nature portfolio

Corresponding author(s): Hyunghoon Cho, Jean-Pierre Hubaux, Bonnie Berger

Last updated by author(s): Nov 20, 2024

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	For data processing and baseline association analysis, we used PLINK v2.0 (https://www.cog-genomics.org/plink/2.0/) and the scikit-learn package v1.3.0 (with Python 3.8) for PCA-based pipeline and REGENIE v2.2.4 (https://rgcgithub.github.io/regenie/) for LMM-based pipeline. Michigan Imputation Server v1 used Minimac4 for imputation, which we utilized to obtain imputed genotypes for AMD analysis. GrafPop 1.0 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/Software.cgi) was used for ancestry group assignment. Our code for secure and federated GWAS is available as open source software (v0.1.2) at: https://github.com/hhcho/sfgwas (DOI: 10.5281/zenodo.14726447).

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

The three datasets used for comparison with the prior work on Secure GWAS (Cho et al., Nature Biotechnology, 2018) are available via NIH dbGaP with accession numbers phs000716.v1.p1 (lung cancer), phs000346.v2.p2 (bladder cancer), and phs001039.v1.p1 (AMD). The eMERGE consortium data is also available via dbGaP (phs000888.v1.p1). Data access applications for the UK Biobank data can be submitted at: https://www.ukbiobank.ac.uk/.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	With the exception of the lung cancer dataset, all datasets in our study includes individuals of both sexes. Where available, we included self-reported sex as a covariate in the association analysis. The lung cancer dataset is obtained from a targeted study of never-smoking East Asian women and includes only females. Sharing of individual-level data for all our datasets is consented by the participants through the informed consent procedure of each respective study.
Reporting on race, ethnicity, or other socially relevant groupings	All our experiments analyze cohorts that were recruited by the original studies. For a subset of experiments considering only individuals of European ancestry, the ancestry label was determined based on self-reported ethnicity for UK Biobank and GrafPop-inferred ancestry for eMERGE and IAMDGC (P_e > 90%). Confounding due to ancestry backgrounds in our analysis is controlled by either including genotype principal components or using linear mixed models.
Population characteristics	Lung cancer dataset: never-smoking East Asian women in all age groups, including 5,088 lung cancer cases and 4,090 controls, recruited by Female Lung Cancer Consortium in Asia (FLCCA). Bladder cancer dataset: individuals of European descent, including both sexes and all age groups, 6,211 cases with histologically confirmed primary carcinoma of the urinary bladder and 6,849 controls, pooled from 16 studies conducted in Spain, Finland, and USA. AMD dataset: 9,648 cases with geographic atrophy (GA), choroidal neovascularization (CNV), or mixed GA/CNV and 13,035 controls, recruited by International Age-Related Macular Degeneration Genomics Consortium, including both sexes and all age groups. eMERGE dataset: a prospective cohort including individuals with electronic medical records (EMRs) and linked genomic data, recruited by individual participating sites in the eMERGE Network (both sexes, aged 3-75). UK Biobank dataset: a prospective cohort of individuals in the UK with EMR and linked genomic data, recruited by the UK Biobank (both sexes, aged 40-69).
Recruitment	See above.
Ethics oversight	Access to all datasets were approved by respective data access committees through NIH dbGaP and UK Biobank Access Management System.

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

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	No sample size calculation was performed to predetermine the sample sizes as we used the largest available public datasets and did not generate new data. We chose a variety of datasets to demonstrate our methods on a range of sample sizes.
Data exclusions	We adopted the following standard quality control filters: genotype missing rate per SNP <0.1, minor allele frequency (MAF) >0.05, and Hardy–Weinberg equilibrium chi-squared test statistic <23.928 (p-value >10^{-6}). We used the same set of filters for the UKB and cross-biobank AMD data, except for MAF >0.001 to account for the larger size of these datasets. These thresholds were established prior to the experiments.
Replication	We confirmed that our experiments can be reproduced based on the code we provide.
Randomization	This is not relevant to our study, as our analyses do not involve distinct experimental groups. Instead, for each dataset, a single analysis is conducted encompassing all samples.

Blinding	This is not relevant to our study, as our analyses do not involve distinct experimental groups. Instead, for each dataset, a single analysis is conducted encompassing all samples.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*