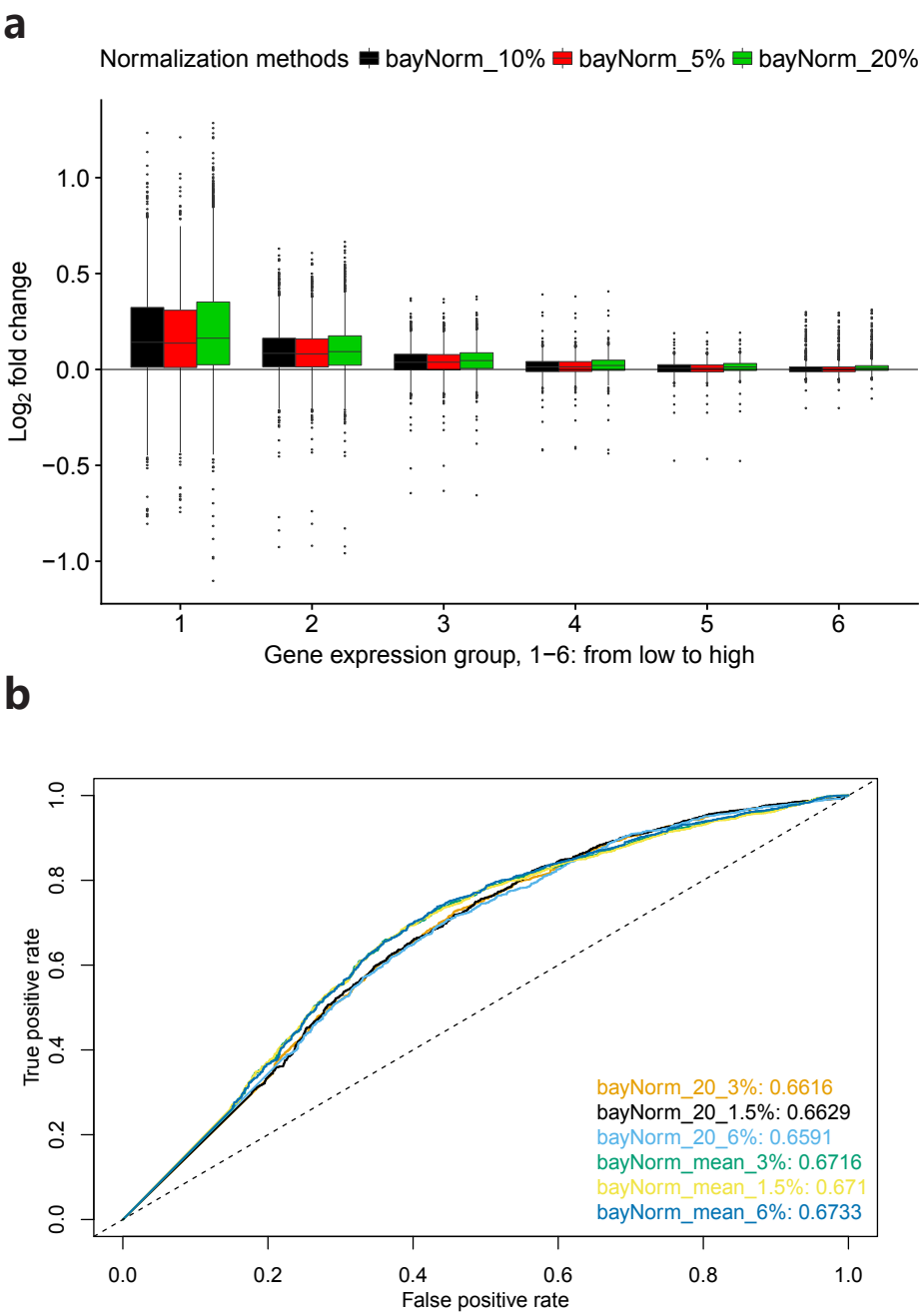
Supplementary Figure 18: Impact of global or local priors on the DE detection for simulated samples with different sequencing depths. (a-b) Simulation analysis, SIM noDE study I, with no DE genes. (c-d) SIM DE studies I-IV where 2000 out of 10000 genes were simulated to be differentially expressed in the first group. (c) and (d) are based on 10 samples (3D array) and mean (2D array) output from bayNorm respectively.

