Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mitochondrial mutation, drift and selection during human development and ageing

Peter Campbell (pc8@sanger.ac.uk)
Wellcome Trust Sanger Institute https://orcid.org/0000-0002-3921-0510

Michael Spencer Chapman
Wellcome Trust Sanger Institute

Moritz Przybilla
Wellcome Sanger Institute https://orcid.org/0000-0001-5645-9492

Andrew Lawson
Wellcome Sanger Institute https://orcid.org/0000-0003-3592-1005

Emily Mitchell
Wellcome Sanger Institute

Kevin Dawson
The Cancer, Ageing and Somatic Mutation Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA

Nicholas Williams
Wellcome Trust Sanger Institute https://orcid.org/0000-0003-3989-9167

Luke Harvey
Wellcome Sanger Institute

Anna Maria Ranzoni
Wellcome Sanger Institute

Ana Cvejic
Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, Cambridge, UK https://orcid.org/0000-0003-3204-9311

Krishna Mahbubani
University of Cambridge https://orcid.org/0000-0002-1327-2334

Kourosh Saeb-Parsy
University of Cambridge https://orcid.org/0000-0002-0633-3696

Anthony Green
University of Cambridge

Jyoti Nangalia
Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK https://orcid.org/0000-0001-7122-4608

Elisa Laurenti
University of Cambridge https://orcid.org/0000-0002-9917-9092

Inigo Martincorena
Wellcome Sanger Institute https://orcid.org/0000-0003-1122-4416

Article

Keywords:

Posted Date: June 26th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3083262/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations:
Yes there is potential Competing Interest. I.M and P.J.C. are co-founders, stockholders and consultants for FL86 Inc.

Supplementary Tables are not available with this version.

Mitochondrial mutation, drift and selection during human development and ageing

Authors:

Michael Spencer Chapman[1,2,3]*, Moritz J. Przybilla[1]*, Andrew R. J. Lawson[1], Emily Mitchell[1,2,3], Kevin Dawson[1], Nicholas Williams[1], Luke M. R. Harvey[1], Anna Maria Ranzoni[1,2,3], Ana Cvejic[1,2,3,4], Krishnaa Mahbubani[5,6], Kourosh Saeb Parsy[5,6], Anthony R Green[2,3], Jyoti Nangalia[1,2,3], Elisa Laurenti[2,3], Iñigo Martincorena[1], Peter J Campbell[1,2]

* These authors contributed equally to this work

Affiliations:

1. Cancer, Ageing and Somatic Mutation Programme, Wellcome Sanger Institute, Hinxton, UK.
2. Wellcome–MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, Cambridge, UK.
3. Department of Haematology, University of Cambridge, Cambridge, UK.
4. Biotech Innovation and Research Centre, University of Copenhagen, Copenhagen, Denmark
5. Department of Surgery, University of Cambridge, Cambridge, UK.
6. Cambridge Biorepository for Translational Medicine, NIHR Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK.

Address for correspondence:

Dr Peter Campbell,
Cancer, Ageing & Somatic Mutation Programme,
Wellcome Sanger Institute,
Hinxton CB10 1SA,
United Kingdom
e-mail: pc8@sanger.ac.uk

Abstract:

Progressive dysfunction of mitochondria, organelles responsible for energy provision to cells, is a major hallmark of ageing. How the dysfunction steadily accrues over a lifetime remains unclear, given the transient nature of free radical damage and the high turnover of mitochondria. Here, we leveraged whole-genome sequencing data from single-cell derived foetal and adult haematopoietic stem/progenitor cell colonies to study the clonal dynamics of turnover and selection in mitochondrial genomes throughout life. We found that genetic drift and independent convergence of mitochondrial mutations complicate their use as lineage marks in single-cell sequencing experiments. Point mutations accrued linearly with age at an average rate of 0.007 mutations/genome/year. Using the distribution of mtDNA allele frequencies with age, we infer that a cell's complement of mitochondrial genomes is replicated every 4-19 weeks in adults, with faster turnover rates during foetal development. Clock-like accumulation of mutations coupled with rapid turnover induces complex evolutionary dynamics in mitochondria as individuals age. Nonsense mutations are disadvantageous regardless of their heteroplasmy level. Missense variants, however, are positively selected at low heteroplasmy but negatively selected at high heteroplasmy, suggesting that some mutations improve fitness of individual mitochondria, even though fitness of the whole cell deteriorates if their expansion proceeds too far. Thus, age-related decline in mitochondrial function can arise from preferential cellular accumulation of selfish mitochondrial clones whose superior fitness ultimately disadvantages the host cell.

Main text:

Introduction

Mitochondria have a codependent relationship with eukaryotic cells, being responsible for the regulation of the Krebs cycle and oxidative phosphorylation[1]. They have maintained their own genome, which in humans is ~16.6kb in size, typically present at 1,000–10,000 copies in each cell. The numbers of mitochondria in a cell increase (mitochondrial fission) or decrease (mitophagy, mitochondrial fusion) throughout the cell cycle – at mitosis, mitochondria, and therefore their genomes, then assort to the two daughter cells[2]. Drift of mitochondrial clones over time can arise from uneven partitioning of mitochondria between daughter cells at cell division (known as vegetative segregation) or through uneven turnover outside cell division (termed relaxed replication), the latter occurring even in post-mitotic cells[3]. Mutations in individual copies of the mitochondrial genome may be acquired somatically, predominantly arising through errors during replication of mitochondrial DNA (mtDNA)[4], reported to be 10-100-fold higher than that of the nuclear genome[5–8]. The high mutation rates of mtDNA, coupled with the complex dynamics of mitochondrial turnover, mean that a given mutation may be present in all mtDNA copies of the cell (homoplasmy), or in a subset (heteroplasmy), with heteroplasmic variants being subject to genetic drift and selection.

Mitochondrial dysfunction is a major hallmark of ageing in humans and other species, associated with decreased efficiency of the respiratory chain[9]. Reactive oxygen species produced by the electron transport chain directly damage mitochondria, but have a half-life measured in milliseconds. To produce age-related deterioration of mitochondrial function, accumulating over decades, such damage must cause long-term, heritable changes in the cell, especially given the active destruction and replenishment of mitochondria in all cells. Several hypotheses for the nature of these long-term changes have been advanced – reactive oxygen species may elicit long-lived, progressive cellular responses that lead to age-related deterioration[10]; oxidative stress at low levels may induce cellular defence mechanisms that are initially adaptive but at higher levels contribute to cellular damage and ageing (so-called mitohormesis)[11]; mitochondria with mutations that slow the respiratory chain may expand within a cell, leading to cumulative dysfunction[12].

Germline mitochondrial mutations, inherited through the maternal oocyte, cause human diseases, with >100 mtDNA single nucleotide variants (SNVs) identified in clinical disorders, together affecting ~1 in 8,000 individuals[13]. Due to the high functional percentage of the mitochondrial genome, with 13 protein coding, 22 tRNA and 2 rRNA genes, many variants could be subject to selective pressures[14] and, as a result, the dynamics of heteroplasmic drift and selection have been thoroughly investigated for germline mtDNA variants[15–17]. Recent large-scale sequencing studies have investigated the patterns of heteroplasmic mutations in kinships[18] as well as across a broad variety of sample sources in health and disease[19], aiming to understand the transmission to offspring and their prevalence in the population.

Mutations in mtDNA may also accumulate somatically, evidenced by studies in either normal somatic tissues[20–22] or in cancers[4,23–25]. These studies have demonstrated a progressive accumulation of somatic mitochondrial mutations with age, occurring with the same distinctive mutational spectrum as observed in the germline and suggesting complex dynamics of positive and negative selection on individual variants. However, the estimates and analyses described in these studies rely on assessments in bulk tissues, which typically demonstrate a

polyclonal mix of cells – this has the difficulty that heteroplasmy levels in individual clones cannot be measured accurately, making inference of drift dynamics challenging. To be able to investigate the true degree of somatic mutation, drift and selection, analyses at the level of individual cells or clones are necessary.

Recently, we published large-scale genomic studies of single haematopoietic stem and progenitor cells (HSPCs)[26,27]. Here, using these data, we combined whole-genome sequences from 4,217 colonies across 12 subjects, leveraging the unprecedented resolution enabled by single-cell-derived colonies to quantify clone-specific rates of mitochondrial somatic mutation, turnover, drift and selection. We demonstrate 270-fold higher rates of somatic mutation in mtDNA than the nuclear genome; a rapid rate of mitochondrial genome turnover every 4-19 weeks; and antagonistic selection pressures acting at the levels of individual mitochondria versus the whole somatic cell. Together, these forces shape the evolution of mitochondrial genomes over the human lifespan, explaining how decline in mitochondrial function can both progressively accumulate with age and be heritable from mother to daughter cell.

Results

The dataset comprises whole-genome sequencing (WGS) data from 4,217 single HSC/HPC-derived colonies of 12 haematopoietically normal individuals: two foetuses (8 and 18 post-conception weeks [pcw])[27], two neonates, and eight adults[26] (Supplementary Table 1). In brief, HSCs/MPPs were isolated from multiple anatomical locations and cultured on cytokine-supplemented medium. Genomes were sequenced to a mean nuclear genome coverage of 14.6×, enabling somatic mutation calling from the nuclear genome with established variant-calling algorithms[28,29].

SNVs in the nuclear genome were used to construct high-confidence phylogenetic trees, which were stable across phylogeny-building algorithms and bootstrapping approaches[26,27]. The mitochondrial genome was represented at high coverage across all samples, with mean coverage of 6,008x and >99% of samples achieving >1,000× (Extended Data Fig. 1a). Although we observed uneven coverage at some positions in the mitochondrial genome, the median coverage was consistently over 1000× at all positions enabling reliable somatic mutation calling (Extended Data Fig. 1b).

Mitochondrial copy number and point mutations throughout life

We inferred the average mtDNA copy number (CN) in each colony, using the ratio of mitochondrial to nuclear genome coverage[4,25,30]. This revealed consistent median values between individuals (range of median: 192-896), with no age correlation (Fig. 1a, Extended Data Fig. 1c; Supplementary Table 2). However, we observed broad within-individual variation, reflected by a log-normal distribution. These differences were related to cellular composition of colonies (erythroid, myeloid or monocytic), with monocytic or myelomonocytic colonies having 2.5-fold higher CN values than erythroid colonies on average (Fig. 1b; Supplementary Table 3). In keeping with this, white cells have previously been demonstrated to influence mtDNA CN measurements in peripheral blood[31]. No robust phylogenetic signal for CN was evident after correction for multiple hypothesis testing (Extended Data Fig. 1d).

We identified potential somatic mutations in the mitochondrial genome using a site-specific error model, removing artefacts arising from sample contamination or mis-mapping of reads from the nuclear genome (Extended Data Fig. 1e). Previous studies of mitochondrial mutations have been limited to variant allele fractions (VAFs) >1% [4,25,32], but the high coverage of mtDNA and single-cell origin of colonies in our study meant that calling of variants with lower VAFs was possible. The consistency of the mutational signature in mitochondria across individuals, with a predominance of C>T and T>C transitions and significant heavy-/light-chain strand bias, as previously described, enabled the precise assessment of true somatic variants using signature decomposition (Extended Data Fig. 2a: 'N1')[4,25,33]. Reassuringly, mutations attributable to this characteristic signature were evident down to 0.1% VAF, suggesting that variants at very low heteroplasmy can be identified. All other mutational signatures identified were strongly enriched at very low VAFs, suggesting they are signatures of sequencing and amplification errors (Extended Data Fig. 2a-b). For this reason, the final dataset of mtDNA point mutations comprises the 9,200 mutations for which the assigned signature was the canonical 'N1' signature.

We estimated the average number of mutations per individual copy of the mitochondrial genome for each colony, calculated as the sum of VAFs of mtDNA mutations for that colony (Fig. 1c). Interestingly, within a given individual, there was broad variation in mutation burden,

mirroring observations for the CN variation, evenly following a log-normal distribution. There was a gradual increase in mean mutation burden per mitochondrial genome with age (p=4.7×10^{-9}, R^2 = 0.97, linear regression; Fig. 1d). Low VAF mutations were present in most colonies even at birth (64% in cord blood), while homoplasmic mutations were rare until old age (<6% below 40 years of age, ~20% by 75-80 years) (Fig. 1e).

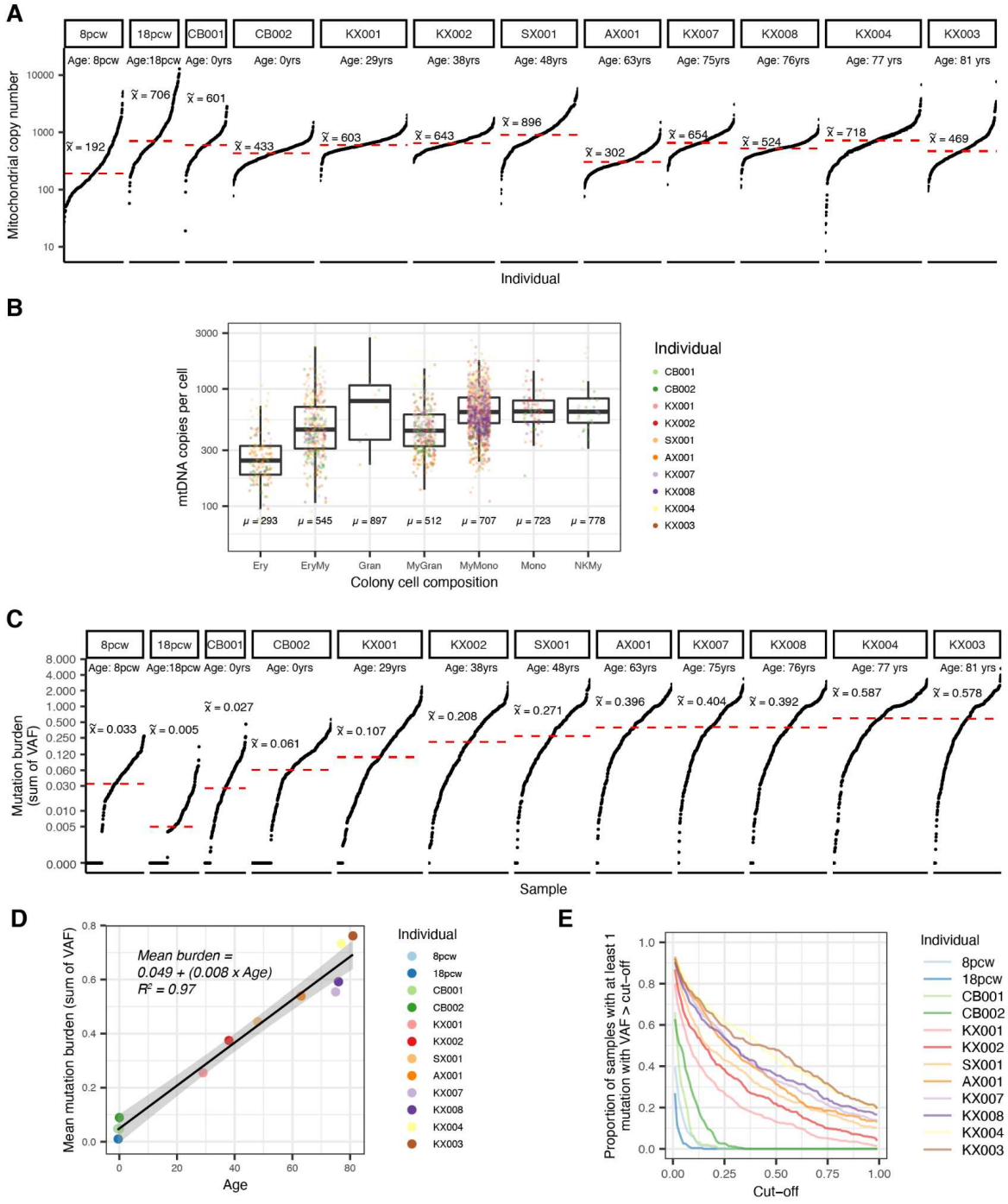


Fig. 1| Genomic variation in the mitochondrial genome throughout life. a, Mitochondrial copy number (log10, y-axis) across all samples per patient (x-axis). Patients have been ordered by increased age. Horizontal red line indicates the median copy number across all samples within a patient annotated above. **b**, Mitochondrial copy number (y-axis) for each HSC-derived colony where the colony cell type composition was available. **c,** Mitochondrial mutation burden represented as the sum of variant allele fraction (VAF) across all samples per patient. Patients are ordered by increasing age. **d,** Correlation between median mitochondrial mutation burden (y-axis) and age (x-axis). The black line highlights a linear regression with gray boundaries indicating the standard error. Each dot is coloured according to the patient identity. **e,** Proportion (y-axis) of samples per patient with at least 1 mutation at a VAF higher than our detection limit. The x-axis indicates the VAF level of a respective mutation. (Ery - Erythroid, Gran - Granulocytic, EryMy - Erythroid/ myeloid, MyGran - Myeloid/ Granulocytic, MyMono - Myeloid/ Monocytic, Mono - Monocytic, NKMy - Natural Killer/ Myeloid)

