Whole-genome sequencing of rare disease patients in a national healthcare system

The NIHR BioResource, on behalf of the 100,000 Genomes Project*

*A list of authors and their affiliations appears at the end of the manuscript and a list of collaborators and their affiliations appears at the end of Supplementary Information.

Most patients with hereditary rare diseases do not receive a molecular diagnosis and the aetiological variants and mediating genes for half such disorders remain to be discovered. We implemented whole-genome sequencing (WGS) in a national healthcare system to streamline diagnosis and to discover unknown aetiological variants, in the coding and non-coding regions of the genome. In a pilot study for the 100,000 Genomes Project, we generated WGS data for 13,037 participants, of whom 9,802 had a rare disease, and provided a genetic diagnosis to 1,040 of the 7,065 patients with detailed phenotypic data. We identified 99 Mendelian associations between genes and rare diseases, of which at least 80 are confirmed aetiological. Using WGS of UK Biobank, we showed that rare alleles can explain the presence of some individuals in the tails of a quantitative red blood cell (RBC) trait [1]. Finally, we reported novel non-coding variants which cause disease through the disruption of transcription of *ARPC1B*, *GATA1*, *LRBA* and *MPL*. Our study demonstrates a synergy by using WGS for diagnosis and aetiological discovery in routine healthcare.

Hereditary rare diseases affect approximately 1 in 20 people, but only a minority of patients receive a genetic diagnosis [2]. Approximately 7,500 rare diseases are known, but only half have a resolved genetic aetiology. Even when the aetiology is known, the prospects for diagnosis are severely diminished by a fragmentary approach to phenotyping and the restriction of genetic testing to candidate genes. On average, a molecular cause is determined after three misdiagnoses and 16 physician visits over a “diagnostic odyssey” lasting more than two years [3]. However, recent developments in WGS technology mean it is now possible to perform comprehensive genetic testing systematically in an integrated national healthcare system. The large-scale implementation of WGS for diagnosis will also enable the discovery of new genetic aetiologies, through the identification of novel causal mutations in the coding and non-coding parts of the genome.

In a pilot study for the 100,000 Genomes Project supported by the National Institute for Health Research (NIHR), we have performed WGS of 13,037 individuals enrolled at 57 National Health Service (NHS) hospitals in the United Kingdom and 26 hospitals in other countries (Fig. 1a., Extended Data Fig. 1a, Supplementary Table 1) in three batches, to clinical standard (Fig. 1b). Participants were approximately balanced between sexes (Extended Data Fig. 1b) and their distribution across ethnic groups closely matched that reported in the UK census (Fig. 1c; https://www.ons.gov.uk/census/2011census). In total, 9,802 individuals (75%) were affected with

a rare disease or had an extreme measurement of a quantitative trait, 9,024 of which were probands. Each participant was assigned to one of 18 domains (Table 1): 7,388 individuals to one of 15 rare disease groups, 50 individuals to a control group, 4,835 individuals to a Genomics England Limited (GEL) group and 764 individuals to a group of UK Biobank participants with extreme red blood cell indices (Supplementary Information). The rare disease domains covered pathologies of a wide range of organ systems and each had pre-specified inclusion and exclusion criteria (Table 1, Supplementary Information). We subsequently collected detailed phenotypic information, through web-based data capture applications, in the form of Human Phenotype Ontology (HPO) terms for 13 of the rare disease domains (Fig. 2a,b, Extended Data Fig. 1c). Patients with a diversity of disease were enrolled to the GEL domain, together with healthy family members, but only the affection status of these participants was available for this study. In addition, HPO-coded phenotypes were not collected for Leber Hereditary Optic Neuropathy (LHON) and Hypertrophic Cardiomyopathy (HCM) patients. In total, 19,605 HPO terms were selected to describe patient phenotypes. Quantitative data were transcribed to HPO terms using domain-specific rules, while free text was transcribed manually.

Domain name	Acronym
Bleeding, Thrombotic and Platelet Disorders	BPD
Process Controls	CNTRL
Cerebral Small Vessel Disease	CSVD
Ehler-Danlos and Ehler-Danlos-like Syndromes	EDS
100,000 Genomes Project–Rare Diseases Pilot	GEL
Hypertrophic Cardiomyopathy	HCM
Intrahepatic Cholestasis of Pregnancy	ICP
Inherited Retinal Disorders	IRD
Leber Hereditary Optic Neuropathy	LHON
Multiple Primary Malignant Tumours	MPMT
Neurological and Developmental Disorders	NDD
Neuropathic Pain Disorders	NPD
Pulmonary Arterial Hypertension	PAH
Primary Immune Disorders	PID
Primary Membranoproliferative Glomerulonephritis	PMG
Stem cell and Myeloid Disorders	SMD
Steroid Resistant Nephrotic Syndrome	SRNS
UK Biobank – Extreme Red Cell Traits	UKBio

Table 1. Domain names and their acronyms.

