

# Sources of Variation in Cell-Type RNA-Seq Profiles

**Johan Gustafsson** (✉ [gustajo@chalmers.se](mailto:gustajo@chalmers.se))

Chalmers tekniska hogskola <https://orcid.org/0000-0001-5072-2659>

**Felix Held**

Mathematical Sciences, University of Gothenburg and Chalmers University of Technology

**Jonathan Robinson**

Chalmers University of Technology, Department of Biology and Biological Engineering

**Elias Björnson**

University of Gothenburg, Department of Molecular and Clinical Medicine/Wallenberg Laboratory for Cardiovascular and Metabolic Research

**Rebecka Jörnsten**

Mathematical Sciences, University of Gothenburg and Chalmers University of Technology

**Jens Nielsen**

Chalmers University of Technology, Department of Biology and Biological Engineering

---

## Research article

**Keywords:** Sources of Variation, Cell-Type RNA-Seq Profiles

**Posted Date:** December 18th, 2020

**DOI:** <https://doi.org/10.21203/rs.2.23415/v2>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published on September 21st, 2020. See the published version at <https://doi.org/10.1371/journal.pone.0239495>.

## Sources of Variation in Cell-Type RNA-Seq Profiles

Johan Gustafsson<sup>1,2,\*</sup>, Felix Held<sup>3</sup>, Jonathan Robinson<sup>1,2</sup>, Elias Björnson<sup>1,4</sup>, Rebecka Jörnsten<sup>3</sup> and Jens Nielsen<sup>1,2,5</sup>

<sup>1</sup> Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, Gothenburg, Sweden.

<sup>2</sup> Wallenberg Center for Protein Research, Chalmers University of Technology, Kemivägen 10, Gothenburg, Sweden.

<sup>3</sup> Mathematical Sciences, University of Gothenburg and Chalmers University of Technology, Chalmers tvärgata 3, Gothenburg, Sweden

<sup>4</sup> Department of Molecular and Clinical Medicine/Wallenberg Laboratory for Cardiovascular and Metabolic Research, University of Gothenburg, Blå Stråket 5B, Gothenburg, Sweden

<sup>5</sup> BioInnovation Institute, Ole Maaløes Vej 3, DK2200 Copenhagen N, Denmark

\* Corresponding author

## Abstract

Cell-type specific gene expression profiles are needed for many computational methods operating on bulk RNA-Seq samples, such as deconvolution of cell-type fractions and digital cytometry. However, the gene expression profile of a cell type can vary substantially due to both technical factors and biological differences in cell state and surroundings, reducing the efficacy of such methods. Here, we investigated which factors contribute most to this variation. We evaluated different normalization methods, quantified the variance explained by different factors, evaluated the effect on deconvolution of cell type fractions, and examined the differences between UMI-based single-cell RNA-Seq and bulk RNA-Seq. We investigated a collection of publicly available bulk and single-cell RNA-Seq datasets containing B and T cells, and found that the technical variation across laboratories is substantial, even for genes specifically selected for deconvolution, and has a confounding effect on deconvolution. Tissue of origin is also a substantial factor, highlighting the challenge of applying cell type profiles derived from blood on mixtures from other tissues. We also show that much of the differences between UMI-based single-cell and bulk RNA-Seq methods can be explained by the number of read duplicates per mRNA molecule in the single-cell sample. Our work shows the importance of either matching or correcting for technical factors when creating cell-type specific gene expression profiles that are to be used together with bulk samples.

## Introduction

RNA Sequencing is a well-established method for comparing the transcriptome between different cell types, conditions and cell states(1). Cell types can be separated from samples using fluorescence-activated cell sorting (FACS)(2) before sequencing, and recent advances have made it possible to use RNA-Seq at the single-cell level and to sequence hundreds of thousands of cells(3). The ever-growing collection of publicly available data enables integrative data analysis across many datasets, making it possible to discover system-wide phenomena. Such analyses are however made difficult by systematic batch effects across laboratories and technologies, posing a large challenge for data analysis.

Single-cell RNA-Seq facilitates the study of distinct cell types. However, the number of patients involved in such experiments is usually small compared to datasets containing bulk data from biopsies, such as The Cancer Genome Atlas (TCGA). It is therefore desirable to be able to conduct studies on bulk data with mixed cell types, with the help of mathematical tools that can help extract similar information as is available in single-cell data. One example of such a tool is cell type deconvolution, which estimates the fractions of different cell types in a mixed sample from the RNA-Seq data. This is implemented in for example CIBERSORTx(4), EPIC(5) and CPM(6). Most implementations require gene expression profiles (GEPs) for each of the cell types into which the mixed sample is to be deconvolved. Some methods work with single-cell data(6), others with a representative expression profile(5), but in general they need a representative expression of the cell types in the sample. Other tools that also need gene expression profiles for cell types include an extension of deconvolution often referred to as digital cytometry, which is implemented in CIBERSORTx(4), and automatic cell type annotation of cell types in single-cell data(7,8).

Representative RNA-Seq gene expression profiles for cell types can be created from either FACS-sorted bulk samples or single cell datasets where the average expression of cell populations can be used. However, the RNA-seq profiles for a cell type can vary substantially, both due to biological differences between samples and technical biases. It is therefore challenging to construct universal gene expression profiles for cell types that work well in all conditions. The variation in RNA-Seq originating from sequencing at different laboratories has previously been examined(9,10). In these studies, each laboratory has been given detailed sample preparation and sequencing instructions, enabling estimation of technical variation introduced from distinct technical factors. These studies, however, do not describe the typical case encountered for methods such as deconvolution of cell type fractions. In such cases, the cell type profiles are typically generated once and reused for many studies, often using different laboratory procedures, or even generated from single-cell data(4). In addition, the cell type profiles and samples may be generated from different tissues. An example of such a case is using immune cell profiles from blood for estimation of cell type fractions in bulk tumor data from TCGA, which is made available by xCell(11).

There are a number of methods for measuring the impact of different factors, such as cell type, tissue of origin, or dataset on gene expression. ANOVA is commonly used to estimate such differences, for example. Another example is variancePartition(9), a method developed by Hoffman et al. that is based on mixed effects modeling and specifically adapted for RNA-Seq data. Hoffman et al. showed that variancePartition outperforms ANOVA for such analyses, supporting it as a strong candidate for the analysis of cell type profiles in this study.

Normalization is a crucial step when analyzing RNA-Seq data. In the beginning of the RNA-Seq era, library size normalization, for example FPKM(12) and TPM(13), was commonly employed. It was later discovered that the library size is often strongly affected by a few highly expressed genes with

stochastic behavior, which is remedied by methods such as the trimmed mean of M values (TMM)(14) and the normalization performed by DESeq2(15). These methods are designed to operate directly on gene counts and work under the assumption that most genes are not differentially expressed across samples. The restriction to work on gene counts makes it more difficult to normalize single-cell data based on unique molecular identifiers (UMIs) together with bulk RNA-Seq. Bulk counts need to be normalized by transcript length, such as in FPKM and TPM normalization, to yield a representative gene expression, while UMI-based single-cell data does not require length normalization. Consequently, the methods working on counts cannot directly account for transcript length, while library size normalization, which allows for such length correction, is sensitive to highly expressed noisy genes.

A common technique used to overcome technical biases is to computationally remove batch effects. The batch effect removal tool ComBat(16) implements a strategy where differences across batches in mean and dispersion of each gene are removed, regardless if the source of the variation is technical or biological. However, ComBat and similar tools require overlapping samples with similar biological traits across the datasets. This overlap does not exist between expression profiles of distinct cell types and biopsies containing a mix of cell types. CIBERSORTx employs a batch correction strategy using ComBat where mixed samples are created *in silico* by mixing the cell type expression profiles at different fractions. The batch correction applied on the synthetic mixed samples is then projected back to the cell types. A drawback of this approach is that it introduces a bias depending on the cell fractions selected in the synthetic mixture. Acquiring representative cell type profiles for a given dataset of mixed samples remains a challenge. In this study, we sought to quantify the relative importance of the most important factors that cause undesired variation between gene expression profiles of individual cell types for typical cases encountered in deconvolution and other methods that use cell type profiles. We assembled a collection of public datasets and evaluated normalization and batch correction methods, quantified different sources of variation between samples for different sets of genes, and investigated technical biases between single-cell and bulk RNA-Seq. In addition, we evaluated the impact of using different cell type profiles in deconvolution.

## Results

### Data Preparation

To limit the complexity of the study, only B and T cell profiles were used. We gathered 74 publicly available RNA-Seq bulk samples of B and T cells from 5 different sources, and pooled 31 cell populations from single-cell data from 7 public datasets, for a total of 105 samples (Table 1, Table S1). The data sources were chosen to span different laboratories and contain samples with different cell subtypes and tissues of origin. In addition, we investigated a dataset(17) containing both bulk and single-cell data from the same samples to examine biases between single-cell and bulk RNA-seq.

### Normalization and Batch Effects

First, we investigated data normalization approaches, which is challenging since our dataset contains both bulk data and pooled single-cell samples from UMI-based methods. This prevents the typical use of normalization methods that operate on counts, since the counts are not directly comparable between these sequencing technologies. In short, the bulk samples need to be corrected for transcript length, whereas the drop-seq based samples should not. We therefore decided to test three well-established methods that can operate on non-count data: library size normalization (TPM/CPM)(13), Trimmed Mean of M-values (TMM)(14), and quantile normalization(18). TMM was originally designed to work on counts with a known library size; we therefore scaled the TPM values to pseudo-counts (Methods). To avoid the stochasticity from lowly expressed genes, we only analyzed genes with a mean expression over 1 TPM across all samples.

Figure 1 shows the relative log expression across all genes and samples for three normalization methods, and highlights the inadequacy of library size normalization when comparing these samples. The drop-seq-based pooled single-cell samples (all single-cell samples except SC Melanoma) are especially problematic; a large portion of the genes are lowly expressed compared to the bulk samples. TPM normalization between bulk samples also fails to scale the samples properly, which has been shown previously(14). TMM and quantile normalization succeed in overcoming most of the normalization issues, both in terms of mean and variation of the relative log expression. The advantage of TMM over quantile normalization is that it only scales the samples, minimizing the introduction of additional technical biases, while quantile normalization replaces all expression values. On the other hand, TMM assumes that most genes are not differentially expressed. This is a reasonable assumption here, but it may be inappropriate when comparing mixed bulk samples and GEPs for cell types. For the analyses in this study, we selected TMM as the normalization method unless otherwise noted.

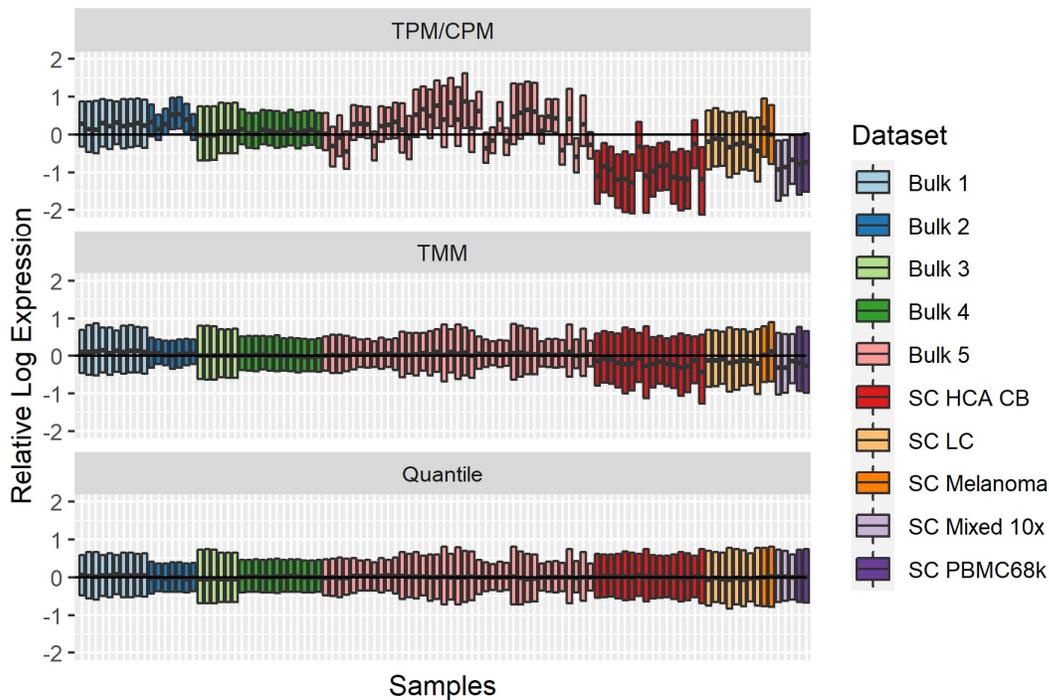


Figure 1 Evaluation of normalization methods for combined single-cell and bulk data. Each bar is a boxplot over all genes describing the  $\log_2$  fold change in gene expression between the sample and the mean over all samples. 12,072 genes were included for each sample.

