QClus: Robust and reliable preprocessing method for human heart snRNA-seq

Eloi Schmauch[1,2,3]*, Johannes Ojanen[1,2,3]*, Kyriakitsa Galani[1,2], Juho Jalkanen[4], Maija Hollmen[7], Jarmo Gunn[4,5], Tuomas Kiviniemi[4,5,6], Minna Kaikkonen-Määttä[3], Manolis Kellis[1,2], and Suvi Linna-Kuosmanen[1,2,3].

1. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA.
2. Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA.
3. A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, 70211 Kuopio, Finland.
4. Heart Center, Turku University Hospital, 20521 Turku, Finland.
5. Department of Medicine, University of Turku, 20500 Turku, Finland.
6. Cardiovascular Medicine and Network Medicine Division, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.
7. Medicity Research Laboratories, University of Turku, 20500 Turku, Finland.

* These authors contributed equally to this work

Abstract

Single nuclei RNA sequencing (snRNA-seq) is widely used to study tissues and diseases. However, the technique remains challenging, as cytoplasmic RNA often contaminates nuclei-containing droplets and may even complicate the removal of empty droplets. Incomplete removal of contaminated background signal masks cell type-specific signal and interferes with differential gene expression analysis. Thus, removing empty droplets and highly contaminated nuclei is of paramount importance in snRNA-seq analysis. This is especially the case in solid tissues such as the heart. Here, we present QClus, a novel nuclei filtering method targeted to human heart samples. In the human heart, most of the contamination originates from cytoplasmic RNA, stemming from cardiomyocytes. Therefore, we use specific metrics such as splicing, mitochondrial gene expression, nuclear gene expression, and non-cardiomyocyte and cardiomyocyte marker gene expression to cluster nuclei and filter empty and highly contaminated droplets. This approach combined with other filtering steps enables for flexible, automated, and reliable cleaning of samples with varying number of nuclei, quality, and contamination levels. To confirm the robustness of our method, we ran QClus on our dataset of 32 human heart samples and compared results to six alternative methods. None of the methods resulted in the same filtering quality as QClus, each of them failing significantly in some samples. As snRNA-seq is predominantly used on unique human tissue samples, sample failure due to inadequate data processing seems unacceptable. Our study shows that, instead of using universal preprocessing methods, challenging tissue types, such as heart tissue, may benefit from methods that are more specific to the tissue under study, as QClus outperforms the other methods primarily due to improved integration of sample type-specific features.

Introduction

Single-cell RNA sequencing enables the quantification of the transcriptomes of large, heterogeneous collections of cells[1,2]. Single-nucleus RNA sequencing (snRNA-seq) uses the same principle, but isolates nuclei instead of cells[3], thus being more suited to solid and frozen tissues[4].

A critical component of RNA-seq workflows is quality control. Droplet-based snRNA-seq works by encapsulating single nuclei in droplets, where each droplet contains RNA originating from one nucleus. However, other cellular RNA, such as cytoplasmic or cell-free RNA from the input solution can contaminate the droplets. This contamination is often an important issue for solid tissues since it can lead to spurious cell type and cell state identification[5]. In addition to droplets containing both nuclei and ambient RNA contamination, snRNA-seq experiments generate droplets that contain only the ambient contamination without nuclei. Accurate removal of both empty and highly contaminated droplets from the data is important for improving results of downstream analyses, such as differential expression analysis, which can be sensitive to noise.

Standard methods for empty droplet removal in snRNA-seq are often a combination of UMI distribution filtering (usually included in preprocessing programs, such as Cell Ranger) and of mitochondrial fraction-based filtering [5–10]. The filtering is based on visualizations that rely on clear separation of droplets containing nuclei from empty droplets, which is usually true for samples with normal levels of contamination. However, there are many tissue types for which the level of cytoplasmic contamination is too high for standard approaches to easily remove empty droplets[5]. To answer this, a variety of bioinformatics tools have been developed in recent years which either identify and remove empty droplets and highly contaminated cells or profile and remove contamination directly[5,6,11–13].

In this study, we sought to find the best way to process cardiac data, a tissue type that is especially challenging due to contamination problems. After examining existing, published methods[5,6,11–15], and testing them with our heart dataset, we decided to develop a novel approach for contamination-based nuclei filtering in snRNA-seq for cardiac tissue to improve the data quality. Here, we present Quality Clustering, QClus, a novel algorithm for nuclei filtering that is tailored for cardiac tissue and based on unsupervised clustering of general and cell-type-specific quality metrics. Comparison of QClus to the common practice of *a priori* inclusion criteria and the previously published algorithms, revealed improved results on quality metrics, cell type heterogeneity and robustness, across all samples.

Results

QClus efficiently removes empty droplets and highly contaminated nuclei

To improve the quality of cardiac snRNA-seq data, we set out to develop a versatile method that robustly removes empty droplets and highly contaminated nuclei, accommodating sample-to-sample variation in contamination levels and in resulting number of good quality nuclei. The resulting droplet filtering method QClus can automatically process a heart dataset (*see methods*).

