nature research

Corresponding author(s): Seunggeun Lee
Cristen Willer

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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	SAIGE is implemented as an open-source R package available at https://github.com/weizhouUMICH/SAIGE/.
Data analysis	BOLT-LMM v2.3 https://data.broadinstitute.org/alkesgroup/BOLT-LMM/ and GMMAT v0.7 https://github.com/hanchenphd/GMMAT were used in both simulation and real data analysis. In addition, GEMMA 0.96 https://github.com/genetics-statistics/GEMMA was used in the performance comparison.

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 upon request. 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

SAIGE is implemented as an open-source R package available at https://github.com/weizhouUMICH/SAIGE/.
The GWAS results for > 1400 binary phenotypes with the PheCodes constructed based on ICD codes in UK Biobank using SAIGE are currently available for public download at https://www.leelabsg.org/resources
Information for all the phenotypes can be found in the website above. We also display the results in the Michigan PheWeb http://pheweb.sph.umich.edu/UKBiobank, which consists of Manhattan plots, Q-Q plots, and regional association plots for each phenotype as well as the PheWAS plots for every genetic marker.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	We analyzed publicly available UKBiobank data of white British samples (N=408961). We expect this large sample size can provide enough power to detect small and moderate effects of genetic associations. Since we analyzed a large number of phenotypes with wide ranges of case-control ratios, we didn't perform sample size calculation. Power analysis with a much smaller sample size (N=10,000) was performed to compare the performance of different methods, and show that the proposed approach is more powerful than the existing approaches.
Data exclusions	Due to QC issues in non-HRC imputed markers, we restricted our analysis to directly genotyped or HRC imputed markers. Non White British samples were excluded from the analysis.
Replication	We searched GWAS catalog to check whether GWAS significant loci were known (replicated) or novel. Experimental replication was not attempted.
Randomization	Randomization of experimental groups were not required to this study. Participants were allocated into experimental groups according to their disease status.
Blinding	We used coded public data, and hence were blinded.

Reporting for specific materials, systems and methods

Materials & experimental systems

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

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	Participant in the UK Biobank studies (http://www.ukbiobank.ac.uk/), coded publicily available data

Recruitment	UK Biobank recruited 500,000 people aged between 40-69 years in 2006-2010 from across UK. They have undergone measures, provided blood, urine and saliva samples for future analysis, detailed information about themselves. More details can be found at http://www.ukbiobank.ac.uk/