A genetic bottleneck of mitochondrial DNA during human lymphocyte development

Zhongjie Tang[1, #], Zhaolian Lu[2, #], Baizhen Chen[1], Weixing Zhang[1], Howard Y. Chang[3,4,5, *], Zheng Hu[2, *], Jin Xu[1, *]

[1] State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, China.

[2] CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.

[3] Center for Personal Dynamic Regulomes, Stanford University, United States.

[4] Departments of Dermatology and Genetics, Stanford University School of Medicine, United States.

[5] Howard Hughes Medical Institute, Stanford University, United States.

[*] Correspondence to H.Y.C. at: howchang@stanford.edu, Z.H. at: zheng.hu@siat.ac.cn and J.X. at: xujin7@mail.sysu.edu.cn

[#] These authors contributed equally to the work.

ABSTRACT

Mitochondria are essential organelles in eukaryotic cells that provide critical support for energetic and metabolic homeostasis. Mutations that accumulate in mitochondrial DNA (mtDNA) in somatic cells have been implicated in cancer, degenerative diseases, and the aging process. However, the mechanisms used by somatic cells to maintain proper functions despite their mtDNA mutation load are poorly understood. Here, we analyzed somatic mtDNA mutations in more than 30,000 human single peripheral and bone marrow mononuclear cells and observed a significant overrepresentation of homoplastic mtDNA mutations in B, T and NK lymphocytes despite their lower mutational burden than other hematopoietic cells. The characteristic mutational landscape of mtDNA in lymphocytes were validated with data from multiple platforms and individuals. Single-cell RNA-seq and computational modeling demonstrated a stringent mitochondrial bottleneck during lymphocyte development likely caused by lagging mtDNA replication relative to cell proliferation. These results illuminate a potential mechanism used by highly metabolically active immune cells for quality control of their mitochondrial genomes.

INTRODUCTION

Mitochondrial DNA (mtDNA) encodes genes involved in oxidative phosphorylation that are essential for eukaryotic cells[1]. There are typically hundreds to thousands of copies of mtDNA molecules in each cell and the germline mtDNA is predominantly maternally inherited and does not undergo recombination[2]. mtDNA accumulates mutations at a rate that is five to ten times higher per site than the nuclear genome because the lack of DNA repair systems[3,4] and frequent contact with mutagenic reactive oxygen species (ROS)[5]. More than 500 pathogenic mtDNA mutations have been identified as causative genetic defects of various human diseases[6]. According to the theory known as "Muller's ratchet," continuous accumulation of deleterious mutations in the absence of purifying selection will lead to a decline in population fitness and will ultimately result in mutational meltdown[7]. To avoid this outcome, the animal germline has evolved a mitochondrial genetic bottleneck, wherein only a small subset of mtDNA is transmitted to the next generation, thus resulting in significant removal of deleterious mutations[8-10]. Population studies have also revealed an increase in mtDNA heteroplasmy in blood cells as part of the normal aging process[11] and the accumulation of pathogenic mtDNA mutations has been reported in cancers and neurodegenerative disorders[12,13]. However, the transmission and clonal dynamics of somatic mtDNA mutations along tissue development are largely unknown, due to the technical difficulties of detecting heteroplasmic mutations in single cells.

We and others have recently developed a single-cell lineage tracing method leveraging the somatic mtDNA mutations detected in single-cell assay for transposase-accessible chromatin with high-throughput sequencing (scATAC-seq) and/or RNA-seq (scRNA-seq) data[14,15]. Using this method, a recent study had shown that a pathogenic mutation 3243A/G, the cause of mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episode (MELAS)[16], was remarkably purified in T cells as compared to other blood cells from peripheral blood mononuclear cells (PBMCs), with unknown mechanisms. These results inspired our investigation of the mtDNA mutational landscape in a large population of single cells in order to understand the clonal dynamics of mtDNA in the development of somatic cell lineages.

RESULTS

Somatic mutational landscape of mtDNA at single-cell resolution

In this study, we focused on human hematopoietic system where the cellular differentiation lineages have been well documented. We first identified somatic mtDNA mutations in a previously reported mitochondrial scATAC-seq (or mtscATAC-seq) dataset including more than 20,000 blood cells from a healthy 47-year-old individual[17] (**Fig. 1a-b, Extended Data Fig. 1a**, **Methods**). We summarized the numbers of mutations and the variant allele frequency (VAF, also referred to as mtDNA heteroplasmic ratio) in each cell in order to compare the VAF distribution in a population of different cell types. Interestingly, we found cells of the mature lymphocyte lineages--specifically B, T, NK cells--carried a significantly lower mtDNA mutational burden as compared to those identified in hematopoietic progenitor cells, including hematopoietic stem cells (HSCs), multipotent progenitors (MPPs), lymphoid-primed multipotent progenitors (LMPPs), common lymphoid progenitors (CLPs), common myeloid progenitors (CMPs), and granulocyte-macrophage progenitors (GMPs) (**Fig. 1c and Extended Data Fig. 1b**, Wilcoxon test, $p < 2.2e^{-16}$). The mtDNA mutational burden was also lower in lymphocytes as compared to the myeloid and erythroid lineages (**Fig. 1c**, Wilcoxon test, $p < 2.2e^{-16}$). As anticipated, most somatic mtDNA mutations were detected at low VAF in individual cells in all cell types (**Fig. 1d**). However, the distribution of homoplastic mutations (i.e., those at VAF~1) varies substantially among the different cell types. For instance, progenitor cells, including HSCs, MPPs, LMPPs, CLPs, CMPs, and GMPs, exhibit the typical monotonic decline in the number of mutations with increasing VAF (**Fig. 1d**). While this pattern was also true in both the myeloid and erythroid lineages (e.g., monocytes and erythrocytes), we observed an unanticipated increase in the number of homoplastic mutations in B, T and NK cells (**Fig. 1d**).

In addition to the mtscATAC-seq dataset from PBMCs, we analyzed another mtscATACseq dataset of 10,327 bone marrow mononuclear cells (BMMCs) from an independent healthy donor[18] (**Fig. 2a**). As the observations in PBMCs, lymphocytes in BMMCs also carried a lower mtDNA mutational burden with a characteristic overrepresentation of homoplastic mutations (**Fig. 2b and Extended Data Fig. 1c**). In fact, these lymphocyte-specific characteristics were also verified by additional scATAC-

seq or scRNA-seq data from 7 independent individuals (**Extended Data Fig. 2**), indicating a general and unique process of clonal dynamics of mtDNA in lymphocyte development.

Asynchronous replication of mitochondrial and nuclear genome during B cell development

To examine whether the distinct VAF distribution between lymphoid cells and myeloid/erythroid cells is due to the variation of mtDNA copy number per cell, we estimated the relative number of mtDNA copies in each cell type according to the fraction of sequencing reads mapped to the mitochondrial genome relative to the total number of reads in each cell (**Fig. 3a and Extended Data Fig. 3a**). Although mature lymphocytes and progenitor cells had similar mtDNA copy numbers**,** pro-B and pre-B cells—the earliest lineage-committed cells in B cell development—exhibited a significantly lower number of mtDNA copies (Wilcoxon test, pro-B/pre-B versus HSC/MPP, $p < 2.2e^{-16}$; pro-B/pre-B versus B, $p < 2.2e^{-16}$). Of note, the CLPs also showed significantly fewer mtDNA copies than earlier progenitors (Wilcoxon test, CLP versus HSC/MPP, $p < 2.2e^{-16}$; CLP versus LMPP, $p < 2.2e^{-16}$), thus indicating a remarkable mtDNA copy number reduction in early lymphocyte development. Therefore, we hypothesized that the characteristic mutational spectra in lymphocyte mtDNA (**Fig. 1c-d and Fig. 2b**) might result from a mitochondrial genetic bottleneck.

To address this possibility, we examined the mtDNA replication machinery to gain insight into the regulation of mtDNA copy number along the lymphocyte differentiation trajectory. Since T cells matures in thymus and their progenitor, pre-T cells, are not available in by PBMCs, we focused on the B cell lineage. DNA polymerase γ is the only known mitochondrial DNA polymerase in animals[19]. DNA polymerase γ has both a catalytic (*POLG*) and a binding subunit (*POLG2*) and can catalyze the polymerization of deoxyribonucleotides. High levels of DNA polymerase γ activity have been detected in cell cycle phases S and G2 to maintain stable numbers of mtDNA during cell division[19-21]. To determine whether the expression of DNA polymerase γ increases with cell proliferation during B cell development, we projected the developmental trajectory of cell subpopulations from HSCs to mature B cells via a pseudo-time analysis with scRNA-seq data (**Fig. 3b and Extended Data Fig. 3b**). We observed up-regulation of

133 G1/S phase-specific genes (such as DNA polymerase δ, *POLD1–3*) in both pro-B and
134 pre-B cell populations, thus suggesting high activation of cell proliferation in these cell
135 types (**Fig. 3c-d and Extended Data Fig. 3c-e**). In contrast, the expression of DNA
136 polymerase γ was not coupled with cell proliferation (**Fig. 3e-f and Extended Data Fig.**
137 **3c-e**). Unexpectedly, the expression of the DNA polymerase γ binding subunit (*POLG2*)
138 was significantly diminished in the highly proliferative pro-B and pre-B cell
139 subpopulations (**Fig. 3g**). Together, these results imply a genetic bottleneck during B
140 cell development which might have resulted from limited replication of mtDNA, thus
141 diluting the mtDNA copy number throughout cell division.

142

143 **Quantification of mtDNA genetic bottleneck by computational modeling**

144 To test our hypothesis and quantify the extent of the mitochondrial genetic bottleneck,
145 we developed a computational model of an mtDNA dilution process based on
146 population genetics theory (**Fig. 4a**). In this model, we assumed that only a proportion
147 of mtDNA molecules (denoted by α) replicates during each cell cycle. This process
148 continues for T_d cell cycles until the mtDNA copy number recovers to the initial levels
149 (~500 copies per cell estimated by Ryan et.al[22]). Using the approximate Bayesian
150 computation (ABC) method, we estimated the model parameters for B, T and NK cell
151 populations by using a constant mtDNA mutation rate of 10^{-7} per site per cell division[23]
152 (**Fig. 4b and Extended Data Fig. 4a**). The model estimations showed the minimal
153 mtDNA copy number were 21 (95% confidence interval [CI] =13–56), 13 (95% CI=12–
154 19) and 14 (95% CI=12–21) in each B, T and NK cell, respectively. These values were
155 20–40-fold lower than the normal mtDNA levels. The VAF distribution simulated with
156 these parameter estimations recapitulated the observed data, showing a characteristic
157 overrepresentation of homoplastic mutations (i.e., VAF~1) and a reduced overall
158 mutational burden (**Fig. 4c and Extended Data Fig. 4b-c**). Notably, this pattern cannot
159 be achieved by random genetic drift alone with a constant number of mtDNA copies.

160

161 **The consequence of mtDNA genetic bottleneck**

162 Collectively, our integrative genomic data analysis and computational modeling
163 demonstrated the existence of a stringent mtDNA genetic bottleneck that resulted from
164 replicative dilution during lymphocyte development. This mechanism strengthens the
165 genetic drift toward a lower mtDNA mutational burden and lower genetic diversity within

each cell. We wondered whether the genetic bottleneck during lymphocyte development might have the same purifying selection effects as those in the germline. We thus examined the VAF distribution in various genomic regions (loop, tRNA, rRNA and coding) or mutation types (synonymous and nonsynonymous), as well as the dN/dS ratio (the ratio of the normalized number of nonsynonymous substitutions - *dN* to the normalized number of synonymous substitutions - *dS*) (**Extended Data Fig. 5a** and **Extended Data Fig. 6a**). We observed no significant differences in the VAF distribution for mutations in different genomic regions or substitution types among the various cell types. Moreover, the calculated *dN*/*dS* ratios revealed a pattern of generally neutral evolution (i.e., *dN*/*dS*~1) in all categories in most of the cases examined (**Extended Data Fig. 5b** and **Extended Data Fig. 6b**).