Step 1: Cell Ranger. The first step of processing is accomplished with Cell Ranger[6,7] which removes a good amount of empty droplets using UMI count distribution in moderately contaminated samples without over-filtering nuclei (**Fig 1A**). We used a sample from our dataset as an example to illustrate our findings. After Cell Ranger filtering, the nuclei were found to form a star shape on the dimension reduction embedding (UMAP in **Fig 1A**), where all cell types seemed to originate from one common cluster. We hypothesized that this star center is composed of empty droplets and highly contaminated nuclei, as the cell types in the adult human heart are differentiated and therefore expected to form independent clusters.

Step 2: Initial filtering and quality metrics calculation. The second filtering step is a nuclei removal based on the detected number of genes and mitochondrial fraction. The gene-level filtering ensures that all nuclei have enough genes and reads per gene to be of significant biological interest in the downstream analyses, whereas a high mitochondrial threshold is used to remove only a handful of nuclei, which are clear outliers.

After the second step for our sample, the star shape was retained. Next, we measured key contamination-related and cell-type-specific metrics (plotted on **Fig 1B**). Identification of the six metrics for each nucleus is a key step of QClus, as they will be used in the next filtering step. The first metric is the fraction of unspliced genes observed in the nuclei. As nuclei contain higher levels of unspliced transcripts than the cytoplasm, this fraction is expected to anti-correlate with contamination. The second metric is the mitochondrial fraction, which correlates with contamination, as mitochondria are specific to the cytoplasm and their signal should not be found within the nuclei. The third metric is a nuclear-enriched gene expression fraction, which is expected to be anticorrelated with contamination, as nuclear enriched genes have a relative expression that should be lower in droplets containing more cytoplasmic RNA compared to less contaminated nuclei. The expression patterns detected from these three metrics confirmed our hypothesis that the star center (**Fig 1B**) is composed of empty and highly contaminated droplets.

Our cardiac snRNA-seq dataset showed that cardiomyocytes are one of the most abundant cell types, containing high levels of RNA, in absolute measures. Cardiomyocytes as a cell type have larger cytoplasms and higher numbers of mitochondria than any other cell in the human body, comprising 25-30% of the cardiomyocyte's volume[16–18]. Therefore, we hypothesized that the contamination originates mostly from cardiomyocyte cytoplasm and is characterized by high levels of mitochondrial RNA and RNA aligning to cardiomyocyte genes that are enriched in the cytoplasm. Thus, the fourth metric is the fraction of non-cardiomyocyte

marker gene expression, which is expected to anticorrelate with contamination in nuclei that are not cardiomyocytes.

The fifth metric is cardiomyocyte-specific gene expression fraction, which distinguishes cardiomyocyte and non-cardiomyocyte nuclei. In addition, the metric correlates with contamination in non-cardiomyocyte nuclei and reveals empty droplets, which should have high cardiomyocyte-specific gene expression fraction. The sixth metric is nuclear-enriched cardiomyocyte marker gene expression fraction, which anticorrelates with contamination in droplets containing cardiomyocyte nuclei.

Step 3: K-means clustering. The six metrics from step 2 are used as input for k-means clustering to identify four clusters (**Fig 1C**), namely 1) high quality non-cardiomyocyte (CM) nuclei, which have low contamination, low CM expression, and high nonCM expression, 2) good quality CM nuclei which have low contamination, high CM expression nuclei and cytoplasm, and low nonCM expression, 3) highly contaminated nuclei, and 4) empty droplets, both of which have high contamination, high cytoplasmic CM expression, low nuclei CM genes expression, and low nonCM expression (**Fig 1B-C**). The recommended settings instruct QClus to then remove the empty droplets cluster, but the user can choose to also remove the highly contaminated nuclei cluster, on a sample-by-sample basis.

Step 4: Outlier filtering. The fourth filtering step removes highly contaminated nuclei in a more adjustable, continuous manner by identifying outliers based on the splicing distribution and mitochondrial fractions within the sample.

Step 5: Doublet filter. In the final step, we apply the Scrublet algorithm[19], which achieves two objectives: 1) it removes remaining highly contaminated nuclei as doublets, i.e. nonCMs that have high level of contamination, as they contain RNA from both nonCMs and CMs; 2) it removes nuclei doublets.

After QClus preprocessing, the final UMAP showed a clear separation of the 11 major cell types observed in the sample (**Fig 1A, step 5**), whereas removed nuclei displayed clear signals of contamination (**Fig 1B** and **1D-F)**. No single metric alone predicted which nuclei were removed and which kept, suggesting that a multidimensional approach such as QClus should be taken during the quality control to maximize the biological signal for downstream analyses.

