nature research

Corresponding author(s): Bonnie Berger

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

We chose our benchmark GWAS data sets from the dbGaP repository such that their sample sizes are representative of large-scale GWAS. In order to demonstrate the scalability of our method to sample sizes larger than what was available to us, we simulated larger data sets containing up to 100K individuals, which provided us enough data to support our extrapolation to 1 million individuals.

2. Data exclusions

Describe any data exclusions.

Our quality control filters include: genotype missing rate per individual < 0.05 and per SNP < 0.1, individual heterozygosity rate > 0.25 and < 0.30, minor allele frequency > 0.1, and Hardy-Weinberg equilibrium test chi-squared statistic < 28.3740 (p-value < 10E-7). We excluded the heterozygosity filter for the bladder cancer and AMD data sets due to the distribution of heterozygosity rates being considerably different in these data sets.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Our computational experiments can be reliably reproduced based on the code we provide.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

There were no group allocations in this study.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

There were no group allocations in this study.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

Nature Biotechnology: doi:10.1038/nbt.4108

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	□	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☒	□	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	□	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

We provide a link to the C++ implementation of our method (secure GWAS protocol), which we used to obtain the results presented in our study. Our code is also provided as Supplementary Code.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials are used in this study.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies are used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No cell lines are used.

b. Describe the method of cell line authentication used.

No cell lines are used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No cell lines are used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No cell lines are used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No research animals are used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	All individuals in the lung cancer data set are non-smoking females in East Asia. Cases are histologically confirmed lung cancer. For the bladder cancer data set, cases are histologically confirmed primary carcinoma of the urinary bladder. The AMD data set includes advanced AMD cases with GA and/or CNV in at least one eye and age at first diagnosis >= 50 years and intermediate AMD cases with pigmentary changes in the RPE or more than five macular drusen greater than 63µm and age at first diagnosis >= 50 years.

Nature Biotechnology: doi:10.1038/nbt.4108