nature research

Corresponding author(s): Ernest Turro, F Lucy Raymond, Willem H Ouwehand

Last updated by author(s): May 1, 2020

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.

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	Illumina Isaac aligner (v.SAAC00776.15.01.27); Illumina Starling variant caller (v.2.1.4.2); Illumina Manta (v.0.23.1); Illumina Canvas (v.1.1.0.5); Illumina HiSeq Analysis Software (v.2.0); BWA (v.0.7); VILMAA (https://github.com/mh11/VILMAA); CellBase (v.4.5); VEP (Ensembl API 89); OpenClinica (https://www.openclinica.com/); CiviCRM (https://civicrm.org/).
Data analysis	R (v.3.1 to v.3.5); CrossMap (v.0.2.7); samtools (v.1.3 to v.1.9); verifyBamID (v.1.1.3); bedtools (v.2.26.0); picard (v.1 to v.2); Apache Spark (v.2.5); plink (v.1.9); PRIMUS (v.1.7); Prism (v.7); RedPop (v.1; https://gitlab.haem.cam.ac.uk/et341/redpop); Blueprint DCC ChIP-Seq Analysis Pipeline; Sapientia(TM) (v.1.0 to v.1.9); IGV (v.2, v.3); F-Seq (v.1.84); deepTools plotFingerprint (v.2.3.5); MACS2 (v.2.1.1); MatInspector (https://www.swmath.org/software/21812); Genalice (http://www.genalice.com/); VILMAA (https://github.com/mh11/VILMAA) ;CellBase (v.4.5); VEP (Ensembl API 89); BWA (v.0.7); Kaluza Analysis Software (v.2.1). R packages: BeviMed, biomaRt, Biostrings, cowplot, data.table, doParallel, dplyr, egg, foreach, gdsfmt, GENESIS, GenomicRanges, GGally, ggpubr, ggplot, ggplot2, ggrepel, ggthemes, grid, gridExtra, Gviz, GWASTools, hexbin, httr, jsonlite, magrittr, MASS, Matrix, methods, ontologyIndex, parallel, plotly, plot3D, plyr, png, RColorBrewer, reshape2, SNPRelate, rtracklayer, R.utils, scales, scatterplot3d, stringr, taRifx, tibble, tidyr, VGAM, viridis, xml2, xyloplot.

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. 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

Genotype and phenotype data from the 4,835 participants enrolled in the NIHR BioResource for the 100,000 Genomes Project–Rare Diseases Pilot can be accessed by seeking access via Genomics England Limited following the procedure outlined at: https://www.genomicsengland.co.uk/about-gecip/joining-research-community/. The genotype data for the 764 UK Biobank samples will be made available through a data release process which is being overseen by UK Biobank (https://www.ukbiobank.ac.uk/). The phenotype data from UK Biobank participants are available from UK Biobank using their access procedures.
Subject to ethical consent, the genotype data of the remaining 7,438 NIHR BioResource participants are available from the European Genome-phenome Archive (EGA) at the EMBL European Bioinformatics Institute under access procedures managed by EGA. The domain specific accessions are as follows: BPD: EGAD00001004519, CSVD: EGAD00001004513, EDS (EGAD00001005123), HCM: EGAD00001004514, ICP: EGAD00001004515, IRD: EGAD00001004520, LHON (EGAD00001005122), MPMT: EGAD00001004521, NDD: EGAD00001004522, NPD: EGAD00001004516, PAH: EGAD00001004525, PID: EGAD00001004523, PMG: EGAD00001004517, SMD: EGAD00001004524, SRNS: EGAD00001004518. Access to detailed phenotype data of the NIHR BioResource participants can be requested by contacting the NIHR BioResource Data Access Committee at dac@bioresource.nihr.ac.uk.

The ATAC-seq and H3K27ac ChIP-seq data to support the generation of the regulomes are available from GEO, EGA, ENCODE, or referenced to their publication. For transcription factor ChIP-seq: MK (GATA1, GATA2, TAL1, FLI1 - PMID: 21571218; MEIS1 - PMID: 25258084; CTCF - EGAD00001002362); EB (GATA1, KLF1, NFE2, TAL1 - PMID: 25521328; CTCF - EGAD00001002377); MONO (ENCSR000ATN); B (ENCSR000AUV). For H3K27ac ChIP-seq: MK (EGAD00001002362); EB (EGAD00001002377); MONO (ERR829362 (ERS257420), ERR829412 (ERS222466), ERR493634 (ERS214696), BLUEPRINT consortium); B (ERR1043004, ERR1043129, ERR928206, ERR769436, BLUEPRINT consortium); aCD4(PMID:28870212); rCD4 (PMID:28870212). For ATAC-seq: MK (EGAD00001001871); EB (SRR5489430 (GSM2594182)); MONO (EGAD00001006065); B (SRR2126769 (GSE71338)); aCD4 (GSE124867); rCD4 (GSE124867).

