nature genetics

Article

https://doi.org/10.1038/s41588-023-01433-8

Single-cell multi-omics of mitochondrial DNA disorders reveals dynamics of purifying selection across human immune cells

Received: 21 January 2022

Accepted: 24 May 2023

Published online: 29 June 2023

Check for updates

Caleb A. Lareau[1,2,3,4,5] ✉, Sonia M. Dubois[5,21], Frank A. Buquicchio[1,21], Yu-Hsin Hsieh[6,21], Kopal Garg[4,5], Pauline Kautz[6,7,8], Lena Nitsch[6,7,9], Samantha D. Praktiknjo[6,7], Patrick Maschmeyer[6,7], Jeffrey M. Verboon[4,5], Jacob C. Gutierrez[1], Yajie Yin[1], Evgenij Fiskin[4], Wendy Luo[4], Eleni P. Mimitou[10,18], Christoph Muus[4,11], Rhea Malhotra[4], Sumit Parikh[12], Mark D. Fleming[13], Lena Oevermann[14], Johannes Schulte[14], Cornelia Eckert[14], Anshul Kundaje[3,15], Peter Smibert[10,19], Santosha A. Vardhana[16], Ansuman T. Satpathy[1,2], Aviv Regev[4,17,20] ✉, Vijay G. Sankaran[4,5] ✉, Suneet Agarwal[5] ✉ & Leif S. Ludwig[4,5,6,7] ✉

Pathogenic mutations in mitochondrial DNA (mtDNA) compromise cellular metabolism, contributing to cellular heterogeneity and disease. Diverse mutations are associated with diverse clinical phenotypes, suggesting distinct organ- and cell-type-specific metabolic vulnerabilities. Here we establish a multi-omics approach to quantify deletions in mtDNA alongside cell state features in single cells derived from six patients across the phenotypic spectrum of single large-scale mtDNA deletions (SLSMDs). By profiling 206,663 cells, we reveal the dynamics of pathogenic mtDNA deletion heteroplasmy consistent with purifying selection and distinct metabolic vulnerabilities across T-cell states in vivo and validate these observations in vitro. By extending analyses to hematopoietic and erythroid progenitors, we reveal mtDNA dynamics and cell-type-specific gene regulatory adaptations, demonstrating the context-dependence of perturbing mitochondrial genomic integrity. Collectively, we report pathogenic mtDNA heteroplasmy dynamics of individual blood and immune cells across lineages, demonstrating the power of single-cell multi-omics for revealing fundamental properties of mitochondrial genetics.

Mitochondria are complex organelles essential for metabolism and carry their own genome. Characterized by a high mutation rate, cell cycle-independent (relaxed) replication and variable copy number, mitochondrial DNA (mtDNA) possesses distinct genetic properties compared to nuclear DNA. In human cells, mitochondrial genomes are present in high copy numbers (100–1,000s), and mutations in mtDNA may vary in their level of heteroplasmy (proportion of mitochondrial genomes carrying a specific variant) in and across individual cells[1,2]. Notably, mtDNA-related disorders affect approximately 1 in 4,300 individuals, many of which present with heterogeneous phenotypes, cell-type-specific defects and variable severity that may correlate with heteroplasmy of pathogenic mutations[3]. Similarly, the age-related

A full list of affiliations appears at the end of the paper. ✉e-mail: clareau@stanford.edu; aviv.regev.sc@gmail.com; sankaran@broadinstitute.org; suneet.agarwal@childrens.harvard.edu; leif.ludwig@bih-charite.de

accumulation of somatically mutated mtDNA molecules in human cells and tissues may contribute to a variety of complex human diseases[1,2,4,5]. While germline single nucleotide variants (SNV) have been studied in human tissues, the effects of a major class of mutations and large mtDNA deletions have been examined to a lesser extent. Notably, single large-scale mtDNA deletions (SLSMDs) have been implicated in a continuum of congenital disorders, including Pearson syndrome (PS), Kearns–Sayre syndrome (KSS) and chronic progressive external ophthalmoplegia (CPEO)[6].

Recently, we and others have demonstrated the utility of single-cell genomics for mtDNA genotyping in combination with cellular state characterization[7,8]. The droplet-based mitochondrial single-cell assay for transposase accessible chromatin by sequencing (mtscATAC-seq) technique enables the scalable, concomitant profiling of accessible chromatin and mtDNA[9,10]. Further innovations enable additional single-cell measurements alongside chromatin accessibility and mtDNA genotyping, including antibody-based quantification of protein expression (ATAC with selected antigen profiling by sequencing (ASAP-seq)) and gene expression (DOGMA-seq)[11]. The application of these approaches has revealed the high prevalence of somatic mtDNA mutations, many of which are stably propagated and facilitate clonal/lineage tracing studies[7–9,11]. Moreover, these assays facilitate the study of pathogenic mtDNA variants associated with human disease. In patients with mitochondrial encephalomyopathy lactic acidosis and stroke-like episodes (MELAS) caused by the m.3243A > G mutation, we demonstrated a previously unappreciated purifying selection against pathogenic mtDNA in particular T cells, suggesting a link between heteroplasmy and cell state[10].

Here we use a series of multi-omics single-cell approaches and introduce mgatk-del, a computational approach to assess heteroplasmy of mtDNA deletions with high sensitivity and specificity, in single cells from patients with SLSMD. By examining primary hematopoietic cells in the peripheral blood and the bone marrow (n = 206,663 primary cell profiles), we reveal the distribution of pathogenic mtDNA deletions in hematopoietic lineages and its depletion or persistence in specific cell types indicative of distinct metabolic vulnerabilities. We identify context-dependent alterations in cell state as assessed by transcriptional, accessible chromatin, and protein expression profiling. Collectively, this study underscores the power of single-cell multi-omics to interrogate congenital mitochondriopathies, revealing metabolic requirements, vulnerabilities and cell-type-specific means of compensation.

Results

Single-cell quantification of mtDNA deletions

We have previously demonstrated that mtscATAC-seq yields relatively uniform coverage across the mitochondrial genome and can robustly quantify pathogenic SNVs in single cells[9,10]. Here we sought to assess this approach for detecting and quantifying large mtDNA deletions that underlie PS and related SLSMD. These large mtDNA deletions have been hypothesized to occur due to strand displacement errors in mtDNA replication between the heavy (O_H) and light (O_L) origins of replication and occur very early in development or oogenesis (Fig. 1a)[12,13]. To examine these deletions in single-cell data, we conducted mixing experiments by pooling in vitro cultured fibroblasts derived from two healthy donors and three patients with PS carrying three distinct mtDNA deletions for mtscATAC-seq (Fig. 1b). Following sequencing, cells from each donor were demultiplexed using private SNVs (Fig. 1b; Methods). Pseudobulk summaries of high-quality cells per donor revealed distinct dips in coverage along the mtDNA genome corresponding to the specific deletions at variable levels of heteroplasmy (Fig. 1c).

Although the software has been developed to analyze mtDNA deletions in bulk sequencing data[14,15], these workflows do not ensure valid estimation of deletion heteroplasmy, particularly in lower-coverage libraries such as individual cells, and do not readily scale to thousands of cells from a mtscATAC-seq library. Thus, we developed a computational approach, mgatk-del, that uses aligned mtDNA sequencing reads that result from the CellRanger-ATAC preprocessing (Extended Data Fig. 1a). To achieve precise heteroplasmy estimation, we reasoned that base-resolution breakpoints in sequencing reads (encoded as soft-clips in the alignments) could be used to infer deletion junctions, which could be corroborated with per-read secondary alignments reported from BWA (Extended Data Fig. 1b,c). Deletion heteroplasmy could then be estimated as a ratio of reads supporting or contradicting a deleted junction sequence. To benchmark this approach, we evaluated heteroplasmy estimation as a function of two hyperparameters using grid-searching simulated synthetic data (Extended Data Fig. 1c–f; Methods), yielding a method to accurately estimate heteroplasmy for each deletion.

Having established the computational approach, we quantified single-cell heteroplasmy for all three investigated pathogenic deletions. Following donor demultiplexing, our clipped-read heteroplasmy estimates revealed variation in deletion heteroplasmy across the population of cells (Fig. 1d), consistent with our previous observations of heterogeneity in cells derived from patients with mitochondriopathy caused by SNVs[9,10] and exceeding the variation that could be explained from variable single-cell coverages under a null model (Extended Data Fig. 1g). Furthermore, nonzero heteroplasmy was highly specific for each PS cell line. Conversely, a coverage-based estimate of heteroplasmy (ratios of read depths within and outside the deleted region) showed greater nonspecific heteroplasmy at deletions discordant from the originating PS patient cells although both methods were overall concordant (Fig. 1e and Extended Data Fig. 1h–j; Methods). Together, our analyses demonstrate the ability of mgatk-del to map mtDNA deletions at base-pair resolution and quantify their heteroplasmy in single cells.

Purifying selection of mtDNA deletions in T cells

We then used mtscATAC-seq and mgatk-del to analyze mtDNA deletions and heteroplasmy in primary patient cells. We obtained peripheral blood mononuclear cells (PBMCs) from three cases, including a 7-year-old male with PS/KSS ('PT1'), a 4-year-old female with PS ('PT2') and a 4-year-old male with PS and chromosomal 7q deletion (del7q) myelodysplastic syndrome (MDS, 'PT3'). Each patient presented with a distinct SLSMD (Supplementary Table 1). PBMCs from all three patients were profiled using both mtscATAC-seq and 10× 3′ scRNA-seq to quantify heterogeneity of mtDNA deletion heteroplasmy, chromatin accessibility and transcriptional profiles (Fig. 2a). Application of mgatk-del revealed the base-resolution breakpoints corresponding to the deleted regions for each patient (without prior knowledge) and enabled the quantification of deletion heteroplasmy in single cells (Fig. 2b and Extended Data Fig. 2a). Among cells passing quality control (n = 15,064; mean 81.5× coverage), we observed marked variation of heteroplasmy in PBMCs, including hundreds of cells that had no detectable mtDNA deletion heteroplasmy (Fig. 2c).

As mtDNA genotypes are paired with single-cell chromatin accessibility data, we sought to examine heteroplasmy variability and dynamics as a function of cell state. We performed a dictionary-based reference mapping of all cells to a previously annotated atlas of PBMCs (Fig. 2d; Methods)[16]. Notably, our analyses revealed clusters of T cells consistently depleted of mutant mtDNA relative to other immune or T-cell populations across all three donors, including effector/memory CD8 T cells (CD8.TEM) and mucosal-associated invariant T cells (MAIT) that could not be explained by variation in sequencing coverage (Fig. 2e-h, Extended Data Fig. 2b,c and Supplementary Table 2). These results are reminiscent of the previously described purifying selection of pathogenic mtDNA in T cells from patients with MELAS[10] but add nuance by revealing multiple subpopulations of affected T cells. To corroborate this inference, we performed the same dictionary-based reference annotation for MELAS cells. Indeed, MAIT and CD8.TEM

https://doi.org/10.1038/s41588-023-01433-8

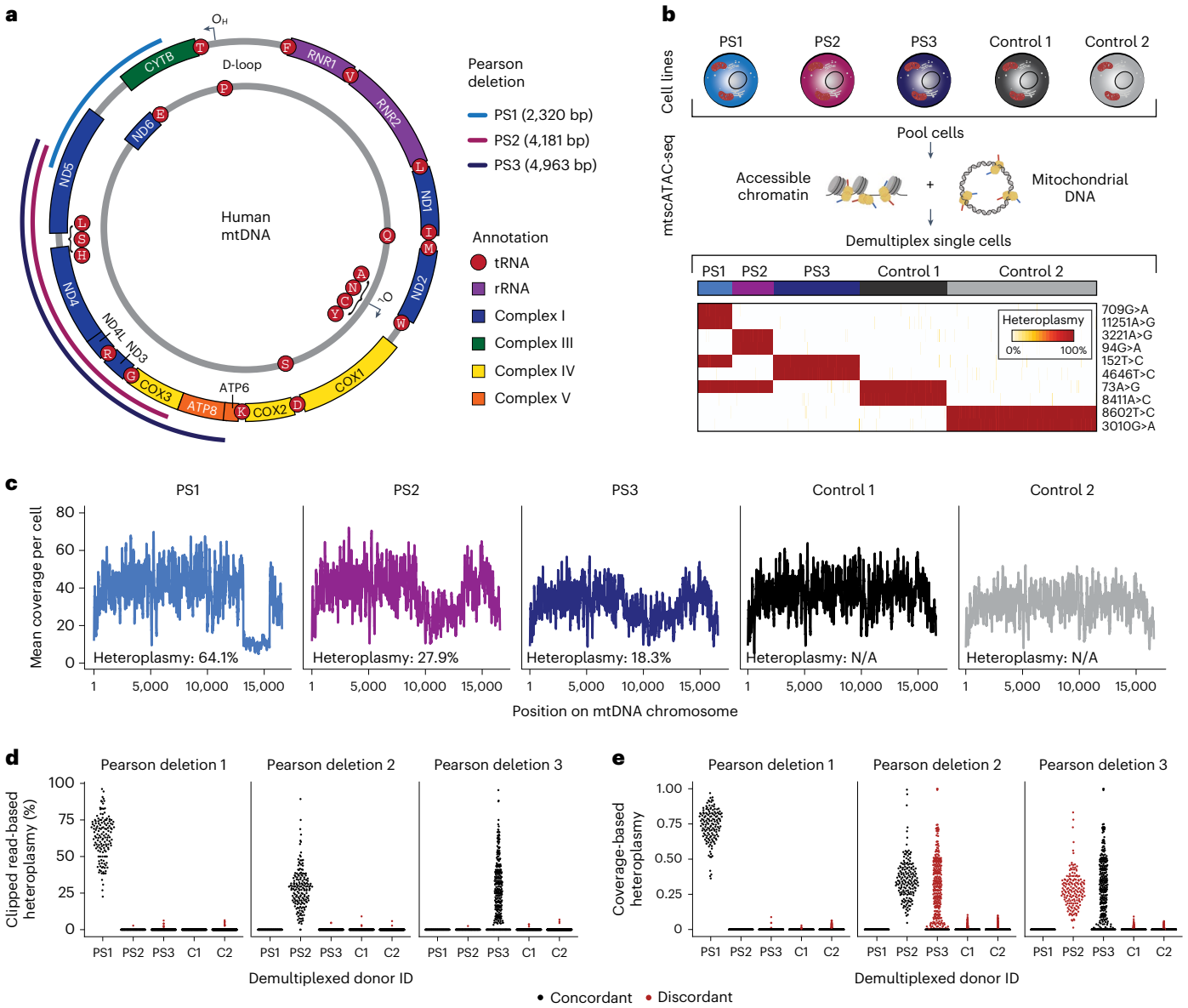


Fig. 1 | Identification and quantification of heteroplasmic pathogenic mtDNA deletions in single cells. a, Schematic of mtDNA in humans with PS-related deletions relevant for the cell lines examined in **b**. O_H and O_L represent the heavy and light chain origins of replication, respectively. PS1, PS2 and PS3 represent three different mtDNA deletions identified in three independent donors from which the cell lines were derived. Size and location of deletions are indicated. **b**, Summary of cell line mixing experiment and demultiplexing using mtDNA haplotype-derived SNVs in single cells. Heatmap depicts homoplasmic SNVs that facilitate the separation of cells from distinct donors. **c**, Mean coverage plots per cell across the mitochondrial genome for the demultiplexed donor cell identities. Drops in the coverage are indicative of large mtDNA deletions. **d**,**e**, Estimates of single-cell heteroplasmy using (**d**) clipped-read enumeration and (**e**) coverage-based approaches. Red dots represent false-positive heteroplasmy assessments from the deletion/donor pair ('discordant'). Each dot represents estimated heteroplasmy from a single cell for each respective deletion. C1, C2 = Control 1 and Control 2 as in **b** and **c**.

cells displayed reduced heteroplasmy of mutant mtDNA (Fig. 2i,j and Extended Data Fig. 2d,e). Overall, our results suggest that MAIT and CD8.TEM cells are both under specific selection pressures that are conserved between different classes of pathogenic mtDNA genotypes and diagnoses, thus resulting in a refined model of the degree of purifying selection in immune cells in the context of congenital mitochondrial disease (Fig. 2k).

In vitro T-cell models corroborate purifying selection

We sought to investigate pathogenic mtDNA dynamics during in vitro activation and differentiation of T cells from PS donors (Fig. 3a). We observed retention of naive-like marker CD45RA, reduced expansion of T cells, particularly $CD8^+$ T cells, relative to healthy controls, consistent with the stronger selective pressure of CD8.TEM cells observed in vivo (Fig. 3b–d and Extended Data Fig. 3a). Conversely, parallel cultures of healthy adult and pediatric T cells did not show impairment either in total proliferation or in the ratio of $CD8^+$:$CD4^+$ T cells (Fig. 3d and Extended Data Fig. 3b). To link cell surface markers to pathogenic mtDNA heteroplasmy, we performed proteogenomic characterization via ASAP-seq at days 14 and 21 of culture, observing the percentage of cells with zero heteroplasmy increasing to 75%, compared to 23% in ex vivo PBMCs (Fig. 3e). Unsupervised dimensionality reduction and cell state annotation revealed heteroplasmy to be mostly restricted to naive-like $CD45RA^+$ and Th_{17}-like T cells (Fig. 3f–h and Extended Data Fig. 3c). In contrast, $CD4^+$ and $CD8^+$ effector-like T cells (marked by $CD45RO^+$) had mostly selected against pathogenic mtDNA,