Benchmarking QClus

To evaluate the efficacy of QClus compared to adapted standards, we ran nuclei filtering on our 32-sample dataset with QClus and 6 other methods[5,11–13,15]. Four of the methods focus on removing empty droplets and highly contaminated nuclei[5,11–13], similar to QClus, the fifth method performs RNA decontamination providing a contamination score that can be used for nuclei filtering[15]. The last method is a simple mitochondrial threshold filtering.

Strikingly, all 6 methods failed to produce satisfying results for some of the samples in our dataset, in a clear and visual way (**Fig 2**). For a fair comparison, we did not run Scrublet in the QClus pipeline for the benchmarking, as it is not specific to our method, and we always selected

the same k-means clusters (i.e., removed the empty droplet cluster, but kept the highly contaminated one, so these could be automatically removed by the outlier filtering). This was done to get a fully unsupervised filtering algorithm. Similarly, we ran the other methods and set thresholds so that the data would be comparable, but similarly unsupervised (i.e., no manual settings of thresholds at the sample level, except for DIEM, which highly recommends setting a sample-level threshold based on a visual distribution).

The first methods we compared to QClus were EmptyNN[11], DropletQC[13], and SampleQC[12]. The two first are specifically designed to remove empty nuclei, while SampleQC is an outlier detection algorithm. However, the filtering did not remove enough nuclei for some samples: Highly contaminated nuclei / empty droplets remained (**Fig 2A**), resulting in a similar star shape that we observed for the first steps of QClus in **Fig 1A**-**B**, with very high mitochondrial percentage near the center. These highly contaminated droplets were removed with QClus, resulting in clear cell type separation, and much lower mitochondrial fraction (**Fig 2A**).

The next methods we tested, a modified version of DecontX[15] and mitochondrial fraction filtering, did not succeed. DecontX is designed to evaluate and remove contamination from single nuclei transcriptomics expression and is not per se a filtering method. The method can however be used to extract a contamination metric for each nucleus, which we then use as a threshold-based nuclei filtering method. Mitochondrial fraction filtering is often used for moderate contamination filtering. Both methods require a filtering threshold to be set. To be able to compare the resulting quality of the compared methods (modified DecontX, mitochondrial filtering, QClus), we set the threshold such that the total number of nuclei retained was the same for the whole dataset. Both DecontX and mitochondrial filtering over-filtered nuclei in some samples, removing entire cell types. For example, in one sample (**Fig 2B**), fibroblasts and macrophages were removed with DecontX, and fibroblasts and most vascular endothelial cells were filtered out with the mitochondrial method. Across samples, we discovered multiple cases where DecontX recovered fewer cell types than QClus (**Fig 2C**).

The last method we tested was DIEM[5], which is especially tailored to remove empty droplets and highly contaminated nuclei. With the default settings of 20 clusters, the method failed in several samples by recovering far less nuclei (3701 vs. 9757), and cell types (3 vs 11) (**Fig 2D-F**). Additionally, the recovered nuclei exhibited lower overall quality.

Taken together, we observed a reliability and robustness with QClus that was not achieved with any other method.

Discussion

In this study, we have established a preprocessing method, QClus, that successfully filters and cleans snRNA-seq data from heart samples, despite significant levels of contamination. The method can accommodate heterogeneous datasets, in terms of number of nuclei, overall quality, and contamination, as it automatically adapts to the sample characteristics. It requires very little supervision and, as illustrated by our benchmarking, can even run successfully without supervision. The method is specifically tailored for heart data and uses a reasonable hypothesis that most of the contamination originates from cardiomyocyte cytoplasms, which was confirmed by our analysis.

QClus is highly customizable, as its parameters can be adjusted at the global level to change the stringency of the outlier filtering. This adjustment, when done at the level of the whole dataset, permits the filtering to be more suited to the downstream analyses and the associated biological questions which may require as many nuclei as possible, at the cost of contamination level, or vice-versa. Moreover, all parameters can be adjusted at the sample level, if the user wishes for a more precise, albeit time-consuming approach, which is more suited for a smaller number of samples.

A clear limitation of the benchmarking comparison we ran between the methods was that QClus works unsupervised at the sample level, as it is designed for automated processes that are required for larger datasets and more powerful and scalable pipelines. It is likely that some of the methods used here would have benefitted from sample-to-sample quality checks for optimal adjustment of the parameters. However, in our case, the goal was to find the best method that is suited for large datasets, thus requiring as little supervision as possible. While running the comparison for methods that required threshold setting, we considered setting the threshold based on the number of nuclei filtered by QClus in each sample instead of over the whole dataset as we chose to do. However, this would not be a relevant comparison, as the number of high-quality nuclei varies significantly between samples, and is therefore a key result of nuclei filtering methods. A method which needs this number to function does not answer the problem which QClus seeks to resolve.

There are many potential ways to improve QClus. Currently, the method runs agnostic of the identified cell types, although there is evidence that cell types are affected differently by contamination. Therefore, a finer approach that treats cell types cell-type-specifically, especially in the last steps, namely outlier and doublet removal, is called for. In addition, QClus does not remove contamination from the nuclei that are not filtered out. Combining QClus with a decontamination method, from which the calculated contamination metric could be added to the quality metrics used as QClus input, could further improve the results.