Evidence of negative and positive selection in mitochondria

To measure selection acting on the mitochondrial genome, we calculated d*N*/d*S* ratios, while accounting for the trinucleotide spectrum and replication strand asymmetry of mtDNA[34] (Fig. 2a; Extended Data Fig. 3a-b; Supplementary Table 4). d*N*/d*S* is a normalised ratio of non-synonymous to synonymous mutations such that a score of 1.0 represents overall neutrality (or exact balance between positive and negative selection), whereas a ratio <1.0 suggests an excess of negative selection over positive selection and vice versa for d*N*/d*S* >1.0. We found strong negative selection on nonsense mutations across all age groups with d*N*/d*S* ratios consistently close to 0, compatible with rapid elimination of mitochondrial genomes containing protein-truncating mutations. In contrast, for missense mutations, we found low d*N*/d*S* ratios in cord blood (d*N*/d*S*=0.5, $CI_{95\%}$=0.4-0.6), but high ratios in adult blood (d*N*/d*S*=1.2, $CI_{95\%}$=1.1-1.3; Fig. 2a). Notably, when looking at gene-level d*N*/d*S* estimates for adult and cord blood respectively, *MT-CO1, MT-ND1* and *MT-ND4* demonstrated significant positive selection across missense mutations, while *MT-CYB, MT-ND3* and *MT-ND4* showed significant negative selection (q≤0.05, Extended Data Fig. 3c).

To evaluate these findings further, we repeated the d*N*/d*S* analysis splitting SNVs into distinct heteroplasmy groups in the adult blood samples and excluding very low VAF (<1%) variants (Fig. 2b-c; Extended Data Fig. 3b). While we found consistent negative selection on nonsense mutations at all VAF levels (d*N*/d*S*<1), our results for missense mutations showed positive selection on low VAF mutations (1-5%: d*N*/d*S*=1.4, $CI_{95\%}$=1.2-1.6), which progressively declined with increasing VAF, only to show negative selection at high heteroplasmy levels (≥50%: d*N*/d*S*=0.8, $CI_{95\%}$=0.7-1.0). This signal could not be attributed to any particular gene in the mitochondrial genome and was spread across many genes (Fig. 2d).

Taken together, these data suggest that some missense mutations improve the fitness of individual mitochondria but become deleterious to the fitness of the whole cell as they approach homoplasmy – evidence of antagonistic selection landscapes at the scale of organelles versus the scale of cells.

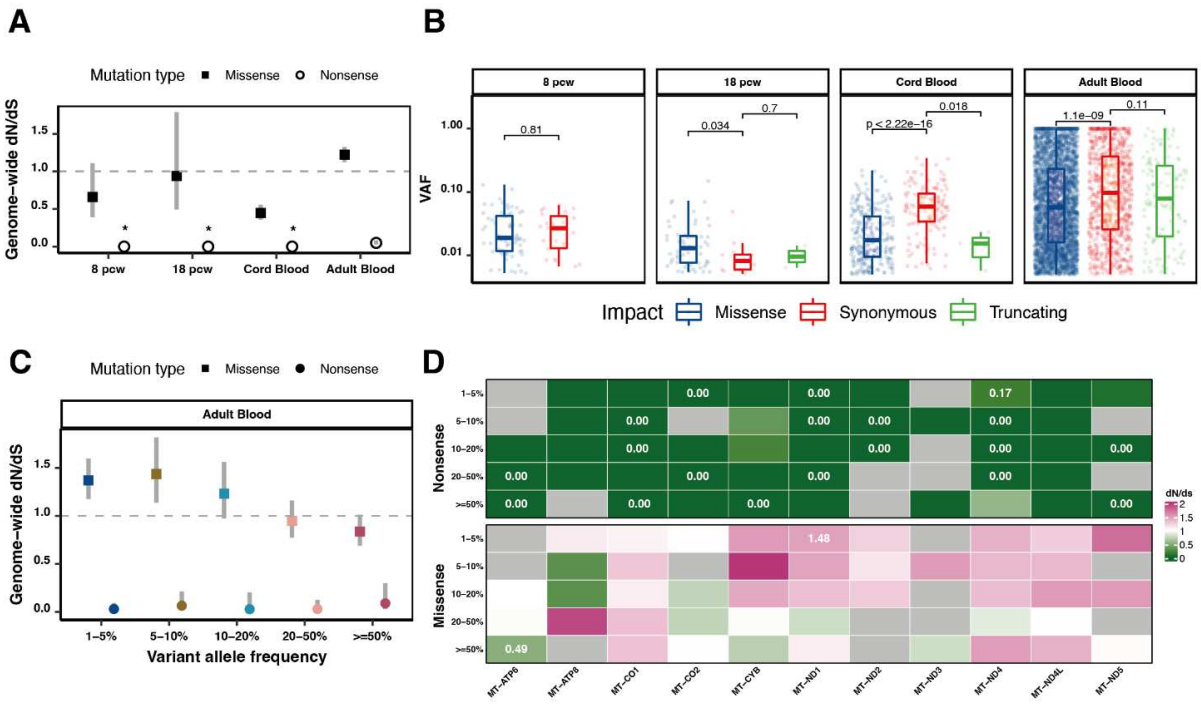


Fig. 2| Functional somatic mutations in the mitochondrial genome. a, Mitochondrial genome-wide d*N*/d*S* ratio for missense and nonsense mutations in the overarching age groups. Asterisks highlight d*N*/d*S* ratios where the confidence intervals from dndscv were infinitive. **b,** Boxplots highlighting the variant allele fraction (y-axis) of all mutations per mutation type (x-axis) across age groups. Statistical annotations denote the results of a Wilcoxon rank sum test using synonymous mutations as a reference. Of note, the 8pcw fetus did not have truncating mutations therefore not showing any data for this mutation type. **c,** Mitochondrial genome-wide d*N*/d*S* ratio for missense and nonsense mutations of specific VAF levels in the adult blood. **d**, Gene-level d*N*/d*S* ratios for missense and nonsense mutations for all coding mitochondrial genes on the heavy strand across VAF levels. Note that *MT-ND6* is on the light strand and was not considered for this analysis. Only point estimates which have a significant q-value as provided by dndscv (q-value ≤ 0.05) are highlighted in cells.

Heteroplasmic mtDNA mutations in the oocyte

Heteroplasmic mtDNA mutations may be inherited from the mother if present in the oocyte. If deleterious, such mutations have the potential to cause disease in offspring, even if present at benign levels in the mother[35], and dramatic shifts in mutation heteroplasmy from mother to child have been observed[15,36,37]. Previous studies have found no more than 2 variants in any one oocyte, although their sensitivity to detect mutations was restricted to VAFs >1-2% [38,39]. Given that oocytes have several hundred thousand copies of the mitochondrial genome[3], this means mutations must be present in >1,000 mtDNA molecules to have a chance of detection. The frequency of variants below this threshold is unknown. We reasoned that the foetal and cord blood somatic phylogenies provide a sensitive way to detect such mutations, even if present at low VAFs in the oocyte, since low VAF variants may drift up to detectable levels in some lineages, and the phylogenetic relationships of mutation-positive samples would imply their origin in the oocyte.

Across the four foetal/neonatal phylogenies, we found strong evidence for 21 heteroplasmic mutations that were likely present in the oocyte (range: 1-8 per individual, mean=5.25; Fig. 3). For each mutation, the most recent common ancestor (MRCA) of colonies with that mutation was the zygote, with most demonstrating significant phylogenetic signal. Independent

acquisition of the same mutation in different somatic lineages is unlikely at this early stage of life given the low mutation burdens overall. In all cases, the mutation was present in ≥2 samples and had drifted up to a VAF >1% in at least one of these, although the mean VAF across all samples was generally low (range: 0.01%-3.9%, median: 0.25%). The mean VAF represents the maximum likelihood value for the heteroplasmy level in the oocyte, and would therefore have been below the 1% detection threshold of previous studies for 16/21 mutations. These results suggest that low-level heteroplasmic mutations in the oocyte may be more common than previously appreciated.

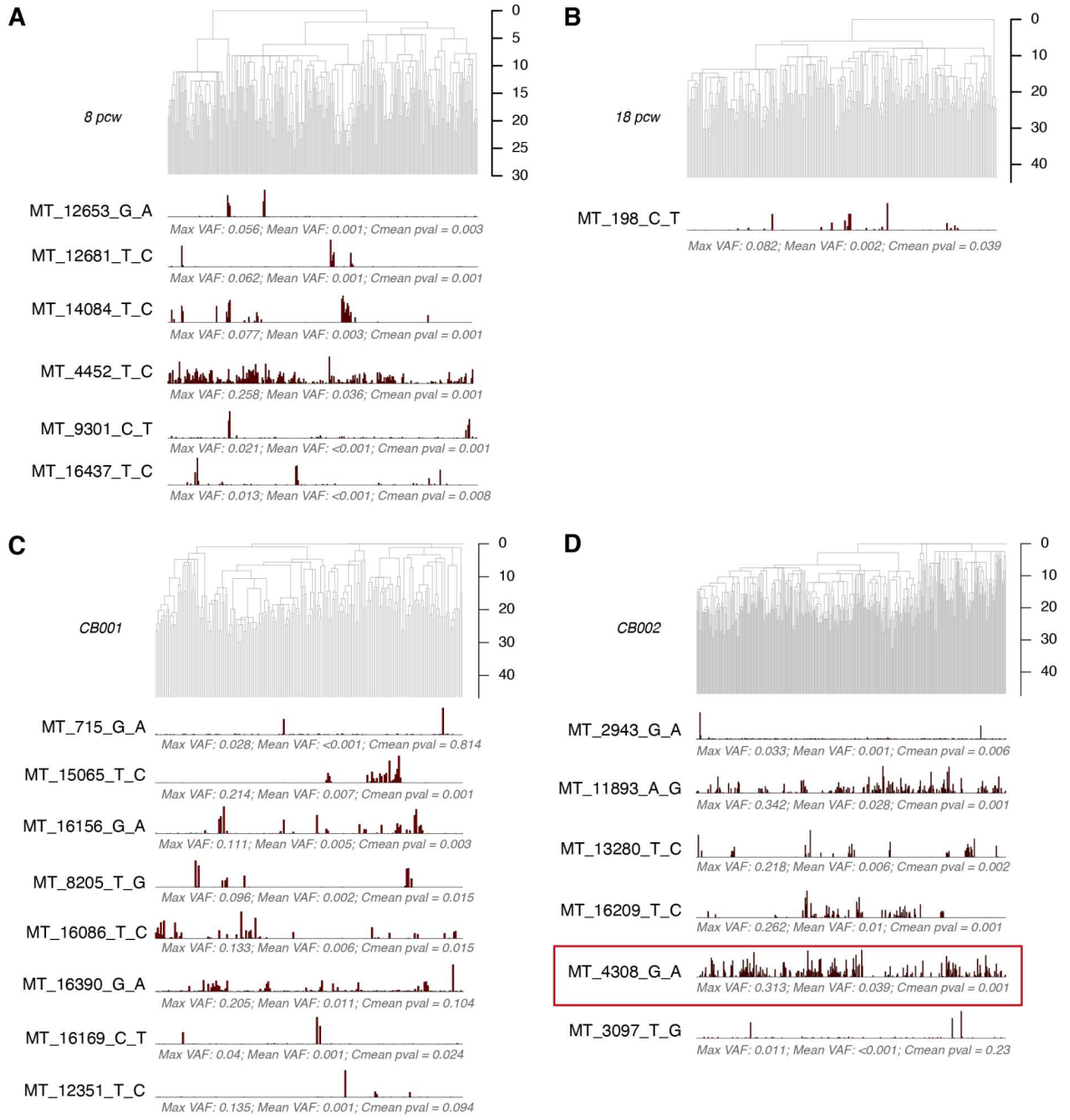


Fig. 3| Heteroplasmic mutations during early development. a-d, Bar plots below the phylogenetic trees for the 8pcw foetal (a), 18pcw foetal (b), CB001 cord blood (c) and CB002 cord blood (d) samples, showing the VAF for shared mutations (present at >1% VAF in >1 sample) across samples. Printed below each bar plot are the maximum mutation VAF, the mean VAF across all samples and the p-value for the Cmean measure of phylogenetic signal. The phylogeny illustrates the clonal relationships between samples. Only shared mutations for which the inferred MRCA was the zygote are depicted, with the exception of the highly asymmetric 18pcw phylogeny, for which the depicted mutation has an inferred MRCA of the 2-cell stage embryo. The red box highlights a known pathogenic mutation

associated with chronic progressive external ophthalmoplegia. pcw, post-conception weeks; MRCA, most recent common ancestor; VAF, variant allele fraction.

Inferring heteroplasmic drift from VAF distributions

The rate at which the heteroplasmy of mtDNA mutations drifts up or down is currently unknown *in vivo* in human somatic lineages. We found that the VAF distribution of mtDNA somatic mutations changes with increasing age (Fig. 4a), and reasoned that the rate of heteroplasmic drift within HSCs could therefore be estimated. We modelled drift using a Wright-Fisher (WF) population model[4,40] in which mutations were introduced into a population at a fixed rate per mitochondrial genome per generation, and allowed to drift over subsequent generations. Drift rates are determined by the effective population size (number of mitochondrial copies) multiplied by the generation time (time to replicate the population), with lower values corresponding to more rapid drift. As predicted, the simulated VAF distributions mirrored the changing distributions observed with increasing age in the data (Fig. 4b). Importantly, this highlighted that after many generations, low VAF mutations were only recently acquired, while high VAF mutations had invariably been acquired during early generations, indicating that mutation VAF is predictive of time of acquisition (Extended Data Fig. 4), in line with previous modelling[41].