https://doi.org/10.1038/s41588-023-01433-8

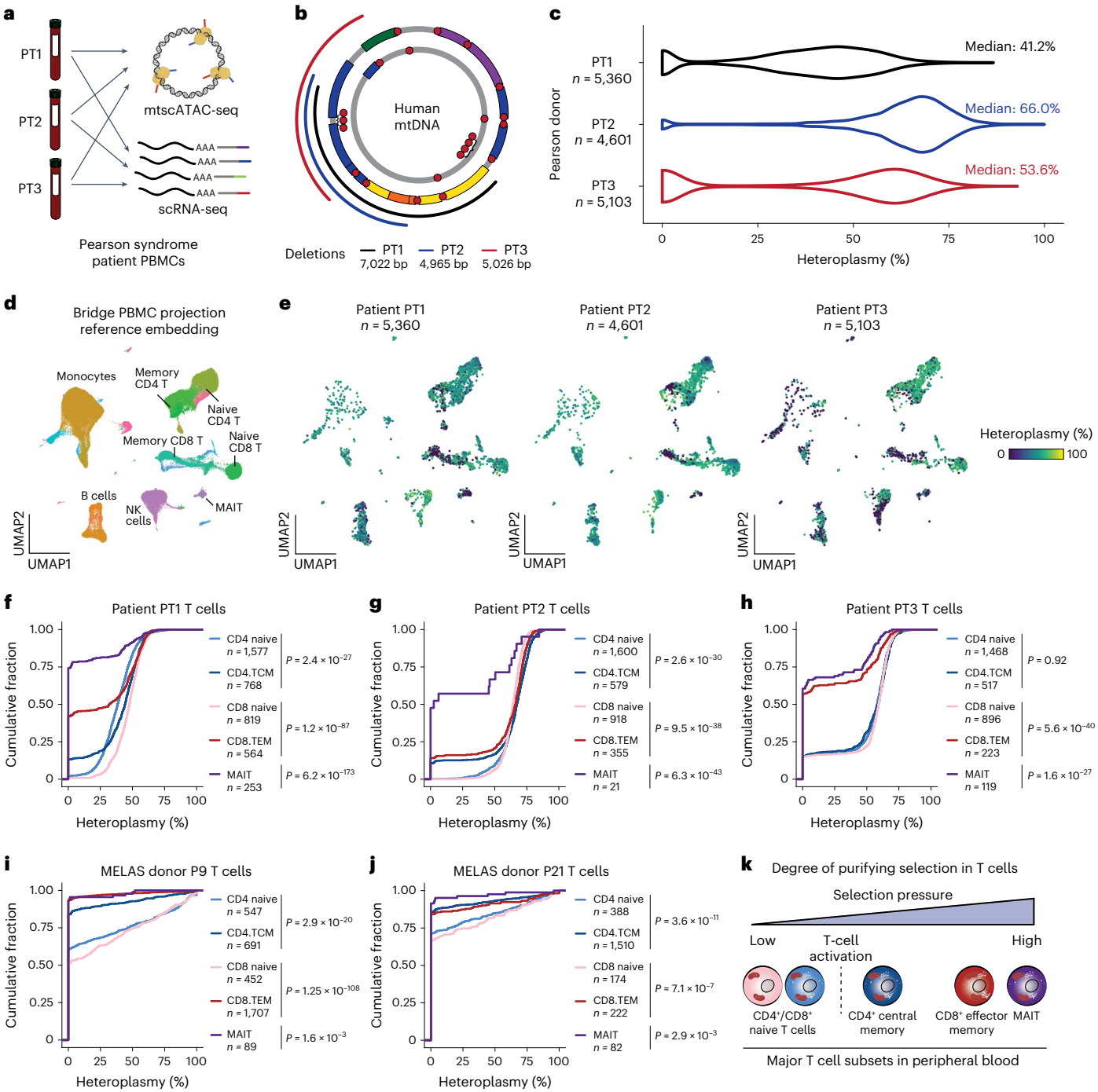


Fig. 2 | Purifying selection against pathogenic mtDNA deletions in peripheral blood MAIT and CD8 T cells in PS. a, Schematic of single-cell genomics data generation. PBMCs from three patients (PT1, PT2 and PT3) with PS were collected and processed via scRNA-seq and mtscATAC-seq. **b**, Depiction of mtDNA deletions from three investigated patients with PS as determined by mgatk-del. Location and size of deletions are indicated. **c**, Violin plots of single-cell heteroplasmy across indicated patients and respective mtDNA deletions are indicated in **b**. Median heteroplasmy (%) and profiled cell numbers are indicated for each patient. **d**, Reduced dimensionality projection and joint clustering of PBMCs from three patients with PS and one healthy control are shown. Major cell types and clusters are annotated in the same color. **e**, Reduced dimensionality projection as in **d** split by patient with PS and colored by respective mtDNA deletion heteroplasmy. **f**–**h**, Cumulative distribution plots of heteroplasmy stratified by T-cell subset derived from PT1, PT2 and PT3. Each comparison is a two-sided binomial test for the proportion of 0% cells comparing CD4.TCM to CD4 naive, CD8.TEM to CD8 naive and MAIT to other T-cell subsets. **i**,**j**, Purifying selection of the m.3243A>G allele in individuals with MELAS as previously reported[10]. The cell annotations and statistical tests are the same as in **f**–**h** but for the m.3243A>G allele. **k**, A refined model of purifying selection in T cells with the relative ordering of cells based on the proportion and frequency of 0% heteroplasmic cells observed in these donors between the two disease cohorts. DC, dendritic cells; pDCs, plasmacytoid dendritic cells; NK, natural killer cells.

suggesting that mitochondrial genetic integrity is essential to acquire these T-cell states. Identical observations were made upon extension of culture to day 21, which further revealed clonality of expanded T-cell populations as indicated by somatic mtDNA mutations (for example, m.12631T>C and m.4225A>G; Extended Data Fig. 3d–h). Culture of T cells from patient with PS (PT1) replicated these findings, including

https://doi.org/10.1038/s41588-023-01433-8

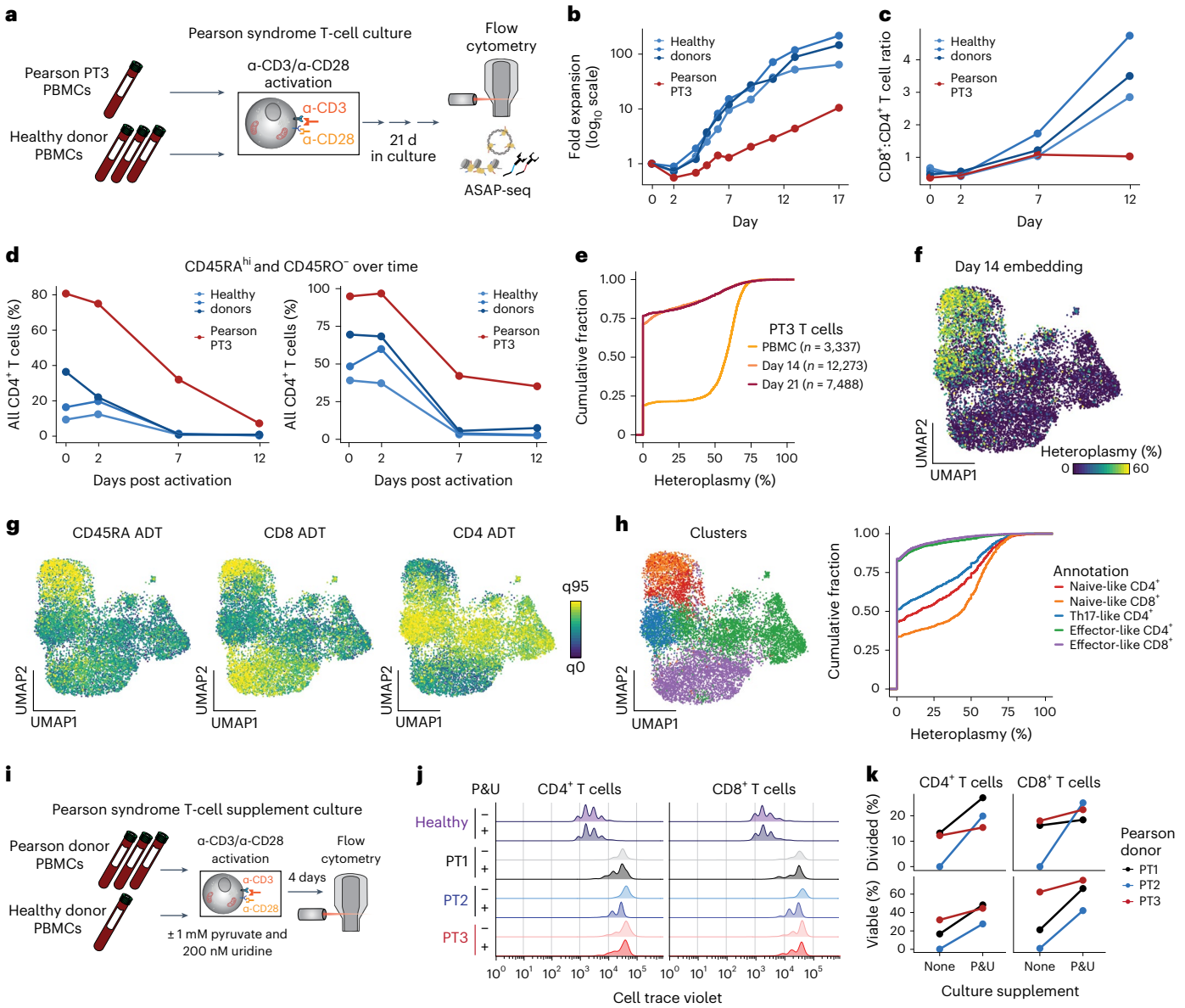


Fig. 3 | Altered CD8⁺ T cell expansion and purifying selection of pathogenic mtDNA deletions in vitro. a, Schematic of experimental design. PBMC-derived T cells were cultured and activated via α-CD3 and α-CD28 for ~3 weeks. ASAP-seq and flow cytometry measures were collected longitudinally. **b**, Fold expansion for all cells in culture for the healthy donors (blue) compared to the Pearson donor (red). **c**, Dynamics of CD8⁺ to CD4⁺ T-cell ratios in culture over time, implicating deficient CD8⁺ expansion of PS cells. **d**, Percentage of CD45RA^{hi}/CD45RO⁻ among CD4⁺ (left) and CD8⁺ (right) cells from the four donors over 12 d. **e**, Per-sample distribution of heteroplasmy of T cells from PBMCs, after 14 d in culture, and after 21 d in culture demonstrating most selection to have occurred by day 14 of culture relative to PBMCs. **f**, Day 14 ASAP-seq embedding annotated by heteroplasmy (%). **g**, The same embedding in **f** but colored by ADTs for CD45RA, CD8 and CD4, allowing for annotation of cluster-specific cell states. **h**, Cell state clusters of the day 14 ASAP-seq embedding. Per-cluster heteroplasmy is depicted via a continuous distribution plot. **i**, Schematic of in vitro culture experiment with and without 1 mM pyruvate and 200 nM uridine (P&U) supplementation. **j**, Cell trace violet plots showing cell division traces of a healthy donor and three donors with PS stratified by CD4⁺ or CD8⁺ T cells. **k**, All comparisons between none and P&U were not significantly different at an α value of 0.05 using a Student's paired t-test.

limited expansion, reduced CD8⁺:CD4⁺ ratio and depletion of heteroplasmy specifically in effector-like CD8⁺ T cells (Extended Data Fig. 3i–k). Finally, we hypothesized that the fitness deficit of T cells in vitro may be mitigated through the supplementation of pyruvate (to accept electrons instead of oxygen) and uridine (to enable pyrimidine synthesis in the absence of DHODH activity) in the culture media[17,18]. To explore this, we repeated the T-cell expansion cultures with and without pyruvate and uridine (P&U) and assessed proliferation and viability via flow cytometry (Fig. 3i; Methods). Indeed, after 4 d of culture, we observed an increase in viability and proliferation (Fig. 3j,k), suggesting that restoring OXPHOS function may partially restore T-cell function. In total, our in vivo and in vitro results suggest that pathogenic mtDNA deletions compromise the proliferation and differentiation of naive to effector T-cell states, with a particular vulnerability of the CD8.TEM lineage.

Deletion heteroplasmy in adults with SLSMD

While PS may be lethal early in life, individuals with CPEO and KSS, other diseases caused by SLSMD, more commonly live into adulthood. To study the dynamics of purifying selection of SLSMD in the immune

https://doi.org/10.1038/s41588-023-01433-8

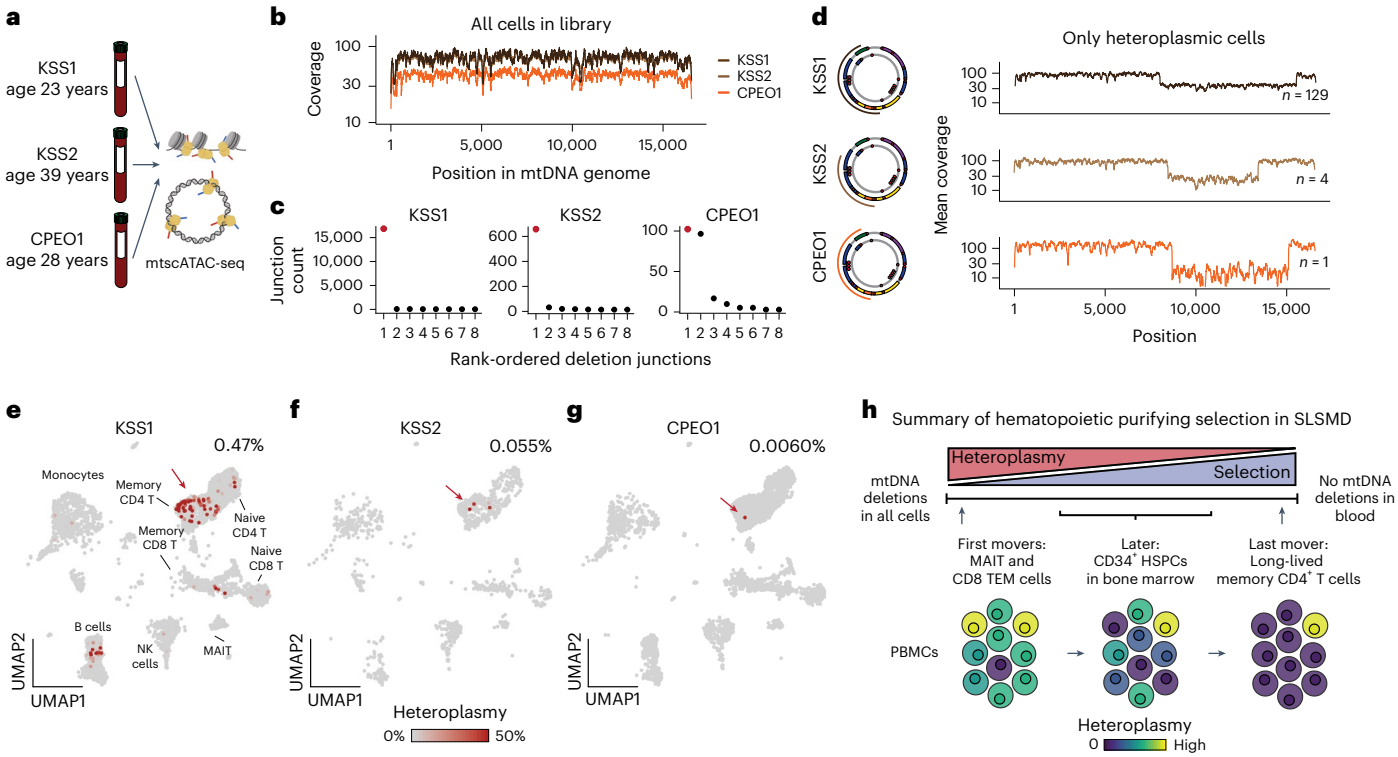


Fig. 4 | Purifying selection and retention of pathogenic mtDNA heteroplasmy across the peripheral blood in adults with SLSMD. a, Three donors with either KSS or CPEO were profiled using mtscATAC-seq. **b**, Pseudobulk coverage of three donors across the mtDNA genome. **c**, Results of deletion calling from mgatk-del find. Each dot represents a pair of junctions. The junction pair marked in red represents the pathogenic deletion. **d**, Validation of deletions in cells from pseudobulk coverage estimates in cells with nonzero heteroplasmy. The coordinates of the deletions are noted in the miniature diagrams of the mtDNA deletions. **e**–**g**, Reference projection of cell states from mtscATAC-seq (as in Fig. 2d) with heteroplasmy annotated for the indicated mtDNA deletions. The red arrow points to the memory CD4$^+$ T-cell compartment. Only a single cell harbored the deletion for the patient with CPEO (heteroplasmy = 68.8%; 11 reads supporting the deletion; 5 reads supporting wild-type mtDNA). **h**, Schematic of the lifetime dynamics of purifying selection against pathogenic mtDNA across PS (as in Fig. 2k) and other SLSMDs.

system after decades of life, we profiled three donors aged 23–39 years with mtscATAC-seq (Fig. 4a). While an initial examination of mtDNA coverage did not reveal obvious deletions (Fig. 4b), application of clipped-read analysis via mgatk-del revealed specific large deletions between 4,965 bp and 7,514 bp in all three donors (Fig. 4c,d), including deletions with 0.0060% pseudobulk heteroplasmy. Here all reads supporting the deletion came from the same cell (68.8% heteroplasmy; 11 unique reads supporting the deletion), whereas other software[19,20] failed to uncover this rare event. For all three patients, we observed that these heteroplasmic cells were enriched in the CD4$^+$ central memory and regulatory T-cell compartments (53.0% of nonzero heteroplasmy cells versus 25.0% of all cells; Fisher's exact $P = 9.9 \times 10^{-7}$; Fig. 4e–g). Thus, analysis of adult patients with SLSMD provides a lens into the lifetime dynamics of pathogenic mtDNA deletions, indicating CD8. TEM and MAIT cells are initially most sensitive to the selection, which over time extends to CD34$^+$ hematopoietic stem and progenitor cells (HSPCs) and all descending cells, leaving long-lived CD4$^+$ cells with residual heteroplasmy (Fig. 4h).

Identification of mosaic del7q cells

For PT3 with PS, the clinical evaluation revealed a mosaic del7q, a chromosomal abnormality consistent with the development of MDS on the backdrop of a congenital bone marrow failure syndrome (Fig. 5a)[21]. Notably, the acquisition of monosomy 7 has recently been reported in a case of PS[22]. We applied mtscATAC-seq to PT3 bone marrow mononuclear cells (BMMNCs) with and without CD34$^+$ enrichment (Extended Data Fig. 4a) to analyze the association of mtDNA deletion heteroplasmy and the nuclear del7q abnormality at single-cell resolution. To assess the distribution of del7q cells, we first examined the abundance of fragments overlapping the deleted region, which revealed a clear multimodal distribution, that could be classified using a Gaussian mixture model (Fig. 5b and Extended Data Fig. 4b–d; Methods). As del7q was most abundant in CD34$^+$ HSPCs, we examined the association between del7q and mtDNA heteroplasmy in these cells. Notably, we observed a striking association where del7q cells had substantially higher levels of mutant mtDNA, suggesting the acquisition of del7q in HSPCs with the most compromised mitochondrial function (Fig. 5c and Extended Data Fig. 4e).

