



# scMitoMut for calling mitochondrial lineage-related mutations in single cells

Wenjie Sun , Daphne van Ginneken , Leïla Perié\*

Institut Curie, Université PSL, Sorbonne Université, CNRS UMR168, Physique des Cellules et Cancer, 16 rue Pierre et Marie Curie, 75005 Paris, France

\*Corresponding authors. Wenjie Sun, Institut Curie, Université PSL, Sorbonne Université, CNRS UMR168, Physique des Cellules et Cancer, 16 rue Pierre et Marie Curie, 75005 Paris, France. E-mail: [sunwjie@gmail.com](mailto:sunwjie@gmail.com); Leïla Perié, Institut Curie, Université PSL, Sorbonne Université, CNRS UMR168, Physique des Cellules et Cancer, 16 rue Pierre et Marie Curie, 75005 Paris, France. E-mail: [leila.perie@curie.fr](mailto:leila.perie@curie.fr)

## Abstract

Tracing cell lineages has become a valuable tool for studying biological processes. Among the available tools for human data, mitochondrial DNA (mtDNA) has a high potential due to its ability to be used in conjunction with single-cell chromatin accessibility data, giving access to the cell phenotype. Nonetheless, the existing mutation calling tools are ill-equipped to deal with the polyploid nature of the mtDNA and lack a robust statistical framework. Here we introduce scMitoMut, an innovative R package that leverages statistical methodologies to accurately identify mitochondrial lineage-related mutations at the single-cell level. scMitoMut assigns a mutation quality  $q$ -value based on beta-binomial distribution to each mutation at each locus within individual cells, ensuring higher sensitivity and precision of lineage-related mutation calling in comparison to current methodologies. We tested scMitoMut using single-cell DNA sequencing, single-cell transposase-accessible chromatin (scATAC) sequencing, and 10× Genomics single-cell multiome datasets. Using a single-cell DNA sequencing dataset from a mixed population of cell lines, scMitoMut demonstrated superior sensitivity in identifying a small proportion of cancer cell line compared to existing methods. In a human colorectal cancer scATAC dataset, scMitoMut identified more mutations than state-of-the-art methods. Applied to 10× Genomics multiome datasets, scMitoMut effectively measured the lineage distance in cells from blood or brain tissues. Thus, the scMitoMut is a freely available, and well-engineered toolkit (<https://www.bioconductor.org/packages/devel/bioc/html/scMitoMut.html>) for mtDNA mutation calling with high memory and computational efficiency. Consequently, it will significantly advance the application of single-cell sequencing, facilitating the precise delineation of mitochondrial mutations for lineage-tracing purposes in development, tumour, and stem cell biology.

**Keywords:** mitochondrial mutation; lineage tracing; single-cell sequencing

## Introduction

Cellular DNA barcode lineage tracing uses heritable DNA sequences as markers to track descendants from the same ancestral cells and has become a valuable tool for studying biological processes [1, 2]. Cellular barcoding in humans uses retrospective methods based on genomic mutations. Using somatic genetic mutations as the markers with single-cell read-outs of genome sequencing has provided a valuable insight into numerous biological fields, including development, ageing, and the onset of cancer [3–6]. However, this method is costly and fails to concurrently capture omics data related to cellular phenotypes, such as transcriptomics and epigenomics.

As an alternative, mutations in the mitochondrial DNA (mtDNA) have been used [7–9]. This has numerous advantages over somatic mutation-based cellular barcoding methods, including the short size of mtDNA (~16 600 bp), the presence of multiple mtDNA copies (of 100s–1000s in a blood cell [10]), and the much higher mutation rate of mtDNA (which is 100–1000 times that of nuclear genomic sequences) [11, 12]. Further, as mtDNA lacks chromatin, it can be detected with a single-cell transposase-accessible chromatin sequencing (scATAC-seq) technology [8, 9]. scATAC-seq can be used in conjunction with

single-cell multiome sequencing to provide cellular phenotype information via transcriptomes or via chromatin accessibility data. However, the unique properties of mtDNA make it difficult to use methods designed for calling single-cell nuclear DNA mutations [13, 14]. Additionally, searching for lineage-related mutations is not equivalent to searching for somatic mutations, as not all mtDNA mutations are lineage related [15]. Thus, better bioinformatics tools are necessary to analyse single-cell somatic mutations in mtDNA.

Current methods use the allele frequency (AF) of a single-nucleotide variant, and the number of cells containing the variant, to identify lineage-informative mutations [16, 17]. However, these tools then use a threshold based on the AF to call mutations in a single cell, yet AF is affected by sequencing depths and noise [18], which could compromise sensitivity and specificity (Fig. 1a). To address this, statistical models such as beta-binomial distribution (BBD) have been proposed for scRNA-seq or bulk variant calling [19, 20], but this has not yet been proposed for analysing mtDNA mutations.

In this work, we have developed a framework to call mtDNA mutations in single cells based on BBD (Fig. 1a and b; Supplementary Fig. 1a and b). Testing this method using data

Received: August 20, 2024. Revised: December 11, 2024. Accepted: February 13, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

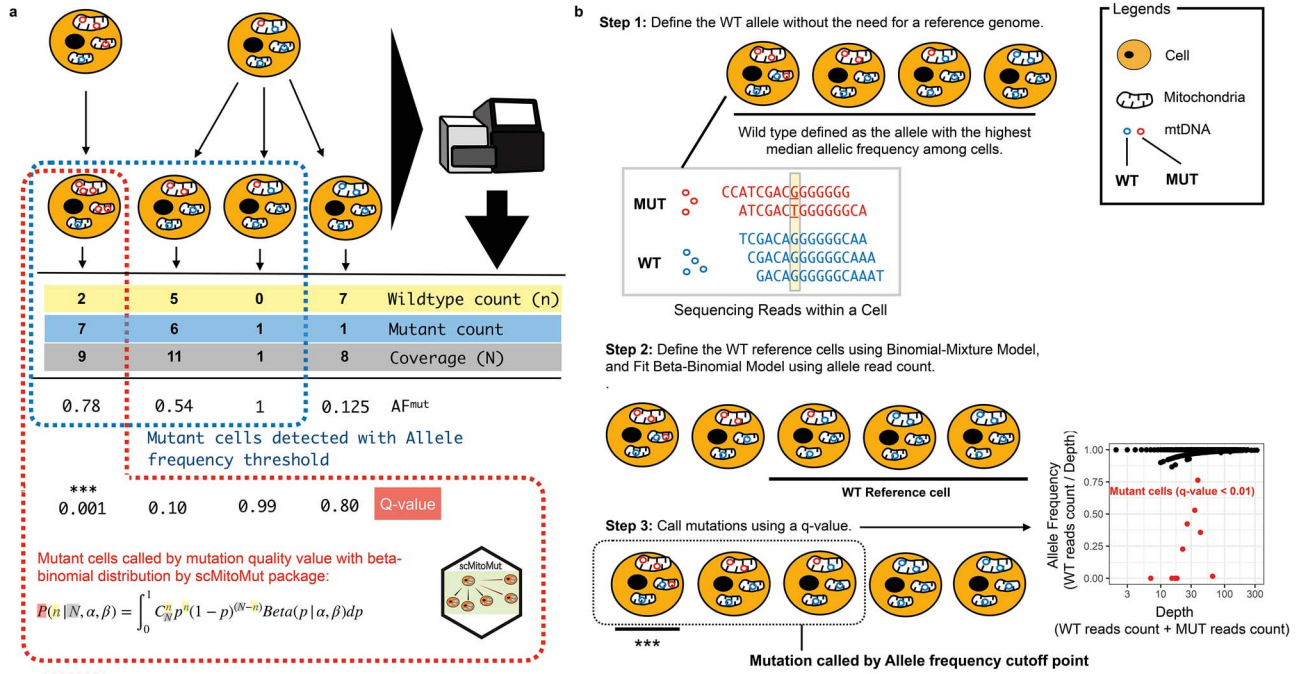


Figure 1. Statistical framework for calling lineage-informative mtDNA mutations. (a) Comparison of BBD and AF-based single-cell mtDNA mutation calling strategies. The diagram shows examples of mtDNA sequencing results from cells originating from two different common ancestor cells with varying mtDNA mutation heteroplasmy. The scMitoMut framework provides a statistical mutation calling approach, with the mutant cell identified by red dash-line, using BBD to assess the confidence ( $q$ -value) of lineage-informative mtDNA mutations by considering sequencing depth and WT read count. In contrast, the conventional AF threshold method, with identified mutant cells highlighted by blue dash-line, relies on a simple AF threshold and/or read count threshold to identify mutations. (b) Step 1 of scMitoMut consists of identifying the WT allele for each locus without the need of a reference genome. In brief, each cell contains multiple mitochondria, each with multiple mtDNA molecules (represented by circles). The read sequences obtained from the first cell, which has two alleles (in blue and red), are displayed below. Of all the cells, the blue allele has a higher median frequency (2/5, 2/4, 3/4, and 4/4 in the four example cells, respectively) than the red allele for the locus highlighted in yellow in the example sequences. This allele is defined as the WT allele. Step 2 consists of defining the WT reference cells using a binomial-mixture distribution classifier; the parameters of the beta-binomial distribution are then fitted to these cells for each locus using the sequencing depth and WT read count. Step 3 uses the beta-binomial model fitted in Step 2 to call mutations for each locus in each cell. For a given sequencing depth and WT read count, the confidence score of the presence of a mutation ( $q$ -score) is calculated from the BBD, taking into account multiple comparisons using the FDR. \*\*\* indicates cells with a  $q$ -value below a specified threshold are identified as mutant cells. The scatter plot on the right displays an example, in which AF is plotted on the y-axis and depth on the x-axis. Each dot represents a cell, with  $q$ -values < 0.01 in red, and other cells, in black.

with ground truth and more complex data, we find that it has a better sensitivity as compared to existing methods. Notably, as this method gives a confidence score for each mutation in each cell, it provides improved criteria for selecting lineage-informative mutations for lineage analysis. We also have developed an R package (called scMitoMut, for single-cell mitochondrial mutation) that is available in Bioconductor, which uses H5F files to store process data (thereby reducing memory usage) and accelerated with C++ (reducing computational consumption), which can be used on personal computers. The implementation of the BBD model and scMitoMut in R will serve as a valuable tool for enhancing mitochondrial mutation analyses and for using mitochondrial mutations in lineage tracing and mitochondrial genetic studies.

