Genome Biology

RESEARCH

Open Access

Clonal expansion dictates the efficacy of mitochondrial lineage tracing in single cells

Xin Wang[1†], Kun Wang[2†], Weixing Zhang[3†], Zhongjie Tang[3], Hao Zhang[1], Yuying Cheng[1,4], Da Zhou[2,5], Chao Zhang[6], Wen-Zhao Zhong[6], Qing Ma[1,7*], Jin Xu[3*] and Zheng Hu[1*]

†Xin Wang, Kun Wang and Weixing Zhang contributed equally to this work.

*Correspondence: qing.ma@siat.ac.cn; xujin7@mail.sysu.edu.cn; zheng.hu@siat.ac.cn

[1] State Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2] School of Mathematical Sciences, Xiamen University, Xiamen, China
[3] State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, China
[4] School of Life Sciences, Henan University, Kaifeng, China
[5] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
[6] Guangdong Lung Cancer Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China
[7] Shenzhen Key Laboratory of Synthetic Genomics, Guangdong Provincial Key Laboratory of Synthetic Genomics, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Abstract

Background: Mitochondrial DNA (mtDNA) variants hold promise as endogenous barcodes for tracking human cell lineages, but their efficacy as reliable lineage markers are hindered by the complex dynamics of mtDNA in somatic tissues.

Results: Here, we use computational modeling and single-cell genomics to thoroughly interrogate the origin and clonal dynamics of mtDNA variants across various biological settings. Our findings reveal that the majority of mtDNA variants which are specifically present in a cell subpopulation, termed subpopulation-specific variants, are pre-existing heteroplasmies in the first cell instead of de novo somatic mutations during divisions. Moreover, subpopulation-specific variants demonstrate limited discriminatory power among different genuine lineages under weak clonal expansion; however, certain subpopulation-specific variants with consistently high frequencies among a subpopulation are capable of faithfully labeling cell lineages in scenarios of stringent clonal expansion, such as strongly expanded T cell populations in diseased conditions and clonal hematopoiesis in aged individuals. Inspired by our simulations, we introduce a lineage informative score, facilitating the identification of reliable mitochondrial lineage tracing markers across different modalities of single-cell genomic data.

Conclusions: Combining computational modeling and single-cell sequencing, our study reveals that the performance of mitochondrial lineage tracing is highly dependent on the extent of clonal expansion, which thus should be considered when applying mitochondrial lineage tracing.

Keywords: Lineage tracing, mtDNA variants, Clonal dynamics, Single-cell genomics

Background

Mitochondrial DNA (mtDNA) mutations naturally occur and accumulate over cell divisions, constituting an alternate source of heritable markers to somatic nuclear genomic mutations for tracing human cell lineages in vivo. Compared to nuclear genomic mutations, mtDNA mutations arise at 10 to 100-fold higher rate [1], indicating a possible higher resolution in reconstructing cell lineages. Although lineage

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

tracing techniques based on gene editing have achieved high resolution at a large scale [2–7], their application in direct clinical samples is largely limited due to their dependence on genetic engineering of synthetic DNA barcodes. In recent years, mtDNA mutations have been utilized to study the clonal dynamics underlying normal development and disease [8–12]. Rapid development of methods for simultaneous single-cell profiling of mtDNA mutations and chromatin accessibility (e.g. mtscATAC-seq [11, 13]) or transcriptome (e.g. MAESTER [14]) or both (e.g. ReDeeM [9]) has greatly accelerated and expanded the application of mitochondrial lineage tracing technologies [15].

However, the complex processes of mitochondrial replication over cell divisions introduces a high degree of dynamics, thus complicating the use of mtDNA mutations as clonal lineage markers [16]. For example, mitochondrial bottleneck, a process wherein effective mitochondrial content is drastically reduced, has been reported during oogenesis and embryogenesis in various species [17]. This process is critical for mitochondrial quality control, but at the same time, it naturally causes a dramatic shift of heteroplasmies of mtDNA mutations, undermining the propagation of mtDNA mutations over cell divisions. Additionally, random segregation of mitochondrial DNA into daughter cells could lead to the loss of mtDNA mutations in the progeny, which further introduces more complexities into mitochondrial lineage tracing system [18, 19]. Moreover, clonal expansion of certain cell lineages within a cell population rapidly changes the clonal structure and simultaneously alters the mtDNA mutational dynamics, which brings even more uncertainties to mitochondrial lineage tracing. Besides these biological complexities, the detection of low-frequency mtDNA mutations at single-cell level is still technically challenging, hindering mtDNA mutation-based lineage reconstruction. Collectively, the complexity of mtDNA mutational dynamics and limited detectability for lineage-informative variants raised doubts about the efficacy of mitochondrial lineage tracing [16]. Therefore, a systematic study on characterizing the genuine clonal mtDNA mutations under different biological scenarios is warranted.

To address this, we combined computational simulations and single-cell genomic data to thoroughly investigate the mtDNA mutational dynamics and the effectiveness of mtDNA mutations in tracking clonal lineages under different biological scenarios. First, our simulations results showed that many of the subpopulation-specific variants (SSVs) were already present (pre-existing) in the initial cell, which is a distinct feature of mtDNA unlike nuclear genomic variants-based lineage tracing. Importantly, we found that mitochondrial lineage tracing failed to reconstruct true cell lineages in context with weak clonal expansion. However, certain mtDNA mutations characterized by a high average variant allele frequency (VAF) but low variance did show a better performance under strong clonal expansion in both simulations and experimental datasets. Lastly, we developed a metric, lineage informative score (LIS), for assessing the robustness of mtDNA mutations on tracing cell lineages. We then applied it to real single-cell data with different modalities and successfully navigated high-confidence mtDNA variants for inferring clonal structure. Overall, our study interrogated the dynamics of mtDNA mutations and demonstrated scenarios with significant clonal expansions as optimal biological settings for applying mitochondrial lineage tracing methods.

Results

Computational simulation of mtDNA mutation dynamics over cell divisions

To better understand the dynamics of mtDNA mutations, we developed a computational framework to simulate the replication and random segregation of mitochondrial genomes in a dividing cell population with distinct intensities of clonal expansion over cell divisions (Fig. 1a-b, Additional file 1: Fig. S1a-b, Methods). Briefly, mitochondrial genomes in the first cell were initialized with pre-existing mutations where the initial heteroplasmy levels (namely VAF) followed a power-law distribution [20] (Fig. 1c). During the replication of mtDNA, de novo mutations arise at a specific rate ($\mu = 5 \times 10^{-8}$ per replication per base pair [21]) (Fig. 1b). Subsequently, all mitochondrial genomes, regardless of whether they are newly generated or not, were randomly segregated into two daughter cells during cell division (total mutation rate per

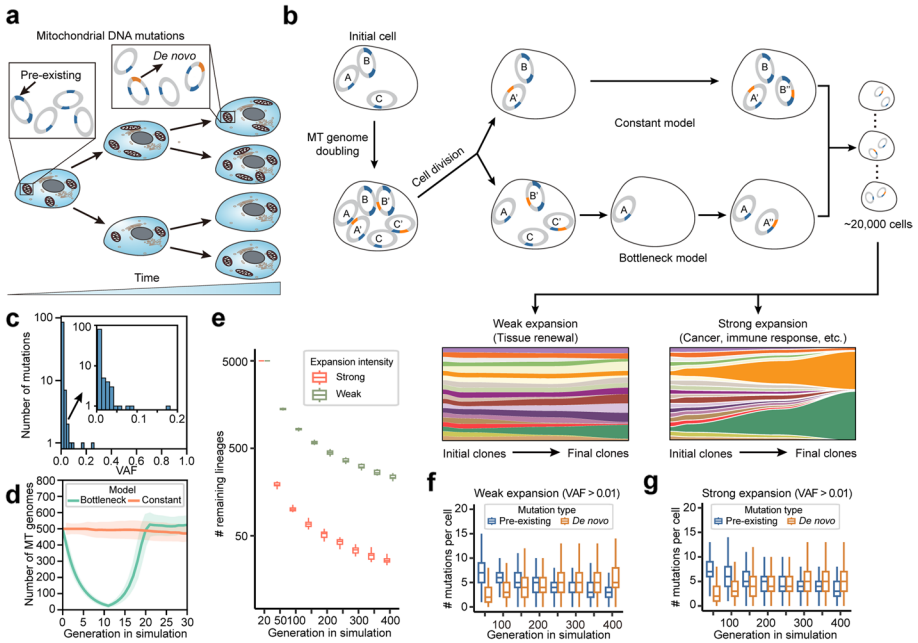

Fig. 1 Computational simulation of mtDNA mutation dynamics over cell divisions. **a** Schematic of mitochondrial lineage tracing. Mitochondria that carry pre-existing mutations (blue bar) in the initial cell are replicated and randomly distributed into daughter cells, and de novo mutations (orange bar) arise during the following cell divisions. **b** Schematic of the simulation framework. The first cell was initialized with 500 mitochondrial genomes carrying around 100 pre-existing mutations (blue). During the mitochondrial DNA replication (e.g. A′ is a replicate of A), de novo mutations (orange) were generated and accumulated in the cell. After replication of MT genomes, all MT genomes were randomly segregated into two daughter cells. This replication-segregation process was repeated along with the cell divisions. In constant model, MT genome copies remained stable (~ 500) along the cell divisions. However, in the bottleneck model, the number of MT genomes first decreased to ~ 25 and then recovered back to around 500. After the recovery of MT genome copies, both constant model and bottleneck model experienced the same replication segregation of MT genomes until the population size reached 20,000. Next, 5,000 randomly sampled cells were subjected to weak expansion model or strong expansion model. For weak expansion model, many cell lineages persisted, whereas in strong expansion model, much fewer lineages were kept for future divisions, resulting a strong clonal expansion. **c** Histogram showing the VAF distribution of pre-existing mtDNA mutations in the initial cell. **d** Line plot showing the number of MT genomes over the simulation generation. The shaded area represents standard deviation. **e** Boxplot showing the number of remaining lineages over the simulation generation. **f-g** Boxplot showing the number of pre-existing and de novo mutations in each cell over the simulation generation in weak expansion model (**f**) and strong expansion model (**g**)

