bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Missing Data and Technical Variability in Single-Cell RNA-Sequencing Experiments

Stephanie C. Hicks[1,2] , F. William Townes[2,] Mingxiang Teng[1,2], Rafael A. Irizarry[1,2,*]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute,

[2]Department of Biostatistics, Harvard School of Public Health

[*]Corresponding Author

Emails:

Stephanie C. Hicks, shicks@jimmy.harvard.edu

F. William Townes, will.townes@gmail.com

Mingxiang Teng, mxteng@jimmy.harvard.edu

Rafael A. Irizarry, rafa@jimmy.harvard.edu

Abstract

Until recently, high-throughput gene expression technology, such as RNA-Sequencing (RNA-seq) required hundreds of thousands of cells to produce reliable measurements. Recent technical advances permit genome-wide gene expression measurement at the single-cell level. Single-cell RNA-Seq (scRNA-seq) is the most widely used and numerous publications are based on data produced with this technology. However, RNA-Seq and scRNA-seq data are markedly different. In particular, unlike RNA-Seq, the majority of reported expression levels in scRNA-seq are

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

zeros, which could be either biologically-driven, genes not expressing RNA at the time of measurement, or technically-driven, gene expressing RNA, but not at a sufficient level to detected by sequencing technology. Another difference is that the proportion of genes reporting the expression level to be zero varies substantially across single cells compared to RNA-seq samples. However, it remains unclear to what extent this cell-to-cell variation is being driven by technical rather than biological variation. Furthermore, while systematic errors, including batch effects, have been widely reported as a major challenge in high-throughput technologies, these issues have received minimal attention in published studies based on scRNA-seq technology. Here, we use an assessment experiment to examine data from published studies and demonstrate that systematic errors can explain a substantial percentage of observed cell-to-cell expression variability. Specifically, we present evidence that some of these reported zeros are driven by technical variation by demonstrating that scRNA-seq produces more zeros than expected and that this bias is greater for lower expressed genes. In addition, this missing data problem is exacerbated by the fact that this technical variation varies cell-to-cell. Then, we show how this technical cell-to-cell variability can be confused with novel biological results. Finally, we demonstrate and discuss how batch-effects and confounded experiments can intensify the problem.

Keywords

sparsity, censoring, missing not at random (MNAR), genomics, single-cell RNA-Sequencing, confounding

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Introduction

Single-cell RNA-Sequencing (scRNA-seq) has become the primary tool for profiling the transcriptomes of hundreds or even thousands of individual cells in parallel. In contrast to the standard RNA-seq approach, which is applied to samples containing hundreds of thousands of cells and therefore measures average gene expression level across cells, scRNA-seq measures gene expression in a single cell. To distinguish these two technologies we refer to the latter as *bulk RNA-seq*. Today scRNA-seq is increasingly being used across a diverse set of biomedical applications such as profiling the transcriptomes of differentiated cell types[1-4], profiling the changes in cell states[5, 6], identifying allele-specific expression[7, 8], spatial reconstruction[9, 10] and the classification of subtypes[11-13].

While scRNA-seq data provides a new level of data resolution, it also results in a larger number of genes reporting the expression level to be zero, or practically zero, as compared to using bulk RNA-seq[14]. A gene reporting the expression level to be zero can arise in two ways: (1) the gene was not expressing any RNA at the time the cell was experimentally isolated and processed prior sequencing (referred to as structural zeros[15]) or (2) the gene was expressing RNA in the cell at the time of isolation, but not at a sufficient level to be detected in the experimental procedure to capture and process the RNA prior to sequencing (referred to as dropouts[15, 16]). While the former is a type of biological event, the latter is purely technical as it stems from the limitations of current experimental protocols to detect low amounts of RNA in a cell, referred to as capture efficiency[17].

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Batch effects are commonly found in high-throughput data[18] and given the way that scRNA-seq experiments are conducted, there is much room for concern regarding confounding[18]. Specifically, batch effects in scRNA-seq experiments occur when cells from one biological group or condition are cultured, captured and sequenced separate from cells in a second condition (Figure S1). However, due to the nature of certain experimental scRNA-seq protocols, which restrict the way cells are captured and sequenced separately, sometimes standard balanced experimental designs are not possible[14, 18-20]. This reality makes it particularly important to be cautious about the potential for correlated variability induced by technical factors.

The unwanted variability introduced by batch effects can be particularly troublesome in scRNA-seq data because one of the most common applications has been the use unsupervised learning methods[21-26] such as data exploration after dimensionality reduction or clustering to identify novel or rare subpopulations of cells[11-13]. Although a diverse set of techniques are used in these papers, both linear dimensionality reduction techniques, such as principal component analysis (PCA)[21], and non-linear ones, such as t-Stochastic Neighbor Embedding (t-SNE)[26], rely on computing distances between the cell expression profiles. Given that the majority of genes in a cell report the expression level to be zero and that the proportion of zeros varies greatly from cell to cell, it is not surprising that the distance estimates between cells are greatly influenced by the proportion of zeros[27, 28]. However, it remains unclear to what extent this cell-to-cell variation is being driven by technical rather than biological variation.

