SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data

Matthew D. Young[*1] **and Sam Behjati**[*1,2,3]

[1] *Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK*
[2] *Cambridge University Hospitals NHS Foundation Trust, Cambridge, CB2 0QQ, UK*
[3] *Department of Paediatrics, University of Cambridge, Cambridge, CB2 0QQ, UK*

April 18, 2018

Droplet based single cell RNA sequence analyses assume all acquired RNAs are endogenous to cells. However, any cell free RNAs contained within the input solution are also captured by these assays. This sequencing of cell free RNA constitutes a background contamination that has the potential to confound the correct biological interpretation of single cell transcriptomic data. Here, we demonstrate that contamination from this "soup" of cell free RNAs is ubiquitous, experiment specific in its composition and magnitude, and can lead to erroneous biological conclusions. We present a method, SoupX, for quantifying the extent of the contamination and estimating "background corrected", cell expression profiles that can be integrated with existing downstream analysis tools. We apply this method to two data-sets and show that the application of this method reduces batch effects, strengthens cell-specific quality control and improves biological interpretation.

1 Main

Droplet based single cell RNA sequencing has enabled quantification of the transcriptomes of hundreds of thousands of cells in single experiments (Zilionis et al., 2016; Zheng et al., 2017). This technology underpins recent advances in understanding normal and pathological cell behaviour (Hashimoto et al., 2017; Bach et al., 2017; Daniszewski et al., 2018; Stephenson et al., 2018; Chen et al., 2017; Alberti-Servera et al., 2017). Large scale efforts to create a "Human Cell Atlas" also critically depend on the accuracy and cell specificity of the transcriptional readout produced by droplet based single cell RNA sequencing (Regev et al., 2017; Rozenblatt-Rosen et al., 2017).

A core assumption of all droplet based scRNA-seq is that each droplet, within which tagging and reverse transcription takes place, contains mRNA from a single cell. If this assumption is violated it may distort biological interpretation of mRNA sequencing data. Doublets, where a droplet contains multiple cells, and empty droplets are the most obvious example of this. Attempts to detect and to remove doublets are an active area of research (Gayoso, Shor, and Brand, 2017; Ilicic et al., 2016).

Another way in which non-endogenous mRNAs can contaminate a droplet is via the sequencing of cell free RNA admixed with cells in the input solution. It has been recognised that such non-endogenous RNAs are present in even the most ideal data sets (Zheng et al., 2017). No systematic effort has been made to quantify, and compensate for, their contribution. That is, the strategy for correcting for cell free RNA has been to assume that their contribution is negligible.

Here, we show that this "soup" of cell free mRNAs

is ubiquitous, experiment specific and can lead to misleading and inaccurate, and thus misleading biological interpretation. We present a SoupX, a method for quantifying the extent of soup contamination whilst deconvoluting the true, cell specific, signal from the observed mixture of cellular and exogenous mRNAs.

To understand the nature of cell free mRNAs in scRNA-seq and validate our method we have used two data sets. The first we refer to as the "discovery" data set consists of Chromium 10X sequencing of an input solution containing one mouse and one human cell line, mixed in roughly equal proportions (Zheng et al., 2017). This allows us to directly and unambiguously identify which mRNAs are specific to the cell within a droplet and which are cell free contamination. The second, real-life data set, which we refer to as "validation", consists of 9 independent tumour biopsies from 7 patients with the most common types of kidney cancer: Wilms' tumour, clear cell renal cell carcinoma (ccRCC) and papillary renal cell carcinoma (pRCC) (see Table S1).

Counts of observed unique molecular identifiers (UMIs) per gene in each cell are the basic output produced by droplet based single cell RNA-seq. These raw counts are then used to infer the fraction of expression derived each gene in each cell, either by library size normalisation (Satija et al., 2015; Butler and Satija, 2017; Wolf, Angerer, and Theis, 2018) or the inclusion of an offset term in a generalised linear model (**monocle**). We refer to these normalised expression values as a cell's "expression profile". If background contamination is present, the measured expression profile represents a mixture of the true expression profile of the cell, mixed with the expression profile of the background contamination (see Figure 1a and Equation 2).

Our method aims to infer each cell's true expression profile by removing the contribution from the cell free mRNA "soup" expression profile. This is done in three steps:

1. Estimate the soup expression profile from empty droplets.
2. Measure the contamination fraction, the fraction of UMIs originating from background, in each cell.
3. Calculate the cell specific expression profile from the observed mixture of cellular and soup expression.

We include code to input the resulting cell specific expression profiles into popular downstream analysis packages (Seurat Satija et al., 2015; Butler and Satija, 2017 and SCANPY Wolf, Angerer, and Theis, 2018).

Figure 1b shows the relative abundance of human and mouse mRNAs in each droplet in the discovery data set. The groups of droplets on the right of this plot, which represent droplets containing human (top) and mouse (bottom) cells, show that roughly 1% of observed transcripts in droplets containing mouse cells come from human genes (and visa versa). This demonstrates that cell free mRNA contamination is present even in highly controlled experiments.

To investigate the composition of the cell free mRNAs we compared the expression profile derived by aggregating all droplets containing cells to all droplets containing fewer than 10 UMIs (assumed to contain only background mRNAs). We found these two profiles to be highly correlated in both the discovery (correlation coefficient 0.99; Figure 1c) and validation (correlation coefficient 0.74 to 0.98; median 0.93; Figure S1) data sets . This correlation implies that cell free contamination is derived from the input solution, channel specific and must be corrected independently for each channel. By estimating the background expression profile from different groups of cells in the discover data set (where we know a priori which genes are contamination in each droplet) we found that the expression profile of the cell free contamination is invariant for cells derived from the same channel (Figure S3).

Next we estimated the contamination fraction, the fraction of expression derived from cell free mRNA background in each cell, using genes which we could assume to be unexpressed in most cells. Without spike-ins (which are never expressed by a cell) an appropriate set of genes known to be unexpressed in specific cell types must be determined. Such genes will have a characteristic bimodal expression pattern (Figure S7) with each cell either expressing the gene in abundance or only containing cell free RNA derived copies. To prevent over estimating the contamination fraction, we algorithmically identify likely candidates from which the most appropriate genes are selected using prior biological knowledge (see Methods). For the discovery data set, we use all mouse transcripts to estimate the contamination fraction from human cells and visa versa. For the validation data set, we use the haemoglobin genes and exclude any red bloods cells from the estimation.

The distribution of the contamination fraction revealed that the greatest contamination occurs in droplets with the lowest number of UMIs (Figure 1d). This trend for increasing contamination with decreasing number of UMIs per droplet was found across data sets (Figure S4) and is consistent with an approximately constant number of cell free mRNAs in each droplet, with the contamination fraction being determined by the number of cell endogenous molecules present. Furthermore, we found a large range in sample averaged contamination fraction (1% to 50%, median 8%) with the highest contamination found in the most necrotic samples (as determined by clinical pathological evaluation).

To demonstrate the biological utility of our method, we analysed both data sets with and without correction for cell free RNA. Our analysis consisted of normalisation, feature selection, dimension reduction, clustering and marker gene detection using the Seurat package (Satija et al., 2015; Butler and Satija, 2017).

