



Bottleneck and selection in the germline and maternal age influence transmission of mitochondrial DNA in human pedigrees

Arslan A. Zaidi^{a,1,2}, Peter R. Wilton^{b,1}, Marcia Shu-Wei Su^{a,1,3}, Ian M. Paul^c, Barbara Arbeithuber^a, Kate Anthony^a, Anton Nekrutenko^d, Rasmus Nielsen^{b,e,4}, and Kateryna D. Makova^{a,4}

^aDepartment of Biology, Penn State University, University Park, PA 16802; ^bDepartment of Integrative Biology, University of California, Berkeley, CA 94720; ^cDepartment of Pediatrics, Penn State College of Medicine, Hershey, PA 17033; ^dDepartment of Biochemistry and Molecular Biology, Penn State University, University Park, PA 16802; and ^eDepartment of Statistics, University of California, Berkeley, CA 94720

Edited by Anne C. Stone, Arizona State University, Tempe, AZ, and approved October 17, 2019 (received for review April 12, 2019)

Heteroplasmy—the presence of multiple mitochondrial DNA (mtDNA) haplotypes in an individual—can lead to numerous mitochondrial diseases. The presentation of such diseases depends on the frequency of the heteroplasmic variant in tissues, which, in turn, depends on the dynamics of mtDNA transmissions during germline and somatic development. Thus, understanding and predicting these dynamics between generations and within individuals is medically relevant. Here, we study patterns of heteroplasmy in 2 tissues from each of 345 humans in 96 multigenerational families, each with, at least, 2 siblings (a total of 249 mother–child transmissions). This experimental design has allowed us to estimate the timing of mtDNA mutations, drift, and selection with unprecedented precision. Our results are remarkably concordant between 2 complementary population-genetic approaches. We find evidence for a severe germline bottleneck (7–10 mtDNA segregating units) that occurs independently in different oocyte lineages from the same mother, while somatic bottlenecks are less severe. We demonstrate that divergence between mother and offspring increases with the mother's age at childbirth, likely due to continued drift of heteroplasmy frequencies in oocytes under meiotic arrest. We show that this period is also accompanied by mutation accumulation leading to more de novo mutations in children born to older mothers. We show that heteroplasmic variants at intermediate frequencies can segregate for many generations in the human population, despite the strong germline bottleneck. We show that selection acts during germline development to keep the frequency of putatively deleterious variants from rising. Our findings have important applications for clinical genetics and genetic counseling.

mitochondrion | heteroplasmy | bottleneck

The human mitochondrial genome (mtDNA) is a circular ~16.5 kilobase-long nonrecombining DNA present in multiple copies per cell. It is inherited almost exclusively from the mother (1, 2). MtDNA has a higher mutation rate than the nuclear genome (3, 4). Mutations can result in the presence of a mixture of mtDNA haplotypes within an individual, a phenomenon called heteroplasmy. A healthy human carries, on average, 1 heteroplasmy with minor allele frequency (MAF) ≥ 0.01 (5). Heteroplasmies can lead to mitochondrial diseases (6), which occur in, at least, 1 in 5,000 people (7), and the severity of such diseases depends on the frequency of the pathogenic allele in a tissue (8, 9). Changes in heteroplasmy allele frequency between generations are thought to be facilitated by the germline mtDNA bottleneck—a reduction in the effective mtDNA content during oogenesis common in many species (reviewed in ref. 10) among which it differs in size (*SI Appendix, Table S1*). The germline mtDNA bottleneck size has been studied by applying population genetic theory to changes in heteroplasmy allele frequencies between mothers and their offspring, and/or by directly measuring changes in the number of mtDNA molecules during oogenesis (10). Population genetic approaches infer the number of mtDNA segregating units consistent with the observed genetic drift (i.e., the effective bottleneck

size), which was found to be ~10–30 segregating units in humans (5, 11). Direct measurement of mtDNA content in humans suggests that a reduction in mtDNA copies in the germline (~1,500) (12) is not as drastic as that inferred from heteroplasmy allele frequency shifts. Importantly, the human effective bottleneck size is relevant to genetic counseling when predicting the chances of an offspring to inherit the mitochondrial disease from a carrier mother. Thus, estimating the germline bottleneck size and understanding other aspects of mtDNA transmission is important for human health.

In addition to the germline bottleneck, heteroplasmy allele frequencies can be affected by the potential decrease in the

Significance

Mitochondria frequently carry different DNA—a state called heteroplasmy. Heteroplasmic mutations can cause mitochondrial diseases and are involved in cancer and aging, but they are also common in healthy people. Here, we study heteroplasmy in 96 multigenerational healthy families. We show that mothers effectively transmit very few mitochondrial DNA to their offspring. Because of this bottleneck, which intensifies with increasing maternal age at childbirth, mutation frequencies can change dramatically between a mother and her child. Thus, a child might inherit a disease-causing mutation at high frequency from an asymptomatic carrier mother and might develop a disease. We also demonstrate that natural selection acts against disease-causing mutations during germline development. Our study has important implications for genetic counseling of mitochondrial diseases.

Author contributions: I.M.P., R.N., and K.D.M. designed research; A.A.Z., P.R.W., M.S.-W.S., B.A., and K.A. performed research; A.A.Z., P.R.W., A.N., and R.N. contributed new reagents/analytic tools; A.A.Z., P.R.W., and K.D.M. analyzed data; and A.A.Z., P.R.W., and K.D.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

This open access article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Data deposition: Heteroplasmy allele frequencies are provided in [Dataset S2](#). The raw sequencing reads have been deposited to the Sequence Read Archive under BioProject ID [PRJNA565594](https://ncbi.nlm.nih.gov/bioproject/PRJNA565594). Code is available at https://github.com/makovalab-psu/Mt_heteroplasmy and <https://github.com/makovalab-psu/opl>.

¹A.A.Z., P.R.W., and M.S.-W.S. contributed equally to this work.

²Present address: Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19103.

³Present address: Department of Biotechnology and Laboratory Science in Medicine, School of Biomedical Science and Engineering, National Yang Ming University, Taipei, 112-21 Taiwan.