We begin this article by describing the publicly available scRNA-seq data sets we used, which includes studies with only scRNA-seq data and studies with scRNA-seq and a matched bulk

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RNA-seq sample measured on the same population of cells. In the next section, we survey a large number of published scRNA-seq studies and illustrate the wide range of variation in the proportion of genes reporting the expression level to be zero across cells and studies (Section 3.1). Then, we present evidence that some of these reported zeros are driven by technical variation by demonstrating that scRNA-seq produces more zeros than expected and that this bias is greater for lower expressed genes (Section 3.2). In addition, we show that the consequences of this missing data problem are exacerbated by the fact that the technical variation of the probability of a gene being detected varies from cell to cell. Then, we illustrate that the proportion of genes reporting the expression level to be zero is a major source of cell-to-cell variation and this variability is partly driven by a mathematical artifact related to the transforming data in the original scale, but computing distances in the log scale (Section 3.3). Finally, we consider several case studies showing how differences in the detection rates can be driven by batch effects, which in turn can result in the false discovery of new groups (Section 3.4).

2. Data Description

A scRNA-seq experiment typically involves randomly sampling and capturing single cells from a population of cells, isolating the mRNA from the individual cells, reverse transcribing the RNA into cDNA, and sequencing the cDNA using massively parallel sequencing technologies[19, 29, 30]. Strengths and weaknesses of different scRNA-seq experimental protocols vary[31-33] in the cost per cell, the sensitivity to capture and convert RNA to cDNA and the accuracy to quantify the concentration of RNA, leading to differences in the number of cells sequenced per study and the number of features detected per cell. This experimental process is particularly challenging and laboratory protocols are still under intense development.

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.1 scRNA-seq data sets

We examine fifteen publicly available scRNA-seq data sets that included at least 200 samples with preprocessed and normalized expression data available on GEO[34] (Table 1). These data sets were created using six different scRNA-seq protocols for sequencing[1, 12, 35-39] and five studies include the use of unique molecular identifiers[40] (UMIs) for counting specific cDNA molecules. For the ten studies not using UMIs we examine the data as submitted to GEO, with one exception[11]. These ten studies reported measurements in either Transcripts per Million[41] (TPM), Reads Per Kilobase of transcript per Million mapped reads[42] (RPKM) or Fragments Per Kilobase of transcript per Million mapped reads[43] (FPKM), so each sample was corrected for gene length and library size. The one exception[11] uploaded data that was de-trended so that measurements for each gene averaged to zero across cells. For this particular study, we downloaded the raw sequencing files data from the Sequence Read Archive (SRA)[44] and computed expression in TPM units using Kallisto[45]. In the studies that used UMIs for molecule counting[1, 3, 9, 12, 46], the data uploaded to GEO was not normalized for library size, so to assure that these data were in similar units to the rest of our studies, we followed a published procedure[1] that normalizes each gene or transcript count by dividing by the total number of UMIs per cell and multiplies by a scaling factor (10^6). We refer to this unit as Counts Per Million (CPM).

Although details of the experimental protocols, which can help define groupings that may lead to technical batch effects, are not always included in the annotations that are publicly available, one can extract informative variables from the raw sequencing (FASTQ) files[47]. Namely, the sequencing instrument used, the run number from the instrument and the flow cell lane. Although the sequencing is unlikely to be a major source of unwanted variability, it serves as a

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

surrogate for other experimental procedures that very likely do have an effect, such as the starting amount of RNA in a cell, capture efficiency, PCR amplification reagents/conditions, and cell cycle stage of the cells[6, 48-50]. Here we will refer to the resulting differences induced by different groupings of these sources of variability as *batches*.

2.2 scRNA-seq data sets with matched bulk RNA-seq data

To help determine if the increased proportion of zero in scRNA-seq is explained by biology or technical biases, we examined three publicly available scRNA-seq data sets[5, 51, 52] that included a matched bulk RNA-seq sample measured on the same population of cells with preprocessed and normalized expression data available on GEO. One of these studies[5] is one of the 15 studies described in the previous subsection. The other two studies[51, 52] sequenced only 18 and 96 cells, respectively, thus were not included in the 15 large studies.

3. Results

3.1 The proportion of reported zeros varies from cell to cell and from study to study

We define the *detection rate* as the proportion of genes in a cell reporting the expression levels greater than a predetermined threshold δ. In this paper, we used δ=1 based on exploratory data analysis as this revealed two clear modes in the gene expression distribution (Figure S2), which we interpreted to be associated with background noise and signal respectively, with the lower mode defined as values below or equal to a TPM, FPKM, RPKM or CPM threshold of δ=1. This threshold has been previously used by Shalek et al. (2014)[6] and accommodates the bimodality[6, 16, 51, 53, 54] of scRNA-seq data that is not found in bulk RNA-Seq. We found wide variation in the detection rate across cells in all studies: from <1% detected to 65% (Figure 1). Similar results

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were obtained if we set δ=0 (Figure S3). For studies including groups known to have different gene expression profiles, we stratified by biological group to minimize the possibility of a biological explanation and also found wide variation (Figures S4-S9). We also note the detection rate is not necessarily dependent on sequencing depth (Figure S10).

3.2 scRNA-seq data contains more zeros than what is expected from biological variation

To demonstrate that there are more zeros in scRNA-seq data than what is expected from biological variation, we examined the gene expression of cells measured both on scRNA-seq and bulk RNA-seq. We found a bias consistent with a technical explanation. The details follow.

