RESEARCH ARTICLE

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Fine Time Scaling of Purifying Selection on Human Nonsynonymous mtDNA Mutations Based on the Worldwide Population Tree and Mother–Child Pairs

Bruno Cavadas,[1,2] Pedro Soares,[2,3] Rui Camacho,[4,5] Andreia Brandão,[1,2,6] Marta D. Costa,[2] Verónica Fernandes,[1,2] Joana B. Pereira,[1,2] Teresa Rito,[2] David C. Samuels,[7] and Luisa Pereira[1,2,8]*

[1] *Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto 4200-135, Portugal;* [2] *Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto 4200-465, Portugal;* [3] *Department of Biology, CBMA (Centre of Molecular and Environmental Biology), University of Minho, Braga 4704-553, Portugal;* [4] *INESC TEC, Porto 4200-465, Portugal;* [5] *Departamento de Engenharia Informática, Faculdade de Engenharia da Universidade do Porto, Porto 4200-465, Portugal;* [6] *Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto (ICBAS), Porto 4050-313, Portugal;* [7] *Vanderbilt Genetics Institute, Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, Tennessee 37232-0700;* [8] *Faculdade de Medicina da Universidade do Porto, Porto 4200-319, Portugal*

Communicated by Peter Oefner

Received 8 May 2015; accepted revised manuscript 20 July 2015.
Published online 7 August 2015 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22849

ABSTRACT: A high-resolution mtDNA phylogenetic tree allowed us to look backward in time to investigate purifying selection. Purifying selection was very strong in the last 2,500 years, continuously eliminating pathogenic mutations back until the end of the Younger Dryas (~11,000 years ago), when a large population expansion likely relaxed selection pressure. This was preceded by a phase of stable selection until another relaxation occurred in the out-of-Africa migration. Demography and selection are closely related: expansions led to relaxation of selection and higher pathogenicity mutations significantly decreased the growth of descendants. The only detectible positive selection was the recurrence of highly pathogenic nonsynonymous mutations (m.3394T>C-m.3397A>G-m.3398T>C) at interior branches of the tree, preventing the formation of a dinucleotide STR (TATATA) in the *MT-ND1* gene. At the most recent time scale in 124 mother–children transmissions, purifying selection was detectable through the loss of mtDNA variants with high predicted pathogenicity. A few haplogroup-defining sites were also heteroplasmic, agreeing with a significant propensity in 349 positions in the phylogenetic tree to revert back to the ancestral variant. This nonrandom mutation property explains the observation of heteroplasmic mutations at some haplogroup-defining sites in sequencing datasets, which may not indicate poor quality as has been claimed.

Hum Mutat 36:1100–1111, 2015. © 2015 Wiley Periodicals, Inc.

KEY WORDS: mtDNA; worldwide phylogenetic tree; mother–child pairs; massive parallel sequencing; heteroplasmy

Introduction

Current large-scale genomic surveys on the human population are revealing an astounding catalog of rare mutations, which were generated by the recent demographic expansion [Fu et al., 2013]. Many of these rare polymorphisms are potentially pathogenic, as demonstrated in the analysis of the 1000 Genomes dataset [Abecasis et al., 2012]: for SNPs with frequencies lower than 0.5%, each individual carries 130–400 nonsynonymous, 10–20 loss-of-function, two to five damaging, and one to two SNPs identified previously in cancer screenings. Efforts have been applied to establish the age of these mutations, as this information would allow us to understand the evolutionary history of the species [Fu et al., 2013], and, in particular, would provide insights into the selection forces acting upon genetic diversity [Pereira et al., 2011; Soares et al., 2013]. Dating mutational events in the nuclear DNA (nDNA) is challenging due to the fast decrease of linkage disequilibrium through recombination, adding uncertainty to the reconstruction of the evolutionary history of any particular region. This problem has led in the recent past to very different age estimates for pathogenic mutations, for instance, the hemochromatosis associated p.Cys282Tyr mutation in HFE gene was dated as having occurred 59 [27–161] [Ajioka et al., 1997] or 138 [88–156] [Toomajian et al., 2003] generations ago. The current availability of exomes and complete genomes is enabling further attempts to establish chronologies of a large number of nDNA polymorphisms [Fu et al., 2013]. These authors dated 1,146,401 protein-coding SNPs with known ancestral state by using a simulation approach to generate coalescent trees under a given demographic model, including under neutral, selection, constant, exponential, and migration conditions. They found that the average age of conservatively defined deleterious mutations (14.4% of the total) was 5,200 ± 300 years for European Americans (EA) and 10,100 ± 600 years for African Americans (AA), although the

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Luisa Pereira, IPATIMUP, R. Dr. Roberto Frias s/n, Porto 4200-465, Portugal. E-mail: lpereira@ipatimup.pt

Contract grant sponsors: FCT, the Portuguese Foundation for Science and Technology (PTDC/IVC-ANT/4917/2012, SFRH/BD/78990/2011, and IF/01641/2013); FEDER, COMPETE, and FCT (FCOMP-01-0124-FEDER-029291, PEst-C/SAU/LA0003/2013, and PEst-OE/BIA/UI4050/2014); Luso American Foundation (FLAD).

vast majority of these mutations (86.4%) appeared in the past 5,000 years (91.2% and 77.0% for EA and AA, respectively). The authors also described that three and 18 genes in EA and AA, respectively, have a significant excess of deleterious variants that arose within the last 5,000 years. This result is in apparent contradiction to their main conclusion that EA had an excess of deleterious variants in essential and Mendelian disease genes, due to the bottleneck associated with the out-of-Africa dispersal leading to less efficient purging of weakly deleterious alleles. More recently, Fu et al. (2014) call attention to the fact that population bottlenecks and expansions have opposing effects on patterns of variation in populations and individuals: bottlenecks lead to skews toward common variation, and expansions result in skews toward rare variation; but individuals from populations that have experienced bottlenecks tend to carry more deleterious alleles, whereas the ones from expanding populations carry slightly fewer derived deleterious alleles. By focusing on nDNA genes that code mitochondrial proteins, Pereira et al. (2014) observed that the proportion of individuals having at least one potential pathogenic mutation was significantly lower in Europeans than in Africans and Asians, and hypothesized that this difference may reflect recent demographic asymmetries in population expansions (the main postglacial expansion in Europeans preceded the main Neolithic expansions in Africans and Asians). The authors comment that focusing the demographic models in the out-of-Africa bottleneck event, which took place probably 60,000 years ago [Soares et al., 2012], is inappropriate to model the fine-scale demography across continents that impacted the emergence of rare mutations, since it is likely that these rare mutations formed much more recently.

In relation to the influence of selection upon rare mutations, Schaibley et al., (2013) argue that variants with frequencies below 0.1% and estimated ages of 250 years (or less, depending on the rate of population growth considered in the model) are typically less affected by natural selection. We confirmed [Pereira et al., 2014] the absence of selection in SNPs with frequency lower than 0.1% (for 104 nDNA genes coding mitochondrial proteins studied in 1,092 individuals from the 1000 Genomes dataset), but the class with frequency 0.1%–1% already displayed signs of selection. This renders very rare mutations (minimum allele frequency; MAF<0.1%) the most appropriate resource for studying the spectrum and genomic distribution of mutations in the absence of selection, at the population level. Two promising datasets to ascertain mutation under neutrality, without the problems of the population datasets, are: (1) the somatic mutation pool identified in cancer surveys [Pereira et al., 2012; Ju et al., 2014], and (2) the set of mutations transmitted to the offspring in analysis of mother–father–offspring trios. Both these datasets represent recent mutations, generated in at most a few cell divisions.