## Results

### Applying beta-binomial distribution to call lineage-informative mitochondrial mutations

We applied a hypothesis-testing framework to call mutations, which can be seen as a task to decide whether the observed read counts are generated by technical noise [e.g. Polymerase chain reaction (PCR) errors, sequencing errors, sample contamination, or other unknown factors]. After fitting a statistical distribution to the data, the framework compares the observed data to the

fitted distribution, to evaluate the probability that the observed read counts are derived from noise. If this is unlikely, we reject the null hypothesis and accept the alternative hypothesis, thus calling it a mutation. Using the probability value, we can quantify the fidelity of a cell containing a specific mutation by considering the magnitude of the noise (Fig. 1a and b).

As an input, the framework needs a matrix with allele count per cell per locus per base. To start, we defined the wild-type (WT) allele for each biological individual. In contrast to previous methods that defined the reference genome base as the WT, we now defined the majority base in most cells as the WT, removing the effect of potential individual polymorphism compared to the reference genome. Once the WT allele was identified, we defined the mutant reads as the total number of reads per locus minus the number of WT reads at the same locus for each cell (Fig. 1b; Supplementary Fig. 1a and b). As in normal-depth mtDNA sequencing results, two mutations are rarely observed at the same site (Supplementary Table 1); this allowed us to reduce the statistical test needed for each cell, thereby increasing the statistical power with multi-test type I error correction.

To select the reference WT cells for identifying mutations in subsequent steps, we classified cells into WT or mutant by fitting a binomial-mixture distribution to the mtDNA reads and computed the probability of specific cells belonging to the WT subset. Using this probability, we selected a cell subset assumed

to have no mutations [by default, false discovery rate (FDR) > 0.05] (Fig. 1b).

In the final step, we fitted the BBD to each locus using the WT cell subset (Supplementary Fig. 1a). The fitted BBD model was then used as the null distribution to compute the probability of observing a specific WT allele, in a given cell and locus (Supplementary Fig. 1b). To control the multiple testing, the adjusted P-value, referred to as the mutation *q*-value, served as a quantitative measure of confidence in calling a mutation in a given cell and locus (Fig. 1b; Supplementary Fig. 1b).

Overall, this statistical framework using BBD allows to call mtDNA mutations in single cells based on its *q*-value.

## Implementation of the scMitoMut package

We created an R package that we termed scMitoMut (freely available in Bioconductor) that includes functions for data importing, mutation calling, data export, and visualization (Fig. 2a). The input of scMitoMut is the preprocessed read count matrix, which is independent of the sequencing protocol and platform and can be obtained from tools such as CellSNP-lite [21] or mtGATK [22]. At the start of the pipeline, the mapped BAM file is used to extract read count matrix using CellSNP-lite [21] or mtGATK [22]. The read count of the four bases (A, T, C, and G) per cell per locus is used by scMitoMut as an input for calling the lineage-informative mutations per cell per locus. These lineage-informative mutations can then be jointly analysed with single-cell omics data (Fig. 2b).

As a typical input matrix for scMitoMut included tens of thousands of cells, each of which has thousands of alleles and associated *q*-value, scMitoMut keeps all the raw input and intermediate output on the hard disk by using an HD5F database associated with an R object and functions for importing data, model fitting, and exporting results. Thus, by reducing the memory usage, scMitoMut can run on a personal computer and is able to analyse tens of thousands of cells without slowing down the system (Fig. 2c).

We also boosted the mutation calling process by implementing expectation maximization (EM) and maximum likelihood estimation (MLE) algorithms in C++ separately for fitting the binomial-mixture distributions, and BBD. This significantly increased the speed of distribution fitting, achieving a maximum ~170× improvement over the existing R function in the mixtools package [23] with the same fitting accuracy (Fig. 2d; Supplementary Fig. 2a–d). Our BBD fitting achieved a fitting speed that was 6× faster than the VGAM package [24] with the same accuracy and around 1.5× faster than the bbmle package with better accuracy (Fig. 2e; Supplementary Fig. 3a–c). In scMitoMut, the fitting process can also run in parallel to leverage the multicore capabilities of central processing unit (CPUs).

The scMitoMut is a well-engineered toolkit for mtDNA mutation calling with high memory and computational efficiency that is freely accessible for researchers willing to study mtDNA mutation for lineage tracing.

## scMitoMut is more precise at identifying small clones from a scDNA-seq dataset of mixed cell lines

To compare the ability of scMitoMut to detect mutations as compared to other methods, we first tested the performance of scMitoMut using an *in vitro* experimental dataset with ground truth. It's a mix of BJ cells (human skin fibroblasts) and MKN45 cells (human gastric cancer cells) at a ratio of 99:1. MKN45 cells have copy number variation (CNV), but BJ cells do not, allowing us to distinguish the two types of cells. We identified 11 MKN45 cells and 1045 BJ cells, which served as ground

truth for later benchmarking for mtDNA mutation calling (Fig. 3a). The mtDNA sequencing depth distribution was relatively homogeneous over the loci, with a median depth of around 10 (Supplementary Fig. 4a). We then selected the cells with median depth >5, giving a final set of 962 cells (of which, 11 MKN45 cells). This set was used to test the performance of scMitoMut with its different parameters (Supplementary Table 2).

To determine how strictly we should call a mutation in a cell, we tested the mutation quality *q*-value (FDR adjusted P-value) thresholds of 0.5, 0.01, 0.005. Decreasing the *q*-value threshold led to a decrease in the number of observed mutations (Fig. 3b). We calculated the precision for each mutation, defined as the percentage of MKN45 mutant cell numbers versus total mutant cell numbers. The majority of mutations exhibited a precision rate of 100%, with a statistically significant enhancement in precision observed when comparing *q*-value thresholds of 0.01 and 0.05 (Fig. 3c). We saw no statistically significant improvement of precision when we used more stringent thresholds (from 0.01 to 0.005) (Fig. 3c). When aggregating all mutations, the overall precision was 95.1% using a *q*-value threshold of 0.05, 98% with a *q*-value of 0.01, and 98.6% with a *q*-value of 0.005 (Supplementary Fig. 4b). Notably, changing the *q*-value threshold did not statistically change the number of cells detected per mutation (Supplementary Fig. 4c). Thus, scMitoMut precisely called mutations per cell and a *q*-value of 0.01 can be used as default for calling a mutation in a cell.

To determine the minimum number of cells required to call a mutation, we tested the impact of different thresholds of cells per mutation on the precision and the numbers of lineage-informative mutations. We observed no statistically significant differences in precision when increasing the minimum number of cells per mutation, despite an expected decrease in the total number of mutations called and an increase in clone size (number of cells per mutations) (Supplementary Fig. 4d–g). The calling of lineage-informative mutation was robust to the other parameters of scMitoMut (Supplementary Table 2), which are the mean mtDNA depth per cell for selecting good quality cell and binomial-mixture distribution classifier FDR threshold for WT cells preselection to avoid overfit of the BBD (Supplementary Fig. 4h–n). Thus, the precision of the mutation calling results was mainly affected by the *q*-value threshold. In the scMitoMut package, we therefore defined the default parameters as a minimum count of five mutant cells and a predetermined *q*-value threshold of 0.01; note, however, that these values can be adjusted for different biological questions.

We next compared scMitoMut with the two state-of-the-art (SOTA) methods, Signac, and MQuad [8, 16, 17]. Calling the mtDNA mutations in MKN45 cells using Signac, MQuad, or scMitoMut (parameters are shown in Supplementary Fig. 5a), we identified a total of 44 mtDNA mutations in MKN45 cells, of which scMitoMut detected all 44, Signac detected 14, and MQuad detected 2 (Fig. 3d). The mutations identified by scMitoMut alone have the same qualities as the mutations also identified by Signac and MQuad. (Fig. 3e; Supplementary Fig. 5b). The mutations identified in MKN45 cells had an AF ranging from 0.137 to 1 (Supplementary Fig. 5c). Employing the previously defined precision (i.e. the proportion of MKN45 mutant cells to the total number of mutant cells), we found no statistically significant differences between mutations detected by scMitoMut and Signac (Supplementary Fig. 5d). However, the median clone size of the Signac result is 10.5, which is statistically smaller than that of the scMitoMut result, of 11.0 (Supplementary Fig. 5e). Thus, given that there are 11 MKN45 cells in this dataset, scMitoMut