Figure S19



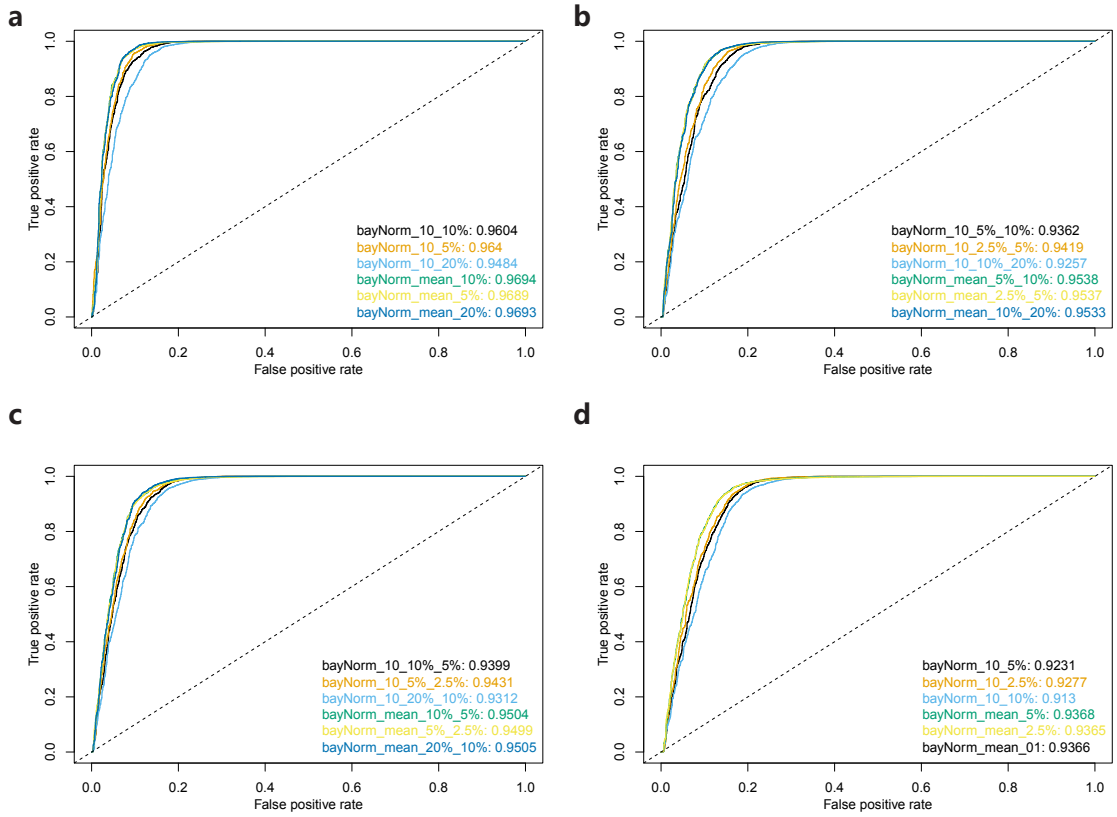
Supplementary Figure 19: Impact of different DE methods and different types of output from bayNorm on samples with different sequencing depths. M stands for MAST, R stands for ROTS and W stands for Wilcoxon test. Mean (2D array), mode (2D array) and array (3D array) stands for the three different types of output from bayNorm (see Fig S1). For the 3D array output from bayNorm, each DE method was applied on each one of 10 samples from the posterior distribution, and the median of P-values was used. (a) H1 hESC data from the Bacher study. (b) Simulated data, SIM noDE study I. (c) Simulated data, SIM noDE study II. See Supplementary note 1 and 2 for details about simulation protocols. Genes were categorized into 6 groups according to their mean expression (1-low 6-high).

Figure S20



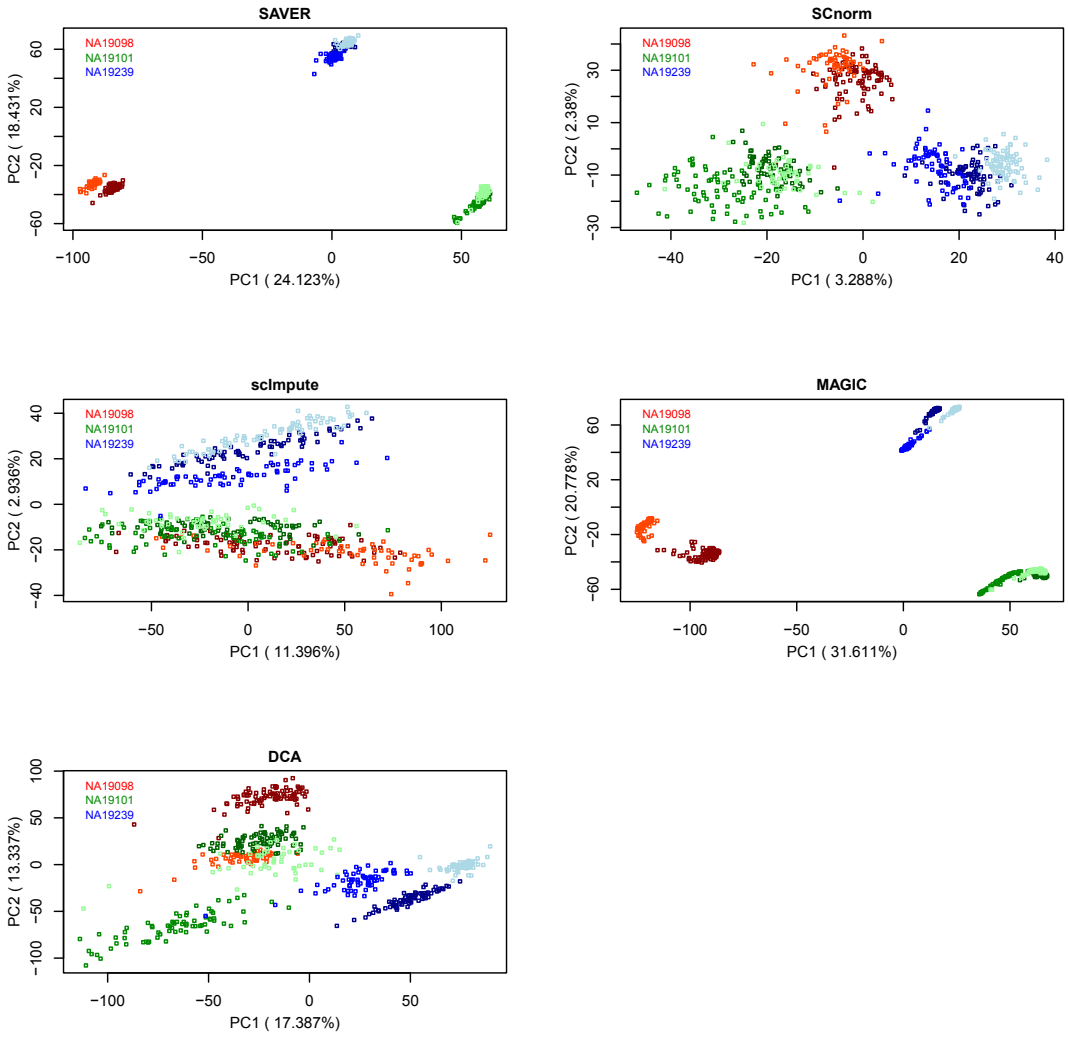
Supplementary Figure 20: Impact of different mean capture efficiencies on DE analysis based on Bacher and Islam studies. (a) For data from the Bacher study, mean capture efficiencies were set to 5%, 10% or 20%. Results are based on mean of posterior output (2D array) from bayNorm. The result of DE detection was not shown as no genes were called DE at threshold 0.05. (b) Islam study. Mean capture efficiencies were set to 1.5%, 3% and 6%. mean (2D array) or 20 samples (3D array) were used for DE detection.

