An atlas of mitochondrial DNA genotype-phenotype associations in UK Biobank

Table of Contents

Supplementary Notes

Nuclear genome quality control and principal component analysis

A genetically European group of UK Biobank participants was defined using principal component analysis (PCA, FlashPCA2)[1] calculated on SNPs (MAF > 0.05) with: overall call rate ≥ 99%, HWE P ≥ 10^{-5}, and $R^2 < 0.2$. Regions of the genome known to exhibit long-range linkage disequilibrium (LD) were removed (chr6: 25–33.5Mb, chr8: 8–12Mb, chr17: 40.4–42.4Mb) to ensure the nuclear principal component (nucPCs) were picking up ancestry and not LD. Then, the method of Astle et al.[2] was adopted to identify ancestral outliers for removal. This was followed by batch level variant and sample quality control (QC) and QC over all batches. During the former, variants were removed if: call rate ≤ mean (call rate) - [3 x SD (call rate)]; HWE *P*-value < $1x10^{-12}$ (MAF < 0.01) or HWE *P*-value < $1x10^{-6}$ (MAF ≥ 0.01), and individuals were removed if: call rate < mean (call rate) - [3 x SD (call rate)] or Heterozygosity > (mean +/- 3SD). During the latter, variants that failed QC in > 48 batches (UKBB array) or > 3 batches (UKBL array), and individuals who's genetic sex and phenotypically defined sex did not match, were excluded.

After variant and sample QC across all batches, a second PCA was performed (FlashPCA2) as above and genetic distance measure of 0.175 calculated using the first 8PCs was used to remove additional individuals of non-European ancestry. A subset of unrelated participants was defined using the kinship information provided by UK Biobank that lists the kinship coefficient of pairs of individuals up to 3rd degree relatives[3]. Pairs that shared individual(s) were aggregated into families and only the individual with the highest call rate from each family was retained. The nucPCs used in the present article were calculated on this final set of variants (N=782,205) and individuals (N=449,771) as above.

Design of a bespoke workflow for quality control, re-calling and imputation of mitochondrial variants

The workflow we designed (**Fig. 1**) consists of four stages: 1. pre-recalling QC, 2. manual re-calling, 3. pos-recalling and 4. imputation; and makes use of the files routinely provided after standard genotyping: 1. file with genotypes per individual (plink, vcf or oxford formatted), 2. intensity files (either pre-extracted or CEL files), 3. array manifest files

which contain information on the probes and the reference sequence (rCRS; NC_012920) to which those were mapped (including strand information). In addition, we used reference panels of whole mitochondrial genomes for imputation purposes and to estimate allele frequency in reference populations.

Stage 1: Pre-re-calling QC procedures

First, within each genotyping array, we determined the per mtSNV and per sample call rates (PLINK 1.9[4]). Similar to the nuclear genome, number of missing genotypes per variant was low (median call rate of 0.998 and 0.997 for the UKBB array and the UKBL array, respectively), and, for each variant, sample were missing at random. However, we observed that the per-sample call rate was much lower in comparison to the one observed for autosomal variants (mean of 0.89 and 0.96, respectively). Therefore, at that stage, we selected for recalling variants with call rate <0.990 (N=7) and variants with lower call rates than in the 150,000 UK Biobank release (N=8). Next, we excluded 54 samples who were outliers in terms of mean intensities over all SNVs, irrespective of batch, as well as 2,054 samples, because of plate effects (**Supplementary Fig. 3 and 4**). Next, in the remaining individuals, we calculated allele frequencies and compared those to allele frequencies in three reference datasets (GenBank, 1000 Genomes, WTCCC). These datasets are not ideal references due to either size (1000 Genomes, WTCCC) or origin (a.k.a. majority of sample are non-British: 1000 Genome, GenBank). The GenBank is the biggest cohort available and, consequently, it is also the most heterogeneous. Therefore, variants with discordant MAF (> 3%), compared to at least one of the reference data sets, were selected for re-calling, only if they also showed suboptimal clustering. Finally, to address the issue of wrongly assigned calls, cluster plots for each variant were produced and visually inspected. This was done per array or, when more resolution was needed to determine cluster edges, per-batch (**Supplementary Table 1**). Variants with more than 2 clusters, or where clusters were poorly separated or where genotypes were missing for entire clusters (in the middle or at cluster edges) were also selected for re-calling.

Stage 2 and 3: Re-calling procedures and post-recalling QC

Re-calling was done using a bespoke R script (https://lighthouse.mrc-mbu.cam.ac.uk/gogs/cc926/UKBiobank) by E.Y-D and C.C. To ensure re-calling quality, a set of SNVs was re-called by both re-callers and the concordance in calls was 99.9%. Of the 135 variants, 53 (39.3%) variants were re-called per array for all batches together, while for 66 (48.9%) variants the re-calling was done both per array and per batch. Furthermore, 2 variants (1.5%) that showed overlapping clusters to an extent where re-calling was not possible and 15 monomorphic variants (10.7%) were excluded, resulting in a set of 248 (**Supplementary Tables 1-4**). To check the quality of the re-calling procedure, for each of the recalled variants cluster plots (per array and per batch) were generated and visually inspected by E.Y-D and C.C., and in cases of disagreement (N = 7), additionally by J.M.M.H. Finally, we re-calculated the per sample call rate (within each array) using the final set of SNV and excluded samples with call rate ≤ 0.97 (N = 2,643).

Stage 4: Imputation

Prior to attempting imputation, we explored whether imputation of mtDNA variants was at all possible. We split the GenBank reference panel described in Wei W *et al*.[5] (30,506 sequences, 6,580 biallelic SNVs) in half and imputed one half (a test set) against the other (a reference set). We then deleted different number of variants from the test set and kept 50%, 12.5% and 1.3% of variants in common between the two sets (**Supplementary Fig. 5**). Furthermore, we also tested our high-quality set of 248 variants that were present in the UK Biobank. The imputation was done using IMPUTE2[6] and both haploid and diploid settings were tested and found to perform equally well (**Supplementary Fig. 5**). When the two sets shared ≥ 10% of SNVs, an imputation quality score (INFO) cut-off of 0.7 was sufficient to assure high concordance for the minor allele (≥ 90%), irrespective of MAF (**Supplementary Fig. 5**). However, when the number of shared mtSNVs was lower than 10% of those present in the reference panel, as is the case with UK Biobank, an additional MAC cut-off was also required (**Supplementary Fig. 5**). Finally, as with nuclear variants, when the reference and imputation panel shared very few SNPs (≤ 1%), only common variants could be imputed accurately (**Supplementary Fig. 5**). We therefore

chose the most conservative score cut-off for filtering the imputation results and set out to use only imputed SNVs with INFO score ≥ 0.7 and mac ≥ 10.

We then imputed the UK Biobank dataset against GenBank, described in Wei W *et al*.[5] As partial sequences may present with alignment errors, to further increase accuracy we excluded those from the reference data set, resulting in 17,815 GenBank complete genomes and we only used biallelic SNVs (N = 5,271). The GenBank genomes was the largest data-set available to us and included European (58%) and non-European haplogroups (42%). Imputation was performed separately on each array (N = 49,945 for UKBL and N = 438,377 for UKBB). Although homoplasmic mtDNA variants are in phase, IMPUTE2 cannot handle missing calls. Thus, we used the IMPUTE2 prephasing step to fill in those, following standard procedures. The post-imputation QC parameters were set as above (**Supplementary Fig. 5**). In cases where mtSNVs were genotyped on one array and imputed on the other, we took forwards the imputed genotypes only if their INFO score was ≥ 0.7.

Functional annotation of mtSNVs in UK Biobank

With the exception of a few rare variants linked to mitochondrial diseases such as LHON, the 265 mtSNVs probed in UK Biobank were mostly selected to broadly capture haplogroup variability, rather than variant pathogenicity (**Supplementary Table 5-6**). Similarly, the imputation procedure we used was agnostic of the functional consequences of mtSNVs, hence we assessed functional annotations of the genotyped and imputed SNVs that passed QC (N = 473) in the EUR set and calculated variant allele fractions, in both the Full (N=483,626) and the EUR set (N=358,916), in all groups of variants (N = 265 before recalling, N = 248 after recalling, N = 471 imputed) (**Supplementary Tables 1,3,4**) to further compare these with the three reference datasets (GenBank, 1000 Genomes and WTCCC). Allele frequencies and MAC were calculated using QCTOOL v2 (https://www.well.ox.ac.uk/~gav/qctool_v2/).

Half of the 473 SNVs were synonymous (N=238, 50.3%), while non-coding SNVs (N = 125, 26.4%) and non-synonymous SNVs (N=110, 23.3%) were present in similar proportions (**Supplementary Figure 10, Supplementary Table 5**). Only a small percentage (N = 25; ~5%) of mtDNA SNVs were shown to be pathogenic based on

previous studies (Methods), of which only 6/24 (25%) had a confirmed pathogenic status according to MITOMAP[7] **(Supplementary Table 5**). Notably, known pathogenic mutations, like the m.3243A>G[8] and m.3460G>A[9] were not in the final set of 473 SNVs because they were monomorphic (**Supplementary Table 1, Supplementary Figure 10**), and those with confirmed pathogenic status were very rare (MAF < 1%). We compared allele frequencies of genotyped and imputed UK Biobank SNVs, using Spearman correlation, with those calculated in the reference datasets (UK Biobank N = 483,626; GenBank N = 17,815; 1000 Genomes N = 2,419; WTCCC N = 763), repeating also the comparison within the set of Europeans with nuclear-mitochondrial matched ancestry (UK Biobank N = 358,916; GenBank N = 6,593; 1000 Genomes N = 986; WTCCC N = 747). This comparison showed that the correlation was stronger when comparing European subgroups rather then the whole sets with no further selection for matched ancestries (**Supplementary Table 7**).

We further annotated the haplogroups tagged by each of the 473 mtSNVs (**Supplementary Table 6**) according to the Phylotree build 17[10]. The majority of SNVs (N = 412, 87%) were observed at least once in the human phylogeny, prevalently tagging European haplogroups (H, J, K, T, U, I, V, W, X). Despite the fact that the ancestry of this subset was exclusively European, African and Asian alleles (both genotypes and imputed) were still present, but at very low MAFs (mean MAF < 1%), suggesting that the UK Biobank sequencing data could be a valuable resource to re-evaluate the current human mitochondrial phylogenetic tree.