Based on this, we used Approximate Bayesian Computation (ABC) to quantify the number of generations and mtDNA mutation rate for each individual. Our results demonstrated a linear increase in the estimated total number of mitochondrial generations with age (Fig. 4c), suggesting a constant mitochondrial turnover rate across life. We estimated a drift parameter of 16,000 mitochondria days ($CI_{95\%}$ 13,000-19,000). This drift rate may represent various combinations of effective population size and generation time (Fig. 4d). In practice, the effective population size is likely to be between the mtDNA copy number (600, the median in adults), and the total number of mitochondria (~120, at ~5 mtDNA genomes/mitochondrion [32]). Therefore, this range would correspond to a generation time of between 27 and 133 days, suggesting that the set of mitochondrial genomes within a given HSC will be replicated approximately once every 4-19 weeks. These estimates are of the same order as directly measured mitochondrial turnover rates in rat tissues[42].

The inferred mtDNA mutation acquisition rates clustered around $4\text{-}6\times10^{-4}$ mutations/mtDNA genome/generation, corresponding to rates of 3×10^{-8}/base/generation (Fig. 4e), similar to previous modelling[43]. Given estimates of 0.7 mutations/division in the diploid nuclear genome[27,44], our mtDNA mutation rate estimate indicates that the mtDNA mutation acquisition rate is ~270x higher than the nuclear genome, although in any given lineage most mtDNA mutations will in fact be lost through drift[7,8,45].

Lastly, we aimed to assess the rate of heteroplasmic drift in early human development. We leveraged the six heteroplasmic oocyte mutations detected in the 8pcw foetus using a similar ABC approach (Fig. 3a, Methods). With this framework, we modelled drift of heteroplasmic oocyte mutations of the 8pcw foetus down independent lineages, comparing the simulated VAF distributions to the data. Our combined results across mutations estimated the average drift parameter during this 8-week period to be ~1,200 mitochondria days ($CI_{95\%}$ 1,100-1,300 mitochondria days) (Fig. 4a). This implies a 13-fold faster drift compared to adult HSCs, which likely relates to the rapid cell division at this point of development[46]. However, one might expect this to be offset by the extremely high mitochondrial copy number during blastulation[38].

Therefore, this finding supports the concept of a somatic mitochondrial bottleneck, whereby a transient drop in mitochondrial copy number accelerates neutral drift[32].

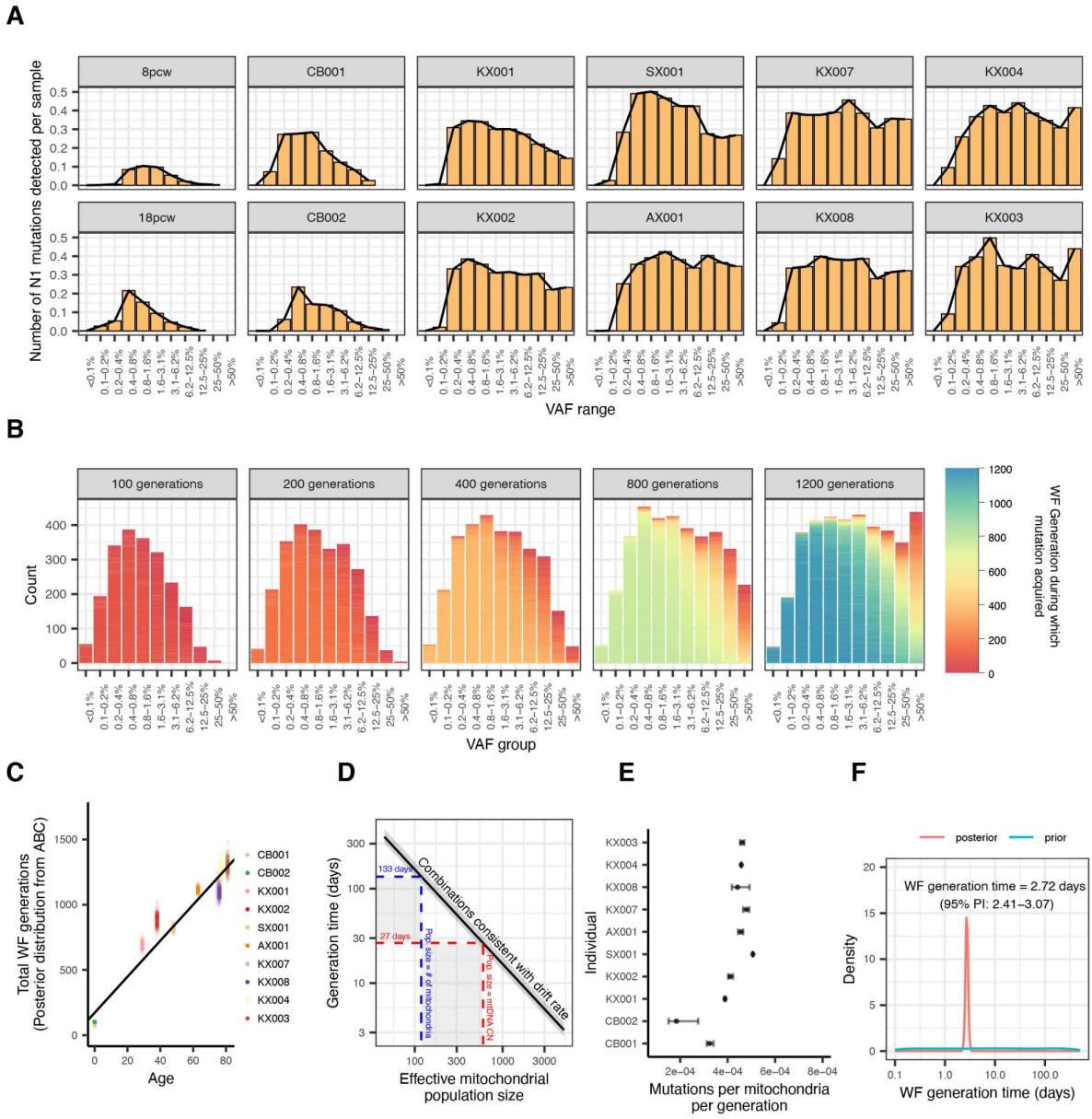


Fig. 4| Shifting variant allele fraction distributions of mtDNA mutations through life. a, The average number of mutations per sample at different VAF levels in different individuals. Includes only those mutations that most likely derive from the genuine 'N1' signature. **b**, Histograms of simulated VAF distributions after different numbers of WF generations, assuming constant rate of mutation acquisition in each generation and a population size of 600. Each mutation in the barplot is coloured according to the WF generation in which it was acquired. **c**, Dot plot showing the posterior distribution of the number of WF generations (y-axis) for each individual, against individual age (x-axis). The black line shows the inferred relationship between WF generations and age (linear mixed effects regression). **d,** Plot illustrating the combinations of population size and generation time consistent with the inferred drift rate in the WF model (black line). Also shown, are the corresponding generation times assuming the true effective population size to be the average mtDNA copy number (red dashed line) or the number of mitochondria (blue dashed lines) which may be considered the upper and lower bounds for this value. **e,** Dot plot showing the inferred mitochondrial mutation acquisition rate per mitochondria, per WF generation (x-axis) for each patient. Error bars represent the 95% posterior interval on these values. **f,** Posterior distribution for the inferred average WF generation time during the first 8 weeks of

development. ABC, Approximate Bayesian Computation; WF, Wright-Fisher; VAF, variant allele fraction; PI, Posterior Interval.

Mitochondrial mutations as lineage-tracing markers

The potential to utilise SNVs in the mitochondrial genome as lineage-tracing markers has garnered major interest in recent years[8,33,47]. We assessed the power of mtDNA mutations to report on known clonal relationships between cells. For this purpose, we selected the four individuals over 75 years of age, whose SNV-based phylogenies showed the richest clonal structure (Fig. 5, Extended Data Fig. 5). Across these four individuals, there were 58 expanded clones (Extended Data Fig. 6a), defined as an ancestral lineage contributing at least 1% of colonies.

Each expanded clone was tested for mtDNA mutations present at >1% VAF in any colony. In total, 24 of 58 clones had no mtDNA mutation shared by more than one colony. More importantly, in most cases, the best potential lineage marker was present in only a small proportion of colonies within the expansion (median=0.25; range=0.07-1). However, 7 of 58 expansions had a mtDNA marker mutation present in all colonies within the expansion (Fig. 5a). Five of these seven mutations were at homoplasmic levels (Fig. 5a, Extended Data Fig. 6b), suggesting that the mutations were already homoplasmic in the most recent common ancestor of the expansion. Given that mtDNA mutations take time to attain homoplasmy, we reasoned that clonal expansions with later ancestors may be more likely to have useful mtDNA marker mutations. Indeed, a significant correlation was observed (linear model, $p=0.04$, $R^2=0.056$) (Extended Data Fig. 6b).

Next, we assessed all 348 mtDNA SNVs shared by more than one colony for phylogenetic signal, which can be considered a measure of how accurately shared mtDNA mutations report true phylogenetic structure. Overall, 72% of shared mutations (249/348) showed no phylogenetic signal (Fig. 5a, Extended Data Fig. 5, Extended Data Fig. 6c). While most uninformative mutations were shared by just two samples (214/249), some were shared by up to 7 samples and sometimes at near-homoplasmic levels. Interestingly, the MRCA of colonies with uninformative mutations often occurred in embryogenesis, frequently being the zygote (Fig. 5a, Extended Data Fig. 5). This can be explained either by independent mutation acquisition in multiple lineages (violation of infinite sites assumption) or by low-level heteroplasmy of the mutation in the oocyte, stochastically drifting upwards in independent lineages. Using these mutations to infer clones could therefore lead to incorrect conclusions.

Mutations present in more colonies and at higher levels of heteroplasmy were more likely to show phylogenetic signal (Fig. 5a; Extended Data Fig. 5-6). Both of these measures are captured in the ‘global VAF’, which represents the aggregated VAF across all samples in an individual. This can be thought of as a surrogate for the mutation VAF in a bulk sample. All mutations with a global VAF >1% and 41/51 mutations with a global VAF >0.5% showed significant signal (Extended Data Fig. 6e). However, many mutations with a high phylogenetic signal still demonstrated imperfect fit with the true phylogeny (Fig. 5c), therefore being poor indicators of subclonal phylogenetic structure. This arises because heteroplasmic mutations stochastically increase or decrease within different subclones.

In line with this, we hypothesised that VAF distributions of specific heteroplasmic mutations across clonal expansions (Fig. 5c) would provide a complementary approach to estimate drift rates in HSCs. We designed a phylogeny-specific WF-based model, wherein drift was

explicitly modelled along different ancestral lineages in the phylogeny. An ABC approach based on this model estimated the drift parameter as 12,000 mitochondria days ($CI_{95\%}$ 9,000-27,000; Extended Data Fig. 7), highly concordant with our previous approach.

Lastly, we assessed the utility of mtDNA mutations for inferring clonal relationships between colonies using a previously described nearest neighbours approach[33]. We restricted mutations to those present (1) at >1% VAF in more than one sample, and (2) at a global VAF >0.5%. Utilising these variants resulted in correct recognition of clones with consistent mtDNA marker mutations (Extended Data Fig. 8a-b). However, as anticipated based on our previous results, the majority of mtDNA clones related to only a minor proportion of the true clones, with some mtDNA mutations identifying colonies that were in fact clonally unrelated (Extended Data Fig. 8). Hence, 4-12% of 'singleton' colonies, which have no clonally related colonies after the time of embryogenesis, were inappropriately assigned to the same clones (Extended Data Fig. 8d). Hierarchical clustering to reconstruct putative phylogenies of the inferred clones (Extended Data Fig. 8e) sometimes correctly grouped related clones (e.g. clones 2, 8 and 10 for KX003), but were generally uninformative.

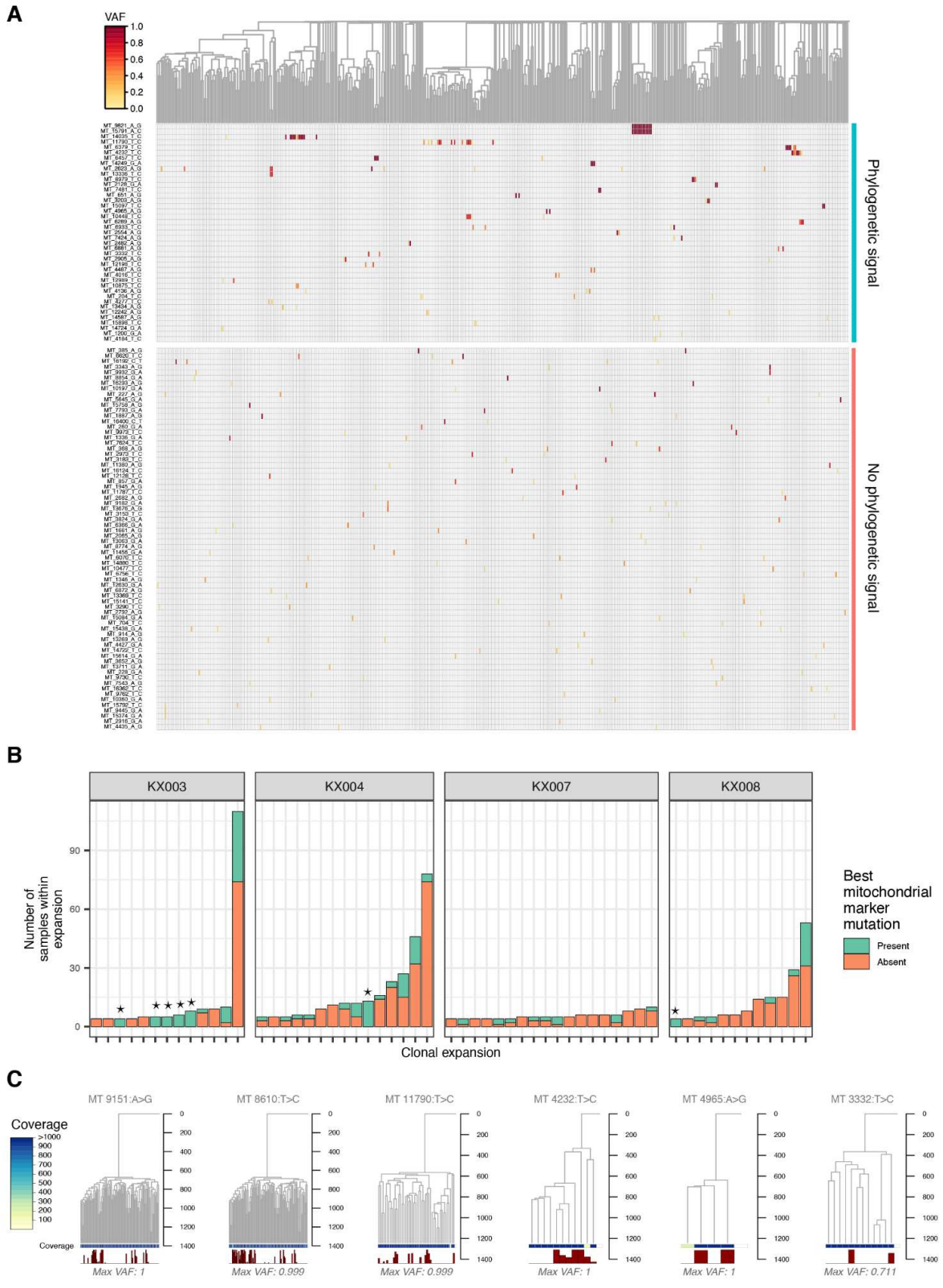

A
VAF
1.0
0.8
0.6
0.4
0.2
0.0
Phylogenetic signal
No phylogenetic signal
B
KX003
KX004
KX007
KX008
Number of samples within expansion
90
60
30
0
Clonal expansion
Best mitochondrial marker mutation
Present
Absent
C
Coverage
>1000
900
800
700
600
500
400
300
200
100
MT 9151:A>G
MT 8610:T>C
MT 11790:T>C
MT 4232:T>C
MT 4965:A>G
MT 3332:T>C
Max VAF: 1
Max VAF: 0.999
Max VAF: 0.999
Max VAF: 1
Max VAF: 1
Max VAF: 0.711

Fig. 5| Potential usage of mtDNA mutations as lineage markers. a, Heatmap of all shared mitochondrial mutations for KX004, with the SNV-based phylogeny shown above. Mutations that show phylogenetic signal - i.e. those that might be informative for inferring clones - are grouped separately from those that do not. **b,** Barplot indicating the number of samples in each expanded clone with the presence (green) or absence (orange) of the best mitochondrial marker mutation. Asterisks indicate an expansion with a consistent mitochondrial mutation in all samples. **c,** Selected mutations that show phylogenetic signal and are present exclusively within a clonal expansion but are not present within all colonies of that clone. Mutations demonstrate poor correlation with the sub-clonal phylogenetic structure. The height of the red bars underneath the clone phylogeny indicates the mutation VAF within each colony. The blue-green squares indicate the site-specific mitochondrial coverage in each colony, showing that the lack of mutation in some samples is unrelated to coverage.