Reported alleles and their clinical interpretation have been deposited with ClinVar under the study names "NIHR_Bioresource_Rare_Diseases_13k", "NIHR_Bioresource_Rare_Diseases_Retinal_Dystrophy", "NIHR_Bioresource_Rare_Diseases_MYH9" and "NIHR_Bioresource_Rare_Diseases_PID".

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	13,187 samples. As our study piloted WGS of rare disease patients on a national scale, we had to accept recruitment of patients with a wide range of diseases and diverse aetiologies. Under certain realistic scenarios concerning penetrance and genetic architecture, only a small number of cases (< 10) with a shared genetic aetiology and several hundred non-cases are required to identify a genetic association. Previous WES studies with comparable sample sizes had been shown to be well powered, by replication and biological follow up.
Data exclusions	150 samples that failed quality control, as detailed in Supplementary Information. The data exclusion criteria were established over time as WGS data were generated, but were applied uniformly to the final dataset. The exclusion criteria were not informed by the phenotypes of the participants, to minimise the possibility of exclusion generating confounding.
Replication	Experimental replication was not attempted.
Randomization	For logistical reasons, recruitment and WGS were performed concurrently. Consequently, it was not possible to randomise the order of individuals to sequencing over time and, thus, over the three successive read length batches. However, we found that the variation in read length did not pose any difficulty in practice thanks to the stringent quality control imposed by our variant filters.
Blinding	Our study was not an intervention study and therefore blinding was not required. However, WGS quality control was performed without reference to the phenotypes.

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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

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

Antibodies

Antibodies used

For the functional analysis of the GATA1 enhancer/HDAC6 deletion: Rabbit HDAC6 (clone D2E5, cat.no 7558S, staining 1:50, blot 1:1000, Cell Signaling technology, Danvers, MA, USA), mouse anti-acetylated tubulin antibody (clone 6-11B-1, cat. no T7451, staining 1:50, blot 1:1000, Sigma, St Louis, MO, USA), mouse anti-alpha-tubulin (clone 236-10501, cat. no A11126, staining 1:250, blot 1:1000, Thermo Fisher Scientific, Waltham, MA, USA), rabbit VWF (cat. no A0082, staining 1:50, Dako Agilent Technologies, Leuven, BE), mouse CD63 and rat GATA1 N6 (cat. nos sc-5275 and sc-265 respectively, clones MX-49.129.5 and N6 respectively, staining 1:50 -mouse- and blot 1:1000 -rat-, Santa Cruz Biotechnology, Dallas, TX, USA), rabbit GATA1 (NF that was produced against recombinant N-terminal zinc finger, blot 10 µg/ml, PMID:19924028), rabbit GAPDH (clone 14C10, cat. no 2118S, blot 1:1000, Cell Signaling) and integrin beta3 (clone H96, cat. no sc-14009, blot 1:1000, Santa Cruz Biotechnology).
For MPL expression on platelets: APC-labelled IgG1 against CD42b (clone HIP1, cat. no 551061, staining 1:5, BD Pharmingen, number: 551061), PE-labelled IgG1 against CD110 (clone REA250, cat. no 130-101-648, staining 1:5, Miltenyi Biotec) and a PE-labelled isotype control (clone MOPC-21, cat.no 555749, BD Pharmingen); the staining was the same for all: add antibodies to 5ul of whole blood - make up to 12.5ul with PBS.

Validation

For the functional analysis of the GATA1 enhancer/HDAC6 deletion: The rabbit GATA1 antibody was produced against the N-terminal zinc finger of GATA1 (see PMID:19924028). This antibody was validated by immunoblot analysis against full length GATA1 expressed in HEK293 cells in parallel with the commercial GATA1 N6 antibody and both do generate bands of comparable sizes that are absent from lysates of non-transfected cells.
All the other antibodies used have been published by others as specified in the datasheets from the suppliers mentioned above.
For the MPL expression on platelet: all antibodies were used according to manufacturer's instruction.

Human research participants

Policy information about studies involving human research participants

Population characteristics

Age: birth to 95 years old. Gender: Male and Female. Patients with rare disorders across 15 disease domains, and relatives. Wide range of diagnosis and treatment categories, as detailed in Supplementary Information.