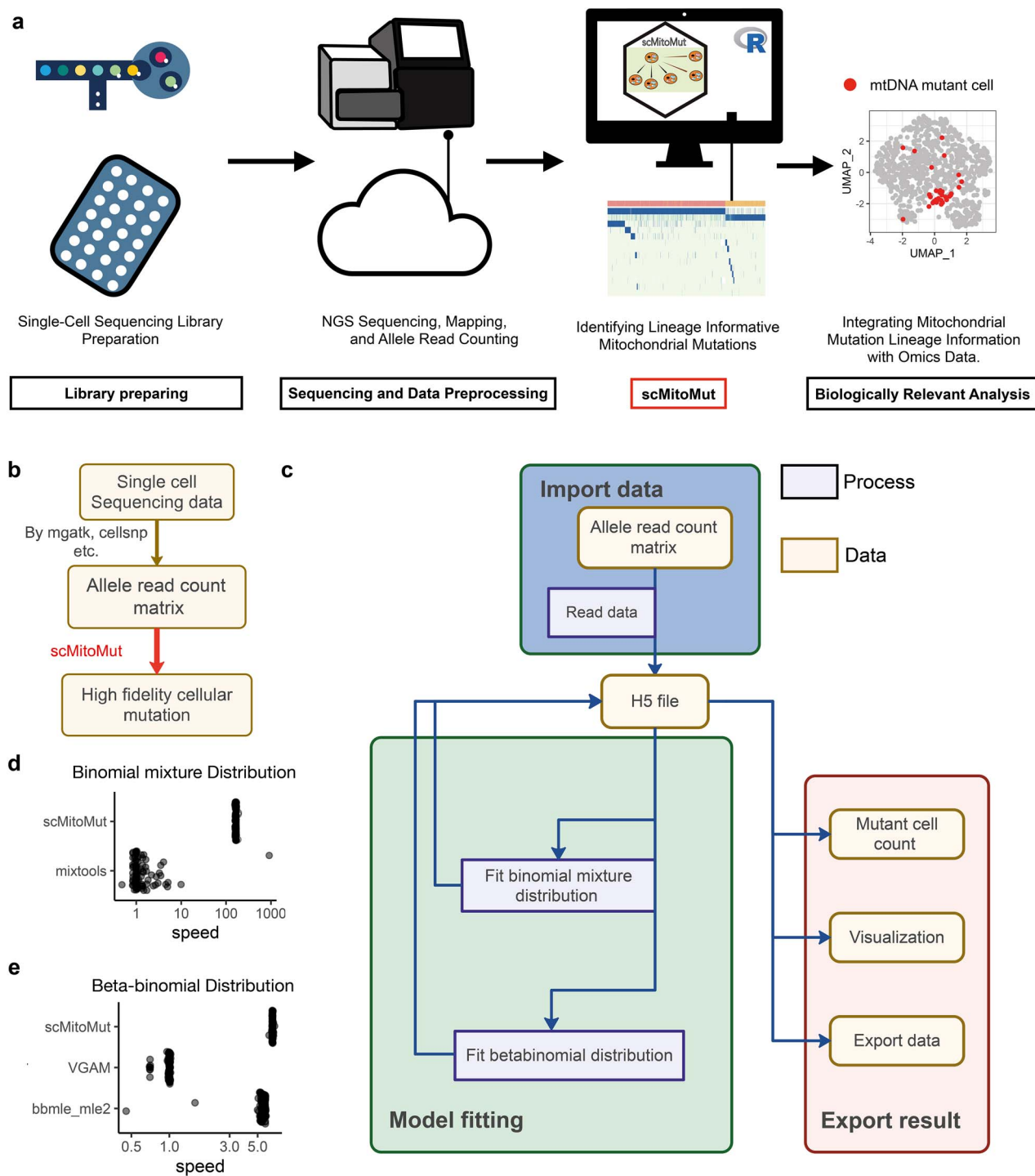


Figure 2. Application and implementation of scMitoMut. (a) After single-cell library preparation, sequencing data are preprocessed to generate the allele count matrix generally using high-performance computers or cloud computers. scMitoMut is then used to identify high-confidence single-cell mtDNA mutations using a personal computer. Finally, the mtDNA mutation results from scMitoMut are combined with cell phenotypes data for biological analysis. (b) Tool chain of using scMitoMut to call mtDNA mutations. The boxes represent the data and are linked by arrows representing the data operations. Tools used for the data operations are listed next to the arrows. (c) The diagram illustrates the organization of the scMitoMut R package. The yellow box represents data connected by arrows symbolizing data operations. These operations describe the data operation processes (purple box). Rounded corner frames are used to emphasize groups of data and processes, showcasing the three main modules in scMitoMut: data import (blue), model fitting (green), and result export (pink). All three modules interact with HDF5 files data for reading and writing raw and intermediate data stored on the hard disk. (d) Relative speed of each run divided by the median run time of mixtools for the R packages mixtools and scMitoMut. Each point is a run, with 100 runs tested per tool with the following parameters of the binomial mixture distribution:  $\theta=0.5$ ;  $\pi_1 = 0.992$ ;  $\pi_2 = 1$ ;  $n$  follows a log-normal distribution with log-mean = 2 and log-SD = 1 rounded up to the nearest non-negative integer; sample size is 1000. (e) Relative speed of each run divided by the median run time of VGAM for the R packages VGAM, bbmle, and scMitoMut. Each point is a run, with 100 runs tested per tool with 5000 observations and the following parameters of the beta-binomial distribution: mean probability parameter  $\theta$  of 0.5 and a super-dispersion parameter  $\phi$  of 20.



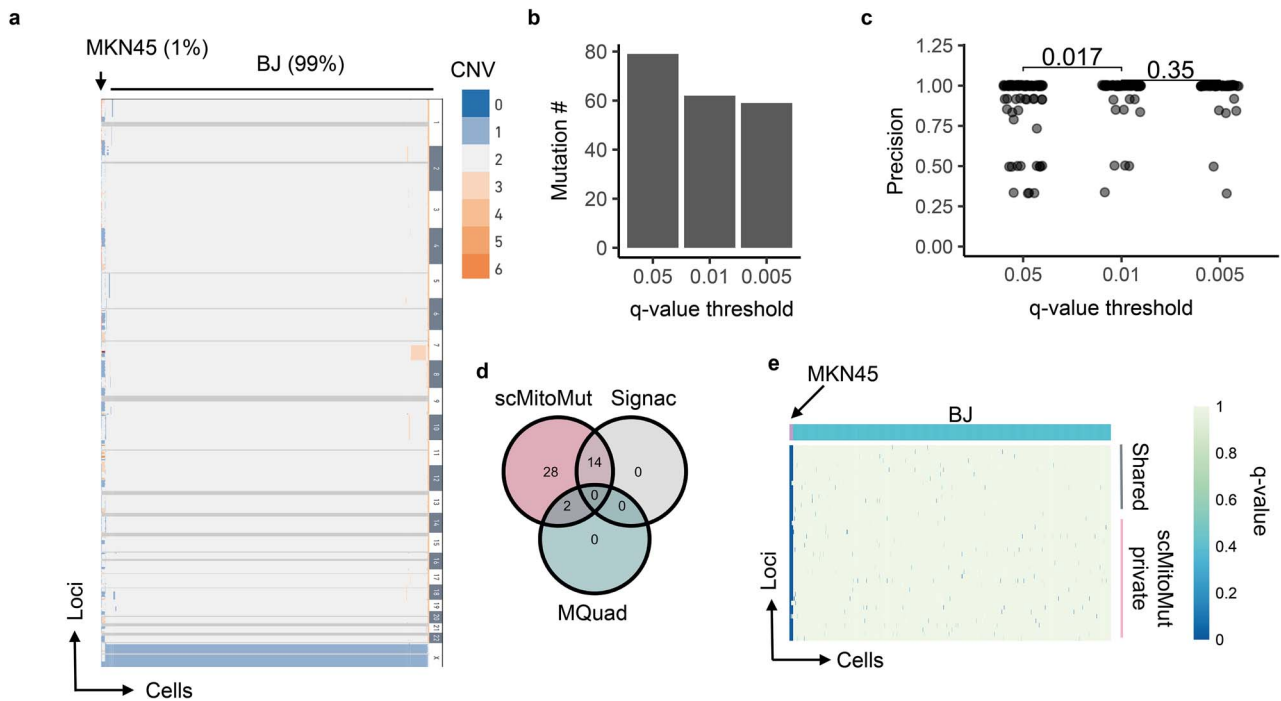


Figure 3. Benchmarking mtDNA mutation calling on a mixed-cell line dataset. (a) Heatmap of the CNV from single-cell genome sequencing data of a mixture of 11 MKN45 cells and 1045 BJ cells from a 10× Genomics demo dataset. Each row is a CNV segment, and each column is a cell. The colour within the heatmap represents the CNV variation according to the colour key on the right side of the heatmap, whereby the dark grey colour indicates missing data and light grey indicates diploidy. The cell types with and their percentages in the mixture are indicated at the top of the heatmap. (b) Number of mtDNA mutations present in MKN45 cells for various quality thresholds ( $q$ -values of 0.05, 0.01, and 0.005). The mutation was counted if its  $q$ -value was below the corresponding  $q$ -value threshold in at least one MKN45 cell. (c) Mutation calling precision for various  $q$ -value thresholds. The mutation precision was calculated as the ratio of mutant MKN45 to total mutant cells for each mutation. Each dot represents a mutation;  $P$ -values were calculated using the two-sided Wilcoxon test.  $n$  is 79, 62, and 59 for  $q$ -value threshold of 0.05, 0.01, and 0.005, respectively. (d) Number of mtDNA mutations that were called in scMitoMut, Signac, or MQuad analysis. Only mutations present in at least one MKN45 cell were counted. (e) Heatmap displaying the mutation quality  $q$ -values of scMitoMut. Each row represents a locus, and each column represents a cell. The darker the blue colour, the lower the  $q$ -value, indicating higher confidence in the mutation, while the white indicates missing values. MKN45 and BJ cells were annotated in pink and blue respectively above the heatmap. The mutations only detected by scMitoMut (28 total) are highlighted in pink, while shared mutations (16 total: 2 with MQuad and 14 with Signac) are indicated in grey.

had better recall. We also compared scMitoMut with Mitoclone2 [25], a tool designed for calling mtDNA mutations in scRNA-seq data using conventional read threshold filtering and removing mutations found by more than one individual. scMitoMut showed also better performance in terms of precision and sensitivity compared to Mitoclone2 [25] (Supplementary Fig. 5a and f–i). Overall, scMitoMut outperformed the SOTA methods.

### scMitoMut package detects more mutations than other state-of-the-art methods in scATAC-seq data from human colorectal cancer cells

We next tested the scMitoMut and other SOTA methods using a human colorectal cancer (CRC) scATAC-seq dataset with enriched mtDNA [22], which consists of epithelial-derived CRC cells and blood cells (Supplementary Fig. 6a). We compared the lineage mutation identified by scMitoMut with the ones from MQuad and Signac. We used the default parameters for Signac and ScMitoMut (parameters are shown in Supplementary Fig. 6b) and the ones from the original publication for MQuad [17]. Of note, the original work used less stringent parameters for MQuad than its default settings or than those used in the two other methods [17]. Of the 14 mutations identified by scMitoMut, Signac and MQuad both identified nine mutations respectively, whereby four mutations were identified only by scMitoMut and one only by MQuad (no specific mutations were identified by Signac) (Fig. 4a, b). Thus, scMitoMut allowed us to identify

more mutations, even as compared to MQuad, which used a less-stringent mutation filtering threshold. The MQuad-specific mutation was not identified by scMitoMut due to its clone size of 4, which was below the threshold of 5 (Supplementary Fig. 6c), but it would have been identified if the same threshold had been applied to scMitoMut as in the original publication using MQuad [17]. Two of the scMitoMut-specific mutations, chrM.200 and chrM.310, were filtered by Signac by the VMR and strand concordance threshold, respectively (Supplementary Fig. 6d–f).

We then evaluated the lineage precision of mutations within a particular lineage of CRC and blood cells. Epithelial and blood cells diverge early in development, originating from the endoderm [26] and mesoderm [27], respectively, and should not share many mutations. Thus, defining lineage precision as the ratio of mutant cells in the dominant lineage of cancer epithelial or blood cells as compared to all mutant cells, we expect a 100% lineage precision if the mutations were exclusively present in one lineage. The lineage precision was similar between scMitoMut and Signac (Fig. 4c), indicating that the detected mutations by scMitoMut were informative for lineage identification. While scMitoMut found more mutations, the number of cells detected per mutations was similar between scMitoMut and Signac (Fig. 4d), suggesting that scMitoMut detected mutations with a similar cell abundance as previous methods. All four of the scMitoMut-specific mutations are lineage-informative mutations that are restricted to one lineage of cancer epithelial or blood cells (Fig. 4b).

