nature portfolio

Corresponding author(s): Eric Jorgenson

Last updated by author(s): Sep 11, 2022

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

No software was used for data collection.

Data analysis

The is publicly available and can be found at https://github.com/rgcgithub/regenie. The REGENIE software for whole genome regression, which was used to perform all genetic association analysis, is available at https://github.com/rgcgithub/regenie. GCTA v1.91.7 was used for approximate conditional analysis. SHAPEIT4.2.0 was used for phasing of SNP array data. Imputation was completed with IMPUTE5. Somatic calling was done with Mutect2 (GATK v4.1.4.0). We use Plink1.9/2.0 for genotypic analysis as well as for constructing polygenic risk scores. FINEMAP was used for fine-mapping, and genetic correlations were calculated using LDSC version 1.0.1 with annotation input version 2.2. Beyond standard R packages, visualization tools, and data processing libraries (e.g. dplyr, ggplot2, data.table), we used the survival (version 3.2.13) and survminer (version 0.4.9) packages for survival analyses, the MendelianRandomization package for Mendelian Randomization (version 0.6.0), and the winnerscurse package (version 0.1.1, https://amandaforde.github.io/winnerscurse/) to adjust GWAS effect size estimates for the effects of Winner's Curse.

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

Individual-level sequence data, CHIP calls, and polygenic scores have been deposited with UK Biobank and will be freely available to approved researchers, as done with other genetic datasets to date9. Individual-level phenotype data are already available to approved researchers for the surveys and health-record datasets from which all our traits are derived. Instructions for access to UK Biobank data is available at https://www.ukbiobank.ac.uk/enable-your-research. Summary statistics from UK Biobank trait are available in the GWAS Catalog (accession IDs are listed in the tables description sheet available in the supplementary data tables excel file). As described in Backman et al.9, the HapMap3 reference panel was downloaded from ftp://ftp.ncbi.nlm.nih.gov/hapmap/, GnomAD v3.1 VCFs were obtained from https://gnomad.broadinstitute.org/downloads, and VCFs for TOPMED Freeze 8 were obtained from dbGaP as described in https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-8. Data used for replication, such as DiscovEHR exome sequencing and genotyping data, and derived CHIP calls, can be made available to qualified, academic, non-commercial researchers upon request via a Data Transfer Agreement with Geisinger Health System (Contact person: Lance Adams, ljadams@geisinger.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

Sample size was not predetermined. Association analyses were restricted to the intersection of samples with both exome sequence and array genotypes available after QC. See methods section "Exome sequencing" for details on QC performed. All samples that pass genotype QC and with non-missing phenotype data were included in association analyses. Sample sizes represent all available samples from both UKB and GHS, which together represent a ten-fold increase in sample size relatively to prior publications in the literature.

Data exclusions

Phenotype selection and QC was performed as described in methods section "Health- and behavior-related phenotypes." Variant level QC was performed as described in methods section "Exome sequencing." Variants with minor allele count less than five were excluded from association testing. The minor allele count threshold was pre-determined based on extensive simulations performed with REGENIE. See https://www.nature.com/articles/s41588-021-00870-7 for additional details.

Replication

Replication was attempted for all significant variant-trait associations available for follow-up in the DiscovEHR study. As noted in the manuscript, we estimated that we had sufficient power in GHS to detect 19.99 true and directionally consistent associations across lead SNPs from the 24 loci we identified in UKB, and achieved nominally significant ($p<0.05$) replication for 15 SNPs (Table S2).

Randomization

Randomization was not required for the analyses completed in this study. To control for confounding, we performed association analysis with the following covariates included in the regression model: age, age-squared, sex, age-x-sex, 10 ancestry-informative principal components, six exome sequence batch indicator variables, and 20 principal components derived from exome variants with a MAF between 2.6x10-5 and 1%.

Blinding

Blinding was not required for the analyses completed in this study. Participant recruitment and phenotype collection were obtained without prior knowledge of sample genotypes. Association analyses were performed with all available samples, without any filtering based on sample genotypes.

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	The UK Biobank is a prospective cohort study previously described in detail by Bycroft et al, Nature 2018 (https://www.nature.com/articles/s41586-018-0579-z). Briefly, 94.7% of sequenced participants are of European ancestry, 54.2% are female, the average age at assessment is 58, and the mean BMI is 26. 45% of participants report a history of smoking, and each participant reports 8 inpatient ICD10 3D codes, on average. See supplementary table 1 for additional details.
Recruitment	Please see Bycroft et al, Nature 2018.
Ethics oversight	Ethical approval for the UK Biobank was previously obtained from the North West Centre for Research Ethics Committee (11/NW/0382). The work described herein was approved by UK Biobank under application number 26041. Approval for DiscovEHR analyses was provided by the Geisinger Health System Institutional Review Board under project number 2006-0258. Informed consent was obtained for all study participants.

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