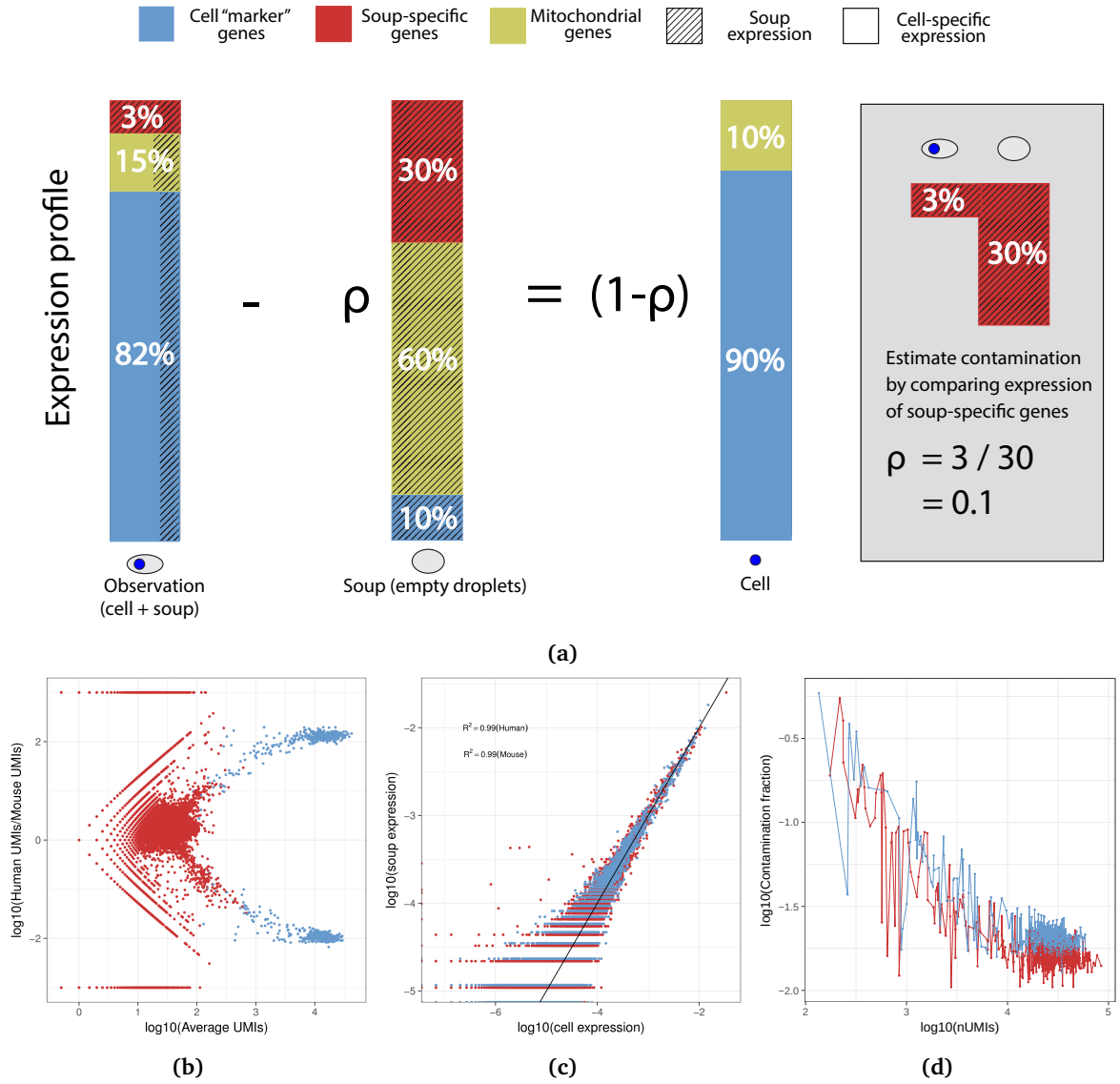

Figure 1: *The properties of the cell free mRNA soup and an overview of the SoupX method. Panel a shows a schematic overview of our method for removing background contamination. The three bars represent the relative mRNA composition of the observed data (left), the background (middle) and the cell alone (right) and their relationship to one another (see Equation 2). The box on the right shows a toy example demonstrating how the background contamination fraction is estimated. Panel b shows the* $\log_{10}$ *ratio of the number of UMIs mapping to human and mouse mRNAs for each droplet in the discovery data set. Droplets that are determined to contain cells are marked in blue. Panel c shows the correlation of the* $\log_{10}$ *fraction of expression from all human (or mouse) cells comapred to* $\log_{10}$ *fraction in soup for all human (red) or mouse (blue) transcripts. Panel d shows the estimated contamination fraction as a function of number of UMIs in each droplet in individual cells in the discovery data set. Red (Blue) dots denote droplets containing human (mouse) cells.*

As the background expression profile is experiment specific, we expected that one of the effects of background contamination would be to increase batch effects. That is, two identical cells captured in different experiments will appear different due to differences in their cell free RNA composition. Consistent with this interpretation, we found that background correction decreased batch effects in our validation data set (Figure 2a and Figure S5).

It is common practice to filter scRNA-seq data to exclude droplets where the fraction of mitochondrial gene expression exceeds some value, typically in the range 5%-20% (Daniszewski et al., 2018; Chen et al., 2017; Bach et al., 2017). We found that the fraction of mitochondrial gene expression in the cell free background varied considerably (2% to 15%, median 7%). Figure 2b shows the consequences of this: the uncorrected mitochondrial expression fraction can over or under estimate the endogenous expression by as much as 5%, leading some cells to erroneously pass/fail this quality control filter.