cell division is $u = 5 \times 10^{-8} \times 16{,}569 \times 500 \approx 0.4$, when assuming 500 mtDNA copies per cell [22]).

To interrogate the effects of mitochondrial bottleneck, a process in which the mtDNA copy number within a cell dramatically decreases [19], we designed two models wherein mtDNA copy number sharply drops and resumes (referred as bottleneck model hereafter) or maintain constant (referred as constant model hereafter) (Fig. 1d). When the cell population reached approximately 20,000 cells, 5,000 cells of both bottleneck model and constant model were subject to two distinct scenarios of clonal dynamics with different intensities of clonal expansion (clonal expansion coefficient, τ), including weak expansion ($\tau = 0.1$) and strong expansion ($\tau = 0.9$) (Methods). Weak expansion corresponds to normal developmental processes, where many cell lineages remain over stem cell self-renewal. In contrast, strong expansion corresponds to lineage turnover commonly observed in cancer progression, immune responses, etc.(Fig. 1b). In scenarios with strong clonal expansion, the cell population was dominated by a few cell lineages (< 1% of total) whereas significantly more cell lineages (~ 10%) persisted in the weak expansion model (Fig. 1e). As the simulation proceeded, we checked the mutation dynamics over the simulations and found that the number of detectable pre-existing mutations (VAF > 1%) per cell constantly decreased while the detectable de novo mutations (VAF > 1%) gradually accumulated in both weak expansion (Fig. 1f) and strong expansion (Fig. 1g). This is expected because a pre-existing heteroplasmy can be lost in descendant cells due to random drift caused by mitochondrial segregation [23]. Taken together, we developed a computational framework and successfully captured the dynamics of mtDNA mutations, enabling further investigations of the efficacy of mitochondrial lineage tracing.

The origin and efficacy of lineage-informative mtDNA variants

A subpopulation-specific variant (SSV) is defined by their unique presence in a cell subpopulation [13, 24]. To evaluate the efficacy of SSV on tracking real clonal lineages, we first adopted the method developed by MAESTER method [14] for the identification of SSVs from the simulated data (Fig. 2a, Methods). Based on the presence of these SSVs, cells could be assigned to various subpopulations ideally with closer lineage relationships. Given the availability of genuine cell-division history in our simulations, lineage relationship within each of these SSV-defined subpopulations could be assessed. We observed that cells within each SSV-defined subpopulation failed to aggregate in the ground-truth cell lineage tree in weak expansion model (Fig. 2b). Interestingly, strong clonal expansion resulted in a prominent pattern of phylogenetic aggregation for SSV-defined cell subpopulations, indicating high efficacy of these SSVs in distinguishing different cell lineages. This was further supported by a higher clone aggregation score (CAS), a metric for quantitatively assessing the accuracy of clone assignment (Methods), in strong versus weak expansion model (0.75 versus 0.47, Fig. 2b-c). Interestingly, we found that SSVs in weak expansion model were mostly pre-existing mutations even after 400 cell divisions (~ 99%, Fig. 2d), whereas some SSVs in strong clonal expansion model were indeed de novo mutations (~ 30%, Fig. 2d), indicating de novo mtDNA mutations expanded along with clonally expanded cells under stronger selection. Notably, in nuclear

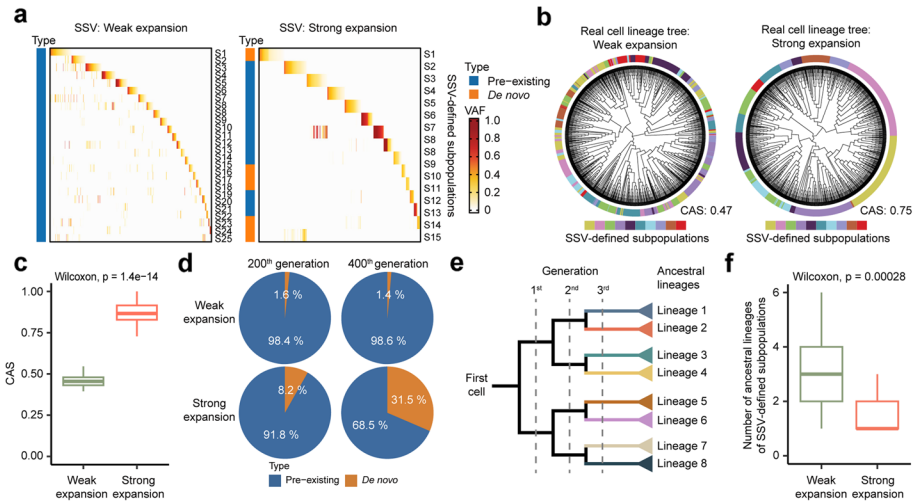

Fig. 2 The origin and efficacy of informative mtDNA variants. **a** Identification of subpopulation-specific variants (SSVs) in weak expansion model (left panel) and strong expansion model (right panel). Heatmap showing the VAF of SSVs in single cells. Each row represents an SSV, and each column represents a cell. The type of mutations was color coded. **b** Cell division tree (downsampled to 1,000 cells) showing the phylogeny of top 10 SSV-defined subpopulations in weak expansion (left panel) and strong expansion model (right panel). Clone aggregation score (CAS) was labeled next to the tree. **c** Boxplot showing the comparison of CAS between weak expansion and strong expansion. **d** Pie charts showing the origin of SSVs in weak expansion and strong expansion model from 200th generation (left panel) and 400th generation (right panel). **e** Schematic of the approach for assessing the lineage composition of SSV-defined clones. The 3rd generation ancestor of each cell is traced to analyze the lineage composition of SSV-defined subpopulations. **f** Boxplot showing the comparison of ancestral lineage composition of each SSV-defined subpopulation between weak expansion and strong expansion model

genomic mutation-based lineage tracing, germline mutations are stably propagated within the cell population and present in every single cell, thus losing lineage tracing capacity. However, pre-exsting mtDNA mutations can exist at different heteroplasmies due to the high copy number of mtDNA in single cells, so they can be passed on to a subpopulation and still retain the capacity for clonal tracking.

To further quantify the lineage tracing capacity of SSVs, we empirically defined cells in the third generation of divisions from a common progenitor cell as lineage ancestors (termed Lineage 1–8, Fig. 2e) and traced the origin of each SSV-defined subpopulation back to these eight lineages (Methods). In line with aforementioned findings, SSV-defined subpopulations in weak expansion model consisted of a large number of ancestral lineages (median = 3), thus displaying a highly mixed lineage composition. However, strong expansion model showed a small number of ancestral lineages (median = 1), suggesting better performance of SSVs in the context of strong clonal expansion (Fig. 2f). To comprehensively assess the efficacy of mitochondrial lineage tracing in more diverse contexts, we conducted the same analysis for mitochondrial constant model (without mitochondrial bottleneck, Additional file 1: Fig. S2), and the results were consistent with mitochondrial bottleneck model as shown in Fig. 2. Collectively, these results demonstrated that unlike nuclear genomic mutations, lineage-informative mtDNA variants can be pre-existing and they show better performance under contexts with strong clonal expansions.

