On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data

Stephanie C. Hicks[1,2] , 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

Mingxiang Teng, mxteng@jimmy.harvard.edu

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

Abstract

Single-cell RNA-Sequencing (scRNA-Seq) has become the most widely used high-throughput method for transcription profiling of individual cells. Systematic errors, including batch effects, have been widely reported as a major challenge in high-throughput technologies. Surprisingly, these issues have received minimal attention in published studies based on scRNA-Seq technology. We examined data from all fifteen published studies including at least 200 samples and found that systematic errors can explain a substantial percentage of observed cell-to-cell expression variability. Specifically, we found that the proportion of genes reported as expressed explains a substantial part of observed variability and that this quantity varies systematically across experimental batches. Furthermore, we found that experimental designs that confound outcomes of interest with batch effects are common. Finally, we propose a simple experimental design that can ameliorate the effect of theses systematic errors have on downstream results.

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. Our experience with high-throughput genomic data in general, is that well thought-out data processing pipelines are essential to produce meaningful downstream results[1-3]. We expect the same to be true for scRNA-seq data. Here we show that while some tools developed for analyzing bulk RNA-Seq can be used for scRNA-Seq data, such as the mapping and alignment software, other steps in the processing, such as normalization, quality control and quantification, require new methods to account for the additional variability that is specific to this technology.

One of the most challenging sources of unwanted variability and systematic error in high-throughput data are what are commonly referred to as *batch effects*. Given the way that scRNA-Seq experiments are conducted, there is much room for concern regarding batch effects[4]. Specifically, batch effects occur when cells from one biological group or condition are cultured, captured and sequenced separate from cells in a second condition. Although batch information is not always included in the experimental annotations that are publicly available, one can extract surrogate variables from the raw sequencing (FASTQ) files[5]. 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 surrogate for other experimental procedures that very likely do have an effect, such as starting material, PCR amplification reagents/conditions, and cell cycle stage of the cells[6-8]. Here we will refer to the resulting differences induced by different groupings of these sources of variability as *batch effects.*

In a completely confounded study, it is not possible to determine if the biological condition or batch effects are driving the observed variation. In contrast, incorporating biological replicates across in the experimental design and processing the replicates across multiple batches permits observed variation to be attributed to biology or batch effects (Figure 1). To demonstrate the widespread problem of systematic bias, batch effects, and confounded experimental designs in scRNA-Seq studies, we surveyed several published data sets. We discuss the consequences of failing to consider the presence of this unwanted technical variability, and consider new strategies to minimize its impact on scRNA-Seq data.

Batch and outcomes of interest are confounded in published scRNA-Seq experiments

We examined all publicly available scRNA-Seq data sets including at least 200 samples to investigate the extent of confounding biological variation with batch effects. These data sets were created using six different scRNA-Seq protocols for sequencing[9-15] and five studies include the use of unique molecular identifiers[16] (UMIs) for counting cDNA molecules. For each study, we downloaded the processed data available on GEO[17] and reconstructed the study design from the sequence identifiers provided in the FASTQ files. We used the standardized Pearson contingency coefficient to assess the experimental design between processing batches and outcome of interest and found values ranging from 82.1% to 100% (perfect confounding) in eight studies (Table 1). Note that with this level of confounding it is nearly impossible to parse batch effects from biological variation. We could not assess the experimental design in the other seven studies because no biological outcomes of interest were defined, the purpose of the study was the discovery of novel cell types or the paper described a new technology. We note that in one of these studies for the discovery of novel cell types, the predicted cell type was provided and we found that the confounding percentage between the reported cell type and batch was 80.5%[18].

Proportion of detected genes is a major source of technical cell-to-cell noise

Most, if not all, published studies using scRNA-Seq rely explicitly or implicitly on computing distances between the cell expression profiles. Principal component analysis is used explicitly to quantify biological or molecular distance[19] or implicitly to approximate distance between individual cells. We used the processed expression data available on GEO, applied principal components analysis on the log (base 2) transformed values (adding 1 to avoid logs of 0), and computed the proportion of detected genes from the same data set with the exception of one study[19]. In this exception, the processed expression data available on GEO excluded most non-detected genes and the values for each gene were centered by removing the average. For this case, we computed the proportion of detected genes from the raw data. We found wide variation in the proportion of detected genes across cells in all studies: from 1% detected to 65%. Furthermore, we found strong correlation between the first principal component and the proportion of detected genes within each data set (Figure 2, Figure S1).