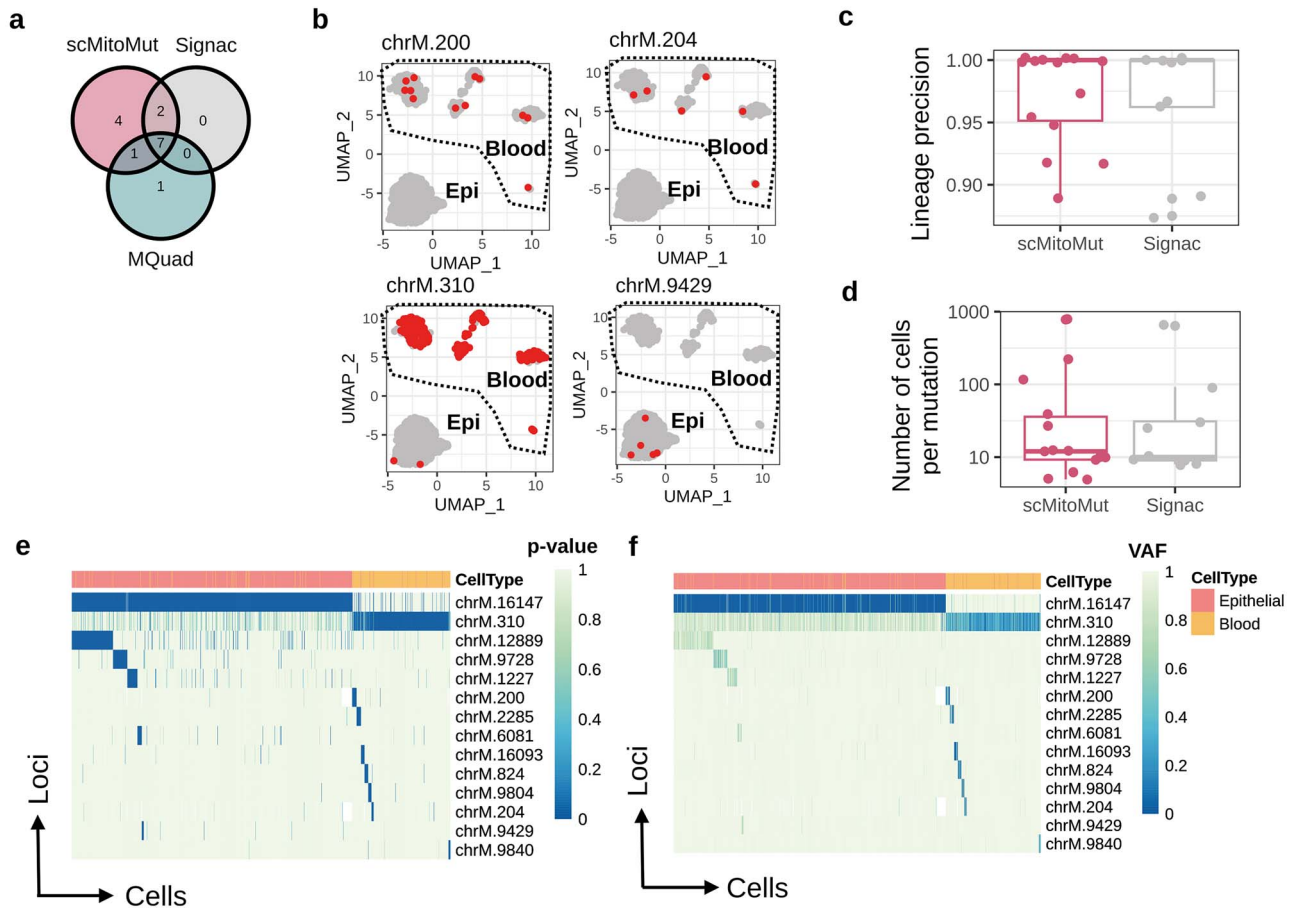


Figure 4. The beta-binomial distribution provides better lineage precision for distinguishing epithelial and blood lineages in the CRC dataset. (a) Number of mtDNA mutations detected by scMitoMut, Signac, and MQuad in the CRC dataset from [22]. (b) The UMAP projections of the scATAC-seq data for 4 mtDNA loci exclusively detected by the scMitoMut. In each UMAP, the CRC and blood cell clusters are annotated. Each dot represents a cell, with red dots indicating high-confidence mutant cells called by scMitoMut with  $q$ -value  $< 0.01$ . (c) Mutation calling precision for each mutation in the scMitoMut or Signac results. The precision was calculated by dividing the count of cells in the predominant lineage (blood cells or cancer epithelial cells) by the total number of cells with a given mutation.  $n = 14$  mutations in scMitoMut and  $n = 9$  mutations in Signac. Centre line, median; box limits, upper, and lower quartiles; whiskers,  $1.5 \times$  interquartile range; points, outliers are displayed. Two-sided Wilcoxon test was used to compare precision between the two methods ( $P = 0.89$ ). (d) Number of mutant cells identified by scMitoMut or Signac. Each dot represents a mutation,  $n = 14$  mutations in scMitoMut results and  $n = 9$  mutations in Signac results. Centre line, median; box limits, upper, and lower quartiles; whiskers,  $1.5 \times$  interquartile range; points, outliers are displayed. Two-sided Wilcoxon test to compare both methods ( $P = 0.92$ ). (e) The mutation heatmap displays the scMitoMut mutation  $q$ -value per cell (columns) and per locus (rows). The deeper blue within the heatmap represents the smaller mutation  $q$ -value that corresponds to higher confidence in the calling, while white indicates missing values. An annotation bar above the heatmap indicates the cell lineage, with cancer epithelial cells in red and blood cells in orange. (f) The mutation heatmap showing the mutation AF per cell (columns) and per mutation (rows). The shading of blue within the heatmap represents the WT AF. An annotation bar above the heatmap indicates the cell type with yellow of blood and red of cancer epithelial cells.

In the mutation  $q$ -value heatmap generated from scMitoMut results, the mutations chrM.16147 and chrM.310 are predominantly confidently detected (with small  $q$ -value) in CRC and blood cells, respectively (Fig. 4e). Other mutations appeared to be clustered in regions on either chrM.16147 or chrM.310, indicating that the chrM.16147 and chrM.310 mutations appeared earlier (Fig. 4e). The AF heatmap is consistent with the  $q$ -value heatmap but with a lower signal-to-noise ratio between mutant cells and WT cells (Fig. 4f); this is especially seen for the later mutations clustered on chrM.16147 and chrM.310 (Fig. 4e and f), demonstrating the added value of using the  $q$ -value to call and visualize lineage-informative mutations.

### mtDNA mutations identify lineage branching in two 10× multiome human datasets

After determining that scMitoMut exhibited higher sensitivity and specificity as compared to the SOTA methods Signac and MQuad in single-cell, whole-genome sequencing and

mtDNA-enriched scATAC-seq datasets, we next tested scMitoMut on a dataset for which both the chromatin accessibility and the transcriptome were recovered from the same cells using the 10× Genomics multiome kit. We first analysed the 10× Genomics single-nuclear multiome dataset from human peripheral blood mononuclear cells (PBMC). After analysing the transcriptomes of the cells, we divided cells into three categories: monocytes, T cells, and B cells (Fig. 5a). Even though single-nucleus sequencing was used, mtDNA reads can still be found in the scATAC-seq results with median depth of 3, which indicates that mtDNA could be detected (Supplementary Fig. 7a). Thus, we included the cells with mean mtDNA depth of  $> 5$ , resulting in median mtDNA sequencing depth of 10 (Supplementary Fig. 7b); this is comparable to the mtDNA profile of single-cell genome sequencing of a mixture of cells that we analysed before (see Supplementary Fig. 4a). Finally, there were no noticeable differences in mtDNA sequencing depth for monocytes, T cells, or B cells (Supplementary Fig. 7c–e).

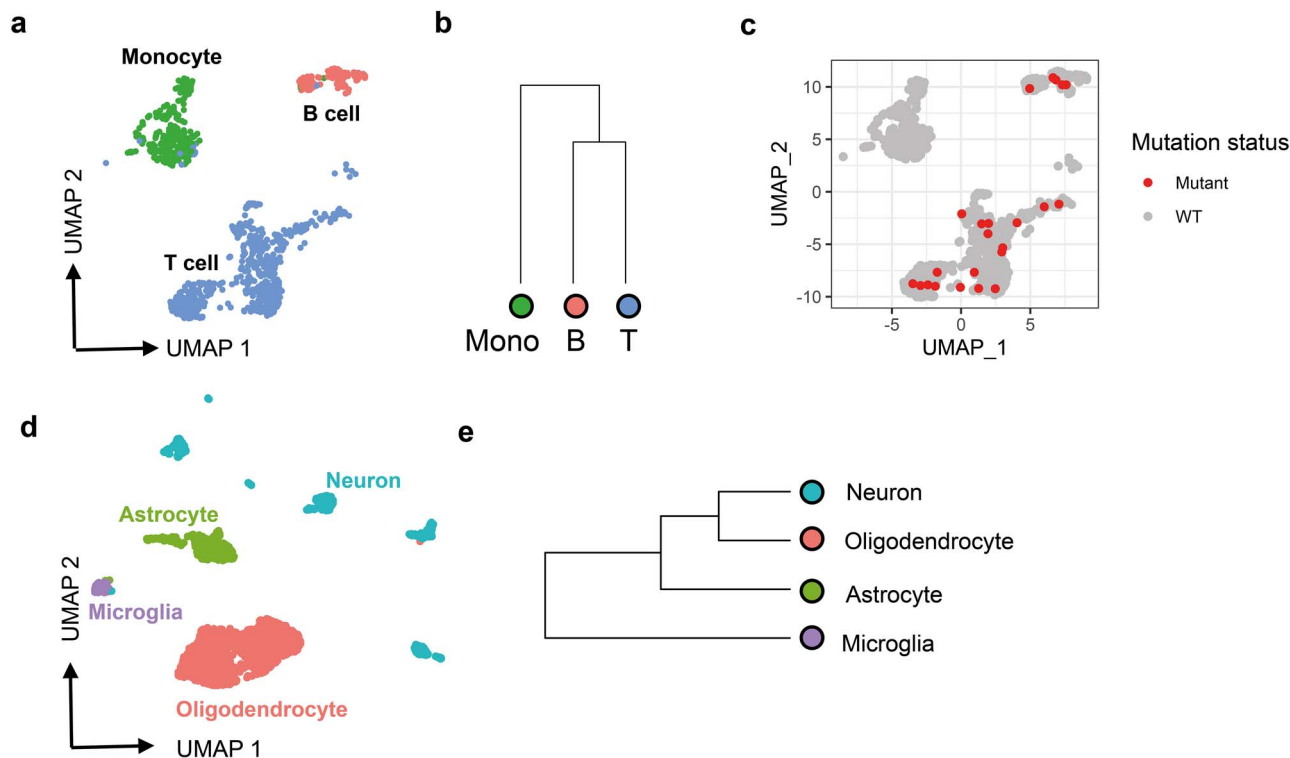


Figure 5. scMitoMut identifies lineage-informative mitochondrial mutations in the 10× multiome dataset. (a) UMAP visualization of the PBMC 10× multiome dataset with default parameters using Seurat. Each dot represents a cell, colour-coded by cell type: red for B cells, blue for T cells, and green for monocytes. (b) Hierarchical clustering of monocytes (shown in green), B cells, and T cells, based on mutant cell counts normalized by total mutant count per cell type cluster, using Euclidean distance and complete linkage clustering. (c) UMAP visualization of locus chrM.16499. Each dot represents a cell, and red dots indicate mutant cells ( $q$ -value  $< 0.01$ ) identified using scMitoMut. (d) UMAP visualization of brain data from 10× multiome dataset. Each dot represents a cell; oligodendrocytes are shown in red, microglia in purple, neurons in green, and astrocytes in blue. (e) Hierarchical clustering analysis of four brain cell types conducted by the method described in (b).