The impact of clonal expansion on mitochondrial lineage tracing

To better characterize how clonal expansion affects the efficacy of mitochondrial lineage tracing, we conducted a systematic comparison between different intensities of clonal expansions. The clonal structure of the final cell population in weak expansion model is comprised of various weakly expanded cell populations from the initial clones (Fig. 3a). Moreover, cells carrying the same SSV showed mixed clone identities, failing to reconstruct the clonal structure (Fig. 3b). We calculated Shannon entropy to quantitatively assess the clonal composition of each SSV-defined subpopulation. A higher entropy indicates a higher diversity for the clonal composition, representing a poor lineage tracing performance of the corresponding SSV. We found that in the weak expansion model, all SSV-defined subpopulations showed high clonal diversity, demonstrating generally poor effectiveness of SSVs in the weak expansion model (Fig. 3c). However, in a cell population that is dominated by strongly expanded clones (Fig. 3d), SSV-defined subpopulations showed a much better concordance with the true clonal structure (Fig. 3e) and the clonal diversity was also lower (Fig. 3f), suggesting a higher efficacy of SSVs.

In addition to simulations, we also included real single-cell datasets to strengthen our findings. TCR (T-cell receptor) sequence is expanded along with the proliferation of T cells under immune responses, thus serving as a reliable clonal marker independent of mtDNA mutations. We made use of a publicly available dataset which profiled T cells isolated from treatment-naïve colorectal cancer (CRC) patients using Smart-seq2 [25]. The high sequencing coverage of mitochondrial genome and TCR region enabled us to simultaneously detect mtDNA mutations and assemble TCR sequences (Additional file 1: Fig. S3a-b), thus providing independent lineage information for evaluating the efficacy of mitochondrial lineage tracing. With the recovered TCR sequences, we grouped T cells into two subpopulations: cells within top 10 clones (in terms of clone size) were classified as strongly expanded and the remaining cells were classified as weakly expanded, resembling simulated strong expansion and weak expansion, respectively (Fig. 3g). In line with the simulations, SSVs identified

(See figure on next page.)

Fig. 3 The distinct efficacy of mitochondrial lineage tracing under weak and strong expansions. **a** Schematic of simulation under weak expansion. **b** Heatmap showing the VAF of SSVs under weak expansion. Different colors in the colorbar above the heatmap represent different real clones. **c** Bar plot showing the lineage diversity of SSV-defined subpopulations in weak expansion model as quantified by Shannon entropy. **d** Schematic of simulation under strong expansion. **e** Heatmap showing the VAF of SSVs under strong expansion. Different colors in the colorbar above the heatmap represent different real clones. **f** Bar plot showing the lineage diversity of SSV-defined subpopulations in strong expansion model as quantified by Shannon entropy. **g** Schematic of T cell isolation from colorectal tumor tissue, single-cell sequencing and TCR clonotyping. The strongly expanded group included cells from the top 10 TCR clones and the rest of cells were grouped as weakly expanded T cells. **h** Identification of SSVs in weakly expanded T cells from CRC P0825 sample. The colors in the colorbar above the heatmap represent different clones defined by TCR. **i** Bar plot showing the TCR diversity of each SSV-defined subpopulation in weakly expanded T cells as quantified by Shannon entropy. **j** Identification of SSVs in strongly expanded T cells from CRC P0825 sample. The colors in the colorbar above the heatmap represent different clones defined by TCR. **k** Bar plot showing the TCR diversity of each SSV-defined subpopulation in strongly expanded T cells as quantified by Shannon entropy. **l** Schematic of $CD34^+$ HSPC isolation, HSPC colony culture and single-cell sequencing. **m** Identification of SSVs in strongly expanded HSPC colonies from Donor 1. The colors in the colorbar above the heatmap represent different colonies. **n** Bar plot showing the colony diversity of each SSV-defined subpopulation in strongly expanded HSPC colonies as quantified by Shannon entropy

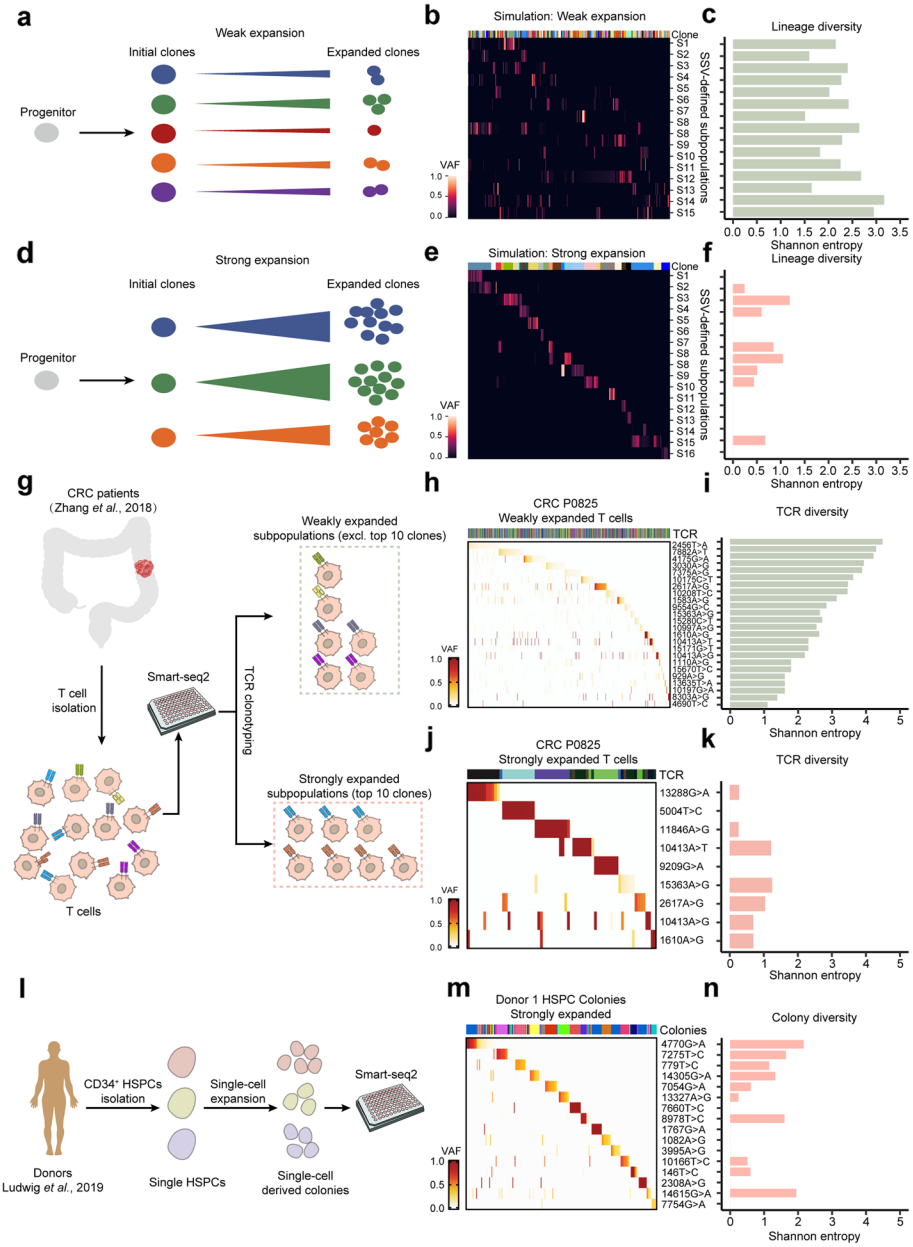

Fig. 3 (See legend on previous page.)

from weakly expanded T cells failed to reconstruct clonal relationships (Fig. 3h) and the TCR diversity of each SSV-defined subpopulation, like weak expansion model (Fig. 3c), was also high (Fig. 3i). In contrast, SSVs in the strongly expanded T cells showed greater potential for marking TCR clones (Fig. 3j). For example, cells carrying mtDNA mutations 5004T > C and 9209G > A aligned perfectly with two different TCR clones (Fig. 3j). Additionally, the TCR diversity of these SSV-defined subpopulations were much lower compared with weakly expanded T cell population (Fig. 3k), displaying a better performance of SSVs in strongly expanded T cells. These findings were confirmed by another independent sample (Additional file 1: Fig. S3c-f).

To further validate these findings, we exploited another dataset where the true clonal relationships were available [8]. $CD34^+$ hematopoietic stem and progenitor cells (HSPCs) were collected from healthy donors and these HSPCs were cultured in parallel to generate single HSPC-derived colonies for downstream single-cell transcriptomic profiling using Smart-seq2 method (Fig. 3l, Additional file 1: Fig. S3g-h). The expansion of each single HSPC into an HSPC-derived colony is a typical clonal expansion process, so the final cell population comprised single HSPC-derived colonies and thus representing a good example for strong clonal expansions. We identified SSVs in these strongly expanded HSPC colonies and found that each SSV-defined subpopulation was in agreement with the prior colony information yielded during the Smart-seq2 library preparation (Fig. 3m, Additional file 1: Fig. S3i). The colony diversity of each SSV-defined subpopulation was also low (Fig. 3n, Additional file 1: Fig. S3j), consistent with our simulations (Fig. 3e-f) and experimental data (Fig. 3j-k). Taken together, these results proved that mitochondrial lineage tracing showed a greater potential under scenarios with strong clonal expansions.