⁴To whom correspondence may be addressed. Email: rasmus_nielsen@berkeley.edu or kdm16@psu.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1906331116/-DCSupplemental>.

First published November 22, 2019.

content of mtDNA during embryonic development by multiple mitoses randomly partitioning variable mtDNA between somatic cells during a lifetime (13) as well as by selection. Deciphering the timing and relative strength of germline vs. somatic bottlenecks has important implications for the efficacy of preimplantation diagnosis of mitochondrial disease (13). Recently, we estimated the effective bottleneck sizes during the embryonic development to be substantially less severe than the germline bottleneck (14). Selection against deleterious mtDNA mutations was shown to be an important force in the mouse germline (reviewed in ref. 13) and in *Drosophila* (15). In humans, studies examining selection at heteroplasmic variants have been lacking (but see ref. 11).

Here, we investigate heteroplasmy allele frequencies in 96 human multigenerational maternal lineages with no known mitochondrial disease with, at least, 2 generations per family, at least, 2 offspring per mother, and 2 tissues per individual. This design enables us to infer the timing of mtDNA mutations and allele frequency changes with greater resolution than in previous studies limited to either single-child families (5) or single tissues (11, 16). Moreover, examination of healthy families ensures that our results are not biased by ascertainment of clinically affected individuals (17). Using 2 complementary population-genetic approaches accounting for the phylogenetic relatedness among sampled tissues and individuals (presented in this paper and in ref. 14), we provide a comprehensive analysis of the timing, severity, and dependence on age of genetic drift, mutation, and selection acting during germline and somatic ontogenetic processes that influence heteroplasmy allele frequency segregation in the human body.

Results

Heteroplasmy Discovery. Single-nucleotide heteroplasms were called in 345 individuals from 96 families. Each family consisted of 1 mother and, at least, 2 (and up to 5) children and up to 4 generations (a total of 249 mother-child transmissions; *SI Appendix, Fig. S1*). These included 39 mother-1-child pairs analyzed by us

previously (5). For the remainder, we sequenced buccal and blood samples to high depth (mean $\sim 11,400\times$, *SI Appendix, Fig. S2*). After filtering (see *Materials and Methods* and *SI Appendix, Fig. S3*), we identified 668 heteroplasms with $\text{MAF} \geq 0.01$ (*Dataset S1*) in 690 tissue samples (345 individuals \times 2 tissues), amounting to approximately 1 heteroplasmy per sample, corroborating previous findings (5). Most heteroplasms were present at low frequency (median $\text{MAF} = 0.029$; *SI Appendix, Fig. S4*). A total of 133 out of 150 heteroplasms with $\text{MAF} \geq 0.1$, 133 were successfully validated with Sanger sequencing (see *Materials and Methods*). The remaining 17 high-frequency heteroplasms occur in or near low-complexity regions; thus, we removed all such sites from further analyses (*SI Appendix, Fig. S3*). From 668 heteroplasms, we inferred 346 independent mutations assuming that heteroplasms at the same site in a family represent a single mutation (if this assumption is incorrect, we would underestimate the number of independent mutation events, but this should not substantially affect our conclusions regarding the mutation spectrum). These mutations were present at 312 mtDNA sites; 288 sites segregated in 1 family only, 18 sites—in 2 families and 6 sites (195, 214, 234, 12,488, 16,129, and 16,172)—in 3 or more families. All mutations at sites that occurred in 2 or more families were observed at different haplogroups among families, suggesting that these are mutation hotspots. Among the 346 mutations, we observed a high transition-to-transversion ratio (of 30.5) as was shown previously (4, 5). Most variants were $C > T/G > A$ (130) and $T > C/A > G$ (202) transitions (*SI Appendix, Fig. S5A*), likely arising from polymerase γ replication errors (18). Spontaneous deamination of methylated cytosines might have contributed to transitions as well because $C > T$ mutations were overrepresented in CpG context (2-sided binomial test $P = 1.57 \times 10^{-03}$, *SI Appendix, Fig. S5B*); although methylation in human mtDNA is disputed (e.g., ref. 19). Mutations were enriched in the D-loop (2-sided binomial test $P = 4.1 \times 10^{-18}$, *SI Appendix, Table S2*), confirming its high mutation rate (20). The overall patterns of heteroplasmy

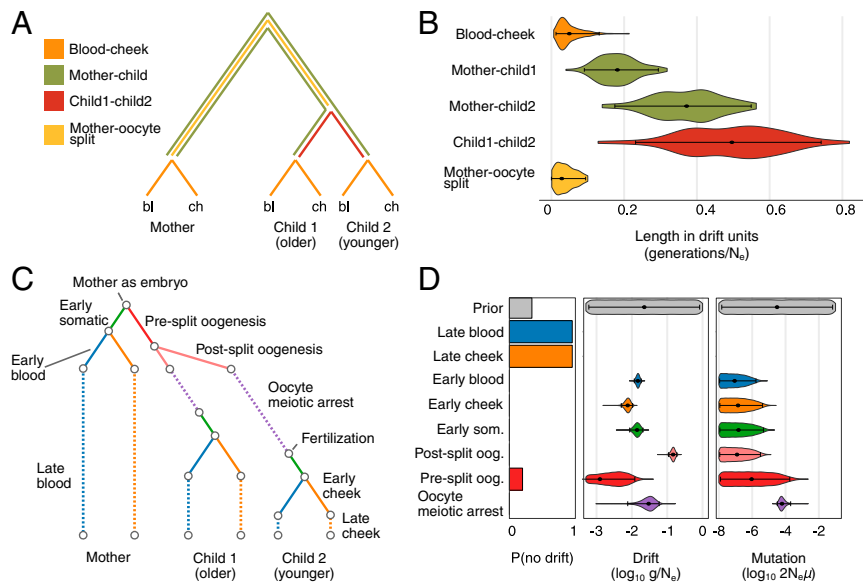


Fig. 1. Drift in heteroplasmy frequency estimated for different stages of ontogenetic development with the BLS (A and B) and OPL (C and D) approaches. (A) Assumed phylogeny for an example 1-mother-2-children family showing branch lengths that can be estimated in the BLS framework. Branches (bl = blood, ch = cheek, child 1 = older child, child 2 = younger child) are not drawn to scale. (B) BLS-estimated distribution of mean lengths (in drift units on the x axis) for each branch in the phylogeny (y axis). The colors correspond to branches in A. (C) Ontogenetic phylogeny assumed under the OPL model for the same 1-mother-2-children pedigree as in A. Each solid branch represents an ontogenetic process assumed to impart the same amount of genetic drift and mutation in every individual. Each dashed line represents an ontogenetic process in which genetic drift and mutation are assumed to accumulate linearly from birth until sampling or childbirth. Branches are not drawn to scale. (D) OPL-estimated posterior distribution of genetic drift and mutation rates. The latter are scaled by an effective population size N_e for different ontogenetic processes. The colors correspond to processes in C.