Denote the expression level for the g^{th} gene and i^{th} cell as x_{gi} where $i = 1, \ldots, n$. The expression for the g^{th} gene in bulk tissue composed of these cells will then be:

$$e_g = \sum_{i=1}^{n} x_{gi}.$$

Bulk RNA-seq produces an estimate proportional to this quantity and includes measurement error (ε_g):

$$Y_g = K_{bulk} e_g + \varepsilon_g.$$

Here K_{bulk} is a normalizing constant needed to account for the fact that experimental protocols and normalization procedures are adjusted to assure that the average or sum of measurements from each experiment are approximately the same. Since a tissue sample will have millions of cells, we consider *n* to be large enough to be treated as infinity. Note that some of these x_{gi} can be zero even when e_g is a large number. In fact, this is part of the biological explanation for why

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

single cell measurements have more zeros: an gene appearing expressed in bulk RNA-seq need not be expressed in every single cell at the time the cells were isolated and measured.

In a single cell experiment, we take a random sample of N cells from the population. We denote the expression values for these as X_{gi} where $i = 1, \ldots, N$. Using scRNA-seq technology, we obtain measurements:

$$Z_{gi} = K_{SC} X_{gi} + \eta_{gi} \qquad \text{if } X_{gi} > 0 \text{ and 0 otherwise.}$$

Here K_{SC} is the normalizing constant and η_{gi} is measurement error. Because the single cell data is a random sample, it follows that

$$E\left[\sum_{i=1}^{N} X_{gi}\right] = e_g$$

and therefore, if there is no biased induced from dropouts,

$$E\left[\sum_{i=1}^{N} Z_{gi} | \, Y_g = e\right] = \beta_0 + \beta_1 e$$

is a linear function with β_0 and β_1 determined by the normalization constants and the variance of the measurement error, which has been reported to be relatively low. While in a typical scRNA-seq experiment, bulk RNA-seq measurements from the same tissue is not available, the three studies described in Section 2.2 with both bulk RNA-seq and scRNA-Seq from the same biological specimens permits us to check if this relationship holds.

As evidence that scRNA-Seq technology is working as expected, previous publications plot $\frac{1}{N}\sum_{i=1}^{N} Z_{gi}$ versus Y_g for each gene to show it generally follows a linear relationship with reported correlations around 0.80 (see, for example, Figure 1C in Shalek et al. (2013)[51] which we

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reproduced in Figure 2A). However, a closer look at this plot reveals a problem: the linear relationship does not appear to hold for lowly expressed genes (Figure 2B). This same pattern is observed in the other two studies with bulk and scRNA-seq data (Figure S11).

To further explore this apparent bias, we stratified the values of Y_g and estimated the conditional expectations of $E\left[\sum_{i=1}^{N} Z_{gi} | \, Y_g = e\right]$ by averaging the scRNA-seq data in each strata. Plotting these against each other revealed a bias that increases as e becomes closer to zero (Figure 3). These results are very much consistent with the theory that some of the observed zeros are due to technical and not biological differences with the actual relationship being:

$$E\left[\sum_{i=1}^{N} Z_{gi} | \, Y_g = e\right] = p(e) * (\beta_0 + \beta_1 e)$$

with $p(e)$ the probability of a gene with expression e being detected. A crude estimate of $p(e)$ can be obtained by

$$\hat{p}(e) = \hat{E}\left[\sum_{i=1}^{N} Z_{gi} | \, Y_g = e\right] / \left(\hat{\beta}_0 + \hat{\beta}_1 e\right)$$

This estimate suggests $p(e)$ follows a logistic function (Figure S12) as others have previously noted[16]. In other words, genes with a lower expression e are less likely to be detected, which suggests the zeros can be considered missing not at random as the probability of the missing value depends on the level of expression.

Motivated by Figure S12 we fit a logistic curve to determine the relationship between $Z_{gi} > \delta$ and Y_g for each cell i. We found that the biases induced by this missing data problem is exacerbated by the fact that the probability of a gene being detected varies cell to cell, as the

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

estimates for the logistic curve's intercept parameter are highly related to the detection rate (Figure S13). We also note that the slope estimates are between 0.53 and 1.31. For example, the slopes estimated using the Trapnell et al. (2014)[5] data has an average of 0.82 and a standard deviation of 0.6 demonstrating the strong effect overall expression has on detection: note for example that a slope of 0.82 means that if the expression level is cut in half, the detection odds decrease by more than two fold since $e^{0.82} = 2.27$.

3.3 Detection rate is a major source of cell-to-cell variation

Finak et al. (2015)[53] showed that detection rates correlate with the first two principal components (PCs) in two scRNA-seq data sets[6, 53]. We confirmed this relationship on the fifteen publicly available scRNA-seq datasets we studied (Figures S14-S15). From this strong correlation it follows that estimated distances between cells are affected by differences in detection rate. We note that for five of these studies[3, 8, 55-57], the primary variation along the first two principal components was correlated strongly with the biological groups known to have different gene expression patterns. For these studies, we stratified the data into these groups and found the same strong relationship between detection rates and first principal component. In this section, we present results that demonstrate that (1) this variability is partly driven by a mathematical artifact related to scaling the original data but computing distances in the log transformed data and that (2) that differences in detection rate can be completely driven by technical reasons which can in turn result in false discoveries.