Furthermore, we found that the estimated contami-

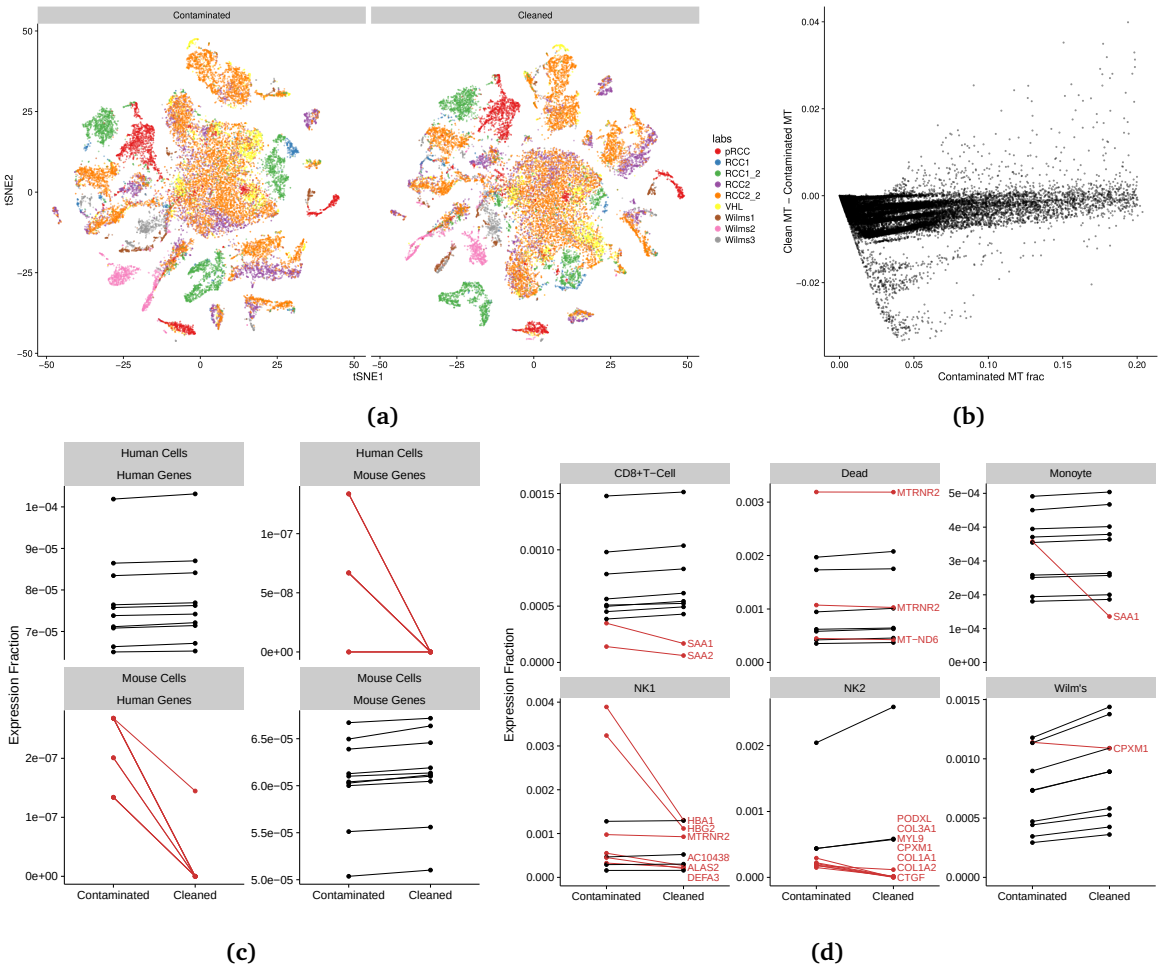

Figure 2: *The effect of removing cell free RNA on QC, downstream analyses and biological interpretation. Panel a shows tSNE plots of the validation data set with and without background correction. Panel b shows how the MT fraction changes with background correction for each droplet in the validation data. Panels c and d show the change in expression fraction for the top 10 cluster specific marker genes after background correction for the discovery and validation data sets, respectively. In Panel d only those clusters where at least one marker gene decreases in expression are shown. Those genes that decrease in expression are labelled in red.*

nation fraction can itself be used as an additional quality control (QC) filter. In the discovery data, where the contamination fraction can be reliably estimated in individual droplets, we designated droplets with significantly less than 10% background expression as containing cells (p-value < 0.01; binomial test against null of $\rho \leq 0.1$). We found a significant overlap between droplets identified in this way and those identified using state-of-the-art cell detection method EmptyDrops (Lun T. L. et al., 2017) (1072 in both, 8 found only using contamination fraction filtering, 64 found only by EmptyDrops; p < 0.001; hypergeometric test).

Unless spike-ins have been used, this approach is unlikely to be useful for cell specific QC in other data sets. However, channel level estimates of the contamination fraction can still provide useful information about the quality of an experiment. For example, the Wilm's tumour channels were derived from tissue that has already undergone chemotherapy and are highly necrotic, and perform poorly in terms of cell yield per channel. We found these channels to have amongst the highest contamination fraction in our validation data set.

To test the effect that background correction has on the biological interpretation of clusters of cells, we calculated the change in expression fraction for the top 10 cluster specific genes for each cluster in the discovery and validation data sets. Figure 2c shows the results for the discovery data set and reveals that the effect of background correction is to increase the ranking of human transcripts in human cells (by a small amount) and decrease the ranking of mouse transcripts (by a large amount) and vice versa. After background correction only 0.1% percent of expression was assigned to human/mouse transcripts in mouse/human cells, respectively.

In the validation data set, correcting for background contamination also increased the expression of cluster specific markers in most circumstances (see Figure S6). However, there were 6/32 clusters (shown in Figure

2d) for which the expression of cluster markers was decreased for at least one gene. These clusters highlight instances where failure to correct for background correction confounds the biological interpretation.

For example, SAA1 expression is elevated in inflammation and is known to be secreted by macrophages (Meek, Eriksen, and Benditt, 1992). Its expression has also been reported in tumour cells of the RCCs with poor prognosis (Paret et al., 2010). As such, it is plausible that it is being secreted by the macrophages and cytotoxic T-cells in these tumours, as the uncorrected expression in Figure 2d suggests. However, correcting for background contamination shows that SAA1 is predominately expressed by RCC cells (Cluster 14, Figure S6) and only appears expressed in these T-cells and macrophages due to contamination. That is, it's expression is significantly decreased in these clusters by background correction. As another example, the cluster of natural killer cells marked NK2 in Figure 2d expresses the collagen genes COL1A1, COL1A2 and COL3A1 before background correction. This could be interpreted as evidence for their tissue residency. However, background correction completely removes all expression of these genes in this cluster. The other clusters for which expression of marker genes decrease are unlikely to be misinterpreted biologically, but still highlight that the failure to correct for background contamination can falsely identify genes as being marker genes for a cluster of cells.

We have shown that cell free RNA is ubiquitously present in droplet based single cell RNA-seq data and proposed a method to identify, quantify and remove its contaminating effect. We find that estimating contamination can itself be used as a QC measure and that correcting for contamination reduces batch effects, makes existing QC measures more accurate and improves biological interpretation.

The size of the correction will depend on the nature of the experiment, with solid tissues, particularly highly necrotic samples or those requiring extensive processing, likely to result in the highest contamination fractions. Estimating the contamination fraction relies on the presence of genes that can be unambiguously identified as contaminated expression. In solid tissues, haemoglobin genes perform this function perfectly, making our method particularly suited to application to experiments on solid tissues.

In plate based protocols, exogenous spike-in RNAs are often included to aid with data normalisation (Picelli et al., 2014; Bacher and Kendziorski, 2016). In contrast, they have not been widely used in droplet based scRNA-seq due to the additional costs and the difficulties controlling input concentration. Although these costs will remain prohibitive in many settings, the utility of being able to accurately identify background contamination rates, both to improve biological interpretation and as a QC measure, provides an additional argument in favour of their inclusion. Furthermore, the concentration of any spike-ins could be estimated directly from the data, a situation which contrasts with the use of spike-ins for plate based experiments, where the utility of spike-ins depends on accurately controlling the spike-in abundance (Lun et al., 2017; Robinson and Oshlack, 2010).

The background correction method we propose can be easily applied to standard data with negligible computational cost relative to downstream analyses. The corrected expression values it produces can be easily incorporated with essentially no modification into any analyses based on library size normalised data (such as Seurat Satija et al., 2015; Butler and Satija, 2017 or SCANPY Wolf, Angerer, and Theis, 2018). Integration with count based analyses (such as the "simple single cell workflow" Lun, McCarthy, and Marioni, 2016) will require larger changes as explicit calculation of the expression profile for a cell destroys the mean-variance relationship of the data. We provide an R package, SoupX, which can be used to estimate the background composition and contamination rate, remove this contamination from droplets containing cells and input the corrected expression values into downstream analysis tools. We envision background correction forming a standard part of droplet based single cell RNA-seq analyses pipelines.

2 Acknowledgments

We thank William Heaton and Valentine Svensson for discussions about droplet sequencing; Sarah Teichmann for discussions about the methodology; Sarah Teichmann, Aaron Lun and Manasa Ramakrishna for comments and review of the manuscript and Manasa Ramakrishna for improvements to the figures and their layout.

3 Author contribution

M.D.Y. conceived the project, developed the method and wrote the manuscript. S.B. supervised the project.

4 Data availability

The data set referred to as the "discover data set" was the mixture of the human cell line 293T and mouse cell line 3T3 described in Zheng et al., 2017. We used the data mapped and quantified using Cell Ranger 1.1.0 from `https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/293t_3t3`.

Raw sequencing data for the validation data set have been deposited in the European Genome-phenome Archive (EGA) under study ID EGAS00001002325. Sample specific identifiers can be found in Table S1.

5 Code availability

An implementation of the method described here, which we call SoupX, as well as tools to integrate the results with downstream analysis tools, can be found here `https://github.com/constantAmateur/SoupX`.

6 Supplementary Materials

6.1 Processing of data sets

For both data sets, we used the EmptyDrops (`https://github.com/TimothyTickle/hca-jamboree-cellidentification/tree/master/src/poisson_model`) method to identify droplets likely to contain cells using a FDR cut-off of 0.05. For the discovery data set we removed any droplet that contained 1000 or more UMIs from both human and mouse genes as being likely doublets. For the validation data set we further excluded any droplet that contained UMIs from fewer than 50 distinct genes.