Figure S21



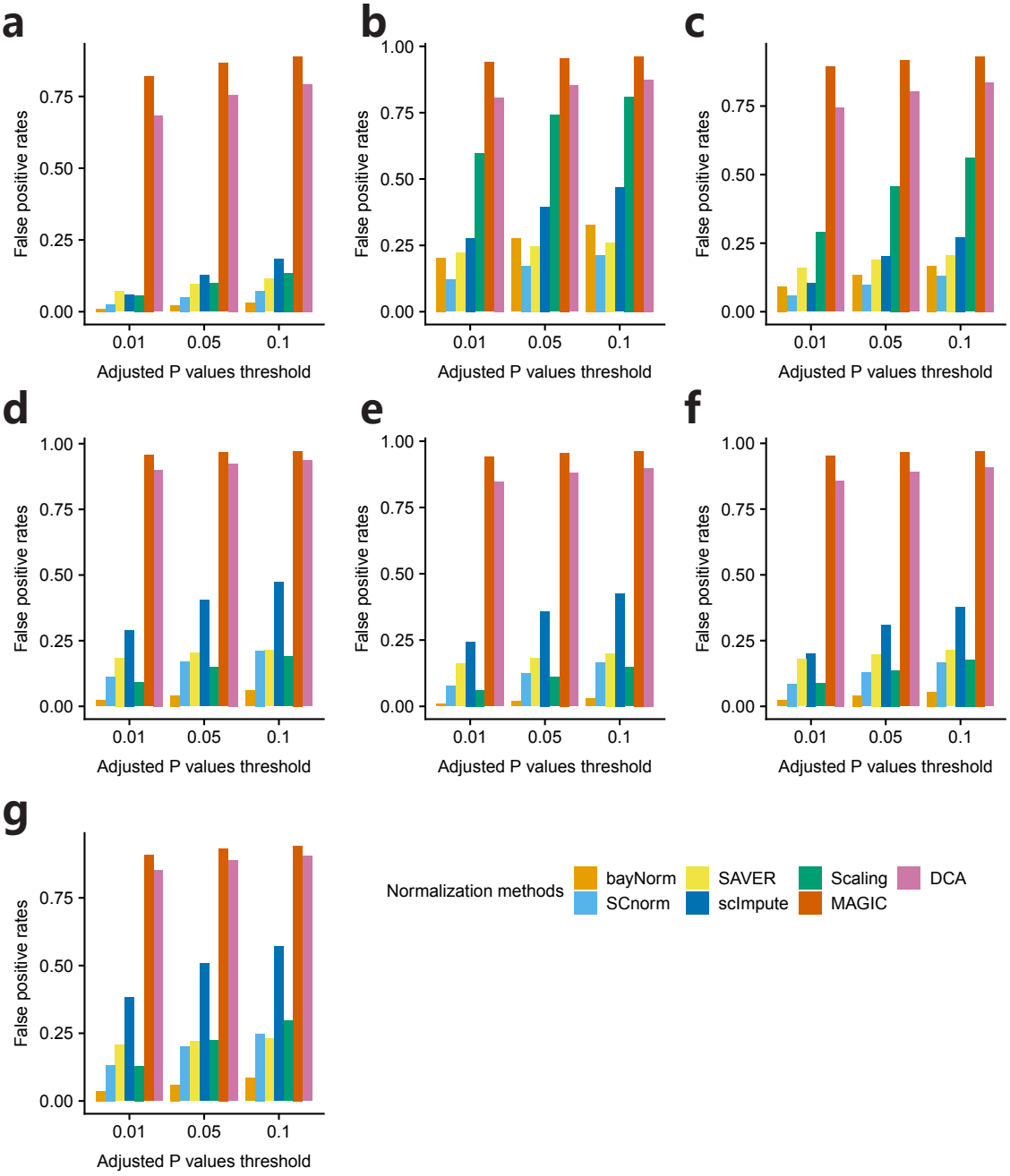
Supplementary Figure 21: Impact of different mean capture efficiencies on DE detection in simulated studies (see Supplementary note 1 and 2). (a) SIM I from our SIM DE study. Mean capture efficiencies were set to 5%, 10% or 20% using either mean of posterior (2D array) or 10 samples generated from the posterior distributions (3D array) as normalized data. (b) SIM II from our SIM DE study. Mean capture efficiencies were set to twice or half of the original magnitude. (c) SIM III from our SIM DE study, Mean capture efficiencies were set to twice or half of the original magnitude. (d) SIM IV from our SIM DE study. Mean capture efficiencies were set to 2.5%, 5% or 10%. mean stands for the mean versions output from bayNorm (2D array). Otherwise the number indicates the number of samples generated from posterior distribution (3D array). DE was performed on each sample, the median of MAST P-values were used.