Recruitment

Patients were recruited from 83 hospitals in the UK and worldwide, as detailed in Supplementary Information. The patient populations at these hospitals differ with respect to genetic ancestry, which may have induced a degree of selection bias. This potential bias was mitigated by enrolling as widely as possible across different hospitals (see Extended Data Figure 1a) and by accounting for coarse ancestry in the association analyses.

Ethics oversight

East of England Cambridge South national research ethics committee (REC) reference 13/EE/0325 or separate local ethics, as detailed in Supplementary Information.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

No ChIP-seq data were generated, we used publicly available data.

Transcription factor ChIP-seq:
MK - GATA1, GATA2, TAL1, and FLI1: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24674;
MK - MEIS1: https://www.ebi.ac.uk/ega/datasets/EGAD00001000745;
MK - CTCF: https://www.ebi.ac.uk/ega/datasets/EGAD00001002362;
EB - GATA1, KLF1, NFE2 and TAL1: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59801;
EB - CTCF: https://www.ebi.ac.uk/ega/datasets/EGAD00001002377;
MONO - CTCF: https://www.encodeproject.org/experiments/ENCSR000ATN;
B - CTCF: https://www.encodeproject.org/experiments/ENCSR000AUV.

H3K27ac ChIP-seq:

MK: https://www.ebi.ac.uk/ega/datasets/EGAD00001002362;
EB: https://www.ebi.ac.uk/ega/datasets/EGAD00001002377;
MONO: BLUEPRINT Consortium website (http://www.blueprint-epigenome.eu) with accession IDs ERR829362 (ERS257420), ERR829412 (ERS222466), ERR493634 (ERS214696);
B: BLUEPRINT Consortium website (http://www.blueprint-epigenome.eu) with accession IDs ERR1043004, ERR1043129, ERR928206, ERR769436;
aCD4: https://www.ebi.ac.uk/ega/datasets/EGAD00001002686;
rCD4: https://www.ebi.ac.uk/ega/datasets/EGAD00001002686.

Files in database submission	n/a
Genome browser session (e.g. UCSC)	n/a

Methodology

Replicates	As in publications (MK - PMID:25258084 and PMID:21571218; EB - PMID:25521328; aCD4 and rCD4 - PMID:28870212) or the ENCODE website (MONO - https://www.encodeproject.org/experiments/ENCSR000ATN/; B - https://www.encodeproject.org/experiments/ENCSR000AUV).
Sequencing depth	As in publications (MK - PMID:25258084 and PMID:21571218; EB - PMID:25521328; aCD4 and rCD4 - PMID:28870212) or the ENCODE website (MONO - https://www.encodeproject.org/experiments/ENCSR000ATN/; B - https://www.encodeproject.org/experiments/ENCSR000AUV).
Antibodies	As in publications (MK - PMID:25258084 and PMID:21571218; EB - PMID:25521328; aCD4 and rCD4 - PMID:28870212) or the ENCODE website (MONO - https://www.encodeproject.org/experiments/ENCSR000ATN/; B - https://www.encodeproject.org/experiments/ENCSR000AUV).
Peak calling parameters	TFs and H3K27ac peaks were called with MACS2, significance threshold was set to qvalue<1e-5, narrow option was used.
Data quality	Low quality reads (-q 15), multi-mapped and duplicate reads were marked and removed with samtools and picard respectively. ChIP-seq efficiency was assessed with deepTools fingerPrint.
Software	BWA (v.0.7); picard (v.1 to v.2); deepTools plotFingerprint (v.2.3.5); MACS2 (v.2.1.1)

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	For MPL expression on platelets: the level of MPL protein on the platelet membrane was measured by flow cytometry (Beckman Coulter FC500) using the monoclonal antibodies: APC-labelled IgG1 against CD42b (clone HIP1, BD Pharmingen, cat. no 551061), PE-labelled IgG1 against CD110 (clone REA250, Miltenyi Biotec, cat. no 130-101-648) and a PE-labelled isotype control (clone MOPC-21, BD Pharmingen, cat. no 555749). In short, a sample of EDTA anticoagulated blood was incubated with anti-CD110 (or control) and anti-CD42b for 30 minutes.
Instrument	Beckman Coulter FC500
Software	Kaluza Analysis Software from Beckman (Version 2.1)
Cell population abundance	n/a
Gating strategy	Platelets were gated based on size using forward scatter and size scatter. The median fluorescent intensity of the CD110 PE-antibody was calculated for all CD42b positive platelets.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.