Supporting Information

Rebolledo-Jaramillo et al. 10.1073/pnas.1409328111

SI Materials and Methods

Numt Reads Simulation. We obtained the hg19 numt genomic coordinates from the numt track of the University of California, Santa Cruz (UCSC) Genome Browser (1), extended them by 1 kb on each side and downloaded the corresponding sequences in FASTA format from the UCSC Genome Browser (http://genome.ucsc.edu/). We used these numt sequences to simulate 1 million paired-end reads of different lengths (50, 100, 150, 200, 250, and 300 bp) with wgsim (2). The numt-derived paired-end reads were aligned to the human reference genome hg19 (chrM replaced by the revised Cambridge Reference Sequence, NC_012920) with bwa (3). Reads were simulated and aligned with 1,000 simulations at each read length. The number of reads mapped to the mitochondrial genome was recorded.

Likelihood Function for Evaluating Point Heteroplasmies. We used the tool ANGSD (4) (www.github.com/ANGSD) to obtain a likelihood ratio test statistic (LRT) for each site in each sample. The tool implements the log likelihood function for the allele frequencies $f = (f_A, f_C, f_G, f_T)$, $f_A + f_C + f_G + f_T = 1$, for k reads covering a position, given by

$$l(\mathbf{f}) = \sum_{i=1}^{k} Log\left(\sum_{b\in S} \Pr(x_i|B_i = b) f_b\right),$$

where $\Pr(x_i|B_i = b)$ is the position-specific probability of the data read, x_i, from the i^{th} read covering the position, given that the true nucleotide in read i, B_i, equals b, $b \in S$={A,C,G,T}. The tool calculates the LRT for the hypothesis H_0: $f_2 = f_3 = f_4 = 0$, where f_j is the frequency of the allele with the j^{th} highest allele frequency. Strong statistical evidence against H_0 indicates that the site is heteroplasmic. The tool was applied independently to each sample's filtered reads. The LRT statistic was transformed into a P value using the χ^2 approximation with 3 degrees of freedom.

Germ-Line Bottleneck Size and Mutation Rate Estimation Accounting for Mitotic Segregation. One can argue that the estimates of the germ-line bottleneck size do not account for the variance owing to developmental bottlenecks and mitotic mtDNA segregation, and this can lead to a lower germ-line bottleneck size. To take this into account, we subtracted the variance between two tissues for the same individual from our estimate of the genetic variance for each quartet, that is, we computed $\sigma^2_{gen} = ((MAF_{m1} - MAF_{c1})^2 + (MAF_{m1} - MAF_{c2})^2 + (MAF_{m2} - MAF_{c1})^2 + (MAF_{m2} - MAF_{c2})^2 - 2(MAF_{m1} - MAF_{m2})^2 - 2(MAF_{c1} - MAF_{c2})^2)/4$, where m is mother, c is child, 1 is buccal tissue, and 2 is blood tissue. Applying this approach to the same 51 quartets as above (while excluding two quartets with negative N) led to the median estimate of $N = 35.0$ (with interquartile range 10.0–138.0; Fig. S15*B*), a value slightly higher but still very similar to our other estimates of N (discussed in the main text). Although this will have to be evaluated in further studies in more detail, our results argue for a smaller effect of developmental bottlenecks and mitotic mtDNA segregation than of the germ-line bottleneck on determining heteroplasmy levels in tissues.

Indel Analysis. The reads used to assemble were first mapped to hg19, rCRS, pUC18, and PhiX174 as described in Fig. S4. Only nonduplicate read pairs that mapped to the rCRS were used as input to the assembler, SPAdes (5). Assemblies were curated by aligning the contigs to the rCRS, then discarding contigs that fell entirely within the alignment of a larger contig. One assembly was chosen to use as the reference for mapping all four samples in each family. The assembly was chosen by first discarding assemblies with no LASTZ hits to the rCRS, ones with over 500 contigs, and ones with an erroneous full-genome duplication. Then the assemblies were narrowed to those without contigs with non-rCRS flanks, and finally the remaining assembly with the lowest number of contigs was chosen.

The same reads used to build the assemblies were mapped back to them using the same bwa version and options as in Dataset S1, Table S19. The only difference is that we did not limit the number of mismatches to the quartet-specific reference, which otherwise could have biased against indels. The alignments for each quartet were merged into one file, with samples marked with read groups. Then, several of the filtering steps described above were applied, specifically PicardTools MarkDuplicates, selecting reads properly mapped in a pair and above a minimum read length, and removing chimeric alignments. The Naive Variant Caller was used to find indels and their read counts, using the same settings as above, except the region restriction. The resulting indels were filtered by quartet to eliminate ones that were below 0.75% frequency or 1,000× coverage in all members. Then indels were eliminated that were above 1.0 strand- or mate-bias in all samples in the quartet. The strand-bias metric used was the "SB" formula in Guo et al. (6) The mate-bias metric used the same formula, but with the first and second mate in the pair replacing forward and reverse strand reads. Indels in the low-complexity regions of 302–316 and 16,183–16,193 were excluded. Finally, indels above 1.0% frequency in any sample, with less than 1.0 strand- and mate-bias were considered real.

To estimate the background noise in indel frequency estimation, the same pipeline was applied to the PCR-amplified pGEMTeasy-derivative Z1-1 clone described below. The frequency cutoff was lowered to 0.2% to observe spurious indel calls. Minor allele frequencies at microsatellites similar to those containing our putative indel heteroplasmies were used as a baseline error rate.