To determine if the variability in the proportion of detected genes was biologically or technically driven we compared the variability across biological groups to the variability across batches. For the

cases in which the experimental design permitted this comparison (Tables S1-S6), we found that batch explained more variability than biological group (Figures S2-S7). We note that for three of the studies[20-22] we split the defined biological groups into subgroups because the differences between biological groups were pronounced, and we found similar results. In the two studies for which batch was completely confounded with biological group (Table S7 and S8) we also observed variability across batch (Figure S8). However, in these cases it is impossible to separate variability due to biology or to batch. In seven of the studies, there were no biological outcomes of interest defined or the purpose of the study was for discovery of novel cell types (Figs S9-S15).

Batch effects lead to differences in detection rates, which lead to apparent differences between biological groups

To illustrate potential down stream effects, we examined 430 single-cells from five biological groups of interest for one of the studies in which at least one biological condition was split across two batches (Table S1). We confirmed that cells cluster by biological group (Figure 3A) as reported in the original publication. However, four of the five biological groups were confounded with batch (Table S1) and batches can also explain the clustering (Figure 3B). The one group that was not confounded with batch confirms the high level of variability explained by processing cells in different batches (Figure 3C). As expected from the previous results, different batches lead to different proportions of detected genes (Figure 3D), which we have shown to be correlated with the first principal component and may be driving the observed biological variation across groups.

Detection rate has indirect effects on reported gene expression measurements

Using the processed scRNA-Seq data available on GEO, we computed the median of the non-zero measurements for each cell. For each study, we noticed a strong non-linear relationship between the median expression and the proportion of detected genes (Figure 4, Figure S16). We also found that the entire distribution of the non-zero genes changed with the proportion detected genes (Figure S17). The overall level of expression changing with the proportion of detected genes could be biologically driven, but there is a reasonable explanation of how it can be a technical artifact, which we explain using statistical notation. Let X_{ij} be the unobserved expression level for sample i and gene j. Let us consider only expressed genes ($X_{ij} > 0$). In the sequencing experiment, each expressed gene has a probability of being amplified, which means we observe a quantity proportional to $X_{ij}Z_{ij}$ where

$Z_{ij} = 1$ if the gene was expressed at a high enough level to be detected and amplified and 0 otherwise. This implies that the expected amount of RNA we will obtain is a quantity proportional to

$$E(X_{ij}Z_{ij}) = \Pr(Z_{ij} = 1)\, E(X_{ij}|Z_{ij} = 1)$$

Because we know the technical variation affects the probability detection $\Pr(Z_{ij} = 1) = p_i$, which varies between samples, this quantity should be calculated for each sample i. We assume that, within a homogenous population for example, the expected level of a gene that is detected, $E(X_{ij}|Z_{ij} = 1)$, is the similar across cells. Then, the total amount of RNA is proportional to the detection rate of the expressed genes p_i. Now because experimentally we amplify to have roughly the same total amount of RNA for each sample, this implies that we have to amplify in a way that is equivalent to multiplying by $\frac{K}{p_i}$, with K the total we aspire to reach. Thus the median expression of the expressed genes will be proportional to $\frac{1}{p_i}$, which is consistent with the data (Figure 4, Figure S16). To confirm this, we implemented an adjustment based on these derivations. Specifically, we used the five studies that count cDNA molecules based on UMIs and followed a standard normalization procedure[15], which scales the raw UMI counts Y_{ij} for the jth gene (or transcript) in the ith cell by the total number of UMIs N_i in the sample. Using the normalized UMI data, we multiplied the ith cell by an adjustment factor $\left(\frac{p_i}{\text{median}(p_1, p, \ldots, p_I)}\right)$ motivated by the derivations above. This adjustment removed much of the dependence between the median expression values and the proportion of detected genes (Figure S18). We found similar results using normalized read counts from scRNA-Seq data sets based on TPM, RPKM and FPKM (results not shown).