We used PCA to investigate the batch effects between different labs and between single-cell and bulk data. Figure 2 shows that without any batch correction, samples group by dataset, and the first principal component mainly describes technical variation between datasets. The drop-seq-based 10x datasets tend to cluster together, and the same holds for the bulk samples. The only Smart-Seq2 single-cell data present (SC Melanoma) seem to be more similar to bulk than the other single-cell samples. The normalization method clearly matters; the first component explains less of the variation with a better normalization method, and the batch effects are more pronounced due to the reduced technical noise. We applied ComBat(16) from the sva(19) R package to remove batch effects between datasets, with the instructions to preserve cell type differences. ComBat effectively removes all systematic differences in mean and dispersion between datasets, except for those specified in the model matrix (cell type in our case), however it does not distinguish between biological and technical variation. Applying ComBat will thus remove any biological differences across datasets not specified in the model matrix, which may affect downstream analysis.

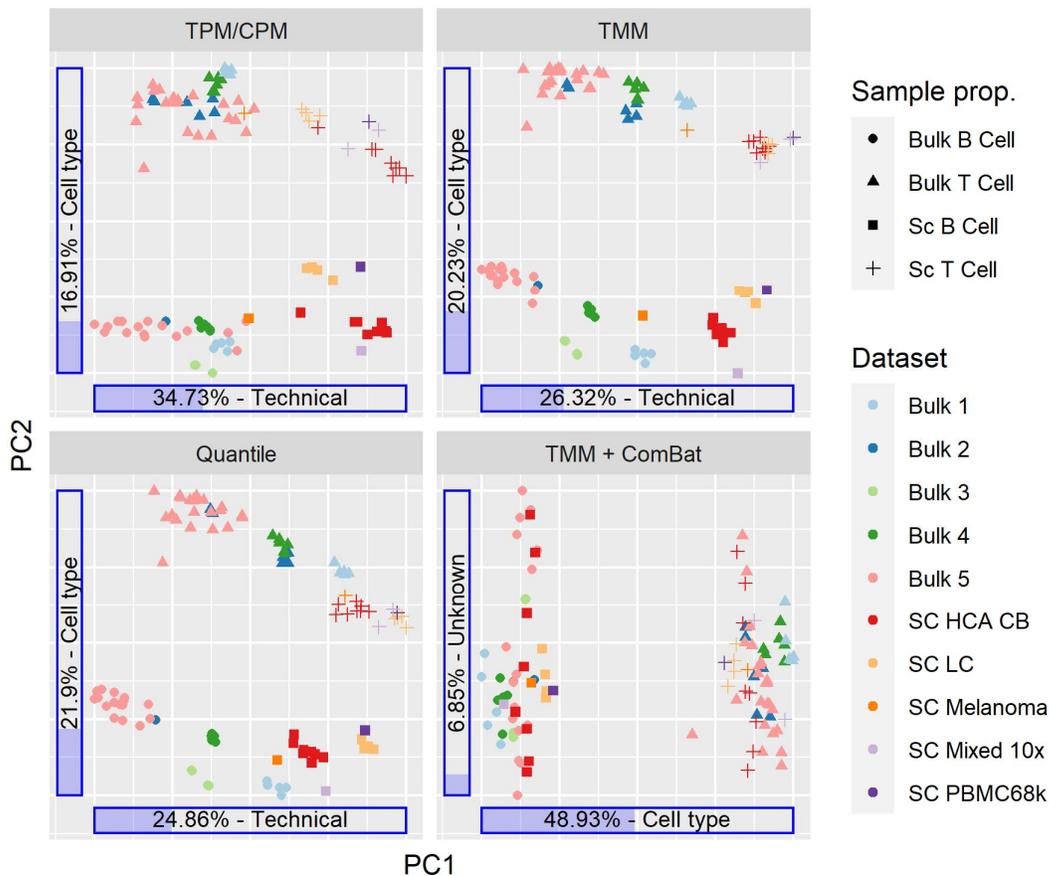


Figure 2 Visualization of normalization and batch effects using PCA. 12,072 genes were included for each sample.

### Sources of Variation Between Samples

To quantify sources of variation in gene expression across samples we investigated to what extent different factors could explain the variance in gene expression. Factors examined were laboratory (which here includes different laboratory procedures and sequencing methods), UMI-based single-cell vs bulk, cell type, cell subtype (distinguishing between different types of B and T cells), tissue of origin, individual, and whether samples are technical replicates. We normalized the 105 samples using TMM as described above and used the R package variancePartition(9) to estimate the importance of the factors to be investigated for each gene.

Figure 3 shows the variation induced by the different factors across bulk samples, investigating the factors laboratory, cell type, cell subtype, and tissue of origin. Figure 3A shows the explained variance for factors when including all genes. Figures 3B-D show how the explained variance changes for different gene sets: housekeeping genes, genes used in deconvolution of immune cells (the genes selected for deconvolution when using the LM22 cell type profiles, developed by the CIBERSORT(20) team), and a subset of LM22 genes that exhibit substantial expression differences between B and T cells (absolute  $\log_2$  fold change  $> 1$ , henceforth named LM22S). When including all genes, lab is the dominating factor, highlighting the importance of working on a reduced gene set for methods such as deconvolution. For housekeeping genes, defined as genes having a stable expression across tissues and cell types(21), lab explains more of the variance compared to the case where all genes were used, which is consistent with the expectation of lower biological variation in

this gene set. As expected, cell subtype is the most important factor for explaining variance among the LM22 genes, and even more so for the genes in the LM22S subset. Interestingly, cell subtype still only explains roughly half of the variation in the LM22S gene set, posing a challenge for deconvolution and similar methods. Figure 3E shows the explained variances for the LM22S genes if cell type is used as a factor instead of cell subtype. The lower explained variance indicates that many of the LM22S genes differ in expression between cell subtypes. This result is consistent with the formulation of the LM22 gene set (which LM22S is derived from), which was optimized for separating immune cell subtypes including those of T and B cells. To verify that the results were not dominated by outlier samples, we repeated the analysis for bulk samples excluding one sample each iteration. The maximum change for any factor was less than 2 percentage points, suggesting a negligible outlier effect.

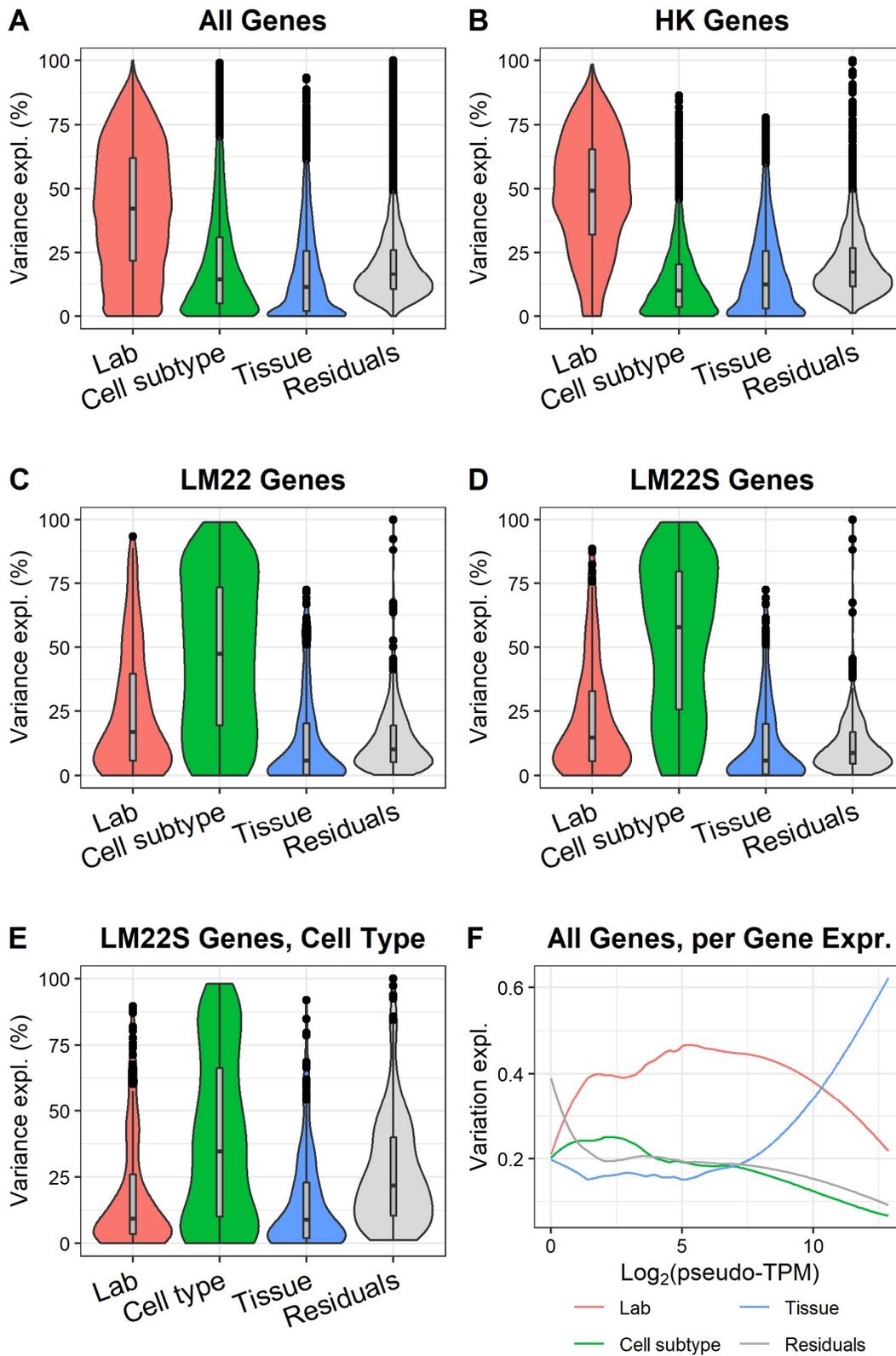


Figure 3 Explained variance in gene expression for bulk RNA-seq samples. A. All genes (12072 genes). B. Housekeeping genes (3393 genes). C. LM22 (395 genes). D. LM22S genes. E. Identical to D, with the difference that cell subtype is replaced

*with cell type (B/T). F. Explained variance per gene expression. The plot shows how the explained variance by the different factors change with gene expression (Loess fit, span = 0.3)*

To investigate the influence of gene expression on explained variance, we estimated explained variance as a function of gene expression for all genes (Figure 3F). As expected, the residuals explain more of the variance for lowly expressed genes, since sampling noise lead to higher variation for such genes. Although lab remains the dominant factor for much of the gene expression range, tissue becomes more important for highly expressed genes.

We also investigated the variance reduction from using samples from the same individual (taken at different time points) and technical replicates (the same sample sequenced multiple times) (Fig. S1E-F, Supplementary Information). Our results indicate that technical replicates, although here represented by only a few samples, exhibit lower variance in gene expression than different biological samples. Samples from the same individual taken at different time points however show similar variance compared to samples taken from different individuals, which is contradictory to previous results indicating the importance of individuals for explaining variance(9). In that study, however, the samples from the same individual were technical replicates, which show a lower variation in our measurements as well. Given the small number of samples, our results here should be interpreted with caution, and further studies are needed to fully characterize this source of variation.