Preparation of Artificial Heteroplasmy Standards from PCR Amplicons and Clonal DNA. To determine the heteroplasmy detection thresholds of Sanger sequencing and ddPCR, we created artificial mixtures with known allelic ratios. To do so we mixed mtDNA amplicons from a sample M9 (7), who is heteroplasmic at site 8,992 (C = 65.9%, T = 34.1%), with mtDNA amplicons of a sample MSu homoplasmic at that same site (C = 100%, T = 0%). Several mixtures were prepared with the frequency of T allele ranging from 0 to 34%. The resulting mixtures were analyzed with ddPCR and Sanger sequencing (Dataset S1, Tables S5 and S7 and Fig. S8 *A* and *B*).

To determine the measurement error ($\sigma^2_{measure}$) for computing the bottleneck size we cloned the D-loop in pGEM-T-Easy (clone Z1-1). Next, we ran two independent PCR reactions amplifying the whole clone (with primers located next to each other but facing opposite directions). Each of such PCR fragments was then sequenced twice in two independent MiSeq runs. We computed the averaged (among sites) squared difference in MAFs for all sites between the runs corresponding to the two PCR reactions. Results are shown in Dataset S1, Table S20.

1. Calabrese FM, Simone D, Attimonelli M (2012) Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics* 13(Suppl 4):S15.

2. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

3. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
4. Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R (2013) Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics* 14:289.
5. Bankevich A, et al. (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19(5):455–477.
6. Guo Y, et al. (2012) The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* 13:666.
7. Goto H, et al. (2011) Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol* 12(6):R59.

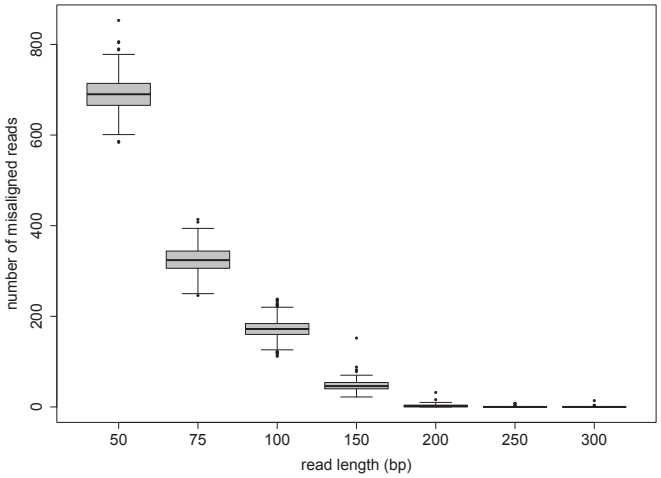

Fig. S1. The number of numt-derived reads aligning to the mtDNA reference genome decreases with read length. One million read pairs were simulated from the known numt sequences annotated in the UCSC Genome Browser (*SI Materials and Methods*). The boxplots show the distribution of the number of numt-derived reads (out of 1 million read pairs) erroneously mapping to the mtDNA reference genome in 1,000 simulations per read length. At 250 bp (the read length used in this study) virtually no numt-derived reads align to mtDNA.

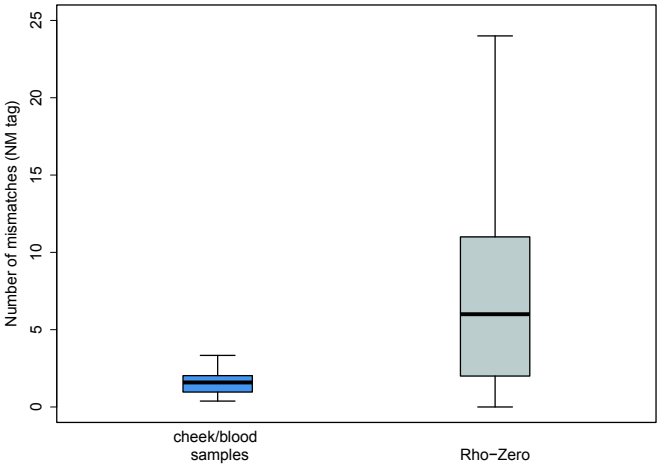

Fig. S2. The distribution of the average number of mismatches to the reference human mtDNA sequence compared between our regular cheek/blood sample reads (~1 million read pairs) and reads obtained from human 143B Rho0 cells (260,000 read pairs). Because the Rho0-derived reads mapping to mitochondrial genome originate from numts, filtering on the number of mismatches can effectively eliminate numt contamination. Amplification of mtDNA from Rho0 cells did not produce the expected 9-kb bands, and sequencing the short products of this amplification resulted in a 10-fold decrease in the number of reads, with only 2% of them mapping to the reference mtDNA. Such reads were numt-derived because of a large number of mismatches (see plot). In contrast, most reads from our regular samples mapped to mtDNA with a small number of mismatches. Thus, the contribution of numts to sequencing of our regular samples was minimal.