Following bioinformatic quality control (QC) and data analysis (Extended Data Fig. 2–5), we identified 172,005,610 short variants, of which 157,411,228 (91.5%) were single nucleotide

variants (SNVs) and 14,594,382 (8.5%) were indels up to 50bp long. 48.6% and 40.8% of the SNVs and indels, respectively, were absent from all major variant databases (Fig. 1e). 54.8% of the variants were observed in only one family, of which 82.6% were novel. Only 9.08% of novel variants were observed in more than one unrelated individual, typically in sets of individuals with recent common ancestry (Fig. 1f). SNVs and indels common in our dataset were well represented in genetic databases but, in accordance with theory, the vast majority of the variants we observed were very rare and most were uncatalogued. We called 24,436 distinct large deletions (>50bp) by synthesising inferences from two algorithms across individuals. We also called more complicated types of structural variant, such as inversions, but evidence could not be reliably aggregated across individuals (Supplementary Information). We used the WGS data to determine that only 13 (0.1%) individuals had non-standard sex chromosomal karyotypes (Extended Data Fig. 3). Using the high quality variant calls, we inferred a wide range of bioinformatically estimated family sizes, in keeping with differences in enrolment strategies (Supplementary Information), of which most comprised singletons (Fig. 1d).

We issued clinical reports for 1,107 distinct causal variants (733 SNVs, 263 indels, 104 large deletions, 6 other structural variants) affecting 304 genes. Of those which were SNVs or indels, 299 (30.0%) were absent from the Human Gene Mutation Database (HGMD). We identified strong evidence (posterior probability (PP) > 0.75) for 99 genetic associations between rare variants and groupings of patients with similar phenotypes using the Bayesian genetic association method, BeviMed [4]. Of these 99 associations, 61 are consistent with firmly established evidence and a further 18 have been reported in the literature since 2015, either by us or by other researchers. We also showed that genetic associations with the extremes of a quantitative trait can identify genes in which mutations cause Mendelian pathologies. Finally, we used a novel method, RedPop, to call cell-type specific regulatory elements (REs) from open chromatin and histone modification data. We combined these calls with cell-type specific transcription factor binding information to identify four pathological rare non-coding variants that cause disease by disrupting the proper regulation of gene expression.

Summary of clinical findings

For each of the 15 rare disease domains, we established a list of diagnostic-grade genes (DGGs) and lists of their corresponding transcripts on the basis of the scientific literature (Supplementary Information). The number of DGGs for each domain ranged from two for Intrahepatic Cholestasis of Pregnancy (ICP) to 1,423 for Neurological and Developmental Disorders (NDD). The DGGs lists were not mutually exclusive because defects of some genes manifest as distinct pathologies compatible with the enrolment criteria of multiple domains (Fig. 2c). A set of 12 multidisciplinary teams (MDTs) with domain-specific expertise examined the rare variants observed in DGGs in the context of the HPO phenotypes of the carriers. They categorised a subset of the variants as *pathogenic* or *likely pathogenic* following standard guidelines [5] and assessed their allelic contribution to disease as *full* or *partial*. A conclusive molecular diagnosis was returned for 1,140 of the 7,065 (16.1%) patient records reviewed and those diagnoses featured 1,106 unique causal variants. One quarter of the reports featured variants in *BMPR2*, *ABCA4* and *USH2A* and a further quarter featured variants in a group of 18 DGGs. The remaining half of the clinical reports concerned variants spread across 306 DGGs, which often featured in a single report (Fig. 2d,

Extended Data Fig. 6). The diagnostic yield by domain ranged from three patients out of 184 (1.6%) for Primary Membranoproliferative Glomerulonephritis (PMG) to 391 patients out of 725 (53.9%) for Inherited Retinal Disease (IRD). The variability of diagnostic yield was attributable to heterogeneity in: phenotypic and genetic pre-screening before enrolment, the genetic architecture of diseases and prior knowledge of genetic aetiologies. However, clinical reporting was enhanced by the use of clinical-grade WGS instead of whole-exome sequencing (WES). Of the 955 SNVs and indels in clinical reports, 96 had insufficient coverage in aggregated WES data (Extended Data Fig. 7) [6]. For example, a causal SNV encoding a start loss of *HPS6* in a case with Hermansky-Pudlak syndrome was identified by WGS but not identified by WES prior to the study. Similarly, deletions spanning only a few exons or part of a single exon are not reliably called by WES and we reported 104 unique large deletions between 203bp and 16.80Mb in length (mean 786.33Kb; median 15.91Kb) [7, 8].