In this work, we investigated the dynamics of very rare (including pathogenic) nonsynonymous mutations, gaining insights into the time frame for selection action. We aimed to do this in an empirical dataset, avoiding the confusing influence of simplistic and inappropriate demographic models used so far to date mutations through simulation approaches. The mitochondrial genome (mtDNA), maternally inherited as a nonrecombining block, is the best genetic system to reconstruct the phylogeny of the human species and to date the most recent common ancestors. Being approximately only 16.6 kb, mtDNA has been intensively sequenced across the globe, so that the population structure for the diversity of this DNA molecule is well known [Torroni et al., 2006; Pereira et al., 2009; van Oven and Kayser, 2009]. Despite recent developments in dating techniques for nDNA diversity, recombination erases the possibility of dating events of admixture older than 4,000 years old [Hellenthal et al., 2014], and mtDNA remains the most informative genetic system to infer past migrations and expansions and estimate their fine-scale time frames [Fernandes et al., 2015]. Of course, even for mtDNA, different mutation rates lead to diverse date estimates, but confidence intervals are highly overlapping (see discussion in [Soares et al., 2009]). The mtDNA molecule is present in hundreds of thousands of copies per cell, and its high replication rate leads to the emergence of mutations and the generation of two or more populations of mtDNA within a mitochondrion, cell, tissue, or individual, a phenomenon known as heteroplasmy. Heteroplasmy is found in both germinal and somatic mutations, and these mutations will then become fixed or lost through genetic drift, bottlenecks, and/or selection—population genetics theory can be applied to the mtDNA diversity within an individual and not only at the population level [Chinnery and Samuels, 1999; Elson et al., 2001; Wonnapinij et al., 2008]. At the somatic level, it was shown that mtDNA nonsynonymous mutations accumulate according to the neutral model expectations in cancer tissues (for which we have more information about somatic mutations); and following expectations of the neutral theory, they tend to homoplasmy (fixation within the individual) over time [Pereira et al., 2012; Ju et al., 2014]. In the female germinal line, according to the mitochondrial genetic bottleneck hypothesis, the mtDNA passes through a strong bottleneck in oogenesis, which reduces the number of copies of mtDNA in each reproductive cell to a relatively small number (~200 copies); the subset of the mtDNA molecules that is transmitted through the bottleneck then becomes the founder population of the offspring's mtDNA. Thus, heteroplasmic dynamics can be a reliable indicator about the recentness of a mutation and its pathogenicity effect. Recently, by using massively parallel sequencing (MPS) data, Ye et al. (2014a) showed that around 90% of healthy individuals carry at least one heteroplasmy when using a 1% MAF threshold, and that the heteroplasmies tend to show high pathogenicity. But these results were questioned [Just et al., 2014] by the observation that among the 15 samples with 20 or more heteroplasmies, all appeared to be a mixture of at least two distinct individuals, and a minimum of 81% of the 584 heteroplasmies occurred at positions that are known to be population variants (haplogroup defining). These authors checked that haplogroup-defining heteroplasmies remained until the minimum heteroplasmy cut-off was raised to 15%–20%, concluding that it remains unclear whether MPS can reliably detect mtDNA heteroplasmy present in frequency less than 5%–10%, even when coverage depths are very high (see reply by [Ye et al., 2014b]). In order to provide additional information on the level of secure cut-off for heteroplasmy by using MPS, Gardner et al. (2015) screened a known stereotyped mutational motif and statistically validated a heteroplasmy threshold of 0.22% when a minimum of 1,500 coverage was performed.

In this study, we combine information from next-generation sequencing of mother–offspring pairs and from a large-scale global human phylogenetic tree. By considering both types of data together, we can investigate selection on mtDNA variation on a wide range of time scales, from a single generation to a few thousand years and ultimately to ~200,000 years. Our comparison shows a surprising unity between the selection at the single generation level and at the broad range of time scales represented by the full phylogenetic tree. First, we applied a robust pipeline for the detection of heteroplasmies in MPS data from the 1000 Genomes database in mother–child pairs (fathers were also considered, to check the quality of sequencing), generated in the Complete Genomics platform. We further characterized the pathogenicity of the nonsynonymous heteroplasmic substitutions using the MutPred algorithm (quantitative pathogenicity score ranging from 0 to 1, with a higher value indicating a greater likelihood of being pathogenic; [Li et al., 2009]),

and analyzed the distributions among de novo, lost, and shared nonsynonymous mutations. Thus, we were able to investigate the dynamics of the selection force acting in one generation time. We then focused on a worldwide fine-resolved mtDNA phylogeny (over 18,000 complete sequences), which allowed us to define precisely (with sharp confidence intervals) the time scale for mutations, and to infer how long selection takes to remove deleterious mutations from the population, and how population expansions and bottlenecks affect that selection.

Material and Methods

mtDNA From Trios and Heteroplasmy Validation

Sequences of 117 trios (mother, father, and child) and seven mother–child pairs from five populations (21 African Yoruba from Nigeria; 32 Western European descendants residing in Utah; 15 South Asian Punjabi from Lahore, Pakistan; 30 East Asian Southern Han Chinese; and 26 American Peruvians from Lima, Peru) were downloaded from the Sequence Read Archive (SRA), based on the familial information provided in the 1000 Genomes Website. mtDNAs mapped to the revised Cambridge reference sequence (rCRS; [Andrews et al., 1999]) were extracted using sam-dump provided by the SRA Toolkit 2.3.4. The filtering process was carried out in three stages: (1) only reads that mapped exclusively to the mtDNA reference were recorded (exclusion of Numts); (2) duplicated reads were removed by Picard; (3) only sequences with a Phred quality score $\geq$20 were selected. Step (2) did not influence the final results.

For a candidate site to be classified as heteroplasmic, a set of quality control requirements was taken into account (adapted from Ye et al. [2014a]): (1) the frequency of the minor allele should be equal or higher to 1% in both strands; (2) have more than 200 reads on each strand; (3) the log likelihood ratio [Picardi and Pesole, 2012], which measures the confidence of the heteroplasmy, must be equal to or higher than five. Heteroplasmic validation was further refined using the father's genome as a control—any position sharing low levels of heteroplasmy in all three individuals was considered as an artifact that slipped through the quality control steps, and was therefore removed. Highly recurrent mutations that appear in poly-C regions, such as at positions 302, 310, 16,182, and 16,183, were also removed. The position 16,189 in the middle of a poly-C region was not removed as it is haplogroup defining in different parts of the mtDNA phylogeny. The polymorphisms at positions 4,769 and 6,502 were also removed because they were widely present as low-level heteroplasmy in almost all samples, and we considered them as artifacts; 4,769 is H2-haplogroup defining, and since sequences are compared versus rCRS, this could indicate some issue with this fact; however, the same is not true for 6,502, which must be a rare mutation, being absent in our worldwide population dataset. Haplogroup assignment was performed with HaploGrep [Kloss-Brandstatter et al., 2011], and heteroplasmic mutations were classified as occurring at haplogroup defining or private positions. Apparent heteroplasmies falling on known population variant sites could possibly be due to sample contamination, and are thus more suspect.

The percentage of heteroplasmy (Supp. Table S1) was reported as a function of the minor allele frequency in mothers. The heteroplasmic events were organized in three sets: de novo mutations, absent in mother and observed in child; lost mutations, present in mother and absent in child; and shared mutations, present in both. The difference in percentage of heteroplasmy between mother and child was calculated for each mutation. Notice that for the lost heteroplasmies the heteroplasmy difference value refers to the real loss of the allele. If it was the major allele of the mother that was lost, the percentage of heteroplasmy represents that loss.

To assess the reliability of our pipeline, we compared 55 individuals sequenced by the two methods (Complete Genomics and Illumina from 1000 genomes) with the same pipeline presented by Ye et al. (2014a). The pipeline consisted as follows: reads extracted from the 1000 genomes project were remapped to the mitochondrial genome using GSNAP, in order to minimize the presence of Numts (copies of the mtDNA found in the nuclear genome). Ye et al. (2014a) analyzed 1,085 individuals, and performed a final QC step of checking in the global dataset that positions had less than 10x coverage (in both strands) in 95% of the individuals; these positions were removed from the paper. As we did not analyze all the individuals, we did not perform this step and kept all the positions.

The Human Worldwide mtDNA Tree

A total of 18,471 complete mtDNA sequences were used to reconstruct a reliable human worldwide mtDNA phylogenetic tree. Most of these sequences are published and deposited in GenBank, or come from the 1000 Genomes Project, and a few are our own and published elsewhere (Supp. Table S2). The additional new mtDNA sequences helped to resolve parts of the tree for which published data were still poor.

We constructed the mtDNA phylogenetic tree by using the reduced median algorithm [Bandelt et al., 1995], resolving reticulations by hand on the basis of the relative frequency of the mutations involved [Soares et al., 2009]. We used the GeneSyn [Pereira et al., 2009] software to convert files, and to identify and classify polymorphisms in comparison with the rCRS [Andrews et al., 1999], and checked haplogroup assignment with HaploGrep [Kloss-Brandstatter et al., 2011]. To a very high degree, our tree agrees with the one reported in the PhyloTree Website [van Oven and Kayser, 2009], and we refer readers to that Website (build 16) in order to obtain a visual perception of the phylogeny, as it would be impossible to present it here in a compatible publication format. For each node on the tree, we calculated the rho value [Forster et al., 1996], that is, the average number of sites differing between a set of sequences and a specified common ancestor; we estimated standard errors as in Saillard et al. (2000). rho values are a systematic way of representing the depth of a node in the tree, and reflect the age of the node. To convert rho values to ages, we used the mutation rate estimate for the complete mtDNA sequence of one substitution in every 3,624 years, which is further corrected for purifying selection [Soares et al., 2009].

A XML file of the tree, reproducing the hierarchical organization of the nodes and identifying mutations along branches, was obtained, which allowed easy automation to estimate variables in each node: rho, age, number of branches, number and types of polymorphisms in above and lower branches, MutPred predicted pathogenicity values, and so on.

The ages of mutations defining a branch can only be determined to lie between the ages of the nodes immediately above and below that branch. In some of the analyses, for which high certainty on the time intervals was mandatory, only the branches for which the lower and upper nodes were within a certain time frame, for instance, in the rho interval between 1 and 2, were included. This required precision resulted in the exclusion of a high quantity of polymorphisms for those particular analyses, an unavoidable restriction.

