miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data

Ariel A. Hippen[1], Matias M. Falco[2], Lukas M. Weber[3], Erdogan Pekcan Erkan[2], Kaiyang Zhang[2], Jennifer Anne Doherty[4], Anna Vähärautio[*2], Casey S. Greene[5], and Stephanie C. Hicks[*3]

[1] *Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, PA, USA*

[2] *Research Program in Systems Oncology, Research Programs Unit, Faculty of Medicine, University of Helsinki, Helsinki, Finland*

[3] *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, MD, USA*

[4] *Huntsman Cancer Institute and Department of Population Health Sciences University of Utah, UT, USA*

[5] *Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, CO, USA*

March 3, 2021

Abstract

Motivation: Single-cell RNA-sequencing (scRNA-seq) has made it possible to profile gene expression in tissues at high resolution. An important preprocessing step prior to performing downstream analyses is to identify and remove cells with poor or degraded sample quality using quality control (QC) metrics. Two widely used QC metrics to identify a 'low-quality' cell are (i) if the cell includes a high proportion of reads that map to mitochondrial DNA (mtDNA) encoded genes and (ii) if a small number of genes are detected. Current best practices use these QC metrics independently with either arbitrary, uniform thresholds (e.g. 5%) or biological context-dependent (e.g. species) thresholds, and fail to jointly model these metrics in a data-driven manner. Current practices are often overly stringent and especially untenable on lower-quality tissues, such as archived tumor tissues.

Results: We propose a data-driven QC metric (miQC) that jointly models both the proportion of reads mapping to mtDNA genes and the number of detected genes with mixture models in a probabilistic framework to predict the low-quality cells in a given dataset. We demonstrate how our QC metric easily adapts to different types of single-cell datasets to remove low-quality cells while preserving high-quality cells that can be used for downstream analyses.

Availability: Software available at `https://github.com/greenelab/miQC`. The code used to download datasets, perform the analyses, and reproduce the figures is available at `https://github.com/greenelab/mito-filtering`.

Contact: Stephanie C. Hicks (shicks19@jhu.edu) and Anna Vähärautio (anna.vaharautio@helsinki.fi)

[*]Co-corresponding authors

1 Introduction

Recent advances in single-cell RNA-sequencing (scRNA-seq) technologies have enabled genome-wide profiling in thousands to millions of individual cells [1]. As these technologies are relatively costly, researchers are eager to maximize the information gain and subsequent statistical power from each sample and experiment [2]. scRNA-seq is also extremely sensitive to poor or degraded sample quality, which is a particular concern for tissues, such as tumors, obtained during extensive surgeries or other long-duration procedures [3, 4]. It is crucial that 'compromised' cells ('low-quality' or 'failed' cell libraries in the library preparation process or cells that were dead at the time of tissue extraction) be removed prior to downstream analyses to mitigate against discovered results stemming from a technical artifact instead of meaningful biological variation [5, 6]. These considerations have inspired a wealth of research into best practices for quality control (QC) in scRNA-seq [7–9].

One widely used QC metric to identify a compromised cell is if the cell includes a high proportion of sequencing reads or unique molecular identifier (UMI) counts that map to mitochondrial DNA (mtDNA) encoded genes. Mitochondria are heavily involved in cellular stress response and mediation of cell death [10]. These sorts of cellular stresses can be products of the vigorous process of cell dissociation, and the inclusion of these transcriptionally-altered cells can affect downstream analysis outcomes [5]. Additionally, a high abundance of counts mapping to mtDNA genes can indicate that the cell membrane has been broken, and thus cytoplasmic RNA levels are depleted relative to the mRNA protected by the mitochondrial membrane [11]. For these reasons, it is standard practice to remove cells with a large percentage of reads or UMI counts mapping to mtDNA genes using some arbitrary and uniform thresholds, for example greater than 5% [12, 13]. However, recent work has shown these thresholds can be highly dependent on the organism or tissue, the type of scRNA-seq technology used, or the protocol specific decisions made as part of the disassociation, library preparation, and sequencing steps [9, 14, 15]. For instance, evidence suggests that cells that have been treated with certain RNA-preserving reagents prior to library preparation have a much higher mitochondrial fraction compared to fresh tissues [16].

Other widely used QC metrics to identify compromised cells are the total number of sequencing reads or UMI counts in a sample and the number of unique genes that those reads or counts map to, also known as library complexity [17, 18]. For example, if all observed UMI counts in a cell map to only a few genes, this also suggests that the mRNA in the cell may have been degraded or lost in one or more protocol steps. A standard approach to filter out these cells is to filter out cells with an *ad hoc* threshold of less than a certain number of unique genes detected, such as less than 100 genes. An alternative approach is to rank the cells by their total UMI count and visually inspect for a knee point in the data, but this approach is often arbitrary and difficult to reproduce [19].

Current best practices use all these QC metrics independently and commonly use uniform thresholds, sometimes species-dependent thresholds, which can lead to arbitrary cutoffs that may not be appropriate for a given dataset. These cutoffs are often conservative, leaving only a small number of the remaining cells, which offers a non-representative sample of the tissue and constrains downstream analyses.

Here, we propose an alternative approach to enable researchers to make data-driven decisions about which population a given cell comes from with respect to both mtDNA fraction and library complexity, which is adaptive across scRNA-seq datasets. Throughout the rest of the text, we use the terms (i) a *compromised* cell to refer to a low-quality cell that is expected to have few unique genes represented and a high mitochondrial fraction and (ii) an *intact* cell to refer to cells that are of a high-quality (e.g. with an intact cell membrane) with a low proportion of mitochondrial reads and should be included in downstream analyses [11, 20]. For a given scRNA-seq sample, we model the cells using latent variable model with a true and unknown (or hidden) latent factor representing the two populations of cells. We fit a finite mixture of models in a probabilistic framework and remove cells based on the posterior probability of coming from the compromised cell distribution. By modeling distributions of parameters for each tissue sample, biological and technical variation can be accounted for in a highly adaptive, sample-specific manner, while still providing a consistent set of principles for inclusion. We demonstrate that across a variety of tissues and experiments, our method preserves more intact cells post-QC than uniform mitochondrial thresholds, which can be used in downstream analyses. Our data-driven methodology for QC is available in a R/Bioconductor software package at `https://github.com/greenelab/miQC`.

2 Results

2.1 miQC: a data-driven metric for quality control in scRNA-seq data

To motivate the need of a data-driven approach, we first explored the use of commonly used QC thresholds to remove compromised cells in a high-grade serous ovarian cancer (HGSOC) tissue sample (sample ID 16030X4 from [21] and described in Section 3.1). For each cell, we calculated the percent of counts mapping to mtDNA genes and the number of unique genes those counts map to (or library complexity). As stated above, based on previous biological knowledge, we expect intact cells to have low percent of counts mapping to mtDNA genes and moderate to high library complexity. In contrast, compromised cells are expected to have a large percent of counts mapping to mtDNA genes and a low library complexity. In our cancer sample, we observed a peak of counts mapping to mtDNA genes at 13% and a wide range of the number of unique genes found (**Figure 1A**). However, as the percent of counts mapping to mtDNA genes increases, the number of unique genes decreases significantly, suggesting these are compromised cells. These are the two population of cells we aim to discover in a data-driven manner.

Using this cancer sample, when we remove cells using a uniform and *ad hoc* QC threshold, for example greater than 10% cell counts mapping to mtDNA genes as suggested by [15], we remove 5828 cells (88.1%) from the sample. As ovarian cancer samples presented the most abundant mtDNA copy numbers in a broad pan-cancer comparison across 38 tumor types [22], mitochondrial transcript content is expected to be relatively high also for intact ovarian cancer cells. Thus, the QC based on an arbitrary limit of 10% is overly aggressive and renders most of the data from a sample unusable. Alternatively, if we use a more data-driven approach that only considers the percent of counts mapping to mtDNA genes and removes cells with greater than 3 median absolute deviations (MADs) [23, 24], we remove no cells from the sample, resulting in an overly permissive QC. Both of these approaches fail in this scenario because they are designed for extremely high-quality datasets with only a trivial number of compromised cells. These analyses motivated our proposed approach that is designed to discriminate between compromised and intact cells that is adaptive to a spectrum of data quality in scRNA-seq data, described in the next section.

2.1.1 Probabilistic classifications for scRNA-seq data quality using mixtures of linear models

Because of the limitations of uniform and *ad hoc* QC thresholds, we aimed to use a probabilistic framework that jointly models two QC metrics to predict the compromised cells in a given dataset. We assume that for any cell i, there is a latent variable that we do not observe $Z_i = 1$ if the cell is considered a compromised cell and should be removed from downstream analyses, and $Z_i = 0$ if the cell library is intact. We denote π_1 as the probability $\Pr(Z_i = 1)$ and $\pi_0 = 1 - \pi_1 = \Pr(Z_i = 0)$. We also define Y_i as the percent of counts in the i^{th} cell that map to mtDNA genes and X_i is the number of unique genes detected or found for the i^{th} cell. Then, we assume that conditional on Z_i and X_i, the expected percent of mtDNA counts Y_i is

$$E[Y_i|Z_i = z, X_i = x_i] = f_z(x_i) \tag{1}$$

We note that f_z represents a different function estimated for the two states $z = \{0, 1\}$ (intact or compromised cells, respectively). We assume the errors are modeled with different variance components $\varepsilon_{iz} \sim N(0, \sigma_z^2)$ for the two z states. By default, we assume the function $f_z(x_i)$ takes the form of a standard linear regression model $f_z(x_i) = \beta_{0z} + \beta_{1z}x_i$ where β_{0z} represents the mean level of percent of mtDNA counts for the two states $z = \{0, 1\}$ and β_{1z} represents the corresponding coefficient, which is also estimated differently for each of the two states $z = \{0, 1\}$ (**Figure 1B**). This finite mixture of linear regression models is also known as latent class regression [25]. However, our approach can also use a more flexible model such as $f_z(x_i) = \mu_z + g_z(x_i)$ where μ_z is again the mean level and $g_z(x_i)$ is a nonparametric smooth function that can be estimated with, for example a B-spline basis matrix (**Figure S1**).