Our recent genetic discoveries have informed treatment decisions: 27 patients with early-onset dystonia due to variants in *KMT2B* can be treated by deep brain stimulation [9]; cases with *DIAPH1-related* macrothrombocytopenia and deafness [10] can have their platelet count restored to a safe level in a preoperative setting with Eltrombopag [11]; and a case of severe thrombocytopenia accompanied by myelofibrosis and bleeding caused by a gain-of-function variant in *SRC* [12] was cured by an allogeneic haematopoietic stem cell transplant. In addition, our diagnoses have helped stratify patient care: patients with Primary Immune Disorders (PID) due to variants in *NFKB1*, which we have shown are the commonest monogenic cause of combined variable immunodeficiency (CVID) [13], have unexplained splenomegaly and an increased risk of cancer; 27 cases from the Bleeding, Thrombotic and Platelet Disorders (BPD) domain with isolated thrombocytopenia caused by variants in *ANKRD26*, *ETV6* or *RUNX1* have an increased risk of malignancy [14, 15, 16] compared to 19 cases with benign thrombocytopenia due to variants in *ACTN1*, *CYCS* or *TUBB1* [17]; and the prognosis for patients with Pulmonary Arterial Hypertension (PAH) caused by mutations in *ATP13A3*, *AQP1*, *GDF2* and *SOX17*, genes which we have recently reported as aetiological [18], is better than the prognosis for patients with mutations in *BMPR2* [19] or *EIF2AK4* [20].

Quantitative intermediate phenotypes can contain information that is useful for understanding genetic aetiology in difficult to diagnose patients. We examined WGS read alignments for patients with complete absence of a protein encoded by a DGG but carrying an explanatory variant call on only one haplotype. Two patients with a severe unexplained bleeding disorder due to a lack of the αIIbβ3 integrin on their platelet membranes carried two different complex variants in intron 9 of *ITGB3*: a tandem repeat and an SVA retrotransposon which was not called by either of the two structural variant callers we employed, but was discernible due to an excess of improperly mapped reads (Extended Data Fig. 8a–e). The third patient had an absence of RhD and RhCE proteins on the membrane of her red cells leading to severe haemolytic anemia. This was due to a large tandem repeat in *RHAG*, which encodes the Rh-associated glycoprotein (Extended Data Fig. 8f).

Discovery of rare variants associated with rare diseases

Several cases with similar aetiologies are typically needed to make a novel discovery in rare disease genetics. Cases can be aggregated across siloed studies, using services such as Matchmaker Exchange (MME) [21]. We used MME to identify novel aetiologies for *SLC18A2* and *WASF1* (Supplementary Information). However, in the context of a study of a unified healthcare system, it is possible make discoveries by statistical analyses of large patient collections.

We applied the statistical method BeviMed [4] to identify genetic associations between gene loci and rare diseases under various modes of Mendelian inheritance (Supplementary Information). We defined a set of phenotypic tags for each domain to determine a set of case/control groupings for BeviMed. Groups of cases were assigned the same tag if their phenotypes were *a priori* judged compatible with a shared genetic aetiology of disease (Supplementary Table 3). The number of unrelated cases in each tag group ranged from three for Roifman syndrome to 1,101 for PAH. For each gene-tag pair, we compared the genotypes at rare variant sites between unrelated individuals with the tag (cases) and unrelated individuals without the tag (controls). We considered a PP of association greater than 0.75 to be strong evidence supporting a genetic aetiology. Additionally, for each analysis BeviMed inferred a conditional PP over the mode of inheritance, a conditional PP over the molecular consequence class of variants mediating disease risk (e.g. 5' UTR variants or predicted loss-of-function variants) and conditional PPs of pathogenicity for each specific variant. These quantities were used to compare established to inferred modes of inheritance and to estimate the number of cases attributable to variants in each gene [4].

We inferred strong evidence for association between 29 phenotypic tags, spanning nine domains, and 99 genes. These included 62 established DGGs, 18 DGGs discovered since 2015 [18, 22, 23, 24, 25, 26, 13, 27, 28, 29, 8, 17, 30, 31, 25, 32, 9, 33, 10] and 19 candidates requiring further investigation (Fig. 3). Thus, 80 of 99 genetic associations are confirmed. We estimated that 606.6 cases are attributable to rare variants contributing to the 80 confirmed associations, 94.8 of which were attributable to the association between variants in *BMPR2* and PAH. For one gene (*GP1BB*), the mode of inheritance inferred by BeviMed differed from that established in the literature, challenging long-held assumptions [29]. These results show that a unified analysis of standardised homogeneously collected genetic and phenotypic data from large cohorts of different rare disease domains is a powerful approach for genetic discovery.

Rare variants associated with extremes of a quantitative trait in UK Biobank

Several rare diseases (e.g. familial hypercholesterolaemia, CVID**,** thrombocytopenia, von Willebrand disease) are diagnosed and clinically characterised by reference to a quantitative trait that acts as a causal intermediate (or close proxy) for pathology and symptoms. Mutation-selection equilibrium ensures strong negative selection in the tails of heritable quantitative traits, so individuals in the extreme tails should have lower fecundity, perhaps due to greater risk of disease. We sought to identify genes likely to carry mutations causing RBC pathologies by computing a univariate quantitative summary of baseline RBC full blood count (FBC) traits in the

UK Biobank participants of European ancestry. We aimed to develop a score capturing as much rare-variant heritability as possible. To achieve this, we used the joint distribution of GWAS-estimated effect sizes for associations between variants with MAF < 1% and four mature RBC FBC traits as a model for the effect of causal rare alleles identified by WGS [34] (Fig 4a). We successfully sequenced 764 participants, 383 of which were extreme for the left tail of the score, representing a low RBC count (RBC#) and a high mean cell volume (MCV), and 381 of which were extreme for the right tail of the score, representing a high RBC# and a low MCV (Fig. 4b,c).