allele frequency were inconsistent with paternal inheritance. We observed an average of ~33 fixed sites among a pair of unrelated individuals in our sample and, thus, would expect this many heteroplasmies at high frequencies in a single individual due to paternal inheritance—this pattern was not found. However, we could not examine this phenomenon explicitly because samples from the fathers were not collected.

Approaches Used to Study Heteroplasmy Transmission. We studied the dynamics of heteroplasmy transmission between generations and among tissues using 2 complementary population-genetic approaches. First, we introduce an approach to study the transmission dynamics of heteroplasmies in families, which is based on the branch-length framework underlying locus-specific branch length (21), population-branch statistic (22), and ancestral branch statistic (23). Namely, we estimate branch lengths along a fixed phylogenetic tree (Fig. 1 *A* and *B*) using pairwise divergence measures based on F_{ST} (21–23). This approach (hereafter referred to as the branch-length statistic [BLS]; see *Materials and Methods*) provides a framework for inferring changes in allele frequency at different life stages in a 1-mother–2-children pedigree (Fig. 1*A*) from the pairwise divergence between different groups of tissues. BLS assumes that the divergence between any 2 nodes on the tree is influenced by genetic drift and potentially by selection but not by mutation.

Second, in the ontogenetic phylogeny likelihood (OPL) approach (14), we estimate posterior distributions of genetic drift and mutation rate parameters in a full likelihood model of heteroplasmy frequency segregation along the branches of ontogenetic phylogenies. The OPL method explicitly defines several ontogenetic processes on the phylogeny (e.g., early oogenesis, pre- and postgastrulation somatic developments, and adult hematopoiesis; shown for a 1-mother–2-children pedigree in Fig. 1*C*) and models genetic drift and mutation rate parameters for each of these processes (Fig. 1*D*). It models both de novo and recurrent mutations separately from genetic drift but, in its current implementation, does not consider selection.

For both approaches, we used the same timing of ontogenetic events to construct phylogenies for 1-mother–2-children pedigrees (Fig. 1 *A* and *C*) and for other pedigree structures present in our data (*SI Appendix*, Figs. *S1* and *S6*). Overall, our conclusions were qualitatively similar between the 2 approaches.

Drastic mtDNA Bottleneck in the Germline. Both OPL and BLS indicate stronger genetic drift for heteroplasmies in the germline than somatic processes, i.e., for internal rather than external branches of the ontogenetic phylogeny (Fig. 1). With BLS, the divergence between mother and offspring (i.e., green branch of Fig. 1*A*) is 0.283 generations per effective population size or g/N_e (median; 95% bootstrap CI 0.098–0.574; Fig. 1*B*). With OPL, the genetic drift between mother and offspring (sum of red, pink, purple, and green branches in Fig. 1*C*) is 0.208 g/N_e (posterior median; 95% highest posterior density [HPD], interval 0.174–0.245; Fig. 1*D*). These values are considerably higher than the genetic drift during somatic processes (Fig. 1 *B* and *D*). The similarity in genetic drift estimates between the BLS and the OPL approaches suggests that mutation is a much weaker force than drift in shaping heteroplasmy frequencies; BLS drift estimates assume that any change in heteroplasmy frequency is due to drift alone, while OPL models the effect of both mutation and drift.

We use the above estimates of genetic drift to quantify the germline bottleneck using the approximation $B = 2/D$, where B is the effective bottleneck size and D is the amount of genetic drift in units of g/N_e (*SI Appendix*, *SI Materials and Methods*). Using the BLS approach, we estimate an effective bottleneck size of ~7.1 mtDNA segregating units (mean; 95% bootstrap CI = 3.5–20.4) for the divergence between the mother’s tissues and her offspring’s tissues (green branch of Fig. 1*A*). Using the OPL framework, we quantify 2 effective germline bottleneck sizes. The first—the

“oogenic bottleneck”—represents the genetic drift occurring during early oogenesis, prior to the onset of meiotic arrest (red and pink branches of Fig. 1*C*). The second—the “combined germline bottleneck”—measures genetic drift that occurs after the establishment of the mother’s germline but before the divergence of the somatic cell lineages at gastrulation in the offspring’s embryo, including drift during the oogenic bottleneck, oocyte meiotic arrest, and fertilization and pregastrulation embryogenesis in the offspring (red, pink, purple, and green branches of Fig. 1*C*). With OPL, we estimate the oogenic and combined germline bottleneck sizes to be 13.4 (posterior median; 95% HPD 10.2–17.3) and 10.3 (8.5–12.3) mtDNA segregating units, respectively. Applying a published method (5, 24, 25) to the effective bottleneck of 10.3 and the observed median of 71 germline (shared by 2 somatic tissues) heteroplasmies among 96 unrelated individuals (1 individual per family sampled at random), we estimated the germline mutation rate as 4.72×10^{-7} mutations per site per generation (95% bootstrap CI: 3.93 – 5.52×10^{-7}).