To estimate the parameters $\theta = (\pi_z, f_z)$ for the two states $z = \{0, 1\}$, we use an Expectation Maximization (EM) algorithm [26] implemented in the *flexmix* [27] R package. Using the estimated parameters from the EM algorithm, we calculate the posterior probability of a compromised cell as

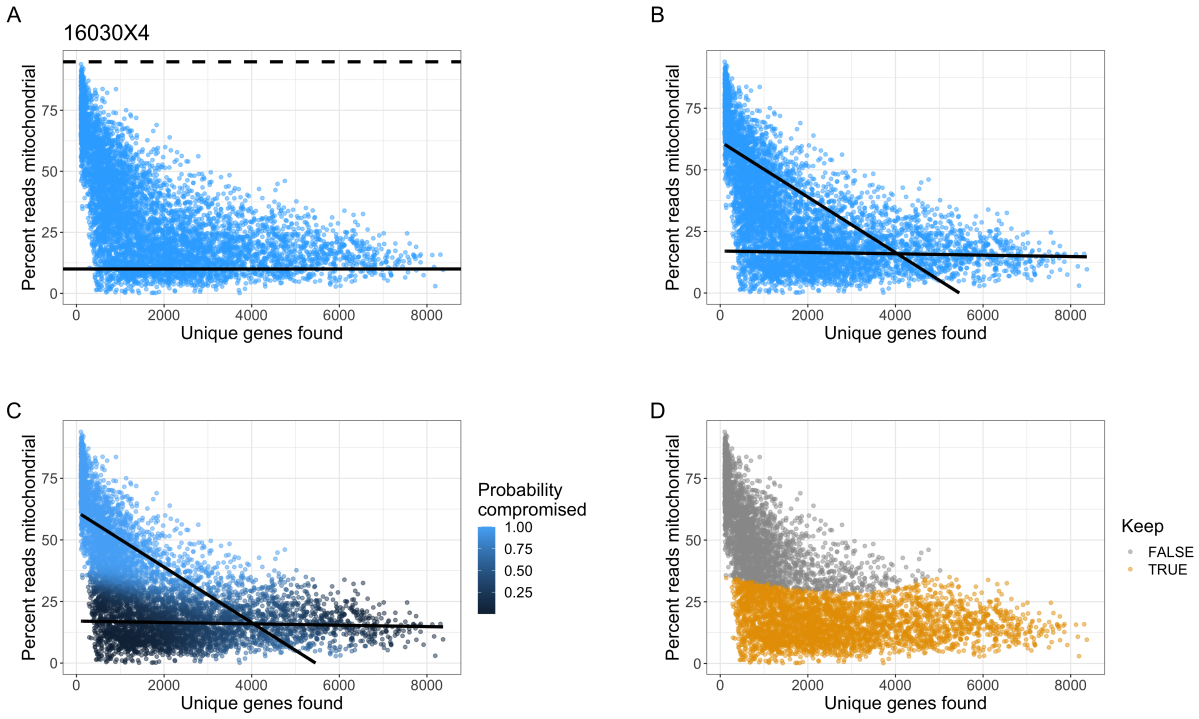

Figure 1: **Uniform and data-driven quality control (QC) thresholds for scRNA-seq data.** Cells (N=6618) from one high-grade serous ovarian cancer (HGSOC) tissue sample (Sample ID: 16030X4) with the number of unique genes found (x-axis) and percent of cell counts mapping to mitochondrial (mtDNA) genes (y-axis). **(A)** Illustration of removing cells with a uniform QC threshold of greater than 10% cell counts mapping to mtDNA genes (solid black line) and a more data-driven threshold of greater than 3 median absolute deviations (MADs) of the percent of counts mapping to mtDNA genes (dotted black line). **(B)** Using our data-driven approach (miQC), we fit a finite mixture of standard linear regression models with two lines (black lines) to calculate a posterior probability of being a compromised cell. **(C)** Cells shaded by their posterior probability of being compromised. **(D)** Discarding all cells with $\geq$ 75% probability of being compromised creates a data-driven QC threshold for scRNA-seq data.

$$\gamma_{Z_i}(1) = \Pr(Z_i = 1 | Y_i, X_i = x_i, \theta) = \frac{\pi_1 N(Y_i | f_1(x_i), \sigma_1^2)}{\pi_1 N(Y_i | f_1(x_i), \sigma_1^2) + \pi_0 N(Y_i | f_0(x_i), \sigma_0^2)} \tag{2}$$

where $N(\cdot)$ represents the probability density function of a Gaussian distribution with mean $f_z(x_i)$ and variance σ_z^2 for the two states $z = \{0, 1\}$.

We use the posterior probability $\gamma_{Z_i}(1)$ as the data-driven threshold to exclude (or keep) cells (**Figure 1C**). In our analyses, we remove cells with a greater than 75% probability of belonging to the compromised cell distribution, in order to maximize the number of potentially informative cells while still removing the cells most likely to confound downstream analyses (**Figure 1D**). In the next section, we demonstrate how this threshold is adaptive across species, tissues and experimental protocols. However, the `posterior` argument in the miQC package can be used to adjust the posterior probability threshold, depending on the needs of a given experiment.

2.2 miQC is adaptive across species, tissues, and experimental protocols

Previous work has demonstrated that the expected amount of mitochondrial activity varies across species and tissue types. For example, one study concluded that filtering all cells above 5% cell counts mapping to mtDNA genes is appropriate for mouse samples, but that a cutoff of 10% is preferable in human samples [15]. However, it has also been demonstrated that certain tissues, especially those with high energy requirements such as brain

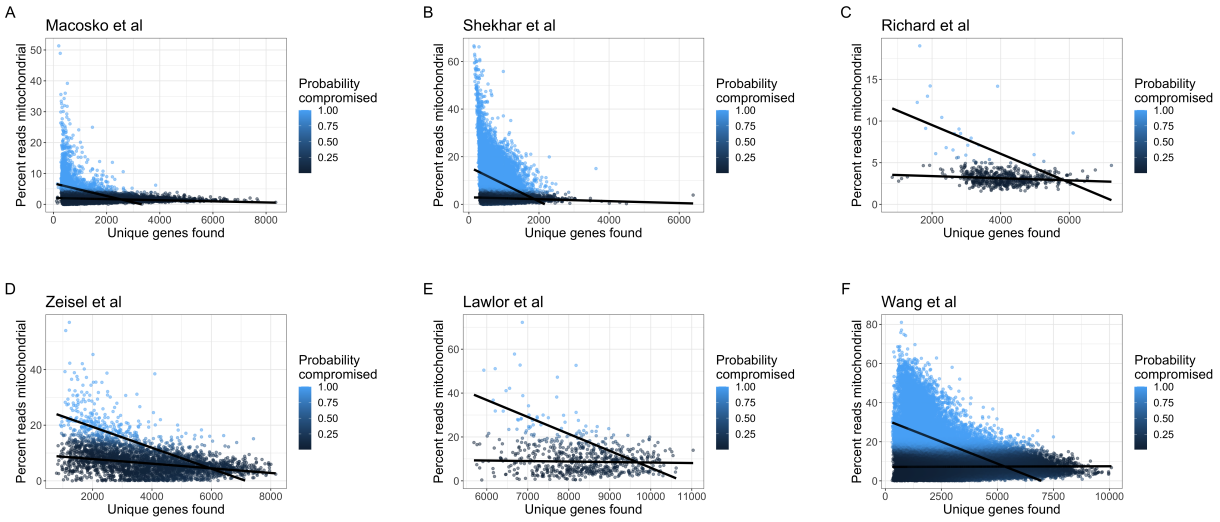

Figure 2: **miQC is adaptive across species, tissues, and experimental protocols.** Using publicly available scRNA-seq data from non-cancer tissues, we calculated the number of detected genes (x-axis) and percent of cell counts mapping to mitochondrial (mtDNA) genes (y-axis) for each cell. Black lines represent the finite mixture of linear regression models estimated from miQC to adaptively identify compromised cells across datasets. Color represents the posterior probability of a given cell being a compromised cell (light blue is a high probability and dark blue is a low probability). The data span both **(A-D)** mouse and **(E-F)** human scRNA-seq cells from non-cancer **(A-B)** retinal [29, 30], **(C)** immune [31], **(D)** brain [32], **(E)** pancreas [33], and **(F)** menstrual blood tissues [34]. In addition, the data span **(A, B, F)** droplet-based **(C)** plate-based and **(D, E)** microfluidic-based protocols.

and heart, have a higher baseline mitochondrial expression [28], and that mtDNA copy numbers highly vary across tissue and cancer types [22].