We treated each of the two tail groups as a set of cases in a BeviMed analysis, identifying 12 genes showing stronger evidence for association than the moderate PP threshold 0.4 (Fig. 4d). *HBB* and *TFRC* can be considered positive controls, as they are known to carry mutations causing Mendelian microcytic anaemias. Other genes, including *CUX1* and *ALG1* are biologically plausible candidates. These results (Supplementary Table 3) indicate that the analysis of quantitative extremes in apparently healthy population samples may identify medically relevant loci unidentified by GWAS for quantitative traits [34, 35].

Aetiological variants in regulatory elements

Recent statistical modelling suggests that only a small proportion of the burden of heritable neurodevelopmental disorders can be attributed to *de novo* pathogenic SNVs in non-coding elements [36]. Nevertheless, rare variants in REs are known to cause disease by disrupting transcription or translation [37, 38, 39]. We searched for aetiological variants in the REs of 246 DGGs implicated in recessive haematopoiesis-related disorders. Firstly, we defined a set of active REs we named a 'regulome' for each of six blood progenitor and mature blood cell types. We achieved this by merging transcription factor binding sites identified by ChIP-seq with genomic regions called by RedPop, a new detection method exploiting the anti-covariance of ATAC-seq and H3K27ac ChIP-seq coverage in REs (Supplementary Information). We linked the REs to genes on the basis of genomic proximity and promoter capture Hi-C data [40]. Secondly, we assigned each regulome to one or more of the BPD, PID and Stem Cell and Myeloid Diseases (SMD) domains, depending on the relevance of the corresponding cell types to these domains (Supplementary Table 3). Finally, we searched for cases carrying a rare homozygous or hemizygous deletion of an RE active in a cell type assigned to the domain of the case and which was linked to a DGG of that domain. We also searched for heterozygous deletions meeting these criteria that were in compound heterozygosity with a rare coding variant in a DGG linked to the deleted element (Fig. 5a). This explained three cases: a PID patient carrying a deletion overlapping the 5' UTR region of *ARPC1B* in compound heterozygosity with a frameshift variant in the same gene (Thaventhiran *et al*, under review), a nine year old boy with autism spectrum disorder and thormbocytopenia carrying a hemizygous deletion of a *GATA1* enhancer on the X chromosome, and a male with several autoimmune-mediated cytopenias carrying a homozygous deletion of intronic CTCF binding sites [41] of *LRBA*.

The X-linked deletion in the autistic boy (Extended Data Fig. 9a–b) removed an element regulating *GATA1* as well as exons 1-4 of *HDAC6*. He had a persistently low platelet count (52×10^9/l; normal range 150×10^9/l–450×10^9/l) and a mean platelet volume in the 99.9th percentile of the distribution for UK Biobank males (Fig. 5b) [42]. Electron microscopic imaging of his platelets showed reduced

α-granule content (Extended Data Fig. 9c–e). Culture of the child's stem cells recapitulated ineffective formation of platelets by megakaryocytes (Extended Data Fig. 9f–k). Macrothrombocytopenia, reduced α-granule content and ineffective platelet formation are all characteristic of patients with pathogenic coding mutations of *GATA1* [43,] 28082341). His platelets contained reduced GATA1 (Fig. 5g), consistent with reduced transcription due to deletion of the *GATA1* enhancer [44]. HDAC6 is the major deacetylase for removing the acetyl group from Lys40 of α-tubulin, which is located in polymerized microtubules [45]. The absence of HDAC6 in the child was accompanied by extremely high expression levels of acetylated α-tubulin in his platelets (Fig. 5e), concordant with observations of *Hdac6* knockout mice [46]. This aberrant acetylation is associated with bleeding [46] and altered emotional behaviour [47] in mice. Thus, the reduced expression of *GATA1* and the absence of HDAC6 jointly cause a new syndrome of macrothrombocytopenia accompanied by neurodevelopmental problems.

The patient with a homozygous deletion of a CTCF binding site in the first intron of *LRBA* (Fig. 5h) presented with a pancytopenia, characterised mostly by neutropenia and anaemia, and complicated by periods of thrombocytopenia. These cytopenias were mediated by autoantibodies due to a loss of tolerance for multiple autoantigens, which is characteristic of patients with reduced *LRBA* function [48].