The most striking difference between QClus and the other methods is its specificity to a tissue type and the resulting superior filtering. This makes QClus a tailored and tested method for heart data. Nevertheless, its customizability allows potential adaptation to other tissue types,

such as skeletal muscle, by replacing cardiomyocyte-specific genes with skeletal muscle cell marker genes.

In conclusion, our study suggests that challenging tissue types, such as the heart tissue, may benefit from methods that are more specific to the tissue at hand, as contamination profiles can be very tissue- and disorder-specific.

Methods

We process the heart samples using cellranger 5.0 and run the bam files with velocyto. The results are output as .loom files which are read using the loompy package.

QClus Initial filter
We still observe a high amount of contamination after using the 10x Genomics filtering algorithm. This is evidenced by a significant amount of cells containing a very high fraction of mitochondrial reads, or either very low or very high library complexity. Given the high likelihood of these droplets being either empty or multiplets, we set the lower bound of the number of detected genes to be 500, the upper bound of the number of detected genes to be 6000, and the cutoff for mitochondrial fraction to be 40%.

QClus Clustering features: Unsplicing fraction
After annotating reads as “spliced”, “ambiguous”, or “unspliced” using velocyto, we calculate a “fraction_unspliced” metric for each cell. The metric is calculated by dividing the total number of unspliced reads by the total amount of reads (spliced, unspliced and ambiguous) in each cell.

QClus Clustering features: Mitochondrial fraction
We calculate the fraction of reads aligning to the mitochondrial genome (*MT-ND1, MT-ND2, MT-CO1, MT-CO2, MT-ATP8, MT-ATP6, MT-CO3, MT-ND3, MT-ND4L, MT-ND4, MT-ND5, MT-ND6, MT-CYB*). The metric is calculated using the scanpy function calculate_qc_metrics() on raw counts.

QClus Clustering features: Nuclear fraction
We calculate the fraction of reads aligning to the following genes: *MALAT1, NEAT1, FTX, FOXP1, RBMS3, ZBTB20, LRMDA, PBX1, ITPR2, AUTS2, TTC28, BNC2, EXOC4, RORA, PRKG1, ARID1B, PARD3B, GPHN, N4BP2L2, PKHD1L1, EXOC6B, FBXL7, MED13L, TBC1D5, IMMP2L, SYNE1, RERE, MBD5, EXT1, WWOX*. These were chosen because they exhibit high correlation with *MALAT1*, which is known to be highly expressed in the nucleus. The metric is calculated using the scanpy function calculate_qc_metrics() on logarithmized counts. Logarithmization is performed using the scanpy function log1p().

QClus Clustering features: Nuclear CM genes
Using differential expression analysis from our heart dataset, we selected a set of genes (*RBM20, TECRL, MLIP, CHRM2, TRDN, PALLD, SGCD, CMYA5, MYOM2, TBX5, ESRRG, LINC02248, KCNJ3, TACC2, CORIN, DPY19L2, WNK2, MITF, OBSCN, FHOD3, MYLK3, DAPK2, NEXN*) that are highly specific to CM nuclei. Cells with a high level of expression of these genes are expected to contain CM nuclei. The metric is calculated using the scanpy function calculate_qc_metrics() on raw counts.

QClus Clustering features: Cytoplasm CM genes

CM specific genes enriched in the cytoplasm are selected (*TTN, RYR2, PAM, TNNT2, RABGAP1L, PDLIM5, MYL7, MYH6*). These are genes found to be highly specific of CMs but also present in high numbers in other cell types, as contamination. The metric is calculated using the scanpy function calculate_qc_metrics() on raw counts.

QClus Clustering features: Cell type specific fractions
For each of the eleven remaining cell types, we select a set of genes, using prior knowledge, that are highly specific to that cell type. We then calculate the fraction of reads aligning to those sets of genes. These metrics are calculated using the scanpy function calculate_qc_metrics() on raw counts. Next, for each cell, we select the maximum value out of those 11 metrics (**Table 1**)

Table 1: List of nonCM cell type specific genes used to generate the clustering feature.