Implications of drift for lineage tracing

To better understand the implications of our findings for mtDNA mutation-based lineage tracing, we used the inferred drift parameters (1,200 mitochondria days in foetal development and 15,500 mitochondria days thereafter) to inform simulations of heteroplasmic mutations (1) in the oocyte and (2) in the MRCA of clonal expansions. For the oocyte simulations, we initiated a heteroplasmic mutation with a VAF of 3.6%, mirroring the highest mean heteroplasmic mutation VAF observed in the 8 pcw foetus. We then modelled initial drift with the rate inferred in embryogenesis, subsequently followed by drift according to the rate inferred from adult HSCs. Consistent with well-established properties of the WF model[48], we observed initial diffusion, and subsequent polarisation of the VAF distribution, such that by 80 years most cells had either lost the initial mutation (VAF < 0.1%, 95.4% of cells) or fixed the mutation at homoplasmy (VAF > 99%, 2.1% of cells) (Fig. 6a). This results in patchy presence of the mutation in adult cell populations, with a pattern that is a poor indicator of the true phylogeny. To illustrate this further, we simulated drift of a theoretical heteroplasmic oocyte mutation at a starting VAF of 3.6% in the phylogeny of KX004. Indeed, we observed that the mutation became homoplasmic in several unrelated clones (Fig. 6b). Examples for these mutations have clearly been seen in other published somatic phylogenies[44,49]. Heteroplasmic germline mutations should therefore be taken into consideration when attempting lineage tracing across somatic phylogenies.

We also utilised the estimated HSC drift parameters to simulate heteroplasmic drift in clonal expansions with a range of dynamics. At one extreme, we simulated a gradual clonal expansion over 35 years, while at the other, we investigated a rapid expansion over 2 years (Fig. 6c,d). We modelled the drift with different starting levels of mutation heteroplasmy in the MRCA. Measuring the proportion of samples within expansions that had detectable levels of mutations (≥1% VAF) clearly highlighted that shorter periods of time, and higher starting VAFs led to more consistent marking of samples (Fig. 6d). Rapid expansions over 2 years led to relatively consistent marking even with a starting VAF as low as 5%. Conversely, gradual expansions over 35 years only had consistent marking when the starting VAF was > ~40%. Similar to the experimental data, presence or absence of the mutation within expansions correlated poorly with the sub-clonal phylogenetic structure (Fig. 6c).

In summary, these findings indicate that the effectiveness of mtDNA mutations as lineage markers will vary considerably as a consequence of heteroplasmic drift. In settings with recent and rapid growth dynamics, such as disease relapse following treatment, this may be appropriate and useful for distinguishing clonal populations.

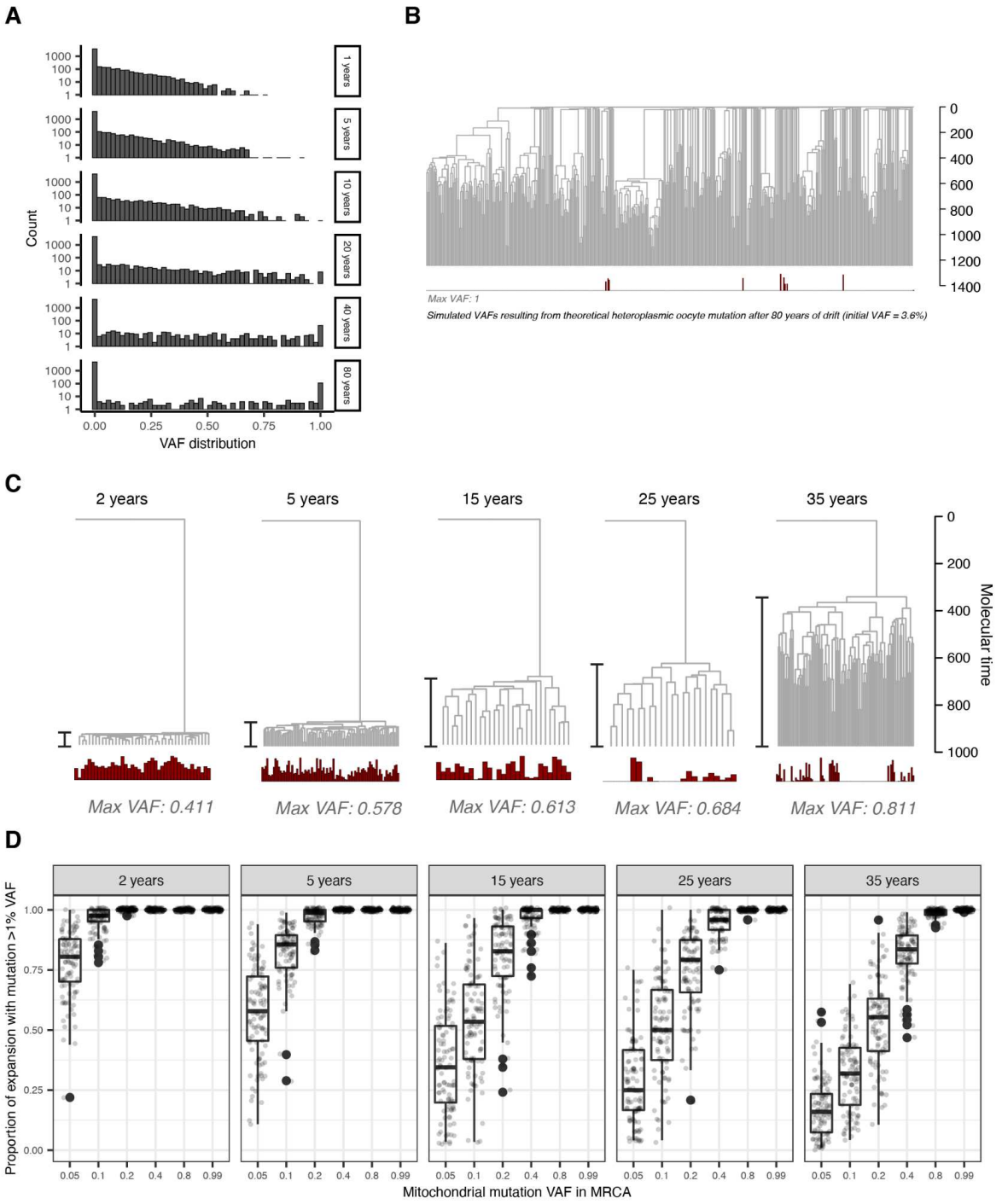


Fig. 6| Understanding the implications of drift through simulation. a, Simulation of the VAF distribution for a heteroplasmic mutation present in the oocyte at 3.6% VAF through lifetime. By old age, the mutation has been either lost or fixed in most lineages. **b**, Simulated VAF distribution of a heteroplasmic mutation present in the oocyte at 3.6% VAF that has undergone drift through the KX004 phylogeny (used for illustrative purposes only). The mutation has been fixed in multiple unrelated lineages. **c**, Illustrative heteroplasmic drift simulations for clonal expansions arising with distinct dynamics (2 yrs = rapid growth, 35 yrs = slow growth). The mtDNA copy number used is 1000, and the generation time 20 days, corresponding to the highest density posterior estimates from the ABC. In each simulation, VAF in the MRCA is 0.2. **d**, Boxplot derived from simulations (*N*=100 for each boxplot) indicating the proportion of clonal expansions anticipated to have detectable levels (>1% VAF) of a mitochondrial marker mutation depending on the specific dynamic of an expansion (indicated above each panel), and mutation VAF in the MRCA. Jittered raw data is plotted overlying. The boxes indicate the median and IQR and the whiskers extend to the largest and smallest values no more than 1.5× IQR

from the box. Outlying points are plotted individually. ABC, Approximate Bayesian Computation; VAF, Variant allele fraction; MRCA, Most recent common ancestor.

Discussion

Collectively, our data suggest that age-related decline in mitochondrial function is, at least in part, due to the accumulation and preferential expansion of selfish mitochondria with mutations slowing their respiratory function. The "survival of the slowest" hypothesis[12,50] posits that such mitochondria would accumulate free radical damage more slowly than their fully functional counterparts. Mutated mitochondria would survive longer by delaying turnover triggered by free radical damage, thus giving a selective advantage over their wild-type counterparts. Despite its articulation >25 years ago, this hypothesis has proved challenging to test due to the absence of methods to infer drift dynamics and selection on mtDNA mutations in specific somatic lineages over the lifespan. Here, the combination of single-cell colony sequencing and detailed lineage trees of stem cell clones enabled us to quantify the ingredients required for this Darwinian model of mitochondrial ageing.

Genetic variation in mitochondria accumulates at 270× the rate in the nuclear genome – thus, about half of all mitochondrial genomes in haematopoietic stem cells would have at least one mutation by ~50 years, even in the absence of selection. We estimate that an HSC replaces its complement of mitochondria every 4-19 weeks, 2-10× faster than the rate of cell division[51]. Such high turnover rates mean that mitochondria with selective benefits can preferentially expand to meaningful proportions over decades. The dN/dS ratio for missense mutations at low heteroplasmy (1-10%) was estimated at ~1.4 – this implies that about a quarter of all missense mutations in mtDNA confer selective benefits on their mitochondria. In contrast, at high levels of heteroplasmy (>50%), the dN/dS was ~0.8 – this implies that a fifth of missense mutations affect energy provision so adversely as to disadvantage the whole cell when present in over half of mitochondria in that cell.

Recent studies have demonstrated that somatic tissues are characterised by the preferential expansion of clones of cells with driver mutations in the nuclear genome[26,52–55], although it remains unproven whether these clones contribute to organ-wide phenotypes of ageing. The data here have striking parallels shrunk to the scale of the internal microcosm of the cell, but with the additional morsel that expansions of selfish mitochondria can, when sufficiently large, impair the fitness of the whole cell. This phenomenon may play a particularly important role in post-mitotic cells such as neurons – whereas cells with impaired respiratory function will be out-competed by cells with more effective energy provision in replicating tissues such as bone marrow, this cannot apply in a tissue where cells never turn over. This, then, could be a general paradigm of ageing, explaining how stochastic damage in individual organelles or cells propagates over decades, through preferential replication of selfish functional units, to generate age-related dysfunction system-wide.

Acknowledgements

This work was supported by the WBH Foundation. Investigators at the Sanger Institute are supported by a core grant from the Wellcome Trust. Work in the A.R.G. laboratory is also supported by the Wellcome Trust, Cancer Research UK and the Alborada Trust. Work in E.L. laboratory is supported by a Wellcome Trust Sir Henry Dale Fellowship, BBSRC and a European Haematology Association Non- Clinical Advanced Research Fellowship. The E.L. and A.R.G. laboratories are supported by a core support grant from the Wellcome Trust and Medical Research Council to the Cambridge Stem Cell Institute. K.M. is supported by the Chan-Zuckerberg Initiative. We would like to thank the Cambridge Blood and Stem Cell Biobank; the Cambridge Biorepository for Translational Medicine for access to human bone marrow and matched peripheral blood; the Cambridge NIHR BRC Cell Phenotyping Hub for their flow cytometry services and advice. We acknowledge further assistance from the National Institute for Health Research Cambridge Biomedical Research Centre and the Cambridge Experimental Cancer Medicine Centre. We are grateful to the donors, donor families and the Cambridge Biorepository for Translational Medicine for the gift of their tissue.

Competing Interests

I.M and P.J.C. are co-founders, stockholders and consultants for FL86 Inc.

References and notes

1. Spinelli, J. B. & Haigis, M. C. The multifaceted contributions of mitochondria to cellular metabolism. *Nat. Cell Biol.* **20**, 745–754 (2018).
2. Westermann, B. Mitochondrial fusion and fission in cell life and death. *Nat. Rev. Mol. Cell Biol.* **11**, 872–884 (2010).
3. 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).
4. Ju, Y. S. *et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *Elife* **3**, 1–28 (2014).
5. Wallace, D. C. *et al.* Sequence analysis of cDNAs for the human and bovine ATP synthase beta subunit: mitochondrial DNA genes sustain seventeen times more mutations. *Curr. Genet.* **12**, 81–90 (1987).
6. Brown, W. M., Prager, E. M., Wang, A. & Wilson, A. C. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.* **18**, 225–239 (1982).
7. Kang, E. *et al.* Age-Related Accumulation of Somatic Mitochondrial DNA Mutations in Adult-Derived Human iPSCs. *Cell Stem Cell* **18**, 625–636 (2016).
8. Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325-1339.e22 (2019).
9. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
10. Hekimi, S., Lapointe, J. & Wen, Y. Taking a “good” look at free radicals in the aging process. *Trends Cell Biol.* **21**, 569–576 (2011).
11. Bárcena, C., Mayoral, P. & Quirós, P. M. Mitohormesis, an Antiaging Paradigm. *Int. Rev. Cell Mol. Biol.* **340**, 35–77 (2018).
12. de Grey, A. D. A proposed refinement of the mitochondrial free radical theory of aging. *Bioessays* **19**, 161–166 (1997).
13. Wallace, D. C. Mitochondrial genetic medicine. *Nat. Genet.* **50**, 1642–1649 (2018).
14. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).
15. Olivo, P. D., Van de Walle, M. J., Laipis, P. J. & Hauswirth, W. W. Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA D-loop. *Nature* **306**, 400–402 (1983).
16. Cree, L. M. *et al.* A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. *Nat. Genet.* **40**, 249–254 (2008).
17. Upholt, W. B. & Dawid, I. B. Mapping of mitochondrial DNA of individual sheep and goats: rapid evolution in the D loop region. *Cell* **11**, 571–583 (1977).
18. Liu, C. *et al.* Presence and transmission of mitochondrial heteroplasmic mutations in human populations of European and African ancestry. *Mitochondrion* **60**, 33–42 (2021).
19. Laricchia, K. M. *et al.* Mitochondrial DNA variation across 56,434 individuals in gnomAD. *Genome Res.* **32**, 569–582 (2022).
20. Taylor, R. W. *et al.* Mitochondrial DNA mutations in human colonic crypt stem cells. *J. Clin. Invest.* **112**, 1351–1360 (2003).
21. Hübner, A. *et al.* Sharing of heteroplasmies between human liver lobes varies across the mtDNA genome. *Sci. Rep.* **9**, 11219 (2019).
22. Li, M., Schröder, R., Ni, S., Madea, B. & Stoneking, M. Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 2491–2496 (2015).
23. Larman, T. C. *et al.* Spectrum of somatic mitochondrial mutations in five cancers. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 14087–14091 (2012).
24. Stewart, J. B. *et al.* Simultaneous DNA and RNA Mapping of Somatic Mitochondrial Mutations across Diverse Human Cancers. *PLoS Genet.* **11**, e1005333 (2015).
25. Yuan, Y. *et al.* Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet.* **52**, 342–352 (2020).
26. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).

