
Supplementary information

Whole-genome sequencing of patients with rare diseases in a national health system

In the format provided by the authors and unedited

Enrolment, research ethics and consent	3
Domain descriptions	3
Overview of the collection of participants	8
Clinical and laboratory phenotype data	9
DNA sequencing	9
Analysis of genotyping data	10
Overview of data processing pipeline	10
Alignment and SNV/indel calling	11
Sample provenance checks	11
Variant quality checks at the sample level	12
Coverage assessment	12
Computation of sex chromosome karyotypes	12
Ancestry and relatedness estimation	13
Variant normalisation and loading of variants into HBase	15
SNV and indel annotations	15
SNV and indel quality control	15
Concordance in duplicates and twins	16
Overview of SNVs and indels in the dataset	17
Integrative Variant Analysis (IVA) web application	18
Large deletion calling and quality control	18
Clinical reporting	21
Gene lists and transcript selection	21
Variant filtering	21
Web application for variant assessment	22
Variant interpretation in MDT meetings	22
Sensitivity of WGS to detect known pathogenic variants	23
Overview of diagnostic results	24
Upload of reported variants to public repositories	26
Reported variants informing changes in clinical management	26
Structural variants in patients with null protein phenotypes	27
Genetic association between rare variants and rare diseases	28
Statistical approach	28
Case/control groupings using phenotypic tags	29
Selection of variants affecting transcript sequences	29

Post-processing of BeviMed results	30
BeviMed results for genes	31
Observed false positive rate	31
Positive predictive value	32
Estimated false positive rate	33
Polygenic and rare variant associations with phenotypic extremes in the UK Biobank cohort	34
Selection of participants	34
Polygenic associations with tails	34
WGS GWAS	35
BeviMed analysis	36
Matchmaker Exchange	36
Regulome analysis	37
Ethics	37
Definition of cell type specific regulomes	37
Regulatory element detection using patterns of peaks (RedPop)	38
Transcription factor binding peak calling	39
Identifying possibly causal deletions of elements regulating diagnostic-grade genes	39
Deletion of GATA1 enhancer and HDAC6 open reading frame	40
Details on the functional analysis of the GATA1 enhancer/HDAC6 deletion	41
Homozygous deletion of CTCF binding sites in the first intron of LRBA	44
Identifying causal SNVs in elements regulating diagnostic-grade genes	44
An SNV in the promoter of MPL, combined with deletion of exon 10 of MPL	44
Details of the luciferase reporter assay	45
Alternative variant datasets for versions 37 and 38 of the human reference genome	46
Appendix	46
Appendix 1: Neuropathic pain disorders	46
Appendix 2: Extreme red cell traits in UK Biobank	47
References	51

Enrolment, research ethics and consent

As part of a pilot study for the 100,000 Genomes Project, the NIHR BioResource (NBR) enrolled a group of 9,742 study participants comprising patients with rare diseases and some of their close relatives. Participants were enrolled into one of 15 rare disease domains approved by the NBR's Sequencing and Informatics Committee. Each domain corresponded to a group of clinically related rare diseases and their recruitment criteria covered pathologies of various body systems. The domains adopted differing approaches to recruitment in order to address particular research objectives (see below, Domain descriptions). Enrolment took place between December 2012 and March 2017. In addition to the 15 rare disease domains recruited by the NBR, the study included three additional domains: a second rare diseases pilot study coordinated by Genomics England Ltd (GEL), a domain comprising a set of control samples and a domain comprising samples from UK Biobank (UKB)¹. Consequently, there were 18 domains in total.

The enrolment of NBR participants was coordinated by the University of Cambridge. Although most participants were recruited at NHS Hospitals in the UK, some recruitment occurred at overseas hospitals (**Supplementary Table 1 (Enrolment by hospital), Extended Data Figure 1a**). All 13,187 participants provided written informed consent, either under the East of England Cambridge South national research ethics committee (REC) reference 13/EE/0325 or under local institutional review board approval and governance. It was the responsibility of the respective principal investigators at the enrolling hospitals to obtain consent for overseas samples. The NBR retained de-identified versions of the consent forms from overseas participants and a material transfer agreement was applied to regulate the exchange of samples and data between the donor institutions and the University of Cambridge.

Domain descriptions

Supplementary Table 1 (Domain metrics) contains quantitative summaries of each domain, including the distributions of participants by sex, ethnicity and phenotype, while in this section we set out a brief description of each domain in prose.

BPD (Bleeding, Thrombotic and Platelet Disorders). Participants in this domain were enrolled at 31 hospitals. The probands had one of: a platelet disorder (with or without pathological bleeding); a pathological bleeding disorder not explained by a platelet function defect; a coagulation factor defect; or multiple thrombotic events at a young age. To enrich for inherited aetiologies, probands with a positive family history and/or early onset and/or presence of syndromic features were enrolled preferentially. Where participants had informative pedigrees, affected and unaffected individuals were recruited if possible. The clinical characteristics of a subset of the participants in this domain have been reported previously². With some exceptions, patients with a BPD of known molecular aetiology (e.g. Haemophilia A and B) were excluded from enrolment.

CNTRL (Technical Controls). A set of DNA control samples from consented participants: 27 healthy individuals from the NBR were enrolled under the REC-approved study *BLUEPRINT*

(REC 12/EE/0040), five healthy individuals enrolled under the REC-approved study *GENES & PLATELETS Healthy* (REC 10/H0304/65) and 18 patients (9 individuals with BMI > 40 and 9 individuals with partial lipodystrophy) collected under the REC-approved study *Inherited platelet conditions (BPD)* (REC 10/H0304/66). These samples were recruited primarily in order to test our laboratory procedures and our data transfer and analysis pipelines.

CSVD (Cerebral Small Vessel Disease). The participants in this domain were enrolled at 10 hospitals. Patients were eligible if they were suspected to have had familial CSVD from an early age (typically < 60 years), based on clinical features such as lacunar stroke or cognitive impairment. Evidence consistent with SVD from magnetic resonance imaging (MRI) of the brain, such as lacunar infarcts and white matter hyperintensities, was also required. Data on features typical of monogenic SVD were collected, including clinical history of migraines, encephalopathy and psychiatric disturbance. Clinical information and MRI results for all recruits were assessed by consultant neurologists and databases of previous stroke admissions were screened. Individuals with causal variants in *NOTCH3* were excluded from enrolment.

EDS (Ehlers-Danlos and Ehlers-Danlos-like Syndromes). The participants in this domain were enrolled at five hospitals. Patients with EDS subtypes of known molecular aetiology were excluded from enrolment. Patient participants met the Villefranche criteria for hypermobility EDS³. Probands had generalised joint hypermobility, skin hyperextensibility, connective tissue fragility and chronic pain. At the time of recruitment, patients had previously undergone ophthalmological and cardiac assessments to exclude Marfan syndrome and other non-EDS hereditary disorders of connective tissue. Where possible, close relatives were also recruited. The participants had a family history suggestive of autosomal dominant inheritance.

GEL (100,000 Genomes Project–Rare Diseases Pilot). The main purpose of this domain was to pilot the recruitment processes for the 100,000 Genomes Project and to assess the ability of National Health Service (NHS) hospitals to enrol rare disease patients and their close relatives, under the governance framework of the NHS. Enrolment was coordinated by GEL in partnership with the NBR. The recruitment strategy differed from that of the NBR domains by enrolling trios and pedigrees if possible. Patients exhibiting clinical symptoms of one of 161 rare inherited disorders and their close relatives (preferably both parents) were enrolled at eight NHS hospitals in England. Patients with known causal mutations were not eligible. Detailed clinical characteristics were collected on the participants using the human phenotyping ontology (HPO). All participants will be followed over their life course using electronic health data from primary care (general practitioners), secondary care (hospitals) and relevant registries. A full analysis of this domain will be published in due course by GEL. For the present study, data from this domain were used principally to increase the number of genomes from genetically independent individuals, thus increasing the power of genetic association analyses (see below). Only the binary affection status of these participants were available to researchers working on the present study.

HCM (Hypertrophic Cardiomyopathy). The participants in this domain had an unequivocal diagnosis of HCM made at one of four UK NHS centres providing specialist HCM care. HCM is characterised by primary left ventricular hypertrophy and has an estimated population prevalence of approximately 1 in 500. It is the leading cause of arrhythmia and sudden death in athletes and young adults aged under 35 years. All participants were diagnosed below the age of 70, or were affected relatives. To qualify for enrolment, patients had to have undergone clinical genetic testing on an HCM gene panel, with no pathogenic mutation identified.

ICP (Intrahepatic Cholestasis of Pregnancy). Severe ICP is defined as gestational pruritus in association with maternal serum bile acids $\geq 40 \mu\text{mol/L}$. It is associated with adverse pregnancy outcomes, including spontaneous preterm labour, fetal asphyxia and intrauterine death. ICP typically presents in the third trimester. Affected women and their relatives are at increased risk of biliary disease in later life, e.g. drug-induced cholestasis, gallstones, and hepatic fibrosis. The ICP domain comprises women with disease onset before 32 completed weeks of gestation, which affects approximately 1 in 3,000 pregnant women in the UK. Individual cases with severe, early onset ICP were recruited from 14 UK consultant-led antenatal NHS clinics and from three international units in Argentina, Australia and Sweden. Women were excluded from the study if they had other known causes of hepatic dysfunction such as haemolysis, elevated liver enzymes and low platelets (HELLP) syndrome, preeclampsia, acute fatty liver of pregnancy, acute viral hepatitis, confirmed primary biliary cholangitis or any cause of biliary obstruction on ultrasound. No genetic pre-screening for causal variants in known genes was applied before enrolment.

IRD (Inherited Retinal Disorders). IRD describes a phenotypically heterogeneous group of conditions consequent upon dysfunction and/or degeneration of the neural retina or retinal pigment epithelium, resulting in visual impairment. It is the most common cause of severe visual impairment. The participants were enrolled at five NHS hospitals, most at the Moorfields Eye Hospital, London. Most individuals had undergone some previous genetic testing using routine diagnostic approaches. The analysis of the whole genome sequencing (WGS) data of a fraction of the participants in this domain has been reported previously⁴.

LHON (Leber Hereditary Optic Neuropathy). LHON causes subacute sequential bilateral visual loss which is usually irreversible. UK patients with an LHON diagnosis previously known to carry one of four causal mitochondrial DNA (mtDNA) mutations (m.11778G>A, m.3460A>G, m.14484T>C, m.14495A>G) were enrolled in this domain. These mutations are found in ~1 in 300 of the UK population, but the prevalence of blindness due to LHON mtDNA mutations is approximately 1 in 40,000. Environmental factors undoubtedly influence the clinical penetrance, but segregation analyses implicate an interaction between a nuclear modifier locus and the mtDNA mutation. The LHON domain was included to test the hypothesis that nuclear mutations are required for the blindness phenotype in those carrying mtDNA risk alleles.

MPMT (Multiple Primary Malignant Tumours). Participants in this domain were mostly (> 95%) recruited in the UK, through the NHS regional clinical genetics services. In each kindred there was a clinical suspicion of a cancer predisposition syndrome, but routine genetic assessment/testing had not identified a genetic cause at the time of enrolment. Most (95%) of cases analysed had developed MPMT (defined as more than one primary malignant tumour by age 60 or more than two primary malignant tumours by 70) but a minority had developed a single primary malignant tumour and had a first-degree relative with MPMT. Tumours in the same tissue type and organ were considered separate primary tumours if, in the case of paired organs, they occurred bilaterally or if the medical record clearly denoted them as separate. The International Agency for Research on Cancer criteria for defining separate primary tumours were also used⁵. For 95% of pedigrees, only the DNA of a single family member was sequenced.

NDD (Neurological and Developmental Disorders). The participants in this domain were adults and children with a previously undiagnosed neurological condition. Referrals were from clinical specialists in paediatrics, adult neurology or clinical genetics. The cases enrolled had early or severe epilepsy, encephalopathy, dystonia, spasticity, intellectual disability or metabolic disorders. In many cases, pre-screening for known genetic aetiologies was performed but this was not systematic. The majority were singleton cases but some trios of proband, mother and father were included for paediatric conditions.

NPD (Neuropathic Pain Disorders). The NPD domain enrolled individuals with extreme neuropathic pain phenotypes (both sensory loss and gain) at secondary care clinics located in six NHS hospitals. In order to be recruited, cases had to be over 18 years of age with a history of life-style altering sensory disorder (either pain or loss of sensation) for greater than three months. Patients with a known underlying genetic cause of chronic pain (e.g. Fabry's disease [ORPHA:324] and SCN9A erythromelalgia [ORPHA:1956]), intellectual disability and/or autistic spectrum disorder so severe that they could neither give consent or engage in additional pain phenotyping were excluded. The Neuropathic Pain Special Interest Group of the International Association for the Study of Pain grading for neuropathic pain was used to assess the participants⁶. Further details are available in **Appendix 1: Neuropathic pain disorders** and **Supplementary Table 1 (NPD Criteria – Diagnostic criteria, NPD Criteria – Outcome measures)**.

PAH (Pulmonary Arterial Hypertension). In PAH, adverse remodelling of the pulmonary vasculature causes narrowing and obliteration of the capillary arteries in the lung, resulting in elevated resting mean pulmonary artery pressure and right heart dysfunction. A mean pulmonary artery pressure of 25 mm Hg or above, with a pulmonary capillary wedge pressure of less than 15 mm Hg is indicative of PAH. The diagnosis of idiopathic PAH was based on the exclusion of other associated forms of PAH. Participants with idiopathic PAH were recruited at 10 NHS UK National Pulmonary Hypertension hospitals and four international hospitals. All recruits had a clinical diagnosis of idiopathic PAH, heritable PAH, drug-associated PAH, or pulmonary veno-occlusive disease/pulmonary capillary haemangiomatosis (PVOD/PCH)

established by their expert centre. Analyses of this domain have been recently reported^{7,8,9}. No systematic genetic pre-screening for causal variants in previously established causal genes was applied before enrolment.

PID (Primary Immune Disorders). Cases were recruited by specialists in clinical immunology (either trained in paediatrics or internal medicine) at 21 NHS hospitals in the UK and a small number of hospitals in the Netherlands. The eligibility criteria included the following: clinical diagnosis of common variable immunodeficiency disorder (CVID) according to internationally established criteria¹⁰, extreme autoimmunity, or recurrent (and/or unusual) severe infections suggestive of defective innate or cell-mediated immunity. Patients with known secondary immunodeficiency (e.g. due to cancer or HIV infection) were excluded. Genetic screening for known causes of PID prior to enrolment was encouraged but not applied systematically. Within a broad range of phenotypes, CVID is the most common disease category, comprising 46% of the cases. Participants in this domain consisted predominantly of singleton cases, but the DNA samples from additional affected and/or unaffected relatives of some of the patients were also sequenced (see the accompanying article by Thaventhiran et al. in this issue of *Nature*).

PMG (Primary Membranoproliferative Glomerulonephritis). PMG refers to kidney disease in which a biopsy shows increased glomerular mesangial matrix and cellularity with thickening of the capillary walls and there is an absence of an underlying infectious, neoplastic or autoimmune disorder. Participants in this domain were enrolled at 10 NHS paediatric renal units in the UK (64 patients) and 18 NHS adult renal centres (120 participants, of whom 21 had paediatric onset of disease). IC-PMG refers to PMG where there is deposition of complement C3 and immunoglobulins in the glomeruli, and 'C3 glomerulopathy' (C3G) refers to PMG where there is deposition of complement C3 without significant deposition of immunoglobulins. The C3G category is further subdivided by electron microscopic appearance into 'C3 glomerulonephritis' (C3GN) and 'dense deposit disease' (DDD). PMG has an estimated incidence of 3–5 in 1 million and has a poor renal prognosis¹¹. Where available, kidney biopsies were reviewed centrally to confirm classification into IC-PMG, C3GN or DDD. No genetic pre-screening for causal variants in known genes was applied before enrolment.

SMD (Stem Cell and Myeloid Disorders). Cases were enrolled into this domain if they presented with an inherited bone marrow failure of unknown molecular aetiology in childhood or an inherited cytopenia (including erythroid lineage) for which non-genetic and autoimmune aetiologies could be excluded. In addition, patients presenting with a myeloproliferative phenotype and a positive family history of a related haematological disorder were eligible for enrolment. Recruitment into the domain was performed by paediatric and adult haematologists from NIHR/National Cancer Research Network Primary Treatment Centres. In addition, cases and their close relatives were enrolled at centres in Canada, Denmark, Egypt, Israel, Norway, Sri Lanka, South Africa, Sweden, Turkey and the USA. Most probands had undergone previous genetic analysis using targeted sequencing with phenotype-specific gene panel tests by which no pathogenic variants were identified.

SRNS (Steroid Resistant Nephrotic Syndrome). The prevalence of SRNS is estimated at 2–4 in 100,000 people. There are two major clinical subsets: primary SRNS, defined as no response to high dose steroid treatment after four weeks in children or after four months in adults, and secondary SRNS, where initial steroid sensitivity is lost and treatment resistance evolves as a secondary event either rapidly or over time. Participants in this domain were enrolled through the UK NephroS study at tertiary NHS paediatric nephrology centres and adult renal units across the UK. Patients with either sporadic or familial SRNS were eligible for enrolment. Cases with SRNS secondary to systemic disease or obesity were excluded from enrolment. The majority of patients had a histological renal biopsy diagnosis that allowed a second tier of stratification based on histology as well as drug response. The majority (70%) of the cases had undergone previous genetic testing by whole-exome sequencing and tested negative for causal variants in diagnostic-grade SRNS genes.

UKB (UK Biobank – Extreme Red Cell Trait Score). The UK Biobank is a biomedical cohort of approximately half a million participants recruited in the UK between 2006 and 2010¹. The participants, 54% of whom are women, were aged between 37 and 73 years at their date of recruitment. Each participant underwent a baseline assessment at one of 21 centres distributed across Great Britain, during which EDTA anticoagulated blood was collected for full blood count (FBC) analysis¹². A subset of UK Biobank participants likely to carry rare alleles perturbing erythropoiesis or red cell survival/clearance mechanisms were selected for WGS by considering a univariate quantitative phenotype derived from the FBC-measured mean corpuscular volume (MCV) and red blood cell count (RBC#) values. 383 and 381 participants from the extreme left and extreme right tails of the quantitative distribution passed quality control (QC) after being successfully sequenced. Further details of the approach used for selecting these participants are given in **Appendix 2: Extreme red cell traits in UK Biobank**.

Overview of the collection of participants

Sex (derived from X and Y chromosome karyotypes), ethnicity and relatedness for all samples were inferred computationally from WGS data (as detailed below). The age of the participant on the date his/her DNA sample was collected was available for 97.9% of the 13,037 samples that underwent WGS and passed QC (see below). These variables highlight some of the differences in the disease presentation and the recruitment approaches between the domains (**Extended Data Figure 1b, Supplementary Table 1 (Domain metrics)**). The ICP domain comprised only female participants because the phenotype is related to pregnancy. The proportion of families for which WGS data were only generated for the proband was > 50% for all domains except GEL, CSVD and SMD, which largely comprised parent-child trios and mother-child duos. The overall ethnic breakdown for the study collection closely matched that of the 2011 UK census¹³, suggesting equality of access (**Supplementary Table 1 (Domain metrics)**). DNA sample collection generally occurred in childhood for recruits with a paediatric onset (NDD, SMD), at childbearing age for ICP recruits, after the age of 50 for the UKB recruits (due to the enrolment criteria of UK Biobank), and throughout adulthood for the other disorders (**Extended Data Figure 1b**).

Clinical and laboratory phenotype data

Staff at the hospitals responsible for enrolment were provided with the eligibility criteria for the domains for which they were recruiting patients. Different levels of pre-screening to exclude previously known disorders were used across the 15 domains. ICP, PAH, and PMG had virtually no genetic pre-screening. BPD, CSVD, EDS, IRD, MPMT, NDD, NPD, PID, SMD and SRNS had some pre-screening based on clinical presentation and the results of non-DNA based laboratory and imaging tests. HCM and LHON had extensive genetic screening in the relevant diagnostic-grade genes. The clinical and laboratory phenotype data were captured through case report forms (CRF) by paper questionnaires or by online CRF data capture applications and deposited in the NBR study database managed by the University of Cambridge. A web-based phenotype capture system allowed the reporting clinical staff to record Human Phenotype Ontology (HPO) terms at the enrolling hospitals^{2,14,15}. Additional phenotypic information recorded on paper questionnaires were transformed into HPO terms by the NBR study coordination office. Free text entries were transformed into HPO terms where feasible. An overview of the HPO data obtained for the 15 domains is depicted in **Extended Data Figure 1c**.

