Sources of Variation in Cell-Type RNA-Seq Profiles

Johan Gustafsson[1,2,*], Felix Held[3], Jonathan Robinson[1,2], Elias Björnson[1,4], Rebecka Jörnsten[3] and Jens Nielsen[1,2,5]

[1] Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, Gothenburg, Sweden.

[2] Wallenberg Center for Protein Research, Chalmers University of Technology, Kemivägen 10, Gothenburg, Sweden.

[3] Mathematical Sciences, University of Gothenburg and Chalmers University of Technology, Chalmers tvärgata 3, Gothenburg, Sweden

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

[5] BioInnovation Institute, Ole Maaløes Vej 3, DK2200 Copenhagen N, Denmark

* Corresponding author

Abstract

Background

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.

Results

We evaluated different normalization methods, quantified the magnitude of variation introduced by different sources, and examined the differences between UMI-based single-cell RNA-Seq and bulk RNA-Seq. We applied methods such as random forest regression to 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 of the same magnitude as the biological variation across cell types. Tissue of origin and cell subtype are less important but still substantial factors, while the difference between individuals is relatively small. 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.

Conclusions

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.

Background

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).

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.

Normalization is a crucial step when analyzing RNA-Seq data. In the beginning of the RNA-Seq era, library size normalization, for example FPKM(7) and TPM(8), 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)(9) and the normalization performed by DESeq2(10). These methods are designed to operate directly on gene counts and work well under the assumption that most genes are not differentially expressed across samples. Being restricted to working on counts makes it more difficult to normalize single-cell data based on unique molecular identifiers (UMIs) together with bulk. Bulk counts need to be divided by transcript length, such as in FPKM and TPM normalization, to get a representative gene expression, while UMI-based single-cell data should not. Consequently, the methods working on counts cannot directly account for gene length, while library size normalization, which allows for such correction, fails to properly handle the problem with highly expressed noisy genes.

A common technique used to overcome technical biases is to computationally remove batch effects. The batch effect removal tool ComBat(11) 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 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 different factors that cause undesired variation between gene expression profiles of individual cell types. We evaluated normalization and batch correction methods, quantified different sources of variation between samples, and investigated technical biases between single-cell and bulk RNA-Seq.

Results

Data Preparation

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). In addition, we investigated a dataset(12) containing both bulk and single-cell data from the same samples to examine biases between single-cell and bulk sequencing.

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)(8), Trimmed Mean of M-values (TMM)(9), and quantile normalization(13). 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(9). 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 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.

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(11) from the sva(14) 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.

Sources of Variation Between Samples

To quantify sources of variation in gene expression across samples we investigated the impact of several factors on gene expression. Factors examined were laboratory, cell type, subcellular type, tissue of origin, individual and whether samples are technical replicates. We analyzed the 105 samples in pairs, normalized using TMM as described above, and collected information about the factors to be investigated. We selected pairs that only differed in maximum one factor and compared them using the LFCSD (Log Fold Change Standard Deviation) metric (Methods). Figure 3 shows the variation induced by the different factors.

Technical replicates, although represented by only a few samples, exhibit lower differences in expression than different biological samples. Pairs where the samples come from the same individual taken at different time points show similar differences as pairs from different individuals, indicating that differences between individuals are small compared to other factors. Differences in tissue of origin are greater than for subcellular type, but both introduce less variation than difference in cell type. Pairs where the samples originate from different labs exhibit similar variation as pairs with different cell type, but pairs from pooled single-cell data from different labs or a combination of single-cell and bulk show a larger variation. Pooled single-cell profiles seem to have more variation than their bulk counterpart, which is potentially a combination of higher technical variability and presence of misclassified cells. We also repeated the pairwise investigation using Pearson correlation between samples instead of LFCSD (Supplementary Information). Although such a comparison may be more intuitive to some readers, it fails to include normalization issues, which is a technical factor that should ideally be included in the comparison.

As an alternative method to quantify sources of variation in gene expression, we employed random forest regression to PCA-transformed data (Methods). Figure 4 shows pairwise comparisons for different factors. In general, the random forest regression confirms the results from the LFCSD method. The random forest method is sensitive to unbalanced datasets, which can explain the divergence from the LFCSD method for subcellular type versus tissue of origin. For B cells, lab 5 has only a few samples with a specialized subcellular type, while it is specialized for all T samples. This makes the subcellular type more important for the T cells, while this does not necessarily reflect real differences in variation between the cell types.

Differences Between Single-Cell and Bulk

To unravel the sources of variation between pooled single-cell data and bulk, we investigated the EVAL dataset (Methods), which contains bulk and single-cell (10x) data generated from the same samples(12). We used the cortex 1 and cortex 2 samples, originating from mouse brain, to see if we could 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 (both UMI counts and total counts are available in this dataset). We defined the UMI copy fraction as

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

for each gene, representing the fraction of the counts that are filtered as UMI duplicates. The UMICF used for cortex 1 was calculated from cortex 2 and vice versa, to reduce the effect of any

dependency between UMICF and gene expression originating from sampling. Figure 5A shows a clear negative correlation between the difference in gene expression and UMICF. 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.

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.