To investigate the factors explaining the variance between profiles from pooled single-cell data, we used variancePartition on the single-cell samples to analyze the factors lab, cell type, and tissue. The SC melanoma dataset was excluded from the analysis because it was the only dataset using the Smart-Seq2 technology, and consequently had a high impact on the variance explained by lab (Fig. S1C-D). We evaluated the effect of outliers the same way as described above for the bulk samples, and similarly found that the maximum change for any factors obtained by removing any one sample was less than 2 percentage points. Figures 4A and B show the explained variance when including all genes or LM22S genes, respectively. The fraction of unexplained variance (residuals) is larger for single-cell than bulk, indicating that there is much variation that we fail to model, and it is therefore likely more challenging to work with single-cell data for deconvolution and other computational methods that require cell type profiles. Potential sources of the increased unexplained variation include higher technical noise in single-cell data and the presence of misclassified cells. Figure 4C-D shows the explained variance in LM22S across all samples, both bulk and single-cell (still excluding the Smart-Seq2 samples). Specifically, the figures highlight the difference in explained variance between lab (where each lab is modeled separately) or by just separating single-cell and bulk samples (SC/B). Although there seem to be systematic differences in gene expression between single-cell and bulk, a large part of the variance is still explained by lab-specific factors.

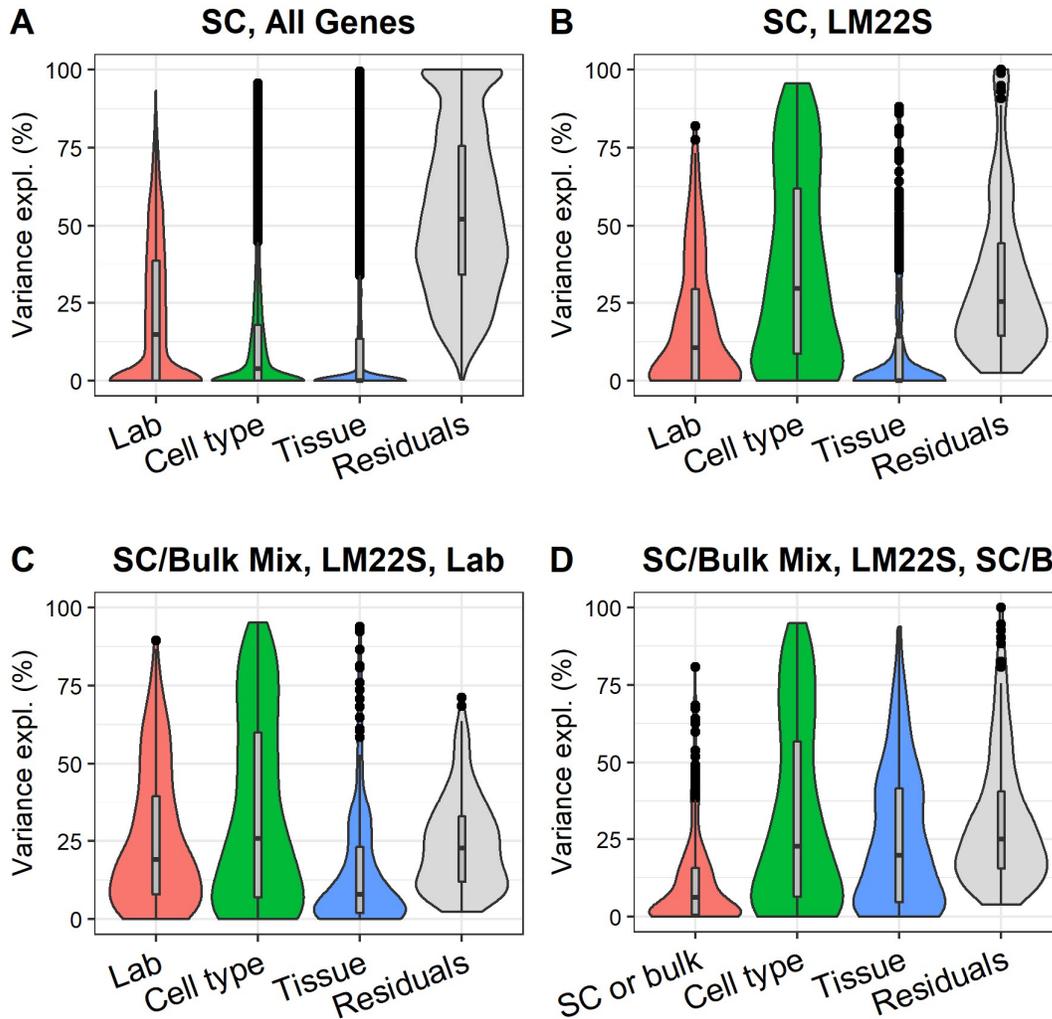


Figure 4 Explained variance for single-cell samples. A. Only single-cell samples, all genes. B. Only single-cell samples, LM22S. C. Mix of single-cell and bulk samples, LM22S. D. Identical to C, with the exception that lab is replaced by a flag indicating if the samples are single-cell or bulk.

### Differences Between Single-Cell and Bulk

To unravel the sources of variation between pooled single-cell data and bulk, we investigated the EVAL dataset which contains bulk and single-cell (10x) data generated from the same samples(17). We used the cortex 1 and cortex 2 samples, originating from mouse brain, to determine if it was possible to identify technology-driven differences between 10x data and bulk. We first pooled the single cells for each cortex and normalized them together with the bulk data using TMM (FPKM data and UMI counts, as described earlier). A small set (232) of outlier genes were filtered (Methods).

We defined difference in gene expression per gene as the log2 fold change between the 10x pool and bulk data (TMM normalized as described above) and investigated to what extent that difference could be explained by different technical covariates across genes. First, we calculated the number of discarded UMI duplicates per UMI for each gene. We defined the UMI copy fraction as

$$UMICF = \frac{\text{total counts} - \text{UM counts}}{\text{total counts}}$$

for each gene, representing the fraction of the counts that are filtered as UMI duplicates. Figure 5A shows a clear negative correlation between the difference in gene expression and UMICF ( $p < 2.2 * 10^{-16}$ , F Test). There could be several explanations for this effect. The covariate could represent differences in PCR amplification between genes. We reason that although such biases are mostly removed in the 10x data due to discarding of UMI duplicates, they are present in the bulk data, and the PCR amplification could be similar for the same gene in both cases. An alternative explanation could be that copies of the same molecules are assigned to different genes with similar sequence, such that they are counted as different molecules in the single-cell data. Such an effect would increase the total count for genes sharing copies of the same original molecule. A third possible reason for the negative slope could be that for some genes more reads are discarded due to alignment failure or quality filtering. Since there are often several copies per UMI, such an effect would be limited in the single-cell data but would have a large impact for bulk gene quantification.

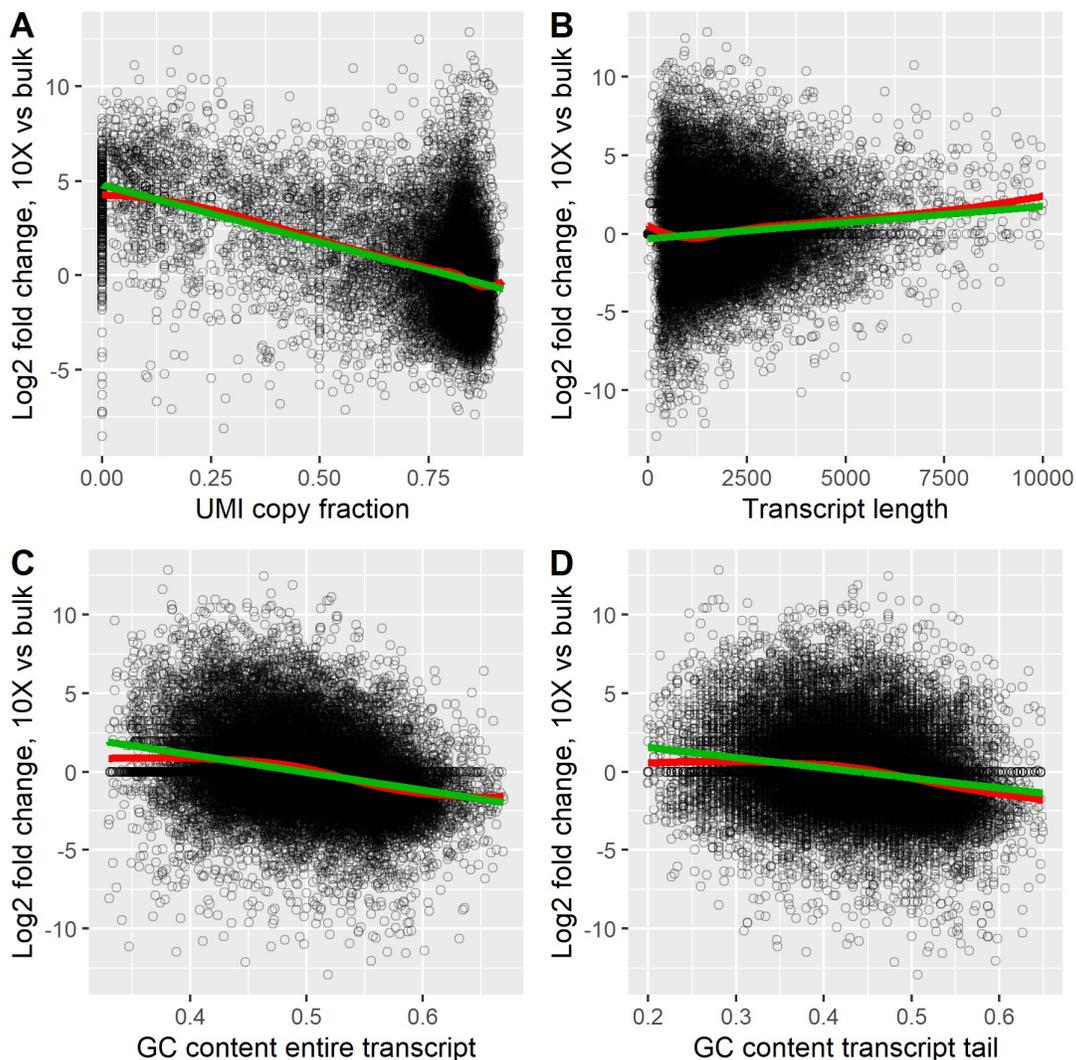


Figure 5. Log2 fold change between 10x data and bulk for each gene, plotted as a function of different covariates. The green line represents a linear fit, whereas the red line shows a Loess fit. The data shown is from cortex 1 of the EVAL dataset. A. UMI copy fraction. Only genes with more than 5 molecules available for calculating UMICF are shown. B. Transcript length. C. The GC content of the entire transcript. D. The GC content of the 150 bases closest to the transcript tail.

Second, we investigated if transcript length introduces a bias due to differences in the sequencing protocol between 10x and bulk. Bulk reads can come from the entire transcript whereas the 10x reads originate from the sequences close to the polyA tail and is thereby less affected by transcript length. Figure 5B indicates that this covariate introduces a bias, where longer transcripts in general seem to be over-penalized by the division of gene length in bulk ( $p < 2.2 * 10^{-16}$ , F Test).

Third, we investigated the effect of the GC content of genes since this is a known source of bias in RNA-Seq(22). We investigated two covariates, the GC content of the entire transcript and the GC content of the 150 base pairs closest to the polyA tail, to see if those better could explain the variability between 10x data and bulk. Figure 5C-D shows that a higher GC content in general gives a higher expression in bulk compared to 10x ( $p < 2.2 * 10^{-16}$  for both covariates, F Test), which could be related to PCR amplification biases.