DNA sequencing

DNA from 58.8% of enrolled participants was extracted at the receiving hospital or research laboratory and sent to Cambridge for plating. The DNA for the remaining samples was extracted from EDTA-treated whole bloods at a central DNA extraction and QC laboratory in Cambridge. Samples were tested for adequate DNA concentration (Picogreen), DNA degradation (gel electrophoresis) and purity (OD 260/280 QC(Trinean)) before selection for WGS. DNA samples were prepared at a minimum concentration of 30 ng/μl in 110 μl, visually inspected for degradation and had to have an OD 260/280 between 1.75 and 2.04. They were then prepared in batches of 96 and shipped on dry ice to the sequencing provider (Illumina Inc, Great Chesterford, UK).

Illumina measured the DNA concentration of each sample they received. A sample was excluded from WGS if its corresponding measurement was < 30 ng/μl. They also genotyped each sample using an Illumina Infinium Human Core Exome microarray. A sample was excluded from WGS if its genotyping call rate was < 0.99. The extent of cross contamination by human DNA was estimated by computing the fraction of homozygous alternative allele calls exhibiting at least one high quality read discordant with the alternative allele. Samples with a contamination measure > 0.05 were excluded. Finally, samples for which there was a mismatch between genetically inferred and clinician reported sex that could not be resolved by further investigation, or for which consent had been withdrawn, were also excluded from WGS. In total, 59 DNA samples sent to Illumina were excluded from WGS. Comparisons of the microarray and WGS genotyping data did not reveal discrepancies which could not be explained by genotyping errors.

To prepare each DNA sample for sequencing, 0.5 μg of DNA was fragmented using Covaris LE220 (Covaris Inc., Woburn, MA, USA) to obtain fragments with a mean length of 450 base pairs (bp). The fragmented DNA samples were processed using the Illumina TruSeq DNA

PCR-Free Sample Preparation kit (Illumina Inc., San Diego, CA, USA) on the Hamilton Microlab Star (Hamilton Robotics, Inc, Reno, NV, USA). The concentrations of the final libraries were measured using the Roche LightCycler 480 II (Roche Diagnostics Corporation, Indianapolis, IN, USA) with KAPA Library Quantification Kits (Kapa Biosystems, Inc, Wilmington, MA, USA).

Over the course of the project (February 2014 to June 2017) samples were sequenced at three different read lengths: 100bp (377 samples), 125bp (3,154 samples) and 150bp (9,656 samples). Samples sequenced with 100bp and 125bp reads used three and two lanes of an Illumina HiSeq 2500 instrument, respectively. Samples sequenced with 150bp reads used a single lane of a HiSeq X instrument. The changes in read length over the course of the project were imposed by the sequencing service provider. At least 95% of the autosomal genome had to be covered at 15X and a maximum of 5% of insert sizes had to be less than twice the read length. Following sample and data QC at Illumina, 13,187 sets of WGS data files were received at the University of Cambridge High Performance Computing Service (HPC) for further QC.

Analysis of genotyping data

Overview of data processing pipeline

WGS data for 13,187 samples were sent to Cambridge University by Illumina. There, they underwent a series of processing steps (**Extended Data Figure 2**), which are summarised in this paragraph and described in detail in the following sections. We estimated the karyotypes of the sex chromosomes and computed pairwise kinship coefficients between samples. This information was used to identify repeat participant submissions and DNA sample swaps. Additionally, four further QC procedures were applied to ensure the single nucleotide variant (SNV) and short (≤ 50 bp) insertion/deletion (indel) calls for each sample were of a high standard. In total, the sequence data of 150 samples (1.1%) were removed, leaving a dataset of 13,037 participants for downstream analysis. The 13,037 individuals were assigned to one of the following ethnicity clusters: European, African, South Asian, East Asian or Other. Pairwise kinship coefficients adjusted for population stratification were then estimated. These were used to identify networks of closely related participants and to identify a maximal set of 10,259 unrelated participants (MSUP).

The variants called in the 13,037 individuals were left-aligned and normalised with bcftools¹⁶, and loaded into an HBase database¹⁷. The sex chromosome karyotypes, the ethnicities and the relatedness estimates were used, along with enrolment information, to annotate the samples and variants. Participants were annotated with:

- affection status (affected/unaffected),
- family membership indexed by proband identifier,
- an indicator for membership of the MSUP,
- ethnicity and
- sex chromosome karyotype.

Variants were annotated with:

- consequence predictions,
- Human Gene Mutation Database (HGMD)¹⁸ information where available and

- population-specific allele frequencies.

Alignment and SNV/indel calling

Reads were aligned with the Illumina Isaac aligner version SAAC00776.15.01.27¹⁹ to the reference human genome build GRCh37²⁰. This produced BAM files²¹ the mean sizes of which were approximately 95GB for 100bp samples, 60GB for 125bp samples and 65GB for 150bp samples. SNVs and indels were called from all BAM files using the Illumina Starling software version 2.1.4.2²². Variants were called independently across samples and sample genotypes were stored in both a standard VCF and a genome VCF (gVCF) file²³. The gVCF file format stores information on read coverage and alignment quality in homozygous reference regions as well as variant positions, which allows a PASS filter to be applied to all positions in the genome for each sample.

Sample provenance checks

As a means of identifying duplicates, approximate pairwise kinship coefficients were computed between all samples as data arrived at the Cambridge HPC. A set of 8,872 single nucleotide polymorphisms (SNPs) was chosen at the start of the project for this purpose at random from the common SNPs typed by Roche on their targeted microarray platforms. The kinship coefficients were computed using PLINK²⁴ after each data delivery. More precise relatedness estimation was performed at a later stage (see below). Any kinship coefficient > 0.99 triggered an investigation to determine whether there had been an accidental duplicate enrolment, a laboratory duplication or an enrolment of monozygotic twins. When a pair of samples resolved to the same individual, only the higher quality WGS data were retained, where quality was determined firstly by read length and secondly by autosomal callability (defined below). 124 duplicates and 1 triplicate were found. The 22 pairs of monozygotic twins we identified were retained in the resource. Eight unresolved duplicates were excluded. Most duplicate samples were explained by submission of the same patient by different consulting clinicians recruiting into the same domain. However, one duplicate spanned two domains because a patient with a phenotype consistent with the recruitment criteria for two domains was enrolled by two clinicians with different specialties.

Participant-declared and genetically-determined sex were compared for concordance following each WGS data delivery. For each sample, we counted the number of heterozygous, homozygous and hemizygous QC-passing SNVs in the non-pseudoautosomal region of the X chromosome. We defined the H-ratio as the number of heterozygous variants divided by the number of homozygous and hemizygous variants. We made initial sex calls based on the H-ratio as follows: individuals with H-ratio < 0.02 were declared putatively male and the others were declared putatively female (**Extended Data Figure 3e**). Comprehensive sex chromosome karyotyping was performed at a later stage (see below). Based on the initial sex calls, two samples with discordance between declared and estimated sex were excluded from further analysis.

Variant quality checks at the sample level

We performed the following variant quality checks:

- We computed the proportion of positions with a known base in the reference autosomes that passed quality filters in the gVCFs. This proportion is called the autosomal callability. Samples for which fewer than 95% of bases in the reference genome passed were removed (**Extended Data Figure 3a**).
- We identified a set of common SNPs by downloading and merging the genome and exome data from gnomAD²⁵ and computing the minor allele frequencies (MAFs) in each ethnically homogeneous cluster of study participants (except for the South Asian cluster, which is absent from the gnomAD genome set; see below for details of clustering by ethnicity). The set of SNVs that had a MAF > 5% in each population was identified. We then determined the proportion of the positions of these SNVs in each gVCF that failed the variant caller's internal QC. Samples for which this proportion exceeded 0.55% were excluded (**Extended Data Figure 3b**).
- Transition-transversion ratios (Ts/Tv) are commonly used to determine SNV calling accuracy²⁶. We calculated Ts/Tv ratios for all samples and observed that the distribution of these ratios varied depending on the read length; samples with a Ts/Tv ratio more than 3.5 times the read length specific interquartile range plus the read length specific median were excluded (**Extended Data Figure 3c**).
- We estimated the degree to which a DNA sample was contaminated by any other DNA sample using verifyBamID²⁷. Samples with an estimate of contamination (FREEMIX) exceeding 3% were excluded (**Extended Data Figure 3d**).

In total, 14 samples were excluded due to these four genotyping QCs.

Coverage assessment

The mean (over samples) of the mean autosomal coverage by unique reads was 41.4, 37.9 and 35.3, and the mean of the 10th percentile of autosomal coverage was 31.0, 25.7 and 26.2 for the 100bp, 125bp and 150bp read length batches, respectively (**Figure 1b**). The mean coverage was greater than 30X in all samples and 90% of the reference genome was covered at least to 19X in all samples. In addition, all samples were covered to at least 15X in at least 95% of the reference autosomes. The average read duplication rate was 1.5% for read pair inserts in the HiSeq 2500 (100bp and 125bp read lengths). However the rate was 18% for the HiSeq X-generated 150bp data, which meant that extra reads were required to make up a similar level of coverage.

Computation of sex chromosome karyotypes

Once the WGS dataset was complete, we estimated the sex chromosome karyotypes using the BAM and VCF files. For each sample and chromosome, we normalised the number of aligned reads by dividing it by the length of the chromosome in the reference genome, excluding positions with an unknown base. The X/Auto and Y/Auto ratios were defined as the normalised read counts on X and Y divided by the median of the normalised read counts on the autosomes (Auto). The median was used because it is robust to large copy number alterations. For the putative males and putative females (defined based on the H-ratio, as described above), the

mean and standard deviation of X/Auto and Y/Auto were computed (**Extended Data Figure 3f**). Let us define these values as $\text{mean}_z(\text{X/Auto})$, $\text{sd}_z(\text{X/Auto})$, $\text{mean}_z(\text{Y/Auto})$ and $\text{sd}_z(\text{Y/Auto})$, where $z \in \{m, f\}$ denotes whether the summaries were conditioned on the putative males or the putative females. We defined the following five gates to classify individuals according to their specific X/Auto and Y/Auto values, denoted x and y , respectively:

- XY gate (shown in blue in **Extended Data Figure 3f**):
 - $\text{mean}_m(\text{X/Auto}) - 10 \text{sd}_m(\text{X/Auto}) < x < \text{mean}_m(\text{X/Auto}) + 10 \text{sd}_m(\text{X/Auto})$
 - $\text{mean}_m(\text{Y/Auto}) - 10 \text{sd}_m(\text{Y/Auto}) < y < \text{mean}_m(\text{Y/Auto}) + 10 \text{sd}_m(\text{Y/Auto})$
- XYY gate (shown in green in **Extended Data Figure 3f**):
 - $\text{mean}_m(\text{X/Auto}) - 10 \text{sd}_m(\text{X/Auto}) < x < \text{mean}_m(\text{X/Auto}) + 10 \text{sd}_m(\text{X/Auto})$
 - $y > \text{mean}_m(\text{Y/Auto}) + 10 \text{sd}_m(\text{Y/Auto})$
- XX gate (shown in red in **Extended Data Figure 3f**):
 - $\text{mean}_f(\text{X/Auto}) - 10 \text{sd}_f(\text{X/Auto}) < x < \text{mean}_f(\text{X/Auto}) + 10 \text{sd}_f(\text{X/Auto})$
 - $\text{mean}_f(\text{Y/Auto}) - 10 \text{sd}_f(\text{Y/Auto}) < y < \text{mean}_f(\text{Y/Auto}) + 10 \text{sd}_f(\text{Y/Auto})$
- XXY gate (shown in purple in **Extended Data Figure 3f**):
 - $\text{mean}_f(\text{X/Auto}) - 5 \text{sd}_f(\text{X/Auto}) < x < \text{mean}_f(\text{X/Auto}) + 5 \text{sd}_f(\text{X/Auto})$
 - $\text{mean}_m(\text{Y/Auto}) - 5 \text{sd}_m(\text{Y/Auto}) < y < \text{mean}_m(\text{Y/Auto}) + 5 \text{sd}_m(\text{Y/Auto})$
- XXX gate (shown in orange in **Extended Data Figure 3f**):
 - $1.5 \text{mean}_f(\text{X/Auto}) - 5 \text{sd}_f(\text{X/Auto}) < x < 1.5 \text{mean}_f(\text{X/Auto}) + 5 \text{sd}_f(\text{X/Auto})$
 - $\text{mean}_f(\text{Y/Auto}) - 5 \text{sd}_f(\text{Y/Auto}) < y < \text{mean}_f(\text{Y/Auto}) + 5 \text{sd}_f(\text{Y/Auto})$

One participant had an H-ratio of 0.01 and was thus declared initially to be a male. However, the participant was located within the XX gate. The two X chromosomes were almost identical in sequence to one another, explaining the low H-ratio, and further investigation showed that this was due to consanguinity. This and other anomalies are labelled in **Extended Data Figure 3g**. Where permitted by the study ethics, we approached the referring clinician to confirm the non-standard sex chromosome karyotypes ($n = 5$). The observed frequencies of non-standard sex chromosome karyotypes, which were present in 13 of the 13,037 individuals (5 XXY, 4 XYY, 2 XXX, 2 XO), were comparable with those reported by a previous study²⁸. The genotypes for the samples with non-standard sex chromosome karyotypes (except for XXX) were recalled to take into account the correct ploidy in the sex chromosomes (**Extended Data Figure 2**).

Ancestry and relatedness estimation

We estimated the degree of relatedness between participants and clustered the participants by their nuclear ancestry into the ethnicity categories European, African, East Asian, South Asian and Other, as follows.

We selected a set of SNPs for ancestry and relatedness analyses by starting from the 292,878 autosomal SNPs typed by three widely used Illumina genotyping arrays (HumanCoreExome-12v1.1, HumanCoreExome-24v1.0 and HumanOmni2.5-8v1.1). Presence of a SNP on all three arrays indicates that it can be reliably measured by microarrays, a good prior that it will be reliably called by WGS. To ensure this, we removed SNPs with a missing WGS genotype in at least one individual or with a minimum overall pass rate (OPR; defined

below) < 0.99 . We also removed SNPs at genomic positions in which more than two distinct alleles had been observed in the 1000 Genomes Phase 3 dataset²⁹ or the NBR dataset to ensure that all genotypes could be coded unambiguously as the count of the number of copies of the unique alternative allele carried by the individual. We then removed SNPs with a MAF $< 30\%$ in our dataset. Finally, we pruned the SNPs using PLINK v1.9²⁴ so that no pair of variants had an $r^2 \geq 0.2$. The final set contained 32,875 SNPs.

We implemented a pipeline that partitions a given set of participants into sets of related individuals and derives an MSUP. The inputs to the pipeline are a matrix of genotypes for the 32,875 SNPs encoded in a VCF file and a list of individuals with a high prior of being unrelated (the *reference set*). The pipeline uses PC-AiR³⁰ to perform a principal component analysis (PCA) on the standardised genotypes of the reference set and to project the standardised genotypes of the individuals outside the reference set onto the fitted principal components (PCs). The PC-AiR output object is then passed to the PC-Relate function to compute a kinship matrix that accounts for the population structure captured by the leading 20 PCs. Finally, this kinship matrix is passed to the PRIMUS³¹ function to obtain clusters of related participants and to derive a MSUP on the basis that pairs of individuals with a kinship coefficient > 0.09 are related.

We used the pipeline firstly to obtain a set of unrelated non-admixed individuals within the 1000 Genomes Phase 3 data. To achieve this, we first excluded the 347 participants labelled as “Admixed American”. We then defined a reference set of 2,110 putatively unrelated individuals within the remaining 2,157 participants using the `snpgdsIBDKING` function from the SNPRelate R package^{32,33}. We used the pipeline secondly to compute the relatedness amongst the 13,037 individuals in our dataset. In this second run, we merged the genotypes of the 13,037 individuals in our dataset with those of the 2,110 individuals from the 1000 Genomes data. We specified the 2,110 individuals as the reference set. The pipeline identified an MSUP comprising 10,259 individuals (**Supplementary Table 1 (Domain metrics)**) and a set of 9,244 pedigree networks ranging from singletons to a network of size 15 were computed (**Figure 1f**).

We partitioned the final set of unrelated individuals into the equivalence classes: non-Finnish Europeans, Finns, Africans, South Asians and East Asians using their 1000 Genomes population code annotations. We modelled the score vectors of the leading five PCs by a multivariate normal distribution within each class of the partition and estimated the mean vector and covariance matrix for each population. Subsequently, we projected the genotypes from each of the 13,037 participants onto the vector space spanned by the leading five components of the 1000 Genomes PCAs (**Extended Data Figure 3h**). We computed the likelihood of the projected data of each individual under the five multivariate normal models estimated from the 1000 Genomes scores and assigned each individual to the population corresponding to the model that yielded the highest likelihood, provided it was greater than 10^7 . If no such population existed, the individual was classified as “Other” (**Extended Data Figure 3i**). The resource is 82% European. However, approximately 7% of the participants are of South Asian ancestry, a

population for which limited whole-genome sequencing (WGS) reference data are presently publicly available through resources such as gnomAD (**Extended Data Figure 3j**).

Variant normalisation and loading of variants into HBase

Variants in the gVCF files were processed using the bcftools norm command with the -cs option, which left aligns and normalises indels and sets/fixes incorrect or missing reference alleles³⁴. The variants were then transformed from gVCF format into Google *proto* format³⁵ and stored as binary *proto* objects spanning 1000bp of contiguous sequence in the reference genome using OpenCGA³⁶. During the transformation, the anchoring reference base of insertions and deletions was removed. The *proto* objects of each sample were then copied from the respective *proto* file to the HBase³⁷ *Archive* table using the 'PUT' operation of the HBase application programming interface. The resulting *Archive* table held the *proto* objects by sample in columns and grouped by position in rows. Header information from the original gVCF files was stored for each sample. Any multiple variant calls overlapping the same position within an individual were resolved by only retaining the variant with the highest genotype quality. The list of participants corresponding to each of the genotype calls and whether these calls had a PASS filter value were recorded for different genomic positions in an HBase *Allele count* table using the 'APPEND' operation. As homozygous reference and PASS values were the most frequent values of genotype and filter data, we stored data as differences from these. The variants were then transferred from the HBase *Allele count* table to the HBase *Analysis* table.

SNV and indel annotations

All variants called across the 13,187 samples were annotated with deleteriousness scores and conservation scores and with various summary statistics, which were computed for different subsets of samples. These summary statistics included allele count, allele number, genotype count, minor allele frequency, HWE statistics (obtained using HTSJDK³⁸), call rate, pass rate and overall pass rate (defined below). Subsequently, using the RESTful annotation service provided by CellBase³⁹, the variants stored in the HBase *Analysis* table were annotated with consequence terms, HGVS descriptions, deleteriousness scores such as the combined annotation dependent depletion (CADD) score⁴⁰, SIFT⁴¹ and PolyPhen-2⁴², and conservation scores such as genomic evolutionary rate profiling (GERP)⁴³, PhastCons⁴⁴ and PhyloP⁴⁵. The genotype and variant annotation data were then exported to AVRO files. These files were modified to add additional annotations from external reference datasets, including 1000 Genomes Project Phase 3⁴⁶, UK10K project (version 2016-05)⁴⁷, the ExAC / gnomAD project (version r2.0.2)⁴⁸, Trans-Omics for Precision Medicine Program (TOPMed, freeze5)⁴⁹ and the HGMD (version PRO 2018.1)¹⁸ using Apache Spark⁵⁰. Since the TOPMed data are only available with reference to GRCh38, these data were first mapped back onto GRCh37 using CrossMap⁵¹.

SNV and indel quality control

The HBase database described above holds records of whether a genotype could be called reliably by Starling from the aligned read data at each position of the reference genome in each sample. The call rate for a position, for a given set of the samples, is the proportion of the

samples for which the variant caller made a call. The pass rate for a position, for a given set of samples, is the proportion of the samples with a genotype called, for which the call passed QC. We defined the overall pass rate (OPR) as the call rate multiplied by the pass rate. Only female and male samples were used to calculate the OPRs for variants on the X and the Y chromosome, respectively. In subsequent QC steps, we considered the minimum OPR, taken over the three read length batches.