Currently the most widely used unit for reporting expression values is the Transcripts per Million (TPM) unit. Using this unit guarantees that the sum of gene expression measurements are

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

constant. This is also true for CPM and approximately true with the RPKM[42] and FPKM[43] units. However, distance calculations are performed after log transformations and cell expression profiles are not always reported as being centered (centered by removing overall mean expression from the i^{th} cell) in published analyses[1, 2, 57, 58]. We can show, mathematically, that if we normalize expression profiles to have the same mean across cells, the mean after the $f(x) = \log(x + c)$ transformation used for RNA-Seq data will not be the same and it will depend on the detection rate (Figure S16).

$$E\left[\log\left(X_{gi} + c\right)\right] \approx (1 - p_i)\log(c) + p_i\left[\log\left(\frac{M}{G * p_i} + c\right)\right] \qquad (1)$$

where X_{gi} is the expression value for the g^{th} gene and i^{th} cell, c is a pseudo count, G is the number of genes (or features), M is sequencing depth, and p_i is the marginal probability of detection for the i^{th} cell (mathematical details are provided in the Supplement). The implication of this result is that although the means are constant using across cells i, the means of the log-transformed data depend on the detection rate. In fact, when the sequencing depth is large, the mathematical relationship above is approximately a linear function of the detection rate. Because these mean values affect the entire vector, they can result in large overall variability and therefore be correlated with the first principal component PC. Therefore, the result in Figures S14-S15 can be explained by differences in mean values that correlated with the detection rate.

Not surprisingly, if we center the data before computing the PCs, then the correlation between the detection rate and the first PC decreases. However, despite the decrease, even after the centering the correlation is strong (Figures S17-S18). This also is not surprising given that not

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

only the mean expression depends of the detection rate, but also the entire distribution of the non-zero measurements using (Figure 4).

To demonstrate that technical variability can lead to differences in detection rates, which in turn can lead to false discoveries, we used a dataset composed of a group of cells from the same biological specimen, but processed at different times. Specifically, we used a subset of 118 single cells data from Patel et al. (2014), which were isolated from one tumor, but processed in two different sequencing instruments[11]. For these data, there are no biological reasons for the two groups, defined by the sequencing instrument used, to be different since the cells were randomly selected from the same tumor. If we apply an unsupervised clustering algorithm to these data, two clusters strong clusters appear (Figure S19) even after removing the cell mean before computing distances. A PCA plot shows that a batch effect drives the clustering (Figure 5A). We then note that the first PC strongly correlates with the detection rate (Figure 5B), which is substantially different between the two batches (Figure 5C). In addition, the differences in detection rates are highly related to the logistic curve's intercept parameter when estimating the probability of a gene being detected, which varies cell to cell (Figure S20). Therefore, we see how a batch effect can produce differences in detection rate that drive distances between transcription profiles and leads to false discoveries.

3.4 The impact of the detection rate in applications of unsupervised learning methods

In the previous sections, we demonstrated how the detection rate is a major source of observed cell-to-cell variation, which can be driven by technical variation. For example, we considered cells from the same biological specimen for which we knew no differences should be discovered,

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

but we found a batch effect that had differences in the detection rate and could drive the variation between the cells. In this section, we examine detection rates in data used in published studies as evidence for the discovery of new cell types.

In the first case-study[57], we found differences in the detection rate between two groups of discovered cell types (Figure S21A), which was associated with how the cells were processed in different batches (Figure S21B) (Fisher's Exact test: $p = 0.002$). For example, 12 out of 13 cells from one discovered cell types were processed in the same batch (Figure S21C), which had a smaller median detection rate than the other batches. Furthermore, the detection rate was associated with the first principal component (Figure S21D), which could be partly driving the variation across the two groups of discovered cell types. Similarly, we found differences in the detection rate between groups of discovered cell types in another other study[3] (Figure S22), which was associated with how the cells were processed in different batches (Chi-squared test: $p < 0.001$).

4. Discussion

We have demonstrated how detection varies substantially across scRNA-seq experiments. We presented evidence that part of this variability is technically driven. Given the logistics of how scRNA-Seq experiments are performed and that fact that this technology is being used to discover new cell-types, batch effects are of particular concern. Specifically, when two groups of cells are cultured, captured and sequenced separately from another group of cells in a second condition, correlated measurements may lead to the incorrect conclusion that these groups have different expression profiles. This experimental limitation presents a challenge in distinguishing

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

biologically driven differences from technical ones because it is logistically difficult to avoid processing cells from different biological specimens in different batches. For the eight studies that were interested in comparing predefined biological groups, we used the standardized Pearson contingency coefficient to assess the level of confounding between the run number from the sequencing instrument and outcome of interest and found values ranging from 82.1% to 100% (perfect confounding) (Table 1; Tables S1-S8). Note that with this level of confounding it difficult to impossible to parse technical from biological variation.

Furthermore, explicitly modeling confounding factors as in published batch correction methods[59-61] is not appropriate in this context because the biological variation or signal of interest is often confounded with the unwanted technical variability. For the specific application of differential expression, a proposed solution is to account for differences in the proportion of detected genes by explicitly including it as a covariate in a linear regression model[53]. However, given the current levels of confounding, this approach will not be able to distinguish biological from technical effects. For example, some studies have demonstrated cells with different biological phenotypes can express a different number of genes[62].

An experimental design solution is to use biological replicates, namely independently repeating the experiment multiple times for each biological condition (Figure S1). This approach allows for multiple batches of cells to be randomized across sequencing runs, flow cells and lanes as in bulk RNA-Seq. With this design we can then model and adjust for batch effects due to systematic experimental bias. A more detailed discussion of how these factors affect the experimental design has been recently published[14, 18, 19].