To evaluate how much of the differences between 10x data and bulk can be explained by the technical covariates, we measured the improvement in correlation between 10x and bulk data after regressing out the covariates. To exclude the possibility that the differences originate from stochasticity, i.e. lack of reproducibility of data, we first checked the log space correlation between cortex 1 and 2. Although the samples originate from slightly different parts of the brain, the different samples had a Pearson correlation of 0.989 for bulk and 0.981 for 10x, showing low stochasticity. Figure 6 shows the correlation improvement after regressing out different combinations of covariates, using both a linear and loess fit. It is evident that regressing out the UMICF covariate increases the correlation more than the other covariates ( $p < 2.2 * 10^{-16}$ , both Cortex 1 and 2, both loess and linear), though the other covariates explain some variation on their own. When all covariates are combined, the tail GC content does not add information (the correlation is identical down to the 4<sup>th</sup> decimal) and transcript length has a neglectable effect on correlation improvement (mean correlation improvement is  $7.7 * 10^{-4}$ , although significant,  $p < 2.2 * 10^{-16}$ ); the combination of UMICF and GC content is therefore a good choice (significant improvement vs UMICF alone,  $p < 2.2 * 10^{-16}$ ). The differences between regressing out a loess or linear fit for a covariate are generally small, although loess performs slightly better.

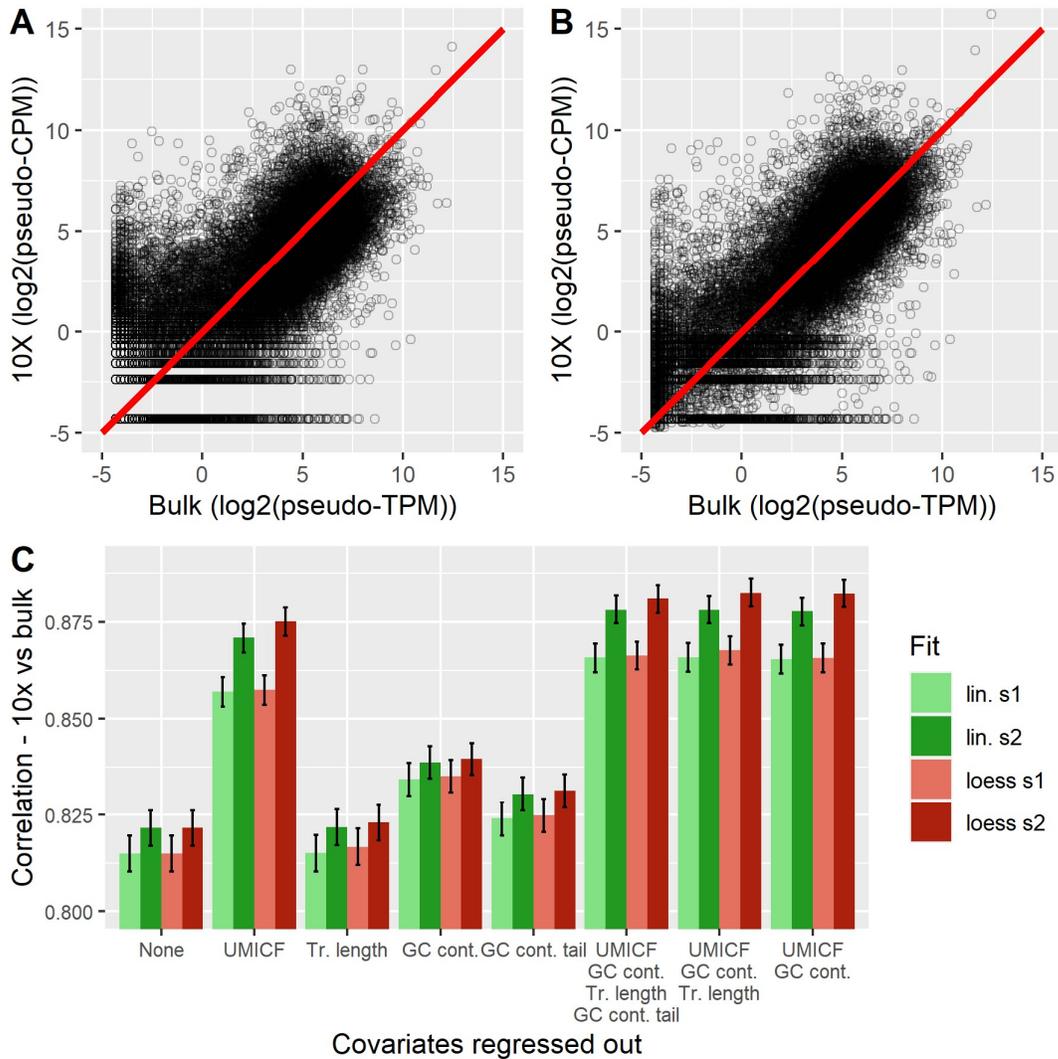


Figure 6 A. Gene expression for cortex 1 from the EVAL dataset plotted as 10x vs bulk. The red line represents a perfect correlation. B. Gene expression for cortex 1 from the EVAL dataset after regressing out the differences in UMICF and GC content between 10x and bulk using a loess fit, which improves the correlation. C. Average Pearson correlation coefficient between 10x data and bulk in log scale after regressing out technical covariates (UMI copy fraction, transcript length, GC content and GC content tail), using linear or loess regression. The correlation shown is the average of the correlations from cortex 1 and 2 of the EVAL dataset. The error bars represent the confidence interval, based on bootstrapping of genes to include in the correlation.

## Deconvolution of Cell Type Fractions

To assess the impact of technical sources of variation on deconvolution of cell type fractions, we conducted a qualitative comparison of deconvolution performance across profiles generated from different datasets. We generated in total 40 synthetic mixed bulk samples containing 50% from a B cell sample and 50% from a T cell sample, both produced at the same lab. We further generated in total 12 pairs of cell type profiles (B cells and T cells from the same lab), originating from both single-cell and bulk data. Figure 7 shows the deconvolution performance from CIBERSORTx, which varies largely across profiles. As expected, the best results are obtained when using profiles and in-silico mixtures from the same lab, with no difference in tissue of origin or cell subtypes. The general conclusion that can be drawn is that the uncertainty in deconvolution performance is large unless the profiles and mixtures are generated under similar conditions, using similar technologies and

laboratory procedures. Our results also suggest that using cell type profiles developed from single-cell data is more challenging. Furthermore, the performance varies much between sets of cell type profiles, where profiles generated from some datasets, such as HCA CB, performs much better than others (e.g. profiles generated from PBMC68k). Some of the single-cell profiles come from different tissues, which could partly explain the reduced deconvolution performance in those cases. However, the SC PBMC68k profiles originate from blood and contain a mix of cell subtypes. The lower performance for those profiles as compared to bulk profiles from similar cells (e.g. Bulk 4) can thus not be attributed to differences in tissue of origin or cell subtype only. CIBERSORTx supports two batch correction methods to remedy such issues; B-mode for cell type profiles from bulk and S-mode for single-cell profiles. We found that the S-mode batch correction seems to improve the performance considerably for single-cell profiles, while the B mode batch correction has a less pronounced effect.

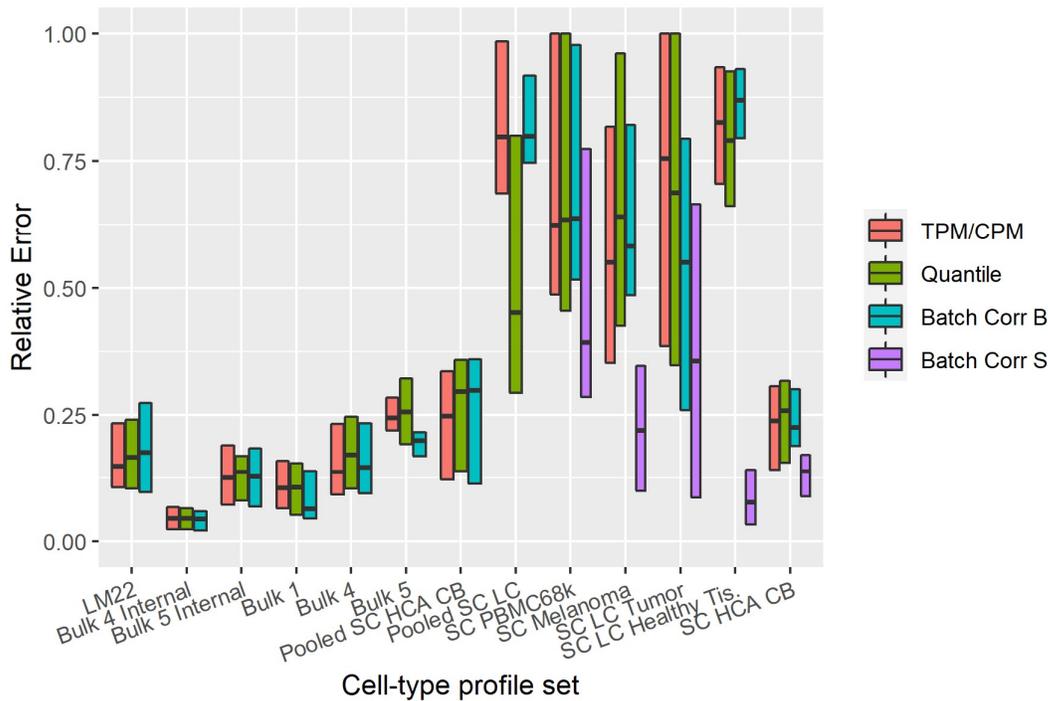


Figure 7 Deconvolution performance for different cell type profiles tested against in-silico generated bulk mixtures with 50% T cells and 50% B cells. The figure shows the relative error of the B cell fraction estimated by CIBERSORTx. The Bulk 4/5 Internal profiles were generated from the Bulk 4/5 lab and were tested against mixtures generated from other samples from the same lab. The remaining profiles were tested against 40 mixtures generated from Bulk 1, 2, 4, and 5 (Bulk 3 has no T cell samples).

## Discussion

Cell-type specific gene expression profiles are useful for analyzing bulk RNA-Seq samples containing mixed cell types, since it enables the use of advanced computational methods such as deconvolution of cell type fractions and digital cytometry. In this study, we investigated the impact of different sources of variation on cell-type specific gene expression profiles. We evaluated normalization methods and the effect of batch correction and used variancePartition to quantify the contribution of variance originating from differences in lab, cell-type, cell subtype and tissue of origin for different gene sets. Furthermore, we investigated the biases between UMI-based single-cell and bulk RNA-Seq. Finally, we examined the effect of cell type profile features on the performance of deconvoluting cell type fractions.

We found that the variance explained by different factors is heavily dependent on the gene set, where lab explains the most variance when considering all genes. For deconvolution of cell type fractions and similar methods, the genes that separate cell types are of greater interest. Although cell type/cell subtype is the factor that explains most variance here, other factors such as lab and tissue of origin also have a substantial effect on gene expression for these genes. Consistent with our findings, large variation across experiments together with small differences between some cell subtypes has previously been shown(23). Our deconvolution results similarly show that indeed the performance of deconvolution decreases substantially when cell type profiles and mixed samples originate from different labs, and even more so if the profiles are generated from single-cell data. Hoffman et al has previously shown that the variance explained by lab is low(9). In that study, however, they used data where all laboratories were given precise instructions on how to perform the experiment. Such an experiment does not reflect the usual use case for deconvolution and other such methods; that the cell profiles are generated beforehand using one method, and that the mixtures are generated by another lab, potentially using other protocols, or by using a different technology. Our results show that for the typical use case, the technical factors are substantial.

Normalization and batch correction are important steps in the analysis of RNA-Seq data. In this study, we have shown that while library size normalization is inadequate, TMM(14) applied to pseudo counts and quantile normalization(18) both work well for normalizing between bulk and single-cell data. We also show that ComBat(16) effectively removes technical batch effects. These methods are however limited in that they either assume that most genes are not differentially expressed or require some biological overlap across samples. Neither of these criteria are generally fulfilled when normalizing and correcting bulk data from mixed samples with cell-type specific gene expression profiles. CIBERSORTx(4) implements a method for using ComBat on in-silico mixtures of cell type profiles together with mixed bulk samples (S-Mode batch correction) which clearly improves the deconvolution performance, but there are still challenges remaining to obtain accurate deconvolution results when the technical platforms are not matched.