Finally, the non-synonymous m.10398A>G in *MT-ND3* (rs2853826) which we found associated with an increased risk of multiple sclerosis in the UK Biobank cohort (**Table 1**) was previously found associated with a decreased risk of MS (OR = 0.87; $P = 1.5 \times 10^{-2}$) in a mtGWAS cohort with a larger number of cases (N = 9,985)[11]. The same variant was found associated with a similar decreasing risk for Parkinson's disease, Ischemic Stroke and Ankylosing Spondylitis[11]. However, the mitochondrial *P*-value threshold adopted in Hudson et is not properly accounting for the actual number of independent mtSNVs and/or distinct haplogroup branches tested (**Methods**), thus is far above the mitochondrial threshold ($P < 5 \times 10^{-5}$) proposed here. We did not confirm any of the previously observed pleiotropic effects of this variant even though for the majority of

traits we have over 70% power to detect association of the same effect size at the *P*-value (*P* = 0.01) reached by Hudson et al.[11] (Multiple Sclerosis: 0.59; Ankylosing Spondylitis: 0.81; Parkinson's disease: 0.75; Ischemic Stroke: 0.97).

Modelling the geographic distribution of mtDNA variation in Great Britain

We modelled the geographic distribution of mitochondrial variation in Great Britain and its relationship to the nuclear genome variation at different levels of granularity. We analysed macro-haplogroups, mtDNA principal components (mtPCs) or mtSNVs to study mtDNA variation, while we used nucPCs or clusters derived using those to study nuclear variation (**Supplementary Tables 10-13**). In terms of geographic parameters, we used either postcode of birth longitude (#130) and latitude (#129) on their own or regional units derived from Nomenclature of Territorial Units for Statistics level 2 (NUTS2) (https://geoportal.statistics.gov.uk/datasets/nuts-level-2-january-2018-names-and-codes-in-the-united-kingdom) (**Supplementary Table 11**, **Fig. 2**). Detailed information on haplogroup predictions and nucPCs derivation is provided in the main text and above. To further reduce dimensionality and make the geographic comparison between macro-haplogroups and nucPCA easier, we also performed k-means clustering using the first 10 nucPCs, establishing the number of clusters as equal to 8. This number was chosen using the Elbow method, which determines the number of clusters at which the total within-cluster sum of square is minimised.

In UK Biobank, mtPCs were calculated with FlashPCA2[1] using three different sets of genotyped variants: (1) a set of 35 common LD-pruned (MAF > 0.01, $R^2 < 0.2$) mtSNVs; (2) a set of 123 common mtSNVs (MAF > 0.01), without LD-pruning; (3) a set including all 248 post-recalling mtSNV (**Extended Data Fig. 2**). LD-pruning was performed using PLINK[4] and the whole mt-genome was treated as a single window. The mtPCs resulting from the first set were used for the detection of mt-related genetic structure in UK Biobank, while the mtPCs from the other two set were used to explore whether those strategies would capture long-range LD patterns and recapitulate the haplogroup predictions. For comparison we also calculated mtPCs in the reference datasets (GenBank, 1000

Genomes, WTCCC) in the same way as in UK Biobank (**Supplementary Fig. 8 and Extended Data Fig. 3)**.

The birth coordinates (#130, #129) were intersected with NUTS2 geographical locations as follows. The reverse_geocoder Python module (v. 1.5.1) was used to retrieve administrative regions (UK region, county and city), which were subsequently intersected with UK NUTS level 2 data available at the GRID database (2019-02-17 release) (https://www.grid.ac/). Level 2 was chosen as it provides a more detailed classification than level 1 and more power per category than level 3. Individuals with missing birth coordinates or where NUTS category could not be assigned were excluded from the analyses and territory units with less than 1000 participants were merged to a neighboring unit. This resulted in 327,665 individuals in the EUR set, classified in 34 units (**Supplementary Table 11**).

We first explored the distribution of the European macro-haplogroups across NUTS2 and found small but significant differences ($P < 5x10^{-5}$) in frequencies for macro-haplogroups J, W and I across the country (**Fig. 2**, **Extended Data Fig. 1**, **Supplementary Table 11**). The first 10 nucPCs explained 23% (longitude) and 15% (latitude) of variance in birth location and the cluster derived from those were present at different frequencies across NUTS2 (**Fig. 2**). For example, fewer people fell within clusters 1 and 4 and those were approximately equally distributed across the country, while cluster 5 was most common in the two NUTS2 units mapping to Wales. People born in the South and central England predominantly fell within clusters 2 and 6, while cluster 3 was most common in the North of England. People born in Scotland fell predominantly in clusters 7 and 8, while people from the North East and West of England fell within clusters 2, 6 and 8. Furthermore, clusters common in Scotland and Wales were more likely to occur in individuals belonging to macro-haplogroups I, J and K (**Supplementary Table 10**). We also explored the association between longitude or latitude and common mtSNV (MAF > 0.01). This replicated what we observed with the macro-haplogroups and NUTS2 but did not yield additional information (**Supplementary Table 11**).

Next, we explored the relationship between mtSNV, mtPCs and nucPCs. Three hundred and thirty-two mtSNVs were associated with at least one of the first 10 nucPCs ($P < 5x10^{-5}$, **Extended Data Fig. 2**). We further show both in UK Biobank and the

reference datasets that mtPCs, when calculated without pruning, reflect long range LD-patterns and not geographic structure. In all cases, the correlation between nucPCs and mtPCs was low and mtPCs or haplogroups were less informative in terms of population structure compared to the nucPCs. This is in line with a previous study that showed that macro-haplogroup membership provides limited information about either continental ancestry or geographical origin of individuals[12].

GenBank

For reference and imputation purposes, we used 17,815 complete GenBank human genomes[5], accounting for 5,271 biallelic mtSNVs. The SNVs present in the reference dataset were identified by aligning the complete human genomes downloaded from GenBank to rCRS (NC_012920.1), as described in Wei et al. 2017[5]. We retained only homoplasmic variants (i.e. not showing ambiguity in the genome sequence) and variants tagging biallelic single nucleotide variants. With the exception of three SNVs not present in the reference dataset, 99% of the high quality re-called UKBB mtSNVs (N = 245/248) were observed at least once in GenBank (**Supplementary Table 3**). We cross-checked haplogroup predictions and nuclear ancestry to identify a subset of 6,593 (37%) GenBank genomes with European origin.

1000 Genomes

We used mitochondrial variant calls for 2,419 individuals from Phase 1 and 3 of the 1000 Genomes Project[13], to calculate reference allele frequencies and haplogroup predictions. We downloaded the VCF file with mtDNA calls generated by the MToolBox pipeline[14] from the *sourceforge* repository of the tool (https://sourceforge.net/projects/mtoolbox/files/1000Genomes_data/) and re-mapped the variant to the rCRS reference sequence. To calculate allele frequencies and to perform haplogroup predictions we further retained only variants that were homoplasmic or nearly homoplasmic (i.e. with heteroplasmic fraction≥0.8, N=3,608). These we used for quality check purposes and cross-checked with nuclear ancestry to identify 498 (21%) 1000 Genomes individuals with European origin. The 1000 Genomes consortium subjects

provided their consent as explained in https://www.internationalgenome.org/sample_collection_principles.

Wellcome Trust Case Control Cohort

The Wellcome Trust Case Control Cohort (WTCCC)[15] individuals were residents of Great Britain, self-identified as white Europeans. We sequenced 763 individuals of the WTCCC with Illumina Hiseq 2000 using an amplicon-based library preparation and Illumina sequencing. In brief the Fluidigm Access Array™ technology was used to generate tagged and indexed amplicons, (roughly ~100 per sample of 150-200bp), with sample-specific barcodes and Illumina adaptor sequences. We checked the resulting PCR products for quality using the Agilent 2100 Bioanalyzer and then pooled together in equal volumes. The PCR product library was purified using AMPure XP beads and quantified with PicoGreen prior to loading for Illumina sequencing. Prior to mapping we check the quality of raw sequencing fastq files with FastQC. We mapped read using the MToolBox pipeline[14] with default arguments, using rCRS as mitochondrial reference genome and the hg19 genome build as nuclear reference to remove possible nuclear-mitochondrial DNA sequences (NumtS) contaminations. Sequencing coverage (% of rCRS covered by the mapped reads) was>95% in all samples and mean read depth per sample was 1004X (sd=223X; min=514X, max=1903X). Mitochondrial variant calling was performed with the MToolBox pipeline, using the default options (minimum read depth per alternative allele ≥ 5 and minimum quality score per base ≥25). Only variants homoplasmic or nearly homoplasmic (i.e. with heteroplasmic fraction≥0.8, N=825) were further retained to calculate allele frequencies and to perform haplogroup predictions. These were used for quality check and to identify 747 (98%) participants with matched nuclear-mitochondrial European origin. The Wellcome Trust Case-Control Consortium subjects gave written informed consent and the project protocols were approved by the relevant research ethics committees in the UK. A list of investigators who contributed to the generation of the data is available from www.wtccc.org.uk.

Haplogroup predictions in the reference datasets

In order to compare mtSNV allele frequencies and haplogroup distributions, with the three reference datasets, Haplogrep2[15] predictions for the 17,815 GenBank complete mitochondrial genomes were downloaded from *Wei W* et al, 2017[5], and calculated also for 2,419 1000 Genomes individuals and 763 WTCCC individuals, using only the subset of homoplasmic variants identified by the MToolBox variant calling[14] (**Supplementary Tables 8-9**). In this case, only variants homoplasmic or nearly homoplasmic (i.e. with heteroplasmic fraction ≥ 0.8) were retained to calculate allele frequencies and to perform haplogroup predictions.

Individuals were considered European if they belonged to the nine major European haplogroups (H,V,U,J,T,K,I,W,X) and were also European as inferred by nuclear PCA (1000 genomes, WTCCC). Based on this, in the 1000 Genomes data we identified 986 (41%) individuals belonging to the nine major European haplogroups, of which 498 (21%) were also of European origin in terms of their nuclear genome. Of the 763 WTCCC participants, 747 (98%) WTCCC individuals had matched nuclear-mitochondrial European origin.