Figure S22



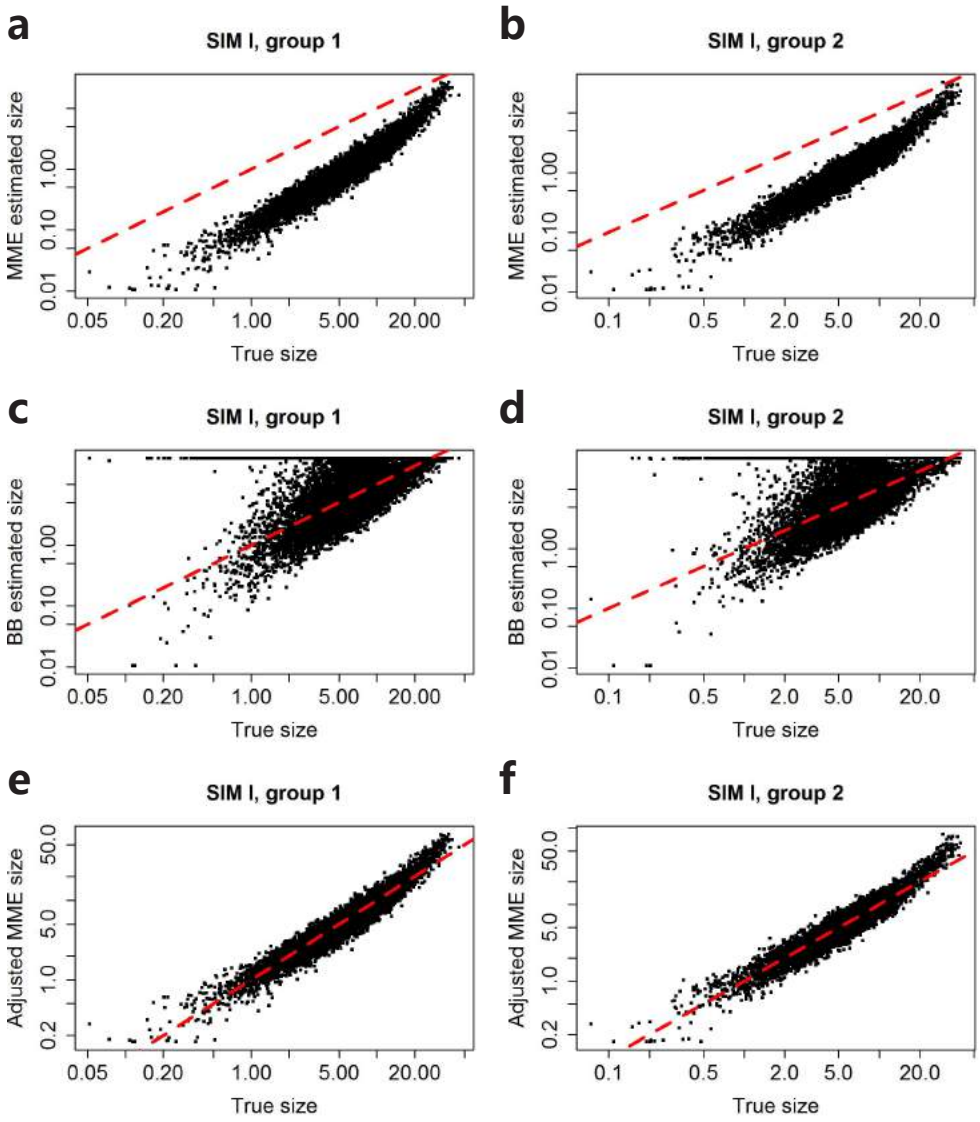
Supplementary Figure 22: PCA plots of scRNA-seq data normalised using different methods (Tung study). PCA plot of SAVER normalized data is based on the mean versions output (2D array). Different colours represent different individuals. Different shades of the same colour stands for a specific batch within each individual.

