nature portfolio

Corresponding author(s): Slavé Petrovski

Last updated by author(s): Jul 11, 2021

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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	Single-sample processing, on Amazon Web Services (AWS) cloud compute platform. * Conversion of sequencing data in BCL format to FASTQ format and the assignments of paired-end sequence reads to samples based on 10-base barcodes; bcl2fastq v2.19.0 https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html * read alignment and variant calling performed on Illumina DRAGEN Bio-IT Platform Germline Pipeline v3.0.7 to align the reads to the GRCh38 genome reference and perform small variant SNV and indel calling. SNVs and indels were annotated using SnpEFF v4.3 against Ensembl Build 38.92. We further annotated all variants with their gnomAD minor allele frequencies (gnomAD v2.1.1 mapped to GRCh38). * For ancestry we used PEDDY v0.4.2 with the ancestry labelled 1K Genomes Project reference sequence data for genetic ancestry predictions. * For relatedness we use ukb_gen_samples_to_remove() function from the R package ukbtools v0.11.3.
Data analysis	* PheWAS and exWAS association tests were performed using a custom built frame PEACOK (PEACOK 1.0.7), which is an extension and enhancement of PHESANT. PEACOK 1.0.7 can be found: https://github.com/astrazeneca-cgr-publications/PEACOK/versions/1.0.7 * Large-scale compute was done using AWS Batch computing environment. * exWAS association tests for Chapter IX binary traits across all autosomes were performed using SAIGE (SAIGE v0.43 https://github.com/weizhouUMICH/SAIGE) and REGENIE (REGENIE v 2.0.2, https://github.com/rgcgithub/regenie). * Using exome sequence-derived genotypes for 43,889 biallelic autosomal SNVs located in coding regions as input to the kinship algorithm included in KING v2.2.3. * Various downstream analysis and summarization were performed using R v3.4.3 https://cran.r-project.org. R library data.table (v1.12.8), MASS (7.3-51.6), tidyr (1.1.0) and dplyr(1.0.0) were also used.

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All WES data described in this paper are publicly available to registered researchers through the UKB data-access protocol. Additional information about registration for access to the data are available at http://www.ukbiobank.ac.uk/register-apply/. Data for this study were obtained under Resource Application Number 26041. Furthermore, a web portal to interact with the 200K version of these analyses is now publicly available at www.azphewas.com

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 the 302,355 UK Biobank participants for whom exome sequencing were available in December 2019. Further subsetting was applied as described in the manuscript. No sample size calculations for power were performed.
Data exclusions	At the sample level, predefined exclusion criteria as detailed in the manuscript were: did not pass sequencing quality control thresholds, and are related.
Replication	Identified signals were also compared to existing catalogues and other biobank-based PheWAS (namely Finngen public r5, ClinVar and OMIM) and were highly successful with 88.6% of the binary trait collapsing analysis gene-phenotype examples overlapping with these replication sets.
Randomization	This study is observational. Randomization was not applicable to this study.
Blinding	This study is observational, using coded de-identified data. Blinding was not applicable to this study.

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 and archaeology
☒	☐	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

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	Of the 302,355 UK Biobank participants for whom exome sequencing were available in December 2019, the average age at recruitment was 56.5 years, 54% were female, and 95% were of European ancestry. Additional population characteristics of the overall UK Biobank cohort are available to the public at http://www.ukbiobank.ac.uk/
Recruitment	Participants were recruited to the UK Biobank on a voluntary basis. Approx 500K individuals 40-69 years of age in 2006-2010 volunteered. Informed consent was obtained for all participants. It has previously been observed that participants are less

	likely to live in socioeconomically deprived areas than non-participants, and they tend to be healthier than non-participants, which may impact some of the reporting rates in comparison to what could be observed through random sampling from the UK population. Fry et al (10.1093/aje/kwx246).
Ethics oversight	The protocols for UK Biobank are overseen by The UK Biobank Ethics Advisory Committee (EAC), for more information see https://www.ukbiobank.ac.uk/ethics/ and https://www.ukbiobank.ac.uk/wp-content/uploads/2011/05/EGF20082.pdf

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