Lineage informative score (LIS) facilitates the application of mitochondrial lineage tracing

SSVs identified in different biological scenarios appeared to show distinct efficacies in marking cell lineages, presenting an urgent need for the identification of genuine informative mtDNA mutations for lineage tracing. Inspired by the findings from simulations, we next sought to introduce a quantitative metric for identifying reliable lineage-informative mtDNA variants in single-cell genomic data. Clearly, if an mtDNA variant is fixed (VAF ~ 100%) in a cell subpopulation (referred as subclonal homoplasmy), this variant stably labels the corresponding cell subpopulation over time, just like a nuclear genomic variant. However, subclonal homoplasmies are rare because most subclonal variants are heteroplasmic within a cell. We therefore defined a metric to quantify the reliability of SSVs as reliable lineage marker,

$$\text{Lineage informative score (LIS)} = \frac{\text{Mean}(VAF)}{1 + \text{Var}(VAF)} \quad (1)$$

where Mean(VAF) and Var(VAF) denote the average and variance of VAFs, respectively, observed in the cells with the detected variant. A higher LIS indicates higher reliability of this mtDNA variant as cell lineage marker. To establish a practical threshold for the LIS, we generated a precision-versus-cutoff curve (Methods). Our analysis revealed that setting the LIS cutoff at approximately 0.6 (0.58) achieved an 80% precision rate in our simulation data (Fig. 4a). Above this cutoff, there was a strong diminishing return for the increase of precision rate.

Notably, the ancestral generation of an SSV-defined subpopulation with a high LIS was significantly larger than that with low LIS (Fig. 4b, Additional file 1: Fig. S2f), suggesting that this cutoff (0.6) indeed could be used for categorizing SSVs. Subsequently, we tested the robustness of the LIS cutoff by generating more simulations with different parameters, including varying mtDNA copy numbers (500, 750, 1000), mtDNA mutation rates per cell division per base pair ($\mu = 10^{-8}, 5 \times 10^{-8}$, or 10^{-7}) and the expansion intensities (τ=0.1, 0.5, and 0.9). The results demonstrated that the current LIS cutoff (0.6) achieved high precision (> 80%) under different settings, showing great robustness

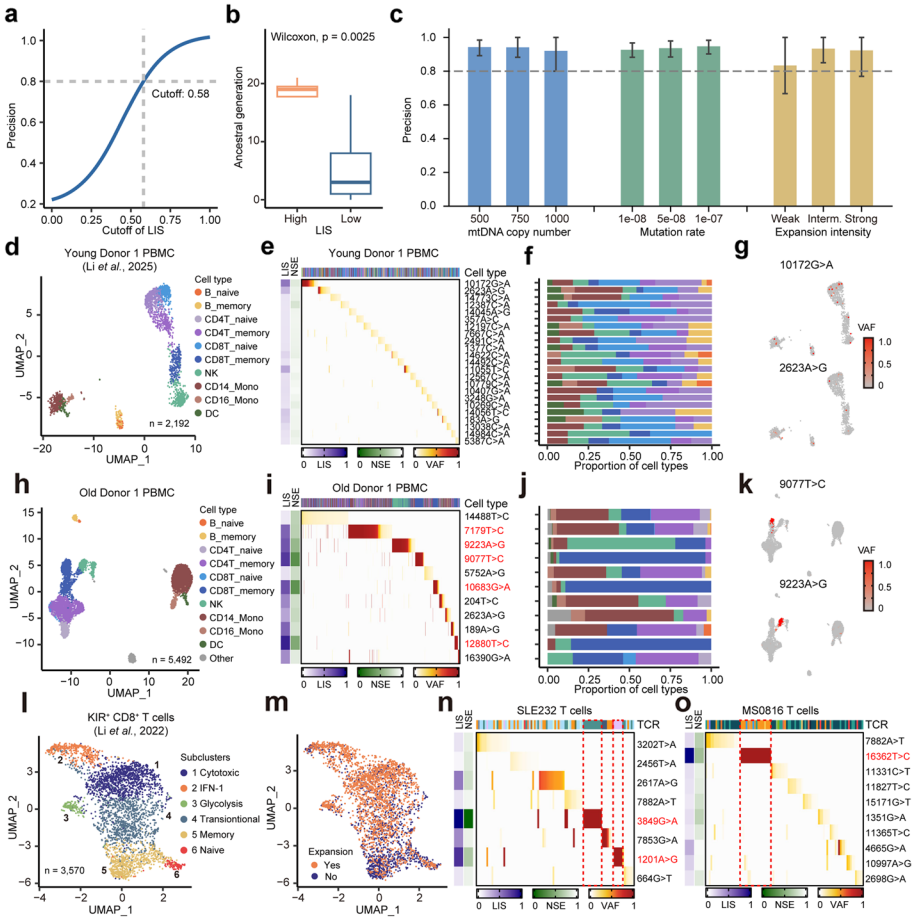

Fig. 4 Lineage informative score (LIS) accurately estimates the lineage tracing capacity of SSVs. **a** Line plot showing the determination of LIS cutoff. **b** Boxplot showing the comparison of the ancestral generation of high-LIS and low-LIS SSV-defined subpopulations. **c** Bar plot showing the robustness analysis of LIS cutoff under scenarios with varying mtDNA copy numbers (left panel), mutation rates (middle panel) and expansion intensities (right panel). **d** UMAP-embedded chromatin accessibility profile of PBMC sample from Young Donor 1. **e** Identification of SSVs in Young Donor 1. LIS and normalized Shannon entropy (NSE) are indicated by colorbars on the left. SSVs with LIS exceeding 0.6 are colored red. **f** Bar plot showing the cell type composition of each SSV-defined subpopulation. **g** The distribution of cells within 10172G > A-defined subpopulation (top panel) and 2623A > G-defined subpopulation (bottom panel) on UMAP-embedded chromatin accessibility profiles. **h** UMAP-embedded chromatin accessibility profile of PBMC sample from Old Donor 1. **i** Identification of SSVs in Old Donor 1. LIS and NSE are indicated by colorbars on the left. SSVs with LIS exceeding 0.6 are colored red. **j** Bar plot showing the cell type composition of each SSV-defined subpopulation. **k** The distribution of cells within 9077T > C-defined subpopulation (top panel) and 9223A > G-defined subpopulation (bottom panel) on UMAP-embedded chromatin accessibility profiles. **l** Single-cell transcriptomic profiles, generated with Smart-seq2, of KIR5$^+$ CD8$^+$ T cells from healthy and autoimmune diseases. **m** Clonal expansion status is color-coded and determined by TCR sequence. **n–o** Identification of SSVs in a systemic lupus erythematosus (SLE) sample—SLE232 (**n**) and multiple sclerosis (MS) sample – MS0816 (**o**). LIS and NSE are indicated by colorbars on the left. SSVs with LIS exceeding 0.6 are colored red

(Fig. 4c, Additional file 1: Fig. S4a-c). Therefore, LIS $\geq$ 0.6 was used to select good SSVs in downstream single-cell genomic data analyses.

We next sought to apply LIS for identifying potential lineage-informative markers in real data. We first analyzed a single-cell multi-omics dataset that we previously generated from peripheral blood mononuclear cells (PBMC) samples of four young donors

(aged 22–24) and obtained transcriptomics, chromatin accessibility and mtDNA mutation profiles for single PBMCs [26]. Since these donors were relatively young, with a low likelihood of clonal hematopoiesis (a typical clonal expansion process during hematopoiesis [27]), this dataset served as an example of weak expansion. The mtDNA mutations were detected at single-cell resolution with ~ 20 × coverage of mitochondrial genome for each cell (Additional file 1: Fig. S5). Following that, we utilized chromatin accessibility information to annotate cell types and different cell types commonly seen in PBMCs could be annotated (Fig. 4d, Additional file 1: Fig. S6). We identified SSVs and calculated LIS for each SSV-defined subpopulation (Fig. 4e). Because the real lineage information was lacking in this dataset, we therefore assayed the cell type similarity within each SSV-defined subpopulation. In these young donor samples, we failed to find substantial SSVs exceeding the LIS cutoff (0.6), consistent with lower clonal expansions in PBMCs. Of note, high diversity of cell types was observed within each SSV-defined subpopulation (Fig. 4e-g, Additional file 1: Fig. S6). Similar results were also seen in all three donors (Additional file 1: Fig. S6), while we also observed a few high-LIS SSVs (e.g. 2 out of 35 SSVs in Young Donor 4), indicating a clonal expansion even at a young age.

