Accelerated expansion of pathogenic mitochondrial DNA heteroplasmies in Huntington's disease

Yiqin Wang[a,†], Xiaoxian Guo[a,†], Kaixiong Ye[a,#], Michael Orth[b], Zhenglong Gu[a,*]

[a] *Division of Nutritional Sciences, Cornell University, Ithaca, New York 14853, USA;*

[b] *Department of Neurology, Ulm University Hospital, Ulm, Germany.*

†These authors contributed equally to this work

#Current address: Department of Genetics, University of Georgia, Athens, GA 30602-7223

*Corresponding author: Zhenglong Gu, zg27@cornell.edu, +1 607-254-5144

SUPPORTING INFORMATION

Contents

Supplemental Discussion 2
1. Possibly fixed pathogenic mitochondrial genome (mtDNA) heteroplasmies in lymphoblasts 2
2. Very-low-fraction pathogenic mtDNA heteroplasmies in blood samples 3
3. mtDNA content in lymphoblast and blood samples 4
4. mtDNA haplogroup in lymphoblast and blood samples 5
Supplemental Methods 7
1. Read duplicates identification and consensus sequence calling 7
2. Nuclear mitochondrial DNA segment (NUMTS) identification and filtering 8
3. mtDNA variant detection and quality control procedures 11
4. mtDNA content quantification and quality control procedures 15
5. Statistical modeling of age-dependent changes of mtDNA heteroplasmies in lymphoblasts 17
Supplemental References 19
Supplemental Figures 21
Supplemental Tables 37
Supplemental Data (Titles) 57

Supplemental Discussion

1. Possibly fixed pathogenic mitochondrial genome (mtDNA) heteroplasmies in lymphoblasts

Theoretical analyses of the expansion of mtDNA heteroplasmies indicate that a new mutation can drift to high fractions, or even become fixed in mtDNA of a subset of cells within decades well under the life span of humans (1–3). We thus sought to investigate to what extent these fixed heteroplasmies would influence our conclusions regarding the heteroplasmies in lymphoblasts of Huntington's disease (HD) patients.

Among the 181 HD patients with longitudinal blood samples, 180 had lymphoblast mtDNA which was sequenced and passed quality control for homoplasmy (variant allele fraction, VAF≥0.99) analysis in the current study. We used the mtDNA sequence obtained from the blood samples as the background sequence to infer fixed heteroplasmies (VAF≥0.99, the same VAF criterion used to define homoplasmies) in the derived lymphoblast cell line. As a result, we found that only 5 out of the 3623 homoplasmies detected in these lymphoblast samples were absent in blood samples from the same individual, suggesting that fixation of mtDNA heteroplasmies is possible but still a rare event in lymphoblasts of the current study. We further determined pathogenic homoplasmies by using the same criteria as for heteroplasmies, although this definition might include some private mtDNA polymorphisms from the remaining 1369 HD lymphoblasts and 182 control lymphoblasts for which we did not have mtDNA data on the parental tissue.

Remarkably, 30 predicted pathogenic homoplasmies were detected in 1549 HD lymphoblasts, while only one was detected in 182 control lymphoblasts (logistic regression, OR=3.0, P=0.3). Adding these homoplasmies into our analyses of predicted pathogenic variant incidence further strengthened the significance level of the difference in high-fraction variants (**Fig. S11a**) between HD and control

lymphoblasts. Including these homoplasmies in the calculation of pathogenic variant dosages did not qualitatively change their associations with HD stages (**Fig. S11b**), clinical phenotypes and *HTT*-related genetic burden (**Fig. S11, c-f)**, indicating that our results regarding predicted pathogenic heteroplasmies in HD lymphoblasts were not affected by the presence of possibly fixed heteroplasmies.

2. Very-low-fraction pathogenic mtDNA heteroplasmies in blood samples

Previous cell studies have indicated that there could be massive variability in mtDNA sequences at a cellular level, even though their fractions at a tissue level were below the detection threshold of mtDNA heteroplasmies (4–7). Therefore, our conservative choice of VAF≥0.5% for heteroplasmy detection in blood samples would lead to an underestimate of heteroplasmy incidence in blood, which might impact our conclusions regarding the changes in heteroplasmies during clinical progression of HD.

To partially address this issue, we lowered the minimum VAFs used for detecting heteroplasmies, from 0.5% to 0.25%, and for defining the share of heteroplasmies between the baseline and follow-up blood samples from 0.2% to 0.1%. As a result, we raised the number of pre-existing pathogenic heteroplasmies detected among the 169 HD patients (**Table 2**) from 72 to 240, and the number of individuals carrying at least one pre-existing pathogenic heteroplasmy from 51 to 118 (**Table S11**). Given that the median sequencing coverage at 6100-fold(X) in blood samples did not guarantee enough power to identify heteroplasmies at very low VAFs (<0.5%), we postulate that at least 70% (118/169) of HD patients may carry predicted pathogenic mtDNA heteroplasmies in the hematopoietic system.

We further verified that the severity of HD clinical phenotypes, assessed by using total functional capacity (TFC) scores, total motor scores and symbol digit modalities test (SDMT) scores at follow-up, was significantly associated with the VAF changes of the 240 pathogenic heteroplasmies (linear mixed-

effects model, P=0.00023-0.0072, **Table S11**), but not with those of other, non-pathogenic heteroplasmies detected in the same individuals. Therefore, prevalent low-fraction or very-low-fraction mtDNA heteroplasmies and their VAF changes at a tissue level, measured by using a refined sequencing method, may be a sensitive biomarker for monitoring mitochondrial quality in somatic cells during disease progression or aging.

3. mtDNA content in lymphoblast and blood samples

After performing quality control on the mtDNA content datasets in lymphoblast and blood samples of the current study (**Supplemental Method 4**), we obtained STAMP-CN measurements in 1252 HD lymphoblasts, 182 control lymphoblasts, and 362 blood samples (**Table S1**), as well as qPCR-CN measurements in 1285 HD lymphoblasts, 116 control lymphoblasts and 318 blood samples, of which 306 involved both baseline and follow-up blood samples from 153 HD patients (**Table S12**).

Since both STAMP-CN and qPCR-CN represent mtDNA content relative to nuclear DNA (nDNA), rather than the actual numbers of mtDNA copies in cells, and rely on different nDNA regions for normalizing read numbers or signal intensity, we performed analyses with both STAMP-CN and qPCR-CN in the current study. mtDNA content was compared between control and HD lymphoblasts, as well as between blood samples collected at baseline and follow-up visits of HD patients.

We found consistently decreased mtDNA content in HD lymphoblasts (logistic regression adjusted for age, sex and sequencing coverage, $P \leq 3.3\times10^{-15}$, **Fig. S12**) for all the disease stages, which recapitulates a decline of oxidative phosphorylation (OXPHOS) and ATP production previously observed in lymphoblasts of HD patients (8). However, unlike heteroplasmies, which showed a disease stage-related increase of pathogenic variant dosages, no significant association was found between

mtDNA content and disease stages in HD lymphoblasts (linear regression adjusted for age, sex and sequencing coverage, $P \geq 0.1$, **Fig. S12**). Moreover, mtDNA content in lymphoblasts did not correlate with the variant dosages of predicted pathogenic heteroplasmies (linear regression, $P \geq 0.22$, **Table S7**) and was unable to account for the pathogenic variant dosage differences between HD lymphoblasts and control lymphoblasts (logistic regression of disease status with additional adjustment for mtDNA content, $P = 0.0011$).

In blood samples of HD patients, we noted a significant increase of mtDNA content (paired t-test, $P \leq 1.3 \times 10^{-5}$, **Fig. S13 a, b**) at follow-up. This longitudinal increase agrees with previous observations of an elevation of mtDNA content in lymphocytes of symptomatic HD patients, compared to control individuals (9). However, in contrast to the VAF changes of pre-existing pathogenic mtDNA heteroplasmies, the degree of the longitudinal increase of mtDNA content in blood did not differ between HD patients with and without a progression of disease stage during the follow-up (t-test, $P \geq 0.48$, **Fig. S13 c, d**). mtDNA content in blood of HD patients was also not associated with the expansion of pre-existing pathogenic heteroplasmies (linear mixed-effects model, $P \geq 0.25$, **Table S8**). Therefore, changes in mtDNA content in lymphoblast and blood samples of HD patients were unlikely to account for the expansion of pathogenic mtDNA heteroplasmies during disease progression.

4. mtDNA haplogroup in lymphoblast and blood samples

We used haplogrep2 (v2.1.1) (10) to determine haplogroups of the major mtDNA sequences in lymphoblast and/or blood samples from 1798 individuals that had median sequencing coverage of consensus reads over 100-fold (X) in the current study (**Table S1**). Consistent with self-reported ethnicity, 98.9% of the 1754 individuals who claimed to have Caucasian ancestry had mtDNA falling

into one of the common mtDNA macro-haplogroups in Europe (**Table S13)**. The most prevalent macro-haplogroup identified was H (frequency=46.2%), followed by the macro-haplogroups U, T, J, K, V, and I, all of which had frequency greater than 2% in the current study (**Table S14**). Among the 18 individuals who self-reported with African, American or Asian ancestry, their mtDNA also belonged to common macro-haplogroups in the corresponding regions (**Table S14**). Moreover, identical haplogroups were identified using mtDNA in either lymphoblast or blood samples from the same individuals (**Table S15)**.

By comparing the frequencies of mtDNA haplogroups between HD patients and control individuals, we did not find that any of the common mtDNA macro-haplogroups were significantly associated with disease status, in relation to either the most prevalent macro-haplogroup H or all the other haplogroups in the current study (logistic regression, $P \geq 0.12$, **Table S14**). The frequencies of mtDNA haplogroups also did not significantly differ between control individuals and HD patients in the prodromal or early stages ($P \geq 0.1$), as well as between control individuals and HD patients in the middle or late stages ($P \geq 0.05$, **Table S14**).

Furthermore, including mtDNA macro-haplogroup as an additional covariate in the analyses of pathogenic mtDNA variant dosages in lymphoblasts (**Fig. 2a**) did not abolish their associations with disease status (logistic regression, $P=0.0013$) as well as advancing disease stages (linear regression, $P=0.0021$), indicating that the observed associations of pathogenic mtDNA heteroplasmies with HD were not modulated by common genetic variations in mtDNA at a population level.

Supplemental Methods

1. Read duplicates identification and consensus sequence calling

Due to biased PCR amplification of mitochondrial genome (mtDNA) fragments (11, 12), retaining read duplicates in calling variants may affect the estimation of the variant allele fractions (VAF) of mtDNA heteroplasmies. To resolve this issue, we relied on the molecular barcode in the ligation probe of STAMP (**s**equencing by **t**argeted **a**mplification of **m**ultiplex **p**robes) to infer the identity of the capturing event, and to filter out duplicate paired-end reads that were produced from the same mtDNA fragment. Among paired-end reads with the same molecular barcode, instead of choosing the read pairs with the highest overall quality to represent the sequence of the DNA fragment (the consensus read sequence), we used a Bayesian approach to call each nucleotide and its associated quality score in the consensus read by incorporating information from all reads (13). The posterior probability of having a nucleotide, such as "A", at a certain position in the consensus read sequence can be represented using the equation below,

$$P(A|\text{all reads}) = \frac{\prod_{i=1}^{n} P(\text{read}_i|A) \times P(A)}{\sum_{NT} \prod_{i=1}^{n} P(\text{read}_i|NT) \times P(NT)}$$

, where $P(NT)$ is prior probability and $\prod_{i=1}^{n} P(\text{read}_i|NT)$ is the estimated likelihood under the assumption that all paired-end reads in a read family are independent. To simplify calculation, we used equal prior probability for all nucleotides. The likelihood of a nucleotide in each read can be approximated using the base quality score as

$$P(read_i|NT) = \begin{cases} 1\text{-}10^{-\frac{BAQ}{10}}, & NT="A" \\ \frac{1}{3} \times 10^{-\frac{BAQ}{10}}, & NT\neq"A" \end{cases}$$

We took the nucleotide with the highest posterior probability (P_{max}) to construct the consensus read and assigned a quality to this nucleotide by using the phred score of its probability as $-10\log_{10}(1-P_{max})$. The quality scores of the consensus read were rounded to the nearest integers in a bam file with ASCII characters from 33 to 126. So the maximum phred quality score of a nucleotide is 93, representing an error rate of $<10^{-9}$. By using this statistical framework to determine consensus read sequences, we estimated the rates of variant alleles that remained after removing bases with BAQ (base alignment quality) under 30. These variant alleles might reflect very-low-fraction mtDNA heteroplasmies, as well as errors introduced in the capturing step or during the early rounds of PCR amplification of STAMP (13). We found that the variant alleles occurred at rates of about 0.01% and 0.03% per base of the consensus reads, constructed with and without duplicate paired-end reads respectively, in lymphoblast and blood samples of the current study (**Fig. S2a**). Overall, the rates of variant alleles among all consensus reads were estimated to be 0.02% per base, which set an upper bound for the background error rate of STAMP in the current study. Assuming an error rate of 0.02% for consensus reads, STAMP has over 99.5% power to identify true variants, such as mtDNA heteroplasmies, at VAF of 1%, 0.5% and 0.2%, with over 1000, 3000 and 10000 consensus reads, respectively (**Fig. S2b**).

2. Nuclear mitochondrial DNA segment (NUMTS) identification and filtering

After trimming the barcode and arm sequences of mtDNA probes from the paired-end reads, we aligned these reads by using "bwa mem" (v0.7.17) (14) to the latest human reference genome containing both nuclear DNA (nDNA, genome assembly GRCh38) and mtDNA (Revised Cambridge Reference Sequence, rCRS) sequences (downloaded from ftp://ftp.1000genomes.ebi.ac.uk). Paired-end reads that were aligned to nDNA with MAPQ (mapping quality) ≥10 were marked as potential NUMTS in the alignment file (**Fig. S10**). After determining consensus read sequences, we compared all consensus read

sequences to a collection of known NUMTS sequences in the reference genome. This collection of NUMTS sequences comprises those obtained from BLASTN search of the 46 mtDNA segments captured by mtDNA probes in STAMP, as well as variants of those sequences harboring common polymorphisms identified in the 1000 Genomes project (minor allele frequency≥1%, retrieved from the *commonSNP147* track from the UCSC genome browser). A consensus read was marked as potential NUMTS if it had a lower pairwise edit distance to NUMTS sequences than to the sample's major mtDNA sequence (fraction>50%), or it was constructed from paired-end reads already annotated as NUMTS according to the BWA alignment (**Fig. S10**) (13).

We found that 90.4% and 97.5% of the NUMTS annotations for the consensus reads agreed between the alignment method and the sequence distance method in blood samples and lymphoblast samples, respectively, indicating that most NUMTS captured by mtDNA probes were correctly aligned to nDNA as per design (**Fig. S14c**). The average proportion of NUMTS reads among consensus reads was low, at about 0.37% in lymphoblasts and about 2.0% in blood samples, mirroring the difference in mtDNA content between these two types of samples. The nDNA origin of these NUMTS reads was further revealed by the significant negative correlations between their proportions among consensus reads and the relative mtDNA content measured in lymphoblasts and blood samples (r=-0.4 and -0.71, $P \leq 1.5\times10^{-56}$; **Fig. S14 e** and **f**).

Since not all NUMTS have been successfully annotated in the latest human reference genome (assembly GRCh38) and some NUMTS are even polymorphic at a population level, misclassification of reads derived from these sequences as mtDNA reads may lead to excessive low-fraction heteroplasmies called at certain mtDNA sites (15). To address this issue, we searched for consensus reads that contain alleles that were shared between samples with distinct major mtDNA sequences (**Fig. S14a**). Such evidence may suggest that these minor alleles of consensus reads co-segregate with individuals' nDNA

instead of their mtDNA. Given that blood samples had greater proportions of NUMTS reads and higher sequencing coverage than lymphoblast samples in the current study, we relied on the consensus reads from the 362 blood samples in this analysis. Moreover, we counted alleles by using only consensus reads constructed from duplicate paired-end reads and bases with BAQ≥30, in order to reduce the influence of sequencing errors.

As a result, we found 52 minor alleles of the consensus reads which occurred ≥5 times and had a fraction ≥0.5% (similar to the criteria for heteroplasmy calling) in at least 7 out of the 362 blood samples (a ratio similar to a minor allele frequency of 1% for nDNA; **Fig. S14b** and **Table S16**). As expected, NUMTs already annotated in the human genome assembly GRCh38 contributed to 38 of the 52 minor alleles (72%). Another 11 (21%) overlapped with previously reported polymorphic NUMTS (15, 16), indicating that these reads might be improperly aligned to mtDNA due to incomplete information in the reference genome. The remaining 3, captured by EL probes A6/A7, A11, A12, contained three discriminating substitutions at mtDNA sites nt2623, nt4216, and nt4560, which were detected in less than 20% of the blood samples (**Fig. S14b)**. All of these substitutions were transition changes and overlapped with non-pathogenic polymorphisms in the coding region of mtDNA, which could be due to recent insertion of mtDNA fragments into the human genome. We then added these 52 sequences into the collection of NUMTS and used it to annotate consensus reads in the alignment file for potential NUMTS again.

Finally, we examined the influence of low- and rare-frequency NUMTS on heteroplasmy identification in the current study. If a large proportion of NUMTS reads remained in our data set after we filtered out those with a population frequency≥1% in the blood samples, it would give rise to spurious heteroplasmies, with a predominant mutation pattern resembling that of nDNA, as well as a negative correlation with mtDNA content. Both signatures were revealed in the analyses of known

NUMTS in samples of the current study (**Fig. S14 d-f**). However, these two signatures did not stand out in the analyses of heteroplasmies in both lymphoblasts and blood samples: we found that the heteroplasmies detected in the current study showed a characteristic pattern of mtDNA replication errors with high transition to transversion ratios (**Fig. S7**), in contrast to the mutation pattern of NUMTs in nDNA (**Fig. S14d**). The overall incidence and dosages of mtDNA heteroplasmies detected in lymphoblasts were not significantly associated with mtDNA content measured in the same samples (linear regression adjusted for age, sex, and STAMP sequencing coverage, $P>0.12$). For pre-existing mtDNA heteroplasmies detected in longitudinal blood samples, their fraction changes did not significantly correlate with mtDNA content measured in the baseline and follow-up samples (linear mixed-effects model with adjustment for baseline age, sex and years of the follow-up, $P>0.12$). Taken together, our results indicate that the mtDNA-targeted design of probes in STAMP and the computational methods for NUMTS identification can correct for the contamination of reads derived from NUMTS in the current study.

3. mtDNA variant detection and quality control procedures

Reliable detection of heteroplasmies in mtDNA is dependent on numerous factors, such as the correctness of read alignment, the depth of sequencing coverage, and the rates of sequencing and PCR amplification errors. In the current study, we applied a series of quality filters on the raw reads of STAMP to reduce false positive errors in calling mtDNA variants that may arise from each of these factors (listed in **Table S10** and discussed below).

For proper read alignment, we required that the raw paired-end reads should possess a high-quality molecular barcode (at least 9 bases with BAQ ≥15) and appropriate DNA sequences from one of the 46

mtDNA probe pairs in the complementary regions of the ligation and extension arms (a maximum of 3 mismatched bases allowed), and DNA sequences in between should be aligned to the target mtDNA region indicated by the probe pair with bwa (v0.7.17) (14) (MAPQ≥20). Reads passing these quality filters were then subject to local realignment and base quality recalibration by using freebayes (v1.1.0)(17) and samtools (v1.6) (18), respectively, before calling consensus read sequences. Reads from NUMTS and reads showing an excess of mismatches (>5 in the coding region and >8 in the D-loop region) compared to the individual's major (fraction>50%) mtDNA sequence were excluded from variant calling (13).

