Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Walker MA, Lareau CA, Ludwig LS, et al. Purifying selection against pathogenic mitochondrial DNA in human T cells. N Engl J Med 2020;383:1556-63. DOI: 10.1056/NEJMoa2001265

SUPPLEMENTARY APPENDIX

Purifying Selection Against Pathogenic Mitochondrial DNA in Human T cells.

Walker MA, Lareau CA, Ludwig LS et al.

TABLE OF CONTENTS

References page 16

LIST OF INVESTIGATORS

Melissa A. Walker, Caleb A. Lareau, Leif S. Ludwig, Amel Karaa, Vijay G. Sankaran, Aviv Regev, Vamsi K. Mootha

SUPPLEMENTARY METHODS

Oversight

This study was approved by the Massachusetts General Hospital Institutional Review Board (Protocol 2016P001517).

Single cell accessible chromatin and mitochondrial genotyping

Venous blood was collected from patients at clinical baseline using sodium heparin CPT tubes (BD Biosciences #362753) and peripheral blood mononuclear cells (PBMCs) were purified per manufacturer instructions. PBMCs were cryopreserved prior to use. Upon thawing, cells were stained with a fixable viability dye (Zombie Green, Biolegend #423111) and APC-conjugated anti-hCD45 (Biolegend #304012). After washing, PBMCs were fixed in 1% formaldehyde (FA; ThermoFisher #28906) in PBS for 10 min at RT, quenched with glycine solution to a final concentration of 0.125M before washing cells once with PBS supplemented with 0.4% bovine serum albumin, and subsequently in PBS alone via centrifugation at 400g, 5 min, 4C. Fluorescence-Activated Cell Sorting (FACS) was then performed to exclude dead and non-leukocyte cells.

MtscATAC-seq libraries were generated using the 10x Chromium Controller and the Chromium Single Cell ATAC Library & Gel Bead Kit (#1000111) according to the manufacturer's instructions (CG000169-Rev C; CG000168-Rev B) but with the following modifications: 1.5ml – 2ml DNA LoBind tubes (Eppendorf) were used to wash PBMCs in PBS and downstream processing steps. Cells were subsequently treated with lysis buffer (10mM Tris-HCL pH 7.4, 10mM NaCl, 3mM MgCl2, 0.1% NP40, 1% BSA) for 3 min on ice, followed by adding 1ml of chilled wash buffer and inversion (10mM Tris-HCL pH 7.4, 10mM NaCl, 3mM MgCl2, 1% BSA) before centrifugation at 500g, 5 min, 4C. The supernatant was discarded, and cells were diluted in 1x Diluted Nuclei buffer (10x Genomics) before counting using Trypan Blue and a Countess II FL Automated Cell Counter. If large cell clumps were observed a 40µm Flowmi cell strainer was used

prior to 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 polymerase chain reaction (PCR) as described in the 10x User Guide. After breaking the GEMs, the barcoded tagmented DNA was purified and further amplified to enable sample indexing and enrichment of scATAC-seq libraries. The final libraries were quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and a High Sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent). Paired-end sequencing was performed using an Illumina NextSeq 500 platform using 2x 72 base reads.

Data Analysis

Raw sequencing reads were demultiplexed and aligned to the hg19 reference genome using the CellRanger-ATAC v1.0 software. We identified cells as barcodes that met the following criteria: (1) presence of at least 1,000 unique fragments mapping to the nuclear genome; (2) at least 40% of nuclear fragments overlapping a previously-established chromatin accessibility peak set in the hematopoietic system[1]; and (3) a mean mtDNA coverage of at least 20x. From the output of the CellRanger-ATAC call, we quantified heteroplasmy at all loci, including A3243G, in the mitochondrial genome using the mgatk package, which is available at https://github.com/caleblareau/mgatk. For heteroplasmy analyses, cells with less than 20x coverage at position m.3243 in the mtDNA and outliers with m.3243 coverage of >1.5 interquartile ranges above the third quartile were excluded to avoid artefactual sequencing multiplets.