We also collected PBMC samples from two elderly donors (one male and one female) over 70 years old because clonal hematopoiesis were more frequently seen in elderly individuals [28]. We performed mtscATC-seq of PBMC samples from these two individuals and the cell types were also annotated using chromatin accessibility of single cells (Fig. 4h). With an average of 20-40X coverage of mitochondrial genome per cell, we identified mtDNA mutations and subsequently performed SSV analysis in both samples (Fig. 4i, Additional file 1: Fig. S7). The average LIS of SSVs identified in old donors were higher compared with young donors (Additional file 1: Fig. S6). We also found that 5 out of 11 mtDNA variants from Old Donor 1 can be selected as good SSVs (LIS $\geq$ 0.6, shown in red). Interestingly, compared with SSVs with low LIS, reliable SSVs showed a biased cell type composition. For example, $CD8^+$ T memory cells made up 89.3% of cell subpopulations defined by 9077T > C and NK cells made up 66.0% of cell subpopulations defined by 9223A > G (Fig. 4j). Furthermore, cells within the 9077T > C or 9223A > G subpopulation exhibited a significantly closer relationship in terms of chromatin accessibility, indicating that those $CD8^+$ T memory cells or NK cells possibly arose from a common progenitor cell by clonal expansions (Fig. 4k). Analysis of another independent sample, Donor 2, showed similar results as Donor 1 (Additional file 1: Fig. S7). Overall, these results, unlike the young PBMC samples (Fig. 4d-g, Additional file 1: Fig. S6), strongly suggested that SSVs identified in the context of strong clonal expansion, particularly those with high LIS, were promising lineage markers, further supporting our simulation results.

Although these results agreed with the simulations, using the similarity of chromatin accessibility alone for testing mitochondrial lineage tracing could lead to false conclusions due to the inconsistency between the similarity of chromatin accessibility and lineage relationship under certain circumstances (e.g. lineage plasticity). Therefore, we took advantage of another Smart-seq2 dataset including T cells isolated from patients with autoimmune diseases (Fig. 4l) [29]. We again leveraged the clonal structure defined by TCR to evaluate the efficacy of mitochondrial lineage tracing in immune responses with significant clonal expansions (Fig. 4m). We first performed mtDNA mutation calling in

sample SLE232, a patient with systemic lupus erythematosus, and MS0816, a patient with multiple sclerosis (Additional file 1: Fig. S8a-d). Identification of SSVs in both samples showed several SSVs with high LIS, which included 3849G > A and 1201A > G in SLE232 and 16362T > C in MS0816 (Fig. 4n). Importantly, we compared the subpopulations defined by SSVs and TCR sequences and found a high concordance between these SSV-defined subpopulations and TCR-defined subpopulations (Fig. 4n-o, Additional file 1: Fig. S8e-h), suggesting that these SSVs with high LIS could be used to identify real clones. It was also worth noting that only these three SSVs (3849G > A, 1201A > G and 16362T > C) showed low TCR diversity as quantified by NSE, again indicating that SSVs with high LIS could be used as faithful lineage markers. Combining simulations and multiple single-cell genomics datasets, we demonstrated that LIS could be used as a reliable metric for identifying genuine informative mtDNA variants for lineage tracing at single cell level.

Discussion

Mitochondrial lineage tracing utilizes naturally occurred mtDNA mutations for tracking in vivo cell lineages and has shown great potential in delineating cellular clonal lineages in native human tissues. In this study, we combined computational simulations and single-cell genomic data to examine the mtDNA mutation dynamics and evaluate the efficacy of mitochondrial lineage tracing under various biological scenarios. Our simulations demonstrated that a significant proportion of mtDNA variants were pre-existing in the initial cell. A recent study examined mtDNA mutations in normal somatic cells of different tissues and concluded that germline mtDNA mutations were not rare and could be transmitted to daughter cells [1], constituting a potential source of mtDNA mutations for lineage tracing. Unlike nuclear germline mutations, these pre-existing heteroplasmic mtDNA mutations still retain the possibility of marking cell lineages, which is a unique feature introduced by the high copy number of mtDNA.

As for the optimal biological settings for mitochondrial lineage tracing, our study revealed that the efficacy of mitochondrial lineage tracing was highly context-dependent. Specifically, the vast majority of mtDNA mutations have limited efficacy to reconstruct lineage history in contexts of no significant clonal expansions as they failed to distinguish clonal lineages. However, they did show promise in tracing cell lineages in scenarios with strong clonal expansion as demonstrated by our simulations, HSPC colonies, human PBMCs and TCR datasets. Therefore, applications of mitochondrial lineage tracing techniques should be primarily applied to scenarios undergoing significant clonal expansions. For example, tumor relapse following treatment (e.g. chemotherapy) is often accompanied by clonal expansion of a subgroup of cells with higher fitness [30, 31], representing a suitable setting for mitochondrial lineage tracing. Indeed, a recent study applied mitochondrial lineage tracing to tumor relapse after chemotherapy treatment in chronic lymphocytic leukemia (CLL) and successfully captured the clonal dynamics during the disease progression [10]. Another recent study by Champan et al. also suggested that the effectiveness of mitochondrial lineage tracing varied significantly under different contexts and biological settings with rapid growth dynamics (e.g. a strong clonal expansion) could be ideal situations for applying mitochondrial lineage tracing [16].

Importantly, our study provided a novel method (LIS) for identifying reliable mtDNA mutations for lineage tracing. Mitochondrial mutations with higher LIS had a better performance on distinguishing clonal relationships, suggesting a good performance of LIS on categorizing informative mtDNA mutations (Fig. 4, Additional file 1: Fig. S6-8). Application of this method to emerging datasets generated by mitochondrial-enriched single-cell sequencing methods, e.g. mtscATAC-seq [11, 13], MAESTER [14] and ReDeeM [9], could lead to novel biological discoveries.

In this study, we mainly used the method originally developed by Miller et al. [14] to identify mtDNA mutations for clonal tracking and downstream analyses. This method has provided a framework for the identification of informative mtDNA mutations, accelerating the application of mitochondrial lineage tracing. However, we also noticed that this method could be limited by setting a threshold of VAF, which may miss variants with lower heteroplasmies. These variants could also provide valuable information for reconstructing clonal relationships at a finer scale, especially under a rapid clonal expansion. Additionally, pre-existing heteroplamic mutations could exist in independent cell lineages, so clonal reconstruction solely relying on single mutations may lead to false conclusions. In light of this, using the combination of multiple mtDNA variants would improve the clonal reconstruction. Therefore, the development of a more comprehensive framework for robustly identifying informative mtDNA variants or their combinations should be warranted.

It is also worth noting that we used Shannon entropy to evaluate the performance of mtDNA variants in real data without ground-truth cell lineage information. However, the effectiveness of this method could be limited in some complex scenarios. For example, a high Shannon entropy of a clone may result from a multipotent cell lineage differentiating into diverse cell types, leading to the false exclusion of genuine lineage markers. This limitation can be overcome by incorporating other lineage tracing markers, such as TCR and epimutations. In fact, a recent study has successfully utilized DNA methylation to accurately trace lineages across diverse contexts, presenting a feasible alternative to mitochondrial lineage tracing for tracing cell lineages in native human tissues [32].

Finally, future work that aims to obtain high sequencing coverage of mitochondrial genome could improve the detectability of low frequency mutations, which may represent informative mtDNA variants or combinations for reconstructing cell lineages. In addition to mtDNA mutations, the simultaneous profiling of multi-modal information [9], including spatial location, transcriptome and epigenome, would bring novel insights into mtDNA mutation dynamics. Along with the development of new computational tools, these new methods may collectively lead to a better understanding of mitochondrial lineage tracing and clonal dynamics in human somatic cells.

Conclusions

Our study, which combined computational modeling and single-cell genomic analysis, systematically interrogated the origin and efficacy of mtDNA variants for single-cell lineage tracing. The results demonstrated that unlike nuclear germline mutations, many mtDNA variants are pre-existing but still retain the capacity for tracking clonal relationships. The efficacy of mitochondrial lineage tracing shows a high dependency on biological contexts, demonstrating better performance in the scenarios of stringent clonal

expansions. Furthermore, lineage informative score (LIS) provides a reliable metric for assessing the lineage tracking efficacy of mtDNA variants. Overall, our study deepened the understanding of somatic dynamics of mtDNA variants and provided practical guidance on applying mitochondrial lineage tracing.

Methods

Computational simulation of mitochondrial lineage tracing

The simulation of mitochondrial lineage tracing data is divided into two parts, simulation of cell division history and simulation of mitochondrial genome replication along with cell divisions.

Simulation of cell division history

As our previously published study [33], we first simulate the cell divisions and population growth using Gillespie algorithm [34]. By conceptualizing cell division as components of continuous-time Markov process, we designate the reaction rate of cell division as $p(t)$ given as follows:

$$p(t) = r\left(1 - \frac{1}{1 + e^{-k(t-t_0)}}\right)$$

Where the three parameters r, k and t_0 jointly determine the change of cell growth rate with time.

We then simulate the cell population growth with the given division rate from 1 initial cell until the population size reached 20,000 cells. 5,000 cells were randomly sampled to obtain their division history (Fig. 1b).

