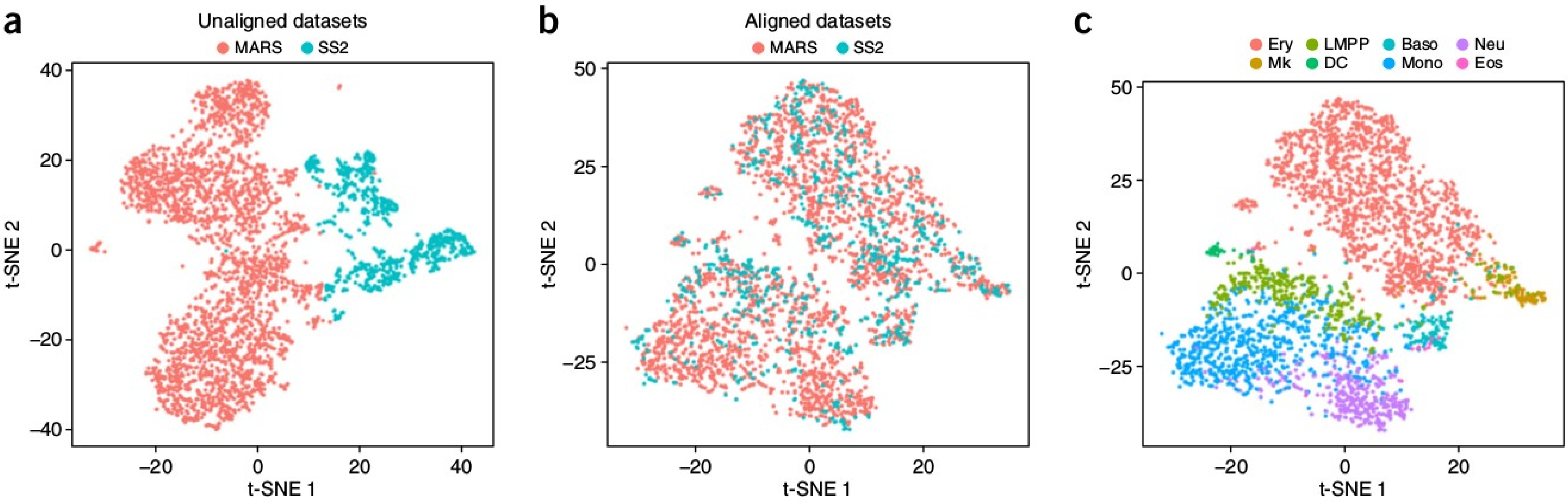
Removing Technical Variation in scRNAseq Data

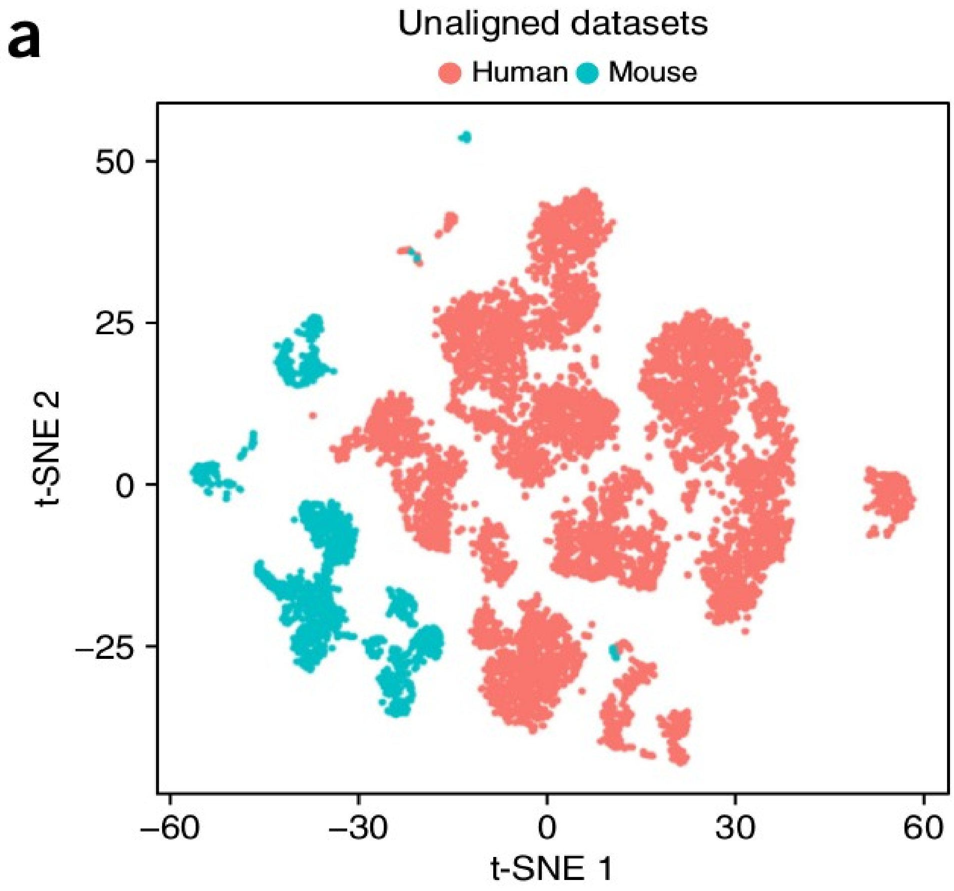
Nikolay Oskolkov
NBIS Long-Term Support (WABI)



scRNAseq course, Stockholm, 05.02.2019

Why to remove technical variation?

In order to facilitate discovering biological signal



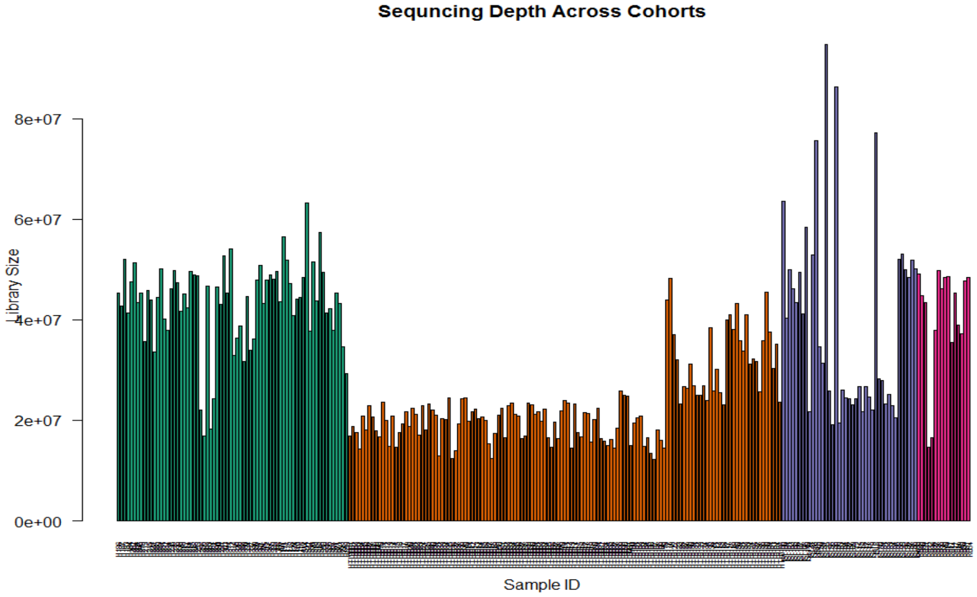
Batch-effects:
1) dates of sequencing
2) people done sequencing
3) flow-cells / plates
4) chemistry / protocol
5) lanes
6) read length
7) labs produced data
8) organisms
9) etc.

100% confounding: put cases
and
controls on different flow-cells

Normalization: correct for systematic variation in sequencing experiment
1) between samples (e.g. sequencing depth bias)
2) between features (e.g. gene length or GC content)

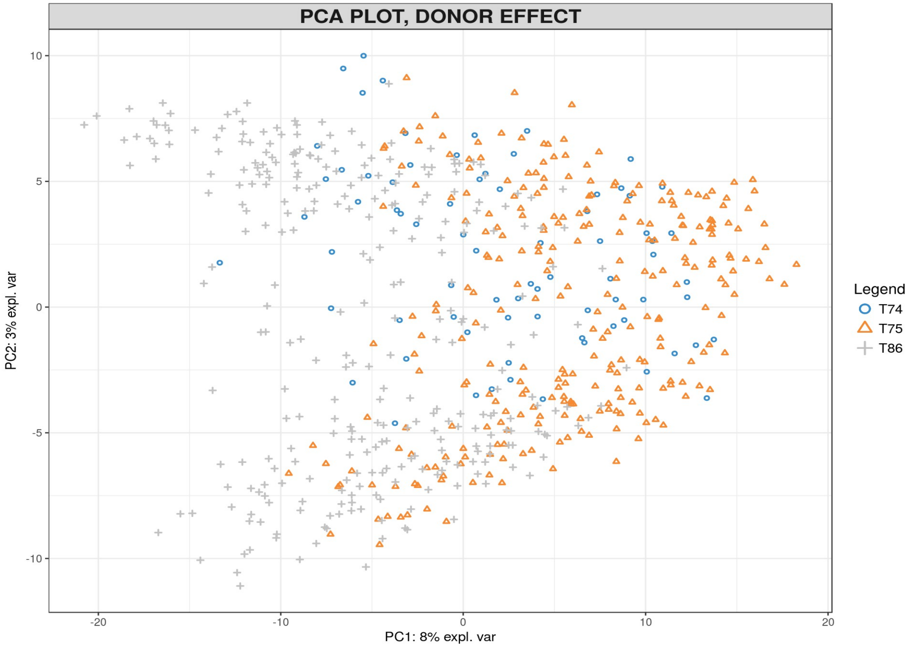
How to detect technical variation?

Difference in sequencing depth:



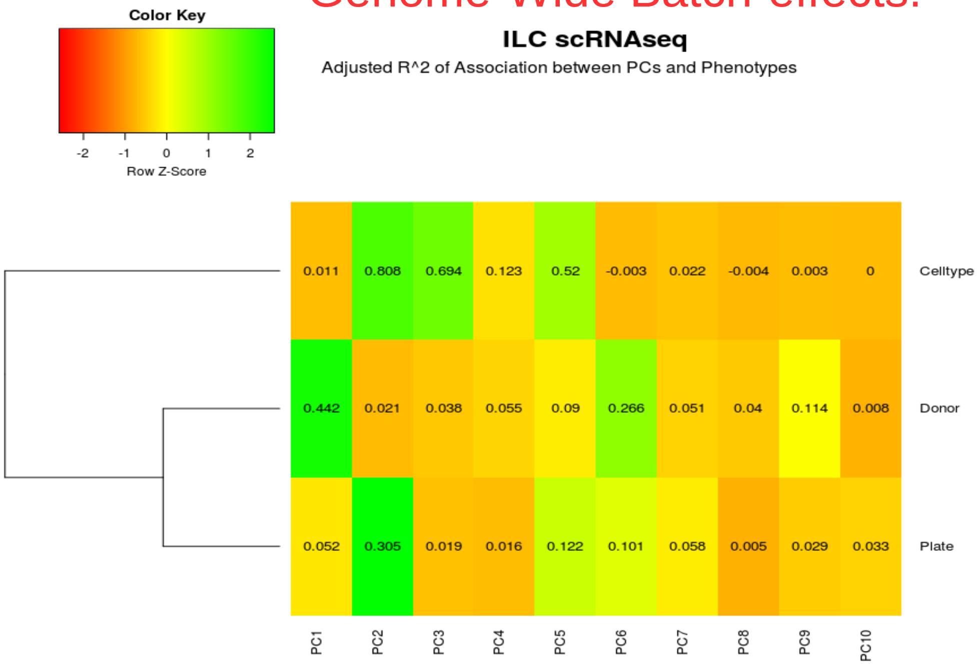
Sequncing Depth Across Cohorts

Genome-Wide Batch-effects:



PCA PLOT, DONOR EFFECT

Genome-Wide Batch-effects:

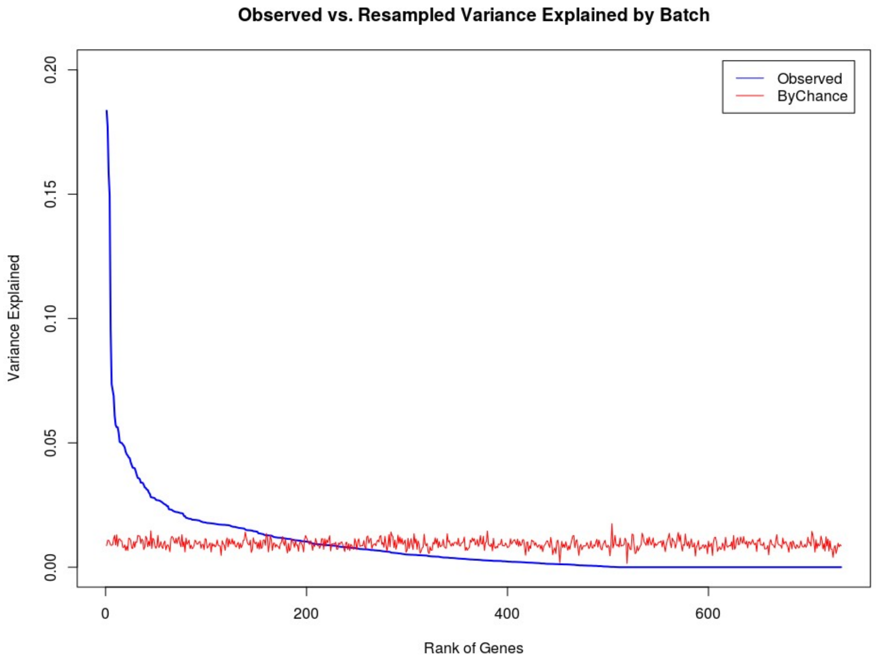


ILC scRNAseq

Adjusted R^2 of Association between PCs and Phenotypes

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Celltype	0.011	0.808	0.694	0.123	0.52	-0.003	0.022	-0.004	0.003	0
Donor	0.442	0.021	0.038	0.055	0.09	0.266	0.051	0.04	0.114	0.008
Plate	0.052	0.305	0.019	0.016	0.122	0.101	0.058	0.005	0.029	0.033

Per-Gene Batch-effects:



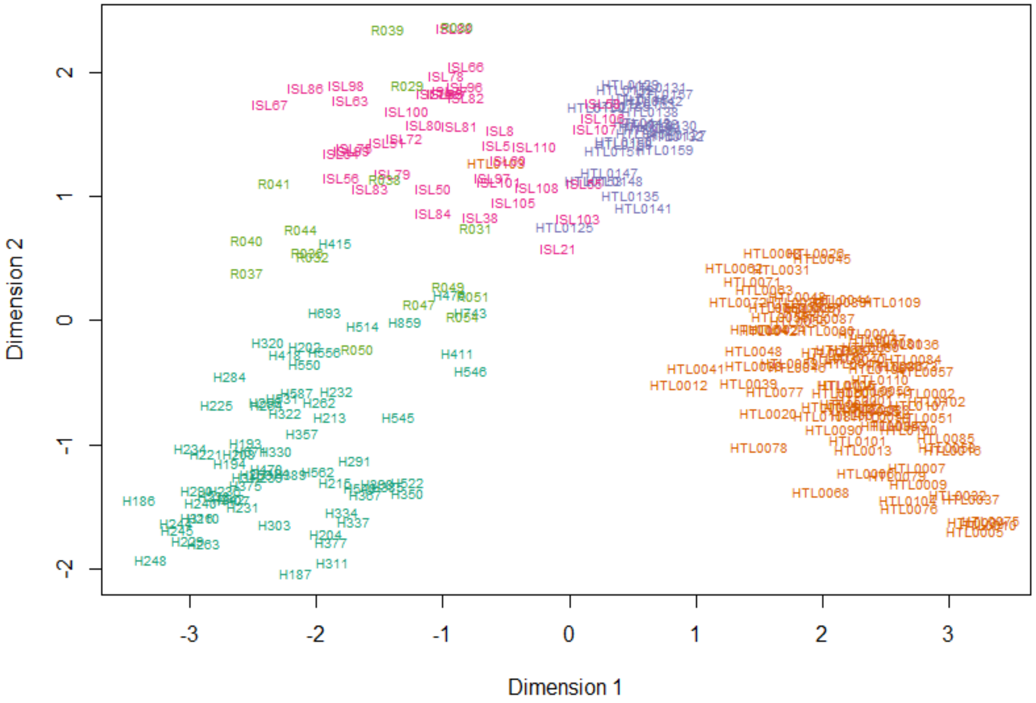
Observed vs. Resampled Variance Explained by Batch

How to correct for technical variation?