Thus, our results showed that the entire mtDNA genome was evolving under a neutrality-like process. However, this is likely due to linkage of whole mitochondrial genome with strong Hill–Robertson interference, leading to a pattern of quasi-neutrality as in cancer evolution[24]. Therefore, we checked individual mutation sites to look for the signals of purifying selection and indeed observed several mutations that were specifically eliminated in lymphocytes compared to myeloid lineage (**Fig. 5a-b**). For example, the mutations, 2636G/A and 3209A/G, underwent the most profound decrease in prevalence (**Fig. 5c**) in lymphocytes. Intriguingly, these two sites are all located at MT-RNR2, which encode 16S rRNA and Humanin, a peptide playing protection roles in multiple mitochondrial diseases (**Fig. 5d**)[25]. Furthermore, we queried MITOMAP, a human mitochondrial genome database, and found that mtDNA variants reported on MT-RNR2 were highly associated with sepsis ($p<2.2e^{-16}$, **Fig. 5e**)[26,27], suggesting MT-RNR2 may play important roles in immune functions to protect from infections. These data indicate purifying selection in lymphocytes indeed occurs for specific mtDNA mutation sites.

DISCUSSION

Collectively, we observed an unanticipated lower mutational burden and accumulation of homoplastic mtDNA mutations in lymphocytes that depicted a stringent genetic bottleneck and purifying selection of mtDNA. Gene expression data and computational modeling suggest a dilution process, based on the rate of mtDNA replication relative

to the nuclear genome. Although the single-cell data derived from PBMCs cannot capture the full developmental trajectory of T cells because pre-T cells develop in the thymus. Our single cell data and computational inference indicates the genetic bottleneck in T and NK cells might be as stringent as that in B cells (**Fig. 1d, Fig. 2a, Fig. 4b**). Further systematic study of T cell precursors in the thymus may provide further insight on how genetic bottleneck occurs during T cell development. Also, based on our observations and simulations, we hypothesize that the regulation of lymphocyte specific genetic bottleneck may start from CLP stage, instead of subsequent lineage commitment for B, T and NK cells. The effect of this regulation was likely enhanced via the active proliferation of progenitor cells. We knew that during lymphocyte development, multipotent T and B progenitor cells undergo a series of maturation steps that include positive selection for functional T-cell receptors (TCRs) or immunoglobulins and negative selection to eliminate cells with a high affinity for self-associated peptides or antigens[28]. Only a small proportion of T lymphoid cells will survive from the negative and positive selections. Moreover, mitochondrial function is important for T cell development and their functional activation[29,30]. The metabolic responses characteristic of lymphocytes development and activation are both well-regulated at transcriptional and post-transcriptional levels[31]. For example, several groups have shown that T or B cell activation leads to mitochondrial remodeling and dramatic shifts in cell metabolism, as part of their role in eliminating pathogens[32-36]. Meanwhile the selection against pathogenic mutations 3243 was stronger in T cells than B and NK as shown by Walker et.al[16]. All these evidences suggested that the mtDNA genetic bottleneck may be one of several potential mechanisms in the regulation of mitochondrial genome in different lineages.

Our novel discovery of a somatic mtDNA bottleneck specifically within the lymphoid lineage may play a role in the quality control of mitochondrial genomes, in parallel to the selection of immunoreceptor genes in the nuclear genome. Thus, a robust population of mtDNA may be crucial for lymphocyte-mediated immune responses. These findings provide new insight into immune degeneration and related diseases. The causing and the consequence of the somatic mtDNA genetic bottleneck require extensive efforts to explore.

CODE AVAILABILITY

Code used for single cell data analysis and computational modeling are available at https://github.com/tangzhj/Bottleneck

ACKNOWLEDGEMENTS

We thank Weiwei Zhai (Institute of Zoology, Chinese Academy of Sciences) for constructive comments and suggestions on the manuscript. This work was supported by National Natural Science Foundation of China (32070644) to J.X., Guangdong Basic and Applied Basic Research Foundation (2019A1515110387, 2019A1515110387) to J.X., the Natural Science Foundation of Guangdong (2021B1515020042) to Z.H., NIH grants RM1-HG007735 to H.Y.C. and China Postdoctoral Science Foundation (E12503) to Z.L. H.Y.C. is an Investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

J. X., Z.H and H.Y.C. designed and conceived the study. Z.T., B.C. and W.Z. collected and analyzed the data. Z.H. and Z.L. designed the dilution model and performed the simulations. J. X., Z.H., Z. T. and Z.L. wrote the manuscript with inputs from all authors. All authors read and approved the final manuscript.

COMPETING INTERESTS

H.Y.C. is a co-founder of Accent Therapeutics, Boundless Bio, Cartography Biosciences, and is an advisor to 10x Genomics, Arsenal Biosciences, and Spring Discovery. The other authors declare no conflict of interest.

ONLINE METHODS

Data collection

The mtscATAC-seq dataset generated through evaluation of hematopoietic and PBMCs was retrieved from a recent study evaluating samples from a healthy 47-year-old donor[17]. The mtscATAC-seq dataset from human bone marrow from 25-year-old healthy donor was obtained from Mimitou et.al[18]. The scATAC-seq data from $CD4^+$ T cells were obtained from the study published by Satpathy et al.[37]. The scATAC-seq dataset for hematopoietic stem cells (HSCs), multi-potent progenitors (MPPs),

266 lymphoid-primed multipotent progenitors (LMPPs), common lymphoid progenitors
267 (CLPs), common myeloid progenitors (CMPs), granulocyte-macrophage progenitors
268 (GMPs) and plasmacytoid dendritic cells (pDCs) derived from $CD34^+$ bone marrow was
269 obtained from Buenrostro et al.[38] (**Extended Data Fig. 2**). The scRNA-seq dataset
270 generated from an evaluation of healthy $CD34^+$ PBMCs, bone marrow mononuclear
271 cells (BMMCs) and total PBMCs was downloaded from the study published by Granja
272 et al.[39] These datasets were used to analyze mtDNA replication and gene transcription
273 (**Methods**). The scRNA-seq dataset of 70 effector memory T cells (Tem cells), 70
274 central memory T cells (Tcm cells) and 142 $CD4^+$ regulatory T cells (Treg cells) from
275 healthy human colon tissue were downloaded from Array Express (E-MTAB-6072)[40].
276 Detailed information on data resources is provided in **Supplementary Table 1**.
277

278 **Single-cell (sc)ATAC-seq data pre-processing and annotation of the cell**
279 **populations**

280 Raw data from GSE142745 were processed with Cell Ranger ATAC (version 2.0.3; 10x
281 Genomics, https://www.10xgenomics.com/products/single-cell-atac) with default
282 parameters. Reads were aligned to the reference hg19 human genome
283 (https://support.10xgenomics.com/single-cell-atac/software/downloads/latest). In each
284 cell, 40% of fragments overlapping a compendium of DNase hypersensitivity peaks
285 and 1,000 unique nuclear fragments were filtered. From the output of the Cell Ranger
286 Software calls, we performed a computational annotation of the cell types on the basis
287 of chromatin accessibility. Clustering and gene activity scores were determined
288 through standard processing via ArchR[41]. Clustering was performed with the
289 "addClusters" and "addUMAP" functions (resolution=0.8, neighbors=10, minDist=0.1).
290 To identify marker genes according to gene scores, we used the "getMarkerFeatures"
291 function with useMatrix "GeneScoreMatrix" and generated a reproducible peak set in
292 ArchR by using the "addReproduciblePeakSet" function. By default, ArchR attempts to
293 identify peaks by using the MACS2 algorithm[42]. Because common cell markers are
294 sometimes not suitable for classification with "GeneScoreMatrix", we used enhancer
295 accessibility to define the cell type. For example, we identified myeloid cells according
296 to the unique accessibility of enhancers at +85 kb and +87 kb in the interferon
297 regulatory factor (*IRF8*) locus. Plasmacytoid dendritic cells (pDCs) were identified on
298 the basis of the unique accessibility of +54 kb and +56 kb enhancers, as described by
299 Satpathy et al.[37]. Furthermore, to label scATAC-seq clusters with scRNA-seq

300 information, we used the "addGeneIntegrationMatrix" function, which integrates
301 scATAC-seq with scRNA-seq. Specific marker genes used to identify individual cell
302 types in scATAC-seq datasets of healthy $CD34^+$ hematopoietic cells and PBMCs are
303 documented in **Supplementary Table 2**.
304
305 **Mitochondrial DNA variants identified in single-cell ATAC-seq datasets**
306 Paired-end raw reads from each sample were aligned to the human reference genome
307 (hg19) with Cell Ranger ATAC after adapter sequences were trimmed. First, the reads
308 mapped to multiple sites or the nuclear genome, and duplicates were also removed.
309 The remaining reads were realigned to correct the potential mapping errors around
310 indels according to the process from GATK[43]. Bam files for each cell type were merged
311 to identified germline mtDNA variants (bulk VAF >90%). Variants with VAF >90%
312 shared among more than 90% cells were also considered germline mutations. Then
313 mtDNA variants were called for each individual cell with VarScan2[44] with "--min-var-
314 freq 0.01" and "--min-reads2 2". To identify high confidence somatic variants in single
315 cell, the following filter steps were applied.
316
317 First, the germline mutations identified in the merged bam file were removed.
318 Second, the following sites were explicitly removed because of the large numbers of
319 homopolymers in the revised Cambridge Reference Sequence (rCRS) and sequencing
320 errors in the reference genome[13]:
321 Misalignment due to ACCCCCCCTCCCCC (rCRS 302–315), including
322 302A/C, 309C/T, 311C/T, 312C/T, 313C/T and 316G/C;
323 Misalignment due to GCACACACACACC (rCRS 513–525), including
324 514C/A, 515A/G, 523A/C and 524C/G;
325 Misalignment due to 3107N in ACNTT (rCRS 3105–3109), including
326 3106C/A, 3109T/C and 3110C/A.
327 Third, sequencing errors can significantly affect the identification of somatic variants.
328 Therefore, sequencing errors known to be associated with a high error rate according
329 to Illumina NextSeq and sequence errors (G→T and C→A) from DNA damage were
330 removed.
331 Fourth, strand balance was required for confident somatic variants. For the given
332 variant site, we required the reads mapped to the forward strand to be above 30% but
333 below 70% of the total mapped reads for the variant allele.