We applied a computational strategy to identify cell types independent of possible alterations in chromatin accessibility caused by the pathogenic allele. This was achieved by first defining axes of variation in a healthy individual and then projecting new (patient) cells onto this existing space, utilizing Latent Sematic Indexing (LSI) and Uniform Manifold Approximation and Projection (UMAP) as previously described[2]. Specifically, we first generated a binarized matrix of chromatin accessibility peaks for ~10,000 PBMCs derived from a healthy donor[3], which were reduced into 25 dimensions via LSI followed by further reduction to 2 dimensions via UMAP for visualization. Using the 25 dimensions in LSI space we constructed a *k* nearest neighbors graph (*k*=20), and

obtained twelve data-driven clusters by a Louvian community clustering on this graph, which we annotated into five major cell types expected to be observed in PBMCs.

The selection of k=20 was chosen as it serves as a default value consistently used in common single-cell analyses tools, including the statistical frameworks used herein[2,4]. To verify that the results are not sensitive to this choice of parameter, we computed the Adjusted Rand Index (ARI) for values of k = 10, 15, 20, 25, and 30 to compare the clustering results under variable choice of this parameter. An ARI value of 0 is indicative of no concordance between clusters (random) whereas a value of 1 represents perfect concordance. When analyzing these in the context of our data, we found that for all values of k, the ARI to the definitions used in the manuscript exceed 0.9, reflective of very robust results irrespective of the choice of parameter for this value.

Next, we classified all patient cell types by projecting chromatin accessibility data onto this 25-dimensional space and assigning cell types based on minimum distance to cluster medoids. Finally, two dimensional representations of patient data were produced by projecting the 25 LSI dimensions onto the pre-trained UMAP model as previously reported[2]. In our assignment of cells to their closest reference cluster, we used the minimum Euclidean distance between the reference medoid and the individual cell in the reduced dimension space defined by the LSI components. While we did not require a minimum distance for the classification, we did observe a mean 2-fold distance between the individual cells and closest reference cluster medoid (0.011) compared to the second closest cluster medoid (0.025). These results support that the classification was robust in this high-dimensional space.

To test for correlations between A3243G heteroplasmy and our proxy of mtDNA copy number (the ratio of reads aligning to the mitochondrial and nuclear genomes), we calculated Spearman rank correlation coefficients for each dataset in R using cor.test (Package stats version 3.5.1 Index). We estimated 95% confidence intervals from the distributions of the test statistic from 10,000 datasets generated from our observed dataset by bootstrapping with replacement. These computations were performed using the boot function (Package boot version 1.3-23) and the boot.ci function, basic 95% confidence intervals (Package boot version 1.3-23). We calculated

critical values (r_s) for Spearman rank correlation coefficients for $\alpha = 0.05$ as follows: $r_s = \pm z/(\sqrt{n-1})$.

Bulk sequencing and heteroplasmy analysis

We stained cryopreserved PMBCs with anti-human CD45-APC (Biolegend #304012), OKT3 antihuman CD3e -FITC (Biolegend #317305), and antihuman CD56 Pacific Blue™ clone HCD56 (Biolegend #318325). FACS was then used to purify T cell and T cell-depleted PBMC populations from which DNA was extracted (Qiagen #69504). Small amplicons containing the m.3243 locus and surrounding region were generated by PCR using Phusion Mastermix (NEB) per manufacturer instructions and amplifying for 35 cycles at an annealing temperature of 60°C. We then used to generate libraries for sequencing on an Illumina MiSeq platform at Massachusetts General Hospital or using a commercial vendor (Genewiz). Heteroplasmy was called from this data using Samtools[5]. Primer sequences were 5'-CGCCTTCCCCCGTAAATGA-3' (forward), 5'-GGGGCCTTTGCGTAGTTGT-3' (reverse) for amplicon amplification and next generation sequencing.