We then analysed mtDNA mutations in single-cell ATAC-seq data using scMitoMut; this identified a total of 17 mutations based on a minimum mutant cell number of  $> 10$  and a mutation quality value  $q$ -value threshold of 0.01 (Supplementary Fig. 8a and b). In comparison, MQuad and Signac detected fewer mutations using all their default parameters except the minimum mutant cell number requirement of  $> 10$  (Supplementary Fig. 8a and b). This demonstrated that scMitoMut had a higher sensitivity also for multiome data.

Taking advantage of this higher sensitivity, we conducted hierarchical clustering on the three cell clusters based on the normalized mutant cell count per cell type per mutation. The results indicated that the two lymphoid cell types (T and B cells) clustered together and were separated from the monocytes (Fig. 5b and c). This aligns with our current understanding of blood cell development, as T and B cells originate from the lymphoid progenitor, while monocytes derive from the myeloid progenitor. It also demonstrated that the increase sensitivity of scMitoMut allows cells to be correctly assigned to their lineage of origin.

Studying the central nervous system development in mice has revealed that microglia are derived from yolk sac cells in embryo, while neurons, astrocytes, and oligodendrocytes are derived from embryonic neural tube in embryos [28]. To explore the lineage distance between cells in the human central nervous system, we analysed brain cell 10× multiomics data, which included 3000 brain cells, with scATAC-seq and scRNA-seq (Supplementary Fig. 9a). We annotated neurons, astrocytes, oligodendrocytes, and microglia using known cell type markers (Fig. 5d; Supplementary Fig. 9c). Using the same parameters as

applied in the above 10× Genomics multiome PBMC dataset, scMitoMut identified 14 mutations and MQuad identified four mutations. However, none of these mutations passed the threshold set by Signac (Supplementary Fig. 9b and d). Using the 14 mutations identified by scMitoMut, we calculated the mutation frequency for each cell type and conducted hierarchical clustering based on these frequencies. This analysis revealed that human microglia are less closely related to neurons and oligodendrocytes, similar to the known patterns of cell development in the central nervous system of mice (Fig. 5e) [29].

Thus, by applying scMitoMut to two human datasets, we demonstrated that it can be used to assess lineage relationship between cells, which is of a particular importance for understanding human lineage, as most of it is unknown.

## Discussion

Here we present scMitoMut, an R package that we developed to facilitate the identification and analysis of mitochondrial lineage informative DNA mutations at a single-cell resolution. Within the scMitoMut package, we implemented a BBD-based statistical framework to call the mtDNA lineage-informative mutation. Using this statistical framework, scMitoMut provides a  $q$ -value for the confidence of the presence of the mutation that is used to call if the mutation is present in a single cell. This approach allows the user to select lineage-informative mutations based on the tools from statistical testing.

The statistical framework demonstrated heightened sensitivity in mutation detection as compared to the other SOTA methods.

Signac [16] uses a threshold on the variant mean ratio (VMR) to select mutations with high AF variation between cells to remove germline variations and a threshold on the strand consistency to filter sequencing noise. However, it is not straightforward how to choose this threshold, especially as there could be variation between loci. Mquad [17] tried to overcome this by using a binomial-mixture distribution that assumes there are two cell populations with each a different mtDNA heteroplasmy. As a consequence, it might fail to capture the mtDNA heteroplasmy variation within each cell population compared to the beta-binomial distribution used in scMitoMut. In our statistical framework, the  $q$ -value can more accurately assess mutation certainty by integrating multiple pieces of information when calling a single-cell mutation, rather than applying simple AF threshold.

In practice, scMitoMut includes an efficient implementation using an HDF5 file database for high-speed data access and reduced runtime memory usage. To enhance running speed, we used Rcpp for distribution fitting and parallelised tasks with CPU multi-threading. This package will thus serve as an invaluable tool for enhancing mitochondrial mutation analysis, as well as for promoting the use of mitochondrial mutation in lineage tracing and mitochondrial genetic studies in large dataset, such as the atlas dataset [30]. Compared to other SOTA methods [8, 16, 17], scMitoMut uses fewer parameters, thus providing an easier-to-use tool.

The flexibility of the beta-binomial distribution makes it more robust at learning the noise distribution if there is a small proportion of mutant cells in the reference cells used for fitting. However, the presence of a high proportion of mutant cells can cause overfitting of the beta-binomial and thus weaken the ability of the beta-binomial distribution to detect mutant cells. To overcome this issue, we used a binomial-mixture distribution classifier for unsupervised classification. The benchmark shows that utilizing a binomial-mixture preselection improves significantly the sensitivity of mutation detection (Supplementary Fig. 4l–n).

By analysing single-cell mitochondrial mutations of known lineages, we found that the precision of mutations depends on the  $q$ -value threshold rather than any other parameters of scMitoMut. However, in practice, the total number of false-positive mutations is a product of both the mutation precision and the total number of detected mutations. Lowering the threshold of the mutant cell number leads to a higher mutant cells number, which increases the total number of false-positive mutant cells. We recommend adjusting the threshold of the mutant cell number based on usage purposes.

Mitochondrial somatic mutations have recently emerged as an optimizing lineage genetic marker, which hold great potential in human lineage tracing [8, 31]. However, some issues are still seen as hindrances, including how the high mutation occurrence and the fast turnover rate of mutations in mtDNA affect the lineage information [15] and how the RNA editing issues affect calling lineage-informative mutation [8]. While more work is required to investigate the impact of these issues on the use of scMitoMut, we currently recommended using DNA sequencing (e.g. scATAC-seq) as the input value for scMitoMut to avoid RNA editing.

Other limitations of scMitoMut are the assumption that there is one mutation per locus. This assumption seems reasonable with the current sensitivity of mtDNA detection. Indeed, our simulation shows that for a total of 50 mutations detected per dataset, the probability is 0.0010 for a specific locus to have multiple mutant alleles (Supplementary Table 1). However, if more

mtDNA mutation would be detectable this assumption would lead to false negative and would need to be reassessed. Improvement to increase the number of mtDNA mutants are in progress with the use of mtDNA enrichment steps and error correction systems [32]. However, it comes with the detection of low heteroplasmic mtDNA mutations, which are more susceptible to fast turnover and are less informative for lineage tracing. How to best use mtDNA mutations, especially the low heteroplasmic ones, for lineage reconstruction should be investigated.

In summary, in the newly developed scMitoMut package, we have established a robust statistical framework for identifying lineage-informative mtDNA mutations, with an easy-to-use interface in R. This tool enhances the utilization of mtDNA somatic mutations to trace lineages.

## Materials and Methods

### Applying beta-binomial distribution to call lineage informative mitochondrial mutations

#### Defining the wild-type allele frequency

Single-cell sequencing data provided allele counts per locus  $j = 1, 2, \dots, K$  per cell  $i = 1, 2, \dots, S$ . In the single nuclear variation case, for a specific locus and cell, the allele counts  $n$  for four nucleotides are  $n_{A,ij}$ ,  $n_{T,ij}$ ,  $n_{C,ij}$ , and  $n_{G,ij}$ . The sequencing depth  $N_{ij}$  was defined as

$$N_{ij} = n_{A,ij} + n_{T,ij} + n_{C,ij} + n_{G,ij}. \quad (1)$$

To define the WT allele, we used the dominant allele  $W_j$  for locus  $j$  over all cells:

$$W_j = \arg(\max_{x \in A, T, C, G}) (M_{x,j}), \quad (2)$$

where  $M_{x,j}$  was the median count of allele  $x$  for locus  $j$  across all cells.

$$M_{x,j} = \text{median}(\vec{n}_{x,j}) \quad (3)$$

In equation (3), the  $\vec{n}_{x,j}$  was the count of allele  $x$  at locus  $j$  across all cells. Specially,

$$\vec{n}_{x,j} = n_{x,1j}, n_{x,2j}, \dots, n_{x,Sj}. \quad (4)$$

For a cell  $i$  and locus  $j$ , let  $m_{ij}$  represent the allele count for the WT allele  $W_j$ , which was used to calculate WT AF  $wAF_{ij}$ :

$$m_{ij} = n_{x,ij}, \text{ for } x = W_j \quad (5)$$

$$wAF_{ij} = \frac{m_{ij}}{N_{ij}}. \quad (6)$$

We used  $wAF_{ij}$  as a proxy of the heteroplasmy level of the WT allele at locus  $j$  of cell  $i$ .

Based on the simulation (Supplementary Table 1), we assumed that each locus was unlikely to have more than one mutation. Therefore, each locus could either be WT or bear a single mutation, representing two possible states. These states were modelled using binomial related distributions. A mutation is identified for locus  $j$  in cell  $i$  when the AF  $wAF_{ij}$  is significantly lower.



### Fitting mitochondrial heteroplasmy sequencing results with beta-binomial distribution

Since mutation calling for each locus was independent of each other in our strategy, we explained the model fitting process for a single locus, omitting the subscript  $j$  for conciseness.