We also sought to examine in detail the technical biases present for cell-type specific gene expression profiles derived from UMI-based single-cell RNA-Seq when compared to bulk RNA-Seq data. We found that the number of duplicate reads per mRNA molecule in the single-cell data (UMICF) can explain a substantial fraction of the differences between single-cell and bulk, whereas transcript length biases are generally small in comparison to other effects. GC content has previously been reported to introduce bias(24), and it is likely that, to a large extent, this bias is caused by PCR(22). We show that the UMICF covariate can explain more of this bias than GC content, and that these effects are partly correlated, suggesting that they at least partly describe the same underlying phenomenon. However, the combination of both covariates explains more of the variation between single-cell and bulk. A potential explanation is that GC content also provides information for genes with few reads for which duplicate reads per mRNA molecule is poorly estimated, or that GC content can explain other technical effects that are not captured by UMICF.

This study is limited in that we only investigated two cell types. We can thus not claim that the technical variation is of the same magnitude as differences across cell types in general, but only between B and T cells. Furthermore, the study was not fully balanced; some sources of variation are represented more strongly than others, which can have a slight effect on the results. For example, most samples originate from blood, increasing the influence of that specific environment when estimating the importance of the tissue of origin variation factor.

Our work suggests that estimating the number of duplicate reads per mRNA molecule can help in predicting and correcting for technical bias and thereby yield more comparable samples, both across

bulk samples, single-cell samples and between bulk and single-cell. These results need to be further validated in more datasets, and the factors introducing this bias need to be investigated in more detail. Although such factors may differ across experiments, it is possible that a library of factor patterns aggregated from many single-cell experiments could be used for a more generalized prediction and correction of bias in bulk data. Such a method would be useful for a broad range of applications extending beyond the generation of gene expression profiles for deconvolution or digital cytometry.

## Conclusions

In this study, we investigated the sources of variation in cell-type specific gene expression profiles. We demonstrated that technical effects resulting from different laboratory procedures and data types explains much of the variance across samples and confounds analyses such as deconvolution, but also that biological traits such as cell subtype and tissue of origin are important to consider when generating cell-type specific gene expression profiles. These results provide valuable insight to users of computational methods such as deconvolution of cell type profiles and digital cytometry, highlighting the importance of matching both technical protocols and biological traits between cell type profiles and bulk data samples.

## Methods

### Data Preparation

We downloaded the publicly available RNA-Seq datasets listed in Table 1, in total 74 bulk samples of B and T cells in addition to 8 single-cell datasets.

**Table 1**

*Table 1. List of RNA-Seq datasets used in this study.*

ID	Description	Data Type	Source
HCA CB	Umbilical cord blood PBMCs from the Human Cell Atlas; in total ~254,000 cells from 8 patients.	Single cell, 10x genomics, UMI counts.	Li et al(32), Rozenblatt-Rosen et al(33). The data can be downloaded from <a href="https://data.humancellatlas.org/">https://data.humancellatlas.org/</a> , Census of immune cells.
LC	~39,000 cells from the tumor microenvironment of lung cancers and ~13,000 cells from adjacent healthy tissue. The cells originate from 5 patients.	Single cell, 10x genomics, UMI counts.	Lambrechts et al(34). The data is available in in ArrayExpress under accessions E- <a href="#">MTAB-6149</a> and E- <a href="#">MTAB-6653</a> .
PBMC68k	~68,000 PBMCs from blood, one patient.	Single cell, 10x genomics, UMI counts.	Zheng G.X.Y. et al(3). The data is available at <a href="#">10x Genomics' home page</a> .
B10k	~10,000 FACS-sorted CD19+ B cells from blood, one patient.	Single cell, 10x genomics, UMI counts.	Zheng G.X.Y. et al(3). The data is available at <a href="#">10x Genomics' home page</a> .
CD4TMEM	~10,000 FACS-sorted CD4+/CD45RO+ Memory T Cells, one patient.	Single cell, 10x genomics, UMI counts.	Zheng G.X.Y. et al(3). The data is available at <a href="#">10x Genomics' home page</a> .

TCD8	~10,000 FACS-sorted CD8+ T cells from the blood of a single patient.	Single cell, 10x genomics, UMI counts.	Chen et al(35). The data is available for download on GEO data repository, accession number <a href="#">GSE 112845</a> .
MEL	~4,600 cells from the tumor microenvironment of Melanoma, 19 patients.	Single cell, SMART-Seq2, TPM	Tirosh et al(36). The data is available for download on GEO data repository, accession number <a href="#">GSE 72056</a> .
EVAL	Dataset produced for evaluating the performance of existing single-cell technologies. Data from mouse brain, PBMC and cell lines. Data includes 7 single-cell technologies and bulk, all performed on the same samples.	Single cell data from 7 different technologies and corresponding bulk samples, counts/ UMI counts/ TPM	Ding et al(17). The data is available for download at the Single Cell Portal, id <a href="#">SCP425</a> .
BULK 1	In total 6 bulk samples from B cells of varying origin.	Bulk RNA-Seq, FASTQ files	The ENCODE Consortium(37,38), Gingeras. The samples can be downloaded individually from <a href="#">ENCODE</a> .
BULK 2	In total 7 bulk samples from B cells (1) and T cells (6) of varying origin.	Bulk RNA-Seq, FASTQ files	The ENCODE Consortium(37,38), Stamatoyannopoulos and Weng. The samples can be downloaded individually from <a href="#">ENCODE</a> .
BULK 3	In total 12 bulk samples from B cells (6) and T cells (6) of varying origin.	Bulk RNA-Seq, FASTQ files	The functional annotation of the mammalian genome 5 (FANTOM5)(39,40). The data can be downloaded from <a href="#">FANTOM5</a> .
BULK 4	In total 39 bulk samples from B cells (16) and T cells (23) of varying origin.	Bulk RNA-Seq, FASTQ files	The BLUEPRINT Epigenome Project(41). The samples can be downloaded individually from <a href="#">BLUEPRINT</a> .
BULK 5	In total 10 PBMC bulk samples from B cells (5) and T cells (5).	Bulk RNA-Seq, RPKM/counts	Pabst et al(42), <a href="#">GSE 51984</a> .

We downloaded fastq files for BULK 1-4 to reduce the technical variability across datasets induced by the computational pipeline, and processed them using Kallisto(25) (v. 0.45.0). We pseudo-aligned to the HG38 (version GRCh38.p12) genome with the parameters “kallisto quant -i transcripts.gtf.gz -o [output folder] -b 1 [fastq file 1] [fastq file 2]”. For BULK 5, we did not have access to fastq files and instead used the RPKM expression values produced by the authors, converted to TPM.

For single-cell datasets, cell type classifications were retrieved from the authors of the study in cases where it was not publicly available. For Smart-Seq2 data (the MEL dataset), we used the TPM values produced by the authors, and pooled the cells within a cell population by calculating the average expression per gene. For 10x data, we pooled the cells by first summing the counts from all cells for each gene, and then scaled the expression to a total sum of  $10^6$  for all genes. For simplicity, only genes that existed in all datasets and could be properly converted to HGNC were used in this study. The datasets B10k, CD4TMEM and TCD8 have been treated as if they have been produced in the

same laboratory (called “SC Mixed 10x data”) even though they have not, which is motivated by that they have used similar techniques and contain too few samples to be treated as separate labs.

All samples are described in more detail in Table S1, including the design matrix and detailed information of each sample, for example the number of cells in each pooled single-cell sample.

### Normalization, PCA and Batch Correction

TPM normalization was performed according to  $TPM_i = \frac{10^6 * E_i}{\sum_j E_j}$ , where  $TPM_i$  represents the normalized expression for gene  $i$  and  $E_x$  is the expression of gene  $x$  before normalization.

TMM normalization was performed using the `calcNormFactors(14)` function in the `edgeR` package(26) (version 3.26.7). TMM was originally designed to work on counts, and needs to know the library size, but can work with non-integer data. The TPM values were scaled to pseudo-counts, where the sum of all gene expression values equals the original library size. The pseudo-counts differ from the original counts in that they are corrected for transcript length, but with identical library size.

For quantile normalization we used the function `normalize.quantiles` in the `preprocessCore` package(27) (version 1.46.0).

PCA was performed using the R function `prcomp` with the parameter `scale` set to `FALSE`.

For batch correction we used the `ComBat` function in the `SVA` package, specifying that differences related to cell type should be preserved (in the `model.matrix`, using “~1 + cellType”). As batch, we used dataset id with one modification; the datasets PBMC68k, B10k and CD4TMEM were treated as the same dataset since they had too few samples to be batch corrected separately. We deemed that this was reasonable, since the data is produced in a similar way and published in the same publication.

### Log Transformation

We applied log transformation for many analyses to make the expression data more normally distributed. The log transform is applied according to  $L_i = \log_2(E_i + b)$ , where  $L_i$  is the log transformed expression of gene  $i$ ,  $E_i$  is the expression of gene  $i$  in pseudo-TPM, and  $b$  is a constant set to 1, which is added to avoid taking the logarithm of zero values.

We use the term “log<sub>2</sub> fold change” (LFC) throughout the results to compare the expression of a gene between two samples. This is calculated as  $LFC_i = \log_2\left(\frac{E_{i,1}+b}{E_{i,2}+b}\right)$ , where  $E_{i,1}$  and  $E_{i,2}$  represent the expression of gene  $i$  in the two samples which are to be compared.

### Measuring the Explained Variance of Factors

We used the functions “`fitExtractVarPartModel`” and “`plotVarPart`” from the R package `variancePartition`(9) to calculate and plot the variance of factors, using TMM-normalized data. All factors were modeled as random effects.

### Deconvolution with CIBERSORTx

For generation of cell type profiles used for deconvolution with CIBERSORTx, the bulk profiles were generated by inputting data sample-wise, TPM normalized, in a data matrix to CIBERSORTx. The pooled single-cell profiles were handled in a similar way, where each pooled sample from the design matrix was treated as a single bulk sample. Profiles were created using only cells from datasets within the same lab. The final calculation of profiles was done by CIBERSORTx, using the “Create

Signature Matrix” function (specifying RNA-Seq). For single-cell data, the cells of each type were sorted into 100 bins for disk quota reasons in CIBERSORTx. Each bin was given the gene expression of the mean of all the cells in the bin, TPM normalized. The final profiles were then generated using the CIBERSORTx “Create Signature Matrix” function (specifying sc RNA-Seq).

For generating mixtures used for deconvolution, the expression values of one T-cell sample multiplied by 0.5 was added to a similarly treated B-cell sample, both TMM normalized. The mixture samples were then TPM-normalized before being sent to CIBERSORTx. The reason for this procedure is that TPM-normalization before mixing the data could confound the mixture, creating a mixture where one cell type is given more weight, since library size normalization is sometimes inaccurate.

For the LM22 results, the fraction of B cells is the sum of the fractions for “B cells naive”, “B cells memory”, and “Plasma cells”. The pooled samples are pooled and input to CIBERSORTx as bulk profiles, while the SC samples for disk quota reasons are pooled into 100 B cells and 100 T cells for each profile pair, and input as single-cell profiles. For evaluation of profiles generated from bulk data, the mixtures generated from the same lab as the profiles were removed to avoid using mixtures and profiles generated from the same data.

CIBERSORTx deconvolution was run using the function “Impute Cell Fractions”, specifying batch correction (B or S) and quantile normalization to produce such results. For quantile normalization, separate cell type profiles were produced, specifying that the cell type profiles should be quantile normalized.

CIBERSORTx was used from the CIBERSORTx web page (<https://cibersortx.stanford.edu/>), accessed 2020-06-05.

### Retrieval of Transcript Length and GC Content

Transcript length was retrieved using the GenomicFeatures(28) R package (version 1.36.4) together with the biomaRt(29) package (version 2.40.0). We used the biomaRt ENSEMBL\_MART\_ENSEMBL (version 98) and the dataset mmusculus\_gene\_ensembl (version GRCm38.p6). We calculated GC content by using the R package BSgenome.Mmusculus.UCSC.mm10(30) (version 1.4.0), together with GenomicFeatures and Biostrings(31).

### The UMICF Covariate

The EVAL dataset contains both UMI counts and total counts, which makes it possible to calculate the number of copies per molecule. The UMICF covariate is calculated as

$$UMICF = \frac{total\ counts - UMI\ counts}{total\ counts}.$$

### Regressing out Covariates

To regress out one or more covariates, a linear or loess (R package stats v3.6.1 using default parameter values) curve was first fitted to the  $\log_2$  fold change between 10x and bulk in the covariates space. The curve was then regressed out of the 10x gene expression in log space as  $L_{corr,i} = L_{orig,i} - p_i + mean(p)$ , where  $L_{corr,i}$  is the corrected gene expression for gene  $i$ ,  $L_{orig,i}$  is the original gene expression,  $p_i$  is the predicted value of gene expression from the fit and  $mean(p)$  is the mean of all predicted values from the fit. The UMICF covariate was set to NA unless more than 5 unique UMIs were available for the gene, to avoid the noise induced by too few measurement points.