Author Contributions: MAW designed and performed experiments, analyses, provided clinical insights, and wrote manuscript, CAL, and LSL designed and performed experiments, analyses, and wrote manuscript, AK provided clinical insights, VGS, AR, and VKM designed and supervised experiments, analyses, and wrote manuscript

Supp. Fig. 1

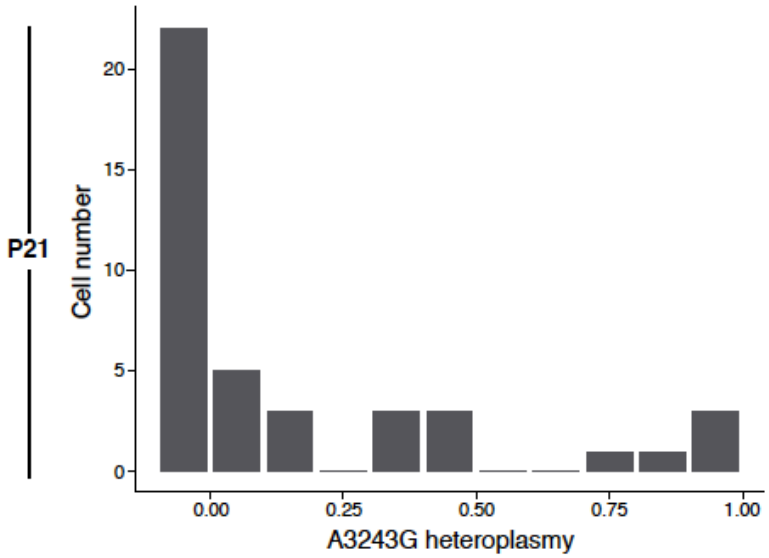

Figure S1. A3243G heteroplasmy across all cell types in patient P21, restricting to cells with ≥ 100x mtDNA. 41 cells in the P21 dataset have ≥ 100x and < 1.5 interquartile ranges above the third percentile coverage at m.3243.

Supp. Fig. 2

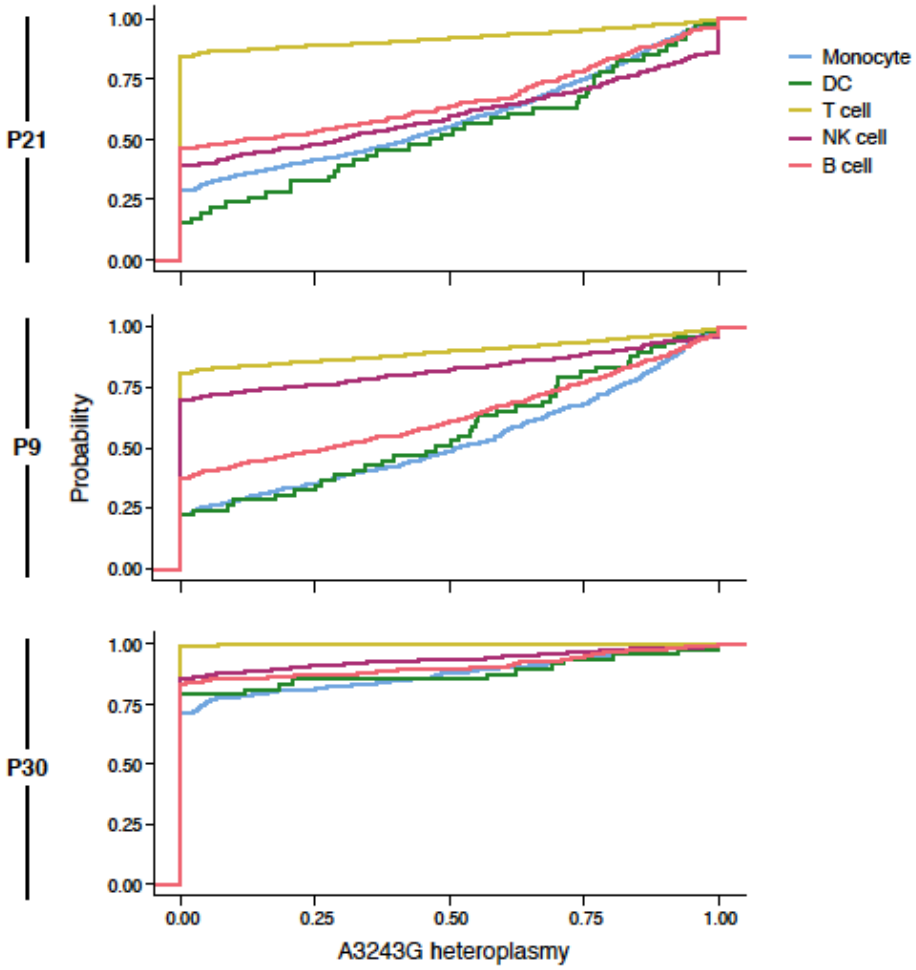

Figure S2. Cumulative distributions of A3243G heteroplasmy in MELAS patients. Cumulative distributions are stratified by cell type for the three indicated patient PBMCs profiled with mtscATAC-seq (DC = dendritic cell, NK= natural killer).