We adapted our approach to solving cases caused by non-coding deletions to search for non-coding SNVs with a CADD [49] score > 20, in the presence of a high-impact coding variant in compound heterozygosity in the assigned DGG. This approach identified two potentially aetiological SNVs in elements assigned to *AP3B1* and *MPL*, and we studied the 10 year old male patient carrying the latter mutation in more detail. *MPL* encodes the receptor for the megakaryocyte growth factor thrombopoietin [50]. Loss of *MPL* causes chronic amegakaryocytic thrombocytopenia in humans [51] and *Mpl* knockout mice have severe thrombocytopenia [52, 53]. The SNV (chr1:43803414 G>A) was in an RE detected by RedPop, the activity of which is specific to megakaryocytes in blood cell physiology (Fig. 5i), had a CADD score of 21.8, was absent from gnomAD, and was in compound heterozygosity with a deletion of exon 10 of *MPL,* which was inherited from the patient's mother (Extended data, Fig. 10a,b,c). As a result, platelet MPL levels were significantly reduced in the patient compared to controls (Extended data, Fig. 10d), suggesting *MPL* transcription on the haplotype inherited from the father is less efficient, probably because of disruption to a binding site for the transcription factor HIF1A. In contrast to MPL-null patients [54], who are extremely thrombocytopenic because their bone marrow is almost devoid of megakaryocytes and eventually suffer haematopoietic stem cell exhaustion, this 10 year old boy had platelet counts which stabilised around 45×10^9/l and a marrow that was only moderately depleted of megakaryocytes. As the regulatory SNV does not abolish MPL completely (Extended Fig. 10c), the boy has a milder clinical phenotype than MPL-null cases.

Discussion

Before now there has been limited integration between clinical genetic testing services and aetiological studies of rare diseases. We have shown that WGS in a unified national healthcare system can tackle these two objectives concurrently (Fig. 1a). This synergy can only be achieved if sequencing data from explained cases (Fig. 2), unexplained cases and unaffected individuals

are analysed jointly and if consent to contact participants for follow-up studies has been obtained at enrolment. We have demonstrated the utility of data aggregation and sharing through the number of genetic associations we have found across a diversity of rare diseases (Fig. 3). This study follows on from large-scale whole-exome and shallow genome sequencing studies in the UK [55, 56] and has been the blueprint for the UK's 100,000 Genomes Project, which recently completed sequencing. We have initiated WGS of UK Biobank participants to study individuals with extreme values for a quantitative phenotype. Extreme trait values may be the result of measurement error, extreme polygenic loads [35] or rare genetic variation and such individuals are typically excluded from GWAS studies. We have shown that genetic associations with the tail of a quantitative distribution can identify genes mediating Mendelian pathologies in the same domain of human biology (Fig. 4). The forthcoming WGS of 0.5M UK Biobank participants provides an opportunity to study other traits following similar approaches. Finally, we have provided examples of rare variants causing disease by disrupting non-coding REs of the genome. The reliability and affordability of WGS and the availability of cell-type specific epigenetic data make the exploration of the non-coding genome (Fig. 5) a promising focus for future research in unresolved rare disorders.

References

1. 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
2. Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. Am J Hum Genet. 2017 May 4;100(5):695-705
3. Vissers LELM, van Nimwegen KJM, Schieving JH, Kamsteeg EJ, Kleefstra T, Yntema HG, et al. A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. Genet Med. 2017 Sep;19(9):1055-1063
4. 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
5. 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
6. 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
7. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A. 2015 Apr 28;112(17):5473-8
8. 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
9. 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

10. 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
11. 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
12. 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
13. 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
14. Noris P, Favier R, Alessi MC, Geddis AE, Kunishima S, Heller PG, et al. ANKRD26-related thrombocytopenia and myeloid malignancies. Blood. 2013 Sep 12;122(11):1987-9
15. 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
16. 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
17. 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
18. Gräf 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
19. 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
20. Hadinnapola C, Bleda M, Haimel M, Screaton N, Swift A, Dorfmuller 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
21. 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
22. Bariana TK, Labarque V, Heremans J, Thys C, De Reys M, Greene D, et al. Sphingolipid dysregulation due to lack of functional KDSR impairs proplatelet formation causing thrombocytopenia. Haematologica. 2018 Nov 22;
23. Berrou E, Soukaseum C, Favier R, Adam F, Elaib Z, Kauskot A, et al. A mutation of the human EPHB2 gene leads to a major platelet functional defect. Blood. 2018 Nov 8;132(19):2067-2077
24. Revel-Vilk S, Shai E, Turro E, Jahshan N, Hi-Am E, Spectre G, et al. GNE variants causing autosomal recessive macrothrombocytopenia without associated muscle wasting. Blood. 2018 Oct 25;132(17):1851-1854

