nature portfolio

Corresponding author(s): Bibo Jiang, Dajiang Liu

Last updated by author(s): Nov 19, 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	We did not collect new data
Data analysis	We developed GPS software: https://github.com/wangc29/gps We used the following other softwares for data-analysis, Lassosum (0.4.5): https://github.com/tshmak/lassosum bigsnpr (1.12.2): https://github.com/privefl/bigsnpr PRS-CS (1.1.0): https://github.com/getian107/PRScs MTAG (2017-04-07):https://github.com/JonJala/mtag TL-PRS (1.0.0): https://github.com/ZhangchenZhao/TLPRS multivariate Lassosum (version 1.0.0): https://github.com/abureau/multivariateLassosum ADMIXTURE (1.3.0):https://dalexander.github.io/admixture/download.html PLINK (2.0): https://www.cog-genomics.org/plink/2.0/ REGENIE (2.2.4): https://rgcgithub.github.io/regenie/ PheWAS (2023-5-31):https://github.com/PheWAS/PheWAS We also used the following R packages: psychometric version 2.3, ROCR version 1.0-11, MASS version 7.3-51.4, data.table version 1.14.0, ggplot version 3.3.5

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

1. GWAS summary statistics for autoimmune diseases are from the following publications (PubMed ID): 26502338,28714469,18204446,18204098,34594039,29848360,34594039,33310728
2. The EHR lab test, diagnosis and genotype data of patients from the All of Us (https://www.researchallofus.org) biobank can be accessed by approved researchers for the All of Us study.
3. The EHR lab test, diagnosis and genotype data of patients from the Vanderbilt biobank (https://victr.vumc.org/biovu-description/) can be accessed via Vanderbilt University.

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	Sex is used as a covariate when conducting GWAS analysis.
Reporting on race, ethnicity, or other socially relevant groupings	Ancestries for participants from the Vanderbilt biobank are determined by ADMIXTURE software using their genotype data. Ancestries for participants from the All of US biobank are determined by utilizing pre-calculated genetic ancestry.
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

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. The study uses all existing datasets we can access. 1. Summary statistics of GWAS studies of autoimmune diseases are retrieved from previous publications. The resulting sample size is 37,828 for Rheumatoid Arthritis (RA) and 16,654 for Systemic Lupus Erythematosus (SLE). 2. The sample size for patients with preclinical and full-blown RA and SLE are calculated based on extracted data from the Vanderbilt biobank and the All of Us biobank, European samples: Vanderbilt Biobank: preclinical RA, 414; full-blown RA 329 Vanderbilt Biobank: preclinical SLE, 2784; full-blown RA 394 All of Us Biobank: preclinical RA, 397, full-blown RA 126 All of Us Biobank: preclinical SLE, 2150; full-blown RA 184 Non European samples: All of US Biobank's African ancestry samples: preclinical RA, 283, full-blown RA 72 All of US Biobank's African ancestry samples: preclinical SLE 735 full-blown SLE 159 All of US Biobank's American ancestry samples: preclinical SLE 734 full-blown SLE 150
Data exclusions	Patients from non-European ancestry groups in Vanderbilt Biobank are excluded due to their limited sample size for constructing polygenic risk scores. Patients from non-European ancestry groups with sample size less than 100 in All of US Biobank are excluded due to limited sample size for model evaluation.
Replication	We trained our model using data from the Vanderbilt biobank and further tested the trained model using the All of US biobank.

Randomization	This is not a clinical trial. No randomization was performed.
Blinding	No blinding applicable. Our study utilized existing datasets that are de-identified.

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	NA
Novel plant genotypes	NA
Authentication	NA