To refine our analysis, we projected CD34$^+$ PT3 data onto a healthy donor reference of sorted CD34$^+$ cells to define the continuous differentiation trajectory of these progenitors via patterns of chromatin accessibility (Fig. 5d; Methods)[23,24]. Relative to healthy control cells, PT3 displayed a stark depletion of cells annotated as hematopoietic stem cells (HSCs) and multipotent progenitors (MPPs) as well as an enrichment of granulocyte–monocyte progenitors (GMP) and monocytes in peripheral blood, resulting in a markedly different estimated composition of the entire HSPC compartment (Fig. 5e and Extended Data Fig. 4f–h). We observed the pronounced presence of del7q in PT3 GMPs and multipotent erythroid progenitors (MEP), consistent with the MDS phenotype (Fig. 5f). Notably, the PT3 common lymphoid progenitor (CLP) population was mostly wild-type for chr7 and depleted of pathogenic mtDNA (Fig. 5g,h). Together, these analyses reveal the complexity of lineage commitment and differentiation in the presence of pathogenic mtDNA deletion heteroplasmy and the onset of MDS within the early hematopoietic progenitor compartment of the patient with PS.

https://doi.org/10.1038/s41588-023-01433-8

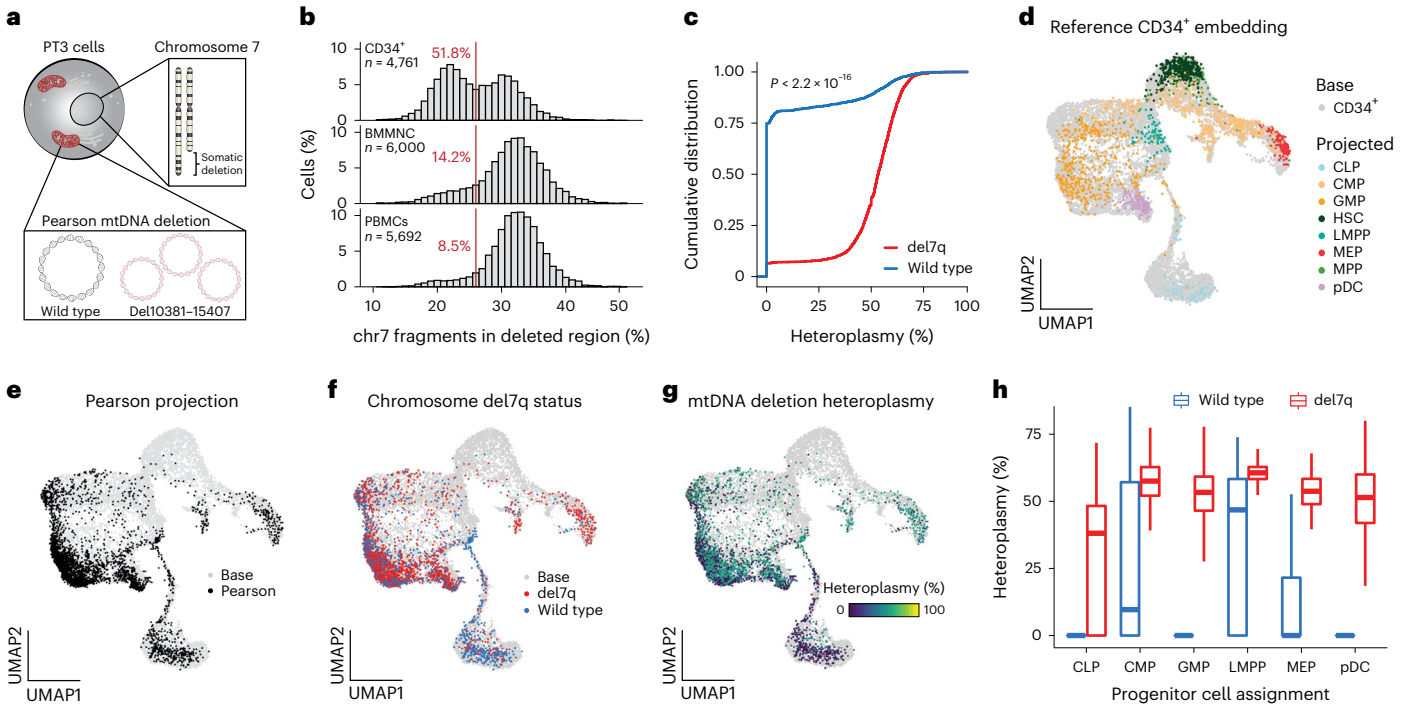


Fig. 5 | Myelodysplastic cells resolved by a chromosomal 7q deletion in CD34$^+$ HSPCs in a case of PS. a, Schematic depicting the del7q and mtDNA deletion of PT3 in the same cell. **b**, Histograms showing the percentage of fragments on chromosome 7 mapping to the somatically deleted region. A consistent cutoff of 26% with the percentage of cells below this threshold is indicated in red. *n*, cells assayed for each indicated population of CD34$^+$, BMMNC and PBMCs. **c**, Cumulative distribution curves of mtDNA heteroplasmy stratified by the del7q status per cell. Statistical test—two-sided Kolmogorov–Smirnov test. **d**, UMAP of a reference CD34$^+$-based embedding (base; gray) with sorted cell populations projected (color coded) onto the reference. All cells were derived from healthy donors. CLP, common lymphoid progenitor; CMP, common myeloid progenitor; GMP, granulocyte–monocyte progenitors; HSC, hematopoietic stem cell; LMPP, lymphoid primed multipotent progenitor; MEP, megakaryocyte erythroid progenitor; MPP, multipotent progenitor; pDC, plasmacytoid dendritic cell. **e**, Projection of PS CD34$^+$ cells onto the same base embedding as in **d**. **f**, Annotation of del7q status for each PS CD34$^+$ cell, indicating diploid (blue) and del7q (red) status. **g**, Annotation of single-cell mtDNA deletion heteroplasmy per PS CD34$^+$ cell. **h**, Heteroplasmy (%) stratified based on annotated CD34$^+$ progenitor cell state and by del7q ploidy status. Data from one biologically independent sample and experiment. Boxplots—center line, median; box limits, first and third quartiles; whiskers, 1.5× interquartile range.

Purifying selection in hematopoietic development

To further investigate the interplay of del7q and mtDNA deletion heteroplasmy dynamics across the hematopoietic compartment, we applied ASAP-seq to PT3-derived BMMNCs, including the profiling of 242 surface antigens, yielding 20,580 high-quality cells with quantification across four distinct modalities per cell (chromatin accessibility, nuclear chromosomal aberrations, mtDNA genotypes and surface protein abundance; Fig. 6a). We revealed variation in heteroplasmy across hematopoietic lineages with the proteogenomic measurements facilitating more highly resolved inferences of cell type/state (Fig. 6b–d and Extended Data Fig. 5a). Furthermore, our mixture model approach identified del7q to be primarily in myeloid and erythroid cells and largely absent in lymphocytes (Fig. 6e).

We performed integrative analyses to determine surface markers overexpressed on del7q compared to chr7 wild-type monocytic and erythroid cells (Fig. 6f). For both comparisons, we observed an enrichment of surface proteins CD15, CD56, CD64 and HLA-DR, all markers that have previously been reported to be upregulated in patients with MDS[25,26]. Unlike CD56, other markers of NK cells such as CD335 were not expressed on MDS-associated cells, but were present exclusively on NK cells (Extended Data Fig. 5b). In addition, we corroborated our observation of the depletion of phenotypic HSCs as revealed in the CD34$^+$ projection analysis (Fig. 5d,e). We compared the distribution of HSPC populations and protein markers in PS to healthy donor bone marrow ASAP-seq data[11] and were unable to detect CD34$^+$c-Kit$^+$CD71 cells in PT3 despite the clear presence of these cells in healthy BMMNCs (Extended Data Fig. 5c,d). Noting that HSCs do not express CD71, our integrated analysis confirms the apparent relative depletion of phenotypic HSCs in PT3, which may further present a consequence of the pathogenic mtDNA deletion and/or the MDS phenotype.

Then, we investigated two subpopulations of CD4$^+$ and CD8$^+$ T cells that were depleted of pathogenic mtDNA heteroplasmy. Differential gene accessibility revealed markers associated with recent thymic emigrants (RTEs)[27], including *ADAM23*, *IKZF2*, *TOX* and *ZNF462* (Fig. 6g–i and Extended Data Fig. 5e). Differential protein expression of the same populations showed a relative enrichment of CD21 and CD35 (Fig. 6j), which are both upregulated on RTEs[27]. Notably, we verified the presence of RTE-heteroplasmy-depleted cells in peripheral blood by reclustering PBMCs from PT3, which did not separate RTEs in our previous reference projection analyses (Fig. 2) and confirmed the purifying selection of CD8.TEM cells in the bone marrow (Extended Data Fig. 5f–j). Overall, integrating our observations of populations depleted of pathogenic mtDNA deletions— including CLPs in the CD34$^+$ compartment (Fig. 5d,g), subpopulations of CD4$^+$/CD8$^+$ RTEs and CD8.TEM cells in multiple hematopoietic compartments— suggests multiple distinct modes of purifying selection at distinct stages of lymphopoiesis (Fig. 6k).

As we have previously demonstrated somatic mtDNA SNVs to identify clonal subsets in hematopoietic populations of adults[7,9], we sought to determine the prevalence of these mutations at 4 years of age. Application of mgatk revealed 69 somatic mtDNA SNVs that were enriched in expected nucleotide substitution patterns[9,11], located

largely outside of the mtDNA deleted region and present in both del7q and wild-type cells (Extended Data Fig. 5k–m). For example, variants m.1719G>A and m.7836T>C refined clones within the del7q compartment, whereas variants m.12242A>G, m.14476G>A, m.5557T>A and m.13970G>A were predominantly found in cells with wild-type chr7 (Fig. 6l,m and Extended Data Fig. 5n). Notably, the m.5557T>A variant was observed in both $CD4^+$ and $CD8^+$ RTEs and in the myeloid compartment, suggesting that the HSPC carrying this variant is capable of multi-lineage output, whereas variants m.12242A>G and m.14476G>A were only identified in lymphoid cells. Together, our analyses indicate that the utility of mtDNA-based lineage tracing extends to pediatric patients and clonal myeloid disorders.

Selection dynamics in erythropoiesis

A hallmark feature of PS is severe macrocytic sideroblastic anemia, characterized by erythroblasts accumulating iron deposits around their mitochondria, which is frequently detectable in the neonatal period and often in the context of evolution to pancytopenia (a significant reduction in the number of all blood cells)[28–30]. Given this phenotype, we sought to understand the selection dynamics and altered gene expression programs underlying defective erythropoiesis. To realize this, we performed pseudotime trajectory inference for 1,511 cells from the BMMNC ASAP-seq data along the erythroid pseudotime axis (Fig. 7a; Methods). Our trajectory corroborated known cell state markers associated with early-to-late erythroid transitions from both surface protein and chromatin accessibility, including the *GATA1* and *TMCC2* loci (Fig. 7b)[31]. Along this axis, the proportion of cells harboring del7q and mtDNA deletion heteroplasmy was greatly reduced (Fig. 7c,d), suggesting selection during late erythropoiesis. Differentiating erythroid cells also showed high OXPHOS module scores relative to other BMMNCs, indicative of a high metabolic demand (Extended Data Fig. 6a). These findings support a model of the high vulnerability of the erythroid lineage and its altered output in PS, analogous to our observations of increased OXPHOS demand and resulting selection during T-cell proliferation and differentiation.

To further corroborate the observed in vivo phenotypes, we differentiated PT3 BMMNCs and healthy control cells in vitro in the presence of erythropoietin (EPO), collecting cells at days 6 and 12 of culture, before processing with scRNA-seq and mtscATAC-seq (Fig. 7e). Phenotypically, PS cells displayed poor proliferation and clear signs of impairment during erythroid differentiation as assessed by surface markers and cytology (Extended Data Fig. 6b–d). Assessment of mtDNA deletion heteroplasmy and del7q status revealed a relative increase in the proportion of del7q cells at days 6 and 12, with no notable selection against mtDNA heteroplasmy (Fig. 7c–h). These results, however, may reflect a low abundance of late-stage erythroblasts at the sampled time points.

Finally, we investigated the altered gene expression programs that may underlie the anemic phenotype in PS. We performed unsupervised dimension reduction of 28,783 high-quality cells, which revealed a trajectory of erythroid differentiation (Fig. 7i–k and Extended Data Fig. 6e; Methods). Differential gene expression and pathway enrichment analyses comparing erythroid cells from the PS donor to the healthy control were robust despite MDS-associated cells, suggesting alterations to be primarily attributable to the mtDNA deletion (Extended Data Fig. 6f). Most notably, we observed genes of the serine and glycine biosynthesis pathway, including *PHGDH*, *PSAT1*, *PSPH* and *SHMT2* to be upregulated in PS erythroblasts (Fig. 7l,m). Serine metabolism, reported to be altered in response to mitochondrial dysfunction, may aid in maintaining cellular one-carbon availability to provide essential precursors for synthesizing urines, phospholipids and the antioxidant glutathione (GSH), a scavenger of reactive oxygen species (ROS)[32–34]. Conversely, the heme biosynthesis pathway, including genes *UROS*, *CPOX*, *FECH*, *UROD*, *HMBS* and *PPOX*, and cholesterol biosynthesis pathways were substantially downregulated in PS (Fig. 7l,m and Extended Data Fig. 6g–i). In total, our multi-omic analyses nominate numerous perturbed genes and pathways, the deregulation of which likely contributes to the characteristic anemia in PS (Fig. 7n), and suggests avenues for additional functional follow-up.

Discussion

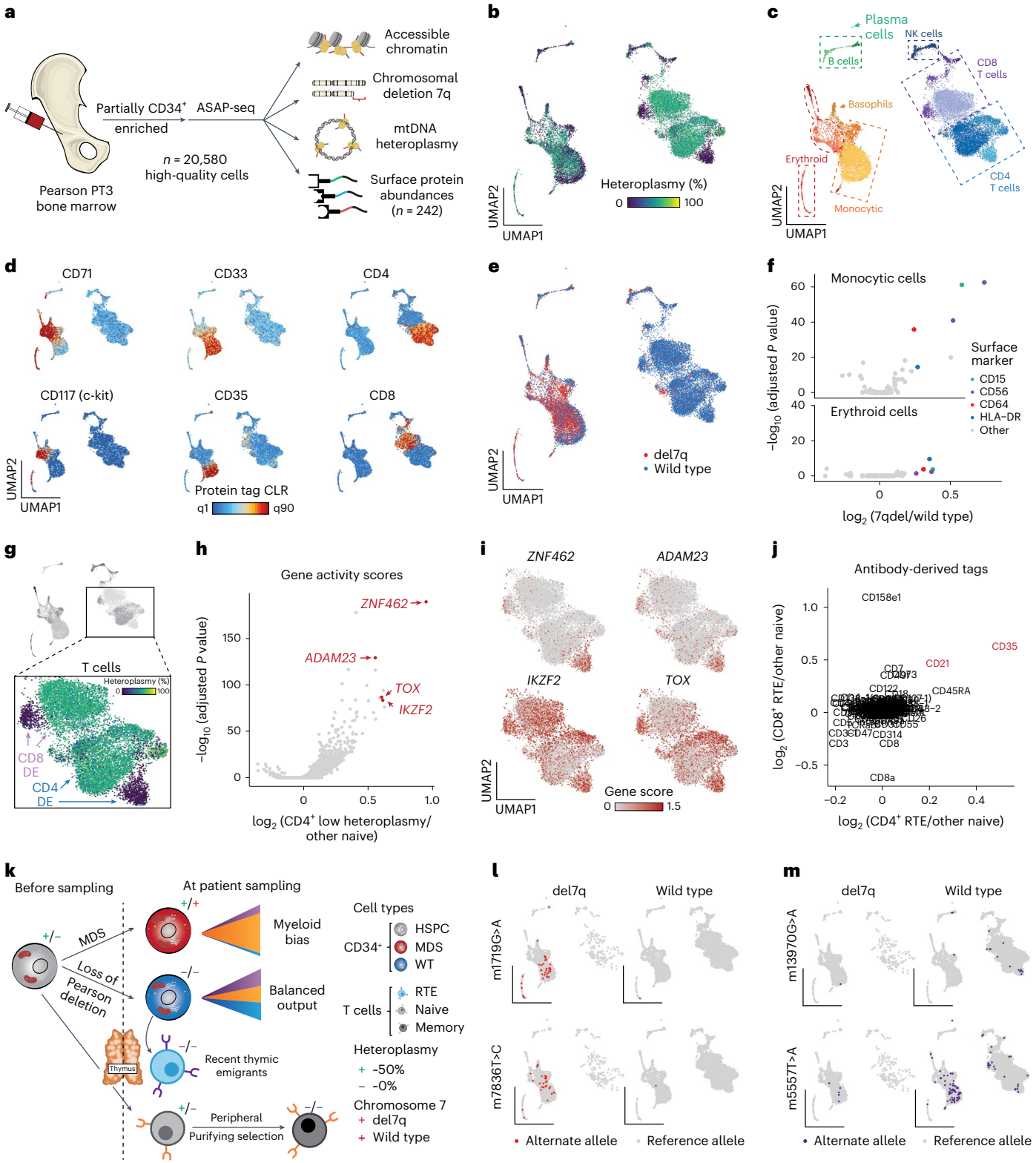
Multi-omic approaches provide complementary and orthogonal measurements to more holistically characterize the cellular circuits underlying perturbed cellular phenotypes in disease[35,36]. Here we charted genomic alterations across five modalities (that is, transcriptome, accessible chromatin, cell-surface markers, mtDNA genotypes and nuclear chromosomal aberrations) resulting from large mtDNA deletions across ~200,000 primary patient cells. In particular, we demonstrate how mgatk-del in conjunction with mtscATAC-seq[9,10], ASAP-seq[11] or DOGMA-seq (Supplementary Note)[11] enables the sensitive identification and quantification of large mtDNA deletions in single cells, alongside concomitant readouts of cell state. MtDNA copy number and heteroplasmy can be present at highly variable levels across a population of cells and cell states, thereby emphasizing the utility of our multi-omic advances. Our approach will aid in studying the phenotypic effects of somatically arising mtDNA mutations, which may be selected against in individuals with cancer, aging-related degenerative diseases and healthy tissues[1,2,37–39]. We note reports of the accumulation of mtDNA deletions in postmitotic cells, including in single muscle fibers[40] and neurons in Parkinson's disease[41], whereas SNVs appear more common in mitotic cells[42], indicative of distinct evolutionary pressures underlying these two classes of mutations. Future studies at the single mitochondrion level[43] or that study the full mtDNA or RNA molecule via long-read sequencing technologies[39] will complement our cell state inferences of heteroplasmy.

By studying pathogenic mtDNA deletion dynamics in vivo and in vitro, we observed multiple instances of purifying selection, including in MAIT, CD8.TEM and RTEs, indicative of metabolic vulnerabilities at distinct stages of T-cell maturation. Notably, we observed that CD8.