Statistical Evaluation

The pathogenic scores for all possible nonsynonymous substitutions predicted with the MutPred algorithm [Li et al., 2009] version 1.2 were used. In previous publications [Pereira et al., 2012; Pereira et al., 2014], we compared the MutPred prediction scores with the ones obtained by using PolyPhen and SIFT, and concluded there was a good correlation between these algorithms, especially between MutPred and SIFT. For this reason and due to the fact that MutPred is quantitatively more discriminating than the other two methods allowing a more reliable statistical evaluation, we decided to perform all tests in this work based only in MutPred.

The simple choice of the set of all possible nonsynonymous variants produced by single-nucleotide changes as our null hypothesis (referred to as the all possible set) might be criticized due to the well-known bias in de novo mutations to transitions over transversions [Kennedy et al., 2013; Williams et al., 2013]. To test whether our analysis is sensitive to this bias, we also constructed the distribution of predicted pathogenicity for all possible variants weighted by a 40 to 1 transition to transversion ratio. The pathogenicity distribution in the biased set is not greatly different than that in the simpler all-possible-variants set (Supp. Fig. S1).

Comparisons of mean MutPred pathogenicity values were conducted by two-tailed *t*-tests assuming unequal variances, performed in Origin 7 software (http://www.originlab.com).

Multivariate general linear regression analyses for the variables rho, the number of branches extending from a node, and the summed MutPred pathogenicity scores for variants defining branches immediately below and above the node were performed in R. The variables rho and number of branches were natural log transformed and normalized according to the formula $Y = \log10(Y + 1 - \min[Y])$, as rho values of zero (tree tips) were included. The offset of 1 within the log is needed to avoid infinite values. Three models were tested individually:

Model 1a: (log number of branches out of node) = A + B1 (average MutPred score in branch above node) + B2 (log rho)
Model 1b: (log number of branches out of node) = A + B1 (highest MutPred score in branch above node) + B2 (log rho)
Model 2: (average MutPred in branches directly below node) = A + B1 (log number of branches directly below node) + B2 (log rho).

Analysis of variance (ANOVA) *F*-tests were used to compare the distributions of predicted pathogenic scores between two different time intervals (for all pairwise comparisons). The calculations were performed in R.

To evaluate the relative mutation rate of back mutations, we selected mutations that appeared in the worldwide mtDNA tree at least 3,800 years ago, in the time frame of one mutation every 3,624 years [Soares et al., 2009], thus allowing for the occurrence of back mutation to be observed in the population. A final amount of 3,664 mutations fulfilling this age criterion was obtained. For each of these mutations, we calculated its relative mutation rate in the tree and compared it with the relative mutation rate of its reversion (according to Supp. Fig. S2), applying a 2×2 contingency Fisher exact test to evaluate biases in back mutation versus mutation rate.

For each of the 349 mutations showing a significant higher relative back mutation rate, we performed a linear regression model to evaluate its randomness in the mtDNA molecule.

The secondary structures for the HVS-I region (between positions 16,024 and 16,383) for rCRS and the sequence containing the transition at position 16,217 (Supp. Fig. S3) were inferred using the mfold online tool (http://mfold.rna.albany.edu/?q=mfold/DNA-Folding-Form).

Results

Quality Checking of the Heteroplasmy in Trios

We used the data from the Complete Genomics platform, which provides a high coverage (Supp. Table S3), about 8,862 reads per position per individual across populations. The Peru samples had a lower number of reads than the other populations (average 810). This allowed us to establish a strong quality threshold of 200 reads on each strand.

To check the quality of our methodology, we compared the heteroplasmy inference on a total of 55 (mothers) in common with Ye et al. (2014a). These authors used Illumina sequences instead of the sequences from the Complete Genomics platform, and the former contain a lower (average of 2,000 reads per position per individual across populations) number of reads than the later. Supp. Table S4 reports the comparisons we made for these individuals, based on the following tests: (a) Ye et al. (2014a) data on Illumina sequences; (b) our results from the Complete Genome platform; and (c) our analysis based on the Illumina sequences and a similar pipeline to Ye et al. (2014a). Although there is variability between the three tests of heteroplasmy calling, the MutPred pathogenicity score distributions are similar: $n = 89$ and mean = 0.618 for (a); $n = 77$ and mean = 0.612 for (b); and $n = 110$ and mean = 0.631 for (c). It was expected that our test (c) would allow more heteroplasmies than (a) as we did not remove some positions as explained in *Material and Methods*. Our dataset based on (b) seems to be more conservative, as expected for being based on a higher number of reads.

Focusing on the dataset from the Complete Genome platform, after applying the quality filters, a total of 635 heteroplasmic events, >1% threshold, were detected throughout the mitochondrial genome in the 124 mother–child pairs (full information provided in Supp. Table S1). Nine pairs did not present any heteroplasmy, these individuals being from diverse haplogroup backgrounds (two B2, and one from each B2b, H1e1a, H2a2a1, J1b, L2a1a, L3b1a7a, and L3b1a8). A total of 115 pairs had at least one heteroplasmic event in at least one of the members (Supp. Table S1), and most of these heteroplasmies were present in very low percentages both in mothers and children (Fig. 1A): 17.4% with percentage of heteroplasmy between 1% and 2%; 23.7% with percentage between 2% and 3%; and 10.1% with percentage between 3% and 4%. The majority of the events (80.5%) presented heteroplasmy lower than 10%. A total of 97 mother–child pairs had less than 10 events in total, 16 pairs had between 10 and 16 heteroplasmic events, and two pairs in L haplogroups were outliers with 26 and 30 heteroplasmic events. Attending to the fact of the heteroplasmic mutations being either on haplogroup-defining sites or private, these two outlier pairs had a high number of haplogroup-defining site heteroplasmic variants: 24 out of 26 and 17 out of 30, respectively. Due to the possibility of contamination in these samples or poor quality (although they passed the editing quality controls), we decided to remove these two pairs from further analyses, leaving a total of 579 heteroplasmic events.

We further confirmed that even pairs sharing a low amount of heteroplasmic mutations had in some cases mostly or totally haplogroup-defining variant sites, such as six out of seven in the pair HG02260–HG02261; four out of five in HG02734–HG02735; all four in HG02301–HG02303, all two in HG02784–HG02785. This occurred in diverse haplogroups, in all the three main subtrees, testifying that it is not an issue of the rCRS working poorly for aligning distant haplogroups. The value of haplogroup defining versus private heteroplasmic mutations decreases as the heteroplasmic threshold increases: 119 versus 460 for threshold >1% (0.26);

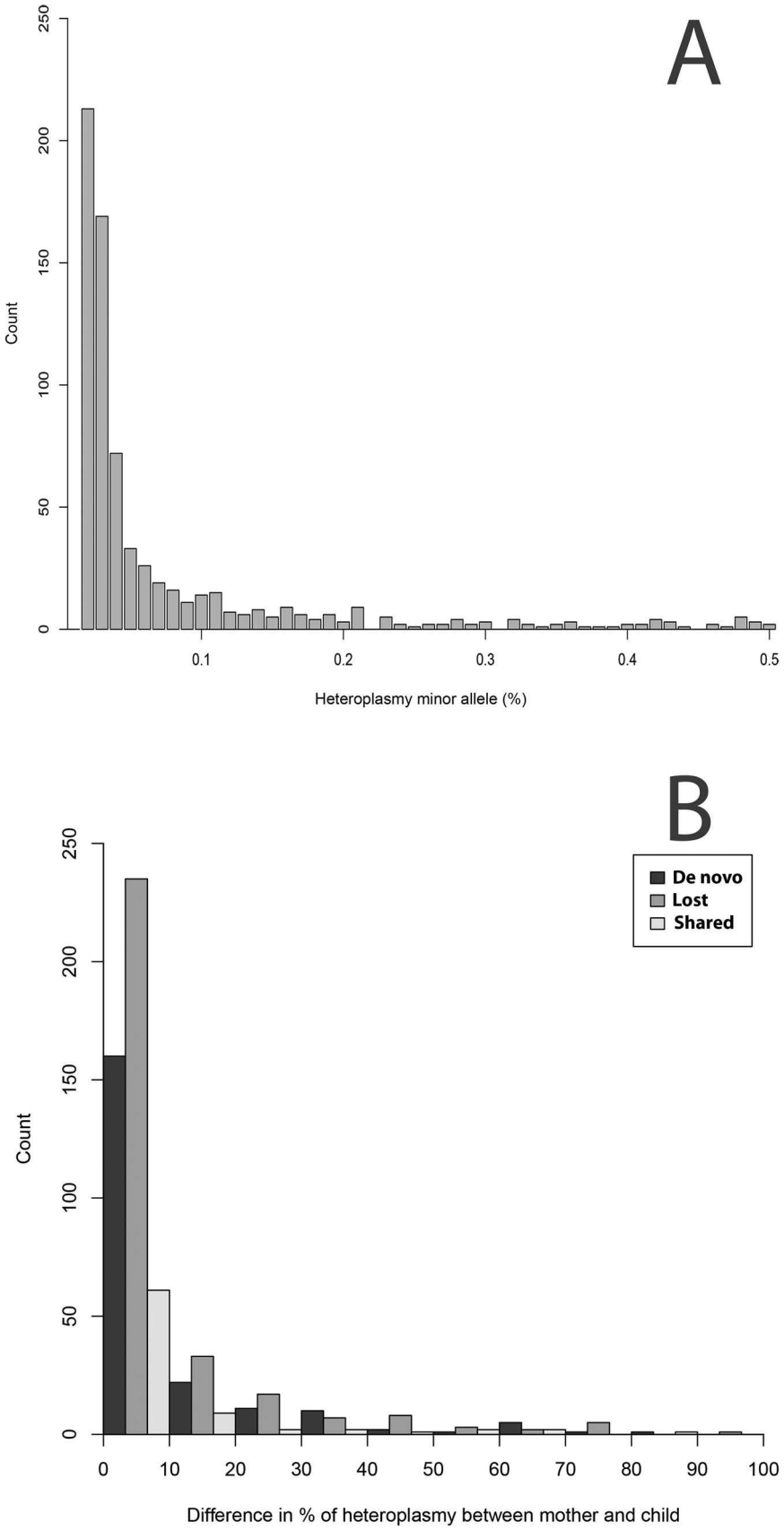

Figure 1. Histograms for the (**A**) number of heteroplasmic sites in mothers and children combined, distributed over the percentage of heteroplasmy of the minor allele, and (**B**) absolute value of the shift in percentage of heteroplasmy between mother and child, split into those heteroplasmic sites present in both mother and child (shared), those present only in the mother (lost), and those present only in the child (de novo).