We generated histograms of P -values from tests of departure from the null hypothesis of HWE for common SNPs and indels (MAF > 5%) amongst 8,511 unrelated Europeans over different intervals of minimum OPR ((0.97,0.98], (0.98,0.99], (0.99,1]) and MAF ((5%,6%], (6%,7%], (7%,8%], (8%,9%], (9%,10%], (10%,20%], (20%, 30%], (30%, 40%], (40%, 50%]). Variants with minimum OPR > 0.99 were found to have genotype proportions generally consistent with HWE in all MAF bins for SNVs, short insertions and short deletions (**Extended Data Figure 4a–c**). For variants with minimum OPR in (0.98,0.99], we observed approximately double the number of P -values < 0.05 relative to that expected under the null, uniformly across MAF bins, suggesting that deviation from HWE was present in a small proportion of variants in this interval of minimum OPR. We thus only released variants with a minimum OPR > 0.99 through the variant browser (described below). However, for rare variant analyses, where maintaining high sensitivity is important and potentially pathogenic variants are routinely confirmed by Sanger sequencing, we used a threshold of minimum OPR > 0.98.

Concordance in duplicates and twins

We selected pairs of samples from the 124 duplicated samples and from the 22 monozygous twins to evaluate the reproducibility of genotypes called by WGS. From all resolved duplicates and twins, we chose pairs where both samples were sequenced with the same read length and the participant was one of the 13,037 participants included in the final data release (one pair sequenced with 100bp reads, 17 pairs sequenced with 125bp reads and 88 pairs sequenced with 150bp reads).

For each sample, we selected all the autosomal SNVs and small insertions and deletions with a minimum OPR > 0.99 and for which the genotype was either heterozygote for the reference and principal alternative alleles, or homozygote for the principal alternative allele. **Extended Data Figure 4d** illustrates all possible combinations of genotypes in a pair of samples. We use d_{ij} (in the cases that $i \neq j$) to denote the number of variants in which sample 1 has the i th genotype and sample 2 has the j th genotype, where i and j take values in {0,1,2,3} representing, respectively, the homozygote reference genotype, the heterozygote genotype for the reference and principal alternative alleles, the homozygote genotype for the principal alternative allele, and any other genotype. We use c_{ii} , where i takes values in {1,2} to denote the number of variants in which both samples have genotype i . We use c and d to denote the total numbers of variants in the cells corresponding to concordant ($c_{11}+c_{22}$) and discordant calls ($d_{01}+d_{02}+d_{10}+d_{12}+d_{13}+d_{20}+d_{21}+d_{23}+d_{31}+d_{32}$) respectively. The mutual non-reference concordance can be defined as $(c / (c + d)) * 100\%$. The distribution of this statistic is shown in **Extended Data Figure 4e**.

The probability of observing a heterozygous variant in sample 2 given it is heterozygous in sample 1 is $c_{11} / (d_{10} + c_{11} + d_{12} + d_{13})$. The probability of observing a homozygous alternative variant in sample 2 given it is homozygous alternative in sample 1 is $c_{22} / (d_{20} + d_{21} + c_{22} + d_{23})$ (**Extended Data Figure 4f–g**). Symmetrical formulae apply when the sample ordering is reversed. Hence, for each pair of duplicates/twins there are two probabilities of each type, the smaller of which is shown in red and the greater of which is shown in blue in **Extended Data Figure 4f–g**.

Overview of SNVs and indels in the dataset

We called 172,005,610 short variants with a minimum OPR > 0.99 in the MSUP, of which 157,411,228 (91.5%) were SNVs and 14,594,382 (8.5%) were indels. We annotated all short variants with Sequence Ontology (SO) terms based on their predicted consequences with respect to transcripts in Ensembl 75. A single primary consequence was selected for each variant following a series of filtering steps. Firstly, only consequences associated with transcripts annotated with a biotype in the most biologically significant biotype group (ranked by protein coding, pseudogene, long non-coding and short non-coding)⁵² were chosen. Secondly, only consequences associated with transcripts with the best source of curation were retained (ranked by HGNC, VEGA, Ensembl, UniProt and miRBase/RFAM). Thirdly, only consequences with the strongest Ensembl-determined impact were retained⁵³ (ranked by HIGH, MODERATE, LOW and MODIFIER). Fourthly, only consequences with the greatest severity were retained (ranked by the order provided by Ensembl⁵³). Fifthly, only consequences associated with transcripts annotated as being canonical by Ensembl were retained, if any existed. Finally, the transcript with the longest coding/cDNA sequence length was chosen.

On the basis of primary consequences, 61.7% of the short variants in the MSUP were labelled genic (SO:0001564, gene_variant), while the remainder were labelled intergenic (SO:0001628, intergenic_variant), regulatory (SO:0001566, regulatory_variant) or other. 80.0% of the genic variants were intronic (SO:0001627, intron_variant) and only 4.4% were exonic (SO:0001791, exon_variant) or affected a splice site (SO:0001568, splicing_variant). The ratio of coding (SO:0001968, coding_transcript_variant) to UTR (SO:0001622, UTR_variant) variants was 45:55 for SNVs and 21:79 for indels. 33.7% of the 1,853,212 coding SNVs were synonymous (SO:0001819, synonymous_variant) and 64.2% were missense (SO:0001583, missense_variant). 38,886 SNVs (2.1% of coding SNVs) led to a start loss (SO:0002012, start_lost), stop loss (SO:0001578, stop_lost) or stop gain (SO:0001587, stop_gained). The majority (62.1%) of the 67,746 coding indels introduced a frameshift (SO:0000865, frameshift), while 34.5% introduced an inframe insertion or deletion (SO:0001817, inframe) and only 2,287 indels (3.4% of coding indels) led to a start loss (SO:0002012, start_lost), stop loss (SO:0001578, stop_lost) or stop gain (SO:0001587, stop_gained) (**Extended Data Figure 5**).

48.6% and 59.2% of the SNVs and indels in the MSUP, respectively, were absent from the following databases: 1000 Genomes, UK10K, gnomAD, TOPMed (lifted back to GRCh37) and HGMD. 54.8% of the variants were observed in only one member of the MSUP. 82.6% of these

variants were novel. 36.7% of the variants observed in two members of the MSUP were novel. A small proportion (1.90% of SNVs and 9.01% of indels) of the variants observed in more than two members of the MSUP were novel (**Figure 1d**). Thus, common variants are well represented in genetic databases but the majority of genetic variants are very rare and are mostly absent from the available databases.

In coding regions, indels that introduce frameshifts are, on average, selected against due to their deleterious effects on reproductive fitness⁵⁴, and our analysis corroborated this notion (**Extended Data Figure 3k**). In contrast, in non-coding regions of the genome, indel lengths were biased towards a multiple of two bases (**Extended Data Figure 3l**). We have made a similar observation in the gnomAD dataset (data not shown) to ensure this is not a technical artefact specific to our bioinformatic pipeline. The pattern is specific to repetitive regions of the genome (**Extended Data Figure 3m**), suggesting it can in part be attributed to tandem duplications.

Integrative Variant Analysis (IVA) web application

The SNVs and indels have been loaded into the IVA web application⁵⁵. The data can be browsed freely online at ref. 56.

Large deletion calling and quality control

We used both Manta version 0.23.1⁵⁷ and Canvas version 1.1.0.5⁵⁸ to call deletions larger than 50bp, independently across samples. Manta calls deletions on the basis of split read alignments and larger-than-expected insert sizes of read pair alignments. Canvas calls deletions on the basis of sustained reductions in coverage in a region of the reference genome. Manta is optimised for calling deletions of 50bp–10Kb while Canvas is optimised for calling deletions > 10Kb.

For each sample, deletion calls were merged across the two methods if the reciprocal overlap between them (the smaller of the two values obtained by dividing the length of the intersection of the two deletion calls by the length of each of the two calls) was > 0.7. Merged deletions were annotated with the Manta breakpoints as they are more precise than Canvas breakpoints. After merging, each deletion call in each sample was assigned to one of the three QC categories “Fail”, “Possible” or “Confident”. These categories are defined according to the rules below.

“Fail”

Any of the following criteria are met:

- the length of the call is > 50Mb;
- the genotype is heterozygous and the call is on the non-PAR of chromosome X in a male;
- the call is on the mitochondrial or the Y chromosome;
- the call was made by Canvas only and CANVAS_QUAL < 10;
- the call was made by Manta only and MANTA_FILTER is not “PASS”.

“Confident”

The call does not meet the “Fail” criteria and none of the following additional criteria are met:

- > 70% of the call overlaps a flagged region (see below);
- the call was made by Canvas only and SNPS_DENSITY_HET_PASS > 0.5;
- the call was made by Canvas only and the number of Canvas calls in the sample is greater than the mean plus five times the standard deviation across all samples (as such extreme values are probably due to generalised and artefactual non-uniformities in coverage leading to inflation in the number of false positive calls);
- the call was made by Canvas only, the sample had more than 50 Canvas deletions and the proportion of the genome which was deleted was greater than the 99.7% percentile across all samples;
- the call was made by Manta only and MANTA_QUAL < 20th batch-specific percentile;
- the call was made by Manta only and MANTA_GQ < 20th batch-specific percentile;
- the call was made by Manta only and MANTA_IMPRECISE = “TRUE”;
- the call was made by Manta only and length > 1Mb;
- the call was made by Manta only and DUKE0_START = “TRUE” or DUKE0_STOP = “TRUE”;
- the call was made by Manta only and SNPS_N_HET_PASS > 0;
- the call was made by Manta only and the genotype was heterozygous and READS_MEAN_MAPQ < 45.

The parameters in capitals above were generated by the Manta and Canvas software, except for DUKE0_START, DUKE0_STOP, SNPS_N_HET_PASS, SNPS_DENSITY_HET_PASS and READS_MEAN_MAPQ, which were generated using custom code. DUKE0_START and DUKE0_STOP evaluate to “TRUE” if a deletion call's start or end positions lie within 10bp of the breakpoints of a Duke non-unique region (see definition below); SNPS_N_HET_PASS is the number of heterozygous SNPs with a minimum OPR > 0.99 within the boundaries of a deletion call; SNPS_DENSITY_HET_PASS is the ratio of SNPS_N_HET_PASS to the length of the deletion in Kb; READS_MEAN_MAPQ is the mean mapping quality of the read alignments within a deletion's boundaries. The flagged regions in the reference genome were the following:

- low mappability regions: DAC Blacklisted Regions (Encode, Accession: wgEncodeEH001432) and Human Mappability Blacklist (Encode, Accession: wgEncodeEH000322),
- centromeres⁵⁹,
- telomeres⁵⁹,
- IG loci (NCBI, Gene IDs: 50802, 3492, 3535),
- HLA loci (NCBI, Accession: NC_000006),
- segmental duplications^{60,61},
- Duke non-unique regions (Encode, Accession: wgEncodeDukeMapabilityUniqueness35bp).

“Possible”

The deletion call does not meet either the “Fail” or the “Confident” criteria. By visual inspection of the reads around 100 deletion calls from the Possible and Confident deletion sets using

Integrative Genomics Viewer (IGV)⁶², we estimated the false call rates to be 24% and 10% respectively. To control the rate of false calls, whilst retaining calls in the Possible class which have a high chance of being real due to overlap with calls in the Confident set, any deletion in the Possible set which did not have a reciprocal overlap of at least 0.7 with any of the Confident deletions was removed.

In order to select a set of rare deletions for use in the downstream association analyses, the deletion calls required annotating with internal cohort allele frequencies. The deletion calls from all samples were partitioned into groups corresponding to distinct latent deletions. The partitioning was performed as follows. First, we used the 'hclust' function in R to construct a dendrogram of all deletion calls. One minus the reciprocal overlap was treated as the distance between each pair of calls, and the complete linkage option was used (i.e. the distance between two clusters corresponds to the maximum distance between two elements chosen from the two clusters). The dendrogram was cut at the position corresponding to a reciprocal overlap of 0.7 to create the partition, thus ensuring that all pairs of deletion calls within the same cluster had a reciprocal overlap of at least 0.7. To boost the number of real deletions retained, the partitioning algorithm was applied to the Confident deletions alone, and Possible deletions were subsequently assigned to the same groups as the Confident deletions with which they shared the highest reciprocal overlap. Applying this procedure resulted in 201,986 distinct latent deletions.

Some sets of overlapping Manta deletion calls, which were probably generated by the same underlying deletion, had very different variances of left-hand and right-hand breakpoints across samples. This caused the partitioning procedure to fail to assign such calls to the same latent deletion. To address this, we employed the following remedy. First, we constructed a dendrogram using hierarchical clustering as described above on the 201,986 latent deletions. The start and end points of each latent deletion were estimated by taking the mean of the start points and the mean of the end points respectively across the Manta calls that mapped to the latent deletion, or if there were no Manta calls mapped, by taking the means across the Canvas calls. For each node in the dendrogram for which the mean reciprocal overlap of its descending latent deletions exceeded 0.5, we evaluated whether the spatial densities of start points or end points were significantly greater than expected by chance. We did this by performing a binomial test under the null hypothesis that the probability of observing the start or end point of a deletion at any given base was the same in the interval as in the entire chromosome. If all *P*-values (one for each internal node in the dendrogram) were greater than a threshold, the procedure was halted, otherwise the deletions descending from the node with the lowest *P*-value were merged, and the procedure was then repeated. We selected a threshold of 0.05 which gave high specificity whilst removing a large number of highly overlapping clusters which appeared on manual inspection to represent the same latent deletions. Application of this merging step reduced the number of deletions by 24,436, resulting in 177,550 distinct deletions being retained for final analysis.

Retained deletion calls were annotated with cohort allele frequencies based on the number of other deletions in their groups. We computed *P*-values under the null hypothesis of HWE for all common deletions. The distribution of *P*-values was highly skewed (82% of deletions in unrelated Europeans with a cohort allele frequency of at least 0.05 had a *P*-value < 0.01) and we therefore concluded that the calling of common deletion genotypes was unreliable but retained rare deletions for downstream analysis (see below).

Clinical reporting

Diagnostic multi-disciplinary team (MDT) meetings reported pathogenic or likely pathogenic variants to referring clinicians for cases in all 18 domains except EDS, CNTRL, GEL, HCM, LHON and UKB. The reasons that no reports were made for these domains were as follows: the MDTs for EDS did not identify any likely pathogenic or pathogenic variants in any of the 26 EDS samples; LHON cases were all known to carry pathogenic mitochondrial variants prior to enrollment, as the scientific motivation for recruitment was to identify risk-modifying nuclear DNA variants; HCM enrolled patients who had already been found negative in a screening of HCM diagnostic-grade genes; CNTRL and UKB participants did not include rare disease patients; clinical reporting for GEL participants via the NHS Genomic Laboratory Hubs is ongoing and the results will be reported separately in due course.

Gene lists and transcript selection

For each of the 15 rare disease domains (i.e. all domains except CNTRL, GEL and UKB) a list of genes was created by domain-specific experts. Only genes with an established causal role (more than three independent families reported with causal variants in a gene, or two families reported plus additional functional studies and/or a mouse model) were included in these lists. The lists for the different domains overlapped (**Figure 2b**). The 2,478 gene/domain pairs, encompassing 2,073 distinct genes across all domains, were manually curated and annotated with the relevant RefSeq and/or Ensembl transcript identifiers to support variant reporting. Transcripts were selected based on, by order of priority: community input, presence in the Locus Reference Genomic (LRG) resource⁶³ and designation as canonical in Ensembl (**Supplementary Table 2 (Diagnostic-grade genes)**). The genes in the 15 lists were submitted to LRG, whose role is to determine the relevant transcripts for clinical reporting. The gene lists were also loaded into the Sapientia™ web application (Congenica Inc, Cambridge, UK) for use by the MDTs reporting causal variants to referring clinicians. During the course of the project, gene lists were updated and reversioned when new evidence caused genes to meet the threshold for inclusion. A final re-analysis of all samples was performed using gene lists that were fixed at the time of the data freeze. We refer to the genes in the final lists as ‘diagnostic-grade genes’.

Variant filtering

Variants (SNVs, indels) were shortlisted if (i) their MAF in control populations²⁵ was < 1/1,000 for putative novel causal variants and < 25/1,000 for variants listed as disease-causing in HGMD, (ii) their predicted impact according to the Variant Effect Predictor (VEP)⁶⁴ was “HIGH” or “MODERATE” or if the consequences with respect to the designated transcript included one of

“splice_region_variant” or “non_coding_transcript_exon_variant” if the variant was in a non-coding gene, and (iii) the variant affected a gene with a known aetiological role in the patient’s disease. Variants with more than three alternative alleles or an internal MAF $\geq 10\%$ were discarded to guard against errors in repetitive regions and remove potential systematic artefacts, respectively. The above filtering criteria were applied universally to all domains, except for ICP, which adopted a higher MAF threshold of 3% for both novel and previously reported variants. The higher threshold accounted for causal variants being present in the male and non-child bearing female population. This strategy reduced the number of variants for review by the MDT from about 4 million per individual to fewer than 10, while retaining known regulatory or moderately common pathogenic variants.

Web application for variant assessment

For each case with prioritised variants, the variant calls, HPO-coded phenotype and the relevant metadata (unique study numbers; referring clinician and hospital; self-declared and genetically inferred sex, ancestry, relatedness to other participants, and consanguinity level) were transferred to Congenica for visualisation in the Sapientia™ web application during MDT meetings. Sapientia™ displays variant information such as predicted effect, location in the protein, MAFs in reference cohorts (e.g. ExAC, UK10K, and ESP^{25,47,65}), conserved regions, splice sites and links to external resources (e.g. HGMD¹⁸, ClinVar⁶⁶, OMIM⁶⁷ and PubMed⁶⁸), as well as patient data such as phenotype information in the form of HPO terms¹⁴. Sapientia™ also allows MDTs to annotate each variant with its likely level of pathogenicity, its contribution to the disease phenotype (full or partial, the latter meaning the variant was either one of two in compound heterozygosity causal of a recessive disorder or was the cause of only a subset of the main phenotypic abnormalities), and to generate tailored research reports for the referring clinicians.

Variant interpretation in MDT meetings

MDTs brought together experts from different hospitals across the UK and overseas, and typically consisted of an experienced clinician with domain-specific knowledge, a scientist with experience in clinical genomics, a clinical bioinformatician and a member of the reporting team. Assignment of the level of pathogenicity to variants followed the American College of Medical Genetics (ACMG) guidelines⁶⁹ and variants were marked in Sapientia™ as pathogenic, likely pathogenic or of uncertain significance (VUS). Only pathogenic and likely pathogenic variants were systematically reported and VUSs were reported at the MDT’s discretion. As per the REC-approved study protocol, secondary findings (e.g. pathogenic variants in *BRCA1* in patients not presenting with a relevant cancer phenotype) were not reported. Deletions involving domain-relevant diagnostic-grade genes were reviewed by the MDT and reported if deemed likely pathogenic or pathogenic. Complex rearrangements were analysed only in IRD and NDD participants⁷⁰. The reports recommended clinical grade confirmation in an accredited laboratory before feedback of the results to the patient or modification of clinical management. Although Sanger sequencing was not performed systematically to confirm the > 1000 reports issued, no failures to confirm the variants were reported back by the referring clinicians. In addition, 177 of

the variants reported in cases in the IRD domain were confirmed by Sanger sequencing (SNVs, indels or structural variants) or microarray (structural variants only)⁴.

Sensitivity of WGS to detect known pathogenic variants

The mean depth of coverage by WGS was 36.07X across all samples. In order to assess the relative sensitivity of WGS and WES, we compared the coverage of WGS and WES at the sites of 116,449 variants in the autosomal diagnostic-grade genes catalogued as “DM”/“DM?” in HGMD Pro 2018.1, or as “likely pathogenic”/“pathogenic” in ClinVar version 2018-07-29, excluding variants with at least one “benign” interpretation. We also assessed the sensitivity of WES at the 938 distinct SNVs and indels included in our pertinent findings reports. Only indels of length ≤ 50 bp were considered. We computed coverage statistics for 1,000 representative male samples (i.e. with an XY sex chromosomal karyotype) from our WGS 150bp read data and 1,000, 1,000, 248 and 553 representative male samples respectively from each of: UK Biobank¹, INTERVAL⁷¹, the Columbia University exome sequencing study for chronic kidney disease IDT xGen Exome Research Panel V1 samples and the Columbia University exome sequencing study for chronic kidney disease Roche samples⁷². Specifically, we computed mean coverage and the 1st and 99th percentiles of coverage over each set of samples. Because UK Biobank and INTERVAL sequencing reads had been aligned to human genome GRCh38, we realigned them to human genome GRCh37 with BWA mem⁷³ and marked duplicate reads using Picard MarkDuplicates⁷⁴. We computed coverage using samtools depth after discarding secondary alignments, reads not passing quality controls, duplicates and supplementary alignments. Only bases with a base call quality ≥ 10 and reads with a mapping quality ≥ 20 were counted. The coverage of a deletion was defined to be the mean coverage of the deleted bases and the coverage of an insertion was defined to be the mean coverage of the two bases flanking the insertion breakpoint. A nominal cutoff on the mean coverage of 20X was used to distinguish between detectable and undetectable variants to diagnostic standard.