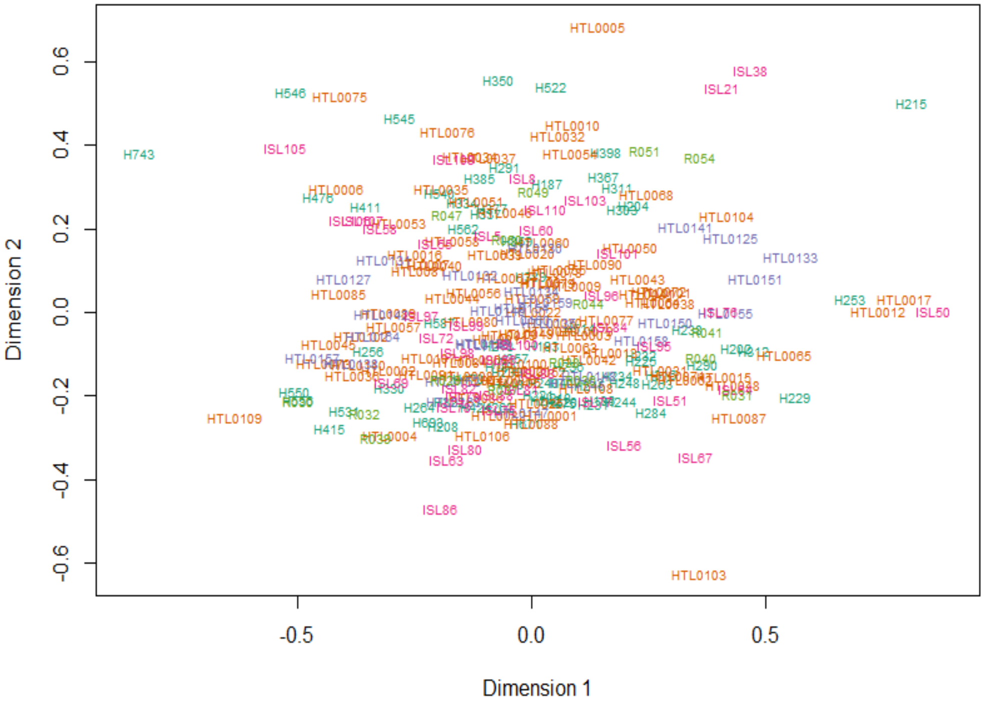
Normalization: normalize by library size (other choices: RPKM, SCnorm, Deconvolution)

Batch-effects: ComBat (supervised), SVA (unsupervised) etc.

Before ComBat



After ComBat



I do not recommend unsupervised batch-effects correction for scRNAseq data

I do not recommend library size normalization for any type of data

ComBat has a lot to do with more modern BASICS

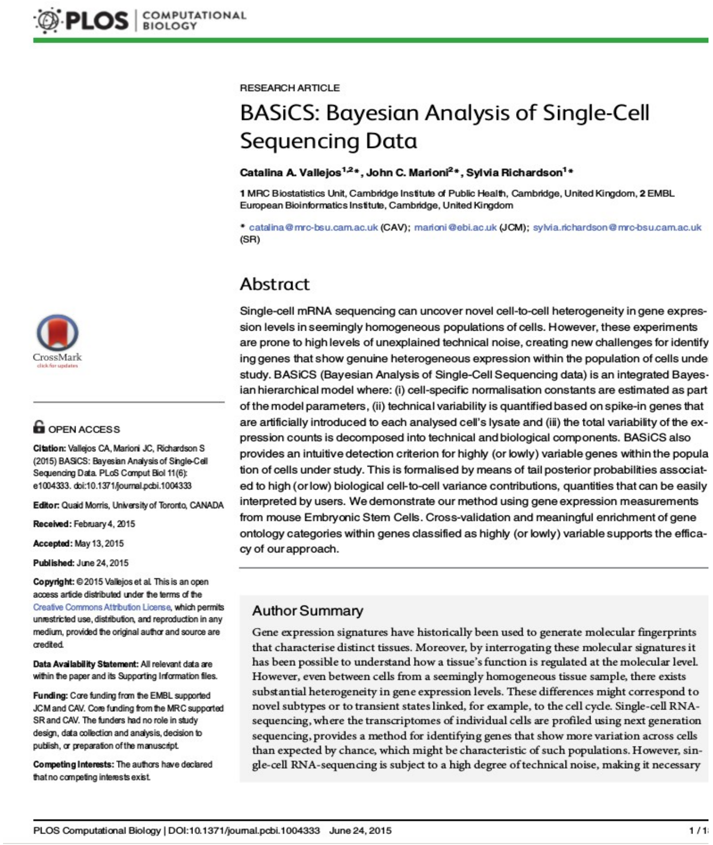
Bayesian framework for scRNAseq analysis:

1) normalization

2) batch correction

3) differential gene expression

4) detection of highly variable genes



PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

BASiCS: Bayesian Analysis of Single-Cell Sequencing Data

Catalina A. Vallejos[1,2]*, John C. Marioni[2]*, Sylvia Richardson[1]*

1 MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, United Kingdom, 2 EMBL European Bioinformatics Institute, Cambridge, United Kingdom

* catalina@mrc-bsu.cam.ac.uk (CAV); marioni@ebi.ac.uk (JCM); sylvia.richardson@mrc-bsu.cam.ac.uk (SR)

Abstract

Single-cell mRNA sequencing can uncover novel cell-to-cell heterogeneity in gene expression levels in seemingly homogeneous populations of cells. However, these experiments are prone to high levels of unexplained technical noise, creating new challenges for identifying genes that show genuine heterogeneous expression within the population of cells unde study. BASiCS (Bayesian Analysis of Single-Cell Sequencing data) is an integrated Bayesian hierarchical model where: (i) cell-specific normalisation constants are estimated as part of the model parameters, (ii) technical variability is quantified based on spike-in genes that are artificially introduced to each analysed cell's lysate and (iii) the total variability of the expression counts is decomposed into technical and biological components. BASiCS also provides an intuitive detection criterion for highly (or lowly) variable genes within the population of cells under study. This is formalised by means of tail posterior probabilities associated to high (or low) biological cell-to-cell variance contributions, quantities that can be easily interpreted by users. We demonstrate our method using gene expression measurements from mouse Embryonic Stem Cells. Cross-validation and meaningful enrichment of gene ontology categories within genes classified as highly (or lowly) variable supports the efficacy of our approach.

OPEN ACCESS

Citation: Vallejos CA, Marioni JC, Richardson S (2015) BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. PLoS Comput Biol 11(6): e1004333. doi:10.1371/journal.pcbi.1004333

Editor: Quaid Morris, University of Toronto, CANADA

Received: February 4, 2015

Accepted: May 13, 2015

Published: June 24, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

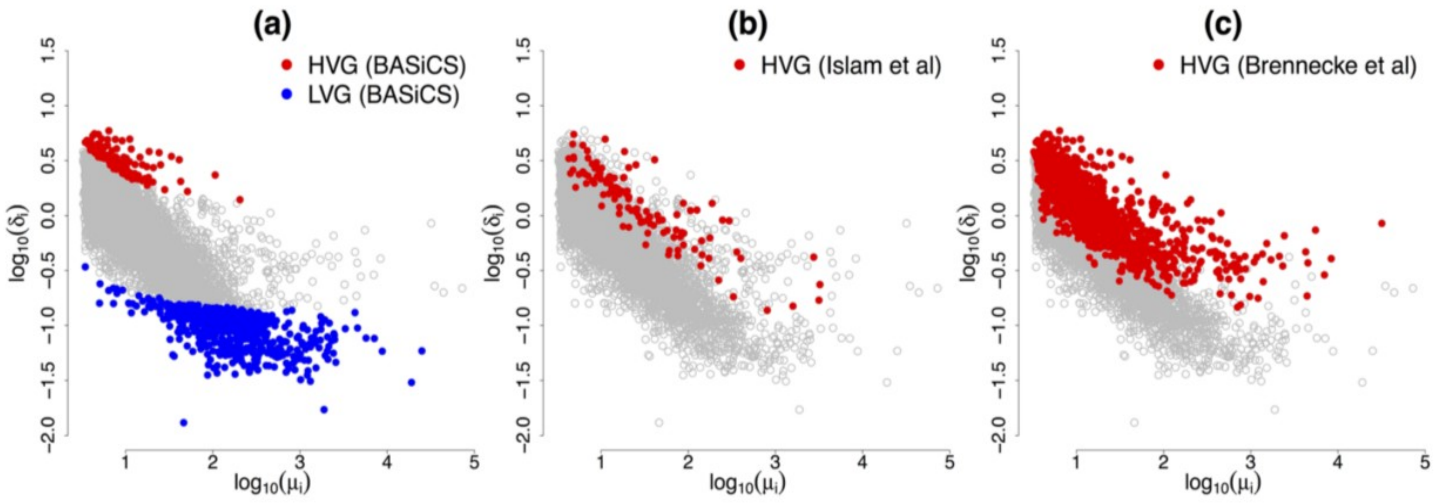
Funding: Core funding from the EMBL supported JCM and CAV. Core funding from the MRC supported SR and CAV. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Author Summary

Gene expression signatures have historically been used to generate molecular fingerprints that characterise distinct tissues. Moreover, by interrogating these molecular signatures it has been possible to understand how a tissue's function is regulated at the molecular level. However, even between cells from a seemingly homogeneous tissue sample, there exists substantial heterogeneity in gene expression levels. These differences might correspond to novel subtypes or to transient states linked, for example, to the cell cycle. Single-cell RNA-sequencing, where the transcriptomes of individual cells are profiled using next generation sequencing, provides a method for identifying genes that show more variation across cells than expected by chance, which might be characteristic of such populations. However, single-cell RNA-sequencing is subject to a high degree of technical noise, making it necessary

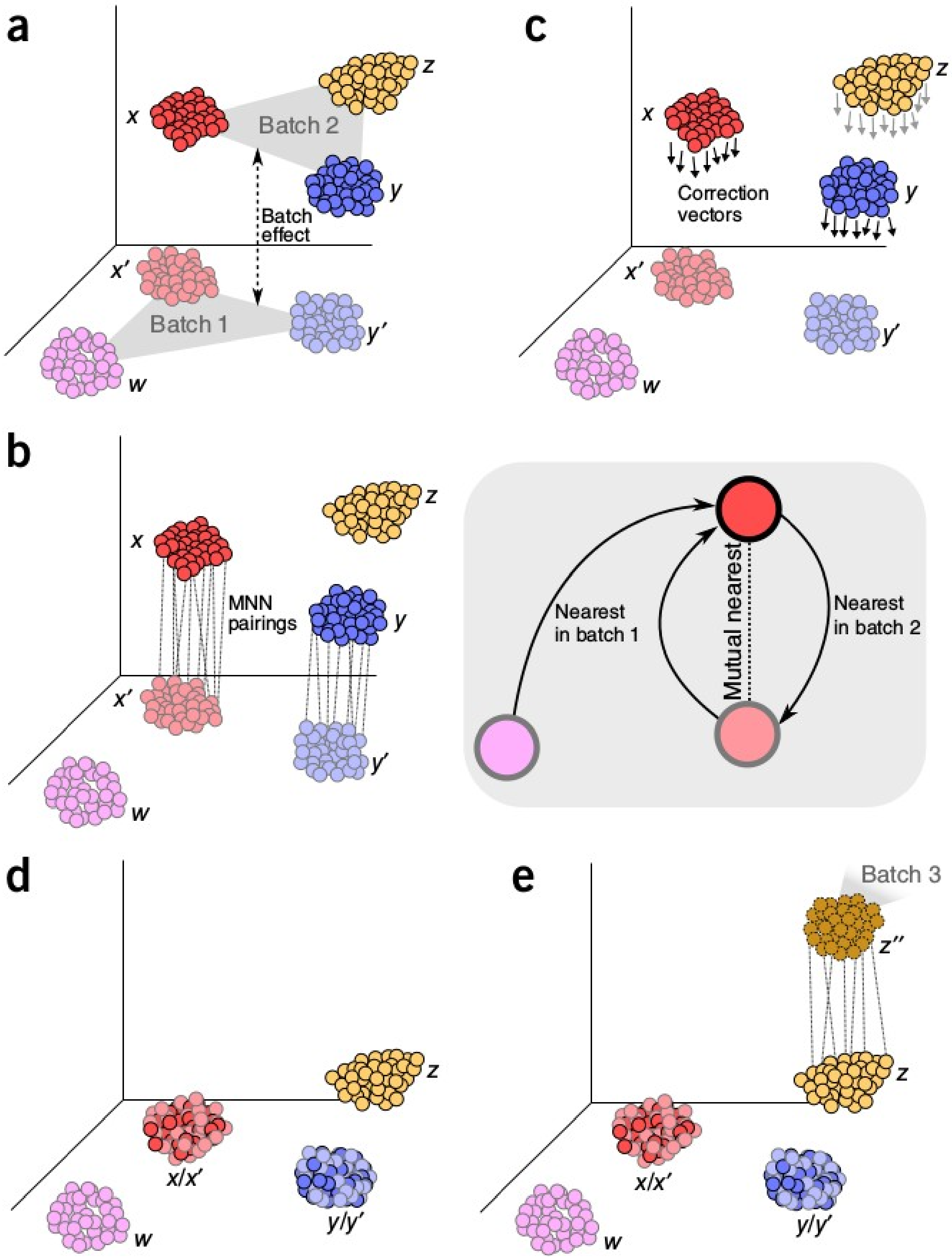
PLOS Computational Biology | DOI:10.1371/journal.pcbi.1004333 June 24, 2015 1 / 1