Single-cell sequencing can be modelled as a random sampling process of mtDNA from a cell. Assuming the sampled mtDNA number is small compared to the total mtDNA copy number in a cell, the sampling process of sequencing mtDNA then corresponds to a series of Bernoulli trials. In this Bernoulli framework, sequencing the WT allele is defined as success with the probability  $\pi$ , which corresponds to the mtDNA heteroplasmy level of the cell.  $\pi$  is inferred from AF using the resulting successful trial number  $m_i$  (WT allele count of cell  $i$ ) and the total trial number  $N_i$  (sequencing depth of cell  $i$ ). Let random variable  $X$  represent the WT allele count, it then follows a binomial distribution when sequencing a cell with heteroplasmy level  $\pi$  and total allele count  $N_i$ .

$$X \sim B(N_i, \pi) \quad (7)$$

Mitochondrial turnover can create inter-cell heterogeneity, which create heterogenous mtDNA heteroplasmy  $\pi$  between cells. To account for these factors, we introduce the beta-binomial distribution that allows  $\pi$  to be a random variable. In the beta-binomial distribution, we consider the mtDNA heteroplasmy level as a continuous random variable  $\Pi \in [0, 1]$  and is described by the beta distribution:

$$\Pi \sim \text{Beta}(\alpha, \beta). \quad (8)$$

Let  $X$  be the WT allele count and a random variable that follows a beta-binomial distribution (BBD):

$$X \sim \text{BBD}(N, \alpha, \beta). \quad (9)$$

The BBD also can be described in this format:

$$X \sim \text{BBD}(N, \theta, \phi), \quad (10)$$

where

$$\theta = \frac{\alpha}{\alpha + \beta} \quad (11)$$

$$\phi = \frac{1}{(\alpha + \beta + 1)}. \quad (12)$$

The parameter  $\theta$  describes mean of  $\Pi$ , while  $\phi$  is the dispersion parameter for describing the variance. If  $\phi = 0$ , the BBD degrades into binomial distribution, with a probability parameter of  $\theta$ . To avoid fitting the BBD on mutant cells (which will cause overfitting), cells were clustered using binomial-mixture distribution classifier to select the likely WT cells for model fitting. After fitting the BBD, the hypothesis test was used to get probability of a mutation using beta-binomial PMF function:

$$H_0 : \pi_i = \theta \quad (13)$$

$$H_1 : \pi_i < \theta \quad (14)$$

$$P_{\text{BBD}} = P(X < n_{W,i} | N_i, \alpha, \beta) = \sum_{n=1}^{n_{W,i}} \int_0^1 C_N^n p^n (1-p)^{(N-n)} \text{Beta}(p|\alpha, \beta) dp \quad (15)$$

We got  $P_{\text{BBD}}$  per locus per cell. For each locus, we performed the FDR correction for multiple tests across multiple cells and nominated the FDR value as  $q$ -value for evaluating the quality of lineage-informative mutation.

### Clustering cells with binomial mixture distribution

The binomial mixture distribution assumes existing populations with different mtDNA mutation heteroplasmy but no variation of heteroplasmy among cell populations. It oversimplified the mtDNA sequencing results. But due to its simplicity, it has less overfitting issues. Thus, it is used to preselect WT cells for beta-binomial distribution fitting.

We defined the WT allele heteroplasmy level in WT cell  $\pi_W$  and in mutant cell  $\pi_M$  and the percentage of WT cells  $\theta_W$  and mutant cells  $\theta_M$  in the population. Let the random variable  $X$  of the WT allele count follow a binomial mixture distribution:

$$X \sim \text{BMM}(N, \pi_W, \pi_M, \theta_W, \theta_M). \quad (16)$$

After fitting the distribution to the data, we calculate the probability of cell under investigation having a WT allele is calculated; this is denoted as  $P_{\text{BMM}}$ .

$$P_{\text{BMD}} = P(X < n_{W,i} | N_i, \pi) = \sum_{n=0}^{n_{W,i}} \binom{N_i}{n} \pi_W^n (1 - \pi_W)^{(N-n)} \quad (17)$$

The value of the probability  $P_{\text{BMD}}$  is used to classify the cells to get potential WT cells for fitting the beta-binomial model. The likely WT cells are selected by binomial-mixture distribution classifier with an FDR-adjusted  $P_{\text{BMD}}$  value larger than 0.05. While using scMitoMut, the user can choose not to apply the preselection or customized the binomial-mixture distribution classifier FDR threshold.

### Rcpp-based distribution fitting algorithm reimplemention and benchmarking

We reimplemented the model fitting algorithm of BBD and BMD using Rcpp in R [33]. This reimplemention was intended to improve CPU efficiency. The BBD was fitted using Maximum likelihood estimation (MLE), which finds the parameters of the BBD that maximize the likelihood function that represents the probability of the observed data given the model parameters. The BMD was fitted with expectation-maximization (EM) fitting, which is an iterative method to find maximum likelihood of parameters in models with latent variables with iterations of expectation (E) step and maximization (M) step. Detailed mathematical formulations, derivations, and additional implementation details can be found in the supplementary data.

We compared the speed of fitting BMD between scMitoMut and the multmixEM function in the mixtools R package [23], by fitting simulated BMD with two binomial distributions with  $\theta = 0.5$ ,  $\pi_1 = 0.992$ , and  $\pi_2 = 1$ ;  $n$  follows a log-normal distribution (log-mean=2 and log-SD=1) rounded up to the nearest non-negative integer. Each simulation includes 1000 observations. The simulation was done in R with a custom function using rlnorm and rbinom R random number generator. We benchmarked the running speed with the microbenchmark function from the microbenchmark R package with default parameters [34]. The accuracy of model fitting was tested using the simulations follow the parameter:  $\pi_1 = 0.5$ ,  $\pi_2 = 1$ ,  $\theta = 0.01, 0.1, 0.2$ , or  $0.4$ , and  $n$  following a log-normal distribution (log-mean=2, log-SD=1) rounded to the nearest non-negative integer; each simulation contains 5000 observations.

We benchmarked the model fitting speed of BBD between `scMitoMut`, `vglm` function from `VGAM` R package [24] and `mle2` function from `bbmle` R package [35], by fitting the simulated data with: mean probability parameter  $\theta$  of 0.5; super-dispersion parameter  $\phi$  of 20; each simulation includes 5000 observations. The simulation was done by the '`rbetabinom.ab`' function from `VGAM` R package [24]. The `mle2` function is a generated maximum likelihood fitter, we need to provide the target function that was the `dbetabinom` function from the `emdbook` R package [36]. The accuracy of the fitting was tested using a simulated BBD with parameters specified in `rbetabinom.ab`: mean parameter  $\theta$  of 0.1, 0.5, 0.9, or 0.99; dispersion parameter  $\phi$  of 20, 40, 80, 160;  $n$  adheres to a log-normal distribution with a log-mean of 2 and log-SD of 1, subsequently rounded up to the nearest non-negative integer; each simulation encompasses 5000 observations.

### Probability of multiple mutant alleles at a locus

To assess the likelihood of multiple mutations occurring at a single locus when identifying mtDNA mutations, we simulated the mutation detection of 16-kb human mitochondria genome, assuming that each mtDNA locus has an equal chance of mutation. For the 16-kb mitochondrial genome, we simulated the 48 000 possible mutation events (16 k genome size multiplied by three potential mutant bases). We then randomly sampled 25, 50, and 100 mutations from the 48 000 possibilities and repeated 100 000 times and calculated the probability that multiple mutant alleles occurred at a locus. The probability was calculated by dividing the total number of loci with multiple mutant alleles to the total simulated mutations.

### Mixed-cell line single-cell genome sequencing Identifying the MKN45 and BJ cells in the mixed-cell line dataset using copy number variation

The dataset consists of single-cell genome sequencing data for a cell line-mixture containing 1% MKN45 cells and 99% BJ cells from the 10× Genomics demo. In the dataset, the MKN45 cells, a cancer cell line with chromosome instability, were distinguished from the normal BJ cell line using CNV analysis. A Loupe scDNA Browser 1.1.0 was used to analyse the '`bj_mkn45_1pct_dloupe.dloupe`' output from Cell Ranger DNA version 1.1.0, which was downloaded from 10× Genomics Demo dataset website. Using Loupe scDNA Browser, we generated CNV heatmap and extracted cell clusters based on CNV. Based on CNV, we identified a MKN45 cell cluster with 11 cells (1%), which has high CNV variations, from 1045 cells. The cellular barcodes of MKN45 cells were exported from Loupe scDNA Browser and used in the future as ground truth of the MKN45 cell identity for benchmarking the mtDNA lineage-informative mutation calling.

### Calling mtDNA mutation using `scMitoMut`

The `mgatk` version 0.7.0 was used to analyse the Bam file to count mtDNA AF. The `mgatk` ran in '`tenx`' mode options '`-keep-duplicate`' and no alignment-quality constraint '`-alignment-quality -1`'. The Bam file '`bj_mkn45_1pct_possorted_bam.bam`' was downloaded from 10× Genomics website together with the good quality cell list '`bj_mkn45_1pct_per_cell_barcode.tsv`'. The '`coverage.txt.gz`' file within the `mgatk` output was used to assess coverage distribution per locus for all cells. After selecting cells with an average mtDNA coverage greater than 5, `scMitoMut` calculated the  $q$ -value for mutation quality per cell per locus using the allele count matrix from `mgatk`.

We identified mutations with a  $q$ -value below a threshold per cell per locus, and required the number of mutant cells to

meet a threshold. We assessed the specificity of mtDNA lineage-informative mutations using various  $q$ -value thresholds (0.05, 0.01, and 0.005) and mutant cell number thresholds (1, 5, and 9). The identified mutation number and mutation precision were evaluated using the mutations with at least one mutant MKN45 cell.

For each mutation, we defined the precision of mutation calling as the percentage of MKN45 cells. For a specific mutation  $i$ , the mutant cells number in MKN45 cells is  $n_i$  ( $n_i > 0$ ) and  $m_i$  for total mutant cells. We defined the precision  $y_i$  by dividing mutant MKN45 cell to the total mutant number:

$$y_i = \frac{n_i}{m_i}. \quad (18)$$

We calculated the overall precision  $y_{pre}$  by dividing the total number of mutant MKN45 cells by the total number of mutant cells after merging all identified mutations.

$$y = \frac{\sum_i n_i}{\sum_i m_i} \quad (19)$$

We used a  $q$ -value of 0.01 and minimum mutant cell number of 5, when comparing the mutation calling results with other SOTA methods.

### Calling mtDNA mutation using `Signac`

The `Signac` R Package is designed for scATAC-seq analysis, which also equipped mtDNA mutation calling function by providing the `ReadMGATK`, `IdentifyVariants` functions [16]. We used the `ReadMGATK` function to read in the `mgatk` output, which was also used in `scMitoMut` mutation calling process.

For the `Signac`, we followed the vignette and default parameters: VMR larger than  $10^{-2}$ ; strand accordance at least 0.65; mutant cells at least five, where the mutant cells were defined by at least two mutant reads in both fwd and rev strands.