Supp. Fig. 3

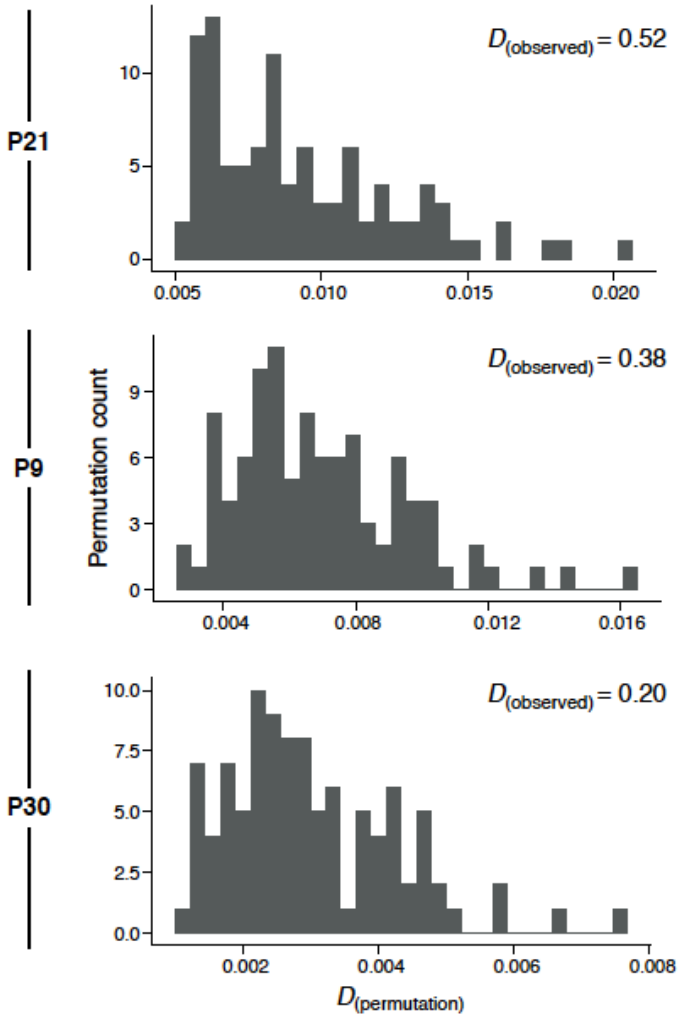

Figure S3. Permutation analysis of the two sample Kolmogorov-Smirnov D statistic. We permutated the cell type label (i.e., T cell or not T cell, preserving the proportion of T cells observed in the respective patient). For each permuted dataset we computed the two-sample K-S test statistic for the heteroplasmy CDF of "T cells" versus "all cells" under the permutation. We repeated this procedure 100 times to generate a null distribution of K-S statistics that can be compared to the statistic obtained with the real data (D_{obs}).

