

# Mit oclone2: an R package for elucidating clonal structure in single-cell RNA-sequencing data using mitochondrial variants

**Journal Article****Author(s):**

Story, Benjamin; Velten, Lars; Mönke, Gregor; Annan, Ahrmad; Steinmetz, Lars

**Publication date:**

2024-09

**Permanent link:**

<https://doi.org/https://doi.org/10.3929/ethz-b-000690204>

**Rights / license:**

[Creative Commons Attribution 4.0 International](#)

**Originally published in:**

NAR Genomics and Bioinformatics 6(3), <https://doi.org/10.1093/nargab/lqae095>

# Mitoclone2: an R package for elucidating clonal structure in single-cell RNA-sequencing data using mitochondrial variants

Benjamin Story<sup>1,2</sup>, Lars Velten<sup>3</sup>, Gregor Mönke<sup>4</sup>, Ahrmad Annan<sup>5</sup> and Lars Steinmetz<sup>1,6,7,\*</sup>

<sup>1</sup>Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

<sup>2</sup>Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

<sup>3</sup>Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

<sup>4</sup>Developmental Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

<sup>5</sup>Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

<sup>6</sup>Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

<sup>7</sup>Stanford Genome Technology Center, Palo Alto, CA, USA

\*To whom correspondence should be addressed. Tel: +49 62213870; Email: [lars.steinmetz@embl.de](mailto:lars.steinmetz@embl.de)

## Abstract

Clonal cell population dynamics play a critical role in both disease and development. Due to high mitochondrial mutation rates under both healthy and diseased conditions, mitochondrial genomic variability is a particularly useful resource in facilitating the identification of clonal population structure. Here we present mitoClone2, an all-inclusive R package allowing for the identification of clonal populations through integration of mitochondrial heteroplasmic variants discovered from single-cell sequencing experiments. Our package streamlines the investigation of this phenomenon by providing: built-in compatibility with commonly used tools for the delineation of clonal structure, the ability to directly use multiplexed BAM files as input, annotations for both human and mouse mitochondrial genomes, and helper functions for calling, filtering, clustering, and visualizing variants.

## Introduction

Genetic variability between clonal populations within an individual presentation of cancer, termed intratumor heterogeneity (IH), is a hallmark of most cancers and considered a major driver of cancer progression and relapse (1). Similarly, studies in healthy tissues aimed at lineage tracing in the context of development can also benefit from the ability to distinguish between the progeny of different founder cells. A historical genomics hurdle in identifying intra-sample genetic variation was the difficulty in distinguishing coexisting clonal populations from bulk sequencing data. With the advent of high-throughput single-cell sequencing technologies, scientists are now better able to capture and delineate cell-to-cell variability across -omics technologies.

In the context of cancer, a quantitative analysis of IH allows for the creation of cancer phylogenetic trees (CPTs) which trace the path of a tumor as it evolves from healthy cells. Uncovering the key steps involved in this pathogenic transformation is essential to understanding cancer initiation/relapse and can expose genetic changes that are vulnerable to therapeutic intervention. Furthermore, beyond applications in disease, the ability to track cells as they proliferate and differentiate is of great relevance to the field of developmental biology given that most methods for lineage tracing are costly and invasive (2). Using endogenous mitochondrial genetic variability, which is easily accessible, may help overcome these obstacles and provide a clearer picture of phenomena such as stem cell proliferation and tissue regeneration.