Here, we consider publicly available scRNA-seq datasets and demonstrate that our miQC approach identifies adaptive QC thresholds across species, tissue types, and experimental protocols. Specifically, we explore $N = 6$ datasets (described in detail in Section 3.1) ranging from hundreds to tens of thousands of cells from (i) two species (mouse and human), (ii) five non-cancer tissue types (retinal, immune, brain, pancreas, menstrual blood), (iii) one cancer tissue type (HGSOC), and (iv) two experimental protocols (plate-based and droplet-based single cell protocols).

2.2.1 Using non-cancer tissues

Using mouse scRNA-seq data from retinal [29, 30] and immune [31] cells measured on the Drop-seq and Smart-seq2 platforms, we found that miQC identifies a similar QC threshold to using the 5% threshold found in a previous study [15] (**Figure 2A-C**). However, using mouse scRNA-seq data from brain cells measured on the Fluidigm C1 platform [32], we found that miQC proposes a less stringent QC threshold compared to the 5% threshold suggested by [15] (**Figure 2D**). In this case, if a 5% threshold was used, there would be $N = 1948$ cells (or 64.8%) removed from the sample, which are likely to contain intact and biologically informative cells.

In humans, previous work has shown pancreas typically expresses a large fraction of mtDNA genes [15]. Using human scRNA-seq data from pancreas measured on the Fluidigm C1 platform [33], our miQC approach agrees with this result and the model suggests excluding $N = 48$ cells (or 7.5%) from the sample (**Figure 2E**) in contrast to removing $N = 290$ cells (or 45.5%) if using a 10% threshold as suggested for human tissues. In addition, we found using human scRNA-seq data from menstrual blood measured on the 10x Chromium platform [34] that our miQC approach excludes fewer cells ($N = 13158$ or 18.5%) from the sample (**Figure 2F**) in contrast to removing $N = 29129$ cells (or 41%) if using a 10% threshold as suggested by [15].

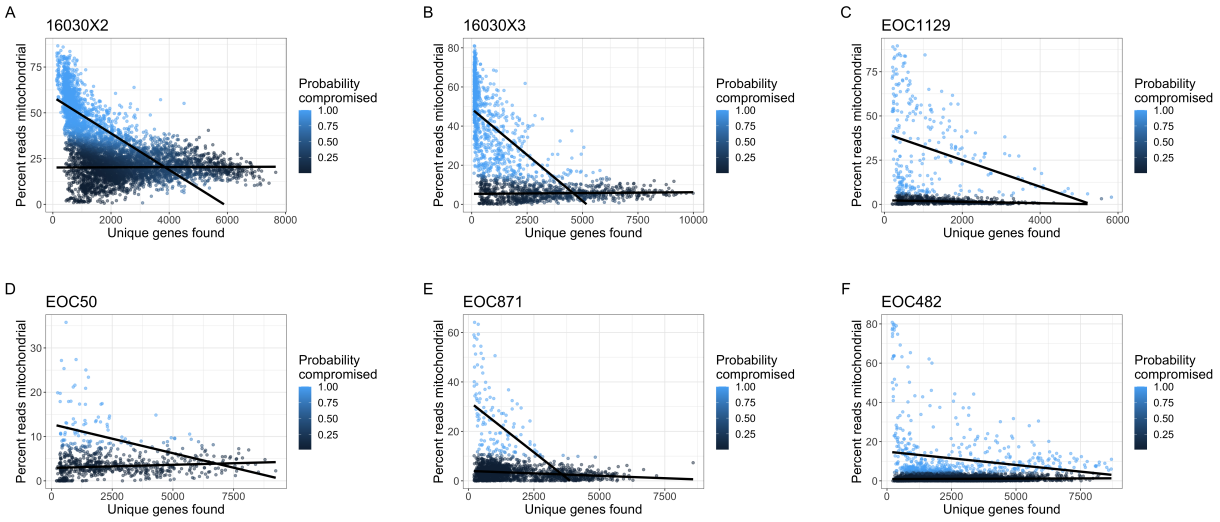

Figure 3: **miQC is adaptive across different samples within same cancer type**. scRNA-seq data from human HGSOC samples from **(A-B)** the Huntsman Cancer Institute (Sample IDs: 16030X2, 16030X3) and **(C-F)** the University of Helsinki (Sample IDs: EOC1129, EOC50, EOC871, EOC482). (Data from remaining Huntsman Cancer Institute sample, ID 16030X4, is depicted in Figure 1.) We calculated the number of detected genes (x-axis) and percent of cell counts mapping to mitochondrial (mtDNA) genes (y-axis) for each cell. Black lines represent the finite mixture of linear regression models estimated from miQC to adaptively identify compromised cells. Color represent the posterior probability of a given cell being a compromised cell (light blue is a high probability and dark blue is a low probability).

2.2.2 Using cancer tissues

A major advantage of our data-driven miQC approach is the use of the posterior probability threshold for inclusion, because it allows for a consistent QC metric to be applied across all samples in a set of experiments, while still flexibly accommodating differences in samples or tissues. This is important for experiments leveraging data collected from across different experimental laboratory settings or at multiple times where these factors have been shown to contribute differences in batch effects [35] or percent of counts mapping to mtDNA genes. This is particularly true in application of scRNA-seq cancer samples, where the high heterogeneity of tumor composition and cancer cell behavior make it especially challenging to assign one cutoff metric for all samples.

Here, we apply miQC to a set of scRNA-seq data derived from multiple human high-grade serous ovarian tumors (HGSOC) [21] ($N = 7$ tumor samples described in detail in Section 3.1, with Sample 16030X4 depicted and discussed in Section 2.1) using the 10X Chromium experimental protocol. For each cell in a HGSOC tumor sample, we calculate the number of detected genes and the percent of cell counts mapping to mitochondrial genes, similar to Figure 2, which resulted in wide variation of what might be a compromised cell. However, we found that our approach miQC is able to adaptively find QC thresholds across different tumor samples all within the same cancer type (**Figure 3A-F**). Specifically, we found using miQC removes $N = 1387$ (28.1%), 792 (47.8%), 254 (29.2%), 78 (11.3%), 132 (8.9%), 508 (13.2%) cells in contrast to 4683 (94.8%), 911 (55.0%), 200 (23.0%), 50 (7.2%), 131 (8.5%), 185 (4.8%) cells if using a 10% threshold as suggested by [15].

2.3 miQC is adaptive across choice of reference genome used in a data analysis

In addition to the biological factors that can affect baseline mitochondrial expression, there are additional technological and experimental factors that can change the observed number of counts mapping to mtDNA genes as well. For example, we found one crucial component is the choice of the reference genome used for quantification of cell reads or UMI counts. The mitochondrial genome has been annotated for decades and genic content is known to be highly conserved across animal species: 37 genes, coding for 13 mRNAs, 2 rRNAs, and 22 tRNAs [36]. However, some reference genomes include all 37 genes where others only include the 13 protein-coding genes.

We investigated this technological confounding factor within one of the HGSOC tumor samples (Sample ID: EOC871). We considered the scRNA-seq cell counts that were quantified using (i) Cell Ranger [1] with the human genome reference GrCh38 (version 2020-A) filtered to remove pseudogenes, and (ii) salmon alevin [37] with the unfiltered human genome reference GENCODE (Release 31) [38]. We found that when quantifying reads with these two different reference genomes, the cell counts that would have mapped to the "missing" mitochondrial tRNA and rRNA genes (in the GENCODE reference genome) are instead assigned to mitochondrial-like pseudogenes on the chromosomes (in the GrCh38 reference genome). This results in a non-uniform shift and technological inflation in the percent of cell counts mapping to mitochondrial genes (**Figure 4**). These results agree with the findings of Brüning et al that using a filtered transcriptome annotation causes an increase in number of reads mapping to mitochondrial genes irrespective of quantification software used [39]. While we compared GrCh38 and GENCODE annotations, the authors of [39] compared Ensembl annotations with and without cellranger's *mkgtf* function applied, indicating the effect on mitochondrial reads is present across several references. This highlights the importance of accounting for this potential confounding factor to consider if, for example, researchers are performing quality control on cell counts derived with differently derived reference genomes, which we anticipate to become more relevant as cancer atlases grow. Also, mitochondrial reference genomes may diverge further as additional non-coding RNAs and pseudogenes are discovered and characterized [40].