For all analyses, we removed a few outliers with extreme values for transcript length and GC content, in total 232 genes. UMICF was set to NA for all genes for which we had five or fewer UMIs (5321 and 6126 genes for cortex 1 and 2, respectively), since we deemed the amplification measure to be too noisy otherwise. All excluded genes were left untouched by the regression. In total (including genes for which UMICF was set to NA) 26,563 genes were used in the calculations for both Cortex 1 and Cortex 2. The genes where UMICF was set to NA were still used in the correlation calculation.

### Statistical Tests

For estimating the confidence intervals in figure 6, we bootstrapped the genes to include in the correlation for 10,000 iterations. The confidence intervals were then calculated as the values located at 2.5% and 97.5% in the sorted vector of correlations.

We used the function R function `wilcox.test` for the Wilcoxon signed rank tests for investigating if correlation differences between covariates were significant, with the parameters `alternative = "greater"`, `paired = TRUE`. The same bootstraps were used for all covariates, making it possible to compare them with a paired test.

### Software

The data was analyzed using R version 3.6.1 and MATLAB R2018b. MATLAB was used for assembling the single-cell data and exporting the pooled samples to a text file; the rest of the analysis was done in R. The MATLAB code uses the component `SingleCellToolbox` for importing public single-cell datasets (<https://github.com/SysBioChalmers/SingleCellToolbox>). The processed data and source code are available at: [doi.org/10.5281/zenodo.3900226](https://doi.org/10.5281/zenodo.3900226)

## Acknowledgements

This study makes use of data generated by the BLUEPRINT Consortium. A full list of the investigators who contributed to the generation of the data is available from [www.blueprint-epigenome.eu](http://www.blueprint-epigenome.eu). Funding for the project was provided by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 282510 – BLUEPRINT.

## References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009 Jan;10(1):57–63.
2. Picot J, Guerin CL, Le Van Kim C, Boulanger CM. Flow cytometry: retrospective, fundamentals and recent instrumentation. *Cytotechnology.* 2012 Mar;64(2):109–30.
3. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun.* 2017 Jan 16;8:14049.
4. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol.* 2019 Jul;37(7):773–82.
5. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife [Internet].* [cited 2018 May 13];6. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5718706/>
6. Frishberg A, Peshes-Yaloz N, Cohn O, Rosentul D, Steuerman Y, Valadarsky L, et al. Cell composition analysis of bulk genomics using single-cell data. *Nat Methods.* 2019 Apr;16(4):327.
7. Shao X, Liao J, Lu X, Xue R, Ai N, Fan X. scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *iScience.* 2020 Mar 27;23(3):100882.
8. Cao Y, Wang X, Peng G. SCSA: a cell type annotation tool for single-cell RNA-seq data. *bioRxiv.* 2020 Mar 25;2019.12.22.886481.
9. Hoffman GE, Schadt EE. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics.* 2016 Nov 25;17(1):483.
10. 't Hoen PAC, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat Biotechnol.* 2013 Nov;31(11):1015–22.
11. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* 2017 Nov 15;18:220.
12. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 2008 Jul;5(7):621–8.
13. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 2012 Dec;131(4):281–5.
14. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 2010 Mar 2;11:R25.

15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014 Dec 5;15:550.
16. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007 Jan 1;8(1):118–27.
17. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat Biotechnol.* 2020 Apr 6;1–10.
18. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature.* 2010 Apr;464(7289):768–72.
19. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012 Mar 15;28(6):882–3.
20. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.* 2015 Mar 30;12(5):nmeth.3337.
21. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. *Trends Genet TIG.* 2013 Oct;29(10):569–74.
22. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 2012 May;40(10):e72.
23. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep.* 2019 Feb 5;26(6):1627-1640.e7.
24. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics.* 2011 Dec 17;12(1):480.
25. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol.* 2016 May;34(5):525–7.
26. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma Oxf Engl.* 2010 Jan 1;26(1):139–40.
27. Bolstad B. preprocessCore: A collection of pre-processing functions version 1.46.0 from Bioconductor [Internet]. [cited 2019 Oct 24]. Available from: <https://rdrr.io/bioc/preprocessCore/>
28. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. *PLoS Comput Biol.* 2013;9(8):e1003118.
29. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc.* 2009;4(8):1184–91.
30. The Bioconductor Dev Team. BSgenome.Mmusculus.UCSC.mm10: Full genome sequences for *Mus musculus* (UCSC version mm10). R package version 1.4.0. [Internet]. Bioconductor. 2014

[cited 2019 Oct 24]. Available from:  
<http://bioconductor.org/packages/BSgenome.Mmusculus.UCSC.mm10/>

31. Pagès H, Aboyoun P, Gentleman R, Debroy S. Biostrings: Efficient manipulation of biological strings version 2.52.0 from Bioconductor [Internet]. 2019 [cited 2019 Oct 24]. Available from: <https://rdrr.io/bioc/Biostrings/>
32. Li B, Kowalczyk MS, Dionne D, Ashenberg O, Tabaka M, Tickle T, et al. Census of Immune Cells [Internet]. Human Cell Atlas Data Portal. 2018 [cited 2019 Feb 19]. Available from: <https://data.humancellatlas.org/explore/projects/cc95ff89-2e68-4a08-a234-480eca21ce79>
33. Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. The Human Cell Atlas: from vision to reality. *Nat News*. 2017 Oct 26;550(7677):451.
34. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat Med*. 2018 Aug;24(8):1277–89.
35. Chen J, Cheung F, Shi R, Zhou H, Lu W, Candia J, et al. PBMC fixation and processing for Chromium single-cell RNA sequencing. *J Transl Med*. 2018 Jul 17;16(1):198.
36. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016 Apr 8;352(6282):189–96.
37. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012 Sep 6;489(7414):57–74.
38. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res*. 2018 04;46(D1):D794–801.
39. The FANTOM Consortium and the RIKEN PMI and Clst (dgt), Forrest ARR, Kawaji H, Rehli M, Kenneth Baillie J, de Hoon MJL, et al. A promoter-level mammalian expression atlas. *Nature*. 2014 Mar;507(7493):462–70.
40. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol*. 2015 Jan 5;16(1):22.
41. Blueprint Epigenome Project, 2016. [Internet]. [cited 2019 Mar 4]. Available from: <http://dcc.blueprint-epigenome.eu/#/home>
42. Pabst C, Bergeron A, Lavalley V-P, Yeh J, Gendron P, Norddahl GL, et al. GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. *Blood*. 2016 Apr 21;127(16):2018–27.

## Supplementary Information Legends

Supporting Text S1. Supplementary methods and figures.

Table S1. Sample Information

# Figures

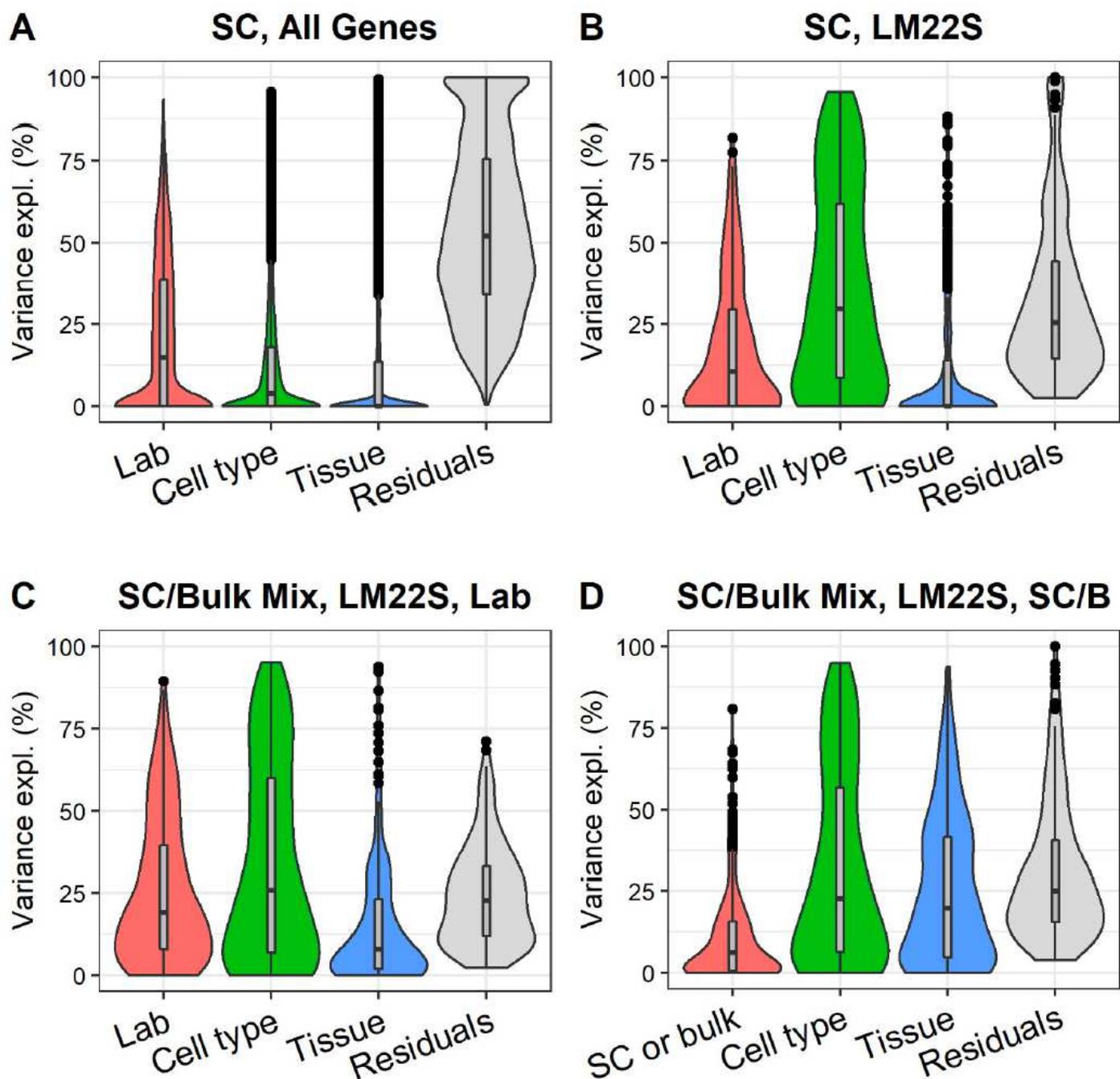
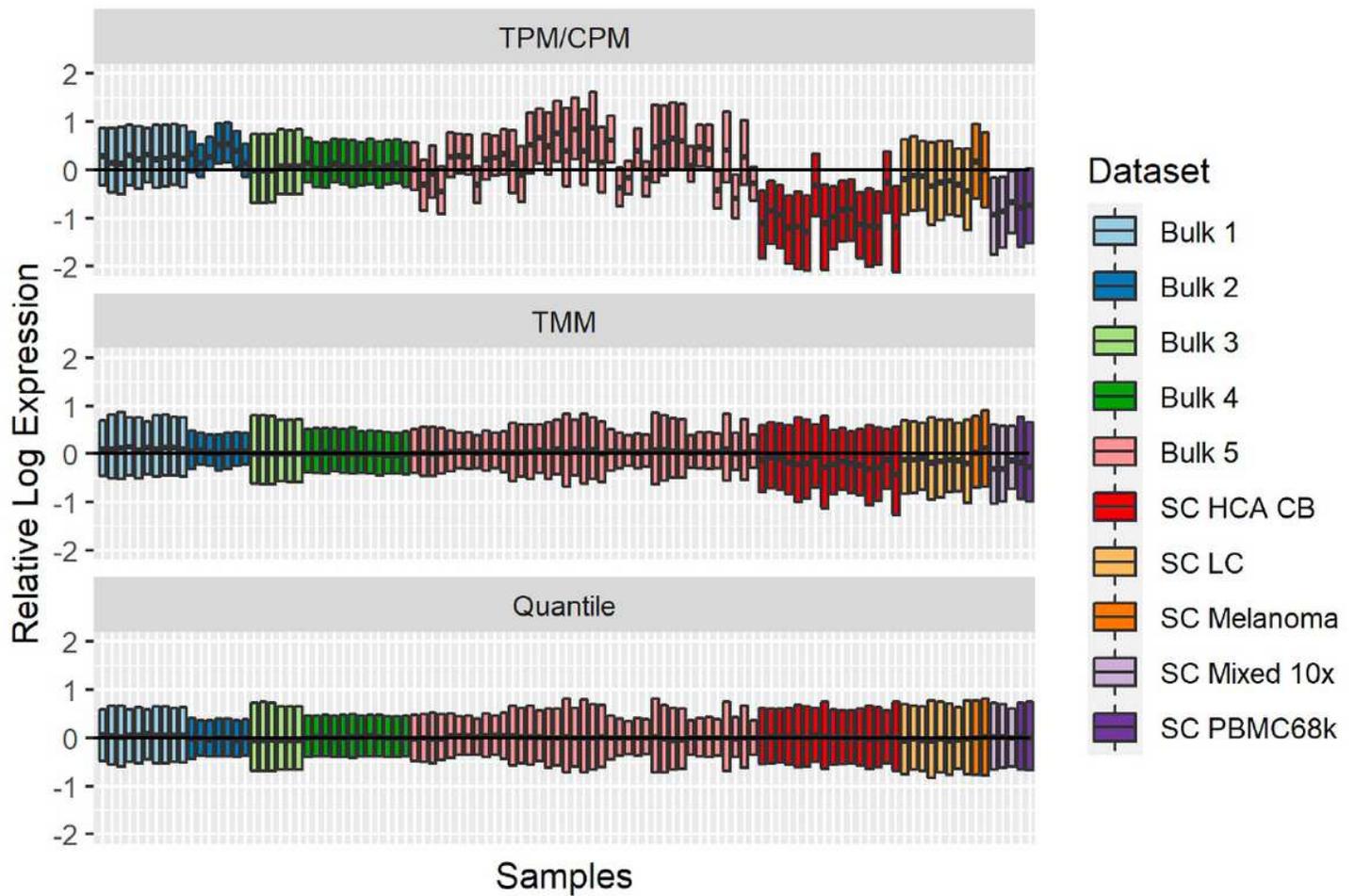