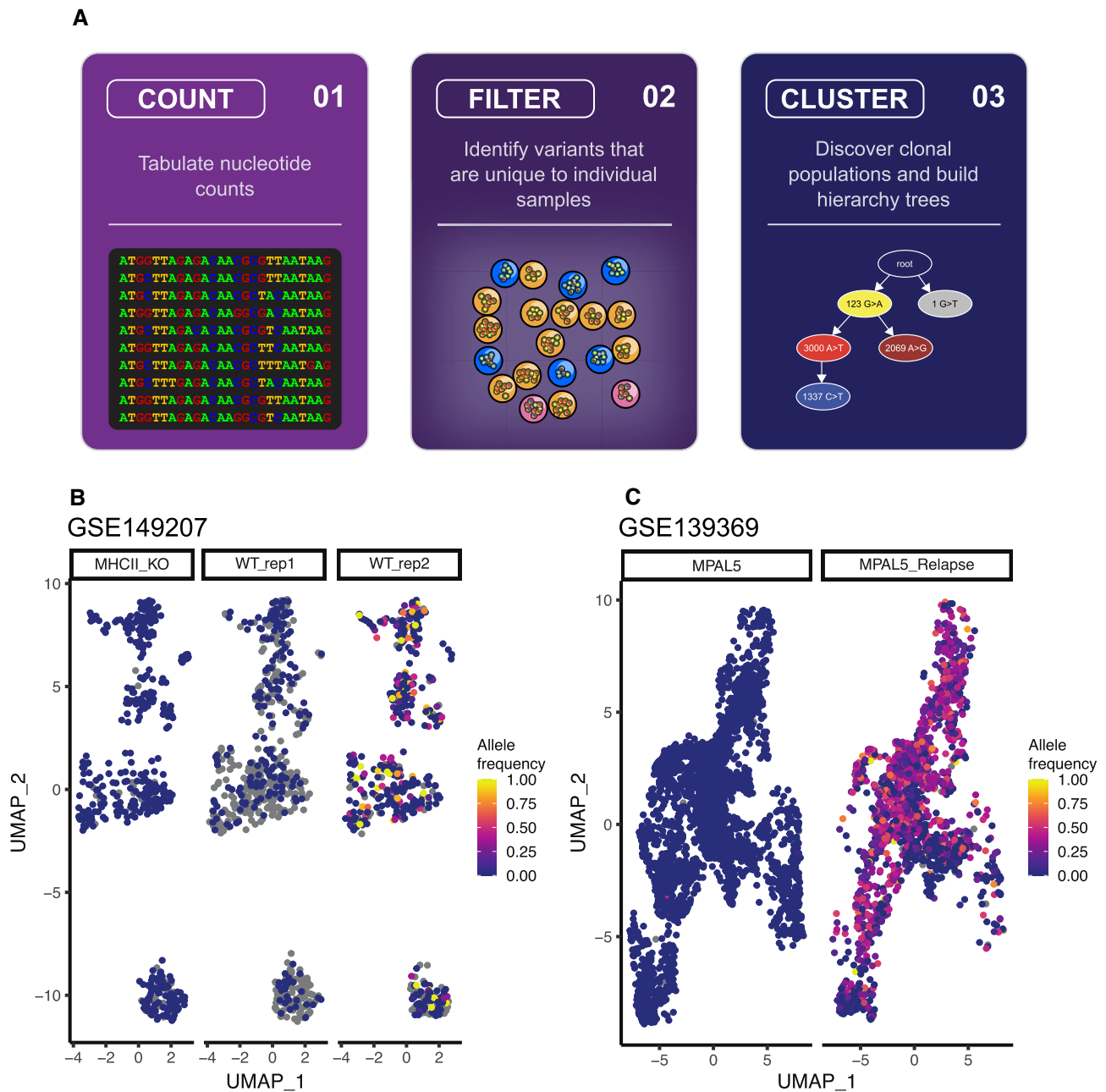
Despite the abundance of sequencing technologies, methods that computationally mine single-cell datasets for the presence of somatic nuclear variants are plagued by a lack of coverage and the noise inherent to single-cell sequencing. And although methods exist for genotyping single cells at sites of interest, doing so in tandem with molecular profiling methods such as RNA-seq is challenging, especially in scenarios marked by a low mutational burden within the nucleus. However, recent studies have shown that mitochondrial reads are a potential treasure trove for uncovering clone-specific mutations (3,4). The mutation rate of the mitochondrial genome is drastically elevated compared to the nuclear genome and the RNA coverage in single-cell sequencing experiments is high (5).