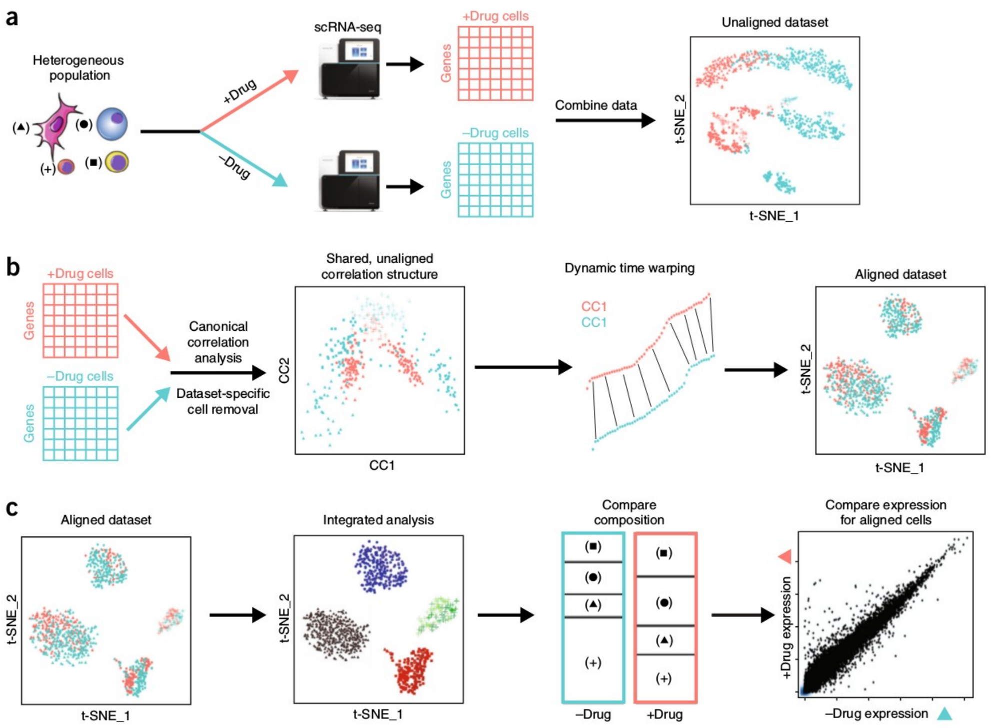
Other Methods for Batch Effects Corrections: Mutual Nearest Neighbors (MNN)

1) For each cell in batch 1 find a nearest neighbor in batch 2 and vice versa

2) Systematic difference in expression between MNN from batch 1 and 2 are to be removed

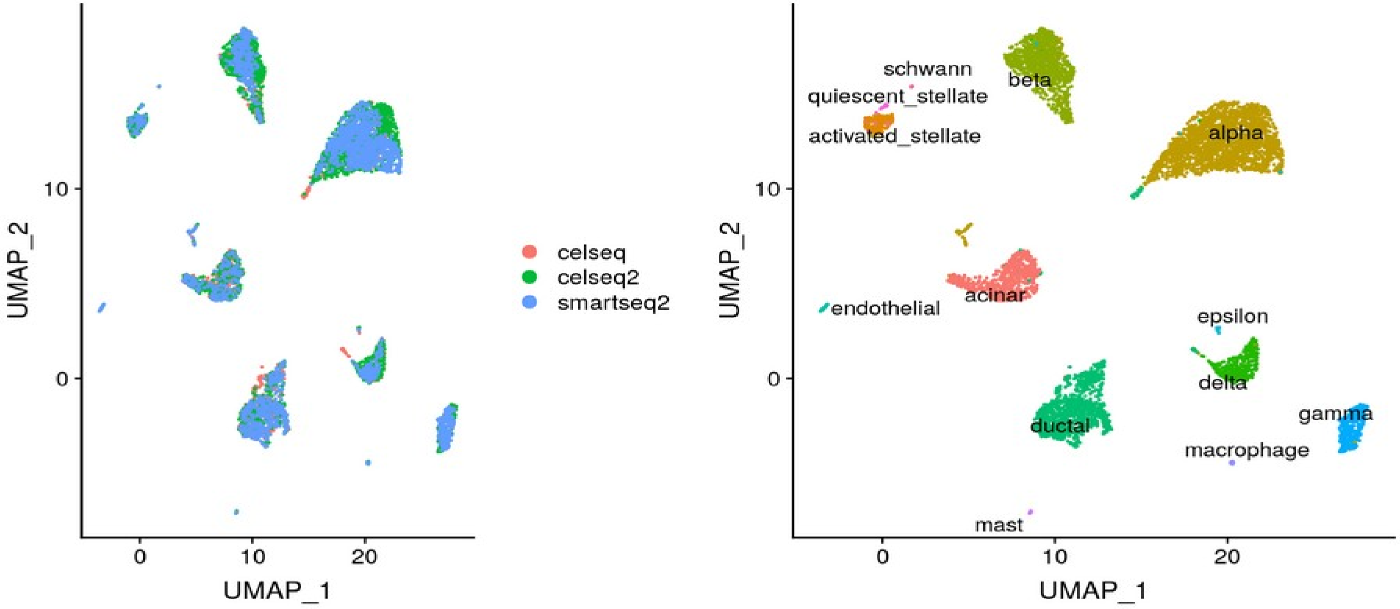


Other Methods for Batch Effects Corrections: Seurat and Canonical Correlation Analysis (CCA)

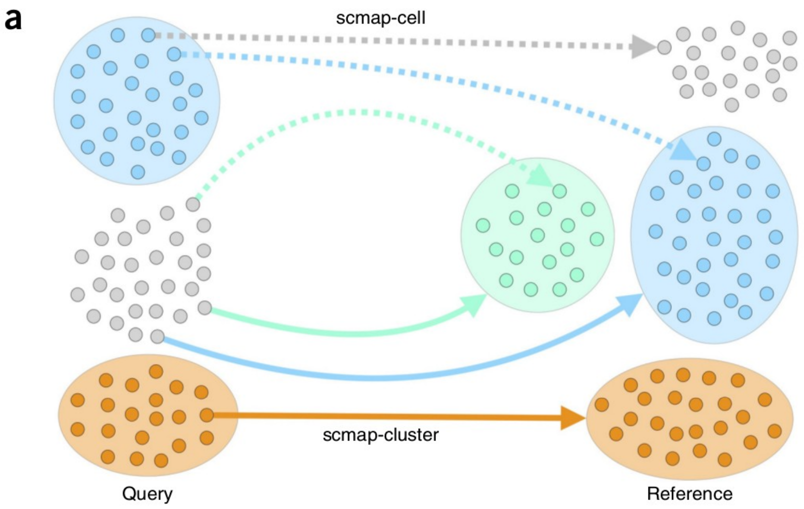


Find latent vectors that maximize correlation between data sets

Human Pancreatic Islets: 3 different technologies

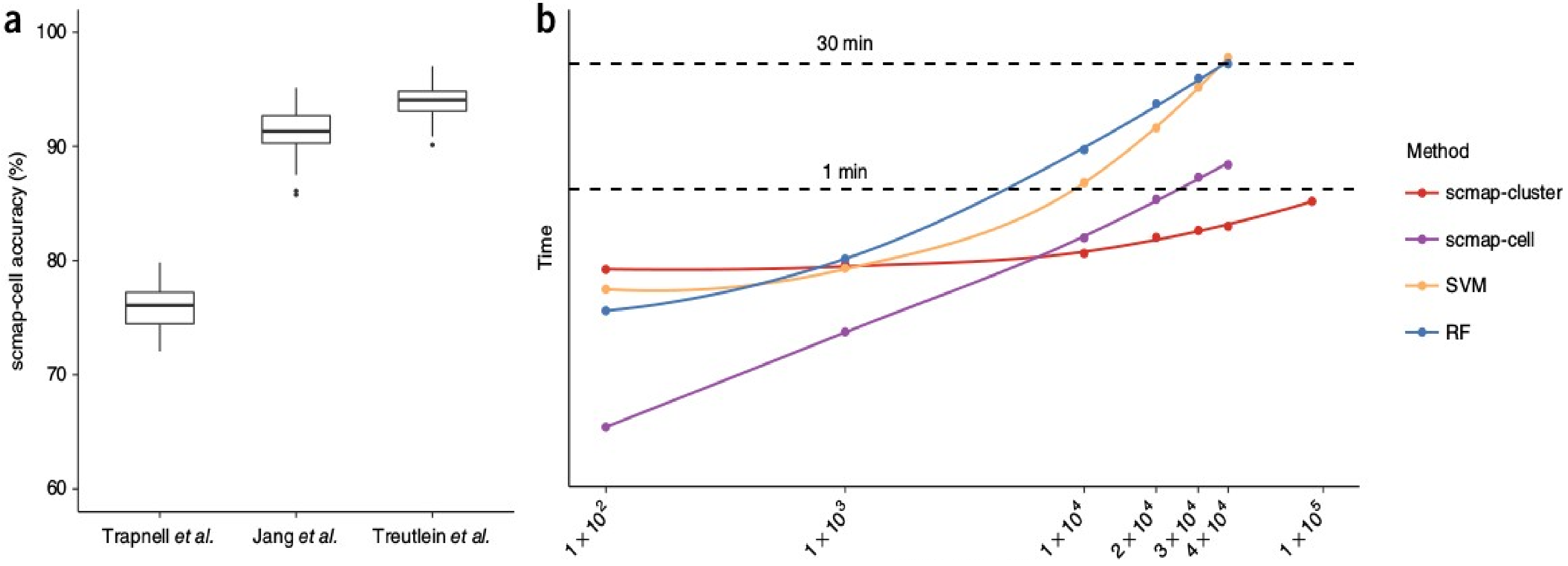


Other Methods for Batch Effects Corrections: Projection



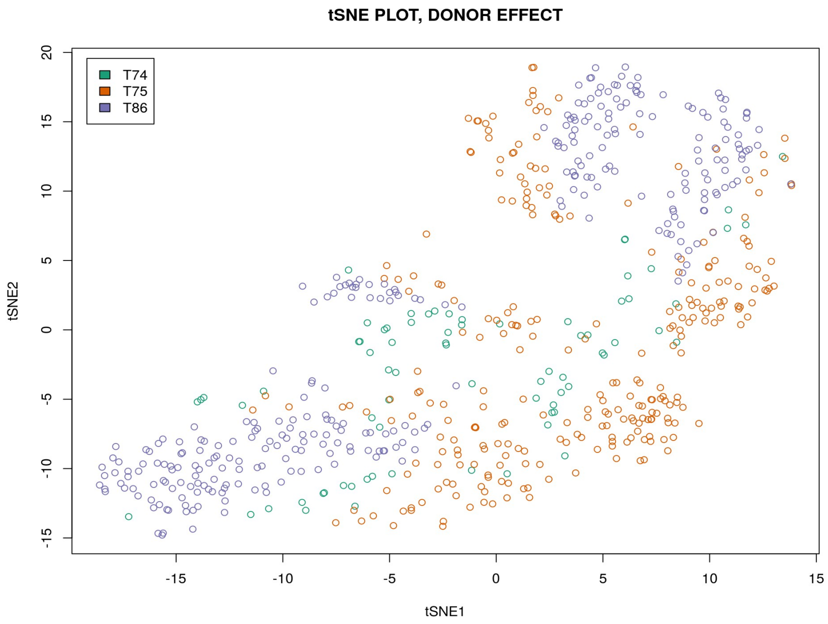
Based on Machine Learning principles

Not areal batch-effects correction but projection of cells from batch 1 to cells from batch 2

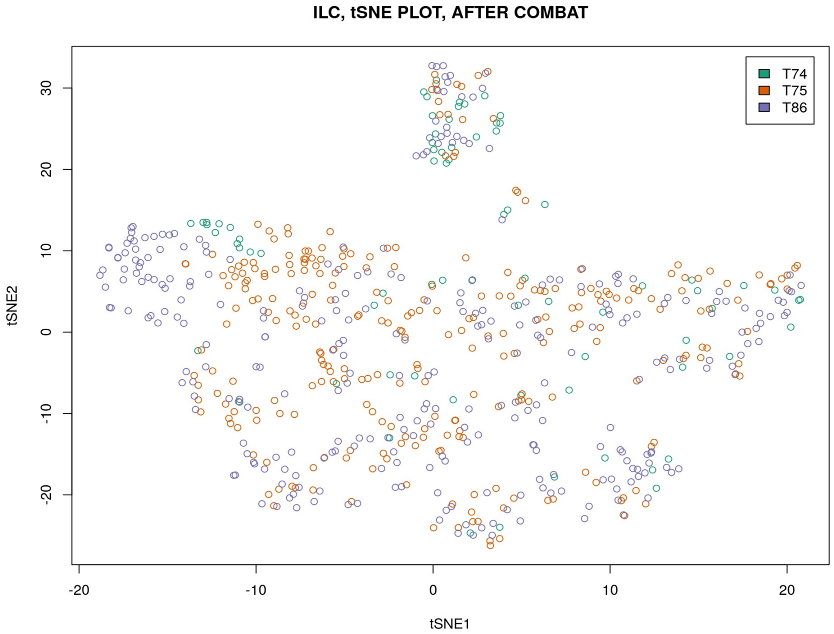


Other Methods: Do They Work?