Experimental design solutions

Batch effects are not limited to scRNA-Seq data and have been shown to be a widespread problem in high-throughput experiments[3, 5], but there are currently no published general statistical solutions to the problem of batch effects in scRNA-Seq data. 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[23]. However, given the current levels of confounding and experimental designs, 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[24]. However, due to the nature of the experimental

protocol needed to run scRNA-Seq experiments imposed by the way cells are captured and sequenced in batches, standard balanced experimental designs are not possible[4, 25, 26]. An experimental design solution is to use biological replicates, namely independently repeating the experiment multiple times for each biological condition (Figure 1). 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. Other considerations for the design of a scRNA-Seq experiment such as the number of cells that should be generated can be determined using a power analysis, which will depend multiple factors including the goal of the study, the library complexity or sequencing depth required to detect most expressed genes, the protocol used to quantify the expression values such as using unique molecular identifiers[16], and the success rate for failed single-cell libraries due to degraded RNA or low amplification efficiency. A more detailed discussion of how these factors affect the experimental design has been recently published[4, 26].

Discussion

Batch effects and unwanted technical cell-to-cell noise remains a challenge in the analysis of scRNA-Seq data. The challenge is more complex than in previous sequencing experiments since experimental batches lead to different detection rates, which in turn lead to different transcription level estimates. In addition, detection of a gene or transcript in scRNA-Seq experiments is heavily dependent on the experimental protocol, which leads to systematic differences in the proportion of detected genes between batches of cells. The development of statistical methods that account for these systematic biases will therefore be essential in the analysis of scRNA-Seq data. Incorporating biological replicates in the experimental study design provides a solution to reducing confounding between biological condition and batch effects and will permit modeling of the technical variability that relates to processing the cells in different batches.

Acknowledgements

We thank Bradley Bernstein who provided insightful comments that greatly improved the manuscript. This research was supported by NIH R01 grants GM083084, RR021967/GM103552 and HG005220.

Competing Interests

The authors declare no competing interests.

Supplementary Material

Supplementary materials are available in a single PDF.

References

1. Akey, J.M., Biswas, S., Leek, J.T. & Storey, J.D. On the design and analysis of gene expression studies in human populations. *Nature genetics* **39**, 807-808; author reply 808-809 (2007).
2. Baggerly, K.A., Edmonson, S.R., Morris, J.S. & Coombes, K.R. High-resolution serum proteomic patterns for ovarian cancer detection. *Endocrine-related cancer* **11**, 583-584; author reply 585-587 (2004).
3. Leek, J.T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics* **11**, 733-739 (2010).
4. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nature reviews. Genetics* **16**, 133-145 (2015).
5. Gilad, Y. & Mizrahi-Man, O. A reanalysis of mouse ENCODE comparative gene expression data. *F1000Research* **4**, 121 (2015).
6. Shalek, A.K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-369 (2014).
7. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods* **10**, 1093-1095 (2013).
8. 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).
9. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6**, 377-382 (2009).
10. Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome research* **21**, 1160-1167 (2011).
11. 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).
12. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10**, 1096-1098 (2013).
13. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports* **2**, 666-673 (2012).
14. 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).
15. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
16. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods* **9**, 72-74 (2012).

17. 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).
18. 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).
19. Patel, A.P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).
20. 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).
21. Guo, F. et al. The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell* **161**, 1437-1452 (2015).
22. Kumar, R.M. et al. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* **516**, 56-61 (2014).
23. Finak, G. et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA-seq data. *bioRxiv* (2015).
24. 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).
25. 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).
26. Grun, D. & van Oudenaarden, A. Design and Analysis of Single-Cell Sequencing Experiments. *Cell* **163**, 799-810 (2015).
27. 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).
28. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371-375 (2014).
29. 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).
30. Leng, N. et al. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nature methods* **12**, 947-950 (2015).
31. Bose, S. et al. Scalable microfluidics for single-cell RNA printing and sequencing. *Genome biology* **16**, 120 (2015).
32. 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).
33. 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).