Simulation of mitochondrial DNA replication over cell divisions

To model the mitochondrial lineage tracing within a specified phylogenetic tree, an initial population of mitochondrial DNA was allocated to the progenitor cell. The replicative dynamics of mitochondrial DNA were simulated using the Gillespie stochastic simulation algorithm, with division and death rates at 1 and 0.1, respectively. This simulation was terminated when the mitochondrial genome count reached 500 copies. During the initial phase of mitochondrial genome generation, the mutation rate per division was set to follow a Poisson distribution with an expectation of 0.1 mutations per division, resulting around 100 pre-existing mutations in the initial cell [35] (Fig. 1b, c).

After establishing the initial population of mitochondrial DNA, the model was extended to simulate the accumulation of de novo mutations within the mitochondrial genome as cellular division occurred and lineages diverged. Mitochondrial DNA replication depended on cellular division, with these conditions: for mitochondrial genome doubling index $di = 1$, a single replication event for all mitochondrial DNA; for di within the interval $[1, 2]$, a single replication for all mitochondrial DNA followed by a secondary replication of a random sample of $di - 1$ mitochondrial DNA; and for $di < 1$, a single replication of a randomly selected di mitochondrial DNA. In the simulation of constant model, we set the doubling index to $di = 1$. In the simulation of bottleneck model, we set $di = 0.52$ for the first 10 cell divisions and $di = 1.85$ for the latter (Fig. 1d). When the cell divides, the replicated mitochondrial DNA molecules are segregated to the two daughter

cells, which follows binomial distribution. It has been estimated that per-mitosis mutation rate for mitochondrial genome is 10–100 higher than nuclear genome per site [36]. Assuming the somatic mutation rate of nuclear genome as $\sim 10^{-9}$ per mitosis per site [37], 50-fold increase in mtDNA mutation rate relative to nuclear genome per site and 500 mtDNA copies per cell, each mtDNA copy acquires $10^{-9} \times 16{,}569 \times 50 \approx 0.0008$ mutations per replication and each cell acquires $0.0008 \times 500 \approx 0.4$ mtDNA mutations per replication on average. Hence, we set the number of mitochondrial mutations generated within the cell after each division to follow a Poisson distribution within an expectation of 0.4 mutations per division.

To simulate varying intensities of clonal expansion, we defined a clonal expansion coefficient, τ. During the turnover phase, the probability of each cell retaining one offspring after division is $1 - \tau$, while the probability of retaining either two offspring or none is $0.5\,\tau$. We modeled weak and strong expansion by setting τ to 0.1 and 0.9, respectively (Fig. 1b, Additional file 1: Fig. S1a, Additional file 2: Table S1).

Benchmarking the robustness of the LIS cutoff

To test the robustness of the LIS cutoff, we generated additional simulations with varying parameters. Specifically, we varied the mtDNA copy numbers (500, 750, and 1,000), the mutation rates of mtDNA per cell division ($\mu = 10^{-8}$, 5×10^{-8}, or 10^{-7}), and the expansion intensities ($\tau = 0.1$, 0.5, and 0.9). We then demonstrated the robustness of the LIS cutoff by comparing the precision of SSVs with LIS values greater than the cutoff.

Calculation of clone aggregation score (CAS)

To quantify the accuracy of reconstructing cell lineages from mtDNA mutations, we define the Clone Aggregation Score (CAS). First, we obtain the lineage information of cells from the ground truth tree. For the reconstructed tree $\mathcal{T}$, we define the clone switch value s. As we traverse all leaf nodes of the cells from left to right, s increments by 1 whenever there is a lineage change in the ground truth tree. Clearly, the range of s is between the number of lineages (s_{min}) and the number of cells (s_{max}). Therefore, we normalize s to a range of 0 to 1 and take the logarithm to linearize it, resulting in the CAS:

$$CAS = 1 - \log\left(\frac{s_{min}}{s}\right) / \log\left(\frac{s_{min}}{s_{max}}\right)$$

For the calculation of CAS, we first subset the top 10 SSV-defined subpopulations and downsample the cell population to 1,000 cells. The real cell division history and the SSV-defined clonal structure of the downsampled cells were then used to calculate the CAS.

mtscATAC-seq of PBMCs

Peripheral blood samples from two elderly individuals (Donor 1: male, 73 years old; Donor 2: female, 79 years old) were collected from Guangdong Provincial People's Hospital. The study protocol was approved by the local ethics committee, and written informed consent was obtained from all donors. The blood samples were first diluted 1:1.5 in 0.9% physiological saline and peripheral blood mononuclear cells (PBMCs) were then isolated with a Lymphoprep™ density gradient (density 1.077 g/mL) according to the manufacturer's instructions (STEMCELL).

scATAC-seq libraries were generated using the 10X Chromium Controller and the Chromium Next GEM Single Cell ATAC kit according to the manufacturer's instructions. To retain mtDNA, permeabilization was done using 10 mM Tris–HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP40 and 1% BSA as previously described [13]. PBMCs were then fixed with 1% formaldehyde (28906, Thermo Fisher) in PBS for 10 min at room temperature, followed by quenching with 125 mM glycine solution. Fixed PBMCs were then washed twice with PBS and centrifuged at 400g for 5 min at 4 °C. Subsequently, PBMCs were lysed with lysis buffer (10 mM Tris–HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP40, 1% BSA) for 3 min, followed by adding 1 ml of ice-cold wash buffer (10 mM Tris–HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA). Then lysed cells were centrifuged at 400g for 5 min at 4 °C and the supernatant was discarded. The pellet was resuspended in $1\times$ Diluted Nuclei buffer (10X Genomics) for counting using Trypan Blue. The downstream procedure, including tagmentation, single-cell Gel Bead-In-Emulsions (GEMs) preparation and library amplification, was performed according to the standard protocol provided by 10X Genomics. The final libraries were quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and the DNA fragment was analyzed using Bioanalyzer 2100 system (Agilent).

mtscATAC-seq data preprocessing and cell type annotation

Raw reads were mapped to hg38 reference genome using cellranger-atac (2.1.0) with default settings and the resulting fragment files were subject to ArchR (1.0.2) for downstream analysis. Cells with TSS enrichment score > 5, fragment number $\geq$ 1000, BlacklistRatio < 0.1 and PromoterRatio > 0.1 were kept for downstream analysis. Then doublet score for each cell was calculated and potential doublets were removed. Dimensionality reduction and clustering were carried out using addIterativeLSI, addClusters and addUMAP from ArchR package. After that, markers of each cluster were identified with cutoff FDR < 0.01 and log2FC $\geq$ 1 and these markers were later used for curation of cell types. To better annotate the cell type, a publicly available single-cell RNA-seq dataset of PBMC [38] was integrated with the scATAC-seq data. Based on the integration results, we then manually curated the cell type annotation using marker gene lists from ScType database [39] and relevant literature [19].

Single-cell multi-omics data analysis

For the analysis of the mitochondrial genome-enriched single-cell multi-omics data [26] (Young PBMC samples from healthy individuals aged 22–24), raw reads were mapped to reference genome hg38 using the cellranger-arc (2.0.2). Notably, this library mixed PBMCs from four young donors and homoplasmic germline mtDNA mutations were used to demultiplex the PBMCs. The chromatin accessibility was used to annotate cell types and call mtDNA mutations as previously described in the mtscATAC-seq session.

Smart-seq2 data preprocessing

Raw reads of Smart-seq2 datasets were first aligned to hg38 reference genome using STAR (2.7.10a) with default parameters. Bam files were then sorted using samtools (1.15.1) and duplicates were removed using MarkDuplicates from GATK toolkit

(V4.2.0.0). Resulting bam files were then used for downstream mitochondrial mutation calling.

TCR assembly

TRUST4 [40] tool was employed to assemble the TCR sequence of single T cells isolated from CRC patients. Briefly, bam files of aligned Smart-seq2 data were used as input files and candidate reads were then extracted for the assembly of TCR sequences. The assembled TCR sequences were then annotated and assessed. Default settings of TRUST4 were used. The TCR sequences of T cells collected from autoimmune diseases (MS0816 and SLE232) were provided by the published study [29].

Identification of mtDNA mutations

Mitochondrial DNA mutation calling was performed as previously described [19] with minor modifications. For mtscATAC-seq data, samtools (1.15.1) was used to subset chrM-specific alignments from the original bam file generated by cellranger-atac. The resulting bam was then subject to IndelRealigner to correct potential mapping errors around indels. After realignment, mutations were called using varscan (2.4.4) with –min-var-freq 0.01 –min-reads2 2 to identify germline mutations (bulk VAF cutoff 90%). For identification of mitochondrial DNA mutations in single cells, the realigned bam file was first split into separate bam files. For each bam file, duplicates were removed, and mutations were called using varscan with the same parameter for bulk sample calling. Apart from the germline mutations identified previously, mutations that are present in more than 90% of cells with VAF > 5% were also considered as germline mutations. To identify high-confidence mutations, germline mutations were first removed. Second, mutations within the range of 302–316, 514–524 and 3106–3110 were also removed due to a large number of homopolymers, which potentially cause misalignment. Third, mutations must meet the following criteria: 1) variant counts ≥ 4; 2) sequencing depth of the site ≥ 20; 3) VAF $\geq 5\%$; 4) ratio of reads mapped to forward and reverse strand must be in the range of 0.3–0.7. After the identification of high-confidence variants in each individual cell, all high-confidence variants were merged. Following that, a mutation recalling procedure was done in the cell population. More specifically, for each high-confidence variant site, VAF was calculated for all cells. At last, cells with a mean sequencing depth < 10 were discarded.