Independent Germline Bottleneck in Each Oocyte Lineage. With, at least, 2 children per mother, we can distinguish whether the bottleneck occurs prior or following the divergence of 2 oogonia. If the bottleneck occurs before the oogonia diverge, the change in allele frequency between a mother and her offspring would be similar among siblings. Alternatively, if the germline bottleneck occurs after oogonia diverge, changes in allele frequencies between mother and offspring would be largely uncorrelated among siblings. We observe weaker correlations for tissue-averaged heteroplasmy allele frequencies between 2 siblings ($r = 0.319$) than between a mother and her child ($r = 0.551$; *SI Appendix*, Fig. *S7*). Furthermore, the BLS approach shows that the internal divergence is greater between 2 children (BLS median = 0.501 drift units for the red branch; 95% CI = 0.231–0.761) than between a mother and her children (0.283 drift units for the green branch in Fig. 1*A*; 95% CI = 0.098–0.574). It also demonstrates that 85.3% (median; 95% CI = 71.2–100%) of the divergence between a mother and her children can be attributed to changes in allele frequency after oocyte lineages split (i.e., [green–yellow]/green branches in Fig. 1*A*). According to OPL estimates, the fraction of genetic drift after oogonial divergence is 99.1% (posterior median; 95% HPD 93.4–100%) of the total genetic drift during germline development (*SI Appendix*, Fig. *S8*). These results strongly argue for the germline bottleneck occurring after divergence of oogonia.

Divergence Between a Mother and Her Child Increases with the Age of the Mother at Childbirth. Such a relationship would be expected if heteroplasmies diverge in frequency during oocyte meiotic arrest as a result of de novo mtDNA mutations and/or mitochondrial turnover via mitophagy and biogenesis. We found that the

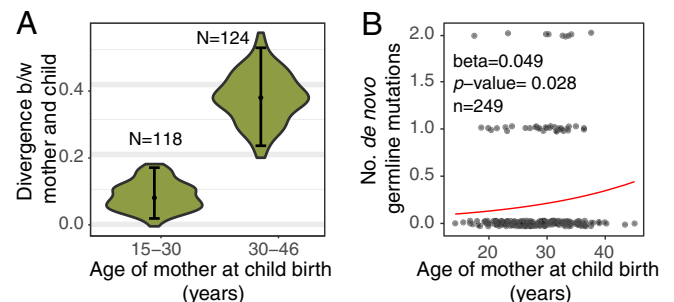


Fig. 2. Effect of age of the mother at childbirth on (A) the divergence (in drift units) of heteroplasmies (for 118 and 124 mothers who gave birth to 15–30 and 30–46 y of age, respectively), and (B) the number of putative germline de novo mutations in her children (50 mutations overlapping between the OPL and the heuristic approaches were included, 249 mothers were analyzed).

correlation in somatic allele frequency is lower between older mothers (age ≥ 30 y) and their children ($r = 0.67$) than between younger mothers and their children ($r = 0.92$, *SI Appendix, Fig. S9*). Moreover, the BLS shows a much higher mother–child divergence (green branch of Fig. 1A) in allele frequency for mothers who gave birth at an older (≥ 30 y; median = 0.361 drift units; 95% bootstrap CI = 0.225–0.505) vs. younger (median = 0.076 drift units; 95% bootstrap CI = 0.017–0.162) age (Fig. 2A). Lastly, using the OPL framework, we estimate the rate of accumulation of genetic drift during meiotic arrest (i.e., purple branch of Fig. 1C) to be 0.001 drift units per year (95% HPD 1.9×10^{-4} – 1.9×10^{-3} ; posterior $P[\text{drift} > 0] = 1.0$).

Mitochondrial Heteroplasmy Can Persist for Many Generations.

With >2 generations sampled in 9 families (*SI Appendix, Fig. S1*), our sample provides an opportunity to characterize heteroplasmy persistence across multiple generations. Within these families, 5/10 heteroplasmies present in a grandmother persisted in, at least, 1 of her grandchildren. In the 2 families with 4 generations, both heteroplasmies present in the great-grandmother persisted for the subsequent 3 generations (*Dataset S1*). Thus, heteroplasmies may commonly be inherited across multiple generations. We then estimated (within the OPL framework, see *SI Appendix, SI Materials and Methods*) the number of generations for which a heteroplasmy persists before fixation or loss (*SI Appendix, Fig. S10A*). Low-frequency heteroplasmies tend to be quickly lost from the germline with those at MAF = 0.02 in the mother's germline persisting, on average, 0.748 generations. In contrast, a heteroplasmy at MAF = 0.5 in the germline will persist for 7.26 generations on average. Because of a high number of low-frequency heteroplasmies in our dataset (*SI Appendix, Fig. S10B*), the average heteroplasmy lifespan is estimated to be just 0.171 generations. However, several heteroplasmies have MAF = 0.25–0.50 (*SI Appendix, Fig. S4*) and, thus, may be 5–10 generations old.

De Novo Germline Mutations. We used 2 approaches to identify de novo germline mutations in our dataset. First, heuristically, we identified heteroplasmies present at MAF ≥ 0.01 in both tissues of an offspring but absent (i.e., present below noise level of MAF < 0.002 , corresponding to 3.16 SE above the nominal error for a Phred score of 30 at a coverage of 10,000 \times) in all other samples from the same family. This resulted in a total of 78 putative de novo mutations (*SI Appendix, Table S3*), which could have arisen in the mother's germline or in the embryo before the blood and buccal tissues diverged. All 5 out of 5 randomly selected putative de novo germline mutations were successfully validated with droplet digital PCR ([ddPCR], see *Materials and Methods*).

Second, we used the OPL framework to calculate the posterior probability that a given heteroplasmy is the product of a de novo germline mutation somewhere on the ontogenetic phylogeny (see *Materials and Methods*). The distribution of these probabilities across heteroplasmies was bimodal with some strongly indicating de novo mutation and others supporting ancestral polymorphism as the source of heteroplasmy (*SI Appendix, Fig. S11*). We identified 86 heteroplasmies (*SI Appendix, Table S3*) with a posterior probability of being de novo exceeding a visually informed threshold of 0.8. All such heteroplasmies were predicted to arise during oocyte meiotic arrest (purple branch in Fig. 1C and *SI Appendix, Fig. S11*) consistent with the inferred mutation rate being 100–1,000 \times higher during oocyte meiotic arrest than any other ontogenetic process (Fig. 1D).