Supp. Fig. 4

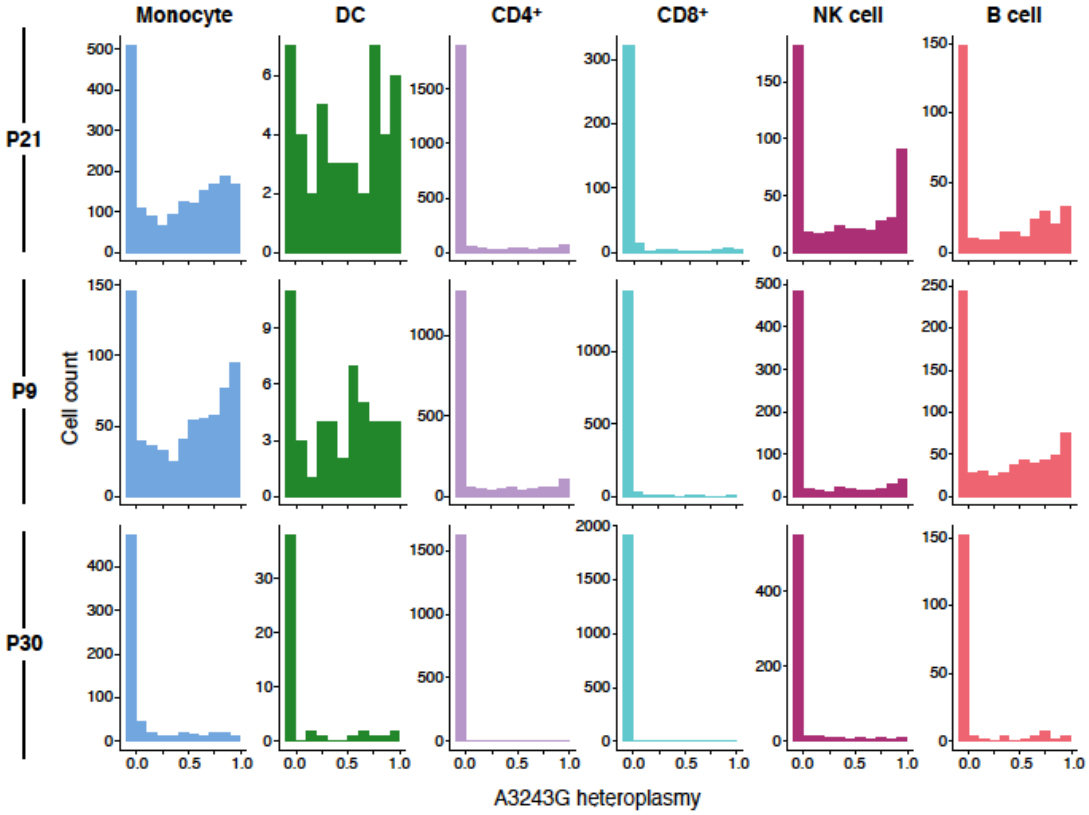

Figure S4. Subdivision of T cell lineages. Histograms show per cell A3243G heteroplasmy fraction in CD4+ and CD8+ T cells compared to other populations (DC = dendritic cell, NK= natural killer).