Raw counts of UMIs per gene in each cell were then normalised using the Seurat (`http://satijalab.org/seurat/`) "NormalizeData" function. This function transforms the data to,

$$x_{gc} = \log(1 + 10,000 F_{gc}) \tag{1}$$

where x_{gc} is the normalised value. To use the decontaminated expression profiles in the downstream analysis, it is sufficient to replace F_{gc} with f_{gc} in Equation 1.

Variable genes were identified using the FindVariableGenes function with default parameters. We then calculate the first 30 principle components using the normalised data values for the variable genes scaled to have mean 0 and standard deviation 1. tSNE embedding was calculated using a perplexity of 30 and clusters tere identified using the community identification algorithm implemented in the "FindClusters" function with a resolution parameter of 1.

To identify genes specific to each cluster, we ranked genes using an adaptation of the "tf-idf" metric widely used in natural language processing (Rajaraman and Ullman, 2011). Specifically, we ranked genes in each cluster by $\lambda_{gc} log(\beta_g)$ where λ_{gc} is the fraction of cells in cluster c expressing gene g and β_g is the fraction of all cells expressing g. We then took the top 10 genes in this list for each cluster, or all genes for which the genes was present in a cluster more than expected if expressed genes were randomly distributed at a p-value cut-off of 0.01 using a hypergeometric test.

These markers were then manually inspected and each cluster was assigned a cell type based on the comparison of these markers to the literature (particularly Chabardès-Garonne et al., 2003; Habuka et al., 2014; Lee, Chou, and Knepper, 2015).

6.2 Detailed description of the SoupX method

The number of observed unique molecular identifiers (UMIs) for gene g in cell c depends on the abundance of gene g relative to all other genes. That is, by sequencing transcripts we aim to infer the fraction of expression for each gene f_{gc}, where $\sum_g f_{gc} = 1$. The relationship between the observed abundance for a gene g in a single cell c and the true distribution of that cell are given by:

$$F_{gc} = \rho_c f_{gs} + (1 - \rho_c) f_{gc} \tag{2}$$

where F_{gc} is the observed proportion of UMIs in droplet c, gene g, f_{gc} is the true proportion for the cell contained within this droplet and f_{gs} is the fraction of UMIs from gene g in the cell free background. ρ_c is the fraction of UMIs in droplet c that originate from the soup (i.e, the contamination fraction).

The method for quantifying the composition and abundance of cell free mRNAs and correcting cell specific mRNA expression profiles for their presence is a three part procedure. This consists of:

- Calculating an expression profile for the cell free background.
- Estimating the contamination fraction for each cell.
- Removing contamination from each cell using the above information.

6.2.1 Background expression profiles

To calculate the expression profile of cell free mRNAs, we assume that those droplets with a very low number of UMIs contain only cell free mRNAs and no cell. As the number of droplets with low numbers of UMIs is very large compared to numbers of cells ($\sim 1,000,000$ droplets versus $\sim 10,000$ cells) there is typically abundant power to accurately calculate the expression profile of the soup. We estimate the background expression for gene g as,

$$f_{gs} = \frac{n_{gs}}{\sum_g n_{gs}} \tag{3}$$

where n_{gs} is the number of UMIs mapping to gene g in those droplets with total UMIs in the range $a < UMIs < b$. We set $a = 1$ to exclude errors in the droplet barcodes contaminating our estimate of the background (although we find no evidence this is a problem for chromium 10X data). We set $b = 10$ based on comparisons of the background profile estimated from droplets with different numbers of UMIs to the true background profile measured directly from droplet containing cells in the discovery data set (see Figure S2).

6.2.2 Estimating the contamination fraction

To estimate the contamination fraction, we observe that when $f_{gc} = 0$ equation 2 reduces to,

$$F_{gc} = \rho_c f_{gs} \tag{4}$$

The maximum likelihood estimator (assuming a Poisson distribution) for ρ_c from a single cell is then just,

$$\rho_c = \frac{\sum_g F_{gc}}{\sum_g f_{gs}} \tag{5}$$

where the sum is taken over all genes for which it can be assumed that $f_{gc} = 0$.

Unfortunately, there is often insufficient power to accurately estimate ρ from an individual cell. To overcome this limitation, we assume that cells with approximately equal total numbers of UMIs have approximately equal contamination fractions (an assumption justified by the discovery data Figure S3). Different cells will have different genes for which it can be assumed that $f_{gc} = 0$. For example we can safely assume haemoglobin genes have zero expression in macrophages, but not in red blood cells. As such, the maximum likelihood estimator for ρ from a group of cells (again assuming a Poisson distribution) is

$$\rho = \frac{\sum_c \sum_g n_{gc}}{\sum_c (N_c \sum_g f_{gs})} \tag{6}$$

where $N_c = \sum_g n_{gc}$ is the total number of UMIs in cell c. The inner sum over genes is taken over a different set of genes for each cell depending on which genes can be assumed to be soup specific for that cell.

For our validation data set we identified excluded from the estimation of ρ any cell for which we could reject the null hypothesis that $\rho_c \leq 1$ using a Poisson test with p-value cut-off 0.05. That is, we identified and removed those cells for which the haemoglobin gene expression exceeded what we would expect if a droplet contained nothing but cell free mRNAs.

6.2.3 Selecting genes for estimating ρ

To identify which genes to use when estimating the contamination fraction in each channel, we calculated the distribution of each gene's expression across all cells in channel, relative to the expression in the soup. That is, we calculate

$$\frac{F_{gc}}{f_{gs}} \tag{7}$$

for all cells in a channel. To identify those genes with a bimodal pattern of expression across all cells, where one mode represents all cells truly expressing the gene and the other mode all cells where the only expression for the gene comes from the background. Such genes are ideal for estimating ρ as it is clear which cells "seed" the mRNAs in the soup to begin with and which merely incorporate it indirectly. To obtain candidates genes, we exclude any gene that has fewer than 10% of cells with $F_{gc} < f_{gs}$ and then sort genes by their mean squared value of $\log(F_{gc}/f_{gs})$.

An example of this approach is shown in Figure S7 for one channel of data from a pRCC biopsy. This shows that HBB, HBA2 and TPSB2 are the most useful in estimating ρ. The selection of genes likely to be unexpressed genes is also aided by knowledge of the biology of the cells in each sample. In this example, HBB and HBA2 are haemoglobin genes which should either be expressed in abundance in red blood cells or completely absent in other cell types. Likewise, this biopsy contains a large number of MAST cells and TPSB2 is a highly specific marker of MAST cells, which can be assumed to be expressed only in MAST cells.

6.2.4 Correcting cell expression profiles

Having calculated the expression profile for the background f_{gs} and the contamination fraction for each cell ρ_c, we estimate the cell endogenous expression by,

$$f_{gc} = \frac{F_{gc} - \rho_c f_{gs}}{1 - \rho_c} \tag{8}$$

After this correction, values of $f_{gc} < 0$ are set to 0 and the resulting expression profile is renormalised to sum to 1. It is sometimes useful to assume that a group of cells share the same expression profile and estimate this profile using all the cells simultaneously. A typical example of this is estimating an expression profile for a cluster of cells assumed to represent the same cell type. In these cases cells within a cluster may have different values of ρ_c or f_{gs} (when the cluster contains cells from multiple experiments) and we estimate the joint expression profile by numerically maximising the Poisson log-likelihood (we provide R code for this purpose).

Bibliography

Alberti-Servera, Llucia et al. (2017). "Single-cell RNA sequencing reveals developmental heterogeneity among early lymphoid progenitors." In: *The EMBO journal* 36.24, pp. 3619–3633.

Bach, Karsten et al. (2017). "Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing." In: *Nature communications* 8.1, p. 2128.

Bacher, Rhonda and Christina Kendziorski (2016). "Design and computational analysis of single-cell RNA-sequencing experiments." In: *Genome biology* 17.1, p. 63.