25. 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
26. Hofmann I, Geer MJ, Vogtle T, Crispin A, Campagna DR, Barr A, et al. Congenital macrothrombocytopenia with focal myelofibrosis due to mutations in human G6b-B is rescued in humanized mice. Blood. 2018 Sep 27;132(13):1399-1412
27. Westbury SK, Canault M, Greene D, Bermejo E, Hanlon K, Lambert MP, et al. Expanded repertoire of RASGRP2 variants responsible for platelet dysfunction and severe bleeding. Blood. 2017 Aug 24;130(8):1026-1030
28. Pleines I, Woods J, Chappaz S, Kew V, Foad N, Ballester-Beltran J, et al. Mutations in tropomyosin 4 underlie a rare form of human macrothrombocytopenia. J Clin Invest. 2017 Mar 1;127(3):814-829
29. Sivapalaratnam S, Westbury SK, Stephens JC, Greene D, Downes K, Kelly AM, et al. Rare variants in GP1BB are responsible for autosomal dominant macrothrombocytopenia. Blood. 2017 Jan 26;129(4):520-524
30. Helbig KL, Lauerer RJ, Bahr JC, Souza IA, Myers CT, Uysal B, et al. De Novo Pathogenic Variants in CACNA1E Cause Developmental and Epileptic Encephalopathy with Contractures, Macrocephaly, and Dyskinesias. Am J Hum Genet. 2018 Nov 1;103(5):666-678
31. Fiorentino A, Yu J, Arno G, Pontikos N, Halford S, Broadgate S, et al. Novel homozygous splicing mutations in ARL2BP cause autosomal recessive retinitis pigmentosa. Mol Vis. 2018;24:603-612
32. Khan KN, El-Asrag ME, Ku CA, Holder GE, McKibbin M, Arno G, et al. Specific Alleles of CLN7/MFSD8, a Protein That Localizes to Photoreceptor Synaptic Terminals, Cause a Spectrum of Nonsyndromic Retinal Dystrophy. Invest Ophthalmol Vis Sci. 2017 Jun 1;58(7):2906-2914
33. 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
34. 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
35. Natarajan P, Peloso GM, Zekavat SM, Montasser M, Ganna A, Chaffin M, et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. Nat Commun. 2018 Aug 23;9(1):3391
36. Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. Nature. 2018 Mar 29;555(7698):611-616
37. Giardine B, Borg J, Viennas E, Pavlidis C, Moradkhani K, Joly P, et al. Updates of the HbVar database of human hemoglobin variants and thalassemia mutations. Nucleic Acids Res. 2014 Jan;42(Database issue):D1063-9
38. Albers CA, Paul DS, Schulze H, Freson K, Stephens JC, Smethurst PA, et al. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat Genet. 2012 Feb 26;44(4):435-9, S1-2
39. Maas SA, Fallon JF. Single base pair change in the long-range Sonic hedgehog limb-specific enhancer is a genetic basis for preaxial polydactyly. Dev Dyn. 2005 Feb;232(2):345-8

40. 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
41. Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet. 2014 Apr;15(4):234-46
42. Grube JW, Morgan M, Kearney KA. Using self-generated identification codes to match questionnaires in panel studies of adolescent substance use. Addict Behav. 1989;14(2):159-71
43. 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
44. 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
45. Skultetyova L, Ustinova K, Kutil Z, Novakova Z, Pavlicek J, Mikesova J, et al. Human histone deacetylase 6 shows strong preference for tubulin dimers over assembled microtubules. Sci Rep. 2017 Sep 14;7(1):11547
46. 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
47. 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
48. 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
49. 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
50. 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
51. 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
52. Gurney AL, Carver-Moore K, de Sauvage FJ, Moore MW. Thrombocytopenia in c-mpl-deficient mice. Science. 1994 Sep 2;265(5177):1445-7
53. Alexander WS, Roberts AW, Nicola NA, Li R, Metcalf D. Deficiencies in progenitor cells of multiple hematopoietic lineages and defective megakaryocytopoiesis in mice lacking the thrombopoietic receptor c-Mpl. Blood. 1996 Mar 15;87(6):2162-70
54. Ballmaier M, Germeshausen M. Congenital amegakaryocytic thrombocytopenia: clinical presentation, diagnosis, and treatment. Semin Thromb Hemost. 2011 Sep;37(6):673-81
55. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015 Mar 12;519(7542):223-8
56. 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

Methods

Enrolment, research ethics and consent The NIHR BioResource (NBR) enrolled patients with rare diseases and their close relatives as part of a pilot study for the 100,000 Genomes Project. For this study, 15 rare disease domains were approved after review by the Sequencing and Informatics Committee of the NBR. Enrolment of participants started in December 2012 and was completed in March 2017. In addition, samples from a second rare diseases pilot study, coordinated by Genomics England Ltd (GEL) are included together with a number of control samples and samples from the UK Biobank cohort [57]. The NBR–Rare Diseases study was coordinated by the University of Cambridge. Participants were recruited mainly at NHS Hospitals in the UK, but also at overseas centres (Supplementary Table 1, Extended Data Fig. 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 alternative REC-approved studies. Obtaining consent for overseas samples was the responsibility of the respective principal investigators at the enrolling hospitals. The NBR retained blank 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.

Clinical and laboratory phenotype data Staff at hospitals responsible for enrolment were provided with the eligibility criteria for their respective domains as described above in the domain descriptions. 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. Online data capture allowed for the entry of Human Phenotype Ontology (HPO) terms [58] by staff at the enrolment centre and data from paper questionnaires were transformed into HPO terms by the study coordination office. Free text entries were transformed into HPO terms where feasible. An overview of the HPO data obtained for the 15 NBR rare disease domains is depicted in Extended Data Fig. 1c,d.