Fig. 6 | Multimodal characterization of PS BMMNCs with ASAP-seq. a, Schematic of ASAP-seq experiment from PS BMMNCs derived from PT3 with MDS. **b**, Dimensionality reduction and embedding for high-quality BMMNCs with heteroplasmy colored. **c**, The same embedding as in **b** is annotated by major cell type clusters. **d**, Selected lineage-defining surface protein markers are shown on the reduced dimension space as in **b** and **c**. **e**, Projection of annotated del7q status onto the UMAP space as in **b** and **c**. **f**, Volcano plots of differentially expressed protein surface markers inferred from antibody barcodes for del7q versus wild-type cells annotated as erythroid or monocytic from **c**. Markers with distinct colors were significantly upregulated in both comparisons (logFC > 0.1 and Wilcoxon test with Bonferroni correction $P < 0.01$). **g**, Schematic illustrating $CD4^+$ and $CD8^+$ T-cell clusters used for differential gene score expression (DE) analyses to identify markers distinguishing low-heteroplasmy cell populations. **h**, Volcano plot showing differential gene activity scores for comparison of $CD4^+$ T-cell clusters as illustrated in **g**. Genes in red (*ZNF462*, *ADAM23*, *IKZF2* and *TOX*) indicate marker genes for RTEs. **i**, Projected gene scores for indicated marker genes onto UMAP space as highlighted in **g**. **j**, Differentially expressed proteins for the comparisons of $CD4^+$ and $CD8^+$ T-cell populations as illustrated in **g**. CD21 and CD35, shown in red, are known surface markers for RTEs. A total of three markers (CD21, CD35 and CD45RA) were significantly upregulated in both $CD4^+$ and $CD8$ T^+ RTEs (logFC > 0.1 and Wilcoxon test with Bonferroni correction $P < 0.01$). **k**, Schematic of multifaceted clonal output and purifying selection in PT3 with PS and MDS. Major cell transitions are depicted as a function of 7qel status and mtDNA deletion heteroplasmy. **l**, Projection of somatic mtDNA mutations m.1719G>A and m.7836T>C enriched in cells carrying the del7q. **m**, Projection of somatic mtDNA mutations m.13970G>A and m.5557T>A enriched in wild-type cells (diploid for chr 7), including in RTEs.

https://doi.org/10.1038/s41588-023-01433-8

TEMs are more sensitive to selection compared to their effector CD4. TCM counterparts, corroborating distinct mitochondrial/metabolic demand following TCR-specific activation and conversion of CD8$^+$ T cells to effector/memory cells[44]. Such a model is consistent with mitochondrial respiratory capacity to be a critical regulator of CD8 memory T-cell development[45]. Furthermore, our approach also allowed the characterization of rare populations of cells, including MAIT, which experienced similar strong selection pressures to CD8.TEM, presumably requiring efficient mitochondrial capacity for cellular differentiation and function[46,47]. Together, these results provide insights into our emerging understanding of the effects of pathogenic mtDNA on the development of T-cell states and their life-long dynamics across the hematopoietic compartment (Fig. 4). Furthermore, through a rare case of PS with MDS, we show the del7q alteration to be restricted to cells carrying the mtDNA deletion, indicating that it arose in a mitochondrially/metabolically dysfunctional background and was



https://doi.org/10.1038/s41588-023-01433-8

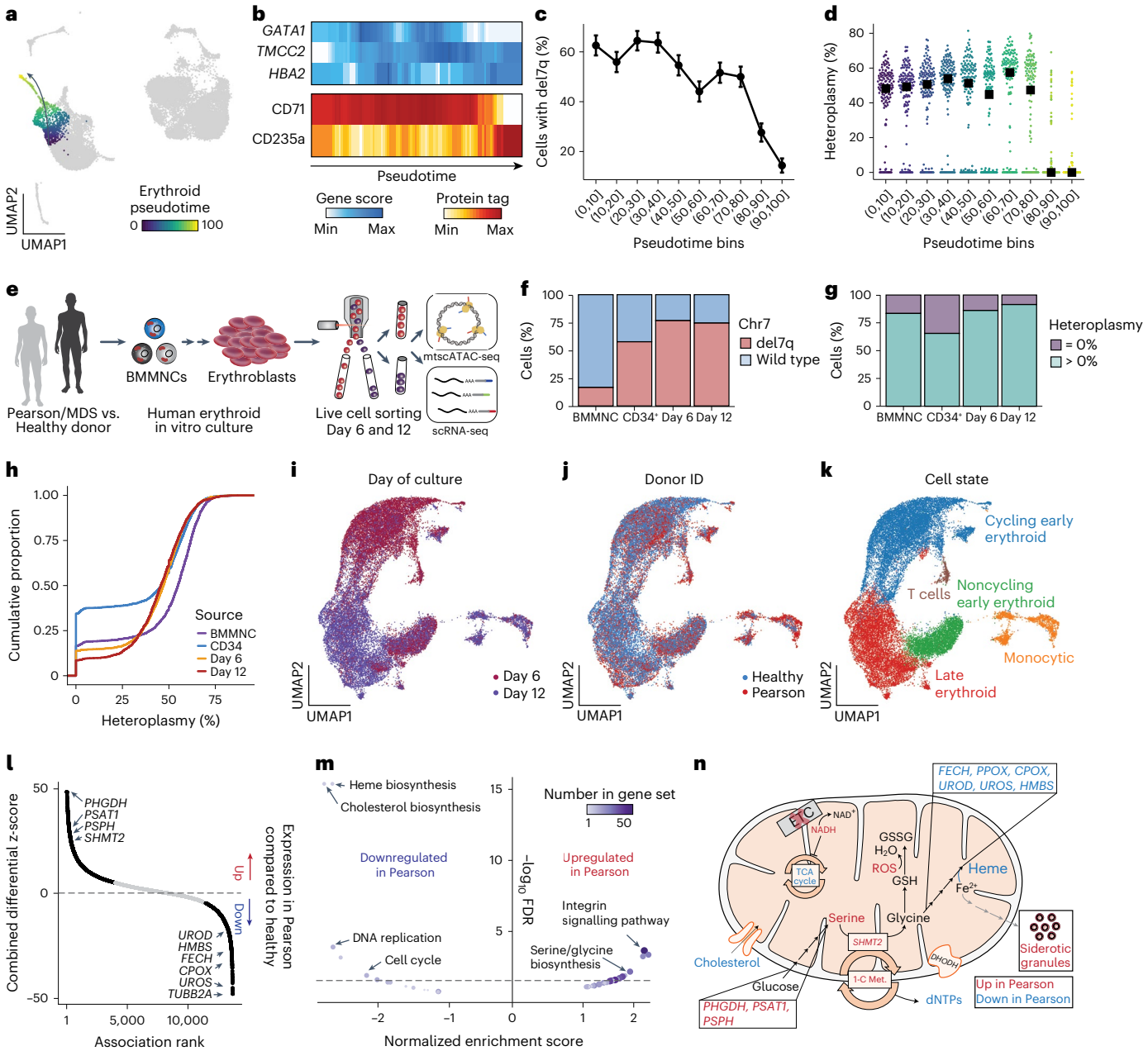


Fig. 7 | Altered erythroid differentiation and selection in PS. a, Erythroid pseudotime trajectory of in vivo ASAP-seq data. The color bar represents the annotated pseudotime for 1,511 cells, with the arrow orienting the inferred trajectory for the embedding in Fig. 6c. **b**, Summary of cell state features along erythroid pseudotime axis. **c**, Abundance of del7q cells along erythroid pseudotime axis. The mean of each pseudotime bin is noted ±s.e.m. **d**, Distribution of heteroplasmy across erythroid pseudotime bins. Each cell is plotted in color matching (**a**) with the per-bin median noted in black. **e**, Schematic of experimental design. BMMNCs were derived from PT3 with PS/MDS and healthy controls and differentiated toward erythroblasts in vitro. Patient and healthy cells were harvested on day 6 and day 12 and jointly processed via mtscATAC-seq or scRNA-seq. **f**, Stacked bar graph of cells annotated as wild-type or del7q across indicated cell populations, including at day 6 and day 12 of in vitro culture. **g**, Same as **f** but showing the proportion of cells with exactly 0% and >0% heteroplasmy of the mtDNA deletion as determined by mgatk-del. **h**, Cumulative distribution graphs of mtDNA deletion heteroplasmy across the indicated four cell populations. **i**–**k**, UMAP embedding of 28,783 high-quality cells profiled via scRNA-seq annotated by (**i**) day of culture collected, (**j**) healthy or disease state and (**k**) annotated cell state/cluster. **l**, Rank-sorted differentially expressed genes across erythroid populations. Selected top genes overexpressed and downregulated in PS are annotated. Black dots ($n = 6,577$) represent statistically significant genes at a Bonferroni-adjusted significance threshold of <0.01. **m**, Volcano plot of pathway enrichment analysis results via erythroid differential gene expression comparisons of the PS to healthy control cells. Selected top pathways are annotated. The dotted line represents the threshold for consideration at an FDR < 0.1. **n**, Schematic overview of altered (metabolic) genes and pathways in PS relative to the healthy status. Genes and pathways upregulated in PS are shown in red and when downregulated shown in blue. Note—not all biochemical steps necessarily take place in mitochondria and the schematic has been simplified for illustrative purposes.

coincidental with an expanded myeloid pool and reduction of the phenotypic HSC pool. While profiling additional patients is required to verify selection dynamics in HSPCs, our observation of purifying selection in progenitor cells in the bone marrow compartment explains pan-lineage purifying selection across all mitochondriopathies studied herein.

Leveraging in vivo pseudotime trajectory analysis and in vitro models, we further assessed genomic and mtDNA features during erythroid differentiation to study the anemia characteristic of PS (Fig. 7). Our data suggest that serine/glycine biosynthesis is upregulated in PS cells to maintain DNA production and other critical components of the cell in states of mitochondrial dysfunction[48–50]. Mitochondrial one-carbon metabolism appears to be less sensitive to product inhibition by increased NADH:NAD^+ ratios, which are associated with mtDNA-related diseases due to an impaired electron transport chain[34]. These downstream perturbations may contribute to the downregulation of heme biosynthesis[51], which is necessary for adequate hemoglobin production during red blood cell generation[52]. We hypothesize that glycine may be redirected to synthesize one-carbon precursors for DNA replication in highly proliferative erythroblasts and/or GSH to scavenge increased ROS levels resulting from mitochondrial dysfunction. Correspondingly, we observed the downregulation of heme biosynthesis genes, which may lead to excess iron accumulation and granular depositions, ultimately forming characteristic sideroblastic cells in PS.

In sum, our multi-omic methods revealed unique genomic alterations in response to pathogenic mtDNA in distinct cellular compartments throughout the hematopoietic system. While mitochondria are ubiquitous, they nevertheless fulfill distinct roles depending on cell type and cell state. This emphasizes the need to ideally study patient-derived cellular specimens to fully capture alterations resulting from mitochondrial dysfunction attributable to germline or somatic mtDNA mutations. In this light, we demonstrate how comprehensive single-cell multi-omic approaches provide biologically important insights into the molecular alterations of primary mitochondrial defects.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01433-8.

References

1. Stewart, J. B. & Chinnery, P. F. Extreme heterogeneity of human mitochondrial DNA from organelles to populations. *Nat. Rev. Genet.* **22**, 106–118 (2021).
2. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.* **16**, 530–542 (2015).
3. Wallace, D. C. & Chalkia, D. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harb. Perspect. Biol.* **5**, a021220 (2013).
4. Gorelick G, A. N. et al. Respiratory complex and tissue lineage drive recurrent mutations in tumour mtDNA. *Nat Metab.* **3**, 558–570 (2021).
5. Smith, A. L. M. et al. Age-associated mitochondrial DNA mutations cause metabolic remodeling that contributes to accelerated intestinal tumorigenesis. *Nat. Cancer* **1**, 976–989 (2020).
6. Goldstein, A. & Falk, M. J. *Mitochondrial DNA Deletion Syndromes* (University of Washington, 2023).
7. Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339 (2019).
8. Xu, J. et al. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *eLife* **8**, e45105 (2019).
9. Lareau, C. A. et al. Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat. Biotechnol.* **39**, 451–461 (2020).
10. Walker, M. A. et al. Purifying selection against pathogenic mitochondrial DNA in human T cells. *N. Engl. J. Med.* **383**, 1556–1563 (2020).
11. Mimitou, E. P. et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* **39**, 1246–1258 (2021).
12. Krishnan, K. J. et al. What causes mitochondrial DNA deletions in human cells? *Nat. Genet.* **40**, 275–279 (2008).
13. Pitceathly, R. D. S., Rahman, S. & Hanna, M. G. Single deletions in mitochondrial DNA–molecular mechanisms and disease phenotypes in clinical practice. *Neuromuscul. Disord.* **22**, 577–586 (2012).
14. Lujan, S. A. et al. Ultrasensitive deletion detection links mitochondrial DNA replication, disease, and aging. *Genome Biol.* **21**, 248 (2020).
15. Hjelm, B. E. et al. Splice-Break: exploiting an RNA-seq splice junction algorithm to discover mitochondrial DNA deletion breakpoints and analyses of psychiatric disorders. *Nucleic Acids Res.* **47**, e59 (2019).
16. Hao, Y. et al. Dictionary learning for integrative, multimodal, and scalable single-cell analysis. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01767-y (2023).
17. Battaglia, S. et al. Uridine and pyruvate protect T cells' proliferative capacity from mitochondrial toxic antibiotics: a clinical pilot study. *Sci. Rep.* **11**, 12841 (2021).
18. King, M. P. & Attardi, G. Human cells lacking mtDNA: repopulation with exogenous mitochondria by complementation. *Science* **246**, 500–503 (1989).
19. Basu, S. et al. Accurate mapping of mitochondrial DNA deletions and duplications using deep sequencing. *PLoS Genet.* **16**, e1009242 (2020).
20. Goudenège, D. et al. eKLIPse: a sensitive tool for the detection and quantification of mitochondrial DNA deletions from next-generation sequencing data. *Genet. Med.* **21**, 1407–1416 (2019).
21. Kardos, G. et al. Refractory anemia in childhood: a retrospective analysis of 67 patients with particular reference to monosomy 7. *Blood* **102**, 1997–2003 (2003).
22. Nishimura, A. et al. Acquisition of monosomy 7 and a RUNX1 mutation in Pearson syndrome. *Pediatr. Blood Cancer* **68**, e28799 (2021).
23. Granja, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
24. Buenrostro, J. D. et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535–1548 (2018).
25. Chung, J.-W. et al. A combination of CD15/CD10, CD64/CD33, CD16/CD13 or CD11b flow cytometric granulocyte panels is sensitive and specific for diagnosis of myelodysplastic syndrome. *Ann. Clin. Lab. Sci.* **42**, 271–280 (2012).
26. Kussick, S. J. et al. Four-color flow cytometry shows strong concordance with bone marrow morphology and cytogenetics in the evaluation for myelodysplasia. *Am. J. Clin. Pathol.* **124**, 170–181 (2005).
27. Pekalski, M. L. et al. Neonatal and adult recent thymic emigrants produce IL-8 and express complement receptors CR1 and CR2. *JCI Insight* **2**, e93739 (2017).
28. Farruggia, P., Di Marco, F. & Dufour, C. Pearson syndrome. *Expert Rev. Hematol.* **11**, 239–246 (2018).
29. Gagne, K. E. et al. Pearson marrow pancreas syndrome in patients suspected to have Diamond-Blackfan anemia. *Blood* **124**, 437–440 (2014).
30. Cherry, A. B. C. et al. Induced pluripotent stem cells with a mitochondrial DNA deletion. *Stem Cells* **31**, 1287–1297 (2013).

31. Ludwig, L. S. et al. Transcriptional states and chromatin accessibility underlying human erythropoiesis. *Cell Rep.* **27**, 3228–3240 (2019).
32. Bao, X. R. et al. Mitochondrial dysfunction remodels one-carbon metabolism in human cells. *eLife* **5**, e10575 (2016).
33. Yang, M. & Vousden, K. H. Serine and one-carbon metabolism in cancer. *Nat. Rev. Cancer* **16**, 650–662 (2016).
34. Yang, L. et al. Serine catabolism feeds NADH when respiration is impaired. *Cell Metab.* **31**, 809–821 (2020).
35. Macaulay, I. C., Ponting, C. P. & Voet, T. Single-cell multiomics: multiple measurements from single cells. *Trends Genet.* **33**, 155–168 (2017).
36. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
37. Trifunovic, A. et al. Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature* **429**, 417–423 (2004).
38. Kujoth, G. C. et al. Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging. *Science* **309**, 481–484 (2005).
39. Vandiver, A. R. et al. Nanopore sequencing identifies a higher frequency and expanded spectrum of mitochondrial DNA deletion mutations in human aging. *Aging Cell* **22**, e13842 (2022).
40. Lehmann, D. et al. Understanding mitochondrial DNA maintenance disorders at the single muscle fibre level. *Nucleic Acids Res.* **47**, 7430–7443 (2019).
41. Nido, G. S. et al. Ultradeep mapping of neuronal mitochondrial deletions in Parkinson's disease. *Neurobiol. Aging* **63**, 120–127 (2018).
42. Lawless, C., Greaves, L., Reeve, A. K., Turnbull, D. M. & Vincent, A. E. The rise and rise of mitochondrial DNA mutations. *Open Biol.* **10**, 200061 (2020).
43. Morris, J. et al. Pervasive within-mitochondrion single-nucleotide variant heteroplasmy as revealed by single-mitochondrion sequencing. *Cell Rep.* **21**, 2706–2713 (2017).
44. Jones, N. et al. Metabolic adaptation of Human $CD4^+$ and $CD8^+$ T cells to T-cell receptor-mediated stimulation. *Front. Immunol.* **8**, 1516 (2017).
45. van der Windt, G. J. et al. Mitochondrial respiratory capacity is a critical regulator of CD8 T cell memory development. *Immunity* **36**, 68–78 (2012).
46. Hinks, T. S. C. & Zhang, X.-W. MAIT cell activation and functions. *Front. Immunol.* **11**, 1014 (2020).
47. Lisci, M. et al. Mitochondrial translation is required for sustained killing by cytotoxic T cells. *Science* **374**, eabe9977 (2021).
48. Korge, P., Calmettes, G. & Weiss, J. N. Increased reactive oxygen species production during reductive stress: The roles of mitochondrial glutathione and thioredoxin reductases. *Biochim. Biophys. Acta* **1847**, 514–525 (2015).
49. Sharma, R. et al. Circulating markers of NADH-reductive stress correlate with mitochondrial disease severity. *J. Clin. Invest.* **131**, e136055 (2021).
50. Enns, G. M. et al. Degree of glutathione deficiency and redox imbalance depend on subtype of mitochondrial disease and clinical status. *PLoS ONE* **9**, e100001 (2014).
51. De Franceschi, L. et al. Oxidative stress modulates heme synthesis and induces peroxiredoxin-2 as a novel cytoprotective response in β-thalassemic erythropoiesis. *Haematologica* **96**, 1595–1604 (2011).
52. Sankaran, V. G. & Weiss, M. J. Anemia: progress in molecular mechanisms and therapies. *Nat. Med.* **21**, 221–230 (2015).

[1]Department of Pathology, Stanford University, Stanford, CA, USA. [2]Parker Institute of Cancer Immunotherapy, San Francisco, CA, USA. [3]Department of Genetics, Stanford University, Stanford, CA, USA. [4]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [6]Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin Institute for Medical Systems Biology (BIMSB), Berlin, Germany. [7]Berlin Institute of Health at Charité—Universitätsmedizin Berlin, Berlin, Germany. [8]Technische Universität Berlin, Institute of Biotechnology, Berlin, Germany. [9]Department of Biology, Chemistry, Pharmacy, Freie Universität Berlin, Berlin, Germany. [10]Technology Innovation Lab, New York Genome Center, New York City, NY, USA. [11]Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [12]Center for Pediatric Neurosciences, Mitochondrial Medicine, Cleveland Clinic, Cleveland, OH, USA. [13]Department of Pathology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [14]Department of Pediatric Oncology, Charité-Universitätsmedizin Berlin, Campus Virchow Klinikum, Berlin, Germany. [15]Department of Computer Science, Stanford University, Stanford, CA, USA. [16]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [17]Department of Biology and Koch Institute, Massachusetts Institute of Technology, Cambridge, MA, USA. [18]Present address: Immunai, New York City, NY, USA. [19]Present address: 10x Genomics, San Francisco, CA, USA. [20]Present address: Genentech, San Francisco, CA, USA. [21]These authors contributed equally: Sonia M. Dubois, Frank A. Buquicchio, Yu-Hsin Hsieh. ✉e-mail: clareau@stanford.edu; aviv.regev.sc@gmail.com; sankaran@broadinstitute.org; suneet.agarwal@childrens.harvard.edu; leif.ludwig@bih-charite.de

Methods

Our research complies with all relevant ethical and regulatory guidance, including the Institutional Review Board at Boston Children's Hospital, ethics approval from Charité Universitätsmedizin Berlin, Germany, and a research agreement with the North American Mitochondrial Disease Consortium (NAMDC).