In the case of GenBank, autosomal genotypes were not available. Therefore, we retrieved the geographical origin of each genome by running a programmatic API query of the HmtDB database[16] genome cards for each GenBank accession id. For 4,642 accessions with undefined geographical origin or not present in HmtDB we attempted the retrieval by fetching further metadata from their GenBank genome entries (i.e. publication journal, title and sample isolation), using the Bio.Entrez Biopython module (https://biopython.org/docs/dev/api/Bio.Entrez.html). When the GenBank entry did not provide a clear indication on the geographical origin, we manually reviewed the correspondent literature available for that sample. Based on haplogroup predictions already available[5] and on geographical origin, retrieved as described, we identified 6,593 (37%) complete European human GenBank genomes. Additionally, to further compare allele frequencies of the UK Biobank Full Set with GenBank, as shown in **Fig. 1**, we calculated the number of Asian (AS) and African (AFR) matched nuclear-mitochondrial ancestries in UK Biobank, using the ethnicity background field (#21000) (AS, N = 888; AFR, N = 2,012), and in GenBank (AS, N = 3,587; AFR, N = 704), adding these to the

European group, to obtain 361,816 UK Biobank and 10,884 GenBank matched genomes in total.

Phenotype definitions and model specifications sensitivity analyses

Binary traits

We used the ICD-10 category (#41270 code for primary and #41204 code for secondary), self-reported category (#20002) and several categories from health and medical history records (codes under category #100036) as binary traits, as of the August 2017 freeze. Number of informative participants for this analysis were identified as those with at least one primary or secondary ICD-10 record (N=279,179 European unrelated individuals). Participants with no annotation for a specific ICD-10 code have been considered as controls. Regarding the non-cancer self-reported illness (N=166), we mapped them to ICD-10 chapters and considered as cases participants with at least one record of non-cancer self-reported illness (N=271,332 European unrelated individuals), while participants with no self-reported records have been used as controls.

We have also tested 22 traits belonging to health and medical history records, assigning to missing values those individuals answering to the touch screen questionnaire with “Do not know” and “Prefer not to answer” categories. The full list of categorical traits tested is available in **Supplementary Table 14**.

We performed a trait-related sensitivity analysis for a subset of phenotypes of interest, already linked to mitochondrial dysfunction. The sensitivity analysis included merging ICD-10, ICD-9 and self reported codes together, excluding individuals with co-morbidities, as well as testing UK Biobank algorithmically defined traits and bespoke phenotypes definitions (**Supplementary Table 15**). The traits of interest were: type 2 diabetes, Parkinson’s disease, epilepsy, multiple sclerosis, ulcerative colitis, rheumatoid arthritis, ankylosing spondylitis, polymyalgia, fibromyalgia, Crohn disease and dyspepsia, CVD-related and traits related to airways function. The codes used to create these bespoke categories, number of cases and controls and genomic inflation factor per trait are available in **Supplementary Table 14.**

Post code of birth place (northness and eastness) (#130, #129) were transformed using rank inverse normal transformation and tested for association with mtSNVs or

haplogroups (factorial variable), adjusting for age, age squared, sex, array and the first 10nucPCs (**Supplementary Table 13**).

Quantitative traits

We selected 126 quantitative traits for the PheWas analysis, including BCTs, SBs, ATs and few other traits, as summarised in **Supplementary Table 16.** When necessary, traits were split by sex before outlier removal and normalisation. Unless stated otherwise, traits were transformed using quantile-inverse-normal transformation prior to the association analysis. The effects of additional covariates used for further sensitivity analysis are available in **Supplementary Table 22**. For a selected group of traits (anthropometric traits and blood cell traits) we performed additional adjustments, as described in the sections below.

Impedance measured anthropometric traits (iATs): We studied 21 iATs (**Supplementary Table 16**). Prior to the analyses, phenotypes were prepared using standardized protocol in a sex-specific manner[17]. Briefly, we removed outliers (values greater than 1st and 99th percentile) and normalised the iATs using rank inverse normal transformation. Then the iATs were regressed on age and age squared and the resulting residuals were standardized to have a mean of 0 and an SD of 1. Finally, the females and males standardised residuals were combined. Furthermore, pregnant women or individuals with COPD or cancer diagnoses were excluded.

Blood cell traits (BCTs): We studied 33 hematological traits (**Supplementary Table 16**) that were either measured in standard clinical full blood counts (FBCs, N = 15) or derived from the measurements taken during those counts (N = 18), including red blood cell traits (N = 12), platelet traits (N = 4) and myeloid and lymphoid white blood cell traits (N = 17). The FBCs were obtained using Beckman Coulter LH700 Series instruments. To increase the power to detect genetic associations, an extensive quality control was performed as previously described[2]. Briefly, sources of technical and non-genetic biological variation were identified and removed. To improve accuracy, FBCs measured more than 36 hr after venipuncture and samples that fell within the 96th percentile of mean platelet volume were removed[2]. Further technical covariables such as the time between venipuncture and FBC analysis, FBC instrument drift, calibration events, episodes of malfunction, and seasonal

effects were adjusted for. Individuals suffering from blood cancers or other blood disorders were removed, and the following biological covariables were adjusted for: age, sex, menopause status, height, weight, smoking, and alcohol consumption[2]. Observations by trait for which there was a large difference between the raw measured trait value and the adjusted trait value were removed. Finally, the adjusted BCTs were standardized and transformed using quantile-inverse-normal transformation.

Sensitivity analyses of model specifications

Trait and covariate definitions: For phenotypes already linked to mitochondrial dysfunction ICD-10, ICD-9 and self-reported codes were reviewed by a clinician and codes with similar meanings were merged. Individuals with disorders causing a similar phenotype, or at high risk of developing the disorder of interest were excluded from controls, for example individuals with optic neuritis were excluded from controls for the multiple sclerosis analysis. Besides bespoke phenotypes definitions, we have also included UKBB algorithmically defined traits (**Supplementary Table 15**). We also explored additional covariate (*eg*. age, geographic parameters) that, for the purpose of increasing power, were not part of the initial model In all these analyses, a change of estimate (absolute difference is Z scores (beta/standard error)) was deemed significant if it exceeded 10% irrespective of changes in *P*-values.

Due to the sensitive of binary traits to model misspecification, especially in the case of imbalanced case-control analyses, we explored three different tests (score, Wald tests and likelihood ratio test) to ensure robustness of the results. These tests were applied on the 29 significant binary associations and produced similar results (**Supplementary Tables 21**). We focused on the results from the Wald test as those were the most conservative. In addition, we explored whether treating the genotypes at each mtSNV as factorial variable altered the results (STATA 14.2, likelihood ratio test). For this step we converted the genotype probabilities to hard-call genotypes at cut-off of 0.6. The factorial and additive models did not differ, again showing the robustness of the results and that models, usually used for autosomal variants, are applicable for mitochondrial variants as well.

We also explored the effects of adding different covariates to the models. We first tested the standard set of covariate such as age, age squared, sex, array and the addition of nuclear principal components. All those were retained in subsequent models for the quantitate traits as they were statistically significantly associated with the traits of interest (**Methods**). In order to increase power, age and age squared were not included in the models for binary traits, but their effect was tested in the sensitivity analyses (**Supplementary Tables 21**). The observed binary associations were valid to the addition of the effects of age.

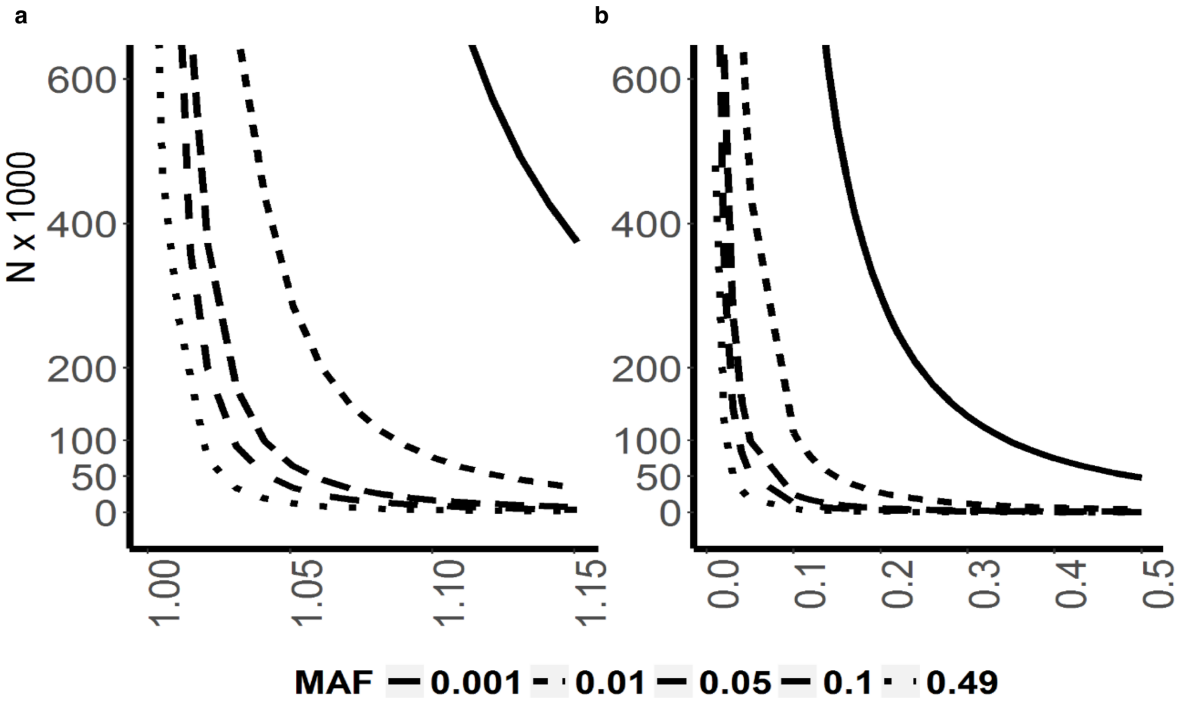
We also explored additional covariate that, for the purpose of increasing power, were not part of the initial model: variables that account for geographic differences in variant and haplogroup frequencies (easthnes and northness of birth, region of birth), variables that adjust for the effect of drugs proposed to influence mitochondrial function (*e.g*. antibiotics, metformin, antidepressants), and, when possible, adjusting for the effect of nuclear genetic variants known to be associated with the outcome (**Supplementary Tables 21-22**). The observed associations were valid to the addtion of all these different covariates even when addting those covariated redused the sample size avaialbe for analyses (e.g. covariates derived from birth coordiantes).