DNA sequencing Samples were received as either DNA extracted from whole blood or as whole blood EDTA samples that were extracted at a central DNA extraction and QC laboratory in Cambridge. Samples were tested for adequate concentration (Picogreen), DNA degradation (gel electrophoresis) and purity (OD 260/280 quality control (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). Further sample QC was performed by Illumina Inc to ensure that the concentration of the DNA was > 30 ng/ul and that every sample generated high quality genotyping results (Illumina Infinium Human Core Exome microarray). Samples with a repeated array genotyping call rate < 0.99, high levels of cross-contamination, mismatches with the declared gender that could not be resolved by further investigation, or for which consent had been withdrawn, were excluded from WGS (n=59). The genotyping data were also used for positive sample identification and sample identity was verified before data delivery. In short 0.5 µg of the DNA sample was fragmented using Covaris LE220 (Covaris Inc., Woburn, MA, USA) to obtain an average size of 450 base pair (bp) DNA fragments. 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 final libraries were checked using the Roche LightCycler 480 II (Roche Diagnostics Corporation, Indianapolis, IN, USA) with KAPA Library Quantification Kit (Kapa Biosystems, Inc, Wilmington, MA, USA) for concentration. From February 2014 to June 2017 three read lengths were used: 100bp (377 samples), 125bp (3,154 samples) and 150bp (9,656 samples). Samples sequenced with 100bp and 125bp reads utilised three and two lanes of an Illumina HiSeq 2500 instrument, respectively. Samples sequenced with 150bp reads utilised a single lane of a HiSeq X instrument. 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.

WGS data processing pipeline The WGS data for the 13,187 samples returned by the sequencing provider underwent a series of processing steps (Extended Data Fig. 2), described in detail in the Supplementary Information. Briefly, the samples were sex karyotyped and pairwise kinship coefficients were computed. This information was used to check for repeat sample submissions and sample swaps. Additionally, four further QC checks were applied to ensure the SNVs and short insertions/deletions (indels) were of a high standard. Overall, 150 samples (1.1%) were removed, leaving a dataset of 13,037 samples for downstream analysis. The 13,037 individuals were assigned one of the following ethnicities: European, African, South Asian, East Asian or Other. Pairwise relatedness adjusted for population stratification was then computed and used to generate networks of closely related individuals and to define a maximal set of 10,259 unrelated individuals. The variants in the 13,037 individuals were left-aligned and normalised with bcftools, loaded into our HBase database and filtered on their overall pass rate (OPR), defined in Supplementary Information. The sex karyotypes, the ethnicities and the relatedness estimates were used, along with enrolment information, to annotate the samples and variants. Samples were annotated with: affected/unaffected status, membership of the set of probands, membership of the maximal unrelated set, ethnicity and sex karyotype. Variants were annotated with CellBase consequence predictions, HGMD information where available and population-specific allele frequencies.

Pertinent findings For each of the 15 rare disease domains (i.e. all domains except UKB, CNTRL and GEL) a gene list was generated by domain-specific experts. Genes were included in the lists if there was a high enough level of evidence in the literature for gene-disease association. The 2,497 gene/domain pairs, encompassing 2,073 unique 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) resource [59] or designation as canonical in Ensembl. Variants (SNVs, indels) were shortlisted if (i) their MAF in control populations [6] 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 [60] 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, (iii) the variant affected a gene relevant to the patient's disease. Variants with more than 3 alleles or a MAF >= 10% in the diseases cohort were discarded to, respectively, guard against errors in repetitive regions and remove potential systematic artefacts. 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 person to fewer than 10, while confidently retaining known regulatory or moderately common pathogenic variants. For each affected participant 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 gender, ancestry, relatedness, and consanguinity level) were transferred to Congenica Inc (Cambridge, United Kingdom) for visualisation in the Sapientia™ web application during MDT meetings. MDTs brought together experts from different hospitals across the UK and abroad, 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 followed the American College of Medical Genetics guidelines [5] and variants (V) were marked in Sapientia™ as pathogenic, likely pathogenic or of unknown significance (VUS). Only pathogenic and likely pathogenics ones were systematically reported and VUSs were reported at the MDT's discretion. As per REC-approved study protocol, secondary findings (e.g. breast cancer pathogenic variants in *BRCA1* in patients not presenting with this phenotype) were not reported.

Genetic association testing in genes We used the BeviMed statistical method [4] to identify genetic associations with rare diseases in our dataset. Each run of BeviMed requires the definition of a set of cases and controls, all of which should be unrelated with each other, and a set of rare variants to include in the inference. To achieve adequate power, the cases should be chosen such that they potentially share a common genetic aetiology (e.g. because the phenotypes are similar) and the rare variants should be chosen such that they potentially share a mechanism of action on phenotype (e.g. because they are predicted to have a similar effect on a particular gene product). BeviMed computes PP values of no association, dominant association and recessive association and, conditional on dominant or recessive association, it computes the PP that each variant is pathogenic. We can impose a prior correlation structure on the pathogenicity of the variants that reflects competing hypotheses as to which class of variant is responsible for disease. These classifications typically group variants by their predicted consequences. The class of variant responsible can then be inferred by BeviMed, thereby suggesting a particular mechanism of disease. The methodology is described in further detail in Supplementary Information and in reference [4].

