nature portfolio

Corresponding author(s): Damian Smedley

Last updated by author(s): Jan 8, 2025

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	STAR v2.7.6a; Salmon v1.4.0; no software was used for genetic data collection of the 100KGP participants (multi-sample VCF files were made available within the Genomics England Research Environment, GeL RE). Phenotypic and demographic data were retrieved using the desktop application and corresponding API Labkey within the GeL RE.
Data analysis	Exomiser version 12.1.0 with default settings and latest 2007* (July2020) databases; Perl version 5.30.2; R v4.2.1 (R packages: tidyverse v2.0.0; data.table v1.15.4; reshape2 v1.4.4; biomaRt v2.54.1; ggplot2 v3.5.1; ggrepel v0.9.5; httr v1.4.7; drawProteins v1.18.0; ensembldb v2.22.20; AnnotationHub v3.6.0; ontologyIndex v2.12; ontologyPlot v1.7); open-source geneBurdenRD pipeline available at https://github.com/whri-phenogenomics/geneBurdenRD; Mutplot (https://github.com/VivianBailey/Mutplot); CoExp (https://rytenlab.com/coexp).

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

Access to the genetic and phenotypic data for the 100KGP participants is open to all through the Genomics England Research Environment (GeL RE) and via application at https://www.genomicsengland.co.uk/research/academic/join-gecip to become a member of the Genomics England Research Network. Multi-sample VCF files and PED files used to run Exomiser can be found under /genomes/analysis/rare_disease in the GeL RE. PanelApp gene panels and evidence of disease associations were obtained using the PanelApp API available at https://panelapp.genomicsengland.co.uk/api/docs/ (March 2021). Data used for UNC13A gene in Figure 2: UniProt accession code: Q9UPW8; data used for RBFOX3 gene in Figure 2: UniProt accession code: A6NFN3 and dbGaP accession code: phs000424.v10.p2; data used for ARPC3 gene in Figure 3: UniProt accession code: O15145; data used for POMK gene in Figure 3: UniProt accession code: Q9H5K3; GEO accession code: GSE41616 and ENA accession code: E-MTAB-1498; data for GPR17 gene in Figure 3: UniProt accession code: Q13304.

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	Breakdown of phenotypic sex for the gene burden analysis cohort (n = 34,851) as available in the Genomics England Research Environment: 18,007 male and 16,844 female.
Reporting on race, ethnicity, or other socially relevant groupings	Breakdown of self-reported ethnicity for the gene burden analysis cohort (n = 34,851) as available in the Genomics England Research Environment is provided in Table S1 (63% 'White: British'/'White: Irish'/'White:Any other White background').
Population characteristics	Cohort of 34,851 rare disease probands plus their affected and unaffected family members (total participants: 72,690; recruited as described below). Mean age (and standard deviation) of probands: 30.3 years (22.8); ages ranged between 0 and 99 with a median of 25.
Recruitment	Patients with rare diseases and affected and unaffected family members were enrolled to the 100KGP through one of the 13 NHS Genomic Medicine Centres (GMCs) across England, Northern Ireland, Scotland and Wales. The recruiting clinicians assigned each proband to a specific disease (according to a hierarchical disease classification available within the project (226 'specific diseases' [level 4]; the 'specific diseases' are in turn grouped into less specific 91 'disease sub groups' [level 3], each of which corresponds to one of 20 broad 'disease groups' [level 2]).
Ethics oversight	The Health Research Authority (HRA) Research Ethics Committee (REC) East of England – Cambridge South (Ref 14/EE/111) gave ethical approval for the 100KGP. Consent was obtained from all participants to the 100KGP.

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	The 100KGP is unique in its scale (i.e. one of the main goals was to achieve the sequencing of 100,000 genomes from rare disease and cancer patients). To our knowledge, formal sample size and power calculations were not performed for the 100KGP. Sample size and statistical power of gene burden analyses in rare diseases are difficult to determine and depend on many factors, including selection criteria for variants to test, background variation rates, locus heterogeneity, mode of inheritance, prevalence and penetrance. Guo et al. (AJHG, 2016) have shown that achieving 80% power in many scenarios - particularly those with high locus heterogeneity requires tens, hundreds, or even thousands of cases. However, there are scenarios where only a few cases, or even tens of cases, may be sufficient. We used an arbitrary threshold of at least five cases for the specific disease tested and at least four probands, of which at least two were a case, per each of the four gene-based proband's genotype scenarios described in the main text and methods.
Data exclusions	An initial cohort of 74,061 genomes (35,548 single probands and larger families) from the rare disease pilot and main programme of the 100KGP (Data Release v.11) was available for analysis (March 2021). Quality control performed by Genomics England highlighted that 81 of the probands had been recruited and sequenced twice and these duplicates were removed from our cohort. In addition, the required data for our Exomiser-based gene burden analysis, e.g. recruited disease category and phenotypic terms, was not available for 16 families and these were also excluded from our cohort. The Exomiser analysis did not return any candidate variants for 29 families, generally for larger families

with multiple affected individuals where no rare, putative disease-causing variants remained after filtering, leading to an interim dataset size of 35,422 single probands and larger families. Further variant frequency filtering led us to discard data from 41 additional families as detailed in the main text. Potentially digenic probands with more than one recruited disease category were discarded from the analysis, leading to a dataset of 35,008 families (40,584 probands and affected family members and 32,434 unaffected family members). Finally, 157 probands with cryptic relatedness to other probands were removed, leading to a final input analysis dataset of 34,851 single probands and larger families (40,402 probands and affected family members and 32,288 unaffected family members)

Replication

Due to the unique features of the 100KGP, formal replication of the statistically significant signals from the whole-cohort gene burden analysis was not performed. The 0.5% FDR-adjusted disease-gene signals (306) were the object of a thorough in silico and clinical expert triage (165 previously known; from 141 potentially novel - not reported in OMIM and not green genes in PanelApp, to: 27 ClinGen moderate, 42 ClinGen Limited), leading to 5 highlighted disease-gene associations which are detailed in the main text (plus 5 interim confirmations via collaboration or independent study). Gene and variant look up were performed in available independent rare disease cohorts for renal disease and corneal abnormalities associated genes as described in the methods and Table S4.

Randomization

Since recruitment and sequencing were performed concurrently across rare disease categories, the order in which individuals were sequenced with respect to phenotype was random.

Blinding

This is an observational genetic association study, not a clinical trial. Since sequencing followed enrolment, both the participants and any investigators were unaware of the participant's genotype produced by the 100KGP at the time of the enrolment.

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

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.