Table 1. Distribution of the Three Types of Heteroplasmies (De Novo, Lost, and Shared) Throughout the mtDNA in Trio Data for Private Heteroplasmies

All mutations	De novo	Lost	Shared
D-loop	28	34	21
tRNA	14	11	0
rRNA	33	43	4
Syn protein	31	44	7
Non-Syn protein	63	113	5
Intergenic	0	1	8
Total	169	246	45

78 versus 337 for threshold >2% (0.23); 42 versus 242 for threshold >3% (0.17); and 11 versus 112 for threshold >10% (0.10).

More interestingly, a few haplogroup-defining sites were highly heteroplasmic in certain haplogroups, as position 16,217 appearing in six out of 10 B2 and one out of three B4 pairs (B2 is a subhaplogroup of B4), and 14,182 in all three U5b pairs. These two heteroplasmic variants, occurring only in haplogroups where those sites are nonreference alleles, seem to be evidence of some type of pressure to revert the sequence back to the original. This pressure is not necessarily selection, but may be a structural instability that the variant causes in the new DNA sequence leading to back mutation. We did not find signs of structural change induced by the mutation at position 16,217 (Supp. Fig. S3), but the transition at the nearby haplogroup-defining position 16,247 reported inside the Polynesian haplogroup B4a1a1a1 creates a hairpin in the control region [Duggan and Stoneking, 2013] and may be a good example of high back mutation due to structural instability.

Although it is unlikely that all heteroplasmic mutations present in one individual are in haplogroup-defining positions, and this criterion could be used for quality control, it does not mean that all heteroplasmies at haplogroup-defining positions are false heteroplasmies, especially if they are fast evolving sites. Supporting this is the fact that a total of 58 out of 119 (48.7%) of the haplogroup-defining variants observed here as heteroplasmic sites are in the D-loop, against 83 in 460 (18%) for the private ones, since the D-loop has a higher substitution rate than the coding region. Even so, we decided to be conservative, and limited further analyses to the private heteroplasmic mutations observed in the 113 mother–child pairs.

None of the de novo mutations observed in children could be explained by paternal transmission in the 117 trios for which the father's genome was available.

The Germinal Transmission—One Generation

About 53.5% of the total 460 private heteroplasmies detected are lost in the transmission from mother–child, whereas 36.7% appear de novo (in the child only) and only 9.8% are shared between mother and child (Table 1). The greatest contributor for lost mutations is the amount of nonsynonymous mutations, which is nearly double compared with the proportion in the de novo class, although this is not quite statistically different (Fisher's exact test $P = 0.086$). The amount of synonymous mutations is almost equal in both de novo and lost groups (18.3% to 17.8%). It is also interesting to notice that in the shared heteroplasmies, the proportion occurring in the D-loop represents 47% of the mutations, compared with 17% in the de novo class and 14% in the lost class. Most of the mutations (351 in 460) have a difference in percentage of heteroplasmy between mother and child below 10% (Fig. 1B), but a few cases (10 in 460) are observed with a difference of 90% or more.

We further investigated the pathogenicity of the nonsynonymous mutations using the MutPred algorithm. As can be seen in Figure 2, the set of lost mutations is very similar to all possible nonsynonymous single-nucleotide mutations that can occur in the human mtDNA (mean values of 0.640 for all possible mutations and 0.639 for lost mutations). Nevertheless, in the child, random mutation leads to the emergence of de novo mutations, which are slightly less pathogenic than the ones that are lost (mean of 0.588). This dynamic shows that in one generation time, purifying selection removes the extremely high pathogenic mutations, but random mutation creates new ones; the two distributions, de novo and lost mutations, are not statistically different from each other ($P = 0.11$), but the de novo group is statistically different from all possible mutations (de novo vs. all possible P value = 0.05; lost vs. all possible P value = 0.97).

In relation to the shared heteroplasmic sites between mothers and offspring, there were only five nonsynonymous mutations, all with MutPred pathogenicity scores below 0.7 (which we have previously demonstrated as being a reasonable threshold for deleterious mutations; [Pereira et al., 2011]), indicating that these transmitted heteroplasmies are unlikely to be pathogenic.

The Worldwide Human Tree

The human worldwide mtDNA phylogenetic tree, based on 18,471 samples (Supp. Table S2), has 7,563 nodes, with estimated ages ranging from 0 to 193,451 years (rho = 57.61) at the root of the tree. Many nodes (63%) have age lower than 7,800 years (distributed as 1,993 nodes between 0 and 2,500 years, 1,571 for 2,500 and 5,200 years, and 1,193 for 5,200 and 7,800 years), a reasonable amount (29%) between 7,800 and 20,400 years, decreasing fast (8%) toward the out-of-Africa period at around 60,000 years, and remaining 48 (0.6%; African only) nodes until the root.

If the tree is split into the macrohaplogroups L(xM,N), M, and N, which basically represent the African, Asian, and Asian/European populations (some subgroups inside N haplogroups are characteristic of Amerindian populations, and M1 is observed in Africa), there is still a bias in the amount of information for the N subtree. 2,515 samples and 1,124 nodes are in L; 3,923 samples and 1,825 nodes are in M; and 12,033 samples and 4,614 nodes are in N. This is probably an observational bias due to the large number of studies in European populations (contained within the N macrohaplogroup). The main branches of the L and N trees are already well defined, but it is reasonable that more branches derived from the root of the M tree will still be discovered with more sequence data from Asia.

A strong sign of population expansions is the presence of star-like nodes, harbouring many derived branches. The first moderate signal of population expansion occurs in the L3 haplogroup, at around 70,000 years ago, leading to the emergence of five African branches and the non-African N and M macrohaplogroups [Soares et al., 2012]. Then, the M haplogroup had a very large expansion around 50,000 years ago, generating 60 branches identified so far. The expansion at around the same time (55,000 years) for the sister-clade N was more restrictive, leading to the appearance of 16 descendants, but its close descendant R (48,000 years) led to additional 22 branches. Macrohaplogroup N displays the greatest expansions in the more recent haplogroups. The most extreme expansion is in the western European haplogroup H (122 known branches so far; forming at 15,000 years) together with its derived clades (H1 with 127 at 11,000 years, H3 with 90 at 9,800 years, H5 with 33 at 12,000