Cell lines and cell culture

Biological samples for cell lines were procured under protocols approved by the Institutional Review Board at Boston Children's Hospital, after obtaining written informed consent in accordance with the Declaration of Helsinki. PS fibroblasts were derived from the patient bone marrow (PS1 and PS2) or skin (PS3). Control fibroblasts were derived from healthy donors' skin (Control 1 and Control 2). All fibroblasts were grown in DMEM containing 15% fetal bovine serum (FBS), L-glutamine, nonessential amino acid and penicillin/streptomycin. Cells were incubated at 37 °C with 5% carbon dioxide (CO_2).

Healthy donor and patient samples

Primary human peripheral blood and bone marrow samples were collected under Institutional Review Board-approved protocols and with written informed consent for genomic sequencing. Primary hematopoietic samples were collected from three previously diagnosed patients, including a 7-year-old male with PS/KSS ('PT1'), a 4-year-old female with PS ('PT2') and a 4-year-old male with PS and del7q MDS ('PT3'). Peripheral blood and BMMNCs were isolated using Ficoll Paque Plus solution and density gradient centrifugation using SepMate tubes (StemCell Technologies). PBMCs from adult female donors ('KSS1', 'KSS2' and 'CPEO1') were obtained in collaboration with NAMDC[53]. Both donors with KSS presented with pigmentary retinopathy and various neurological symptoms, as KSS1 was diagnosed with hearing loss and KSS2 was diagnosed with ataxia and dementia over the disease course. There were limited relevant clinical annotations for CPEO1. None of the patients had defined postmitotic heteroplasmy levels following the diagnosis available in the NAMDC records. All PBMC samples were stored in vapor-phase liquid nitrogen after cryopreservation with 10% dimethyl sulfoxide until analysis.

Healthy donor BMMNCs were obtained from StemCell Technologies. Healthy adult $CD34^+$ HSPCs were obtained from the Fred Hutchinson Hematopoietic Cell Processing and Repository. The $CD34^+$ samples were de-identified, and approval for use of these samples for research purposes was provided by the Institutional Review Board and Biosafety Committees at Boston Children's Hospital. For healthy pediatric controls for the differential gene expression analyses and T-cell culture experiments, pseudonymized samples from bone marrow donors of 5 ('Ped1') and 14 ('Ped2') years of age were obtained at Charité Universitätsmedizin Berlin, Germany, following approval of the local ethics commission (EA2/144/15). Informed consent was obtained from parents/legal guardians for all pediatric material.

Statistics and reproducibility

No statistical method was used to predetermine the sample size. No data were excluded from the analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessments. All custom codes used to replicate analyses are available as part of the code availability.

Human T-cell activation cultures

PBMCs or isolated T cells were cultured in RPMI-1640 medium supplemented with 10% FBS, penicillin and streptomycin, as well as 10 ng ml^{-1} IL-2 (PeproTech) at 37 °C and 5% CO_2. Cells were in vitro activated with plate-bound anti-CD3 antibody (5 µg ml^{-1}, clone OKT3, BioLegend) plus soluble anti-CD28 antibody (1 µg ml^{-1}, clone 28.2; BioLegend, 302901). Upon thawing (defined as 'day 0'), cells were resuspended at a concentration of ~10^6 cells per ml in culture medium plus anti-CD28 antibody, and ~150,000 cells were plated into 96-well plates precoated with anti-CD3 antibody. After 48 h of activation (defined as 'day 2'), cells were transferred into uncoated plates and maintained at a density of 1–2 × 10^6 cells per ml. Cell counts were determined every 2–3 d.

The following culturing conditions were used for the proliferation assay: RPMI-1640 medium supplemented with 5 mM glucose, 2 mM stable glutamine, 10% FBS, penicillin and streptomycin, as well as 10 ng ml^{-1} IL-2 (PeproTech) with or without the presence of 1 mM pyruvate and 200 nM uridine at 37 °C and 5% CO_2. PBMCs were labeled with CellTrace Violet (2.5 µM; Thermo Fisher Scientific, C34557) before activation according to the manufacturers' instructions. Cell proliferation was measured on day 4 after activation as measured by the dilution of CellTrace Violet upon activation. In vitro activated T cells were stained using 1:100 AF488-conjugated CD3 (clone OKT3; BioLegend, 317310), 1:100 PE-conjugated CD4 (clone PRA-T4; BioLegend, 300508), 1:100 APC-conjugated CD8 (clone SK1; BioLegend, 344721), 1:100 BV785-conjugated CD45RA (clone HI100; BioLegend, 304139) and 1:2,000 Fixable Viability Dye eFluor 780 (eBioscience, 65-0865-18).

Human erythroid in vitro cell culture

BMMNCs or $CD34^+$ HSPCs from healthy donors or PT3 were differentiated into mature erythroid cells using a three-phase culture protocol[54,55]. Cells used for scRNA-seq and mtscATAC-seq experiments were derived from two independent cultures using PT3 cells, but two different healthy control donors were used for each culture. In phase 1 (days 0–7), cells were cultured at a density of 10^5–10^6 cells per ml in IMDM supplemented with 2% human AB plasma, 3% human AB serum, 1% penicillin/streptomycin, 3 IU ml^{-1} heparin, 10 µg ml^{-1} insulin, 200 µg ml^{-1} holo-transferrin, 1 IU EPO, 10 ng ml^{-1} stem cell factor and 1 ng ml^{-1} IL-3. In phase 2 (days 7–12), IL-3 was omitted from the medium. In phase 3 (days 12–18), cells were cultured at a density of 1 × 10^6 cells per ml, with both IL-3 and SCF omitted from the medium, and the holo-transferrin concentration was increased to 1 mg ml^{-1}. Cells were cultured at 37 °C and 5% CO_2.

Flow cytometry analysis and sorting

For flow cytometry analysis and sorting, cells were washed with FACS buffer (1% FBS in PBS) before antibody staining. In vitro cultured primary erythroid cells were stained using 1:50 APC-conjugated CD235a (glycophorin A, clone HIR2; eBioscience, 50-153-69) and 1:50 FITC-conjugated CD71 (clone OKT9; eBioscience, 14-0719-82) for 15 min on ice. PT3 bone marrow-derived $CD34^+$ cells were stained using 1:40 APC-conjugated CD34 (clone 581; BioLegend, 343509). In vitro activated T cells were stained using 1:200 AF488-conjugated CD3 (clone OKT3; BioLegend, 317310), 1:200 PE-conjugated CD4 (clone PRA-T4; BioLegend, 300508), 1:100 APC-conjugated CD8 (clone SK1; BioLegend, 344721), 1:200 PE-Cy7-conjugated CD45RO (clone UCHL1; BioLegend, 304229), 1:200 BV785-conjugated CD45RA (clone HI100; BioLegend, 304139), 1:50 BV605-conjugated CCR7 (clone G043H7; BioLegend, 353223) and 1:50 APC-H7-conjugated CD27 (clone MT271; BD Bioscience, 560223). For mtscATAC-seq and ASAP-seq experiments, residual granulocytes were excluded by staining cells using 1:50 PE-conjugated CD66b (clone G10F5; BioLegend, 305102). For live/dead cell discrimination, Sytox Blue was used at a 1:1,000 dilution according to the manufacturer's instructions (ThermoFisher Scientific, S34857). FACS analysis was conducted on BD Bioscience Fortessa flow cytometers at the Whitehead Institute Flow Cytometry Core and at the Berlin Institute of Health and Berlin Institute for Medical Systems Biology. Cell sorting was conducted using the Sony SH800 sorter with a 100-µm chip at the Broad Institute Flow Cytometry Facility. The data were analyzed using the FlowJo software v10.4.2.

May–Grünwald-Giemsa staining

Harvested cells were washed once at 300*g* for 5 min, resuspended in 200 µl of FACS buffer and spun onto poly-L-lysine-coated microscope slides with a Shandon 4 (ThermoFisher Scientific) cytocentrifuge at

300 rpm for 4 min. Visibly dry slides were transferred into the May–Grünwald solution (Sigma-Aldrich) for 5 min, rinsed four times every 30 s in water and transferred to Giemsa solution (Sigma-Aldrich) for 15 min. Slides were washed as described previously, dry mounted with coverslips and examined. All images shown were taken using a Metafer slide scanning platform and software (Metasystems) at 63× magnification.

Mitochondrial single-cell ATAC-seq (mtscATAC-seq)

MtscATAC-seq libraries were generated using the 10× Chromium Controller and the Chromium Single Cell ATAC Library & Gel Bead Kit (1000111) according to the manufacturer's instructions (CG000169-Rev C and CG000168-Rev B) as outlined below and previously described to increase mtDNA yield and genome coverage[9]. Briefly, 1.5 ml or 2 ml DNA LoBind tubes (Eppendorf) were used to wash cells in PBS and downstream processing steps. After washing, cells were fixed in 0.1 or 1% formaldehyde (FA; ThermoFisher Scientific, 28906) in PBS for 10 min at room temperature, quenched with glycine solution to a final concentration of 0.125 M and washed twice in PBS via centrifugation at 400*g* for 5 min at 4 °C. Cells were subsequently treated with lysis buffer (10 mM Tris–HCL, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP40 and 1% BSA) for 3 min for primary cells and 5 min for cell lines on the ice, followed by addition of 1 ml of chilled wash buffer and inversion (10 mM Tris–HCL, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$ and 1% BSA) before centrifugation at 500*g* for 5 min at 4 °C. The supernatant was discarded, and cells were diluted in 1× diluted nuclei buffer (10× Genomics) before counting using trypan blue and a Countess II FL Automated Cell Counter. If large cell clumps were observed, a 40-µm Flowmi cell strainer was used before processing cells according to the Chromium Single Cell ATAC Solution user guide with no additional modifications. Briefly, after tagmentation, the cells were loaded on a Chromium Controller Single-Cell Instrument to generate single-cell gel bead-in-emulsions (GEMs), followed by linear PCR, as described in the protocol using a C1000 Touch Thermal Cycler with 96-Deep Well Reaction Module (BioRad). After breaking the GEMs, the barcoded tagmented DNA was purified and further amplified to enable sample indexing and enrichment of scATAC-seq libraries. All genomic libraries were quantified using a Qubit dsDNA HS Assay Kit (Invitrogen) and a high-sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent).

ATAC with selected antigen profiling by sequencing (ASAP-seq)

PT3-derived BMMNCs were stained with a 242 TSA-conjugated antibody panel (BioLegend; Supplementary Table 3 for a list of antibodies, clones and barcodes used for ASAP-seq) as previously described[11]. To enable flow cytometry-based enrichment of $CD34^+$ cells, the sample was co-stained using an APC-conjugated CD34 (clone 581; BioLegend, 343509) to sort live CD66b-$CD34^+$ and otherwise CD66b-BMMNCs, which were then pooled after sorting and processed for ASAP-seq as previously described[11] and outlined online at https://cite-seq.com/asapseq/. Briefly, following sorting, cells were fixed in 1% FA and processed as described for the mtscATAC-seq workflow described previously, with the modification that during the barcoding reaction, 0.5 µl of 1 µM bridge oligo A (BOA for TSA) was added to the barcoding mix. Silane bead elution and SPRI cleanup steps were modified as described to generate the indexed protein tag library[11].

scRNA-seq

scRNA-seq libraries were generated using the 10× Genomics Chromium Controller and the Chromium Single Cell 3′ Library Construction Kit v2 according to the manufacturer's instructions. Briefly, the suspended cells were loaded on a Chromium Controller Single-Cell Instrument to generate single-cell GEMs, followed by reverse transcription and sample indexing using a C1000 Touch Thermal Cycler with 96-Deep Well Reaction Module (BioRad). After breaking the GEMs, the barcoded cDNA was purified and amplified, followed by fragmenting, A-tailing and ligation with adaptors. Finally, PCR amplification was performed to enable sample indexing and enrichment of scRNA-Seq libraries.

10× Genomics multiome

Single-cell multiome libraries of healthy pediatric control PBMCs were generated using the 10× Genomics Chromium Next GEM Single Cell Multiome ATAC + Gene Expression reagent bundle (100285) and the Chromium controller according to the manufacturer's instructions (CG000338-Rev E). Briefly, following the sorting of live and CD66b-negative cells, 1.5 ml DNA LoBind tubes (Eppendorf) were used to wash cells in PBS and for downstream processing steps. After washing, cells were lysed for 3 min in lysis buffer (10 mM Tris–HCL, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20, 0.1% NP40, 0.01% digitonin, 1% BSA, 1 mM DTT and 1 U μl^{-1} RNase inhibitor). Following lysis, cells were washed three times with 1 ml wash buffer (10 mM Tris–HCL, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA, 0.1% Tween-20, 1 mM DTT and 1 U μl^{-1} RNase inhibitor) before centrifugation at 500*g* for 5 min at 4 °C. The supernatant was discarded, and cells were diluted in 1× diluted nuclei buffer (10× Genomics) before counting using trypan blue and a Countess II FL Automated Cell Counter. If large cell clumps were observed, a 40-µm Flowmi cell strainer was used before processing cells according to the Chromium Next GEM Single Cell Multiome ATAC + Gene Expression user guide with no further modifications. Briefly, after transposition and chip loading, the cells were loaded into the Chromium Controller instrument to generate single-cell GEMs, followed by incubation as described in the protocol using a C1000 Touch Thermal Cycler with 96-Deep Well Reaction Module (BioRad). After breaking the GEMs, the barcoded DNA was purified and further amplified before separate ATAC and cDNA library construction. For the ATAC part, the purified DNA was further amplified to enable sample indexing and enrichment of the DNA. The cDNA was further amplified, purified and quantified using a high-sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent). The cDNA was subsequently fragmented, PCR-amplified and purified as depicted in the Chromium Next GEM Single Cell Multiome ATAC + Gene Expression user guide with no further modifications.

DOGMA-seq

For Pearson PT1, PBMCs were processed with DOGMA-seq as described previously[11]. Briefly, PBMCs were stained with total-seq A antibody panels (TotalSeq-A Human Universal Cocktail, v1.0; BioLegend, 399907; no dilution), PE-conjugated CD66b and Sytox Blue, and dead and $CD66b^+$ cells were removed via sorting. Sorted live cells were fixed at 0.1% FA for 5 min at room temperature, and subsequent lysis/permeabilization steps were analogous to the multiome protocol except for Tween-20, and digitonin was omitted from the lysis and wash buffers. Permeabilized cells were then processed as described for the multiome workflow above with the modification of adding 1 µl of 0.2 µM antibody-derived tag (ADT) additive primer (CCTTGGCACCCGAGAATT*C*C). After SPRI cleanup of the preamplification PCR product, the beads were eluted in 100 µl of EB buffer. Notably, 25 µl of the eluate were used for ATAC-seq library processing, and 35 µl were each used as input for cDNA and antibody tag amplification[11], respectively. Libraries for MAS-ISO-seq were constructed using the cDNA from the RNA modality as previously described[56]. While the MAS-ISO-seq libraries were of high quality, the low mtRNA capture in the DOGMA-seq assay limited further analysis as we detected less than one fusion or relevant wild-type mtRNA per cell from these libraries (Supplementary Fig. 2). Furthermore, we note the fixation step as part of the DOGMA-seq workflow that reduces the average cDNA fragment size.

Sequencing

All libraries were sequenced using the Illumina NextSeq550 and NovaSeq6000 sequencing platforms. 10× Genomics scATAC-seq

and ASAP-seq libraries were sequenced with paired-end reads (2 × 72 cycles). 10× Genomics multiome and 3′ scRNA-seq libraries were sequenced as recommended by the manufacturer. For DOGMA-seq and in vitro T-cell stimulation libraries (Fig. 3), libraries were sequenced on the NovaSeq6000 platform with a 2 × 150 bp of 2 × 100 paired-end read configuration, which were trimmed to be 2 × 72 cycles for compatibility with the optimized deletion calling workflow originally implemented on the NextSeq. MAS-ISO-seq libraries were sequenced using two SMRT cells using the PacBio Sequel II as previously described[56]. Both cells yielded >30 M reads, and >99% of reads had a valid barcode and UMI from the 10× multiome/DOGMA-seq design.

mtDNA deletion calling and heteroplasmy estimation in single cells

Although large mtDNA deletions have been well-documented in a variety of next-generation sequencing datasets, we observed that coordinates associated with deletions (for example, the 'common' deletion) may be incorrect at base-pair resolution. These differences are primarily due to differences in the coordinates of the mitochondrial reference genome and variations in the results of sequencing read alignment tools, particularly near homomorphic sequences at deletion junctions. Thus, we recommend using the sequencing data from the particular sequencing experiment to identify the deletion junction within the primary sequencing data that is being analyzed. As part of our software solution, mgatk-del in 'find' mode takes a.bam file and compiles a list of key summary statistics, including the number of clipped reads per position, secondary alignment bases and overall coverage, to identify deletions. The outputs of mgatk-del find include plots (for example, Extended Data Fig. 1b) and tables that facilitate identifying the specific base pairs associated with the mtDNA deletions.

After the precise breakpoints have been determined, single-cell mtDNA deletion heteroplasmy can be estimated using the second step in the mgatk-del pipeline. Here PCR-deduplicated single-cell bam files (emitted as intermediate output in the standard mgatk pipeline) serve as the primary input, yielding an estimation of mtDNA deletion heteroplasmy per user-specified deletion. This metric is determined by using the ratio of reads overlapping the deletion junction that either support (via clip) or provide no evidence of the deletion (contiguous alignment over the read window, as depicted in Fig. 1c,d). Notably, each paired-end read contributes only once to the heteroplasmy metric. As a comparison, coverage-based heteroplasmy (Fig. 1e) was estimated via one minus the ratio of mean per-base coverage within the deleted region over the mean per-base coverage outside the deleted region. For negative values (when the within-deleted region coverage exceeded that of the outside region), values were adjusted to 0% heteroplasmy for display purposes (coverage-based heteroplasmy was not used in any downstream analyses).