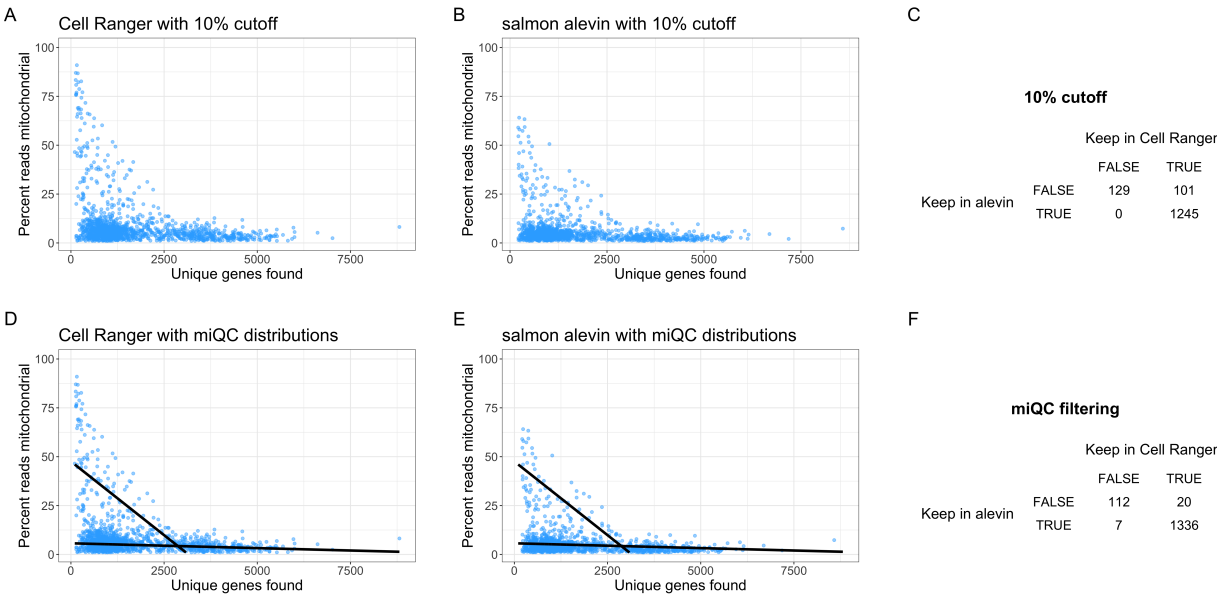

Figure 4: **miQC is adaptive across the choice of reference genome used**. Cells (N=1733) from one high-grade serous ovarian cancer (HGSOC) tissue sample (Sample ID: EOC871) with the number of unique genes found (x-axis) and percent of cell counts mapping to mitochondrial (mtDNA) genes (y-axis). Quantification of cell counts was performed with **(A)** Cell Ranger using the GrCh38 reference genome (version 2020-A) and **(B)** salmon alevin with the GENCODE reference genome (release 31). In both (A-B), we use a standard 10% mtDNA threshold for this human cancer sample to remove compromised cells. **(C)** A confusion matrix of how cells are filtered differently by this uniform threshold using the different reference genomes, resulting in a large number of cells removed after QC using either reference genome, but differences in which cells are removed depending on the choice of reference genome. **(D-E)** Using the same tissue sample as (A-C), but here we use our miQC approach to fit a finite mixture of standard linear regression models with two lines (black lines) to calculate the posterior probability of being a compromised cell. **(F)** A confusion matrix of what cells are filtered by our miQC approach which not only results in a larger number predicted intact cells after QC, but also results in a smaller number of discrepancies between which cells are included after QC.

Using a uniform 10% QC threshold to identify compromised cells (**Figure 4A,B**), we found this removes either $N = 101$ and $N = 230$ (or 6.8% and 14.6%) using Cell Ranger and salmon alevin, respectively, when using two different reference genomes, despite these being being the exact same cell libraries, just being quantified with

two different reference genomes (**Figure 4C**). Interestingly, we also found differences in which cells are removed depending on the choice of reference genome with a greater fraction removed by Cell Ranger. In contrast, we found our miQC approach (**Figure 4D,E**) is able to flexibly identify different QC thresholds when using two different reference genomes, removing a more similar set of cells: $N = 119$ cells and $N = 132$ cells (or 8.1% and 8.9%) using Cell Ranger and salmon alevin, respectively (**Figure 4F**). This demonstrates our data-driven approach is able to adjust for differences in this technological confounding factor of diverging mitochondrial annotation the quantification step of the analysis of scRNA-seq data.

2.4 miQC minimizes cell type-specific sub-population bias

A standard downstream scRNA-seq data analysis is identifying cell types in a tissue or tumor sample and detecting differences between cell types [24]. A crucial component of this analysis is to have sufficient statistical power to detect differences between cell types, which depends on having appropriate sample sizes of measured cells – and the choice of QC metrics and thresholds directly impacts the number of cells employed in these downstream analyses. For example, in application of unsupervised clustering if a large number of cells are removed post-QC, the number of cells per cluster, and even the number of clusters discovered, can be affected. Therefore, it is important to evaluate whether the choice of QC metric and corresponding threshold do not significantly negatively impact the unsupervised clustering results. In fact, Germain et al. [41] argued "although more stringent filtering tended to be associated with an increase in accuracy, it tended to plateau and could also become deleterious. Most of the benefits could be achieved without very stringent filtering and minimizing subpopulation bias", where *sub-population bias* is defined as disproportionate exclusion of certain cell populations.

Here, we aimed to investigate whether our miQC approach resulted in minimized sub-population bias, as described by [41], compared to the standard approach of using a uniform QC threshold of 10% of cell counts mapping to mtDNA genes. Using one HGSOC tumor sample (Sample ID 16030X4), we preprocessed and normalized the scRNA-seq data according to [24] followed by applying dimensionality reduction using the Uniform Manifold Approximation and Projection (UMAP) [42] representation. The percent of cell counts mapping to mtDNA genes in this representation is shown in (**Figure 5A**). Using the top 50 principal components, we performed unsupervised clustering using the mini-batch k-means (mbkmeans) algorithm [43] implemented in the *mbkmeans* [44] R/Bioconductor package for unsupervised clustering to identify cell types, which is a scalable version of the widely-used k-means algorithm [45–47] (**Figure 5B**). The number of clusters (k=6) was determined using an elbow plot with the sum of squared errors (**Figure S2**). Using these k=6 clusters, we compared proportions of cells belonging to each predicted cluster using (i) no filtering, (ii) our miQC threshold, and (iii) the uniform QC threshold of 10% of cell counts mapping to mtDNA genes (**Figure 5C-E**).

We found that the cells in cluster 5 (purple bar in Figure 5) were almost entirely removed (cluster 5: 99.3%, 100%) by miQC and the standard 10% threshold, respectively. These cells had an average mitochondrial fraction of 68.9% (**Figure 5A**). We can reasonably infer that those are compromised cells, and as such excluding them from a downstream analysis is appropriate. In contrast, we found that for all other clusters miQC removed far fewer cells than the 10% threshold approach (cluster 1: 5.7%, 83.3%; cluster 2: 4.9%, 76.2%; cluster 3: 52.2%, 96.0%; cluster 4: 1.9%, 54.1%; cluster 6: 74.2%, 97.5%). This suggests miQC preserves more cells within each predicted cluster and minimizes sub-population bias, compared to the uniform threshold approach of a 10% cutoff.

3 Methods

3.1 Datasets

3.1.1 Non-cancer tissue scRNA-seq datasets

We obtained non-cancer tissue scRNA-seq datasets for the studies Macosko et al. [29], Shekhar et al. [30], Richard et al. [31], Zeisel et al. [32], and Lawlor et al. [33] from the the R/Bioconductor data package *scRNAseq* [48]. We obtained the scRNA-seq data from Wang et al. [34] from the Sequence Read Archive (SRA) (accession code SRP135922). All datasets were processed using *scater* [23], as described in Section 3.2.1. Table 1 contains

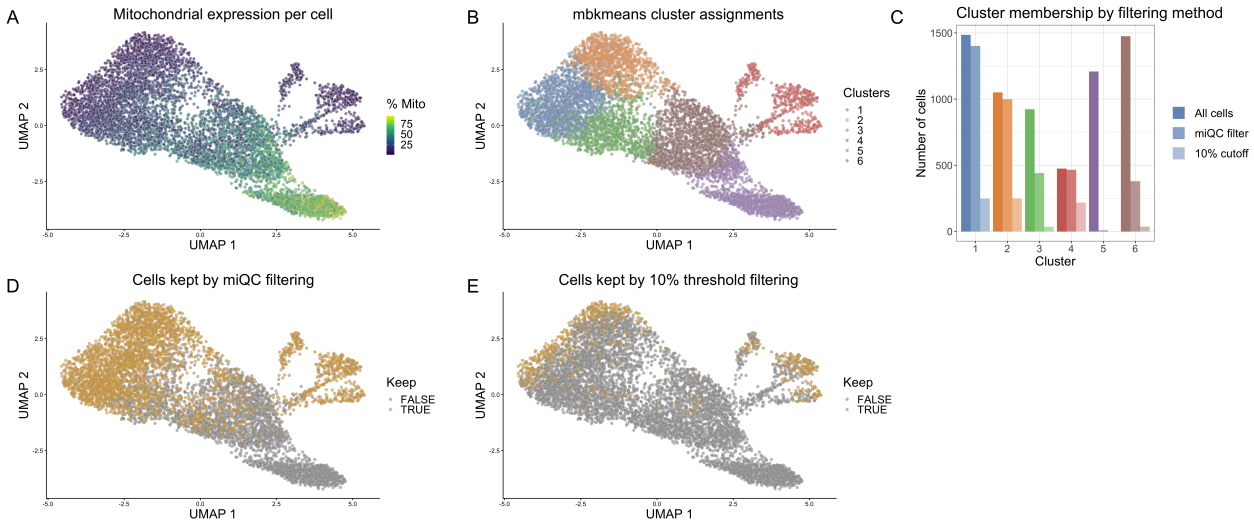

Figure 5: **miQC minimizes sub-population bias in unsupervised clustering**. Cells ($N = 6691$) from one high-grade serous ovarian cancer (HGSOC) tissue sample (Sample ID: 16030X4). **(A)** UMAP representation of cells, colored by percent of cell counts mapping to mtDNA genes. **(B)** Predicted cluster labels for k=6 cell types identified using the mini-batch k-means (mbkmeans) algorithm for unsupervised clustering. **(C)** Total number of cells preserved within each predicted cluster group using three filtering approaches: no filtering ('All cells'), miQC filtering, or removing all cells greater than 10% mtDNA content. **(D)** UMAP representation of cells colored by whether miQC keeps (gold) or removes (gray) cells post-QC. **(E)** Similar as (D), but using a uniform 10% threshold for mtDNA content.

a summary of the non-cancer datasets used: the organism, the tissue, the experimental protocol, and the number of cells prior to QC for each dataset.