Capitalizing on this observation, we describe a new package, providing an all-in-one R library for the computational investigation and detection of intercellular heterogeneity with a focus on mitochondrial variants extracted from single-cell RNA-sequencing (scRNA-seq). Our package builds on existing methods including an initial prototype (6) and deepSNV (7). We enable biologists with limited programming experience to detect intercellular heterogeneity from mitochondrial reads in their single-cell datasets, all within R. We expand beyond the original mitoClone package by providing extra functionality including fast variant extraction from BAM files containing multiple single cells (i.e., multiplexed), out-of-the-box functionality (i.e., no external dependencies), extended compatibility with both the mouse mm10 and human hg19 genomes, and access to more CPT-building tools (8).

Received: February 11, 2024. Revised: June 14, 2024. Editorial Decision: July 22, 2024. Accepted: July 23, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.

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.** Undetected clonal variants identified in public datasets by applying mitoClone2. **(A)** Illustration of the mitoClone2 workflow. **(B)** UMAP of cells from a single-cell RNA-seq experiment conducted on mouse thymuses, focusing on three mice (one MHCII knock-out mouse and two wild-type mice) that share a consistent C57BL/6 genetic background. Color represents allele frequency of the chrM:13575 T>C mutation which marks a clonal population of cells. **(C)** UMAP of cells from a single-cell RNA-seq experiment of a mixed-phenotype acute leukemia patient showing acquisition of a novel clonal population following cancer relapse. Color represents allele frequency of the chrM:15596 G>C mutation. All gray points represent cells with insufficient coverage to call a variant.

mitoClone2 is now optimized for easy integration into bioinformatic pipelines. Finally, our approach matches or exceeds alternative methods in terms of performance and offers a level of portability that is not found in other software packages (Supplementary Figure S1).

## Materials and methods

mitoClone2 quantifies mitochondrial and nuclear variants directly from input BAM files. The package accommodates both individual BAM files per cell or multiplexed BAM files (e.g., 10× Genomics Chromium/Visium). Variants of interest are

identified through two possible methods. Common to both is that variants present within the population must meet certain quality thresholds that are pre-defined by the user. These include the coverage over specific variant sites, the proportion of the overall and discrete populations allowed to exhibit a variant, the minimum variant allele frequency, and the underlying depth and read-quality of bases at variant sites (Figure 1A).

The first method is used in cases where data from multiple individuals/samples (e.g. different patients or cohorts) are available. All variants are extracted and then filtered to identify mutations unique to individuals. In such cases false-positive events due to biological (e.g. RNA-editing) or techni-

cal (e.g. PCR bias) noise can be excluded. The second method uses an exclusion list which is necessary when only a single sample is available. In this case, a database of problematic variants is provided that allows for the exclusion of sites that are in low-complexity/repeat regions, known RNA-editing sites, or arising due to re-occurring methodological errors (e.g., due to specific aligners or sequencing technologies) (9). Further information on both filtering methods can be found in the tutorial provided in the [Supplementary Section \(Supplementary Figures S3–S8\)](#).

After either of these filtering processes, what remains are candidate variants underlying biological differences (i.e. clonal populations). For cancer samples, we generate matrices of variant genotypes for compatibility with commonly used tools for generating CPTs (8,10). Additionally, instructions are provided for transferring metadata to other commonly-used packages for single-cell analysis, such as Seurat (11).

## Results

To illustrate the functionality of our package, we investigated the presence of mitochondrial variants in two published scRNA-seq datasets. The first was a study of the mouse thymus using Smart-seq2 and SMARTer single-cell RNA sequencing data, GEO accession: GSE149207 (12). The second included 10× Genomics 3′ single-cell RNA sequencing data at multiple time-points from human mixed phenotype leukemia patients, GEO accession: GSE139369 (13). The variants discovered in both cases were, up until now, unpublished (Figure 1B, C, [Supplementary Figure S2](#)). Our demonstration shows that by using mitoClone2, mitochondrial variants demarcating clonal populations are readily identified from scRNA-seq data across species. In the cancer context, we are able to detect novel clonal populations after cancer relapse.

## Discussion

The mitoClone2 package enables the identification of clonal populations with ease by harnessing mitochondrial variants detected using single-cell sequencing technology. Furthermore, by providing a tool that works with scRNA-seq, we allow researchers to retain the ability to profile the transcriptome while simultaneously collecting genotype information (6). Our package makes elucidating clonal structure from any single-cell dataset straightforward and paves the way for downstream characterization of clonal cell populations such as through differential gene expression testing or clinical profiling. The method is applicable not only to other cancer datasets but also to other diseases and model systems.