Lambda simulations

To evaluate the λ distribution we drew 10,000 samples per trait from the test statistics available from the Astle et. al.[2] that matched the MAF distribution in the mitochondrial data and calculated the λ each time (**Supplementary Fig. 11)**. The Astle et. al.[2] utilised smaller number of individuals (~170,000) and triple genomic control. Given this and the polygenicity of BCTs, our simulation is likely to underestimate the inflation that could be observed in a sample of over 350,000 participants. Therefore, we increased by 15% each of the simulated λ_s.

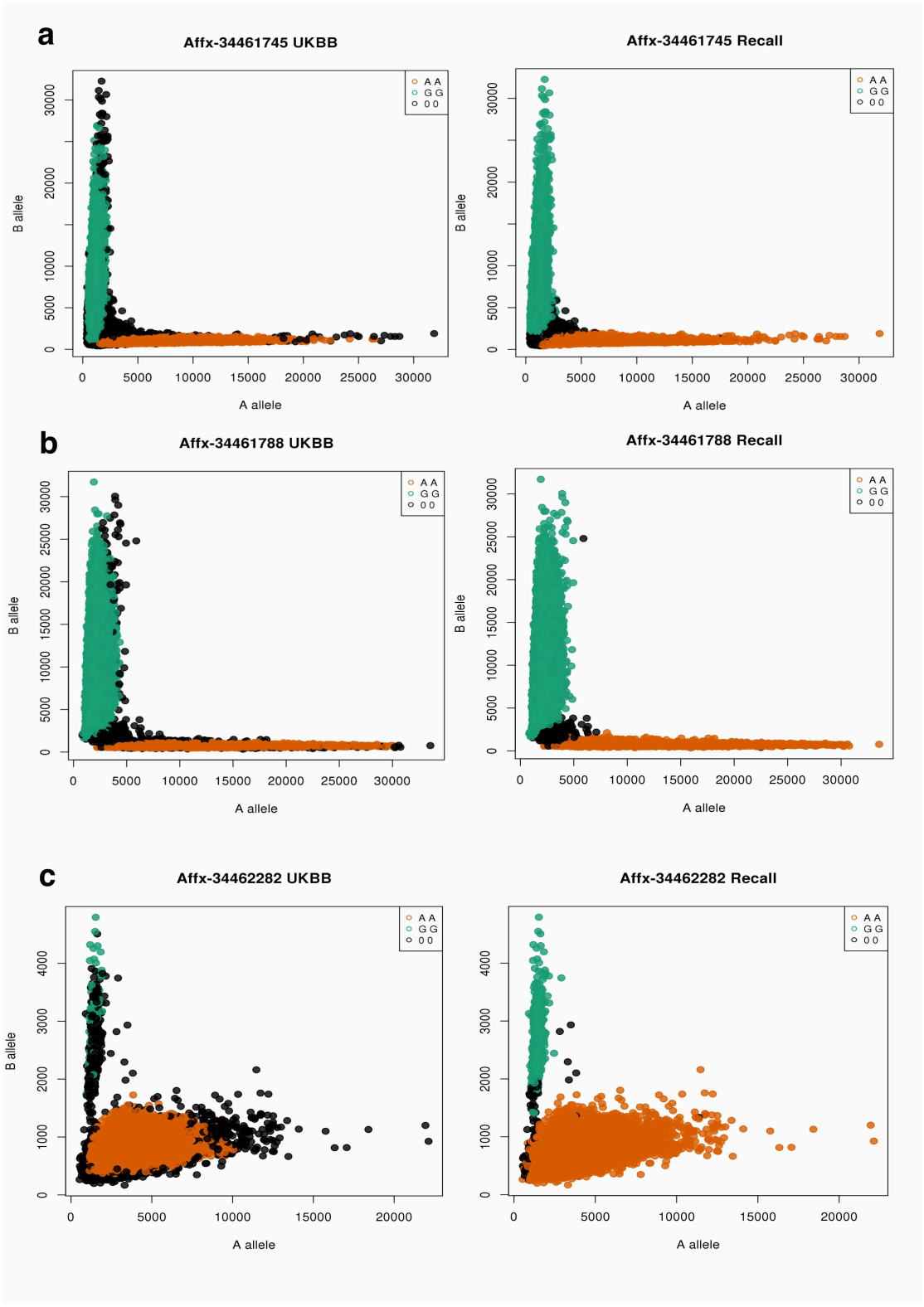
Supplementary Figures

Supplementary Figure 1. ***Power calculations for binary and quantitative traits***



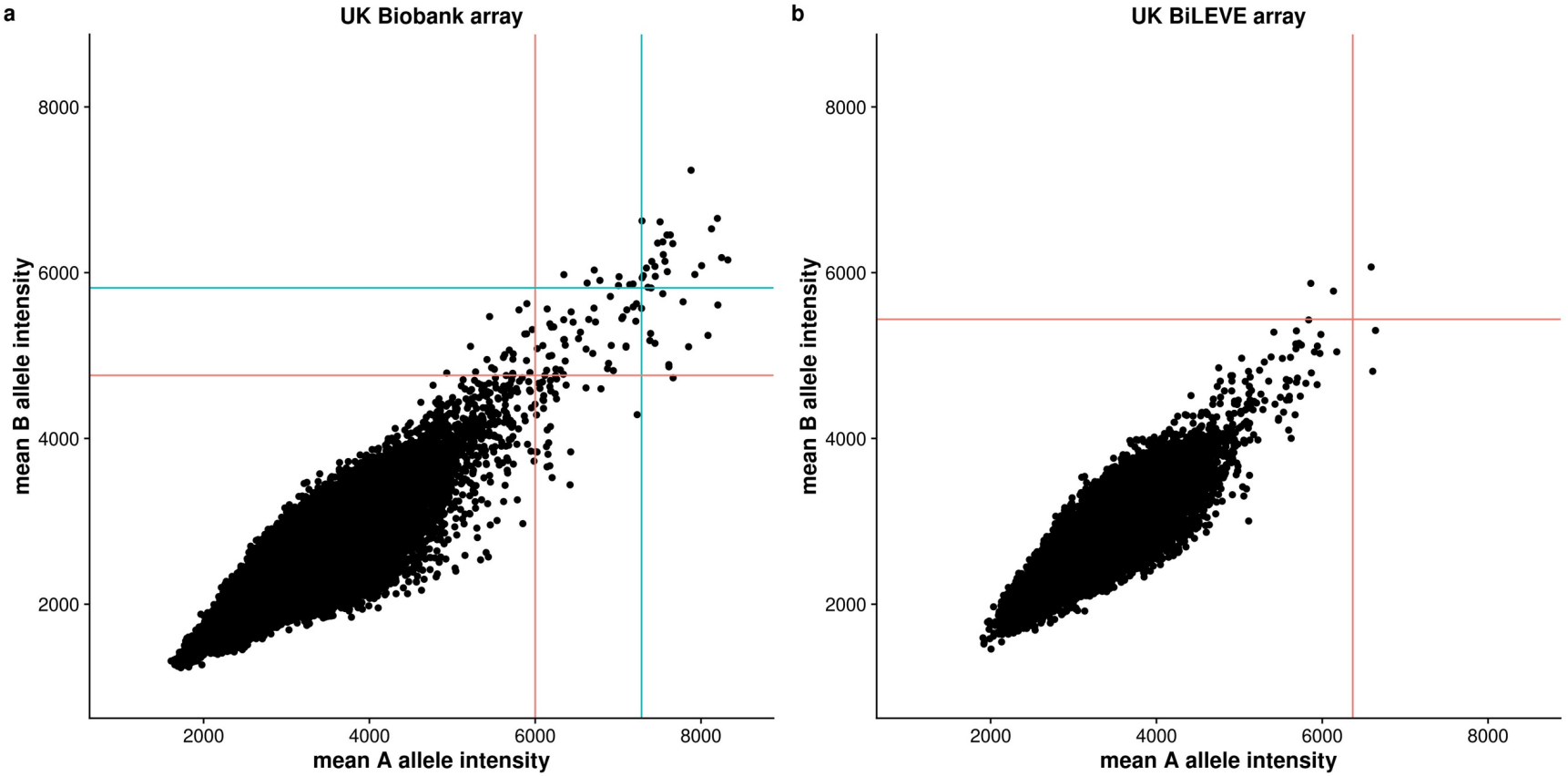
The figure summarises the results of power calculation for binary and quantitative traits across different minor allele frequency (MAF) ranges of mitochondrial variants. The number of participants needed for a study to have 80% power to detect an effect at $\alpha=5x10^{-5}$ for (**a**) binary traits with 7% prevalence and (**b**) continuous traits. x-axes: number of individuals needed to achieve 80% power; y-axes: odd ratio (OR) for (**a**) and beta for (**b**). Power calculations were performed using Quanto[21]

Supplementary Figure 2. Examples of mitochondrial cluster plots of the UK Biobank mtSNVs before and after genotype re-calling



