nature research

Corresponding author(s): Stephen B Montgomery

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Sequencing data was generated with illumina sequencing machines. No specialized software was developed for data collection.
Data analysis	cutadapt (v1.11); bcl2fastq (v2.15.0.4); STAR (v2.4.0j); Picard Tools MarkDuplicates (v1.131); RSEM (v1.2.21); BWA-mem (v0.7.12); GATK (v3); Vcfanno (v0.2.7); BCFtools (v1.8); BEDtools (v2.26.0-112-gd8c0fe4); Vcfanno (v0.2.7); ASEReadCounter (v3.8-0-ge9d806836). Custom code is available at https://github.com/lfresard/blood_rnaseq_rare_disease_paper

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

UDN data is accessible through the UDN Gateway and through dbGaP entry at phs001232.v1.p1. DGN RNA-seq data is available by application through the NIMH

Center for Collaborative Genomic Studies on Mental Disorders. Instructions for requesting access to data can be found at https://www.nimhgenetics.org/access_data_biomaterial.php, and inquiries should reference the "Depression Genes and Networks study (D. Levinson, PI)." The GTEx Analysis V7 release allele-specific expression data is available from dbGaP (dbGaP Accession phs000424.v7.p2). PIVUS RNA-seq data is accessible on the European Genome-Phenome Archive (EGAS00001003583). The Care4Rare data is available through Genomics4RD.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was based on the availability of undiagnosed cases for which we had blood data. This paper is showing that comparing those cases to large control cohorts can lead to diagnosis. We used controls sets for which we had blood transcriptome data and varied sample size to assess its impact on observed results. We investigated 94 cases across 16 diverse disease categories, using over 1500 controls. This is more samples than any publication on the subject to date.
Data exclusions	To control for potential residual technical artifacts impacting outlier expression, we removed samples for which 100 or more genes had normalized expression values of |Z-score| > 4 after SVA correction (54 samples). In general, samples were excluded if not meeting QC requirements specific to analyses. Sample size used are reported for each analysis.
Replication	Cases for which the diagnosis is described as validated were confirmed through independent experiments such as Sanger sequencing or RT-PCR and reflect cases for which the gene was previously known in the literature to be involved in similar phenotypes. All attempts for replication were successful.
Randomization	RNA-sequencing runs were performs based upon data availability. Family members were grouped on the same run. We corrected the data for batch effects (as described in the manuscript) and show the correlation of those batches to the sequencing run before and after correction.
Blinding	No blinding was performed in the context of our study. Phenotypes and family relationship were needed for appropriate diagnostic of the affected samples.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	94 affected individuals together with 49 unaffected family members were recruited for this study. There were 76 women and 67 men between 1 and 80 years old. Among them, there was 43 European descendant, 15 Asian-descendants, 19 Hispanics, 4 persons of multiple origins and 62 for which we don't have the information however, ancestry PCs have been inferred for all.
Recruitment	For the UDN, the recruitment is based on the following criteria from to the consortium (https://commonfund.nih.gov/diseases): - the applicant does not have a diagnosis that explains the objective findings. - the applicant or legal guardian agrees to the storage and sharing of information and biomaterials in an identified fashion amongst the UDN centers and in a de-identified fashion to research sites beyond the network. For the CGS, we requested de-identified specimens and/or subsequently generated data from participants (and, when applicable, their family members) that have been enrolled in the Stanford Clinical Genomics Program's research pilot pro

gram. In this research pilot program, participants were already consented to be involved in research and disease (Stanford IRB-approved protocol #23066) and were referred from healthcare providers at Stanford. These referring physicians have expertise in heritable disorders, and patients selected for the pilot program have conditions with suspected genetic etiology, based on the physician's evaluation of specific clinical or differential diagnosis as well as family history. We requested deidentified specimens and/or subsequently generated data for the purposes of development of RNA seq and other computational/experimental approaches aimed at improving the diagnostic yield of genomic sequencing (Stanford IRB-approved protocol # 38046). Appropriate Biobank/ laboratory staff responsible for processing the biobank specimens provided de-identified specimens/data to our research group.
The CHEO patient recruitment was performed under the Care4Rare Canada consortium protocol (http://care4rare.ca/).

We analyzed all samples received that passed our QC, independently of the disease type or the origin of the sample.