### Calling mtDNA mutation using `MQuad`

We utilized CellSnip-lite Version 1.2.3 to process the bam file and generate the AF matrix using the `-countORPHAN` option. The resulting files '`tag.AD.mtx`' and '`tag.DP.mtx`' were then inputted into `MQuad` Version 0.1.8 with a minimum DP of 5 specified (`-minDP 5`). Mutations were counted based on whether their DeltaBIC value exceeded the Kneel point, as indicated by the '`PASS_KP`' flag in the '`BIC_params.csv`' file. No minimum mutant cell number threshold was applied, making the filtering less stringent than the `Signac` or `scMitoMut` pipeline.

## Colorectal cancer scATAC-seq mitochondrial mutation analysis

### Cell type annotation

We used the CRC scATAC-seq with mtDNA kept (mtscATAC-seq) from Lareau et al. [22]; the scATAC-seq file can be download following the links in the `Signac` package's (v1.9.0) vignettes (<https://github.com/stuart-lab/signac/blob/1.9.0/vignettes/mito.Rmd>).

We followed the instructions provided in the vignette to preprocess the scATAC-seq data. We annotated open chromatin peaks with the `EnsDb.Hsapiens.v75` gene. The high-quality cells fulfilling all the following conditions were retained: `nCount_peaks > 1000`, `nCount_peaks < 50,000`, `pct_reads_in_DNase > 40`, `blacklist_ratio < 0.05`, `TSS.enrichment > 3`, `nucleosome_signal < 4`, and an average mitochondrial sequencing depth  $> 10$ . We normalized peaks with the term frequency inverse document frequency method. Peaks with a minimum of 10 counts are selected and

subjected to singular value decomposition for dimension reduction. Dimensions 2 to 50 were utilized to generate the UMAP and perform nearest neighbour-based clustering with LSI reduction, using a resolution of 0.5 and algorithm 3. Gene accessibility of *TREM1*, *EPCAM*, *PTPRC*, *IL1RL1*, *GATA3*, and *KIT* were employed to annotate cells into epithelial, basophil, myeloid, and T cell categories. These cell clusters were subsequently binarized into epithelial or blood cell categories.

### Comparing the mutation calling between scMitoMut and MQuad Signac

scMitoMut used the allele count matrix output from mgatk, which was downloaded using the link in the Signac vignette (<https://github.com/stuart-lab/signac/blob/1.9.0/vignettes/mito.Rmd>). The mutant cell was called by  $q$ -value  $<0.01$ , then the mutations were filtered with a mutant cell number threshold of 5.

For Signac, we adhered to the guidelines provided in the Signac package (v1.9.0) vignette (<https://github.com/stuart-lab/signac/blob/1.9.0/vignettes/mito.Rmd>). Mutations are called by imposing the following thresholds: VMR larger than  $10^{-2}$ , strand accordance  $>0.65$ , and mutant cells  $\geq 5$ , where the mutant cells were defined by at least two mutant reads in both fwd and rev strands ( $\text{cells\_conf\_detected} \geq 5$ ,  $\text{strand\_correlation} \geq 0.65$ , and  $\text{vmr} > 0.01$ ).

The MQuad result was retrieved from Kwok et al. [17], [https://github.com/aaronkwc/MQuad\\_paper\\_reproduced\\_results/blob/main/CRC\\_mtscATAC/CRC\\_mtscATAC.ipynb](https://github.com/aaronkwc/MQuad_paper_reproduced_results/blob/main/CRC_mtscATAC/CRC_mtscATAC.ipynb). They called all the mutations with  $\text{deltaBIC} < 0$ , which means an informative gain for binomial-mixture versus binomial distribution, without mutant cell number threshold.

For each mutation, we defined the precision of mutation calling as the consistency with cell types. For a specific mutation  $i$ , the mutant cells number in each lineage is  $n_i$  and  $m_i$  for epithelial and blood lineage, respectively. We defined the precision  $y_i$  by dividing the dominant lineage to the total mutant number:

$$y_i = \frac{\max(n_i, m_i)}{n_i + m_i}. \quad (20)$$

The precision  $y_i$  is a rational number in the range of [0.5, 1].

### Peripheral blood mononuclear cell 10× Genomics multiome data Cell type annotation

The 10× Genomics multiome dataset pbmc\_granulocyte\_sorted\_10k was downloaded from 10× Genomics website. We preprocessed the data according to the methods described in multiome data analysis vignette of Signac ([https://github.com/stuart-lab/signac/blob/1.9.0/vignettes/pbmc\\_multiomic.Rmd](https://github.com/stuart-lab/signac/blob/1.9.0/vignettes/pbmc_multiomic.Rmd)). Briefly, we selected the high-quality cells fulfilling all of the following condition:  $\text{nCount\_ATAC} < 100,000$ ,  $\text{nCount\_RNA} < 25,000$ ,  $\text{nCount\_ATAC} > 1000$ ,  $\text{nCount\_RNA} > 1000$ ,  $\text{nucleosome\_signal} < 2$ ,  $\text{TSS.enrichment} > 1$ . With Seurat 4.4.0, the transcriptome data were normalized using SCT transform, followed by Principal Components Analysis (PCA) and UMAP embedding. The reference annotation dataset used was from Hao et al. 2021 [37]. We annotated cells as CD14 Mono, CD16 Mono, CD4 Naive, CD4 TCM, CD4 TEM, CD8 Naive, CD8 TCM, CD8 TEM, Treg, CD4 CTL, gdT, MAIT, B memory, B intermediate, and B naive. We then classified these cell types into three categories: mono, T, or B.

### Calling mtDNA mutation using scMitoMut, Signac, and MQuad

The mgatk version 0.7.0 ran in 'tenx' mode options '-keep-duplicate' and no alignment-quality constrain '-alignment-quality -1', with the Bam file 'pbmc\_granulocyte\_sorted\_10k\_atac\_possorted\_bam.bam', and QC-passed cell list from 'filtered\_feature\_bc\_matrix.tar.gz' file downloaded from 10× Genomics website. The 'coverage.txt.gz' of mgatk output was used to evaluate the coverage distribution per locus for all cells.

For scMitoMut mutation calling, we utilized cells with an average mtDNA depth greater than 5. The mutation calling threshold was set at a  $q$ -value of less than 0.01 and required more than 10 mutant cells.

For Signac, we applied the parameters: VMR larger than  $10^{-2}$ , strand accordance  $>0.65$ , and mutant cells  $>10$ , where the mutant cells were defined by at least two mutant reads in both fwd and rev strands ( $\text{cells\_conf\_detected} \geq 5$ ,  $\text{strand\_correlation} \geq 0.65$ , and  $\text{vmr} > 0.01$ ).

For MQuad, the AF matrix was prepared by CellSnp-lite Version 1.2.3 with options '-UMItag None -countORPHAN -exclFLAG UNMAP -chrom chrM'. Then, MQuad Version 0.1.8 was used with minimum DP of five specified (-minDP 5). Mutations were called based on whether their DeltaBIC value exceeded the Kneel point, as indicated by the 'PASS\_KP' flag in the 'BIC\_params.csv' file, and more than 10 mutant cells per mutation.

### Brain 10× Genomics multiome data Cell type annotation

The 10× Genomics multiome dataset human\_brain\_3k was downloaded from 10× Genomics website. We selected the high-quality cells fulfilling all of the following conditions were retained:  $\text{nCount\_ATAC} < 100,000$ ,  $\text{nCount\_RNA} < 25,000$ ,  $\text{nCount\_ATAC} > 1000$ ,  $\text{nCount\_RNA} > 1000$ ,  $\text{nucleosome\_signal} < 2$ ,  $\text{TSS.enrichment} > 1$ . The transcriptome data were normalized using SCT transform, followed by PCA and UMAP embedding (with 30 PCA dimensions). The cell clusters were identified using FindNeighbours and FindClusters function with default parameters. Then we used the gene expression markers to annotated the cell clustering with MAP2 and SYN1 for neuron [38]; GFAP and ALDH1L1 for astrocyte [39]; P2RY12 and CX3CR1 for microglia [40]; MBP for oligodendrocyte [41].

### Calling mtDNA mutation using scMitoMut, Signac, and MQuad

The mgatk version 0.7.0 ran in 'tenx' mode options '-keep-duplicate' and no alignment-quality constrain '-alignment-quality -1' with the Bam file 'human\_brain\_3k\_atac\_possorted\_bam.bam' and QC-passed cell list in 'filtered\_feature\_bc\_matrix.tar.gz' downloaded from 10× Genomics website. Within the mgatk output, the coverage per locus per cell file 'coverage.txt.gz' was used to evaluate the coverage distribution per locus for all cells.

We called the mtDNA mutation using scMitoMut with the cells with average mtDNA depth larger than 5. The thresholds were set to:  $q$ -value  $<0.01$  and mutant cell number  $>10$ .

For Signac, we applied the parameters: VMR larger than  $10^{-2}$ , strand accordance  $>0.65$ , and mutant cells  $>10$ , where the mutant cells were defined by at least two mutant reads in both fwd and rev strands ( $\text{cells\_conf\_detected} >10$ ,  $\text{strand\_correlation} \geq 0.65$ , and  $\text{vmr} > 0.01$ ).

For MQuad, the AF matrix was prepared by CellSnp-lite Version 1.2.3 using the bam file with options '-UMItag None -countORPHAN -exclFLAG UNMAP -chrom chrM'. Then, MQuad

Version 0.1.8 was used with minimum DP of 5 specified ( $-\text{minDP } 5$ ). Mutations were counted based on whether their DeltaBIC value exceeded the Kneel point, as indicated by the 'PASS\_KP' flag in the 'BIC\_params.csv' file, and more than 10 mutant cells per mutation.

## Statistics

### Statistical test

The two-sided Wilcoxon rank-sum test was used to compare continuous unpaired variables, while the two-sided Fisher exact test was applied to categorical data. The FDR was applied for multiple tests correction. The heatmap is sorted using the 'memo sort' method adapted from <https://gist.github.com/armish/564a65ab874a770e2c26>.

### Clustering cell types using mtDNA mutation

Let  $C$  be the set of cell types and  $M$  be the set of mutations. We defined  $N_{ij}$  as the number of cells of type  $i \in C$  that have mutation  $j \in M$ . For each mutation  $j \in M$ , we normalized the number of mutant cells per cell type. Define  $T_j$  as the total number of cells with mutation  $j$ :

$$T_j = \sum_{i \in C} N_{ij}. \quad (21)$$

The normalized mutation frequency  $F_{ij}$  for cell type  $i$  and mutation  $j$  is given by

$$F_{ij} = \frac{N_{ij}}{T_j}. \quad (22)$$

Using the normalized mutation frequency matrix  $F$ , we calculate the Euclidean distance between cell types based on their mutation frequencies followed by the hierarchical clustering using complete methods in R `hclust` function [42].