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5. Conclusions

Technical variability is considered to be a major challenge in the analysis of data measured on next-generation sequencing platforms. For example, amplification bias leading to batch effects has been shown to induce false positives in differential expression studies with bulk RNA-seq data[63, 64]. By examining three assessment experiments containing both bulk and single cell RNA-Seq data, we demonstrated that technical variability is a challenge in scRNA-Seq as well, with a major problem arising due to differences in capture inefficiencies. Using public data from fifteen studies, we showed that these inefficiencies lead to substantial differences in detection rates that lead to distortion in distance calculations, which in turn can lead to false discoveries when using unsupervised clustering.

Acknowledgements

We thank Bradley Bernstein who provided comments that we used to improve the manuscript. This research was supported by NIH R01 grants GM083084, RR021967/GM103552 and HG005220 and NIH/NHGRI grant K99HG009007.

Competing Interests

The authors declare no competing interests.

Supplementary Material

Supplementary materials are available in a single PDF. All the code for this analysis is available on GitHub (https://github.com/stephaniehicks/scBatchPaper).

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
2. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371-375 (2014).
3. Zeisel, A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142 (2015).
4. Wilson, N.K. et al. Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell stem cell* (2015).
5. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381-386 (2014).
6. Shalek, A.K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-369 (2014).
7. Borel, C. et al. Biased allelic expression in human primary fibroblast single cells. *American journal of human genetics* **96**, 70-80 (2015).
8. Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193-196 (2014).
9. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* **33**, 495-502 (2015).
10. Achim, K. et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature biotechnology* **33**, 503-509 (2015).
11. Patel, A.P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).
12. Jaitin, D.A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776-779 (2014).
13. Usoskin, D. et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature neuroscience* **18**, 145-153 (2015).
14. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome biology* **17**, 63 (2016).
15. Zhu, L., Lei, J. & Roeder, K. A Unified Statistical Framework for RNA Sequence Data from Individual Cells and Tissue. *arXiv* (2016).
16. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nature methods* **11**, 740-742 (2014).
17. Lun, A.T.L., Bach, K. & Marioni, J.C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome biology* **17**, 75 (2016).
18. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nature reviews. Genetics* **16**, 133-145 (2015).
19. Grun, D. & van Oudenaarden, A. Design and Analysis of Single-Cell Sequencing Experiments. *Cell* **163**, 799-810 (2015).
20. Saliba, A.E., Westermann, A.J., Gorski, S.A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic acids research* **42**, 8845-8860 (2014).
21. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philiosophical Magazine* **2**, 559-572 (1901).

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

22. Tipping, M.E. & Bishop, C.M. Probablistic principal components analysis. *JR Stat Soc: Series B (Statistical Methodology)* **61**, 611-622 (1999).
23. Torgerson, W.S. Multidimensional scaling I: Theory and method. *Psychometrika* **17**, 401-419 (1952).
24. Lafon, S. & Lee, A.B. Diffusion maps and coarse-graining: A unified framework for dimension- ality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**, 1393-1403 (2006).
25. Nadler, B., Lafon, S., Coifman, R.R. & Kevrekidis, I.G. Diffusion maps, spectral clustering and the reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis: Special Issue on Diffusion Maps and Wavelets* **21**, 113-127 (2006).
26. van der Maaten, L. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).
27. Finak, G. et al. Mixture models for single-cell assays with applications to vaccine studies. *Biostatistics* **15**, 87-101 (2014).
28. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology* **16**, 241 (2015).
29. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. & Teichmann, S.A. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular cell* **58**, 610-620 (2015).
30. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics* **14**, 618-630 (2013).
31. Combs, P.A. & Eisen, M.B. Low-cost, low-input RNA-seq protocols perform nearly as well as high-input protocols. *PeerJ* **3** (2015).
32. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nature methods* **14**, 381-387 (2017).
33. Ziegenhain, C. et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular cell* **65**, 631-643 e634 (2017).
34. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207-210 (2002).
35. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6**, 377-382 (2009).
36. Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome research* **21**, 1160-1167 (2011).
37. Ramskold, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology* **30**, 777-782 (2012).
38. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10**, 1096-1098 (2013).
39. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports* **2**, 666-673 (2012).
40. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods* **9**, 72-74 (2012).
41. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

42. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628 (2008).
43. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-515 (2010).
44. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database, C. The sequence read archive. *Nucleic acids research* **39**, D19-21 (2011).
45. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34**, 525-527 (2016).
46. Bose, S. et al. Scalable microfluidics for single-cell RNA printing and sequencing. *Genome biology* **16**, 120 (2015).
47. Gilad, Y. & Mizrahi-Man, O. A reanalysis of mouse ENCODE comparative gene expression data. *F1000Research* **4**, 121 (2015).
48. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods* **10**, 1093-1095 (2013).
49. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology* **33**, 155-160 (2015).
50. Tung, P.Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Sci Rep* **7**, 39921 (2017).
51. Shalek, A.K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-240 (2013).
52. Wu, A.R. et al. Quantitative assessment of single-cell RNA-sequencing methods. *Nature methods* **11**, 41-46 (2014).
53. Finak, G. et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA-seq data. *bioRxiv* (2015).
54. Korthauer, K.D. et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome biology* **17**, 222 (2016).
55. Guo, F. et al. The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell* **161**, 1437-1452 (2015).
56. Kumar, R.M. et al. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* **516**, 56-61 (2014).
57. Burns, J.C., Kelly, M.C., Hoa, M., Morell, R.J. & Kelley, M.W. Single-cell RNA-Seq resolves cellular complexity in sensory organs from the neonatal inner ear. *Nature communications* **6**, 8557 (2015).
58. Kowalczyk, M.S. et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome research* (2015).
59. Leek, J.T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic acids research* **42** (2014).
60. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).
61. Risso, D., Ngai, J., Speed, T.P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature biotechnology* **32**, 896-902 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