Vascular endothelial cells	*VWF, ERG, ANO2, PTPRB, EGFL7, PREX2, ADGRL4, FLT1, CYYR1, GRB10, PPP1R16B, DOCK9, SHANK3, PECAM1, PLEKHG1, EMCN*
Pericytes	*RGS5, DACH1, GUCY1A1, ABCC9, BGN, NOTCH3, PDGFRB, FRMD3, RNF152, CCDC102B, NGF*
Smooth muscle cells	*MYH11, KCNAB1, NTRK3, CHRM3, ACTA2, RGS6, DGKG, ITGA8, TBX2, LMOD1, SDK1, GPC6, ANTXR1, FLNA, CLMN, ATP10A, MCAM, TAGLN, CCDC3*
Adipocytes	*PLIN4, PLIN1, PDE3B, GPAM, PTPRS, PPARG, MLXIPL, MGST1, AQP7, SLC19A3, FABP4, TPRG1, DIRC3, LPL, PNPLA2, LIPE, ADH1B, ADIPOQ, PRKAR2B, CIDEA, LINC00278, PFKFB3, LINC02237, LIPE-AS1, SVEP1*
Schwann Cells	*XKR4, AC016766.1, SLC35F1, ZNF536, NCAM2, GPM6B, KIRREL3, SORCS1, ST6GALNAC5, PRKCA, GINS3, PMP22, ALDH1A1, IL1RAPL2, DOCK5, NKAIN3, COL28A1, RALGPS2, PKN2-AS1, KLHL29, PTPRZ1*
Neurons	*CSMD1, SYT1, KCNIP4, CNTNAP2, DLGAP1, PTPRD, LRRTM4, ATRNL1, LRP1B, CTNND2, KCNQ5, NRG3, SNTG1, GRIA2, RIMS2, CSMD3, XIST, KAZN, DPP10, HS6ST3, OPCML*
Endocardial endothelial cells	*PCDH7, PCDH15, LINC02147, LINC02388, MYRIP, GMDS, ADAMTSL1, LEPR, CALCRL, CGNL1, HMCN1, NPR3, POSTN*
Fibroblasts	*DCN, ABCA8, ABCA6, ABCA10, FBLN1, COL15A1, FBN1, C7*
Lymphocytes	*SKAP1, RIPOR2, CD247, IKZF1, BCL11B, SLFN12L, ITGAL, SAMD3, CARD11, CDC42SE2, CCND3*
Epicardial mesothelial cells	*C3, SULF1, AP000561.1, PRG4, GPM6A, CDON, DPP6, CCDC80, EZR, FOS, BNC1, AC245041.2, PRKD1, CYSTM1, TLL1, WT1*

Macrophages	*TBXAS1, SLC9A9, MRC1, MS4A6A, RBM47, DOCK2, MCTP1, SYK, MSR1, ATP8B4, F13A1, CD74, MS4A4E, ADAP2*

Clustering method

Once the clustering features have been computed for the remaining cells, the features are scaled using the MinMaxScaler() function implemented in the scikit-learn package. Next, we use the k-means algorithm, as implemented in the scikit-learn package, to partition the cells into four clusters. By default, the cluster demonstrating the highest mean value of splicing is removed, as this cluster is predicted to be empty droplets. However, the user can remove whichever clusters they choose. Also, the default number of clusters is 4, but that number can be changed, e.g. at the sample level for very highly contaminated samples.

Outlier filtering

Next we calculate threshold values for the fraction of unspliced reads as well as mitochondrial fraction based on their distribution in the nonCM cluster. The unspliced fraction threshold is chosen as the lower quartile minus 0.1. The mitochondrial fraction threshold is chosen as the upper quartile plus 0.05. These values can be adjusted by the user. Each cell is then compared to these values. If the unspliced fraction is lower than the unspliced threshold and the mitochondrial fraction is higher than the mitochondrial threshold, the cell is removed.

Doublet removal

We use the scrublet package with a score threshold of 0.1 to further remove highly contaminated cells as well as doublets. We used the following default parameters for scrublet: (min_counts=2, min_cells=3, min_gene_variability_pctl=85, expected_doublet_rate=0.06).

Competing methods

Mitochondrial filtering

A common practice in cardiac specific single-cell research is to set a threshold for mitochondrial fraction and filter cells that exceed that threshold. We set the threshold such that we get the same total number of cells filtered out in the dataset with QClus, which turned out to be 9.55%. This is done to have a fair comparison with the same total number of cells as outcome, and keep a similar level of non-supervision. Many snRNA-seq pipelines will use a similar threshold, as pure nuclei are not supposed to contain mitochondrial genes.

DIEM

DIEM is an approach that aims to model RNA distributions of contamination as well as cell types using a multinomial distribution. The user first sets a hard count threshold based on a barcode rank plot, and the cells below this threshold are fixed as debris. DIEM then infers the parameters of the model using expectation maximization. A score is computed for each cell,

based on their expression of genes in the debris set. The user can then filter cells using a threshold value for the debris score.

DecontX

DecontX is a bayesian method that attempts to learn the distribution of background contamination as well as the amount of contamination in each cell. We take the estimated amount of contamination and use it as a metric for cell calling. As with mitochondrial filtering, we set a global threshold so that we get the same number of cells across all samples as we do with QClus, which turned out to be 0.43.

EmptyNN

EmptyNN utilizes positive-unlabeled learning to train a neural network to be able to distinguish between empty droplets and nuclei-containing droplets.

SampleQC

SampleQC is a quality control method that aims to reduce bias towards rarer cell types with significantly different values for common quality metrics, such as mitochondrial fraction. The method fits a gaussian mixture model across multiple samples and filters outliers. We use counts, library complexity, mitochondrial fraction, and splice ratio as input.