27. Spencer Chapman, M. *et al.* Lineage tracing of human development through somatic mutations. *Nature* **595**, 85–90 (2021).
28. Jones, D. *et al.* cgpCaVEManWrapper: Simple execution of caveman in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinformatics* **2016**, 15.10.1-15.10.18 (2016).
29. Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1-15.7.12 (2015).
30. Longchamps, R. J. *et al.* Evaluation of mitochondrial DNA copy number estimation techniques. *PLoS One* **15**, e0228166 (2020).
31. Moore, A. Z. *et al.* Influence of cell distribution and diabetes status on the association between mitochondrial DNA copy number and aging phenotypes in the InCHIANTI study. *Aging Cell* **17**, (2018).
32. Floros, V. I. *et al.* Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nat. Cell Biol.* **20**, 144–151 (2018).
33. Lareau, C. A. *et al.* Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0645-6.
34. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e21 (2017).
35. Rebolledo-Jaramillo, B. *et al.* Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15474–15479 (2014).
36. Ashley, M. V., Laipis, P. J. & Hauswirth, W. W. Rapid segregation of heteroplasmic bovine mitodiondria. *Nucleic Acids Res.* **17**, 7325–7331 (1989).
37. Laipis, P. J., Van de Walle, M. J. & Hauswirth, W. W. Unequal partitioning of bovine mitochondrial genotypes among siblings. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 8107–8110 (1988).
38. Boucret, L. *et al.* Deep sequencing shows that oocytes are not prone to accumulate mtDNA heteroplasmic mutations during ovarian ageing. *Hum. Reprod.* **32**, 2101–2109 (2017).
39. Jacobs, L. *et al.* mtDNA point mutations are present at various levels of heteroplasmy in human oocytes. *Mol. Hum. Reprod.* **13**, 149-154* (2007).
40. Wilton, P. R., Zaidi, A., Makova, K. & Nielsen, R. A Population Phylogenetic View of Mitochondrial Heteroplasmy. *Genetics* **208**, 1261–1274 (2018).
41. Elson, J. L., Samuels, D. C., Turnbull, D. M. & Chinnery, P. F. Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. *Am. J. Hum. Genet.* **68**, 802–806 (2001).
42. Gross, N. J., Getz, G. S. & Rabinowitz, M. Apparent turnover of mitochondrial deoxyribonucleic acid and mitochondrial phospholipids in the tissues of the rat. *J. Biol. Chem.* **244**, 1552–1562 (1969).
43. Coller, H. A. *et al.* High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. *Nat. Genet.* **28**, 147–150 (2001).
44. Coorens, T. H. H. *et al.* Extensive phylogenies of human development inferred from somatic mutations. *Nature* 1–6 (2021) doi:10.1038/s41586-021-03790-y.
45. Tao, L. *et al.* Retrospective cell lineage reconstruction in humans by using short tandem repeats. *Cell Reports Methods* **1**, 100054 (2021).
46. Herbert, M., Wolstenholme, J., Murdoch, A. P. & Butler, T. J. Mitotic activity during preimplantation development of human embryos. *J. Reprod. Fertil.* **103**, 209–214 (1995).
47. Xu, J. *et al.* Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *Elife* **8**, e45105 (2019).
48. Kimura, M. SOLUTION OF A PROCESS OF RANDOM GENETIC DRIFT WITH A CONTINUOUS MODEL. *Proc. Natl. Acad. Sci. U. S. A.* **41**, 144–150 (1955).
49. Park, S. *et al.* Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature* **597**, 393–397 (2021).

50. Harman, D. Aging: a theory based on free radical and radiation chemistry. *J. Gerontol.* **11**, 298–300 (1956).
51. Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P. & Abkowitz, J. L. The replication rate of human hematopoietic stem cells in vivo. *Blood* **117**, 4460–4466 (2011).
52. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
53. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
54. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* 1 (2019) doi:10.1038/s41586-018-0811-x.
55. Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, 1DUMMY (2019).
56. Ellis, P. *et al.* Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021).
57. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
58. Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 13 (2014).
59. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
60. Gerstung, M., Papaemmanuil, E. & Campbell, P. J. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198 (2014).
61. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
62. Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
63. Keck, F., Rimet, F., Bouchez, A. & Franc, A. Phylosignal: An R package to measure, test, and explore the phylogenetic signal. *Ecol. Evol.* **6**, 2774–2780 (2016).
64. Williams, N. *et al.* Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162–168 (2022).

Methods

Sample cohort

Whole-genome sequencing data from two previous studies [26,27] were leveraged for the work presented. In both studies, a recently developed low input library preparation method was used to generate libraries from each colony [56]. Together, this included WGS data from 4217 single-HSC/HPC-I derived colonies of 12 haematopoietically normal individuals: two fetuses (8 and 18 post-conception weeks [pcw]) [27], two neonates, and eight adult individuals [26]. In brief, HSCs/MPPs derived from either foetal liver and bone marrow, cord blood, peripheral blood or post-mortem vertebral bone marrow samples were flow sorted and grown into single-cell derived colonies. Samples from the two studies were grown using different protocols. Colonies from Spencer Chapman et al. were grown by sorting single phenotypic HSCs/ HSPCs (for the 8pcw fetus, this included liver-derived Lin− CD34+ CD38+ cells [HPCs], and for the 18pcw fetus, this included bone marrow and liver-derived CD34+ CD38− CD90+ cells [HSCs], and CD34+ CD38+ CD90− CD45RA− CD49f− cells [CMPs and MEPs]) into 96-well plates containing an MS5 feeder layer and cultured for 6–8 days at 37 °C and 5% CO2. Colonies from Mitchell et al were grown by sorting single phenotypic HSC/MPP (Lin−CD34+ CD38− CD45RA−) or HPC (Lin− CD34+ CD38− CD45RA−) cells into 96-well plates containing 100 µl supplemented StemPro medium (Stem Cell Technologies) but no mouse feeder layer, and culturing for 21±2 days at 37 °C. Full details of sorting strategies and culture supplements are available in the original manuscripts. DNA was extracted and sequencing libraries prepared using an enzyme-based fragmentation protocol optimised for low input DNA quantities [56]. In both studies, libraries were sequenced to a mean nuclear genome coverage of approximately 15×, enabling somatic mutation calling from the nuclear genome with established variant-calling algorithms [28,29]. BWA mem was used to align 150bp paired end reads generated to the human reference genome (GRCh37d5).

Immunophenotyping of flow-sorted colonies

For samples from Mitchell et al. (all individuals except the 8 and 18 pcw fetuses), a subset of colonies were immunophenotyped to define the colony mature cellular composition. Following 21 +/- 2 days in culture, ½ of each colony selected by size criteria (> 3000 cells) was harvested into a U bottomed 96 well plate (ThermoFisher). Plates were then centrifuged (500g/5 minutes), media was discarded, and the cells were resuspended in an antibody mix as per table S5 (30 min/4°C). After washing cells were resuspended in 100 ml PBS/3% FBS containing an antibody panel consisting of: CD45/PECy5, CD41/FITC, CD11b/APCCy7, CD14/PECy7, CD15/BV421, GlyA/PE, CD56/APC. Immunophenotyping of the mature cells in each stained colony was performed using a BD Fortessa2 (BD Biosciences) as per the gating strategy in fig. S9A. In the analysis 50 cells were required in a gate to call that cell type present. The one exception was the NK cell gate for which a lower threshold of 30 cells were used. The relationships between the mature cell types produced in the assay are shown in fig. S9B.

Mitochondrial mutation calling

Alignment of reads to human genome assembly GRCh37 was carried out by *BWA-MEM* [57]. Duplicate reads were marked using *biobambam* [58]. Reads mapping to the mitochondrial genome were extracted using *samtools* [59]. *bam2R* (available as part of the *deepSNV* R package [60]) was used to count the number of forward and reverse reads supporting each base at each site in the mitochondrial genome for each sample. Unmapped reads, duplicate reads,

failed reads, secondary and supplementary alignments, reads with mapping quality scores <30 and bases with Phred quality scores <30 were excluded from these nucleotide counts.

Variant calling was performed using *ShearwaterML* [52]. A base-specific error model for the mitochondrial genome was generated from a reference panel of 102 polyclonal normal samples. In order to avoid having reduced calling sensitivity at sites with germline mutations, homoplasmic alternative sites in these polyclonal samples were treated as uninformative (i.e. the coverage at that site was set to 0). For 7 highly polymorphic sites (263, 750, 1438, 4769, 8860, 15326 and 16519) at which the reference base was observed less frequently than an alternative base, the reference base was instead set as uninformative.

In order to generate the background model, nucleotide counts were assumed to be distributed by a beta-binomial distribution, i.e. a binomial distribution in which the probability of success in n trials is not fixed but is instead drawn from a beta distribution defined by shape parameters α and β. The mean and variance of this distribution are defined as follows:

$$\mu = \frac{n\alpha}{\alpha+\beta} = n\pi, \text{ where } \pi = \frac{\alpha}{\alpha+\beta}$$

$$\sigma^2 = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} = n\pi(1-\pi)\frac{\alpha+\beta+n}{\alpha+\beta+1} = n\pi(1-\pi)[1+(n-1)\rho], \text{ where } \rho = \frac{1}{\alpha+\beta+1}$$

The over-dispersion value (ρ) defines the amount of extra variance, as compared with pure sampling errors; for ρ = 0, the model is a simple binomial distribution. As there is no closed-form solution to estimate ρ, we performed a grid-based search within the interval [10^{-6}, $10^{-0.5}$] (using increments of 0.05 for the index) to obtain a maximum likelihood estimate of ρ for each nucleotide at each site in the mitochondrial genome. A conservative lower bound of ρ (2×10^{-4}) was then applied to reduce the risk of artefactual calls being made at sites with insufficient depth in the reference panel for accurate ρ estimation.

In order to call variants, a given sample of interest is tested against this base-specific error model. A true variant will be present on both strands, with an equal rate that must be higher than the background error rates. *P*-values were obtained from each strand using a likelihood-ratio test with one degree of freedom for the extra parameter μ and *P*-values from both strands are combined using Fisher's method, as described previously [52]. Homoplasmic germline variants in a sample were excluded and *P*-values were subject to multiple testing correction using Benjamini & Hochberg's False Discovery Rate [61] and a q-value cut-off of 0.01 was used to call somatic mutations.

Investigating the trinucleotide context of mitochondrial mutations

We leveraged a custom R script relying on the R version of samtools to examine the overall mtDNA substitution signatures in the 96 possible mutation classes. The mutational spectrum remained consistent with previous reports above this threshold (Ju et al., 2014, eLife; Yuan et al., 2020, Nature Genetics), with an absolute dominance of C>T and T>C substitutions. By splitting the mutations between the heavy and light strand, we could also observe extreme replication strand bias. Lastly, we calculated the number of observed compared to expected mutations by considering how often each trinucleotide context is observed in the mtDNA sequence.

Filtering of mutations resulting from contamination or NUMTs

We noted that there were multiple heteroplasmic mutations that were widespread at a low VAF level across samples from a given individual, or across several individuals, but without any phylogenetic signal. We found that many of these mutations had an inverse correlation between the VAF and the mitochondrial copy number, consistent with what might be expected from a constant level of contaminating DNA from other sources mapping to the mitochondrial genome. If the contaminating DNA has a sequence mismatch, this will be recurrently called as a mutation across samples, at a VAF that depends on the total quantity of true mitochondrial DNA in the sample. Such DNA may result from low-level contamination from another individual, or from nuclear DNA of mitochondrial origin (so-call "NUMTs"). We therefore tested all mutations for this inverse correlation between VAF and mtDNA copy number by performing a linear regression, including all samples with VAF greater than 0. This revealed many mutations with a strong correlation (fig. S1e). We performed Benjamini-Hochberg multiple testing correction on all mutations with a positive correlation coefficient and considered mutations with a q-value < 0.01 (n = 316) to commonly result from sample contamination or NUMTs. However, given that such mutations may in some instances be genuine, we retained mutations in specific samples if the implied mutation copy number was above a threshold of 25 (where the mutation copy number is calculated as the mutation VAF multiplied by the total mtDNA copy number). To increase the confidence level in our calls and to exclude samples which were believed to have inter-individual contamination, we utilised VerifyBamID on the somatic mutations called in the nuclear genome of the respective sample [62]. Furthermore, when multiple germline mtDNA sites had multiple reference reads rather than being completely altered we suspected the sample to be contaminated as well and this sample was therefore excluded from the analysis.

Filtering of mutations resulting from other artifacts

Below a VAF of 1% there was contamination by mutations with distinct mutational signatures, suggestive of artifact or *in vitro*-acquired mutations. To better define these processes, and to filter such mutations from the call set, we ran a *de novo* 192-profile mutational signature extraction (incorporating heavy/ light strand) using a hierarchical Dirichlet process, treating mutations present at different VAF levels in different individuals as 'samples'. This extracted the genuine mutational signature 'N1' - accounting for essentially all mutations >1% VAF and a subset of lower frequency mutations - and 8 artefactual signatures that were variably present between individuals (fig. S2). We were then able to assign individual mutations to their most likely causative signature using the mutation VAF, trinucleotide context, and heavy/ light chain information. Mutations assigned to the real 'N1' signature were retained in the final mutation set.

Assessing forces of selection on the mitochondrial genome

To assess acting forces of selection on the mitochondrial genome, we utilised dndscv in R [34]. Initially, we assessed global dN/dS ratios of all mitochondrial genes on the heavy strand in our dataset (excluding MT-ND6, since it is on the light strand). Mutations were split according to age groups and separately utilised as input to dndscv. The output of dndscv was visualised using a custom R script. For gene-level analysis, dN/dS estimates with q-value < 0.05 were highlighted as a heatmap. For the heteroplasmy-level specific analysis, SNVs were split according to VAF groups and leveraged as input for dndscv. In addition, signs of selection were visualised by assessing the distribution of variant allele fraction across all mutations per age group. In principle, the VAF of synonymous mutations thereby serves as a reference point

to compare missense and truncating mutations to. Hence, if the VAF is sufficiently higher across mutations compared to the synonymous mutation VAF, this implies positive selection and vice versa. Significant differences were assessed utilising a Wilcoxon rank sum test across mutation types.

Inferring mitochondrial copy number

To estimate the number of mitochondrial copies within our samples, we applied a formula previously used in the pan-cancer analysis of whole genomes (PCAWG; [25]).

$$CN = \frac{coverage(mtDNA)}{coverage(nDNA)} \; x \; ploidy(sample)$$

where CN is the mtDNA copy number, $coverage_{mtDNA}$ and $coverage_{nDNA}$ are the mean coverage depths for mtDNA and the nuclear genome in individual WGS bam files, respectively, and $ploidy_{sample}$ is 2. We assessed the mean depth for both the nDNA and the mtDNA by applying bedtools genomecov.

Measuring the phylogenetic signal

The phylogenetic signal is a way to assess the heritability of a trait within a known phylogenetic structure. For individual mutations, the phylogenetic signal was calculated using the phyloSignal function from the R package 'phylosignal' (https://cran.r-project.org/package=phylosignal) [63]. This function outputs several measures of the phylogenetic signal and their associated p-values for a set of quantitative traits associated with the tips of a phylogeny - in this case the VAFs of each mutation. We focussed on the 'Cmean', which appeared most informative for our data. We used the same approach to assess the phylogenetic signal of mitochondrial copy number.