For Smart-seq2 datasets, several extra steps were taken to ensure the compatibility with the previous framework. Read groups (RG) of LB and SM were added to bam files using samtools addreplacerg. Moreover, SplitNCigarReads from GATK was used to split reads that contain Ns in their cigar string, which is caused by spanning splicing events of RNA-seq data.

Evaluation of clone-informative variants

To assess the potential of mitochondrial lineage tracing, we identified cell subpopulation-specific variants (SSVs) from simulated data and real data to examine their efficacy in reconstructing cell lineages. SSVs were identified using the method described by Miller et al. [14]. We first calculated the presence of each variant in the cell population under different thresholds of VAF, ranging from 0 to 50%. We then tested various

thresholds of minimal VAF and minimal cell population size to select SSVs. More specifically, for each variant, only cells with VAF exceeding the minimal VAF cutoff are counted and if the number of cells carrying this variant exceeds the minimal cell population size, the variant is considered as an SSV. SSVs identified under a certain combination of thresholds were then sorted by their presence in the cell population and cells were subsequently assigned to SSVs, constituting SSV-defined subpopulations. Small SSV-defined subpopulations (<5 cells in simulation data) were excluded from downstream analysis. Cells within each SSV-defined subpopulation were also sorted by their VAF. Pearson correlation of SSVs was performed and hierarchical clustering was conducted to calculate the distance amongst SSVs. We chose 0.8 as the cutoff for height (output of the previous hierarchical clustering) and SSVs had a distance lower than 0.8 were grouped. If multiple SSVs were clustered together (co-existed in the same subgroup of cells), the SSV with the highest mean VAF was kept. It was also worth noting that mitochondrial mutational profiles of each dataset could be greatly different from each other. For example, some cell populations had a great number of mutations with high VAF (e.g. aged PBMC samples) while others only had a few such mutations, so that the threshold of VAF and cell population size for SSV identification should be customized. Therefore, in our study the parameter for identifying SSVs in each dataset was adjusted accordingly and the final size of the cell population with SSVs should be at least 10% of the examined cell population.

To better assess the lineage tracing potential of SSVs, we examined the statistical properties of SSVs within both real and simulated datasets and observed that SSVs identified high-confidence clones exhibit a higher VAF. Moreover, the VAF distribution across all cells within a given SSV-defined subpopulation demonstrated relative uniformity. Consequently, we proposed the following metric for SSV assessment:

$$LIS = \frac{\mathrm{Mean}(VAF)}{1 + \mathrm{Var}(VAF)}$$

SSVs with elevated LIS are typically indicative of clonal identification capability. To establish a cutoff LI score for clone delineation, we scrutinized simulated datasets derived from two mitochondrial number models, constant and bottleneck. An SSV is deemed efficacious if the common ancestor of at least 60% of the cells with the closest lineage relationship within an SSV-defined subpopulation is later than 10^{th} generation. The precision of an SSV is quantified as the ratio of efficacious SSVs to the aggregate count of SSVs exceeding the cutoff score.

Analysis of the simulation data facilitated the construction of a precision-versus-cutoff curve. An increase in the cutoff correlates with a gradual augmentation in precision, albeit accompanied by a decrease in the total number of retained SSVs. To optimize the selection of LI score cutoff, we fitted a sigmoid function to the precision variation curve against the cutoff,

$$Precision = \frac{0.84}{1 + \mathrm{e}^{-7.33(cutoff - 0.45)}} + 0.19$$

followed by computing the cutoff value that yields a precision of 0.8, which was found to be around 0.6 (0.58). Above this cutoff, there was a strong diminishing return of the

increase of precision rate. Therefore, we chose 0.6 as the optimal cutoff and applied it to downstream single-cell genomic data analyses.

Lineage composition analysis

To better examine the efficacy of SSVs in defining cell lineages in simulation, we developed a method to quantitatively measure the lineage composition of SSV-defined subpopulations. In this method, we first defined the eight cells at the third generation from a common ancestral cell as eight ancestral lineages (termed Lineage 1–8). Then, for each cell within an SSV-defined subpopulation, we traced their ancestor to the third generation (i.e. Lineage 1–8) and calculated the sum of third generation ancestral lineages for each SSV-defined subpopulation. This number serves as an indicator of the performance of SSVs, with values closer to one indicating better performance. For example, if one SSV-defined subpopulation contains ancestors from multiple lineages, the cells within it are not eligible to be considered as one genuine clone, suggesting bad performance of this SSV.

Shannon entropy (SE) was calculated for assessing the diversity of SSV-defined subpopulations. Assuming an SSV-defined subpopulation $\mathcal{S}$ is composed by real clone or TCR/colonies clone $(c_1, c_2, \ldots, c_n)$ with proportion $(p_1, p_2, \ldots, p_n)$. The Shannon entropy was defined as follows

$$SE = -\sum_{i=1}^{n} p_i \log p_i$$

For the ease of visualization and direct comparison across different datasets, normalized Shannon entropy (NSE) was calculated by dividing Shannon entropy with maximum possible entropy, yielding a score from 0 to 1

$$NSE = -\frac{\sum_{i=1}^{n} p_i \log p_i}{\sum_{i=1}^{n} \log p_i}$$

A higher NSE indicates a higher diversity of a cell subpopulation.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03540-7.

Additional file 1: Fig. S1 - S8 and figure legends. Additional results of simulations and single-cell genomics datasets.

Additional file 2: Table S1. Description of models and related parameters.

Acknowledgements

We thank Zhenglong Gu, Weiwei Zhai and Hu laboratory members for constructive discussions. We thank Lei Zhang and Zemin Zhang for providing access to the Smart-seq2 dataset of the T cells isolated from CRC patients.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

Z.H., X.W. and K.W. conceived and designed the study. X.W. analyzed the single-cell genomic data and simulated data. K.W. performed simulation studies. W.Z. performed the mtscATAC-seq experiments. Z.T. and D.Z. provided guidance on data analysis and modeling. C.Z. and W.Z.Z. collected samples. Z.H., X.W, J.X. and Q.M. interpreted the results. X.W., Z.H. and K.W. wrote the manuscript with contributions from all co-authors. Z.H, J.X. and Q.M. supervised the project. All authors read and approved the final manuscript.

Funding
This work was supported by National Natural Sciences Foundation of China (82241236 and 32270693 to Z.H., 32070644 and 32293190 to J.X., 32070870 to Q.M., 32300493 to X.W.), Guangdong Basic and Applied Basic Research Foundation (2021B1515020042 to Z.H.) and China Postdoctoral Science Foundation (2022M723301 to X.W.).

Data availability
The mtscATAC-seq data of two PBMC samples have been deposited in National Genomics Data Center [41], China National Center for Bioinformation (GSA-Human: HRA007291 [42]) that are available at https://ngdc.cncb.ac.cn/gsahuman. The single-cell multi-omics data of four PBMC samples from young donors is available at National Genomics Data Center, China National Center for Bioinformation (GSA-Human: HRA004605 [43]). Single-cell RNA-seq (Smart-seq2) dataset of the T cells isolated from CRC patients has been deposited at EGA under accession EGAS00001002791 [44]. Single-cell RNA-seq (Smart-seq2) dataset of T cells of autoimmune diseases is publicly available at GEO under accession GSE193439 [45]. All custom code used to reproduce the analysis is available under MIT license at GitHub (https://github.com/BoxWong/MT_lineage_tracing [46]) and has been archived on Zenodo (https://zenodo.org/records/14955230 [47]). The mtDNA mutation information and related metadata generated in our study have been made available at Zenodo (https://zenodo.org/records/14955230 [47]).

Declarations

Ethics approval and consent to participate
This study was approved by the local ethics committee of Guangdong Provincial People's Hospital (approval number: KY2023-537), and written informed consent for participation and publication was obtained from all donors. All experimental methods complied with the Declaration of Helsinki.

Competing interests
The authors declare no competing interests.