Butler, Andrew and Rahul Satija (2017). "Integrated analysis of single cell transcriptomic data across conditions, technologies, and species". In: *bioRxiv*, p. 164889.

Chabardès-Garonne, Danielle et al. (2003). "A panoramic view of gene expression in the human kidney." In: *Proceedings of the National Academy of Sciences of the United States of America* 100.23, pp. 13710–13715.

Chen, Ying-Jiun J et al. (2017). "Single-cell RNA sequencing identifies distinct mouse medial ganglionic eminence cell types." In: *Scientific reports* 7, p. 45656.

Daniszewski, Maciej et al. (2018). "Single cell RNA sequencing of stem cell-derived retinal ganglion cells." In: *Scientific data* 5, p. 180013.

Gayoso, Adama, Jonathan Shor, and Ryan Brand (2017). "DoubletDetection: Identifying Technical Error in Single-cell RNA-sequencing Data". In: `https://github.com/JonathanShor/DoubletDetection/blob/master/docs/DoubletDetection.pdf`. URL: `https://github.com/JonathanShor/DoubletDetection/blob/master/docs/DoubletDetection.pdf`.

Habuka, Masato et al. (2014). "The kidney transcriptome and proteome defined by transcriptomics and antibody-based profiling." In: *PLoS One* 9.12, e116125.

Hashimoto, Shinichi et al. (2017). "Comprehensive single-cell transcriptome analysis reveals heterogeneity in endometrioid adenocarcinoma tissues." In: *Scientific reports* 7.1, p. 14225.

Ilicic, Tomislav et al. (2016). "Classification of low quality cells from single-cell RNA-seq data." In: *Genome biology* 17.1, p. 29.

Lee, Jae Wook, Chung-Lin Chou, and Mark A Knepper (2015). "Deep Sequencing in Microdissected Renal Tubules Identifies Nephron Segment-Specific Transcriptomes." In: *Journal of the American Society of Nephrology : JASN* 26.11, pp. 2669–2677.

Lun, Aaron T L, Davis J McCarthy, and John C Marioni (2016). "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor." In: *F1000Research* 5, p. 2122.

Lun, Aaron T L et al. (2017). "Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data." In: *Genome research* 27.11, pp. 1795–1806.

Lun T. L., Aaron et al. (2017). "On the correct detection of empty droplets in droplet-based single-cell RNA sequencing protocols". In: `https://github.com/HumanCellAtlas/hca-jamboree-cell-identification/blob/master/docs/EmptyDrops_group4_report.pdf`. URL: `https://github.com/HumanCellAtlas/hca-jamboree-cell-identification/blob/master/docs/EmptyDrops_group4_report.pdf`.

Meek, R L, N Eriksen, and E P Benditt (1992). "Murine serum amyloid A3 is a high density apolipoprotein and is secreted by macrophages." In: *Proceedings of the National Academy of Sciences of the United States of America* 89.17, pp. 7949–7952.

Paret, Claudia et al. (2010). "Inflammatory protein serum amyloid A1 marks a subset of conventional renal cell carcinomas with fatal outcome." In: *European urology* 57.5, pp. 859–866.

Picelli, Simone et al. (2014). "Full-length RNA-seq from single cells using Smart-seq2." In: *Nature Protocols* 9.1, pp. 171–181.

Rajaraman, Anand and Jeffrey David Ullman (2011). *Mining of Massive Datasets*. Cambridge University Press.

Regev, Aviv et al. (2017). "The Human Cell Atlas." In: *eLife* 6, p. 503.

Robinson, Mark D and Alicia Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." In: *Genome biology* 11.3, R25.

Rozenblatt-Rosen, Orit et al. (2017). "The Human Cell Atlas: from vision to reality." In: *Nature* 550.7677, pp. 451–453.

Satija, Rahul et al. (2015). "Spatial reconstruction of single-cell gene expression data." In: *Nat Biotechnol* 33.5, pp. 495–502.

Stephenson, William et al. (2018). "Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation." In: *Nature communications* 9.1, p. 791.

Wolf, F Alexander, Philipp Angerer, and Fabian J Theis (2018). "SCANPY: large-scale single-cell gene expression data analysis." In: *Genome biology* 19.1, p. 15.

Zheng, Grace X Y et al. (2017). "Massively parallel digital transcriptional profiling of single cells." In: *Nature communications* 8, p. 14049.

Zilionis, Rapolas et al. (2016). "Single-cell barcoding and sequencing using droplet microfluidics". In: *Nature Protocols* 12.1, pp. 44–73.

Donor (study ID)	Experiment	Age	Replicates	Tumour type
Child1	Wilms1	4 years 2 months	3	Wilms'
Child2	Wilms2	8 months	3	Wilms'
Child3	Wilms3	2 years 6 months	3	Wilms'
Adult1	PapRCC	70	2	Papillary cell carcinoma
Adult2	RCC1	67	4	Clear cell carcinoma
Adult3	RCC2	63	4	Clear cell carcinoma
Adult4	VHL_RCC	49	2	Clear cell carcinoma

Table S1: *Sample information for the validation data set. All replicates are technical replicates (same input solution run down different 10X channels), with the exception of RCC1 and RCC2 where replicates 1 and 2 are from one biopsy and 3 and 4 are from a separate biopsy of the same tumours.*