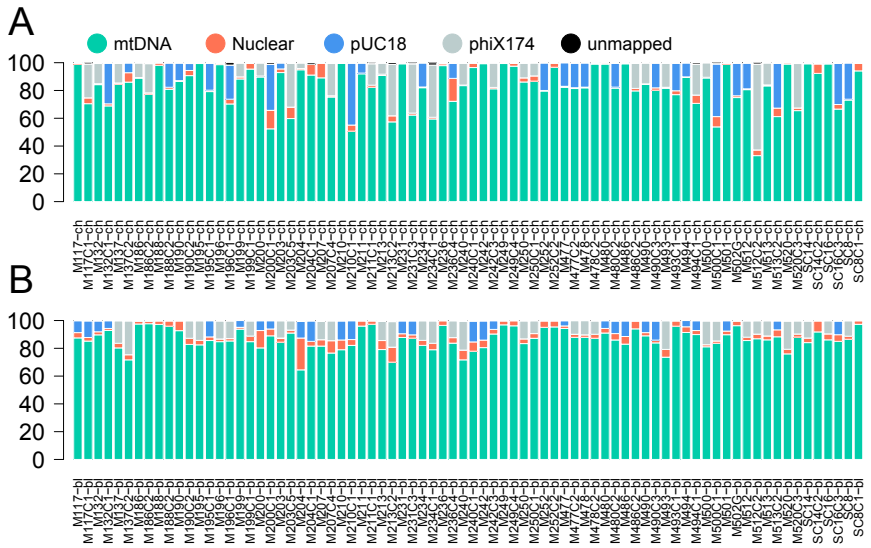

Fig. S3. The summary of read alignments with bwa aligner. The percentage (out of the total number of reads per sample) of reads mapping to a particular reference in (*A*) cheek and (*B*) blood samples is shown. Sample names are indicated on the *x* axis.

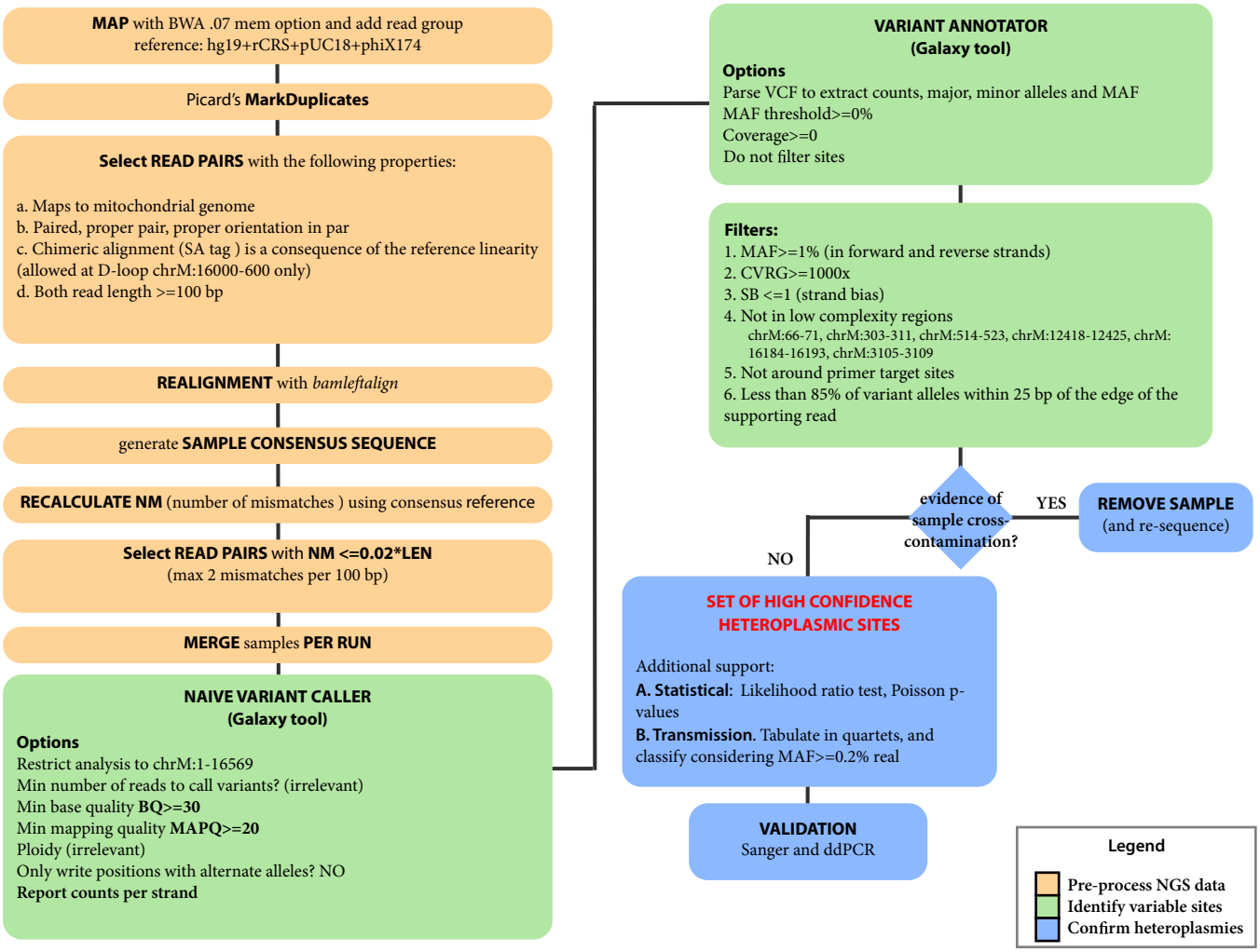

Fig. S4. Sequencing reads processing pipeline for the detection of point heteroplasmies. Reads are processed as pairs taking into consideration several filters to minimize the interference by numts, as well as alignment and sequencing artifacts. Strand bias was computed as in ref. 1. The parameters for each software is given in Dataset S1, Table S19.