Figure S23



Supplementary Figure 23: DE detection between scRNA-seq data for different batches within single individuals. (a-c) Individual NA19101, (d-f) Individual NA19239, (g) Individual NA19098. (a), (d) and (g) show DE detection between batch 1 and batch 3 (batch 2 was not considered as suggested in the Tung study). (b) and (e) show DE detection between batch 1 and batch 2. (c) and (f) show DE detection between batch 2 and batch 3. Results of bayNorm and SAVER are based on 5 samples from posterior distributions (3D array).

Figure S24



Supplementary Figure 24: Estimation of the size factor (dispersion parameter) of the negative binomial prior distribution based on simulation studies (See Supplementary note 1 and 2). (a-b) comparison between the MME estimated size and the true size. (c-d) comparison between the BB estimated size and the true size. (e-f) comparison between the adjusted MME size and the true size. 2000 out of 10000 genes were simulated to be differentially expressed in group 1. Results are similar for other three simulated datasets (SIM DE II-IV).

Supplementary Information

Supplementary Note 1: two simulation protocols with Binomial distribution

"Binomial_Splatter" simulation protocol

We adapted the simulation protocol proposed in the R package Splatter[4] but made two main modifications to that protocol:

1. We do not multiply the mean of the Gamma distribution by the library size factors. Instead, we add cell specific factors (capture efficiencies β_j) at the last stage of simulation: Binomial step.

2. Unlike Splatter, we do not model the dropout rates explicitly. Instead we dropouts are the result of Binomial downsampling at the last stage of the simulation, which leads to a dropout vs mean expression relationship in the simulated data very similar to the one of experimental data (Supplementary Figures S2c, S3c, S4c, S5c and S6c).

The details of the simulation procedure are as follow:

1. We simulate a vector of base mean expressions λ_i' such that

$$\lambda_i' \sim \text{Gamma}(\text{shape} = \alpha_1, \text{rate} = \alpha_2)$$

.

2. We simulate a vector of outlier factors ψ_i such that $\psi \sim \ln\mathcal{N}(\mu^0, \sigma^0)$. For a proportion π^0 of genes, we multiply the base mean expression by outlier factors: $\lambda_i^0 = 1_i^0 \lambda_i \psi_i \text{median}(\lambda_i') + (1 - 1_i^0)\lambda_i'$, where $1_i^0 \sim \text{Ber}(\pi^0)$. The above two steps are the same as those implemented in Splatter.

3. In the simulations with differential expression (SIM DE), the mean expression λ_i^0 for the two groups are the same except that in the first group, we multiply λ_i^0 by a vector of DE factors simulated from the log normal distribution. Conversely, in SIM noDE study, no DE genes were simulated.

4. Then, $\lambda_i \sim \text{Gamma}(1/B_i^2, \lambda_i^0 B_i^2)$, where $B_i = (d + 1/\sqrt{\lambda_i^0})(df/\chi^2(df))^{1/2}$ stands for the Biological Coefficient of Variation (BCV), and d is common dispersion. $1/B^2$ corresponds to the dispersion parameter ϕ in the Negative Binomial distribution.

5. Then the true count $x_{ij}^0 \sim \text{Poi}(\lambda_i)$. So far, no cell-specific factors have been taken into consideration.

6. Then a vector of cell specific capture efficiencies β_j needs to be specified in the simulation. When we compare the simulated data with the real data, the capture efficiencies estimated from the real data are used in the simulation. In all simulation studies, β is simulated from the log normal distribution and normalized to a specific mean capture efficiency (either 0.05 or 0.1).

7. Lastly, we implement the binomial step to obtain the observed count (binomial downsampling):

$$x_{ij} \sim \text{Binom}(x_{ij}^0, \beta_j)$$

"Binomial_bayNorm" simulation protocol

Unlike previous simulation protocol where gene expression parameters were simulated from a specific distribution with several estimated parameters, here we use gene specific priors estimated by bayNorm together with β_j to conduct gene and cell specific simulation. So, this method produces exactly simulated data of the same size as the real data ($P \times Q$).

Let μ_i and ϕ_i be the estimated mean expression and dispersion parameter obtained by bayNorm for the i^{th} gene. Firstly a mean expression matrix (λ'_{ij}) which is of the same dimension as the real data is created, such that $\lambda'_{ij} = \mu_i$ across j. Then sampling from the following distributions leads to the simulated data:

$$\begin{aligned} &\lambda_{ij} \sim \text{Gamma}(\text{shape} = \phi_i, \text{scale} = \lambda'_{ij}/\phi_i) \\ &x^0_{ij}(\text{true count}) \sim \text{Poi}(\lambda_{ij}) \\ &x_{ij}(\text{observed count}) \sim \text{Binom}(x^0_{ij}, \beta_j) \end{aligned} \tag{1}$$

Parameter estimation from the real data ("Binomial_Splatter" simulation protocol)

The parameter estimation methods used in the simulation are basically the same as those in Splatter, except that the input raw data are scaled by β_j, before fitting the mean expression of the scaled data using Gamma distribution to estimate α_1 and α_2.