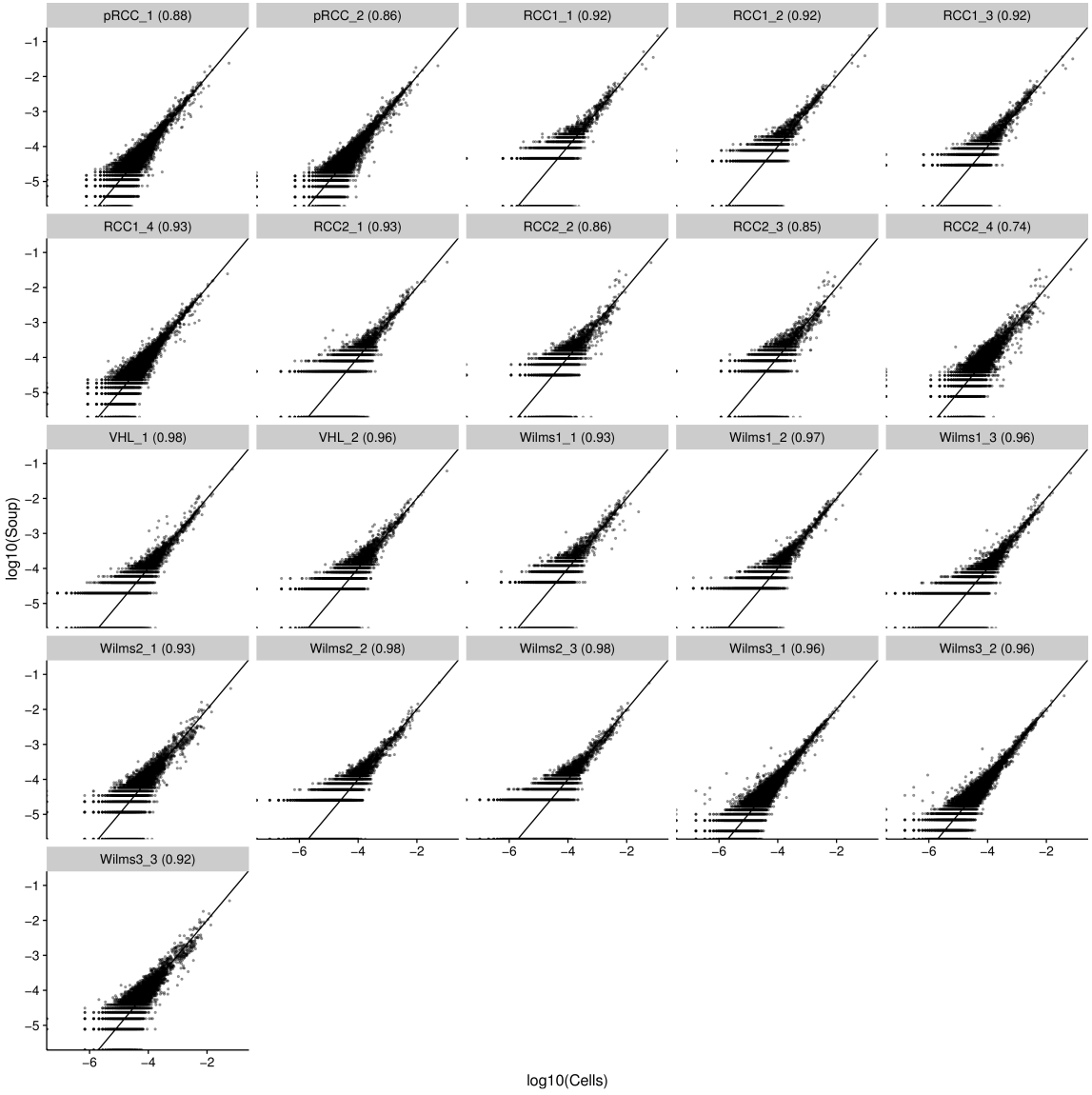

Figure S1: *Correlation of expression profile derived from empty droplets (the soup) with an aggregate expression profile from all cells in a channel. The correlation coefficient is shown in parenthesis in the facet label. The line shows perfect correlation.*

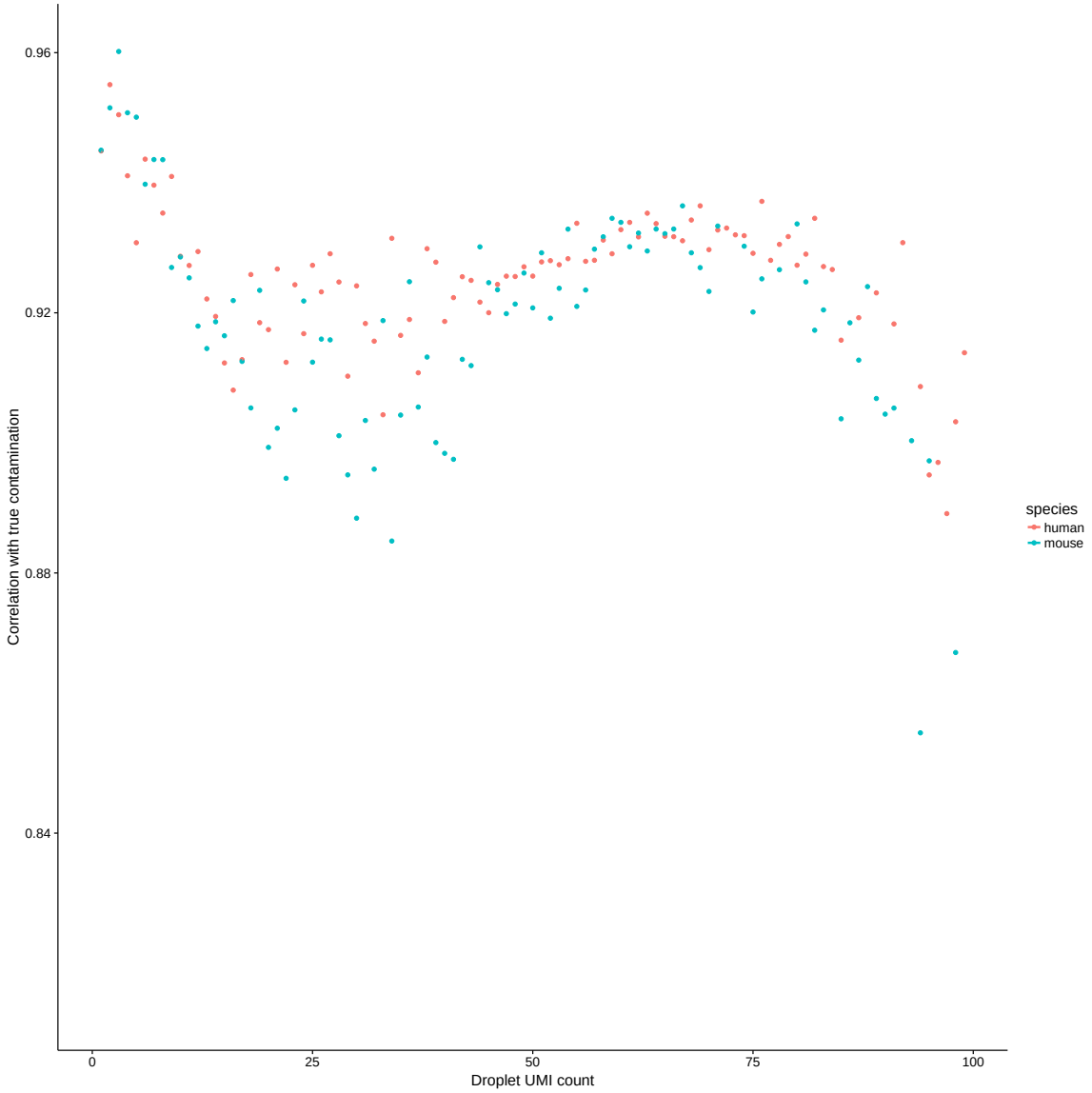

Figure S2: *The correlation between "true background", which is defined by aggregating across mouse transcripts in human cells and visa versa, with the background expression profile derived using only droplets with the total number of UMIs given on the x-axis. This is done independently by comparing the mouse transcripts in the empty droplets with the mouse transcripts in human cells and human transcripts in the empty droplets with human transcripts in mouse cells.*