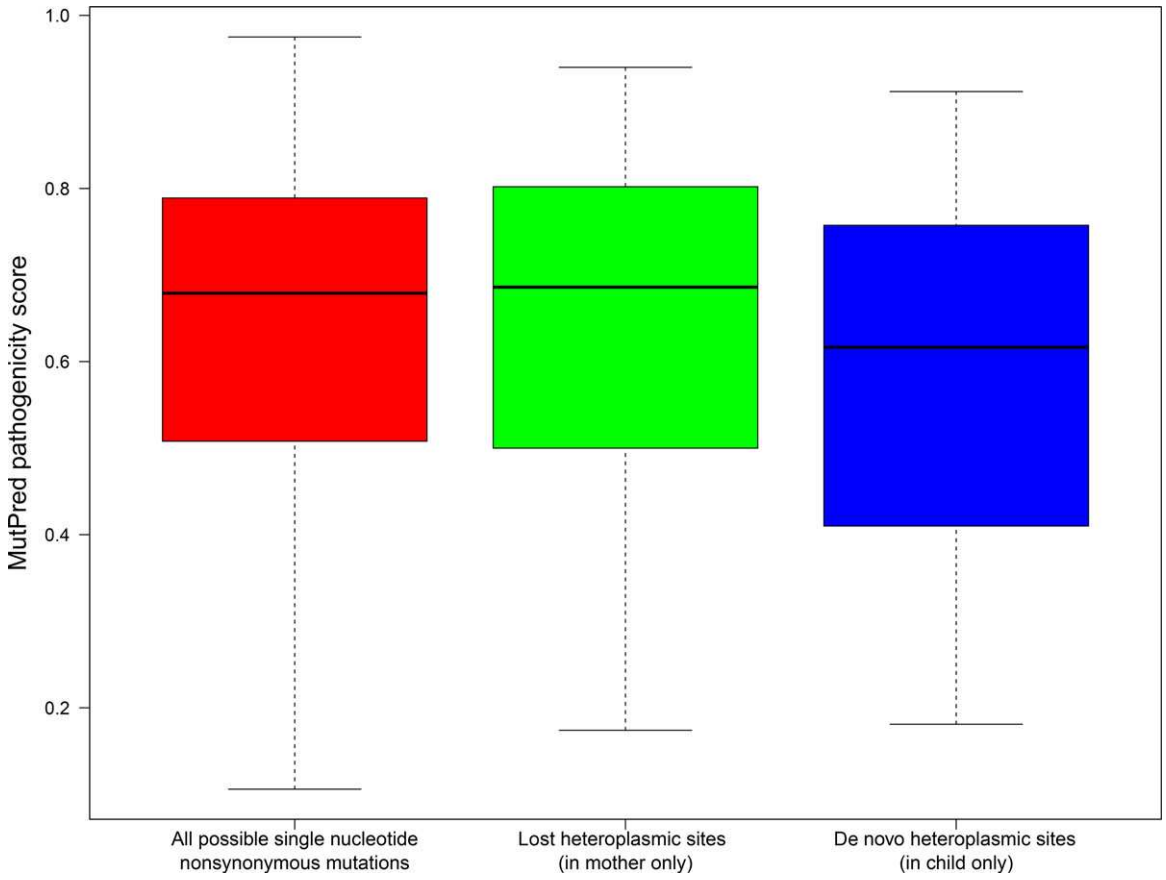

Figure 2. Boxplot of de novo and lost germinal heteroplasmic, as well as all possible, mtDNA amino acid variant pathogenicity scores. The boxes represent the interquartile range and the whiskers are the 5% and 95% quartiles.

years, H5a1 with 47 at 7,000 years, H1c with 42 7,800 years, H1a with 33 at 7,000 years, and many other), which attain a frequency of 40% in most of the European populations—no other case of such a young clade attains such a high frequency over an extended geographic region at the current resolution. The structure of the mtDNA phylogenic tree testifies to the fast population expansion in Europe immediately following the Younger Dryas, whereas African and Asian haplogroups began to display stronger signs of expansion more recently, beginning at 6,000 years ago, as already identified in Bayesian inferences of the population effective size [Atkinson et al., 2008; Soares et al., 2008, 2010, 2012]. This is an important time frame for the potential pathogenic mutations and the pressure of purifying selection acting upon them.

Fine Time Scaling of Purifying Selection

A total of 8,996 nonsynonymous mutation events occurred in the human mtDNA tree, corresponding to 2,012 unique mutations (Fig. 3), meaning that many mutations are observed more than once in the tree. This can be identified in Figure 3 as horizontal lines of points having the same MutPred pathogenicity value (e.g., at $Y = 0.815$, corresponding to mutation m.9438G>A, observed 38 times; the line at $Y = 0.223$ corresponds to four recurrent mutations m.14766C>T, m.9055G>A, m.9966G>A, and m.13889G>A observed 29, 35, 32, and 11 times, respectively). A total of 1,115 (12.4%) of these mutations have MutPred pathogenicity scores equal or higher than 0.7. Of these potential deleterious mutations, the vast majority, 910 (81.6%), occurred in the tips of the tree representing very recent mutations; 99 (8.9%) are located in branches with minimum age (age of the lower node) dating less than 5,000 years; 45 (4.0%) between 5,000 and 10,000 years; 45 (4.0%) between 10,000 and 20,000 years; 15 (1.4%) occurring more than 20,000 years and less than 41,000 years (and not more than 61,000 years if we are conservative and attend to the maximum date); and only one mutation (m.13789T>C; MutPred = 0.731) in one of the L1 branches that emerged as old as 125,000–165,000 years. Again, this confirms the effect of purifying selection acting strongly in the human mtDNA tree, and indicates that 90.5% of the potentially pathogenic mutations are lost within 5,000 years.

We wanted to establish more precise time frames for the loss of these pathogenic mutations (especially at young ages), so we restricted the analyses to mutations located in short branches, which allows a reliable estimation of the age of the mutation's emergence. Although we lose many mutations in doing this, we eliminate uncertainty due to poor resolution of some parts of the tree. Purifying selection is very strong in a time span of a few generation (Fig. 4), reducing the mean pathogenic value from 0.68 (the value of all possible nonsynonymous mutations) to 0.49 in 2,500 years. Then, the decrease is slower until 5,200 years (average MutPred score of 0.46), but a big change occurs in the transition to the period 5,200–7,800 years (average MutPred score of 0.41). The value remains stable until 10,600 years (average MutPred score 0.40), but at this time, the tendency is inverted, with an increase in the average pathogenicity value

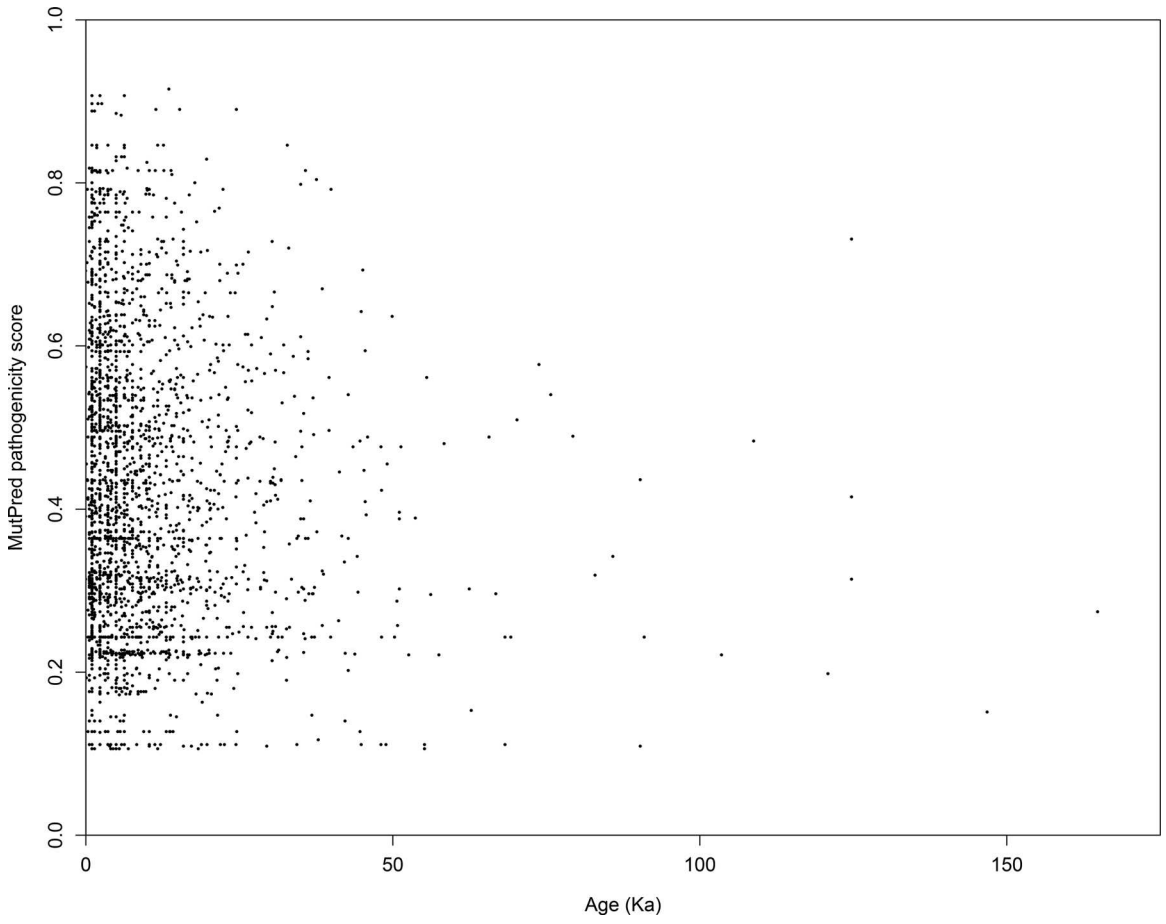

Figure 3. Distribution of MutPred-predicted pathogenicity values of nonsynonymous mutations by age (per 1,000 years, Ka) inferred from the lower node. Each dot represents one nonsynonymous mtDNA mutation in the human phylogenetic tree. Horizontal streaks of dots represent the same nonsynonymous mutation formed independently at many different times in the tree.