The estimation of other parameters like π^0, μ^0, σ^0, B_i were achieved based on library size normalized data as also implemented in Splatter. Non-UMI based data were scaled and rounded before parameters were estimated as explained before.

Supplementary Note 2: Simulation studies using the "Binomial_Splatter" simulation protocol

We estimated parameters from the Klein and Bacher studies (92 H1-4M hESCs) and then generated 6 simulated datasets for comparing different normalization methods for their performance in correcting different capture efficiencies (study without DE genes), and in DE genes detection.

Two simulated datasets are generated without DE genes (homogeneous cells across two groups, but different mean capture efficiencies):

SIM noDE study I (Using parameters estimated from the Klein study): The mean capture efficiencies of the two groups are 0.1 and 0.05 respectively. Simulated data was used in: Supplementary Figures S14, S18a-b, S19b.

SIM noDE study II (Using parameters estimated from 92 H1-4M hESCs of the Bacher study): The mean capture efficiencies of the two groups are 0.1 and 0.05 respectively. Simulated data was used in: Supplementary Figures S15, S19c.

Simulation SIM DE studies are all based on parameters estimated from the Klein study.

SIM DE study I: The mean capture efficiencies for the two groups are both 0.1. Simulated data was used in: Supplementary Figures S12a,e,i, S17a, S18c,d, S21a and S24.

SIM DE study II: The mean capture efficiencies for the two groups are 0.05 and 0.1 respectively. Simulated data was used in: Supplementary Figures S12b,f,j, S17b, S18c,d, and S21b.

SIM DE study III: The mean capture efficiencies for the two groups are 0.1 and 0.05 respectively. Simulated data was used in: Supplementary Figures S12c,g,k, S17c, S18c,d, and S21c.

SIM DE study IV: The mean capture efficiencies for the two groups are both 0.05. Simulated data was used in: Supplementary Figures S12d,h,l, S17d, S18c,d, and S21d.

Genes with 0 counts across two groups were filtered out at the very beginning. No cells were filtered out. Belows are details about parameter settings used in the simulation studies which were estimated from the raw data as discussed above.

Parameters estimated from the Klein study: Most parameters were estimated from the Klein study except β. For each group, 10000 genes and 100 cells were simulated. The base mean expression for both groups were simulated from the Gamma distribution with $\alpha_1 = 1.889$ and $\alpha_2 = 0.1229$. $\pi^0 = 3\%$ across two groups. The outlier factors were simulated from the log normal distribution with μ^0 and σ^0 set to 2.3 and 0.75 respectively. BCV was calculated with $d = 0.12$ and $df = 105$. The estimation of β is discussed in the Supplementary Note 3.

In the SIM noDE study I and the SIM DE studies I-IV, two groups of 100 cells with 10000 genes were simulated using the above parameter settings. β_j are simulated from the log normal distribution within each group with mean and sd (log scale) set to 2.74 and 0.3908 respectively. Within each group, we normalized the β to either 0.1 or 0.05.

In the SIM DE I-IV studies, the DE factors in the first group are simulated from the log normal distribution with log scale mean and sd set to 1 and 0.5 respectively.

Parameters estimated from the Bacher study (based on 92 H1-4M hESCs, 4 million mapped reads per cell): Parameters were estimated from the raw data scaled by 20. $\alpha_1 = 0.4129$ and $\alpha_2 = 0.005766$. Outliers genes were simulated with $\pi^0 = 0.7\%, \mu^0 = 4.745$ and $\sigma^0 = 0.6027$. BCV was calculated with $d = 0.3113$ and $df = 7.6859$.

In the SIM noDE case study II, two groups of 100 cells with 10000 genes were simulated using the above parameter setting. β_j are simulated from the log normal distribution within each group with mean and sd (log scale) to be -2.276 and 0.6886 respectively. Within each group, we normalized the β to either 0.1 or 0.05.

Supplementary Note 3: Publicly available datasets and their preprocessing

Bacher study (non-UMI).

Single-cell RNA-seq expression data were downloaded from GEO GSE85917[5]. In this experiment, two groups of undifferentiated H1 hESCs were sequenced to a depth of 4 million mapped reads per cell and 1 million mapped reads per cell respectively. A similar experiment was done for H9 hESCs cells. The following filtering protocol was used in our study: spike-ins and genes which do not have at least 10 non-zero counts were removed. After filtering, each group of H1 cells had 92 cells and 13181 genes, while groups of H9 cells had 91 cells and 13195 genes. Since these are non-UMI based data, we divided the raw data by a factor 20 for the 4 million mapped reads group and 10 for the other group so that scaled and rounded raw data were closer to the theoretical dropout vs mean curve (Supplementary Figure S9 a-d).

