A Systematic Evaluation of Single Cell RNA-Seq Analysis Pipelines

Library preparation and normalisation methods have the biggest impact on the performance of scRNA-seq studies

Beate Vieth[1,*], Swati Parekh[2], Christoph Ziegenhain[3], Wolfgang Enard[1], Ines Hellmann[1,+]

[1] Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians University, Munich, Germany
[2] Max Planck Institute for Biology of Ageing, Cologne, Germany
[3] Department of Cell and Molecular Biology, Karolinska Institutet, SE-171 65 Stockholm, Sweden

[*] vieth@bio.lmu.de
[+] hellmann@bio.lmu.de

Abstract

The recent rapid spread of single cell RNA sequencing (scRNA-seq) methods has created a large variety of experimental and computational pipelines for which best practices have not been established yet. Here, we use simulations based on five scRNA-seq library protocols in combination with nine realistic differential expression (DE) setups to systematically evaluate three mapping, four imputation, seven normalisation and four differential expression testing approaches resulting in $\sim$ 3,000 pipelines, allowing us to also assess interactions among pipeline steps.

We find that choices of normalisation and library preparation protocols have the biggest impact on scRNA-seq analyses. Specifically, we find that library preparation determines the ability to detect symmetric expression differences, while normalisation dominates pipeline performance in asymmetric DE-setups. Finally, we illustrate the importance of informed choices by showing that a good scRNA-seq pipeline can have the same impact on detecting a biological signal as quadrupling the sample size.

Introduction

Many experimental protocols and computational analysis approaches exist for single cell RNA sequencing (scRNA-seq). Furthermore, scRNA-seq analyses can have different goals including differential expression (DE) analysis, clustering of cells, classification of cells and trajectory reconstruction[1]. All these goals have the first analysis steps leading to a filtered and normalised count matrix in common. Here, we focus on these important first choices made in any scRNA-seq study. As of now, benchmarking studies exist only separately for each analysis step, which are library preparation protocols[2,3], alignment [4,5], annotations[6], preprocessing[7,8] and normalisation[9]. However, the impact of the combined choices of the separate analysis steps on overall pipeline performance has not been quantified. In order to achieve a fair and unbiased comparison of computational pipelines, simulations of realistic data sets are necessary. This is because the ground truth of real data is unknown and alternatives, such as concordance analyses are bound to favour similar and not necessarily better methods.

To this end, we integrated popular methods for each analysis step into our simulation framework powsimR[10]. As the basis for simulations, powsimR uses raw count matrices to describe the mean-variance relationship of gene expression measures. This includes the variance introduced during the experiment itself as well as extra variance due to the first to computational steps of expression quantification. Adding differential expression then provides us with detailed performance measures based on how faithfully DE-genes can be recovered.

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] found up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

Realistic simulations in conjunction with a wide array of scRNA-seq methods, allow us not only to quantify the performance of individual pipeline steps, but also to quantify interdependencies among the steps. Moreover, the relative importance of the various steps to the overall pipeline can be estimated. Hence, our analysis provides sound recommendations regarding the construction of an optimal computational scRNA-seq pipeline for the data at hand.

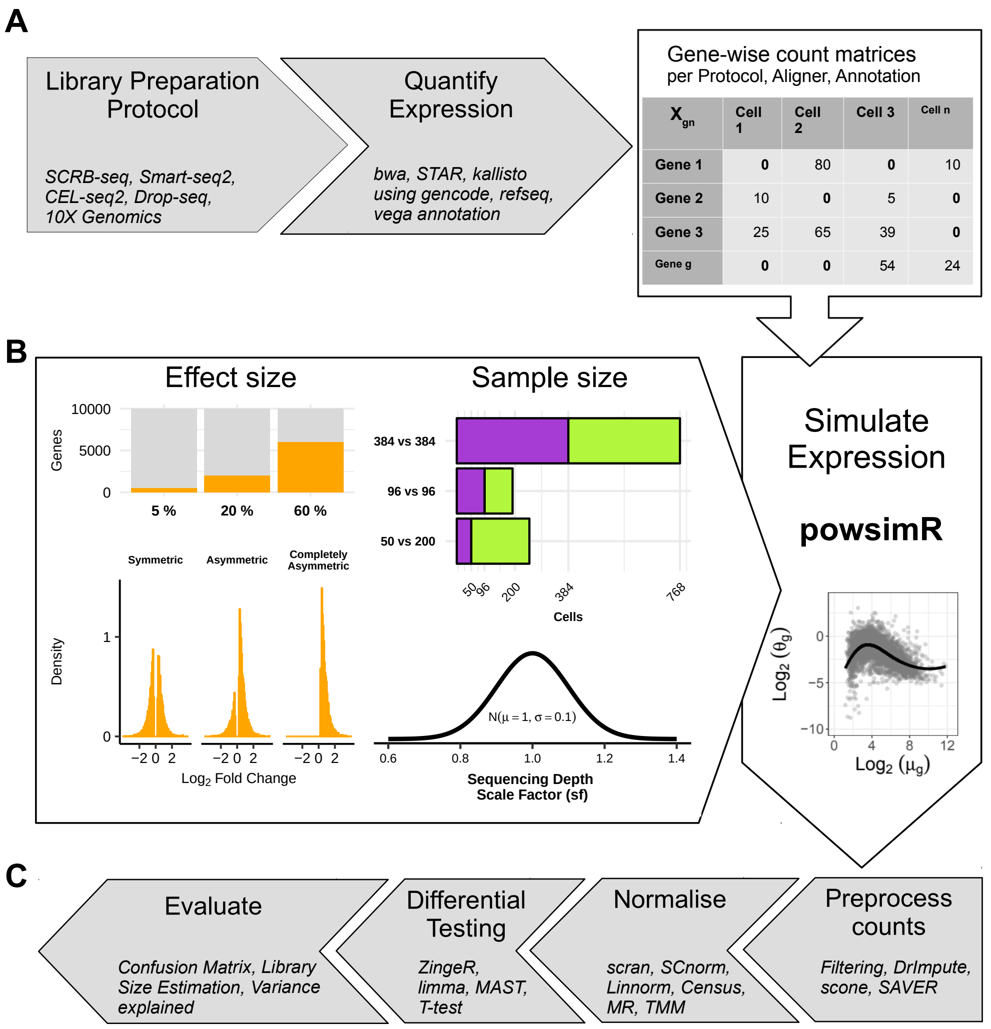
Results

The starting point for our comprehensive pipeline comparison is a representative selection of scRNA-seq library preparation protocols (**Figure 1A**). Here, we included one full-length method (Smart-seq2[13]) and four UMI methods[14,15,2,16], combined with three mapping approaches[17,18,19] and three annotation schemes[20,21,22] resulting in 37 distinct raw count matrices (Online Methods). We simulated 27 distinct DE-setups per matrix, each with 20 replicates, resulting in a total of 19,980 simulated data sets (**Figure 1 B**).