Table 1: **Description of non-cancer tissue scRNA-seq datasets**. Columns from left to right include the source, the organism, the type of tissue, the experimental protocol, and the number of cells prior to QC in each dataset.

Source	**Organism**	**Tissue**	**Protocol**	**Number of cells**
Richard et al. [31]	Mouse	T cells	Smart-seq2	572
Zeisel et al. [32]	Mouse	Brain	Fluidigm C1	3005
Shekhar et al. [30]	Mouse	Retina	Drop-seq	44994
Macosko et al. [29]	Mouse	Retina	Drop-seq	49300
Lawlor et al. [33]	Human	Pancreatic islets	Fluidigm C1	638
Wang et al. [34]	Human	Menstrual blood	10X Chromium	71032

3.1.2 Cancer tissue scRNA-seq datasets

The $N = 7$ HGSOC tumor samples were collected and sequenced at Huntsman Cancer Institute, Utah, USA ($N = 3$) and at University of Helsinki, Finland ($N = 4$).

For the samples from the Huntsman Cancer Institute, raw FASTQ files are available through dbGaP (accession phs002262.v1.p1) and processed gene count tables are available through GEO (accession GSE158937) [21]. Complete details of the experimental protocol and sequencing steps followed for these tumor samples provided in Weber et al. [21], but in brief library prepration was performed using 10x Genomics 3' Gene Expression Library Prep v3, and sequencing was done on an Illumina NovaSeq instrument. Quantification for these samples was performed using salmon alevin [37] with a index genome generated from GENCODE v31 [38].

Genome data for the University of Helsinki samples has been deposited at the European Genome-phenome Archive (EGA) which is hosted at the EBI and the CRG, under accession number EGAS00001005066. The samples were taken as a part of a larger study cohort, where all patients participating in the study provided written informed consent. The study and the use of all clinical material have been approved by The Ethics Committee of the Hospital District of Southwest Finland (ETMK) under decision number EMTK: 145/1801/2015. Immediately after surgery, tissue specimens were incubated overnight in a mixture of collagenase and hyaluronidase to obtain single-cell suspensions. Cell suspensions were passed through a 70-µm cell strainer to remove cell clusters and debris and centrifuged at 300 x g. Cell pellets were resuspended in a resuspension/washing buffer (1X PBS supplemented with 0.04% BSA) and washed three times. scRNA-seq libraries were prepared with the Chromium Single Cell 3' Reagent Kit v. 2.0 (10x Genomics) and sequenced on an Illumina HiSeq4000 instrument. Using the raw FASTQ files, we performed the quantification step with two different methods to obtain two different UMI counts matrices. First, Cell Ranger (version 3.1.0) [1] was used to perform sample de-multiplexing, alignment, filtering, and barcode and UMI quantification. GRCh38.d1.vd1 genome was used as reference and GENCODE v25 for gene annotation. Second, salmon alevin [37] (version 1.4.0) was also used with GRCh38.p13 as reference genome and GENCODE v34 for gene annotation. Table 2 contains a summary of the cancer datasets used, including the source, the organism, the location from where the tumors were obtained, and the number of cells prior to QC in each dataset.

Table 2: **Description of human high-grade serous ovarian cancer (HGSOC) tissue scRNA-seq datasets**. Columns from left to right include the source, the organism, the location from where the tumors were obtained, and the number of cells prior to QC in each dataset. Huntsman Cancer Institute samples were processed with the 10x Genomics 3' Gene Expression Library Prep v3 protocol, and University of Helsinki samples were processed with Chromium Single Cell 3' Gene Expression v2 protocol.

Source	Sample ID	Location	Cell Count
Weber et al. [21]	16030X2	Huntsman Cancer Institute	4939
Weber et al. [21]	16030X3	Huntsman Cancer Institute	1725
Weber et al. [21]	16030X4	Huntsman Cancer Institute	6691
Novel as part of this manuscript	EOC1129	University of Helsinki	1086
Novel as part of this manuscript	EOC50	University of Helsinki	930
Novel as part of this manuscript	EOC871	University of Helsinki	1733
Novel as part of this manuscript	EOC482	University of Helsinki	3879

3.2 Data analysis

3.2.1 Preprocessing scRNA-seq datasets

We processed the gene-by-cell matrix from each dataset using the *scater* [23] R/Bioconductor package, including calculating the number of unique genes represented and the percent of reads or UMI counts mapping to mtDNA genes. At this step, we removed any cells with fewer than 500 total reads or fewer than 100 unique genes represented, which we considered to be unambiguously failed.

To represent the effect of miQC on downstream analyses, we calculated and plotted the Uniform Manifold Approximation and Projection (UMAP) representation of the single-cell expression data using functions in the *scater* package. We chose to highlight how miQC filtering specifically affects clustering results using the *mbkmeans* package, which uses mini-batches to quickly and scalably produce k-means clustering assignments [44]. We ran mbkmeans on a reduced representation of our expression data, the first 50 principal components as calculated via scater. All visual representations and figures were generated using the *ggplot2* R package [49].

3.2.2 miQC software implementation

We used the R package *flexmix* [27] to fit the finite mixture of linear (or non-linear) models, depending on the functional form of $f_z(x_i)$ used. The *flexmix* R packages performs estimation of the parameters using an

Expectation-Maximization (EM) algorithm [26]. Like all implementations of the EM algorithm, *flexmix* is not guaranteed to find a global maximum likelihood, meaning that users should check for convergence across multiple initializations. In our case with a finite mixture two standard linear regression models, we found that *flexmix* converges to extremely similar parameters for each iteration of a given sample, but that the order of distributions given in each iteration is non-deterministic. Therefore, we assumed that the distribution with the greater y-intercept, meaning the for the cells with a low library complexity, we labeled distribution with higher percent of cell counts mapping to mtDNA genes as the compromised cell distribution. The parameters estimated from each mixture model was used to calculate the posterior probability of a cell coming from the compromised cell distribution. Our miQC software is available as an R/Bioconductor package under a BSD-3-Clause License at `https://github.com/greenelab/miQC`. The code used to download datasets, perform the analyses, and reproduce the figures is available at `https://github.com/greenelab/mito-filtering`.

4 Discussion

One critical assumption of our model is that mitochondrial reads are not informative in terms of biological variation. While this is true in many contexts, there are some contexts where high mitochondrial expression is biologically relevant and informative. For instance, scRNA-seq data has shown that aberrant mitochondrial activation is implicated in development of polycystic ovary syndrome (PCOS) [50]. Removing all cells with a large percentage of mitochondrial reads in a PCOS study would therefore hinder much of the downstream analyses. More broadly, metabolic shifts between oxidative phosphorylation and glycolysis, an important indicator of cell proliferation, can also increase or decrease mitochondrial expression [51, 52].

Generally, researchers are able to assess if mitochondrial expression may be relevant to their experimental question at hand. In the majority of cases, cells with a large percentage of mitochondrial reads–especially when paired with few uniquely expressed genes or low numbers of total counts (reads or UMIs)–can be reasonably interpreted as a sign of cell damage and those cells should be discarded.

Our miQC mixture model is designed for scenarios in which there are a non-trivial amount of compromised cells and the amount of compromised cells might vary across samples or experiments. For scRNA-seq data generated from archived tumor tissues, this is often the case. However, in optimal conditions where there are no or few damaged cells, the mixture model may not be able to accurately estimate parameters for the compromised cell distribution, as there might only be a handful of compromised cells. In this case, the model is thus liable to choose very similar parameters for the two distributions, causing the probabilistic assignments for individual cells to be unstable and good cells to be excluded unnecessarily. As an example, our tumor sample EOC50 (**Figure 3D**) had no cells with an extremely high mitochondrial fraction, meaning the intercept for the "compromised" cell distribution was fitted at a much lower value than the other tumors. In this case, miQC actually excluded more cells than a simple 10% mitochondrial threshold did. With this in mind, for cases with no concerns about tissue quality, we recommend using Median Absolute Deviation (MAD) as a data-driven approach for filtering out a small number of damaged cells [53]. We also caution against using miQC on data that has already been filtered by some prior preprocessing step, and recommend users of miQC be aware of any filtering that has been done on their data, especially in the case of public datasets.