334 Variants that passed the multiple filter steps were merged from all individual cells as
335 the final somatic variants. If the variant was sufficiently confident in any given cell, the
336 variant allele frequency was re-counted in all individual cells within the same cell type,
337 without any other constraints.
338

339 **Single-cell RNA-seq data processing and cell-type annotation**

340 Downstream analysis of scRNA-seq dataset was performed with Seurat[45] (version
341 3.2.2; https://satijalab.org/seurat). The following bioinformatic analyses were
342 performed in R software (version 3.6.0; https://www.r-project.org) with default settings
343 unless otherwise stated. Cells with <200 or >2,500 detected genes or with >5%
344 mitochondrial DNA were eliminated from further consideration. Normalization was
345 applied with the MAGIC package (version 2.0.3)[46] by following the Seurat v3 workflow.
346 We next calculated a subset of features that exhibited high cell-to-cell variability by
347 using the "FindVariableFeatures" function and identified 2,000 specific features.
348 Clusters were identified with the "Find-Neighbors" and "FindClusters" functions in
349 Seurat with 45 principal components (PCs) and a resolution of 0.3. The results were
350 annotated to include differential expression of cell type-specific marker genes. Uniform
351 Manifold Approximation and Projection for Dimension Reduction (UMAP)
352 dimensionality reduction was performed with the "RunUMAP" function in Seurat, with
353 45 PCs and other default parameters. The expression of cell type-specific marker
354 genes in PBMCs and BMMCs is shown in **Supplementary Table 3**. We referred to the
355 information and classifications recorded in GSE139369 from the GEO Database to
356 guide our cell type annotations **(Supplementary Table 3)**.
357

358 **Pseudo-time analysis**

359 To construct single-cell differentiation trajectories with scRNA-seq data from HSCs to
360 B cells, we performed a pseudo-time analysis with the Monocle method[47-49]. First, we
361 subdivided scRNA-seq data according to the annotated cell populations revealed by
362 Seurat clustering analysis, according to the common pipeline (http://cole-trapnell-
363 lab.github.io/monocle-release/monocle3/). Re-clustering of selected cell populations
364 was again performed with the "RunUMAP" function. Pseudo-time analysis was

365 conducted on these newly generated clusters with Monocle v3. We delineated
366 expression patterns of G1/S phase-specific and mtDNA replication-related genes along
367 a pseudo-timeline. G1/S phase-specific genes were identified according to a previously
368 annotated list[50] (**Extended Data Fig. 3d**).
369

370 **Mitochondrial DNA variants identified from single-cell RNA-seq data**

371 Mitochondrial DNA variants from single-cell RNA-seq data were processed in the same
372 manner as mtDNA variants from scATAC-seq, with several modifications. Briefly, we
373 used STAR[51] to align reads to the human reference genome (hg19) and to obtain bam
374 files. Germline mutations and mtDNA variants in individual cells were filtered and called
375 in the same manner.
376

377 **Allele frequency spectrum**

378 The allele frequency (heteroplasmic ratio) of each mutation were calculated in each
379 cell and the number of mutations fall in each frequency bin (from 0~1) were counted
380 for each cell types. Somatic mutations arose in the early development stage, which
381 had been fixed in the progenitor cells, were further excluded for the ASF analysis in
382 the mtscATAC-seq from BMMCs.
383

384 **Annotation of mitochondria DNA mutations and calculation of non-**
385 **synonymous/synonymous mutation rates (dN/dS)**

386 The mitochondrial variants were annotated with ANNOVAR[52]. The annotated variants
387 comprised mutations in loops, tRNA, rRNA and mRNA coding regions, including non-
388 synonymous (NS) and synonymous (SY) substitutions according to the variant location
389 (**Extended Data Fig. 5a** and **Extended Data Fig. 6a**). Coding sequences (CDS) within
390 the mitochondrial genome were evaluated with Phylogenetic Analysis of Maximum
391 Likelihood (PAML) to identify all possible synonymous (defined as S) and
392 nonsynonymous (defined as N) substitutions in the human mitochondrial genome[53].
393 On the basis of ANNOVAR's annotations, we identified all observed synonymous
394 (defined as s) and nonsynonymous substitutions (defined as n). The non-synonymous
395 mutation rate (dN)=n/N and the synonymous mutation rate (dS)=s/S, responses to

396 positive, neutral, or negative selection pressure, can be determined by the dN/dS ratio.
397
398 **Computational modeling of the mitochondrial genetic bottleneck**
399 We used the Wright-Fisher model from population genetics to depict the accumulation
400 of mutations and the dynamic frequency of heteroplasmic alleles in mtDNA during
401 lymphoid cell divisions. The Wright-Fisher model assumes discrete generations and
402 random sampling of individuals from the current generation without replacement by
403 reproduction in the following generation. This model has been widely used to model
404 the mtDNA population dynamics in both germline cells and somatic cells, including
405 those that are neoplastic[23,54]. Because normal somatic cells typically contain 100–
406 1,000 copies of mtDNA, we used n=500 as the baseline copy number in our model[22].
407 Results from the scATAC dataset revealed that the relative copy number of mtDNA in
408 NK cells was approximately 60% that detected in B or T cells (**Fig. 3a**); thus, 300
409 (500×0.6) was used as the baseline mtDNA copy number for the NK lymphocyte cohort.
410 We modeled the lymphoid development from lymphoid-primed multipotent progenitor
411 (LMPP) cells, which are the common progenitor cells for all lymphocytes, B, T and NK
412 cells. To model the dilution-based genetic bottleneck, we introduced a dilution rate α,
413 which denotes the fraction of mtDNA molecules in each cell that undergo replication
414 within a single cell cycle, and T_d, which denotes the time of the diluting process. After
415 T_d cell divisions from LMPP, the mtDNA copy number in each cell type rapidly recovers
416 to the baseline level. The minimal mtDNA copy number through the bottleneck can be
417 computed by:
418 $$N_b = N_0 \alpha^{T_d} \qquad \text{Eq (1)}$$
419 where N_0 is the initial number of mtDNA copies. The total number of cell divisions
420 required for the transition from an LMPP to a mature lymphocyte is denoted T_a. The
421 mutation rate at each site within the mitochondrial genome per cell division is denoted
422 μ, which has been estimated to be 10^{-8}–10^{-7} mutations per site for somatic cells[23,55].
423 Thus, the mutation rate for the entire mitochondrial genome during each cell division
424 event will be $u = \mu \times L$, where L=16,569 base pairs (bp), representing the number of
425 potential sites within the mitochondrial DNA length.
426
427 During each cell division, the number of somatic mutations acquired per mitochondrial
428 genome follows a Poisson distribution with a mean of u. Thus, the probability that k
429 mutations occurred in each cell division is as follows:

$$P(x = k) = \frac{u^k e^{-u}}{k!} \qquad \text{Eq (2)}$$

Computational inference of parameters by approximate Bayesian computation

We used the framework of approximate Bayesian computation (ABC) for parameter inference in our computational model of somatic mtDNA population dynamics on the basis of the dilution rate α, the dilution time course T_d and the total number of cell divisions T_a. The minimal mtDNA copy number in each cell can be computed as described by Eq (1) when values for α and T_d are available. The prior uniform distributions used for sampling α, T_d and T_a, were $\alpha \sim U(0,1)$, $T_d \sim U(0,30)$ and $T_a \sim U(10,40)$. To avoid extinction (i.e., minimal mtDNA copy number=0), only the sampled parameter values ensuring $N_b\ (= N_0 \alpha^{T_d})$ >10 were retained. We used a version of ABC based on the acceptance-rejection algorithm[56] to estimate posterior probability distributions for the parameters of interest (i.e., $\boldsymbol{\theta}$ $[\alpha,\ T_d, T_a]$. We used 19 summary statistics (***S***), which included the mtDNA mutation count in each VAF bin as step=0.05 from VAF=0.05 to 1 to fit the simulated to the observed data. The ABC version of rejection sampling is as follows:

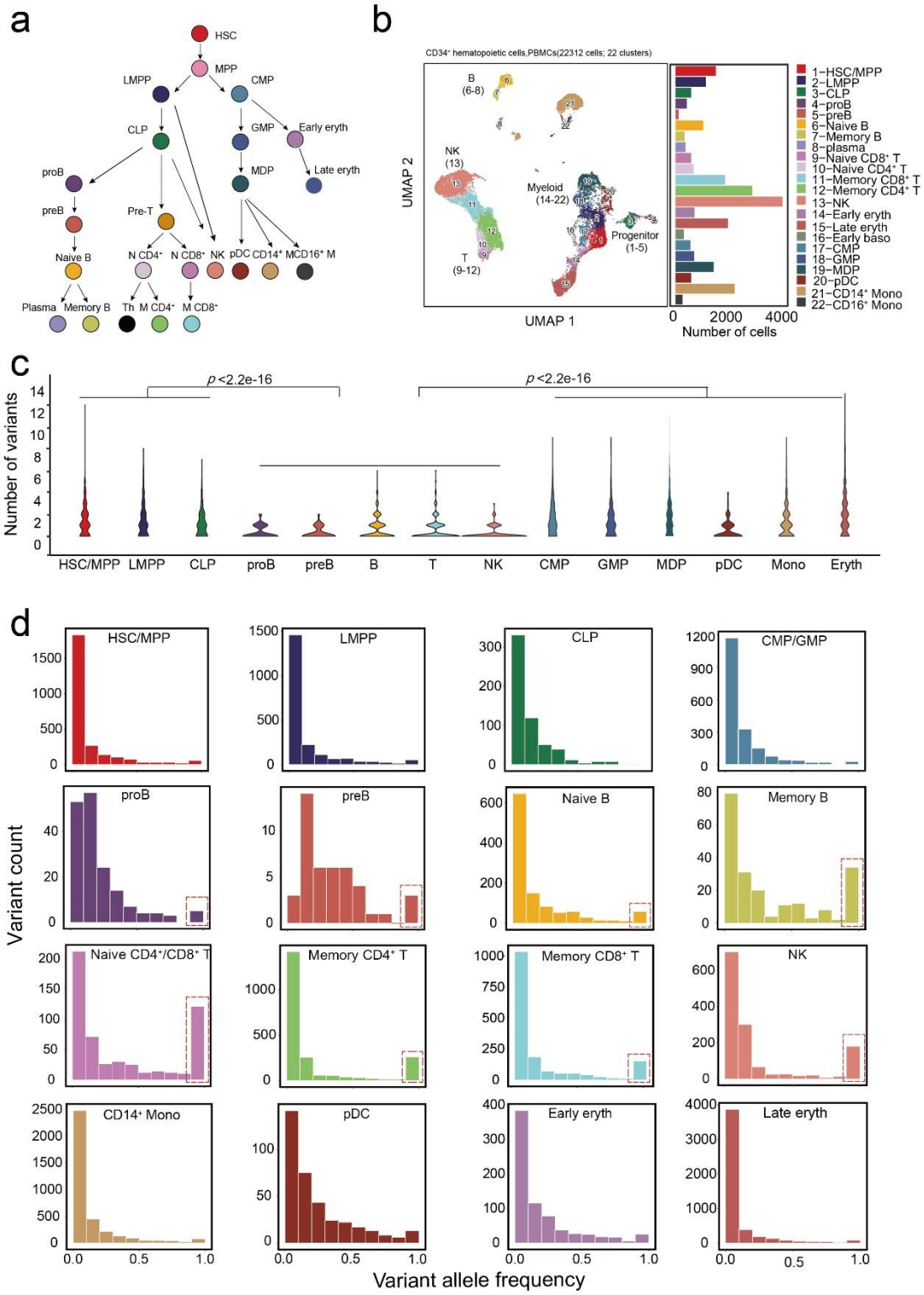
For *i*=1 to *K* simulations:

1. Sample parameters θ' from the prior distribution $\pi(\boldsymbol{\theta})$
2. Simulate data ($\mathbf{D}'$) with the sampled parameters (θ') and summarize $\mathbf{D}'$ as summary statistics ($\mathbf{S}'$).
3. Accept θ' if $d(\mathbf{S}', \mathbf{S}) < \varepsilon$, for a given tolerance rate ε, where $d(\mathbf{S}', \mathbf{S})$ is a measure of the Euclidean distance between $\mathbf{S}'$ and $\mathbf{S}$
4. Return to step 1.

With this scheme, we approximated the posterior distribution by $P(\boldsymbol{\theta} | d(\mathbf{S}', \mathbf{S}) < \varepsilon)$. We used a common variation in ABC[57,58] in which, rather than using a fixed threshold, ε, we sorted all calculated *K* distances by $d(\mathbf{S}', \mathbf{S})$ (see step 3 above) and accepted the θ' that generated the smallest $100 \times \eta$ percentage distances. We used $K=10^6$ and $\eta=0.001$ so that the posterior distribution was composed of $10^6 \times 0.001 = 1{,}000$ data points. We ran the ABC inference procedures for two mutation rates ($\mu=10^{-8}$ and 10^{-7})

460 and performed model selection (**Extended Data Fig. 4**). The mutation rate $\mu=10^{-7}$
461 fitted the data better in all cell types and thus was used for the computational inference.
462 The ABC procedure was performed with the R package *abc*[59].