Genome-mapping quantifies more genes with high accuracy

We first investigated how expression quantification is affected by different alignment methods. For each of the three following strategies we picked one the most popular methods (**Supplementary Figure S2**): 1. alignment of reads to the genome using splice-aware alignment (STAR[17]), 2. alignment to the transcriptome (BWA[18]) and 3. pseudo-alignment of reads guided by a transcriptome (kallisto[23]).We then combined these with three annotation schemes including two curated schemes (RefSeq[20] and Vega [22]) and the more inclusive GENCODE[21] (**Supplementary Table S2**).

First, we assessed the performance by the number of reads or UMIs that were aligned and assigned to genes (**Figure 2A** and **Supplementary Figure S3**). Generally, STAR in combination with GENCODE aligned (82-86%) and assigned (40-63%) the most reads.



X_{gn}	Cell 1	Cell 2	Cell 3	Cell n
Gene 1	0	80	0	10
Gene 2	10	0	5	0
Gene 3	25	65	39	0
Gene g	0	0	54	24

Figure 1. Study Overview
A) The data sets yielding raw count matrices. We use scRNA-seq data sets from Ziegenhain et al.[2] and Zheng et al.[16] representing 5 popular library preparation protocols. For each data set, we obtain multiple gene count matrices that result from various combinations of alignment methods and annotation schemes (see also Supplementary Figure S1 and S2, and Supplementary Table S1 and S2). **B) The Simulation setup**. Using powsimR Vieth et al.[10] distribution estimates from real count matrices, we simulate the expression of 10,000 genes for two groups with 384 vs 384, 96 vs. 96 and 50 vs. 200 cells, where 5%, 20% or 60% of genes are DE between groups. The magnitude of expression change for each gene is drawn from a narrow gamma distribution ($X \sim \Gamma(\alpha = 1, \beta = 2)$) and the directions can either be symmetric, asymmetric or completely asymmetric. **C) The analysis pipeline.** The simulated data sets are then analysed using combinations of four preprocessing, seven normalisation and four DE approaches. The evaluation of these pipelines focuses on the outcome of the confusion matrix and its derivatives (TPR, FDR, pAUC, MCC), deviance in library size estimates (RMSE) and computational run time.

BWA assigned a slightly lower fraction of reads (33-44%), but - suspiciously - these were distributed across more UMIs. As reads with the same UMI are more likely to originate from the same mRNA molecule and thus the same gene, the average number of genes with which one UMI sequence is associated, can be seen as a measure of false mapping. Indeed, we find that the same UMI is associated with more genes when mapped by BWA than when mapped by STAR (**Figure 2B**). This indicates a high false mapping rate, that probably inflates the number of genes that are detected by BWA (**Figure 2C** and **Supplementary Figure S4**). In contrast, the final UMI count matrix obtained with kallisto is more sparse, assigning the smallest number of reads and detecting 20-25% fewer genes than STAR (**Figure 2A,C**).

This said, it remains to be seen what impact the differences in read or UMI counts obtained through the different alignment strategies and annotations have on the power to detect DE-genes.

As already indicated from the low fraction of assigned reads, kallisto has the lowest mean expression and the highest dropout rates (**Figure 2D** and **Supplementary Figure S5**) and, as expected from a high fraction of falsely mapped reads, BWA has the largest variance. To estimate the impact that these statistics have on the power to detect DE-genes, we use the mean-variance relationship to simulate data sets with DE-genes (**Figure 2D,E**). As previously reported[2], UMI protocols have a noticably higher power than Smart-seq2 (**Figure 2F**). Moreover for Smart-seq2, we find that kallisto performs slightly better than STAR, while for UMI-methods STAR performs better (**Figure 2F** and Supplementary Figure S7).

In summary, using BWA to map to the transcriptome introduces noise, thus considerably reducing the power to detect DE-genes as compared to genome alignment using STAR or the pseudo-alignment strategy kallisto, but given the lower mapping rate of kallisto STAR with GENCODE is generally preferable.

Many asymmetric expression changes pose a problem without spike-in data.

The next step in any RNA-seq analysis is the normalisation of the count matrix. The main idea here is that the resulting normalisation factors correct for differing sequencing depths. To begin with, we compare how much the estimated normalisation factors deviate from the truth. As long as there is only a small proportion of DE-genes or if the differences are symmetric, estimated size factors are not too far from the simulated ones and there are no large differences among methods (**Figure 3A** and **Supplementary Figure S8**). However with increasing asymmetry, normalisation factors deviate more and more and the single cell methods scran[24] and SCnorm[25] perform markedly better than the bulk methods TMM[26], MR[27] and Positive Counts as well as the single cell method Linnorm[28]. Census[29] is an outlier in that it has a constant deviation of 0.1, which is due to filling in 1 when library sizes could not be calculated.

To determine the effect of these deviations on downstream analyses, we evaluated the performance of differential expression inference using different normalisation methods (**Figure 3B**). Firstly, the differences in the TPR across normalisation methods are only minor, only Linnorm performed consistently worse (**Supplementary Figures S9**). In contrast, the ability to control the FDR heavily depends on the normalisation method (**Supplementary Figures S10**). For small numbers of DE-genes or symmetrically