1. Guo Y, et al. (2012) The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* 13:666.

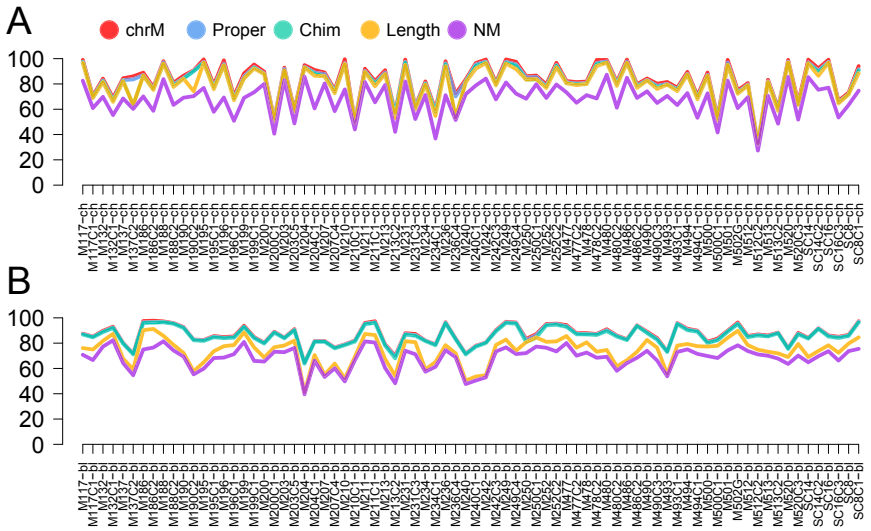

Fig. S5. The percentage of the total number of reads retained after applying consecutive filters (per sample). (*A*) Cheek and (*B*) blood samples. Filtering steps: (1) chrM, the percentage of reads mapping to the mtDNA reference; (2) Proper, the percentage of the reads retained after filter 1 and in which a read and its mate are properly paired; (3) Chim, the percentage of the reads retained after filtering steps 1–2 and having no informed chimeric alignments outside the D-loop region ("SA" tag); (4) Length, the percentage of the reads retained after steps 1–3 and having read length equal to or above 100 bp; and (5) NM, the percentage of the reads retained after steps 1–4 and having less than five mismatches to the mtDNA reference.

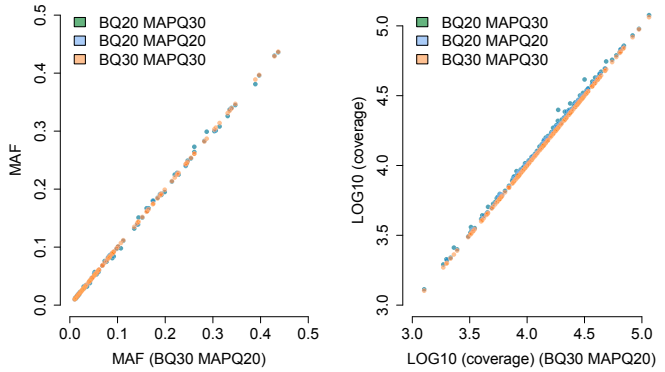

Fig. S6. Minor allele frequency and site coverage are robust to filtering thresholds. MAF (*Left*) and site coverage (CVRG, *Right*) of 172 sites originally identified as high-quality heteroplasmies with BQ ≥30 and MAPQ ≥20 (*x* axis). Different thresholds have no noticeable effect on the estimation of either the minor allele frequency ($R^2 \geq 0.999$) or coverage ($R^2 \geq 0.997$) of the sites.