Study	Organism	Single-cell RNA-Seq protocol	Number of cells	Number of genes or transcripts	Processed data available	Confound-ing (%):	Correlation between PC1 and proportion of detected genes or transcripts	Ref
Deng et al. (2014)	Mouse	SMART-Seq	286	22,958	RPKM	96.6[+]	0.69-0.82[*]	[20]
Guo et al. (2015)	Human	Tang et al. (2009)	154	23,394	FPKM	82.1	0.65-0.90[*]	[21]
Kowalczyk et al. (2015)	Mouse	SMART-Seq	533	8,422	TPM	84.8	0.90	[27]
Kumar et al. (2014)	Mouse	SMART-Seq	361	22,443	TPM	97.1	0.13-0.78[*]	[22]
Patel et al. (2014)	Human	SMART-Seq	430	5,948	TPM	98.9	0.54[**]	[19]
Treutlein et al. (2014)	Mouse	SMART-Seq	198	23,745	FPKM	92.8	0.91	[28]
Shalek et al. (2014)	Mouse	SMART-Seq	383	27,723	TPM	100	0.93	[6]
Trapnell et al. (2014)	Human	SMART-Seq	306	47,192	FPKM	100	0.90	[29]
Burns et al. (2015)	Mouse	SMART-Seq	249	26,585	TPM	NA	0.65-0.96[*]	[18]
Leng et al. (2015)	Human	SMART-Seq	458	19,804	TPM	NA	0.85	[30]
Bose et al. (2015)	Human	CEL-Seq	247	17,450	UMI	NA	0.89	[31]
Jaitin et al. (2014)	Mouse	MARS-Seq	4,466	20,190	UMI	NA	0.89	[14]
Macosko et al. (2015)	Mouse	Drop-Seq	49,300	16,961	UMI	NA	0.75	[15]
Satija et al. (2015)	Zebrafish	SMART-Seq	1,152	13,902	UMI	NA	0.87	[32]
Zeisel et al. (2015)	Mouse	STRT-Seq	3,004	19,972	UMI	NA	0.66-0.68[*]	[33]

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 shows the Pearson correlation between the first principal component of the log transformed data and the proportion of detected genes. Column 9 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.

[*] Differences between the biological conditions in these studies were pronounced; therefore we binned the conditions into separate into groups for this analysis each with its own corresponding correlation coefficient.

[**] The available processed data was previously filtered by the authors and excluded the majority of non-detected genes, which partially explains the lower correlation (0.54 compared to 0.65-0.93).