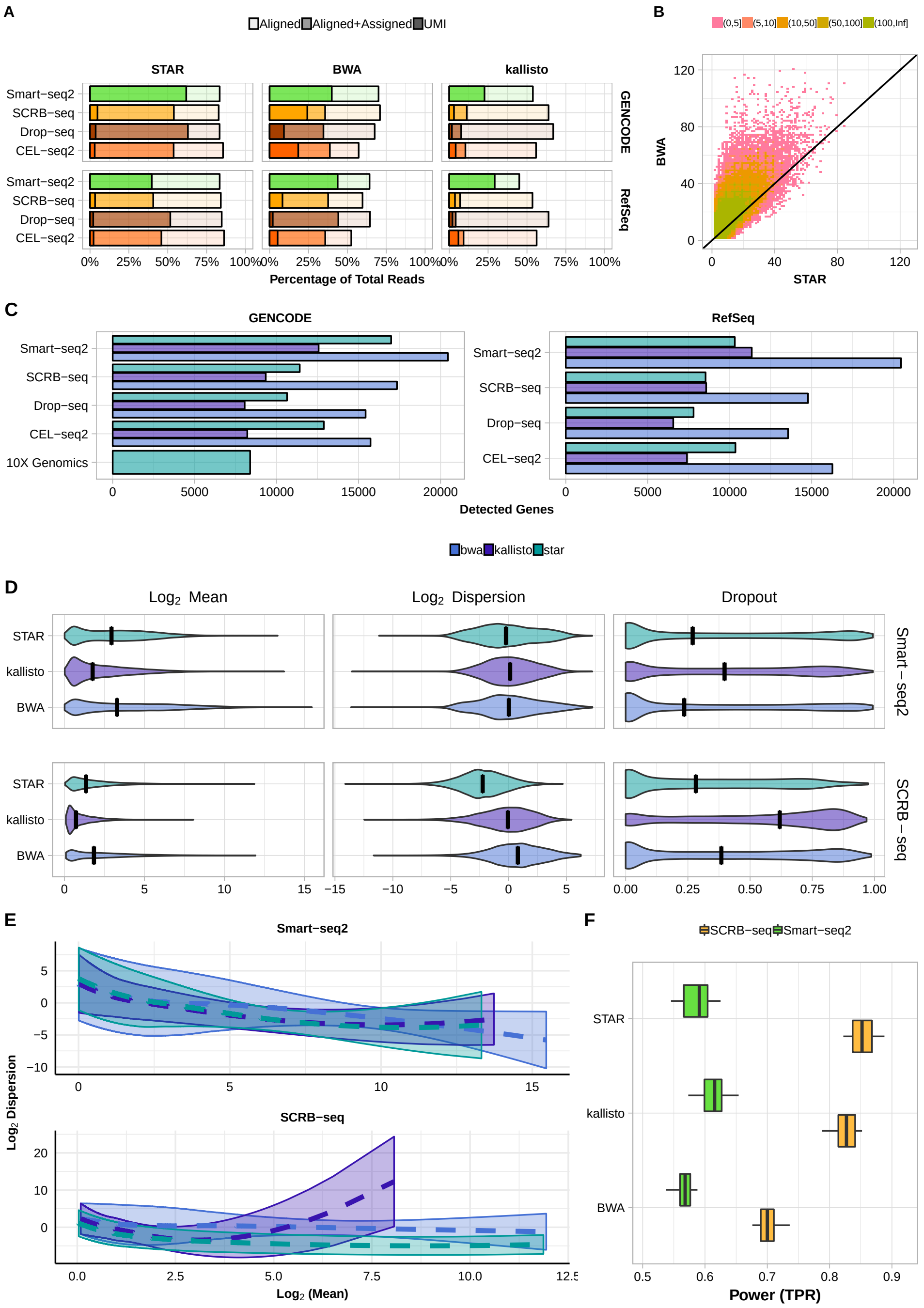


Figure 2. Expression Quantification.
A Read alignment and assignment rates per library preparation protocol stratified over aligner and annotation. The lighter shade represents the percentage of the total reads that could be aligned and the darker shade the percentage that also was uniquely assigned (see also Supplementary Figure S3). **B** Number of genes per UMI with >1 reads for BWA and STAR alignment using the SCRB-seq data set and GENCODE annotation. Colours denote number bins of UMIs. **C** Number of genes detected per Library Preparation Protocol stratified over Aligner and Annotation (i.e. at least 10 % nonzero expression values) (see also Supplementary Figure S4). **D** Estimated mean expression, dispersion and dropout rates for SCRB-seq and Smart-seq2 data using STAR, BWA or kallisto alignments with GENCODE annotation (see also Supplementary Figure S5). **E** Mean-dispersion fitting line applying a cubic smoothing spline with 95% variability bands for SCRB-seq and Smart-seq2 data using STAR, BWA or kallisto alignments with GENCODE annotation (see also Supplementary Figure S6). **F** The effect of quantification choices on the power (TPR) to detect differential expression stratified over library preparation and aligner. The expression of 10,000 detected genes over 768 cells (384 cells per group) were simulated given the observed mean-variance relation per protocol. 5% of the simulated genes are differentially expressed following a symmetric narrow gamma distribution. Unfiltered counts were normalised using scran. Differential expression was tested using limma-trend (see also Supplementary Figure S7).

distributed changes, the FDR is well controlled for all methods except Linnorm. However, with an increasing number and asymmetry of DE-genes, only SCnorm and scran keep FDR control, provided that cells are grouped or clustered prior to normalisation. In our most extreme scenario with 60% DE-genes and complete asymmetry, all methods except Census loose FDR control. SCnorm, scran Positive Counts and MR regain FDR control with spike-ins for 60% completely asymmetric DE-genes (**Supplementary Figure S10**). Given similar TPR of the methods, this FDR control determines the pAUC (**3B,C**).

Since in real data it is usually unknown what proportion of genes is DE and whether cells contain differing levels of mRNA, we recommend a method that is robust under all tested scenarios. Thus, for most experimental setups scran is a good choice, only for Smart-seq2 data without spike-ins, Census might be a better choice.

Imputation has little impact on pipeline performance.

If the main reason why normalisation methods perform worse for scRNA-seq than for bulk data is the sparsity of the count matrix, reducing this sparsity by either more stringent filtering or imputation of missing values should remedy the problem[30]. Here, we test the impact of frequency filtering and three imputation approaches (DrImpute[31], scone[32], SAVER[33]) on normalisation performance.

We find that simple frequency filtering has no effect on normalisation results (**Figure 3D**). Performance as measured by pAUC is identical to using raw counts. In contrast, imputation can have an effect on performance and there are large differences among methods. Imputation with DrImpute and scone rarely increased the pAUC and occasionally as in the case of SCRB-seq with MR normalisation, the pAUC even decreased by 100% and 76%, respectively due to worse FDR control relative to using raw counts (**Supplementary Figure 13**). In contrast, these two imputation methods achieved an appreciable increase in pAUC together with scran normalisation, $\sim$ 28%, 4% and 9% for 10X Genomics, SCRB-seq and Smart-seq2 data, respectively. SAVER on the other hand never made things worse, irrespective of data set and normalisation method but was able to rescue FDR control for MR normalisation of UMI data, even in a completely asymmetric DE-pattern.

These observations suggest that data sets with a high dropout rate might benefit more from imputation than data sets with a relatively low dropout rate (**Supplementary Figure S5**). Nevertheless, if a good normalisation method is used to begin with (e.g. scran with clustering), the improvement by imputation remains relatively small.

Good normalisation removes the need for specialised single cell DE-tools.

The final step in our pipeline analysis is the detection of DE-genes. Recently, Soneson and Robinson[30] benchmarked 36 DE approaches and found that edgeR[26], MAST[34], limma-trend[35] and even the T-Test performed well. Moreover, they found that for edgeR, it is important to incorporate an estimate of the dropout rate per cell. Therefore, we combine edgeR here with zingeR[36].