62. Ramskold, D., Wang, E.T., Burge, C.B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS computational biology* **5**, e1000598 (2009).
63. Lahens, N.F. et al. IVT-seq reveals extreme bias in RNA sequencing. *Genome biology* **15**, R86 (2014).
64. Love, M.I., Hogenesch, J.B. & Irizarry, R.A. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature biotechnology* **34**, 1287-1291 (2016).
65. Leng, N. et al. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nature methods* **12**, 947-950 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Study	Organism	Single-cell RNA-Seq protocol	Number of cells	Number of genes or transcripts	Processed data available	Confounding (%):	Ref
Deng et al. (2014)	Mouse	SMART-Seq	286	22,958	RPKM	96.6[+]	8
Guo et al. (2015)	Human	Tang et al. (2009)	154	23,394	FPKM	82.1	55
Kowalczyk et al. (2015)	Mouse	SMART-Seq	533	8,422	TPM	84.8	58
Kumar et al. (2014)	Mouse	SMART-Seq	361	22,443	TPM	97.1	56
Patel et al. (2014)	Human	SMART-Seq	430	5,948	TPM	98.9	11
Treutlein et al. (2014)	Mouse	SMART-Seq	198	23,745	FPKM	92.8	2
Shalek et al. (2014)	Mouse	SMART-Seq	383	27,723	TPM	100	6
Trapnell et al. (2014)	Human	SMART-Seq	306	47,192	FPKM	100	5
Burns et al. (2015)	Mouse	SMART-Seq	249	26,585	TPM	NA	57
Leng et al. (2015)	Human	SMART-Seq	458	19,804	TPM	NA	65
Bose et al. (2015)	Human	CEL-Seq	247	17,450	UMI	NA	46
Jaitin et al. (2014)	Mouse	MARS-Seq	4,466	20,190	UMI	NA	12
Macosko et al. (2015)	Mouse	Drop-Seq	49,300	16,961	UMI	NA	1
Satija et al. (2015)	Zebrafish	SMART-Seq	1,152	13,902	UMI	NA	9
Zeisel et al. (2015)	Mouse	STRT-Seq	3,004	19,972	UMI	NA	3

Table 1: Description of processed single-cell RNA-Seq data sets. Column 1 shows the publications. Column 2 shows the organism. Column 3 shows the single-cell technology used for sequencing. Column 4 shows the number of cells (samples) included in the study. Column 5 shows the number of genes included in the data uploaded to the public repository. Column 6 indicates the units in which the values were reported. Column 7 shows the level of confounding between biological condition and batch effect quantified using the standardized Pearson contingency coefficient as a measure of association. The percentage ranges from 0% (no confounding) to 100% (completely confounded). Column 8 provides the citation for the study.

[+] The main purpose of this study was to investigate monoallelic gene expression in mouse embryos, but here we consider the different developmental stages (oocyte to blastocyst) as the biological condition as an example.