DropletQC

DropletQC uses splicing fraction and total UMIs detected to distinguish between empty droplets, cells, and damaged cells. Thresholds for these values are estimated and used to flag cells that fall above or below these thresholds. We run the identify_empty_drops function to annotate empty droplets.

References

1. Wang, Y. & Navin, N. E. Advances and applications of single-cell sequencing technologies. *Mol. Cell* **58**, 598–609 (2015).
2. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
3. N, H. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, (2017).
4. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *J. Am. Soc. Nephrol.* **30**, 23–32 (2019).
5. Alvarez, M. *et al.* Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. *Sci. Rep.* **10**, 11019 (2020).
6. Lun, A. T. L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
7. Single-Library Analysis with Cell Ranger count -Software -Single Cell Gene Expression - Official 10x Genomics Support. https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/count.
8. Zeng, W. *et al.* Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. *Nucleic Acids Res.* **44**, e158 (2016).
9. Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**, 792–802 (2020).
10. Wang, M. *et al.* Guidelines for bioinformatics of single-cell sequencing data analysis in Alzheimer's disease: review, recommendation, implementation and application. *Mol. Neurodegener.* **17**, 17 (2022).
11. Yan, F., Zhao, Z. & Simon, L. M. EmptyNN: A neural network based on positive and unlabeled learning to remove cell-free droplets and recover lost cells in scRNA-seq data. *Patterns N. Y. N* **2**, 100311 (2021).
12. Macnair, W. & Robinson, M. D. SampleQC: robust multivariate, multi-celltype, multi-sample quality control for single cell data. 2021.08.28.458012 Preprint at https://doi.org/10.1101/2021.08.28.458012 (2021).
13. Muskovic, W. & Powell, J. E. DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data. *Genome Biol.* **22**, 329 (2021).
14. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. *bioRxiv* 303727 (2018) doi:10.1101/303727.
15. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).
16. Brown, D. A. *et al.* Mitochondrial function as a therapeutic target in heart failure. *Nat. Rev. Cardiol.* **14**, 238–250 (2017).
17. Barth, E., Stämmler, G., Speiser, B. & Schaper, J. Ultrastructural quantitation of mitochondria and myofilaments in cardiac muscle from 10 different animal species including man. *J. Mol. Cell. Cardiol.* **24**, 669–681 (1992).

18. Schaper, J., Meiser, E. & Stämmler, G. Ultrastructural morphometric analysis of myocardium from dogs, rats, hamsters, mice, and from human hearts. *Circ. Res.* **56**, 377–391 (1985).
19. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).

Figure legends

Figure 1: QClus successfully removes contaminated and empty droplets, improving the quality of heart snRNA-seq samples (an example shown here). A. The five main filtering steps which compose the QClus pipeline. Below each step, nuclei that are kept are presented in a UMAP embedding, colored by cell type. **B.** The six quality metrics, used in the k-means clustering (step 3), shown on the UMAP embedding of nuclei after initial filtering (step 2) **C.** UMAP embedding of the example sample after initial filtering (step 2), colored by the clusters identified in the k-means clustering step. **D.** QClus filtering result, plotted on a UMAP embedding of the example sample after the initial filter (step 2). **E.** Relationship of the three main quality features related to contamination (Splicing, mitochondrial fraction, nuclear enrichment), colored by QClus filtering outcome. **F.** Distribution of the three main quality features related to contamination (Splicing, mitochondrial fraction, nuclear enrichment), and total number of UMI (Counts), between nuclei that are filtered in and filtered out by QClus.

Figure 2: QClus Benchmarking. A. Comparison of QClus, SampleQC, EmptyNN, and DropletQC in sample 174. For each filtering method, the resulting UMAP embedding is shown, annotated with the number of nuclei that passed filtering, and colored by cell type (above) and by mitochondrial fraction (bellow). **B.** Comparison of QClus, modified DecontX and mitochondrial-based filtering, in sample 234. For each method, the resulting UMAP embedding is shown, annotated with the number of nuclei that passed filtering, and colored by cell type (colors are same as in panel A). **C.** Comparison between QClus and DecontX of the number of cell types present after filtering, in all samples. **D.** Comparison of Qclus and Diem in sample 170. For both methods, the UMAP embedding resulting after filtering is shown, annotated with the number of nuclei that passed filtering and colored by cell type (colors are same as in panel A). **E.** Comparison of the mean mitochondrial fraction of nuclei after filtering by QClus and by Diem, across samples. **F.** Comparison of the mean unspliced read fraction after filtering by QClus and by Diem, across samples.