To ensure a sufficient number of reads for calling heteroplasmies, we required that mtDNA heteroplasmies be assessed by using only high-quality consensus reads (BAQ≥30) and at sites with 100X depth of coverage. If a site had an excess of low-quality consensus reads (>30% of reads with BAQ<30), which could be attributed to sequencing errors or being close to an indel, this site was excluded from heteroplasmy calling as well. We also excluded sites in the low-complexity regions of mtDNA (nt 302-316, nt 512-526, nt 16184-16193) and another 6 sites in the D-loop region (nt 545, 16224, 16244, 16249, 16255, and 16263) which showed consistently low quality in heteroplasmy calling (average percentage of low-quality reads >30% before quality filtering) from analysis. At a sample level, we required the median sequencing coverage on mtDNA to be over 1000X depth. According to the relative sequencing coverage on the mtDNA region captured by each of the 46 probe pairs in STAMP (**Fig. S15**), this criterion guaranteed that most mtDNA sites would have ≥500X depth of coverage, necessary for calling a variant of VAF at 1% with at least 5 variant alleles.

To reduce false positive variants associated with sequencing errors, we required that the variant allele of a heteroplasmy had to be observed at least 5 times among consensus reads. A log likelihood ratio test, computed with the base quality scores of the consensus reads, was used to evaluate variant

quality (19). Only heteroplasmies with a log likelihood quality score greater than 5 were retained. We further excluded heteroplasmies that exhibited significant differences in the VAFs (Fisher's exact test $P<10^{-4}$ or fold change>5) estimated when using the consensus reads constructed with and without duplicate paired-end reads, which reflect PCR or sequencing errors not predicted by the quality scores. Since the statistical power to detect such a difference may be low for low-fraction heteroplasmies, we required that for heteroplasmies of VAF<1%, a VAF of ≥0.2% had to be detected among the consensus reads constructed with duplicate paired-end reads.

Finally, an exact *Poisson* test was used to estimate the probability of observing more errors than the number of variant alleles at a site. The expected number of errors was computed based on the sequencing depth, and the background rate of errors that could occur in the capturing step or during the early rounds of PCR amplification in STAMP (ε=0.02%, **Fig. S2a**) (13). Heteroplasmies with a *P*oisson test probability ≥ 6×10^{-7} (0.01/16569) were excluded. This criterion was particularly effective in filtering out very-low-fraction errors (i.e. VAF<0.5%) but could be too conservative for the identification of true variants of VAF in this range. Thus, we relaxed the cutoff value of this probability from 6×10^{-7} to 0.001, to distinguish very-low-fraction variants from errors when assessing heteroplasmies at certain sites instead of all sites of mtDNA. For example, we used this relaxed cutoff to determine whether a heteroplasmy could be detected in a sequencing replicate or in another sample, such as the baseline or follow-up blood sample, of the same individual.

The complete list of quality control filters and the associated parameters used is given in **Table S10**. In a pilot study, we applied STAMP and the aforementioned quality control filters to detecting artificial mtDNA variants in the sample mixtures created by combining genomic DNA from two lymphoblast samples at varying relative ratios (from 1:199 (0.5%) to 1:1(50%)) (13). We found that all variants detected at VAF≥0.25% were either at the 58 polymorphic single nucleotide sites, or at the 5 pre-

existing heteroplasmic sites that differed between these two lymphoblast samples (13). High correlations were also found between the fractions of these variants and the ratios of the genomic DNA used to create these sample mixtures (r>0.9, P<0.0046) (13). Of the low-fraction variants at the 58 polymorphic sites in the sample mixtures, we were able to call 84% and 97% by using either the value of the DNA ratio or 50% of this value as the cutoff for the minimum VAF, respectively. Therefore, the overall sensitivity in identifying mtDNA heteroplasmies of VAF≥0.5% was at least 97% if detected at VAF≥0.25%, while the corresponding false positive rate would be well under 10^{-4} per site of mtDNA (<1/16569) (13).

Of the 1731 lymphoblasts that passed quality control for heteroplasmy analysis in the current study, 17 were sequenced for a second time with a median mtDNA sequencing coverage over 1000X in a separate HiSeq run. We found that 89 (98.9%) of the 90 heteroplasmies identified with VAF≥1% in these samples were detectable in their counterparts at VAF≥0.2%, 86 (95.6%) of which had VAF≥0.5% and fulfilled all quality control requirements for heteroplasmy analysis (**Table S3**). The heteroplasmies detected between the 17 samples and their replicates also exhibited a high correlation in their VAFs (r>0.97, **Fig. S3 a** and **c**). The average coefficient of variation between repeated measurements of lymphoblast mtDNA was 17% for VAFs of all heteroplasmies and was 5% for VAFs of medium-to-high-fraction heteroplasmies (VAF≥5%).

Of the 362 blood samples used for heteroplasmy analysis in the current study, 320 were sequenced at least twice with STAMP at a median sequencing coverage over 1000X. By comparing results from two independent sequencing runs performed on the same blood sample (**Table S3**), we found that of the heteroplasmies with VAF≥1% detected in one sample, 97% (448/462) were present in the replicate at VAF≥0.2%, and 89.6% (414/462) had VAF≥0.5% and passed all quality filters in the replicate. The percentage of heteroplasmies that met these two criteria increased to 93.2% when we focused on sites

that had ≥1000 consensus reads in both samples. Similar results were found for heteroplasmies with VAF≥0.5% in blood samples (**Table S3**). Of these heteroplasmies, 87.3% (541/620) and 93.6% (396/423) were identified at VAF≥0.25% and survived quality control filtering in the replicate, when analyzing sites with ≥1500 and ≥2000 consensus reads in both samples, respectively. Over 95.3% were present at VAF≥0.2% in the replicate.

The average coefficient of variation for VAFs of heteroplasmies in repeated measurements was about 20%-23% among all blood samples (**Table S3**) and was 13% among blood samples with a high median mtDNA sequencing coverage at >3500X (N=29 pairs, average median depth of coverage = 5898X). This indicates that increasing read depth can improve accuracy and reliability in detecting low-fraction heteroplasmies and their VAFs. As such, in the current study, we combined sequencing reads from independent STAMP sequencing runs of the same blood sample to call variants. The average median sequencing coverage on mtDNA in blood samples was about 6100X, and 85% had sequencing coverage >3500X. Moreover, the VAFs of the identified heteroplasmies showed high correlations (r>0.96, **Fig. S3 b** and **d**) and nonsignificant differences (Wilcoxon signed rank test, P≥0.13, **Table S3**) between the sequencing replicates of the blood samples. In conclusion, STAMP has low false positive and false negative error rates in calling heteroplasmies of VAF≥0.5%-1% and can assess VAFs and VAF changes of heteroplasmies in mtDNA in the current study.

4. mtDNA content quantification and quality control procedures

Alignment of paired-end reads and identification of consensus reads for the 5 nuclear DNA targets (**Table S1**) were performed as those for the 46 mtDNA targets (13). Of them, two failed to produce sufficient reads (≥5X) in both lymphoblast and blood samples and also had relatively low correlations

with mtDNA reads compared to the other 3 regions (**Table S17**)**.** As such, we focused on the reads from the 3 nDNA regions on chromosomes 8, 14 and 19 for further analysis (13). We calculated mtDNA content in STAMP (hereafter referred to as STAMP-CN) by using a log ratio between the average number of consensus reads from mtDNA and that from nDNA as: **STAMP-CN**=$\log_2$(No. of mtDNA consensus reads)–C×$\log_2$(No. of nDNA consensus reads), where C is the normalization factor for nDNA consensus reads (13). We estimated C by using the coefficient β from the regression of $\log_2$(No. of mtDNA consensus reads) against $\log_2$(No. of nDNA consensus reads) among samples of the current study. The resulting normalization coefficients were 0.53 for lymphoblasts and 0.66 for blood samples (**Table S17**). By using this approach, we computed pair-wise correlations of STAMP-CN between replicates to be at r = 0.82-0.83 ($P \leq 0.00014$), including 15 pairs of lymphoblast samples (2 excluded for no nDNA coverage) and 320 pairs of blood samples (**Fig. S16 a, b**) measured on separate sequencing runs.

Next, we compared STAMP-CN to mtDNA content measured by using a commercially available quantitative PCR-based assay (hereafter referred to as qPCR-CN)(20). The PCR reactions were performed as per manufacturer's instructions (The Detroit R&D, Inc.) with 15ng total genomic DNA, mtDNA or nDNA target primers, and SYBR green PCR master mix, for 1943 samples, which passed quality control for heteroplasmy analysis and still had enough genomic DNA, after STAMP sequencing. The PCR thermal conditions were 10 min at 95°C, followed by 40 cycles of 15 sec at 95°C and 60 sec at 60°C. The resulting C_T values were averaged over the experimental duplicates for mtDNA and nDNA targets, respectively, from a total of 4 PCRs in each sample. The differences between the C_T values of mtDNA and nDNA targets were then normalized to that of a positive control sample (genomic DNA from human MCF10A cells provided by the manufacturer), measured on the same 96-well plate by using the $\Delta\Delta C_T$ method to obtain qPCR-CN (20).

After performing quality filtering on samples that failed in any of the 4 PCRs and/or showed a difference of over 3 cycles in the C_T values between the experimental duplicates, 1401 (88%) lymphoblast samples and 318 (93%) blood samples were retained for the analysis of qPCR-CN. Of these samples, 1552 had STAMP-CN available for comparison (**Table S12**). We found significant positive correlations between STAMP-CN and qPCR-CN at $r = 0.39$ and 0.69 ($P \leq 7.3\times10^{-46}$, **Fig. S16 c, d**) in lymphoblast samples and blood samples, respectively. These values are close to the corresponding results reported in a comparative study on mtDNA content measured with sequencing-based methods and qPCR-based methods (21).

5. Statistical modeling of age-dependent changes of mtDNA heteroplasmies in lymphoblasts

To improve normality of the residuals in the linear models, we normalized the variant dosages and incidence of mtDNA heteroplasmies in lymphoblasts with rank-based inverse normal transformation (INV) and adjusted the resulting values for sex and mtDNA sequencing coverages. This transformation of the raw values of heteroplasmies to a probability scale with z scores in INV also enabled comparisons of the changes of mtDNA heteroplasmies in various functional categories by using standardized effects, regardless of the differences in their initial distributions.

We excluded from analysis six lymphoblasts that came from juvenile HD patients aged under 20 to avoid the influence of extremely early disease onset on the age-dependent changes of mtDNA heteroplasmies in HD. We first evaluated the associations of variant incidence and dosages of mtDNA heteroplasmies with age and CAG repeat length in HD lymphoblasts, by fitting a linear model where the effects of age, CAG-repeat length and their interaction were considered:

$$\mathbf{y \sim \beta_1 \times age + \beta_2 \times cag + \beta_3 \times age \times cag + \alpha} \quad \textbf{(1)}$$

β_1 and β_2 in the above model represent the effects of age and CAG repeat length on y.β_3 is the effect of the multiplicative interaction independent of β_1 and β_2, and α is the intercept. The values of age and CAG repeat length were centered at the population mean. The resulting effect coefficients, and the associated significance levels of the effects computed using the Wald test, are reported in **Table 1.**

In addition, we added the squared terms of age and CAG repeat length into **model (1)** as:

$$\mathbf{y \sim \beta_1 \times age + \beta_2 \times cag + \beta_3 \times age \times cag + \beta_4 \times age^2 + \beta_5 \times age^2 \times cag + \alpha} \quad \mathbf{(2)}$$

$$\mathbf{y \sim \beta_1 \times age + \beta_2 \times cag + \beta_3 \times age \times cag + \beta_6 \times cag^2 + \beta_7 \times age \times cag^2 + \alpha} \quad \mathbf{(3)}$$

$$\mathbf{y \sim \beta_1 \times age + \beta_2 \times cag + \beta_3 \times age \times cag + \beta_4 \times age^2 + \beta_6 \times cag^2 + \beta_9 \times age^2 \times cag^2 + \alpha} \quad \mathbf{(4)}$$

to examine whether there are nonlinear relationships between mtDNA heteroplasmies and age or CAG repeat length. We found that all three **models (2-4)** did not consistently improve predictive ability relative to the base model (versus model 1, F test, P>0.11). As such, we only report the results from **model (1)** in the main text and in **Table 1**.

We further assessed the effects of age on the changes of heteroplasmies for control lymphoblasts by fitting a linear model of the form:

$$\mathbf{y \sim (\beta_{control} + \beta_{hd}) \times age + \alpha_{control} + \alpha_{hd}} \quad \mathbf{(5)}$$

, where y is the normalized variant dosages or incidence, $\beta_{control}$ is the effect of age in control lymphoblasts, and β_{hd} is the effect of age in HD lymphoblasts relative to that of the controls. $\alpha_{control}$ and α_{hd} in the above model are the values of y at birth in the controls and in HD lymphoblasts relative to that of the controls, respectively. The significance levels of $\beta_{control}$ for control lymphoblasts computed using the Wald test for linear regression are listed in **Table S4**.

Supplemental References

1. H. A. Coller, *et al.*, High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. *Nat. Genet.* **28**, 147–50 (2001).
2. P. F. Chinnery, D. C. Samuels, Relaxed Replication of mtDNA: A Model with Implications for the Expression of Disease. *Am. J. Hum. Genet.* **64**, 1158–1165 (1999).
3. J. L. Elson, D. C. Samuels, D. M. Turnbull, P. F. Chinnery, Random Intracellular Drift Explains the Clonal Expansion of Mitochondrial DNA Mutations with Age. *Am. J. Hum. Genet.* **68**, 802–806 (2001).
4. E. Kang, *et al.*, Age-related accumulation of somatic mitochondrial DNA mutations in adult-derived human iPSCs. *Cell Stem Cell* **18**, 625–36 (2016).
5. T.-G. Yao, S. Kajigaya, N. S. Young, Mitochondrial DNA mutations in single human blood cells. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **779**, 68–77 (2015).
6. J. Morris, *et al.*, Pervasive within-Mitochondrion Single-Nucleotide Variant Heteroplasmy as Revealed by Single-Mitochondrion Sequencing. *Cell Rep.* **21**, 2706–2713 (2017).
7. L. S. Ludwig, *et al.*, Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339.e22 (2019).
8. I. S. Seong, *et al.*, HD CAG repeat implicates a dominant property of huntingtin in mitochondrial energy metabolism. *Hum. Mol. Genet.* **14**, 2871–2880 (2005).
9. P. Jędrak, *et al.*, Mitochondrial DNA levels in Huntington disease leukocytes and dermal fibroblasts. *Metab. Brain Dis.* **32**, 1237–1247 (2017).
10. H. Weissensteiner, *et al.*, HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.*, gkw233 (2016).
11. J. A. Casbon, R. J. Osborne, S. Brenner, C. P. Lichtenstein, A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* **39**, e81 (2011).
12. D. Aird, *et al.*, Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
13. X. Guo, Y. Wang, R. Zhang, Z. Gu, STAMP: a multiplex sequencing method for simultaneous evaluation of mitochondrial DNA heteroplasmies and content. *NAR Genomics Bioinforma.* **2**, lqaa065 (2020).
14. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–95 (2010).
15. G. Dayama, S. B. Emery, J. M. Kidd, R. E. Mills, The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Res.* **42**, 12640–12649 (2014).
16. R. S. Just, J. A. Irwin, W. Parson, Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. *Forensic Sci. Int. Genet.* **18**, 131–139 (2015).
17. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing (2012). arXiv:1207.3907 [q-bio.GN].
18. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9

(2009).

19. K. Ye, J. Lu, F. Ma, A. Keinan, Z. Gu, Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 10654–9 (2014).

20. J. M. Santos, S. Tewari, A. F. X. Goldberg, R. A. Kowluru, Mitochondrial biogenesis and the development of diabetic retinopathy. *Free Radic. Biol. Med.* **51**, 1849–60 (2011).

21. P. Zhang, *et al.*, Estimating relative mitochondrial DNA copy number using high throughput sequencing data. *Genomics* **109**, 457–462 (2017).

Supplemental Figures

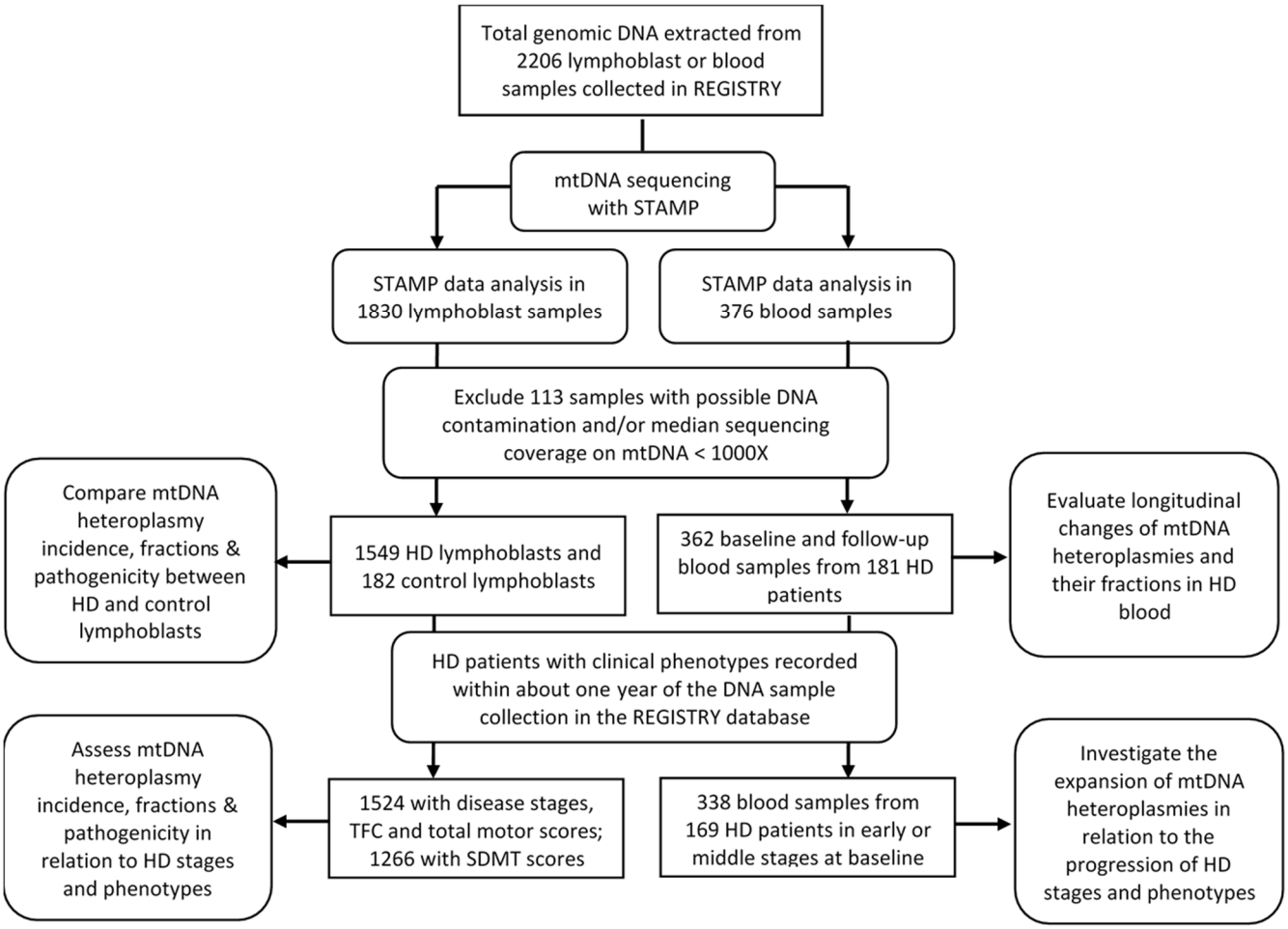


Figure S1. Study flow chart. This study flow chart summarizes the lymphoblast and blood samples of REGISTRY used for mtDNA analyses and the related study aims.

