

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 (<i>n</i>) 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 <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☐	☒ 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ 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 <i>d</i> , Pearson's <i>r</i>), 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	Publicly available analysis software and packages used are described as follows. Variants were annotated using Variant Effect Predictor (VEP) version 107 with the LOFTEE plugin. Unsupervised admixture analysis was performed using ADMIXTURE. FST calculations were performed using vcfTools (version 0.1.16). Principal component analysis was performed using PC-AiR as implemented in the GENESIS package (version 2.36.0). LD-pruning of inputs to PC-AiR, ADMIXTURE, and FST calculations was done with PLINK (v2.00a3.7). Runs of homozygosity calling was performed with the BCFtools (version 1.18) plugin 'roh'. Machine learning-based filtering of predicted splice variants was performed with Pangolin (https://github.com/tkzeng/Pangolin ; v1.0.1). Additional statistical analysis was performed using the Python statsmodels (version 0.13.0 and version 0.14.4) package. Burden PheWAS analysis was performed using Regenie v3.1.3 for batches sequenced at the Broad institute and Regenie v3.1.1 on the DNAnexus Apollo platform for batches sequenced at the Regeneron Genetics Center. Meta-analysis of association results was performed using METAL (2011-03-25 release).

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](#)

1000 Genomes Project (1KG) variant calls and sample manifest including ancestries were downloaded from https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/ and <https://www.internationalgenome.org/data-portal/sample> respectively. Human Genome Diversity Project (HGDP) variant calls were downloaded from https://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/. A recombination map used for runs of homozygosity calling was downloaded from https://storage.googleapis.com/broad-alkesgroup-public/Eagle/downloads/tables/genetic_map_hg38_withX.txt.gz. Mouse essential genes were downloaded from https://ftp.ebi.ac.uk/biostudies/fire/S-BSST/S-BSSTxxx293/S-BSST293/Files/SourceDataFile2_Embryo_windows.txt. A set of genes that are the targets of approved drugs was downloaded from https://github.com/ericminikel/genetic_support/blob/main/data/pp_launched_alltime.tsv. ClinGen disease-associated genes were downloaded from <https://search.clinicalgenome.org/kb/downloads> on June 14th, 2024. The Gene Set Enrichment Analysis (GSEA) Hallmark gene sets were downloaded from <https://www.gsea-msigdb.org/gsea/msigdb/human/genesets.jsp>. Human Protein Atlas (HPA) version 24.0 annotations were downloaded from <https://www.proteinatlas.org/download/proteinatlas.tsv.zip>. PGR. Supplementary Tables include summary statistics for autozygosity, summary statistics for biomarker burden tests, and a full exome-wide list of homozygous loss-of-function (homLoF) variants listed by chromosome, position, reference allele, alternative allele, and homozygous carrier count. Additional summary statistics for all biomarker burden tests, including non-significant associations, are uploaded to figshare. A subset (n=7,298) of the WES samples (accession phs000917.v1.p1) used in this study and all (n=6,678) of the WGS samples (accession phs001569.v1.p1) used in this study were previously made available to researchers in dbGaP2,93. All phenotype data and individual-level exomes and genomes, including those previously deposited in dbGaP, are stored at CNCND, with access protecting participant data. Academic researchers may apply for access to all individual-level data from this study by contacting the CNCND at info@cncndpk.com, with an initial review response time of 45 business days. Each request will require a detailed research plan and data privacy plan, including an agreement to maintain de-identification of PGR participants.

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 PGR, sex is ascertained using whole exome/whole genome sequencing. In the recall by genotype studies, self-reported gender is provided.

Reporting on race, ethnicity, or other socially relevant groupings

Any references to ethnicity (cohort demographics and breakdown of admixture plots by participant ethnicity) refer to self-reported ethnicity. For association analyses within PGR, genetic principal components are included as covariates to control for confounding.

Population characteristics

For PGR whole exome/whole genome data, median age = 50.3 years and 38.8% of participants are female.