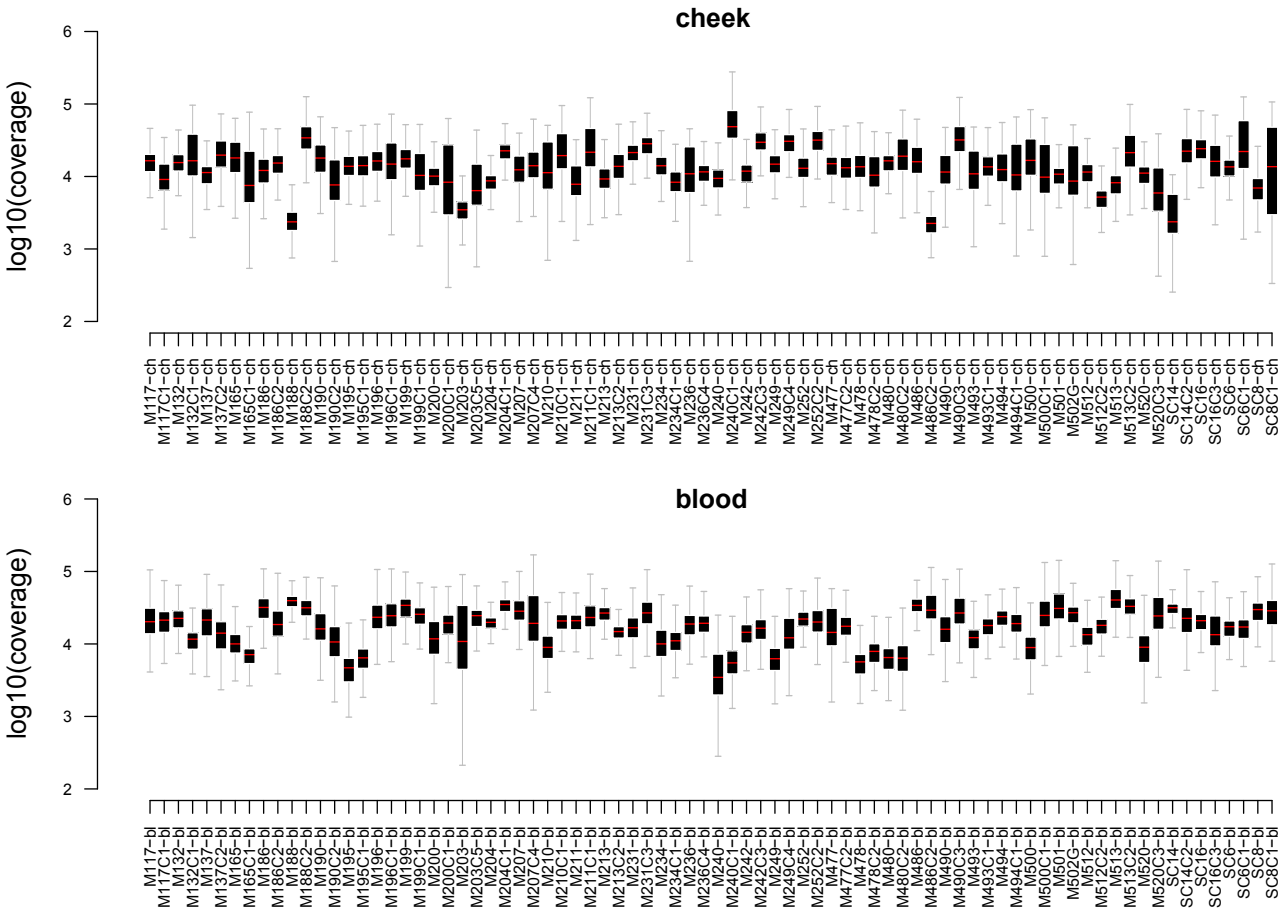

Fig. S7. Cheek and blood sample sequencing depth distribution. Distribution of the number of bases with sequencing quality ≥30 per site after applying several quality filters to the reads (Fig. S5). Red color indicated the median of the distribution. The *y* axis is on the $\log_{10}$ scale.

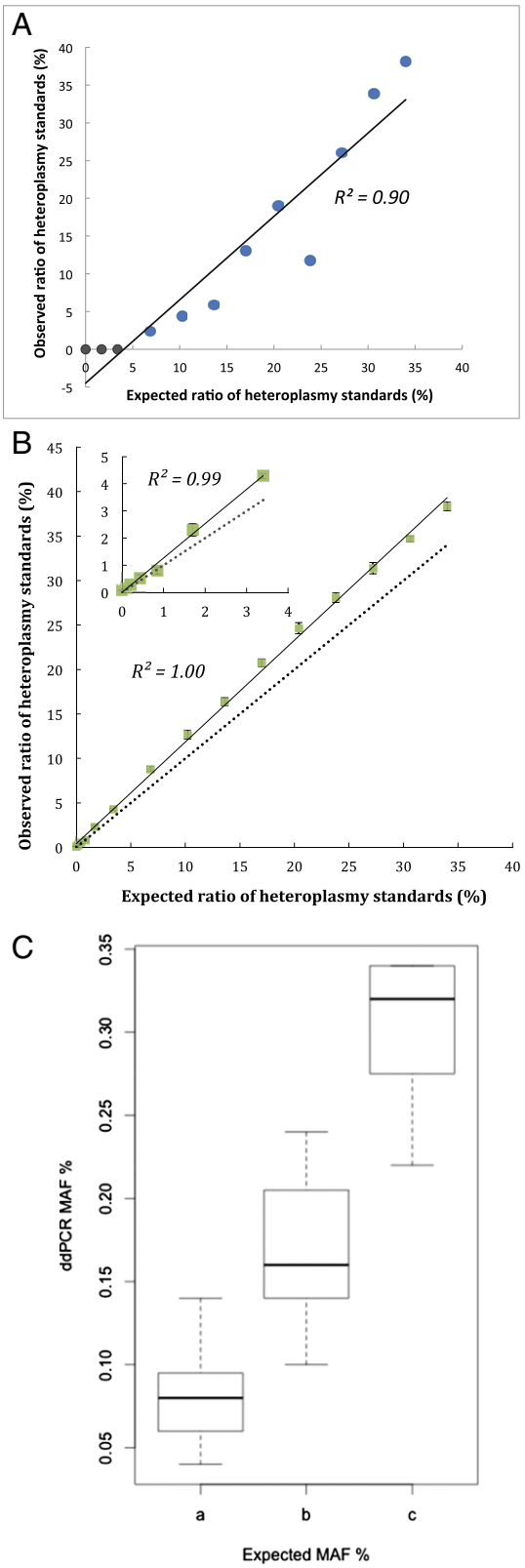

Fig. S8. (*A*) The observed values were obtained from Sanger sequencing of artificially premixed samples heteroplasmic at site 8,992 (from ref. 1). MAF at this site was premixed to be between 0% and 34% (see Dataset S1, Table S5 and *SI Materials and Methods* for details). Blue, heteroplasmy is observed; gray, heteroplasmy is not observed. The data shown are the mean intensities of at least two Sanger sequencing reactions per sample. (*B*) ddPCR detection of point heteroplasmies. The observed values were obtained from ddPCR scoring of artificially premixed samples heteroplasmic at site 8,992 (from ref. 1). MAF at this site was premixed to be between 0% and 34% (see Dataset S1, Table S7 and *SI Materials and Methods* for details). (*Inset*) Samples with expected MAF under 4% in more detail. The data plotted are mean ± SD of eight technical replicates generated from duplicate PCR amplicons per sample. The dotted line shows the

Legend continued on following page

expectation. (*C*) Determination of heteroplasmy detection limit with ddPCR. Observed MAFs were measured with ddPCR in artificial mixtures with known MAFs of (a) 0.0%, (b) 0.11%, and (c) 0.21%. The box plots show the distribution of observed MAFs in each artificial heteroplasmy standard across at least eight technical replicates. The 0% and 0.21% are well separated, and thus 0.21% is considered to be the detection limit.