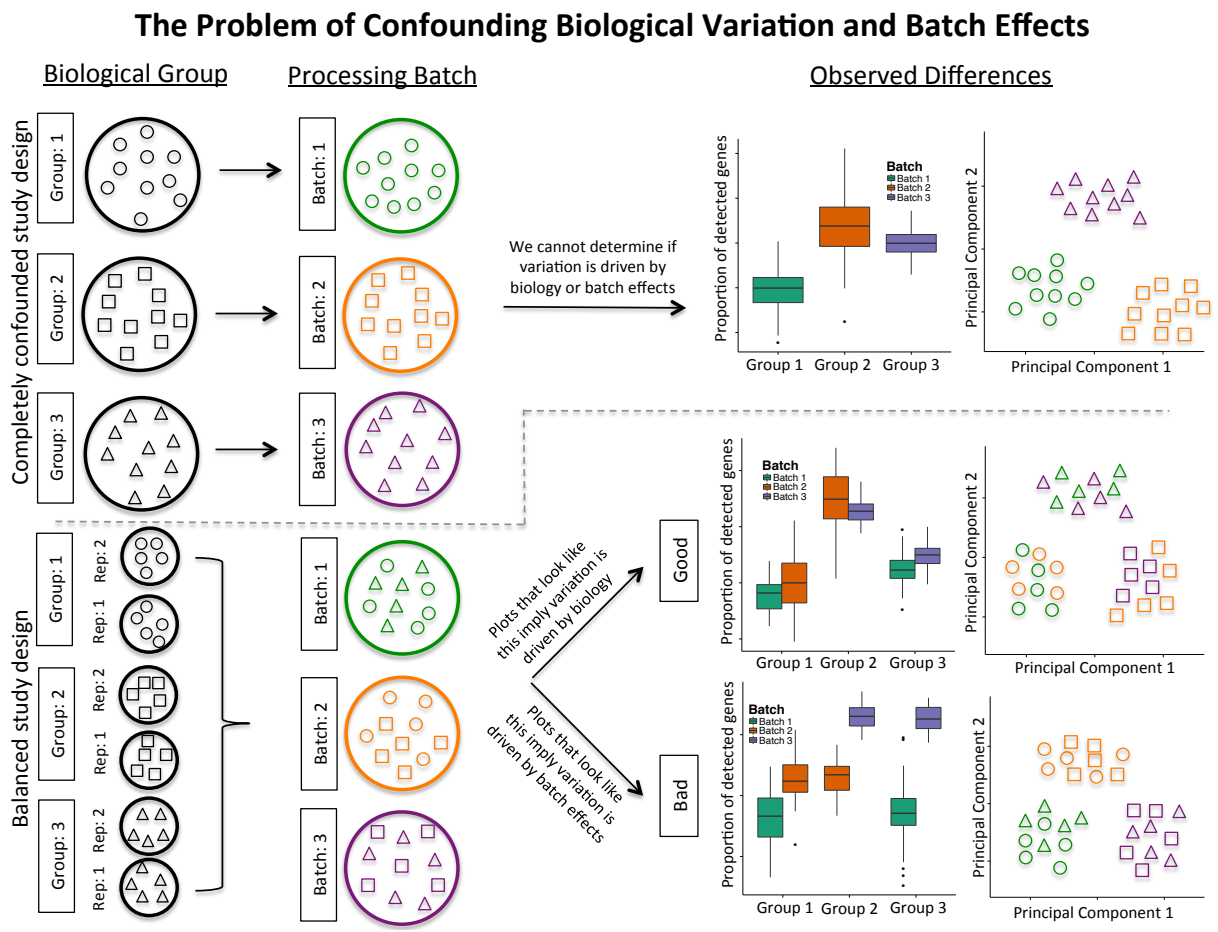

Figure 1: The problem of confounding biological variation and batch effects. The top section depicts a completely confounded study design of processing individual cells from three biological groups (represented by shapes) in three separate batches (represented by colors). In this case, we cannot determine if biology or batch effects drive the observed variation. The bottom section depicts a balanced study design consisting of multiple replicates (rep) split and processed across multiple batches. The use of multiple replicates allows observed variation be attributed to biology (cells cluster by shape) or batch effects (cells cluster by color).