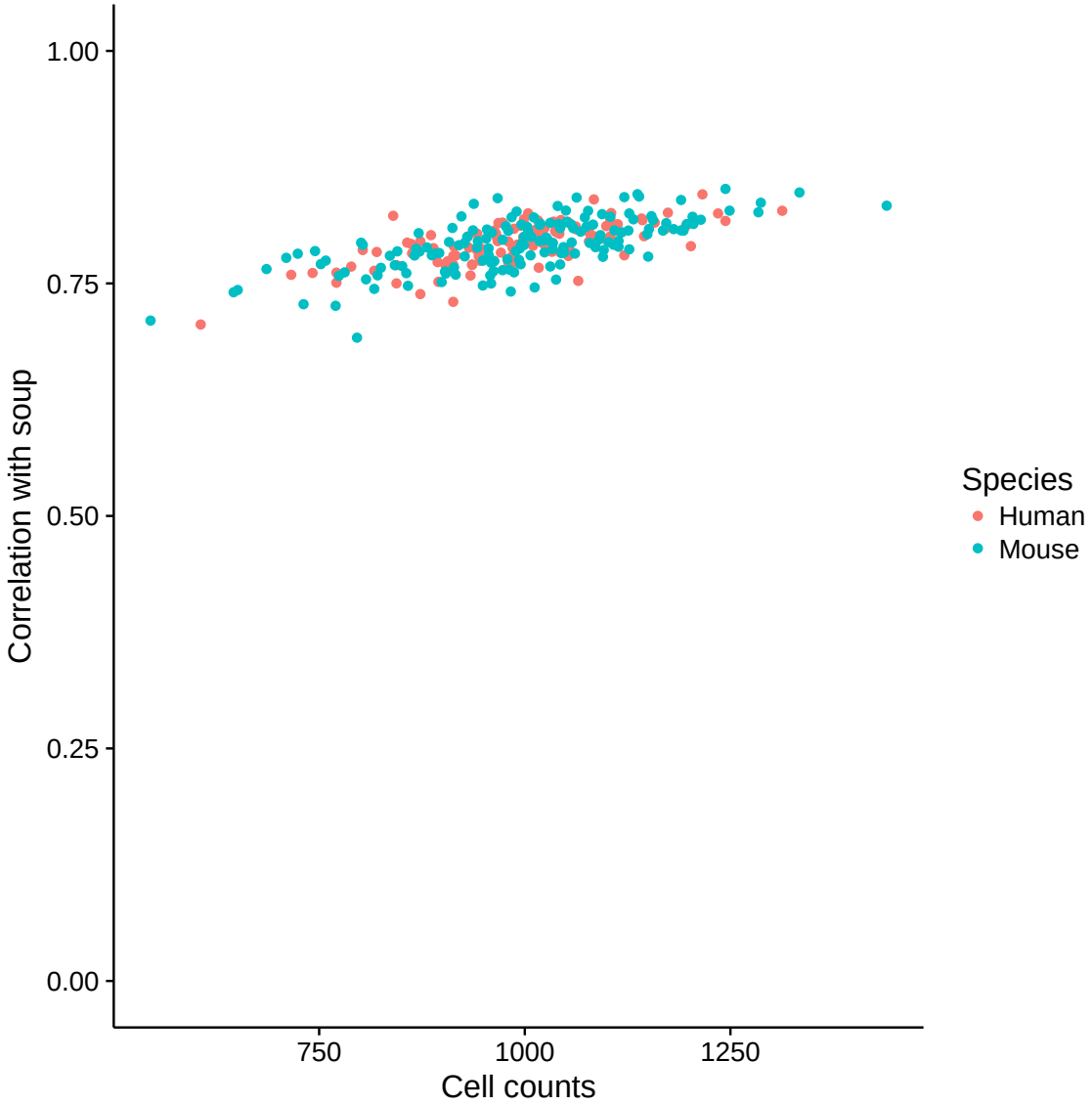

Figure S3: *The correlation between the empty droplet defined background and the background estimated from a small number of cells (binned such that we have $\sim 1,000$ counts with which to estimate the background in each bin). The x-axis gives the number of counts coming from the species that is background (i.e., human for mouse cells and visa versa) for each group of cells. The y-axis gives the correlation between the background for that species and the expression profile using background transcripts from each bin of cells.*

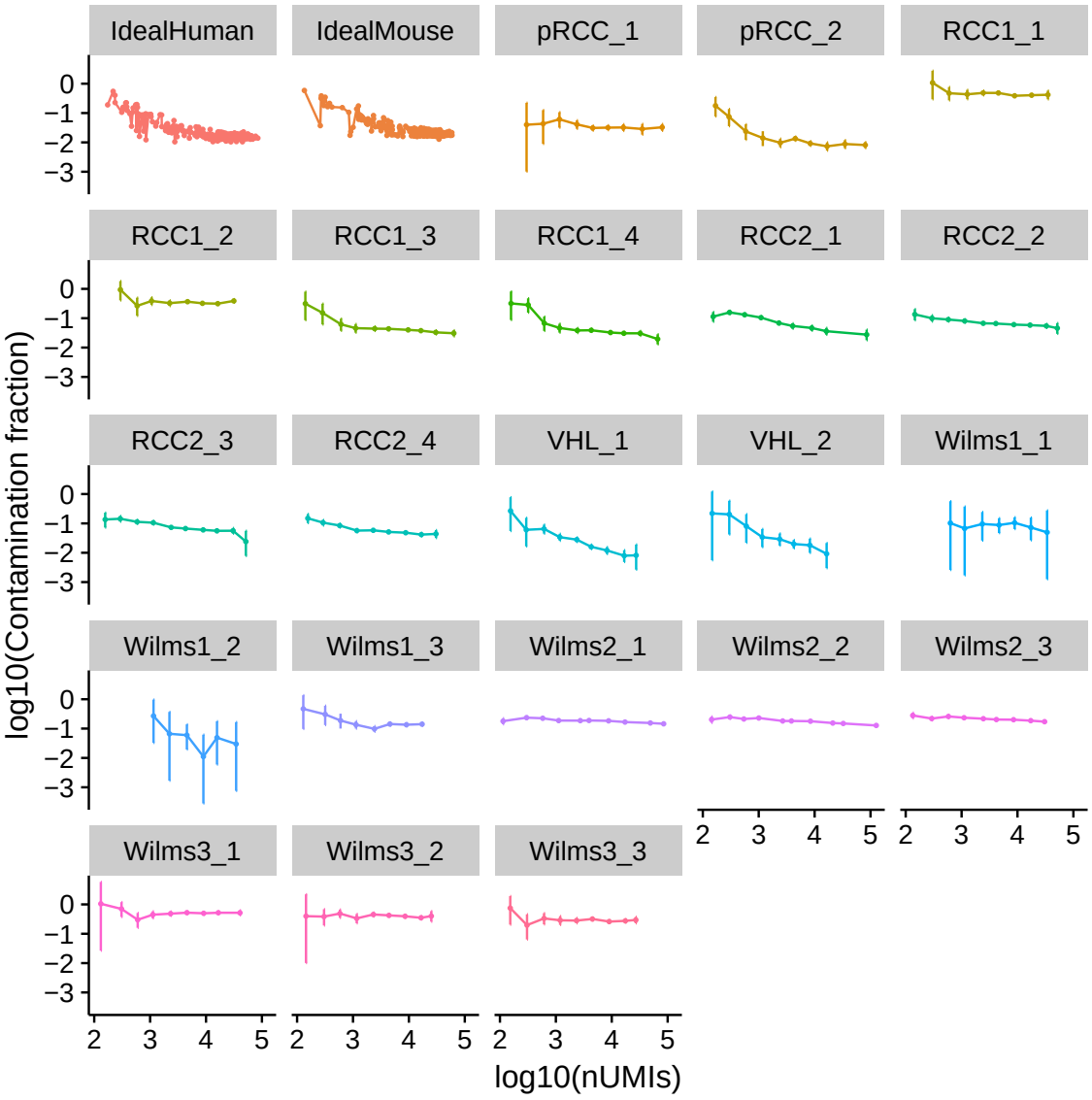

Figure S4: *The contamination fraction, ρ, estimated by binning cells by similar number of UMIs as a function of UMIs in each bin, shown for all samples. The facet label identifies the channel and the vertical bars give binomial 95% confidence intervals for ρ for each bin of cells.*