We find evidence for a positive relationship between the number of de novo germline mutations and the mother's age at childbirth. The number of de novo mutations identified with OPL increases with age at childbirth ($\beta = 0.042$, $P = 0.015$; Poisson regression; *SI Appendix, Fig. S12A*) with an increase of ~ 0.3 mutations transmitted between 20 and 40 y of age. There is a marginally significant positive relationship between age at childbirth and

number of inherited de novo mutations identified heuristically ($\beta = 0.028$, $P = 0.087$; *SI Appendix, Fig. S12B*). Using the 50 mutations overlapping between heuristic and OPL approaches, there is a statistically significant increase in the number of de novo mutations with age at childbirth ($\beta = 0.049$, $P = 0.028$; Fig. 2B) with an increase of ~ 0.2 mutations transmitted between ages 20 and 40.

Limited Drift and Mutagenesis during Somatic Development. Heteroplasmies experience only relatively limited genetic drift since the split of buccal and blood tissue lineages as evidenced from the high correlation in allele frequency ($r = 0.97$, *SI Appendix, Fig. S7C*) and low divergence (BLS: mean = 0.045, 95% CI = 0.009–0.119 drift units; OPL: posterior median = 0.022, 95% CI = 0.019–0.026 drift units) between the 2 tissue types. Using the OPL framework, we find that the effective somatic bottleneck size is smaller for blood (posterior median: 134.9; 95% CI = 106.6–168.0) than for cheek (posterior median: 262.9; 95% CI = 181.0–379.4), consistent with our previous results (14). Nevertheless, each of them is only $\sim 10\%$ of the amount of drift separating a mother and child (see above), supporting a more pronounced bottleneck in the germline than in somatic tissues.

As individuals age, we expect the variance in heteroplasmy allele frequency among their tissues to increase because of genetic drift due to mitotic segregation and mtDNA turnover. Surprisingly, the blood–buccal divergence was not significantly different between older and younger individuals in our data set (*SI Appendix, Fig. S13*). Likewise, the OPL analysis suggests that genetic drift does not accumulate with age in the adult somatic tissues, i.e., posterior $P(\text{drift} = 0) = 1$.

An association between age and number of de novo somatic mutations was tissue dependent. We identified 57 putative de novo somatic mutations (*SI Appendix, Table S4*), which are sites that are present at MAF ≥ 0.01 in only 1 tissue of an individual (14 and 43 de novo mutations were found in blood and buccal tissue, respectively) and are absent (i.e., MAF < 0.002) in all other samples from the same family. The number of de novo mutations in the blood increases with the age of the individual at collection (Poisson regression: $\beta = 0.042$, $P = 8 \times 10^{-5}$; *SI Appendix, Fig. S14*) which corresponds to an increment of ~ 0.4 mutations between birth and 80 y of age. We did not find a significant association between age and the number of de novo mutations in the buccal tissue (*SI Appendix, Fig. S14*). The OPL framework inferred no opportunity for mutation accumulation with age in the adult somatic tissues (posterior probability = 1.0; Fig. 1D).

Purifying Selection Against Nonsynonymous and Pathogenic Mutations in the Germline.

We observe a depletion of mutations (considering the total set of 346) in protein-coding genes ($P = 9.60 \times 10^{-7}$, two-sided binomial test; *SI Appendix, Table S2*), likely due to purifying selection (relative rate of nonsynonymous to synonymous heteroplasmy occurrence of 0.43, $P < 0.001$, Monte Carlo simulations, see *SI Appendix, SI Materials and Methods*). There is also potential depletion of mutations at rRNA sites ($P = 0.061$, two-sided binomial test; *SI Appendix, Table S2*). Of the 312 heteroplasmic sites, 32 were at positions predicted to be pathogenic (see *Materials and Methods*). While the differences in median values were also observed (*SI Appendix, Table S5*), the maximum frequency of pathogenic mutations was strikingly lower for pathogenic than for nonpathogenic mutations in blood (maximum frequencies of 0.052 and 0.916, $P < 1 \times 10^{-04}$, bootstrap test, see *SI Appendix, SI Materials and Methods*) and cheek (maximum frequencies of 0.203 and 0.941, $P = 0.033$, Fig. 3A), respectively, presumably because pathogenic mutations are kept below disease-causing frequencies by selection. The differences in (median and maximum) frequencies of the mutant allele were less apparent between nonsynonymous

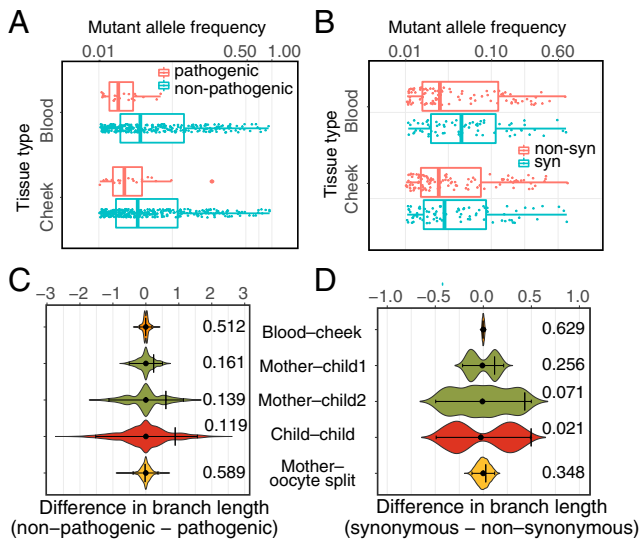


Fig. 3. Effects of selection based on heteroplasmy allele frequencies (*Top row*) and their branch-specific changes (*Bottom row*). (A) Allele frequency distribution for 32 pathogenic vs. 280 nonpathogenic heteroplasms. (B) Allele frequency distribution for 77 synonymous vs. 123 nonsynonymous heteroplasms. See *SI Appendix, Table S5* for summary statistics. (C) Observed differences in branch length between pathogenic and nonpathogenic heteroplasms (vertical black lines) and permuted null distributions of the differences (violin plots). *P* values for each comparison are listed on the right. The colors of the violin plots correspond to branches in Fig. 1A. (D) Like (C) but for synonymous and nonsynonymous mutations.

and synonymous heteroplasms but showed a similar trend (Fig. 3B and *SI Appendix, Table S5*).