1. Goto H, et al. (2011) Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol* 12(6):R59.

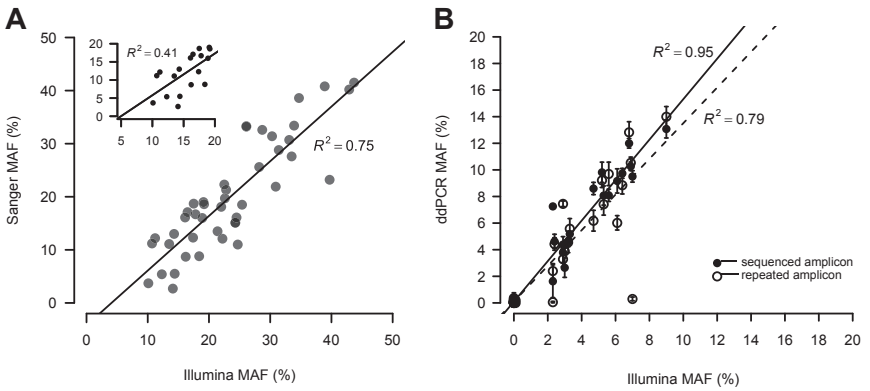

Fig. S9. Comparison of MAFs detected by three distinct experimental approaches: Sanger sequencing, Illumina sequencing (MiSeq instrument), and ddPCR. (*A*) The correlation between MAFs detected with Sanger (*y* axis) and Illumina sequencing (*x* axis). Black circles represent MAFs for 84 sites (21 sites in four samples; for each value the intensity was averaged across at least two sequencing runs). (*Inset*) A magnification of the area with MAFs between 10 and 20%. (*B*) The correlation between MAFs detected with ddPCR (*y* axis) and Illumina sequencing (*x* axis). For sequenced amplicons, the same exact long-range PCR was used to perform MiSeq sequencing and ddPCR (aliquots from the same tube). For repeated amplicons, ddPCR was performed on long-range PCR amplicons produced by an independent reaction (not used for MiSeq sequencing). Data are shown as mean ± SD of four technical replicates generated from PCR amplicons per sample.

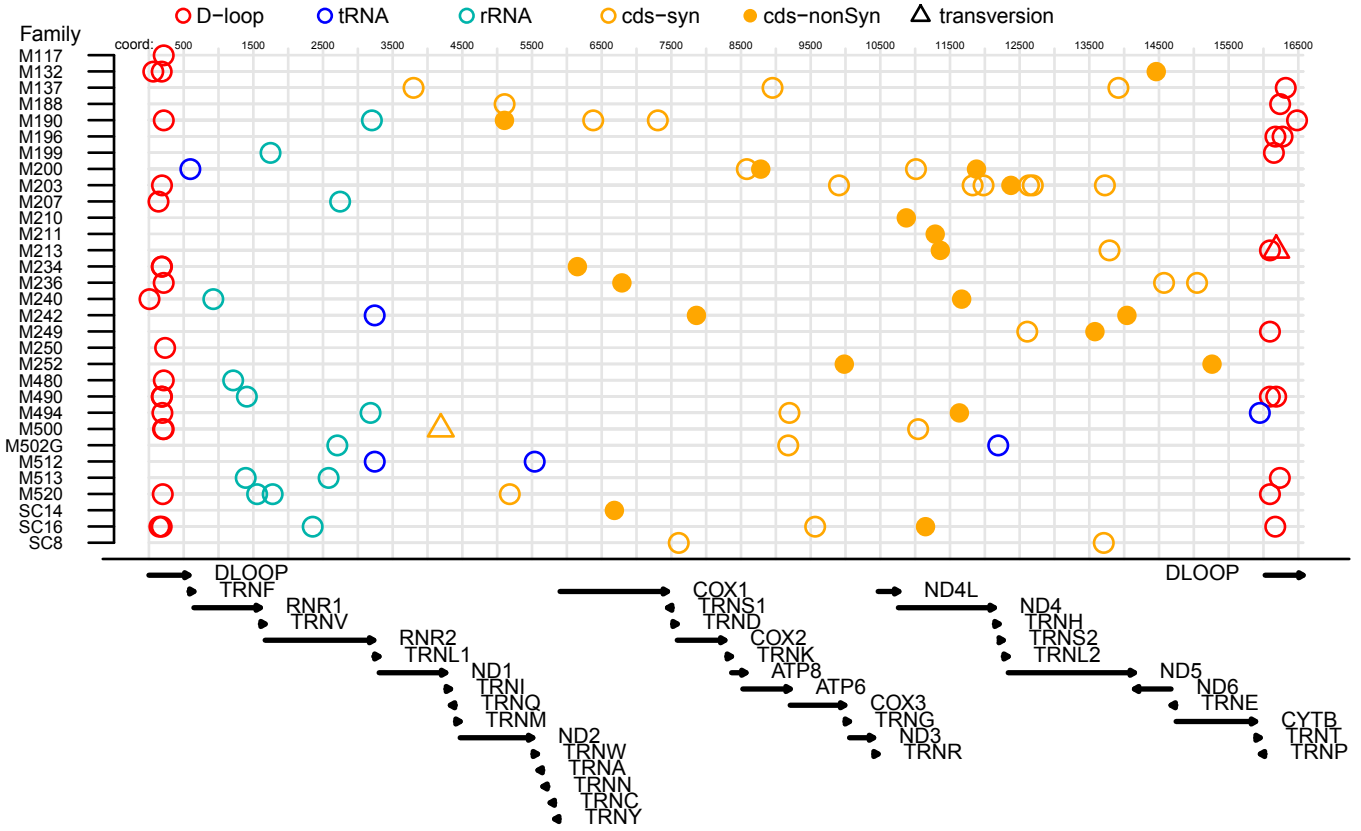

Fig. S10. The distribution of high-quality heteroplasmies from 98 quartets (two tissues from a mother and two tissues from her child) along the mitochondrial genome (*x* axis) stratified by families (*y* axis). Arrows at the bottom of the graph represent strandedness of mitochondrial protein-coding and RNA genes. Circles represent transitions and triangles transversions. cds, protein-coding regions.