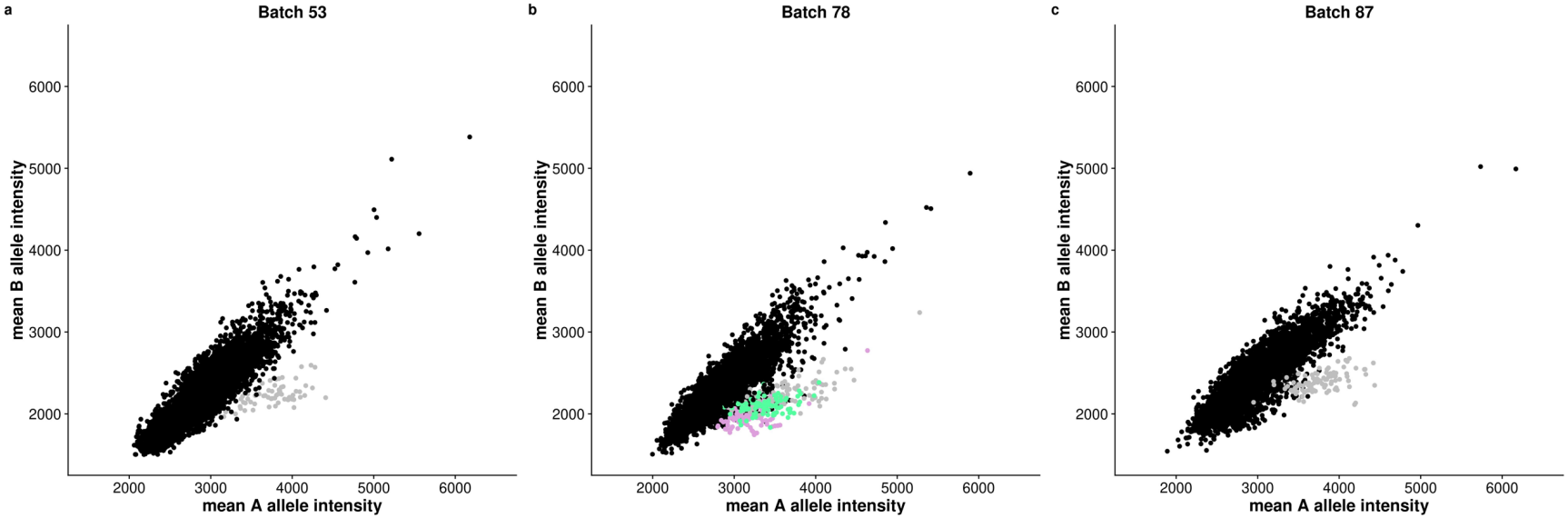
Cluster plots of probe intensities for three mitochondrial SNVs genotyped in UK Biobank in 488,377 individuals (Affx-34461745 (**a**), Affx-34461788 (**b**) and Affx-34462282 (**c**)). Each dot represents a participants and colours correspond to genotype assignment. Black dots indicate missing genotypes. Affx-34461745 and Affx-34462282 (**a**-**b**) are examples of cluster plots with missing calls at high intensities prior to recalling (left), and post recalled genotypes (right). Affx-34462282 (**c**) is an example of a cluster plot with a rare allele (G/G) showing high missingness before recalling (left) which is improved after genotype recalling (right).

Supplementary Figure 3. Mean allele A and allele B intensity outliers across all mtSNVs



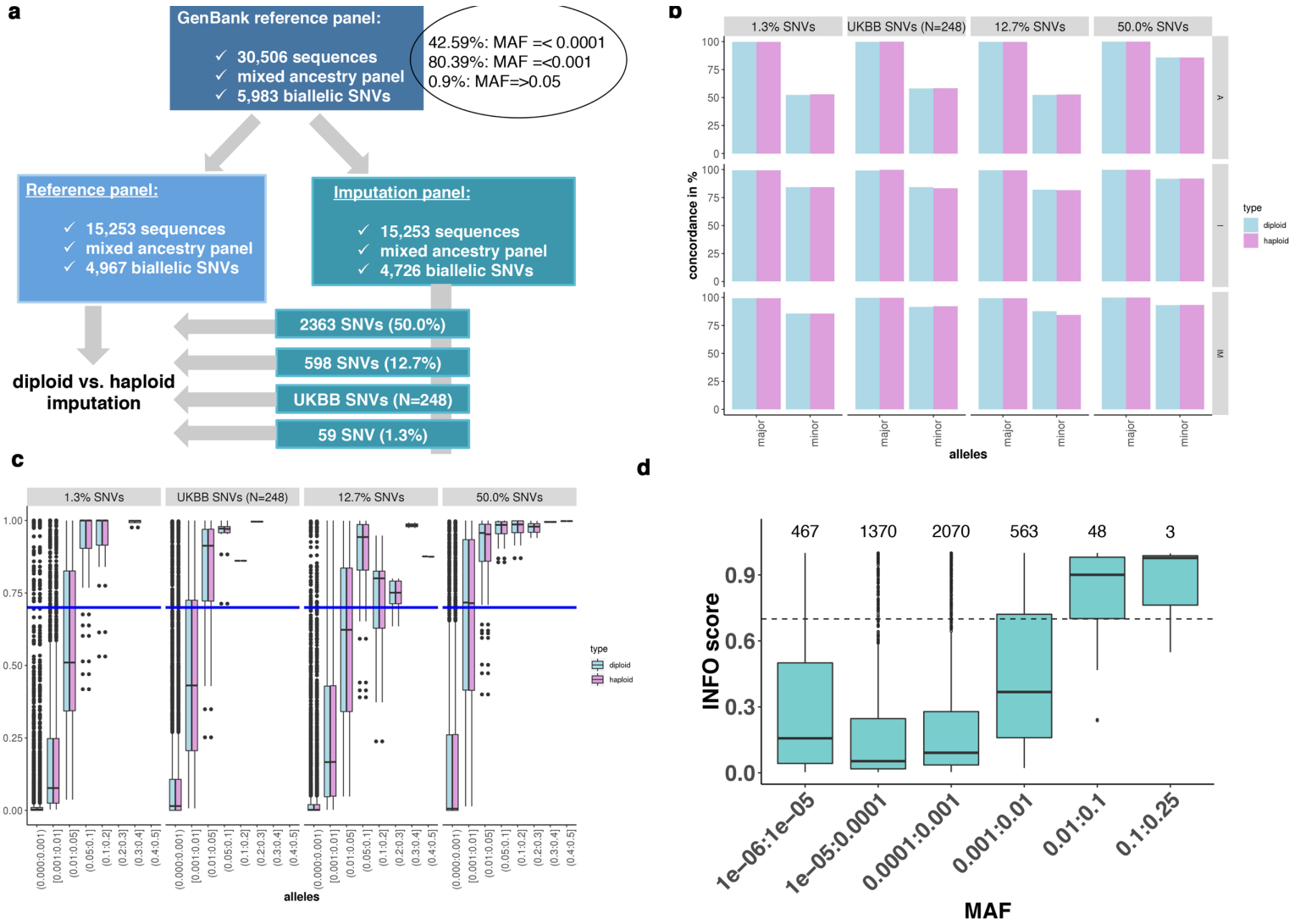
This figure shows the mean allele A and allele B intensities for all mtSNVs (N=265) for the UK Biobank Affymetrix Axiom array (**a**) and UK BiLEVE Affymetrix Axiom array (**b**). Each dot represents a participant: N=438,427 in a) and N=49,950 in b). The magenta and green lines denote 7 and 10 standard deviations (SDs) from the cluster center, respectively. UK Biobank Affymetrix Axiom array samples falling beyond 10 SDs (N=49) and UK BiLEVE array samples falling beyond 7 SDs (N=5) were excluded

Supplementary Figure 4. Examples of plate effects



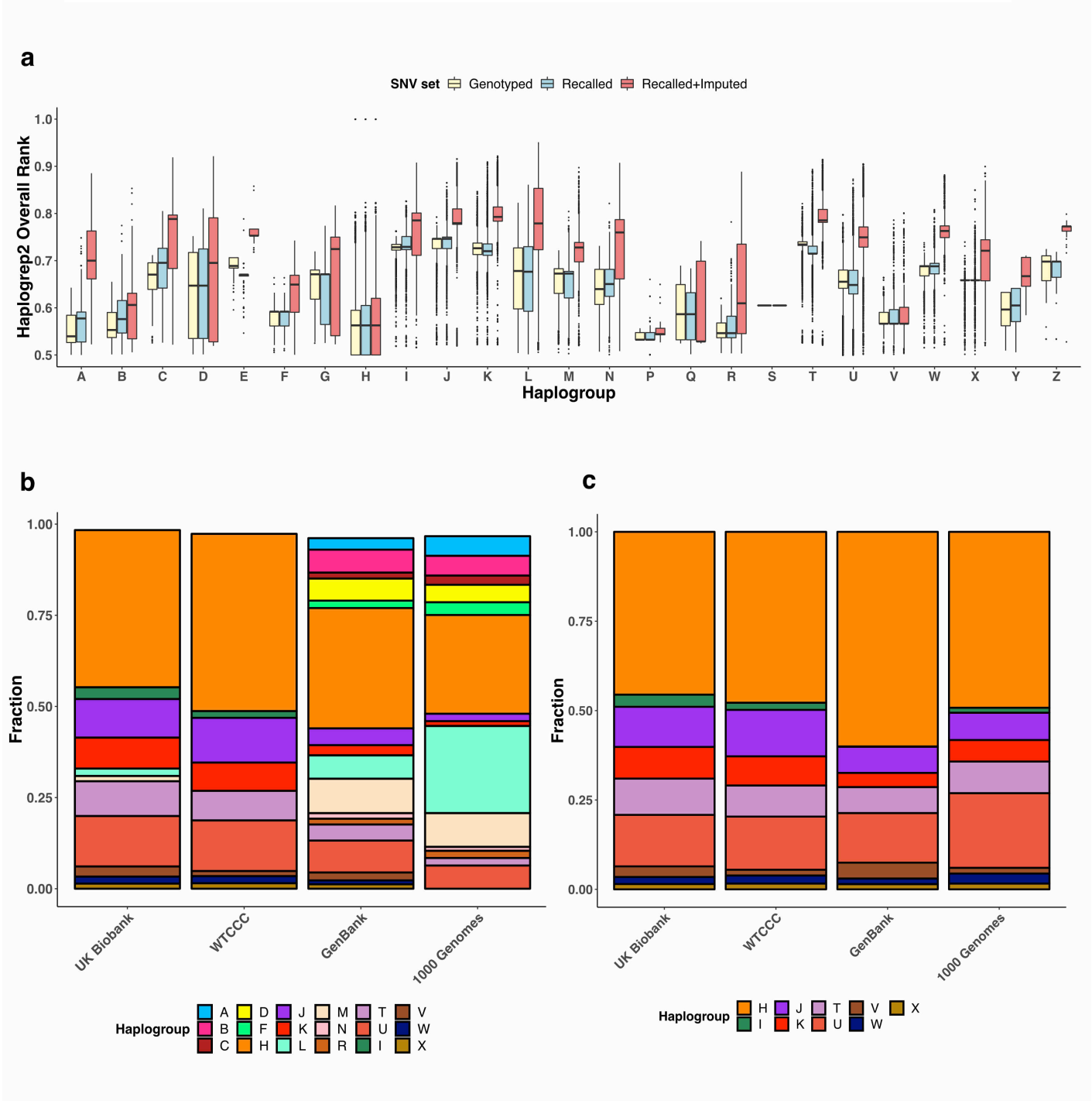
Mean allele A and allele B intensities over all mtSNVs for each participant. Three examples of plate effects for (a) batch 53 (N=4,613), (**b**) batch 78 (N=4,638) and (**c**) batch 87 (N=4,660) are shown. Outlying participants are shaded in grey (N=66, N=91 and N=90, for a-c respectively), green (N=94) and purple (N=90).