The results of our analysis showed that the number of undetectable SNVs and indels ranged from 119 to 318 and 47 to 78 respectively, across the WGS dataset and the four WES datasets. 56–72% and 42–53% of these variants, respectively, map to a small number of genes (*HBA1*, *HBA2*, *VWF* for SNVs; *HBA1*, *HBA2* for indels). 1,997–9,063 of the causal variants (1.71–7.78%), were undetected in WES datasets but only 470 (0.4%) were undetected in the WGS dataset. In keeping with this observation, 25–99 (2.67–10.55%) of the 938 SNVs and indels reported by the MDT showed insufficient mean coverage in the WES datasets and are thus liable not to have been detected by WES (**Extended Data Figure 7b**). The WES data had a much greater variance in coverage across individuals than the WGS data and, consequently, a larger proportion of individuals tended to have coverage < 20 X in WES datasets than in WGS datasets (even after controlling for mean coverage): the proportion of SNV–participant pairs with insufficient coverage ranged from 3.60 to 9.75% across WES datasets and the corresponding proportion of indel–participant pairs ranged from 4.07 to 9.31%, while, for WGS, the proportions were only 1.63% and 1.36% for SNVs and indels, respectively (**Extended Data Figure 7c and 7d**).

Overview of diagnostic results

In total, 1,138 reports were generated by the MDTs, listing 1,103 distinct causal variants (731 SNVs, 264 indels, 102 large deletions, six of other types). 266 (26.7%) of the SNVs and indels were absent from HGMD Pro2018.1 and absent from the set of ClinVar variants with a likely pathogenic/pathogenic assignment and without any benign assignments (**Supplementary Table 2 (SNV and indel list, Large deletion list, Complex structural variant list)**). The reported variants were observed in 329 unique diagnostic-grade genes. We ranked these genes by the number of clinical reports in which they featured. The top three genes (*BMPR2*, *ABCA4* and *TNFRSF13B*) featured in a quarter of all reports. The subsequent 20 genes featured in a further quarter of reports. Most of the remaining 306 genes each featured in a single report (**Figure 2c**).

The minimum number of genes needed to cover 50% of the reports ranged between one and 11 across the domains, excluding NDD. For example, over 80% of the reports issued in the PAH domain featured *BMPR2* and over 50% of the reports issued in the SMD domain featured one of *JAK2*, *SPTA1*, *DNMT3A*, *NF1*, *SBDS* and *SEC23B*. 30 genes were needed to cover at least 50% of the reports issued for the NDD domain. The domain specific numbers of genes required to cover the reports not covered by the minimal sets of genes (for covering at least 50% of reports) ranged between one and 76 genes, depending on the domain (**Extended Data Figure 6a**). Most of the autosomal SNVs and indels included in reports were absent from gnomAD and TOPMed. The probability that a reported autosomal SNV or indel exhibited dominance was inversely correlated with MAF (**Extended Data Figure 6b**). Interestingly, 21 causal variants were reported in diagnostic-grade genes included in two domains (**Extended Data Figure 6c**). For example, two BPD cases with thrombocytopenia, platelet function abnormality accompanied by bleeding and an NDD case with brachycephaly, microcephaly and global developmental delay, carried likely pathogenic missense variants in the phosphatase encoded by *PTPN11*, a gene known to harbour variants causal of Noonan syndrome^{75,76}. Six MPMT cases, two SMD cases and one NDD case carried a protein-truncating variant in *NF1* causative of type 1 neurofibromatosis, manifesting variously as neoplasm of the nervous system or small intestine (MPMT), dilatation and neurofibromas (NDD), and failure to thrive, broad philtrum, xanthomatosis, cafe-au-lait spot, hypotelorism, juvenile splenomegaly, enlarged kidney, hepatomegaly, monocytosis, myelomonocytic leukemia, thrombocytopenia and anaemia (SMD) and dilatation and neurofibromas (NDD). Conversely, some patients had phenotypes caused by variants in several genes. For example, the seven patients with a causal variant in *NMNAT1* or *RPE65* all have retinal dystrophy, as expected, but two also had intellectual disability, likely caused by variant(s) in other genes.

Reports implicating likely pathogenic or pathogenic variants were issued for 1,138 affected individuals. The domain specific diagnostic yield (the percentage of diagnosed patients) ranged from 0% (for PMG) to 53.9% (for IRD) (**Figure 2b**). The wide range in yield can be attributed to variation in the extent of genetic screening before enrolment (see Enrolment section), variation in the genetic architecture of disease and variation in the depth of public domain knowledge of the genetic aetiologies of diseases. For example, the high yield for the IRD domain, in which more than half of the affected individuals received a conclusive report, can be attributed to

detailed retinal image phenotyping, a lack of genetic pre-screening and a predominantly autosomal recessive mode of inheritance. In SRNS, extensive genetic pre-screening by WES⁷⁷ and the apparent contribution of a polygenic component to SRNS risk⁷⁸ led to a relatively low diagnostic yield (11.4%). In BPD, extensive phenotypic pre-screening was performed (e.g. full haemostasis testing). However, certain phenotypic traits, such as bleeding, are strongly influenced by environmental exposures (e.g. trauma, including surgery), and the genetic architecture of BPDs is diverse, resulting in an overall yield of 12.7%. The reason that no causal variants were identified in PMG patients remains unclear, but there is emerging evidence that most cases of PMG can be explained by polygenicity⁷⁹.

Deletion calling by WES remains challenging, particularly the calling of short deletions encompassing only one exon or part of an exon and the calling of deletions of exons with poor coverage⁸⁰. WGS allows more reliable calling of large deletions. 177,550 distinct deletions were retained for analysis after filtering. We excluded deletions with an allele frequency greater than 1 in 4,000 in the MSUP and deletions that did not overlap any diagnostic-grade genes for a case's domain. We visually reviewed the aligned read data of the 524 remaining unique deletions using IGV. Nearly 85% (444) of these deletions were deemed confirmed or likely to be real after this visual review and were presented to the relevant MDTs. Of these, 22.97% (102) were labelled as likely pathogenic or pathogenic (**Supplementary Table 2 (Large deletion list)**). The length of the 104 distinct reported deletions ranged between 203bp and 16.80Mb (mean 786.33Kb; median 15.91Kb) and nearly half (43) were absent from ClinVar.

A total of 28 heterozygous deletions encompassing autosomal recessive genes were returned to the referring clinician, of which 25 were reported in compound heterozygosity with an SNV or an indel on the other haplotype. For example, one SRNS case carried a rare 14.55Kb deletion which removed the last seven exons of *NPHS1* in compound heterozygosity with a splice donor variant. These two variants explain the patient's phenotype. This is the first report of a large deletion in *NPHS1* causing SNRS.

15 of the reported deletions were called as homozygotes. Of these, nine were carried by probands born to consanguineous parents. In another case the homozygosity was illusory due to an overlap between two distinct heterozygous deletions originating from the unrelated parents. In the remaining five cases, the rare homozygous deletions were confirmed by visual inspection in IGV and by additional genotyping; uniparental disomy was excluded as a possible cause. Overall, 102 of the 1,038 reports issued (9.82%) included a large deletion, confirming the clinical importance of this category of variant and the value of using accurate deletion calling algorithms. For the IRD and NDD domains only, we performed an additional analysis of complex structural variants (cxSV), which identified two cases harbouring a cxSV in a pertinent diagnostic-grade gene (see ref. 70 for details).

The penetrances of causal genetic variants differed by domain. In general, causal variants in diagnostic-grade genes for IRD, NDD, BPD and SMD have high penetrance (conditional on a second high penetrance causal allele being present on the other haplotype in the case of

autosomal recessive diseases). In contrast, causal rare variants in *BMPR2*, which are present in approximately 80% of patients with familial PAH and in 20% of patients with idiopathic PAH, have been reported to have a penetrance as low as 14% in males^{8,81}. It is thought that environmental triggers are important in precipitating the disorder. In NPD, penetrance is also modulated by environmental exposures, such as low temperature, which is a causal risk factor for non-freezing cold injury.

Upload of reported variants to public repositories

The reported alleles and their clinical interpretations have been deposited in ClinVar and are available under the study names “NIHR_Bioresource_Rare_Diseases_13k”, “NIHR_Bioresource_Rare_Diseases_MYH9”, “NIHR_Bioresource_Rare_Diseases_PID” and “NIHR_Bioresource_Rare_Diseases_Retinal_Dystrophy”, and also in DECIPHER.

Reported variants informing changes in clinical management

We did not systematically collect information about whether the MDT reports led to changes in clinical management or the repurposing of existing drugs. There have, however, been striking examples of how the latter has changed the care path of patients enrolled in the project. Firstly, 27 patients with early-onset dystonia were shown to carry causal variants in the histone methyltransferase gene *KMT2B*, and many were treated successfully by deep brain stimulation⁸². Secondly, gain-of function variants in *DIAPH1* cause macrothrombocytopenia and deafness⁸³. We have recently shown that the treatment of these patients with Eltrombopag, an FDA-approved drug for the treatment of autoimmune thrombocytopenia, increases the platelet count to a safe level in the perioperative setting⁸⁴, thereby reducing the need to use donor platelet concentrates. Finally, we identified a pedigree with a p.E527K gain-of-function variant in the kinase SRC resulting in juvenile myelofibrosis and severe thrombocytopenia, further complicated by osteoporosis. After discovering that this variant was pathogenic, the proband (case 27 of the published pedigree) was successfully cured of her thrombocytopenia by an allogeneic haematopoietic stem cell (HSC) transplant from her HLA compatible sister, who was negative for the causal variant⁸⁵.

The identification of causal variants has improved the stratification of patient care, including the frequency of clinic visits. We showed that haploinsufficiency of *NFKB1* is the most frequent cause of primary immune deficiency. Patients with loss-of-function *NFKB1* mutations have recurrent and severe infections accompanied by autoimmunity and unexplained splenomegaly and an increased risk of oncological manifestations. These new findings have a direct impact on the care of this genetically defined category of PID patients¹⁰. Similarly, we identified 27 BPD cases with isolated thrombocytopenia caused by variants in *ANKRD26*, *ETV6* or *RUNX1*. These genes encode DNA-binding proteins and the pathogenic variants are associated with an increased risk for haematological cancers, which is particularly severe for patients with variants in *ETV6*^{86,87} and *RUNX1*⁸⁸. Hence, frequent follow-up clinic visits are warranted and allogeneic HSC transplants need to be considered as they can be curative. In contrast, the clinicians of the 19 patients with thrombocytopenia due to variants in *ACTN1*, *CYCS* or *TUBB1* could reassure their patients that their conditions were benign⁸⁹. Such patients do not require regular follow-up

but haematology consultation is required at times of haemostatic challenges (e.g. childbirth or surgical procedures, including dental procedures). Finally, the identification of four new genes (*ATP13A3*, *AQP1*, *GDF2*, *SOX17*) for PAH led to improved diagnostic sensitivity for this condition. The genetic findings have also confirmed that mutations in *BMPR2*⁹⁰ and *EIF2AK4* are associated with a poorer prognosis. In particular, patients carrying causal variants in *EIF2AK4* should be referred early as potential lung transplant recipients⁷.

Structural variants in patients with null protein phenotypes

We recorded comprehensive intermediate phenotype data on most BPD patients. This included measurements relating to haemostasis, including of coagulation factor levels (Factor I (fibrinogen; *FGA*, *FGB*, *FGG*), II (*F2*), VII (*F7*), IX (*F9*), XI (*F11*), von Willebrand Factor (*VWF*)), platelet receptor expression levels e.g. of the fibrinogen receptor (GPIIb/IIIa; *ITGA2B/ITGB3*) and red cell Rh grouping information. These data allowed us to identify 15 unexplained cases who had a complete or near absence of expression of one of these proteins. We refer to such an absence of protein expression as a 'null phenotype.' Genetic defects relating to null phenotypes tend to have recessive inheritance. We identified:

- eight unsolved cases with null phenotypes for coagulation factors;
- six unsolved cases of possible Glanzmann's thrombasthenia (diagnosed from observing impaired platelet aggregation in response to at least two of the five agonists arachidonic acid, collagen, epinephrine, thromboxane analogue and TRAP, and observing normal agglutination in response to ristocetin);
- a patient with a mild bleeding disorder accompanied by an unexplained haemolytic anaemia, who had no Rh group proteins on her red cell membranes.

Three of the 15 cases with null phenotypes had a rare coding variant on one haplotype of the relevant gene while the other haplotype appeared to be wild type according to the standard variant calls. Through visual inspection of the reads in IGV, we were able to resolve two of these three cases.

- The first case had Glanzmann's thrombasthenia and carried a variant in *ITGB3* which encodes a premature stop at amino acid position 242. Visual inspection of the gene body revealed an excess of improperly mapped reads in intron 9, primarily due to alignment of paired-end reads facing away from, rather than towards, each other (**Extended Data Figure 8a**). We counted the number of improperly mapped read pairs in the region in our collection after excluding the GEL domain participants (as we do not hold their detailed phenotypic data), and found this case to have the greatest number of improperly mapped reads out of 6,656 individuals whose samples were sequenced using 150bp chemistry (**Extended Data Figure 8b**). In a separate dataset of 304 WGS samples sequenced with 150bp chemistry, we identified two further individuals (a Glanzmann's thrombasthenia case and his mother) with a similar excess in intron 9, primarily due to read mates aligning to different chromosomes (**Extended Data Figure 8c,d**). The proband's platelets were devoid of GPIIb/IIIa and the mother had a ~50% reduction in expression of GPIIb/IIIa (data not shown). The proband carried a rare paternally inherited missense variant encoding a change from threonine to proline at

amino acid position 456 of *ITGB3*. Sanger sequencing of the *ITGB3* cDNA obtained by reverse transcription of the RNA from the proband's platelets showed exclusive expression of the proline coding allele and provided no evidence of an alternatively spliced transcript (data not shown). In order to resolve the genetic sequence of intron 9, we performed Oxford Nanopore-based sequencing of long-range PCR-amplified target DNA as previously described⁷⁰. The flow cell ran for 3 hours, and the mean coverage was 863,986X. We identified an insertion of an SVA (Alu, SINE-VNTR-Alu) retrotransposon element of 2,270bp at position 45369041 of chromosome 17, predicted to induce nonsense-mediated decay. SVA elements are present throughout the reference genome, explaining the short reads' mates aligning to various chromosomes (**Extended Data Figure 8e**). To date, roughly 130 pathogenic variants caused by retrotransposon elements have been documented in the literature⁹¹, but none of them have been implicated in Glanzmann's thrombasthenia. Here we report the first such example and demonstrate the utility of long-range sequencing reads for identifying complex causal variants suggested by short-read WGS analysis.

- The second case had red cells which serotyped negative for all common antigens of the Rh blood group system. Upon further investigation, the red cells were shown to lack both the RHD and RHCE proteins and also the Rh-associated glycoprotein RHAG. This very unusual Rh_{null} red cell phenotype is known to cause a syndrome characterised by chronic haemolytic anaemia of varying severity⁹². The most frequent genetic explanation for this rare condition is the presence of loss-of-function variants on both *RHAG* haplotypes, because the RHAG protein is essential for the expression of the RHD and RHCE proteins on the red cell membrane⁹³. This case was found to carry a known causal loss-of function splice donor acceptor variant in intron 2 of *RHAG*⁹⁴. The second mutation was identified by visual inspection of the reads of the *RHAG* locus, as a heterozygous tandem duplication spanning exons 2–7.

Genetic association between rare variants and rare diseases

Statistical approach

We used the BeviMed statistical method⁹⁵ to identify genetic associations with rare diseases in our dataset. Each run of BeviMed requires selection of a subset of unrelated study participants partitioned into cases and controls and a set of rare variants to test for association. To achieve adequate power, the cases should be chosen to maximise the chance that they share a common genetic aetiology (e.g. by identifying a group of individuals with similar phenotypes) and the rare variants should be chosen to maximise the chance that they are aetiologically exchangeable (e.g. by identifying variants predicted to have a similar effect on a gene product). BeviMed computes posterior probabilities over the hypotheses:

- no association between any variant and case/control status,
- dominant association between a pathogenic subset of variants and case/control status,
- recessive association between a pathogenic subset of variants and case/control status.

BeviMed also reports the posterior probability that each variant is pathogenic conditional on the mode of inheritance. We impose a prior correlation structure on the probabilities that variants are pathogenic which reflects prior belief about which class of variant is disease causing. This

correlation structure derives from groupings of variants by their VEP consequences. The class of variant responsible for causing disease is inferred by BeviMed, and this inference can suggest a molecular aetiology.

Case/control groupings using phenotypic tags

A set of phenotypic ‘tags’ were defined for each domain that determined the case/control groupings for BeviMed. Cases shared a particular tag if their phenotypic characteristics were considered compatible with a similar genetic aetiology. Some tags were set using logical rules applied to the HPO terms and other data; other tags were set manually. The full set of tags and the corresponding numbers of cases and controls can be found in **Supplementary Table 3 (Phenotypic tags)**. Given a particular tag and a set of rare variants, the corresponding case/control groupings were obtained as follows:

1. cases: unrelated (i.e. at most one per family) affected individuals with the tag who have not been issued with an MDT diagnosis that includes an aetiological variant outside the set of given variants;
2. controls: members of the MSUP who do not have the tag and who are not related to any of the cases.

Selection of variants affecting transcript sequences

For each gene, we selected SNVs/indels for inclusion in the BeviMed analyses. Only variants with MAF less than 0.002 in the MSUP were used. We defined $PMAF_x$ for each variant as the probability that $Y \geq C$, where Y is a random binomial minor allele count for a variant with MAF $1/X$ and C is the fixed observed minor allele count. Variants had to have a $PMAF_{500}$ greater than 0.05 in four subsets of the participants: European probands, non-European probands, European members of the MSUP and non-European members of the MSUP. SNVs/indels were retrieved from the HBase variant database and were filtered as follows:

- a minimum OPR > 0.98 was required;
- unless a variant was in HGMD with class “DM” or “DM?”, it had to have $PMAF_z$ greater than 0.05 in every gnomAD population (only gnomAD males were used to compute the $PMAF_z$ on variants in the non-PAR of X). We used $Z = 1,000$ for recessive association analyses and $Z = 10,000$ for dominant association analyses;
- SNVs with a CADD phred score < 10 were excluded;
- variants in the non-PAR of X that appear only as heterozygotes in males were excluded;
- multi-allelic variants for which the reference allele was the minor allele were excluded;
- unless a variant was in HGMD with class “DM” or “DM?”, it had to have a CellBase-predicted consequence of type transcript_ablation, splice_acceptor_variant, splice_donor_variant, stop_gained, frameshift_variant, stop_lost, start_lost, transcript_amplification, inframe_insertion, inframe_deletion, missense_variant, protein_altering_variant, regulatory_region_ablation, non_coding_transcript_exon_variant or 5_prime_UTR_variant on a transcript of type lincRNA, miRNA, misc_RNA, Mt_rRNA, Mt_tRNA, protein_coding, rRNA, snoRNA, snRNA, TR_C_gene, TR_D_gene, TR_J_gene, TR_V_gene, IG_C_gene, IG_D_gene, IG_J_gene or IG_V_gene (“retained biotypes”).

Large deletions were subjected to the same PMAF₅₀₀ filters as the SNVs and indels.

In order to identify groups of variants that had a high prior probability of being aetiologically exchangeable, we grouped variants into gene specific (non-mutually exclusive) classes labelled “5’ UTR”, “moderate” and “high-impact”:

- “5’ UTR:” there were no VEP consequences with a HIGH or MODERATE impact and at least one transcript was annotated with the VEP ‘5_prime_UTR_variant’ consequence;
- “moderate”: at least one transcript was annotated by VEP with a MODERATE or HIGH impact consequence or with the VEP ‘non_coding_transcript_exon_variant’ consequence;
- “high-impact”: at least one transcript was annotated by VEP with a HIGH impact consequence or the variant was a large deletion that overlapped an exon of a transcript with a retained biotype.