Finally, we investigated at which stages of ontogenetic development purifying selection against deleterious mutations is strongest. Using the BLS approach, we compared the amount of drift experienced by neutral vs. deleterious heteroplasms at different stages of the ontogenetic phylogeny. We find that divergence during germline transmission is higher for nonpathogenic than for pathogenic (see permutation test results in Fig. 3C) and for synonymous than for nonsynonymous heteroplasms (Fig. 3D). This is consistent with our observation that purifying selection prevents deleterious mutations from reaching appreciable frequencies (Fig. 3A). Such patterns were not apparent in branches leading up to somatic tissues (Fig. 3C and D). Thus, purifying selection contributes to shaping the distribution of heteroplasmic allele frequencies during germline development.

Discussion

Strong Germline mtDNA Bottleneck. We employed 2 complementary population-genetic frameworks to quantify the effective human germline bottleneck size to be ~ 10 mtDNA segregating units. This estimate is similar to another recent population genetic estimate of 7–9 (11) but is different from the direct measurement of $\sim 1,500$ molecules per primordial germ cell (PGC)—the lowest mtDNA copy number during oogenesis (12). This discrepancy can be explained by the fact that, whereas direct methods measure mtDNA copy number during oogenesis, population-genetic approaches estimate the effective bottleneck size—a measure of the degree of genetic drift experienced by heteroplasms, which is influenced by several additional factors. First, mtDNA molecules in mammals are compartmentalized into units cosegregating during mitosis, reducing the effective mtDNA population size in a cell (11). Each human PGC contains 5 to 6 mtDNA copies per mitochondrion (12), which may cosegregate in nucleoids during mitosis (26). Thus, counting just mitochondria and assuming no heteroplasmy

within a mitochondrion would produce a bottleneck size estimate of ~ 250 – 300 —still much larger than our estimates. Second, the effective bottleneck size includes heteroplasmy segregation during mitotic cell divisions that occur throughout oogenesis, extending into the adult maternal germline and into the early embryogenesis of the offspring prior to gastrulation and might be affected by replication of a subpopulation of mtDNA molecules after implantation (27–30). The OPL method breaks the germline bottleneck into multiple stages: oogenesis, meiotic arrest prior to ovulation, and pregastrulation somatic development. We estimate the effective oogenic bottleneck size to be equal to only ~ 13 segregating units (vs. ~ 10 for the entire germline transmission), suggesting that most drift occurs during oogenesis, prior to oocyte meiotic arrest. Third, selection acting in the germline will further reduce the effective bottleneck size. Thus, both population-genetic inferences of the germline bottleneck size and direct measurements of the mtDNA copy number and organelle counts are necessary to obtain a detailed view of mtDNA transmission dynamics.

Independent Germline Bottleneck in Oocyte Lineages. Our analysis of heteroplasmy allele frequencies in multiple offspring per mother indicated that siblings experience largely independent genetic drift during the germline bottleneck. This is consistent with higher concordance in allele frequency for monozygotic rather than for dizygotic twins noted for humans (11). Also, in mice, there is high variation in allele frequencies among pups from the same mothers (31). This suggests that the bulk of the germline bottleneck in both humans and mice occurs during oocyte maturation after the split of oocyte lineages. This is a rather wide window in development because oocyte lineages may diverge as early as PGC specification or migration, and mtDNA copy numbers recover only when oocytes fully mature (*SI Appendix, Fig. S15*). In mice, the heteroplasmy variance increases after the development of primary oocytes (29, 31, 32), which places the bottleneck during folliculogenesis rather than oogenesis. Whether this is the case in humans is yet to be determined.

mtDNA Divergence and De Novo Mutation Accumulation in Oocytes with Maternal Age. Our results indicate that heteroplasms experience more divergence in MAF in the germline of older than younger mothers. Because the number of mitotic germline cell divisions is expected to be the same for all mothers (33), this suggests that there is mtDNA turnover in oocytes during meiotic arrest, which leads to increasing drift in heteroplasmy frequencies with age. This result is consistent with, and provides an explanation for, the increased drift with the age of the mother we observed previously (14) and needs to be considered in genetic counseling.

We found a significant positive association with maternal age for putative de novo germline mtDNA mutations. This parallels previous studies showing an increase in de novo germline nuclear mutations with paternal and maternal ages (34–36). The overall germline mutation rate we estimated here is $\sim 75\%$ higher than that in ref. 5, however, the 95% CI of our estimate overlaps with the interquartile range of the estimate in ref. 5 (see the discussion of limitations and comparisons with estimates from other species in ref. 5).

The Persistence of Individual Heteroplasms Across Generations. Our analysis of genetic drift in the OPL framework suggests that most heteroplasms are lost from the germline within a single generation owing to their low allele frequency and severe germline bottleneck. This finding corroborates observations made in Holstein cows (37, 38). We also observe that some heteroplasms can persist for several generations in humans. Modeling of drift with OPL suggests that heteroplasms can segregate for many generations once they reach intermediate frequencies. We find that a heteroplasmic allele at MAF = 0.01, if it becomes fixed, will do so on average after ~ 13 generations (see *Materials and Methods*). Accounting

for reproductive stochasticity does not affect these estimates significantly (*SI Appendix*, Fig. S16).

Somatic Divergence and Mutation Accumulation with Age. We found evidence of the accumulation of de novo mutations with age in blood but not in buccal tissue potentially because the former tissue is more proliferative than the latter. A recent study (39) reported age-related increases in de novo mutations and in heteroplasmy allele frequency at certain mtDNA positions in somatic tissues (39), which we do not observe. One possible explanation for this discrepancy is the older age of individuals in their vs. our study (mean age of 56 vs. 20 y, respectively)—particularly if heteroplasmies diverge in frequency because of age-related degeneration of mitochondrial quality control (40).