Supplementary Figure 5. Imputation procedure validation



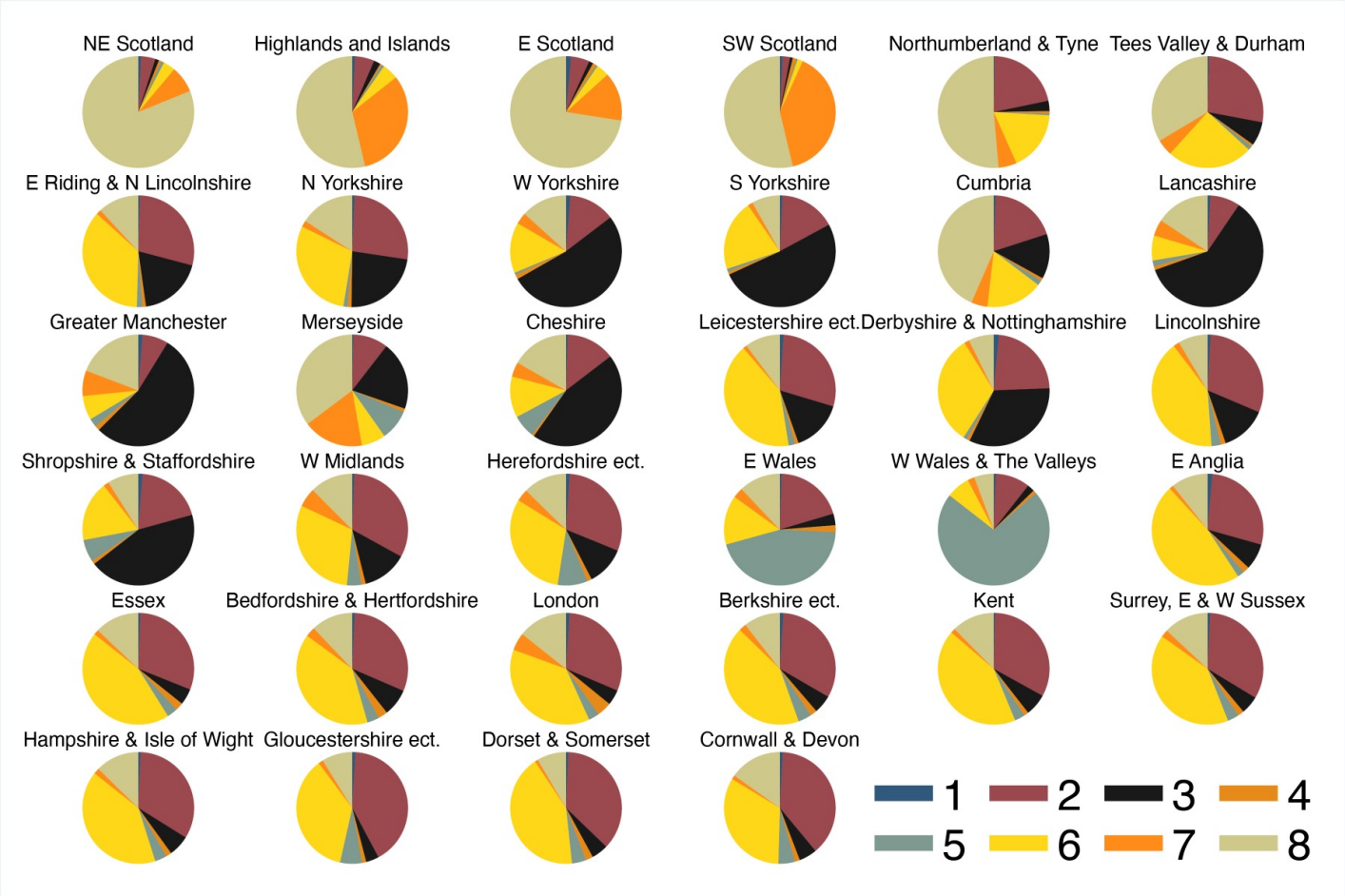
Our validation steps for the imputation of mtSNVs. (**a**) Flow-chart of the validation procedure. The GenBank full data set (30,506 sequences, 5,983 mtSNVs) was split into a reference and an imputation set of equal sizes. Between 50% and 98.7% of mtSNVs were deleted from the imputation set. The missing mtSNVs were next imputed using the reference set to assess the robustness of the imputation approach. The number of mtSNVs in common between the sets at each iteration is reported; (**b**) Concordance (%) between imputed and actual genotypes is reported for the major and the minor alleles separately: diploid imputation (blue); haploid imputation (purple); all mtSNVs (A), mtSNVs with INFO ≥ 0.7 (I); variants with INFO≥0.7 and minor allele count (MAC) ≥10 (IM); (**c**) a boxplot showing the relationship between minor allele frequency (MAF) (divided into 8 categories) and INFO score for both diploid (blue) and haploid (purple) imputation; the blue line represents INFO=0.7. (**d**) Distribution of INFO scores by MAF bins for the UK Biobank imputed mtSNVs. The number of SNVs per bin is indicated above each box. dash line: INFO=0.7 The lower and upper hinges of boxplots in (**c**) and (**d**) correspond to the first and third quartile of the distribution, with median in the center and whiskers spanning 1.5 times the interquartile range. Black dots depict boxplots outliers.

Supplementary Figure 6. Haplogroup predictions in the UK Biobank and the three reference cohorts: WTCCC, GenBank and 1000 Genomes



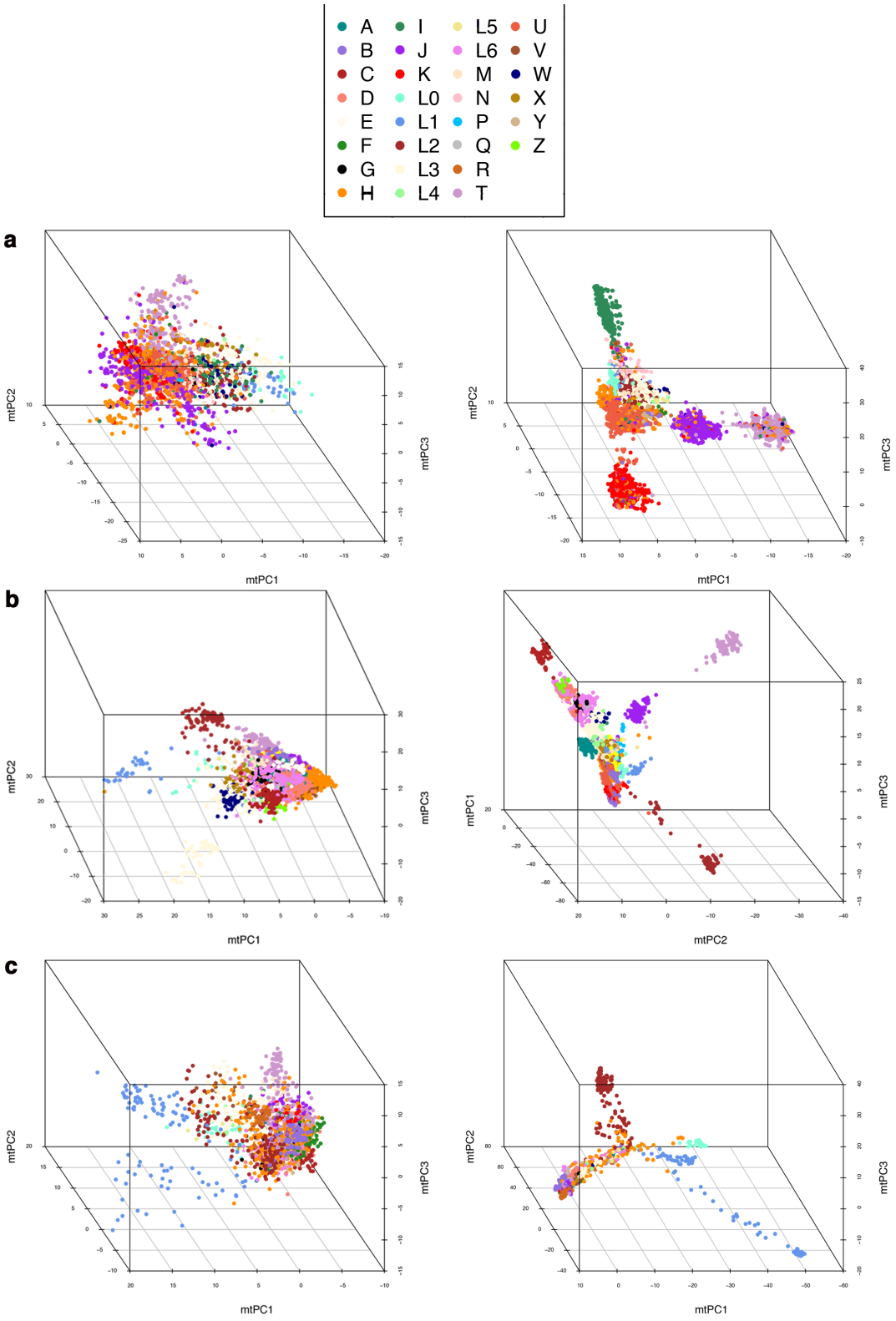
(**a**) Boxplot representing the *per*-individual quality of haplogroup prediction (expressed as Haplogrep2 rank score), based on genotyped, post-recalling and post-imputed SNVs, across all participants that passed the QC in the “Full set”; The lower and upper hinges of boxplots correspond to the first and third quartile of the distribution, with median in the center and whiskers spanning no further than 1.5*interquartile range. Black dots depict boxplots outliers (**b-c**) Fraction of individuals with haplogroup predictions calculated (**b**) over the total number of individuals available per cohort in the “Full set” and **(c**) over the EUR subgroup. Shown are UK Biobank and the three reference cohorts: WTCCC, GenBank and 1000 Genomes. The number of individuals in the “Full sets” are: N=483,626 UK Biobank, N=763 WTCCC, N=17,815 GenBank and N=2,419 1000 Genomes. EUR individuals *per* cohort are: N=358,916 UK Biobank, N=747 WTCCC, N=6,593 GenBank and N=498 1000 Genomes. Haplogroups corresponding to less than 1% of the entire cohort in panel **b**) were not show