To generate simulated sequencing datasets and to benchmark this approach, we used the wgsim tool from within samtools[57] to generate paired-end reads of length 72 bp (same length as for our mtscATAC-seq data), 50 bp or 100 bp, which represent common sequencing configurations. Sequencing reads were simulated from either the revised Cambridge reference sequence (rCRS) or a synthetic mtDNA chromosome that encoded the specified deletion. Simulated sequencing reads were then aligned to the masked reference genome used by CellRanger-ATAC, and the resulting aligned reads from the.bam files were mixed in specific ratios (ten mixtures per deletion) to specify the true heteroplasmy for the given simulation. The estimated heteroplasmy was computed by running the function used in mgatk-del with the search space of possible values in the outer-param and inner-param and then the root-mean-squared error (RMSE) was computed based on the difference. In total, 22 mtDNA deletions were considered, which represented a curated list of the six deletions in our study and 16 additional deletions that were curated from MITOMAP[58]. The default parameters in mgatk-del (outer-param, 9; inner-param, 24) represent values that performed consistently well across a variety of deletions and read lengths. We suggest that mgatk-del can produce reasonable single-cell heteroplasmy estimations from the default parameters. Specifically, we observed a mean 0.93% RMSE difference between default and optimal hyperparameter values across our six PS mtDNA deletions in the cell lines and primary cells. Thus, we suggest that a grid search to determine optimal hyperparameters for accurate heteroplasmy estimation may be useful but typically unnecessary for new datasets.

To quantify that the variance in heteroplasmy was attributable to variation in coverage per single cell (Extended Data Figs. 1g and 2c), the overall mean per comparison was computed, and a permuted heteroplasmy was simulated using the rbinom() function with the overall heteroplasmy and per-cell coverage as inputs. The observed (true data) and null (output of rbinom simulation) are shown for each comparison. To assess the sensitivity and specificity of the heteroplasmy estimation, a threshold of 1% was used for 'detection of deletion' (Extended Data Fig. 1i). For further validation of heteroplasmy estimation as a function of coverage, we used the simulated reads from the wgsim alignments to synthetically create cells with a predetermined heteroplasmy at 15 variable coverages between 10× and 500× coverage (Extended Data Fig. 1j) and subsequently estimated heteroplasmy with the core function in the mgatk-del workflow. Per deletion, we simulated 100 cells and averaged the mean absolute error to quantify the bias associated with the mgatk-del coverage estimates from both coverage and clipped heteroplasmy. Collectively, our benchmarking and simulation analysis indicates that for specific inference near 0%, clipped-read-based heteroplasmy performs better, whereas coverage-based heteroplasmy (once base-pair resolution junctions are inferred) can produce more accurate absolute heteroplasmy estimates, particularly in lower coverage settings, including the DOGMA-seq data shown here. For this study, we consistently use the clipped-read base estimates from mtscATAC-seq data to both accurately infer purifying selection in varied populations and as the mean coverage of our mtscATAC-seq profiles was 81.5×. Full details of the simulations, including code for reproducibility, and additional discussion of the methods are available as part of our online resources.

scATAC-seq analyses

Raw-sequencing data were demultiplexed using CellRanger-ATAC mkfastq. Demultiplexed sequencing reads for all libraries were aligned to the mtDNA blacklist modified[9] hg38 reference genome using CellRanger-ATAC count v2. Deletions in mtDNA were identified per patient library and heteroplasmy was quantified using the exact breakpoints as discussed in the previous section. Downstream analyses of the three PS donors and one healthy donor previously profiled with mtscATAC-seq[9] were performed after identifying cells with a minimum depth of 10× on mtDNA, 1000 ATAC fragments passing filters and 45% of fragments in accessibility peaks from an aggregated peak set. Latent semantic indexing (LSI) was performed, and the 2–30 components were adjusted for donor effects using harmony before producing a two-dimensional embedding and clustering using the harmony components[59]. Gene activity scores were computed and normalized using the Signac workflow[60]. For PBMC cell-type annotations, granular cell-type labels and UMAP coordinates were derived by using the Seurat Dictionary Learning[16] for cross-modality integration. We used Azimuth CITE-seq reference dataset labels[61] with public 10× genomics multiome RNA- and ATAC-seq PBMC data as a cross-modality bridge. Libraries for the MELAS donors[10] were reprocessed with the hg38 scATAC-seq reference and consistently projected using the Seurat reference. Libraries for the CPEO and KSS donors were processed consistently as well, and the base-pair resolution for the deletions was inferred from the '_del_find.clip.tsv' file from mgatk-del-find. For all deletions, the top clipped base pairs were called deletions in Fig. 4 from this analysis. Statistical comparisons of the extent of purifying selection (for example, Fig. 2 cumulative distribution plots) were computed based on the proportion

of cells with less than 1% heteroplasmy using a two-sided proportion test in R. The uncorrected *P* values are shown in each panel.

scRNA-seq analyses

PS cell-derived 10× 3′ scRNA-seq sequencing libraries were demultiplexed and aligned to the hg38 reference with CellRanger v3.0.2. Healthy PBMC datasets were augmented from the public resource of 10× single-cell gene expression. Raw-sequencing reads from two libraries (pbmc4k and pbmc8k) of 10× 3′ v2 chemistry were downloaded and reprocessed consistent with the PS cell libraries. Filtered count matrices from the two 10× 3′ v3 chemistries (pbmc5k and NextGEM) were downloaded from the online resource as they were already aligned to the same reference as the rest of the PS data. We note that the pairs of libraries from each technology were derived from the same biological donor ('H1' for v2 libraries; 'H2' for v3 libraries). Separately, scRNA-seq libraries ('Ped1' and 'Ped2') from the two pediatric donors were aligned to the same hg38 reference and aggregated at the counts-matrix level as the reference transcriptome used for quantification was identical between the libraries.

Using the filtered gene by cell count matrices for all scRNA-seq libraries, we identified and removed putative cell doublets using scrublet[62] with the default parameters and specified a 5% expected doublet rate. Barcodes identified as cell doublets were then filtered. Next, we performed data integration across these seven libraries for the PBMC reference projection via Azimuth via Seurat v4. Differential gene expression summary statistics from scRNA-seq libraries were computed using edgeR (v3.16.0)[63] while adjusting for the scaled number of genes detected per cell (edgeRQLFDetRate[64]). We note that while edgeR was originally introduced for bulk RNA-seq, a comparison of differential expression tools demonstrated good performance for this approach compared to other bulk and single-cell strategies[64]. We performed gene-set enrichment analyses with the Panther Pathway enrichments using the WebGestalt v2019 framework[65] using a rank ordering of genes by the signed *z* score (Supplementary Figs. 3 and 4). Bulk expression data from GTEx were curated from the GTEx online portal for the indicated genes[66]. All other visualizations and analyses for scRNA-seq data were performed using the Seurat framework. Gene module score analyses, including for the oxidative phosphorylation pathway, heme biosynthesis pathway and glycolysis, were performed using gene sets from the PANTHER dataset accessed via WebGestalt[65] and computed using the AddModuleScore in Seurat using the default parameters. For the comparison of within and between donor/state heterogeneity (Supplementary Fig. 3), we considered six major cell types with an ample number of cells from our reference projection annotation. After computing a principal component space using the Seurat defaults for all cells, we randomly subsampled 100 cells within each condition to make comparisons within or between groups. The boxplots in the figure represent the cell–cell distances from all comparisons for ten simulation iterations.

DOGMA-seq analyses

For the DOGMA-seq antibody tag data, per-cell and per-antibody tag counts were enumerated via the kite | kallisto | bustools framework accounting for unique bridging events as previously described[11,67]. Cells called by the CellRanger-arc knee call were filtered based on the abundance of protein (>50 unique molecules), minimal nonspecific antibody binding (<10 molecules associated with isotype control antibodies), total accessible chromatin (>1,000 nuclear fragments), >50% fragments in accessibility peaks and total gene expression (>1,000 UMIs per cell and >500 genes detected per cells), following our prior quality control of cells from DOGMA-seq for use in the 3WNN analyses (Extended Data Fig. 2)[11]. For consistency with other analyses, we used the Seurat reference projection of our dataset with the RNA modality for the DOGMA-seq data and corroborated the cell state classification by performing the same project using the ATAC modality.

For comparison of heteroplasmy, we required either a minimum of 10× total coverage for a coverage-based heteroplasmy estimation or a minimum of ten reads supporting or refuting the deletion for clipped-based heteroplasmy inference (Extended Data Fig. 2f). For the long-rad PacBio sequencing, raw molecules were processed using the standard manufacturer's workflow into full cDNA molecules in the format of unmapped.bam files. Deleted molecules or wild-type molecules from the MAS-ISO-seq libraries were inferred using the base-pair resolution deletion junctions for PT1 between *COX1* and *ND5* or the wild-type sequences of a full gene to derive unique 16-mers (8 bp on either side of the deletion junction), which we determined empirically to be unique strings in the human reference genome for PT1. To quantify heteroplasmy, we filtered reads for a valid barcode and UMI then parsed the full molecules for these unique 16-mers after validating that there were no detectable levels of the deletion junction in other healthy control samples.

Chromosome 7 copy number (deletion) analysis

del7q analyses were performed only for PT3 as the other patients showed no evidence of copy number alterations (either from cytogenetics or sequencing data). To assign cells as either wild-type or del7q, we performed copy number analyses of the accessible chromatin (scATAC-seq) data for all libraries and cells profiled from PT3. From the cytogenetics and sequencing data, we estimated the positions 110,000,000 (on chromosome 7q22) as the approximate break point and computed the fraction of fragments occurring after this coordinate to get an estimate of the copy number changes across the different libraries (shown in Fig. 5b). To call the single-cell del7q status, we used CONICS[68] on the gene activity matrix and specified a custom region spanning the deletion for estimation of the copy number. Because the del7q abundance varied between biological sources (for example, PBMCs and $CD34^+$ cells) and resulted in different maximum-likelihood estimates for the Gaussian distribution parameters, wild-type or single-cell del7q genotype per cell was called using a manual threshold of the predicted probability of the two-component mixture model based on the density of the first component's predicted probability. Explicitly, the thresholds used were 0.5 for the $CD34^+$ data (Fig. 5), 0.3 for the ASAP BMMNC dataset (Fig. 6) and 0.3 for the erythroid differentiation dataset (Fig. 7).

Bone marrow ASAP-seq analyses

For the ASAP-seq antibody tag data, per-cell and per-antibody tag counts were enumerated via the kite | kallisto | bustools framework, accounting for unique bridging events as previously described[11,67]. Cells called by the CellRanger-ATAC knee call were filtered based on the abundance of protein (>150 unique molecules) and accessible chromatin (>1,000 nuclear fragments) as well as accessible chromatin enrichment (>25% fragments in accessibility peaks) and minimal nonspecific antibody binding (<10 molecules associated with isotype control antibodies). Dimensionality reduction and clustering were performed only using the chromatin accessibility modality of the ASAP-seq, and protein expression and gene activity values were used to annotate clusters as previously described[11]. Differential protein and gene activity score calculations (via Signac[60]) were performed using the FindMarkers function in Seurat. Somatic mtDNA mutations were identified by running mgatk on the ASAP-seq cells, exceeding a mean 20× coverage using the default parameters[9]. For $CD34^+$ analysis and projections (Fig. 6d–g), we used the LSI reference projection and reference $CD34^+$ landscape as previously described[23]. For the erythroid pseudotime trajectory (Fig. 7a–d), we used a semi-supervised trajectory inference previously introduced[69,70] connecting the annotated multipotent progenitor populations to the committed erythroid population. Single-cell pseudotime was estimated using the projection of each cell along the axis joined between the per-cluster centroids.

Erythroid single-cell analyses

Raw-sequencing data were demultiplexed and aligned using CellRanger and CellRanger-ATAC as done for the PBMC analyses. To minimize batch effects, PS and healthy cells were pooled for single-cell processing and then computationally deconvolved using donor-specific SNPs. Differential gene expression, via edgeR (v3.16.0)[63], and pathway enrichment (Panther Pathway enrichments using the WebGestalt framework[65]) were conducted using the same workflow as the PBMC data. We computed a per-cell erythroid module score using 99 genes (for example, *GATA1*, *ALAS2* and *HBB*) highly upregulated in erythropoiesis from our previous bulk transcriptomic atlas of cells from this in vitro system[31] using the AddModuleScore function in Seurat with the default parameters.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data associated with this work is available at GEO accession GSE173936.

Code availability

Software and documentation for mitochondrial DNA variant calling, including deletion calling and heteroplasmy estimation, is available via the mgatk package at http://github.com/caleblareau/mgatk as of version 0.6.3. All custom code to reproduce all analyses supporting this paper is available at https://github.com/caleblareau/pearson_syndrome. Code used in this paper is indexed in Zenodo (https://zenodo.org/record/7853604).

References

53. Rosales, X. Q. et al. The North American mitochondrial disease registry. *J. Transl. Genet Genom.* **4**, 81–90 (2020).
54. Hu, J. et al. Isolation and functional characterization of human erythroblasts at distinct stages: implications for understanding of normal and disordered erythropoiesis in vivo. *Blood* **121**, 3246–3253 (2013).
55. Giani, F. C. et al. Targeted application of human genetic variation can improve red blood cell production from stem cells. *Cell Stem Cell* **18**, 73–78 (2016).
56. Al'Khafaji, A. M. et al. High-throughput RNA isoform sequencing using programmable cDNA concatenation. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01815-7 (2023).
57. Li, H. et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
58. Lott, M. T. et al. mtDNA variation and analysis using mitomap and mitomaster. *Curr. Protoc. Bioinforma.* **44**, 1.23.1–1.23.6 (2013).
59. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
60. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
61. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
62. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291 (2019).
63. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
64. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
65. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).
66. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
67. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal RNA-seq quantification with kallisto. *Nat. Biotechnol.* **34**, 525–527 (2016).
68. Müller, S., Cho, A., Liu, S. J., Lim, D. A. & Diaz, A. CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics* **34**, 3217–3219 (2018).
69. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
70. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).

Acknowledgements

We are deeply grateful to the patients and families who made this work possible. Patient samples and data from adult donors with KSS and CPEO were provided by the North American Mitochondrial Disease Consortium (NAMDC), which is funded by a grant from NIH (U54NS078059). We thank A. Shimamura (Boston Children's Hospital) and the Boston Children's Hospital Bone Marrow Failure and Myelodysplastic Syndrome program for their support in this study (supported by NIH grant RC2DK122533). We thank the MDC/BIH Genomics Platform, Berlin (FacilityID=1565, The CoreMarketplace: MDC&BIH Technology Platform Genomics) for technical support relating to flow cytometry and sequencing efforts and Claudia Quedenau for conducting the MAS-ISO-seq experiments. We acknowledge support from the Broad Institute and the Whitehead Institute Flow Cytometry core facilities. This work was supported by NIH K99 HG012579 (to C.A.L.), RC2 DK122533-01 (to M.D.F.), R01 DK103794 (to V.G.S.), R33 HL120791 (to V.G.S.), R01 DK107716 (to S.A.), R33 HL154133 (to S.A.) and UM1 HG012076 (to C.A.L., A.T.S. and L.S.L.). This research was supported by a Stanford Science Fellowship (to C.A.L.), BroadIgnite Award (to L.S.L.), an Emmy Noether fellowship by the German Research Foundation (D.F.G.; LU 2336/2-1 to L.S.L.), the Hector Fellow Academy (to Y.H.H., P.K., L.N. and L.S.L.), a Lloyd J. Old STAR Award from the Cancer Research Institute (A.T.S.), an ASH Scholar Award from the American Society of Hematology (to A.T.S.), a gift from the Lodish Family to Boston Children's Hospital (to V.G.S.), the New York Stem Cell Foundation (NYSCF to V.G.S.), the Howard Hughes Medical Institute and Klarman Cell Observatory (to A.R.), the Champ Foundation (to S.A.) and the Associazione Luigi Comini Onlus (to S.A.).

Author contributions

C.A.L., A.R., V.G.S., S.A. and L.S.L. conceived and designed the project. C.A.L. developed the mgatk-del software and led all analyses. L.S.L. led, designed and performed experiments with F.A.B., Y.H.H., P.K., L.N. and Y.Y., and assistance from S.M.D., W.L. and C.M. E.P.M. and P.S. contributed to reagents and advised on the ASAP-seq experiments. L.O., J.S. and C.E. contributed to healthy pediatric control samples and clinical perspectives. C.A.L., K.G. and J.M.V. developed the mtDNA simulation framework. F.B., Y.H.H., P.K., L.N., S.D.P., J.C.G., E.F., R.M., P.M., SuP, A.K., S.A. and L.S.L. contributed to data interpretation. S.M.D., S.D.P., M.D.F., A.K., S.A.V., A.T.S., A.R., V.G.S., S.A. and L.S.L. supervised various aspects of this work. C.A.L. and L.S.L. wrote the paper with input from all authors.

Competing interests

The Broad Institute has filed for a patent relating to the use of the technology described in this paper where C.A.L., L.S.L., C.M., A.R. and V.G.S. are named inventors (US provisional patent application 62/683,502). C.A.L. and L.S.L. are consultants to Cartography Biosciences. ATS is a founder of Immunai and Cartography Biosciences and receives research funding from Allogene

Therapeutics and Merck Research Laboratories. A.R. is a founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas Therapeutics, and until August 31, 2020, was an SAB member of Syros Pharmaceuticals, Neogene Therapeutics, Asimov and ThermoFisher Scientific. Since August 1, 2020, A.R. has been an employee of Genentech. V.G.S. serves as an advisor to and/or has equity in Branch Biosciences, Novartis, Forma, Cellarity and Ensoma. S.A.V. is an advisor to Immunai and has provided consulting services to Koch Disruptive Technologies and ADC Therapeutics. The remaining authors declare no competing interests.