Third, we investigated the effect of the GC content of genes since this is a known source of bias in RNA-Seq(15). 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, 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. The UMICF and both GC content covariates are correlated with each other. It is evident that the UMICF covariate explains most variation, though the other covariates explain some variation on their own. When all covariates are combined, the tail GC content and transcript length do not contribute much to improving correlation; the combination of UMICF and GC content is therefore a good choice. The differences between regressing out a loess or linear fit for a covariate are generally small, although loess performs slightly better.

Discussion

Cell-type specific gene expression profiles are useful for analyzing bulk RNA-Seq samples containing mixed cell types, since it enables 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 random forest regression to quantify the contribution of variation originating from differences in lab, cell-type, cell subtype and tissue of origin. Furthermore, we investigated the biases between UMI-based single-cell and bulk RNA-Seq. We found that the variation introduced by using data from different labs was of the same magnitude as that between cell types (B and T cells), and that although cell subtype and tissue of origin had less

impact, they were not negligible. Consistent with our findings, large variation across experiments together with small differences between some cell subtypes has previously been shown(16). Furthermore, we evaluated the variation between technical replicates and samples from the same individual. Technical replicates showed less variation than different samples from similar conditions produced at the same lab, which is consistent with previous findings showing high correlation among technical replicates(7). In contrast, samples from the same individual taken at different time points exhibit similar variation as samples from different patients. This suggests that the biological differences in the transcriptome between individuals are small compared to the combined effect of sample-specific technical factors and biological differences within individuals over time.

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(9) applied to pseudo counts and quantile normalization(13) both work well for normalizing between bulk and single-cell data. We also show that ComBat(11) 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. Although CIBERSORTx(4) implements methods for using ComBat on in-silico mixtures of cell type profiles together with mixed bulk samples, there are still biases from the fractions of cell types used in the mixtures, which calls for additional development in this field.

We also sought to examine cell-type specific gene expression profiles derived from UMI-based single-cell RNA-Seq, and specifically their usefulness together with computational methods operating on 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(17), and it is likely that, to a large extent, this bias is caused by PCR(15). 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 effect. However, the combination of both covariates explains more of the variation between single-cell and bulk. A potential explanation is that GC content also provide 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, where the variation between those cell types was used as reference value in relation to other sources of variation. 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. In addition, we did not consider the fact that for some algorithms, only cell-type marker genes are of interest. For such genes, the biological variation is likely higher than the technical, although technical variation will still be present.

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 introduce the largest variation, but also that biological traits such as cell subtype and tissue of origin are also 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.

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(29) (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.

Normalization and Batch Correction

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

TMM normalization was performed using the calcNormFactors(9) function in the edgeR package(30) (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(31) (version 1.46.0).

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 to 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 $\boldsymbol{L_i = log2(E_i + b)}$, where $\boldsymbol{L_i}$ is the log transformed expression of gene $\boldsymbol{i}$, $\boldsymbol{E_i}$ is the expression of gene $\boldsymbol{i}$ in pseudo-TPM, and $\boldsymbol{b}$ is a constant set to 0.05, which is added to avoid taking the logarithm of zero values.

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

Measuring the Importance of Factors for Explaining Differences Between Samples

We employed two techniques to measure the influence of covariates on the variability between gene expression profiles. The first method, called LFCSD (Log Fold Change Standard Deviation), compares samples pairwise, and only uses pairs for which 0 or 1 covariate differs. The second method, random forest regression, regresses out the factors using multiple samples.

The LFCSD Metric for Sample Comparison

The LFCSD (Log Fold Change Standard Deviation) metric measures the dissimilarity of the transcriptome between two samples. The log2 fold change is calculated for all genes and the standard deviation of all fold changes is used as a metric of sample difference. The histogram over fold changes between pairs of samples typically produces a bell-shaped curve (Supplementary Information, Fig. S1), which motivates the use of the standard deviation as a fair quantification of dissimilarity for sample pairs. The metric is calculated from TMM-normalized data as described above. To avoid the noise of lowly expressed genes, only genes with pseudo-TPM >= 1 are used in the calculation.

Random Forest Regression Based on PCA

For the second method, we implemented a variant of Random Forests(32) to measure covariate importance. This analysis did not use technical replicates, leading to the removal of 5 samples. All considered covariates were one-hot encoded. The total number q of resulting predictors was therefore dependent on the number of factor levels of the involved covariates. Three normalizations (TPM, TMM and Quantile) were considered.

All 100 principal components of the remaining 100 log-transformed gene profiles (see above) were computed. These were used in the further analysis instead of the full gene profiles to reduce computational complexity while preserving the variance structure between samples. This can be seen as follows. If $\boldsymbol{Y}$ is the original $n\ x\ p$ (samples times genes) data matrix, then (assuming rows are mean-centred) $\boldsymbol{YY^T}$ is the covariance matrix between samples. Taking the singular value decomposition (SVD) of $\boldsymbol{Y = UDV^T}$ leads to the principal components by setting $\boldsymbol{P = UD}$. For the covariance matrix of the principal components then holds $\boldsymbol{PP^T = UDV^TVDU^T = YY^T}$ since $\boldsymbol{V^TV =}$

$\boldsymbol{I}_p$, the unit matrix of dimension p. The variance (and covariance) structure between the samples is therefore preserved.

Given a subset of the data, the following process was used to measure the impact of covariates on the gene profiles, encoded as their principal components. The package randomForestSRC(33) was used to build a multivariate random forest with 300 trees, minimum leaf size 5, and *mtry* equal to the largest integer less than $q \,/\, 3$. The mean squared error (MSE) for the out of bag (OOB) samples(32) was observed to ensure convergence.