Selection in the Germline. Our study demonstrates lower allele frequencies for deleterious (i.e., pathogenic and nonsynonymous) rather than neutral (i.e., nonpathogenic and synonymous) heteroplasmy alleles in tissues of healthy humans. This pattern is in stark contrast to more uniformly distributed allele frequencies in tissues of patients with mitochondrial diseases (6) and is consistent with negative selection keeping the allele frequency of deleterious variants below a certain threshold in the general population. These results agree with studies in the mouse (reviewed in ref. 13) and 2 recent studies in humans (39, 41) also demonstrating purifying selection acting during germline transmission of mtDNA. The levels [e.g., mitochondrial or cellular (42)], mechanisms, and precise developmental stages of such selection are yet to be investigated. However, the increase in mitochondrial respiration later in oogenesis may facilitate selection during the germline bottleneck by “exposing” rare deleterious alleles, which may be otherwise functionally compensated by wild-type alleles (10, 12). This could explain why selection signals are not swamped out by strong genetic drift during germline development.

Materials and Methods

Sample Collection, DNA Isolation, Amplification, and Sequencing. Buccal and blood samples were collected with informed consent from 96 multigeneration families in Central Pennsylvania. The study was approved by the Penn State College of Medicine (IRB# 30432EP). All samples were de-identified before use, and genomic DNA was isolated as previously described (43). mtDNA was amplified in 2 (43) or 3 (*SI Appendix*, Tables S6–S8) overlapping fragments, which were mixed at an equimolar ratio and spiked with 5% of pUC18 or PhiX174 DNA or not spiked to monitor contamination (43). DNA libraries were prepared with an Illumina TruSeq DNA PCR-free kit and sequenced on a MiSeq platform, generating paired-end reads of 2 × 300 nucleotides (see *SI Appendix*, *SI Materials and Methods* for details).

Heteroplasmy Discovery. Sequenced read pairs were trimmed for adapters (44) and mapped to a reference including human nuclear genome (GRCh37), revised Cambridge reference mtDNA, PhiX174, and pUC18 genomes using bwa mem (45). Reads aligned to the spike-ins in expected proportions (*SI Appendix*, Fig. S17), suggesting no contamination among adjacent samples. Samples sequenced here and from our previous study (5) were analyzed jointly. Sites with MAF ≥ 0.01 and sequenced at ≥ 1,000× depth in an individual’s tissue were called heteroplasmic. We filtered them further using several quality control criteria (*SI Appendix*, Fig. S3) to generate a list of high-confidence heteroplasmies. All 150 heteroplasmies with MAF ≥ 0.1 were also sequenced with the Sanger method (*SI Appendix*, Tables S9–S12 and Fig. S18) as described in ref. 5. From 78 heuristically identified de novo mutations, we selected and validated 5 with ddPCR (*SI Appendix*, Tables S13–S16 and Fig. S19) as described in ref. 5. For each heteroplasmy, we determined the ancestral and mutant allele and determined pathogenicity as described in *SI Appendix*, *SI Materials and Methods*.

Population Genetic Analyses. With BLS, briefly, we calculate all pairwise genetic distances (D_{xy}) among samples in a fixed phylogenetic tree using (46): $D_{xy} = -2 \ln(1 - F_{ST(xy)})$, where $F_{ST(xy)}$ is the F_{ST} between samples x and y calculated using Hudson’s estimator (47). From pairwise values of D_{xy} for each heteroplasmy, we calculate the relative lengths of all external and internal branches of the tree (*SI Appendix*, Fig. S20), which represent the amount of drift (in units of generations per effective population size) in heteroplasmy frequency during various stages of development.

To analyze heteroplasmy frequency data with OPL, we used MOPE (14), which performs Bayesian inference of genetic drift and mutation rate parameters for different developmental processes related by an ontogenetic phylogeny. The method employs the phylogenetic pruning algorithm (48) to calculate likelihoods. To calculate posterior probabilities of a heteroplasmy being de novo in a family, we disallowed ancestral segregation in the allele frequency transition distributions underlying the likelihood calculations. We used the posterior samples of genetic drift parameters to estimate a transition matrix of germline allele frequency changes between generations and quantified the dynamics of heteroplasmy persistence across generations. We use posterior medians as point estimates throughout. The details of BLS, OPL, de novo mutation probabilities, mitochondrial heteroplasmy persistence across generations, and conversion of genetic drift to bottleneck units are in *SI Appendix*, *SI Materials and Methods*.

ACKNOWLEDGMENTS. We are grateful to Jessica Beiler and clinical nurses from PSU College of Medicine Pediatric Clinical Research Office for sample collection, Bonnie Higgins for help with experiments, and Boris Rebollo-Jaramillo for reading the paper. This work was funded by NIH grant R01GM116044. Additional funding was provided by the Office of Science Engagement, Eberly College of Sciences, The Huck Institute of Life Sciences and the Institute for CyberScience at Penn State, as well as, in part, under grants from the Pennsylvania Department of Health using Tobacco Settlement and CURE Funds. The department specifically disclaims responsibility for any analyses, interpretations, or conclusions.

