nature portfolio

Corresponding author(s): Erna V. Ivarsdottir and Kari Stefansson

Last updated by author(s): Sep 17, 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	We used publicly available software together with software/methods developed at deCODE genetics as described in Methods. The publicly available software are: R, version 3.6.0 to analyze data and create plots; Graphtyper version 2, https://github.com/DecodeGenetics/graphtyper; Variant Effect Predictor (release 100), https://github.com/Ensembl/ensembl-vep; IMPUTE2 version 2.3.1, https://mathgen.stats.ox.ac.uk/impute/impute_v2.html; dbSNP version 140, http://www.ncbi.nlm.nih.gov/SNP/; CADD version 1.7, https://github.com/kircherlab/CADD-scripts; dbNSFP version 4.1c, https://sites.google.com/site/jpopgen/dbNSFP; BOLT-LMM version 2.1, http://www.hsph.harvard.edu/alkes-price/software/; STAR software package, version 2.7.10, https://github.com/alexdobin/STAR; Ensembl version 87, https://www.ensembl.org/index.html; LeafCutter version 1, https://github.com/davidaknowles/leafcutter; kallisto version 0.46, https://github.com/pachterlab/kallisto

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

Summary statistics for all burden associations for the 23 cancer phenotypes are available in Supplementary Data. The sequence variants from the Icelandic population whole-genome sequence data have been deposited at the European Variant Archive under accession PRJEB15197. The UK Biobank data were downloaded under application 56270. Other data supporting the findings of this study are available within the article or its Supplementary files.

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	In the association analysis, we adjusted for sex determined from genotype data.
Reporting on race, ethnicity, or other socially relevant groupings	In the association analysis, we used data on individuals of European descent. In the UK Biobank data, principal component analysis was used to identify white British/Irish individuals. To adjust for population stratification in the association analysis, we adjusted for the first 20 principal components in the UK Biobank datasets and the Norwegian dataset, and in the Icelandic dataset we adjusted for county at birth.
Population characteristics	The UK Biobank study is a large prospective cohort study of around 500,000 individuals, who enrolled in the study between 2006 and 2010 throughout the UK and were aged 40-69 years at recruitment and 54% were women. The Icelandic dataset is based on whole-genome sequence data from 63,460 Icelanders participating in various research projects at deCODE genetics. Variants identified through whole-genome sequencing were imputed into 173,025 chip-genotyped Icelanders as well as their untyped close relatives based on genealogy. The Norwegian cancer dataset was derived from a biobank of blood samples collected over a period of 25 years as a patient self-referral initiative for over representation of cancer in families, with both clinical and research intent. The Norwegian controls included samples from Norwegian Blood donors (Oslo University Hospital, Ullevål Hospital), the Norwegian Mother, Father and Child cohort study (MoBa), DemGene, TOP studies, and the Bergen adult ADHD cohort (see Methods for more detail).
Recruitment	For the Icelandic dataset individuals were recruited through various research projects at deCODE genetics. The participants are a large fraction of the adult Icelandic population. Participants in UK Biobank were recruited through assessment centres, designed specifically for this purpose. The Norwegian cancer dataset was derived from a biobank of blood samples collected over a period of 25 years as a patient self-referral initiative for overrepresentation of cancer in families, with both clinical and research intent. The Norwegian controls included samples from Norwegian Blood donors (Oslo University Hospital, Ullevål Hospital), the Norwegian Mother, Father and Child cohort study (MoBa), DemGene, TOP studies, and the Bergen adult ADHD cohort. (see Methods for more detail).
Ethics oversight	In the Icelandic dataset, all genotyped participants signed a written informed consent. The study was approved by the Icelandic Data Protection Authority and the National Bioethics Committee (VSN-18-115, VSN-18-029, VSN-17-204, VSN-20-004, VSN-17-137, VSN-17-138). The UK Biobank data was obtained under application number 56270. All participants provided written informed consent and The North West Research Ethics Committee reviewed and approved the UK Biobank protocol (ref. 06/MRE08/65). In the Norwegian cancer dataset, all participants provided separate written informed consent to the current research, and the study was approved by Regional Committee for Medical and Health Research Ethics (REC) South East C (#2015/2382). The Norwegian controls include samples from Norwegian Blood donors (Oslo University Hospital, Ullevål Hospital), the Norwegian Mother, Father and Child cohort study (MoBa), DemGene, TOP studies, and the Bergen adult ADHD cohort. Written informed consent was obtained, and the Regional Committee for Medical and Health Research Ethics (REC) South East (#2009/2485, #2011/1980, #2016/1226, #270326), Mid Norway (#2014/631) and West (#213/543) approved the studies. In keeping with the ethical approvals, sequence variants in genes listed in the ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing were not called in the Norwegian WGS data and therefore not available in the Norwegian datasets (cases and controls). Danish samples and data were obtained in collaboration with the Copenhagen Hospital Biobank (CHB) and the Danish Blood Donor Study (DBDS). CHB is a biobank for future research, which contains samples obtained during diagnostic procedures on hospitalized and outpatients in the Danish Capital Region hospitals. Data analysis within this study was performed under the CHB protocol "Developing the basis for personalized medicine in cancer – a genetic study on samples from the Danish cancer biobank (CHB-CC). CHB-CC was approved by the Capital Region Data Protection Agency (P-2022-393) and the National Scientific Committee on Health Research Ethics (NVK-2200966). The DBDS cohort is a nationwide study of ~160,000 blood donors. The Data Protection Agency (P-2019-99) and the National Committee on Health Research Ethics (NVK-1700407) approved genetic studies within the DBDS cohort.

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	Sample sizes are reported in the article and correspond to the data that was available.
Data exclusions	No available data was excluded.
Replication	The burden associations was performed on data from 3 different populations (Iceland, UK and Norway) and results were compared across populations. To claim gene associations, we required nominal significance in at least two of the datasets. Furthermore, a dataset including 377,448 Danes was used as a validation dataset. Aiming to replicate 5 gene-cancer associations, 3 were replicated in the Danish dataset.
Randomization	Not applicable. This was a genome wide burden association study, not a randomized trial.
Blinding	Not applicable. This was a genome wide burden association study, not a randomized trial. No blinding is therefore required.

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