The importance of a covariate is measured with a permutation scheme. First, the MSE was measured on the OOB samples for each tree. Then, all one-hot encoded variables belonging to a covariate are permuted to break the relationship between the covariate and the samples. All other covariates remained untouched. The MSE was measured on the modified samples and the change in MSE was recorded. Larger changes indicate larger importance of the covariate. To ensure stability in the estimate, the permutation process was repeated 50 times and the average change in MSE was used as an estimate of covariate importance stemming from a particular tree. The final covariate importance measure is computed by averaging the changes in MSE over all trees.

Retrieval of Transcript Length and GC Content

Transcript length was retrieved using the GenomicFeatures(34) R package (version 1.36.4) together with the biomaRt(35) 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(36) (version 1.4.0), together with GenomicFeatures and Biostrings(37).

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 $\boldsymbol{L_{corr,i} = L_{orig,i} - p_i + mean(p)}$, where $\boldsymbol{L_{corr,i}}$ is the corrected gene expression for gene $\boldsymbol{i}$, $\boldsymbol{L_{orig,i}}$ is the original gene expression, $\boldsymbol{p_i}$ 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. For all analyses including UMICF, we measured UMICF in cortex 2 when analyzing cortex 1, and vice versa. We excluded all genes for which we had five or fewer UMIs in the other cortex, since we deemed the amplification measure to be too noisy otherwise. All excluded genes were left untouched by the regression. In total, 5321 and 6126 genes were excluded for cortex 1 and 2, respectively. The genes excluded due to low UMICF reliability were still used in the correlation calculation.

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).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Means to access the datasets analyzed during the current study are listed in Table 2. The processed data and source code is available at: doi.org/10.5281/zenodo.3631488

Funding

This work was supported by funding from the Knut and Alice Wallenberg foundation (J.N.), the National Cancer Institute of the National Institutes of Health under award number F32CA220848 (J.R.), and the Swedish Foundation for Strategic Research under award number BD15-0088 (R.J.).

Funding for the BLUEPRINT project was provided by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 282510 – BLUEPRINT.

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. Funding for the project was provided by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 282510 – BLUEPRINT.

This work was supported by funding from the Knut and Alice Wallenberg foundation (J.N.), the National Cancer Institute of the National Institutes of Health under award number F32CA220848 (J.R.), and the Swedish Foundation for Strategic Research under award number BD15-0088 (R.J.).

Competing Interests

The authors declare that they have no competing interests.

Authors' Contribution

Conceptualization, J.G., J.R.; Methodology, J.G, J.R., E.B., F.H., R.J.; Software, J.G., F.H., E.B.; Writing – Original Draft, J.G., J.R.; Writing – Review & Editing, J.G., J.R., F.H., E.B., R.J., J.N.; Supervision, J.R, R.J., J.N.; Funding Acquisition, J.R., R.J., J.N.

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. 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.

8. 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.

9. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010 Mar 2;11:R25.

10. 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.

11. 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.

12. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al. Systematic comparative analysis of single cell RNA-sequencing methods. bioRxiv. 2019 May 9;632216.

13. 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.

14. 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.

15. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012 May;40(10):e72.

16. 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.

17. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics. 2011 Dec 17;12(1):480.

18. 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://preview.data.humancellatlas.org/

19. 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.

20. 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.

21. 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.

22. 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.

23. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57–74.

24. 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.

25. 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.

26. 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.

27. Blueprint Epigenome Project, 2016. [Internet]. [cited 2019 Mar 4]. Available from: http://dcc.blueprint-epigenome.eu/#/home

28. Pabst C, Bergeron A, Lavallee 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.

29. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016 May;34(5):525–7.

30. 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.

31. 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/

32. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Science & Business Media; 2009. 757 p.

33. Ishwaran H, Kogalur UB. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC), R package version 2.9.2. 2019.

34. 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.

35. 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.

