nature portfolio

Corresponding author(s): Nicola Whiffin

Last updated by author(s): 24/06/24

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	Data were analysed using custom scripts within the Genomics England secure research environment. We additionally used data from gnomAD v4.0, All of Us (accessed via the publicly available data browser https://databrowser.researchallofus.org/ on 28 March 2023) and the UK Biobank (490,640 genome sequenced individuals release).
Data analysis	Analysis of the 100,000 genomes project and NHS GMS data was performed inside the Genomics England Research Environment. We are happy to share the location of all code to registered users. Code used for analyses outside of Genomics England is available at Github: https://github.com/Computational-Rare-Disease-Genomics-WHG/RNU4-2 and https://github.com/francois-lecoquierre/genomics_shortcuts/blob/main/find_RNU4-2_recurrent_variant.py. The following software and analysis tool were used: bedtools v2.31.0, Ensembl genome annotation v111, R v4.0.2, Illumina's DRAGEN pipeline v3.8.4, FRASER2 and OUTRIDER both run via the DROP pipeline v1.3.3, ggseqlogo R package, GENCODE v31, STAR aligner v.2.4.2a, DEXSeq v1.50.0, and ENCODE ATAC-seq pipeline 0.3.0.

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

Research on the de-identified patient data used in this publication from the Genomics England 100,000 Genomes Project and the NHS GMS dataset can be carried out in the Genomics England Research Environment subject to a collaborative agreement that adheres to patient led governance. All interested readers will be able to access the data in the same manner that the authors accessed the data. For more information about accessing the data, interested readers may contact research-network@genomicsengland.co.uk or access the relevant information on the Genomics England website: https://www.genomicsengland.co.uk/research. Genomic and phenotypic data from the GREGoR consortium (including the RGP cohort) and the UDN are available via dbGaP accession numbers phs003047 and phs001232.v5.p2, respectively, with at least annual data releases. Access is managed by a data access committee designated by dbGaP and is based on intended use of the requester and allowed use of the data submitter as defined by consent codes. The BrainVar data are available through the PsychENCODE Knowledge Portal: syn21557948 on Synapse.org (https://www.synapse.org/#!Synapse:syn4921369). Raw ATAC-seq and ChIP-seq data are available on dbGAP: accession phs002033.v1.p1.

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	We use the term sex to describe individuals.
Reporting on race, ethnicity, or other socially relevant groupings	We record reported ancestry of a subset of participants with detailed clinical information. These data are not used in any analyses.
Population characteristics	49 individuals with RNU4-2 variants had detailed phenotype characterisation. 21 of these individuals (42.9%) were female, with an average age of 10.
Recruitment	Participants were recruited to the Genomics England project based on clinical presentation. There could be biases from accessibility to recruitment centres. Other participants were recruited from other large studies that could have similar biases.
Ethics oversight	The 100,000 Genomes Project Protocol has ethical approval from the HRA Committee East of England Cambridge South (REC Ref 14/EE/1112). This study was registered with Genomics England under Research Registry Projects 354. Clinical data were collected from research participants after obtaining written informed consent from the parents or legal guardians, with the study approved by the local regulatory authority.

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	Primarily 8,841 individuals with undiagnosed neurodevelopmental disorders in GEL. Sample size was not predetermined but was all available data.
Data exclusions	No data were excluded from the analyses.
Replication	The initial analyses were performed in the Genomics England 100,000 Genomes Project dataset. Our findings were then replicated by identification of additional individuals with RNU4-2 variants and matching phenotypes in additional cohorts (including GREGoR, UDN, and NHS GMS).
Randomization	This is an observational study so randomisation is not relevant.
Blinding	This is an observational study so blinding is not relevant.

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	NA
Novel plant genotypes	NA
Authentication	NA