It is possible that not only tissue types may have different baselines of mitochondrial expression, but that the baselines would also vary across the cell types within a heterogeneous tissue, such as a dissociated tumor [15]. This suggests an extension of our miQC approach for future development where the intact/compromised distribution parameters could be estimated for each cell type independently. However, in most scRNA-seq experiments involving tissues, cell type identities are not known *a priori* and cannot be determined without first performing quality control. Stratifying by cell type is thus currently not advisable for the main uses of miQC.

In conclusion, ensuring the quality of scRNA-seq data is essential for robust and accurate transcriptomic analyses. Percent of reads mapping to the mitochondria is a very useful proxy for cell damage, but existing QC methods do not do justice to the myriad of biological and experimental factors relevant to mitochondrial expression. The standard wisdom of removing all cells with greater than 5% (or 10%) mitochondrial counts is unnecessarily stringent in many tissue types, especially cancer tissues, causing a massive loss of potentially informative cells. Our new method, miQC, offers a probabilistic approach to identifying high-quality cells within

an individual sample, based on the assumption that there are both intact and compromised cells within the samples with associated characteristics. This method is flexible and adaptive across experimental platforms, organism and tissue types, and disease states. It is robust to technical differences that alter standard QC metrics, such as differences in reference genome. It also maximizes the information gain from an individual experiment, often preserving hundreds or thousands of potentially informative cells that would be thrown out by uniform QC approaches. miQC is now available as an user-friendly R package available at `https://github.com/greenelab/miQC`, allowing researchers to tailor their QC to the needs of a given scRNA-seq dataset and experiment in a consistent way.

Acknowledgements

We thank John Wherry for consultation on ovarian cancer biology, as well as members of Greene Lab for scientific feedback, particularly Alexandra Lee for code review. We thank the High-Throughput Genomics Shared Resource at the Huntsman Cancer Institute at University of Utah for assistance with data generation. We would also like to thank Johanna Hynninen (Turku University Hospital), as well as Katja Kaipio, Kaisa Huhtinen, Tarja Lamminen and Naziha Mansuri (University of Turku) for the surgery and pre-processing of University of Helsinki HGSOC samples, respectively.

Author Contributions

We use the CRediT taxonomy to define author contributions:

- AAH: Methodology, Software, Formal analysis, Investigation, Data curation, Writing - Original Draft, Review & Editing, Visualization
- MMF: Resources, Software, Investigation, Writing - Review & Editing
- LMW: Software, Investigation, Writing - Review & Editing
- EPE: Resources, Investigation, Writing - Review & Editing
- KZ: Investigation
- AV: Supervision, Funding acquisition, Writing - Review & Editing
- JAD: Resources, Writing - Review & Editing, Funding acquisition
- CSG: Resources, Writing - Review & Editing, Supervision, Funding acquisition
- SCH: Conceptualization, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding acquisition

Funding

AAH, LMW, JAD, CSG, and SCH were supported by the National Institutes of Health grant from the National Cancer Institute R01CA237170. MMF, EPE, KZ, AV were supported by the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 667403 for HERCULES (Comprehensive Characterization and Effective Combinatorial Targeting of High-Grade Serous Ovarian Cancer via Single-Cell Analysis), the Academy of Finland (Projects No.289059, 319243 and 294023), the Sigrid Jusélius Foundation, and the Cancer Foundation Finland.

Competing Interest Statement

The authors declare that they have no competing interests.

References

[1] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017. ISSN 2041-1723. doi:10.1038/ncomms14049.

[2] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, 65(4):631–643.e4, February 2017. ISSN 1097-2765. doi:10.1016/j.molcel.2017.01.023. URL `http://www.sciencedirect.com/science/article/pii/S1097276517300497`.

[3] Mario L. Suvà and Itay Tirosh. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Molecular Cell*, 75(1):7–12, July 2019. ISSN 1097-2765. doi:10.1016/j.molcel.2019.05.003. URL `https://www.cell.com/molecular-cell/abstract/S1097-2765(19)30356-9`.

[4] Michal Slyper, Caroline B. M. Porter, Orr Ashenberg, Julia Waldman, Eugene Drokhlyansky, Isaac Wakiro, Christopher Smillie, Gabriela Smith-Rosario, Jingyi Wu, Danielle Dionne, Sébastien Vigneau, Judit Jané-Valbuena, Timothy L. Tickle, Sara Napolitano, Mei-Ju Su, Anand G. Patel, Asa Karlstrom, Simon Gritsch, Masashi Nomura, Avinash Waghray, Satyen H. Gohil, Alexander M. Tsankov, Livnat Jerby-Arnon, Ofir Cohen, Johanna Klughammer, Yanay Rosen, Joshua Gould, Lan Nguyen, Matan Hofree, Peter J. Tramontozzi, Bo Li, Catherine J. Wu, Benjamin Izar, Rizwan Haq, F. Stephen Hodi, Charles H. Yoon, Aaron N. Hata, Suzanne J. Baker, Mario L. Suvà, Raphael Bueno, Elizabeth H. Stover, Michael R. Clay, Michael A. Dyer, Natalie B. Collins, Ursula A. Matulonis, Nikhil Wagle, Bruce E. Johnson, Asaf Rotem, Orit Rozenblatt-Rosen, and Aviv Regev. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nature Medicine*, 26(5):792–802, 2020. ISSN 1078-8956. doi:10.1038/s41591-020-0844-1. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7220853/`.

[5] Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, March 2015. ISSN 1471-0064. doi:10.1038/nrg3833. URL `http://www.nature.com/articles/nrg3833`. Number: 3 Publisher: Nature Publishing Group.

[6] Peng Jiang. Quality Control of Single-Cell RNA-seq. *Methods in Molecular Biology (Clifton, N.J.)*, 1935: 1–9, 2019. ISSN 1940-6029. doi:10.1007/978-1-4939-9057-3_1.

[7] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, February 2014. ISSN 1548-7105. doi:10.1038/nmeth.2772. URL `http://www.nature.com/articles/nmeth.2772`. Number: 2 Publisher: Nature Publishing Group.

[8] Geng Chen, Baitang Ning, and Tieliu Shi. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. doi:10.3389/fgene.2019.00317. URL `https://www.frontiersin.org/articles/10.3389/fgene.2019.00317/full`. Publisher: Frontiers.

[9] Elena Denisenko, Belinda B. Guo, Matthew Jones, Rui Hou, Leanne de Kock, Timo Lassmann, Daniel Poppe, Olivier Clément, Rebecca K. Simmons, Ryan Lister, and Alistair R. R. Forrest. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biology*, 21(1):1–25, December 2020. ISSN 1474-760X. doi:10.1186/s13059-020-02048-6. URL `https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02048-6`. Number: 1 Publisher: BioMed Central.

[10] Lorenzo Galluzzi, Oliver Kepp, and Guido Kroemer. Mitochondria: master regulators of danger signalling. *Nature Reviews Molecular Cell Biology*, 13(12):780–788, December 2012. ISSN 1471-0080. doi:10.1038/nrm3479. URL https://www.nature.com/articles/nrm3479.

[11] Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C. Marioni, and Sarah A. Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17(1):29, February 2016. ISSN 1474-760X. doi:10.1186/s13059-016-0888-1. URL https://doi.org/10.1186/s13059-016-0888-1.

[12] Soeren Lukassen, Elisabeth Bosch, Arif B. Ekici, and Andreas Winterpacht. Single-cell RNA sequencing of adult mouse testes. *Scientific Data*, 5, September 2018. ISSN 2052-4463. doi:10.1038/sdata.2018.192. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6132189/.

[13] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019. ISSN 1744-4292. doi:10.15252/msb.20188746. URL https://www.embopress.org/doi/full/10.15252/msb.20188746. Publisher: John Wiley & Sons, Ltd.

[14] Aisha A. AlJanahi, Mark Danielsen, and Cynthia E. Dunbar. An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Molecular Therapy. Methods & Clinical Development*, 10:189–196, August 2018. ISSN 2329-0501. doi:10.1016/j.omtm.2018.07.003. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6072887/.

[15] Daniel Osorio and James J. Cai. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics*, August 2020. doi:10.1093/bioinformatics/btaa751. URL https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa751/5896986.

[16] Christian T. Wohnhaas, Germán G. Leparc, Francesc Fernandez-Albert, David Kind, Florian Gantner, Coralie Viollet, Tobias Hildebrandt, and Patrick Baum. DMSO cryopreservation is the method of choice to preserve cells for droplet-based single-cell RNA sequencing. *Scientific Reports*, 9(1):1–14, July 2019. ISSN 2045-2322. doi:10.1038/s41598-019-46932-z. URL https://www.nature.com/articles/s41598-019-46932-z.

[17] Roshan M. Kumar, Patrick Cahan, Alex K. Shalek, Rahul Satija, AJay DaleyKeyser, Hu Li, Jin Zhang, Keith Pardee, David Gennert, John J. Trombetta, Thomas C. Ferrante, Aviv Regev, George Q. Daley, and James J. Collins. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*, 516 (7529):56–61, December 2014. ISSN 1476-4687. doi:10.1038/nature13920.