In order to estimate β, we let the total counts of observed spike-ins in each cell be the scaling factors s_j, and then normalized to 0.1 (based on scaled ERCC data, see Methods). As discussed in the text, within a 2 fold window of mean β, the performance of bayNorm in terms of DE detection is consistent

(Supplementary Figure S20 a). Since cells in the two groups are the same, prior parameters were estimated across two groups when applying bayNorm ("global priors").

Data was used in the Supplementary figures: S7, S9a-d, S13, S19a and S20a.

Islam study (non-UMI).

Raw data (48 ES cells and 44 MEF cells with 7284 genes) and a list of benchmark DE genes were kindly provided by Maria K. Jaakkola[6]. Genes which have zero expressions across all the 92 cells were removed in advance which left us with 5826 genes. Raw data was divided by a factor 10 before applying bayNorm on it as this is a non-UMI data and the scaled and rounded raw data is closer to the theoretical dropout vs mean curve.

For estimating β, scaling factors were estimated using scran[2] and were normalized to 0.03 (see Methods). The impact of different $\bar{\beta}$ can be found in Supplementary Figure S20 b.

Data was used in the Figure 3c, Supplementary figures: S9e-f, S10e and S20b.

Patel study (non-UMI).

Raw data were stored in the R package "patel2014gliohuman"`https://github.com/willtownes/patel2014gliohuman`[7]. Single cell data were scaled by a factor 20 and rounded as this is a non-UMI data and the scaled and rounded raw data is closer to the theoretical dropout vs mean curve. Cells with total counts less than or higher than the tenth percentile of total counts across cells were filtered out. In addition, genes with mean expression less than 1 were also removed. β were estimated using scran[2] with parameter "positive=TRUE". The size factors were normalized to 0.06 (see Methods). Finally, cells with $\beta < 0.01$ were filtered out, leaving 590 cells and 5519 genes in the final datasets. We applied bayNorm on the scaled, rounded and preprocessed dataset. The estimated priors were used as input for Binomial_bayNorm simulation protocol (Supplementary Figure S10f).

Data was used in the Supplementary figures: S10f.

Tung study (UMI).

Filtered molecule count matrix as well as the code for estimating β using spike-ins were downloaded from the GitHub repository: `https://github.com/jdblischak/singleCellSeq` [8]. The list of benchmark DE genes was kindly provided by the author of R package DECENT[3]. Genes with 0 values across all three individuals were filtered out leaving 13058 genes in the final dataset. No cells were filtered out, resulting in 142, 201 and 221 cells for individuals NA19098, NA19101 and NA19239 respectively.

Data was used in the Figure 4, Supplementary figures: S3-5, S8c-d, S10b-d and S22-23.

Grün study (UMI).

Single-cell RNA-seq expression data were downloaded from GEO GSE54695[9]. The smFISH data used in that paper was kindly provided by the author.

The downloaded data was transformed to transcript number. We adapted the code provided in the supplementary material of [10] to convert the data to UMI count. We followed the same criterion as [10] for filtering genes in the 2i and serum data respectively. After filtering, we kept 74 cells and 9377 genes for the 2i medium data. For the serum medium data, we kept 52 cells and 9440 genes.

smFISH data were normalized by scaling factors which were calculated as cell sizes divided by mean of cell sizes.

Total number of input spike-ins was estimated by adapting the code provided in the supplementary information of [10]. We divided the total number of observed spike-ins in each cell by the total number of input spike-ins to obtain scaling factors. We used smFISH data to estimate $\bar{\beta}$ for single cells under 2i and serum medium respectively (0.1212 and 0.1187 respectively, Supplementary Figure S11a-b). To obtain β, we normalized scaling factors (see Methods) to the corresponding $\bar{\beta}$ within each one of the two datasets.

Data was used in the Figure 2a,c,e,g and Supplementary figures: S11a-b.

Klein study (UMI).

Single-cell RNA-seq expression data were downloaded from GEO GSE65525[11]. ES cells data at day 0 (933 cells with 24175 genes) were used for simulations. We did not filter out any genes or cells.

For estimating β, trimmed mean of each cell at 1% was used and was normalized to $\bar{\beta}$ set to 0.06[11] (see Methods).

Data was used in the Figure 1b-e, Figure 3a-b, Supplementary figures: S2 and S8a-b.

Torre study (UMI).

Single-cell RNA-seq expression data were downloaded from GEO GSE99330 (GSE99330_dropseqUPM.txt.gz)[12]. This data matrix was converted to counts using a code kindly provided by the author of SAVER[13].

There are 32287 genes and 8640 cells in the raw data. First cells with less than 2000 genes detected or where the gene "GAPDH" could not be detected were filtered out. Second, genes with mean expression less or equal to 0.01 were removed. Third, the gene "GAPDH" was removed because it is used as a proxy for cell size[13] by normalizing it to the mean β which was estimated using smFISH data (see Methods). The final filtered dataset contained 9289 genes and 1134 cells.

For smFISH data, we filtered out cells with "GAPDH" counts below the bottom 10^{th} percentile or above the top 10^{th} percentile of 'GAPDH" counts. smFISH data were then normalized by the expression of GAPDH divided by GAPDH mean expression[13].