463 **Figures and Legends**



464

465 **Fig. 1 Somatic mutations in the mtDNA of PBMCs.**

466 (**a**) Schematic of human hematopoietic differentiation and lineage commitment. HSC,
467 hematopoietic stem cell; MPP, multipotent progenitor; LMPP, lymphoid-primed

468 multipotent progenitor; CLP, common lymphoid progenitor; CMP, common myeloid
469 progenitor; GMP, granulocyte-monocyte progenitor; MDP, monocyte-dendritic cell
470 progenitor; N CD4, naïve $CD4^+$ T cell; N CD8, naïve $CD8^+$ T cell; M CD4, memory $CD4^+$
471 T cell; M CD8, memory $CD8^+$ T cell; Th, T helper cell; NK, natural killer cell; pDC,
472 plasmacytoid dendritic cell; Eryth, erythrocyte.

473 (**b**) UMAP projection of 22,312 CD34+ hematopoietic cells and PBMCs with
474 mtscATAC-seq data. Dots represent individual cells that have been colored according
475 to cluster identity. The bar plot indicates the number of cells in each cluster (labeled at
476 right).

477 (**c**) Violin plot showing the number of somatic mtDNA variants per cell for various cell
478 types; *P*-values, two-sided Wilcoxon rank-sum test.

479 (**d**) The VAF distribution of somatic mtDNA mutations across different cell types.
480 Homoplastic mutations (VAF ~1) identified in the lymphoid lineage are highlighted with
481 a red box.

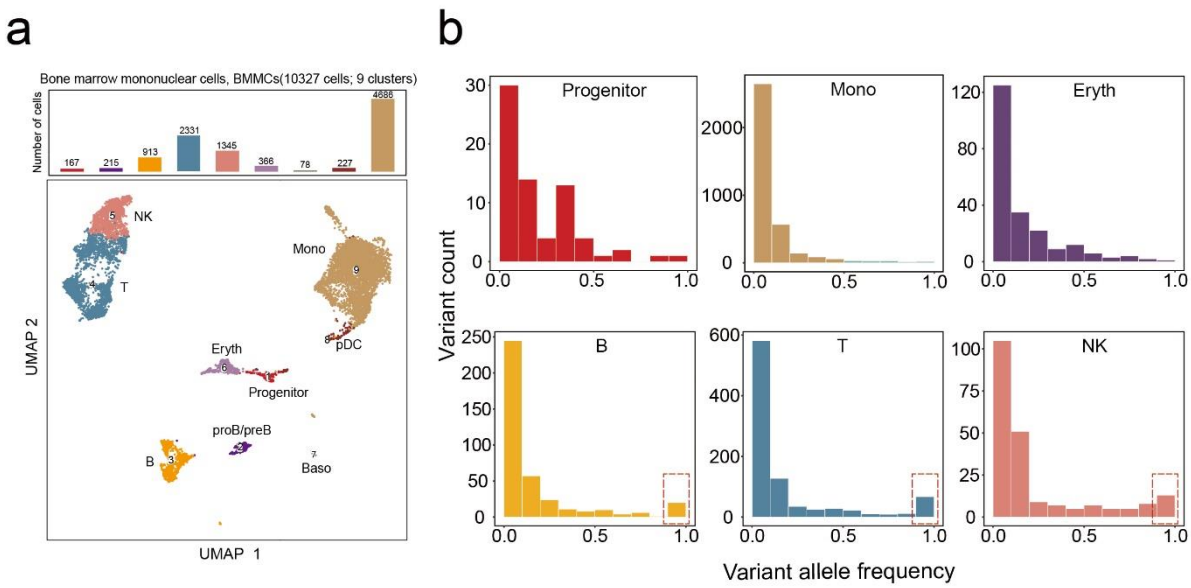


Fig. 2 Somatic mutations in the mtDNA of BMMCs.

(a) UMAP projection of 10,327 mononuclear cells from bone marrow with mtscATAC-seq data. Dots represent individual cells that have been colored according to cluster identify and cell types.

(b) The VAF distribution of somatic mtDNA mutations across different cell types in BMMCs. Homoplastic mutations (VAF ~1) identified in the lymphoid lineage are highlighted with a red box.

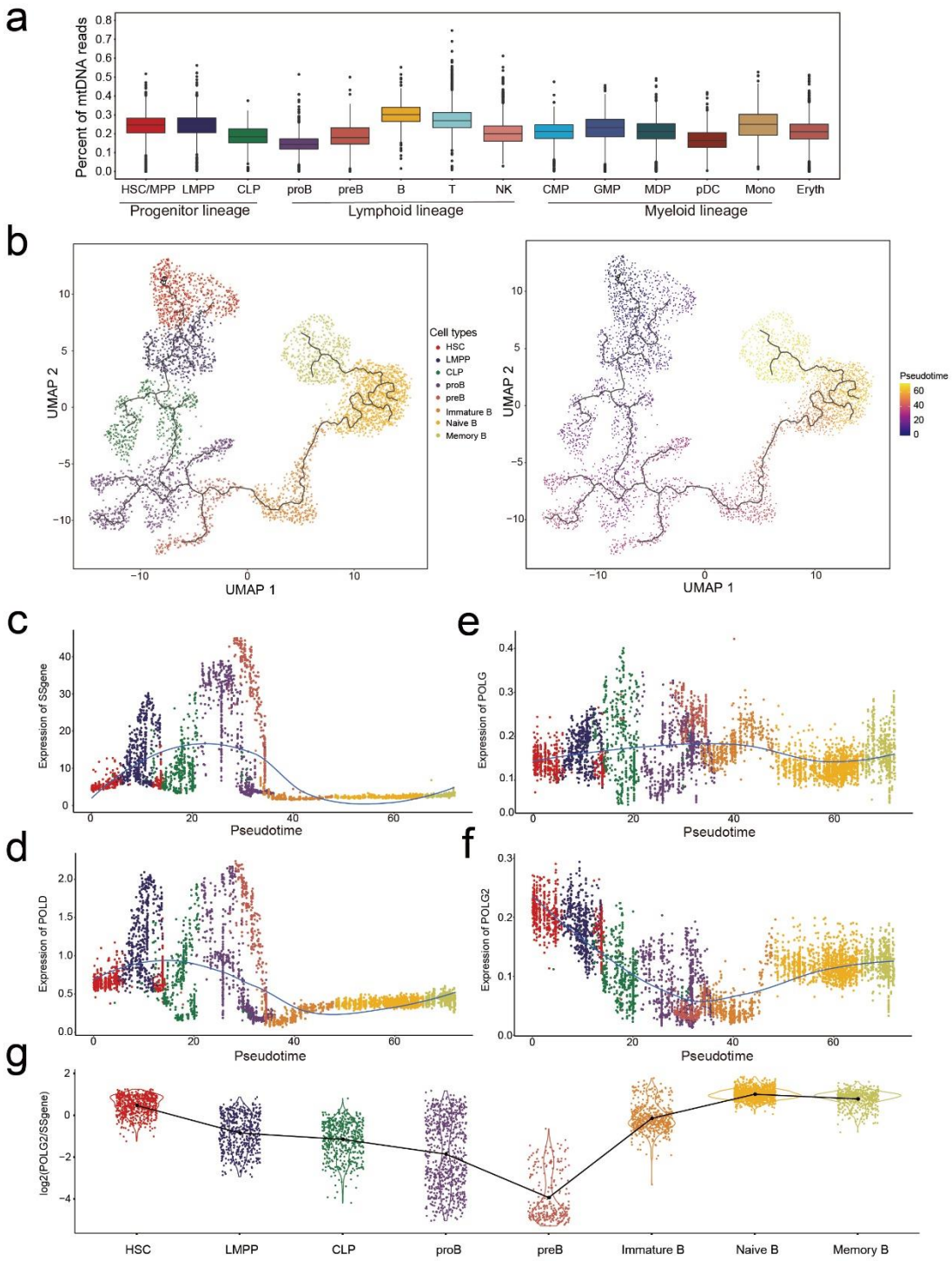


Fig. 3 Replication of mtDNA during B cell development.

(**a**) The relative number of mtDNA copies was determined by the proportion (sequencing reads mapped to mitochondrial genome divided by the total number of reads) in each cell, as identified from the scATAC-seq dataset.

(**b**) Pseudo-time trajectory of B cell differentiation from HSCs by using single-cell RNA-

496 seq data generated from PBMCs (n=4692 cells). Individual colors denote different cell
497 types (top) and developmental stages (bottom) defined by pseudo-time. The solid line
498 represents the fitted trajectory across pseudo-time.
499 (**c-f**) Kinetic plots showing the expression of (**c**) 39 G1/S phase-specific genes
500 (SSgene), (**d**) nuclear DNA polymerase δ (*POLD1–3*), (**e**) mtDNA polymerase γ (POLG)
501 and (**f**) the binding subunit of mitochondrial DNA polymerase γ (*POLG2*) along the B-
502 cell developmental trajectory.
503 (**g**) Violin plot showing the ratio of *POLG2* expression to the mean expression of all
504 G1/S phase-specific genes in each cell associated with B-cell development. The
505 broken line represents the change trend of the mean ratio across different cell types.