[18] Elisabetta Mereu, Atefeh Lafzi, Catia Moutinho, Christoph Ziegenhain, Davis J. McCarthy, Adrián Álvarez Varela, Eduard Batlle, Sagar, Dominic Grün, Julia K. Lau, Stéphane C. Boutet, Chad Sanada, Aik Ooi, Robert C. Jones, Kelly Kaihara, Chris Brampton, Yasha Talaga, Yohei Sasagawa, Kaori Tanaka, Tetsutaro Hayashi, Caroline Braeuning, Cornelius Fischer, Sascha Sauer, Timo Trefzer, Christian Conrad, Xian Adiconis, Lan T. Nguyen, Aviv Regev, Joshua Z. Levin, Swati Parekh, Aleksandar Janjic, Lucas E. Wange, Johannes W. Bagnoli, Wolfgang Enard, Marta Gut, Rickard Sandberg, Itoshi Nikaido, Ivo Gut, Oliver Stegle, and Holger Heyn. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature Biotechnology*, 38(6):747–755, June 2020. ISSN 1546-1696. doi:10.1038/s41587-020-0469-4. URL http://www.nature.com/articles/s41587-020-0469-4. Number: 6 Publisher: Nature Publishing Group.

[19] Marcus Alvarez, Elior Rahmani, Brandon Jew, Kristina M. Garske, Zong Miao, Jihane N. Benhammou, Chun Jimmie Ye, Joseph R. Pisegna, Kirsi H. Pietiläinen, Eran Halperin, and Päivi Pajukanta. Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. *bioRxiv*, page 786285, October 2019. doi:10.1101/786285. URL https://www.biorxiv.org/content/10.1101/786285v2. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[20] Jonathan A. Griffiths, Antonio Scialdone, and John C. Marioni. Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology*, 14(4):e8046, April 2018.

ISSN 1744-4292. doi:10.15252/msb.20178046. URL https://www.embopress.org/doi/abs/10.15252/msb.20178046. Publisher: John Wiley & Sons, Ltd.

[21] Lukas M. Weber, Ariel A. Hippen, Peter F. Hickey, Kristofer C. Berrett, Jason Gertz, Jennifer Anne Doherty, Casey S. Greene, and Stephanie C. Hicks. Genetic demultiplexing of pooled single-cell RNA-sequencing samples in cancer facilitates effective experimental design. *bioRxiv*, page 2020.11.06.371963, November 2020. doi:10.1101/2020.11.06.371963. URL https://www.biorxiv.org/content/10.1101/2020.11.06.371963v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[22] Yuan Yuan, Young Seok Ju, Youngwook Kim, Jun Li, Yumeng Wang, Christopher J. Yoon, Yang Yang, Inigo Martincorena, Chad J. Creighton, John N. Weinstein, Yanxun Xu, Leng Han, Hyung-Lae Kim, Hidewaki Nakagawa, Keunchil Park, Peter J. Campbell, and Han Liang. Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nature Genetics*, 52(3):342–352, March 2020. ISSN 1546-1718. doi:10.1038/s41588-019-0557-x. URL https://www.nature.com/articles/s41588-019-0557-x. Number: 3 Publisher: Nature Publishing Group.

[23] Davis J McCarthy, Kieran R Campbell, Aaron T L Lun, and Quin F Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186, April 2017. ISSN 1367-4803. doi:10.1093/bioinformatics/btw777. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5408845/.

[24] Robert A Amezquita, Aaron T L Lun, Etienne Becht, Vince J Carey, Lindsay N Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, Levi Waldron, Hervé Pagès, Mike L Smith, Wolfgang Huber, Martin Morgan, Raphael Gottardo, and Stephanie C Hicks. Orchestrating single-cell analysis with Bioconductor. *Nat Methods*, Dec 2019. doi:10.1038/s41592-019-0654-x.

[25] Wayne S. DeSarbo and William L. Cron. A maximum likelihood methodology for clusterwise linear regression. *J. Classif.*, 5(2):249–282, 1988. ISSN 0176-4268; 1432-1343/e.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 0035-9246. URL https://www.jstor.org/stable/2984875. Publisher: [Royal Statistical Society, Wiley].

[27] Friedrich Leisch. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, 11(1):1–18, October 2004. ISSN 1548-7660. doi:10.18637/jss.v011.i08. URL https://www.jstatsoft.org/index.php/jss/article/view/v011i08. Number: 1.

[28] Tim R. Mercer, Shane Neph, Marcel E. Dinger, Joanna Crawford, Martin A. Smith, Anne-Marie J. Shearwood, Eric Haugen, Cameron P. Bracken, Oliver Rackham, John A. Stamatoyannopoulos, Aleksandra Filipovska, and John S. Mattick. The Human Mitochondrial Transcriptome. *Cell*, 146(4):645–658, August 2011. ISSN 0092-8674. doi:10.1016/j.cell.2011.06.051. URL http://www.sciencedirect.com/science/article/pii/S0092867411007677.

[29] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015. ISSN 1097-4172. doi:10.1016/j.cell.2015.05.002.

[30] Karthik Shekhar, Sylvain W. Lapan, Irene E. Whitney, Nicholas M. Tran, Evan Z. Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z. Levin, James Nemesh, Melissa Goldman, Steven A. McCarroll, Constance L. Cepko, Aviv Regev, and Joshua R. Sanes. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*, 166(5):1308–1323.e30, August 2016. ISSN 0092-8674, 1097-4172. doi:10.1016/j.cell.2016.07.054. URL https://www.cell.com/cell/abstract/S0092-8674(16)31007-8. Publisher: Elsevier.

[31] Arianne C. Richard, Aaron T. L. Lun, Winnie W. Y. Lau, Berthold Göttgens, John C. Marioni, and Gillian M. Griffiths. T cell cytolytic capacity is independent of initial stimulation strength. *Nature Immunology*, 19(8):849–858, August 2018. ISSN 1529-2916. doi:10.1038/s41590-018-0160-9.

[32] Amit Zeisel, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (New York, N.Y.)*, 347(6226):1138–1142, March 2015. ISSN 1095-9203. doi:10.1126/science.aaa1934.

[33] N Lawlor, J George, M Bolisetty, R Kursawe, L Sun, V Sivakamasundari, I Kycia, P Robson, and ML Stitzel. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Research*, 27(2):208–222, 11 2016. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.212720.116. URL `https://europepmc.org/article/MED/27864352`.

[34] Wanxin Wang, Felipe Vilella, Pilar Alama, Inmaculada Moreno, Marco Mignardi, Alina Isakova, Wenying Pan, Carlos Simon, and Stephen R. Quake. Single-cell transcriptomic atlas of the human endometrium during the menstrual cycle. *Nature Medicine*, 26(10):1644–1653, October 2020. ISSN 1546-170X. doi:10.1038/s41591-020-1040-z. URL `http://www.nature.com/articles/s41591-020-1040-z`. Number: 10 Publisher: Nature Publishing Group.

[35] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, 11(10):733–9, 2010. doi:10.1038/nrg2825.

[36] Jeffrey L. Boore. Animal mitochondrial genomes. *Nucleic Acids Research*, 27(8):1767–1780, April 1999. ISSN 0305-1048. doi:10.1093/nar/27.8.1767. URL `https://doi.org/10.1093/nar/27.8.1767`.

[37] Avi Srivastava, Laraib Malik, Tom Smith, Ian Sudbery, and Rob Patro. Alevin efficiently estimates accurate gene abundances from dscrna-seq data. *Genome biology*, 20(1):65, 2019.

[38] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, January 2019. ISSN 0305-1048. doi:10.1093/nar/gky955. URL `https://doi.org/10.1093/nar/gky955`.

[39] Ralf Schulze Bruening, Lukas Tombor, Marcel H. Schulz, Stefanie Dimmeler, and David John. Comparative analysis of common alignment tools for single cell rna sequencing. *bioRxiv*, 2021. doi:10.1101/2021.02.15.430948. URL `https://www.biorxiv.org/content/early/2021/02/16/2021.02.15.430948`.

[40] Shan Gao, Xiaoxuan Tian, Hong Chang, Yu Sun, Zhenfeng Wu, Zhi Cheng, Pengzhi Dong, Qiang Zhao, Jishou Ruan, and Wenjun Bu. Two novel lncRNAs discovered in human mitochondrial DNA using PacBio full-length transcriptome data. *Mitochondrion*, 38:41–47, January 2018. ISSN 1567-7249. doi:10.1016/j.mito.2017.08.002. URL `http://www.sciencedirect.com/science/article/pii/S1567724917301058`.

[41] Pierre-Luc Germain, Anthony Sonrel, and Mark D. Robinson. pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. *Genome Biology*, 21(1):1–28, December 2020. ISSN 1474-760X. doi:10.1186/s13059-020-02136-7. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02136-7. Number: 1 Publisher: BioMed Central.

[42] James Melville Leland McInnes, John Healy. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, 2020. URL https://arxiv.org/abs/1802.03426.

[43] D. Sculley. Web-Scale k-Means Clustering. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 1177–1178, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi:10.1145/1772690.1772862. URL https://doi.org/10.1145/1772690.1772862.

[44] Stephanie C. Hicks, Ruoxi Liu, Yuwei Ni, Elizabeth Purdom, and Davide Risso. mbkmeans: fast clustering for single cell data using mini-batch k-means. *bioRxiv*, page 2020.05.27.119438, 05 2020. doi:10.1101/2020.05.27.119438. URL https://www.biorxiv.org/content/10.1101/2020.05.27.119438v1.

[45] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, CA, 1967. University of California Press. URL https://projecteuclid.org/euclid.bsmsp/1200512992.

[46] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2346830.