Both edgeR-zingeR and limma-trend in combination with a good normalisation reach similar pAUCs as using the simulated size factors (**Figure 4**). However, in the case of edgeR-zingeR this performance is achieved by a higher TPR paid while loosing FDR

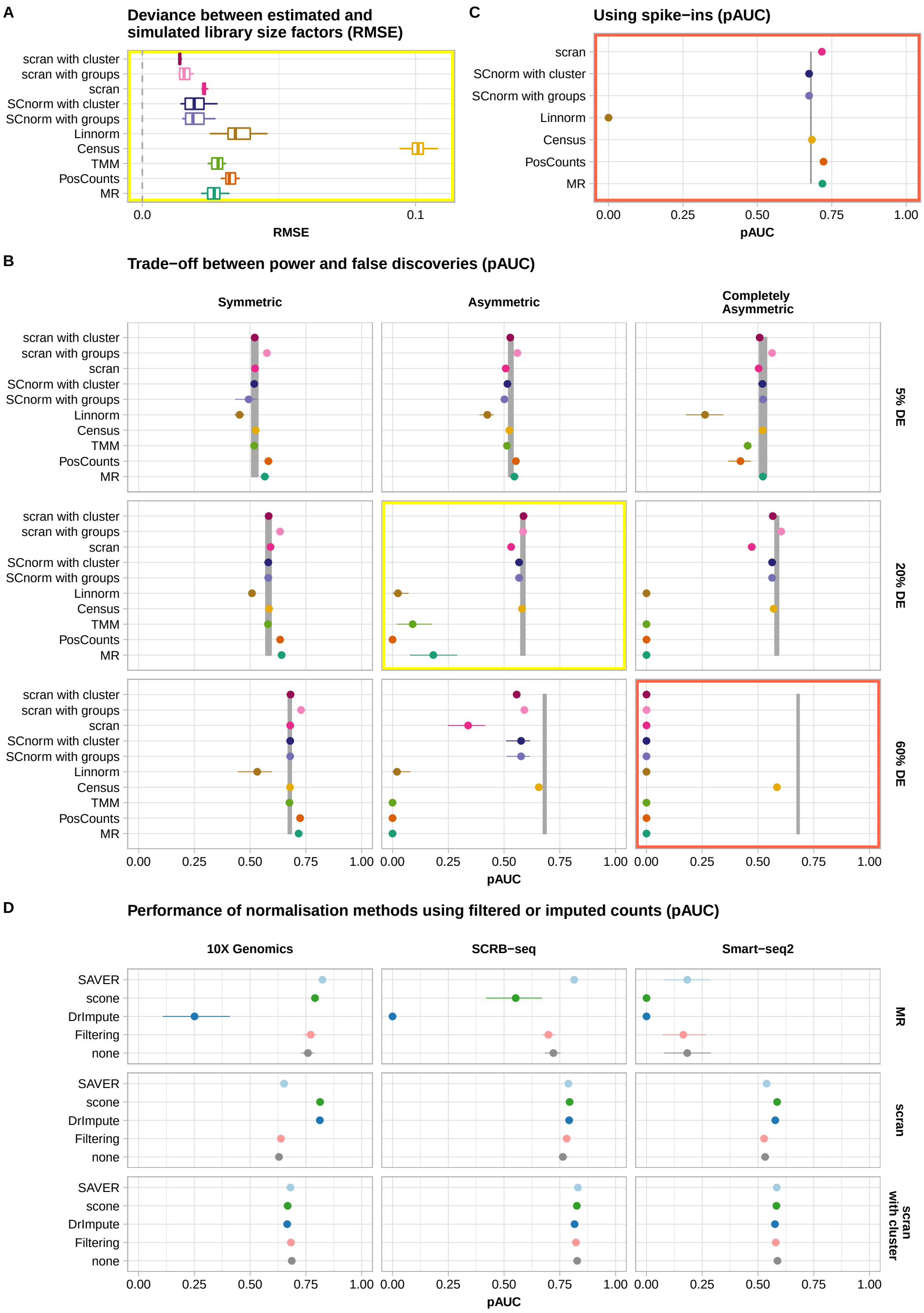


Figure 3. Normalisation choices determines DE-analysis performance, not preprocessing. The data in panels A-C are based on Smart-seq2 data, all panels are based on two groups of 384 cells, STAR alignment with GENCODE annotation was used for quantification. **A** The root mean squared error (RMSE) of estimated scaling factors per normalisation method is plotted for 20% asymmetric DE-genes (see also Supplementary Figure S8). **B** The discriminatory ability determined by the partial area under the curve (pAUC) based on DE testing with limma-trend for normalisation without spike-ins per DE-pattern. The grey ribbon indicates the pAUC given simulated size factors (see also Supplementary Figure S9-S11). **C** Using spike-ins for normalisation for 60% completely asymmetric DE-genes. **D** Effect of preprocessing data for 20% asymmetric DE-genes without spike-ins. Counts were either left asis ('none') or a preprocessing or imputation was applied prior to normalisation. The derived scaling factors were then used for normalisation and DE testing was performed on raw counts using limma-trend (see also Supplementary Figure S12-S14).

control (**see Supplementary Figure S16**), even in the case in symmetric DE-settings (**Supplementary Figure S18-S20**).

Nevertheless, we find that DE-analysis performance strongly depends on the normalisation method and on the library preparation method. In combination with the simulated size factors or scran normalisation, even a T-Test performs well. Conversely, in combination with MR or SCnorm, the T-Test has an increased FDR (**Supplementary Figure S13**). SCnorms bad performance with a T-Test was surprising given SCnorms good performance with limma-trend (**Figure 3B**). One explanation could be the relatively large deviation of SCnorm derived size factors (**Figure 3A** and **Supplementary Figure S8**) which inflate the expression estimates.

Furthermore, we find that MAST performs consistently worse than the other DE-tools when applied to UMI-based data, but -except in combination with SCnorm- it is doing fine with Smart-seq2 data.

In concordance with Soneson and Robinson[30], we found that limma-trend, a DE-tool developed for bulk RNA-seq data showed the most robust performance. Moreover, library preparation and normalisation appeared to have a stronger effect on pipeline performance than the choice of DE-tool.