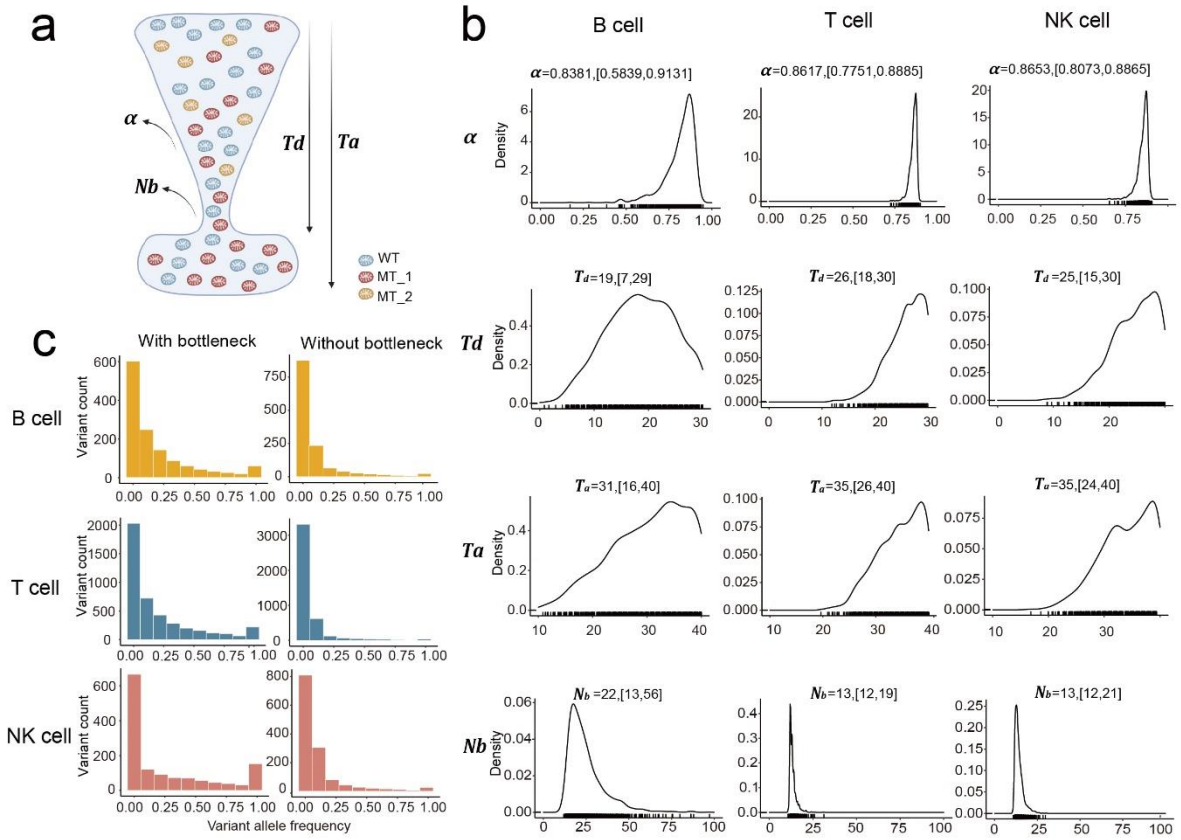


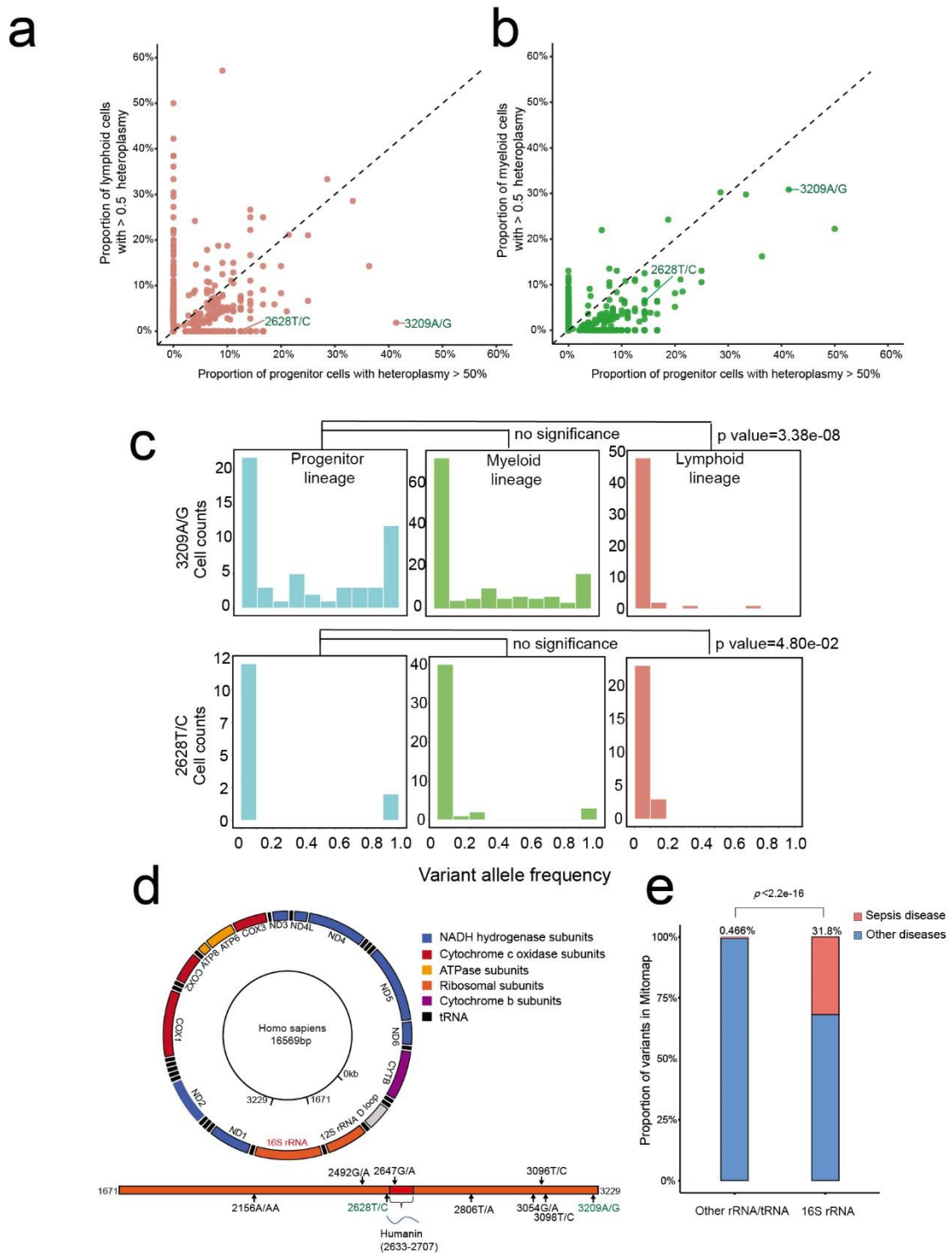
Fig. 4 The dilution model of the mitochondrial genetic bottleneck.

(**a**) Schematic illustration of the dilution model of the mitochondrial genetic bottleneck. In this model, only a fraction of mtDNA molecules (denoted by α) replicate at each cell division. After T_d cell divisions from the LMPP stage, the number of mtDNA copies in each lymphocyte subtype (B, T, and NK cells) undergoes rapid recovery to the baseline level (~500 per cell). N_b denotes the minimal number of mtDNA copies that can be computed as Eq (1). The total number of cell divisions required for the transition from LMPP to mature lymphocyte is denoted as T_a.

(**b**) The distribution of model parameters inferred by the Approximate Bayesian Computation (ABC) algorithm. The mean and 95% confidence interval of each parameter estimation is as shown.

(**c**) Simulations based on the dilution model of mitochondrial genetic bottleneck with the ABC-estimated parameter values recapitulated the lymphocyte-specific overrepresentation of homoplastic mutations and the lower mutation burden (**Extended Data Fig. 3b**). The left and right panels represent the simulations with and

522 without mitochondrial genetic bottleneck, respectively. The average of 100 simulations
523 carried out for each model is as shown. The results of each iteration are shown in
524 **Extended Data Fig. 3c**.

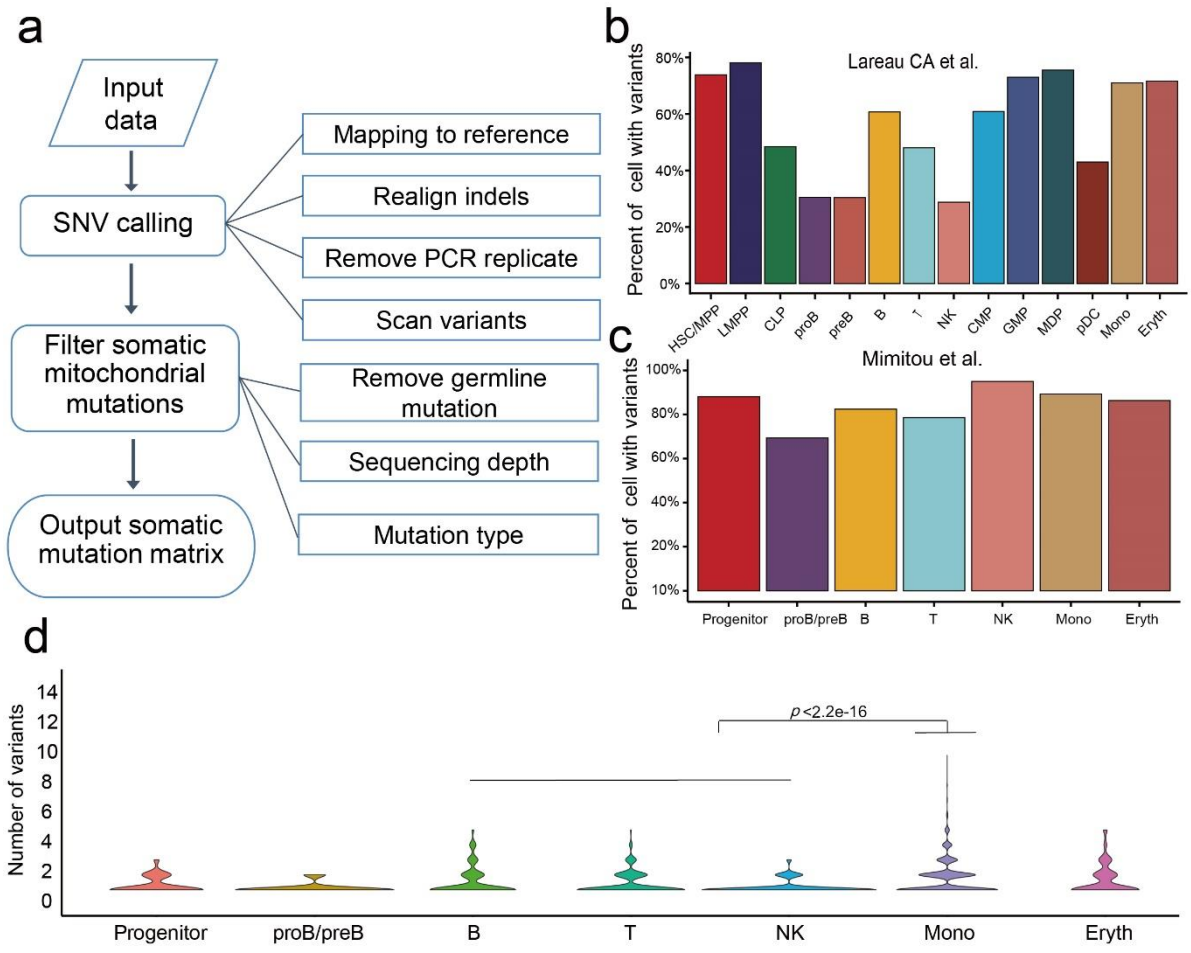


525

526 **Fig. 5 Elimination of specific mtDNA variants in lymphocyte.**

527 (**a**) Scatter plot documenting the percentage of cells with dominant mtDNA mutations
528 **(**VAF>50% in a single cell**)**. Shown are the results from progenitor cells (HSC, MPP,
529 and LMPP) compared to cells from lymphoid (B, T, or NK cells) or myeloid lineages(**b**).

530 (**c**) The distribution of VAF for two individual sites (3209A/G, 2636G/A) in progenitor,
531 myeloid and lymphoid cells, respectively. The *p* values shown were determined by the
532 Chi-square test.
533 (**d**) The location of 16S RNA (MT-RNR2) on the mitochondrial genome and the location
534 of sepsis association variants on MT-RNR2, reported in MITOMAP (in black), or
535 specific eliminated in lymphocytes (in green).
536 (**e**) The proportion of mtDNA variants associated with sepsis disease in 16S RNA
537 versus other rRNA/tRNA genes on mitochondrial genome.



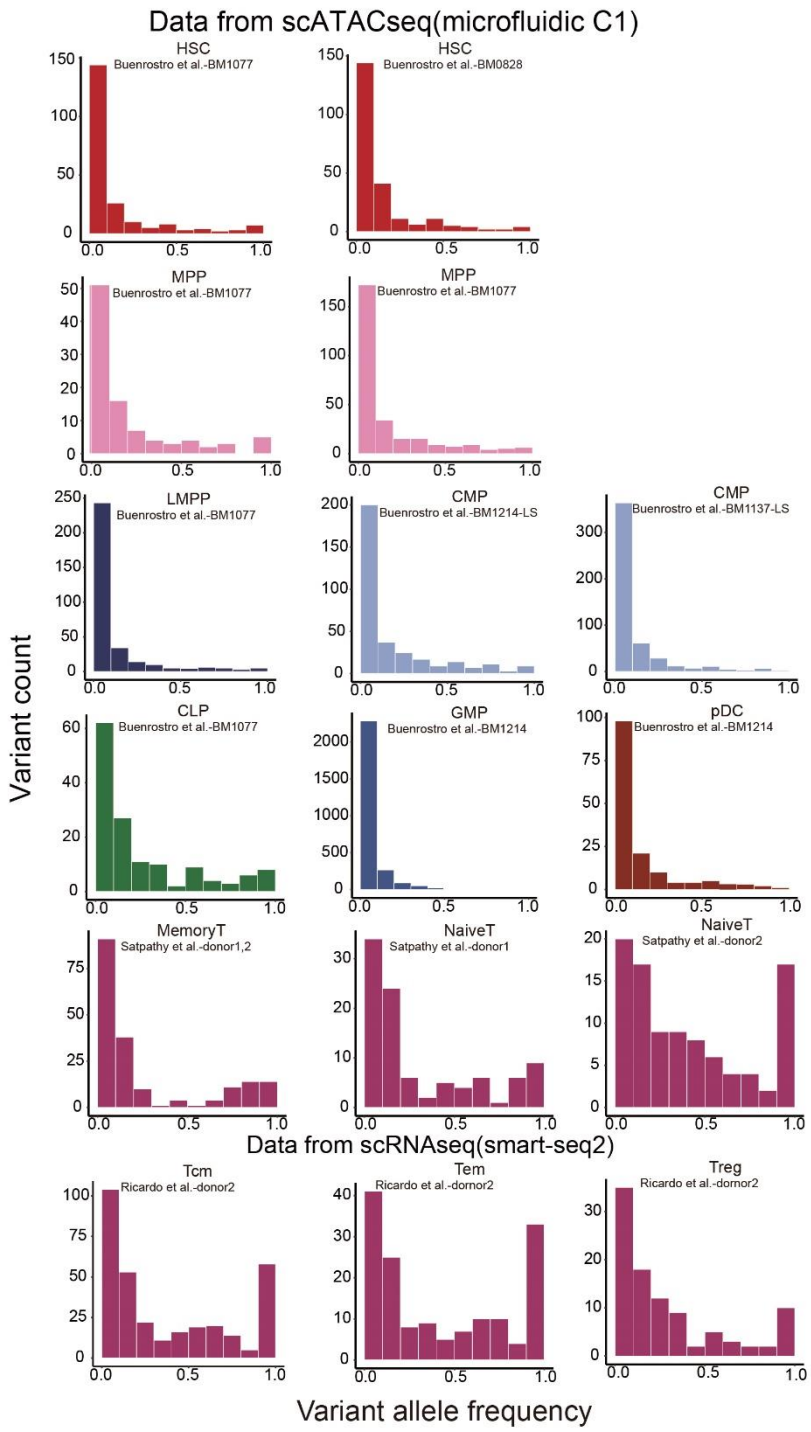
Extended Data Fig. 1 Detection of somatic mitochondrial mutations in single PBMCs with scATAC-seq data or scRNA-seq data.