Thus, all variants in the high-impact class were also in the moderate class, while variants in the 5’ UTR class were not present in any other classes. The rationale for combining missense variants and variants with a HIGH impact consequence label is that both types of variant are capable of inducing loss of function. The conditional prior probabilities of the models that treated each of these three groupings of variants as disease causing were 0.01, 0.495, and 0.495 for the “5’ UTR”, “moderate” and “high-impact” variant classes, respectively. All priors in the BeviMed model were set at the default values, including the prior probability of association accumulated across all association models, which was set to 0.01.

Post-processing of BeviMed results

We applied three post-processing steps to the BeviMed results.

1. Allele frequency datasets are much smaller for non-European ancestry populations than for the European ancestry population. Consequently, the false positive rates of BeviMed association analyses relying on non-European cases can be inflated if variants that are moderately rare in non-Europeans are unintentionally included. To guard against this, when we obtained a BeviMed posterior probability of association (PPA) > 0.75 for a gene–tag pair, we re-ran BeviMed excluding all variants:
 - present in heterozygosity in at least two individuals from the same non-European ancestry (e.g. South Asian or African) group (homozygous genotypes did not contribute towards the exclusion of variants because they are likely to result from consanguinity and be aetiological) and
 - absent from all Europeans.

If the new BeviMed PPA was < 0.25, we removed the result because it was dependent on variants that may be moderately rare in a non-European population.

2. Genomic regions spanning more than one gene that are deleted in multiple cases can potentially induce multiple correlated genetic associations with a particular disease tag. In order to guard against false positive associations due to these correlations, we post-processed the results to ensure that no two genes having a PPA with a given tag > 0.75 both depended on the same set of deletions to obtain a PPA > 0.25. This was done

as follows. For each tag, if at least two genes had a PPA > 0.75, we looped through each pair of such genes, and marked the gene with the lower PPA for removal if, when deletions shared between the genes were excluded, its PPA dropped below 0.25. After completing the loop, the associations corresponding to marked genes were removed.

3. The results with PPA > 0.75 may contain multiple associations between a gene and correlated phenotypic tags. In order to remove such redundant information, we recorded only one association per gene, corresponding to the tag for which the highest PPA was obtained.

BeviMed results for genes

BeviMed was run for each tag/gene pair. For each pair, the PPA was obtained by marginalising the joint posterior over all modes of inheritance (dominant and recessive) and classes of variant (5' UTR, moderate and high-impact). A PPA > 0.75 was considered good evidence of a causal relationship. The 95 genetic associations that remained after post-processing are shown in **Figure 2d** and in **Supplementary Table 3 (BeviMed associations tags)**.

Observed false positive rate

We applied BeviMed to compute the PPA between each of 23,090 genes and each of 29 tags (i.e. we computed 669,610 PPAs). 125 of the PPAs were > 0.75 and 118 of these corresponded to BeviMed analyses passing post-processing QC (steps 1 and 2 above). The 118 PPAs > 0.75 corresponded to 95 unique genes, 79 of which are confirmed. For each of the 95 genes, either all or none of the associations with tags were confirmed. (The lead tag, that is the tag corresponding to the greatest PPA for each of these 95 genes, is shown in **Figure 2d**). The principal reason that some genes were associated with multiple tags is that such tags were assigned to groups of patients which substantially overlapped.

Define a gene-level false positive as a gene with a PPA > 0.75 for a tag for which variation in the gene is not a cause of variation in the tag status. Define a gene-level true positive as any other gene with PPA > 0.75 for a tag. Define a gene-level true negative as a gene with a PPA < 0.75 for all tags and for which variation in the gene is not a cause of variation in the status of any tag. Because the associations with 79 genes are confirmed (and are therefore true positives), there are at most 16 gene-level false positives (**Figure 2d** also shows these 16 unconfirmed associations). The observed gene-level false positive rate (FPR) depends on the number of these 16 genes that correspond to false positives and the number of genes that correspond to true negatives. In principle, the FPR could range between 0 (if and only if all 16 unconfirmed genes correspond to true positives) and 1 (if and only if there are no true negatives). Nevertheless, we can derive an upper bound on the FPR if we make the conservative assumption that the observed true positive rate (TPR) is greater than the FPR (i.e. that the BeviMed analysis of this dataset carries some useful information). Let n be the total number of genes, let x be the unknown number of false positive genes and let y be the unknown number of true negative genes, then

$$FPR = \frac{x}{x+y}; \quad TPR = \frac{95-x}{n-y-x}.$$

By equating y in the above and applying the assumption $FPR < TPR$ we obtain

$$\frac{x}{FPR} - x = n - x - \frac{95-x}{TPR} > n - x - \frac{95-x}{FPR},$$

which implies the upper-bound,

$$FPR < \frac{95}{n}.$$

In total, we tested 23,090 genes for association, which gives an upper bound on the observed false positive rate of $95/23,090 = 4.11 \times 10^{-3}$. However, the upper bound is very conservative because it implies 3,873 true negatives, which is clearly unrealistic. Under the more reasonable assumption that 20,000 genes are true negatives, the FPR can be bounded by $16/20,016 = 7.99 \times 10^{-4}$.

Positive predictive value

We performed a comprehensive analysis of the study-wide positive predictive value (equivalent to 1 minus the false discovery rate). We present four alternative approaches to making inferences about the study-wide positive predictive value (PPV).

Lower bound on observed PPV. Only 16 of the 95 genetic associations we report are unconfirmed, which sets a lower limit on the observed PPV of 83%.

Posterior PPV. The mean posterior probability for the reported associations gives a posterior PPV of 93%.

Synonymous variant PPV. We retrieved the rare synonymous variants from our genotype database using the same criteria as for the reported data analysis but without filtering on CADD score. We then applied BeviMed for each gene with the phenotypic tag-based groupings used for the reported data analysis. This generated 23 results with a PPA > 0.75, corresponding to a point estimate of 76% for the PPV. This estimate is subject to sources of bias that do not apply to the permutation PPV described below. In particular,

- the count distribution of rare synonymous variants in genes differs from the count distribution of the variants used in the real data analysis,
- the synonymous variants have different MAF spectra to the variants in the real analysis,
- for the reported analysis we perform model averaging over aetiological models restricting to different classes of variant (5' UTR, moderate, high-impact, 3' UTR), while there is only one class of synonymous variant.

Permutation PPV. We estimated the study-wide expected number of gene-tag associations with a PPA > 0.75 under the assumption that the null hypotheses of all tests are true. This estimate, divided by the observed number of gene-tag associations with a PPA > 0.75, gives an estimate of the FDR (equivalently, 1-PPV) which is conservative because in practice a proportion of tests are of null hypotheses that are false. Our approach follows closely that proposed by Tusher, Tibshirani and Chu ⁹⁶.

In order to estimate the expected number of gene–tag associations with a PPA > 0.75 under the null, we generated 100 permuted datasets and ran the entire association analysis procedure on each dataset. To ensure that the permuted datasets had a tag correlation and ancestry group structure that closely resembled that of the real dataset, we used the genotype data from the MSUP and outcome data from participants selected within each ancestry group according to the following procedure:

1. Let k_g be the number of participants in the MSUP who belong to ancestry group g .
2. Create a participant (rows) by tag (columns) matrix M_g from the real outcome dataset for participants in ancestry group g , in which elements are coded:
 - 1 to indicate that the participant is a case for the tag,
 - 0 to indicate the participant is a control for the tag,
 - NA to indicate that the participant is to be excluded from the analysis for the tag, because the participant was excluded from the reported analyses for the tag due to relatedness.
3. Row sort the matrix on the row count of 1s, breaking ties by sorting on the row count of 0s.
4. Select the leading k_g rows as outcome data.

On each iteration of the permutation procedure, for each ancestry group g , we randomly assigned the k_g rows of the outcome matrix M_g to the genotype data of the k_g unrelated participants. We applied BeviMed to the ancestry unified dataset of $\sum_g k_g$ rows. For each gene–tag combination, we relabelled as controls the cases who were so relabelled in the actual analysis (because they had been genetically explained by variants in a different gene). We applied the study-wide post-processing filters described above.

The entire process was repeated 100 times. This process, without permutation, generated 95 results with a PPA > 0.75, which is the same number of results obtained using the dataset used for the reported analyses (**Figure 2d**). Over 100 repetitions, the mean number of PPAs > 0.75 was 20.0, giving a permutation PPV of 79%. Increasing the PPA threshold from 0.75 to 1.0 increased the PPV approximately linearly from 79% to 100% (not shown).

Estimated false positive rate

Using the procedure described above to compute the permutation PPV, we estimated the false positive rate as the mean proportion of PPAs > 0.75 before retaining only the top associated tag for each gene. In total, we tested 23,090 genes over 29 tags (i.e. we performed 669,610 statistical tests), of which on average 20.0 had PPA > 0.75, giving an estimated FPR of 3.0×10^{-5} .

Polygenic and rare variant associations with phenotypic extremes in the UK Biobank cohort

Selection of participants

A genome-wide association study (GWAS) of FBC traits measured in approximately 173,000 European ancestry participants in the UK Biobank cohort ($n \sim 133,000$) and the INTERVAL study ($n \sim 40,000$)⁷¹ has previously identified 582 genetic variants independently associated with quantitative properties of mature red blood cells⁹⁷. Most of these variants were common: only 40 had an in-sample MAF lower than 1%. The GWAS design had limited power to identify associations with rare variants, due to the imprecision of genotype imputation at low allele frequencies (**Appendix 2: Extreme red cell traits in UK Biobank**). In order to identify rare variants associated with full blood count measured properties of red cells, we attempted to derive a univariate quantitative phenotype with high rare-variant heritability. To do this, we used 65 previously reported GWAS associations between variants with MAF < 1% and phenotypes of red cells as a model for the effect of rare variants on mature red cell phenotypes (**Figure 3a**). 384 putative European ancestry individuals were selected for WGS from each tail of the quantitative phenotype, of which 383 from the left tail and 381 from the right tail were successfully sequenced and passed QC (**Figure 3b**). Of these, 382 from the left tail and 368 from the right tail were classified as European by our clustering algorithm (see above). The left tail of the score (BeviMed tag “Left”) included individuals with generally lower than average RBC# and generally higher than average MCV, while the right tail (BeviMed tag “Right”) included individuals with generally higher than average RBC# and generally lower than average MCV (**Figure 3c**). A more detailed explanation of the technicalities of the selection procedure can be found in **Appendix 2: Extreme red cell traits in UK Biobank**.

Polygenic associations with tails

We first sought to understand the polygenic variant contribution to the risk of being extreme for the quantitative phenotype, by using variants with published GWAS associations with RBC# or MCV to construct a polygenic predictor of the quantitative phenotype. 448 of the 468 variants reported to be associated with RBC# or MCV in ref. 97 had well called genotypes (minimum OPR > 0.8) in the WGS data. We identified well called high linkage disequilibrium proxy variants for 16 of the remaining 20 variants. Six of these 16 variants had $r^2 = 1.0$ with their proxy variant in the European component of the 1000 Genomes data, nine had $r^2 > 0.95$, 14 had $r^2 > 0.9$ and all had $r^2 > 0.7$. The four variants for which no proxy could be found jointly explain 0.04% of the variance of the phenotype in 116,689 European ancestry participants included in the interim release (May 2015) of the UK Biobank imputed genotype data. We regressed the quantitative selection phenotype jointly on the 464 well called/proxy variants using 298,607 European participants from the full release of the UK Biobank imputation data (July 2017) who were not included in the interim release and were not WGSed for this study. The polygenic score formed by the linear predictor of this regression explained 22% of the variance in the phenotype in sample ($R^2 = 0.219$).

We identified 522 unrelated European participants in study domains other than UKB who can be considered a uniform random population sample with respect to their polygenic score values.

These are rare disease patients whose pathology has been fully explained by rare variants declared as “likely pathogenic” or “pathogenic” in an MDT, or healthy individuals with a patient relative, the phenotype of whom has been fully explained in the same way. We call this group ‘unselected’ because they have not been selected with respect to their polygenic red cell genotype. The distribution of the polygenic score in the individuals with confirmed European ancestry from the two UK Biobank selected tails and in the unselected category is shown in the boxplots in **Figure 3d**. There is strong statistical evidence for an increasing trend in the mean value of the score with an ordinal variable for category membership (0 = left tail, 1 = unselected controls, 2 = right tail; P -value = 2.4×10^{-96} , linear regression).

The central limit theorem implies that the distribution of the polygenic score should be approximately Gaussian in the unselected control group. Under the assumptions 1) that the quantitative selection phenotype (**Figure 3b**) has a Gaussian distribution and 2) that the mean of the phenotype depends linearly on the polygenic score, the expected distribution of the score in each tail group can be derived, conditional on the quantitative trait tail selection thresholds and conditional on the population proportion of phenotypic variance explained by the polygenic score (see **Appendix 2: Extreme red cell traits in UK Biobank** for details). The probability densities of these distributions are displayed as violin plots underlying the boxplots in **Figure 3d**. The selection thresholds chosen correspond to the quantitative phenotype value for the least extreme individual selected from the corresponding tail. (This is a conservative approach, because the selection procedure involved some recalibration of the phenotype for age and sex in the tails of the score). Comparison of the expected distributions with the observed data suggests that tail status is less strongly associated with the polygenic score than predicted by the Gaussian variance components modelling. A number of factors could explain the incongruence between theory and observation. The tails of the quantitative phenotype are heavier than Gaussian and the excess density in the tails could be due to rare events unrelated to polygenicity: for example, haematology analyser measurement error. The excess may also be partly explained by the effects of rare or extremely rare alleles. This motivates analyses designed to search for such alleles and the genes mediating their associations.

WGS GWAS

We performed a WGS GWAS with the aim of identifying genetic variants (particularly moderately rare genetic variants undetected by previous GWAS using genotypes imputed from array data) associated with the ordinal phenotype indicating group membership. We analysed the 11,147,508 genetic variants with a minimum OPR > 0.8 and an in-sample MAF > 0.25% in the 1,272 European ancestry individuals in the two tails and the unselected control category. For each variant, we regressed the ordinal outcome on the alternative allele count using adjacent categories logistic regression, under the assumption that the log-odds for the right tail vs the unselected controls is the same as the log-odds for the unselected controls vs the left tail. Our analysis identified no novel associations at a significance threshold of 1×10^{-7} . However we did replicate two well known common variant red cell associations in the MYB-HBS1L (P -value = 1.4×10^{-10}) and ODF3B-TYMP-SCO2 (P -value = 1.4×10^{-8}) loci⁹⁸.

BeviMed analysis

The participants from each tail who passed QC (including non-European ancestry participants) were treated as a rare disease case group in a BeviMed analysis. We also performed a combined analysis in which the participants in both tails were merged into a single case group.

The inferred posterior probabilities for the genes with the strongest evidence for association (posterior probability greater than 0.4) for the two tail group analyses are displayed in **Figure 3e**. The strength of evidence (respective posterior probabilities 0.997 and 0.575) that rare variants in *HBB* and *TFRC* are associated with the low MCV/high RBC# phenotype (the right tail of the phenotype) is entirely consistent with biological knowledge (**Supplementary Table 3 (BeviMed associations UK Biobank)**) and these associations can be considered 'positive controls'. Similarly, prior biological knowledge suggests that the association identified with the transcription factor *CUX1* is very plausible while the associations identified with *ALG1*, *ZNF407* are plausible. The weight of evidence from a detailed examination of the alleles implicated in the other genes exhibiting a BeviMed PPA > 0.4, together with the expression profiles of the genes in blood cells and other tissue types is less compelling. A list of all variants in the associated genes, for which the posterior probability of pathogenicity is greater than 0.5, is given in **Supplementary Table 3 (BeviMed variants UK Biobank)**. In conclusion, our analysis demonstrates that genomic loci carrying rare alleles causing large deviations in quantitative traits can be identified by applying the BeviMed analysis approach. Since the extreme tails of any heritable quantitative trait are likely to be under negative selection, mutations in such genes may cause rare diseases or be otherwise clinically relevant. The forthcoming WGS of the full UK Biobank cohort offers a further opportunity to search for rare variant associations with other biomedically relevant quantitative intermediate traits, including those blood cell traits not explored here.

Matchmaker Exchange

GeneMatcher⁹⁹, which is part of Matchmaker Exchange¹⁰⁰, was used to identify additional patients, external to the cohort, carrying rare variants in candidate disease genes identified through a manual review process informed by the literature. We were able to find matches for three patients:

1. A BPD patient with a complete lack of serotonin storage in his platelet dense granules had symptoms including hypotonia, mental disability, epilepsy, uncontrolled movements and gastrointestinal problems. The patient had consanguineous parents and was homozygous for a rare allele in *SLC18A2*, which codes for VMAT2, a serotonin transporter¹⁰¹. Consequently, the patient was deficient for VMAT2. We identified a matching child with a similar clinical phenotype. We subsequently became aware of a case report of a patient with VMAT2 deficiency exhibiting a similar neurological phenotype¹⁰².
2. A BPD patient and his sister with a platelet function defect and mild thrombocytopenia had symptoms including psychomotor retardation and epilepsy. Both siblings inherited a rare allele in *MADD* from each of their parents. The siblings matched with 13 patients from 10 different families with similar neurological phenotypes (manuscript in

preparation). However, platelet studies in these other patients have not yet been performed.

3. An NDD patient with intellectual disability, autistic features and seizures was matched to three unrelated patients with a similar phenotype using GeneMatcher. In addition, a fourth patient was matched by personal correspondence¹⁰³. In all five patients, *WASF1* had been independently identified as a strong candidate because it is constrained against loss-of-function variants in ExAC (probability of loss-of-function intolerance (pLI) = 0.91) and is highly and specifically expressed in the adult human brain¹⁰⁴. All variants were *de novo*, predicted to induce a loss of function, and absent from 1000 Genomes, ExAC and gnomAD. Interestingly, they clustered around the WASP-homology 2 (WH2) domain, in the highly conserved C-terminal actin-binding WCA region.

Regulome analysis

Ethics

Blood samples for generating open chromatin and histone modification data for activated CD4+ T cells (aCD4) were collected with written and signed informed consent, approved by the East of England – Cambridgeshire and Hertfordshire REC (reference 05/Q0106/20). The other epigenetic data described in this section have been released as part of previous studies. Information about the consent under which these other data were generated can be found in the publications cited below.

Definition of cell type specific regulomes

Given that WGS provides variant calls across the entire genome, we sought to identify rare variants that exert their effect on phenotype through disruption of non-coding (possibly non-exonic) regulatory elements, such as enhancers. Genetic variation in these elements has not been systematically assayed in previous genetic studies of rare disease, which have principally relied on WES or other targeted sequencing. Regulatory elements are largely cell type specific. We sought to identify regions occupied by regulatory elements in cell types relevant to the biology of a subset of the rare disease domains (**Supplementary Table 3 (Regulome data)**). We call the set of regulatory elements corresponding to a particular cell type a 'regulome'. We determined the regulomes for aCD4 cells, B cells (B), erythroblasts (EB), megakaryocytes (MK), monocytes (MON) and resting CD4+ T cells (rCD4).

For each cell type, we applied the RedPop method (see below) to identify regulatory elements from open chromatin data (ATAC-seq) and histone modification data (H3K27ac). Additionally, we had access to the following transcription factor (TF) chromatin immunoprecipitation sequencing (ChIP-seq) data, which we used to call binding sites (see below) to expand the regulomes: FLI1, GATA1, GATA2, MEIS1, RUNX1, TAL1 and CTCF in MK; GATA1, KLF1, NFE2 and TAL1 in EB; and CTCF in MON and B.

For each cell type, the regulome build process proceeded as follows:

1. call regulatory elements from ATAC-seq/DNase-seq and H3K27ac-seq data using RedPop,

2. call TF binding peaks using ChIP-seq data if available and obtain enrichment scores,
3. discard TF regions with an enrichment score < 10 unless they overlap between at least two different TFs,
4. collapse overlapping features to obtain a single genomic track,
5. merge features within 100bp of each other.

Each regulome feature was assigned a *target gene* label using either gene annotations from Ensembl 75 or a compendium of previously published promoter capture Hi-C (pcHi-C) data¹⁰⁵ as follows:

1. assign a feature to a gene if the feature overlaps the gene or the region up to 10,000bp either side of the gene body,
2. assign a feature to a gene if the feature overlaps the gene's pcHi-C "blind spot" (the region spanning the *HindIII* restriction fragment overlapping the gene's transcription start site and the two adjacent *HindIII* restriction fragments),
3. assign a feature to a gene if the feature overlaps a linked promoter interacting region identified using pcHi-C in the same cell type.