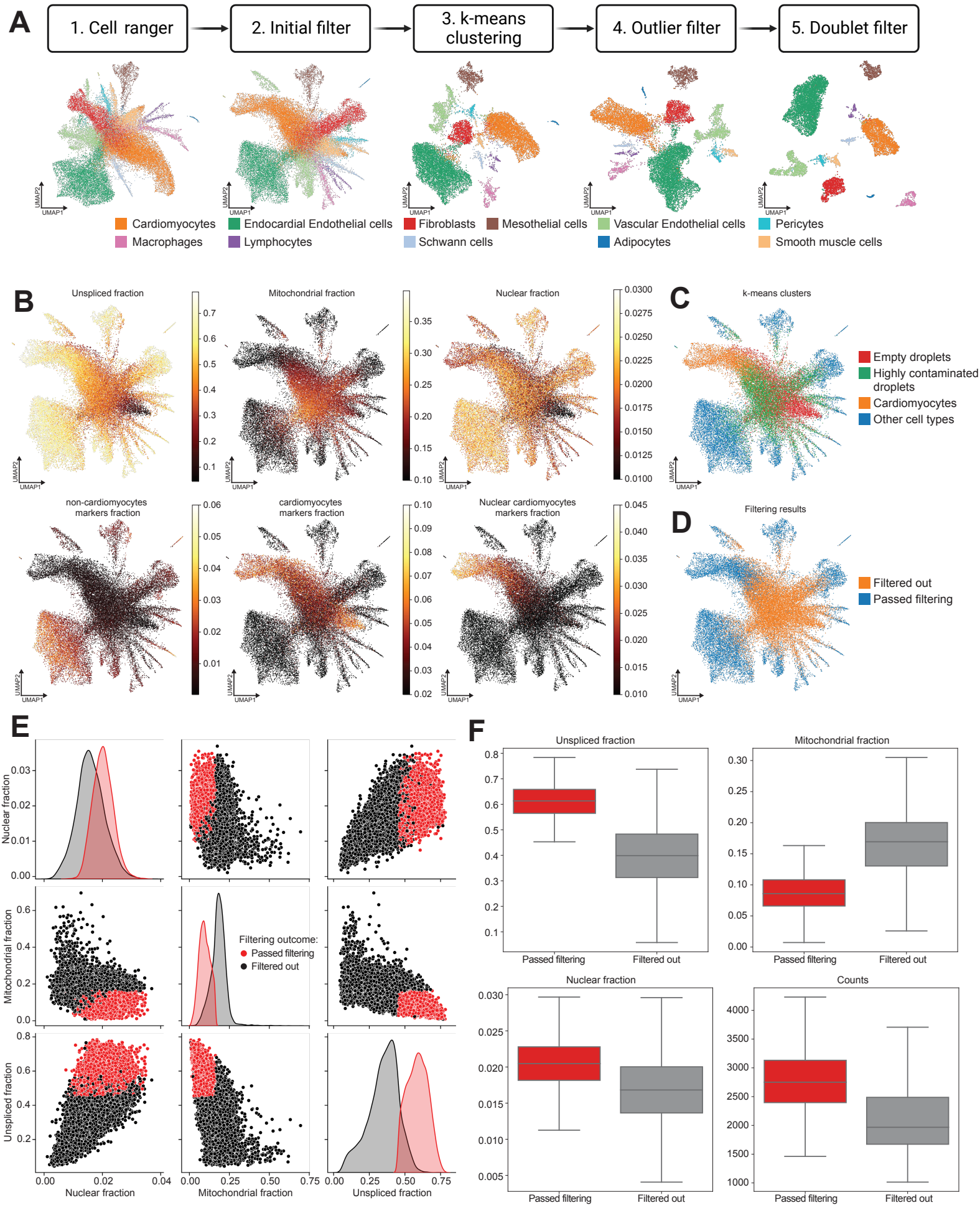
A
1. Cell ranger
2. Initial filter
3. k-means clustering
4. Outlier filter
5. Doublet filter
UMAP1
UMAP2
Cardiomyocytes
Endocardial Endothelial cells
Fibroblasts
Mesothelial cells
Vascular Endothelial cells
Pericytes
Macrophages
Lymphocytes
Schwann cells
Adipocytes
Smooth muscle cells
B
Unspliced fraction
Mitochondrial fraction
Nuclear fraction
non-cardiomyocytes markers fraction
cardiomyocytes markers fraction
Nuclear cardiomyocytes markers fraction
C
k-means clusters
Empty droplets
Highly contaminated droplets
Cardiomyocytes
Other cell types
D
Filtering results
Filtered out
Passed filtering
E
Nuclear fraction
Mitochondrial fraction
Unspliced fraction
Filtering outcome:
Passed filtering
Filtered out
F
Unspliced fraction
Mitochondrial fraction
Nuclear fraction
Counts
Passed filtering
Filtered out