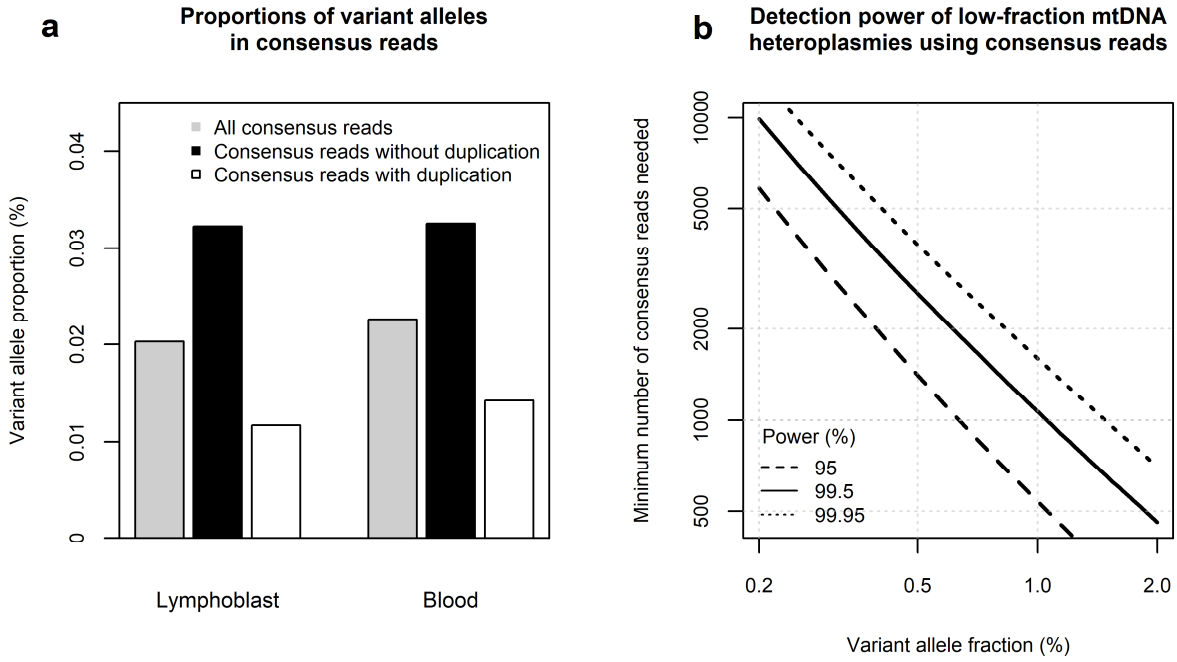


Figure S2. Detection power of low-fraction mtDNA variants using STAMP. (**a**) The proportions of variant alleles per base in the consensus reads used for detecting mtDNA variants that were constructed with and without duplicate paired-end reads in the samples of the current study. (**b**) The numbers of consensus reads needed to achieve 95%, 99.5% and 99.95% power to discriminate real variants from sequencing errors at 16,569 sites of mtDNA, assuming an average error rate of 0.02% per base in the samples of the current study. The results were obtained from one-tailed power calculation for one sample proportion.

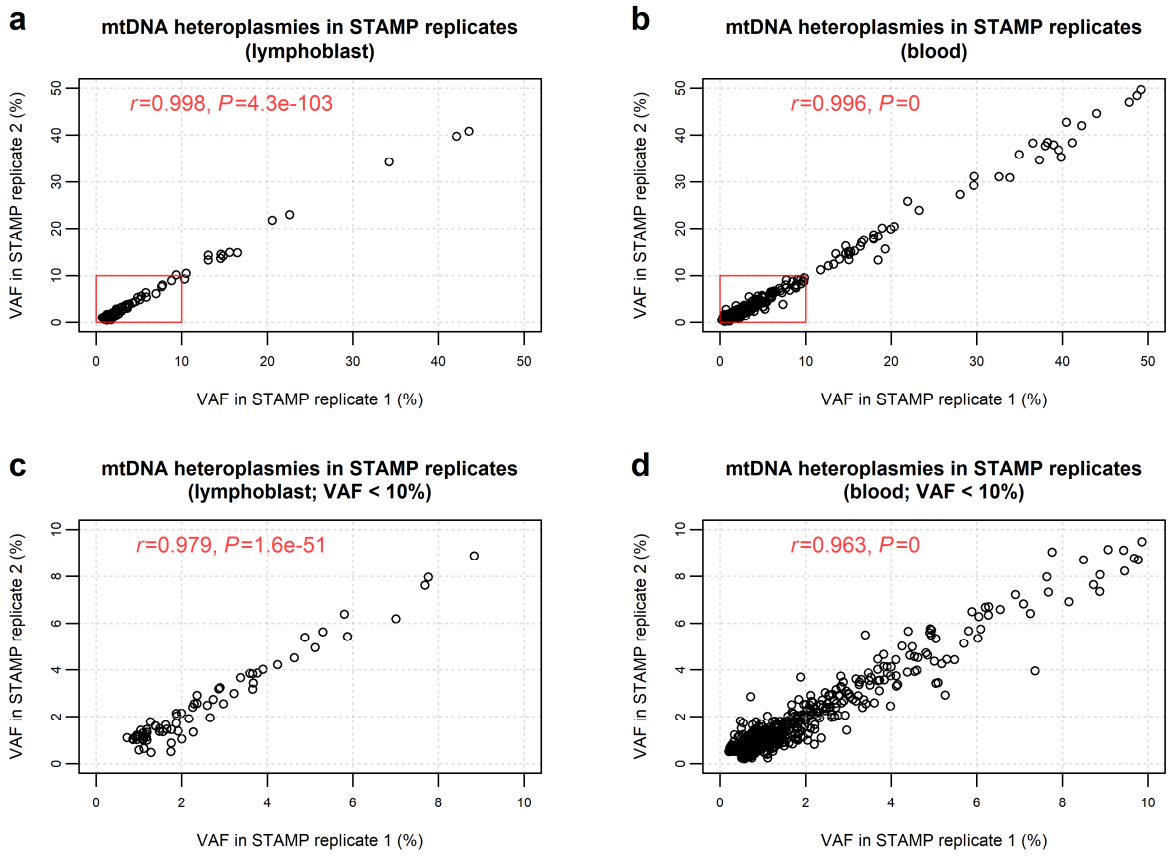


Figure S3. Comparisons of mtDNA heteroplasmies detected in sample replicates. The variant allele fractions (VAF) of mtDNA heteroplasmies detected in replicates of STAMP sequencing performed on 17 lymphoblast samples and 320 blood samples were depicted in panel **a** and panel **b**, respectively. The lower panels **c** and **d** show the VAFs of the heteroplasmies in the red boxes of **a** and **b**, respectively. The minimum VAF used for heteroplasmy detection was 1% in lymphoblast samples and 0.5% in blood samples. The reference allele used to compute VAFs in a pair of sample replicates was the major allele identified in replicate 1.

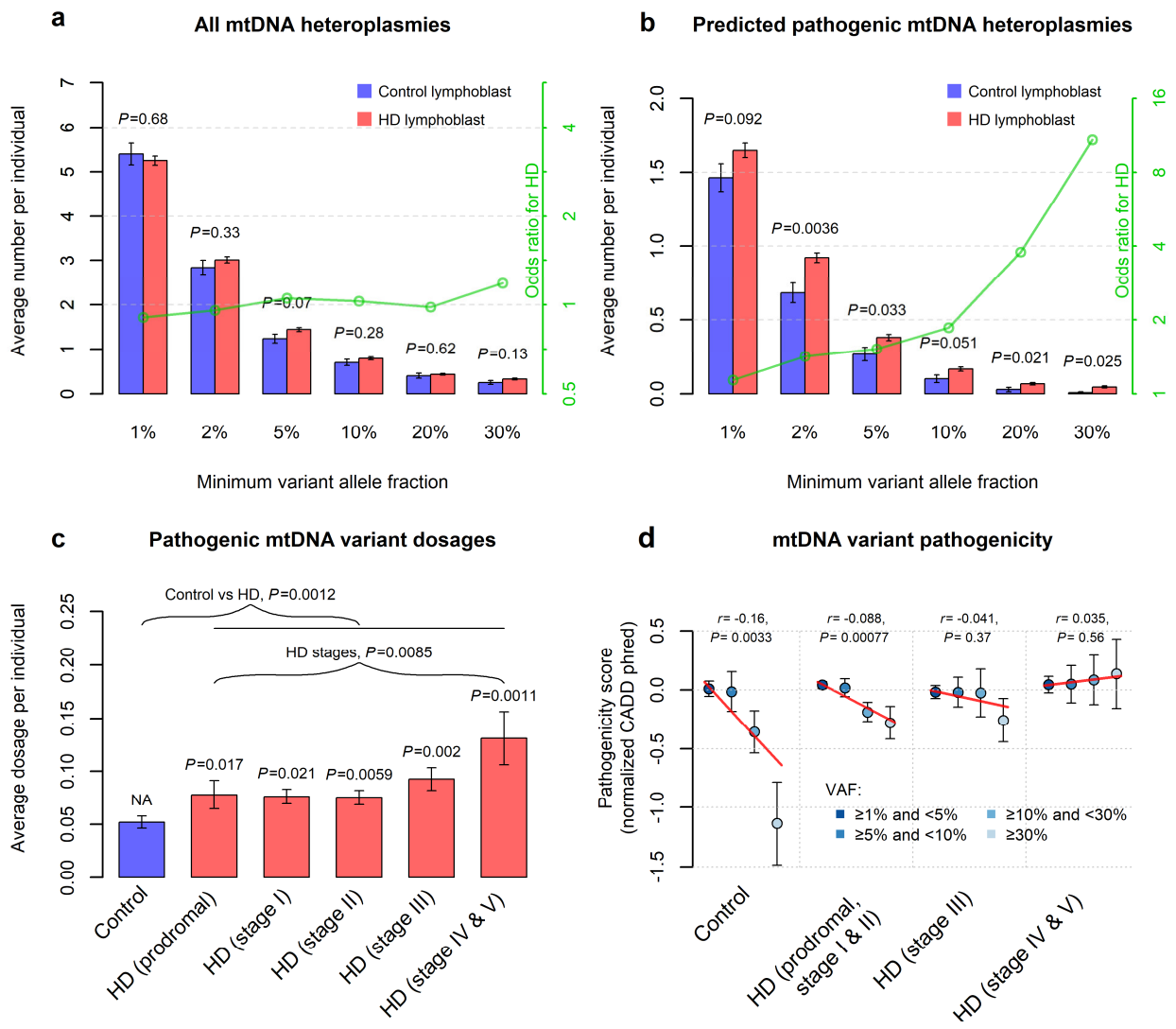


Figure S4. Analysis of mtDNA heteroplasmies in lymphoblasts of young and middle-aged individuals. Individuals aged under 55 years old were used in the analyses. The results are shown in panel **a** for the incidence of all mtDNA heteroplasmies, in panel **b** for the incidence of predicted pathogenic mtDNA heteroplasmies, in panel **c** for the variant dosages of predicted pathogenic mtDNA heteroplasmies, and in panel **d** for the pathogenicity of nonsynonymous heteroplasmies stratified by their variant allele fractions (VAF) in lymphoblasts. The values on the x axes in panels **a** and **b** refer to the minimum VAFs of the heteroplasmies used in the analyses. The *P* values for mtDNA heteroplasmies from the logistic regression analyses of the disease status are shown above the bars in **a** and **b**, with the effects indicated by the green lines and the values on the green y axes on a logarithmic scale. The *P* values for mtDNA variant dosages from the logistic regression analyses of disease status are indicated above the bars representing the corresponding HD stages in panel **c**. NA: not applicable. The Pearson's *r* between the heteroplasmic VAF and the pathogenicity score, as well as the corresponding *P* value, are shown in each panel of **d**. The CADD scores in panel **d** are shown with the inverse normal transformed values.

The red lines in panel **d** represent the fitted regression lines for the VAF categories and the pathogenicity scores. Error bars in all panels represent SEM.

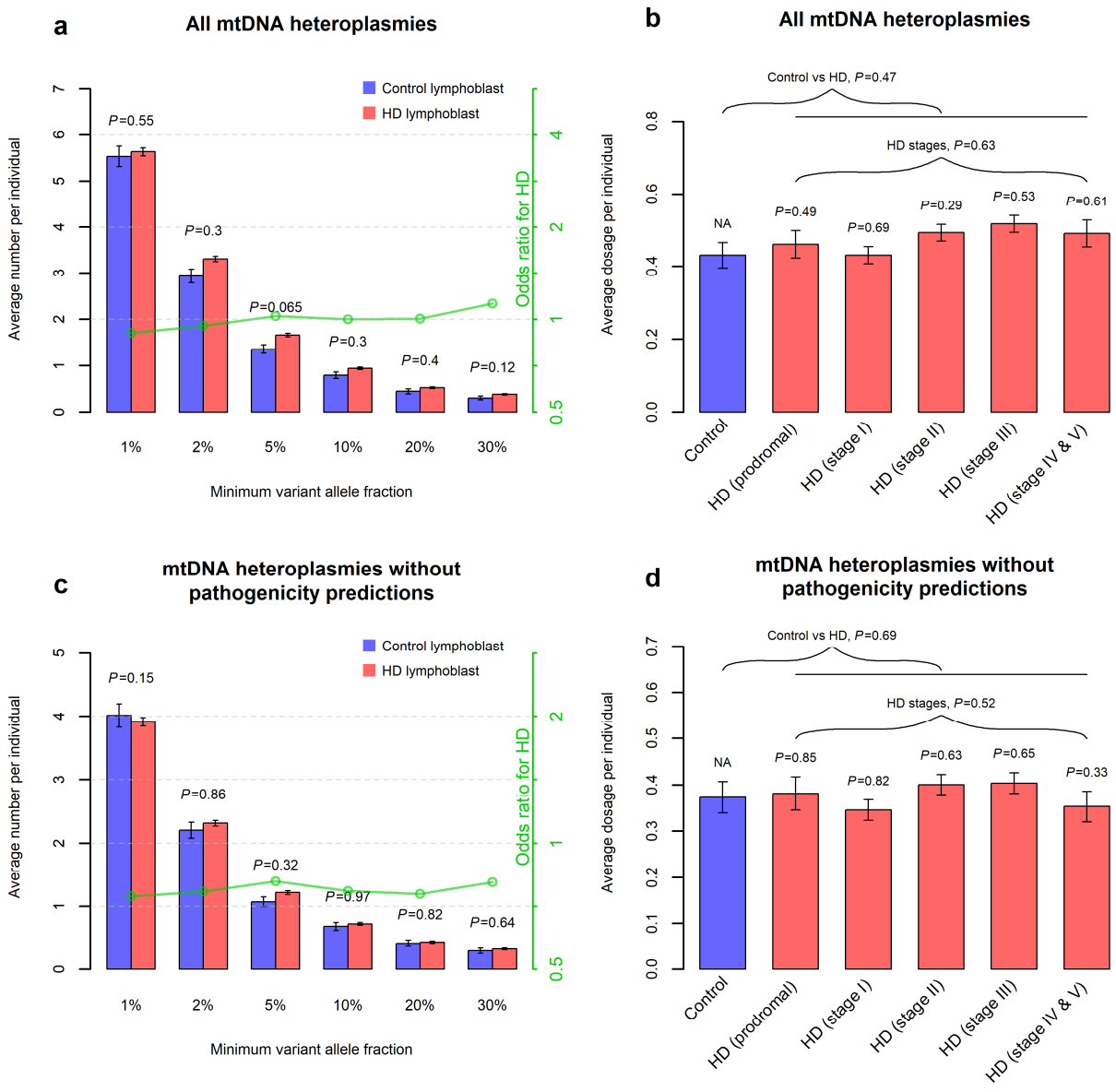


Figure S5. Associations of HD with mtDNA heteroplasmies not predicted with medium or high pathogenicity. (**a**, **b**) All heteroplasmies were used to compute (**a**) variant incidence and (**b**) variant dosages in HD and control lymphoblasts. (**c**, **d**) Only mtDNA heteroplasmies not predicted with medium or high pathogenicity were used to compute (**c**) variant incidence and (**d**) variant dosages. The *P* values for mtDNA variant incidence from the logistic regression analyses of the disease status are shown above the bars in panels **a** and **c,** with the effects indicated by the green lines and the values on the green y axes on a logarithmic scale. The *P* values for mtDNA variant dosages from the logistic regression analyses of disease status are indicated above the bars representing the corresponding HD stages in panels **b** and **d**. NA: not applicable. Error bars in all panels represent SEM.

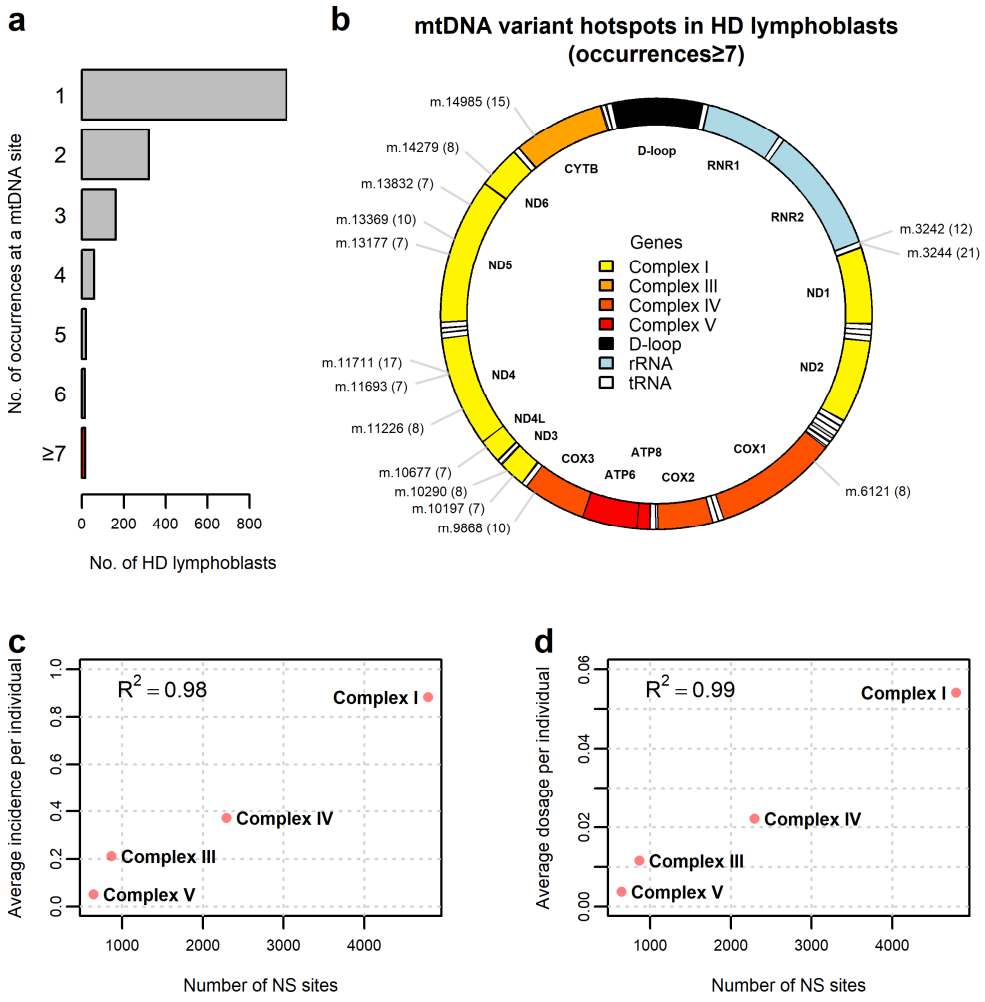


Figure S6. Predicted pathogenic heteroplasmies in HD lymphoblasts. (**a**) Histogram for the number of occurrences of predicted pathogenic heteroplasmies detected at each mtDNA site in HD lymphoblasts. (**b**) The locations of the 15 variant hotspots (the number of occurrences of heteroplasmies ≥7 among HD lymphoblasts) in mtDNA. The number of occurrences of heteroplasmies is indicated in parentheses. (**c**, **d**) Correlations of average (**c**) variant incidence and (**d**) variant dosage of pathogenic heteroplasmies in four oxidative phosphorylation (OXPHOS) complexes with the number of nonsynonymous (NS) sites in genes of OXPHOS complexes in mtDNA. R^2 from linear regression is shown in panels **c** and **d**.