(0.44) in the interval 10,600–13,300 years, which matches the first period of intense human population growth following the Younger Dryas. From this time toward the older parts of the tree, the purifying selection acted as expected, with most pathogenic mutations removed from the population and at the LGM the average is at the level (0.38) of the lowest values. There is a second period of a slight inversion leading to an increase of the mean MutPred pathogenicity value, for the period 27,700–59,000 years (a large interval due to the lower number of branches in this time frame of the tree), which matches the out-of-Africa migration believed to have occurred after the first period of human population explosion. The average MutPred value for the period between 59,000 years and the root of the tree is the lowest one (0.34). The pairwise comparisons are statistically different (Supp. Table S5) for all comparisons involving the all possible group, and for most of the comparisons between the most recent group (0–2,500 years) and earlier groups (except with the classes 2,500–5,000 years and 10,600–13,300 years).

Although the total amount of mutations with MutPred pathogenicity score equal or higher than 0.7 decreases dramatically from the tips to the interior of the tree, the variation in terms of the proportion of these mutations in the total nonsynonymous mutations in the various time frames is not so drastic (Fig. 4). More interestingly, most of these mutations observed along several time frames are highly recurrent ones (Table 2). Mutations m.3394T>C, m.3397A>G, and m.3398T>C are repeated in these branches, and more curiously, together with the unique m.3399A>T mutation (in the range 27,700–59,000 years), they are close together in a repetitive region TATATA in the *MT-ND1* gene.

Evaluating the Effect of Expansions on Purifying Selection Effect

To test the relationships between the tree structure and the severity of nonsynonymous variants, we applied multiple linear regression models to data from the detailed mtDNA phylogenetic tree. We have to keep in mind that the cumulative effects of purifying selection increase with depth in the tree (measured by rho) [Pereira et al., 2011], so rho must always be included as a variable in the models. Despite the better resolution of the N tree, the L and M trees also present examples of population expansions in the form of star-like nodes, so all geographical regions are used in the evaluation we performed.

First, we tested the hypothesis that mutation severity leading into each node of the tree may limit the ability of that node to undergo a population expansion. We tested whether the number of branches below a node is negatively correlated with the average MutPred pathogenicity score of the branch immediately above each node (Model 1a). The results (Table 3) showed a high statistically significant ($P = 4.3 \times 10^{-6}$) negative correlation between the number

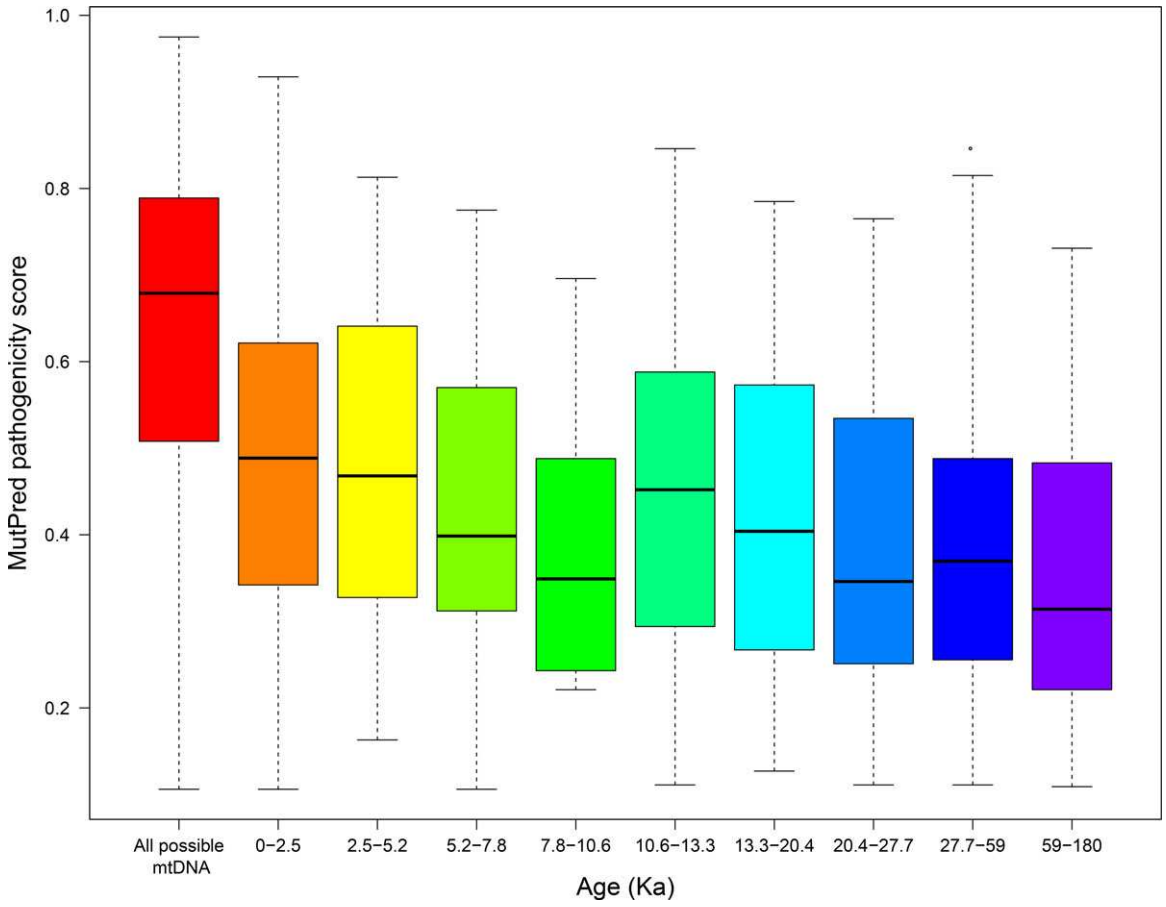

Figure 4. Boxplot showing the MutPred-predicted pathogenicity score distributions in the human population, in intervals of time of emergence of nonsynonymous mutations. The boxes represent the interquartile range and the whiskers are the 5% and 95% quartiles. Outliers are marked as open circles. Ka denotes 1,000 years.

Table 2. Recurrence of the Mutations with MutPred Equal or Higher to 0.7 Observed in the Various Time Frames Analyzed

Period (Ka)	Mutations	Times in the tree	MutPred	Period (Ka)	Mutations	Times in the tree	MutPred
2.5–5.2	m.5979G>A	11	0.813	13.3–20.4	m.3397A>G	25	0.785
	m.4659G>A	18	0.789		m.4680C>A	1	0.743
	m.3398T>C	43	0.775		m.3394T>C	41	0.728
	m.15866A>G	1	0.752		m.7269G>A	30	0.715
	m.14142C>A	3	0.731	20.4–27.7	m.8762T>C	4	0.765
	m.3394T>C	41	0.728		m.11969G>A	18	0.7
	m.15638A>T	1	0.724	27.7–59	m.10750>>G	21	0.846
	m.9234A>G	2	0.712		m.9438G>A	38	0.815
5.2–7.8[a]	m.3398T>C	43	0.775		m.3399A>T	1	0.798
	m.8463A>G	3	0.734		m.3394T>C	40	0.728
10.6–13.3	m.10750A>G	21	0.846		m.15257G>A	12	0.72
	m.3397A>G	25	0.785	59—180	m.13789T>C	3	0.731
	m.14790A>G	2	0.709				

Ka denotes 1000 years.
[a]The interval 7.8–10.6 had no mutations with MutPred equal or higher to 0.7.

of branches and the average MutPred pathogenicity score of the branch above. Results were also statistically significant ($P = 2.1 \times 10^{-6}$) when considering the highest MutPred value in the branch above, instead of mean value (Model 1b). Rho was, as expected, also significantly correlated ($P = 0.0075$ and $P = 0.0077$, respectively) in the models. The highly significant negative correlation shows that the pathogenicity of variants defining a node limits the potential for population expansions of descendants from that node.

Then, we tested the hypothesis that expanding populations experience decreased purifying selection. This was tested by determining whether nodes with a higher number of branches descending from that node (representing population expansions) correlated with an increase in average MutPred scores for the branches in the expansion below the node (representing decreased purifying selection; Model 2). In the regression, the average MutPred score over all branches immediately below the node is the predicted variable and the

Table 3. Values Obtained for the Multilinear Regression Analyses

Model	Variable	Estimated correlation coefficient	*P* value
Model 1a[a]	Average MutPred above	–0.07693	4.29×10^{-6}
	rho	0.031035	0.00751
Model 1b[b]	Highest MutPred above	–0.075406	2.12×10^{-6}
	rho	0.030922	0.00773
Model 2[c]	Number of branches	0.19089	$<2 \times 10^{-16}$
	rho	0.321939	$<2 \times 10^{-16}$

The number of branches was 7,563.
[a]Model 1a: (log number of branches) = A + B1 (average MutPred above) + B2 (log rho).
[b]Model 1b: (log number of branches) = A + B1 (highest MutPred above) + B2 (log rho).
[c]Model 2: (average MutPred below) = A + B1 (log number of branches) + B2 (log rho).

Table 4. Distribution of the Significant and Nonsignificant Back Mutations in Nonreference Alleles Throughout the mtDNA

Region	Number in the significant partition	Number in the nonsignificant partition	Ratio	*P* value (that class vs. other)
D-loop	145	301	0.4751	<2.2e-16
tRNA	17	146	0.1096	0.2235[a]
rRNA	22	218	0.0917	0.4465[a]
Syn	116	1523	0.0729	0.1130[a]
Non-Syn	48	534	0.0880	0.5982[a]

[a]The D-loop portion was not considered in these calculations.

number of those branches and rho were the explanatory variables. As can be seen in Table 3, there is a very strong statistical significant positive correlation ($P = 2 \times 10^{-16}$). The average MutPred pathogenicity score increases as the number of branches out of the same node increases, testifying to a relaxation of the purifying selection when populations are expanding.