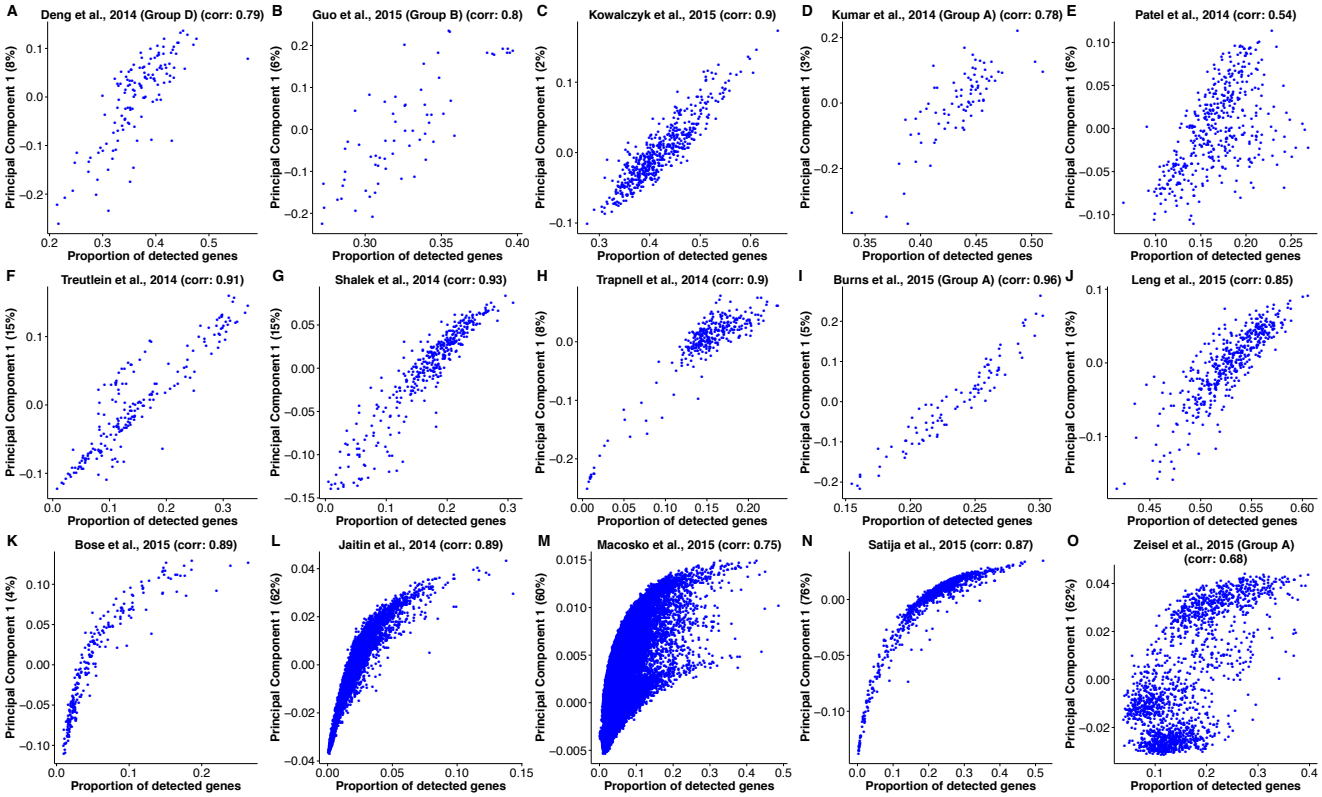

Figure 2: First principal component is strongly correlated with the proportion of detected genes. The principal components were computed using the publicly available processed data available on GEO. The proportion of detected genes was calculated using the same processed data on GEO for all studies except for Patel et al. (2014). In this case, because most non-detected genes were excluded from the publicly available processed data, we computed the proportion of detected genes from the raw data.

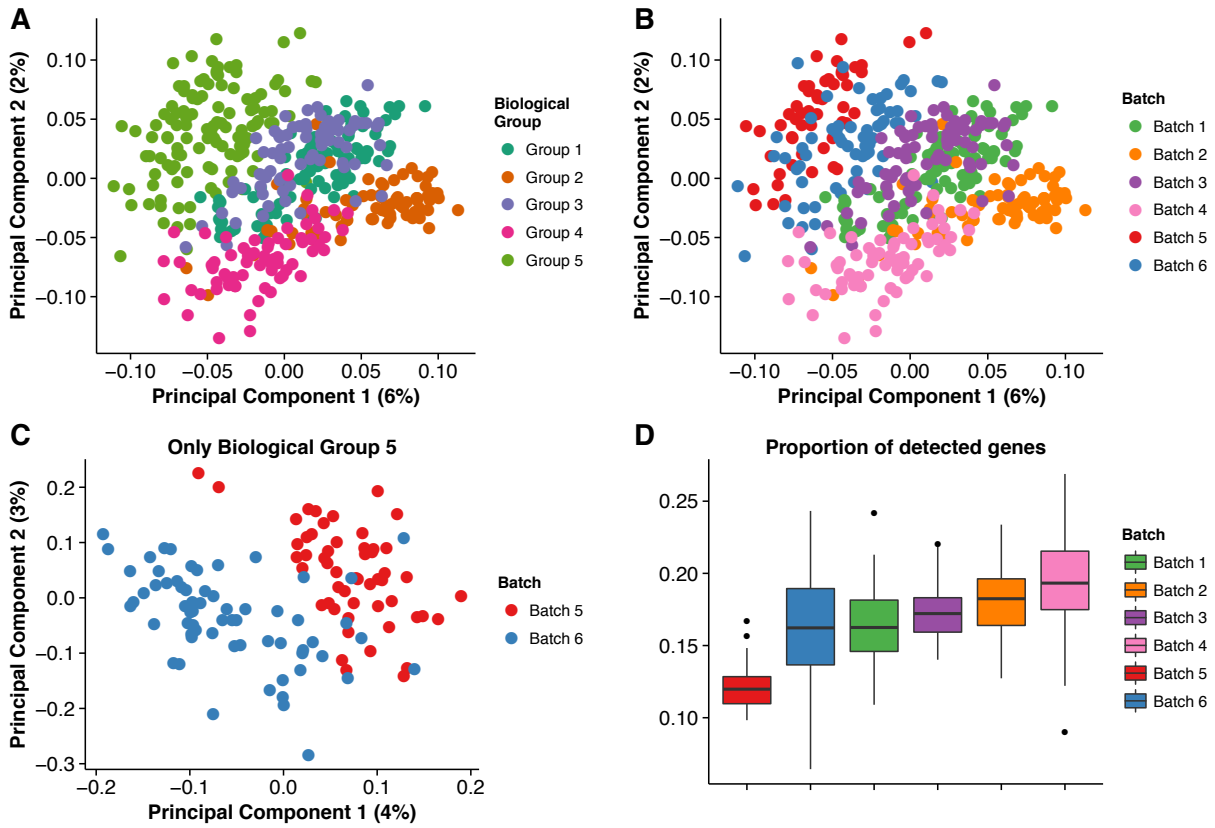

Figure 3: Illustration with public data[19] of how batch effects lead to differences in detection rates, which lead to apparent differences between biological groups. **(A)** Using principal components analysis, scRNA-Seq samples cluster by biological group, but the observed biological variation across groups is confounded with **(B)** technical variation from processing the cells in different batches. **(C)** Within one group (Group 5), the cells cluster by batch. **(D)** Furthermore, individual batches of cells have different proportions of detected genes, which may be driving the observed biological variation across groups.