Regulatory element detection using patterns of peaks (RedPop)

We sought to identify regulatory elements in cell types of relevance to the rare disease domains in this project, in order to search for possible pathogenic variants within them. It is important to locate these elements because they are much more likely to harbour pathogenic variants than other non-coding regions of the genome. Open chromatin data alone do not provide sufficiently high spatial resolution and transcription factor ChIP-seq restricts to regions bound by proteins targeted by specific antibodies. Open chromatin and histone modification (H3K27ac) sequencing data together, on the other hand, can be used to detect regulatory elements with high resolution and in an unbiased fashion. Open chromatin around binding sites typically results in a broad, low-resolution peak of elevated ATAC-seq/DNase-seq coverage. The surrounding nucleosomes of a regulatory element are typically acetylated, leaving two peaks in H3K27ac coverage, spaced a few hundred bp apart. By combining the genomic coverage tracks of an open chromatin and an H3K27ac assay, regulatory elements can be detected with high precision. We developed an algorithm for regulatory element detection using patterns of peaks (RedPop; manuscript in preparation) that utilises these patterns.

First, ATAC-seq/DNase-seq and H3K27ac ChIP-seq reads were aligned to the genome using the BWA⁷³ `aln` command. Open chromatin regions were then called using F-Seq¹⁰⁶. The called open chromatin regions were extended upstream and downstream symmetrically until their lengths were at least 3.2Kb. Any adjacent overlapping segments were merged. Each merged segment was then considered individually. For each segment, the following sets of vectors were generated.

1. *Globally* normalised coverage tracks of open chromatin and H3K27ac. These tracks are position specific read count vectors, which are normalised by the mean coverage genome-wide and smoothed by averaging consecutive segments of 40bp.

2. *Locally* normalised coverage tracks of open chromatin and H3K27ac. These tracks are computed as the globally normalised coverage divided by the mean coverage in the surrounding 800bp region.
3. The covariance track. This track contains the covariance between the open chromatin and the H3K27ac globally normalised coverage tracks, computed in 800bp sliding windows and smoothed by averaging consecutive segments of 800bp.

Any positions for which the value of the smoothed covariance was < -1 and less than the values in the surrounding 160bp were declared local minima. At each position declared a local minimum, the nearest contiguous segment for which the locally normalised open chromatin coverage exceeded the locally normalised H3K27ac coverage, and any other contiguous segments within 100bp of the local minimum, were recorded and extended symmetrically to 400bp. These segments were discarded unless the globally normalised open chromatin coverage track exceeded 47X (a threshold value obtained through a sensitivity/specificity study of ChIP-seq data from MKs). Any overlapping undiscarded segments were merged and recorded as the locations of regulatory elements.

Transcription factor binding peak calling

We applied the BLUEPRINT protocol for ChIP-seq data analysis¹⁰⁷. H3K27ac histone modification sequence reads were mapped to human genome GRCh37 with the BWA⁷³ aligner method. Low-quality reads ($-q\ 15$), multi-mapped reads and duplicate reads were marked and removed with samtools and Picard⁷⁴. We performed QC using deepTools plotFingerprint¹⁰⁸. Files containing aligned read data corresponding to the same cell type were merged. The resulting file was downsampled in proportion to the number of aligned reads. H3K27ac peaks were called by MACS2¹⁰⁹ using the 'narrow' option.

Identifying possibly causal deletions of elements regulating diagnostic-grade genes

In order to identify regulatory elements disrupted in patients with a rare disease, we examined rare deletions called in cases that overlapped regulatory elements called using RedPop or using ChIP-seq analysis. We filtered these deletions to retain only those which delete an element

- with a target gene label that is diagnostic-grade for a recessive disorder encompassed by the domain of the carrier and
- which is called in a cell type relevant to the domain of the carrier (**Supplementary Table 3**).

Furthermore, the deletion had to

- be called homozygous or hemizygous by Manta or Canvas, or
- delete an element the target gene of which contained a rare allele of 'moderate' impact (as described in the *Genetic association testing in genes* section) in the carrier.

The filtering resulted in a list of only four deletions: a heterozygous deletion overlapping the 5' UTR region of the *ARPC1B* found in a PID patient who also carries a frameshift variant in the same gene (see the accompanying article by Thaventhiran et al. in this issue of *Nature*); a hemizygous deletion of a *GATA1* enhancer in a patient with thrombocytopenia, described

below; a homozygous deletion of a CTCF binding site in the first intron of *LRBA* in a PID patient, described below; and a deletion which on manual inspection proved to have been called in error.

Deletion of GATA1 enhancer and HDAC6 open reading frame

The *GATA1* enhancer, called in MKs and EBs, was removed by a 4108bp deletion (X:48,659,245-48,663,353) together with the first four exons of the histone deacetylase 6 gene (*HDAC6*), including its ATG start codon (**Figure 4a**, **Extended Data Figure 9a**). The patient was a 9-year-old boy with macrothrombocytopenia, bleeding symptoms and mild intellectual disability (ID) with an autism spectrum disorder (ASD) (**Extended Data Figure 9b**). His parents were healthy, although mild asymptomatic thrombocytopenia was present in the mother (**Extended Data Figure 9b**). The patient was hemizygous for the deletion while the mother was a heterozygous carrier (**Extended Data Figure 9c–d**). Platelets of the proband expressed no HDAC6 protein (**Figure 4b**). In contrast, HDAC6 expression in platelets from the mother was comparable to expression levels in platelets from unrelated controls and from the father. HDAC6 is the major deacetylase responsible for removing the acetyl group from Lys40 of α -tubulin, which is located in polymerised microtubules¹⁰⁸. Indeed, absence of HDAC6 expression in platelets was accompanied with very high expression levels of acetylated α -tubulin while non-acetylated α -tubulin expression levels were similar to levels in controls (**Figure 4b**). Platelets from the mother but not the father contained at least some platelets with hyperacetylated α -tubulin levels (**Figure 4b**). HDAC6 is extremely intolerant to LOF variants (pLI = 1) and the proband was the only patient in our cohort with a hemizygous high impact variant. **Figure 4a** shows that the deleted region also contains binding sites for several transcription factors important for megakaryopoiesis¹¹⁰. In addition, this region overlaps with a recently identified regulatory element upstream of HDAC6 that was shown to strongly control GATA1 expression in myeloid K562 cells¹¹¹. Interestingly, GATA1 protein was significantly decreased in platelets from the proband and from the mother (**Figure 4c–d**). GATA1 is an important transcription factor that regulates platelet formation, hemizygous mutations of which result in variable degrees of macrothrombocytopenia and dyserythropoiesis or anemia¹¹².

The proband had an elevated mean platelet volume (MPV) and an elevated platelet distribution width (PDW) and his mother had a normal MPV but an elevated PDW (**Extended Data Figure 9b**). Electron microscopy (EM) analysis of platelets from the proband showed his platelets were larger and rounder than usual and had a lower than usual density of alpha granules, which is typical of the platelets of GATA1 deficient patients¹¹³ (**Extended Data Figure 9e–g**). All the red blood cell parameters (red blood cell count, haemoglobin concentration, haematocrit, mean corpuscular haemoglobin, mean corpuscular haemoglobin concentration and red cell distribution width) of the proband were normal, although there were mild signs of dyserythropoiesis. Blood smear analysis revealed the presence of anisocytosis of red blood cells (+), moderate poikilocytosis (++) and slight hypochromasia (+). An elevated LDH level was detected (404 U/L; Normal values < 300), which has been described previously in GATA1 deficient patients¹¹⁴. The red blood cell parameters of the parents of the patient were normal.

HDAC6 deficiency has never been described in humans. Hdac6 knockout mice have no gross defects¹¹⁵ except for altered emotional behaviour¹¹⁶ and enhanced platelet spreading due to hyper-acetylated microtubules. However, their bleeding tendency has not been evaluated¹¹⁷. Structured illumination microscopy (SIM) analysis of acetylated tubulin in combination with F-actin (**Extended Data Figure 9h**) was performed on platelets of the proband under basal and activated conditions. Non-activated platelets showed disturbed marginal bands which were hyperacetylated (**Extended Data Figure 9i**). Quantification of platelet spreading on fibrinogen showed enhanced spreading for the proband while platelets from the parents were similar to those of the control (**Extended Data Figure 9j**). Hdac6 knockout mice have normal platelet counts and exhibit normal megakaryopoiesis. This is in contrast to the results of *in vitro* studies of human MKs depleted of HDAC6 using shRNA or the HDAC6 inhibitor Ricolinostat, which resulted in defective proplatelet formation¹¹⁸. We differentiated peripheral haematopoietic stem cells from the proband, his mother and an unrelated control and differentiated them to MKs. We observed a significant reduction in proplatelet-forming MKs amongst the MKs of the proband and only a mild reduction in the MKs of the mother (**Extended Data Figure 9k**). All MKs from the patient and some cells from the mother showed absence of HDAC6 (**Extended Data Figure 9l**). In contrast to MKs treated with HDAC6 inhibitors, we detected obvious microtubule defects in the MKs of the patient and the mother (**Extended Data Figure 9m**).

Proplatelet formation defects have also been observed in the MKs of GATA1 deficient patients¹¹³. Consequently, it is difficult to distinguish the contributions of the GATA1 and HDAC6 deficiencies to the observed MK defect. In contrast, the platelet phenotypes observed in the proband combine known features of GATA1 and HDAC6 deficiencies:

- low *GATA1* expression leads to dyserythropoiesis, macrothrombocytopenia and reduced density of platelet alpha granules, and
- HDAC6 deficiency causes hyper-acetylation of microtubules, resulting in enhanced platelet spreading.

Details on the functional analysis of the GATA1 enhancer/HDAC6 deletion

PCR and Sanger sequencing to validate the *HDAC6* deletion. Genomic DNA was extracted from peripheral leukocytes. PCR was performed with primers flanking the deletion (*HDAC6*-F: 5'-catcttcaagaggatcagagg and *HDAC6*-R: 5'-catagctagacactgggt), generating a PCR fragment of 358bp when the deletion is present. Sanger sequencing of PCR fragments was performed using the same primer sets.

Antibodies. The following antibodies were used: rabbit HDAC6 (clone D2E5, Cell Signaling technology, Danvers, MA, USA), mouse anti-acetylated tubulin antibody (clone 6-11B-1, Sigma, St Louis, MO, USA), mouse anti-alpha-tubulin (A11126, Thermo Fisher Scientific, Waltham, MA, USA), rabbit VWF (Dako, Agilent Technologies, Leuven, Belgium), mouse CD63 and rat GATA1 N6 (Santa Cruz Biotechnology, Dallas, TX, USA), rabbit GATA1 (NF that was produced against recombinant N-terminal zinc finger¹¹⁹), rabbit GAPDH (14C10, Cell Signaling) and integrin beta3 (sc-14009, Santa Cruz Biotechnology).

Electron and fluorescent microscopy of platelets and immunoblot analysis. Electron microscopy of platelets from the proband was performed as described¹¹². Immunostaining of resting and fibrinogen spread platelets was performed for platelets from the proband, parents and an unrelated healthy control as previously described¹²⁰. Platelet imaging was performed using a structured illumination microscope (SIM, Elyra S.1, Zeiss, Heidelberg, Germany). Images were analysed with ZEN Black (Zeiss, Heidelberg, Germany). Images were analysed with ImageJ software (National Institutes of Health, Bethesda, MD, USA) using the ‘Analyze Particles’ plugin for automated analysis. Total protein lysates were obtained from platelets as described in a previous publication¹²¹. Protein fractions were resolved by SDS–polyacrylamide gel electrophoresis, and blots were incubated with the indicated antibodies. Membranes were next incubated with HRP-conjugated secondary antibody, and staining was performed with the ECL detection reagent (Life Technologies). Chemiluminescent blots were imaged with the ChemiDoc MP imager, and the ImageLab software version 4.1 (Bio-Rad) was used for image acquisition.

Hematopoietic stem cell differentiation assay. CD34+ hematopoietic stem cells (HSC) were isolated by magnetic cell sorting (Miltenyi Biotec) from peripheral blood from the proband, his parents and an unrelated control. The recovered (differentiation day 0) CD34+ stem cells were cultured in StemSpan SFEM medium with StemSpan CC100 ensuring strong expansion of HSC for 3 days (Stem Cell Technologies, Vancouver, Canada). Differentiation was initiated by adding 50ng/ml thrombopoietin (TPO), 25ng/ml stem cell factor and 10ng/ml interleukin 1 β (Peprotech, Rocky Hill, NJ, USA). Proplatelet formation counting (on total differentiation day 12) was performed as described in previous publications^{85,120}. MKs were seeded for 4 hours on fibrinogen-coated coverslips and stained cells were photographed at 63x magnification with a confocal microscope (AxioObserver.Z1, Zeiss, Heidelberg, DE) for immunostaining.

Statistical analysis. For each of the three N6 and NF experiments, we modelled GATA1 expression (y_{ir}) for technical replicate r on individual $i \in \{F, M, P\}$, corresponding to the father, mother and proband, respectively, as:

$$\log y_{ir} = \alpha + 1_{i=F} \beta + 1_{i=M} \gamma + 1_{i=P} \xi + \mu_r + \delta_i + \epsilon_{ir},$$

with priors:

$$\begin{aligned} \alpha &\sim N(0, 1000), \beta \sim N(0, 1000), \gamma \sim N(0, 1000), \xi \sim N(0, 1000), \mu_r \sim N(0, 50), \\ \delta_i &\sim N(0, \psi^2), \epsilon_{ir} \sim N(0, \sigma^2), \\ \psi^2 &\sim \text{Inv-Gamma}(\text{shape} = 10^{-12}, \text{scale} = 10^{-6}), \\ \sigma^2 &\sim \text{Inv-Gamma}(\text{shape} = 10^{-12}, \text{scale} = 10^{-6}). \end{aligned}$$

We performed posterior inference using RStan¹²² after integrating out δ_i, ϵ_{ir} . The posterior predictive densities of $\exp(E \log y_{ir} | \mu = 0)$ for each i are shown as violin plots in **Figure 4d**. For N6-measured expression, the 90% credible intervals of $\exp(\beta), \exp(\gamma), \exp(\xi)$, were respectively (0.86, 1.45), (0.35, 0.59) and (0.37, 0.62); correspondingly, for NF-measured expression the 90% credible intervals were (0.80, 1.05), (0.51, 0.67) and (0.45, 0.60).

We analysed electron microscopy images from the patient P (n=20 images) and 4 controls (n=21, 14, 21 and 20 images). Platelet area ($(\mu\text{m})^2$) and the alpha granule count per unit area ($1/(\mu\text{m})^2$) were computed using ImageJ. For each of the platelet area/granule density outcomes, we modelled the j th observation y_j as:

$$\log y_j = \alpha + 1_{i(j)=P} \beta + \delta_{i(j)} + \varepsilon_j ,$$

with priors:

$$\begin{aligned} \alpha &\sim N(0, 1000), \quad \beta \sim N(0, 1000), \\ \delta_i &\sim N(0, \psi^2), \quad \varepsilon_j \sim N(0, \sigma^2), \\ \psi^2 &\sim \text{Inv-Gamma}(\text{shape} = 10^{-12}, \text{scale} = 10^{-6}), \\ \sigma^2 &\sim \text{Inv-Gamma}(\text{shape} = 10^{-12}, \text{scale} = 10^{-6}). \end{aligned}$$

We performed posterior inference using RStan after integrating out δ_i, ε_j . Posterior predictive densities of $\exp(E \log y_j)$ for each i are shown as violin plots in **Extended Data Figure 9f–g**. The 90% credible intervals for $\exp(\beta)$ were (1.38, 2.03) and (0.15, 0.87) for area and granule density respectively.

For each individual, we computed the mean ImageJ-quantified platelet area in groups of 5 images of F-actin stained platelets at three time points (0, 30 and 60 minutes after spreading on fibrinogen). There was no evidence of a difference in the mean of the mean platelet area in F and M compared to C within time points ($P > 0.12$ for all six two-sided Welch t -tests), so F and M were treated as controls in subsequent modelling. We standardised the data for time points 30 and 60 by dividing the observations for those time points by the geometric mean for time point zero in the corresponding individual. For each of time points 30 and 60 independently, we modelled the standardised outcome for image j of individual $i(j)$ as:

$$\log y_j = \alpha + 1_{i(j)=P} \beta + \delta_{i(j)} + \varepsilon_j ,$$

with priors:

$$\begin{aligned} \alpha &\sim N(0, 1000), \quad \beta \sim N(0, 1000), \\ \delta_i &\sim N(0, \psi^2), \quad \varepsilon_j \sim N(0, \sigma^2), \\ \psi^2 &\sim \text{Inv-Gamma}(\text{shape} = 10^{-12}, \text{scale} = 10^{-6}), \\ \sigma^2 &\sim \text{Inv-Gamma}(\text{shape} = 10^{-12}, \text{scale} = 10^{-6}). \end{aligned}$$

We performed posterior inference using RStan after integrating out δ_i, ε_j . Posterior predictive densities of $\exp(E \log y_j)$ (adjusted to match the unstandardised data) for each i are shown as violin plots in **Extended Data Figure 9j**. The 90% credible intervals for $\exp(\beta)$ were (1.87, 4.56) and (2.07, 3.61) for time points 30 and 60, respectively.

We computed the percentage of proplatelet-forming MKs in 20 images for each individual. There was no evidence of a difference in the mean of the percentage between F and C ($P=0.90$, two-sided Welch t -test), so F was treated as a control in subsequent modelling. After setting

zero values to the minimum non-zero value, we modelled y_j , the % proplatelet-forming MKs in image j of individual $i(j)$, as:

$$\text{logit } y_j = \alpha + 1_{i(j)=M} \beta + 1_{i(j)=P} \gamma + \delta_{i(j)} + \varepsilon_j,$$

with priors:

$$\begin{aligned} \alpha &\sim N(0, 1000), \beta \sim N(0, 1000), \gamma \sim N(0, 1000), \\ \delta_i &\sim N(0, \psi^2), \varepsilon_j \sim N(0, \sigma^2), \\ \psi^2 &\sim \text{Inv-Gamma}(\text{shape} = 10^{-12}, \text{scale} = 10^{-6}), \\ \sigma^2 &\sim \text{Inv-Gamma}(\text{shape} = 10^{-12}, \text{scale} = 10^{-6}). \end{aligned}$$

We performed posterior inference using RStan after integrating out δ_i, ε_j . Posterior predictive densities of $\text{expit}(E \text{ logit } y_j)$ for each i are shown as violin plots in **Extended Data Figure 9k**.

The 90% credible intervals for $\text{expit}(\beta)$ and $\text{expit}(\gamma)$ were (0.32, 0.46) and (0.18, 0.28), respectively.

Homozygous deletion of CTCF binding sites in the first intron of LRBA

LRBA is a member of the family of genes encoding BEACH-domain containing proteins and it has recently been identified as a novel diagnostic-grade gene for the PID domain¹²³. Homozygous coding loss-of-function mutations in *LRBA* cause a syndrome characterised by early onset hypogammaglobulinemia and autoimmunity. We identified an unresolved PID case carrying a homozygous deletion of a CTCF binding site in an element proximal to the *LRBA* promoter. The patient presented with a mild pancytopenia, characterised by neutropenia and autoimmune haemolytic anaemia, occasionally complicated by periods of thrombocytopenia. The clinical features of this PID case are compatible with reduced *LRBA* function and it is thus plausible that the deleted CTCF-binding element is causally implicated in the patient's pathologies.

Identifying causal SNVs in elements regulating diagnostic-grade genes

We sought to identify rare disease-causing non-coding SNVs. To achieve this, we examined rare SNVs with a CADD phred score > 20 overlapping a regulatory element of a diagnostic-grade gene in which variation is known to cause a recessive disorder. Both the cell type in which the element was called and the patient had to be labelled with the same domain (**Supplementary Table 3**). Furthermore, the patient had to carry a high-impact rare allele such as a deletion or premature stop in the body of the diagnostic-grade gene. Application of this procedure yielded two candidate SNVs in elements targeting *AP3B1* and *MPL*. The SNV in the *MPL* element was followed up for further analysis.