Supp. Fig. 5

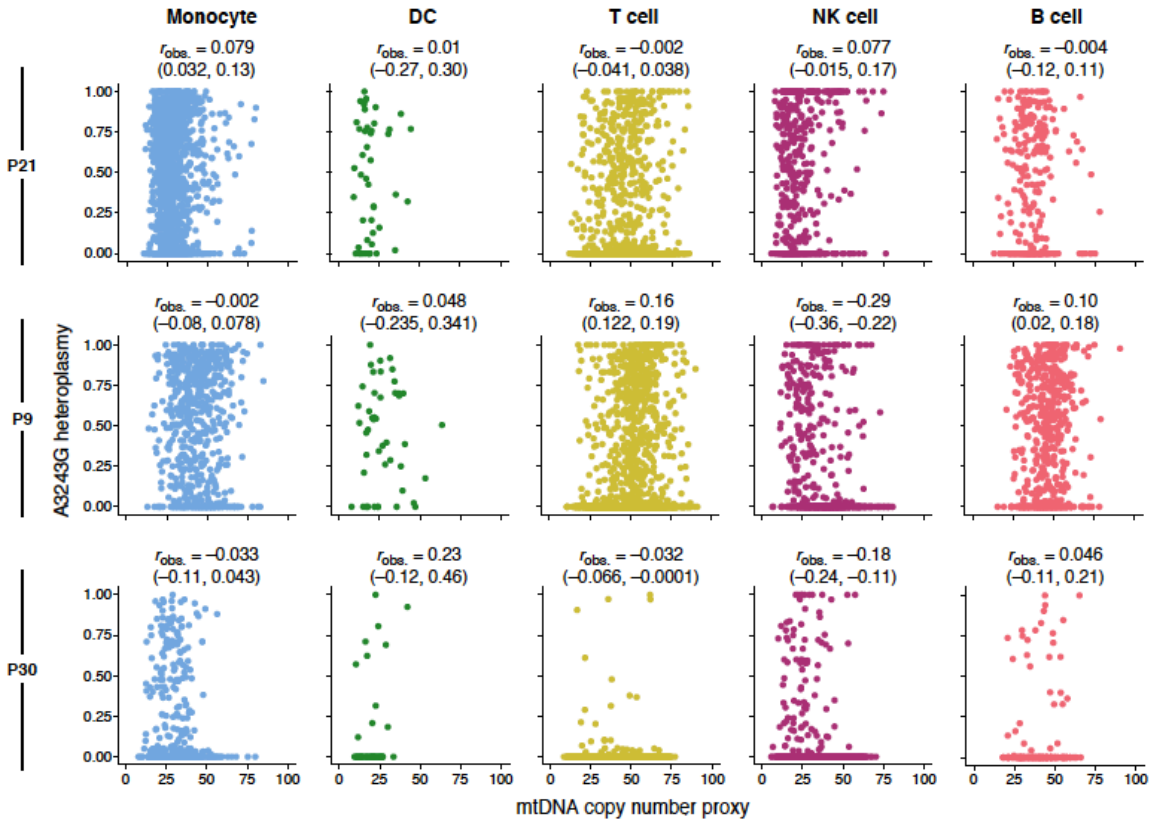

Figure S5. Lack of correlation between A3243G heteroplasmy and mtDNA copy number in major PBMC cell types. For each patient P21, P9, and P30, per cell A3243G percent heteroplasmy (y axis) is plotted against the percentage of reads mapping to the mitochondrial genome (as a proxy of mtDNA copy number (CN) for each patient. Observed Spearman rank correlation coefficients (r_{obs}) are indicated in each panel with bootstrapped 95% confidence intervals shown in parentheses (DC = dendritic cell, NK = natural killer).