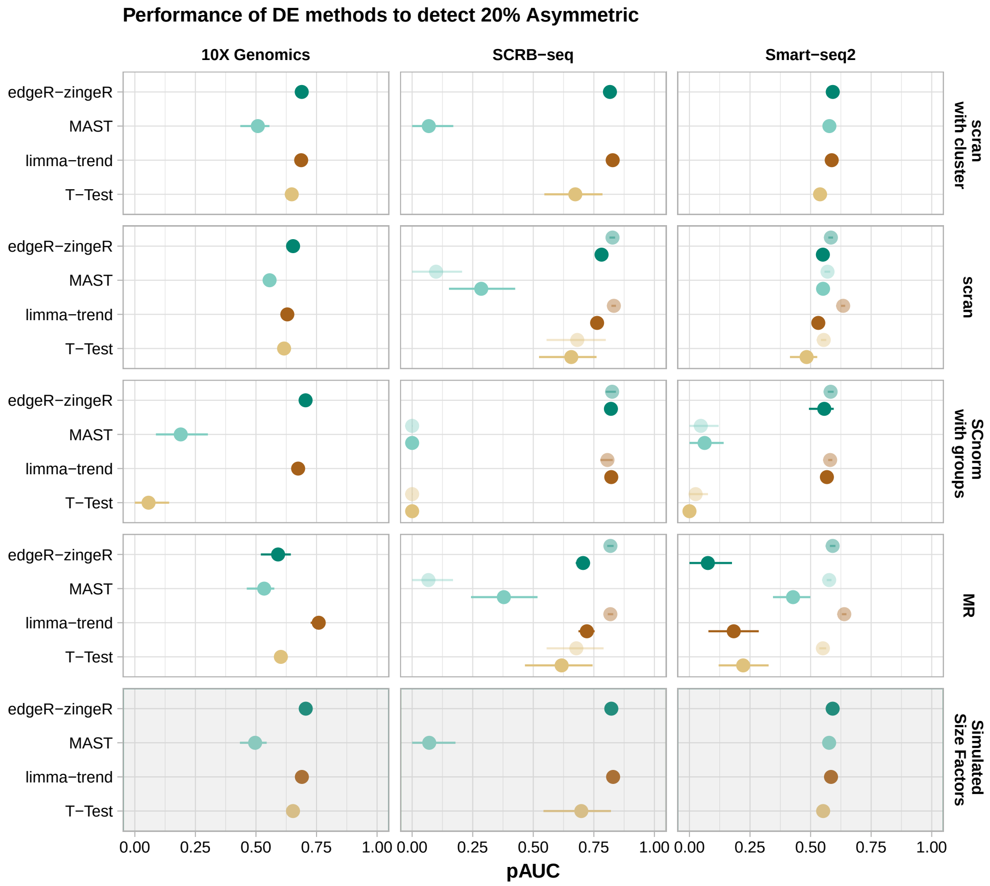


Figure 4. Evaluation of DE tools.
The expression of 10,000 genes over 768 cells (384 cells per group) were simulated given the observed mean-variance relation per protocol. 20% of the simulated genes are differentially expressed following an asymmetric narrow gamma distribution. Unfiltered counts were normalised using simulated library size factors or applying normalisation methods. Differential expression was tested using T-Test, limma-trend, MAST or edgeR-zingeR. The discriminatory ability of DE methods is determined by the partial area under the curve (pAUC) for the TPR-FDR curve (see also Supplementary Figure S15-S17).

Normalisation is overall the most influential step.

Because we tested a nearly exhaustive number of ∼3,000 possible scRNA-seq pipelines, starting with the choice of library preparation protocol and ending with DE-testing, we can estimate the contribution of each separate step to pipeline performance for our different DE-settings (**Figure 1 B**). We used a beta regression model to explain the variance in pipeline performance with the choices made at the seven pipeline steps 1) library preparation protocol, 2) spike-in usage, 3) alignment method, 4) annotation scheme, 5) preprocessing, 6) normalisation and 7) DE-tool as explanatory variables. We used the difference in pseudo-R^2 between the full model including all seven pipeline steps and leave-one-out reduced models to measure the contribution of each separate step to overall performance.

All pipeline choices together (the full model) explain $\sim 50\%$ and $\sim 60\%$ of the variance in performance, for 20% and 60% DE-genes, respectively (**Figure 5A**). Preprocessing choices contribute very little ($\Delta R^2 <= 1\%$). The same is true for annotation ($\Delta R^2 <= 2\%$) and aligner choices ($\Delta R^2 <= 5\%$). For aligner and annotation, it is important to note that these are upper bounds, because our simulations do not include differences in gene detection rates (**Figure 2C**).

Surprisingly, the DE-tools only matters for symmetric DE-setups ($\Delta R^2_{\text{DE}=0.2} = 15\%$; $\Delta R^2_{\text{DE}=0.6} = 11\%$), however the choice of library preparation protocol has an even bigger impact on performance for symmetric DE-setups ($\Delta R^2_{\text{Symmetric}} = 17 - 29\%$) and additionally for 5% asymmetric changes ($\Delta R^2_{5\%\ \text{Asymmetric}} = 17\%$). Normalisation choices have overall a large impact in all DE-settings ($\Delta R^2 = 12 - 38\%$), where the importance increases with increasing levels of DE-genes and increasing asymmetry. Spike-ins are only necessary if many asymmetric changes are expected and have little or no impact if only 5% of the genes are DE or the changes are symmetric (**Figure 5A**). Moreover, for completely asymmetric DE-patterns, the regression model did not converge without normalisation and spike-ins, because their absence or presence alone pushed the MCCs to the extremes.

For the best performing pipeline *SCRB-seq + STAR + GENCODE + SAVER imputation + scran with clustering + limma-trend*, using 384 cells per group instead of 96 improves performance only by 6.5-8%. Sample size is more important if a naive pipeline is used. For *SCRB-seq + BWA + GENCODE + no preprocessing + MR + T-Test* the performance gain by increasing sample size is 10-12% and even worse, for many asymmetric DE-genes, lower sample sizes occasionally appear to perform better (**Figure 5B** and **Supplementary Figure S21**).

In summary, we identify normalisation and library preparation as the most influential choices and the observation that differences in computational steps alone can significantly lower the required sample size nicely illustrates the importance of bioinformatic choices.

Discussion

Here we evaluate the performance of complete computational pipelines for the analysis of scRNA-seq data under realistic conditions with large numbers of DE-genes and differences in total mRNA contents between groups (**Figure 1**). Furthermore, our simulations allow us not only to investigate the influence of choices made at each pipeline step separately, but also to estimate the relative importance and interactions between different steps

