nature portfolio

Corresponding author(s): Farhad Hormozdiari

Last updated by author(s): March 10, 2023

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 is used for data collection. We have used available UK Biobank and COPDGene data.

Data analysis	All previous developed ML algorithms utilized in this work are implemented in TensorFlow (v 2.9.0) including: Adam, ReLu, Early stopping, etc. BOLT-LMM software (v2.3.5): https://data.broadinstitute.org/alkesgroup/bolt-lmm BaselineLD annotations: https://data.broadinstitute.org/alkesgroup/ldscore DeepNull (v0.2.2): https://github.com/Google-Health/genomics-research/tree/main/nonlinear-covariate-gwas GWAS Catalog: https://www.ebi.ac.uk/gwas/ PLINK software (v1.9): https://www.cog-genomics.org/plink1.9 TensorFlow (v 2.9.0): https://www.tensorflow.org UK Biobank study: https://www.ukbiobank.ac.uk UCSC LiftOver: https://genome.ucsc.edu/cgi-bin/hgLiftOver GARFIELD software (v2): https://www.ebi.ac.uk/birney-srv/GARFIELD/ scikit-learn (v1.0.2) https://scikit-learn.org/stable/ S-LDSC (v1.0.1) https://data.broadinstitute.org/alkesgroup/ldscore GREAT (v4.0.4) http://great.stanford.edu Cell-type specific gene expression annotations: https://alkesgroup.broadinstitute.org/LDSCORE/LDSC_SEG_ldscores/ Multi_tissue_gene_expr_1000Gv3_ldscores.tgz Cell-type specific chromatin annotations: https://alkesgroup.broadinstitute.org/LDSCORE/LDSC_SEG_ldscores/ Multi_tissue_chromatin_1000Gv3_ldscores.tgz Vizier (v0.1.1): https://github.com/google/vizier

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

UK Biobank study:https://www.ukbiobank.ac.uk and our access was approved under Application 65275. This research used data generated by the COPDGene study (dbGaP accession phs000179.v6.p2), which was supported by NIH grants U01 HL089856 and U01 HL089897. ICGC (International COPD Genetics Consortium) genome-wide association summary statistics was obtained from dbGaP under accession phs000179.v5.p2. SpiroMeta summary statistics was obtained from LDHub. The full ML-based COPD summary statistics are currently available on our GitHub repository page (https://github.com/Google-Health/genomics-research/releases/tag/v0.1.0-ML-COPD) and uploaded to the GWAS catalog. The raw ML-based COPD liability scores will be returned to UK Biobank.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	NA
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	We use all the samples provided by UK Biobank. We utilize all the European samples in our analysis.
Data exclusions	We removed samples with excess heterozygosity, missingness, or putative sex chromosome aneuploidy as defined by Bycroft et al 2018 (https://www.nature.com/articles/s41586-018-0579-z). We further limited to individuals of European genetic ancestry as defined in Alipanahi

et al 2021 (https://doi.org/10.1016/j.ajhg.2021.05.004)

Replication	Data replication was not apply as we did not collected new data. We did use previous GWAS to preform replication analysis of our novel hits/ loci. We have used GWAS catalog, previous GWAS (Shrine et al. 2019 Nature Genetics; DOI: https://doi.org/10.1038/s41588-018-0321-7 , Sakornsakolpat et al 2019 Nature Genetics; DOI: 10.1038/s41588-018-0342-2, ICGC, GBMI, and SpiroMeta) to replicate our GWAS finding.
Randomization	We use all the samples provided by UK Biobank. We only randomized the data in the process of 2-fold in ML-based COPD However, this is done in our code and we have used the exact seed as provided in the Github so the result is replicated.
Blinding	No blinding has been performed in our analysis as we did not collect any new data. We use all the samples provided by UK Biobank.

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

Methods

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