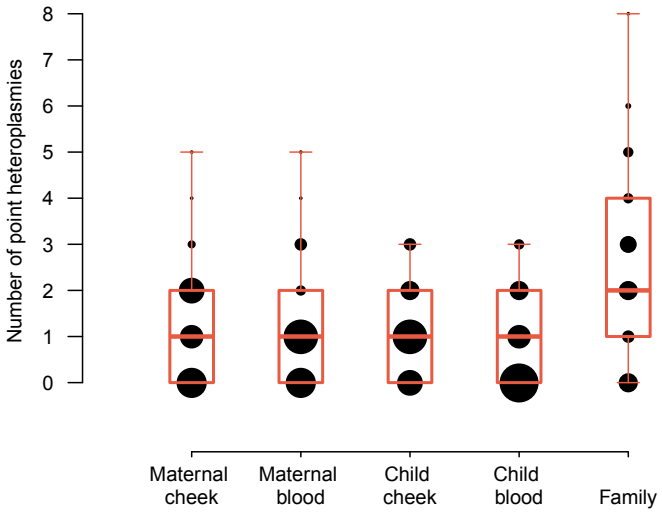

Fig. S11. Distribution of the number of point heteroplasmies per individual or per family. The four leftmost boxplots represent the distribution of the number of heteroplasmies per individual. The rightmost boxplot displays the distribution of the number of heteroplasmies per quartet (two tissues from a mother and two tissues from her child). The size of the dots is proportional to the number of individuals/quartets carrying the corresponding number (*y*-axis coordinate) of heteroplasmies. The average number of point heteroplasmies for maternal blood was 1.13 ± 0.04 (mean ± SE) heteroplasmies per individual. These values were similar for maternal buccal tissue (25 mothers with point heteroplasmies, 1.18 ± 0.04 heteroplasmies per person) and for buccal and blood tissues analyzed in children (28 and 23 individuals with point heteroplasmies, respectively; 1.13 ± 0.02 and 0.97 ± 0.03 sites per person, respectively). The number of heteroplasmic sites per individual was not significantly different between the two maternal or child tissues, between maternal and child blood ($P = 0.279$), or between maternal and child buccal tissue ($P > 0.25$ in each case, two-sided paired *t* test).