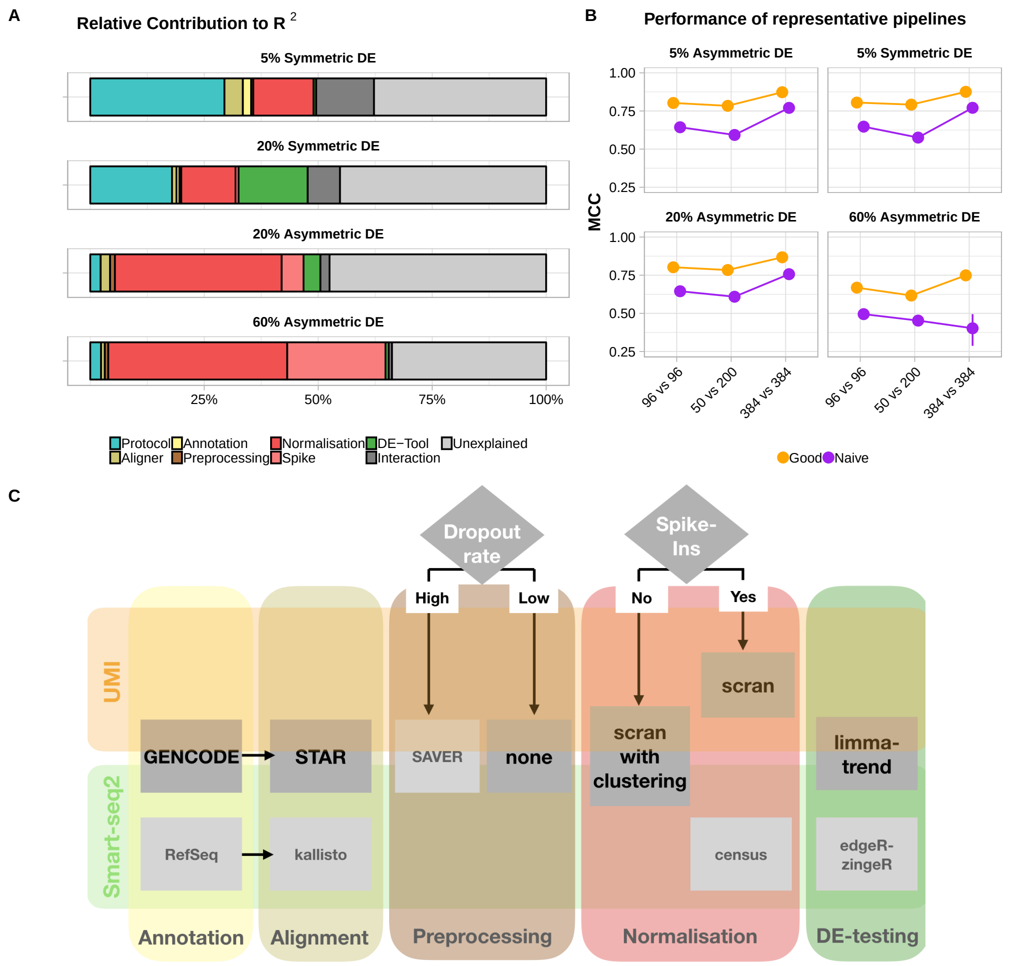


Figure 5. Evaluation of analysis pipeline. The expression of 10000 genes over 768 cells were simulated and 5%, 20% or 60% of the genes were differentially expressed following a symmetric or asymmetric narrow gamma distribution. This simulation setup was applied to protocols, alignments, annotations, preprocessing, normalisation and DE tools. For each analysis set, the Matthew Correlation Coefficient was averaged over 20 simulations and rescaled to [0,1] interval. The MCC was used as a response variable in beta regression models with log-log link function.
A The contribution of each covariate in the full model (Protocol + Aligner + Annotation + Preprocessing + Normalisation + DE-Tool). **B** Performance according to sample size, 1 good and 1 naive pipeline (see also Supplementary Figure S21). **C** Pipeline recommendations for UMI and Smart-seq2 data.

of an entire scRNA-seq analysis pipeline. We implemented all assessed computational methods and more in powsimR, so that users can easily evaluate pipeline performance given their own data and expected DE-settings.

Beginning with the creation of the raw count matrix, we find that transcriptome mapping with BWA[18] appears to recover the largest number of genes. However, many of these are probably due to falsely mapped reads, also increase expression variance which ultimately results in a lower sensitivity (**Figure 2C-F**). In contrast, the pseudo-alignment method kallisto[23] appears to assign reads precisely, but looses a lot of reads or UMIs with 3' UMI-data. One possible explanation is that the 3'end of a gene alone often does not allow to distinguish different transcripts, thus leading to a lower gene detection rate and mean expression. Finally, a genome mapping approach using the splice-aware aligner STAR[17] in conjunction with GENCODE annotation recovers the most genes with the highest accuracy (**Figure 5C**).

Concerning the preprocessing step, we found in concordance with Andrews and Hemberg[37] that in particular for sparse data such as 10X, SAVER[33] imputation before normalisation improves performance, while filtering genes has no effect with our data sets and combinations of normalisation and DE-testing methods.

The choice that had the largest impact on performance throughout all tested DE-settings is the choice of normalisation method. Only for symmetric changes, the choice of library preparation protocol had a slightly larger impact than normalisation. In line with Evans et al.[11], we found that normalisation performance of bulk methods and also some of the single cell methods declined with asymmetry (**Figure 3B**). In particular, for 60% completely asymmetric DE-genes only Census retained FDR control. Unfortunately, Census is not recommended for the use with UMI-counts. Thus, for UMI-counts and 60% completely asymmetric changes, only the use of spike-ins could restore test performance. Thus, in the debate about the usefulness of spike-ins[38,39], we land on the pro side: Our simulations clearly show that spike-ins are useful in DE-testing settings with asymmetric changes which is likely to be a common phenomenon in scRNA-seq data. Due to good performance across DE-settings and its speed (**Supplementary Figure S22**) we would recommend scran with prior clustering as the best choice for normalisation (**Figure 5C**).

The choice in DE-testing method, our final pipeline step had relatively little impact on overall pipeline performance. A good normalisation prior to DE-testing alleviates the need for more complex and thus vulnerable methods, such as for example MASTs hurdle model which implicitly assumes a zero inflated negative binomial distribution of the count data. Indeed, in Vieth et al.[10] we showed that also scRNA-seq data fit a negative binomial distribution rather well and that the previously reported zero-inflation in scRNA-seq data is mainly due to amplification noise which is removed in UMI-data. Hence, it is not surprising that in concordance with Soneson and Robinson[30], we find that relatively straight forward DE-testing methods adapted from bulk RNA-seq perform well with scRNA-seq data.

Finally, we want to remark that paying attention to the details in a computational pipeline and in particular to normalisation pays off. The effect of using a good pipeline as compared to a naively compiled one has a similar or even greater effect on the potential to detect a biological signal in scRNA-seq data as an increase in cell numbers from 96 to 384 cells per group (**Figure 5B**).