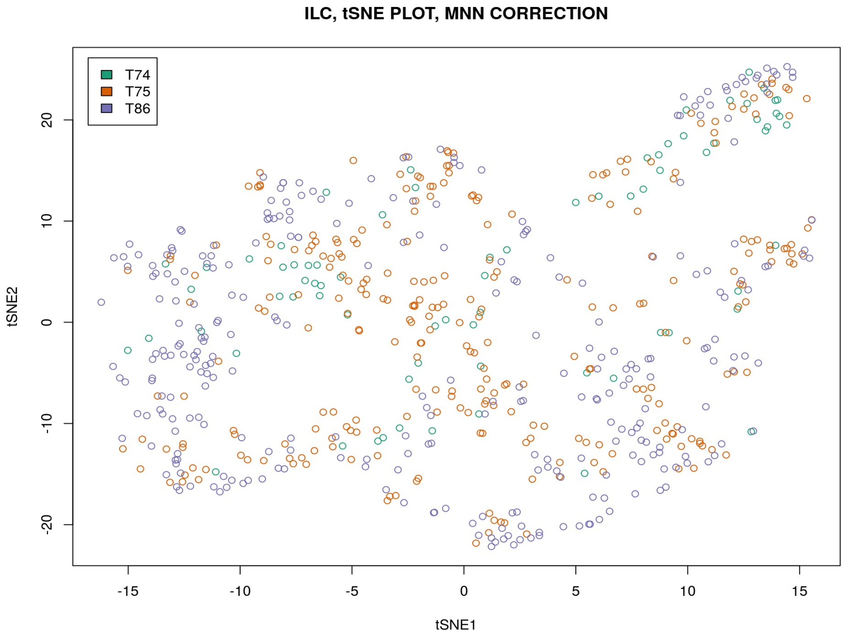
UNCORRECTED



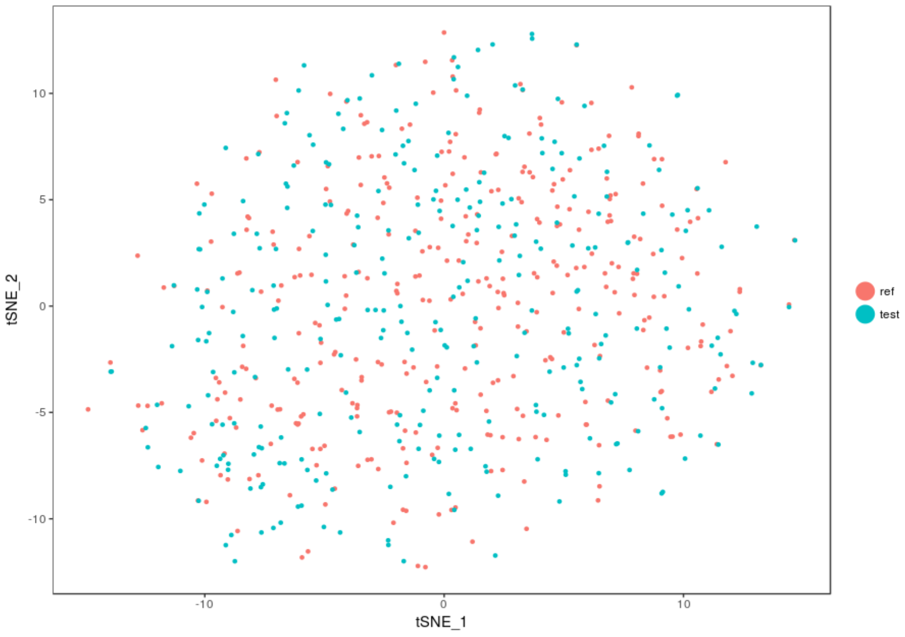
COMBAT



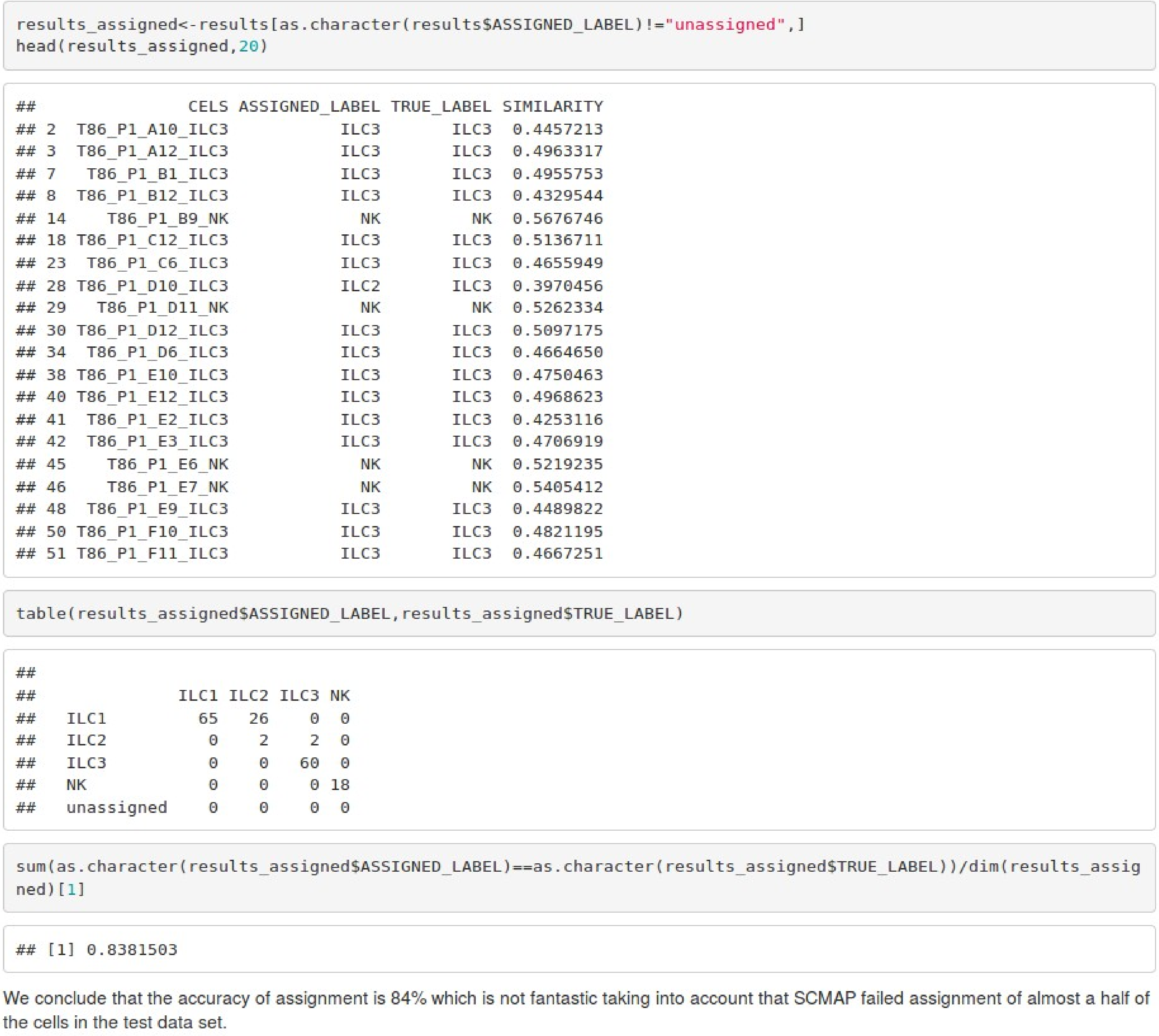
MNN



SEURAT



SCMAP



```
results_assigned<-results[as.character(results$ASSIGNED_LABEL)!="unassigned",]
head(results_assigned,20)
```

```
##               CELS ASSIGNED_LABEL TRUE_LABEL SIMILARITY
## 2  T86_P1_A10_ILC3           ILC3       ILC3  0.4457213
## 3  T86_P1_A12_ILC3           ILC3       ILC3  0.4963317
## 7   T86_P1_B1_ILC3           ILC3       ILC3  0.4955753
## 8  T86_P1_B12_ILC3           ILC3       ILC3  0.4329544
## 14     T86_P1_B9_NK             NK         NK  0.5676746
## 18 T86_P1_C12_ILC3           ILC3       ILC3  0.5136711
## 23  T86_P1_C6_ILC3           ILC3       ILC3  0.4655949
## 28 T86_P1_D10_ILC3           ILC2       ILC3  0.3970456
## 29    T86_P1_D11_NK             NK         NK  0.5262334
## 30 T86_P1_D12_ILC3           ILC3       ILC3  0.5097175
## 34  T86_P1_D6_ILC3           ILC3       ILC3  0.4664650
## 38 T86_P1_E10_ILC3           ILC3       ILC3  0.4750463
## 40 T86_P1_E12_ILC3           ILC3       ILC3  0.4968623
## 41  T86_P1_E2_ILC3           ILC3       ILC3  0.4253116
## 42  T86_P1_E3_ILC3           ILC3       ILC3  0.4706919
## 45     T86_P1_E6_NK             NK         NK  0.5219235
## 46     T86_P1_E7_NK             NK         NK  0.5405412
## 48  T86_P1_E9_ILC3           ILC3       ILC3  0.4489822
## 50 T86_P1_F10_ILC3           ILC3       ILC3  0.4821195
## 51 T86_P1_F11_ILC3           ILC3       ILC3  0.4667251
```

```
table(results_assigned$ASSIGNED_LABEL,results_assigned$TRUE_LABEL)
```

```
##             
##               ILC1 ILC2 ILC3 NK
##   ILC1         65   26    0  0
##   ILC2          0    2    2  0
##   ILC3          0    0   60  0
##   NK            0    0    0 18
##   unassigned    0    0    0  0
```

```
sum(as.character(results_assigned$ASSIGNED_LABEL)==as.character(results_assigned$TRUE_LABEL))/dim(results_assig
ned)[1]
```

```
## [1] 0.8381503
```

We conclude that the accuracy of assignment is 84% which is not fantastic taking into account that SCMAP failed assignment of almost a half of the cells in the test data set.

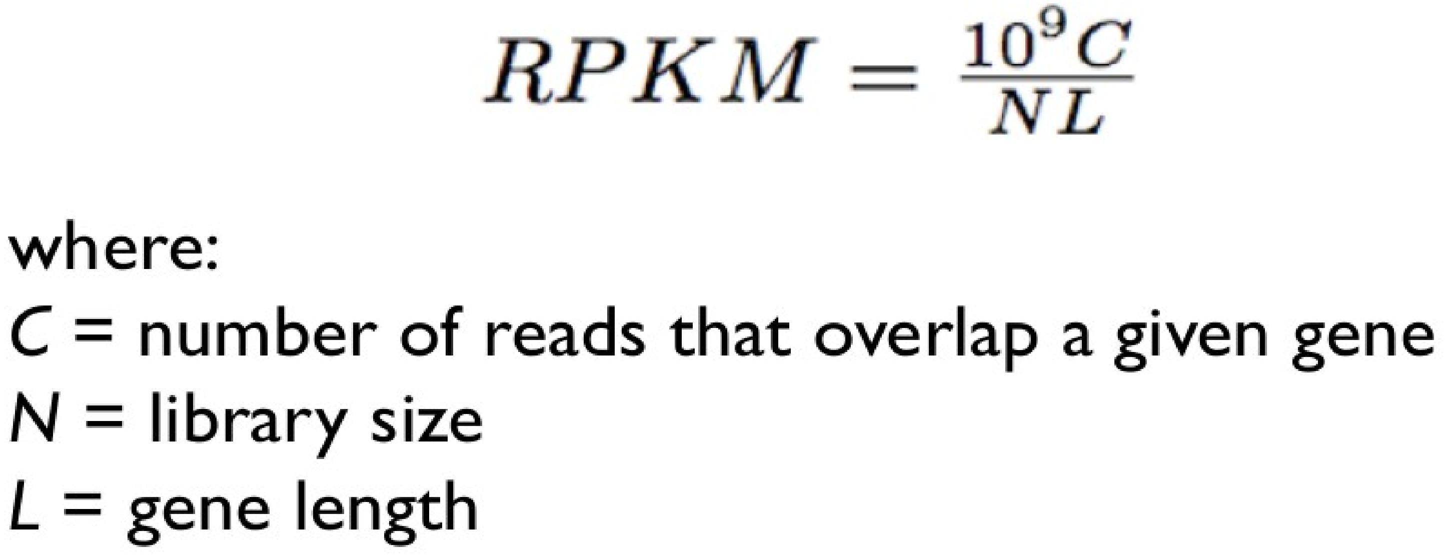
Brief Overview of Bulk RNAseq Normalization Methods:

RPKM, DESeq / TMM

RPKMs (FPKMs)

RPKM normalization is an extension of so-called library size normalization

Library size normalization: scaling such that library size is equal between all libraries



$$RPKM = \frac{10^9 C}{NL}$$

where:
C = number of reads that overlap a given gene
N = library size
L = gene length

Disadvantage: forced equalizing library sizes might eliminate true biological variation

DESeq

DESeq: create reference library based on geometric mean of all libraries, calculate size factors as ratios against the reference library

$$\begin{array}{c|ccc|}
 & s_1 & s_2 & s_k \\
g_1 & c_{11} & c_{12} & c_{1k} \\
g_2 & c_{21} & c_{22} & c_{2k} \\
g_n & c_{n1} & c_{n2} & c_{nk}
\end{array}
\begin{bmatrix} 1/g_{1r} & 1/g_{2r} & 1/g_{nr} \end{bmatrix}
\longrightarrow
\begin{array}{c|ccc|}
 & s_1 & s_2 & s_k \\
g_1 & r_{11} & r_{12} & r_{1k} \\
g_2 & r_{21} & r_{22} & r_{2k} \\
g_n & r_{n1} & r_{n2} & r_{nk}
\end{array}$$

estimate the relative depth of the library

$$S_1 = \text{median}(r_{x1})$$

take the mean for each row to obtain a reference sample

$$\begin{bmatrix} g_{1r} \\ g_{2r} \\ g_{nr} \end{bmatrix}$$

estimate the depth ratio for each gene

$$\begin{array}{c|ccc|}
 & s_1 & s_2 & s_k \\
g_1 & c_{11} & c_{12} & c_{1k} \\
g_2 & c_{21} & c_{22} & c_{2k} \\
g_n & c_{n1} & c_{n2} & c_{nk}
\end{array}
\quad
\begin{bmatrix} S_1 \\ S_2 \\ S_k \end{bmatrix}$$

Disadvantage: DESeq is based on ratio construction

scRNAseq – Specific Normalization Methods:

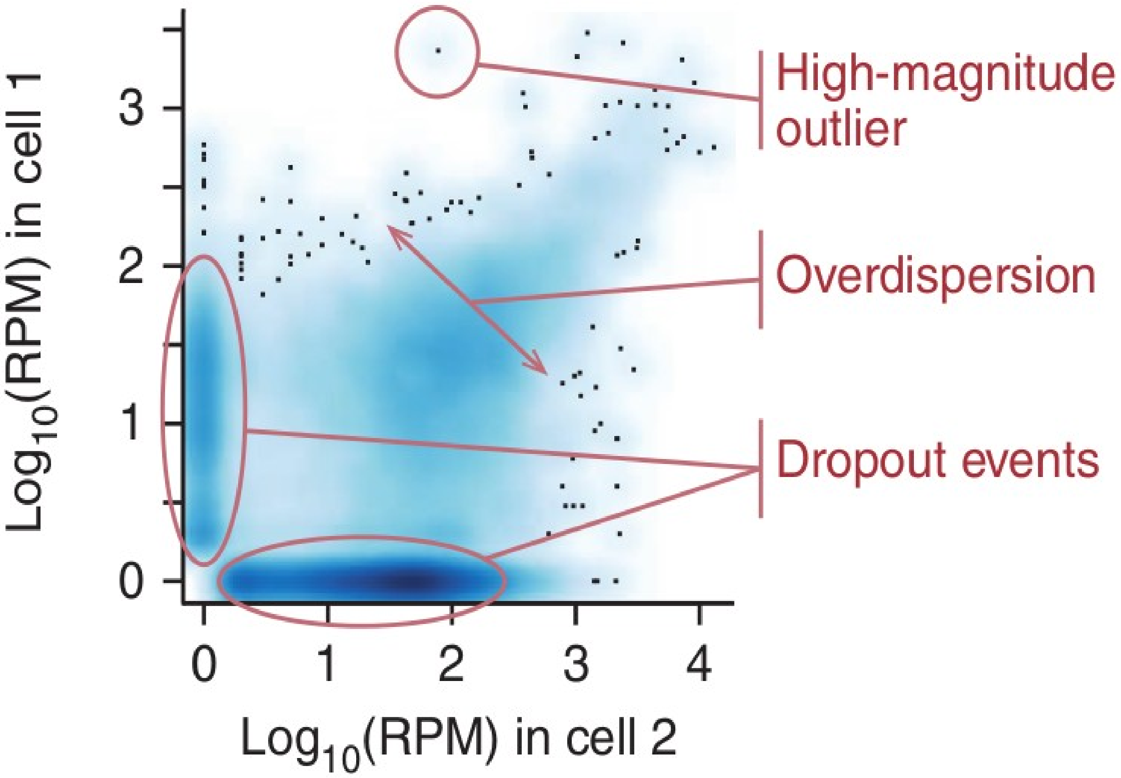
1) Deconvolution (Pooling-Across-Cells)