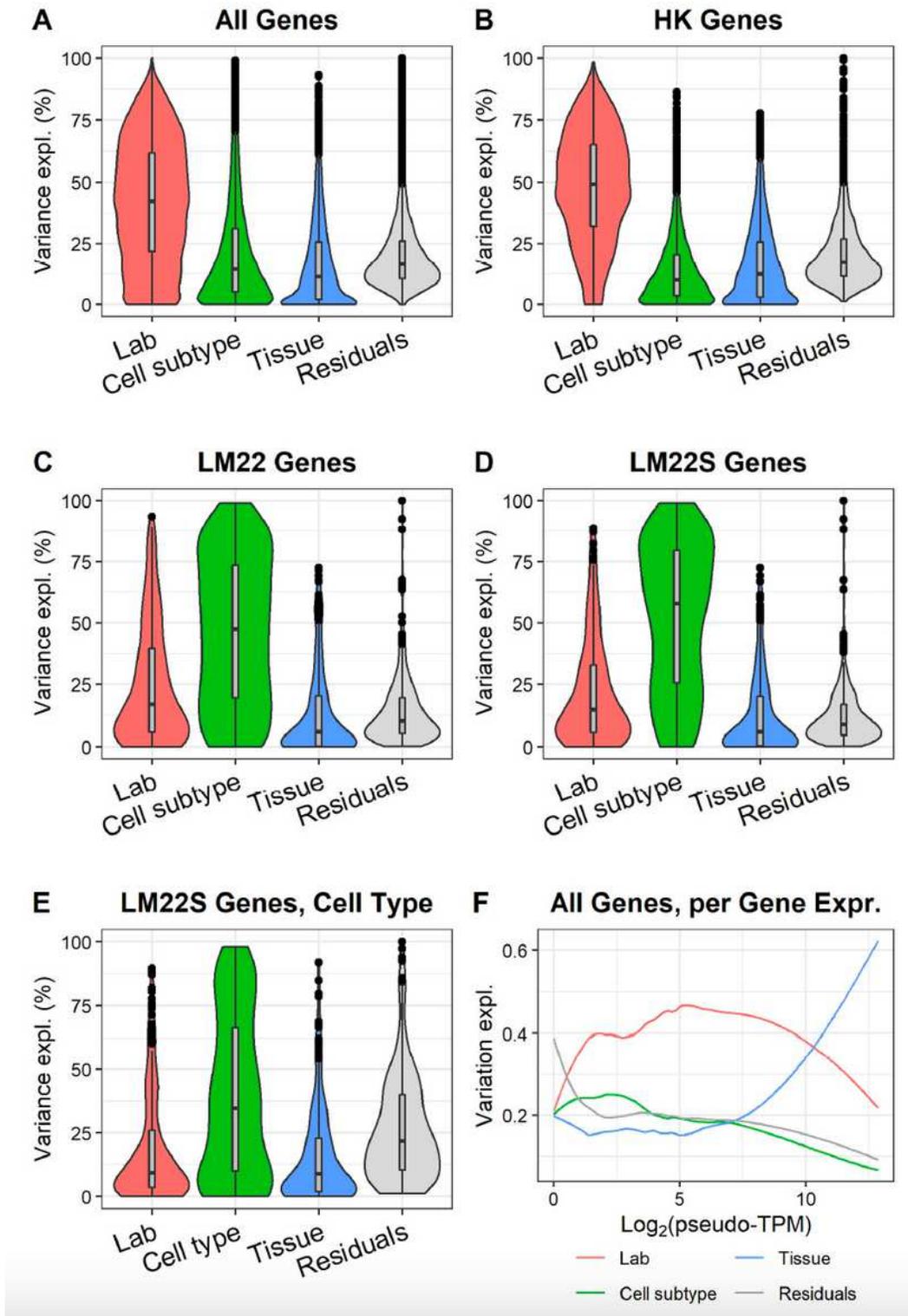
Figure 1

Explained variance for single-cell samples. A. Only single-cell samples, all genes. B. Only single-cell samples, LM22S. C. Mix of single-cell and bulk samples, LM22S. D. Identical to C, with the exception that lab is replaced by a flag indicating if the samples are single-cell or bulk.



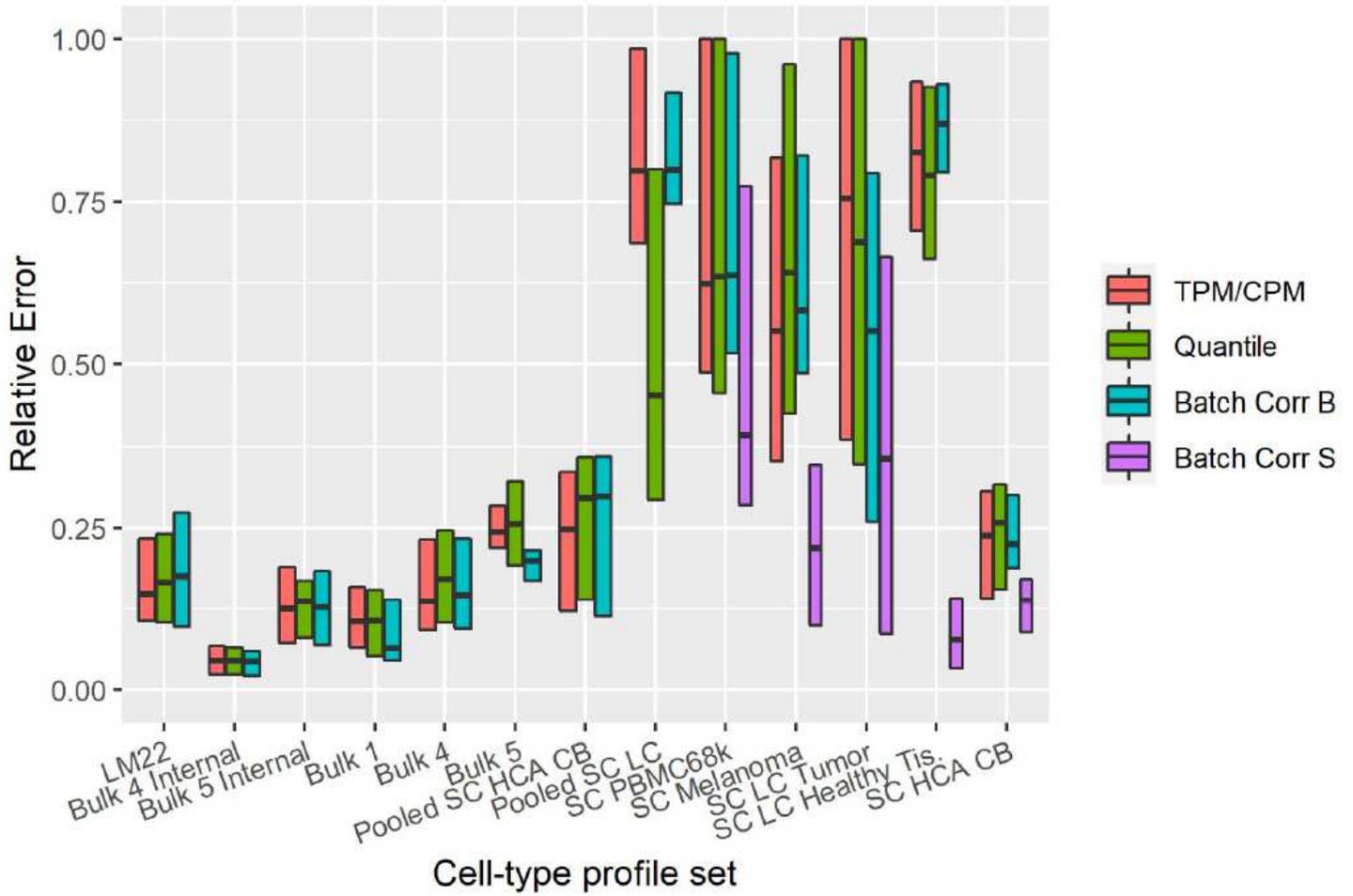
**Figure 1**

Evaluation of normalization methods for combined single-cell and bulk data. Each bar is a boxplot over all genes describing the  $\log_2$  fold change in gene expression between the sample and the mean over all samples. 12,072 genes were included for each sample.