Additional information

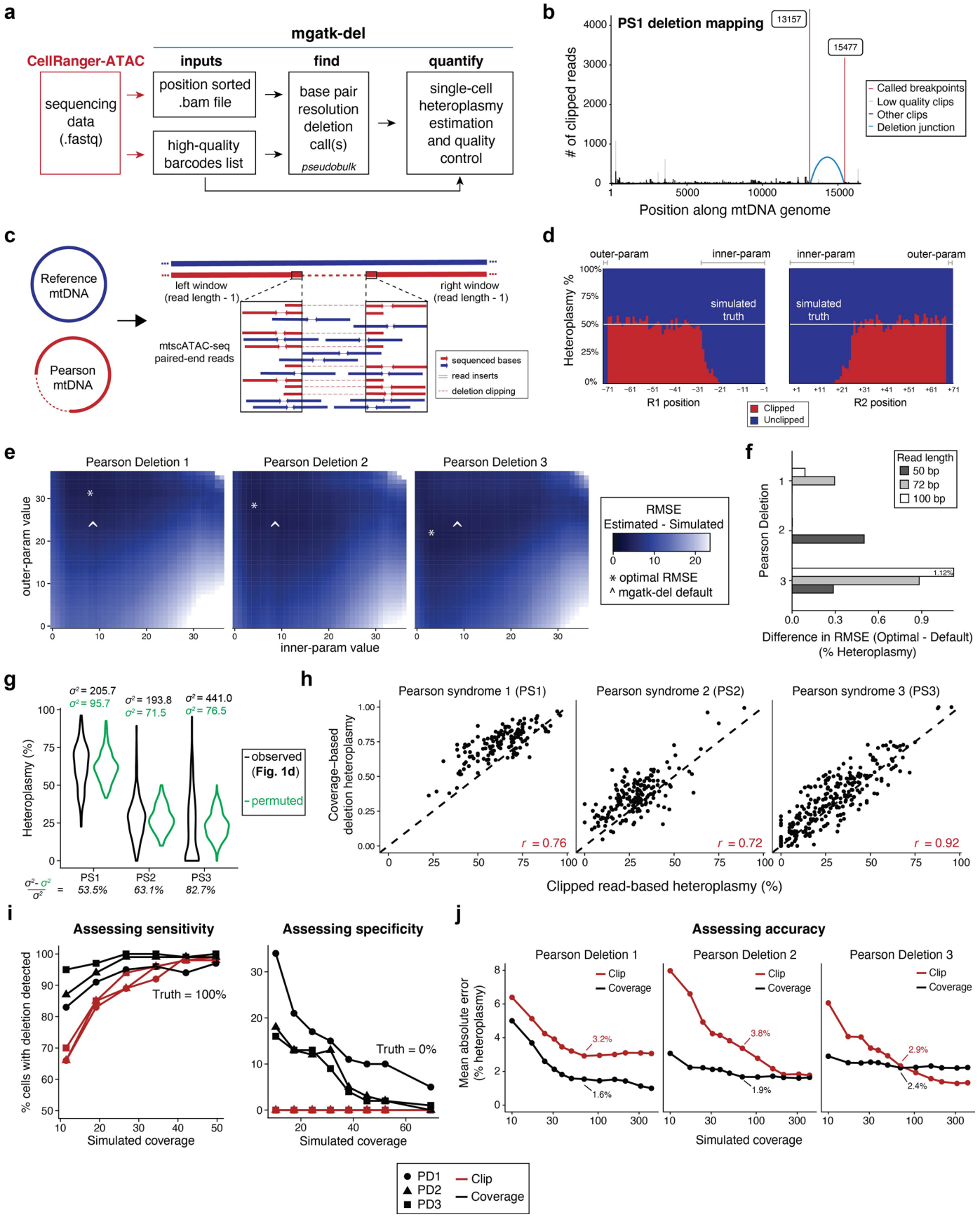
Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01433-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01433-8.

Correspondence and requests for materials should be addressed to Caleb A. Lareau, Aviv Regev, Vijay G. Sankaran, Suneet Agarwal or Leif S. Ludwig.

Peer review information *Nature Genetics* thanks Na Cai, Dan Mishmar, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-023-01433-8

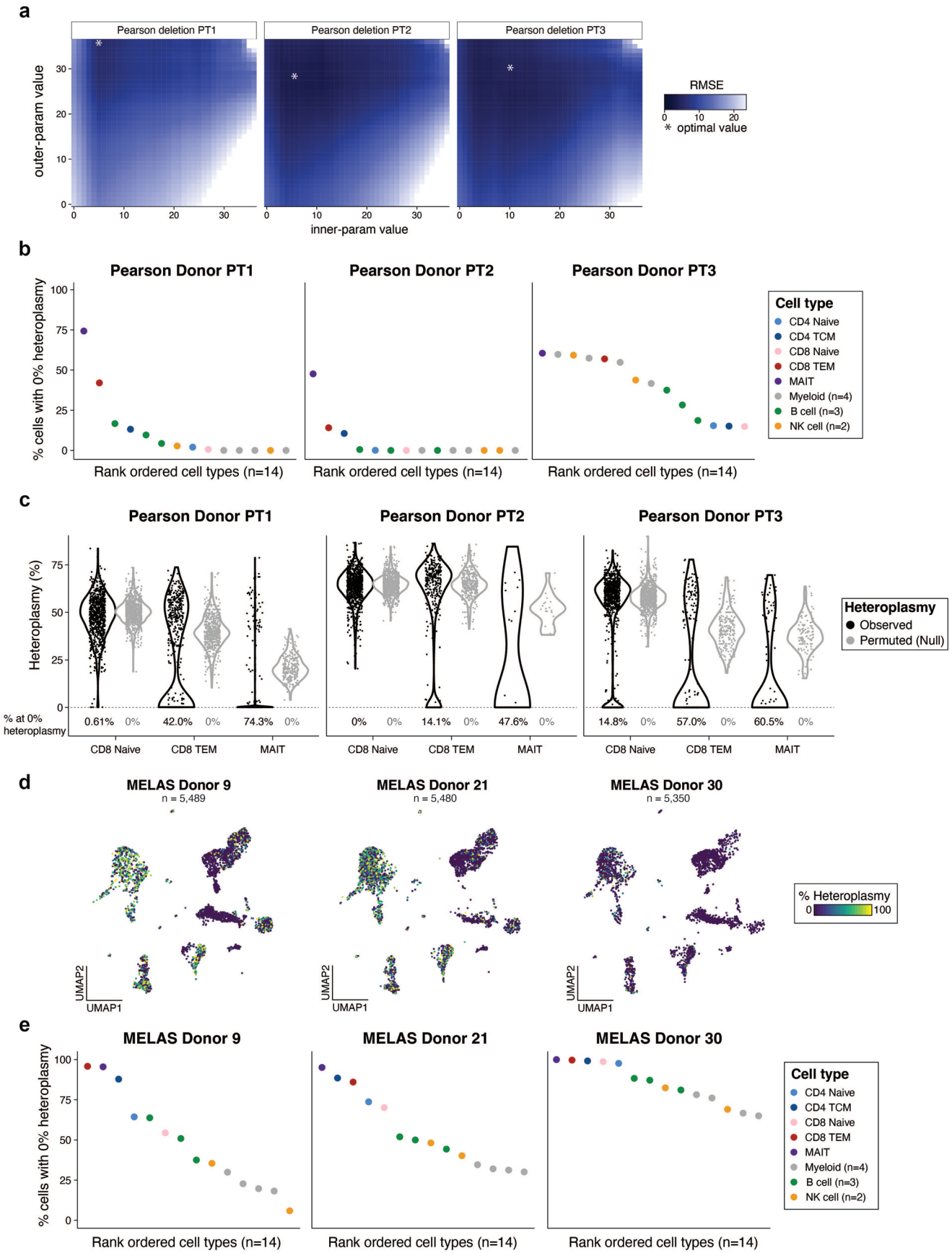


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41588-023-01433-8

Extended Data Fig. 1 | Deletion and heteroplasmy estimation using mgatk-del. (a) Schematic of mgatk-del pipeline, which utilizes the outputs of CellRanger-ATAC. Two critical steps of base-resolution deletion calling ('find') and estimation of single-cell heteroplasmy ('quantify') are illustrated. **(b)** Output of mgatk-del 'find' for Pearson syndrome deletion 1 (PS1). The red vertical lines represent the called deletion breakpoints where the regions were joined (blue arc) via a secondary alignment ('SA' tag in.bam file). **(c)** Schematic of the simulation framework. Synthetic cells with known heteroplasmy were generated via mixtures of reference and PS mtDNA for all previously reported deletions. **(d)** Summary of results from a 50% mix showing the heteroplasmy as estimated from the ratio of clipped to unclipped reads. Parameters 'inner-param' and 'outer-param' define the number of bases that are discarded on the read when estimating the overall heteroplasmy per cell. **(e)** Results of exhaustive simulation for three mtDNA deletions used in the cell mixing experiment. The minimum value of the root mean squared error (RMSE) of the estimated and true heteroplasmy is noted with an asterisk over the grid search. **(f)** Difference in mean estimated heteroplasmy (RMSE) in optimal and default parameters across a variety of settings indicating the stability. **(g)** Decomposition of variance using a permuted model. Black shows the observed variance whereas green shows the spread under a permuted (null) model. The percent of the variance explained by this null model is shown. **(h)** Single-cell correlation of clipped (Fig. 1d) versus coverage-based (Fig. 1e) heteroplasmy estimates for valid deletions per indicated deletion/donor. The Pearson correlation for the three deletions is indicated. **(i)** % of cells with non-zero heteroplasmy for different deletions at different coverages using indicated methods. The left panel assesses sensitivity where the true proportion of cells with the deletion is 100%. The right panel assesses specificity where the true proportion of cells with the deletion is 0%. For both plots, detection of the deletion requires ≥1% heteroplasmy. **(j)** The mean absolute error in heteroplasmy at 50x coverage is indicated by the value shown on the graph for two methods of heteroplasmy estimation, as in (i).

https://doi.org/10.1038/s41588-023-01433-8

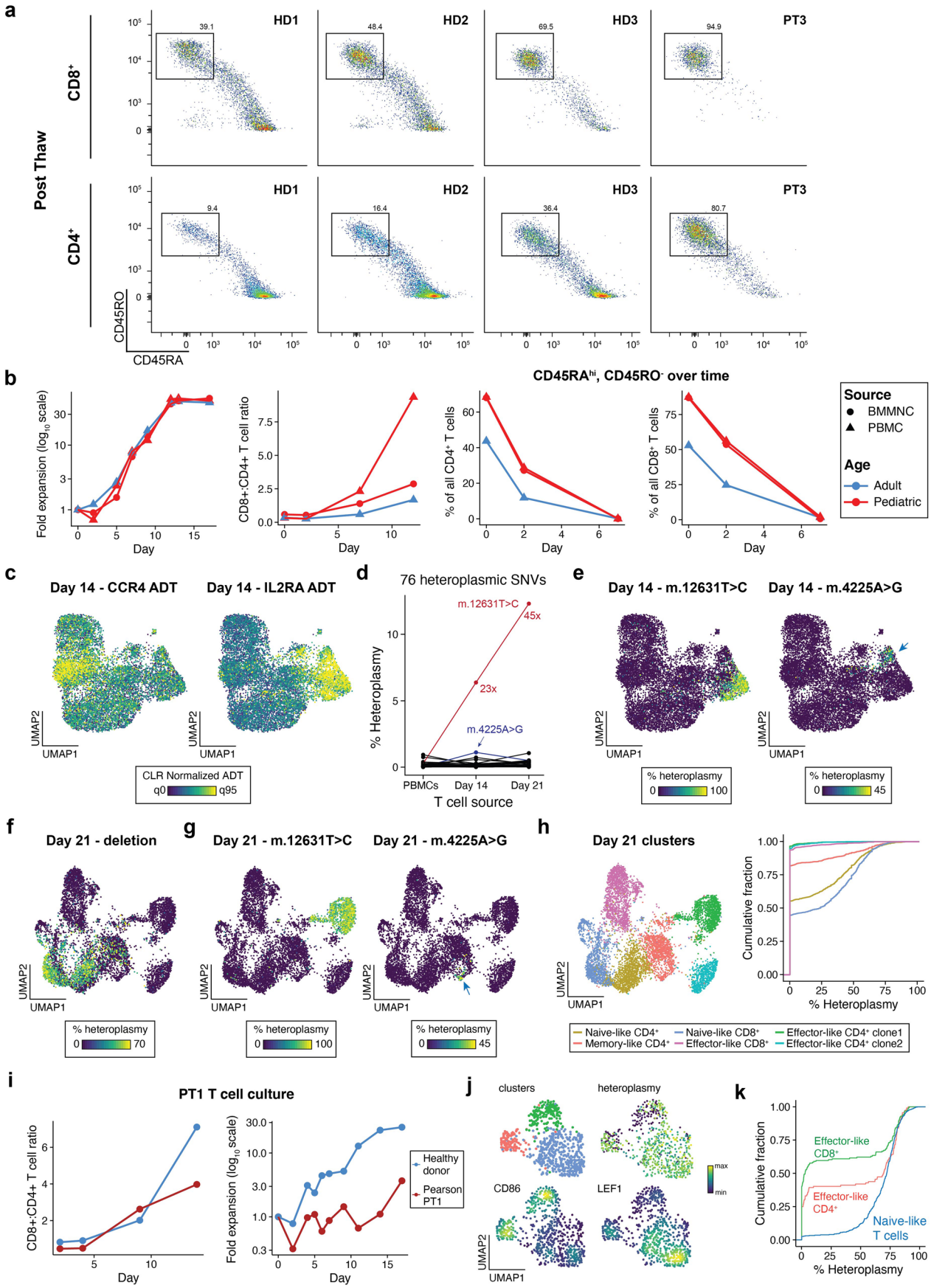


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41588-023-01433-8

Extended Data Fig. 2 | Supporting information for PS PBMC mtscATAC-seq analyses. (a) Result of mgatk-del hyperparameter optimization via a simulation framework. The minimum value of the root mean squared error (RMSE) of the estimated and true heteroplasmy is noted with an asterisk over the grid search. **(b)** Summary of % of cells with 0% heteroplasmy across all hematopoietic cells for the three PS donors. **(c)** Violin plots of each respective mtDNA deletion for all three patients in selected T cell populations. Black indicates the observed data. Gray represents heteroplasmy under a null model of one mean and the variation attributed to differences in coverage per cell. CD8.TEM and MAIT cells have a bimodal distribution (indicating purifying selection) whereas $CD8^+$ naive cells have a distribution that is more consistent with a single mode of heteroplasmy. The percentage of cells with 0% heteroplasmy under observed and null settings are noted for each population below the violins. **(d)** UMAP visualization of MELAS bridge reference projection across three donors previously reported. **(e)** Summary of % of cells with 0% heteroplasmy across all hematopoietic cells for the three MELAS donors[10] with refined cell type annotations from the bridge reference projection.