### Key Points

- We present a statistical method using beta-binomial distribution to accurately detect mitochondrial lineage-related mutations at the single-cell level.
- `scMitoMut` is a well-designed R package for calling mtDNA mutations, offering high memory and CPU efficiency.
- Compared to existing methods, `scMitoMut` shows improved sensitivity and precision in identifying lineage-related mutations.
- The enhanced sensitivity of `scMitoMut` enables broader application of mtDNA mutations-based lineage tracing.

## Acknowledgements

We thank Lionel Chiron and all the members of the Perié laboratory for helpful discussions. We thank Veronica A. Raker for proofreading the text.

## Author contributions

Conceptualization was done by W.S., who also conducted formal analysis with the assistance of D.G. W.S. handled methodology, software, data curation and writing-original draft. L.P. conducted funding acquisition, provided supervision, and writing-review & editing.

## Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors declare no conflict of interest.

## Funding

This work was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme ERC StG 758170-Microbar (to L.P.) and the Institut Thématique Multi-Organismes Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, administered by INSERM (to L.P.).

## Data availability

The 1-k BJ dataset with 1% MKN45 spike-in for single-cell DNA sequencing was obtained from <https://www.10xgenomics.com/cn/datasets/1-k-bj-with-1-percent-mkn-45-spike-in-1-standard-1-1-0>. The human PBMC multiome sequencing data were downloaded from <https://www.10xgenomics.com/cn/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0>. The human brain multiome sequencing data were downloaded from <https://www.10xgenomics.com/cn/datasets/frozen-human-healthy-brain-tissue-3-k-1-standard-2-0-0>. The human Colorectal scATAC-seq dataset was downloaded following the Signac Vignette at: <https://github.com/stuart-lab/signac/blob/1.9.0/vignettes/mito.Rmd>. The PBMC single-cell reference annotation was downloaded following the link in the Vignette at: [https://github.com/stuart-lab/signac/blob/1.9.0/vignettes/pbmc\\_multiomic.Rmd](https://github.com/stuart-lab/signac/blob/1.9.0/vignettes/pbmc_multiomic.Rmd).

## Code availability

The code for all analyses in this study is available on GitHub ([https://github.com/TeamPerie/scMitoMut\\_paper\\_SUN\\_et\\_al](https://github.com/TeamPerie/scMitoMut_paper_SUN_et_al)). The `scMitoMut` package is available on Bioconductor (<https://bioconductor.org/packages/release/bioc/html/scMitoMut.html> or <https://doi.org/10.18129/B9.bioc.scMitoMut>).

## References

1. Sankaran VG, Weissman JS, Zon LI. Cellular barcoding to decipher clonal dynamics in disease. *Science* 2022;**378**:eabm5874. <https://doi.org/10.1126/science.abm5874>
2. Naik SH, Schumacher TN, Perié L. Cellular barcoding: a technical appraisal. *Exp Hematol* 2014;**42**:598–608. <https://doi.org/10.1016/j.exphem.2014.05.003>
3. Lee-Six H, Øbro NF, Shepherd MS. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* 2018;**561**:473–8. <https://doi.org/10.1038/s41586-018-0497-0>
4. Roerink SF, Sasaki N, Lee-Six H. et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* 2018;**556**:457–62. <https://doi.org/10.1038/s41586-018-0024-3>
5. Fabre MA, de Almeida JG, Fiorillo E. et al. The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* 2022;**606**:335–42. <https://doi.org/10.1038/s41586-022-04785-z>
6. Williams N, Lee J, Mitchell E. et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* 2022;**602**:162–8. <https://doi.org/10.1038/s41586-021-04312-6>
7. Ogasawara Y, Nakayama K, Tarnowska M. et al. Mitochondrial DNA spectra of single human CD34+ cells, T cells, B cells, and



- granulocytes. *Blood* 2005;**106**:3271–84. <https://doi.org/10.1182/blood-2005-01-0150>
8. Ludwig LS, Lareau CA, Ulirsch JC. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* 2019;**176**:1325–1339.e22. <https://doi.org/10.1016/j.cell.2019.01.022>
  9. Xu J, Nuno K, Litzenburger UM. et al. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *Elife* 2019;**8**:e45105. <https://doi.org/10.7554/eLife.45105>
  10. Picard M. Blood mitochondrial DNA copy number: what are we counting? *Mitochondrion* 2021;**60**:1–11. <https://doi.org/10.1016/j.mito.2021.06.010>
  11. Khrapko K, Collier HA, André PC. et al. Mitochondrial mutational spectra in human cells and tissues. *Proc Natl Acad Sci* 1997;**94**:13798–803. <https://doi.org/10.1073/pnas.94.25.13798>
  12. Marcelino LA, Thilly WG. Mitochondrial mutagenesis in human cells and tissues. *Mutat Res* 1999;**434**:177–203. [https://doi.org/10.1016/S0921-8777\(99\)00028-2](https://doi.org/10.1016/S0921-8777(99)00028-2)
  13. Malikic S, Jahn K, Kuipers J. et al. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat Commun* 2019;**10**:2750. <https://doi.org/10.1038/s41467-019-10737-5>
  14. Kuipers J, Singer J, Beerenwinkel N. Single-cell mutation calling and phylogenetic tree reconstruction with loss and recurrence. *Bioinformatics* 2022;**38**:4713–9. <https://doi.org/10.1093/bioinformatics/btac577>
  15. Campbell P, Chapman MS, Przybilla M. et al. Mitochondrial mutation, drift and selection during human development and ageing. *Research Square* 2023. <https://doi.org/10.21203/rs.3.rs-3083262/v1>
  16. Stuart T, Srivastava A, Madad S. et al. Single-cell chromatin state analysis with Signac. *Nat Methods* 2021;**18**:1333–41. <https://doi.org/10.1038/s41592-021-01282-5>
  17. Kwok AWC, Qiao C, Huang R. et al. MQuad enables clonal substructure discovery using single cell mitochondrial variants. *Nat Commun* 2022;**13**:1205. <https://doi.org/10.1038/s41467-022-28845-0>
  18. Shin H-T, Choi Y-L, Yun JW. et al. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. *Nat Commun* 2017;**8**:1377. <https://doi.org/10.1038/s41467-017-01470-y>
  19. Cagan A, Baez-Ortega A, Brzozowska N. et al. Somatic mutation rates scale with lifespan across mammals. *Nature* 2022;**604**:517–24. <https://doi.org/10.1038/s41586-022-04618-z>
  20. Gerstung M, Papaemmanuil E, Campbell PJ. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* 2014;**30**:1198–204. <https://doi.org/10.1093/bioinformatics/btt750>
  21. Huang X, Huang Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* 2021;**37**:4569–71. <https://doi.org/10.1093/bioinformatics/btab358>
  22. Lareau CA, Ludwig LS, Muus C. et al. Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat Biotechnol* 2021;**39**:451–61. <https://doi.org/10.1038/s41587-020-0645-6>
  23. Benaglia T, Chauveau D, Hunter DR. et al. Mixtools: an R package for analyzing mixture models. *J Stat Softw* 2010;**32**:1–29.
  24. Yee TW. *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York: Springer; 2015.
  25. Story B, Velten L, Mönke G. et al. Mitoclone2: an R package for elucidating clonal structure in single-cell RNA-sequencing data using mitochondrial variants. *NAR Genom Bioinform* 2024;**6**:lqae095. <https://doi.org/10.1093/nargab/lqae095>
  26. Spence JR, Lauf R, Shroyer NF. Vertebrate intestinal endoderm development. *Dev Dyn* 2011;**240**:501–20. <https://doi.org/10.1002/dvdy.22540>
  27. Cumano A, Godin I. Ontogeny of the hematopoietic system. *Annu Rev Immunol* 2007;**25**:745–85. <https://doi.org/10.1146/annurev.immunol.25.022106.141538>
  28. Ginhoux F, Prinz M. Origin of microglia: current concepts and past controversies. *Cold Spring Harb Perspect Biol* 2015;**7**:a020537. <https://doi.org/10.1101/cshperspect.a020537>
  29. Thion MS, Ginhoux F, Garel S. Microglia and early brain development: an intimate journey. *Science* 2018;**362**:185–9. <https://doi.org/10.1126/science.aat0474>
  30. Regev A, Teichmann SA, Lander ES. et al. Human cell atlas meeting participants. *Elife* 2017;**6**:e27041. <https://doi.org/10.7554/eLife.27041>
  31. Nitsch L, Lareau CA, Ludwig LS. Mitochondrial genetics through the lens of single-cell multi-omics. *Nat Genet* 2024;**56**:1355–65. <https://doi.org/10.1038/s41588-024-01794-8>
  32. Weng C, Yu F, Yang D. et al. Deciphering cell states and genealogies of human haematopoiesis. *Nature* 2024;**627**:389–98. <https://doi.org/10.1038/s41586-024-07066-z>
  33. Eddelbuettel D, Francois R, Allaire JJ. et al. *Rcpp: Seamless R and C++ Integration*. R package version 1.0.11. 2023. Available at: <https://CRAN.R-project.org/package=Rcpp>
  34. Mersmann O. *microbenchmark: Accurate Timing Functions*. R package version 1.5.0. 2024. Available at: <https://CRAN.R-project.org/package=microbenchmark>
  35. Bolker B, R Development Core Team. *bbmle: Tools for General Maximum Likelihood Estimation*. R package version 1.0.25.1. 2023. Available at: <https://CRAN.R-project.org/package=bbmle>
  36. Bolker BM. *Ecological Models and Data in R*. Princeton: Princeton University Press; 2008.
  37. Hao Y, Hao S, Andersen-Nissen E. et al. Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–3587.e29. <https://doi.org/10.1016/j.cell.2021.04.048>
  38. Larson ME, Greimel SJ, Amar F. et al. Selective lowering of synapsins induced by oligomeric  $\alpha$ -synuclein exacerbates memory deficits. *Proc Natl Acad Sci* 2017;**114**:E4648–57. <https://doi.org/10.1073/pnas.1704698114>
  39. Forrest SL, Kim JH, Crockford DR. et al. Distribution patterns of astrocyte populations in the human cortex. *Neurochem Res* 2023;**48**:1222–32. <https://doi.org/10.1007/s11064-022-03700-2>
  40. Bedolla A, McKinsey G, Ware K. et al. Finding the right tool: a comprehensive evaluation of microglial inducible cre mouse models. *bioRxiv* 2023. <https://doi.org/10.1101/2023.04.17.536878>
  41. Tanaka J, Hokama Y, Nakamura H. Myelin basic protein as a possible marker for Oligodendroglioma. *Acta Pathol Jpn* 1988;**38**:1297–303. <https://doi.org/10.1111/j.1440-1827.1988.tb02280.x>
  42. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2023.