[47] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Trans. Information Theory*, 28:129–136, 1982.

[48] Cole M Risso D. *Collection of Public Single-Cell RNA-Seq Datasets*, 2020. URL https://bioconductor.org/packages/scRNAseq.

[49] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

[50] Lingbin Qi, Boxuan Liu, Xian Chen, Qiwei Liu, Wanqiong Li, Bo Lv, Xiaoyu Xu, Lu Wang, Qiao Zeng, Jinfeng Xue, and Zhigang Xue. Single-Cell Transcriptomic Analysis Reveals Mitochondrial Dynamics in Oocytes of Patients With Polycystic Ovary Syndrome. *Frontiers in Genetics*, 11, 2020. ISSN 1664-8021. doi:10.3389/fgene.2020.00396. URL https://www.frontiersin.org/articles/10.3389/fgene.2020.00396/full. Publisher: Frontiers.

[51] Hessel Honkoop, Dennis EM de Bakker, Alla Aharonov, Fabian Kruse, Avraham Shakked, Phong D Nguyen, Cecilia de Heus, Laurence Garric, Mauro J Muraro, Adam Shoffner, Federico Tessadori, Joshua Craiger Peterson, Wendy Noort, Alberto Bertozzi, Gilbert Weidinger, George Posthuma, Dominic Grün, Willem J van der Laarse, Judith Klumperman, Richard T Jaspers, Kenneth D Poss, Alexander van Oudenaarden, Eldad Tzahor, and Jeroen Bakkers. Single-cell analysis uncovers that metabolic reprogramming by ErbB2 signaling is essential for cardiomyocyte proliferation in the regenerating heart. *eLife*, 8:e50163, December 2019. ISSN 2050-084X. doi:10.7554/eLife.50163. URL https://doi.org/10.7554/eLife.50163. Publisher: eLife Sciences Publications, Ltd.

[52] Ed Reznik, Qingguo Wang, Konnor La, Nikolaus Schultz, and Chris Sander. Mitochondrial respiratory gene expression is suppressed in many cancers. *eLife*, 6:e21592, January 2017. ISSN 2050-084X. doi:10.7554/eLife.21592. URL https://doi.org/10.7554/eLife.21592. Publisher: eLife Sciences Publications, Ltd.

[53] Aaron T.L. Lun, Davis J. McCarthy, and John C. Marioni. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, 5, October 2016. ISSN 2046-1402. doi:10.12688/f1000research.9501.2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5112579/.

Supplementary Figures

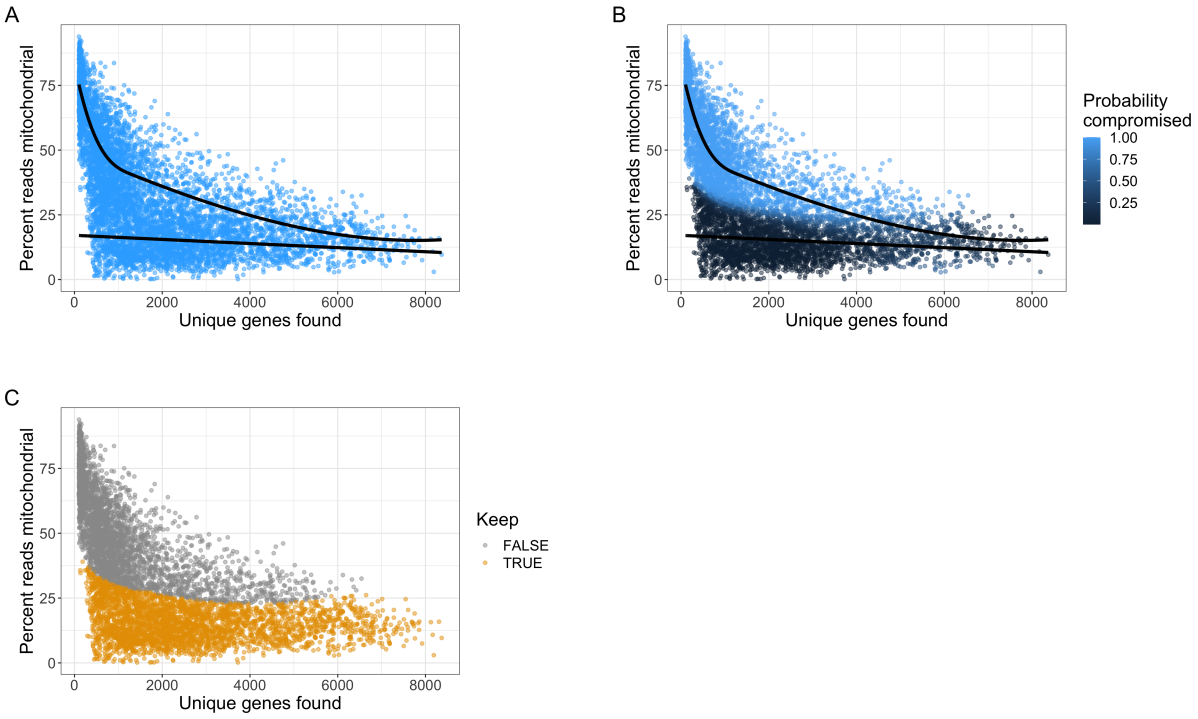

Figure S1: **miQC is extensible to a combination of linear and non-linear models.** **(A)** A mixture model on high-grade serous ovarian tumor data, where the intact cell distribution is modeled linearly and the compromised cell distribution is modeled using a b-spline. **(B)** Posterior likelihood of tumor cells belonging to compromised distribution as fitted with a spline model. **(C)** All cells with greater than 75% posterior probability of being compromised are marked for removal.

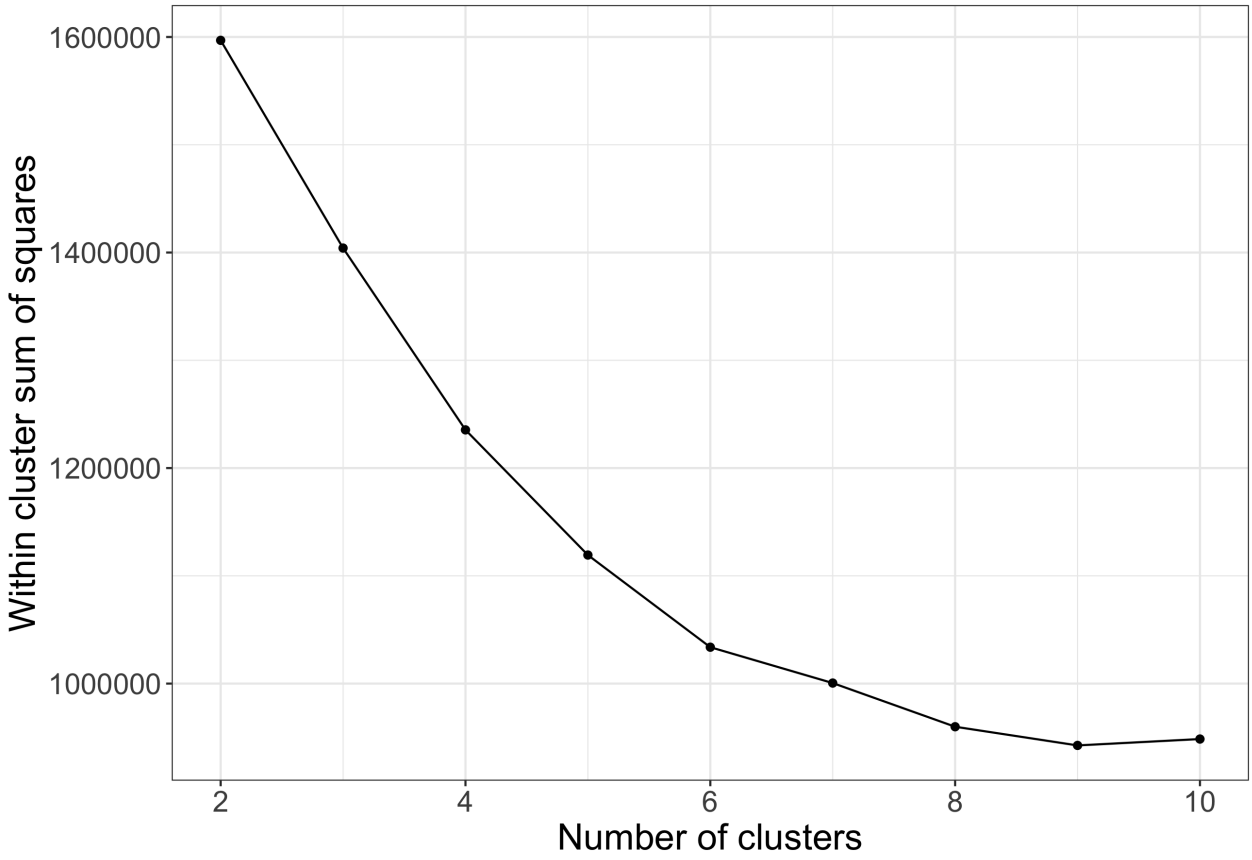

Figure S2: **Selecting an appropriate number of clusters for tumor data** We ran mbkmeans on our tumor data for a range of k from 2 to 10. Based on the within cluster sum of squares (WCSS), we proceeded with 6 clusters.