Inference of mtDNA clones

We aimed to follow the approach outlined as previously by others [33]. However, given our results showing the lack of phylogenetic signal for many mutations present at low global VAF, we restricted the incorporated mutations to somatic mutations that (1) were present at >1% VAF in more than one sample, (2) were present at a global VAF > 0.5%. We then used the VAF matrix for these mutations as inputs to the FindNeighbours/ FindClusters function from the R package 'Seurat', with slight modifications (cosine distance metric and k.param = 5 for the FindNeighbours function, resolution = 10 for the FindClusters functions). Note that we made some additional modifications to the approach used by Lareau et al. Firstly, the resolution value of 10 was much higher than their value of 1. We found that using resolution = 1 led to inappropriate merging of clusters. Secondly, we also used the raw VAF matrix, rather than the square root as we found this gave cleaner results. We also performed clustering using all shared mutations, including those present at global VAF <0.5%, but found that this led to extremely noisy clustering including many clusters that had no relation to the true clonal relationships (results not shown). Therefore, our approach represents an attempt to optimise that used in Lareau et al, using our prior knowledge of the true clonal relationships.

Approximate Bayesian Computation for embryonic development

We developed a simple Wright-Fisher model in which the total mitochondria in a single cell is the population size, and the wild-type/ mutant alleles represent the two alleles within the population. For each generation, the new number of mutant alleles was determined by a

random binomial draw of size *n* (the mitochondrial copy number) and probability *p* (the proportion of mutant alleles in the previous generation). The total number of generations for each cell was decided by rounding the total simulation time (56 days for the 8 post-conception week fetus) divided by the generation time for that simulation e.g. for a generation time of 2 days, there are 28 generations.

We performed 10,000 simulations for each of the 6 heteroplasmic oocyte mutations observed in the 8pcw fetus. For each set of simulations, the starting variant allele fraction was fixed as the mean variant allele fraction across all samples for that mutation. The population size was fixed at 450, the mean mitochondrial copy number for the 8pcw fetus. However, the precise population size used does not affect the final inference of the drift parameter. The $\log_{10}$ of the generation time was drawn from a uniform distribution (minimum = -1, maximum = 2.7) corresponding to a minimum generation time of 0.1 days and a maximum of 500 days. For each simulation, seven summary statistics were collected: mean VAF, minimum VAF, maximum VAF, standard deviation of the VAF and the 5%, 50% and 90% quantiles. The same set of generation time parameters were used for all 6 sets of simulations, and the summary statistics combined across mutations, creating a final set of 42 summary statistics. Generation time parameter inference was performed using the R package 'abc' (https://cran.r-project.org/package=abc) using a neural network regression.

Approximate Bayesian Computation for adult HSCs

Approach 1: overall VAF distributions with age

We first developed a simulation framework of mitochondrial mutation VAF distribution as follows. One thousand theoretical cells were included per simulation, each cell having a fixed mtDNA copy number of 600, approximately the mean mtDNA copy number found for adults in the data. For each simulation run, the total number of Wright-Fisher (WF) generations for that run was picked from between 1 and 2000. For each cell and generation, mutations were introduced into each mitochondria at a constant rate per WF generation. This rate was picked from a prior distribution Uniform(min=0.0001, max = 0.001), these bounds being chosen from some initial test runs. Once introduced into a single mtDNA molecule, each mutation was independently allowed to drift during subsequent generations by sampling with replacement. This process continued until the total number of WF generations was reached. At this point, the total number of mutations in each VAF category (<0.1%, 0.1-0.2%, 0.2-0.4%, 0.4-0.8%, 0.8-1.6%, 1.6-3.1%, 3.1-6.2%, 6.2-12.5%, 12.5-25%, 25-50%, >50%) was recorded and divided by 1000, to get values for the average number of mutations in each category per cell. The number of mutations per VAF category per cell were then used as summary statistics in an approximate Bayesian computation (ABC) to infer the most likely number of WF generations for each individual, using the same summary statistics for each individual as the "target" using the abc function from the R package 'abc' (https://cran.r-project.org/package=abc). For the ABC, the lowest two VAF categories (up to 0.2%) were excluded as potentially unreliable in the data due to coverage and copy number limitations.

To infer the drift rate, we first performed a linear mixed effects regression using the lmer function from the R package 'lme4', with posterior distributions of 'total WF generations' as the dependent variable, 'Age' as a fixed effect and 'Individual' as a random effect. This gives a value for the number of WF generations per year. This was converted to a drift rate by:

$$Drift\ rate\ (mitochondria\ days) = generation\ time\ (days) \times WF\ population\ size$$

$$= \frac{365\ days}{WF\ generations\ per\ year} \times\ WF\ population\ size$$

$$= \frac{365\ days}{13.72} \times\ 600\ =\ 15{,}960$$

This inferred drift rate is thus independent of the precise population size used for the simulations.

Approach 2: distribution of specific mutations across clonal expansions

For this ABC, we devised a phylogeny-specific simulation approach. We selected 10 mutations that were heteroplasmic across a given clonal expansion (fig. S7A, some illustrated in Fig. 3E). For each mutation we simulated independent drift along different branches from a common starting VAF in the MRCA, with the number of Wright-Fisher generations proportional to the branch length. For each simulation the starting VAF in the MRCA was drawn from a uniform distribution (minimum = 0, maximum = 1), the population size was fixed at 675 (the mean mtDNA copy number observed in the adult HSC/MPP colony data), and the generation time drawn from a uniform distribution (minimum = -1, maximum = 2.7) corresponding to a minimum generation time of 0.1 days and a maximum of 500 days. For each branch, the number of generations was then calculated as 17.5 x 365 days x the branch length (in molecular time), divided by the generation time, where 17.5 is the average number of SNVs acquired per year in HSCs [26]. At each node in the clonal expansion, drift continued independently in each daughter cell. We performed 20,000 simulations for each clonal expansion, collecting 6 summary statistics: (1) the median VAF across samples, (2) the number of samples without the mutations (VAF < 0.02), (3) the number of samples with the mutation at homoplasmy (VAF>0.98), (4) the number of samples with the mutation at heteroplasmic levels (0.02<VAF<0.98), (5) the Cmean measure of phylogenetic signal as calculated by the phyloSignal function from the R package 'phylosignal' (https://cran.r-project.org/package=phylosignal) and (6) the sigma parameter calculated by the function phyloCorrelogram from the same package, which is a measure relating to the phylogenetic distance at which the phylogenetic signal is strongest. These summary statistics were then compared with those for the data using the 'abc' R package, using the 'rejection' method, giving a posterior distribution for the generation time.

The posterior distributions for each mutation were combined by multiplying the posterior distribution densities across all mutations, to infer the posterior distribution of parameters that would most likely explain the observed VAF distributions across all 10 clonal expansions.

Mitochondrial drift simulations

Once mitochondrial heteroplasmic drift parameters were inferred using the above ABC approach, we did further simulations - using the same phylogeny-specific approach - to assess the resulting VAF distributions across clonal expansions with different growth dynamics.

The clonal expansions were simulated using the R package 'rsimpop' (https://github.com/NickWilliamsSanger/rsimpop) which has previously been used to model HSC dynamics while recording the phylogenetic structure of the population [26,64]. It also has the ability to introduce driver mutations with specific selection coefficients. Five trees containing clonal expansions were modeled such that the final age of the individual was 50

years, with the origin of the clonal expansion occurring 35, 25,15, 5 and 2 years prior to this. The selection coefficients were altered such that the driver mutation had time to expand to an appreciable size by age 50.

Heteroplasmic drift of a theoretical mitochondrial mutation was then superimposed onto these phylogenetic structures as described above, using various levels of starting VAF in the MRCA of the clonal expansion (5%, 10%, 20%, 40%, 80%, 99%), and the inferred drift parameter in adult HSCs. Drift was simulated 100 times for each starting VAF and tree, with the final proportion of samples with detectable mutation levels (>1% VAF) recorded in each case. An analogous approach was used to assess the consequences of heteroplasmic oocyte mutations. However, in this case the KX004 phylogeny structure was used to illustrate these consequences.

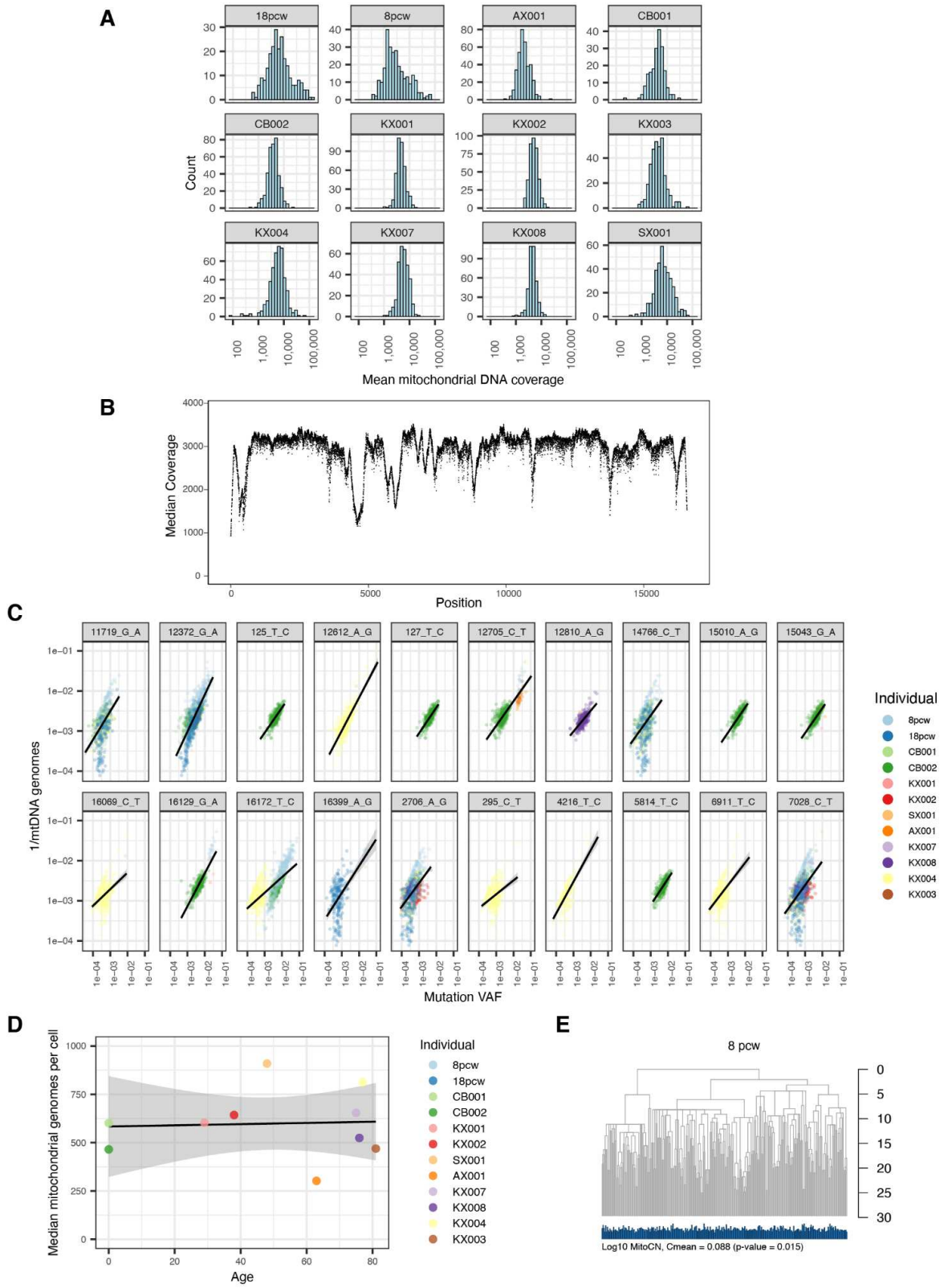
Code availability

The aforementioned computational methods provide a summary of the procedures implemented in various custom-made R and bash scripts. These scripts contain the commands run for the analyses highlighted in this work. To sustain reproducibility, they are publicly available on Github (https://github.com/mspencerchapman/mito_mutations_blood).

Data availability

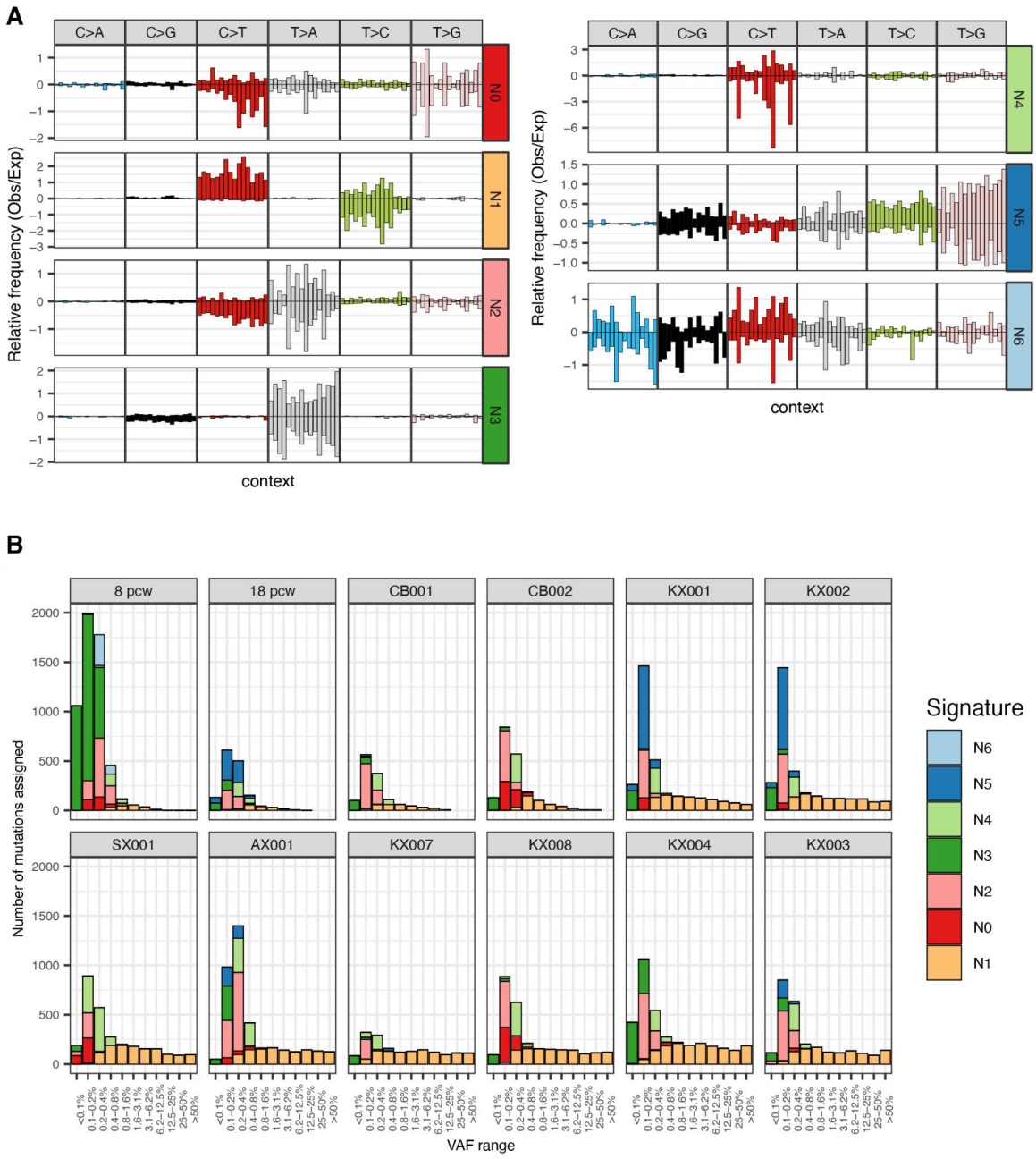
Information on the availability for the sequencing data from Spencer Chapman et al. and Mitchell et al. can be found in frame of the respective publications (https://www.nature.com/articles/s41586-021-03548-6#data-availability and https://www.nature.com/articles/s41586-022-04786-y#data-availability) .