36. 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/

37. 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/

Figure 1

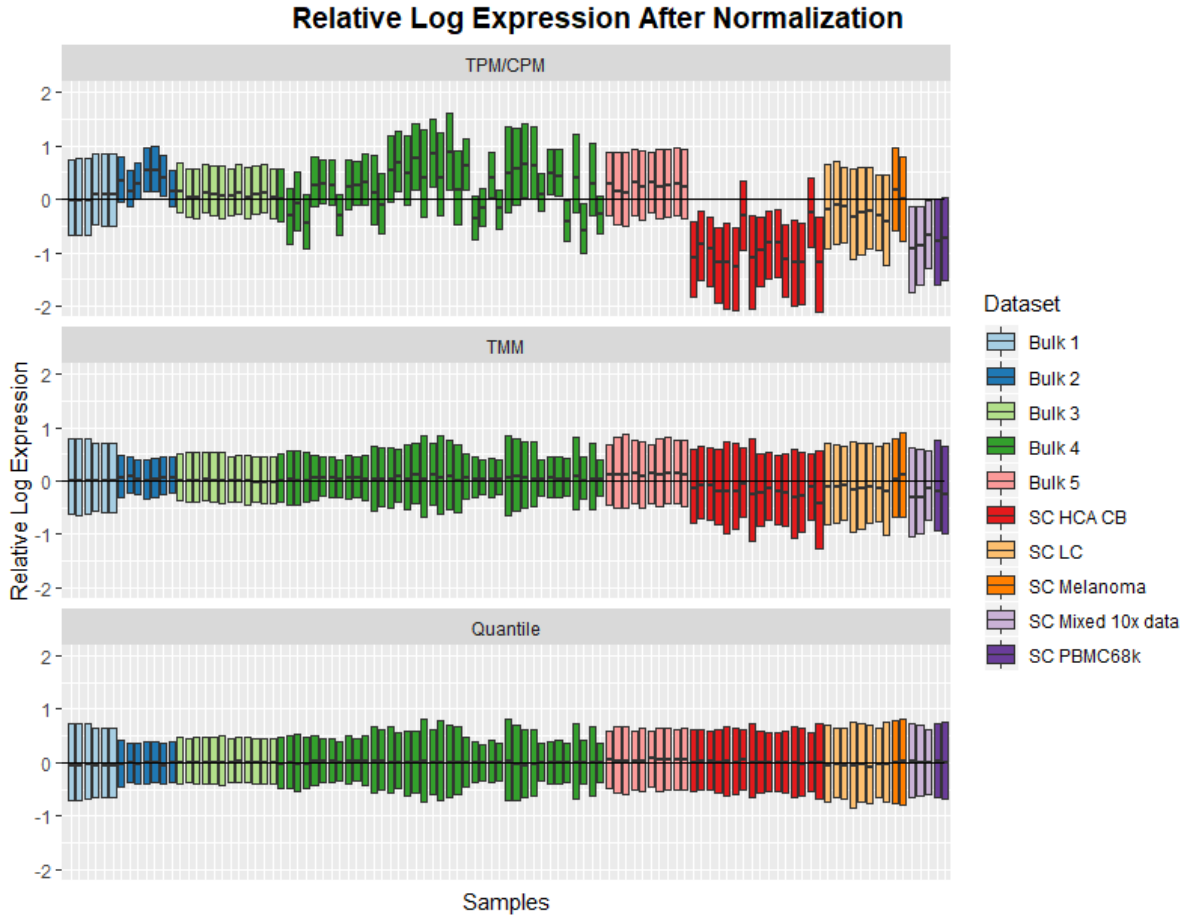

Figure 1 Evaluation of normalization methods for combined single-cell and bulk data. Each bar is a boxplot over all genes describing the log2 fold change in gene expression between the sample and the mean over all samples.

Figure 2

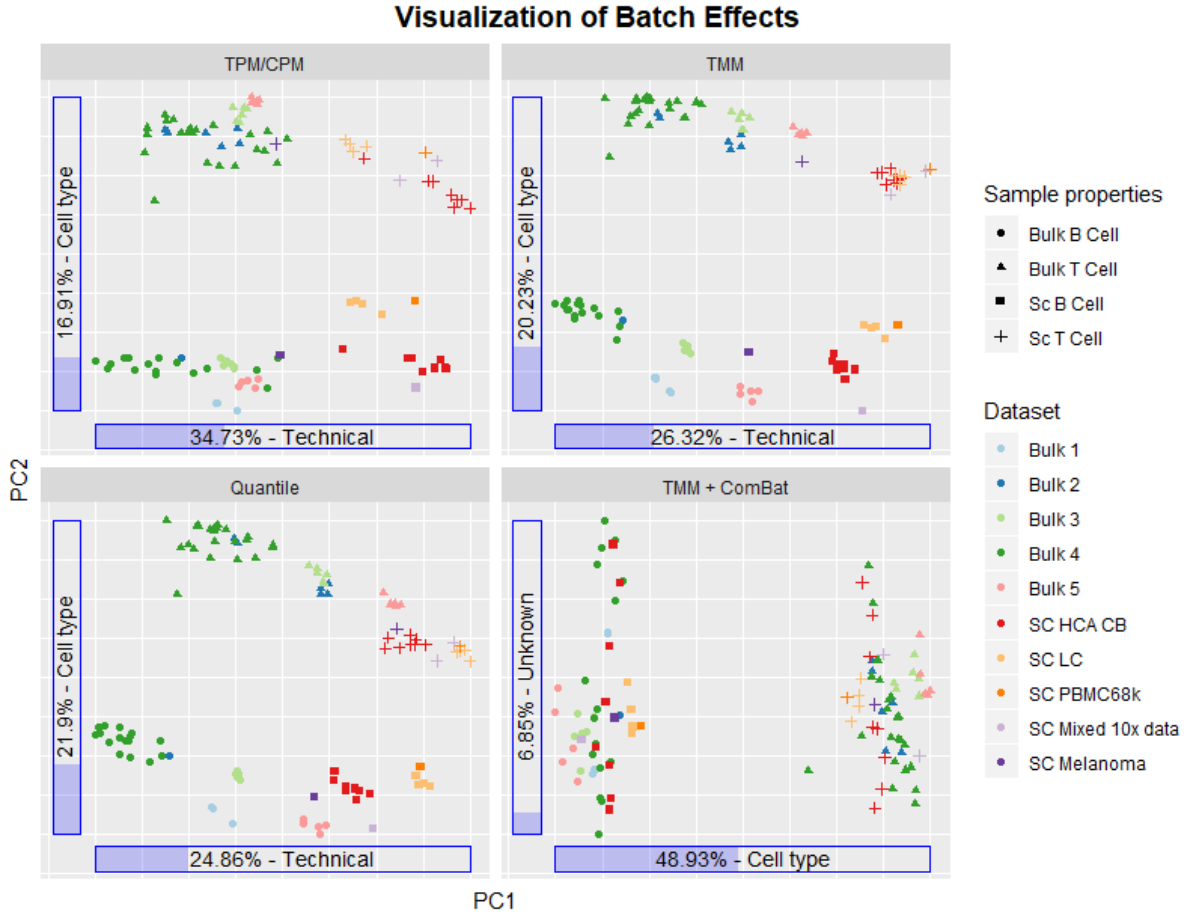

Figure 2 Visualization of normalization and batch effects using PCA.