(**a**) Schematic of mtDNA mutation calling with scATAC-seq (including mtscATAC-seq) or scRNA-seq data.

(**b**) Percentage of cells with at least one somatic mtDNA mutation detected in individual cells for each cell type in the mtscATAC-seq data from Lareau et al.

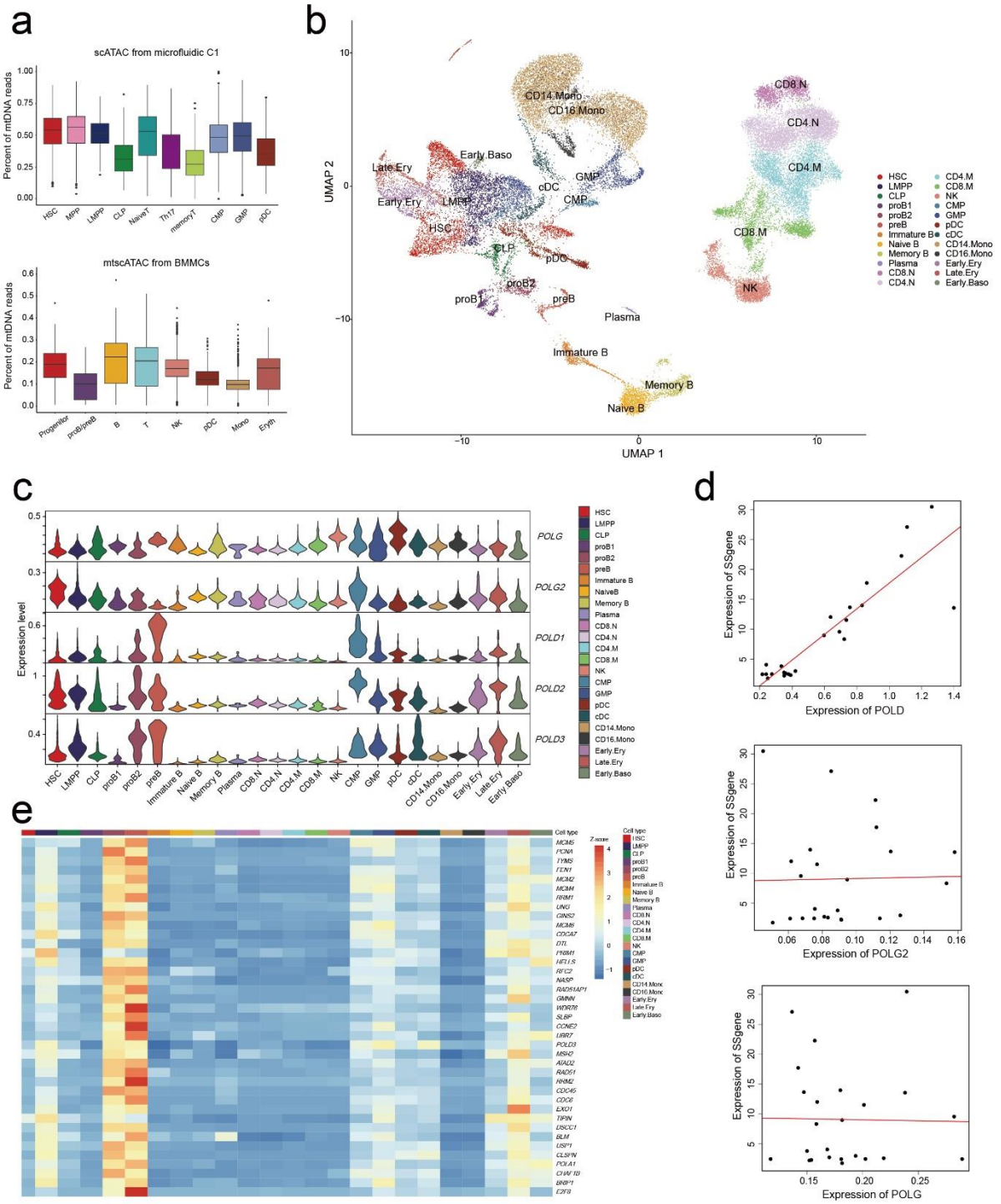
(**c**) Percentage of cells with at least one somatic mtDNA mutation detected in individual cells for each cell type in the mtscATAC-seq data from Mimitou et al.

(d) Violin plot showing the number of somatic mtDNA variants per cell for various cell types; *p* values based on a two-sided Wilcoxon rank-sum test are as shown.



Extended Data Fig. 2

Allele frequency spectrum of somatic mtDNA mutations for different hematopoietic cell types, on the basis of independent scATAC datasets (Buenrostro et al. and Satpathy et al.) and an scRNA-seq dataset (Ricardo et al.).



557

558 **Extended Data Fig. 3 Gene expression of G1/S phase-specific genes in scRNA-**
559 **seq data.**

560 (**a**) Relative mtDNA copies were measured by the percentage of the sequencing reads
561 mapped to the mitochondrial genome out of the total number of reads for each cell
562 types.

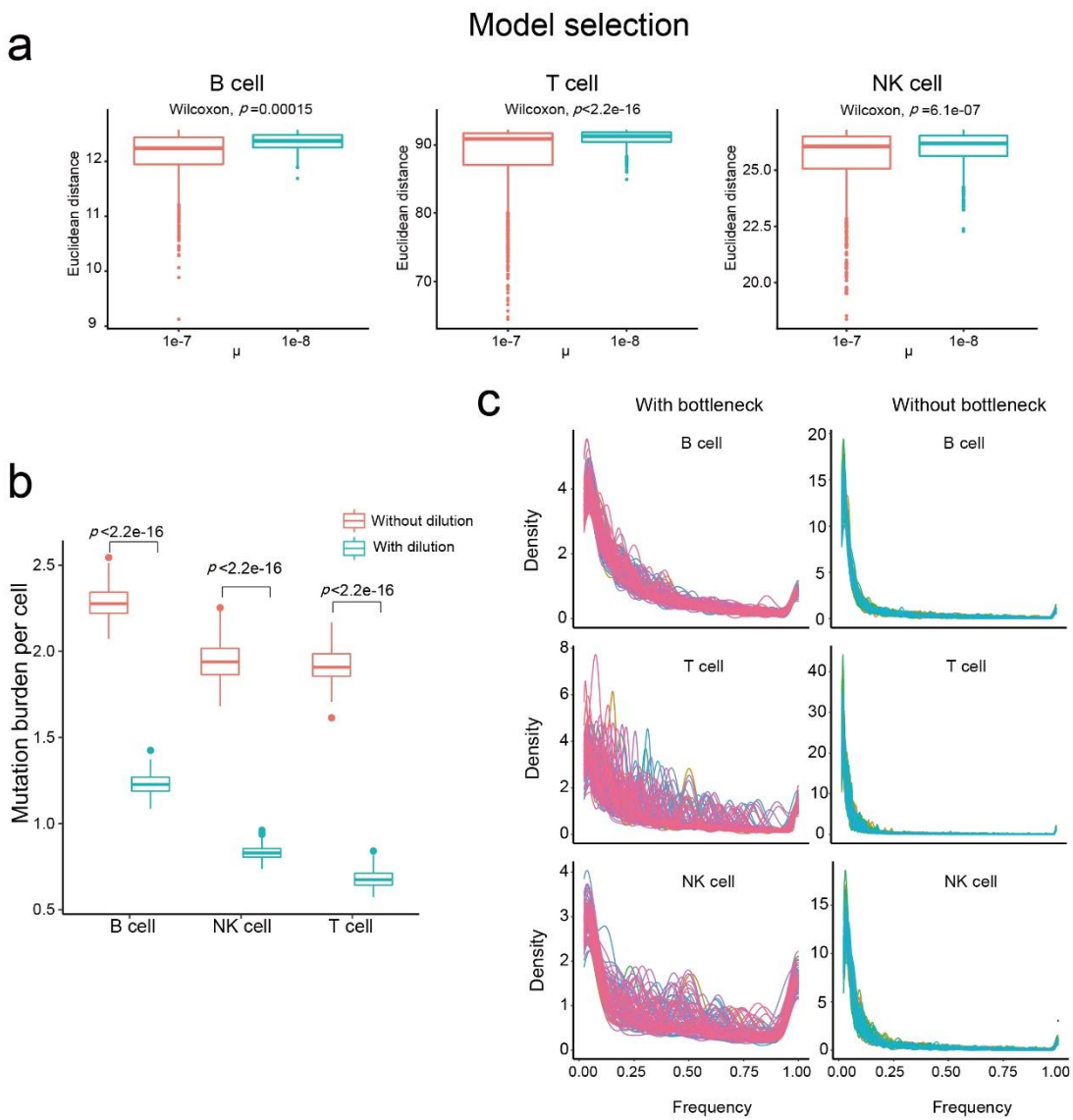
563 (**b**) UMAP projection of PBMCs, BBMCs and CD34^{+} PBMCs with scRNA-seq data.

564 Dots represent individual cells colored by cell types.

565 (**c**) Violin plots showing the expression of mitochondrial DNA polymerase γ (*POLG*)
566 and its binding subunit (*POLG2*) and nuclear DNA replication polymerase genes
567 (*POLD1–3*) from scRNA-seq data.

568 (**d**) Scatter plot showing the correlation of the gene expression of *POLD* (*POLD1–3*),
569 *POLG* and *POLG2* with G1/S phase-specific genes (SSgene).