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

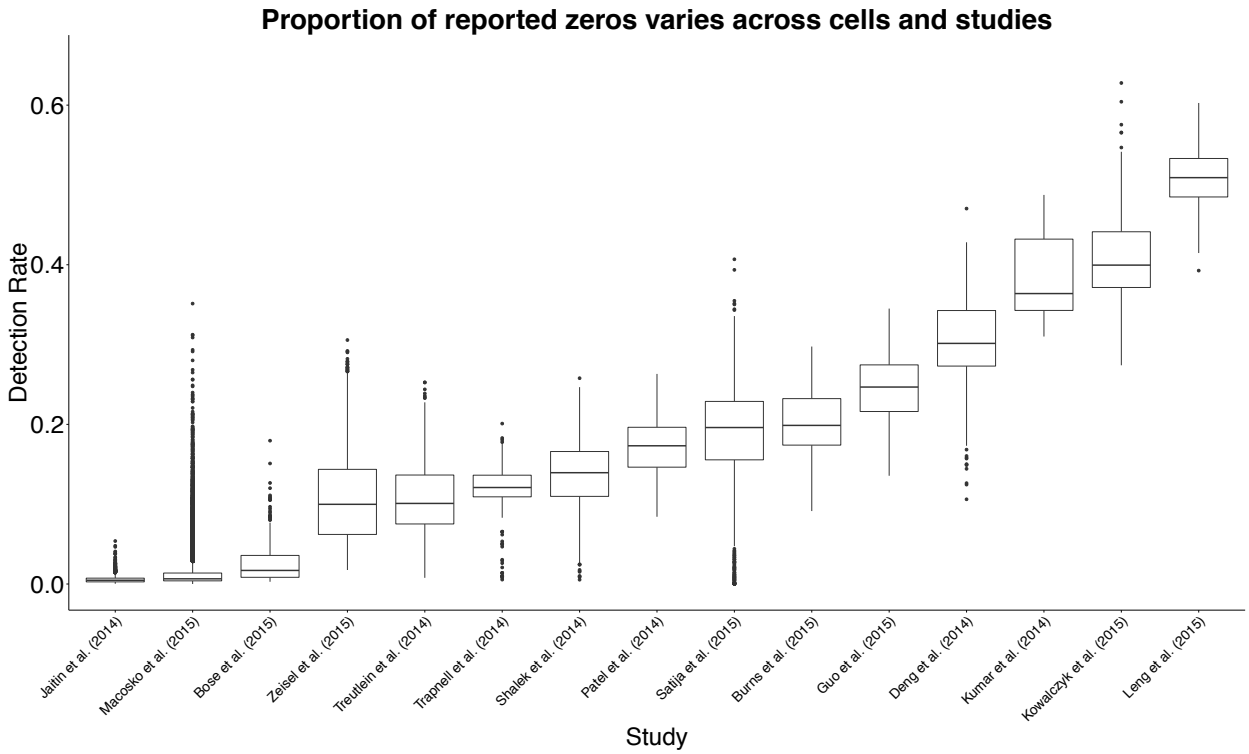

Figure 1: Boxplots of the *detection rate*, or the proportion of genes in a cell reporting expression values greater δ=1 calculated for each cell across fifteen publicly available scRNA-seq studies. The detection rate across cells and studies ranges from less than 1% to 65%.

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

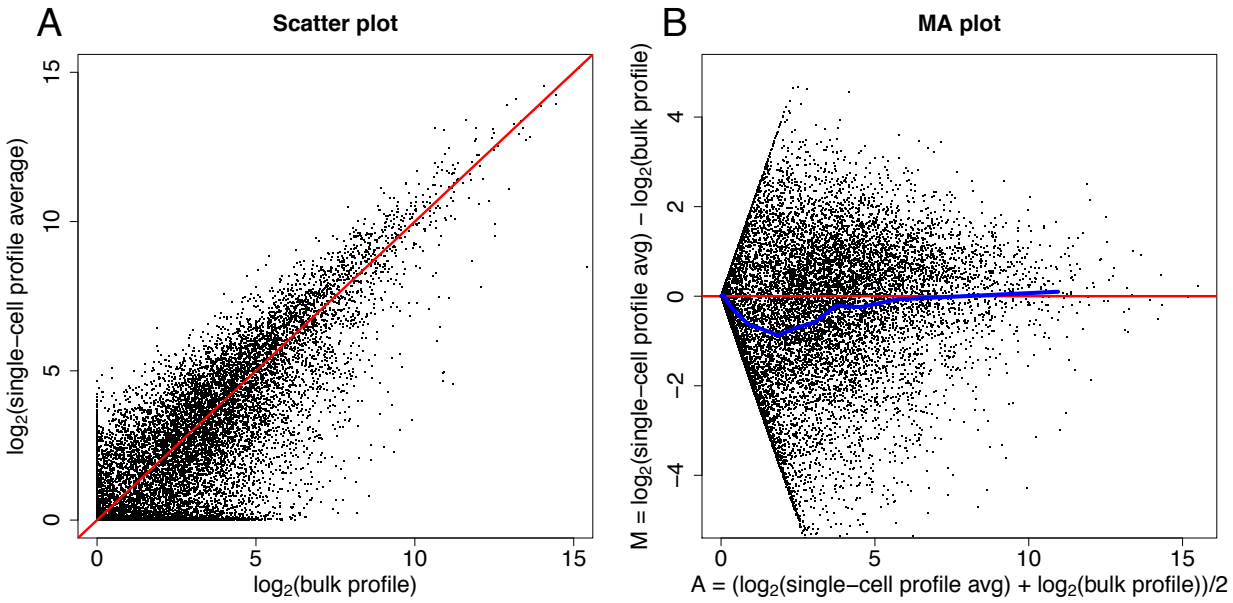

Figure 2: RNA-seq profiles compared to averaged scRNA-seq profile (**A**) Scatter plot comparing a bulk RNA-seq profile and an averaged scRNA-seq profile, which we reproduced from Figure 1C in Shalek et al. (2013)[51]. (**B**) The MA plot demonstrates there is a bias between the bulk profile and the single cell profile averaged across cells as the single cell profile averaged across cells is smaller than the bulk profile for low expressed genes.

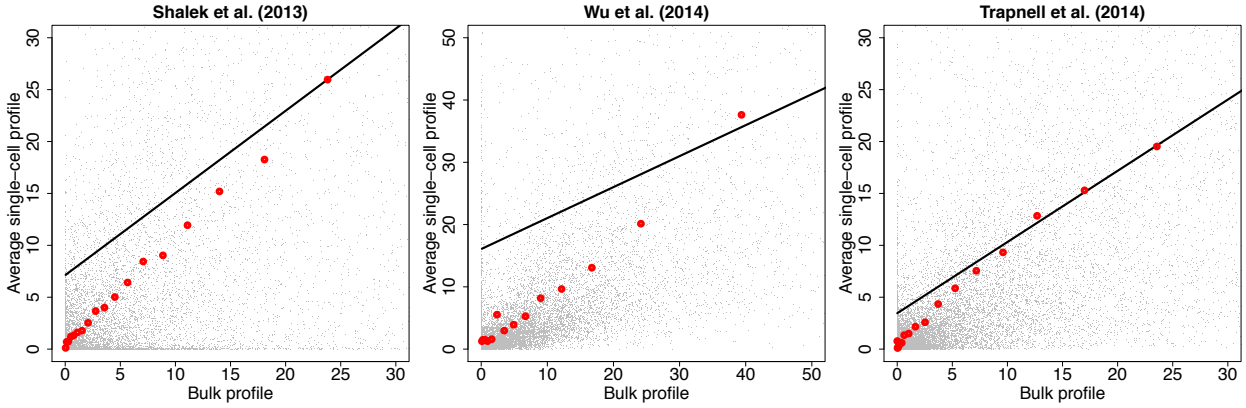

Figure 3: Plots comparing bulk and averaged scRNA-seq profiles that demonstrate evidence of more zeros in in scRNA-seq data for low expressed genes than what is expected. Data was obtained from three publicly available scRNA-seq studies that included a matched bulk RNA-seq sample measured on the same population of cells[5, 51, 52]. The red points are averages of the single cell profiles computed in strata defined by the bulk RNA-seq values. The black solid line is what we expect if there is no bias.