Figure 3

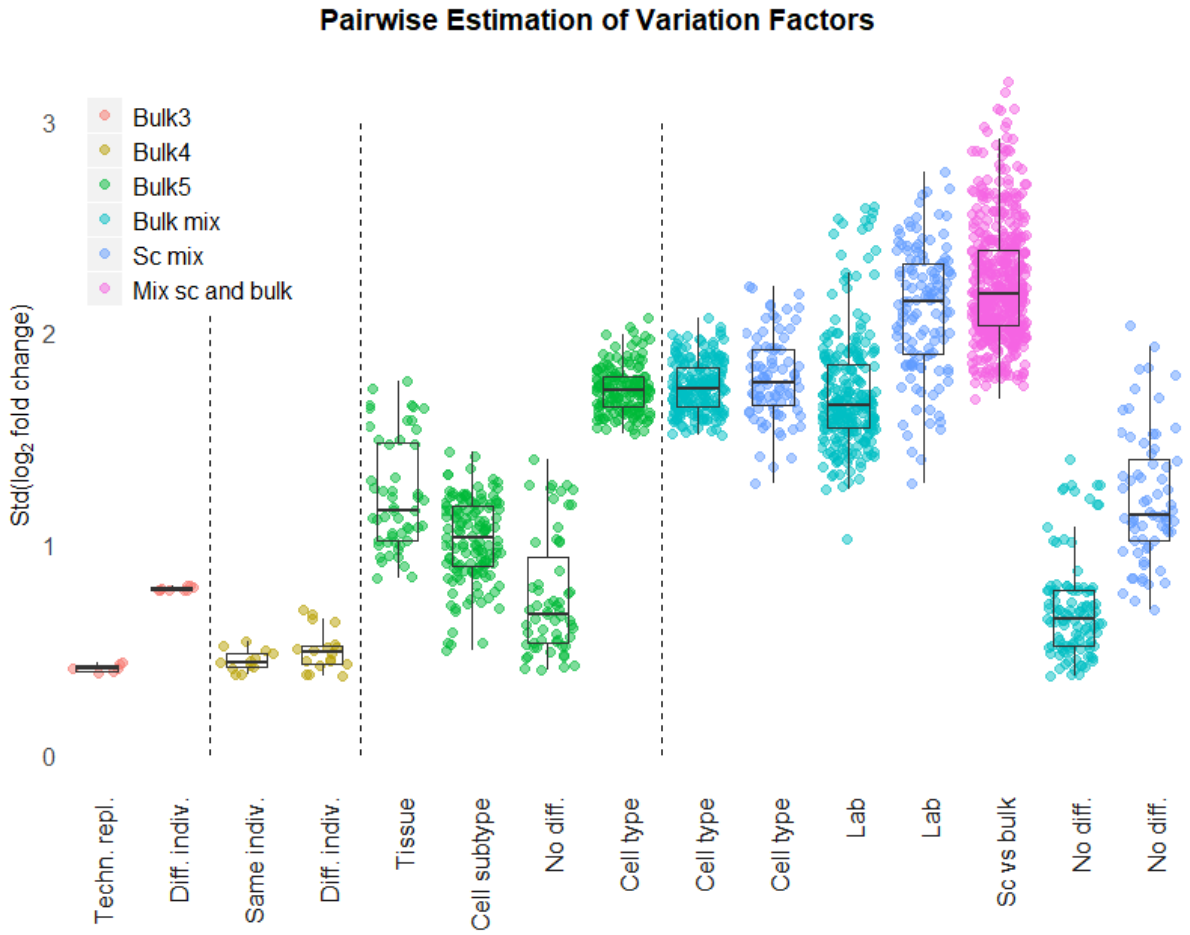

Figure 3 Gene expression difference between sample pairs, measured using the LFCSD method. The technical replicates from lab Bulk3 should be compared to the sample pairs from different individuals from Bulk3. Similarly, the sample pairs from the same individual (but taken at different time points) from Bulk4 should be compared to the pairs from different individuals from the same lab. The factors cell type, cell subtype and tissue of origin from Bulk5 can all be compared with each other and with the pairs where only the individual differs (which represents the case where no factor is present, annotated as "No diff.") from that lab. The sample pairs taken from all bulk or single-cell datasets are generally comparable, although pairs from single-cell data have a larger variation (seen by comparing No diff. for Bulk mix and Sc mix), boosting the difference between such pairs for any factor.