Supplementary Figure 7. Distribution of nuclear clusters across NUTS2 units



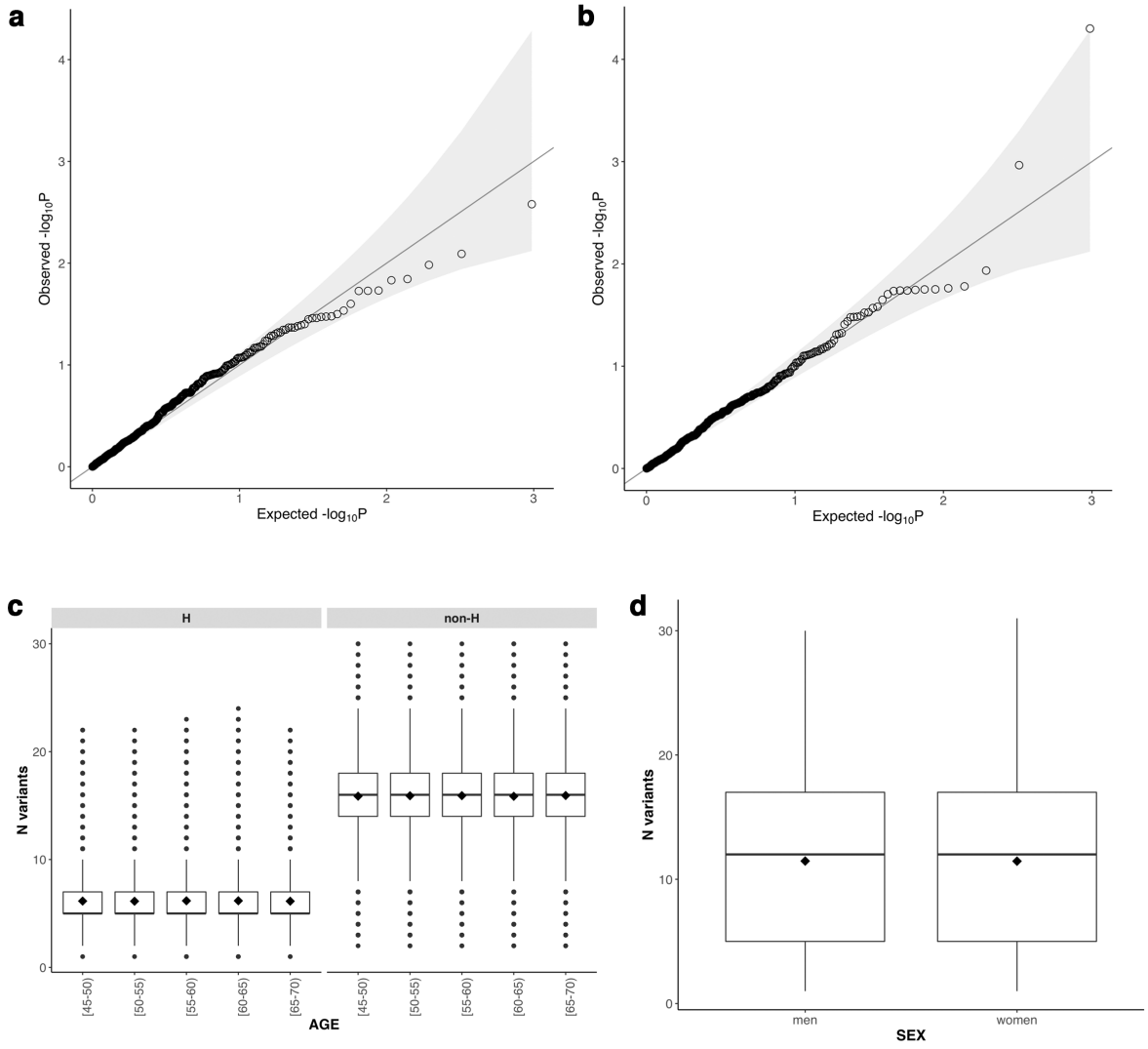
The figure shows the distribution of nuclear clusters (calculated using the first 10 nuclear principal components) across the territorial units (UK regions) from the Nomenclature of Territorial Units for Statistics version 2 (NUTS2) units in 327,665 participants with available birth coordinates.

Supplementary Figure 8. Principal component analysis of the "Full Set" of UK Biobank participants in comparison to GenBank and 1000 Genomes participants



Plots of the first 3 mtDNA principal components (mtPCs) for individuals in: (**a**) the Full Set of UK Biobank (N=483,626), (**b**) GenBank reference set used for imputation (N=17,815) and (**c**) 1000 Genomes individuals (N=2,419). Each dot represents a participant and are colored according to macro-haplogroup carrier status. mtPCs were calculated using genotyped mtSNVs (MAF>0.01). For each of the three data sets, plots on the left-hand side show mtPCs calculated using mtSNVs (with R^2<0.2 for UK Biobank and R^2<0.1 for GenBank and 1000 Genomes) while the plots on the right were generated without pruning correlated mtSNVs. The WTCCC reference set was not considered in this analysis because it includes almost exclusively individuals of European origin (98%).

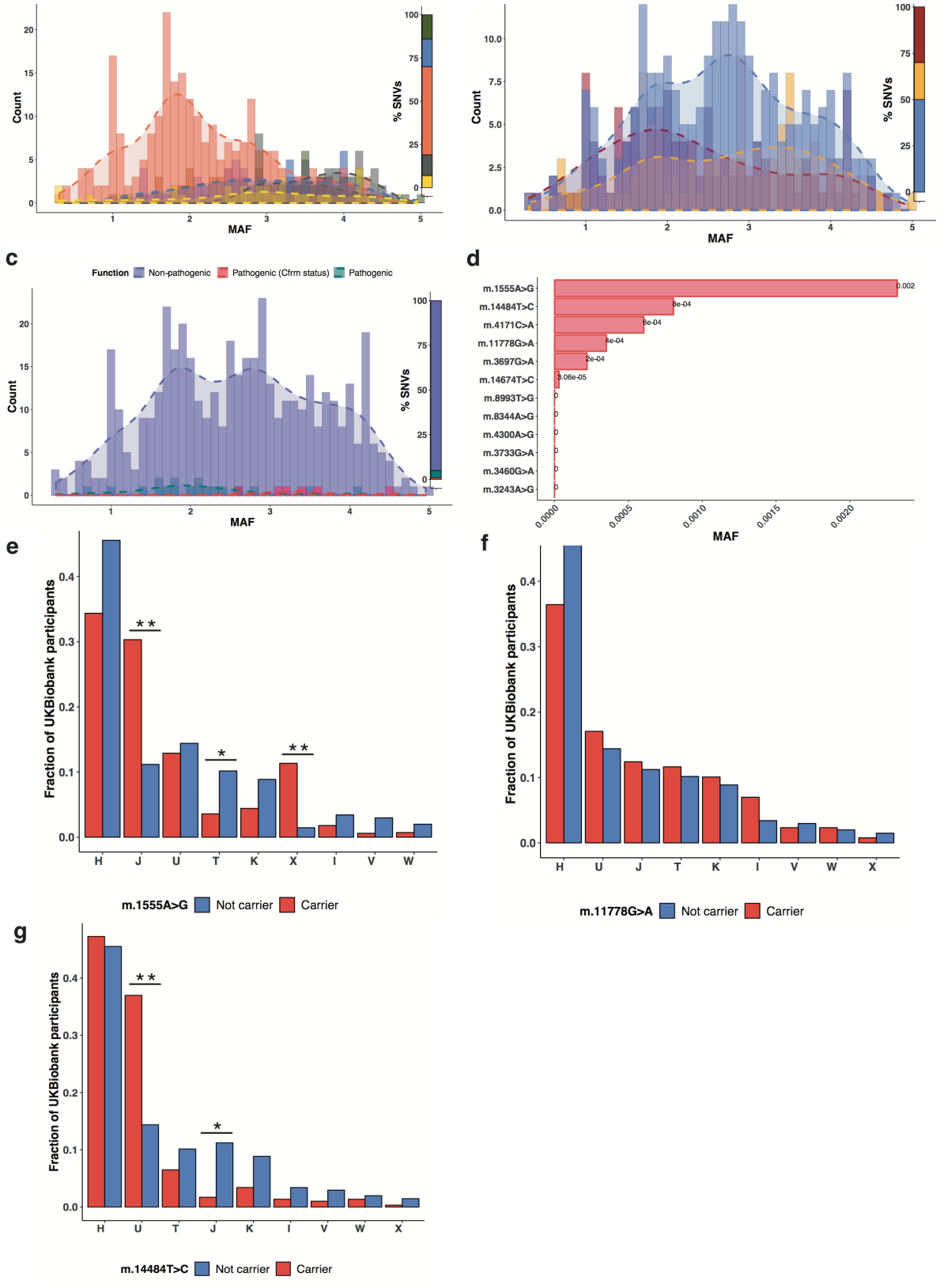
Supplementary Figure 9. Relationship between mtSNVs, age and sex



473 mtSNVs (open circles) were tested for association with age (a) and sex (b) in 358,916 EUR UK Biobank individuals and we provide the quantile-quantile (Q-Q) plots as a an indicator of deviation from normality (**a, b**). We used two-sided *P*-values from score tests adjusting for the first 10 principal components calculated using nuclear variants (nucPCs), sex and array are presented. (**c**) box-plot showing the distribution of the number (N) of mtSNVs (y-axis) a participant carries (out of 473 mtSNVs) stratified by age (categorical; x-axis) and H haplogroup carrier status. (**d**) boxplot showing the distribution of the number