Supp. Fig. 6

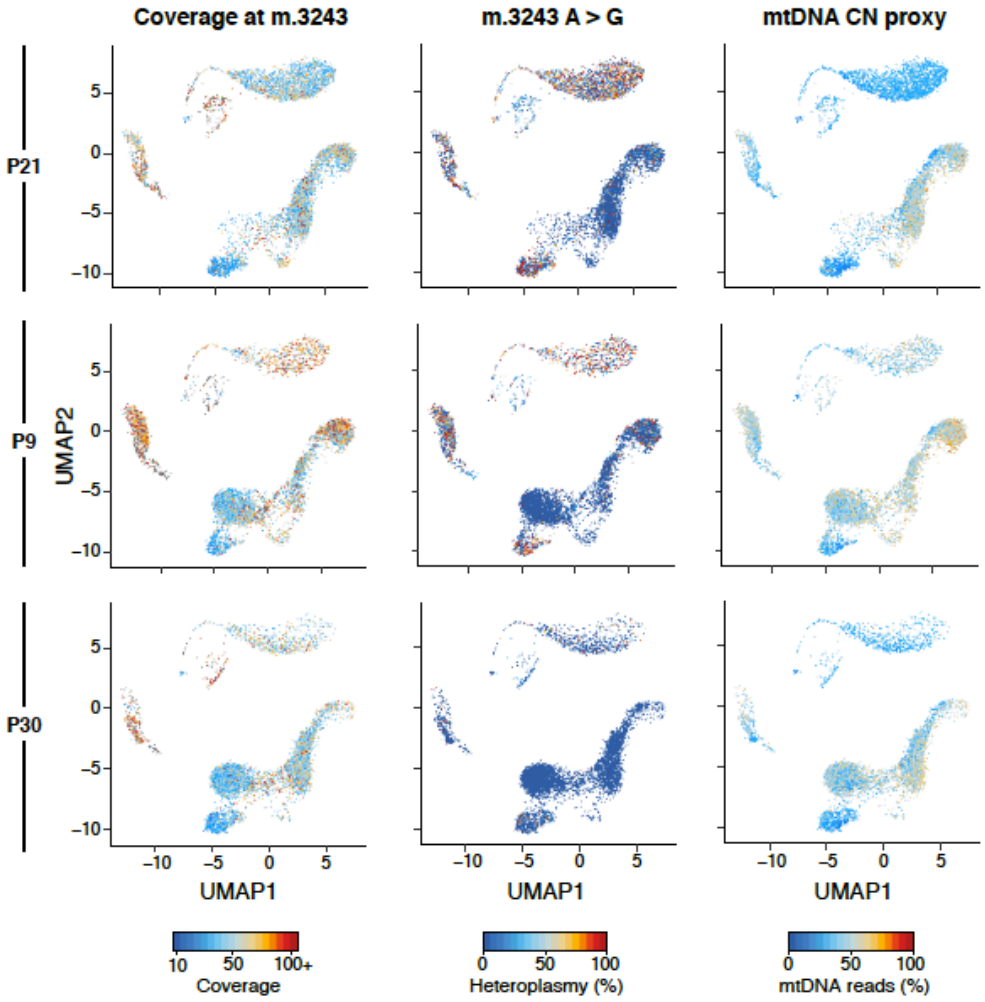

Figure S6. Lack of correlation between A3243G heteroplasmy and mtDNA genome coverage and copy number in PBMCs. UMAPs for each indicated patient's PBMCs are presented colored by mitochondrial genomic coverage at position m.3243 (left column), percentage A3243G heteroplasmy (middle), and percentage of reads mapping to the mitochondrial genome (as a proxy of mtDNA copy number (CN), right).

ID	**Age**	**Sex**	**Blood**	**Oral Rinse**	**Skeletal Muscle**	**Phenotype**
P9	29y	m	39%			stroke, epilepsy, SNHL, urinary dysfunction, cardiomyopathy, HA, ptosis, fatigue
P21	35y	m	+			stroke, FTT, steatohepatitis
P30	60y	m			77%	stroke, cardiomyopathy, ptosis, bilateral SNHL, DM, myopathy
P31	47y	f		25%		SNHL, HA, possible GI dysmotility, autonomic dysfunction, fatigue
P33	65y	f		22.5%		mild myopathy, ptosis, GI dysmotility, deafness, DM, fatigue, exercise intolerance, HA
P36	53y	f	20%			GI dysmotility, HA, burning mouth syndrome, SNHL, fatigue, autonomic dysfunction, myopathy, ptosis
P37	19y	f	46%			seizures, lactate peak on MRS, cardiomyopathy
P38	33y	m	+			DM, hearing loss
P40	35y	m	+			myoclonus, hearing loss

Table S1. Clinical testing results and phenotypes of patients. Clinical heteroplasmy testing results for indicated tissue specimens are summarized (data shown where available). The notation "+" denotes presence of the A3243G mutation by restriction-enzyme based molecular blood testing, without heteroplasmy quantification. Patient clinical phenotypes are summarized. Abbreviations include: m = male, f = female, SNHL = sensorineural hearing loss, HA = headache, FTT = failure to thrive, DM = diabetes mellitus, GI = gastrointestinal, MRS = magnetic resonance spectroscopy.