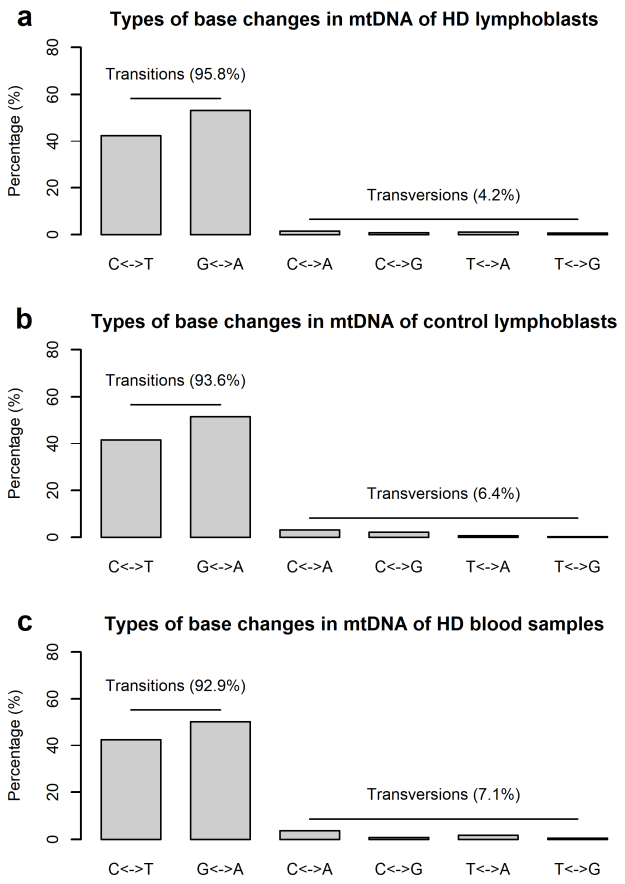


Figure S7. Base changes of mtDNA heteroplasmies detected in lymphoblasts and blood samples. The proportions of different types of base changes are shown for the heteroplasmies detected in lymphoblasts of (**a**) HD patients and (**b**) control individuals, and in (**c**) blood samples of HD patients.

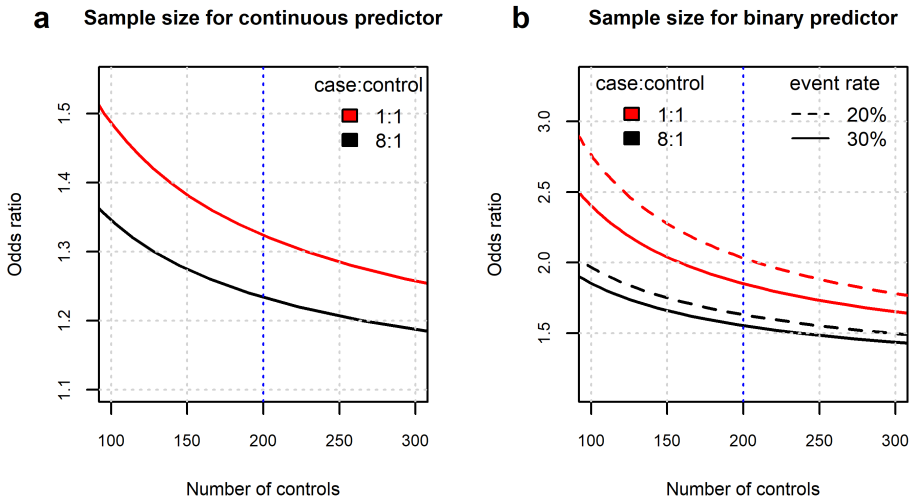


Figure S8. Samples size calculation. We computed sample sizes for logistic regression where we treated mtDNA heteroplasmies as a continuous predictor in panel **a**, and as a binary predictor in panel **b**. In both panels, each line shows the number of controls that guarantees a statistical power of 80% (at alpha=0.05) to identify an effect at odds ratio (OR) indicated by the value on the y axis. Black lines represent the tests involving all samples with a case-to-control ratio of 8 (=1630/200). Red lines refer to the tests with equal numbers of cases and controls, such as those comparing heteroplasmies between controls and HD samples from patients in one disease stage. In panel a, we considered heteroplasmies as a continuous predictor to assess the effect of overall heteroplasmies, including low-fraction ones. In panel b, solid lines and dashed lines represent the tests with the binary predictor having an event rate of 30% and 20%, respectively, which correspond to that of pathogenic heteroplasmies of medium-to-high fraction (VAF≥5%) identified in lymphoblasts. Blue, vertical lines indicate a sample size of 200 for the control group. Therefore, the sample sizes in the current study would guarantee enough statistical power to reveal a moderate increase (OR=~1.3-2.0) in mtDNA heteroplasmies in HD samples relative to controls.

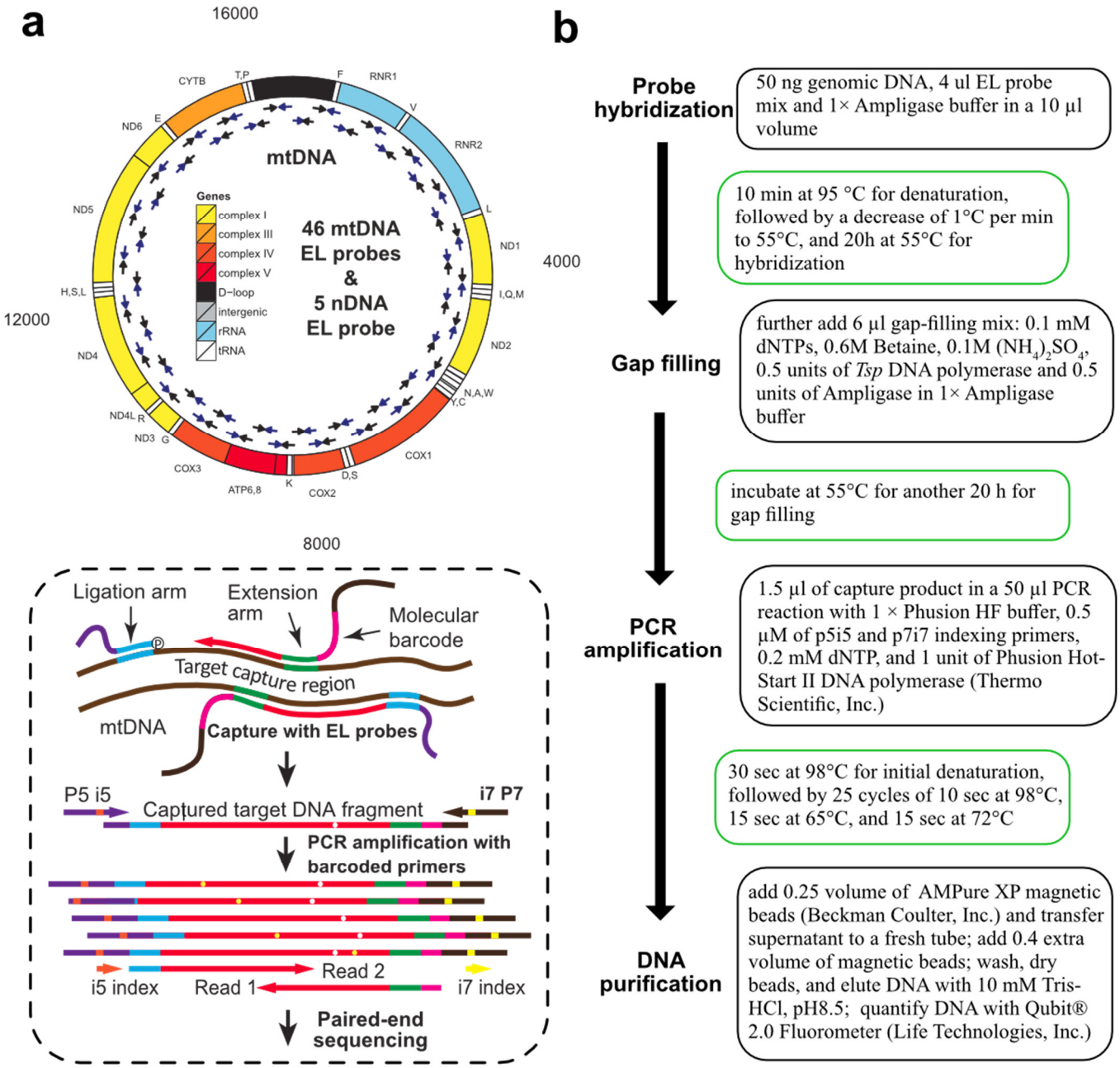
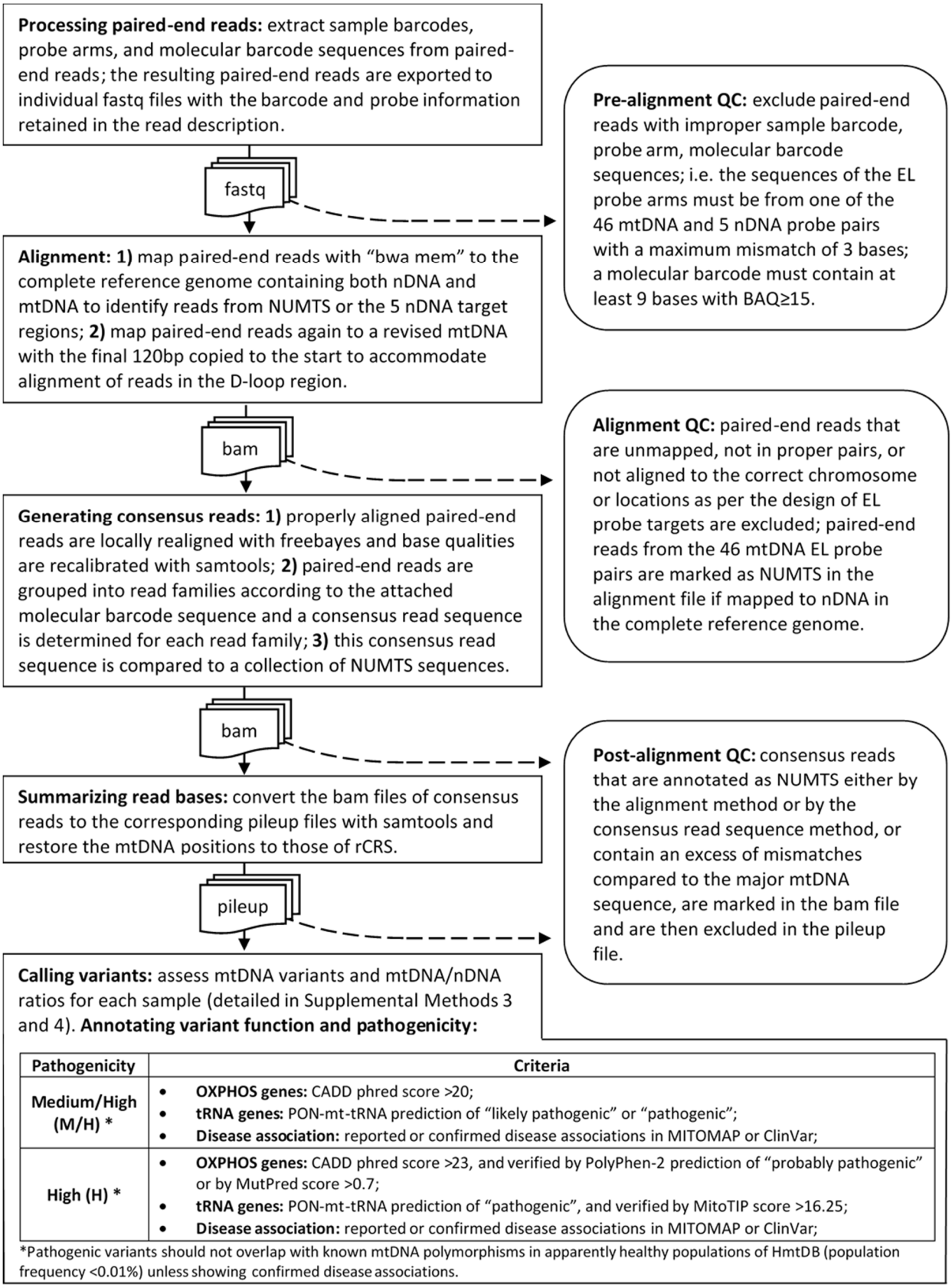


Figure S9. Experimental design and workflow chart for sequencing library preparation in STAMP. (a) The design of mtDNA capturing probes in STAMP and the related gap-filling reactions. The locations of the 46 mtDNA EL (extension-ligation) probes were depicted with arrows next to the mitochondrial genome in the upper panel of a. (**b**) The experimental workflow chart for sequencing library preparation. Information on samples and reagents needed in each step are shown in back boxes and the thermal conditions of the related reaction are shown in green boxes.



Pathogenicity	Criteria
Medium/High (M/H) *	• **OXPHOS genes:** CADD phred score >20; • **tRNA genes:** PON-mt-tRNA prediction of "likely pathogenic" or "pathogenic"; • **Disease association:** reported or confirmed disease associations in MITOMAP or ClinVar;
High (H) *	• **OXPHOS genes:** CADD phred score >23, and verified by PolyPhen-2 prediction of "probably pathogenic" or by MutPred score >0.7; • **tRNA genes:** PON-mt-tRNA prediction of "pathogenic", and verified by MitoTIP score >16.25; • **Disease association:** reported or confirmed disease associations in MITOMAP or ClinVar;

*Pathogenic variants should not overlap with known mtDNA polymorphisms in apparently healthy populations of HmtDB (population frequency <0.01%) unless showing confirmed disease associations.

Figure S10. Analytical workflow chart for read processing in STAMP. This analytical workflow chart summarizes the steps for read alignment, quality filtering, and variant analysis in STAMP. mtDNA: mitochondrial genome; nDNA: nuclear genome; NUMTS: nuclear mitochondrial DNA segment; rCRS: Revised Cambridge Reference Sequence of mtDNA; EL probes: extension-ligation probes; BAQ: base alignment quality; OXPHOS: oxidative phosphorylation.

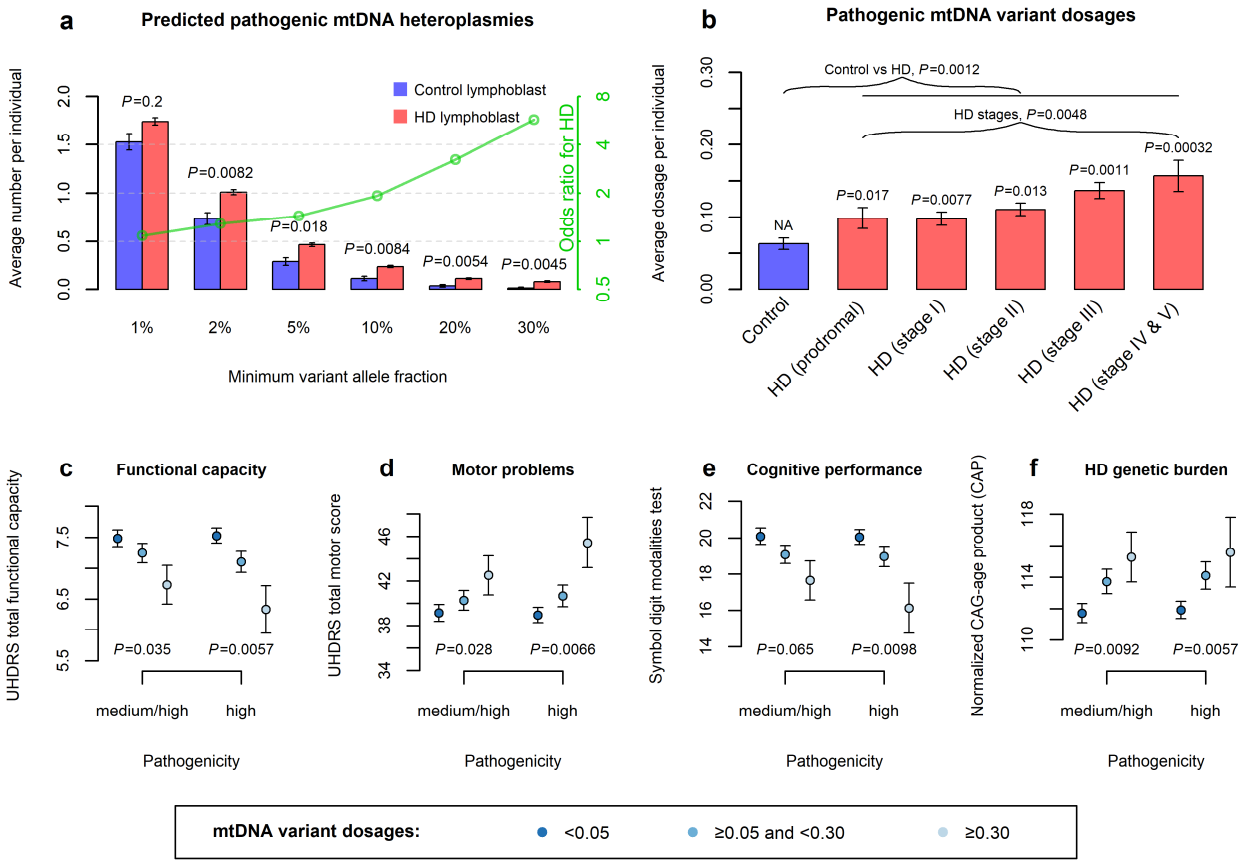


Figure S11. Associations of pathogenic mtDNA variants with HD phenotypes after including possibly fixed heteroplasmies. The results are shown in panel **a** for the incidence of predicted pathogenic mtDNA variants in HD and control lymphoblasts, in panel **b** for the dosages of predicted pathogenic mtDNA variants in HD and control lymphoblasts, and in panels **c-f** for the associations between the dosages of predicted pathogenic mtDNA variants and HD related clinical phenotypes and genetic burden, including (**c**) UHDRS total functional capacity score, (**d**) UHDRS total motor score, (**e**) symbol digit modalities test score, and (**f**) the normalized CAG-age products. The *P* values for mtDNA variants from the logistic regression analyses of the disease status are shown above the bars in panel **a,** with the effects indicated by the green lines and the values on the green y axes on a logarithmic scale. The *P* values for mtDNA variant dosages from the logistic regression analyses of disease status are indicated above the bars representing the corresponding HD stages in panel **b**. NA: not applicable. Error bars in all panels represent SEM.