Inspecting Back Mutations in the Worldwide Phylogenetic Tree

Given the observed cases of back-mutating heteroplasmies in nonreference alleles within certain haplogroups, we inspected the global tree for signs of significant reverting mutations to the ancestral state. For the 3,664 mutations that appeared for the first time at least 3,800 years ago [Soares et al., 2009], most of them (90.5%) do not display a significant difference between the mutation and its reversion rates, meaning there is no bias in the back-mutation rate.

Let us describe the mutations with a significant difference between the rate of mutation and its reversion as significant back-mutating sites. Of the 349 mutations showing a *P* value lower than 0.05 (Supp. Table S6), 204 occurred in the coding region (48 nonsynonymous, 116 synonymous, 22 in rRNAs, 17 tRNAs, and one in noncoding position), and 145 in the D-loop region. Nevertheless, the D-loop has a significantly higher number of back-mutating sites (*P* value of D-loop versus other regions = 7.453×10^{-72}) and also *MT-ATP6* (*P* value of *MT-ATP6* versus other protein genes = 0.000765). *MT-ATP6* has been reported as displaying a high amount of mutations in general [Pereira et al., 2009]; we confirmed that in this case most of the *MT-ATP6* mutations are synonymous (15 vs. seven nonsynonymous). A similar bias for D-loop mutations is observable when evaluating the ratios between the significant and nonsignificant partitions of back mutations in nonreference alleles for each mtDNA region (Table 4). The D-loop has a clearly higher (four to five times) number of back mutations in nonreference alleles when compared with the tRNA, rRNA, synonymous, and nonsynonymous regions (all these are nonsignificant between them when removing the D-loop values, which adds to the argument that this might be caused by instability and not selection pressure). We could argue that the D-loop, although not coding, contains most regions where regulating proteins align, being preferential regions for secondary structures [Pereira et al., 2008]. Supporting this argument, the next largest ratios, although not significant, are for tRNA and rRNA, regions where secondary structure is also of extreme importance.

We confirmed that all the mutations listed in Duggan and Stoneking (2013), of occurrences in the mtDNA phylogeny of repeated back mutations from a mutation that arose only once during the human evolution (m.16247A>G, m.92G>A, m.1703C>T, and m.3780C>T, although in our tree mostly are not observed only once), are in the back-mutating sites. For the positions we detected in our heteroplasmic dataset for mother-child pairs, which seemed to be tending to revert to the ancestral state, only the one at position 16,217 is listed.

More interestingly, and important for the debate of how to establish a reliable heteroplasmic threshold, we verified that nine out of the 11 (82%) haplogroup-defining heteroplasmic position for threshold >10% are significant back-mutating positions in this analysis. This proportion of back-mutating positions in the heteroplasmies at haplogroup-defining sites increases with the heteroplasmy threshold (45% for threshold >1%; 50% for threshold >2%; 62% for threshold >3%). This fact indicates that these are sites caught in the process of forming a back mutation, and not an indication of bad sequencing quality.

Discussion

The mtDNA phylogenetic worldwide tree, given the current extensive number of sequences, is a highly versatile tool to investigate the interplay between demography and natural selection, which have been shaping *Homo sapiens* diversity over the last 200,000 years. The behavior of mtDNA molecules as a population within an individual allows us to inspect how natural selection behaves at the level of mother–children transmission with a sensitivity (through heteroplasmy dynamics) impossible to achieve for nDNA.

Viewing mtDNA selection at the longer time scale, starting from today and moving back in time, purifying selection acted aggressively at the population level until 2,500 years before present, decreasing the mean pathogenic value of nonsynonymous mutations. The pathogenic score continues to decrease into older sections of the tree, until an inversion occurred due to selection relaxation following the Younger Dryas. Values of mutation mean pathogenicity continue to decrease prior to that episode, until another episode of selection relaxation, not so strong, occurred matching the time of the out-of-Africa migration. These two episodes of selection relaxation were previously identified by Loogvali et al. (2009), by measuring purifying selection by the ratio of nonsynonymous versus synonymous mutations in a tree based on 3,057 coding region sequences. The Neolithic has been associated for a long time with the main episodes of extensive population expansions (e.g., Chikhi et al. [1998]), although latter research on mtDNA have shown that, at least for Europe, the post-LGM expansions have been the key demographic episodes for the human population, whereas Neolithic expansions may have been more important in Africa and in Asia [Atkinson et al., 2008; Soares et al., 2008, 2010; Pala et al., 2012; Rito et al., 2013]. The current mtDNA tree is a good illustration of this, as we have described. So, it would be expected that selection relaxations would be observable during the Neolithic for the L and M parts of the

tree, but the current tree resolution does not allow a precise investigation of this. Nevertheless, it is clear that out-of-Africa-centered models being designed for dating and evaluating purifying selection at nDNA diversity are misleading [Fu et al., 2013], and argue for a centering on postglacial expansions instead. Our claim fully explains the observation made by Fu et al. (2014) that while their simulations suggested that the average density of derived deleterious alleles in European descendants is ∼11.3% higher than in African descendants, the empirical data are only ∼1.4% higher in the first.

Loogvali et al. (2009) indicate three possible explanations leading to the accelerated accumulation of pathogenic mutations since the beginning of the Holocene: (1) adaptive shifts and positive selection, (2) insufficient time for purifying selection, or (3) relaxation of selective constraints. Following evidence of others [Kivisild et al., 2006; Ruiz-Pesini and Wallace, 2006], we also do not see evident signs of positive selection for nonsynonymous mutations with high MutPred pathogenicity scores, which could indicate an important functional change. The most curious observation we made related with this was several high predicted pathogenicity nonsynonymous mutations being repeated in older branches, where the ratio of these mutations is low; but most of these are located in a repetitive region (a probable incipient dinucleotide STR, consisting in three repeats TA) in the *MT-ND1* gene, and mutations here may indicate a strong pressure to keep this region as it is, and not begin to increase in size, as that would lead to frameshift mutations. Concerning the second explanation that there has been insufficient time for purifying selection to act on mtDNA, we have shown that purifying selection is very rapid over the time scale of 1,000 years. It is true that a proportion of mildly deleterious mutations was kept for longer periods (a few mutations with MutPred score >0.7 are observable at different time intervals older than 2,500 years), but these probably result from specific episodes in human history, as we have shown here. It is tempting to relate these specific signs with particular environmental factors, as was done previously for adaptation to climate [Mishmar et al., 2003] and to high altitude [Kang et al., 2013]. Nevertheless, it is very difficult to disentangle the strong interplay between demography and selection strength, as the current high level of resolution of the tree allowed us to confirm.

We reinforce that all our analyses were conducted based on a measure of pathogenicity of nonsynonymous mutations, which despite being only an estimator [Li et al., 2009] is nevertheless much more informative about the potential functional implications of mutations than measures based on synonymous/nonsynonymous ratios (which treat all nonsynonymous mutations as equally bad), as extensively used before [Elson et al., 2004; Kivisild et al., 2006; Loogvali et al., 2009]. This fact makes our results even more relevant.

The combination of heteroplasmy calling for MPS data for mother–offspring pairs and phylogenetic information covering longer-scale mtDNA evolution, allows us to contribute to the recent debate on the reliability of MPS inferences of heteroplasmic sites [Just et al., 2014; Ye et al., 2014a, 2014b]. The fact that low thresholds for heteroplasmy calling leads to many haplogroup-defining positions being observed as heteroplasmic has been leading many authors to suggest the use of this proportion of mutations as a quality control measure to define a safe heteroplasmy threshold, which could be as high as 10%–15%. The assumption behind this is that heteroplasmic mutations should be distributed randomly, and thus should be unlikely to occur on haplogroup-defining sites. However, we confirmed that many of these heteroplasmic mutations at haplogroup-defining positions are nonrandom, being back mutations to the ancestral allele in the overall tree, some within a haplogroup background as previously described [Andrew et al., 2011]. We further confirmed in the worldwide tree that a total of 349 out of 3,664 mutations, which had time to mutate back in the human history, do have a significant tendency to mutate back to the ancestral allele. These are accumulating preferentially in the D-loop, indicating a pressure possibly driven by secondary structure that may be responsible for this [Pereira et al., 2008]. The case reported by Duggan and Stoneking (2013) for position 16,247 is a good example of this. We thus showed that mutations at haplogroup-defining sites back to the more frequent allele globally, especially for the D-loop region, are expected to be observed in the heteroplasmy pool. These are not a necessary indication of bad quality of the MPS, and the list provided by us here can be helpful when ascertaining quality control in datasets of heteroplasmy.