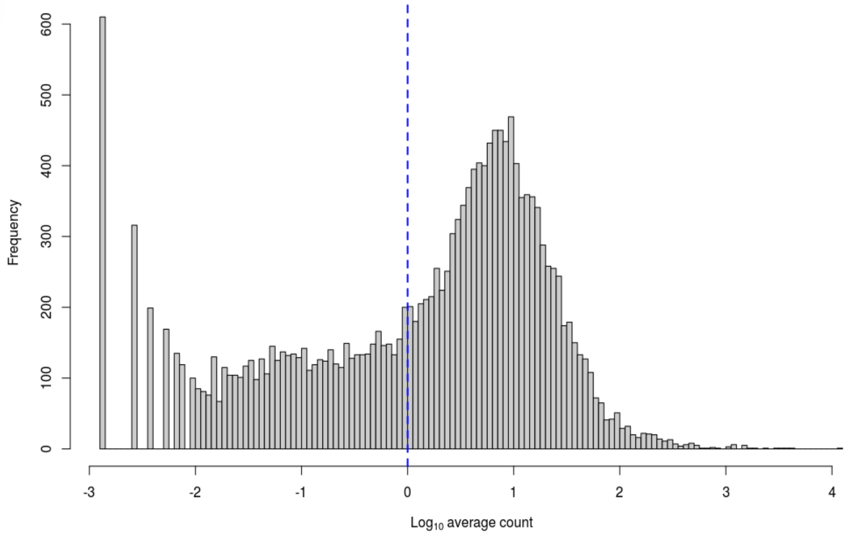
2) SCnorm (Expression-Depth Relation)

Lots of zero-counts is main challenge in scRNAseq

scRNAseq expression counts have typically ~80% of zero-counts

This is due to: 1) low amounts of RNA per cell, 2) RNA capture efficiency





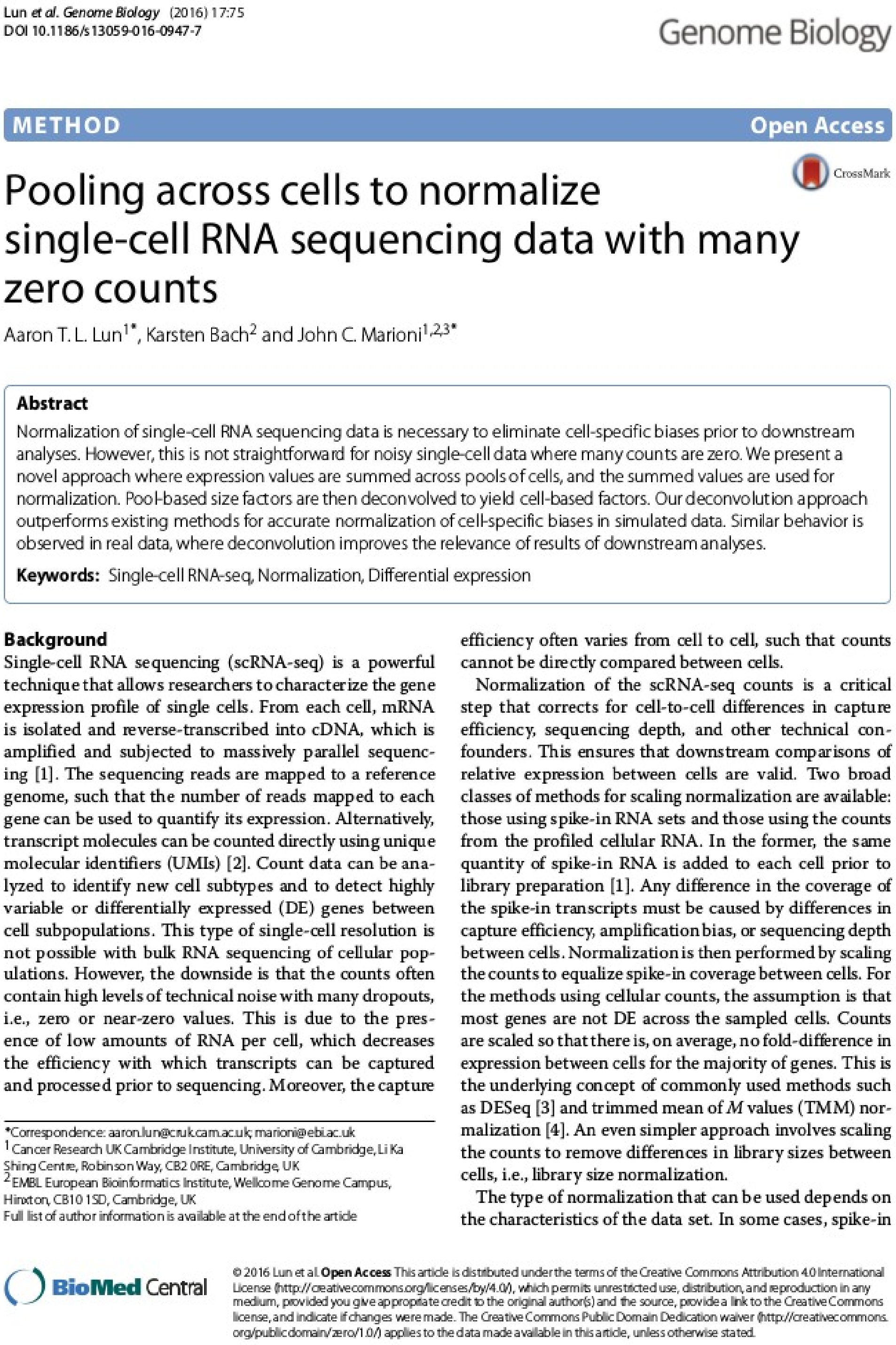
We want to correct for sequencing depth and cell-to-cell difference in RNA capture efficiency

3 common normalization methods used for bulk RNAseq: 1) TMM, 2) DESeq, 3) RPKM

Main assumption of all 3 methods: most of the genes are not differentially expressed

TMM and DESeq rely on ratios of counts, therefore diverge when lots of zero-counts

Deconvolution Normalization Method



Lun *et al. Genome Biology* (2016) 17:75
DOI 10.1186/s13059-016-0947-7

Genome Biology

METHOD Open Access

Pooling across cells to normalize single-cell RNA sequencing data with many zero counts

Aaron T. L. Lun[1*], Karsten Bach[2] and John C. Marioni[1,2,3*]

Abstract

Normalization of single-cell RNA sequencing data is necessary to eliminate cell-specific biases prior to downstream analyses. However, this is not straightforward for noisy single-cell data where many counts are zero. We present a novel approach where expression values are summed across pools of cells, and the summed values are used for normalization. Pool-based size factors are then deconvolved to yield cell-based factors. Our deconvolution approach outperforms existing methods for accurate normalization of cell-specific biases in simulated data. Similar behavior is observed in real data, where deconvolution improves the relevance of results of downstream analyses.

Keywords: Single-cell RNA-seq, Normalization, Differential expression

Background

Single-cell RNA sequencing (scRNA-seq) is a powerful technique that allows researchers to characterize the gene expression profile of single cells. From each cell, mRNA is isolated and reverse-transcribed into cDNA, which is amplified and subjected to massively parallel sequencing [1]. The sequencing reads are mapped to a reference genome, such that the number of reads mapped to each gene can be used to quantify its expression. Alternatively, transcript molecules can be counted directly using unique molecular identifiers (UMIs) [2]. Count data can be analyzed to identify new cell subtypes and to detect highly variable or differentially expressed (DE) genes between cell subpopulations. This type of single-cell resolution is not possible with bulk RNA sequencing of cellular populations. However, the downside is that the counts often contain high levels of technical noise with many dropouts, i.e., zero or near-zero values. This is due to the presence of low amounts of RNA per cell, which decreases the efficiency with which transcripts can be captured and processed prior to sequencing. Moreover, the capture efficiency often varies from cell to cell, such that counts cannot be directly compared between cells.

Normalization of the scRNA-seq counts is a critical step that corrects for cell-to-cell differences in capture efficiency, sequencing depth, and other technical confounders. This ensures that downstream comparisons of relative expression between cells are valid. Two broad classes of methods for scaling normalization are available: those using spike-in RNA sets and those using the counts from the profiled cellular RNA. In the former, the same quantity of spike-in RNA is added to each cell prior to library preparation [1]. Any difference in the coverage of the spike-in transcripts must be caused by differences in capture efficiency, amplification bias, or sequencing depth between cells. Normalization is then performed by scaling the counts to equalize spike-in coverage between cells. For the methods using cellular counts, the assumption is that most genes are not DE across the sampled cells. Counts are scaled so that there is, on average, no fold-difference in expression between cells for the majority of genes. This is the underlying concept of commonly used methods such as DESeq [3] and trimmed mean of *M* values (TMM) normalization [4]. An even simpler approach involves scaling the counts to remove differences in library sizes between cells, i.e., library size normalization.

The type of normalization that can be used depends on the characteristics of the data set. In some cases, spike-in

*Correspondence: aaron.lun@cruk.cam.ac.uk; marioni@ebi.ac.uk
[1] Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, CB2 0RE, Cambridge, UK
[2] EMBL European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, Cambridge, UK
Full list of author information is available at the end of the article

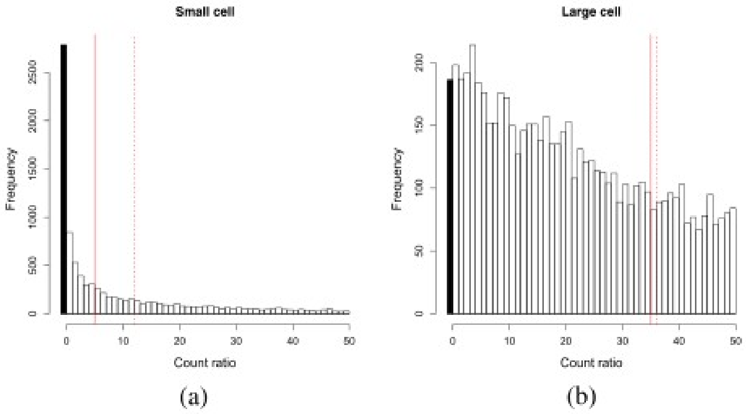


Fig. 2 Illustration of the effect of removing stochastic zeroes (*black*) from the distribution of ratios across all genes. Distributions are shown for cells with **a** small and **b** large θ_j. The estimated median ratio (*dashed*) is increased beyond the true median (*full*) upon removal of zeroes, which results in overestimation of the size factor for the cell. This effect is more pronounced for cells with small θ_j that have greater numbers of zeroes, compared to cells with large θ_j where the estimated and true medians are more similar

of an arbitrary set of cells $\mathcal{S}_k$. Define V_{ik} as the sum of Z_{ij} across all cells in $\mathcal{S}_k$, which has an expectation of

$$E(V_{ik}) = \lambda_{i0} \sum_{j \in \mathcal{S}_k} \theta_j t_j^{-1}.$$

The observed values of V_{ik} across all genes constitute an overall expression profile for the pool of cells corresponding to $\mathcal{S}_k$. Also define U_i as the mean of Z_{ij} across all N cells in the entire data set, which has an expectation of

$$E(U_i) = \lambda_{i0} N^{-1} \sum_{j \in \mathcal{S}_0} \theta_j t_j^{-1}$$

where $\mathcal{S}_0$ refers to the set of all cells in the data set. The observed values of U_i across all genes represent the expression profile for an averaged reference pseudo-cell.

The cell pool k is then normalized against this reference pseudo-cell. Define R_{ik} as the ratio of V_{ik} to U_i for the non-DE gene i. The expectation of R_{ik} represents the true size factor for the pooled cells in $\mathcal{S}_k$, and is written as

$$E(R_{ik}) \approx \frac{E(V_{ik})}{E(U_i)} = \frac{\sum_{\mathcal{S}_k} \theta_j t_j^{-1}}{N^{-1} \sum_{\mathcal{S}_0} \theta_j t_j^{-1}} = \frac{\sum_{\mathcal{S}_k} \theta_j t_j^{-1}}{C} \quad (1)$$

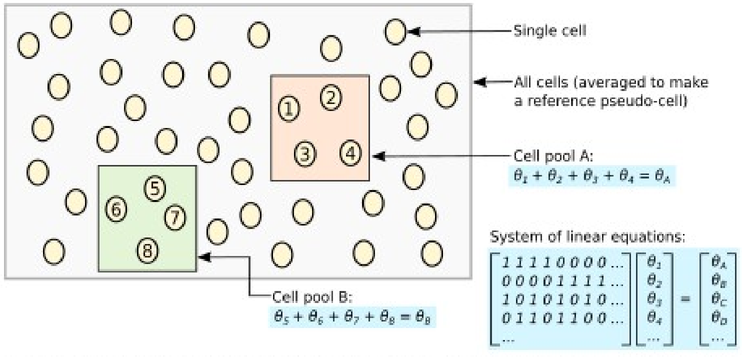

1
2
3
4
5
6
7
8
Single cell
All cells (averaged to make
a reference pseudo-cell)
Cell pool A:
$\theta_1 + \theta_2 + \theta_3 + \theta_4 = \theta_A$
Cell pool B:
$\theta_5 + \theta_6 + \theta_7 + \theta_8 = \theta_B$
System of linear equations:
$\begin{bmatrix} 1\,1\,1\,1\,0\,0\,0\,0 \ldots \\ 0\,0\,0\,0\,1\,1\,1\,1 \ldots \\ 1\,0\,1\,0\,1\,0\,1\,0 \ldots \\ 0\,1\,1\,0\,1\,1\,0\,0 \ldots \\ \ldots \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \ldots \end{bmatrix} = \begin{bmatrix} \theta_A \\ \theta_B \\ \theta_C \\ \theta_D \\ \ldots \end{bmatrix}$

Fig. 3 Schematic of the deconvolution method. All cells in the data set are averaged to make a reference pseudo-cell. Expression values for cells in pool A are summed together and normalized against the reference to yield a pool-based size factor θ_A. This is equal to the sum of the cell-based factors θ_j for cells $j = 1$–4 and can be used to formulate a linear equation. (For simplicity, the t_j term is assumed to be unity here.) Repeating this for multiple pools (e.g., pool B) leads to the construction of a linear system that can be solved to estimate θ_j for each cell j