Figure 1

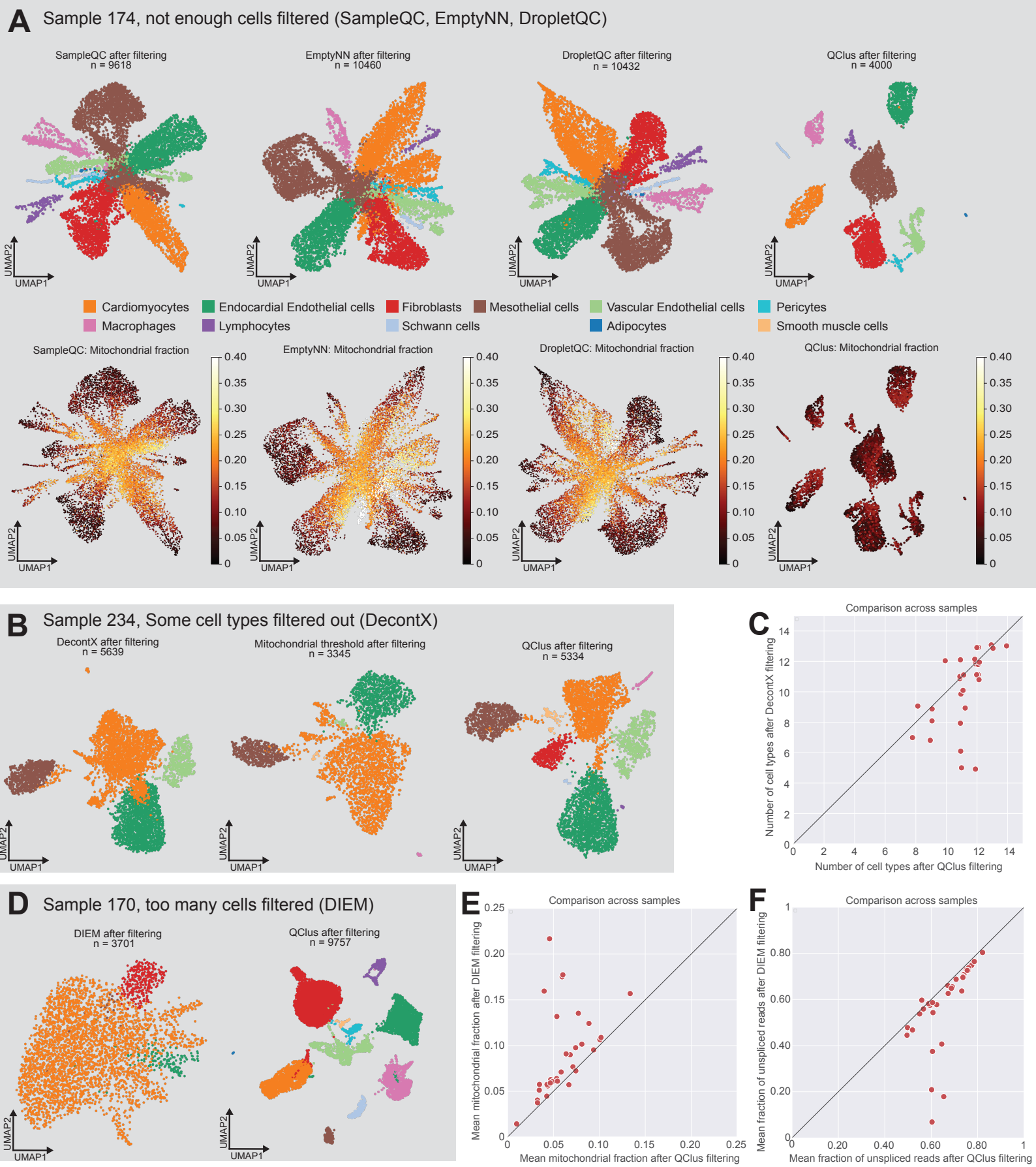
A
Sample 174, not enough cells filtered (SampleQC, EmptyNN, DropletQC)
SampleQC after filtering
n = 9618
EmptyNN after filtering
n = 10460
DropletQC after filtering
n = 10432
QClus after filtering
n = 4000
UMAP1
UMAP2
Cardiomyocytes
Endocardial Endothelial cells
Fibroblasts
Mesothelial cells
Vascular Endothelial cells
Pericytes
Macrophages
Lymphocytes
Schwann cells
Adipocytes
Smooth muscle cells
SampleQC: Mitochondrial fraction
EmptyNN: Mitochondrial fraction
DropletQC: Mitochondrial fraction
QClus: Mitochondrial fraction
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
B
Sample 234, Some cell types filtered out (DecontX)
DecontX after filtering
n = 5639
Mitochondrial threshold after filtering
n = 3345
QClus after filtering
n = 5334
C
Comparison across samples
Number of cell types after DecontX filtering
Number of cell types after QClus filtering
D
Sample 170, too many cells filtered (DIEM)
DIEM after filtering
n = 3701
QClus after filtering
n = 9757
E
Comparison across samples
Mean mitochondrial fraction after DIEM filtering
Mean mitochondrial fraction after QClus filtering
F
Comparison across samples
Mean fraction of unspliced reads after DIEM filtering
Mean fraction of unspliced reads after QClus filtering

Figure 2