Figure 4

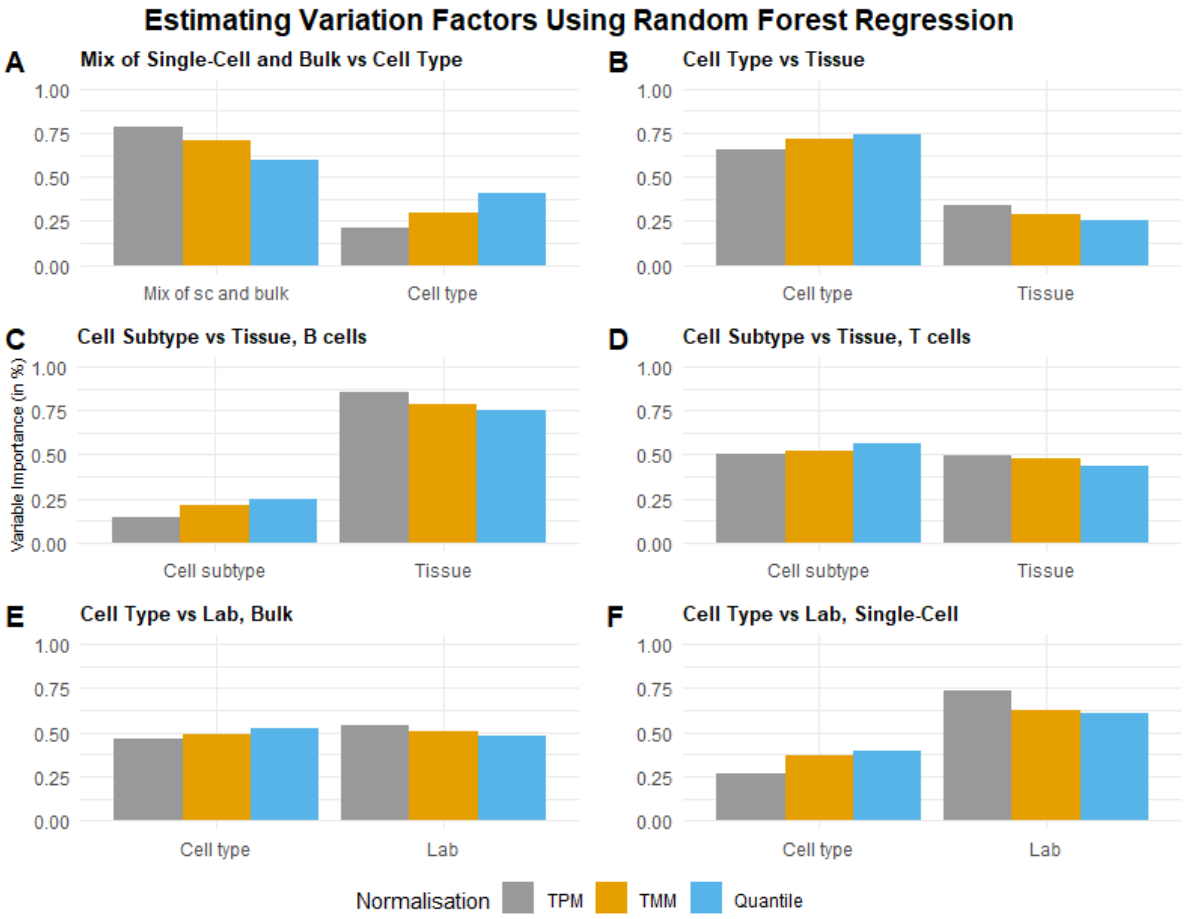

Figure 4 Pairwise comparisons of covariates on the variation between samples as measured by the random forest approach. The importance was computed for three normalizations (bars from left to right), namely TPM (grey), TMM (orange), and Quantile (blue). Importance values are normalized to one in each sub-figure for each normalization technique. A. Comparison of mixing single-cell and bulk vs cell type, using samples from the datasets Bulk5 (only samples from blood) and SC HCA CB. B. Comparison of cell type vs tissue of origin, using samples from the dataset Bulk5. C. Comparison of cell subtype vs tissue, using only B cell samples from the dataset Bulk5. D. Comparison of cell subtype vs tissue of origin, using only T cell samples from the dataset Bulk5. E. Comparison of cell type vs lab using all bulk samples from blood. F. Comparison of cell type vs lab using all single-cell samples from blood.