An SNV in the promoter of MPL, combined with deletion of exon 10 of MPL

MPL encodes the receptor for the MK growth and development factor thrombopoietin¹²⁴. Homozygote or compound heterozygote coding loss-of-function mutations of *MPL* cause chronic amegakaryocytic thrombocytopenia (CAMT)¹²⁵. CAMT is differentiated as 'type 1' or 'type 2' according to the severity of the thrombocytopenia and of the ensuing bone marrow

aplasia. The bioinformatic approach outlined above highlighted a thrombocytopenic 10-year-old male carrying a single exon heterozygous deletion of *MPL* (chr1:43,814,723-43,815,177) and a heterozygous SNV with a CADD score of 21.8 that is absent from gnomAD. The SNV lies in a strong MK-specific regulatory element (**Extended Data Figure 10a**). Motif analysis using MatInspector¹²⁶ predicted binding of HIF1 to the wild type sequence (GGAC**G**TGGGGCT) through the well-characterised recognition site “RCGTG”, but not to the mutant sequence (GGAC**A**TGGGGCT).

The patient presented in the first months of life with a rash and a platelet count of $45 \times 10^9/L$ (reference range $150\text{--}450 \times 10^9/L$). This low count was thought to be secondary to viral infection at the time it was measured and no further investigations were performed. At the age of 4 years, however, a full blood count (FBC) following a consultation for his attention deficit hyperactivity disorder, which at that time was assumed to be due to a delivery-related trauma, revealed that the patient's thrombocytopenia was chronic. A bone marrow aspirate showed a lower than usual number of MKs, adequate erythroid precursor cells, plentiful myeloid precursor cells and no signs of myelofibrosis. A clinical diagnosis of a CAMT-like condition was made but it could not be genetically confirmed because a second coding allele in *MPL* could not be found. The mother of the proband carried the large deletion (**Extended Data Figure 10b-d**). She and the father are healthy and have FBC results within normal ranges. The *MPL* regulatory SNV (chr1:43803414 G>A) is in trans of the large deletion because it is absent in the mother and therefore was inherited from the father or is a *de novo* variant.

A luciferase reporter assay was performed to characterise the activity of an *MPL* promoter fragment (chr1:43,803,336-43,803,488) containing the wild type G allele (MPL-SNV-G) relative to that containing the variant A allele (MPL-SNV-A). The presence of the wild type *MPL* promoter fragment resulted in enhanced luciferase expression that was approximately 50% reduced in the presence of the A allele (**Extended Data Figure 10e**). Measurement of the abundance of *MPL* on platelets of the proband and of the mother by flow cytometry using a specific monoclonal antibody showed markedly reduced levels in the proband compared to levels in controls and in the mother (**Extended Data Figure 10f**).

In conclusion, absence of the *MPL* protein due to coding loss-of-function variants on both *MPL* alleles causes type 1 CAMT. The case reviewed above had a chronic thrombocytopenia but the other blood cell lineages seemed unaffected. This clinical phenotype is compatible with the hypothesis that the residual abundance of *MPL* in this case is sufficient to prevent haematopoietic stem cell exhaustion, which is the hallmark of type 1 CAMT.

Details of the luciferase reporter assay

The *MPL* promoter fragments chr1:43,803,336–43,803,488 containing the G- and A-allele (MPL-SNV-G and MPL-SNV-A, respectively), were cloned in pGL3-luciferase (Promega, Madison, WI, USA). Co-transfection assays in K562 cells were performed with pGL3-empty, pGL3-MPL-SNV-G or pGL3-MPL-SNV-A in combination with pEGFP (Clontech, Mountain View, CA, USA) using the Amaxa electroporation system according to the manufacturer's instructions

(method X-01; Lonza AG, Cologne, Germany). Cell extracts were prepared after 48 hours using 1x reporter lysis buffer and 40µl of cell extract was mixed with 50µl of luciferase assay reagent (Promega, Madison, WI, USA) to determine luciferase activity in a luminescence counter (EG&G Berthold). Each plasmid was assayed in four separate transfection experiments. The firefly luciferase activity was standardised with GFP expression as an internal control. The data were analysed using GraphPad Prism 7.

Alternative variant datasets for versions 37 and 38 of the human reference genome

In order to migrate the 13,187 whole genomes to the human genome reference GRCh38, the Genalix high performance NGS secondary analysis suite¹²⁷ was deployed. The sequencing reads were extracted from the BAM files delivered by Illumina and for comparison mapped against both assemblies, GRCh37 and GRCh38, using Genalix Map (v2.5) using default parameters¹²⁸. The mapped reads were stored in a proprietary file format, called Genalix Aligned Reads (GAR). SNVs and indels were subsequently called using the Genalix Population calling tool in single sample mode. The variants were collated together in a single Genalix Variant Map (GVM) with blocks of reference matching positions and quality metrics (i.e. calling quality, genotype likelihoods, etc.). All genomes were processed against each of the assemblies in fewer than 14 days (read mapping: 12 days; genotyping: 2 days) using 10 compute nodes (Intel(R) Xeon(R) Gold 6142 CPU, 32x 2.60GHz). Variants were then exported in AVRO and standard VCF file format for variant annotation and further downstream analysis.

Appendix

Appendix 1: Neuropathic pain disorders

Neuropathic pain arises as a consequence of a disease or lesion in the somatosensory nervous system¹²⁹. A number of extreme neuropathic pain phenotypes, caused by rare high-impact mutations, have recently been described¹³⁰. Identification of such mutations has implications for diagnosis, genetic counselling and in some cases personalised treatment¹³¹. In broader terms, such mutations help us understand the pathophysiology of neuropathic pain with implications for more common acquired neuropathic pain disorders, such as painful diabetic neuropathy¹³². Loss of sensation can be caused by inherited sensory neuropathies with sensory loss restricted to pain (congenital insensitivity to pain), but it may also include large fibre modalities such as touch (hereditary sensory neuropathy). Mutations in ion channels are increasingly recognised as causes of functional disorders of somatosensation¹³⁰. For instance, homozygous loss-of-function mutations in *SCN9A*, the gene that encodes the sodium channel (Na_v1.7), have been shown to cause congenital insensitivity to pain¹³³. Conversely, heterozygous gain-of-function variants are associated with a number of inherited pain disorders that include inherited erythromelalgia (IEM)¹³⁴ and paroxysmal extreme pain disorder (PEPD)¹³⁵. *SCN9a* variants have also been linked to idiopathic small fibre neuropathy¹³⁶. Some of these variants are relatively common in the general population and their penetrance may depend on environmental factors.

Our goals were to aid genetic diagnosis of patients with NPD within the UK, to determine the prevalence of known mutations associated with NPD in relation to distinct clinical presentations and finally to discover novel mutations causing NPD. The aim of our study was to determine

whether genes previously implicated in neuropathic pain caused their clinical presentation. We recruited singleton individuals with extreme neuropathic pain phenotypes (both sensory loss and gain), all within the UK, from secondary care clinics located in Oxford, London, Salford, and Newcastle. We included participants older than 18 years of age with a proven history of life-style altering sensory disorder (either pain or loss of sensation) for greater than three months. The criteria for case definition for different clinical presentations are shown in **Supplementary Table 1 (NPD Criteria – Diagnostic criteria)**. We excluded patients with a known underlying genetic cause of chronic pain, e.g. Fabry's disease and SCN9A congenital erythromelalgia, although genetic pre-screening for these disorders was not mandatory. Patients with a learning disorder or/and autistic features sufficient to prevent the giving of consent or participation in additional pain phenotyping were also excluded. The outcome measures used for patient phenotyping are shown in **Supplementary Table 1 (NPD Criteria – Outcome measures)**. The Neuropathic Pain Special Interest Group (NeuPSIG) of the International Association for the Study of Pain (IASP)'s grading for neuropathic pain⁶ was used to grade neuropathic pain for all study participants recruited.

A total of 193 study participants with WGS data underwent neuropathic pain grading. We excluded five participants because they were unaffected family members and one participant for whom phenotypic data were unavailable.

1. **No neuropathic pain** – eight (4.1%) participants did not report neuropathic pain.
2. **Neuropathic pain unlikely** – one (0.5%) participant had a history and pain distribution which was not consistent with neuropathic pain.
3. **Possible neuropathic pain** – 11 (5.7%) participants reported an appropriate history of a relevant lesion or disease, AND pain with a distinct neuroanatomically plausible distribution.
4. **Probable neuropathic pain** – 56 (29.0%) participants satisfied criteria for possible neuropathic pain AND had clinical signs in the neuroanatomical distribution of neuropathic pain.
5. **Definite neuropathic pain** – 111 (57.5%) participants satisfied criteria for probable neuropathic pain AND a diagnostic test confirmed a lesion of the somatosensory nervous system.

Appendix 2: Extreme red cell traits in UK Biobank

Selection of participants for sequencing. The UK Biobank is a biomedical cohort of approximately half a million participants recruited in the UK between 2006 and 2010¹. The participants, 54% of whom are women, were aged between 37 and 73 years at their date of recruitment. Each participant underwent a baseline assessment at one of 21 centres distributed across Great Britain, during which 4 ml of EDTA treated peripheral whole blood was collected for FBC analysis¹². These blood samples were stored at 4°C and transported overnight in temperature controlled shipping boxes to the UK Biocentre laboratory in Stockport, Greater Manchester, UK, where FBCs were measured using a bank of four Beckman Coulter LH-700 instruments.

A GWAS of FBC traits measured on approximately 173,000 of the European ancestry participants in the UK Biobank cohort (n~133,000) and the INTERVAL study (n~40,000)⁷¹ has previously identified 582 genetic variants independently associated with quantitative properties of mature red blood cells⁹⁷. Most of these variants were common, with only 40 having an in sample MAF lower than 1%. The GWAS design had limited power to identify associations with rare variants for three reasons. Firstly, it relied on the imputation of rare alleles from the UK10K/1000 Genomes reference panels^{29,47}, which are too small to contain a large proportion of the rare haplotypes carried by the hundreds of thousands of GWAS participants. Secondly, in an attempt to inhibit spurious associations due to statistical model misspecification, participants with extreme phenotype data were deliberately excluded from the association analyses. A genetic association with a rare variant can only be detected with high probability if its effect size is large, which implies that carriers of rare alleles exhibiting detectable associations were more likely than typical study participants to have been excluded from the GWAS analyses. Thirdly, the GWAS relied on univariable genetic analyses to identify allelic associations, and these can have less power than methods such as BeviMed, which are able to model jointly the association of multiple rare variants in a DNA sequence element⁹⁵.

We sought to identify a subset of UK Biobank participants likely to carry rare alleles modulating properties of peripheral red blood cells, which could plausibly be identified by WGS. Our method was to construct a univariable composite quantitative phenotype from the UK Biobank baseline mature red cell FBC measurements, which we thought likely to have high rare-variant heritability. We then selected participants for sequencing from each tail of the distribution of the phenotype.

To construct the phenotype, we used the 65 variants with MAF < 1% that were reported to be significantly ($P < 8.31 \times 10^{-9}$) associated with at least one of 12 quantitative properties of (mature or immature) red cells in ref. 97, as a model for the likely effect of rare alleles on the baseline UK Biobank FBC. **Figure 3a** shows the pairwise relationships between the estimated effect sizes of these variants on the red cell FBC parameters MCV, RBC#, HGB and RDW. This subset of four parameters is minimal in the sense that the other mature red blood cell FBC parameters can be calculated deterministically from it. The estimated effect sizes were reported in ref. 97 as per allele additive differences in the mean of the rank-inverse standard unit normalised trait and are therefore given in units which are comparable across traits. In general, MCV and RBC# exhibit a greater range and variance in absolute effect size than HGB and RDW. Astle *et al.*⁹⁷ reported that, of all the traits they studied, MCV has the highest estimated common variant heritability and that it yielded the second largest number of associated variants with MAF < 1%. It seems reasonable to conjecture from this, that MCV also has a relatively high *rare* variant heritability. There is a strong inverse correlation between the effect sizes of alleles perturbing MCV and RBC#, suggesting that the effect sizes may measure aspects of the same underlying biological mechanism, perhaps the control of the total blood volume proportion of red cells (haematocrit). Since we could not identify any other precise systematic relationship between aspects of the joint distribution of the four estimated red cell trait effect sizes, we decided to restrict our attention to the marginal joint distribution of MCV and RBC# effect sizes

(highlighted by the red square, **Figure 3a**). We used Deming regression to estimate the approximate linear relationship:

$$\beta_{MCV} = -1.69 \times \beta_{RBC\#} \quad (1)$$

between the effect sizes for the two traits. This linear relationship is shown by a red line in **Figure 3a**.

We took the baseline UK Biobank MCV and RBC# parameters and adjusted them to remove the effect of various sources of technical and biological variation. A detailed description of the adjustments can be found in the STAR methods section of ref. 97. In brief, the phenotypes were firstly adjusted to remove differences between instruments, to remove time dependent instrument drifts and to remove the effect of delay time between venepuncture and measurement. Data acquired on days where the instrument mean was an outlier for the corresponding trait were removed. Subsequently, participants who had had a hysterectomy or who had a self-report or medical history containing a record of myelofibrosis, lymphoma, leukemia, malignant lymphoma, multiple myeloma, multiple myelofibrosis or myelodysplasia, chronic lymphocytic leukemia, chronic myeloid leukemia, acute myeloid leukemia, polycythemia vera, polycythemia, a myeloproliferative disorder, essential thrombocytosis, a haematological cancer histology report, an unspecified lymphatic or general haematological neoplasm, a myelodysplastic syndrome, or an unspecified heme malignancy, monoclonal gammopathy, an unspecified hereditary haematological disorder, haemochromatosis, thalassaemia, haemophilia, sickle cell anaemia, neutropenia, lymphopenia or pancytopenia were excluded from analysis. Finally, the phenotypes were adjusted in a second stage to remove the effects of sex, age, menopause status, the interaction of sex and menopause status with age, height, weight and the effects of historical and current habits of smoking and alcohol consumption.

We excluded all participants who did not self report their ancestry as one of “British”, “White”, “Irish” or “Any other white background” in the UK Biobank baseline assessment questionnaire. We also excluded individuals whose genotypes appeared in the interim 2015 genetic data release and who were identified as having non-European ancestry by the principal components approach reported in ref. 97.

We defined the quantitative selection phenotype Q_i for the i th UK Biobank participant as the (signed) Euclidean distance in R^2 between the origin and the point $P(f_i, g_i)$, where P is the orthogonal projection onto the line:

$$g = -1.69 \times f, \quad (2)$$

where $f_i = f(RBC\#_i)$ and $g_i = g(MCV_i)$ for functions f and g that Box-Cox transform, standardise and centre the technically and biologically adjusted traits RBC# and MCV respectively, in the UK Biobank participants not hitherto excluded. In ref. 97 the covariate adjusted traits were rank inverse normalised against $N(0,1)$ before the genetic association analyses. However, here we preferred to work with Box-Cox transformed traits in order to adjust

the central part of the data towards a Gaussian without over-shrinking outliers, which might reduce power. The selection phenotype can be defined equivalently as:

$$Q_i \equiv -1.69 \times g(MCV_i) + f(RBC\#_i)$$

and its centred and standardised distribution in male and post-menopausal female UK Biobank participants is shown in the sub-panel bounded by a dotted line in **Figure 3b**.

We excluded pre-menopausal females as candidates for sequencing because of their high prevalence of anemia and because of the additional component of non-genetic variation in each red cell parameter that is induced by the menstrual cycle. We also excluded individuals with a UK Biobank report of an insufficient DNA quantity (less than 4.7ug) to generate a working stock of 130µl of 36ng/µl (TRINEAN measured concentration < 36ng/µl, PicoGreen measured concentration < 36ng/µl) or with a UK Biobank report of inconsistency between genetic and self-reported sex. Finally, we excluded participants with a platelet count below 75x10⁹/l, a white blood cell count below 0.5 x10⁹/l or a white blood cell count above 13x10⁹/l.

We partitioned the 316,739 male and post-menopausal female participants with a computed phenotype value into six sex specific age groups thresholding at 53.3 years and 62 years. (These age groups divide those participants with a phenotype value, including the pre-menopausal women, into three groups, each of approximately 125,000 participants). Within each of these six sex-age groups, we ranked the study participants according to the value of the phenotype. We selected a total of 384 individuals from each tail of the phenotype, stratifying the selection by sex-age group so that the final selection for each tail sampled each group in proportion to its size. This additional stratification was necessary despite the adjustment of the mean of each trait for age and sex to ensure reasonable age and sex balance in the tails. The full blood count of each selected participant was reviewed by an expert panel of haematologists to exclude any text-book non-genetic or somatic pathologies such as bone-marrow failure, polycythemia vera or essential thrombocytopenia, which might explain the extreme value of the selection phenotype. A small sample of DNA was screened by the Cambridge Blood and Stem-Cell Biobank for the *JAK2* mutation V617F, a common cause of somatic myeloproliferative disorders. Any participant failing the FBC or DNA screen was replaced by the next most extreme individual in the same sex-age subgroup.

DNA samples from a total of 416 male and 352 female UK Biobank participants were retrieved from the central sample archive and sent to Illumina for sequencing. The main panel of **Figure 3b** shows the distribution of the quantitative phenotype of the selected individuals in each tail, while **Figure 3c** is a bivariate scatterplot showing the distribution of RBC# and MCV (after adjustment for technical but not biological variation) in the two tails. Of the 768 individuals sent for WGS one individual from each tail failed Illumina sequencing quality control and two distinct individuals from the right tail failed in-house checks for DNA contamination.

Derivation of distribution of polygenic score in tail-selected participants. Assume that the mean centred, standardised quantitative phenotype Q takes an $N(0, 1)$ distribution in the population and that it may be decomposed as:

$$Q = S + U$$

where S is the polygenic score and U is the part of Q not explained by the score. The central limit theorem suggests that in sufficiently large samples the distribution of S should be $N(\alpha, \phi^2)$ for some α, ϕ^2 . Consequently U should also be normally distributed. An assumption of independence between S and U implies that $0 < \phi^2 < 1$ and that:

$$Q|S = s \sim N(s - \alpha, 1 - \phi^2)$$

Consequently,

$$\begin{aligned} p(q < \tau, s) &= \int_{-\infty}^{\tau} p(q, s) dq \\ &= p(s) \int_{-\infty}^{\tau} p(q|s) dq \\ &= \frac{1}{\phi} \varphi\left(\frac{s-\alpha}{\phi}\right) \Phi\left(\frac{\tau-s+\alpha}{\sqrt{1-\phi^2}}\right) \end{aligned}$$

where φ and Φ are respectively the PDF and CDF of $N(0, 1)$. This implies that,

$$p(s|q < \tau) = \frac{1}{\phi} \varphi\left(\frac{s-\alpha}{\phi}\right) \frac{\Phi\left(\frac{\tau-s+\alpha}{\sqrt{1-\phi^2}}\right)}{\Phi(\tau)}.$$

By a symmetrical argument,

$$p(s|q > \tau) = \frac{1}{\phi} \varphi\left(\frac{s-\alpha}{\phi}\right) \frac{1 - \Phi\left(\frac{\tau-s+\alpha}{\sqrt{1-\phi^2}}\right)}{1 - \Phi(\tau)}.$$

References

1. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015 Mar;12(3):e1001779
2. Westbury SK, Turro E, Greene D, Lentaigine C, Kelly AM, Bariana TK, et al. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. Genome Med. 2015;7(1):36
3. Beighton P, De Paepe A, Steinmann B, Tsipouras P, Wenstrup RJ. Ehlers-Danlos syndromes: revised nosology, Villefranche, 1997. Ehlers-Danlos National Foundation (USA) and Ehlers-Danlos Support Group (UK). Am J Med Genet. 1998 Apr 28;77(1):31-7
4. Carss KJ, Arno G, Erwood M, Stephens J, Sanchis-Juan A, Hull S, et al. Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. Am J Hum Genet. 2017 Jan 5;100(1):75-90
5. International rules for multiple primary cancers (ICD-0 third edition). Eur J Cancer Prev. 2005 Aug;14(4):307-8

6. Finnerup NB, Haroutounian S, Kamerman P, Baron R, Bennett DL, Bouhassira D, et al. Neuropathic pain: an updated grading system for research and clinical practice. *Pain*. 2016 Aug;157(8):1599-606
7. Hadinnapola C, Bleda M, Haimel M, Screaton N, Swift A, Dorfmüller P, et al. Phenotypic Characterization of EIF2AK4 Mutation Carriers in a Large Cohort of Patients Diagnosed Clinically With Pulmonary Arterial Hypertension. *Circulation*. 2017 Nov 21;136(21):2022-2033
8. Graf S, Haimel M, Bleda M, Hadinnapola C, Southgate L, Li W, et al. Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. *Nat Commun*. 2018 Apr 12;9(1):1416
9. Rhodes CJ, Batai K, Bleda M, Haimel M, Southgate L, Germain M, et al. Genetic determinants of risk in pulmonary arterial hypertension: international genome-wide association studies and meta-analysis. *Lancet Respir Med*. 2019 Mar;7(3):227-238
10. Tuijnenburg P, Lango Allen H, Burns SO, Greene D, Jansen MH, Staples E, et al. Loss-of-function nuclear factor kappaB subunit 1 (NFKB1) variants are the most common monogenic cause of common variable immunodeficiency in Europeans. *J Allergy Clin Immunol*. 2018 Oct;142(4):1285-1296
11. Goodship TH, Cook HT, Fakhouri F, Fervenza FC, Fremeaux-Bacchi V, Kavanagh D, et al. Atypical hemolytic uremic syndrome and C3 glomerulopathy: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. *Kidney Int*. 2017 Mar;91(3):539-551
12. Elliott P, Peakman TC. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol*. 2008 Apr;37(2):234-44
13. Office for National Statistics, 2011 Census, <https://www.ons.gov.uk/census/2011census>, 2011.
14. Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*. 2008 Nov;83(5):610-5
15. Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. *N Engl J Med*. 2018 Oct 11;379(15):1452-1462
16. BCFtools, A suite of programs for interacting with high-throughput sequencing data, <http://www.htslib.org>, 2018.
17. Apache HBase, The Hadoop database, a distributed, scalable, big data store, <https://hbase.apache.org>, 2018.
18. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet*. 2017 Jun;136(6):665-677
19. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*. 2013 Aug 15;29(16):2041-3
20. Genome Reference Consortium, GRCh37, https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/, 2009.