Component	Reference Range	P9 Mean	SD	P21 Mean	SD	P30 Mean	SD	P31 Mean	SD	P33 Mean	SD	P36 Mean	SD	P37 Mean	SD	P38 Mean	SD	P40 Mean	SD
WBC	4.5 - 11.0 K/uL	10.63	1.92	10.29	1.36	10.40	0.92	9.55	5.61	5.63		7.21	0.48	9.43	1.54	6.3		8.55	2.69
RBC	4.50 - 5.90 M/uL	4.69	0.23	4.86	0.12	2.82	0.14	4.93	0.11	4.54		4.26		3.80	0.37	5.2		5.19	0.06
Hgb	13.5 - 17.5 g/dL	13.31	0.71	14.30	0.40	8.16	0.42	13.33	0.36	13.70		12.80	0.42	11.39	1.08	15.3		15.55	0.21
HCT	41.0 - 53.0 %	38.57	2.23	41.07	1.33	26.30	1.47	41.30	1.33	39.40		38.45	1.06	34.45	2.61	43.2		45.90	0.71
PLT	150 - 400 K/uL	241.78	29.07	245.00	29.05	311.75	37.74	302.00	15.03	283.00		272.50	0.71	344.45	82.73	219		224.00	1.41
MCV	80.0 - 100.0 fL	82.16	1.84	84.43	1.29	93.15	0.82	83.73	1.20	86.80		93.50	2.12	90.80	2.51	83		88.45	0.35
MCH	26.0 - 34.0 pg	28.36	0.44	29.40	0.85	28.91	0.49	27.03	0.17	30.20		30.80		29.97	0.36	29.3		29.95	0.78
MCHC	31.0 - 37.0 g/dL	34.55	1.09	34.83	0.67	31.03	0.49	32.28	0.52	34.80		33.30	0.14	33.02	0.77	35.4		33.90	0.99
RDW	11.5 - 14.5 %	12.29	0.07	12.17	0.12	17.54	0.45	14.45	0.30	12.10		13.15	0.49	12.02	0.18	13.1		12.45	0.35
MPV	8.4 - 12.0 fl	9.44	0.16	9.70	0.10		0.71	9.70	0.23	9.20				10.09	0.31	9.9		9.90	0.28
NRBC% (auto)	0 - 0.20 /100 WBCs	0.00	0.00	0.00	0.00		0.56	0.00	0.00	0.00		0.00	0.00	0.00	0.00			0.00	0.00
NRBC#, auto	0 - 0.01 K/uL	0.00	0.00	0.00	0.00		0.59	0.00	0.00	0.00		0.00	0.00	0.00	0.00			0.00	0.00
Neutro	40 - 70 %			58.25	7.28	75.74	0.29	65.83	15.43									73.00	
Lymphs	22 - 44 %			23.40	5.52	14.54	0.72	22.50	12.30									21.80	
Monos	4 - 11 %			10.85	0.35	4.71	0.15	8.70	2.36									4.50	
Eos	0 - 8 %			5.10			0.04	2.23	1.46									0.30	
Basos	0 - 3 %			0.50	0.00	1.07	0.08	0.37	0.06									0.10	
Granulo, imm (%)	0.0 - 0.9 %			0.40		7.87	0.03	0.37	0.12										
Neutro#	1.8 - 7.7 K/uL			6.45	1.16	1.51		7.64	6.37									7.63	
Lymph#	1.0 - 4.8 K/uL			2.57	0.46	0.49		1.87	0.34									2.28	
Mono#	0.2 - 1.2 K/uL			1.20	0.11	0.41		0.85	0.32									0.47	
Eos#	0.0 - 0.9 K/uL			0.73	0.21	0.11		0.18	0.06									0.03	
Baso#	0.0 - 0.3 K/uL			0.06	0.01			0.04	0.01									0.01	
Granulo, imm	0.00 - 0.10 K/uL			0.05				0.04	0.04										

Table S2. Patient clinical complete blood cell counts (where available). The mean value of all measured parameters is reported with standard deviation (SD) when multiple measurements were available. WBC = white blood cells, RBC = red blood cells, HGB = hemoglobin, HCT = hematocrit, PLT = platelets, MCV = mean corpuscular volume, MCH = mean corpuscular hemoglobin, MCHC = mean corpuscular hemoglobin concentration, RDW = red cell distribution width, MPV = mean platelet volume, NRBC= nucleated red blood cell, NEUTRO = neutrophils, LMYPHS = lymphocytes, MONOS = monocytes, EOS = eosinophils, BASOS = basophils, GRANULO, IMM = granulocytes, immature, k = thousand, uL = microliter, g = gram, dL = deciliter, fl = femtoliter

REFERENCES:

1. Ulirsch JC, Lareau CA, Bao EL, et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. Nat Genet 2019;
2. Granja JM, Klemm S, McGinnis LM, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat. Biotechnol. 2019;
3. Satpathy AT, Granja JM, Yost KE, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat Biotechnol 2019;
4. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. Cell 2019;
5. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;