1. S. Luo *et al.*, Biparental inheritance of mitochondrial DNA in humans. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 13039–13044 (2018).
2. S. Lutz-Bonengel, W. Parson, No further evidence for paternal leakage of mitochondrial DNA in humans yet. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 1821–1822 (2019).
3. B. Pakendorf, M. Stoneking, Mitochondrial DNA and human evolution. *Annu. Rev. Genomics Hum. Genet.* **6**, 165–183 (2005).
4. S. R. Kennedy, J. J. Salk, M. W. Schmitt, L. A. Loeb, Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLoS Genet.* **9**, e1003794 (2013).
5. B. Rebollo-Jaramillo *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).
6. H. R. Elliott, D. C. Samuels, J. A. Eden, C. L. Relton, P. F. Chinnery, Pathogenic mitochondrial DNA mutations are common in the general population. *Am. J. Hum. Genet.* **83**, 254–260 (2008).
7. Y. S. Ng, D. M. Turnbull, Mitochondrial disease: Genetics and management. *J. Neurol.* **263**, 179–191 (2016).
8. J. B. Stewart, P. F. Chinnery, The dynamics of mitochondrial DNA heteroplasmy: Implications for human health and disease. *Nat. Rev. Genet.* **16**, 530–542 (2015).
9. R. W. Taylor, D. M. Turnbull, Mitochondrial DNA mutations in human disease. *Nat. Rev. Genet.* **6**, 389–402 (2005).
10. H. Zhang, S. P. Burr, P. F. Chinnery, The mitochondrial DNA genetic bottleneck: Inheritance and beyond. *Essays Biochem.* **62**, 225–234 (2018).
11. M. Li *et al.*, Genome of Netherlands Consortium, Transmission of human mtDNA heteroplasmy in the genome of the Netherlands families: Support for a variable-size bottleneck. *Genome Res.* **26**, 417–426 (2016).
12. V. I. Floros *et al.*, Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nat. Cell Biol.* **20**, 144–151 (2018).
13. J. Poulton *et al.*, Transmission of mitochondrial DNA diseases and ways to prevent them. *PLoS Genet.* **6**, e1001066 (2010).
14. P. R. Wilton, A. Zaidi, K. Makova, R. Nielsen, A population phylogenetic view of mitochondrial heteroplasmy. *Genetics* **208**, 1261–1274 (2018).
15. D. M. Rand, Population genetics of the cytoplasm and the units of selection on mitochondrial DNA in *Drosophila melanogaster*. *Genetica* **139**, 685–697 (2011).
16. D. T. Brown, D. C. Samuels, E. M. Michael, D. M. Turnbull, P. F. Chinnery, Random genetic drift determines the level of mutant mtDNA in human primary oocytes. *Am. J. Hum. Genet.* **68**, 533–536 (2001).
17. I. J. Wilson *et al.*, Mitochondrial DNA sequence characteristics modulate the size of the genetic bottleneck. *Hum. Mol. Genet.* **25**, 1031–1041 (2016).
18. J. N. Spelbrink *et al.*, In vivo functional analysis of the human mitochondrial DNA polymerase POLG expressed in cultured human cells. *J. Biol. Chem.* **275**, 24818–24828 (2000).
19. S. Matsuda *et al.*, Accurate estimation of 5-methylcytosine in mammalian mitochondrial DNA. *Sci. Rep.* **8**, 5801 (2018).
20. M. Stoneking, Hypervariable sites in the mtDNA control region are mutational hotspots. *Am. J. Hum. Genet.* **67**, 1029–1032 (2000).
21. M. D. Shriver *et al.*, The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum. Genomics* **1**, 274–286 (2004).
22. X. Yi *et al.*, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
23. X. Cheng, C. Xu, M. DeGiorgio, Fast and robust detection of ancestral selective sweeps. *Mol. Ecol.* **26**, 6871–6891 (2017).

24. C. D. Millar *et al.*, Mutation and evolutionary rates in Adélie penguins from the antarctic. *PLoS Genet.* **4**, e1000209 (2008).
25. M. D. Hendy, M. D. Woodhams, A. Dodd, Modelling mitochondrial site polymorphisms to infer the number of segregating units and mutation rate. *Biol. Lett.* **5**, 397–400 (2009).
26. R. Gilkerson *et al.*, The mitochondrial nucleoid: Integrating mitochondrial DNA into cellular homeostasis. *Cold Spring Harb. Perspect. Biol.* **5**, a011080 (2013).
27. L. Cao *et al.*, The mitochondrial bottleneck occurs without reduction of mtDNA content in female mouse germ cells. *Nat. Genet.* **39**, 386–390 (2007).
28. L. M. Cree *et al.*, A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. *Nat. Genet.* **40**, 249–254 (2008).
29. T. Wai, D. Teoli, E. A. Shoubridge, The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. *Nat. Genet.* **40**, 1484–1488 (2008).
30. L. Cao *et al.*, New evidence confirms that the mitochondrial bottleneck is generated without reduction of mitochondrial DNA content in early primordial germ cells of mice. *PLoS Genet.* **5**, e1000756 (2009).
31. J. P. Jenuth, A. C. Peterson, K. Fu, E. A. Shoubridge, Random genetic drift in the female germline explains the rapid segregation of mammalian mitochondrial DNA. *Nat. Genet.* **14**, 146–151 (1996).
32. I. G. Johnston *et al.*, Stochastic modelling, Bayesian inference, and new in vivo measurements elucidate the debated mtDNA bottleneck mechanism. *eLife* **4**, e07464 (2015).
33. F. Vogel, A. G. Votulsky, *Human Genetics: Problems and Approaches* (Springer-Verlag, 1997).
34. Genome of the Netherlands Consortium, Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
35. A. Kong *et al.*, Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
36. W. S. W. Wong *et al.*, New observations on maternal age effect on germline de novo mutations. *Nat. Commun.* **7**, 10486 (2016).
37. P. D. Olivo, M. J. Van de Walle, P. J. Laipis, W. W. Hauswirth, Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA D-loop. *Nature* **306**, 400–402 (1983).
38. M. V. Ashley, P. J. Laipis, W. W. Hauswirth, Rapid segregation of heteroplasmic bovine mitochondria. *Nucleic Acids Res.* **17**, 7325–7331 (1989).
39. M. Li, R. Schröder, S. Ni, B. Madea, M. Stoneking, 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).
40. R. Szklarczyk, M. Nooteboom, H. D. Osiewicz, Control of mitochondrial integrity in ageing and disease. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130439 (2014).
41. W. Wei *et al.*; NIH BioResource–Rare Diseases; 100,000 Genomes Project–Rare Diseases Pilot, Germline selection shapes human mitochondrial DNA diversity. *Science* **364**, eaau6520 (2019).
42. D. M. Rand, The units of selection on mitochondrial DNA. *Annu. Rev. Ecol. Syst.* **32**, 415–448 (2001).
43. B. Dickens *et al.*, Controlling for contamination in re-sequencing studies with a reproducible web-based phylogenetic approach. *Biotechniques* **56**, 134–141 (2014).
44. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
45. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
46. L. L. Cavalli-Sforza, L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The History and Geography of Human Genes* (Princeton University Press, 1994).
47. G. Bhatia, N. Patterson, S. Sankararaman, A. L. Price, Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
48. J. Felsenstein, Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Biol.* **22**, 240–249 (1973).