Benchmarking: Deconvolution Method Performs Best

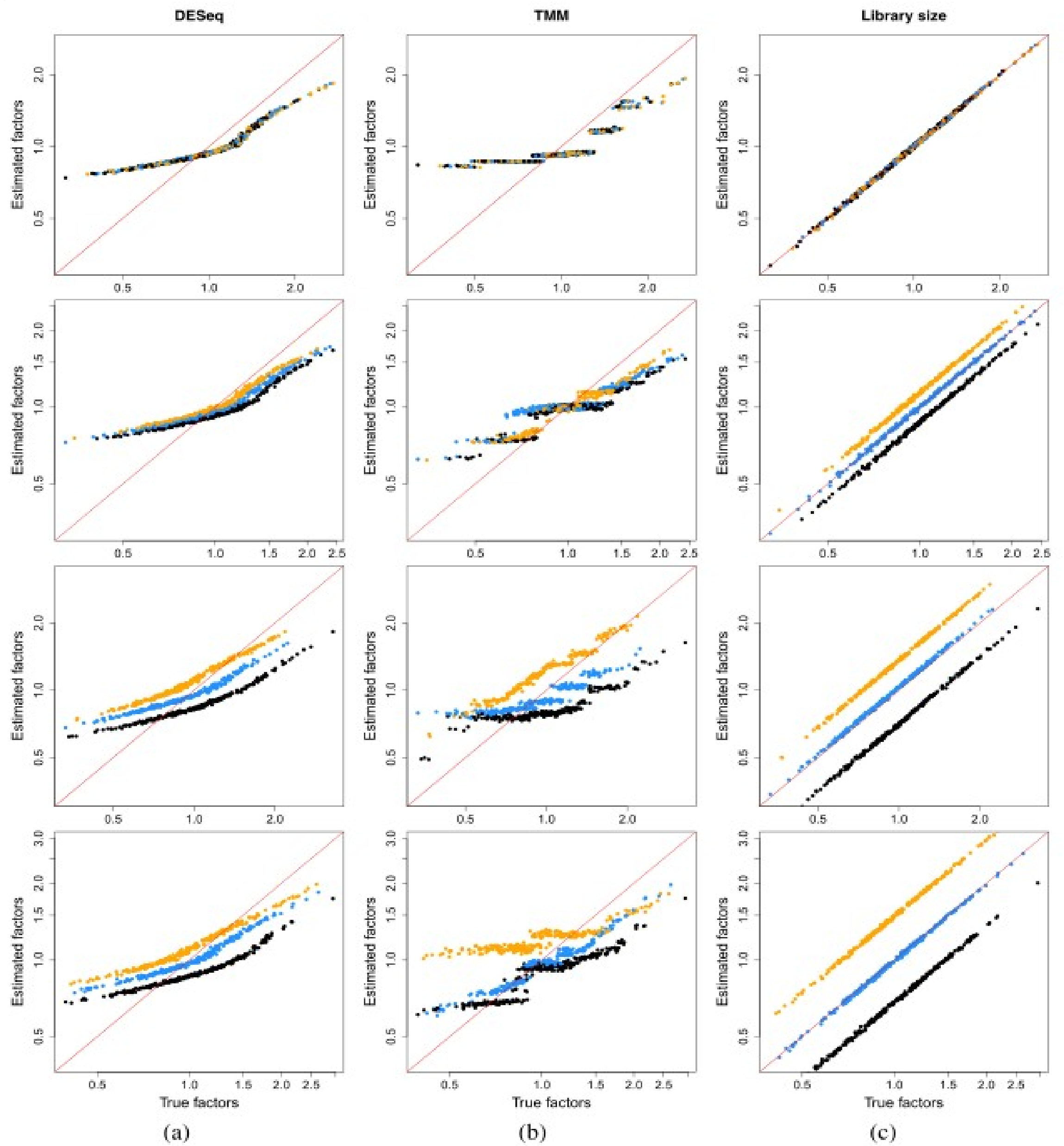
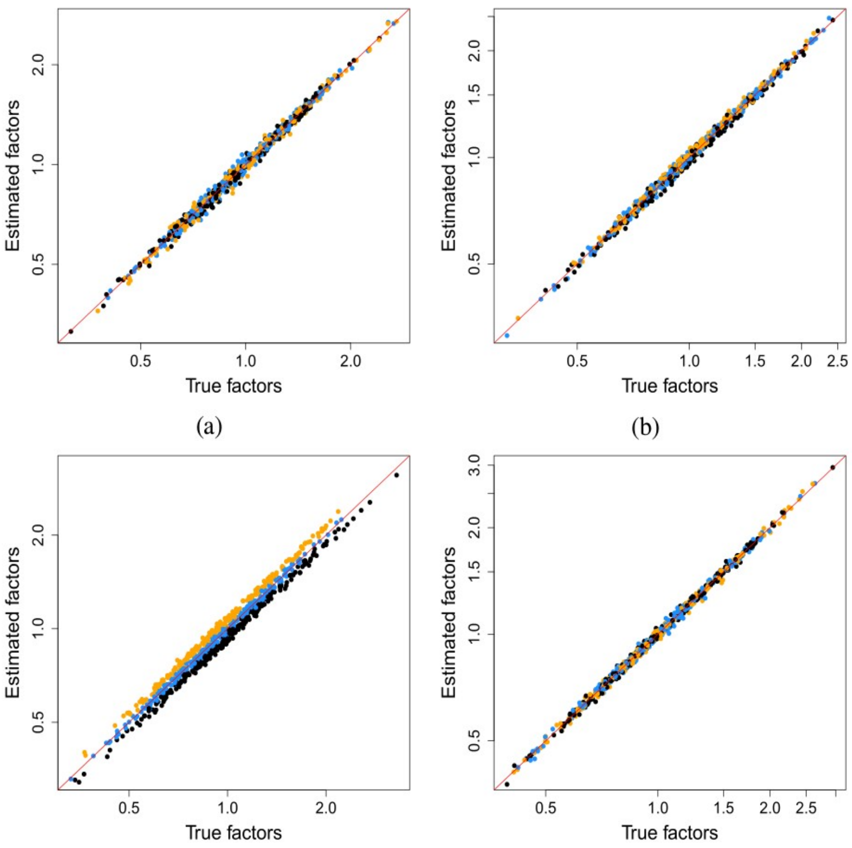
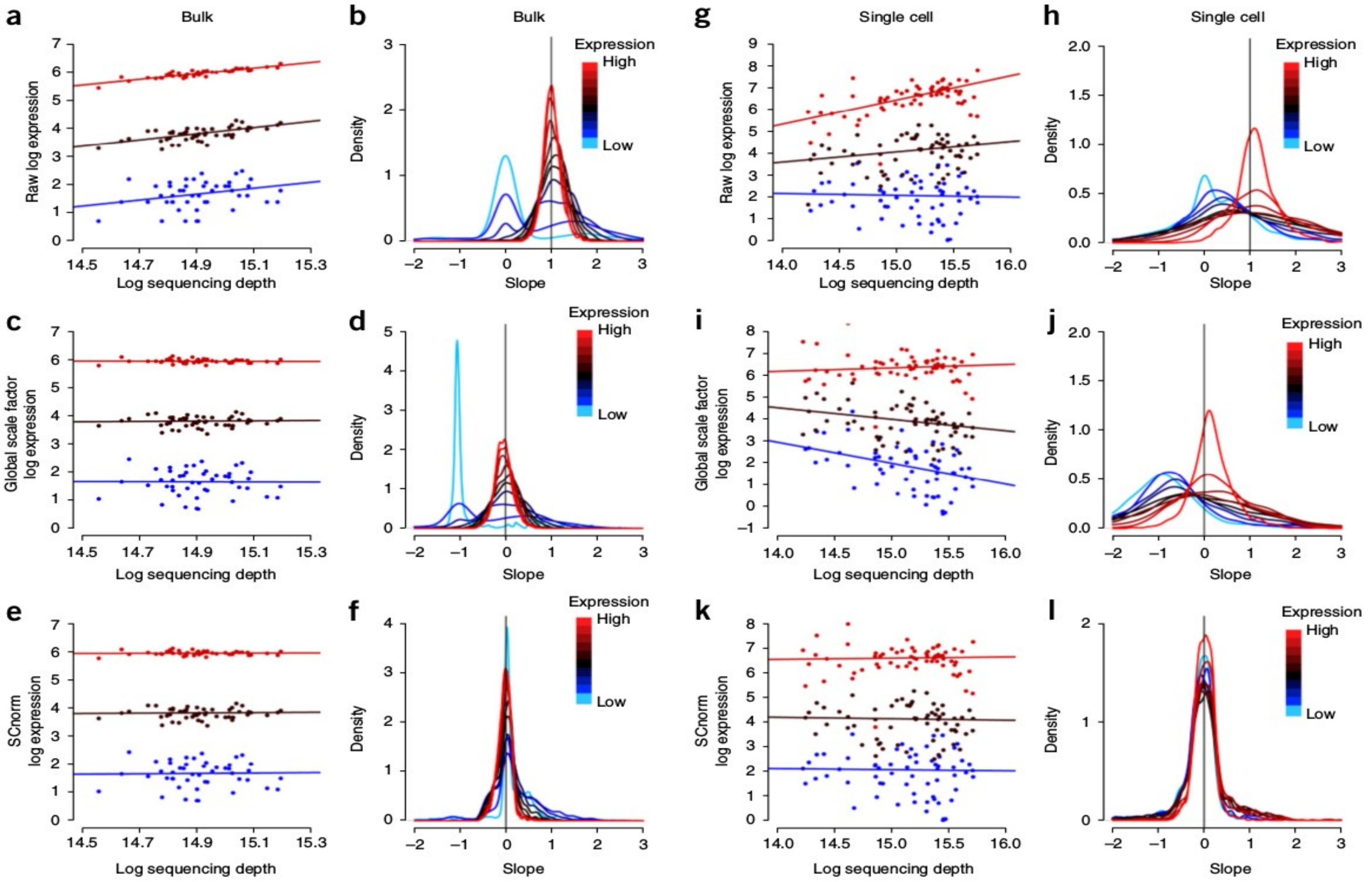


Fig. 1 Performance of existing normalization methods on the simulated data with DE genes and stochastic zeroes. The size factor estimates for all cells are plotted against the true values for **a** DESeq, **b** TMM, and **c** library size normalization. Simulations were performed with no DE (*first row*), moderate DE (*second row*), strong DE (*third row*), and varying magnitudes of DE (*fourth row*). Axes are shown on a log-scale. For comparison, each set of size factors was scaled such that the grand mean across cells was the same as that for the true values. The *red line* represents equality between the rescaled estimates and true factors. Cells in the first, second, and third subpopulations are shown in *black*, *blue*, and *orange*, respectively. *DE* differentially expressed, *TMM* trimmed mean of *M* values

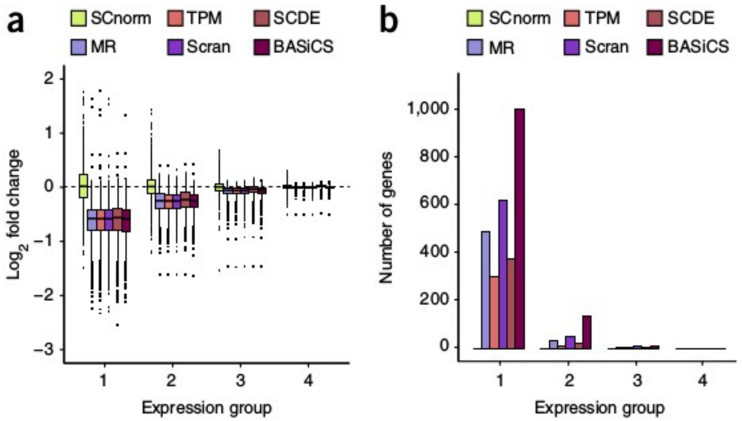
Deconvolution



Scnorm: Expression vs. Depth Bias Correction



Individual size factor per cell per group of genes

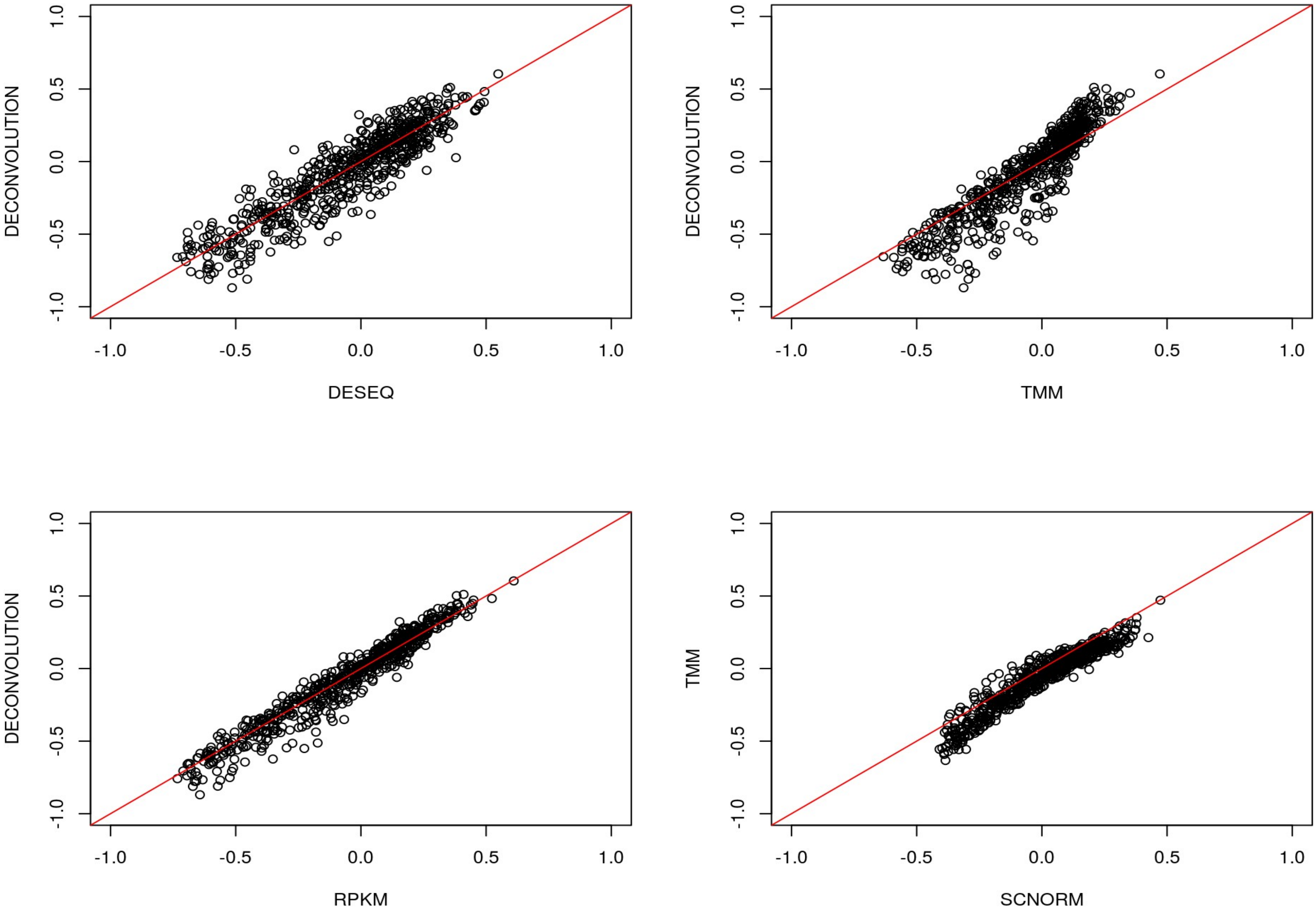


Identical cells in two groups should result in no DE and FC = 1 if normalization was efficient