21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009 Aug 15;25(16):2078-9
22. Illumina, Starling (Isaac Variant Caller), https://support.illumina.com/help/BS_App_TS_Amplicon_OLH_15055858/Content/Source/Informatics/Apps/IsaacVariantCaller_appENR.htm, 2016.
23. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011 Aug 1;27(15):2156-8
24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007 Sep;81(3):559-75
25. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016 Aug 18;536(7616):285-91
26. Gudbjartsson DF, Sulem P, Helgason H, Gylfason A, Gudjonsson SA, Zink F, et al. Sequence variants from whole genome sequencing a large group of Icelanders. *Sci Data*. 2015;2:150011
27. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet*. 2012 Nov 2;91(5):839-48
28. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018 Oct;562(7726):203-209
29. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. *Nature*. 2015 Oct 1;526(7571):68-74
30. Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet*. 2016 Jan 7;98(1):127-48
31. Staples J, Nickerson DA, Below JE. Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genet Epidemiol*. 2013 Feb;37(2):136-41
32. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010 Nov 15;26(22):2867-73
33. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*. 2012 Dec 15;28(24):3326-8
34. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011 Nov 1;27(21):2987-93
35. Google, Protocol Buffers, <https://developers.google.com/protocol-buffers>, 2018.
36. Medina I et al., OpenCGA, <http://docs.opencb.org/display/opencga>, 2017.
37. The Apache Software Foundation, Apache HBase, <https://hbase.apache.org/>, 2018.
38. Handsaker B, HTSJDK, <https://raw.githubusercontent.com/samtools/htsjdk/master/src/main/java/htsjdk/tribble/util/popgen/HardyWeinbergCalculation.java>, 2009.
39. Bleda M, Tarraga J, de Maria A, Salavert F, Garcia-Alonso L, Celma M, et al. CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. *Nucleic Acids Res*. 2012 Jul;40(Web Server issue):W609-14

40. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014 Mar;46(3):310-5
41. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003 Jul 1;31(13):3812-4
42. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet.* 2013 Jan;Chapter 7:Unit7.20
43. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 2005 Jul;15(7):901-13
44. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 2005 Aug;15(8):1034-50
45. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 2010 Jan;20(1):110-21
46. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature.* 2015 Oct 1;526(7571):75-81
47. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. The UK10K project identifies rare variants in health and disease. *Nature.* 2015 Oct 1;526(7571):82-90
48. Song W, Gardner SA, Hovhannisyan H, Natalizio A, Weymouth KS, Chen W, et al. Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. *Genet Med.* 2016 Aug;18(8):850-4
49. NHLBI, Trans-Omics for Precision Medicine (TOPMed) program, <https://www.nhlbiwgs.org/>, 2018.
50. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I, Spark: Cluster Computing with Working Sets, https://www.usenix.org/legacy/event/hotcloud10/tech/full_papers/Zaharia.pdf, 2010.
51. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics.* 2014 Apr 1;30(7):1006-7
52. EMBL-EBI Ensembl FAQ, What do the different biotypes in Ensembl mean?, <https://www.ensembl.org/Help/Faq?id=468>, 2018.
53. EMBL-EBI Ensembl, Ensembl Variation - Calculated variant consequences, http://www.ensembl.org/info/genome/variation/prediction/predicted_data.html, 2018.
54. Taylor MS, Ponting CP, Copley RR. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome Res.* 2004 Apr;14(4):555-66
55. Lopez J, Coll J, Haimel M, Kandasamy S, Tarraga J, Furio-Tari P, et al. HGVA: the Human Genome Variation Archive. *Nucleic Acids Res.* 2017 Jul 3;45(W1):W189-W194
56. NBR Integrative Variant Analysis (IVA) web application, Sequence variant browser, <https://bioresource.nihr.ac.uk/wgs>, 2018.
57. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics.* 2016 Apr 15;32(8):1220-2

58. Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics*. 2016 Aug 1;32(15):2375-7
59. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*. 2007 Nov;17(11):1665-74
60. University of Washington Genome Sciences, Segmental Duplication DB, <http://humanparalogy.gs.washington.edu/build37/build37.htm>, 2009.
61. She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, et al. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*. 2004 Oct 21;431(7011):927-30
62. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011 Jan;29(1):24-6
63. MacArthur JA, Morales J, Tully RE, Astashyn A, Gil L, Bruford EA, et al. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Res*. 2014 Jan;42(Database issue):D873-8
64. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016 Jun 6;17(1):122
65. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012 Jul 6;337(6090):64-9
66. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. 2016 Jan 4;44(D1):D862-8
67. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015 Jan;43(Database issue):D789-98
68. National Center for Biotechnology Information, U.S. National Library of Medicine, PubMed, <https://www.ncbi.nlm.nih.gov/pubmed>, 2018.
69. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015 May;17(5):405-24
70. Sanchis-Juan A, Stephens J, French CE, Gleadall N, Megy K, Penkett C, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med*. 2018 Dec 7;10(1):95
71. Moore C, Sambrook J, Walker M, Tolkien Z, Kaptoge S, Allen D, et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials*. 2014 Sep 17;15:363
72. Cameron-Christie S, Wolock CJ, Groopman E, Petrovski S, Kamalakaran S, Povysil G, et al. Exome-Based Rare-Variant Analyses in CKD. *J Am Soc Nephrol*. 2019 Jun;30(6):1109-1122
73. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009 Jul 15;25(14):1754-60

74. The Broad Institute, Picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF., <http://broadinstitute.github.io/picard>, 2018.
75. Tartaglia M, Mehler EL, Goldberg R, Zampino G, Brunner HG, Kremer H, et al. Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. *Nat Genet.* 2001 Dec;29(4):465-8
76. Tartaglia M, Kalidas K, Shaw A, Song X, Musat DL, van der Burgt I, et al. PTPN11 mutations in Noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity. *Am J Hum Genet.* 2002 Jun;70(6):1555-63
77. Bierzynska A, McCarthy HJ, Soderquest K, Sen ES, Colby E, Ding WY, et al. Genomic and clinical profiling of a national nephrotic syndrome cohort advocates a precision medicine approach to disease management. *Kidney Int.* 2017 Apr;91(4):937-947
78. Sadowski CE, Lovric S, Ashraf S, Pabst WL, Gee HY, Kohl S, et al. A single-gene cause in 29.5% of cases of steroid-resistant nephrotic syndrome. *J Am Soc Nephrol.* 2015 Jun;26(6):1279-89
79. Levine AP, Chan MMY, Sadeghi-Alavijeh O, Wong EKS, Cook HT, Ashford S, et al. Large-Scale Whole-Genome Sequencing Reveals the Genetic Architecture of Primary Membranoproliferative GN and C3 Glomerulopathy. *J Am Soc Nephrol.* 2020 Feb;31(2):365-373
80. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat.* 2014 Jul;35(7):899-907
81. Larkin EK, Newman JH, Austin ED, Hemnes AR, Wheeler L, Robbins IM, et al. Longitudinal analysis casts doubt on the presence of genetic anticipation in heritable pulmonary arterial hypertension. *Am J Respir Crit Care Med.* 2012 Nov 1;186(9):892-6
82. Meyer E, Carss KJ, Rankin J, Nichols JM, Grozeva D, Joseph AP, et al. Mutations in the histone methyltransferase gene KMT2B cause complex early-onset dystonia. *Nat Genet.* 2017 Feb;49(2):223-237
83. Stritt S, Nurden P, Turro E, Greene D, Jansen SB, Westbury SK, et al. A gain-of-function variant in DIAPH1 causes dominant macrothrombocytopenia and hearing loss. *Blood.* 2016 Jun 9;127(23):2903-14
84. Westbury SK, Downes K, Burney C, Lozano ML, Obaji SG, Toh CH, et al. Phenotype description and response to thrombopoietin receptor agonist in DIAPH1-related disorder. *Blood Adv.* 2018 Sep 25;2(18):2341-2346
85. Turro E, Greene D, Wijgaerts A, Thys C, Lentaigne C, Bariana TK, et al. A dominant gain-of-function mutation in universal tyrosine kinase SRC causes thrombocytopenia, myelofibrosis, bleeding, and bone pathologies. *Sci Transl Med.* 2016 Mar 2;8(328):328ra30
86. Noetzli L, Lo RW, Lee-Sherick AB, Callaghan M, Noris P, Savoia A, et al. Germline mutations in ETV6 are associated with thrombocytopenia, red cell macrocytosis and predisposition to lymphoblastic leukemia. *Nat Genet.* 2015 May;47(5):535-538
87. Zhang MY, Churpek JE, Keel SB, Walsh T, Lee MK, Loeb KR, et al. Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy. *Nat Genet.* 2015 Feb;47(2):180-5

88. Song WJ, Sullivan MG, Legare RD, Hutchings S, Tan X, Kufrin D, et al. Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. *Nat Genet.* 1999 Oct;23(2):166-75
89. Poggi M, Canault M, Favier M, Turro E, Saultier P, Ghalloussi D, et al. Germline variants in ETV6 underlie reduced platelet formation, platelet dysfunction and increased levels of circulating CD34+ progenitors. *Haematologica.* 2017 Feb;102(2):282-294
90. Evans JD, Girerd B, Montani D, Wang XJ, Galie N, Austin ED, et al. BMPR2 mutations and survival in pulmonary arterial hypertension: an individual participant data meta-analysis. *Lancet Respir Med.* 2016 Feb;4(2):129-37
91. Hancks DC, Kazazian HH Jr. Roles for retrotransposon insertions in human disease. *Mob DNA.* 2016;7:9
92. Cartron JP. Rh-deficiency syndrome. *Lancet.* 2001 Dec;358 Suppl:S57
93. Mouro-Chanteloup I, D'Ambrosio AM, Gane P, Le Van Kim C, Raynal V, Dhermy D, et al. Cell-surface expression of RhD blood group polypeptide is posttranscriptionally regulated by the RhAG glycoprotein. *Blood.* 2002 Aug 1;100(3):1038-47
94. Kulkarni SS, Vasantha K, Gogri H, Parchure D, Madkaikar M, Ferec C, et al. First report of Rhnull individuals in the Indian population and characterization of the underlying molecular mechanisms. *Transfusion.* 2017 Aug;57(8):1944-1948
95. Greene D, Richardson S, Turro E. A Fast Association Test for Identifying Pathogenic Variants Involved in Rare Diseases. *Am J Hum Genet.* 2017 Jul 6;101(1):104-114
96. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A.* 2001 Apr 24;98(9):5116-21
97. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell.* 2016 Nov 17;167(5):1415-1429.e19
98. van der Harst P, Zhang W, Mateo Leach I, Rendon A, Verweij N, Sehmi J, et al. Seventy-five genetic loci influencing the human red blood cell. *Nature.* 2012 Dec 20;492(7429):369-75
99. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat.* 2015 Oct;36(10):928-30
100. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat.* 2015 Oct;36(10):915-21
101. Padmakumar M, Jaeken J, Ramaekers V, Lagae L, Greene D, Thys C, et al. A novel missense variant in SLC18A2 causes recessive brain monoamine vesicular transport disease and absent serotonin in platelets. *JIMD Rep.* 2019 May;47(1):9-16
102. Rilstone JJ, Alkhatier RA, Minassian BA. Brain dopamine-serotonin vesicular transport disease and its treatment. *N Engl J Med.* 2013 Feb 7;368(6):543-50
103. Ito Y, Carss KJ, Duarte ST, Hartley T, Keren B, Kurian MA, et al. De Novo Truncating Mutations in WASF1 Cause Intellectual Disability with Seizures. *Am J Hum Genet.* 2018 Jul 5;103(1):144-153

104. Dahl JP, Wang-Dunlop J, Gonzales C, Goad ME, Mark RJ, Kwak SP. Characterization of the WAVE1 knock-out mouse: implications for CNS development. *J Neurosci*. 2003 Apr 15;23(8):3343-52
105. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*. 2016 Nov 17;167(5):1369-1384.e19
106. Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*. 2008 Nov 1;24(21):2537-8
107. Blueprint Epigenome, ChIP-Seq Analysis Pipeline, http://dcc.blueprint-epigenome.eu/#/md/chip_seq_grch37, 2016.
108. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 2016 Jul 8;44(W1):W160-5
109. Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137
110. Petersen R, Lambourne JJ, Javierre BM, Grassi L, Kreuzhuber R, Ruklisa D, et al. Platelet function is modified by common sequence variation in megakaryocyte super enhancers. *Nat Commun*. 2017 Jul 13;8:16058
111. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, et al. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*. 2016 Nov 11;354(6313):769-773
112. Freson K, Devriendt K, Matthijs G, Van Hoof A, De Vos R, Thys C, et al. Platelet characteristics in patients with X-linked macrothrombocytopenia because of a novel GATA1 mutation. *Blood*. 2001 Jul 1;98(1):85-92
113. Wijgaerts A, Wittevrongel C, Thys C, Devos T, Peerlinck K, Tijssen MR, et al. The transcription factor GATA1 regulates NBEAL2 expression through a long-distance enhancer. *Haematologica*. 2017 Apr;102(4):695-706
114. Hermans C, De Waele L, Van Geet C, Freson K. Novel GATA1 mutation in residue D218 leads to macrothrombocytopenia and clinical bleeding problems. *Platelets*. 2014;25(4):305-7
115. Zhang Y, Kwon S, Yamaguchi T, Cubizolles F, Rousseaux S, Kneissel M, et al. Mice lacking histone deacetylase 6 have hyperacetylated tubulin but are viable and develop normally. *Mol Cell Biol*. 2008 Mar;28(5):1688-701
116. Fukada M, Hanai A, Nakayama A, Suzuki T, Miyata N, Rodriguiz RM, et al. Loss of deacetylation activity of Hdac6 affects emotional behavior in mice. *PLoS One*. 2012;7(2):e30924
117. Sadoul K, Wang J, Diagouraga B, Vitte AL, Buchou T, Rossini T, et al. HDAC6 controls the kinetics of platelet activation. *Blood*. 2012 Nov 15;120(20):4215-8
118. Messaoudi K, Ali A, Ishaq R, Palazzo A, Sliwa D, Bluteau O, et al. Critical role of the HDAC6-cortactin axis in human megakaryocyte maturation leading to a proplatelet-formation defect. *Nat Commun*. 2017 Nov 27;8(1):1786
119. de Waele L, Freson K, Louwette S, Thys C, Wittevrongel C, de Vos R, et al. Severe gastrointestinal bleeding and thrombocytopenia in a child with an anti-GATA1 autoantibody. *Pediatr Res*. 2010 Mar;67(3):314-9

120. Heremans J, Garcia-Perez JE, Turro E, Schlenner SM, Casteels I, Collin R, et al. Abnormal differentiation of B cells and megakaryocytes in patients with Roifman syndrome. *J Allergy Clin Immunol*. 2018 Aug;142(2):630-646
121. Di Michele M, Thys C, Waelkens E, Overbergh L, D'Hertog W, Mathieu C, et al. An integrated proteomics and genomics analysis to unravel a heterogeneous platelet secretion defect. *J Proteomics*. 2011 May 16;74(6):902-13
122. The Stan Development Team, Open-source software for facilitating statistical inference at the frontiers of applied statistics., <https://mc-stan.org/users/interfaces/rstan>, NA.
123. Lopez-Herrera G, Tampella G, Pan-Hammarstrom Q, Herholz P, Trujillo-Vargas CM, Phadwal K, et al. Deleterious mutations in LRBA are associated with a syndrome of immune deficiency and autoimmunity. *Am J Hum Genet*. 2012 Jun 8;90(6):986-1001
124. Wendling F, Maraskovsky E, Debili N, Florindo C, Teepe M, Titeux M, et al. cMpl ligand is a humoral regulator of megakaryocytopoiesis. *Nature*. 1994 Jun 16;369(6481):571-4
125. Tijssen MR, di Summa F, van den Oudenrijn S, Zwaginga JJ, van der Schoot CE, Voermans C, et al. Functional analysis of single amino-acid mutations in the thrombopoietin-receptor Mpl underlying congenital amegakaryocytic thrombocytopenia. *Br J Haematol*. 2008 Jun;141(6):808-13
126. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, et al. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*. 2005 Jul 1;21(13):2933-42
127. Genalice, Modular high performance NGS secondary analysis suite, <http://www.genalice.com/solutions/products/>, 2018.
128. Pluss M, Kopps AM, Keller I, Meienberg J, Caspar SM, Dubacher N, et al. Need for speed in accurate whole-genome data analysis: GENALICE MAP challenges BWA/GATK more than PEMapper/PECaller and Isaac. *Proc Natl Acad Sci U S A*. 2017 Oct 3;114(40):E8320-E8322
129. Treede RD, Jensen TS, Campbell JN, Cruccu G, Dostrovsky JO, Griffin JW, et al. Neuropathic pain: redefinition and a grading system for clinical and research purposes. *Neurology*. 2008 Apr 29;70(18):1630-5
130. Bennett DL, Woods CG. Painful and painless channelopathies. *Lancet Neurol*. 2014 Jun;13(6):587-99
131. Geha P, Yang Y, Estacion M, Schulman BR, Tokuno H, Apkarian AV, et al. Pharmacotherapy for Pain in a Family With Inherited Erythromelalgia Guided by Genomic Analysis and Functional Profiling. *JAMA Neurol*. 2016 Jun 1;73(6):659-67
132. Blesneac I, Themistocleous AC, Fratter C, Conrad LJ, Ramirez JD, Cox JJ, et al. Rare NaV1.7 variants associated with painful diabetic peripheral neuropathy. *Pain*. 2018 Mar;159(3):469-480
133. Cox JJ, Reimann F, Nicholas AK, Thornton G, Roberts E, Springell K, et al. An SCN9A channelopathy causes congenital inability to experience pain. *Nature*. 2006 Dec 14;444(7121):894-8
134. Yang Y, Wang Y, Li S, Xu Z, Li H, Ma L, et al. Mutations in SCN9A, encoding a sodium channel alpha subunit, in patients with primary erythromelalgia. *J Med Genet*. 2004 Mar;41(3):171-4

135. Fertleman CR, Baker MD, Parker KA, Moffatt S, Elmslie FV, Abrahamsen B, et al. SCN9A mutations in paroxysmal extreme pain disorder: allelic variants underlie distinct channel defects and phenotypes. *Neuron*. 2006 Dec 7;52(5):767-74
136. Faber CG, Hoeijmakers JG, Ahn HS, Cheng X, Han C, Choi JS, et al. Gain of function Nav1.7 mutations in idiopathic small fiber neuropathy. *Ann Neurol*. 2012 Jan;71(1):26-39