Figure 5

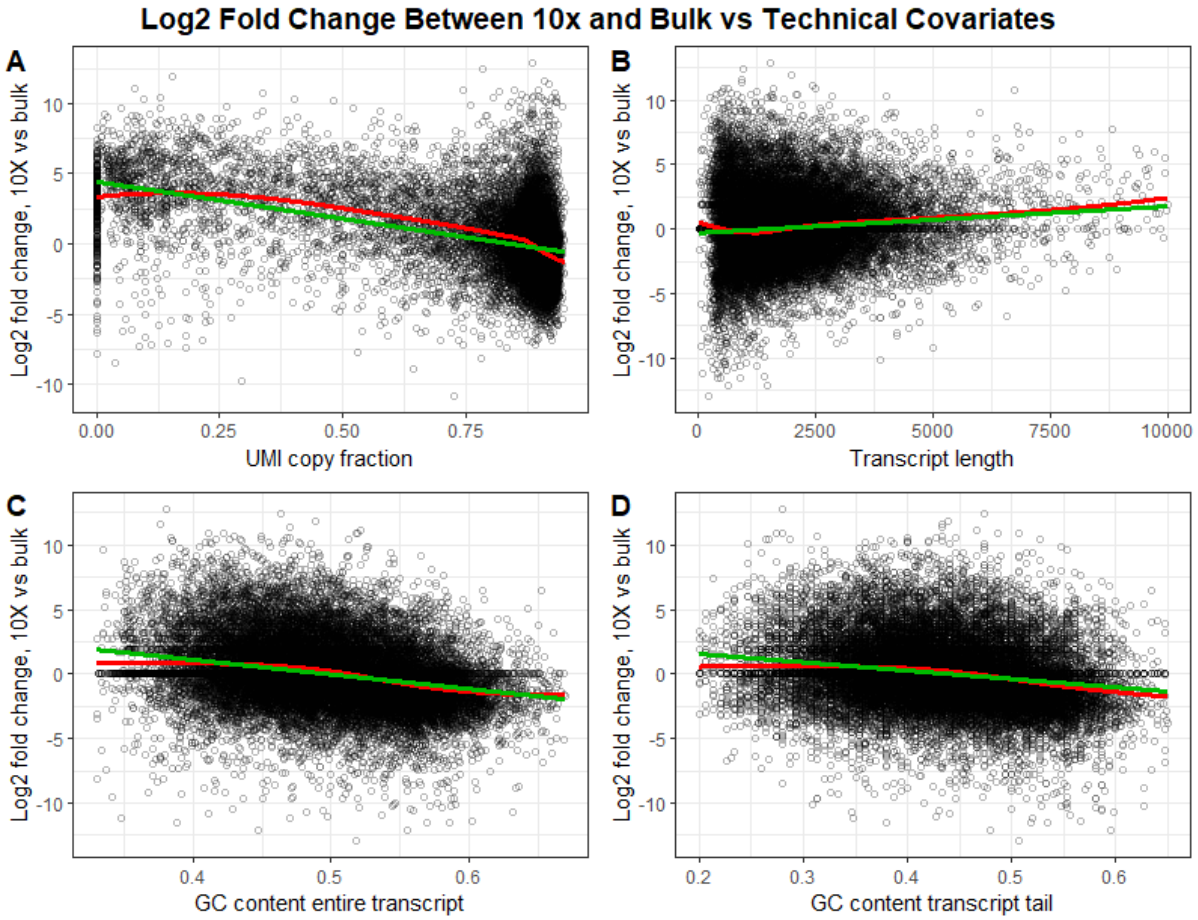

Figure 5. Log2 fold change between 10x data and bulk for each gene, plotted as a function of different covariates. The green line represents a linear fit, whereas the red line shows a Loess fit. The data shown is from cortex 1 of the EVAL dataset. A. UMI copy fraction. Only genes with more than 5 molecules 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.