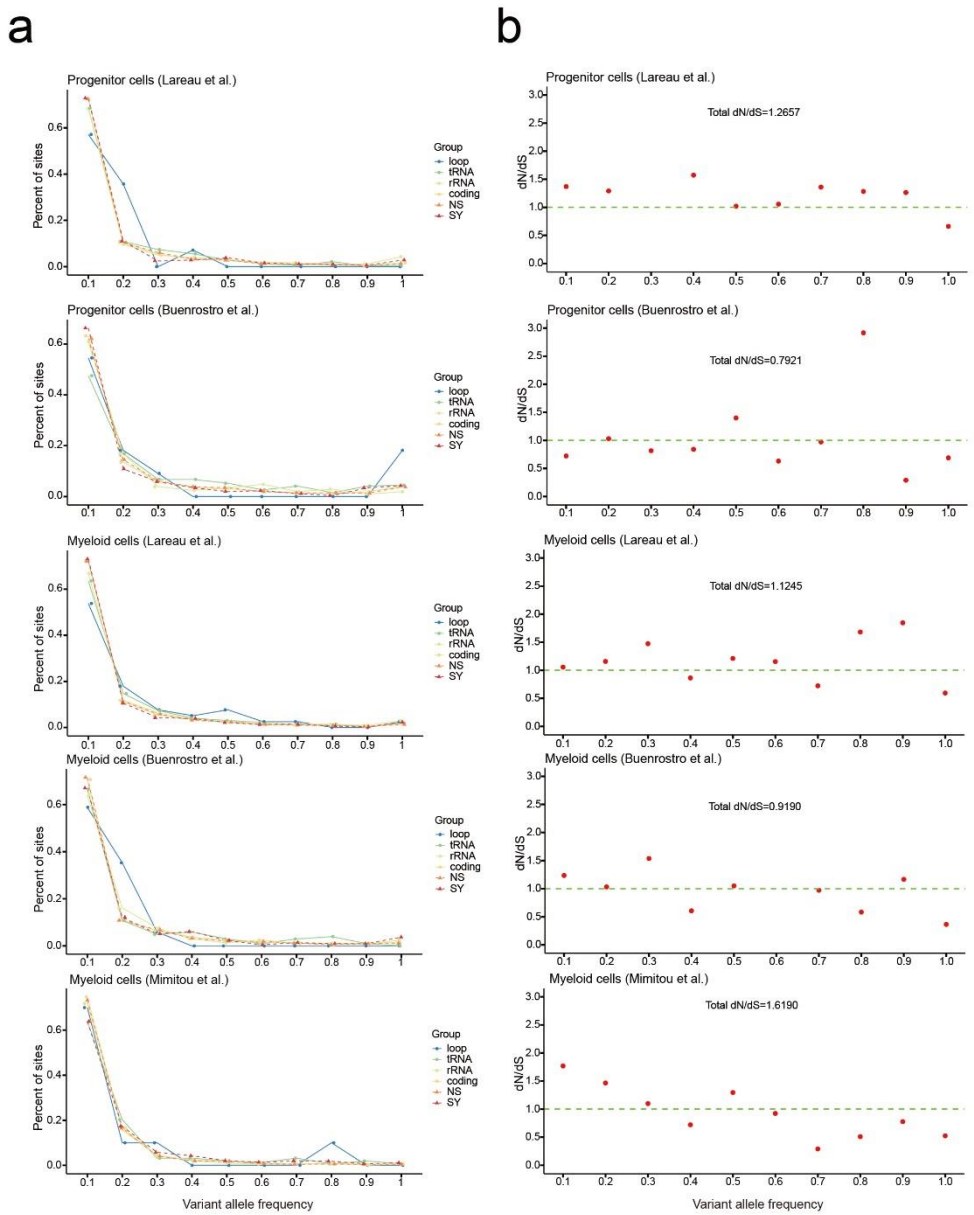
570 (**e**) Heat map showing the expression of 39 G1/S phase-specific genes in 24 cell types.



Extended Data Fig. 4 Details regarding the parameter inference for the dilution model of the mitochondrial genetic bottleneck.

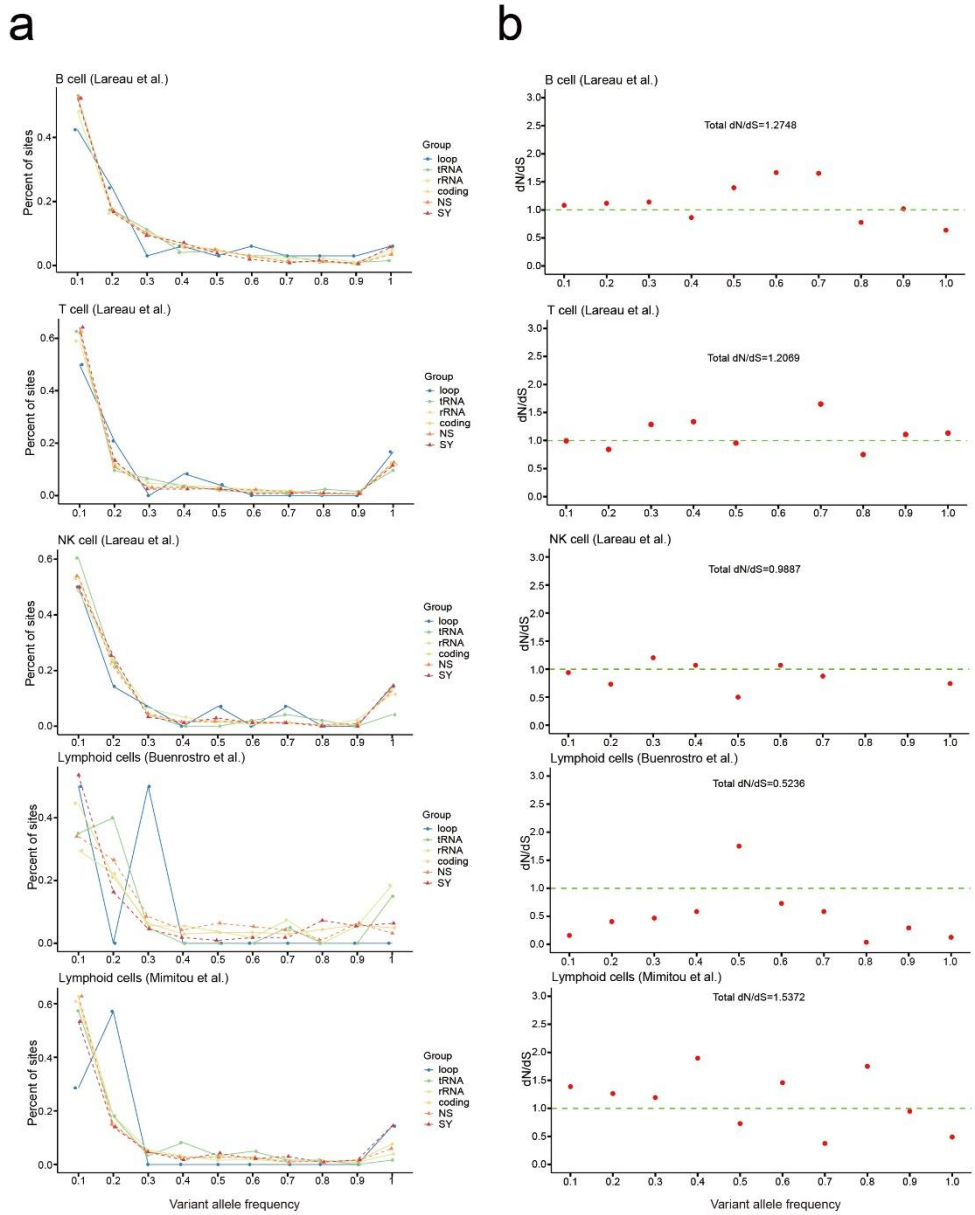
(a) Model selection with respect to the per-site mutation rate μ. We ran the ABC inference procedures for two mutation rates μ=10^{-8} and 10^{-7}, and μ=10^{-7} fitted the data better (smaller Euclidean distance between simulated and observed summary statistics) in all cell types and thus was used for the parameter inference. (**b**) Simulations under the dilution model of the mitochondrial genetic bottleneck with the ABC-estimated parameter values recapitulated the lower mutation burden in B, T and NK cells, as compared with simulations without a mitochondrial genetic bottleneck. (**c**) Simulations (100 times) with inferred parameters from the dilution model under

582 conditions with or without a mitochondrial genetic bottleneck. Each curve represents
583 one simulation.



Extended Data Fig. 5 Allele frequency spectrum of somatic mtDNA mutations for different types and dN/dS in the progenitor and myeloid lineages.

(**a**) Distribution of the VAF for mutations in different mtDNA genomic regions or types in progenitor and myeloid cells. The color code corresponds to mtDNA genomic regions or mutation types, annotated as loop, tRNA, rRNA, coding (coding region), NS (non-synonymous) and SY (synonymous). (**b**) The dN/dS ratio (y-axis) for mutations in different VAF bins (x-axis).



Extended Data Fig. 6 Allele frequency spectrum of somatic mtDNA mutations for different types and dN/dS in the lymphoid lineage.

(**a**) Distribution of VAF for mutations in different mtDNA genomic regions in lymphoid cells (B, T and NK). The color code corresponds to mtDNA genomic regions or mutation types, annotated as loop, tRNA, rRNA, coding (coding region), NS (non-synonymous) and SY (synonymous). (**b**) dN/dS ratio (y-axis) for mutations in different VAF bins (x-axis).

601 **Supplementary Table 1**

TableS1-DataResource

Cell Type	DataSet	library	Platform	# of cells	# of individuals	Notes	Author
naïve T cell	GSE107816	scATACseq+ C1	Illumina NextSeq 500	96	1	healthy	Satpathy et al. 2018
naïve T cell (sample 2)	GSE107816	scATACseq+ C1	Illumina NextSeq 500	192	2	healthy	Satpathy et al. 2018
memory T cell	GSE107816	scATACseq+ C1	Illumina NextSeq 500	192	2	healthy	Satpathy et al. 2018
Th17	GSE107816	scATACseq+ C1	Illumina NextSeq 500	192	2	healthy	Satpathy et al. 2018
Multipotent progenitor cell (MPP)	GSE96769	scATACseq+ C1	Illumina NextSeq 500	192	2	BM1077-frozen, BM0828-frozen	Buenrostro et al. 2018
Lymphoid-primed multipotent progenitor (LMPP)	GSE96769	scATACseq+ C1	Illumina NextSeq 500	96	1	BM0828-frozen	Buenrostro et al. 2018
Common lymphoid progenitor (CLP)	GSE96769	scATACseq+ C1	Illumina NextSeq 500	192	2	BM1077-frozen, BM0828-frozen	Buenrostro et al. 2018
Hematopoietic stem cell (HSC)	GSE96769	scATACseq+ C1	Illumina NextSeq 500	480	3	BM1077-Frozen, BM0106-LS/SIM, BM0828-fresh/frozen	Buenrostro et al. 2018
Common myeloid progenitor (CMP)	GSE96769	scATACseq+ C1	Illumina NextSeq 500	672	4	BM1077-frozen/HYC, BM0828-frozen, BM1214-LS, BM1137-LS	Buenrostro et al. 2018
Granulocyte-monocyte progenitors (GMP)	GSE96769	scATACseq+ C1	Illumina NextSeq 500	288	3	BM1077-frozen, BM0828, BM1214	Buenrostro et al. 2018
Plasmacytoid dendritic cell (pDC)	GSE96769	scATACseq+ C1	Illumina NextSeq 500	192	2	BM1137-LS, BM1214-frozen	Buenrostro et al. 2018
CD34+ hematopoietic cells,PBMCs	GSE142745	mtscATAC-seq	NextSeq 550	22312	1	healthy	Lareau et al. 2020
Effector Memory T(Tem)	E-MTAB-6072	Smart-seq2	Illumina HiSeq 2500	64	1	healthy, organism part:colon	Ricardo et al. 2019
Central Memory T(Tcm)	E-MTAB-6072	Smart-seq2	Illumina HiSeq 2500	61	1	healthy, organism part:colon	Ricardo et al. 2019
CD4+ regulatory(Treg)	E-MTAB-6072	Smart-seq2	Illumina HiSeq 2500	131	1	healthy, organism part:colon	Ricardo et al. 2019
PBMCs, BMMCs,CD34+ bone marrow cells	GSE139369	CITE-seq	Illumina NovaSeq 6000	35434	4	healthy	Granja et al. 2019
Bone marrow cells	GSM4732140	mtscATAC-seq	NextSeq 550	10327	1	healthy	Mimitou et al. 2020