bioRxiv preprint doi: https://doi.org/10.1101/025528; this version posted May 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

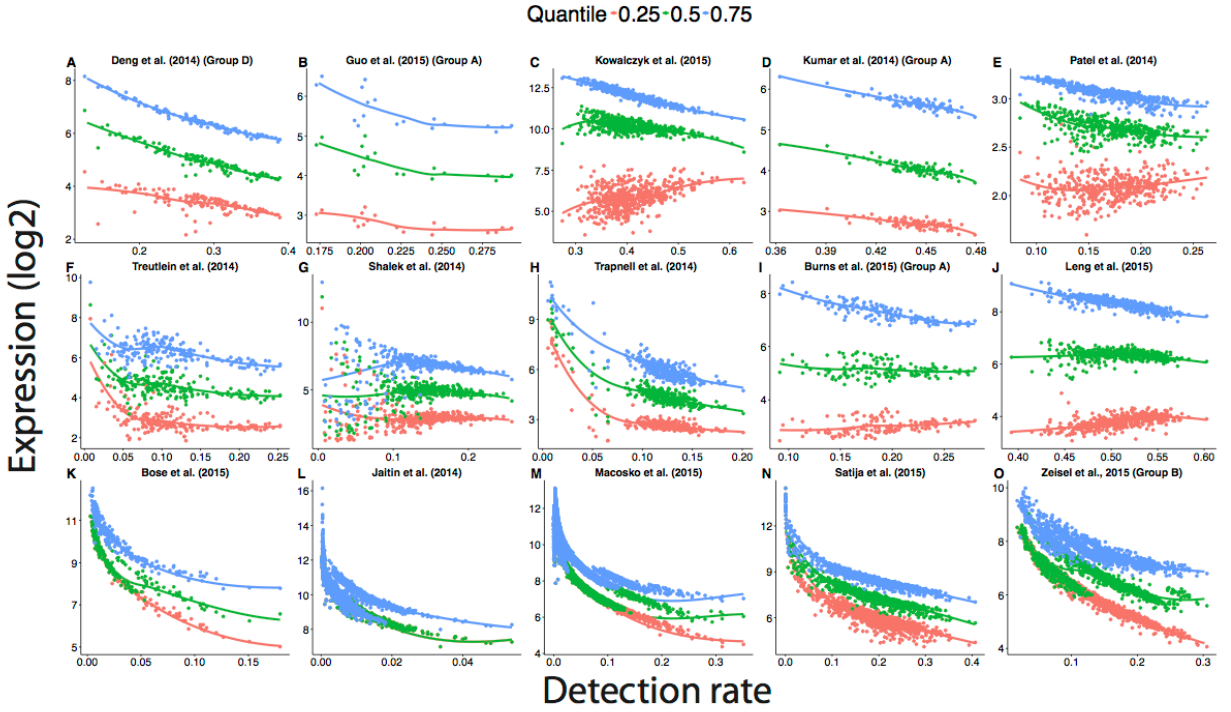

Figure 4: The distribution of gene expression changes with the detection rate using processed scRNA-seq data available on GEO. Failure to account for differences of the proportion of detected genes between cells over-inflates the gene expression estimates of cells with a low detection rate. The curves were obtained by fitting a locally weighted scatter plot smoothing (loess) with a degree of 1. Because the range of detection rate varied from study to study, the range of the x-axis differs across plots.

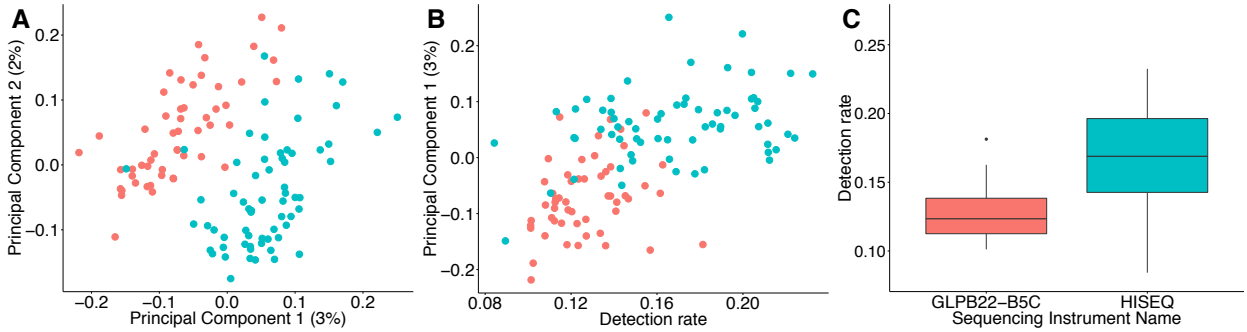

Figure 5: Illustration of how technical variation can lead to differences in detection rates, which in turn can lead to false differences. (A) Boxplots of detection rates from cells stratified by sequencing instrument used to sequence cells. (B) Using principal components analysis, scRNA-seq samples cluster by sequencing instrument. (C) Detection rate is associated with the first principal component.