Figure 6

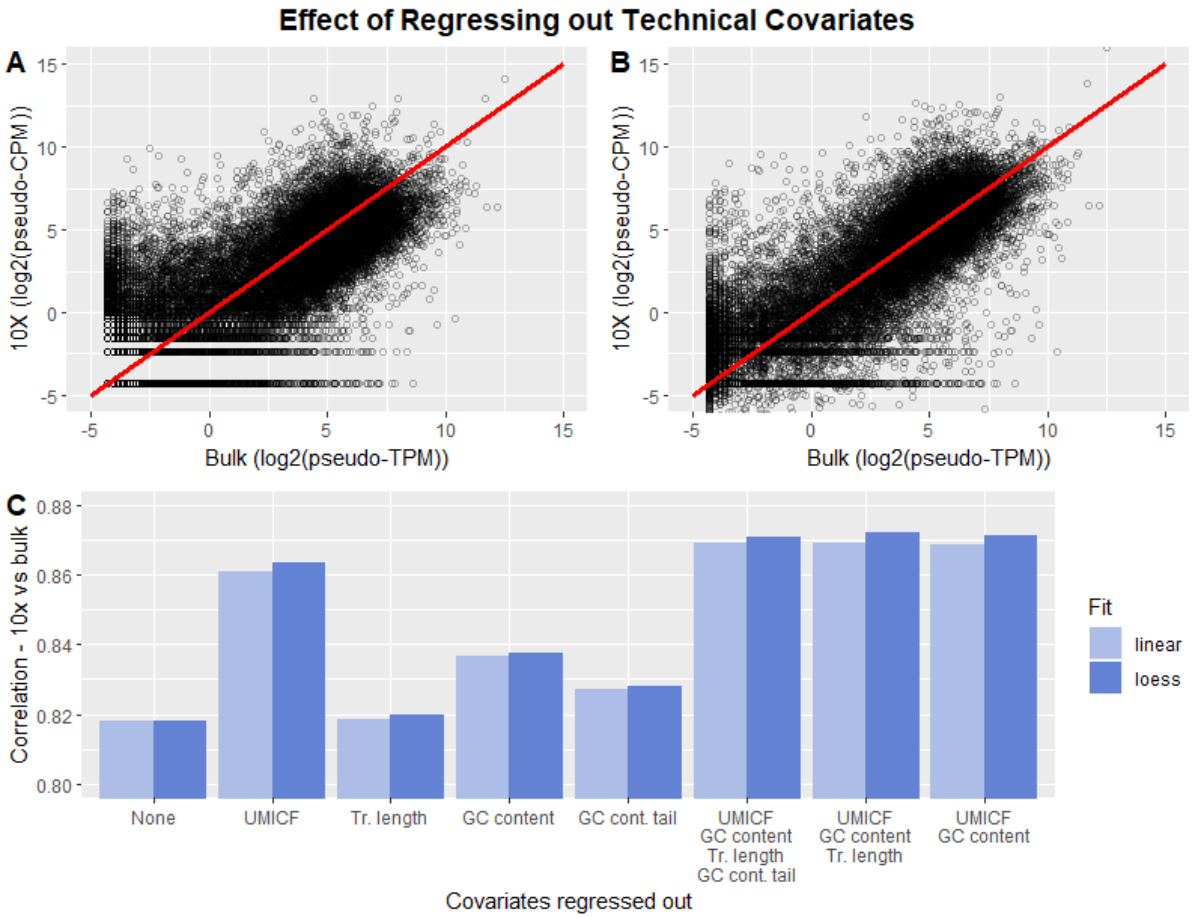

Figure 6 A. Gene expression for cortex 1 from the EVAL dataset plotted as 10x vs bulk. The red line represents a perfect correlation. B. Gene expression for cortex 1 from the EVAL dataset after regressing out the differences in UMICF and GC content between 10x and bulk using a loess fit, which improves 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.

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(18), Rozenblatt-Rosen et al(19)
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(20).
PBMC68k	~68,000 PBMCs from blood, one patient.	Single cell, 10x genomics, UMI counts.	Zheng G.X.Y. et al(3).
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).
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).
TCD8	~10,000 FACS-sorted CD8+ T cells from the blood of a single patient.	Single cell, 10x genomics, UMI counts.	Chen et al(21).
MEL	~4,600 cells from the tumor microenvironment of Melanoma, 19 patients.	Single cell, SMART-Seq2, TPM	Tirosh et al(22).
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(12)
BULK 1	In total 6 bulk samples from B cells of varying origin.	Bulk RNA-Seq, FASTQ files	The ENCODE Consortium(23,24), Gingeras.
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(23,24), Stamatoyannopoulos and Weng.
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)(25,26)
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(27)
BULK 5	In total 10 PBMC bulk samples from B cells (5) and T cells (5).	Bulk RNA-Seq, RPKM/counts	Pabst et al(28), GSE 51984.

Table 2

Table 2 Dataset access information

ID	Source
HCA CB	Li et al(18), Rozenblatt-Rosen et al(18). The data can be downloaded from https://data.humancellatlas.org/, Census of immune cells.
LC	Lambrechts et al(20). The data is available in in ArrayExpress under accessions E-MTAB-6149 and E-MTAB-6653.
PBMC68k	Zheng G.X.Y. et al(3). The data is available at 10x Genomics' home page.
B10k	Zheng G.X.Y. et al(3). The data is available at 10x Genomics' home page.
CD4TMEM	Zheng G.X.Y. et al(3). The data is available at 10x Genomics' home page.
TCD8	Chen et al(21). The data is available for download on GEO data repository, accession number GSE 112845.
MEL	Tirosh et al(22). The data is available for download on GEO data repository, accession number GSE 72056.
EVAL	Ding et al(12). The data is available for download at the Single Cell Portal, id SCP425.
BULK 1, BULK 2	The ENCODE Consortium(23,24). The samples can be downloaded individually from ENCODE.
BULK 3	The functional annotation of the mammalian genome 5 (FANTOM5)(25,26). The data can be downloaded from FANTOM5.
BULK 4	The BLUEPRINT Epigenome Project(27). The samples can be downloaded individually from BLUEPRINT.
BULK 5	Pabst et al(28), GSE 51984.

Figure Legends

Figure 1 Evaluation of normalization methods for combined single-cell and bulk data. Each bar is a boxplot over all genes describing the log2 fold change in gene expression between the sample and the mean over all samples. 15
Figure 2 Visualization of normalization and batch effects using PCA. 16
Figure 3 Gene expression difference between sample pairs, measured using the LFCSD method. The technical replicates from lab Bulk3 should be compared to the sample pairs from different individuals from Bulk3. Similarly, the sample pairs from the same individual (but taken at different time points) from Bulk4 should be compared to the pairs from different individuals from the same lab. The factors cell type, cell subtype and tissue of origin from Bulk5 can all be compared with each other and with the pairs where only the individual differs (which represents the case where no factor is present, annotated as “No diff.”) from that lab. The sample pairs taken from all bulk or single-cell datasets are generally comparable, although pairs from single-cell data have a larger variation (seen by comparing Bulk diff. indiv. to Sc diff. indiv.), boosting the difference between such pairs for any factor. 17
Figure 4 Pairwise comparisons of covariates on the variation between samples as measured by the random forest approach. The importance was computed for three normalizations (bars from left to right), namely TPM (grey), TMM (orange), and Quantile (blue). Importance values are normalized to one in each sub-figure for each normalization technique. A. Comparison of mixing single-cell and bulk vs cell type, using samples from the datasets Bulk5 (only samples from blood) and SC HCA CB. B. Comparison of cell type vs tissue of origin, using samples from the dataset Bulk5. C. Comparison of cell subtype vs tissue, using only B cell samples from the dataset Bulk5. D. Comparison of cell subtype vs tissue of origin, using only T cell samples from the dataset Bulk5. E. Comparison of cell type vs lab using all bulk samples from blood. F. Comparison of cell type vs lab using all single-cell samples from blood. 18
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. 19
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. 20

Supplementary Information Legends

Supporting Text S1. Supplementary methods and figures.

Table S1. Sample Information