Received: 16 May 2024 Accepted: 11 March 2025
Published online: 26 March 2025

References

1. An J, Nam CH, Kim R, Lee Y, Won H, Park S, et al. Mitochondrial DNA mosaicism in normal human somatic cells. Nat Genet. 2024;56(8):1665–77.
2. McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J. Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science. 2016;353(6298):aaf7907.
3. Raj B, Wagner DE, McKenna A, Pandey S, Klein AM, Shendure J, et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nat Biotechnol. 2018;36(5):442–50.
4. Li L, Bowling S, McGeary SE, Yu Q, Lemke B, Alcedo K, et al. A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells. Cell. 2023;186(23):5183-5199.e22.
5. Liu K, Deng S, Ye C, Yao Z, Wang J, Gong H, et al. Mapping single-cell-resolution cell phylogeny reveals cell population dynamics during organ development. Nat Methods. 2021;18(12):1506–14.
6. Yang D, Jones MG, Naranjo S, Rideout WM 3rd, Min KH (Joseph) J, Ho R, et al. Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. Cell. 2022;185(11):1–19.
7. Lu Z, Mo S, Xie D, Zhai X, Deng S, Zhou K, et al. Polyclonal-to-monoclonal transition in colorectal precancerous evolution. Nature. 2024;636(8041):233–40.
8. Ludwig LS, Lareau CA, Ulirsch JC, Christian E, Muus C, Li LH, et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. Cell. 2019;176(6):1325-1339.e22.
9. Weng C, Yu F, Yang D, Poeschla M, Liggett LA, Jones MG, et al. Deciphering cell states and genealogies of human haematopoiesis. Nature. 2024;627(8003):389–98.
10. Penter L, Gohil SH, Lareau C, Ludwig LS, Parry EM, Huang T, et al. Longitudinal single-cell dynamics of chromatin accessibility and mitochondrial mutations in chronic lymphocytic leukemia mirror disease history. Cancer Discov. 2021;11(12):3048–63.
11. Xu J, Nuno K, Litzenburger UM, Qi Y, Corces MR, Majeti R, et al. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. Elife. 2019;8:e45105.
12. Lareau CA, Ludwig LS, Sankaran VG. Longitudinal assessment of clonal mosaicism in human hematopoiesis via mitochondrial mutation tracking. Blood Adv. 2019;3(24):4161–5.
13. Lareau CA, Ludwig LS, Muus C, Gohil SH, Zhao T, Chiang Z, et al. Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. Nat Biotechnol. 2021;39(4):451–61.
14. Miller TE, Lareau CA, Verga JA, DePasquale EAK, Liu V, Ssozi D, et al. Mitochondrial variant enrichment from high-throughput single-cell RNA sequencing resolves clonal populations. Nat Biotechnol. 2022;24:1–5.
15. Penter L, Gohil SH, Wu CJ. Natural barcodes for longitudinal single cell tracking of leukemic and immune cell dynamics. Front Immunol. 2022;12(January):1–11.
16. Chapman MS, Przybilla MJ., Lawson AR., Mitchell E, Dawson K, Williams N, et al. Mitochondrial mutation, drift and selection during human development and ageing. Res Sq. 2023;
17. Zhang H, Burr SP, Chinnery PF. The mitochondrial DNA genetic bottleneck: inheritance and beyond. Essays Biochem. 2018;62(3):225–34.
18. Cree LM, Samuels DC, De Sousa Lopes SC, Rajasimha HK, Wonnapinij P, Mann JR, et al. A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. Nat Genet. 2008;40(2):249–54.

19. Tang Z, Lu Z, Chen B, Zhang W, Chang HY, Hu Z, et al. A genetic bottleneck of mitochondrial DNA during human lymphocyte development. Mol Biol Evol. 2022;39(5):msac090.
20. Guo X, Xu W, Zhang W, Pan C, Thalacker-Mercer AE, Zheng H, et al. High-frequency and functional mitochondrial DNA mutations at the single-cell level. Proc Natl Acad Sci U S A. 2023;120(1):e2201518120.
21. Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, et al. A high observed substitution rate in the human mitochondrial DNA control region. Nat Genet. 1997;15(4):363–8.
22. O'Hara R, Tedone E, Ludlow A, Huang E, Arosio B, Mari D, et al. Quantitative mitochondrial DNA copy number determination using droplet digital PCR with single-cell resolution. Genome Res. 2019;29(11):1878–88.
23. Wonnapinij P, Chinnery PF, Samuels DC. The distribution of mitochondrial DNA heteroplasmy due to random genetic drift. Am J Hum Genet. 2008;83(5):582–93.
24. Kwok AWC, Qiao C, Huang R, Sham MH, Ho JWK, Huang Y. MQuad enables clonal substructure discovery using single cell mitochondrial variants. Nat Commun. 2022;13(1):1205.
25. Zhang L, Yu X, Zheng L, Zhang Y, Li Y, Fang Q, et al. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. Nature. 2018;564(7735):268–72.
26. Li X, Xu Y, Zhang W, Chen Z, Peng D, Ren W, et al. Immunoregulatory programs in anti-N-methyl-D-aspartate receptor encephalitis identified by single-cell multi-omics analysis. Clin Transl Med. 2025;15(1):e70173.
27. Bowman RL, Busque L, Levine RL. Clonal hematopoiesis and evolution to hematopoietic malignancies. Cell Stem Cell. 2018;22(2):157.
28. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. Science. 2019;366(6465):eaan4673.
29. Li J, Zaslavsky M, Su Y, Guo J, Sikora MJ, van Unen V, et al. KIR+CD8+ T cells suppress pathogenic T cells and ar active in autoimmune diseases and COVID-19. Science. 2022;376(6590):eabi9591.
30. Walens A, Lin J, Damrauer JS, McKinney B, Lupo R, Newcomb R, et al. Adaptation and selection shape clonal evolution of tumors during residual disease and recurrence. Nat Commun. 2020;11(1):1–15.
31. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukemia revealed by whole genome sequencing. Nature. 2012;481(7382):506.
32. Chen M, Fu R, Chen Y, Li L, Wang SW. High-resolution, noninvasive single-cell lineage tracing in mice and humans based on DNA methylation epimutations. Nat Methods. 2025;16:1–11.
33. Wang K, Hou L, Wang X, Zhai X, Lu Z, Zi Z, et al. PhyloVelo enhances transcriptomic velocity field mapping using monotonically expressed genes. Nat Biotechnol. 2023;31:1–12.
34. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977;81(25):2340–61.
35. Bi C, Wang L, Fan Y, Yuan B, Alsolami S, Zhang Y, et al. Quantitative haplotype-resolved analysis of mitochondrial DNA heteroplasmy in Human single oocytes, blastoids, and pluripotent stem cells. Nucleic Acids Res. 2023;51(8):3793–805.
36. Wei W, Tuna S, Keogh MJ, Smith KGCCKR, Aitman TJ, Beales PL, et al. Germline selection shapes human mitochondrial DNA diversity. Science. 2019 May 24;364(6442).
37. Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. Differences between germline and somatic mutation rates in humans and mice. Nat Commun. 2017;8(1):1–8.
38. Granja JM, Klemm S, McGinnis LM, Kathiria AS, Mezger A, Corces MR, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat Biotechnol. 2019;37(12):1458–65.
39. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun. 2022;13(1):1246.
40. Song L, Cohen D, Ouyang Z, Cao Y, Hu X, Liu XS. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. Nat Methods. 2021;18(6):627–30.
41. Bao Y, Zhang Z, Zhao W, Xiao J, He S, Zhang G, et al. Database resources of the national genomics data center, China national center for bioinformation in 2024. Nucleic Acids Res. 2024;52(D1):D18-32.
42. Wang X, Wang K, Zhang W, Tang Z, Zhang H, Cheng Y, et al. Clonal expansion dictates the efficacy of mitochondrial lineage tracing in single cells. Datasets. Genome Sequence Archive. 2024. https://ngdc.cncb.ac.cn/gsa-human/browse/HRA007291.
43. Li X, Xu Y, Zhang W, Chen Z, Peng D, Ren W, et al. Immunoregulatory programs in anti-N-methyl-D-aspartate receptor encephalitis identified by single-cell multi-omics analysis. Datasets. Genome Sequence Archive. 2024. https://ngdc.cncb.ac.cn/gsa-human/browse/HRA004605.
44. Zhang L, Yu X, Zheng L, Zhang Y, Li Y, Fang Q, et al. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. Datasets. European Genome-phenome Archive. 2018. https://ega-archive.org/datasets/EGAD00001003910.
45. Li J, Zaslavsky M, Su Y, Guo J, Sikora MJ, van Unen V, et al. KIR+CD8+ T cells suppress pathogenic T cells and are active in autoimmune diseases and COVID-19. Datasets. Gene Expression Omnibus. 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE193439.
46. Wang X, Wang K. Clonal expansion dictates the efficacy of mitochondrial lineage tracing in single cells. GitHub. 2024. https://github.com/BoxWong/MT_lineage_tracing.
47. Wang X, Wang K, Zhang W, Tang Z, Zhang H, Cheng Y, et al. Clonal expansion dictates the efficacy of mitochondrial lineage tracing in single cells. Zenodo. 2025. https://zenodo.org/records/14955230.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.