Supplementary Figures

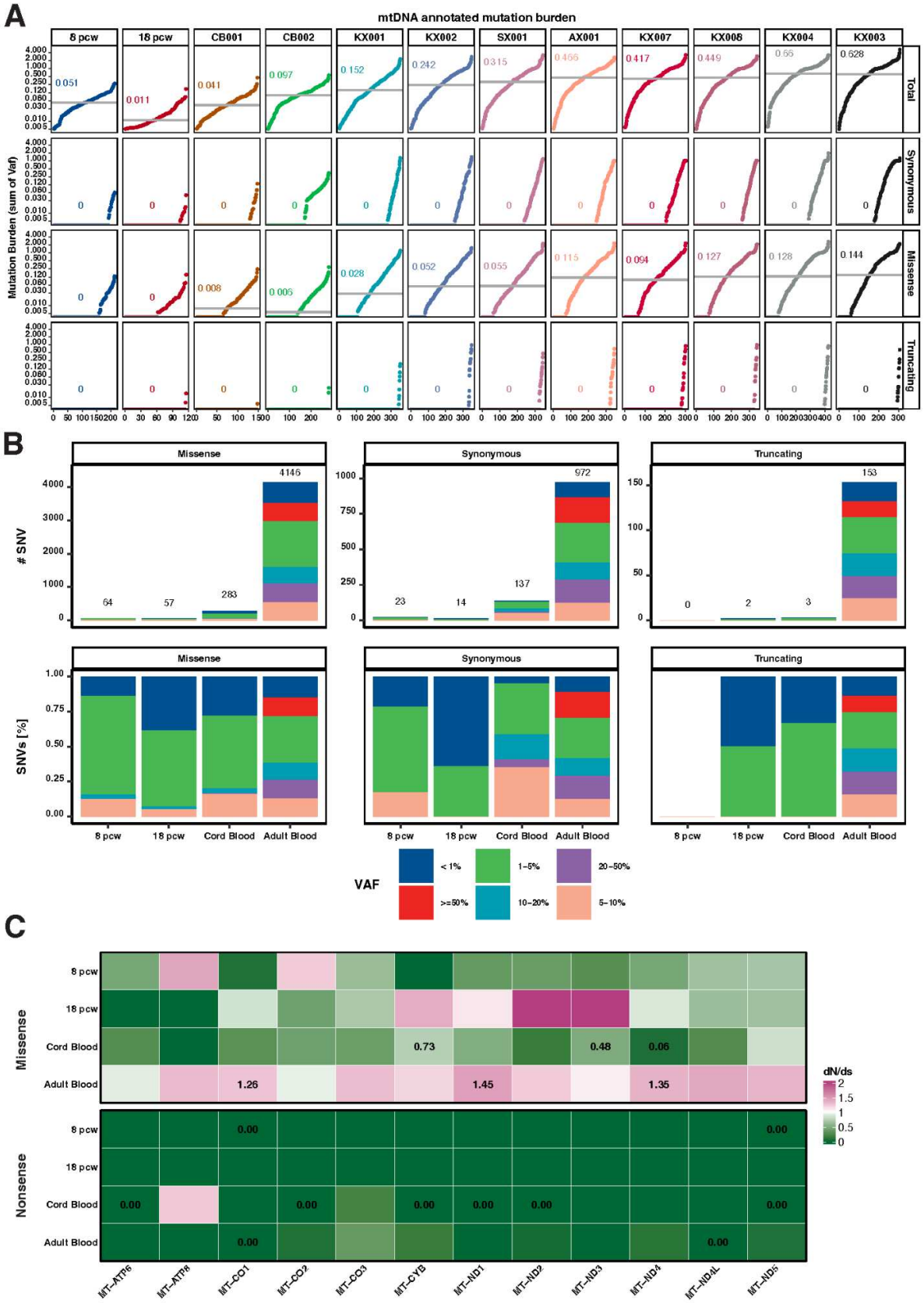


Extended Data Fig. 1| Mitochondrial coverage and copy number throughout life. a, Histograms of the mean mitochondrial coverage per colony separated by individual. **b,** Median coverage by mitochondrial genome position across all the adult samples. **c,** Selected mutations showing significant correlation between mutation VAF and 1/ mtDNA copy number. This pattern is expected for mutations caused by contamination across a sample set, or by mis-mapping of nuclear genome reads. Such

mutations were excluded from further analysis. **d,** Median mitochondrial copy number per colony (including only HSC-derived colonies) by individual (y-axis) plotted against individual age (x-axis), demonstrating no association. The black line highlights a linear regression with gray boundaries indicating the standard error. Each dot is colored according to the patient identity. **e,** Phylogenetic tree for the 8 pcw fetus. Below, the mitochondrial copy number (log_{10}) is shown as a bar plot in blue for each colony. (VAF = variant allele fraction, MitoCN = Mitochondrial copy number, pcw = post-conception weeks)

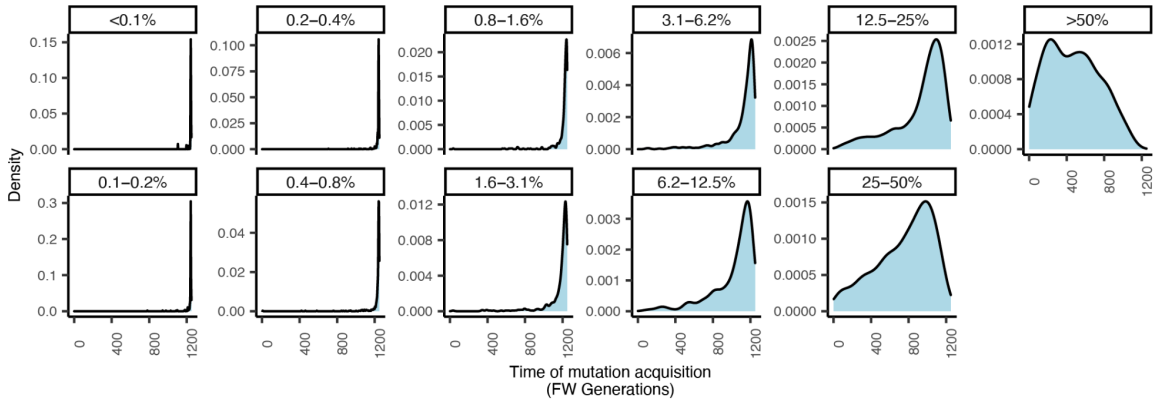


Extended Data Fig. 2| Mitochondrial mutation signature decomposition by variant allele fraction. **a,** 192-profile mutational signatures extracted by a hierarchical Dirichlet process (Methods). Mutations were aggregated according to individual and VAF. Positive bars represent mutations on the heavy strand, and negative bars those on the light strand. Bars are coloured by the substitution type. Signature 'N1' is the genuine signature. **b**, Absolute signature contributions to mutation sets at different VAF levels in different individuals, demonstrating large contributions of artefactual signatures at low VAFs <0.5%, and almost exclusive contributions of the genuine 'N1' signature at higher VAFs above ~1%. VAF, variant allele fraction.

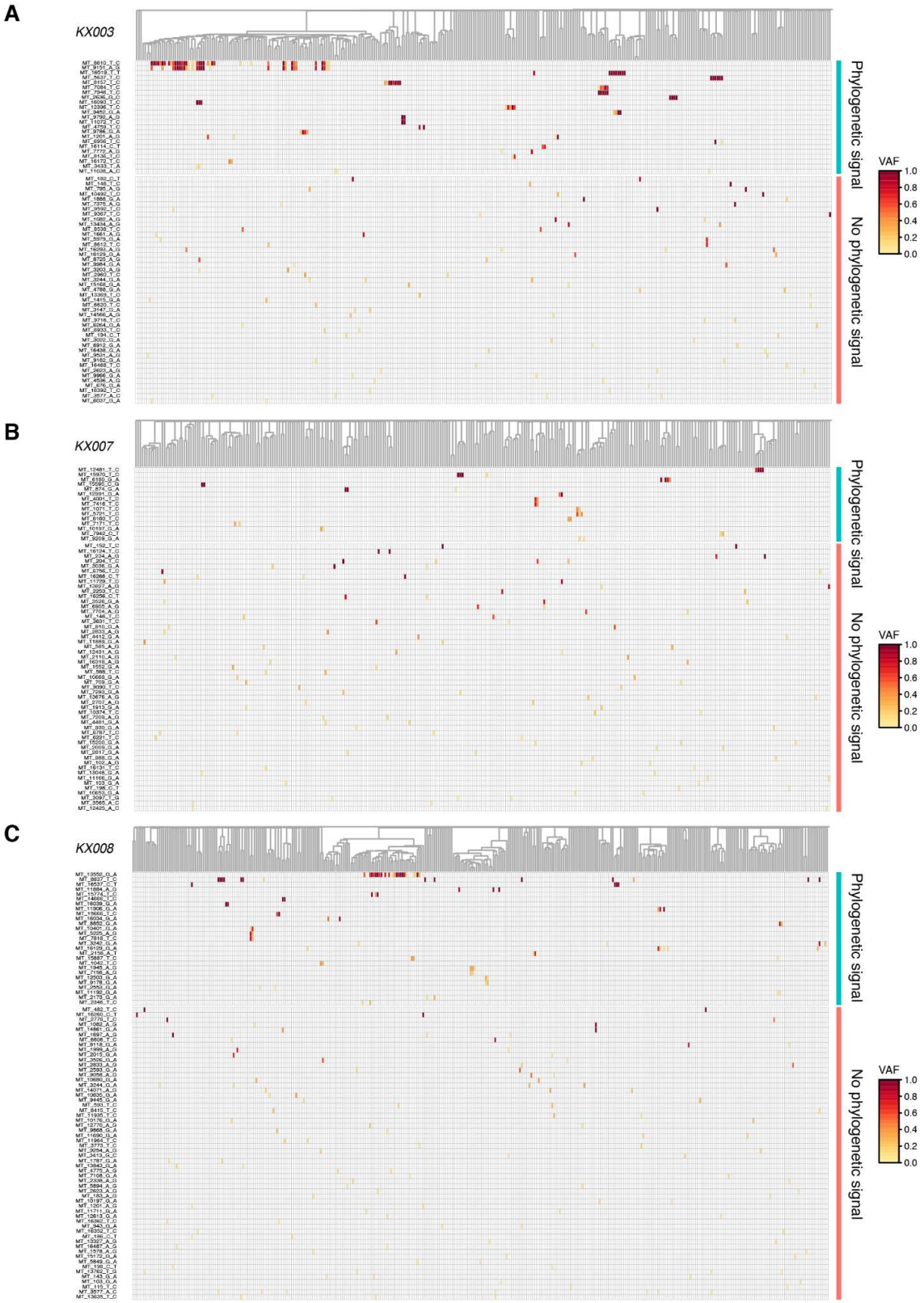


Extended Data Fig. 3| Functional somatic mutations and selection in age groups. a, Functional mitochondrial mutation burden for synonymous, missense and truncating mutations represented as the sum of variant allele fraction (VAF) across all samples per patient equivalent to Fig. 1A. **b,** Bar plot showing the total number and proportion of functional mitochondrial mutations for synonymous, missense and truncating mutations per age groups, coloured according to the variant allele fraction

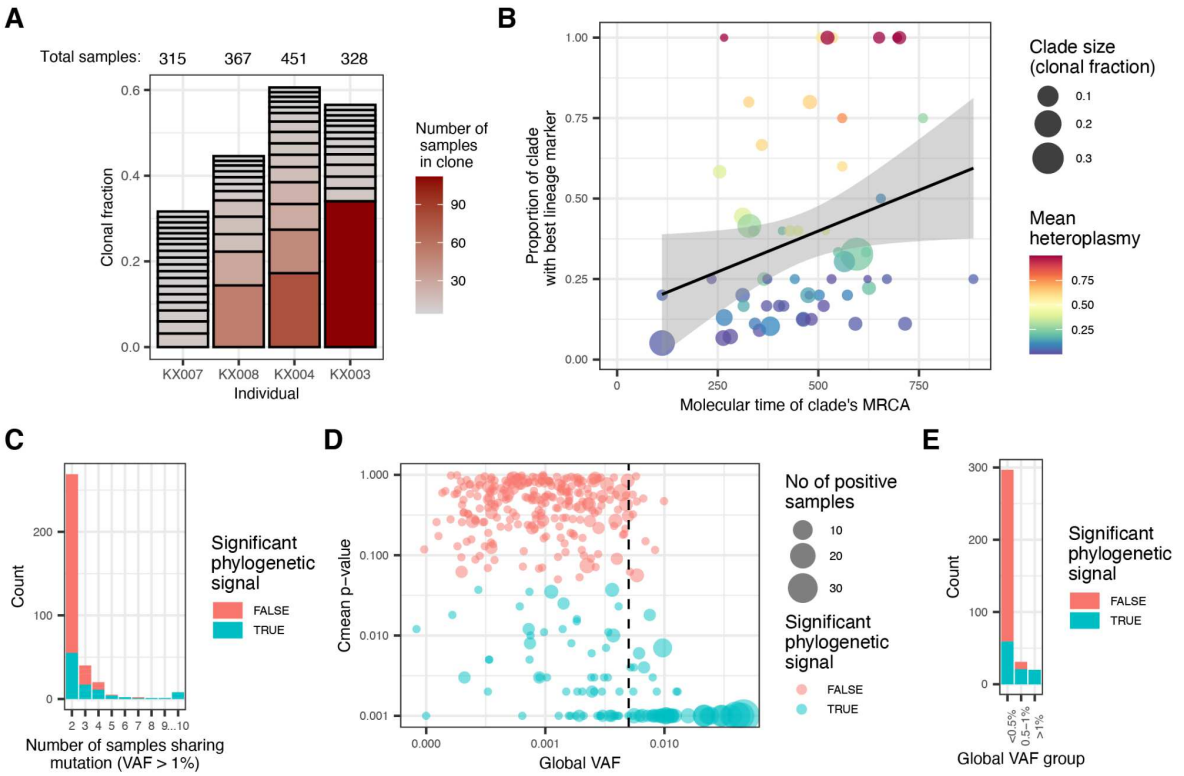
(VAF) category. **c**, Gene-level d*N*/d*S* ratios for missense and nonsense mutations across all coding mitochondrial genes on the heavy strand across age groups. Note that *MT-ND6* is on the light strand and was not considered for this analysis. Only point estimates which have a significant q-value as provided by dndscv (q-value ≤ 0.05) are highlighted in cells.