Online Methods

Single Cell RNA-seq Data Sets

The starting point for our comprehensive pipeline comparison is the scRNA-seq library preparation (**Figure 1 A**). In our comparison, we included the gene expression profiles of mouse embryonic stem cells (mESC) as published in[2] (**Supplementary Figure S1**). We selected four data sets for our comparison: Smart-seq2[13] a well-based full-length scRNA-seq protocol, CEL-seq2[15] a well-based 3' UMI-protocol using linear amplification, SCRB-seq a well-based 3' UMI-protocol with PCR amplification[40,2] and Drop-seq [14] a droplet-based 3' UMI-protocol. In addition, 92 poly-adenylated synthetic RNA transcripts of known concentration designed by the External RNA Control Consortium (ERCCs)[41] were spiked in for all methods except Drop-seq. All raw cDNA sequencing reads were cut to a length of 45 bases and downsampled to one million cDNA reads per cell (**Supplementary Table S1** and **Supplementary Figure S1**).

Finally, due to its popularity, we added a fifth data set from 10X Genomics Support (mouse NIH3T3 cells) generated on the 10X Genomics platform, namely the expression profiles of 1k 1:1 mixture of fresh frozen human (HEK293T) and mouse (NIH3T3) cells generated using the v2 3' gene expression chemistry[16]. We proceeded with $\sim$ 400 mouse cells that had $\sim$ 60,000 reads/cell on average. These choices of library preparation protocols cover the diversity in current protocols without imposing partiality due to biological differences and technical handling.

Gene Expression Quantification

For genome mapping and quantification of the UMI-data, we used the zUMIs[42] (v.0.0.3) pipeline with STAR[17] (v.2.5.3a) and the mouse genome (Mus_musculus.GRm38) together with annotation files for GENCODE (vM15), Vega (VEGA68) and RefSeq (Release 85) (**Supplementary Table S2**). For Smart-Seq2 we use the same pipeline settings as in zUMIs, simply omitting the UMI collapsing step (**Supplementary Table S3**).

For transcriptome alignment, we downloaded transcriptome fasta files for each annotation release. We used BWA[18] (v0.7.12) alignment with assignment of reads to features. Here, we define a feature as a gene including all its associated exons. In the next step reads that aligned equally well to multiple different genes were filtered and the remaining reads were tallied up per gene. For UMI data, the reads were collapsed.

For kallisto[23] (v0.43.1), a transcriptome-guided pseudo-alignment method, we followed the recommended quantification procedure to yield abundance estimates per equivalence class and subsequently back-transformed to estimates per genes using custom R scripts, again filtering out equally good alignments to multiple different genes. For SCRB-seq, CEL-seq2 and Drop-seq libraries, we chose the UMI-aware quantification option. The ERCC spike-in sequences were appended to the genome or transcriptome sequences for quantification.

For the 10X data set we did not download the raw reads, but a UMI count matrix were obtained using the Cell Ranger pipeline. This should yield similar results as zUMIs with STAR as alignment method and GENCODE annotation.

Simulations

We used powsimR to estimate, simulate and evaluate single cell RNA-seq experiments [10]. PowsimR has been independently validated for benchmarking DE-approaches[30]. The gene expression quantification using three different aligners in combination with three annotations per library preparation protocol produced 36 count matrices plus one 10X count matrix. These count matrices are the basis for our estimation in powsimR. Genes needed at least one read or UMI count in at least one cell to be considered in the estimation for simulation parameters. To simulate spike-in data, we added an implementation of the simulation framework for pure technical variation of spike-ins described in Kim et al.[43] to powsimR. The parameters required for these simulations were estimated from 92 ERCC spike-ins in the SCRB-seq, CEL-seq2 and Smart-seq2 data, respectively[2].

For a detailed evaluation of the pipelines, we simulated two groups of cells for pairwise comparisons with the following three sample size setups: 96 vs. 96, 384 vs. 384 or 50 vs. 200 cells (**Figure 1B**). For simplicity, we kept the number of genes that we simulated constant at 10,000. Furthermore, the two groups of cells can have 5%, 20% or 60% differentially expressed genes. To capture the asymmetry of observed expression differences, we considered three setups of DE-patterns: symmetric (50% up- and 50% down-regulated), asymmetric (75% up- and 25% down-regulated) or completely asymmetric (100% up-regulated). The combination of these parameters results in a total of 27 DE-setups that were then applied to the parameter estimates from 37 different count matrices to simulate 20 replicates for each setting, producing a total of 19,980 simulated data sets.

These data sets were then analysed by a nearly exhaustive number of combinations of five filtering and imputation strategies (gene dropout filtering, scone, SAVER, DrImpute) together with seven normalisation approaches (TMM, MR, Linnorm, Census, Linnorm, scran, SCnorm) with or without spike-ins, depending on library preparation protocol and method (**Figure 1C**). The derived normalisation scaling factors were then used in conjunction with the raw count matrices for DE-testing using four representative approaches (T-Test, limma-trend, edgeR-zingeR, MAST). The resulting p-values were corrected for multiple testing with Benjamini-Hochberg FDR and we applied a threshold level of 10% to define positive test results. All these steps were seamlessly implemented into powsimR (github: https://github.com/bvieth/powsimR). In total we analysed 2,979 different RNA-seq pipelines.

Evaluation metrics

To evaluate the normalisation results, we determined the root mean squared error (RMSE) of a robust linear model using the difference between estimated and simulated library size factors as response variable in *rlm()* implemented in R-package MASS[44] (v.7.3-51.1).

All other measures are based on the final results of an entire scRNA-seq pipeline ending with DE-testing. Knowing the identity of the genes that were simulated to show differing expression levels and the results of the DE-testing, we used a number of metrics related to the confusion matrix tabulating the number of true positives, false positives, true negatives and false negatives. We define the power to detect differential expression with the TPR ($TPR = \frac{TP}{TP+FN}$). The false discovery rate is defined as $FDR = \frac{FP}{FP+TP}$.

We combine these two measures in a TPR versus FDR curve to quantify the trade-off between true and false discoveries in a genome-wide multiple testing setup as advocated by[45]. We then summarise these curves by their partial area under curve (pAUC) of TPR versus observed FDR that still ensures FDR control at the nominal level of 10%. This way of calculating the AUC is ideal for data with relatively high true negative rates as the partial integration does not punish methods that are over-conservative, i.e. that stay way below the nominal FDR.

To summarise the whole confusion matrix in one representative value we chose the Matthews Correlation Coefficient ($MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$), because it is a balanced measure ensuring a reliable comparison of method performance across all DE-settings[45,46]. As for the pAUC, we calculated the maximal value of MCC where the cutoff still ensured FDR control at the nominal level of 10%.

To quantify the relative contribution of each step in the analysis pipeline, we used the MCC as a response variable in a beta regression model implemented in R-package betareg (v.3.1-1)[47] with each individual pipeline step. Because the MCC assumes the extremes of 0 and 1 in some DE-settings, we applied the recommended transformation, namely $MCC_{transformed} = \frac{MCC*(n-1)+0.5}{n}$, where n is the sample size[48]. The contribution is then given by the difference between the full model $pseudo - R^2$ containing all covariates versus a model leaving one step out at a time. This is then scaled to the total variance explained to give relative ΔR^2 percentages.