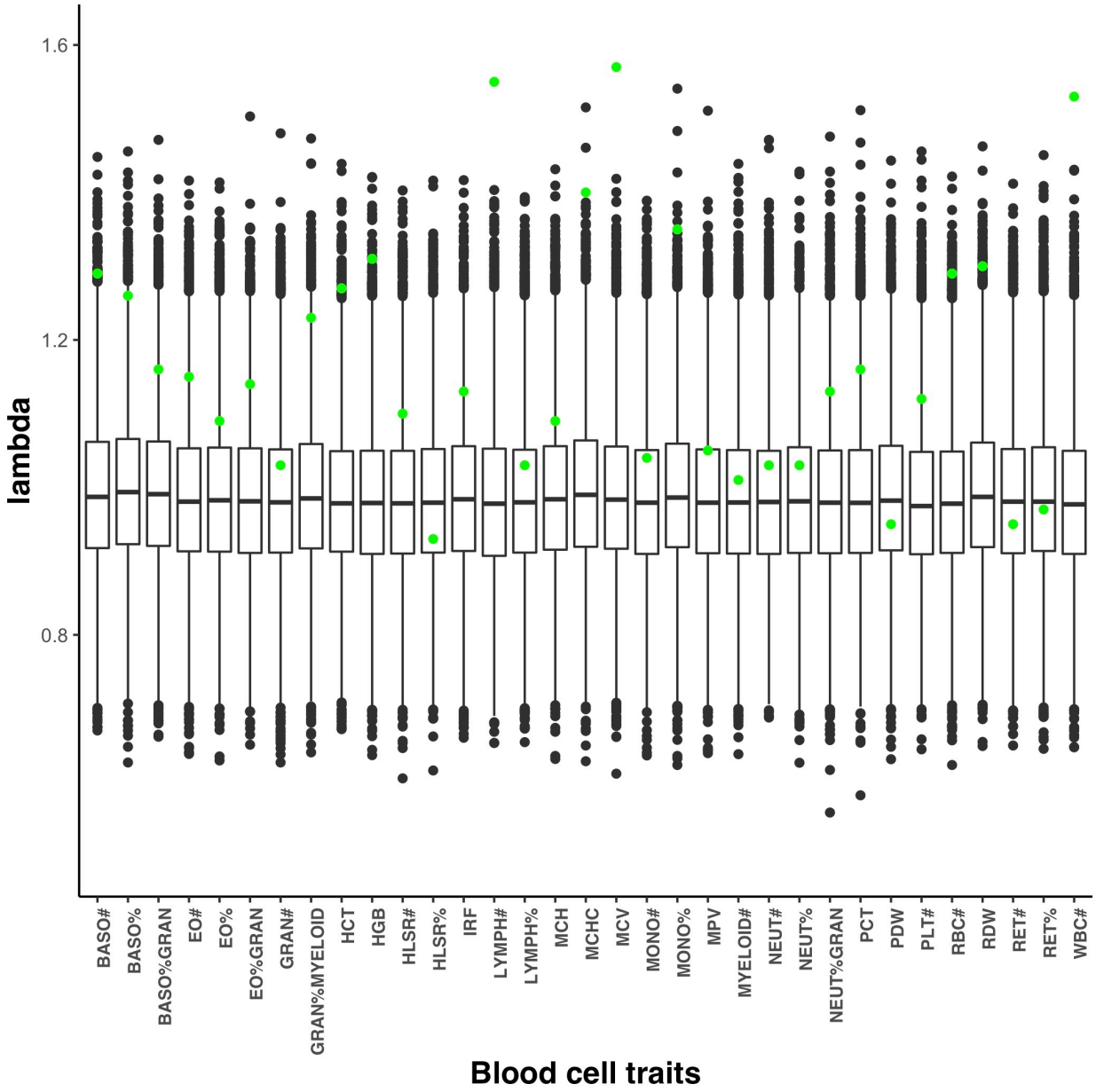
(N) of mtSNVs a participant carries (out of 473 mtSNVs) stratified by sex (x-axis). For (**a**) and (**b**) grey shading denotes the 95% confidence intervals for the observed *P*-values. For (**c**) and (**d**) mean N is denoted by a diamond. The lower and upper hinges of boxplots in (**c**) and (**d**) correspond to the first and third quartile of the distribution, with median in the center and whiskers spanning no further than 1.5*interquartile range. Black dots depict boxplots outliers

Supplementary Figure 10. Functional annotation of mtSNVs in the UK Biobank



Breakdown of the 473 mtSNVs (post-recalling and post-imputation) minor allele frequencies (MAFs) by (**a**) tagged haplogroup, (**b**) nucleotide change, (**c**) function (Cfrm = confirmed pathogenic status according to MITOMAP). Histograms show the distribution of MAFs in each category (dashed line indicates the corresponding smoothed densities). The stacked bar on the right side of each histogram plot shows the percentages of mtSNVs in each category. mtSNVs that are on European haplogroups tag: haplogroups H, J, U, T, K, X, I, V and W; mtSNVs falling on African haplogroups tag haplogroup L; mtSNVs on Asian haplogroups tag: haplogroups A, B, C, D, F, G, M, N9 and R9; SNVs on other haplogroups tag haplogroups E, O, P and Q. (**d**) MAF distribution of the 12 pathogenic mtSNVs variants with confirmed MITOMAP status. MAFs in **a-d** are $-\log_{10}$ transformed. **e**-**g**) Fractions of the UK Biobank individuals carrying the wild type (not carrier) and the mutated (carrier) allele of the three pathogenic mtSNVs, stratified by European haplogroups: **e**) m.1555A>G ($P = 3.7\text{x}10^{-50}$ for J haplogroup; $P = 1.2\text{x}10^{-04}$ for T haplogroup; $P = 2\text{x}10^{-82}$ for X haplogroup); **f**) m.11778G>A; **g**) m.14484T>C ($P = 6.4\text{x}10^{-05}$ for J haplogroup; $P = 2\text{x}10^{-11}$ for U haplogroup). Two-sided *P*-values were calculated with Wald test using multinomial analysis to evaluate significant enrichment or depletion for carriers of pathogenic mtSNVs with a specific haplogroup compared to the reference (H haplogroup) (**Supplementary Table 26**). Significant ($P < 5\text{x}10^{-5}$, denoted by **) and marginally significant enrichments ($P < 0.05$, denoted by *) are indicated.

Supplementary Figure 11. Distributions of inflation factors from the simulation studies



Distributions of inflation factors, lambda, calculated (1000x per trait) by sampling from the effects of nuclear genome SNP associations with blood cell traits reported by Astle et. al. (2016)[2]. The nuclear variants used in the calculation were matched in terms of minor allele frequency to the allele frequencies observed for the mitochondrial genome variants. Green dots represent the lambdas we observed for association of the mtSNVs with the listed blood cell trait using mtSNVs with R^2<0.2. The lower and upper hinges of boxplots

correspond to the first and third quartile of the distribution, with median in the center and whiskers spanning no further than 1.5*interquartile range. Black dots depict boxplots outliers. BASO# = basophil count; BASO% = percentage of basophils; EO# = eosinophil count; EO% = percentage of eosinophils; EO%GRAN = percentage of eosinophils in granulocyte fraction; GRAN# = granulocyte count; GRAN%MYELOID = % of granulocytes in the myeloid fraction; HCT = hematocrit; HGB = hemoglobin; HLSR# = high light scatter reticulocyte count; HLSR% = high light scatter reticulocyte percentage; IRF = immature fraction of reticulocytes; LYMPH# = lymphocyte count; LYMPH% = lymphocyte percentage; MCH = mean corpuscular hemoglobin; MCHC = mean corpuscular hemoglobin concentration; MCV = mean corpuscular volume; MONO# = monocyte count; MONO% = percentage of monocytes; MPV = mean platelet volume; MYELOID# = myeloid white cell count; NEUT# = neutrophil count; NEUT% = percentage of neutrophils; NEUT%GRAN = neutrophil percentage of granulocytes; PCT = plateletcrit; PDW = platelet distribution width; PLT# = platelet count; RBC# = red blood cell count; RDW = red blood cell width; RET# = reticulocyte count; RET% = reticulocyte fraction of red cells; WBC# = white blood cell count;

References

1. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
2. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).
3. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
4. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
5. Wei, W., Gomez-Duran, A., Hudson, G. & Chinnery, P. F. Background sequence characteristics influence the occurrence and severity of disease-causing mtDNA mutations. *PLoS Genet.* **13**, e1007126 (2017).
6. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
7. Kogelnik, A. M., Lott, M. T., Brown, M. D., Navathe, S. B. & Wallace, D. C. MITOMAP: a human mitochondrial genome database. *Nucleic Acids Res.* **24**, 177–179 (1996).
8. Sproule, D. M. & Kaufmann, P. Mitochondrial encephalopathy, lactic acidosis, and strokelike episodes: basic concepts, clinical phenotype, and therapeutic management of MELAS syndrome. *Ann. N. Y. Acad. Sci.* **1142**, 133–158 (2008).

9. Huoponen, K., Vilkki, J., Aula, P., Nikoskelainen, E. K. & Savontaus, M. L. A new mtDNA mutation associated with Leber hereditary optic neuroretinopathy. *Am. J. Hum. Genet.* **48**, 1147–1153 (1991).
10. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386-394 (2009).
11. Hudson, G., Gomez-Duran, A., Wilson, I. J. & Chinnery, P. F. Recent mitochondrial DNA mutations increase the risk of developing common late-onset human diseases. *PLoS Genet.* **10**, e1004369 (2014).
12. Emery, L. S., Magnaye, K. M., Bigham, A. W., Akey, J. M. & Bamshad, M. J. Estimates of continental ancestry vary widely among individuals with the same mtDNA haplogroup. *Am. J. Hum. Genet.* **96**, 183–193 (2015).
13. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
14. Calabrese, C. *et al.* MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics* **30**, 3115–3117 (2014).
15. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
16. Weissensteiner, H. *et al.* HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58-63 (2016).

17. Clima, R. *et al.* HmtDB 2016: data update, a better performing query system and human mitochondrial DNA haplogroup predictor. *Nucleic Acids Res.* **45**, D698–D706 (2017).
18. Tachmazidou, I. *et al.* Whole-Genome Sequencing Coupled to Imputation Discovers Genetic Signals for Anthropometric Traits. *Am. J. Hum. Genet.* **100**, 865–884 (2017).
19. Pacheu-Grau, D. *et al.* Mitochondrial antibiograms in personalized medicine. *Hum. Mol. Genet.* **22**, 1132–1139 (2013).
20. Meyer, J. N., Hartman, J. H. & Mello, D. F. Mitochondrial Toxicity. *Toxicol. Sci.* **162**, 15–23 (2018).
21. Vial, G., Detaille, D. & Guigas, B. Role of Mitochondria in the Mechanism(s) of Action of Metformin. *Front Endocrinol (Lausanne)* **10**, 294 (2019).
22. Gauderman, W. J. Sample size requirements for association studies of gene-gene interaction. *Am. J. Epidemiol.* **155**, 478–484 (2002).