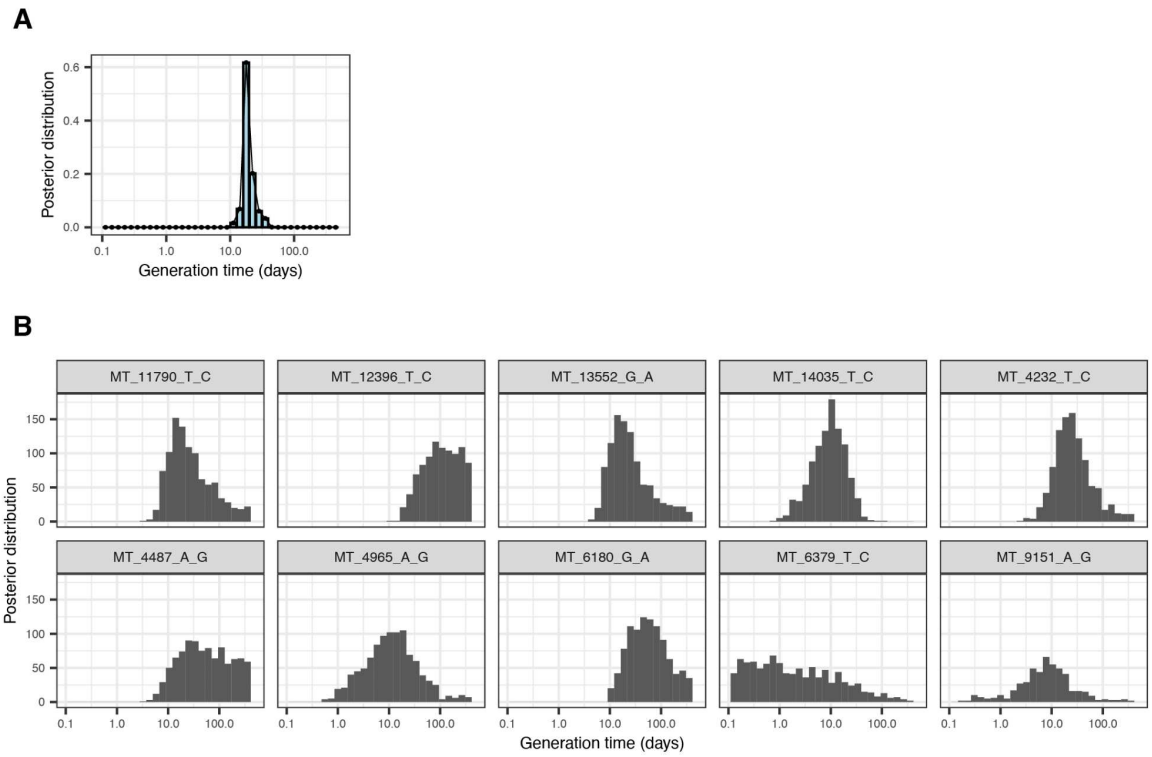
Extended Data Fig. 4| Timing of mutation acquisition for mutations present in different VAF ranges. Wright-Fisher model of heteroplasmic drift illustrates that after 1200 generations (equivalent to an elderly individual of approximately 80 years), mutations in different VAF ranges have been acquired at different times. Low VAF mutations have invariably been acquired in recent generations. High VAF mutations have more variable time of acquisition, but are more likely to have been acquired early in life. (VAF = Variant Allele Fraction)



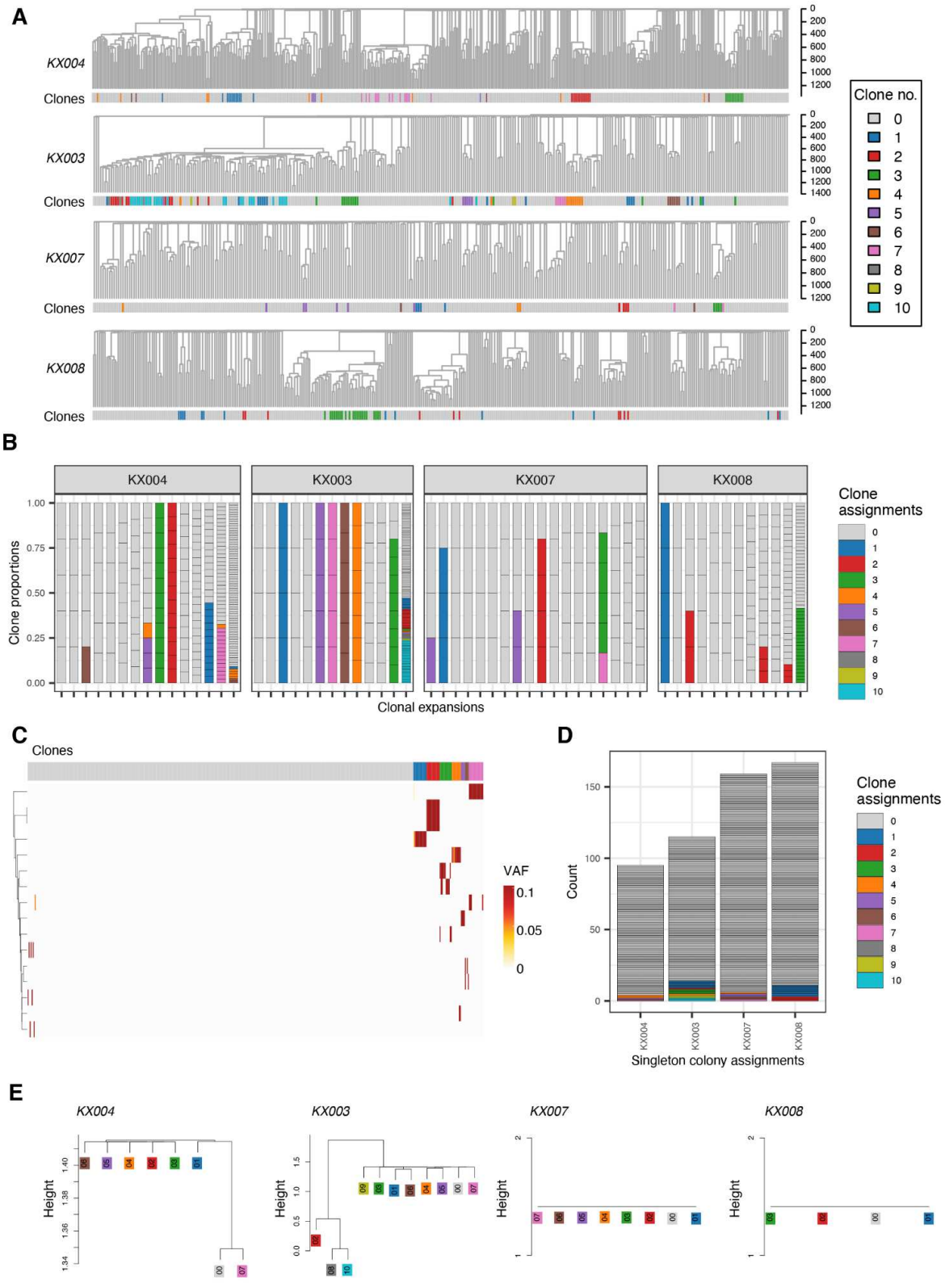
Extended Data Fig. 5| Shared mitochondrial mutations. a - c, Similar to Fig. 5a, heatmaps of shared mitochondrial mutations for KX003 (a), KX007 (b) and KX008 (c). Each row represents a mitochondrial mutation (labeled with the mutation reference in format MT-Position-Reference-Alternative), and each column a sample, ordered by the SNV-based phylogeny (shown above). Mutations are grouped by those that correlate with the SNV phylogeny, and those that do not. SNV, Single nucleotide variant; VAF, variant allele fraction; MRCA, Most recent common ancestor.



Extended Data Fig. 6| Correlation between shared mtDNA mutations and the phylogeny. a, Overview of the clonal expansions in each individual, including all post-embryonic clones with >1% clonal fraction. **b,** Proportion of a clonal expansion with the best mitochondrial marker mutation for that clone vs the molecular time of the clone's most recent common ancestor (MRCA). The black line represents the relationship estimate by linear regression, p-value = 0.04, R^2=0.056. The gray shaded area represents the 95% CI. **c,** Barplot indicating the number of samples sharing each mutation, coloured by whether the mutation correlates with the phylogeny. **d**, Scatter plot of mitochondrial mutation phylogenetic signal p-values (Cmean p-value < 0.05 indicates significant correlation), by the global VAF. Dot size indicates the number of samples sharing the mutation. Vertical dashed line is at global VAF of 0.5%, the cut-off used for inclusion of mutations for inferring clones. **e,** Numbers of mitochondrial mutations with or without phylogenetic signal, divided into three groups by global VAF. mtDNA, mitochondrial DNA; VAF, variant allele fraction; MRCA, Most recent common ancestor; CI, Confidence interval.



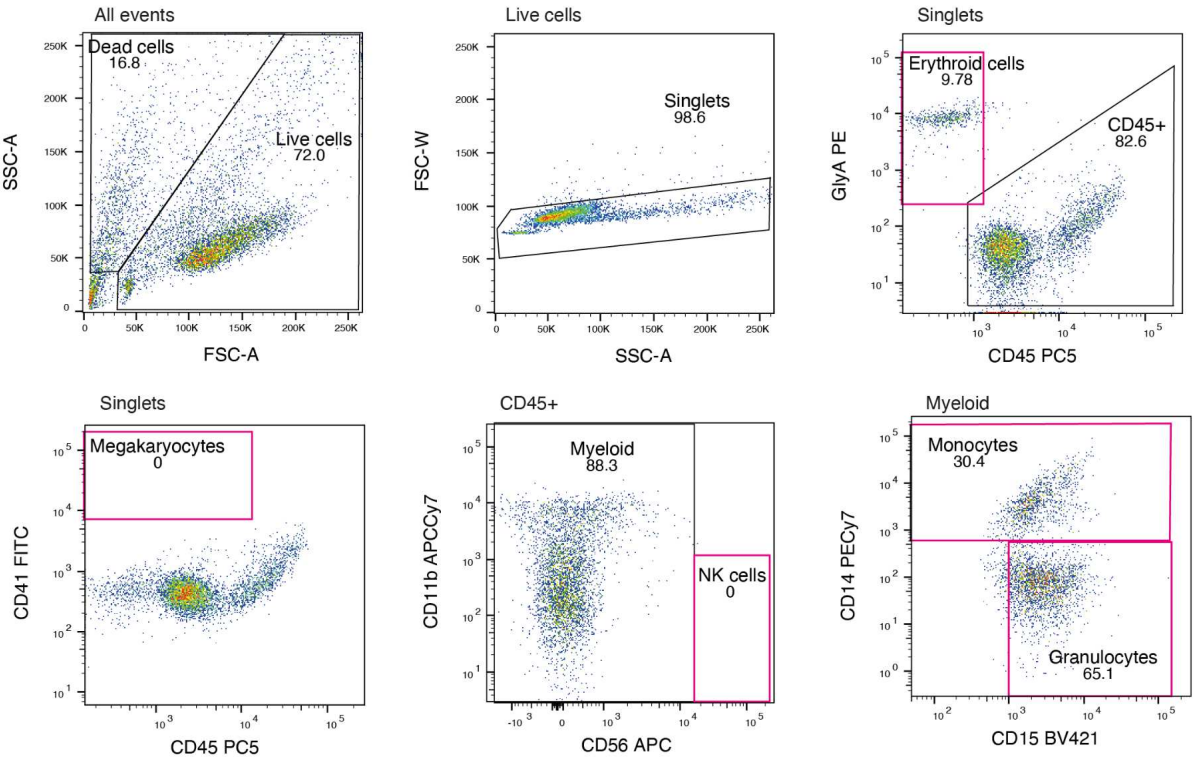
Extended Data Fig. 7| Inference of mtDNA allelic drift through VAF distributions across clonal expansions. a, Posterior distribution for the generation time in adult HSCs assessed using an ABC leveraging the VAF distributions across clonal expansions of specific mutations. The simulations within this ABC were performed using a fixed population size of 675 (the mean mtDNA genome copy number across the 4 individuals considered). The corresponding inferred drift parameter was determined as 12,000 mitochondria days (95% CI, 8,500 - 26,900). **b**, Posterior distributions for the generation time in adult HSCs as inferred from individual mutation distributions that were combined to create (A). mtDNA, mitochondrial DNA; VAF, variant allele fraction; ABC, Approximate Bayesian computation, CI, Confidence interval.



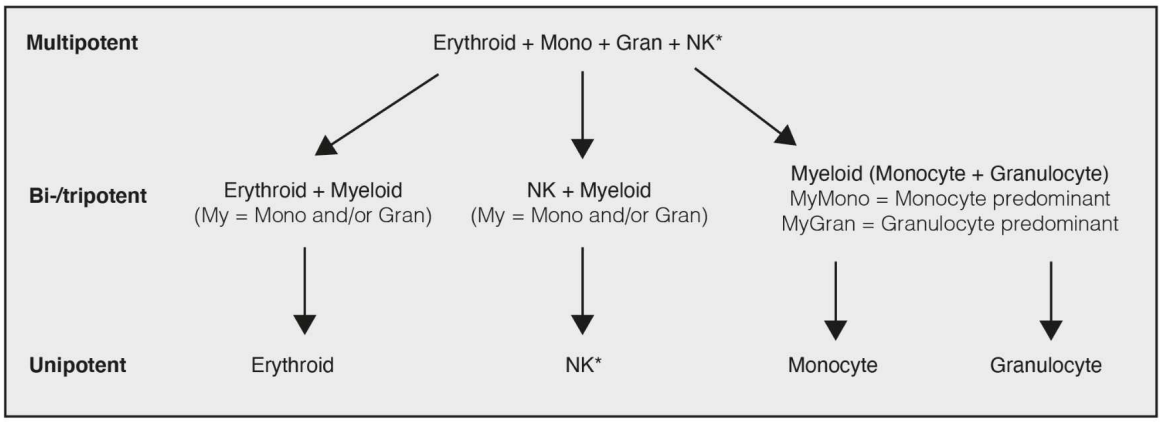
Extended Data Fig. 8| Inferring clones using mitochondrial mutations. a, Clones inferred using mitochondrial mutations for KX004, KX003, KX007 and KX008, highlighted as barplots compared to the high-confidence SNV-based phylogenies above. **b**, Bar plots for all four patients highlighting the proportion of SNV-based clonal expansions overlapping with the inferred mitochondrial clones highlighted in a. **c**, Heatmap of shared mutations in KX004 utilised for clustering (*y* axis) in all samples (*x* axis) coloured according to the VAF and grouped according to clone assignment. **d**, Singleton

colonies that have been assigned to clones by using mitochondrial mutations. In these cases, such assignment does not reflect genuine clonal relationships as these samples are unrelated to any other going right back to embryonic development. **e**, Dendrograms reflecting the clonal relationship between each mitochondrial clone across the four patients highlighted in **a**. VAF, variant allele fraction.

A



B



* No examples of these colony types in this dataset

Extended Data Fig. 9| Mature cell phenotyping of colonies. a, Gating strategy used for immunophenotyping of mature cells in single HSPC-derived colonies from Mitchell et al [26]. The example colony is classified as EryMy (Erythroid + Mono + Gran). **b**, Classification scheme for the different colony phenotypes (note that multipotent colonies and NK-only colonies were not observed in this dataset).

Supplementary Tables

Supplementary Table 1| Patient metadata.

Supplementary Table 2| mtDNA copy number estimates per sample.

Supplementary Table 3| Cell type and immunophenotype information per sample.

Supplementary Table 4| Annotated mtDNA single nucleotide variants per patient.

Supplementary Table 5| Antibody information for immunophenotyping of flow-sorted colonies.