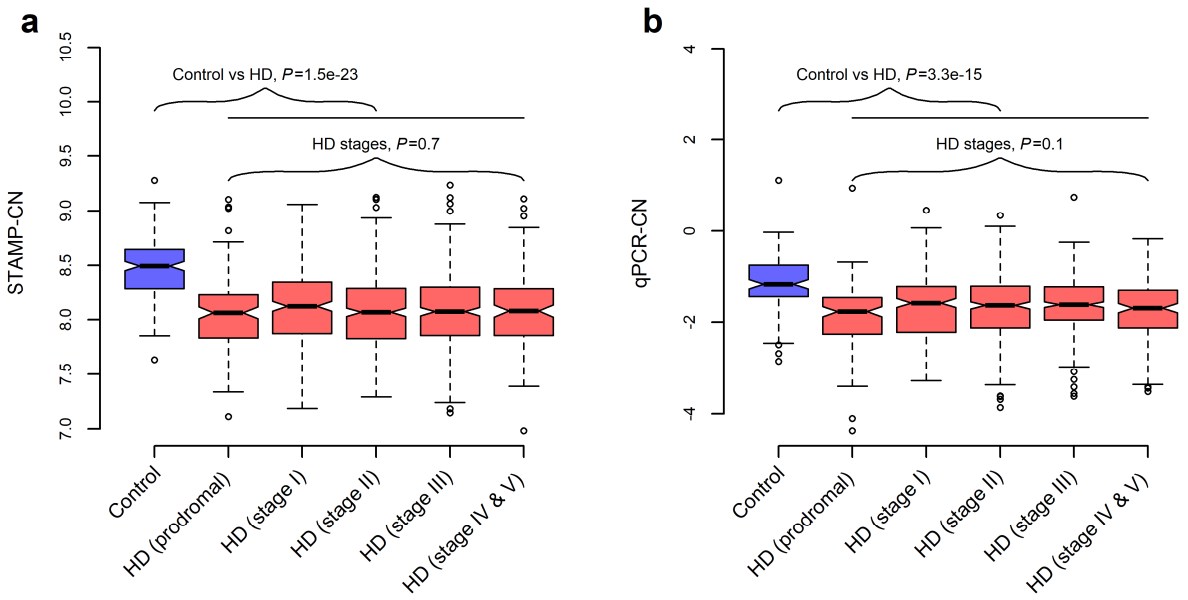


Figure S12. Relative mtDNA content in lymphoblasts of HD patients and control individuals. The box plots of the relative mtDNA content measured by (**a**) STAMP (N=1252 HD samples and 182 control samples) and (**b**) qPCR (N=1285 HD samples and 116 control samples) in lymphoblasts of HD patients and control individuals. The associations of STAMP-CN or qPCR-CN in lymphoblasts with disease status and disease stages were assessed by using logistic regression and linear regression, respectively. The covariates included age and sex, and mtDNA sequencing coverage if STAMP-CN was used. The values of STAMP-CN and qPCR-CN are on a logarithmic scale with base 2.

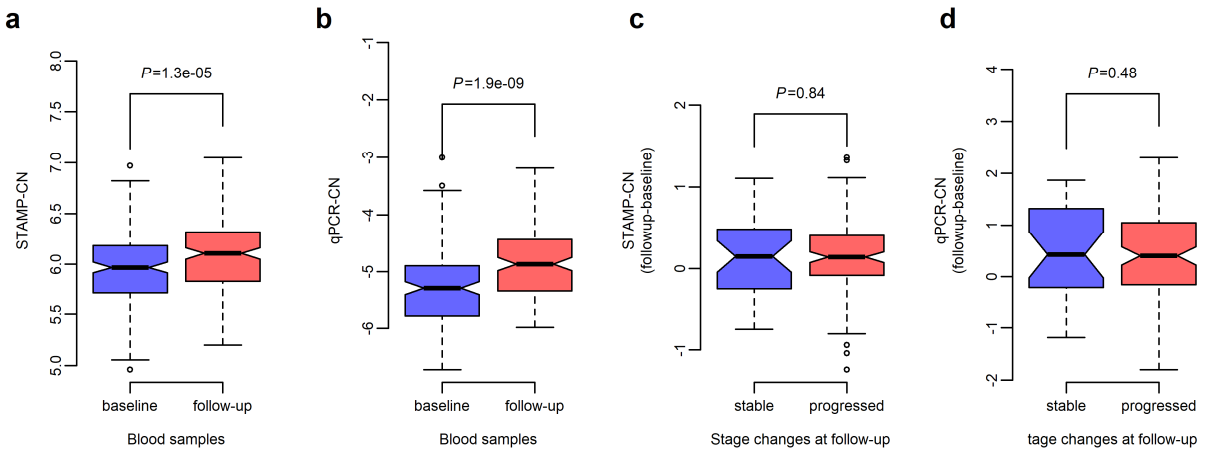


Figure S13. Longitudinal changes of relative mtDNA content in blood of HD patients. (**a, b**) The box plots of the relative mtDNA content measured by (**a**) STAMP (N = 181 patients) and (**b**) qPCR (N = 153 patients) in the baseline blood samples and the follow-up blood samples. The *P* values from paired t-test are shown. **(c, d)** The box plots of the longitudinal changes of relative mtDNA content measured by (**c**) STAMP and (**d**) qPCR in blood of HD patients with a stable stage and a progressed stage during the follow-up. After excluding late-stage patients at baseline, we obtained 35 stable-stage patients and 134 progressed-stage patients for the analysis of STAMP-CN, and 29 stable-stage patients and 113 progressed-stage patients for analysis of qPCR-CN. The t-test *P* values are shown. The values of STAMP-CN and qPCR-CN are on a logarithmic scale with base 2.

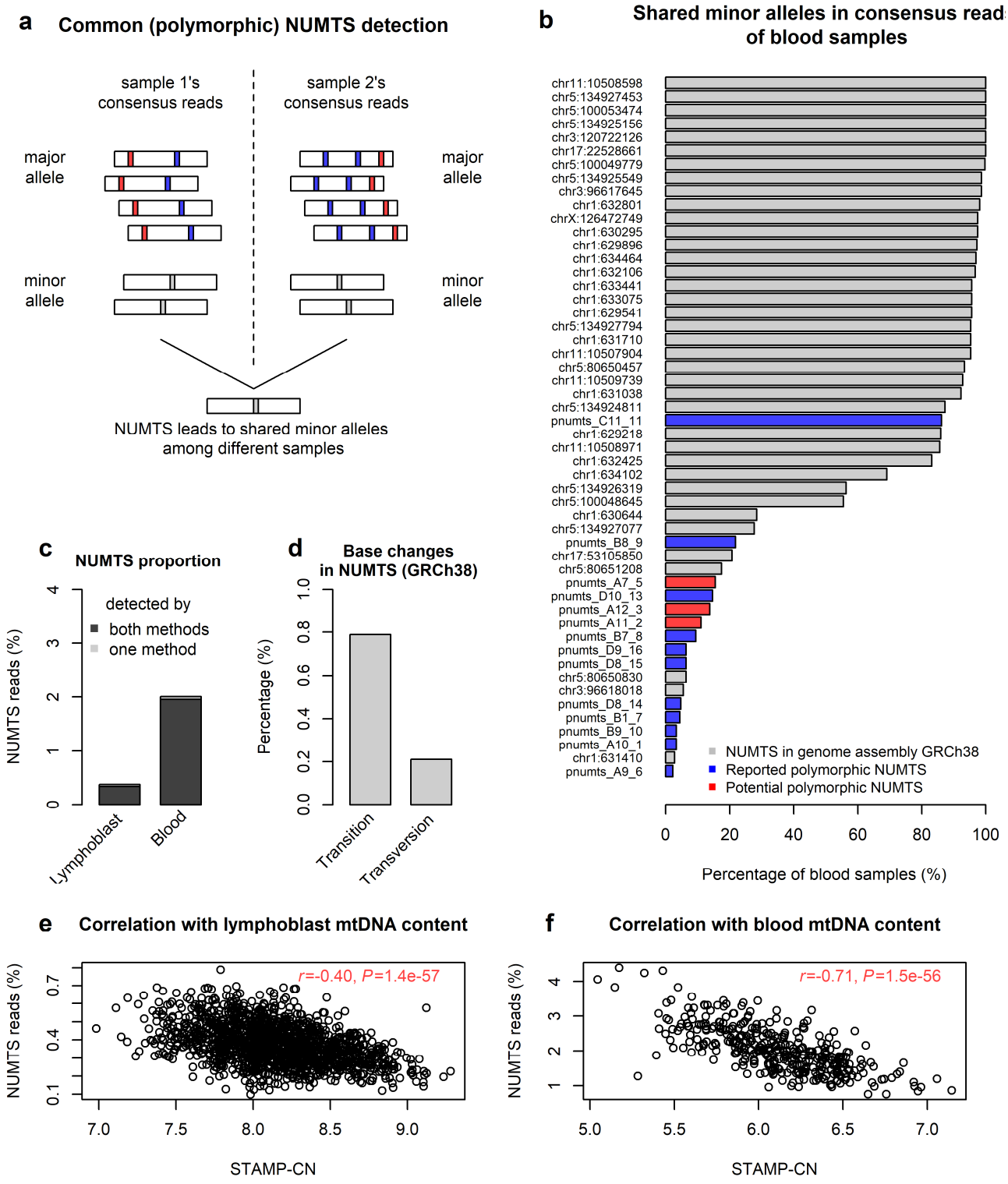


Figure S14. NUMTS identification in STAMP. (**a**) Schematic representation of the method used to identify common polymorphic NUMTS (nuclear mitochondrial DNA segments) based on the sequences of consensus reads. The red, blue and grey rectangles indicate different mtDNA variants. (**b**) The frequencies of the 52 common minor alleles of consensus reads (information provided in Table S16) identified in blood samples of the

current study. (**c**) The percentages of NUMTS reads determined by using the bwa alignment method and the consensus read sequence method. (**d**) The percentages of transition and transversion base changes, including substitutions and single nucleotide polymorphisms, in known NUMTS (Table S9) in the human genome (assembly GRCh38), in comparison to the reference mtDNA sequence. (**e**, **f**) The negative correlations between the proportions of consensus reads annotated as NUMTS and the relative mtDNA content in (**e**) lymphoblasts (N=1434) and (**f**) blood samples (N=362).

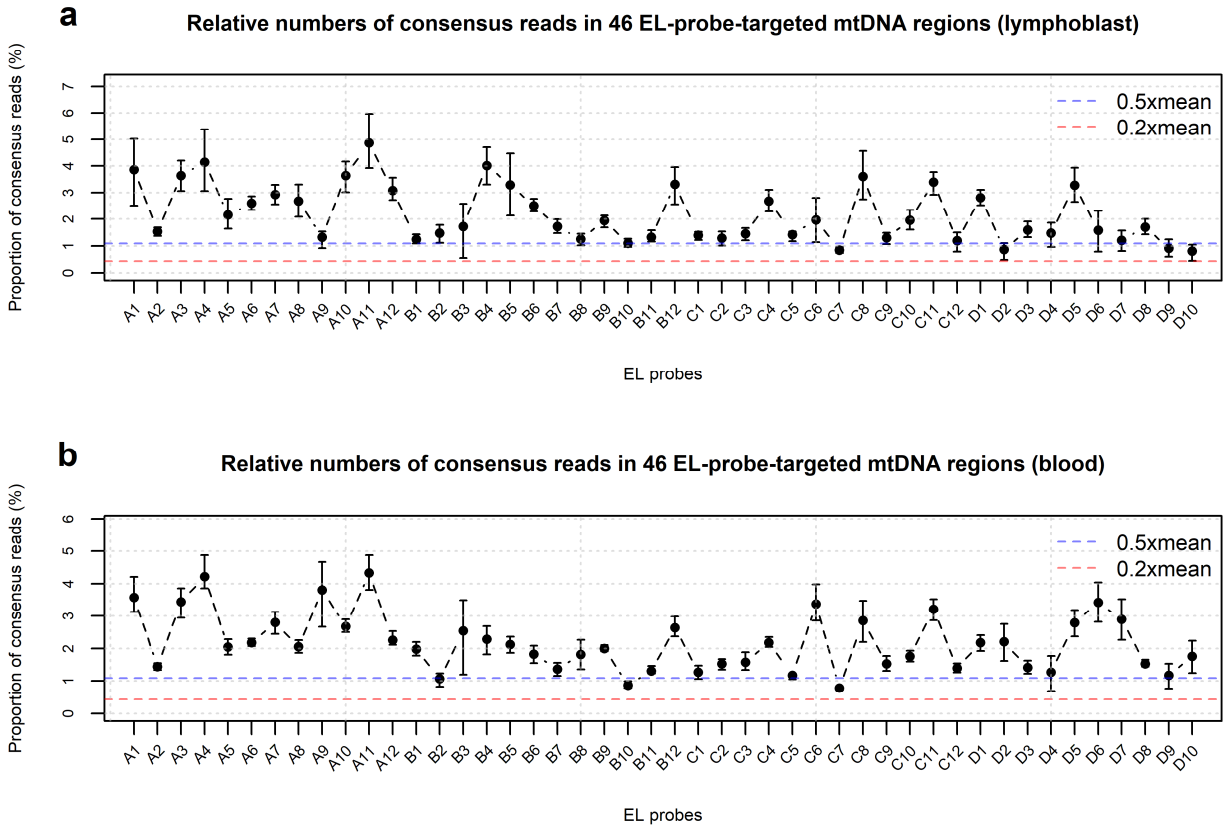


Figure S15. Relative coverages of consensus reads on mtDNA in STAMP. The results are shown in panel **a** for the consensus reads in lymphoblasts and in panel **b** for the consensus reads in blood samples. Each dot refers to the average coverage of consensus reads from one of 46 mtDNA EL probes. Error bars represent the interquartile range. The blue and red dashed lines indicate 50% and 20% of the mean depth of coverage of consensus reads on mtDNA.

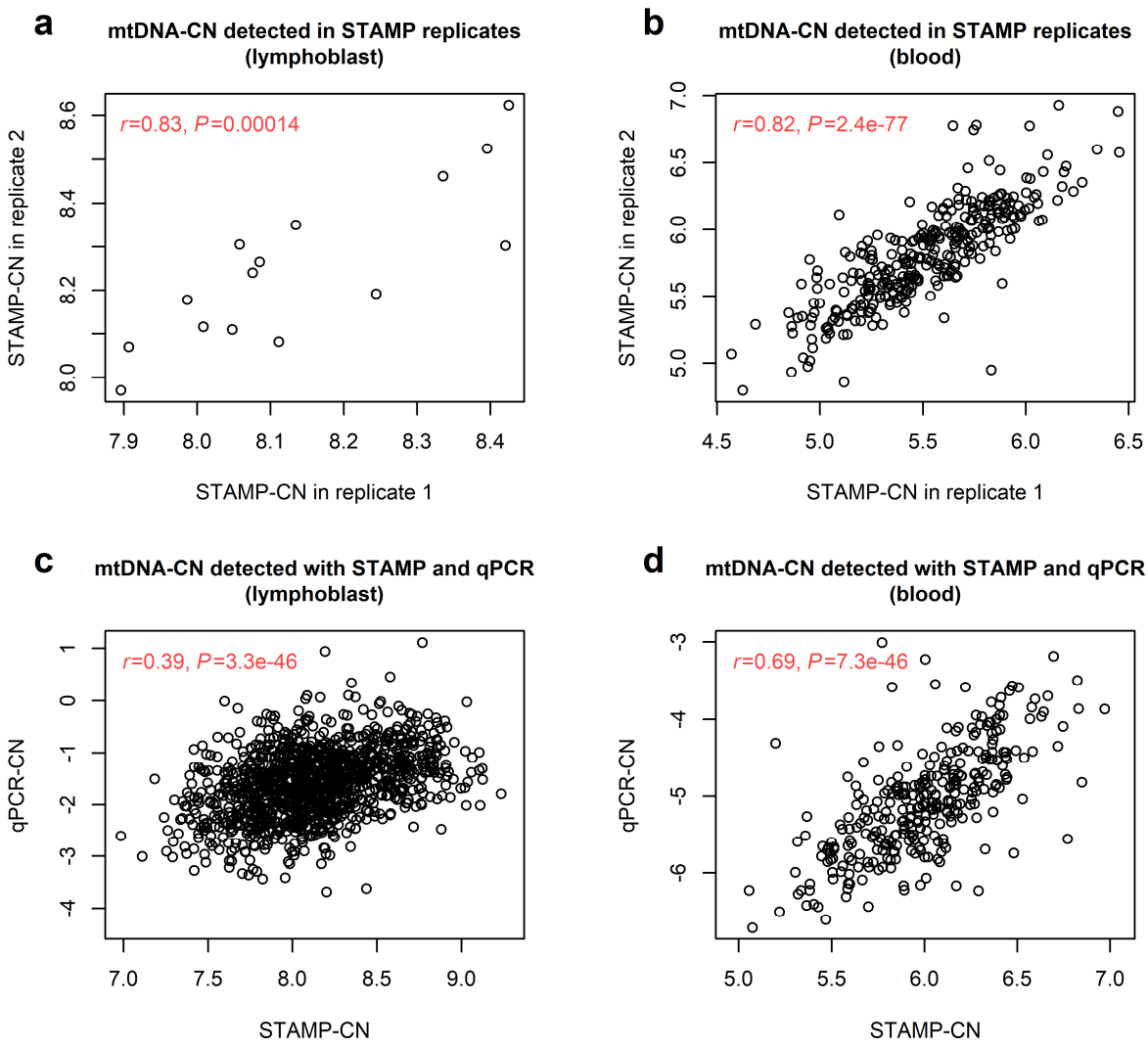


Figure S16. mtDNA content in lymphoblast and blood samples measured by using STAMP and qPCR. (**a, b**) STAMP-CN measured in sequencing replicates of (**a**) 15 lymphoblasts and (**b**) 320 blood samples. (**c, d**) mtDNA content measured by STAMP and qPCR in (**c**) 1234 lymphoblasts and (**d**) 318 blood samples. The values of STAMP-CN and qPCR-CN are on a logarithmic scale with base 2.

Supplemental Tables

Table S1. Summary of STAMP experiments in REGISTRY samples.

Variables	Sample QC criteria	HD lymphoblast samples	Control lymphoblast samples	Blood samples
No. of samples sequenced with STAMP		1630	200	376
No. of samples passing QC for homoplasmy calling	median consensus reads on mtDNA >100X	1614 (99%)	184 (92%)	376 (100%)
No. of samples passing QC for heteroplasmy calling	median consensus reads on mtDNA >1000X	1549 (95%)	182 (91%)	376 (100%)
No. of samples passing QC for heteroplasmy analysis	samples removed due to possible DNA contamination	1549 (95%)	182 (91%)	362 (96%)†
No. of samples sequenced with nuclear DNA probes*		1272 (78%)	182 (91%)	362 (96%)
No. of samples passing QC for STAMP-CN calculation*	≥5X for each of the 3 probes targeting chromosomes 8, 14, 19	1252 (77%)	182 (91%)	362 (96%)

*Numbers are shown for the samples that passed quality control for heteroplasmy analysis. The 5 EL (extension-ligation) probe pairs used to capture nuclear DNA regions were not included in the first 3 sample plates of the current study.

†Both baseline and follow-up samples from the individuals detected with possible DNA contamination were removed.

Table S2. Demographics, clinical phenotypes and mtDNA sequencing characteristics of REGISTRY samples.

Groups	Lymphoblast samples		Blood samples	
	Control	HD	Baseline HD	Follow-up HD
Demographics				
N*	182	1549	181	181
Sex, no. of men (%)	90 (49)	776 (50)	93 (51)	93 (51)
Age, mean (SD)	48.0 (8.8)	52.2 (12.0)	49.9 (9.9)	55.7 (10.0)
Caucasian ethnicity (%) †	176 (97)	1513 (98)	180 (99)	180 (99)
HD-related phenotypes				
CAG repeat length in *HTT*, mean (SD)	-	44 (3.7)	44 (3.3)	44 (3.3)
Prodromal and early stage, N (%) ‡	-	922 (54)	135 (75)	56 (31)
Middle stage, N (%) ‡	-	404 (24)	34 (19)	62 (34)
Late stage, N (%) ‡	-	198 (12)	12 (7)	63 (35)
Total functional capacity score, N, mean (SD)‡	-	1524, 7.3, (3.8)	181, 8.8 (3.5)	181, 4.7 (3.8)
Total motor score, N, mean (SD)‡	-	1524, 39.9, (21.3)	181, 31.5 (17.6)	179, 57.8 (22.7)

Symbol digit modalities test score, N, mean (SD)	-	1266, 19.5 (11.4)	157, 23.3 (11.5)	142, 13.6 (10.1)
STAMP sequencing characteristics				
Median coverage of consensus reads on mtDNA (X), mean (10^{th}, 25^{th}, 75^{th}, and 90^{th} percentiles)	4580 (2418, 3146, 5790, 6966)	3459 (1695, 2439, 4307, 5370)	5695 (3028, 3896, 6991, 8898)	6552 (3369, 4590, 8154, 10300)
Percentage of mtDNA sites with >500X coverage (%), median (10^{th}, 25^{th}, 75^{th}, and 90^{th} percentiles)	100 (98, 100, 100, 100)	100 (94, 98, 100, 100)	100 (98, 100, 100, 100)	100 (98, 100, 100, 100)
Percentage of mtDNA sites with >1000X coverage (%), median (10^{th}, 25^{th}, 75^{th}, and 90^{th} percentiles)	98 (91, 95, 100, 100)	96 (77, 90, 98, 100)	100 (96, 98, 100, 100)	100 (96, 98, 100, 100)

*The number of samples passed quality control for heteroplasmy analysis.