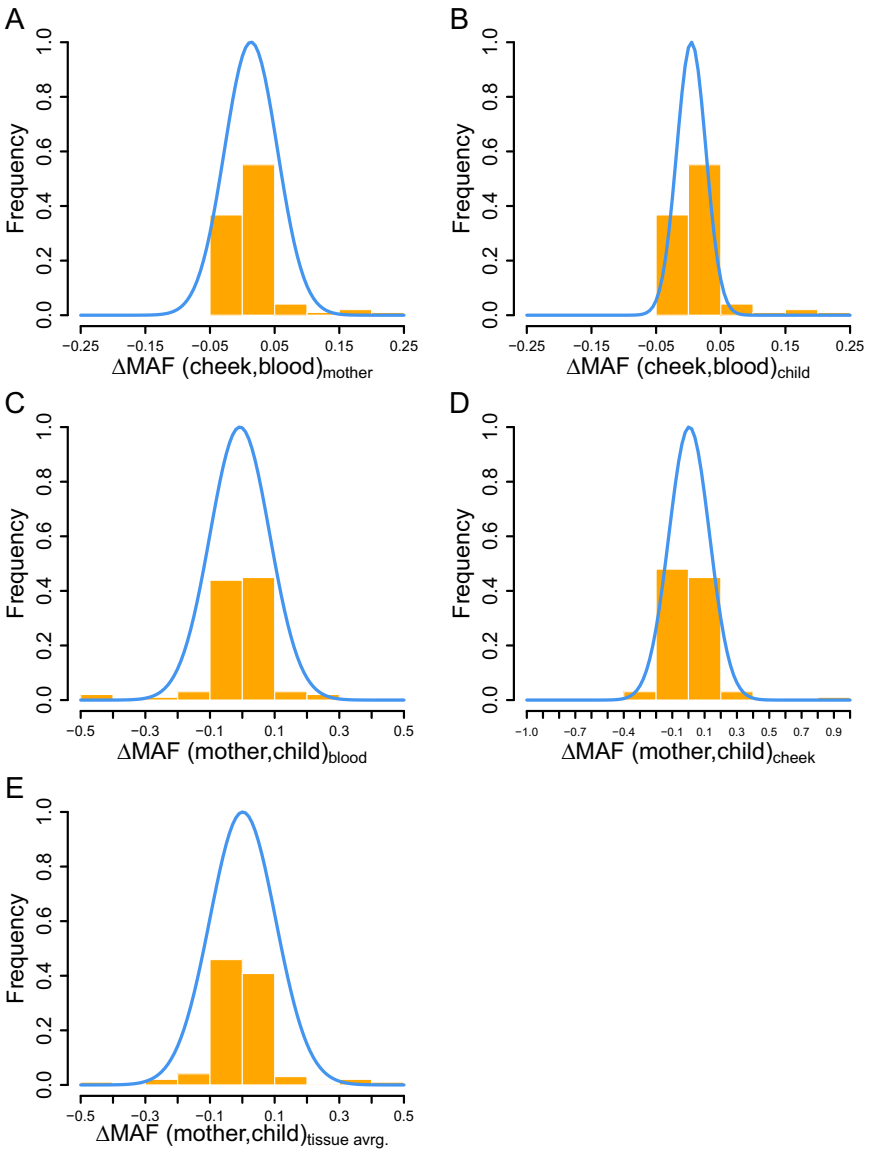

Fig. S12. The distribution of the differences in heteroplasmy allele frequencies between (*A*) maternal buccal tissue and blood, (*B*) child buccal tissue and blood, (*C*) mother and child blood, (*D*) mother and child buccal tissues, and (*E*) the average (between blood and buccal) maternal heteroplasmy allele frequency and the average (between blood and buccal) child heteroplasmy allele frequency. For *C*–*E* we assume maternal major allele to be ancestral for the family.

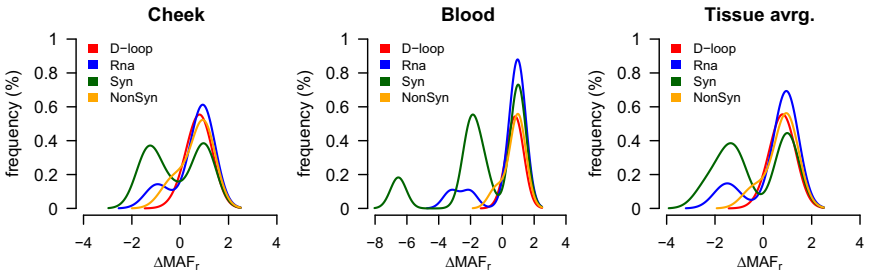

Fig. S13. Relative change (ΔMAF_r) in heteroplasmy allele frequency grouped by categories corresponding to genomic features within mitochondrial DNA: D-loop, RNA (rRNA and tRNA genes), Syn (synonymous sites), and Nonsyn (nonsynonymous sites). Here maternal major allele is assumed to be ancestral for the family. $\Delta MAF_r = (MAF_{mother} - MAF_{child})/MAF_{mother}$.

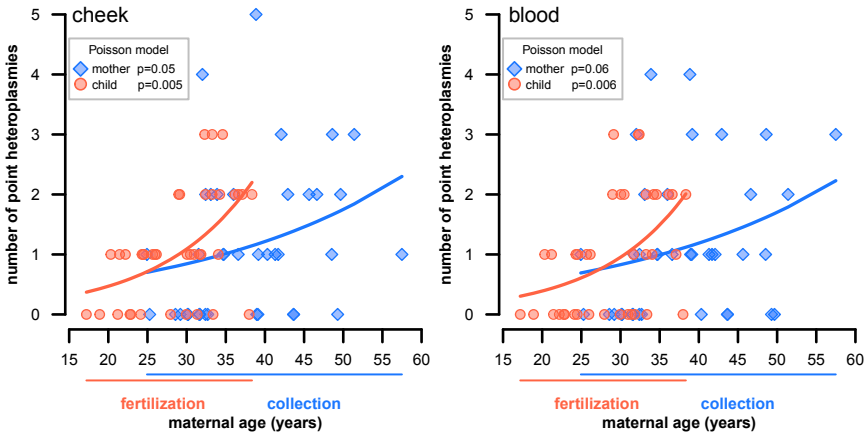

Fig. S14. Maternal age effect per tissue. Age of the mother at the time of tissue collection vs. the number of point heteroplasmies found in the corresponding tissue of the mother (blue) and age of the mother at the time of conception of the child (fertilization age) vs. the number of point heteroplasmies found in her child in the matching tissue (red). Poisson generalized linear models predicting the number of sites in the mother or child and the corresponding *P* values for the predictor (age at collection or fertilization, respectively) are indicated for each comparison.

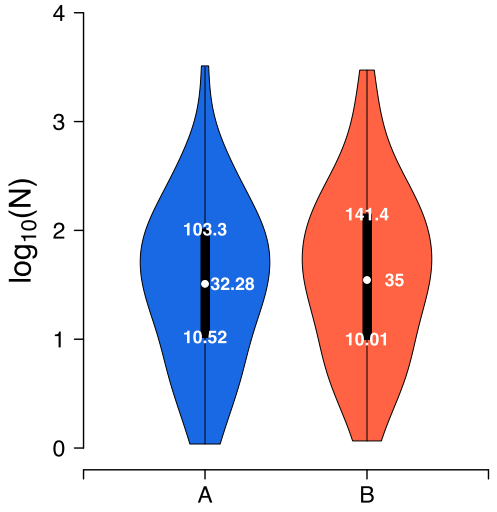

Fig. S15. Bottleneck size estimations. (*A*) The size of the germ-line bottleneck (N) was estimated according to ref. 1 using the minor alleles frequencies of the quartets in the "all," "mother," and "somatic-loss" categories only considering instances when maternal minor allele frequency was above 1% in one tissue and above 0.2% in another tissue (a total of 51 quartets, leftmost violin plot). (*B*) A correction for variance owing to developmental bottleneck and mitotic mtDNA segregation was applied at the rightmost plot. See text for details. The median and interquartile range are indicated inside the violin plots.

1. Millar CD, et al. (2008) Mutation and evolutionary rates in adélie penguins from the antarctic. *PLoS Genet* 4(10):e1000209.

Other Supporting Information Files

Dataset S1 (XLS)