## Data availability

The release-version of the software is freely available to install directly in R and is hosted by Bioconductor: <https://bioconductor.org/packages/release/bioc/html/mitoClone2.html>. The code for all present and past versions is also hosted freely by Bioconductor: <https://code.bioconductor.org/browse/mitoClone2/>. The developmental version is available at: <https://github.com/benstorry/mitoClone2>. Code for the new supplementary section is available at: [https://github.com/benstorry/mitoClone2\\_supplemental](https://github.com/benstorry/mitoClone2_supplemental).

## Supplementary data

[Supplementary Data](#) are available at NARGAB Online.

## Acknowledgements

We would like to acknowledge Mike L. Smith for help in the Bioconductor submission process. Further thanks is due to Daniel Schraivogel, Katie Zeier, Miljan Petrović and Theophilus T. Tettey for reviewing early manuscript drafts.

## Funding

Emerson Collective grant [643577].

## Conflict of interest statement

Lars Steinmetz is co-founder of Levitas Bio, Recombia Biosciences, Sophia Genetics, and a consultant for several companies on genetic analysis. All other authors declare no competing interests.

## References

- Swanton,C. (2012) Intratumor heterogeneity: evolution through space and time. *Cancer Res.*, **72**, 4875–4882.
- Roy,E., Neufeld,Z., Livet,J. and Khosrotehrani,K. (2014) Concise review: understanding clonal dynamics in homeostasis and injury through multicolor lineage tracing. *Stem Cells (Dayton, Ohio)*, **32**, 3046–3054.
- Ludwig,L.S., Lareau,C.A., Ulirsch,J.C., Christian,E., Muus,C., Li,L.H., Pelka,K., Ge,W., Oren,Y., Brack,A., *et al.* (2019) Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell*, **176**, 1325–1339.
- Xu,J., Nuno,K., Litzenburger,U.M., Qi,Y., Corces,M.R., Majeti,R. and Chang,H.Y. (2019) Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *eLife*, **8**, e45105.
- Allio,R., Donega,S., Galtier,N. and Nabholz,B. (2017) Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. *Mol. Biol. Evol.*, **34**, 2762–2772.
- Velten,L., Story,B.A., Hernández-Malmierca,P., Raffel,S., Leonce,D.R., Milbank,J., Paulsen,M., Demir,A., Szu-Tu,C., Frömel,R., *et al.* (2021) Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics. *Nat. Commun.*, **12**, 1366.
- Gerstung,M., Beisel,C., Rechsteiner,M., Wild,P., Schraml,P., Moch,H. and Beerenwinkel,N. (2012) Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.*, **3**, 811.
- Jahn,K., Kuipers,J. and Beerenwinkel,N. (2016) Tree inference for single-cell data. *Genome Biol.*, **17**, 86.
- Picardi,E., Horner,D.S. and Pesole,G. (2017) Single-cell transcriptomics reveals specific RNA editing signatures in the human brain. *RNA (New York, N.Y.)*, **23**, 860–865.
- Malikic,S., Mehrabadi,F.R., Ciccolella,S., Rahman,M.K., Ricketts,C., Haghshenas,E., Seidman,D., Hach,F., Hajirasouliha,I. and Sahinalp,S.C. (2019) PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Res.*, **29**, 1860–1877.
- Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.

12. Karimi,M.M., Guo,Y., Cui,X., Pallikonda,H.A., Horková,V., Wang,Y.-F., Gil,S.R., Rodriguez-Esteban,G., Robles-Rebollo,I., Bruno,L., *et al.* (2021) The order and logic of CD4 versus CD8 lineage choice and differentiation in mouse thymus. *Nat. Commun.*, **12**, 99.
13. Granja,J.M., Klemm,S., McGinnis,L.M., Kathiria,A.S., Mezger,A., Corces,M.R., Parks,B., Gars,E., Liedtke,M., Zheng,G.X.Y., *et al.* (2019) Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.*, **37**, 1458–1465.