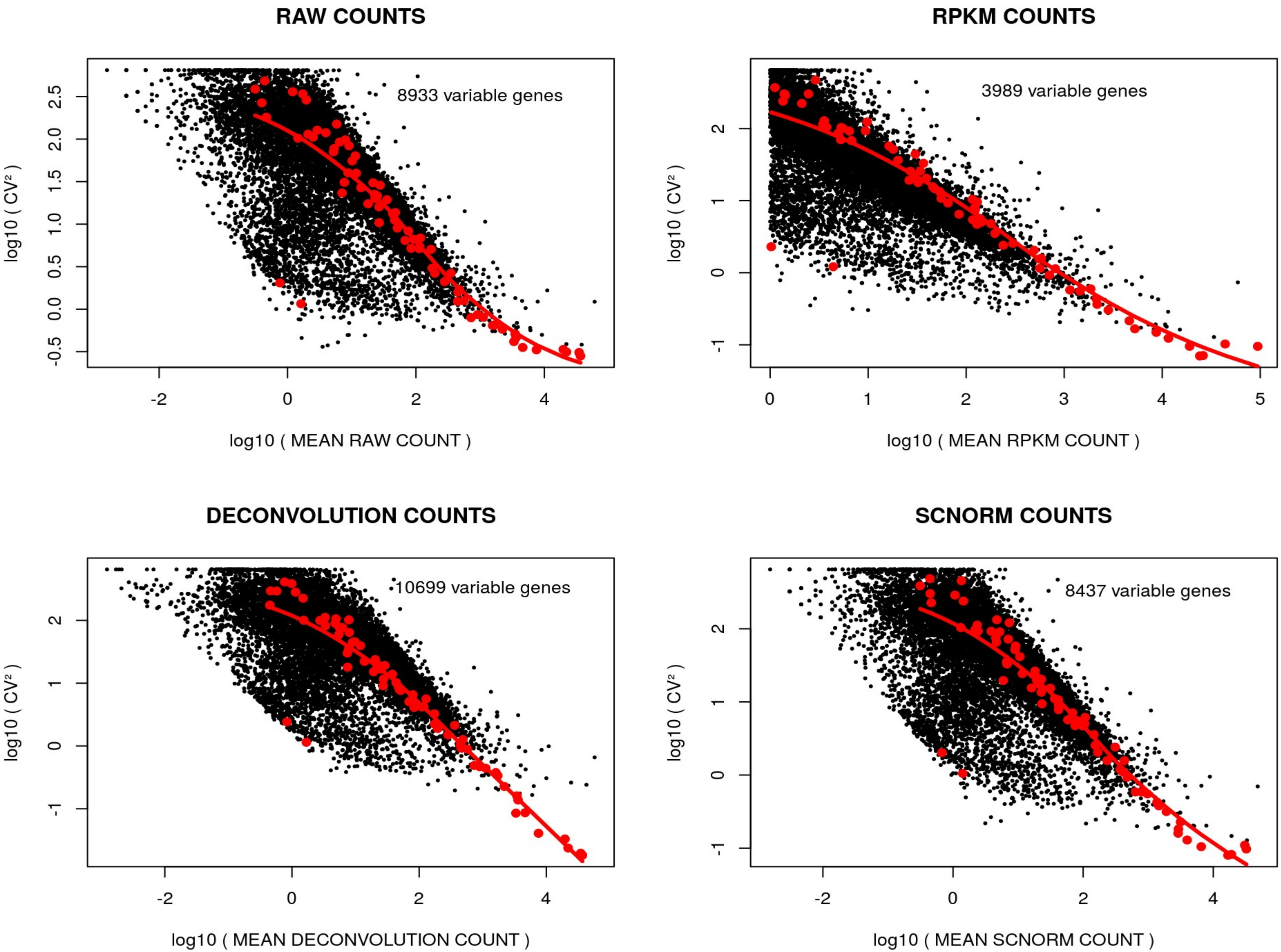
How does deconvolution normalization method compare with RPKM and normalizations by using spike-ins?

Deconvolution vs TMM vs DESeq vs RPKM vs SCnorm: Size Factors

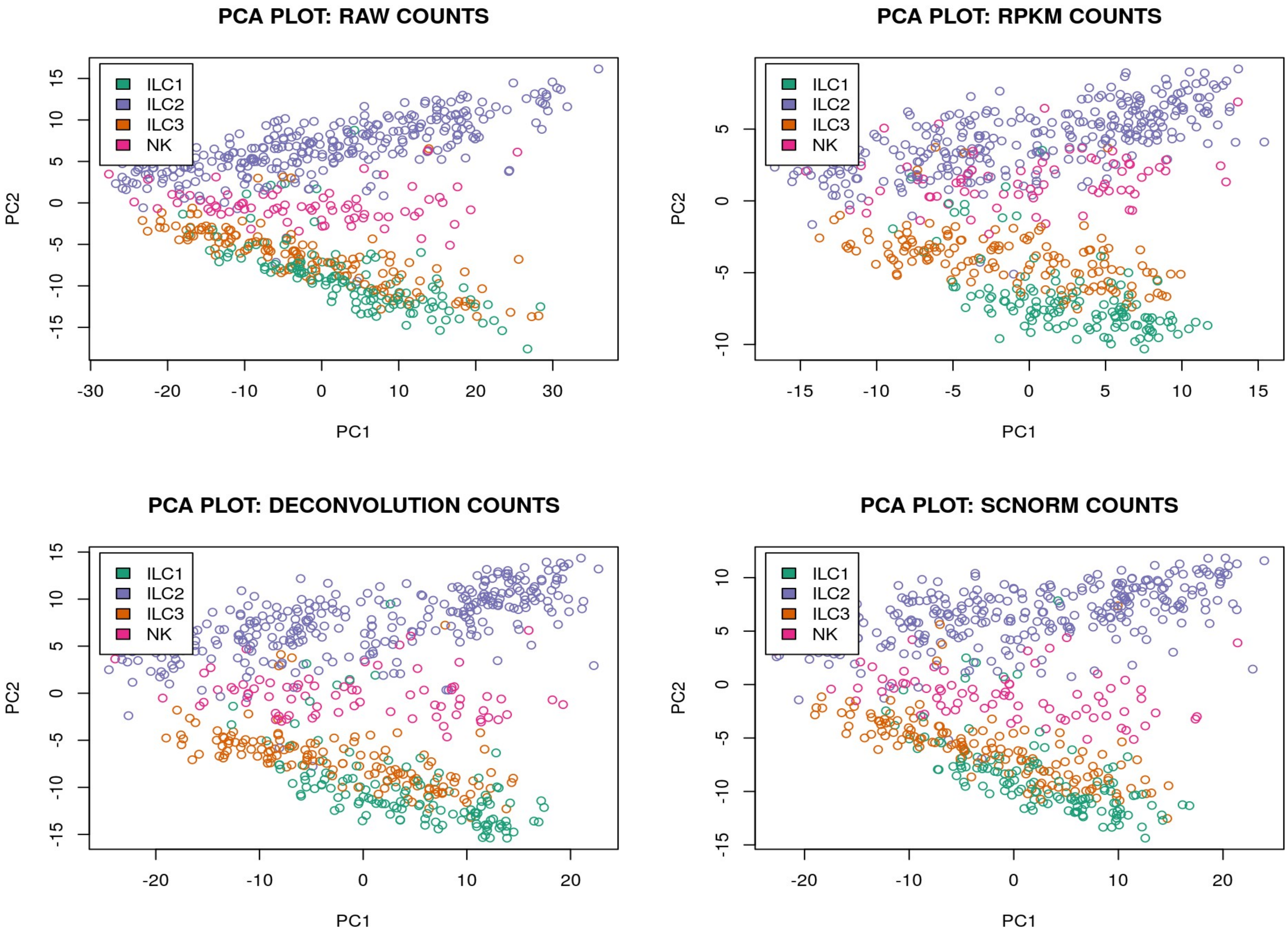
For other data sets it might not look as good as for ILC!



CV^2 vs. Mean Expression Plot



PCA Plot



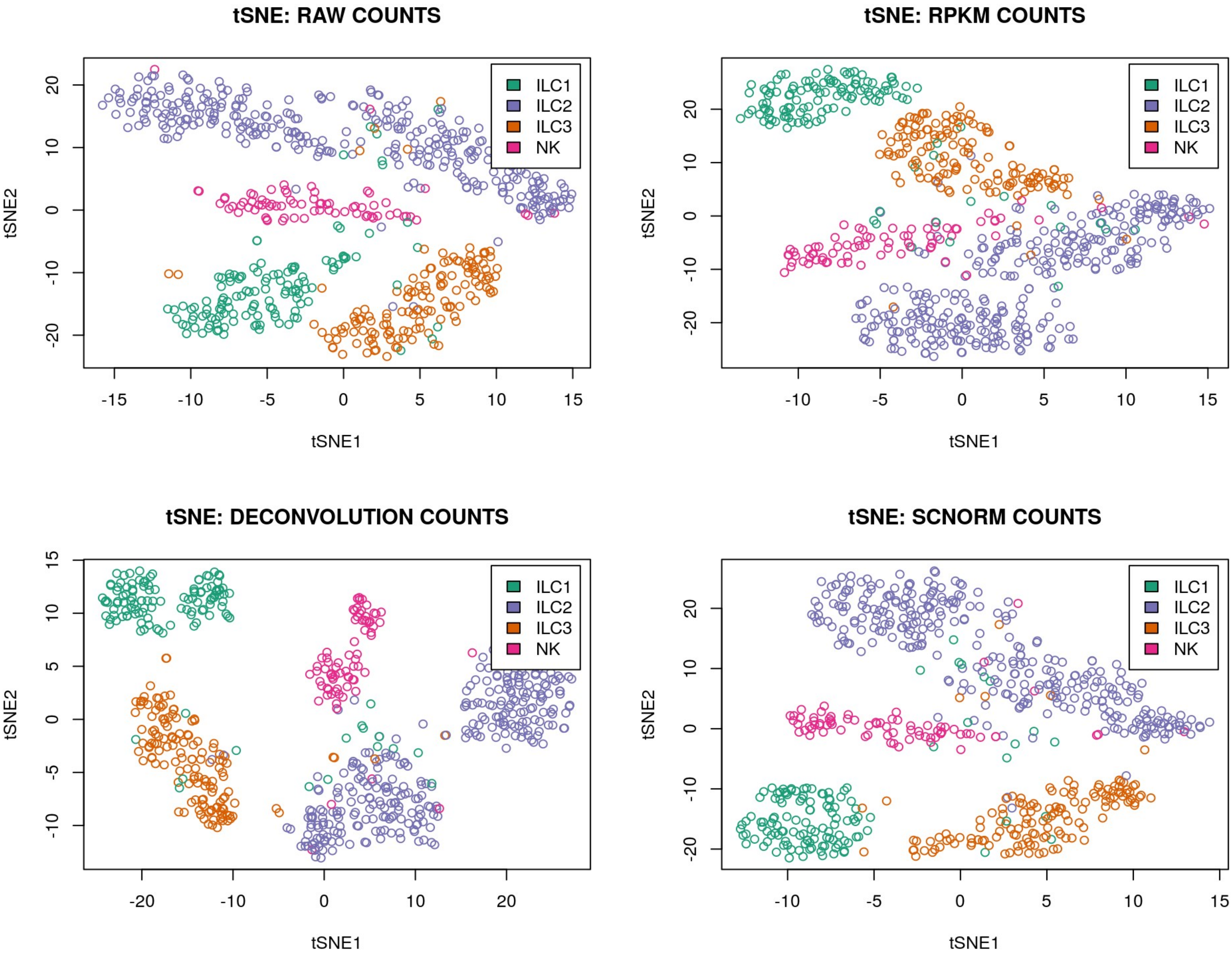
PCA PLOT: RAW COUNTS

PCA PLOT: RPKM COUNTS

PCA PLOT: DECONVOLUTION COUNTS

PCA PLOT: SCNORM COUNTS

tSNE Plot



tSNE: RAW COUNTS

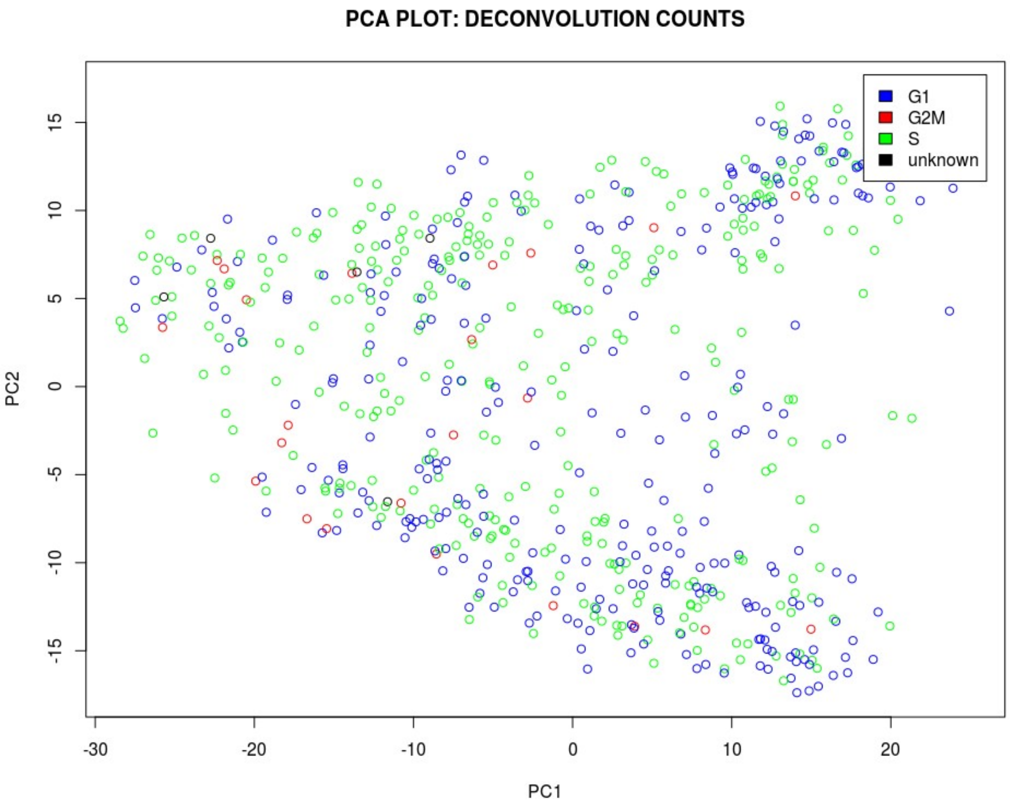
tSNE: RPKM COUNTS

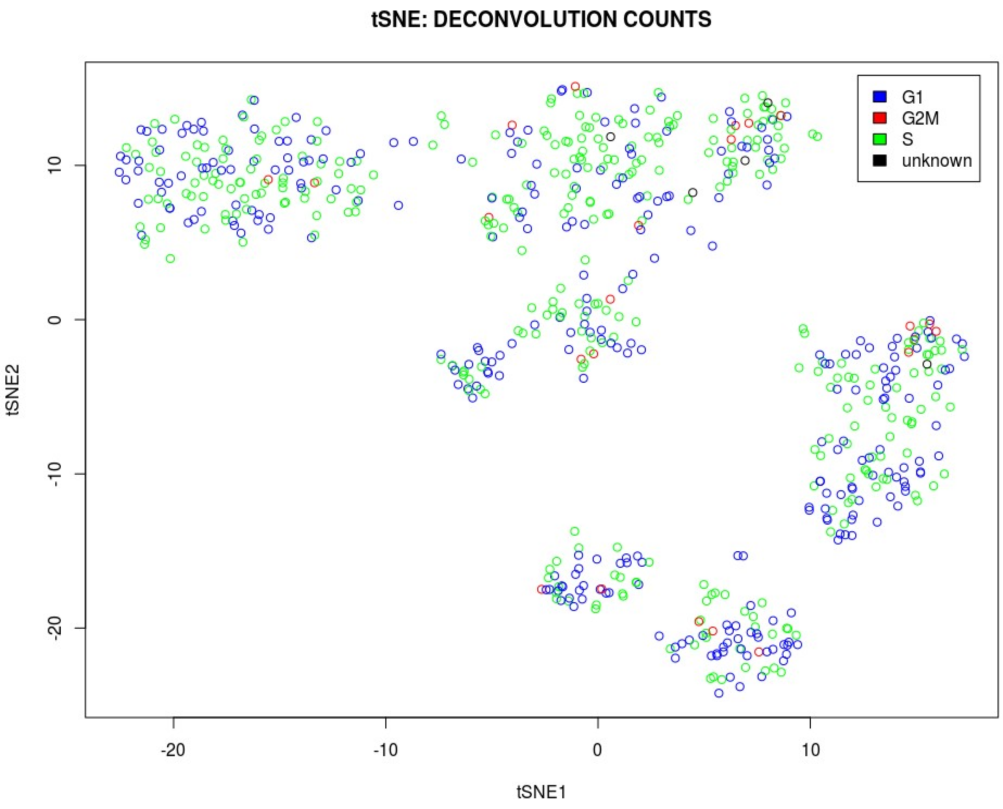
tSNE: DECONVOLUTION COUNTS

tSNE: SCNORM COUNTS

Cell Cycle Phase Assignment

Pre-trained classifier looks at pairs of genes having difference in expression that changes sign from phase to phase of cell cycle