†Self-reported ethnicity or race in the REGISTRY database.

‡Prodromal stage: diagnostic confidence level < 4; early stage: stage I: TFC score≥11 and stage II: 7≤ TFC score<11; middle stage: stage III: 4≤TFC score<7, and late stage: stage IV/V: TFC score<4. Percentages were rounded to the nearest integer and thus may not add up to 100%; 25 HD patients with lymphoblast samples do not have a diagnostic confidence level recorded or a TFC score collected within about 1 year of the sample collection.

Table S3. mtDNA heteroplasmies detected in STAMP sequencing replicates of lymphoblast and blood samples.

Variables	Lymphoblast samples (N=17 pairs)*	Blood samples (N=320 pairs)*			
Average median coverage on mtDNA (-fold; X)	4797 [3062-6537]	3133 [1877-3923]			
Minimum site coverage in the pair of replicates (X)	500	500	1000	1500	2000
Minimum VAF of heteroplasmy calling (%)	1	1	1	0.5	0.5
No. of heteroplasmies detected	90	462	381	620	423
No. of heteroplasmies with VAF≥0.5×minimum VAF and passing all quality control filters† in the replicate (%)	86 (95.6)	414 (89.6)	355 (93.2)	541 (87.3)	396 (93.6)
No. of heteroplasmies with VAF ≥0.2% in the replicate (%)	89 (98.9)	448 (97.0)	378 (99.2)	591 (95.3)	413 (97.6)
Coefficient of variation‡ for VAFs of a heteroplasmy in replicates (%)	17	22	20	23	20
Pearson's *r* between VAFs of a heteroplasmy in replicates§	0.998	0.996	0.996	0.998	0.998
Wilcoxon signed rank test *P* for the VAF fold change of a heteroplasmy between replicates	0.51	0.13	0.77	0.21	0.56

*Pairs of samples with the median STAMP sequencing coverage on mtDNA > 1000X.

†Quality control filters listed in Table S10.

‡Computed using within-subject standard deviation method.

§*P* values for $r < 2.2\times10^{-16}$.

Table S4. Age-dependent changes of mtDNA heteroplasmies in lymphoblasts of control individuals.

Variables	mtDNA variant pathogenicity	Age (control lymphoblast)	
		Beta (SE)	*P*
mtDNA variant dosages	M/H	0.008 (0.008)	0.30
	H	0.005 (0.008)	0.53
	others	0.019 (0.008)	**0.020**
mtDNA variant incidence	M/H	0.004 (0.008)	0.64
	H	0.003 (0.007)	0.73
	others	0.022 (0.008)	**0.0053**

The variant incidence and dosages of mtDNA heteroplasmies in control and HD lymphoblasts were inverse normal transformed and were further adjusted for sex and sequencing coverage. The associations were assessed by using the model: INV dosage/incidence ~ age + age × disease + disease. Beta and *P* refer to the beta coefficient of the term age in the model and its significance level based on the Wald test.

M/H: medium or high pathogenicity; H: high pathogenicity; others: not predicted with medium or high pathogenicity. *P* values <0.05 are highlighted in bold type.

Table S5. Associations of the expansion of mtDNA heteroplasmies in blood with CAG repeat length.

Variables	mtDNA variant pathogenicity (N)*	CAG repeat length	
		Beta (SE)	*P*
Early-stage patients†	M/H (29)	0.39 (0.12)	**0.0034**
	H (20)	0.42 (0.17)	**0.029**
	others (50)	-0.043 (0.10)	0.68
All patients	M/H (76)	0.044 (0.049)	0.38
	H (52)	0.063 (0.056)	0.27
	others (160)	-0.028 (0.037)	0.46

The associations were assessed by using the linear mixed-effects model: $\log_2$(VAF follow-up/ VAF baseline) ~ CAG_length + baseline_age + sex + followup_duration + (1|patient_id).

*The number of mtDNA heteroplasmies used for analysis.

†Among early-stage HD patients with moderate motor symptoms (TFC score ≥7 and total motor score <25).

M/H: medium or high pathogenicity; H: high pathogenicity; others: not predicted with medium or high pathogenicity in HD patients carrying pathogenic heteroplasmies. *P* values <0.05 are highlighted in bold type.

Table S6. Associations of HD stages with variant dosages of pathogenic heteroplasmies in each of the four oxidative phosphorylation complexes encoded by mtDNA.

OXPHOS	mtDNA variant pathogenicity	Beta (SE)	*P*	*P (heterogeneity)*	Power (%)
Complex I	**M/H**	0.40 (0.24)	0.10	0.54	66
	H	0.71 (0.30)	0.017	0.98	66
Complex III	**M/H**	0.78 (0.56)	0.17	0.74	18
	H	0.42 (0.60)	0.48	0.66	22
Complex IV	**M/H**	0.78 (0.37)	0.035	0.63	35
	H	1.00 (0.44)	0.022	0.53	38
Complex V	**M/H**	0.58 (0.73)	0.42	1.00	11
	H	0.50 (0.74)	0.50	0.79	14

The associations were assessed by using linear regression adjusted for age, sex, sequencing coverage, and CAG repeat length among HD lymphoblasts. Variant dosages were computed using heteroplasmies in each of the four oxidative phosphorylation (OXPHOS) complexes encoded by mtDNA. *P* for heterogeneity was obtained from Cochran's Q test by comparing the effect of heteroplasmies from one OXPHOS complex and the effect of all pathogenic heteroplasmies in mtDNA. Power (%) was estimated at two-tailed alpha=0.05 in the linear model by using the effect estimated from all pathogenic heteroplasmies and the relative contribution of each of the four OXPHOS complexes to the variance in variant dosage.

M/H: medium or high pathogenicity; H: high pathogenicity.

Table S7. No correlations between pathogenic mtDNA variant dosages and mtDNA content in lymphoblasts.

Variables	mtDNA variant pathogenicity	STAMP-CN		qPCR-CN	
		Beta (SE)	*P*	Beta (SE)	*P*
All lymphoblasts	M/H	0.009 (0.013)	0.49	0.015 (0.017)	0.38
	H	-0.007 (0.015)	0.63	0.010 (0.020)	0.60
HD lymphoblasts	M/H	0.014 (0.013)	0.29	0.021 (0.017)	0.22
	H	-0.001 (0.016)	0.94	0.017 (0.020)	0.41

The associations were assessed by using linear regression adjusted for age, sex, and sequencing coverage in all lymphoblasts and in HD lymphoblasts. STAMP-CN (N = 182 controls and 1252 HD lymphoblasts) and qPCR-CN (N = 116 controls and 1285 HD lymphoblasts) were standardized to have a mean of 0 and a standard deviation of 1. Beta refers to the effect of every 0.1 increase in the variant dosage of mtDNA heteroplasmies on the standardized mtDNA content.

M/H: medium or high pathogenicity; H: high pathogenicity.

Table S8. No associations between the expansion of predicted pathogenic mtDNA heteroplasmies in blood with mtDNA content.

Variables	Variant pathogenicity (N)	STAMP-CN		qPCR-CN	
		Beta (SE)	*P*	Beta (SE)	*P*
Model 1 (baseline mtDNA content)	M/H (72)	-0.11 (0.09)	0.25	-0.03 (0.10)	0.73
	H (49)	-0.08 (0.12)	0.54	0.01 (0.12)	0.96
Model 2 (follow-up mtDNA content)	M/H (72)	-0.08 (0.10)	0.45	-0.10 (0.11)	0.35
	H (49)	-0.04 (0.13)	0.77	-0.05 (0.13)	0.72

The associations were assessed by using the linear mixed-effects model: $\log_2$(VAF follow-up/ VAF baseline) ~ CN + baseline_age + sex + followup_duration + (1|patient_id). In the analyses of the follow-up mtDNA content (CN), the baseline CN and disease stage was further considered as fixed-effect covariates in the model.

*The number of mtDNA heteroplasmies used for analysis, 7 of which were not included in the analyses of qPCR-CN due to missing qPCR measurements.

M/H: medium or high pathogenicity; H: high pathogenicity.

Table S9. Information of the extension-ligation Probes and the target regions used in STAMP.

CHR: chromosome; Strand: ligation arm strand /extension arm strand; Barcode: maximum barcode length.

Probe ID	CHR	Target region (start-end)	Str-and	Bar-code	Ligation arm sequence / Extension arm sequence	mtDNA polymor-phisms in probe arms*	NUMTS of high sequence similarity to the target region†
A1	chrM	311-753	+/-	15	CTGTGGCCAGAAG CGGGG / TGTTCCTTTTGATC GTGGTGATTTA	C315CC; A750G	chr5:80651964-80652368(ALT=33|SNP=5)
A2	chrM	701-1141	-/+	15	AACAAAACTGCTC GCCAGAAC / CATCCCCGTTCCA GTGAGTT	G709A	chrX:55183468-55183879(ALT=54|SNP=2); chr5:123760949-123761364(ALT=51|SNP=1); chr5:8622215-8622573(ALT=58|SNP=1); chr2:140223662-140224026(ALT=54); chr11:10509718-10510156(ALT=25|SNP=2); chr17:22522185-22522599(ALT=48|SNP=2); chr5:80651577-80652015(ALT=17|SNP=4); chr11:87813584-87813980(ALT=56|SNP=1)
A3	chrM	1095-1535	+/-	15	TTTAACTGTTGAG GTTTAGGGCT / AGGGGTTTTAGTT AAATGTCCTTTG		chr5:123761318-123761758(ALT=50); chr20:30744399-30744813(ALT=51); chr11:10509324-10509764(ALT=10|SNP=4); chr4:155465622-155466041(ALT=54); chr1:237945466-237945877(ALT=48|SNP=5); chr9:64047979-64048396(ALT=47); chr17:22522554-22522995(ALT=37|SNP=2); chrY:11134577-11134994(ALT=52); chrX:55183107-55183512(ALT=53); chr2:94899006-94899420(ALT=45|SNP=1); chr21:8846798-8847202(ALT=50); chr5:80651183-80651623(ALT=10|SNP=3); chr9:33657131-33657574(ALT=50|SNP=4); chr22:11856589-11857001(ALT=53); chr11:103410211-

							103410615(ALT=48|SNP=2); chr2:211775683-211776103(ALT=57|SNP=1); chr4:116298262-116298685(ALT=57); chr22:12135354-12135766(ALT=53); chr17:19599072-19599487(ALT=54|SNP=2); chr20:57360153-57360569(ALT=57)
A4	chrM	1474-1911	-/+	12	CCAAAGCTAAGACCCCCGAAACC / CGTACACACCGCCCGTCAC		chr11:10508948-10509385(ALT=19|SNP=2); chr5:80650807-80651244(ALT=12|SNP=2); chr3:96617995-96618429(ALT=16|SNP=3)
A5	chrM	1854-2283	+/-	15	TTGCAAAGTTATTTCTAGTTAATTCATT / GGTGATAGATTGGTCCAATTGG		chr3:40252583-40253007(ALT=56); chr7:142666452-142666876(ALT=45|SNP=6); chr9:33657892-33658315(ALT=52|SNP=2); chr11:10508576-10509005(ALT=20|SNP=6); chr3:160947648-160948007(ALT=39|SNP=3); chr9:5092647-5093009(ALT=50|SNP=5); chrX:55182343-55182748(ALT=53|SNP=2); chr5:80650435-80650864(ALT=17|SNP=2); chr3:96617623-96618052(ALT=13|SNP=2); chr17:22523282-22523709(ALT=33|SNP=1)
A6	chrM	2206-2646	-/+	15	TGTATGAATGGCTCCACGAGG / AAGCTCAACACCCACTACCT		chr5:80650073-80650512(ALT=32|SNP=2); chr3:96617260-96617700(ALT=22|SNP=2); chr11:10508213-10508653(ALT=26|SNP=2); chr17:22523633-22524075(ALT=41|SNP=3); chr9:33658282-33658681(ALT=50|SNP=2); chrX:55181981-55182379(ALT=55|SNP=1); chr3:40252974-40253366(ALT=58|SNP=1); chr7:142666804-142667244(ALT=58); chr16:3371889-3372280(ALT=55|SNP=1)
A7	chrM	2538-2980	+/-	12	TGTCACTGGGCAGGCGGT / AAACCCTATTGTTGATATGGACTCT		chr9:33658577-33659017(ALT=54|SNP=2); chr7:142667137-142667580(ALT=50); chr11:10507887-10508321(ALT=33|SNP=1); chr17:22523967-22524404(ALT=39|SNP=5)
A8	chrM	2932-3377	-/+	15	ATGGCATTCCTAATGCTTACCGA / GGATAACAGCGCAATCCTATTCT		chr14:84172457-84172840(ALT=55|SNP=3); chr17:22524363-22524800(ALT=52); chr7:141804969-141805339(ALT=49|SNP=1); chr20:57358311-57358746(ALT=55|SNP=3); chr4:92701968-92702367(ALT=55)
A9	chrM	3313-3753	+/-	15	AGGAGTAGGAGGTTGGCCA / AATGATGGCTAGGGTGACTTCA	C3741T	
A10	chrM	3687-4128	-/+	15	ACCTCCCTGTTCTTATGAATTCGA / GCCCTGATCGGCGCACTG		
A11	chrM	4022-4471	+/-	15	ATATGTTGTTCCTAGGAAGATTGTA / ATTAGTACGGGAAGGGTATAACCAA		chr7:57186001-57186415(ALT=56|SNP=1); chr17:19603007-19603416(ALT=50|SNP=2); chr1:629192-629641(ALT=2|SNP=4)
A12	chrM	4348-4797	-/+	15	GCTATAGCAATAAAACTAGGAATAGCCC / CCCATCCCTGAGAATCCAAAAT	A4793G	chr1:629518-629967(ALT=1|SNP=3)
B1	chrM	4705-5150	+/-	12	TGGTTCATTGTCCGGAGAGT / GGTCGTGGTGCTGGAGTTTA	G5147A	chr1:629875-630320(ALT=2|SNP=10)

B2	chrM	5099-5498	-/+	15	CTCCTACCTATCTCCCCTTTTATA / CTAACTACTACCGCATTCCTACTAC		chr1:630269-630667(ALT=3|SNP=7)
B3	chrM	5453-5896	+/-	15	AGCGTGGTAAGGGCGATGAG / GGTGAGGTAAAATGGCTGAGTG	G5460A; G5471A	chr10:69592941-69593321(ALT=53|SNP=1); chr7:69333643-69334004(ALT=45|SNP=2); chr7:64110163-64110546(ALT=51|SNP=2); chr7:141802072-141802466(ALT=49|SNP=2); chr2:131384078-131384460(ALT=52|SNP=2); chrX:55178752-55179164(ALT=54); chr11:103405848-103406247(ALT=49|SNP=3); chr7:57187449-57187834(ALT=47); chr2:130273473-130273891(ALT=57|SNP=2); chr1:630623-631066(ALT=2|SNP=4); chr1:237941074-237941457(ALT=43|SNP=1); chr8:133755411-133755833(ALT=57|SNP=3); chr2:211777824-211778242(ALT=52|SNP=4); chr4:155461249-155461671(ALT=48|SNP=3); chr9:5096226-5096652(ALT=54|SNP=1); chr17:22526876-22527309(ALT=45); chr2:140217252-140217685(ALT=59); chr2:155263778-155264189(ALT=39|SNP=3)
B4	chrM	5843-6286	-/+	15	GCAGGAACAGGTTGAACAGT / CCCCTGTCTTTAGATTTACAGTCC		chr14:32484418-32484859(ALT=28); chr1:631013-631456(ALT=2|SNP=2)
B5	chrM	6218-6658	+/-	15	CAGGAGTAGGAGAGAGGGAGG / CCGAAGCCTGGTAGGATAAGAA	T6221C	chr1:631388-631828(ALT=6|SNP=6); chr2:50589028-50589415(ALT=58|SNP=1); chr14:32484098-32484486(ALT=34)
B6	chrM	6514-6960	-/+	15	TTTCTTTTCACCGTAGGTGGCC / TGGCATCACTATACTACTAACAGAC		chr1:631684-632131(ALT=3|SNP=3); chrX:126471704-126472111(ALT=24|SNP=2); chr5:100054165-100054612(ALT=39|SNP=3)
B7	chrM	6914-7357	+/-	15	AATCCTAGGGCTCAGAGCAC / ACTATTAGGACTTTTCGCTTCGAA		chr5:100053770-100054211(ALT=43|SNP=3); chr17:53105830-53106270(ALT=17|SNP=3); chrX:126472065-126472452(ALT=26|SNP=5); chr1:632085-632528(ALT=3|SNP=9)
B8	chrM	7232-7676	-/+	12	TTTCATGATCACGCCCTCATAATCA / CCCGATGCATACACCACATGA		chr5:100053449-100053893(ALT=43|SNP=1); chr1:632404-632847(ALT=2|SNP=5)
B9	chrM	7608-8036	+/-	15	GGGGAAGTAGCGTCTTGTAGA / GAATGGGGGCTTCAATCGGG		chr1:632779-633207(ALT=3|SNP=9); chr5:100053093-100053517(ALT=42|SNP=4)
B10	chrM	7882-8297	-/+	15	STACCCCCTCTAGAGCCCAC / ATTGGCCACCAATGGTACTGA		chr5:100052827-100053243(ALT=36|SNP=3); chr17:22529285-22529661(ALT=54|SNP=2); chr1:633053-633466(ALT=7|SNP=8)
B11	chrM	8249-8666	+/-	15	TGCTATAGGGTAAATACGGGCC / TGTTGGGTGGTGATTAGTCGG	G8251A; G8269A	chr2:87824890-87825242(ALT=15); chr1:633418-633835(ALT=4|SNP=5); chr5:100052458-100052873(ALT=41|SNP=2)
B12	chrM	8544-8968	-/+	15	CCCATACTAGTTATTATCGAAACCATCA / GCTTCATTCATTGCCCCCAC		chr5:100052156-100052580(ALT=45|SNP=3); chr1:633715-634137(ALT=1|SNP=6)