Author Contributions

B.V. and I.H. conceived the study. B.V. prepared and analysed the scRNA-seq data. B.V. implemented and conducted the simulation and evaluation framework. S.P.and C.Z. helped in data processing and power simulations. W.E. and I.H. supervised the work and provided guidance in data analysis. B.V., I.H., and W.E. wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through LMUexcellent, SFB1243 (Subproject A14/A15) and DFG grant HE 7669/1-1. C.Z. is recipient of an EMBO long-term fellowship (ALTF 673-2017).

References

1. Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, 34(11):1145–1160, November 2016. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.3711.

2. Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, 65(4):631–643.e4, February 2017. ISSN 1097-2765, 1097-4164. doi: 10.1016/j.molcel.2017.01.023.

3. Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods*, March 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4220.

4. Giacomo Baruzzo, Katharina E Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A FitzGerald, and Gregory R Grant. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods*, 14(2):135–139, February 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4106.

5. Douglas C Wu, Jun Yao, Kevin S Ho, Alan M Lambowitz, and Claus O Wilke. Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics*, 19(1):510, July 2018.

6. Shanrong Zhao and Baohong Zhang. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, 16: 97, February 2015. ISSN 1471-2164. doi: 10.1186/s12864-015-1308-8.

7. Tallulah S Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Res.*, 7, November 2018. doi: 10.12688/f1000research.16613.1.

8. Lihua Zhang and Shihua Zhang. Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, June 2018. ISSN 1545-5963, 1557-9964. doi: 10.1109/TCBB.2018.2848633.

9. Catalina A Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods*, 14(6):565–571, June 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4292.

10. Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimr: Power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, July 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx435.

11. Ciaran Evans, Johanna Hardin, and Daniel M Stoebel. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.*, February 2017. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbx008.

12. Amit Zeisel, Ana B Muñoz Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, February 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaa1934.

13. Simone Picelli, Omid R Faridani, Asa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.*, 9(1):171–181, January 2014. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2014.006.

14. Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, May 2015. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2015.05.002.

15. Tamar Hashimshony, Florian Wagner, Noa Sher, and Itai Yanai. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.*, 2(3):666–673, September 2012. ISSN 2211-1247. doi: 10.1016/j.celrep.2012.08.003.

16. Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, January 2017. ISSN 2041-1723. doi: 10.1038/ncomms14049.

17. Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/bts635.

18. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btp324.

19. Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. kallisto. `https://github.com/pachterlab/kallisto/tree/v0.43.1`, August 2017.

20. Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S Joardar, Vamsi K Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M McGarvey, Michael R Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H Rangwala, Daniel Rausch, Lillian D Riddick, Conrad Schoch, Andrei Shkeda, Susan S Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E Tully, Anjana R Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D Murphy, and Kim D Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, 44(D1):D733–45, January 2016. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv1189.

21. Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, 47(D1):D766–D773, January 2019.

22. L G Wilming, J G R Gilbert, K Howe, S Trevanion, T Hubbard, and J L Harrow. The vertebrate genome annotation (vega) database. *Nucleic Acids Res.*, 36(Database issue):D753–60, January 2008.

23. Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527, May 2016. ISSN 1087-0156. doi: 10.1038/nbt.3519.

24. Aaron T L Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, 17:75, April 2016. ISSN 1465-6906. doi: 10.1186/s13059-016-0947-7.

25. 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. *Nat. Methods*, April 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4263.

26. Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, 11(3):R25, March 2010. ISSN 1465-6906. doi: 10.1186/gb-2010-11-3-r25.

27. Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106, 2010. ISSN 1465-6906.

28. Shun H Yip, Panwen Wang, Jean-Pierre A Kocher, Pak Chung Sham, and Junwen Wang. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.*, September 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx828.

29. Xiaojie Qiu, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. Single-cell mRNA quantification and differential analysis with census. *Nat. Methods*, January 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4150.

30. Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, 15(4):255–261, April 2018. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4612.

31. Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J Garry. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, 19(1):220, June 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2226-y.

32. Michael B Cole, Davide Risso, Allon Wagner, David DeTomaso, John Ngai, Elizabeth Purdom, Sandrine Dudoit, and Nir Yosef. Performance assessment and selection of normalization procedures for Single-Cell RNA-seq. May 2018.

33. 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. *Nat. Methods*, June 2018. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-018-0033-z.

34. Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, Peter S Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, 16(1):1–13, December 2015. ISSN 1465-6906, 1474-760X. doi: 10.1186/s13059-015-0844-5.

35. Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, 15(2):R29, February 2014. ISSN 1465-6906. doi: 10.1186/gb-2014-15-2-r29.

36. Koen Van den Berge, Fanny Perraudeau, Charlotte Soneson, Michael I Love, Davide Risso, Jean-Philippe Vert, Mark D Robinson, Sandrine Dudoit, and Lieven Clement. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.*, 19(1):24, February 2018. ISSN 1465-6906, 1474-760X. doi: 10.1186/s13059-018-1406-4.

37. Tallulah S Andrews and Martin Hemberg. Identifying cell populations with scRNASeq. *Mol. Aspects Med.*, July 2017. ISSN 0098-2997, 1872-9452. doi: 10.1016/j.mam.2017.07.002.

38. Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, 12:480, December 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-480.

39. Christoph Ziegenhain, Beate Vieth, Swati Parekh, Ines Hellmann, and Wolfgang Enard. Quantitative single-cell transcriptomics. *Brief. Funct. Genomics*, March 2018. ISSN 2041-2649, 2041-2657. doi: 10.1093/bfgp/ely009.

40. 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*, March 2014. doi: 10.1101/003236.

41. Lichun Jiang, Felix Schlesinger, Carrie A Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.*, 21(9):1543–1551, September 2011. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.121095.111.

42. Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann. zUMIs - a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience*, May 2018. ISSN 2047-217X. doi: 10.1093/gigascience/giy059.

43. Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Tomislav Illicic, Sarah A Teichmann, and John C Marioni. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.*, 6:8687, October 2015. doi: 10.1038/ncomms9687.

44. W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.

45. Charlotte Soneson and Mark D Robinson. iCOBRA: open, reproducible, standardized and live method benchmarking. *Nat. Methods*, 13(4):283, April 2016. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3805.

46. Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PLoS One*, 12(6):e0177678, June 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0177678.

47. Francisco Cribari-Neto and Achim Zeileis. Beta regression in R. *Journal of Statistical Software*, 34 (2):1–24, 2010.

48. Michael Smithson and Jay Verkuilen. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Methods*, 11(1):54–71, March 2006. ISSN 1082-989X. doi: 10.1037/1082-989X.11.1.54.