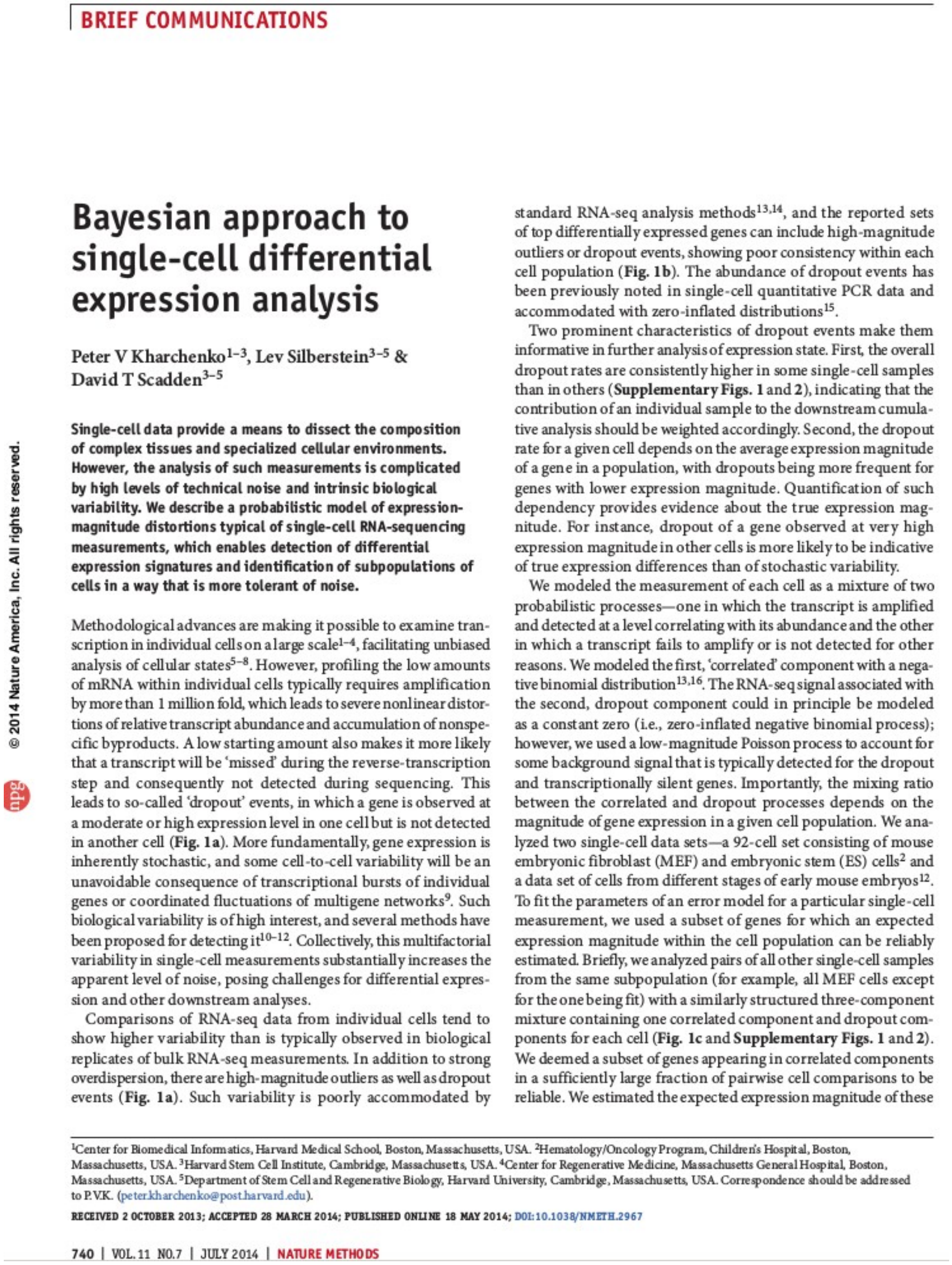




Methods for Testing for Differential Expression without Normalization:

SCDE

Single-Cell Differential Expression (SCDE)



Bayesian approach to single-cell differential expression analysis

Peter V Kharchenko[1–3], Lev Silberstein[3–5] & David T Scadden[3–5]

Single-cell data provide a means to dissect the composition of complex tissues and specialized cellular environments. However, the analysis of such measurements is complicated by high levels of technical noise and intrinsic biological variability. We describe a probabilistic model of expression-magnitude distortions typical of single-cell RNA-sequencing measurements, which enables detection of differential expression signatures and identification of subpopulations of cells in a way that is more tolerant of noise.

Methodological advances are making it possible to examine transcription in individual cells on a large scale[1–4], facilitating unbiased analysis of cellular states[5–8]. However, profiling the low amounts of mRNA within individual cells typically requires amplification by more than 1 million fold, which leads to severe nonlinear distortions of relative transcript abundance and accumulation of nonspecific byproducts. A low starting amount also makes it more likely that a transcript will be 'missed' during the reverse-transcription step and consequently not detected during sequencing. This leads to so-called 'dropout' events, in which a gene is observed at a moderate or high expression level in one cell but is not detected in another cell (**Fig. 1a**). More fundamentally, gene expression is inherently stochastic, and some cell-to-cell variability will be an unavoidable consequence of transcriptional bursts of individual genes or coordinated fluctuations of multigene networks[9]. Such biological variability is of high interest, and several methods have been proposed for detecting it[10–12]. Collectively, this multifactorial variability in single-cell measurements substantially increases the apparent level of noise, posing challenges for differential expression and other downstream analyses.

Comparisons of RNA-seq data from individual cells tend to show higher variability than is typically observed in biological replicates of bulk RNA-seq measurements. In addition to strong overdispersion, there are high-magnitude outliers as well as dropout events (**Fig. 1a**). Such variability is poorly accommodated by standard RNA-seq analysis methods[13,14], and the reported sets of top differentially expressed genes can include high-magnitude outliers or dropout events, showing poor consistency within each cell population (**Fig. 1b**). The abundance of dropout events has been previously noted in single-cell quantitative PCR data and accommodated with zero-inflated distributions[15].

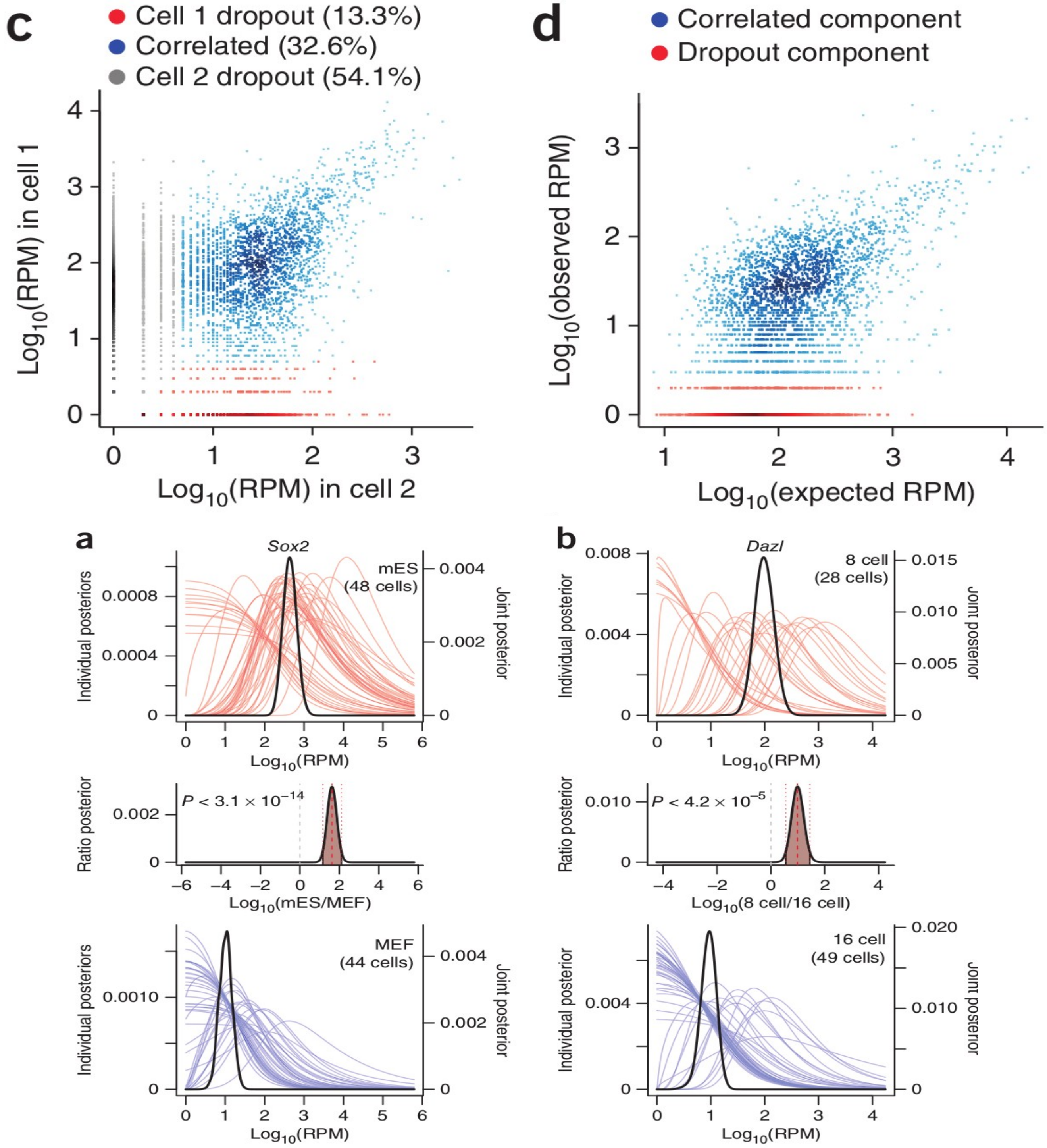
Two prominent characteristics of dropout events make them informative in further analysis of expression state. First, the overall dropout rates are consistently higher in some single-cell samples than in others (**Supplementary Figs. 1** and **2**), indicating that the contribution of an individual sample to the downstream cumulative analysis should be weighted accordingly. Second, the dropout rate for a given cell depends on the average expression magnitude of a gene in a population, with dropouts being more frequent for genes with lower expression magnitude. Quantification of such dependency provides evidence about the true expression magnitude. For instance, dropout of a gene observed at very high expression magnitude in other cells is more likely to be indicative of true expression differences than of stochastic variability.

We modeled the measurement of each cell as a mixture of two probabilistic processes—one in which the transcript is amplified and detected at a level correlating with its abundance and the other in which a transcript fails to amplify or is not detected for other reasons. We modeled the first, 'correlated' component with a negative binomial distribution[13,16]. The RNA-seq signal associated with the second, dropout component could in principle be modeled as a constant zero (i.e., zero-inflated negative binomial process); however, we used a low-magnitude Poisson process to account for some background signal that is typically detected for the dropout and transcriptionally silent genes. Importantly, the mixing ratio between the correlated and dropout processes depends on the magnitude of gene expression in a given cell population. We analyzed two single-cell data sets—a 92-cell set consisting of mouse embryonic fibroblast (MEF) and embryonic stem (ES) cells[2] and a data set of cells from different stages of early mouse embryos[12]. To fit the parameters of an error model for a particular single-cell measurement, we used a subset of genes for which an expected expression magnitude within the cell population can be reliably estimated. Briefly, we analyzed pairs of all other single-cell samples from the same subpopulation (for example, all MEF cells except for the one being fit) with a similarly structured three-component mixture containing one correlated component and dropout components for each cell (**Fig. 1c** and **Supplementary Figs. 1** and **2**). We deemed a subset of genes appearing in correlated components in a sufficiently large fraction of pairwise cell comparisons to be reliable. We estimated the expected expression magnitude of these

[1]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. [2]Hematology/Oncology Program, Children's Hospital, Boston, Massachusetts, USA. [3]Harvard Stem Cell Institute, Cambridge, Massachusetts, USA. [4]Center for Regenerative Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA. [5]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. Correspondence should be addressed to P.V.K. (peter.kharchenko@post.harvard.edu).

RECEIVED 2 OCTOBER 2013; ACCEPTED 28 MARCH 2014; PUBLISHED ONLINE 18 MAY 2014; DOI:10.1038/NMETH.2967

Single-Cell Differential Expression (SCDE) Method


c
● Cell 1 dropout (13.3%)
● Correlated (32.6%)
● Cell 2 dropout (54.1%)
Log10(RPM) in cell 1
Log10(RPM) in cell 2
d
● Correlated component
● Dropout component
Log10(observed RPM)
Log10(expected RPM)
a
Sox2
mES (48 cells)
Individual posteriors
Joint posterior
Log10(RPM)
Ratio posterior
P < 3.1 × 10−14
Log10(mES/MEF)
MEF (44 cells)
b
Dazl
8 cell (28 cells)
P < 4.2 × 10−5
Log10(8 cell/16 cell)
16 cell (49 cells)