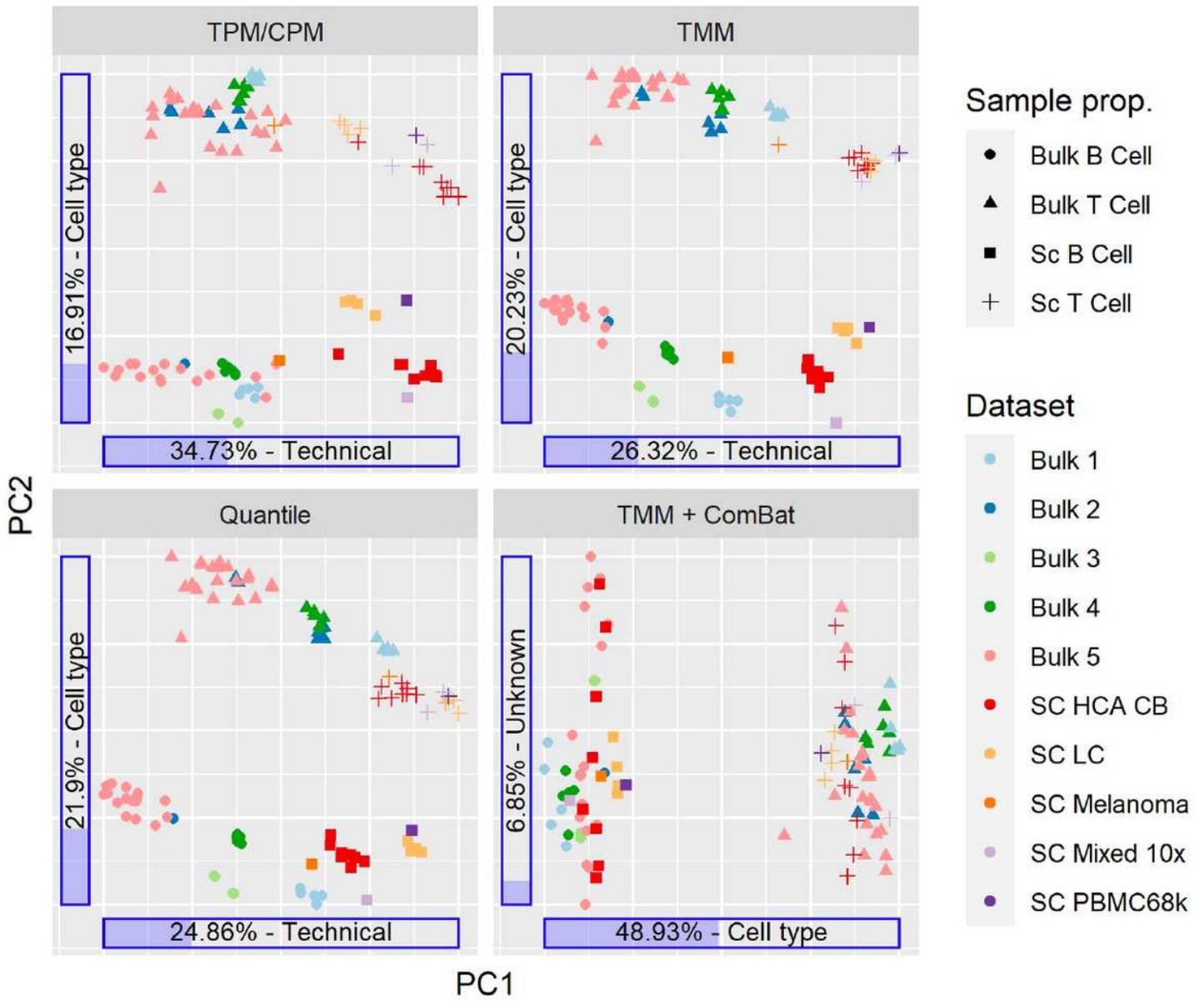
**Figure 1**

Explained variance in gene expression for bulk RNA-seq samples. A. All genes (12072 genes). B. Housekeeping genes (3393 genes). C. LM22 (395 genes). D. LM22S genes. E. Identical to D, with the difference that cell subtype is replaced with cell type (B/T). F. Explained variance per gene expression. The plot shows how the explained variance by the different factors change with gene expression (Loess fit, span = 0.3)



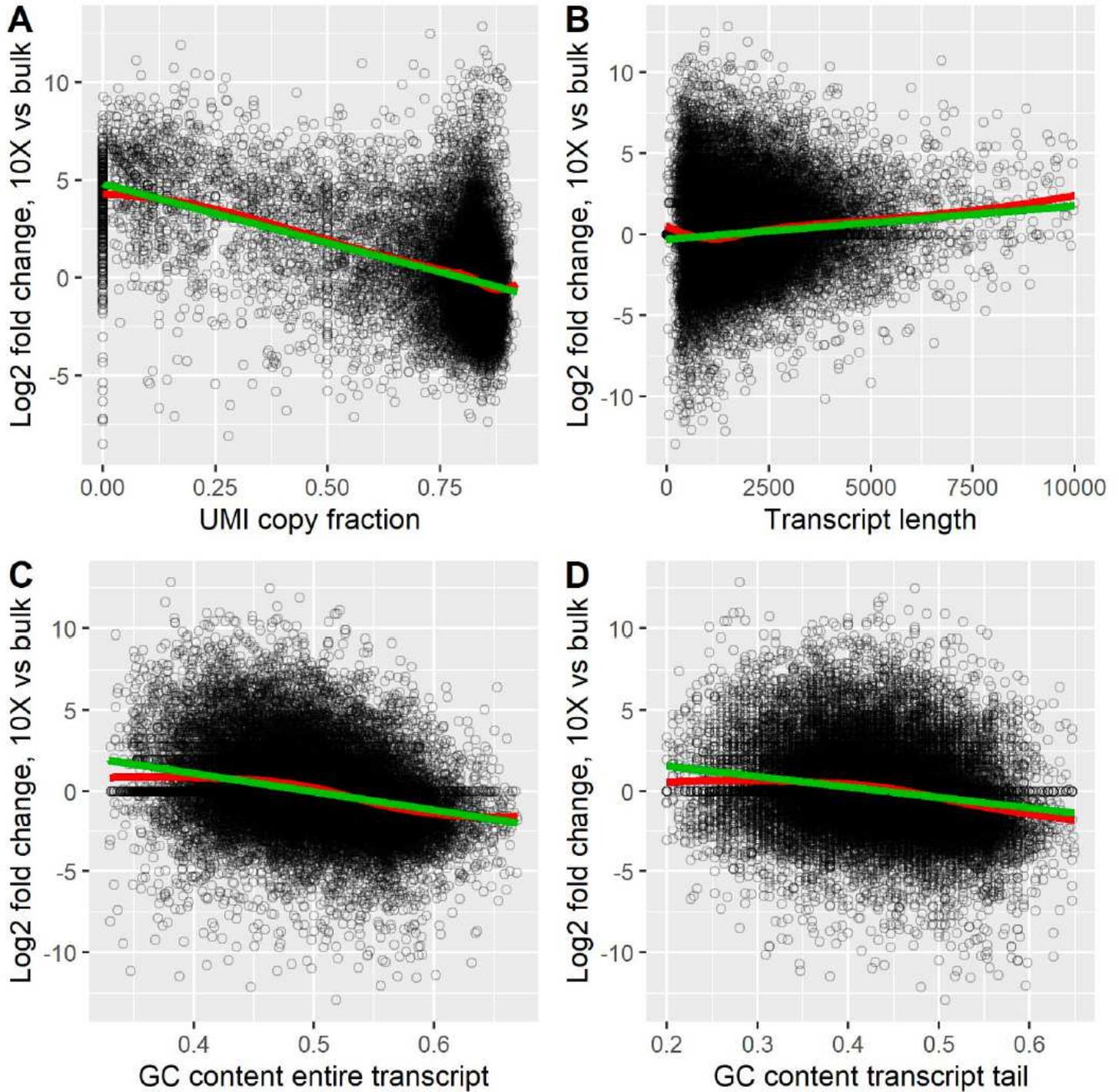
**Figure 1**

Deconvolution performance for different cell type profiles tested against in-silico generated bulk mixtures with 50% T cells and 50% B cells. The figure shows the relative error of the B cell fraction estimated by CIBERSORTx. The Bulk 4/5 Internal profiles were generated from the Bulk 4/5 lab and were tested against mixtures generated from other samples from the same lab. The remaining profiles were tested against 40 mixtures generated from Bulk 1, 2, 4, and 5 (Bulk 3 has no T cell samples).



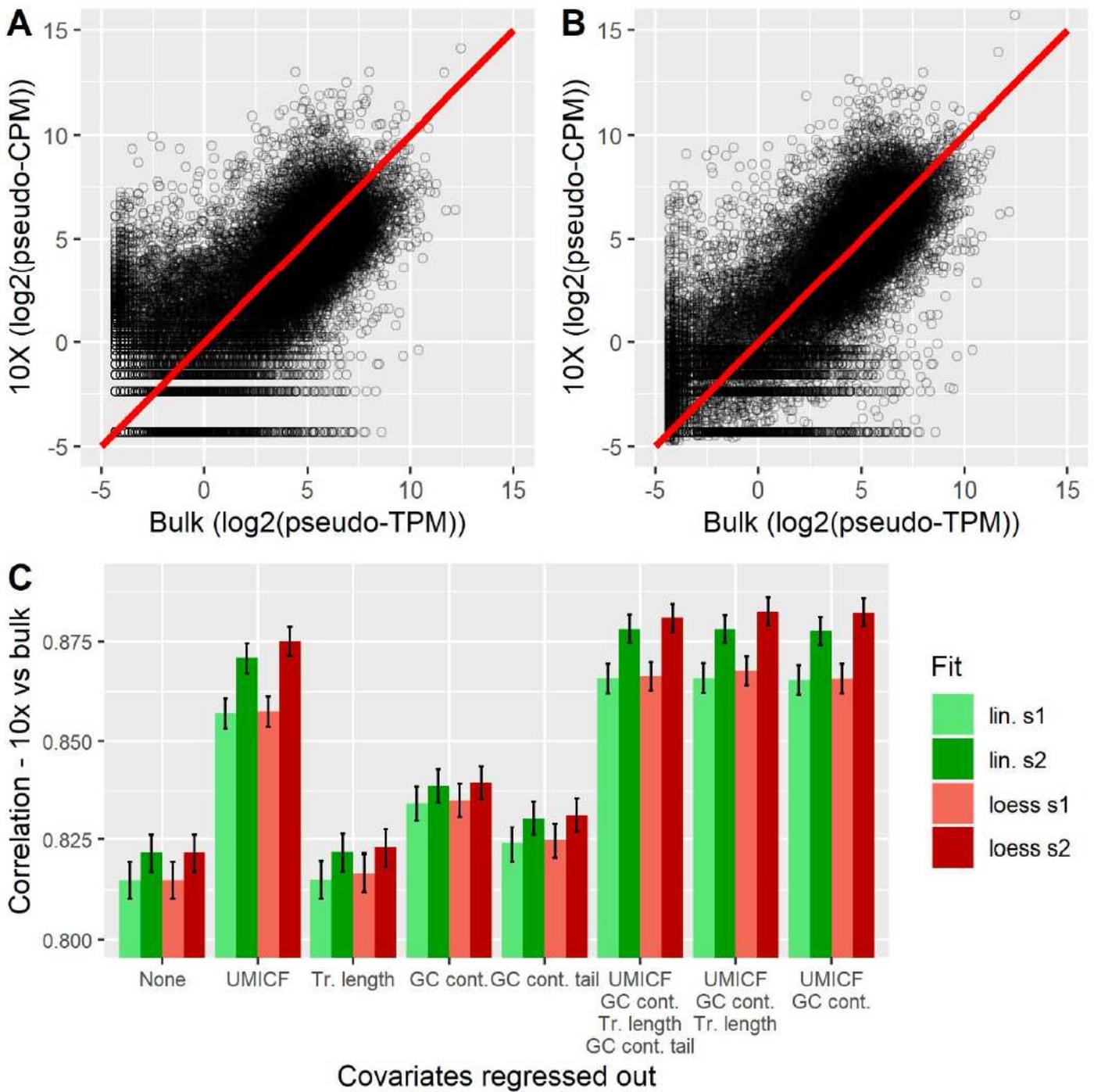
**Figure 2**

Visualization of normalization and batch effects using PCA. 12,072 genes were included for each sample.



**Figure 5**

Log<sub>2</sub> fold change between 10x data and bulk for each gene, plotted as a function of different covariates. The green line represents a linear fit, whereas the red line shows a Loess fit. The data shown is from cortex 1 of the EVAL dataset. A. UMI copy fraction. Only genes with more than 5 molecules available for calculating UMICF are shown. B. Transcript length. C. The GC content of the entire transcript. D. The GC content of the 150 bases closest to the transcript tail.



**Figure 6**

A. Gene expression for cortex 1 from the EVAL dataset plotted as 10x vs bulk. The red line represents a perfect correlation. B. Gene expression for cortex 1 from the EVAL dataset after regressing out the differences in UMICF and GC content between 10x and bulk using a loess fit, which improves the correlation. C. Average Pearson correlation coefficient between 10x data and bulk in log scale after regressing out technical covariates (UMI copy fraction, transcript length, GC content and GC content tail), using linear or loess regression. The correlation shown is the average of the correlations from cortex 1

and 2 of the EVAL dataset. The error bars represent the confidence interval, based on bootstrapping of genes to include in the correlation.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS1.SampleInformation.xlsx](#)
- [SupplementaryInformationv3.docx](#)