nature research

Corresponding author(s): Patrick Chinnery and Joanna Howson

Last updated by author(s): Jan 30, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Data were collected by UK Biobank and we downloaded data from the UK Biobank repository using UK Biobank helper programs (ukb_conv, ukbfetch, ukbgene, ukb_md5, ukb_unpack) July 2017 for genetic data, May 2018 for phenotype data, April 2019 for Biomarker data

Data analysis

We used the following open source softwares (that are provided in the manuscript):
-FastqQC v0.11.5 to check quality of fastq files generated with Illumina Hiseq.
-MToolBox v1.2 to perform mitochondrial variant calling on fastq files.
-Haplogrep2 v2.1.1 to perform haplogroup predictions.
-IMPUTE version 2 to perform imputation using 17,815 GenBank complete genomes and 5,271 biallelic mitochondrial SNVs.
-Gtool v0.7.5 to set genotypes in impute2 files to hard calls in ped files.
-QCTOOL v2 to perform mtSNV allele dosages and MAF calculation and all sorts of data parsing and file conversion on gen and bgen files
-RVTESTS version 0171009 to perform Wald and Score test for association between traits and mitochondrial SNVs, fitting different set of covariates depending on the trait. This tool was also used to generate meta-analysis files that were used as input for the gene level analysis.
-STATA v14.2 to perform likelihood ratio test for association between traits and mitochondrial SNVs, fitting different set of covariates depending on the trait. This tools was also used to perform fine-mapping of loci with multiple mitochondrial SNVs associations.
-PLINK v1.9 to calculate per mitochondrial SNV and per sample call rates, LD-pruning.
-FINEMAP v1.3 to perform fine-mapping of loci with multiple mitochondrial SNVs associations.
-LDSTORE v1 to calculate correlation between mitochondrial SNVs used for fine-mapping.
-rareMETALs R package v6.8 to perform region-based meta-analysis using the SKAT method to compute a two-sided gene-level p-value.
-FlashPCA version 2 to define a genetically European group of UK Biobank participants using principal component analysis.
-reverse_geocoder Python module v1.5.1 to retrieve administrative regions (UK region, county and city), which were intersected with UK NUTS level 2 data, used as covariates for sensitivity analyses.
-Quanto v1.2.4 for power calculation
-https://github.com/clody23/UKBiobank_mtPheWas is a github repository that contains the R script used for variant recalling and several

other R and python scripts used for plotting
- R v3.4, R v3.5, python 2.7, python 3 to run custom R or python scripts

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We have made the full summary statistics available through the Cambridge repository: https://app.box.com/s/vu9ewgd9sv7ga3ua0lsk0cql444xm3ht and on Zenodo: https://doi.org/10.5281/zenodo.4609973; We have also used data from the following publicly available data bases: www.mitomap.org; https://www.ncbi.nlm.nih.gov/clinvar/; https://www.internationalgenome.org/;https://www.hmtvar.uniba.it/. We provide the code used to generate main and supplementary plots, and variant recalling in a Github repository: https://github.com/clody23/UKBiobank_mtPheWas.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used 488,377 UK Biobank participants and performed extensive variant and sample QC. We defined two groups of samples: 483,626 pan-ancestry individuals and a European set of 358,916 unrelated individuals. These are the maximum data set sizes available for any given analyses. Power calculations are provided in Supplementary Figure 1.
Data exclusions	We removed genotyping plates with problematic batch effects, intensity outliers, low sample call rate and non-Europeans and related individuals, exact numbers are provided in Figure 1. Individuals with outlying phenotype measurements were excluded as such individuals can potentially distort the statistical analyses and violate their assumptions. Details on sample exclusion are provided in Supplementary Table 15 and 16 for all the phenotypes tested. Exclusion criteria were pre-selected.
Replication	We did not attempt replication as this is the largest sample size available for these analyses to date.
Randomization	UK Biobank samples were randomised by sex and recruitment centre prior to genotyping.
Blinding	During genetic QC, analysts were blind to phenotype information. During phenotype QC, analysts were blind to genetic information. None of the authors took part in the data collection. We did not have access to group allocation information during data collection.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The description of the UK Biobank, GeneBank, 1000 Genome and WTCCC cohorts used for this analysis is available in the "Study populations" Method's section of this manuscript and in the "GenBank", "1000 Genomes" and "Wellcome Trust Case Control Cohort" sections of Supplementary Information. UK Biobank participants used for association analyses were 358,916, all with mitochondrial and nuclear European ancestry (age range: 39-72 yrs, mean age = 57 yrs; 54% female). GenBank study participants were downloaded as mitochondrial complete genomes from NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) (N=17,815), but no age and sex information are available. The nuclear ancestry was unknown for ~40% of the genomes, while mitochondrial ancestry was known for 100% of the genomes. Of the remaining GenBank genomes with known nuclear-mitochondrial ancestry, 37% were Europeans, 33% Asians, and 30% African. Whole Exome Sequencing of 1000 Genomes individuals were downloaded from the 1000 Genomes project site (http://www.1000genomes.org/) (N=2,419), with information about sex (51% female). The 1000 Genomes participants come from 26 different populations from around the world (37% Asian, 28% African, 21% European, 14% Admixed American). No phenotype data has been collected for these participants. The WTCCC participants included in the study (N=763; 50% female; 98% with mitochondrial and nuclear European ancestry), living within England, Scotland or Wales ('Great Britain') and were sequenced with Illumina Miseq. The cohort we used is part of the control individuals from the 1958 British Birth Cohort (58C) and individuals selected from blood donors recruited as part of the WTCCC project.
Recruitment	The description of how the UK Biobank, GeneBank, 1000 Genomes and WTCCC cohorts were recruited for this analysis is available in the "Study populations" Method's section of this manuscript and in the "GenBank", "1000 Genomes" and "Wellcome Trust Case Control Cohort" sections of Supplementary Information, as follows: The UKBB study is a prospective population study (N=502,682, age range: 39-72, predominantly western-Europeans) of UK residents recruited at 22 assessment centers across the UK between 2006 and 2010. Participants attended a center, where they were deeply phenotyped, including various physical measurements, extensive health and lifestyle questionnaires, and biological samples. 17,815 complete GenBank human genomes were downloaded from the NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) as explained in Wei W et al., 2017 (PMID: 29253894). Requirements for inclusion in the 1000 Genomes project were: individuals to be sequenced in the project had to have been explicitly consented for broad use and public distribution of extensive genotype or sequence data over the Internet. Cell lines from the samples had to be available to the broad research community, to maximize the value of the 1000 Genomes data by facilitating follow-up studies and use of the sequence data to map cellular phenotypes. Further information about 1000 Genomes recruitment is available at The 1000 Genomes Project Consortium, 2010 (PMID: 20981092). The WTCCC cohort is part of a common set of 3,000 nationally-ascertained controls from England, Scotland and Wales. These controls come from two sources: 1,500 are representative samples from the 1958 British Birth Cohort and 1,500 are blood donors recruited by the three national UK Blood Services.
Ethics oversight	UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC). 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.The 1000 Genomes consortium subjects provided their consent as explained in https://www.internationalgenome.org/sample_collection_principles.

Note that full information on the approval of the study protocol must also be provided in the manuscript.