602

603 **Supplementary Table 2**

TableS2 MarkerGenes(scATAC)	
Celltype	**Marker gene for cell type annotation**
HSC/MPP	*AVP,HLF,CRHBP*
LMPP	*FCN2,RUNX1*
CLP	*MME*
proB	*RAG1,CD19,EBF1,IL7R*
preB	*RAG1,CD79B,MS4A1*
Naïve B	*IL4R,PAX5,MS4A1*
Memory B	*PAX5,MS4A1*
plasma	*TACI,BCMA,SDC1,IGKC*
Naive $CD4^+$ T	*ITGA6,CCR7*
Naive $CD8^+$ T	*CD8A,LEF1*
Memory $CD4^+$ T	*CD4,CD52*
Memory $CD8^+$ T	*CD8A,EOMES*
NK	*GNLY,NKG7,FCGR3B,FASLG*
CMP	*TAL1,GATA1*
GMP	*SPI1,MPO*
MDP	*FLT3,MPO*
pDC	*DERL3,FLT3*
Early.ery	*HBB,GATA1*
Late.ery	*GATA1*
Early.baso	*PF4,ITGA2B*
CD14mono	*NCR1,CEBPB*
CD16mono	*FCGR3A,SIGLEC10*

604

605 **Supplementary Table 3**

TableS3 MarkerGene(scRNA)	
Celltype	**Marker genes for cell type annotation**
HSC/MPP	*AVP,HLF,CRHBP*
LMPP	*FCN2,RUNX1*
CLP	*MME,DNTT,IL7R*
proB	*VPREB3,CD79A,CD79B,RAG1*
preB	*MS4A1,EBF1*
Immature B	*CR2,CD19*
Naïve B	*CD22*
Memory B	*CD27,IL4R*
plasma	*TACI,BCMA,SDC1,IGKC*
Naive $CD4^+$ T	*SELL,CCR7,CD95*
Naive $CD8^+$ T	*LEF1,CD8A,IL2RB*
Memory $CD4^+$ T	*IL7R,CD52,CD4*
Memory $CD8^+$ T	*IFNG,EMOES*
NK	*GNLY,CCL5,CD56,NKG7*
CMP	*TAL1,MPO*
GMP/MDP	*ELANE,MPO,PRTN3,CSF1R*
pDC	*SPIB,DERL3,IRF8*
cDC	*CDC1,SPI1*
Early.ery	*HBB,GATA1*
Late.ery	*HBB,BLVRB*
Early.baso	*BCVRB , ITGA2B,PF4,LMO4*
CD14mono	*CD14*
CD16mono	*FCGR3A,SIGLEC10*

606

607 **REFERENCE**

608

609 1. Andersion, S. *et al.* Sequence and organization of the human mitochondrial genome.
610 *Nature* **290**, 457-465 (1981).
611 2. Hagström, E., Freyer, C., Battersby, B.J., Stewart, J.B. & Larsson, N.G. No
612 recombination of mtDNA after heteroplasmy for 50 generations in the mouse maternal
613 germline. *Nucleic Acids Research* **42**, 1111–1116 (2014).
614 3. Kazak, L., Reyes, A. & Holt, I.J. Minimizing the damage: Repair pathways keep
615 mitochondrial DNA intact. *Nature Reviews Molecular Cell Biology* **13**, 659–671 (2012).
616 4. Scheibye-Knudsen, M., Fang, E.F., Croteau, D.L., Wilson, D.M., 3rd & Bohr, V.A.
617 Protecting the mitochondrial powerhouse. *Trends Cell Biol* **25**, 158-70 (2015).
618 5. Lagouge, M. & Larsson, N.G. The role of mitochondrial DNA mutations and free
619 radicals in disease and ageing. *J Intern Med* **273**, 529-43 (2013).
620 6. Ye, K., Lu, J., Ma, F., Keinan, A. & Gu, Z. Extensive pathogenicity of mitochondrial
621 heteroplasmy in healthy human individuals. *Proceedings of the National Academy of*
622 *Sciences of the United States of America* **111**, 10654–10659 (2014).
623 7. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737-56
624 (1974).
625 8. Koehler, C.M. *et al.* Replacement of bovine mitochondrial DNA by a sequence variant
626 within one generation. *Genetics* **129**, 247-55 (1991).
627 9. Jenuth, J.P., Peterson, A.C., Fu, K. & Shoubridge, E.A. Random genetic drift in the
628 female germline explains the rapid segregation of mammalian mitochondrial DNA. *Nat*
629 *Genet* **14**, 146-51 (1996).
630 10. Ghosh, S.S., Fahy, E., Bodis-Wollner, I., Sherman, J. & Howell, N. Longitudinal study
631 of a heteroplasmic 3460 Leber hereditary optic neuropathy family by multiplexed
632 primer-extension analysis and nucleotide sequencing. *Am J Hum Genet* **58**, 325-34
633 (1996).
634 11. Zhang, R., Wang, Y., Ye, K., Picard, M. & Gu, Z. Independent impacts of aging on
635 mitochondrial DNA quantity and quality in humans. BMC Genomics 18, 890 (2017).
636 12. Taylor, R.W. & Turnbull, D.M. Mitochondrial DNA mutations in human disease. Nat
637 Rev Genet 6, 389-402 (2005).
638 13. Ju, Y.S. *et al.* Origins and functional consequences of somatic mitochondrial DNA
639 mutations in human cancer. *Elife* **3**(2014).
640 14. Ludwig, L.S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations
641 and Single-Cell Genomics. *Cell* **176**, 1325–1339.e22 (2019).
642 15. Xu, J. *et al.* Single-cell lineage tracing by endogenous mutations enriched in transposase
643 accessible mitochondrial DNA. *eLife* **8**, e45105 (2019).
644 16. Walker, M.A. *et al.* Purifying Selection against Pathogenic Mitochondrial DNA in
645 Human T Cells. *New England Journal of Medicine* (2020).
646 17. Lareau, C.A. *et al.* Massively parallel single-cell mitochondrial DNA genotyping and
647 chromatin profiling. *Nat Biotechnol* **39**, 451-461 (2021).
648 18. Mimitou, E.P. et al. Scalable, multimodal profiling of chromatin accessibility, gene
649 expression and protein levels in single cells. Nat Biotechnol (2021).
650 19. Radsak, K. & Schutz, E. Changes of mitochondrial DNA polymerase-gamma activity

651 in synchronized mouse cell cultures. *Eur J Biochem* **89**, 3-9 (1978).
652 20. Chatre, L. & Ricchetti, M. Prevalent coordination of mitochondrial DNA transcription
653 and initiation of replication with the cell cycle. *Nucleic Acids Res* **41**, 3068-78 (2013).
654 21. Sasaki, T., Sato, Y., Higashiyama, T. & Sasaki, N. Live imaging reveals the dynamics
655 and regulation of mitochondrial nucleoids during the cell cycle in Fucci2-HeLa cells.
656 *Sci Rep* **7**, 11257 (2017).
657 22. O'Hara, R. et al. Quantitative mitochondrial DNA copy number determination using
658 droplet digital PCR with single-cell resolution. Genome Res 29, 1878-1888 (2019).
659 23. Coller, H.A. *et al.* High frequency of homoplasmic mitochondrial DNA mutations in
660 human tumors can be explained without selection. *Nature Genetics* **28**, 147–150 (2001).
661 24. Chen, B. et al. Tumorigenesis as the Paradigm of Quasi-neutral Molecular Evolution.
662 Mol Biol Evol 36, 1430-1441 (2019).
663 25. Lee, C., Yen, K. & Cohen, P. Humanin: a harbinger of mitochondrial-derived peptides?
664 Trends Endocrinol Metab 24, 222-8 (2013).
665 26. Lott, M.T. et al. mtDNA Variation and Analysis Using Mitomap and Mitomaster. Curr
666 Protoc Bioinformatics 44, 1 23 1-26 (2013).
667 27. Park, J. et al. Mitochondrial gene mutations in pediatric septic shock. Pediatr Res (2021).
668 28. Starr, T.K., Jameson, S.C. & Hogquist, K.A. Positive and negative selection of T cells.
669 Annu Rev Immunol 21, 139-76 (2003).
670 29. Buck, M.D. et al. Mitochondrial Dynamics Controls T Cell Fate through Metabolic
671 Programming. Cell 166, 63-76 (2016).
672 30. Chao, T., Wang, H. & Ho, P.C. Mitochondrial Control and Guidance of Cellular
673 Activities of T Cells. Front Immunol 8, 473 (2017).
674 31. Maciver, N.J., Michalek, R.D. & Rathmell, J.C. Metabolic regulation of T lymphocytes.
675 *Annual Review of Immunology* **31**, 259–283 (2013).
676 32. Sena, L.A. *et al.* Mitochondria Are Required for Antigen-Specific T Cell Activation
677 through Reactive Oxygen Species Signaling. *Immunity* **38**, 225–236 (2013).
678 33. West, A.P. & Shadel, G.S. Mitochondrial DNA in innate immune responses and
679 inflammatory pathology. *Nat Rev Immunol* **17**, 363-375 (2017).
680 34. Angajala, A. *et al.* Diverse Roles of Mitochondria in Immune Responses: Novel Insights
681 Into Immuno-Metabolism. *Front Immunol* **9**, 1605 (2018).
682 35. Lee, J.W. et al. Integrated analysis of plasma and single immune cells uncovers
683 metabolic changes in individuals with COVID-19. Nat Biotechnol (2021).
684 36. Waters, L.R., Ahsan, F.M., Wolf, D.M., Shirihai, O. & Teitell, M.A. Initial B Cell
685 Activation Induces Metabolic Reprogramming and Mitochondrial Remodeling.
686 iScience 5, 99-109 (2018).
687 37. Satpathy, A.T. et al. Transcript-indexed ATAC-seq for precision immune profiling.
688 Nature Medicine 24, 1 (2018).
689 38. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell
690 transcriptomic data across different conditions, technologies, and species. Nat
691 Biotechnol 36, 411-420 (2018).
692 39. Granja, J.M. et al. Single-cell multiomic analysis identifies regulatory programs in
693 mixed-phenotype acute leukemia. Nat Biotechnol 37, 1458-1465 (2019).
694 40. Miragaia, R.J. et al. Single-Cell Transcriptomics of Regulatory T Cells Reveals
695 Trajectories of Tissue Adaptation. Immunity 50, 493-504 e7 (2019).

41. Granja, J.M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet 53, 403-411 (2021).
42. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137 (2008).
43. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-303 (2010).
44. Koboldt, D.C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22, 568-76 (2012).
45. Buenrostro, J.D. et al. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. Cell 173, 1535–1548.e16 (2018).
46. van Dijk, D. et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell 174, 716-729 e27 (2018).
47. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 32, 381-386 (2014).
48. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 14, 979-982 (2017).
49. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature 566, 496-502 (2019).
50. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352, 189-96 (2016).
51. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21 (2013).
52. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164 (2010).
53. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24, 1586-91 (2007).
54. Wilton, P.R., Zaidi, A., Makova, K. & Nielsen, R. A Population Phylogenetic View of Mitochondrial Heteroplasmy. Genetics 208, 1261-1274 (2018).
55. Cabrera, V.M. Human molecular evolutionary rate, time dependency and transient polymorphism effects viewed through ancient and modern mitochondrial DNA genomes. Sci Rep 11, 5036 (2021).
56. Tavaré, S., Balding, D.J., Griffiths, R.C. & Donnelly, P. Inferring coalescence times from DNA sequence data. Genetics 145, 505-18 (1997).
57. Beaumont, M.A., Zhang, W. & Balding, D.J. Approximate Bayesian computation in population genetics. Genetics 162, 2025-35 (2002).
58. Zhao, J., Siegmund, K.D., Shibata, D. & Marjoram, P. Ancestral inference in tumors: how much can we know? J Theor Biol 359, 136-45 (2014).
59. Csilléry, K., François, O. & Blum, M.G. abc: an R package for approximate Bayesian computation (ABC). Methods in ecology and evolution 3, 475-479 (2012).