Recruitment

Participant recruitment into PGR is directed towards assemblage of case-control studies that focus on 50 phenotypes. These phenotypes include stroke, myocardial infarction, angiographically confirmed coronary artery disease, heart failure, age-related macular degeneration, keratoconus, diabetic retinopathy, glaucoma, asthma, chronic obstructive pulmonary disease (COPD), non-alcoholic fatty liver disease (NAFLD), type-2 diabetes, chronic kidney disease, Alzheimer's disease, Parkinson's disease, dementia, progressive multiple sclerosis, autism, Huntington's disease, hematological cancers, breast cancer, ovarian cancer, cancers of head and neck, esophageal cancer, lung cancer, gastric cancer, colorectal cancer, melanoma, cancers of the urinary tract, cervical cancer, prostate cancer, rheumatoid arthritis, systemic lupus erythematosus (SLE), psoriatic arthritis, ankylosing spondylitis, osteoarthritis, scleroderma, juvenile arthritis, systemic sclerosis, inflammatory myositis, alopecia areata, acne rosacea, primary Sjogren's syndrome, sarcoidosis, idiopathic pulmonary cholangitis, idiopathic pulmonary fibrosis, vitiligo, longevity/healthy aging, and previously uncharacterized Mendelian disorders. For each of these phenotypes, screening is carried out at specialized clinical sites across Pakistan by trained research medical officers who review inclusion and exclusion criteria relevant to disease status and approach eligible participants for recruitment into PGR. In a similar manner, controls are frequency-matched to cases for each phenotype based on sex and age (in 5-year bands). Controls are being recruited in the following order of priority: (1) visitors of patients attending the out-patient department; (2) patients attending the out-patient department for routine non-phenotype related complaints, or (3) non-blood related visitors. Following informed consent, both cases and controls are enrolled. To minimize selection bias, enrollment was carried out across both urban and suburban hospitals that also serve surrounding rural populations. Both public and private hospital recruitments were also conducted to minimize biases in standards of care that could exist across such facilities.

Ethics oversight

The Institutional Review Board (IRB) at the Center for Non-Communicable Diseases (IRB: 00007048, IORG0005843, FWAS00014490) approved the study. This study has also been approved by the National Bioethics Committee for Research Pakistan (Reference number: NBC-756). All participants gave written informed consent.

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](https://www.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	The total sample size of the PGR dataset analyzed here is 173,303 participants. No a priori power calculations were used to reach this number of individuals, rather, this represents the total number of individuals who were sequenced from the PGR biobank's enrollment efforts for the past decade and whose sequences and phenotype data passed stringent QC filters. This number represents the largest set of currently sequenced individuals from South Asia.
Data exclusions	For the PGR analysis, individuals not reporting diabetes but having random blood glucose concentrations ≥ 150 mg/dl were excluded from the diabetes analyses. Individuals not reporting hypertension and having diastolic blood pressure ≥ 90 mmHg or systolic blood pressure ≥ 140 mmHg were excluded from hypertension analyses. The LDL-C analyses were subset to individuals who were not on cholesterol lowering drugs. Random glucose concentrations were analyzed for participants who were not on oral antidiabetic drugs. Diastolic and systolic blood pressure analyses were limited to individuals not on antihypertensives.
Replication	In vitro studies were done at least in triplicate, with all attempts successfully replicating. Whole exome/genome sequencing was conducted without replication, except in the in the case of recall by genotype studies in which cases the variant of interest carried by recalled probands was successfully replicated by Sanger sequencing. Sanger sequencing of recall by genotype participants was conducted once as were measurements of relevant clinical assays, such as clinical chemistry panels and cardiac ECHOs.
Randomization	No randomization was conducted for the participant recruitment for observational genetics studies as data were collected on all individuals willing to participate. Regression analyses of data collected from these studies included covariates such as age, age ² , gender, sequencing platform, and the first 10 genetic PCs to account for potential confounders. Randomization was not conducted for in vitro studies as this was not applicable.
Blinding	Recall-by-genotype study coordinators are blinded to genotypes.

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		Methods	
n/a	Involved in the study	n/a	Involved in the study
☒	☐ Antibodies	☒	☐ ChIP-seq
☐	☒ Eukaryotic cell lines	☒	☐ Flow cytometry
☒	☐ Palaeontology and archaeology	☒	☐ MRI-based neuroimaging
☒	☐ Animals and other organisms		
☒	☐ Clinical data		
☒	☐ Dual use research of concern		
☒	☐ Plants		

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	FreeStyle 293-F cells (ThermoFisher; https://www.thermofisher.com/order/catalog/product/R79007?SID=srch-srp-R79007) Expi293F cells (ThermoFisher; https://www.thermofisher.com/order/catalog/product/A14527?SID=srch-srp-A14527)
Authentication	Cell lines were directly obtained from ThermoFisher, which has generated information on the characterization and authentication of the cells. URLs are as follows: FreeStyle 293-F cells ThermoFisher; https://www.thermofisher.com/order/catalog/product/R79007?SID=srch-srp-R79007 Expi293F cells (ThermoFisher; https://www.thermofisher.com/order/catalog/product/A14527?SID=srch-srp-A14527)
Mycoplasma contamination	Cells were negative for mycoplasma contamination

Commonly misidentified lines
(See [ICLAC](#) register)

HEK293 cells

Plants

Seed stocks

n/a

Novel plant genotypes

n/a

Authentication

n/a