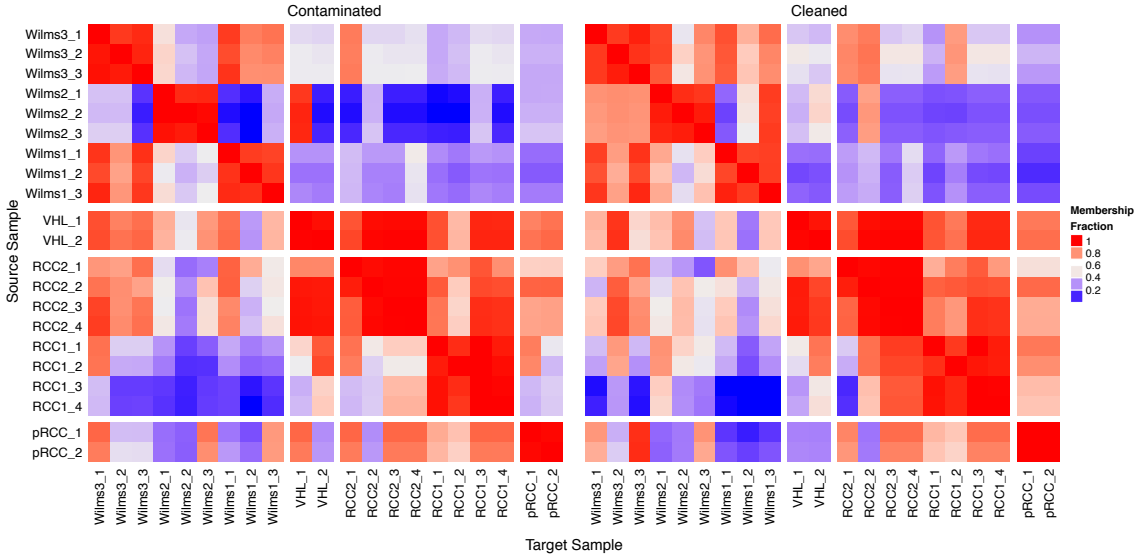

Figure S5: *Comparison of cell membership by cluster before (left) and after (right) background decontamination. For each source channel (rows), we calculate the number of cells that share a cluster with at least one cell from each of the target channels (columns). For example, the first row represents the* 280 *cells from channel Wilms3_1. For each of these* 280 *cells we identify which cluster they belong to before and after background decontamination and for each of these clusters, which of the target channels also has at least one cell in this cluster. We then calculate the fraction of cells in this channel that share a cluster with each of the other channels. In this example, cells from the channel Wilms1_1 share a cluster with cells from channel Wilms1_3 98% of the time before background decontamination and 99% of the time after background decontamination. Rows and columns are split to emphasise biologically related groups of samples.*

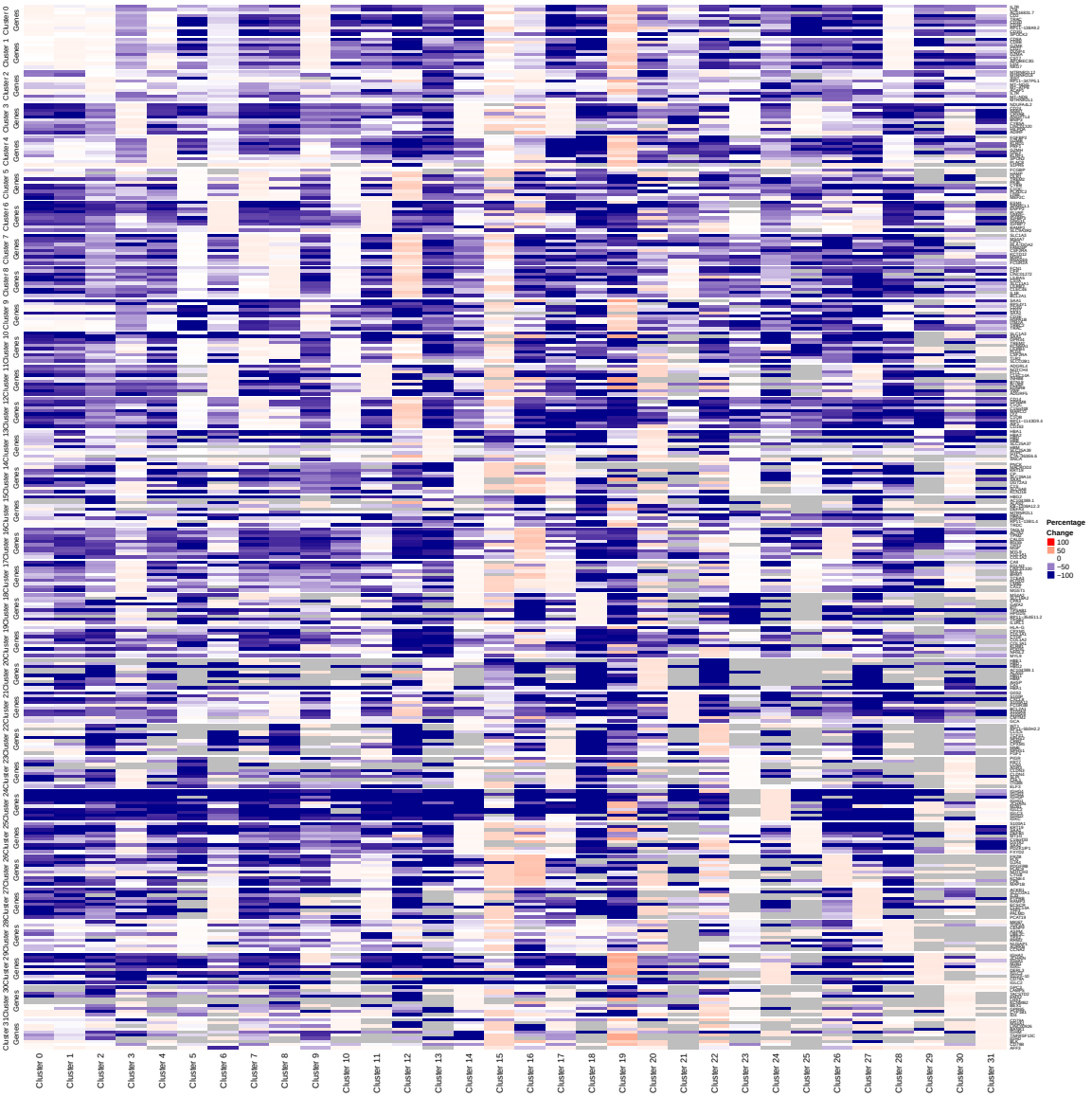

Figure S6: *Overview of how the top 10 cluster specific marker genes change in expression after batch correction. The colour scheme indicates the fractional change in expression following batch correction; genes with zero expression before correction are shown in grey. Rows are split into groups of 10, with each group giving the top 10 cluster specific genes before batch correction.*

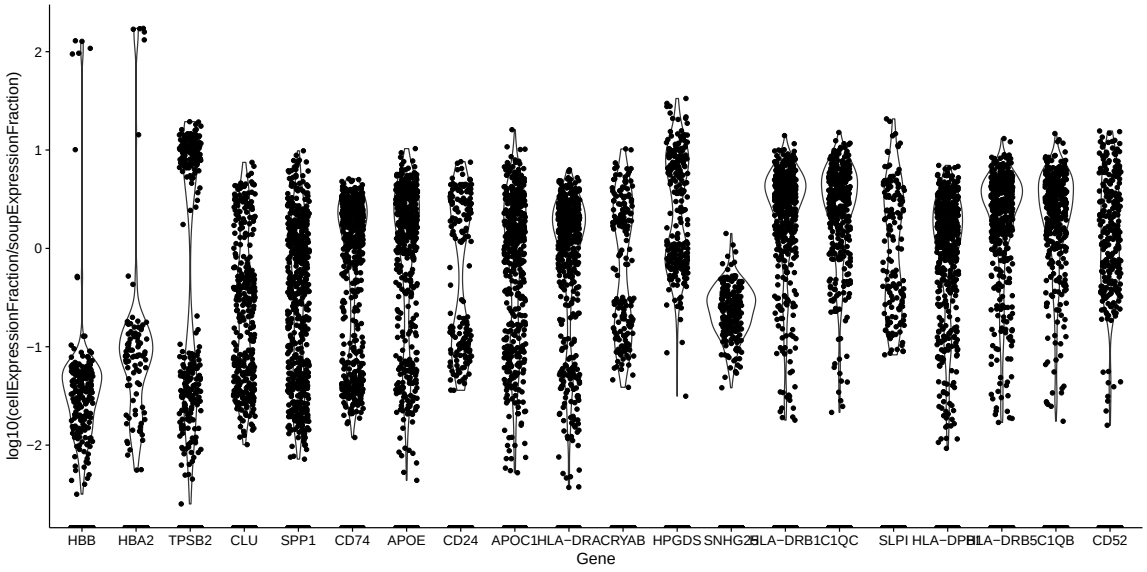

Figure S7: *Distribution of expression relative to background for genes in one of the pRCC channels. These genes are selected as they are most likely to be informative in estimating the contamination fraction ρ (see Methods).*