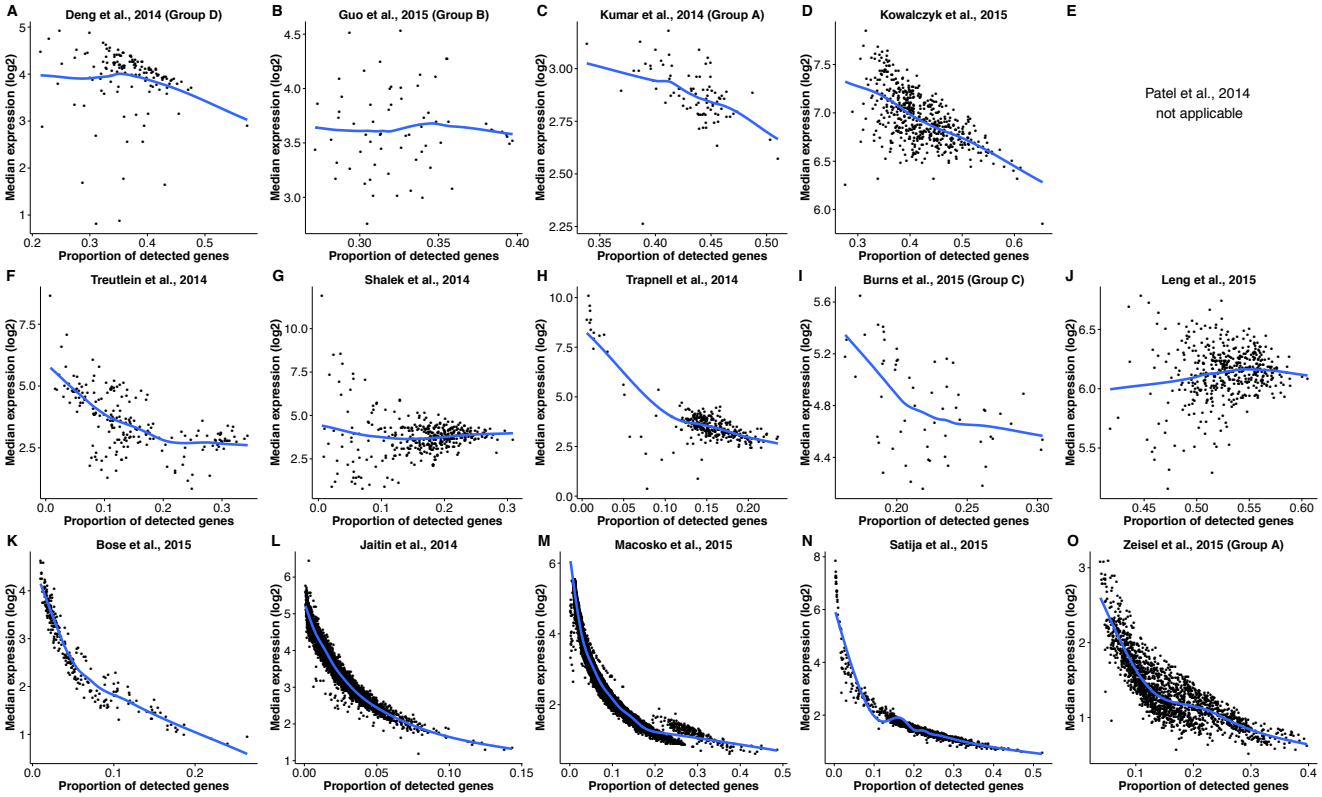

Figure 4: Non-linear relationship between the median gene expression and the proportion of detected genes using processed scRNA-Seq data available on GEO. Failure to account for differences of the proportion of detected genes between cells over-inflates the gene expression estimates of cells with a low proportion of detected genes. The blue curves were obtained by fitting a locally weighted scatter plot smoothing (loess) with a degree of 1 and span of 0.75 for all figures. Because the range of proportion of detected genes varied from study to study, the range of the x-axis differs across plots. **Note:** We could not include Patel et al. (2014) because of the row-standardization applied by the authors in the processed data available on GEO.