Conclusion

Our analysis shows that population expansion leads to relaxation of purifying selection on mtDNA, and that a pathogenic mutation in a branch restricts the descendant population growth. What is most curious is that although purifying selection acts strongly and continuously until the present, periods of population expansion characterized by a relaxation of the selective pressure leave strong traces on the tree that are not erased by the reestablishment of the selective pressures later on, as in the case of period following the Younger Dryas.

The data analyzed in this paper cover two very different time scales. The phylogenetic tree data cover the range of a few hundred to 200,000 years, whereas the mother–offspring data cover a single generation. Despite the wide range of time scales, we observed similarities between the heteroplasmic variants in the mother–offspring pairs and the phylogenetic tree variants. Most strikingly, we found a high level of reversion of haplogroup-defining mutations, in both the phylogenetic dataset and as heteroplasmies in the single-generation dataset. This common result is proof of concept that there is a common dynamic to mtDNA variation selection across this broad time scale.

Acknowledgments

Author contributions: P.S., D.C.S., and L.P. designed research; B.C. performed research; R.C. contributed with bioinformatics expertise; A.B., M.D.C., V.F., J.B.P., and T.R. contributed to phylogenetic analysis; and L.P. wrote the paper with collaboration of all authors.

Disclosure statement: The authors declare no conflict of interest.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Ajioka RS, Jorde LB, Gruen JR, Yu P, Dimitrova D, Barrow J, Radisky E, Edwards CQ, Griffen LM, Kushner JP. 1997. Haplotype analysis of hemochromatosis: evaluation of different linkage-disequilibrium approaches and evolution of disease chromosomes. Am J Hum Genet 60:1439–1447.

Andrew T, Calloway CD, Stuart S, Lee SH, Gill R, Clement G, Chowienczyk P, Spector TD, Valdes AM. 2011. A twin study of mitochondrial DNA polymorphisms shows that heteroplasmy at multiple sites is associated with mtDNA variant 16093 but not with zygosity. PLoS One 6:e22332.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147.

Atkinson QD, Gray RD, Drummond AJ. 2008. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. Mol Biol Evol 25:468–474.

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. Genetics 141:743–753.

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G. 1998. Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. Proc Natl Acad Sci USA 95:9053–9058.

Chinnery PF, Samuels DC. 1999. Relaxed replication of mtDNA: a model with implications for the expression of disease. Am J Hum Genet 64:1158–1165.

Duggan AT, Stoneking M. 2013. A highly unstable recent mutation in human mtDNA. Am J Hum Genet 92:279–284.

Elson JL, Samuels DC, Turnbull DM, Chinnery PF. 2001. Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. Am J Hum Genet 68:802–806.

Elson JL, Turnbull DM, Howell N. 2004. Comparative genomics and the evolution of human mitochondrial DNA: assessing the effects of selection. Am J Hum Genet 74:229–238.

Fernandes V, Triska P, Pereira JB, Alshamali F, Rito T, Machado A, Fajkosova Z, Cavadas B, Cerny V, Soares P, Richards MB, Pereira L. 2015. Genetic stratigraphy of key demographic events in Arabia. PLoS One 10:e0118625.

Forster P, Harding R, Torroni A, Bandelt HJ. 1996. Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59:935–945.

Fu W, Gittelman RM, Bamshad MJ, Akey JM. 2014. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. Am J Hum Genet 95:421–436.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, NHLBI Exome Sequencing Project, Akey JM. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220.

Gardner K, Payne BA, Horvath R, Chinnery PF. 2015. Use of stereotypical mutational motifs to define resolution limits for the ultra-deep resequencing of mitochondrial DNA. Eur J Hum Genet 23:413–415.

Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. Science 343:747–751.

Ju YS, Alexandrov LB, Gerstung M, Martincorena I, Nik-Zainal S, Ramakrishna M, Davies HR, Papaemmanuil E, Gundem G, Shlien A, Bolli N, Behjati A, et al. 2014. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. Elife 3:e02935.

Just RS, Irwin JA, Parson W. 2014. Questioning the prevalence and reliability of human mitochondrial DNA heteroplasmy from massively parallel sequencing data. Proc Natl Acad Sci USA 111:E4546–E4547.

Kang L, Zheng HX, Chen F, Yan S, Liu K, Qin Z, Liu L, Zhao Z, Li L, Wang X, He Y, Jin L. 2013. mtDNA lineage expansions in Sherpa population suggest adaptive evolution in Tibetan highlands. Mol Biol Evol 30:2579–2587.

Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. 2013. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet 9:e1003794.

Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, et al. 2006. The role of selection in the evolution of human mitochondrial genomes. Genetics 172:373–387.

Kloss-Brandstatter A, Pacher D, Schonherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F. 2011. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32:25–32.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744–2750.

Loogvali EL, Kivisild T, Margus T, Villems R. 2009. Explaining the imperfection of the molecular clock of hominid mitochondria. PLoS One 4:e8260.

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace D. 2003. Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci USA 100:171–176.

Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, Reidla M, Tamm E, Karmin M, Reisberg T, Hooshiar Kashani B, Perego UA, Carossa V, et al. 2012. Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. Am J Hum Genet 90:915–924.

Pereira F, Soares P, Carneiro J, Pereira L, Richards MB, Samuels DC, Amorim A. 2008. Evidence for variable selective pressures at a large secondary structure of the human mitochondrial DNA control region. Mol Biol Evol 25:2759–2770.

Pereira L, Freitas F, Fernandes V, Pereira JB, Costa MD, Costa S, Maximo V, Macaulay V, Rocha R, Samuels DC. 2009. The diversity present in 5140 human mitochondrial genomes. Am J Hum Genet 84:628–640.

Pereira L, Soares P, Maximo V, Samuels DC. 2012. Somatic mitochondrial DNA mutations in cancer escape purifying selection and high pathogenicity mutations lead to the oncocytic phenotype: pathogenicity analysis of reported somatic mtDNA mutations in tumors. BMC Cancer 12:53.

Pereira L, Soares P, Radivojac P, Li B, Samuels DC. 2011. Comparing phylogeny and the predicted pathogenicity of protein variations reveals equal purifying selection across the global human mtDNA diversity. Am J Hum Genet 88:433–439.

Pereira L, Soares P, Triska P, Rito T, van der Waerden A, Li B, Radivojac P, Samuels DC. 2014. Global human frequencies of predicted nuclear pathogenic variants and the role played by protein hydrophobicity in pathogenicity potential. Sci Rep 4:7155.

Picardi E, Pesole G. 2012. Mitochondrial genomes gleaned from human whole-exome sequencing. Nat Methods 9:523–524.

Rito T, Richards MB, Fernandes V, Alshamali F, Cerny V, Pereira L, Soares P. 2013. The first modern human dispersals across Africa. PLoS One 8:e80031.

Ruiz-Pesini E, Wallace DC. 2006. Evidence for adaptive selection acting on the tRNA and rRNA genes of human mitochondrial DNA. Hum Mutat 27:1072–1081.

Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67:718–726.

Schaibley VM, Zawistowski M, Wegmann D, Ehm MG, Nelson MR, St Jean PL, Abecasis GR, Novembre J, Zollner S, Li JZ. 2013. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome Res 23:1974–1984.

Soares P, Abrantes D, Rito T, Thomson N, Radivojac P, Li B, Macaulay V, Samuels DC, Pereira L. 2013. Evaluating purifying selection in the mitochondrial DNA of various mammalian species. PLoS One 8:e58993.

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt HJ, Torroni A, Richards MB. 2010. The archaeogenetics of Europe. Curr Biol 20:R174–R183.

Soares P, Alshamali F, Pereira JB, Fernandes V, Silva NM, Afonso C, Costa MD, Musilova E, Macaulay V, Richards MB, Cerny V, Pereira L. 2012. The expansion of mtDNA haplogroup L3 within and out of Africa. Mol Biol Evol 29:915–927.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740–759.

Soares P, Trejaut JA, Loo JH, Hill C, Mormina M, Lee CL, Chen YM, Hudjashov G, Forster P, Macaulay V, Bulbeck D, Oppenheimer S, Lin M, Richards MB. 2008. Climate change and postglacial human dispersals in southeast Asia. Mol Biol Evol 25:1209–1218.

Toomajian C, Ajioka RS, Jorde LB, Kushner JP, Kreitman M. 2003. A method for detecting recent selection in the human genome from allele age estimates. Genetics 165:287–297.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ. 2006. Harvesting the fruit of the human mtDNA tree. Trends Genet 22:339–345.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386–E394.

Williams SL, Mash DC, Zuchner S, Moraes CT. 2013. Somatic mtDNA mutation spectra in the aging human putamen. PLoS Genet 9:e1003990.

Wonnapinij P, Chinnery PF, Samuels DC. 2008. The distribution of mitochondrial DNA heteroplasmy due to random genetic drift. Am J Hum Genet 83:582–593.

Ye K, Lu J, Ma F, Keinan A, Gu Z. 2014a. Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. Proc Natl Acad Sci USA 111:10654–10659.

Ye K, Lu J, Ma F, Keinan A, Gu Z. 2014b. Reply to Just et al.: Mitochondrial DNA heteroplasmy could be reliably detected with massively parallel sequencing technologies. Proc Natl Acad Sci USA 111:E4548–E4550.