C1	chrM	8912-9346	+/-	15	GTGTAGGTGTGCCTTGTGGT / AGGCCTAGTATGAGGAGCGT		chr1:634081-634515(ALT=6|SNP=1); chr5:100051778-100052212(ALT=43|SNP=3)
C2	chrM	9274-9695	-/+	15	CCGAAACCAAATAATTCAAGCACTG / AGCCCTCCTAATGACCTCCG		chr1:634443-634864(ALT=4|SNP=6); chr4:12640294-12640635(ALT=20|SNP=6)
C3	chrM	9500-9940	+/-	15	GCTGGAGTGGTAAAAGGCTCA / ACAAAATGCCAGTATCAGGCG		chr5:100051185-100051624(ALT=40|SNP=6)
C4	chrM	9853-10293	-/+	15	CTTTTACCCCTACCATGAGCCC / TATCTGCTTCATCCGCCAACTA		chr5:100050829-100051269(ALT=48|SNP=1)
C5	chrM	10209-10648	+/-	15	ATAGCTACTAAGAAGAATTTTATGGAGA / GGCACAATATTGGCTAAGAGGG		chr5:134928148-134928527(ALT=20|SNP=1)
C6	chrM	10600-11022	-/+	15	CGCCACTTATCCAGTGAACC / TACTCTCATAACCCTCAACACCC		chrX:126472731-126473147(ALT=26|SNP=1); chr5:134927774-134928194(ALT=17|SNP=6); chr5:100050100-100050522(ALT=41|SNP=1)
C7	chrM	10926-11368	+/-	12	GTTAGGGGGTCGGAGGAA / AAAAGCTATTGTGTAAGCTAGTCAT		chr5:134927428-134927868(ALT=16|SNP=2); chr5:100049755-100050194(ALT=47|SNP=3)
C8	chrM	11298-11744	-/+	15	CTTACATCCTCATTACTATTCTGCC / TCTCACTGCCCAAGAACTATCA	T11299C	chr5:100049378-100049819(ALT=37|SNP=2); chr5:134927052-134927498(ALT=14|SNP=3)
C9	chrM	11687-12127	+/-	15	ATTATGAGAATGACTGCGCCGG / CCCGGTAATGATGTCGGGG		chr5:134926669-134927109(ALT=16|SNP=3)
C10	chrM	12075-12502	-/+	15	CTCTTCCCCACAACAATATTCATGT / ACACCTATCCCCCATTCTCCT	G12501A	chr5:100048622-100049047(ALT=40|SNP=2); chr5:134926296-134926721(ALT=22|SNP=4)
C11	chrM	12448-12894	+/-	12	ATAATAAAGGTGGATGCGACAATGG / TAAGGCGAGGATGAAACCGATA		chr5:134925902-134926348(ALT=29)
C12	chrM	12823-13270	-/+	15	CTCCACTTCAAGTCAACTAGGAC / ATGCCAACACAGCAGCCATT		chr5:94567617-94568064(ALT=53|SNP=2); chr5:100047852-100048299(ALT=51|SNP=2); chr5:134925526-134925973(ALT=20|SNP=1)
D1	chrM	13214-13658	+/-	15	ACGATTTTTTTGATGTCATTTTGTGTA / GTAAGGGTGGGGAAGCGA		chr5:134925138-134925582(ALT=14|SNP=1); chr5:100047465-100047906(ALT=29|SNP=3); chr5:94568008-94568449(ALT=39|SNP=3)
D2	chrM	13592-14008	-/+	15	ACTCCTCCTAGACCTAACCTGAC / AAGCGCCTATAGCACTCGAA		chr5:134924788-134925204(ALT=43|SNP=3)

D3	chrM	13955-14396	+/-	15	TTTTGGCTCGTAAGAAGGCCT / TAGTGGGGTTAGCGATGGAGG	A13966G	chr5:100046726-100047164(ALT=41|SNP=1); chr5:134924400-134924841(ALT=21); chr5:94568763-94569188(ALT=45|SNP=2)
D4	chrM	14348-14773	-/+	15	CCCAATACGCAAAACTAACCCC / ACTCTTTCACCCACAGCACC	C14766T	chr5:94569145-94569563(ALT=34|SNP=2); chr5:134924023-134924448(ALT=20); chr5:100046349-100046774(ALT=34|SNP=3)
D5	chrM	14719-15163	+/-	15	GTGTTCTTGTAGTTGAAATACAACG / TGATATTTGGCCTCACGGGA		chr5:134923634-134924074(ALT=27); chr5:94569512-94569954(ALT=48|SNP=3)
D6	chrM	15092-15535	-/+	15	TATACCCTAGCCAACCCCTTAAAC / GCATTATCCTCCTGCTTGCA		chr5:94569885-94570328(ALT=58|SNP=1); chr5:134923309-134923704(ALT=31)
D7	chrM	15353-15778	+/-	15	TAGGGGGTTGTTTGATCCCG / GCTTACTGGTTGTCCTCCGAT		
D8	chrM	15733-16178	-/+	12	TAGTACATAAAAACCCAATCCACAT / GCAGACCTCCTCATTCTAACC	A16162G; A16163G; T16172C	chr5:94570528-94570918(ALT=47|SNP=1)
D9	chrM	16112-16561	+/-	15	TTTATGGTACCGTACAATATTCATGGT / TGTCTTATTTAAGGGGAACGTGT	T16126C; G16129A; G16129C	
D10	chrM	16499-358	-/+	15	CACATCTCTGCCAAACCCCAA / ATCTGGTTCCTACTTCAGGGTC	T16519C	
EMC1	chr1	19563534-19563970	+/-	15	GTTTTGGGCCGTCAGAGGAT / AAGGAAAACCGGACTTCGCA		
WRN	chr8	33162768-33163207	-/+	15	TGTTTGAAGTCTTGGTTTGGTG / GAAGATAATGGGATTCAGAACTCAG		
SERPINA1	chr14	94439180-94439625	-/+	15	GTGCAGCCTTCATGGTTTCG / CCACAGAGAGCATCGCAAGA		
B2M	chr15	44723613-44724059	-/+	15	TTGTCTGTGATGTAGCCATCA / TGTCAACAACCACTTTCACGG		
AXL	chr19	40978400-40978829	+/-	15	GAATAGCACTCCTCCACAGGG / TCTACCTGAATTCTGGAACAGCCG		

The start and end positions of the target regions are shown with those in rCRS (the reference mtDNA) and nuclear genome (assembly GRCh38). The ligation arm of B10 was designed with a degenerate base S (G/C) to match the mtDNA sequence with an 8271-8279 or 8281-8289 deletion (i.e., in the Asian mtDNA haplogroups B2 or B4).

*mtDNA polymorphisms with a population frequency >1% in the European macro-haplogroups (H, U, J, T, K, R, V, I, W, X, and N) were obtained from the MITOMAP website (https://www.mitomap.org/MITOMAP); the polymorphisms are shown with the first letter indicating the reference allele of rCRS, followed by the position and the variant allele.

[†]Indicates the locations of the nuclear genome regions that cover >85% of the target mtDNA region and have a sequence similarity of >85% to the target region; the number of substitutions (ALT) and polymorphisms (SNP) present in each region according to rCRS are provided in the parentheses; only polymorphisms with a minor allele frequency >1% in the 1000Genomes project (retrieved from the commonSNP147 track from the UCSC genome browser) were counted.

Table S10. Quality filters used to reduce false positive errors in calling mtDNA variants.

Categories	Description	Filter/tool parameters
Alignment errors	1) Paired-end reads do not contain proper molecular barcode information	≥ 9 bases with BAQ (base alignment quality)≥15; bases in the barcode with BAQ<15 are masked as "N" and are not used in comparing barcode sequences between two paired-end reads.
	2) Paired-end reads do not contain arm sequences matching those of the 46 mtDNA and 5 nuclear DNA EL (extension-ligation) probe pairs.	≤ 3 mismatches between the detected E/L arm sequences and the designed E/L arm sequences
	3) Paired-end reads are not aligned to the target region or the correct strand according to the arm sequences identified.	using command "bwa mem -L 100, 5 -M genome_reference.fa"; the argument "-L 100,5" disables soft clips following the trimmed probe arm sequences; reads retained should have mapping quality MAPQ (mapping quality)≥20 and correct alignment locations as per EL probe design.
	4) Properly aligned paired-end reads are locally realigned and base qualities are recalibrated.	Using command "bamleftalign -c -f | samtools calmd -EArb"
	5) Consensus reads are annotated as NUMTS (nuclear mitochondrial DNA segment) in the alignment file.	Consensus reads are not constructed from paired-end reads aligned to nuclear DNA in the complete reference genome (MAPQ≥10); consensus read sequences are more similar to the individual's major mtDNA sequence than a collection of NUMTS sequences provided.
	6) Consensus reads show an excess of mismatches compared to the individual's major mtDNA sequence	≤ 5 mismatches in the coding region of mtDNA and ≤ 8 in the D-loop region of mtDNA
Low read depth	1) Samples have insufficient read coverage on mtDNA for heteroplasmy calling.	After base summarization using command "samtools mpileup -q 20 -Q 0 -B -d 500000 -f mtdna.fa", median coverage of high-quality reads (BAQ≥30) on mtDNA are > 1000X.
	2) Sites have insufficient or low percentage of high-quality reads for heteroplasmy calling.	1) sites with coverage of high-quality reads (BAQ≥30) ≥100X, and ≥500X when they are used to compare allele fractions between two samples; 2) sites having the percentage of high-quality reads (BAQ≥30) ≥70%; 3) sites not located in low-complexity regions of mtDNA or at low-quality sites (nt 302-316, nt 512-526, nt 16184-16193, and nt 545, 16224, 16244, 16249, 16255, and 16263).

Sequencing or PCR errors	1) Minor alleles are rare among consensus reads.	≥ 5 minor alleles from high-quality reads (BAQ≥30)
	2) Allele fractions detected are significantly different between consensus reads constructed from single and duplicate paired-end reads.	1) Fisher's exact test $P \geq 10^{-4}$ or fold change <5 of VAFs detected between consensus reads constructed from single paired-end reads and from duplicate paired-end reads; 2) for variants of VAF <1%, a VAF of ≥0.2% should be detected among consensus reads constructed from duplicate paired-end reads.
	3) VAF (variant allele fraction) detected is indistinguishable from the probability of errors assigned by BAQ	Log likelihood quality score of the variant > 5 computed with read BAQ
Background errors	1) VAF detected is indistinguishable from the background error rate of STAMP	Exact Poisson test, $P<6\times10^{-7}$(0.01/16569) for observing fewer minor alleles than errors that occur at a rate of 0.02% in STAMP; the corresponding *P* is <0.001 when it is used to determine the existence of a heteroplasmy at a certain site in a sequence replicate or another sample of the same individual.

Table S11. Associations of the expansion of predicted pathogenic mtDNA heteroplasmies in blood with HD clinical phenotypes after including very-low-fraction heteroplasmies.

Variables	mtDNA variant pathogenicity (N)*	TFC score		Total motor score		SDMT score	
		Beta (SE)	*P*	Beta (SE)	*P*	Beta (SE)	*P*
Model 1 (baseline phenotypes)	M/H (240)	-0.009 (0.018)	0.62	0.002 (0.004)	0.60	-0.001 (0.005)	0.80
	H (183)	-0.005 (0.021)	0.81	0.003 (0.004)	0.44	-0.002 (0.006)	0.69
	others (684)	-0.003 (0.013)	0.84	-0.002 (0.002)	0.38	0.006 (0.003)	0.077
Model 2 (follow-up phenotypes)	M/H (240)	-0.061 (0.019)	**0.0012**	0.011 (0.003)	**0.00042**	-0.029 (0.008)	**0.00040**
	H (183)	-0.059 (0.022)	**0.0072**	0.014 (0.004)	**0.00023**	-0.030 (0.009)	**0.0015**
	others (684)	-0.005 (0.012)	0.66	-0.001 (0.002)	0.53	-0.006 (0.005)	0.25

The associations were assessed by using the linear mixed-effects model: $\log_2$(VAF follow-up/ VAF baseline) ~ score + age + sex + CAG_length + followup_duration + (1|patient_id). In the analyses of the follow-up phenotype, the baseline phenotype and disease stage were further considered as additional fixed-effect covariates in the model.

*The number of mtDNA heteroplasmies used for analysis; the 240 pathogenic heteroplasmies were detected with VAF≥0.25% and were shared between the baseline sample and the follow-up sample of the same patient at VAF≥0.1%; due to missing phenotypes in either the baseline or the follow-up samples, 2 pathogenic heteroplasmies and 8 non-pathogenic heteroplasmies were not included in the analyses of total motor scores, and 49 pathogenic heteroplasmies and 123 non-pathogenic heteroplasmies were not included in the analyses of SDMT scores.

M/H: medium or high pathogenicity; H: high pathogenicity; others: not predicted with medium or high pathogenicity in HD patients carrying pathogenic heteroplasmies. *P* values <0.05 are highlighted in bold type.

Table S12. Summary of qPCR experiments for mtDNA content in REGISTRY samples.

Variables	QC criteria	HD lymphoblast samples	Control lymphoblast samples	Blood samples*
No. of samples with qPCR (2 duplicated reactions for mtDNA and 2 for nDNA)	-	1474	126	343 (165)
No. of sample that passed QC filters of qPCR experiments	C_T values from all 4 reactions and <3 cycles between duplicates	1285	116	318 (153)
No. of sample with both STAMP-CN and qPCR-CN	≥5X for each of the three probes targeting chromosomes 8, 14, 19 in STAMP	1118	116	318 (153)

*The numbers of individuals with both baseline and follow-up blood samples included are indicated in parentheses.

Table S13. Distribution of mtDNA macro-haplogroups in REGISTRY samples.

Ethnicity† / Haplo-groups*		Caucasian (%)	African North (%)	African American (%)	American Latino (%)	Asian West (%)	Asian East (%)	Others or unknown (%)
Africa	**L**	16 (0.9)	0 (0)	2 (100)	1 (33.3)	0 (0)	0 (0)	2 (10.5)
Asia	**A**	1 (0.1)	0 (0)	0 (0)	1 (33.3)	0 (0)	2 (25.0)	0 (0)
	B	1 (0.1)	0 (0)	0 (0)	1 (33.3)	0 (0)	0 (0)	0 (0)
	C	2 (0.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	D	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (12.5)	1 (5.3)
	F	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (25.0)	0 (0)
Eurasia	**N**	13 (0.7)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	M	4 (0.2)	0 (0)	0 (0)	0 (0)	2 (33.3)	2 (25.0)	2 (10.5)
	R	5 (0.3)	0 (0)	0 (0)	0 (0)	2 (33.3)	0 (0)	2 (10.5)
Europe	**H**	823 (46.9)	3 (50.0)	0 (0)	0 (0)	0 (0)	0 (0)	4 (21.1)
	I	37 (2.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	J	156 (8.9)	0 (0)	0 (0)	0 (0)	2 (33.3)	1 (12.5)	2 (10.5)
	K	142 (8.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	T	178 (10.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	U	256 (14.6)	1 (16.7)	0 (0)	0 (0)	0 (0)	0 (0)	2 (10.5)
	V	64 (3.6)	1 (16.7)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	W	27 (1.5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (5.3)
	X	29 (1.7)	1 (16.7)	0 (0)	0 (0)	0 (0)	0 (0)	3 (15.8)
All		1754 (97.6)	6 (0.3)	2 (0.1)	3 (0.2)	6 (0.3)	8 (0.4)	19 (1.1)

*The continental origin of each mtDNA macro-haplogroup was obtained from the MITOMAP website (http://www.mitomap.org/MITOMAP).

†Self-reported ethnicity or race in the REGISTRY database.

Table S14. No significant associations of mtDNA macro-haplogroups with HD.

mtDNA haplogroups	Frequency (%)*	Reference haplogroup(s)	Control (N=184) vs HD (N=1614)		Control (N=184) vs prodromal/early-stage HD (N= 954)†		Control (N=184) vs middle/late-stage HD (N=635)†	
			OR [95% CI]	*P*	OR [95% CI]	*P*	OR [95% CI]	*P*
H	46.2	H	-	-	-	-	-	-
		Others	1.1 [0.8-1.5]	0.44	1.2[0.9-1.7]	0.20	1.0[0.7-1.4]	0.81
U	14.4	H	0.8 [0.5-1.3]	0.35	0.7[0.5-1.2]	0.21	1.0[0.6-1.6]	0.84
		Others	0.8 [0.6-1.3]	0.40	0.8[0.5-1.2]	0.32	0.9[0.6-1.5]	0.70
T	9.9	H	1.2 [0.7-2.1]	0.59	1.0[0.6-1.9]	0.89	1.5[0.8-2.8]	0.21
		Others	1.3 [0.7-2.2]	0.39	1.2[0.7-2.1]	0.58	1.5[0.8-2.8]	0.17
J	9.0	H	0.7 [0.4-1.2]	0.23	0.6[0.4-1.1]	0.10	0.9[0.5-1.7]	0.83
		Others	0.7 [0.5-1.2]	0.25	0.7[0.4-1.1]	0.14	0.9[0.5-1.6]	0.72
K	7.9	H	0.8 [0.4-1.4]	0.37	0.7[0.4-1.3]	0.25	0.9[0.5-1.6]	0.68
		Others	0.8 [0.5-1.4]	0.44	0.8[0.4-1.4]	0.38	0.8[0.5-1.5]	0.56
V	3.6	H	1.2 [0.5-3.2]	0.65	1.3[0.5-3.5]	0.54	1.1[0.4-3.2]	0.83
		Others	1.4 [0.5-3.5]	0.52	1.5[0.6-4.0]	0.37	1.1[0.4-3.2]	0.80
I	2.1	H	1.3 [0.4-4.2]	0.70	1.3[0.4-4.5]	0.67	1.2[0.3-4.7]	0.74
		Others	1.4 [0.4-4.6]	0.58	1.5[0.4-5.1]	0.52	1.3[0.3-5.0]	0.70
X	1.9	H	3.7 [0.5-27.4]	0.20	3.0[0.4-23.3]	0.29	4.7[0.6-37.1]	0.14
		Others	4.0 [0.5-29.9]	0.17	3.5[0.5-26.4]	0.23	5.0[0.6-38.9]	0.13
W	1.6	H	0.8 [0.2-2.8]	0.76	0.7[0.2-2.5]	0.58	1.0[0.3-3.7]	0.95
		Others	0.8 [0.2-2.6]	0.68	0.7[0.2-2.5]	0.59	0.8[0.2-3.1]	0.77
L	1.2	H	0.4 [0.1-1.3]	0.12	0.5[0.2-1.6]	0.24	0.2[0.0-1.0]	0.05
		Others	0.5 [0.1-1.4]	0.17	0.6[0.2-1.8]	0.33	0.2[0.0-1.0]	0.05
N, M, or R	1.8	H	1.0 [0.3-3.4]	0.97	0.9[0.3-3.2]	0.90	0.9[0.2-3.4]	0.86
		Others	1.1 [0.3-3.7]	0.88	1.0[0.3-3.7]	0.94	0.9[0.2-3.3]	0.82
A, B, C, D, or F	0.7	H	0.6 [0.1-2.7]	0.49	0.4[0.1-2.2]	0.28	1.1[0.2-6.1]	0.91
		Others	0.7 [0.1-3.1]	0.59	0.5[0.1-2.5]	0.37	1.3[0.2-7.0]	0.79

*Frequency in all lymphoblasts (N=1798).

†25 individuals not included due to missing TFC, total motor or diagnosis confidence scores.

OR: odds ratios are shown for the risks of HD computed using logistic regression with adjustment for individual age and sex.

Table S15. Identical mtDNA haplogroups detected in lymphoblast and blood samples of the same individuals.