We fitted a linear regression of the mean expression of filtered dataset on that of the normalized smFISH datset (Supplementary Figure S11c). The coefficient of explanatory variable was then used as $\bar{\beta}$. "GAPDH" expression, which was filtered out previously, was divided by the median and multiplied by $\bar{\beta}$ (see Methods).

Data was used in the Figure 2b,d,f,h, Figure 3a-b, Supplementary figures: S6, S8e-f and S10a.

Soumillon study (UMI).

The dataset was downloaded from GEO GSE53638 (GSE53638_D3_UMI.dat.gz for the single cell data and GSE53638_D3_Bulk_UMI.dat.gz for the bulk data)[14]. DE detection was performed between the stage-3 differentiated cells at day 0 (D3T0) and day 7 (D3T7) (23895 genes and 1949 cells) [3]. Cells with library sizes below the bottom and above the top 5^{th} percentiles were filtered out. Genes with mean expression across two groups greater than 0.05 were retained, resulting in a dataset of 1754 cells with 8586 genes (832 cells and 922 cells belonging to day 0 and day 7 time-points respectively). Using the same 8586 genes in the bulk dataset, the reference DE genes were defined to be the top 1000 genes which have the greatest log fold-change in the corresponding bulk RNA-seq data[3].

Based on the smFISH data, $\bar{\beta}$ is expected to be in the range of $1 - 2\%$[14]. Here we set $\bar{\beta} = 2\%$. Scaling factors were estimated using R package scran[2], and then normalized to 0.02 (see Methods).

Data was used in the Figure 3d and Supplementary figures: S16.

Supplementary Note 4: Normalization methods and relevant R packages

Splatter, R package version 1.4.1

For both UMI and non-UMI data, the default settings of Splatter were used for estimating parameters from the input data and simulating scRNAseq data.

For non-UMI data in Supplementary Figure S7, H1_P24 single cell data were divided by 20 and then rounded as an input for Splatter.

SAVER, R package version 0.4.0

We used the default settings for SAVER throughout the paper. When estimating mean, CV and Gini coeffieicnts, for both bayNorm and SAVER we generated 5 samples for the Torre study and 20 samples the Grün study (3D arrays). The mean, CV and Gini were estimated across the cells and samples. In the 6 simulation studies, 10 samples were generated from posterior for both bayNorm and SAVER. In the Klein (Figure 3a-b), Tung and Soumillon studies, 5 samples were generated. In Tung study, SAVER was applied within each individual.

SCnorm, R package version 1.1.0

We used the default settings in SCnorm except for UMI datasets and simulated data, where "ditherCounts=TRUE" was used as UMI data contains tied counts.

scImpute, R package version 0.0.6

We applied scImpute using its default settings. In the Tung study, scImpute was applied on each individual independently.

MAGIC, R package version 0.1.0

MAGIC was applied using its default settings. In the Tung study, MAGIC was applied within each individual.

DCA, Python package version 0.2.2

DCA was applied using its default settings. In Tung study, since genes with 0 counts across cells within each individual could be filtered out, we applied DCA acorss all cells.

Scaling method

Throughout the paper, the scaling method refers to the modified formula **??** such that $\tilde{x}_{ij} = \frac{x_{ij}}{\beta_j}$. In UMI datasets and simulated data, β used in scaling method are as the same as that used in bayNorm.

For the Bacher study (non-UMI), the scaling factors were set to be the total counts of spike-ins normalized to 0.1 (see Methods).

For the Islam study (non-UMI), the scaling factors were estimated based on raw data using R package scran with "sizes=c(20,30,40,50)" and "positive=TRUE", which were further normalized to 0.03 (see Methods).

References

[1] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 2017.

[2] Aaron TL 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.

[3] Chengzhong Ye, Terence P Speed, and Agus Salim. Decent: Differential expression with capture efficiency adjustment for single-cell rna-seq data. *bioRxiv*, page 225177, 2017.

[4] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

[5] Rhonda Bacher, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski. Scnorm: robust normalization of single-cell rna-seq data. *Nature methods*, 14(6):584, 2017.

[6] Maria K Jaakkola, Fatemeh Seyednasrollah, Arfa Mehmood, and Laura L Elo. Comparison of methods to detect differentially expressed genes between single-cell populations. *Briefings in bioinformatics*, page bbw057, 2016.

[7] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190): 1396–1401, 2014.

[8] Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7:39921, 2017.

[9] Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637–640, 2014.

[10] 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(1):70, 2016.

[11] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[12] Eduardo Torre, Hannah Dueck, Sydney Shaffer, Janko Gospocic, Rohit Gupte, Roberto Bonasio, Junhyong Kim, John Murray, and Arjun Raj. Rare cell detection by single-cell rna sequencing as guided by single-molecule rna fish. *Cell systems*, 6(2):171–179, 2018.

[13] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature Methods*, page 1, 2018.

[14] Magali Soumillon, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, and Tarjei S Mikkelsen. Characterization of directed differentiation by high-throughput single-cell rna-seq. *BioRxiv*, page 003236, 2014.