https://doi.org/10.1038/s41588-023-01433-8

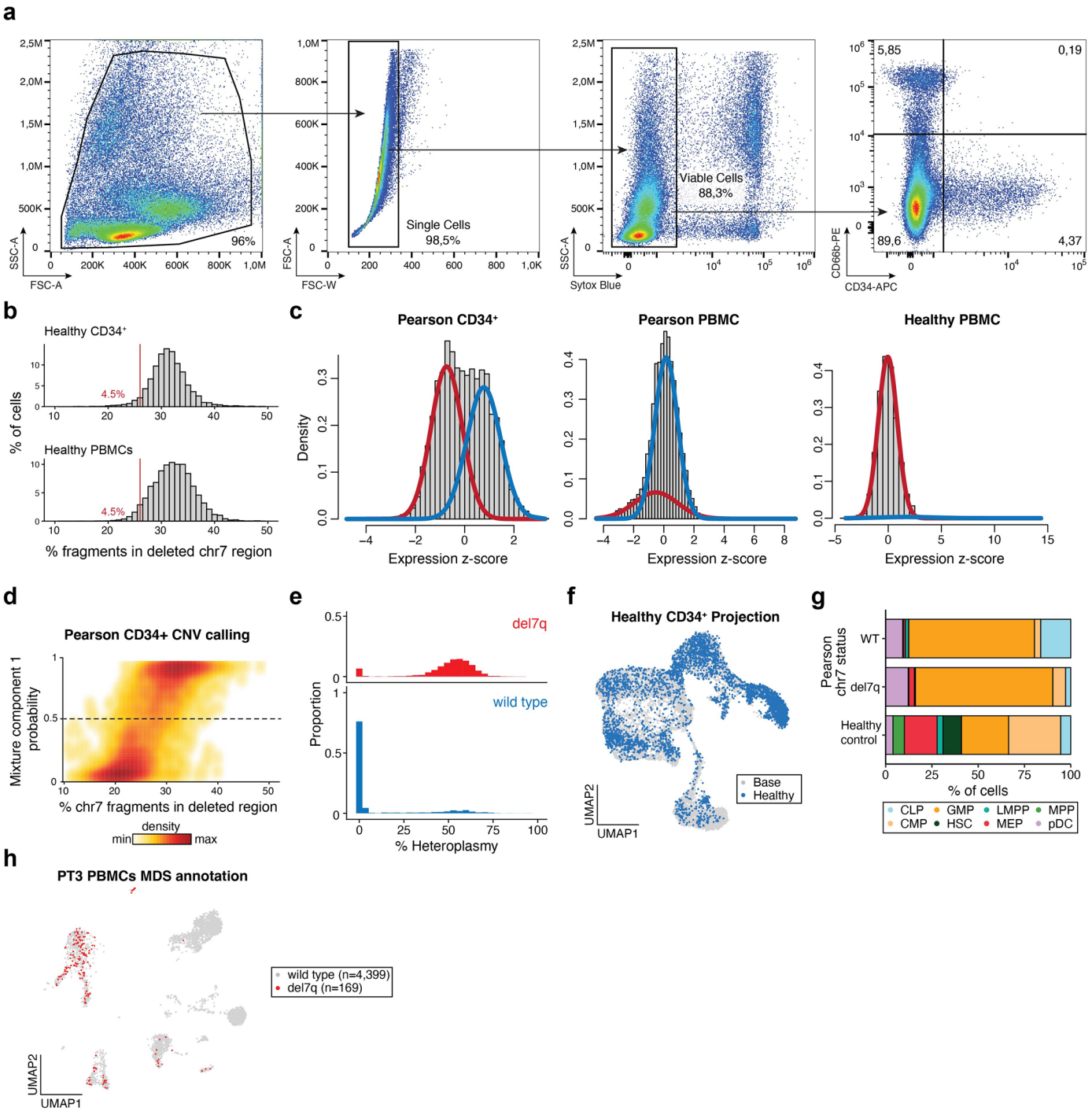


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41588-023-01433-8

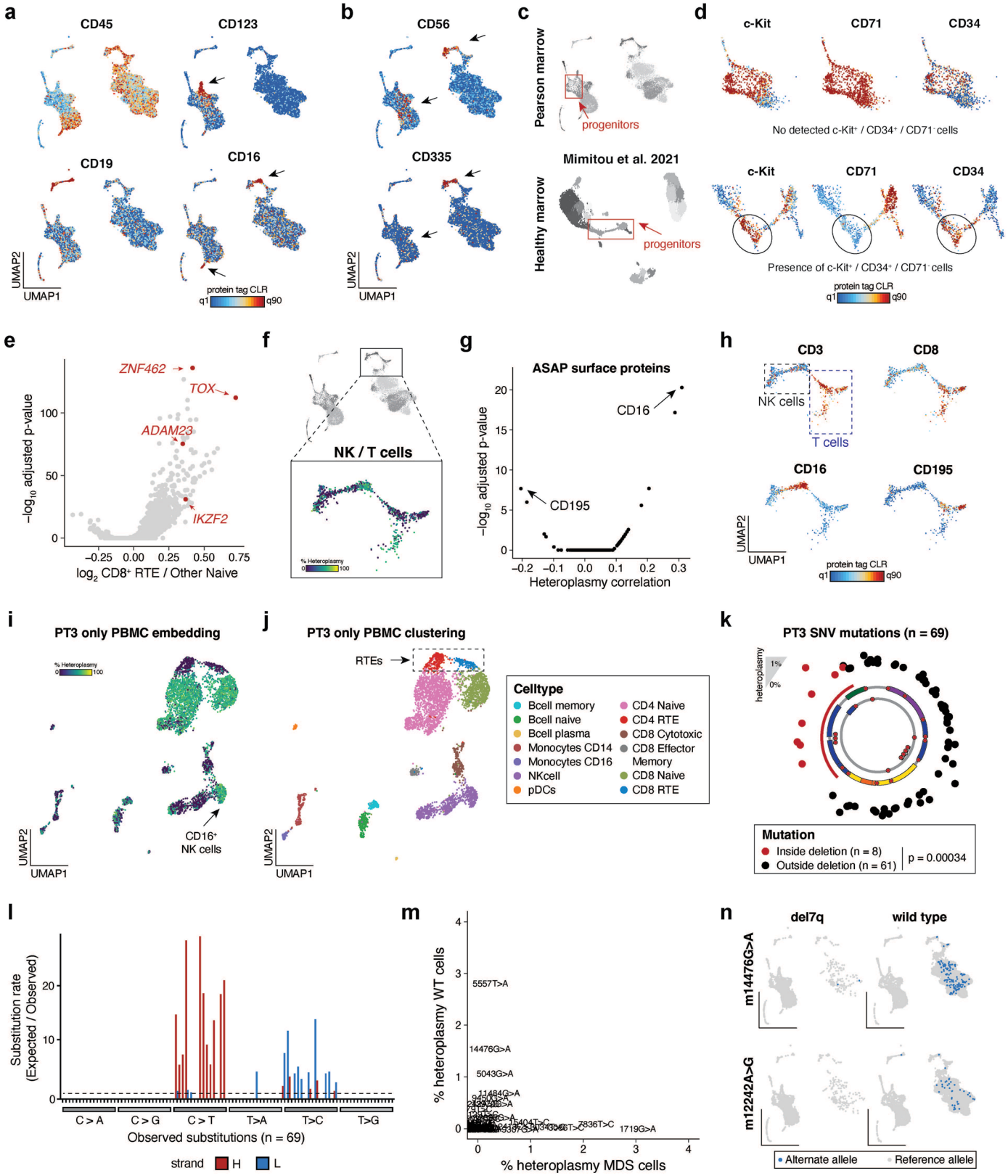
Extended Data Fig. 3 | Supporting analyses for primary PS T cell cultures. **(a)** Representative flow cytometry plot of $CD4^+$ and $CD8^+$ T cells. Cells from the four donors (columns) were assessed for CD45RA and CD45RO expression. The box indicates the proportion of $CD45RA^{hi}/CD45RO^-$ cells summarized in Fig. 3d. **(b)** Summary of flow cytometry data from independent T cell culture comparing cells from a healthy pediatric donor to an adult donor. T cells were derived from both bone marrow mononuclear cells (BMMNCs) and peripheral blood mononuclear cells (PBMCs) for the pediatric donor. **(c)** Embedding of day 14T cells colored by antibody derived tags (ADTs) for CCR4 and IL2RA. **(d)** Dynamics of heteroplasmy for 76 heteroplasmic single nucleotide variants (SNVs) during T cell culture. m.12631T > C and m.4225A > G are highlighted. **(e)** Single-cell heteroplasmy of m.12631T > C and m.4225A > G, variants expanding during T cell culture, annotated in the day 14 embedding. Blue arrow indicates the subpopulation positive for the m.4225A > G variant. **(f)** Single-cell heteroplasmy of the PS deletion annotated on day 21 derived T cells profiled via mtscATAC-seq. **(g)** Single-cell heteroplasmy of m.12631T > C and m.4225A > G, variants expanding during T cell culture, annotated in the day 21 cell embedding. Blue arrow indicates the subpopulation positive for the m.4225A > G variant. **(h)** Day 21 embedding clustering, annotations, and per-cluster heteroplasmy quantified via a cumulative distribution function. **(i)** Independent validation of *in vitro* $CD8^+$ T cell relative expansion (left) and overall T cell proliferation (right) defects from PT1 PBMCs. **(j)** Summary of cell clusters, heteroplasmy, and marker gene scores for PT1 T cells following culture. **(k)** Day 14 PT1 per-cluster heteroplasmy quantified via a cumulative distribution function for PT1 T cells after culture.

https://doi.org/10.1038/s41588-023-01433-8



Extended Data Fig. 4 | Supporting information for del7q calling and CD34$^+$ mtscATAC-seq analyses. (a) Flow cytometry gating strategy for the sorting of live CD66b$^-$CD34$^+$ bone marrow mononuclear cells. **(b)** Summary of 7q fragment abundances in healthy CD34$^+$ and PBMC mtscATAC-seq samples[9]; compare to Fig. 4b with the same cutoff. **(c)** Result of Gaussian mixture model applied to indicated samples. The red trace indicates the first mixture component estimated (lower mean) whereas the blue trace represents the second component with a higher mean. The healthy PBMC sample does not contain a chromosome alteration. **(d)** Graphical density of cells from mixture model (y-axis) and from crude fragment abundance (x-axis; see Fig. 4b). The dotted line indicates the cutoff for wild type and del7q annotations. **(e)** Histograms of mtDNA deletion heteroplasmy proportions (%) stratified on del7q status. **(f)** Projection of a healthy control CD34$^+$ mtscATAC-seq sample onto the reference embedding as shown in Fig. 5d. **(g)** Stacked bar plots of cell type proportions for projected cell types from PT3 with PS/MDS stratified by del7q status (MDS for positive and wild type for negative) and a healthy control donor. **(h)** Annotation of del7q status in PBMCs, which is primarily identified in myeloid, NK, and B-cell populations; see Fig. 2d for cluster annotations.

https://doi.org/10.1038/s41588-023-01433-8

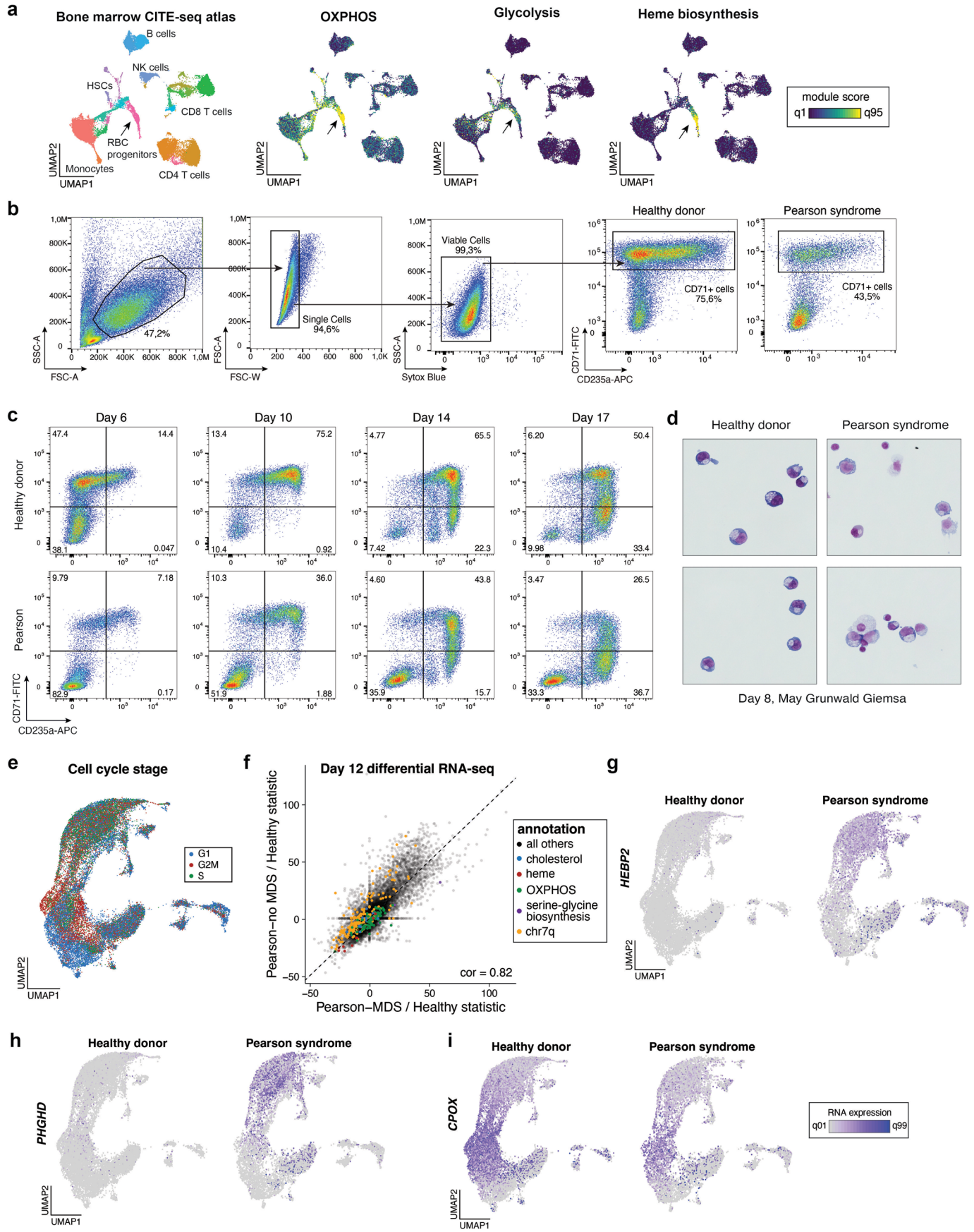


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41588-023-01433-8

Extended Data Fig. 5 | Supporting analyses for PT3 bone marrow mononuclear cell ASAP-seq dataset. (a) Projection of select protein-derived antibody tag abundances for indicated proteins. Select arrows indicate populations positive for the respective marker. UMAP coordinates same as Fig. 6c. **(b)** Projection of protein surface markers CD56 (NK cells and MDS-associated cells) and CD335 (only NK cells) with arrows indicating the two cell populations. **(c)** UMAP of ASAP-seq processed bone marrow mononuclear cells from a PS (top) and a healthy control[11] (bottom) with hematopoietic stem and progenitor cells ('progenitors') indicated in the red boxes. **(d)** Projection of protein tags within the boxed progenitor populations as in (c), contrasting the presence of only $CD71^+$ cells among $CD34^+$/c-Kit^+ cells in PS as compared to the healthy control. **(e)** Volcano plot showing differential gene activity scores for CD8 recent thymic emigrants (RTEs) compared to other CD8 naive T cells. Annotated genes in red represent known marker genes for RTEs. **(f)** Zoom (top) and mtDNA deletion heteroplasmy (bottom) in differentiated CD8 T and NK cells from the BMMNC populations. **(g)** Volcano plot illustrating the association between protein levels and mtDNA deletion heteroplasmy in single cells. P-values were computed from the default two-sided Seurat Wilcoxon test with Bonferroni p-value adjustment. **(h)** Projection of cell state surface markers (CD3, CD8) and top antibody tags (CD16, CD195) as determined in (g). **(i,j)** Reclustering and UMAP depiction of PT3 PBMC mtscATAC-seq data identify **(i)** low heteroplasmy and **(j)** recent thymic emigrants (RTEs). Cell type annotations as indicated. **(k)** Landscape of 69 heteroplasmic somatic mtDNA mutations identified in BMMNC. Statistical test: two-sided Fisher's exact test. **(l)** Substitution rate of mgatk identified heteroplasmic mutations (y-axis) in each class of mononucleotide and trinucleotide change resolved by the heavy (H) and light (L) strands of the mitochondrial genome. **(m)** Scatter plot of 69 somatic mtDNA variants identified in panel (l) stratified based on cells annotated as del7q (x-axis) and wild type for chr7 (y-axis). **(n)** Projection of wild type (diploid chr7)-enriched somatic mtDNA mutations m.14476G > A (50%) and m.12242A > G (25%).

https://doi.org/10.1038/s41588-023-01433-8



Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41588-023-01433-8

Extended Data Fig. 6 | Supporting information for *in vitro* erythroid differentiation experiments. (a) Reference embedding of bone marrow mononuclear cell CITE-seq reference dataset (left) with gene module scores for selected pathways annotated on the UMAP embedding. **(b)** Flow cytometry gating scheme used for sorting of *in vitro* differentiated healthy control and PS cells, related to Fig. 7. **(c)** Flow cytometry plots showing the distribution of CD71 and CD235a surface marker expression of *in vitro* differentiated healthy control and PS cells at indicated days of culture. **(d)** MayGrunwald Giemsa stained cytospins of *in vitro* differentiated healthy control and PS cells at day 8 of culture at 63x magnification. **(e)** UMAP of scRNA-seq data colored by predicted cell cycle state. Cluster annotations as in Fig. 7e–g. **(f)** Comparison of differential gene expression between PT3 donor cells with MDS (x-axis) and without MDS (y-axis) related to healthy control. The Pearson correlation between all genes is annotated (0.82). Genes from relevant pathways or genomic annotations are highlighted in specific colors. **(g-i)** Projection of gene expression of selected differentially expressed genes between PS and healthy control erythroblasts, including **(g)** *PHGDH*, **(h)** *CPOX*, and **(i)** *HEBP2*. Gene expression coloring is scaled for all plots between the first and 99th quantile per gene.

nature portfolio

Corresponding author(s): Lareau, Regev, Sankaran, Agarwal, Ludwig

Last updated by author(s): 3/26/2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	CellRanger-ATAC v2.0; CellRanger v.3.0
Data analysis	Custom code and documentation for mtDNA genotyping, including large deletion genotyping, is available at https://github.com/caleblareau/mgatk. Code to reproduce all analyses is available at https://github.com/caleblareau/pearson_syndrome. Other software for analysis:Seurat v.4.0; Signac v1.0.; FlowJo software v10.4.2; edgeR v3.16

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Sequencing and processed data generated as part of this work is available at GEO accession GSE173936. Databases for gene set enrichment (e.g. GLAD4U; Wikipathways) were accessed via WebGestalt (https://www.webgestalt.org/ v2019)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed a priori. Analyses involved hundreds if not thousands of cells per comparison, providing a robust sample size in-line with similar high-throughput scRNA-seq comparisons and technologies. Nuclei and cell yields were decided based on 10x Genomics manufacturer recommendations for reactions.
Data exclusions	Common scATAC-seq and scRNA-seq quality control cutoffs were applied to identify high quality 'true' cells. For example, cell barcodes failing to meet analysis-specific thresholds were excluded as described in the methods. While the metrics used for exclusion (FRIP; # fragments) were determined ahead of time, the exact thresholds were determined empirically using the density of all single cells to determine appropriate, dataset-specific thresholds.
Replication	A number of the findings were replicated in at least three independent patient derived cell lines or patient specimens from independent donors. In specific, replication across multiple patients was not possible given the unique and rare nature of some of the clinical context (patient with Pearson syndrome and Myelodysplastic syndrome).
Randomization	There were no variables or interventions to randomize in this study. Patient-level covariates were not applicable due to inter-individual variation as the primary source of variation that was analyzed.
Blinding	Blinding is not relevant to our study, as our tools are not dependent on blinding. Our study examined variation of heteroplasmy within an individual, and thus comparisons were not made between individuals for heteroplasmy in cell populations. Investigators could not be blinded during data collection or analysis as there was no intervention. Further, analyses were performed in an exploratory manner where blinding is not possible.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	FITC-conjugated CD71 (clone OKT9, 14-0719-82, eBioscience), APC-conjugated CD235a (Glycophorin A), clone HIR2, 50-153-69, eBioscience), APC-conjugated CD34 (clone 581, 343509, BioLegend), PE-conjugated CD66b (clone G10F5, 305102, BioLegend), AF488-conjugated CD3 (clone OKT3, 317310, Biolegend), PE-conjugated CD4 (clone PRA-T4, 300508, Biolegend), APC-conjugated CD8 (clone SK1, 344721, Biolegend), BV785-conjugated CD45RA (clone HI100, 304139, Biolegend), PE-Cy7-conjugated CD45RO (clone UCHL1, 304229, Biolegend), Fixable Viability Dye eFluor™ 780 (65-0865-18, eBioscience), BV605-conjugated CCR7(clone G043H7, 353223, Biolegend), soluable anti-CD28 (clone 28.2, BioLegend). Cocktail for DOGMA-seq: TotalSeq™-A Human Universal Cocktail, V1.0, #399907, BioLegend)
Validation	Each lot of this antibody is quality control tested by immunofluorescent staining with flow cytometric analysis from Biolegend and eBioscience. Standard methods were used for validation

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Patient derived fibroblast cell lines were obtained at Boston Children's Hospital via standard procedures.
Authentication	We validated the presence of SLSMDs from Pearson Syndrome donors using mgatk-del (via next-generation sequencing).
Mycoplasma contamination	Cell lines are routinely tested for mycoplasma contamination. Results were consistently negative.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used as part of this study.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Primary hematopoietic samples were collected from three previously diagnosed patients, including a 7-year-old male with Pearson syndrome (PS) / Kearns Sayre syndrome ("PT1"), a 4-year-old female with PS ("PT2"), and a 4-year-old male with PS and deletion 7q (del7q) myelodysplastic syndrome (MDS; "PT3"). Other human research subjects were enrolled and collected via the North American Mitochondrial Disease Consortium (NAMDC), which were three female patients aged 23,28,39.
Recruitment	Patients with Pearson Syndrome were recruited to this study by researchers at Boston Children's Hospital. Patients were enrolled with informed consent subject to IRB. As this is an extremely rare disease, no patient-level covariates were considered. Due to a small sample size (n=3 donors), it is unlikely that recruitment biased the interpretation of the data in this study. Other samples from SLSMD were received via a collaboration agreement with NAMDC.
Ethics oversight	Biological samples for cell lines were procured under protocols approved by the Institutional Review Board at Boston Children's Hospital, after written informed consent in accordance with the Declaration of Helsinki. Primary human peripheral blood and bone marrow samples were collected under Boston Children's Hospital Institutional Review Board-approved protocols and written informed consent for genomic sequencing. Healthy CD34+ samples were de-identified and approval for use of these samples for research purposes was provided by the Institutional Review Board and Biosafety Committees at Boston Children's Hospital.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	For flow cytometry analysis and sorting, primary human cells were washed in FACS buffer (1% FBS in PBS) before antibody staining. In vitro cultured primary erythroid cells were stained using 1:50 APC-conjugated CD235a (Glycophorin A, clone HIR2, 50-153-69, eBioscience) and 1:50 FITC-conjugated CD71 (clone OKT9, 14-0719-82, eBioscience) for 15 min on ice. PT3 bone marrow-derived CD34+ cells were stained using 1:40 APC-conjugated CD34 (clone 581, 343509, BioLegend). For mtscATAC-seq and ASAP-seq experiments residual granulocytes were excluded by staining of cells using 1:50 PE-conjugated CD66b (clone G10F5, 305102, BioLegend). For live/ dead cell discrimination Sytox Blue was used at a 1:1000 dilution according to the manufacturer's instructions (Thermo Fisher, S34857).
Instrument	FACS analysis was conducted on a BD Bioscience Fortessa flow cytometer at the Whitehead Institute Flow Cytometry core. Cell sorting was conducted using the Sony SH800 sorter with a 100 µm chip at the Broad Institute Flow Cytometry Facility.
Software	Data was analyzed using FlowJo software v10.4.2.
Cell population abundance	Abundance of cell population was determined by flow cytometry and surface marker analysis and/or via single cell genomic sequencing analysis to determine cell type/state and their relative frequencies within the population.
Gating strategy	Typical gating strategies involved discrimination based on FSC/SSC properties, subsequent exclusion of cell doublets, followed by live/dead cell discrimination and sorting or analysis of surface marker distribution within the live cells.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.