Individual ID	Lymphoblast samples	Blood samples (baseline)	Blood samples (follow-up)
10007	H1b	H1b	H1b
10016	H7d3a	H7d3a	H7d3a
10019	U5b2a1a1	U5b2a1a1	U5b2a1a1
10022	H3b+16129	H3b+16129	H3b+16129
10023	H1c+152	H1c+152	H1c+152
10038	H3+152	H3+152	H3+152
10044	H3h	H3h	H3h
10052	H13a1a1a	H13a1a1a	H13a1a1a
10106	H16+152	H16+152	H16+152
10112	H13b1+200	H13b1+200	H13b1+200
10113	H8b	H8b	H8b
10123	not available	J1c3a2	J1c3a2
10127	J1c6	J1c6	J1c6
10157	K1a	K1a	K1a
10161	H24a	H24a	H24a
10168	H2a3a	H2a3a	H2a3a
10173	U4c1	U4c1	U4c1
10175	H73a1	H73a1	H73a1
10182	X2l	X2l	X2l
10187	H2a1	H2a1	H2a1
10197	H5b1	H5b1	H5b1
10201	T2b33	T2b33	T2b33
10236	U5b2a3a	U5b2a3a	U5b2a3a
10250	H6a1b4	H6a1b4	H6a1b4
10263	U5b2b4	U5b2b4	U5b2b4
10266	K1	K1	K1
10267	T2b4	T2b4	T2b4
10286	H7b	H7b	H7b
10287	J1c9	J1c9	J1c9
10300	HV1a1a	HV1a1a	HV1a1a
10302	H6a1b4	H6a1b4	H6a1b4
10309	H5	H5	H5
10324	H	H	H
10329	V15a	V15a	V15a
10364	T2b	T2b	T2b
10365	H+16129	H+16129	H+16129
10366	H2a5b2	H2a5b2	H2a5b2
10369	J1c2	J1c2	J1c2
10372	H	H	H
10374	H1g1	H1g1	H1g1
10403	H1n1	H1n1	H1n1
10409	H17b	H17b	H17b
10427	H1ab1	H1ab1	H1ab1
10441	V11	V11	V11
10453	HV13	HV13	HV13
10462	C5c1a	C5c1a	C5c1a
10473	H1b1a	H1b1a	H1b1a
10493	J1c2t	J1c2t	J1c2t
10501	H105	H105	H105
10523	H1+152	H1+152	H1+152

10532	K1a4a1	K1a4a1	K1a4a1
10564	H	H	H
10577	H8c	H8c	H8c
10591	U2e1f1	U2e1f1	U2e1f1
10597	H17	H17	H17
10602	H5a1c2	H5a1c2	H5a1c2
10611	H28	H28	H28
10613	K1a5	K1a5	K1a5
10626	H11a	H11a	H11a
10632	H64	H64	H64
10659	H1j8	H1j8	H1j8
10662	U8a1a1a	U8a1a1a	U8a1a1a
10669	H+152	H+152	H+152
10675	V	V	V
10676	H+16291	H+16291	H+16291
10687	H13a1a1a	H13a1a1a	H13a1a1a
10697	W1e1	W1e1	W1e1
10706	H3v+16093	H3v+16093	H3v+16093
10708	H52	H52	H52
10712	H1a1	H1a1	H1a1
10735	H13a1a1a	H13a1a1a	H13a1a1a
10748	T2b21	T2b21	T2b21
10758	H1g1	H1g1	H1g1
10771	V18a	V18a	V18a
10781	J1c3	J1c3	J1c3
10782	V+@72	V+@72	V+@72
10795	K1a1b1g	K1a1b1g	K1a1b1g
10797	H3ap	H3ap	H3ap
10799	H1+16189	H1+16189	H1+16189
10807	H67a	H67a	H67a
10811	H28a	H28a	H28a
10816	V	V	V
10837	H	H	H
10840	U5b1b1a	U5b1b1a	U5b1b1a
10845	H1+16239	H1+16239	H1+16239
10847	J1c5	J1c5	J1c5
10854	I4a	I4a	I4a
10855	T2b	T2b	T2b
10865	H3ap	H3ap	H3ap
10874	H43	H43	H43
10909	U8a1a4	U8a1a4	U8a1a4
10914	U2e1a1	U2e1a1	U2e1a1
10929	T2g	T2g	T2g
10950	H1a1	H1a1	H1a1
10955	H1as	H1as	H1as
10963	H1+16189	H1+16189	H1+16189
10997	J1b1a1	J1b1a1	J1b1a1
10999	U2e1g	U2e1g	U2e1g
11002	H1c3	H1c3	H1c3
11007	R1a1	R1a1	R1a1
11016	H11a2	H11a2	H11a2
11021	H1c	H1c	H1c
11048	K1a5	K1a5	K1a5

11063	H18	H18	H18
11066	H5a1q	H5a1q	H5a1q
11074	H2a1a	H2a1a	H2a1a
11105	H2a1b2	H2a1b2	H2a1b2
11125	H1g1	H1g1	H1g1
11131	V	V	V
11146	J1c6	J1c6	J1c6
11147	U4b	U4b	U4b
11151	N1a1a1a2	N1a1a1a2	N1a1a1a2
11169	H6a1a5	H6a1a5	H6a1a5
11170	U5b2b1a1	U5b2b1a1	U5b2b1a1
11182	K1a4b	K1a4b	K1a4b
11194	T2b	T2b	T2b
11196	H2a2a1	H2a2a1	H2a2a1
11203	H7d	H7d	H7d
11216	J1c16	J1c16	J1c16
11220	H4a1a4b	H4a1a4b	H4a1a4b
11229	J1c	J1c	J1c
11232	L3b1a+@16124	L3b1a+@16124	L3b1a+@16124
11247	U2e1 23	U2e1 23	U2e1 23
11249	H	H	H
11254	V14	V14	V14
11267	T2c1d1	T2c1d1	T2c1d1
11294	K2a10	K2a10	K2a10
11304	H1c3	H1c3	H1c3
11310	H2a	H2a	H2a
11318	H1j4	H1j4	H1j4
11328	K1a4a1	K1a4a1	K1a4a1
11330	T2b1	T2b1	T2b1
11344	HV23	HV23	HV23
11361	W3a1	W3a1	W3a1
11364	H55b	H55b	H55b
11375	H10	H10+(16093)	H10+(16093)
11377	J1b1a1b	J1b1a1b	J1b1a1b
11397	V	V	V
11414	T2b	T2b	T2b
11419	T2a1a	T2a1a	T2a1a
11427	T2b	T2b	T2b
11435	H1c22	H1c22	H1c22
11461	W5a	W5a	W5a
11472	H5a1	H5a1	H5a1
11476	H1c	H1c	H1c
11484	H1b1+16362	H1b1+16362	H1b1+16362
11490	H10	H10	H10
11492	T2c1a2	T2c1a2	T2c1a2
11498	U5a2a1	U5a2a1	U5a2a1
11506	K1c1b	K1c1b	K1c1b
11508	H1e1	H1e1	H1e1
11511	U5a1d2a1	U5a1d2a1	U5a1d2a1
11519	J1b1a1e	J1b1a1e	J1b1a1e
11533	H9a	H9a	H9a
11562	K1a4a1a2b	K1a4a1a2b	K1a4a1a2b
11591	H3+16189	H3+16189	H3+16189

11597	H14	H14	H14
11601	H1b1+16362	H1b1+16362	H1b1+16362
11617	K1c1	K1c1	K1c1
11619	H11a1	H11a1	H11a1
11622	U5a1b	U5a1b	U5a1b
11626	J1c5	J1c5	J1c5
11638	I1b	I1b	I1b
11648	H7c3	H7c3	H7c3
11658	K1a	K1a	K1a
11660	H3ap	H3ap	H3ap
11668	J1c	J1c	J1c
11679	J1c	J1c	J1c
11683	H6a1a	H6a1a	H6a1a
11684	T2b28	T2b28	T2b28
11693	H1+16311	H1+16311	H1+16311
11694	H1q2	H1q2	H1q2
11715	H1a2	H1a2	H1a2
11761	H1+152	H1+152	H1+152
11780	U5a1g	U5a1g	U5a1g
11798	T2b	T2b	T2b
11811	H10e	H10e	H10e
11814	H1u2	H1u2	H1u2
11819	H8c2	H8c2	H8c2
11837	T2c1d1	T2c1d1	T2c1d1
11841	H2a1e1a	H2a1e1a	H2a1e1a

Table S16. Common minor alleles of consensus reads identified in blood samples.

ID	EL probe	Substitutions*	References†
chr11:10509739	A2	754G,756T,773C,804T,813G,825A,843C,859C,930A,(932G),936A,979T,984G,1009T,1018A,1038T,1039G,1040C,1052G,1057A,1106T	GRCh38 (hg38)
chr5:80651208	A3	(1120T),1284C,1292G,1322T,1348C,1376T,1377T,1393A,1405T,1451G	GRCh38 (hg38)
chr11:10508971	A4	1520C,1556T,1654C,1673C,1693T,1708G,1709A,1713G,1715T,1719A,1725T,1761T,1764A,1765T,1766C,1808G,1809C,1824C,1837T,1842G,(1883A)	GRCh38 (hg38)
chr3:96618018	A4	1518T,1554A,1579G,1619T,1654C,1664A,1673C,1685T,1692G,1693T,1709A,1711T,1719A,1761T,1764A,1766C,1824C,1842G,1849T,(1883A)	GRCh38 (hg38)
chr5:80650830	A4	1536G,1556T,1619T,1654C,1664A,1693T,1709A,1719A,1733T,1761T,1764A,1766C,1842G,1883A	GRCh38 (hg38)
chr11:10508598	A5	(1888A),1977C,1978G,(2001T),2005T,2030C,2056A,2071C,2080C,(2113C),2143A,2162T,2168C,(2169C),2213G,2219T,2220G,2221T,2222C,2224T,2242C	GRCh38 (hg38)
chr3:96617645	A5	1900G,2000T,2056A,2059T,2143A,2162T,2221T,2222C,2223G,2224T,2244C,2251G,(2257T)	GRCh38 (hg38)
chr5:80650457	A5	1944T,1977C,2005T,2056A,2071C,2080C,2143A,(2157C),2162T,2219T,2221T,2222C,2224T,2226C,2227G,2245G	GRCh38 (hg38)

chr11:10507904	A7	(2584T),2625T,2667C,2707G,2708T,2710T,2712A,2744C,2750C,(2758A),(2759C),(2762T),(2769G),(2770T),2778C,2788T,2798G,2800G,2831C,2834T,2837G,2849A,2850C,2857C,2858G,2881T,2885C,2887C,2888G,2889T,2891T	GRCh38 (hg38)
pnumts_A7_5	A7/A6	2623G	
pnumts_A9_6	A9		Dayama, G., et al. *Nucleic Acids Research* (2014)
pnumts_A10_1	A10		Dayama, G., et al. *Nucleic Acids Research* (2014)
chr1:629218	A11	4104G,4312T,4318T	GRCh38 (hg38)
pnumts_A11_2	A11	4216C	
chr1:629541	A12	4456T,(4464A),4736C	GRCh38 (hg38)
pnumts_A12_3	A12	4560A	
chr1:629896	B1	4736C,4769G,4856C,(4869G),4904T,4914T,4940T,(4958G),4991A,(5041C)	GRCh38 (hg38)
pnumts_B1_7	B1	4769G	Dayama, G., et al. *Nucleic Acids Research* (2014)
chr1:630295	B2	5147A,5320T,5351G,(5385T),5387T,(5426C),(5437T)	GRCh38 (hg38)
chr1:630644	B3	5498G,5580C,5821A,5840T	GRCh38 (hg38)
chr1:631038	B4	6023A,6221C,6242T	GRCh38 (hg38)
chr1:631410	B5	6266C,6299G,6366A,6383A,6410T,6452T,6483T,6512C,6542T,6569A	GRCh38 (hg38)
chr1:631710	B6	(6542T),6569A,6641C,(6935T)	GRCh38 (hg38)
chr17:53105850	B7	6938T,6944C,6950T,6956C,6962A,7013A,7022C,7040C,7064C,7076G,7133T,7145T,7146G,7169C,7196T,7205T,7211A,7256T,7286C	GRCh38 (hg38)
chr1:632106	B7	7146G,(7195C),(7197A),(7216A),7232T,7256T,7316A	GRCh38 (hg38)
pnumts_B7_8	B7	7028T	Dayama, G., et al. *Nucleic Acids Research* (2014)
chr17:22528661	B8	(7256T),(7259T),(7260T),7283C,7286C,7291A,7299G,7301G,7302C,7316A,7318C,7319C,7325G,7326A,(7327A),7337A,(7356A),(7357A),(7358T),(7361G),(7362C),(7367T),(7372C),(7373T),(7379A),(7385G),(7388G),(7391G),(7400T),(7412T),(7418T),(7419A),(7427A),(7430C),7440C,7462T,7468T,7471T,7474G,7490G,7493T,7497A,7498A,7501C,7503T,7517C,7521A,7528G,7534T,7547C,7559G,7563C,7567T,7572C,7575C,7576G,7594C,7600C,7603G,7604C,7612T,7621C,7624A,7645A,7649G	GRCh38 (hg38)
chr1:632425	B8	7256T,7316A,7521A	GRCh38 (hg38)
chr3:120722126	B8	(7256T),(7259T),(7260T),7283C,7286C,(7287T),(7291A),(7299G),(7301G),(7302C),7316A,7318C,7319C,7325G,(7326A),(7327A),(7331T),7337A,(7340A),(7341T),(7343G),(7349T),(7356A),(7357A),(7358T),(7361G),(7362C),(7367T),(7372C),(7373T),(7377T),(7379A),(7385G),(7386T),(7388G),(7390G),(7391G),(7394C),(7400T),(7412T),(7415G),(7418T),(7419A),(7427A),(7430C),7440C,(7447G),(7462T),(7463C),(7468T),7471T,(7474G),(7475A),(7490G),7493T,7497A,7498A,(7501C),7503T,7517C,7521A,(7528G),(7533T),7534T,(7547C),7559G,7563C,7567T,(7571G),(7572C),7575C,(7576G),7594C,7600C,(760	GRCh38 (hg38)

		3G),7604C,(7609C),7612T,(7618T),7621C,7624A,(7633A),(7635C),7645A,7649G	
chr5:100053474	B8	(7256T),(7259T),7280T,7286C,7302C,7313T,7316A,(7327G),7337A,7340A,7348C,7356A,7358C,7362C,7372C,7373C,7385G,7399T,7400T,7409T,7412T,7415G,7428A,7440C,7447G,(7462G),(7463A),(7468T),7471T,7496C,7497A,7503T,7521A,7559G,7563C,7567T,7572C,7575C,7576G,7600A,(7627T),(7642A),(7645A)	GRCh38 (hg38)
pnumts_B8_9	B8	7299G,7325G,7364G,7473G,7521A,7559G,7610T	Dayama, G., et al. *Nucleic Acids Research* (2014)
chr1:632801	B9	7650T,(7663T),7705C,(7757A),7810T,(7861C),7868T,7891T,(7900T),7912A,(7927T),(8011G)	GRCh38 (hg38)
pnumts_B9_10	B9		Dayama, G., et al. *Nucleic Acids Research* (2014)
chr1:633075	B10	7912A,(7927T),(8011G),8021G,(8038C),(8059T),8065A,(8080T),(8093C),(8114A),(8119C),(8122G),8140T,8152A,(8158G),8167C,8203T,(8206A),(8251A),(8254T)	GRCh38 (hg38)
chr1:633441	B11	8392A,8455T,8461T,8503C,8545A	GRCh38 (hg38)
chr1:634102	C1	8943T,9060A,9075T,9168T,9254G,(9325C)	GRCh38 (hg38)
chr1:634464	C2	9325C,9329C,(9384A),9434G,9527T,9530C,9540C,9545G,9548A,9629G	GRCh38 (hg38)
chr5:134927794	C6	10646A,10652C,10670T,10677A,10679G,10685A,10688A,10721G,10750G,10774T,(10785C),10810C,10846T,10866C,10873C,10885C,10915C,10919T,10920T,10922T,10927C,10945G,10978G	GRCh38 (hg38)
chrX:126472749	C6	10646A,10652C,10653A,10670T,10677A,10679G,(10680A),10685A,10688A,10715T,10750G,10775A,10786C,10801A,10808T,10810C,10822T,10846T,10866C,10873C,10915C,10920T,10922T,10927C,10945G,10975T	GRCh38 (hg38)
chr5:100049779	C7	10954T,10962G,10963G,10966A,10986T,10993A,11009C,11013T,11016A,11017C,11023C,11045G,11061T,(11063T),11065C,11070A,11107T,11113C,11116C,(11129G),11147C,11149A,11155T,(11176A),(11177A),(11179G),(11185T),(11188A),11203T,11212T,11215T,11224G,11233C,11249T,11254C,11255C,11260A,11272G,11284T,11288T,11291T,11299C,11302T,11314G,(11325G),(11332T),(11334G),(11335T)	GRCh38 (hg38)
chr5:134927453	C7	(10945G),10978G,11016A,(11083G),11097T,11147C,11176A,11197T,11233C,11254C,11260C,11281G,11284T,11291T,11302T,11335T	GRCh38 (hg38)
chr5:134927077	C8	11335T,11347G,11353G,11377A,(11383G),11392T,11399C,11404G,11455T,11471T,11527T,11557G,11590G,11662C,11708G	GRCh38 (hg38)
chr5:100048645	C10	12106T,12112T,12115T,12127T,12131A,12136C,12178T,12189C,12192A,12193G,12218T,12236A,12238T,12265G,12285C,12290G,12346T,12348T,12349G,12351C,12354C,12362T,12367G,12372A,12379T,(12390G),12394A,12397G,12403T,12406A,12408C,12423G,12425G,12426T,12438C,12450A,12454A,12456T,12463G	GRCh38 (hg38)
chr5:134926319	C10	12115T,12136C,12189C,(12215G),12218T,12236A,12237T,12285C,12290G,12346T,12349G,12358C,12367G,12372A,12379T,12390T,12406A,12417T,12432T,12441A,12454A	GRCh38 (hg38)
pnumts_C11_11	C11	12684A,12705T	Just, R.S., et al. *Forensic Sci. Int. Genet.* (2015)
chr5:134925549	C12	12873C,12888T,12892C,12904C,12940A,12945C,12950G,12951T,12982T,(13011T),13020C,13023T,13062T,13105G,13111C,13140G,13145A,13164C,13174C,(13207C),13242G	GRCh38 (hg38)

chr5:134925156	D1	13260C,13272T,(13276G),13281C,13359A,13368A,13386C,13440T,13466A,(13476G),13488C,13563G,13581C,13638G	GRCh38 (hg38)
chr5:134924811	D2	13638G,13650A,13651G,13656C,13674C,13707A,13708A,13711A,13712T,13725T,13731G,(13740C),(13743C),(13748G),(13753C),(13754T),(13759A),(13765A),(13768C),(13769C),(13775T),(13776G),13785A,13788T,13792T,13809A,13811G,13812C,13820C,13845T,13869C,13887G,13889A,13890T,13899C,13905T,13908T,13920T,13929T,13934T,13945G,13950T,13968A,13980A	GRCh38 (hg38)
pnumts_D8_14	D8		Dayama, G., et al. *Nucleic Acids Research* (2014)
pnumts_D8_15	D8	16129A	Dayama, G., et al. *Nucleic Acids Research* (2014)
pnumts_D9_16	D9	16218T,16230G,16249C,16259A,16264T,16274A,16278T,16284G,16288C,16290T,16293C,16301T,16311C,16355T,16356C,16368C,16390A,16399G,16444T,16496A,16519C,16527T	Dayama, G., et al. *Nucleic Acids Research* (2014)
pnumts_D10_13	D10	(73G),263G	Dayama, G., et al. *Nucleic Acids Research* (2014)

The 52 minor alleles of the consensus reads identified in blood samples of the current study are listed in order of EL probes.

*Nucleotide substitutions and their positions in the Revised Cambridge Reference Sequence (rCRS); substitutions not present in all the samples detected with this minor allele are shown in parentheses.

†refer to the source of the NUMTS annotations; GRCh38 (hg38): human reference genome.

Table S17. Correlations between consensus reads captured with mtDNA and nuclear DNA probes.

Source	Nuclear DNA (nDNA) target regions*	C	R^2	No. of samples with < 5 consensus reads
Lymphoblast	chr1 (*EMC1*)	0.30	0.37	20
	chr8 (*WRN*)	0.52	0.72	18
	chr14 (*SERPINA1*)	0.47	0.59	1
	chr15 (*B2M*)	0.41	0.49	29
	chr19 (*AXL*)	0.49	0.56	3
	chr8, chr14, and chr19	0.53	0.66	20
Blood	chr1 (*EMC1*)	0.32	0.36	1
	chr8 (*WRN*)	0.60	0.61	0
	chr14 (*SERPINA1*)	0.65	0.68	0
	chr15 (*B2M*)	0.55	0.56	1
	chr19 (*AXL*)	0.64	0.69	0
	chr8, chr14, and chr19	0.66	0.69	0

C and R2 were obtained from linear regression: $\log_2$(No. of mtDNA consensus reads) ~ C×$\log_2$(No. of nDNA consensus reads from the nDNA target region(s) indicated). mtDNA consensus read numbers were estimated using reads from 18 mtDNA probes (A5-A8, B2, B6, B7, B9, C1-C5, C7, C9, C12, D1, D5) that lack common polymorphisms in their arm regions in European populations and showed relatively low variations in consensus read coverage across samples of the current study. *P* values for $C < 2.2\times10^{-16}$.

*Denote the chromosome and the nearest gene in parentheses.

Supplemental Data (Titles)

Data S1. mtDNA heteroplasmies identified in HD and control lymphoblasts.

Data S2. mtDNA heteroplasmies identified in longitudinal blood samples of HD patients.