Regulome analysis We applied the BLUEPRINT protocol for ChIP-seq data analysis (http://dcc.blueprint-epigenome.eu/#/md/chip_seq_grch37). We defined regulomes for activated CD4+ T cells (aCD4), B cells (B), erythroblasts (EB), megakaryocytes (MK), monocytes (MON) and resting CD4+ T cells (rCD4). For each cell type, we used open chromatin data (ATAC-seq or DNAse-seq) and histone modification data (H3K27ac) to identify REs using the RedPop method (see below). Additionally, for MK and EB, we had access to the following transcription factor (TF)

ChIP-seq data, which were used to call peaks (see below) and supplement the regulomes: FLI1, GATA1, GATA2, MEIS1, RUNX1, TAL1 and CTCF for MK; GATA1, KLF1, NFE2 and TAL1 for EB; and CTCF for MON and B. For each cell type, the regulome build process proceeded as follows: 1. Call RedPop regions using ATAC-seq/DNAse-seq and H3K27ac-seq data; 2. Call TF/CTCF 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 gene label using either gene annotations from Ensembl (v75) or a compendium of previously published promoter capture Hi-C data (pcHi-C) [40] as follows: 1. Assign to a gene if the feature overlaps the gene or the region up to 10Kb either side of the gene body; 2. Assign to a gene if the feature overlaps the gene's pcHi-C 'blind' spot. This region is defined by three *Hind*III restriction fragments, incorporating the capture fragment overlapping target gene TSS, and 5' and 3' adjacent fragments; 3. Assign to a gene if the feature overlaps a linked promoter interacting region identified using pcHi-C in the same cell type.

Functional analysis of the *GATA1* enhancer/*HDAC6* deletion The *GATA1* enhancer/*HDAC6* deletion was confirmed by PCR using primers HDAC6-F: 5'-catcttcaagaggatcagagg and HDAC6-R: 5'- catagctagacactggtt. Electron microscopy for platelets was performed as described [43]. Immunostaining of resting and fibrinogen spread platelets was performed as described [33] and analyzed by Structured Illumination Microscopy (SIM, Elyra S.1, Zeiss, Heidelberg, D.E). Total protein lysates were obtained from platelets for immunoblot analysis as described [61]. The following antibodies were used for SIM and immunoblot analysis: rabbit anti-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 anti-VWF (Dako, Aligent Technologies, Leuven, BE), mouse anti-CD63 and rat anti-GATA1 N6 (Santa Cruz Biotechnology, Dallas, TX, USA), rabbit anti-GATA1 (NF that was produced against recombinant N-terminal zinc finger [62], rabbit anti-GAPDH (14C10, Cell Signaling) and anti-β3 integrin (sc- 14009; Santa Cruz Biotechnology).

MPL expression on platelets The level of MPL protein on the platelet membrane was measured by flow cytometry (Beckman Coulter FC500) using the monoclonal antibodies: APC-labelled IgG1 against CD42b (clone HIP1, BD Pharmingen, number: 551061), PE-labelled IgG1 against CD110 (clone REA250, Miltenyi Biotec) and a PE-labelled isotype control (clone MOPC-21, BD Pharmingen, number: 555749). In short, a sample of EDTA anticoagulated blood was incubated with anti-CD110 (or control) and anti-CD42b for 30 minutes. Mean fluorescence intensity (MFI) produced by the anti-CD110 was measured by flow cytometry on cells gated on the CD42b APC signal, side and forward scatter.

Nanopore sequencing Oxford Nanopore-based sequencing of long-range PCR-amplified target DNA was performed as previously described [63] with the aim to resolve the genetic architecture of intron 9 of *ITGB3* in a case with Glanzmann's thrombasthenia. The flow cell ran for 3 hours, and the mean coverage was 863,986X.

Code availability Code to run HBASE is available from https://github.com/mh11/VILMAA. The RedPop software package is available from https://gitlab.haem.cam.ac.uk/et341/redpop/.

Data availability

Genotype and phenotype data will become available from the day that the manuscript has been published in a peer-reviewed journal. It is expected that the manuscript may have been through peer review and revision before September 2019 but this could be also at a later date. The data access procedures outlined below will NOT be active until the date of publication.

The genotype and phenotype data from the 4,835 participants enrolled in the NIHR BioResource for the 100,000 Genomes Project–Rare Diseases Pilot can be accessed by seeking access via Genomics England Limited following the procedure outlined at:
https://www.genomicsengland.co.uk/about-gecip/joining-research-community/
The genotype data for the 764 UK Biobank samples will be made available through a data release process overseen by UK Biobank (https://www.ukbiobank.ac.uk/). The phenotype data from UK Biobank participants are available from UK Biobank using their normal access procedures.
The genotype data from the vast majority of the remaining 7,438 NIHR BioResource participant has been deposited in European Genome-phenome Archive (EGA) at the EMBL European Bioinformatics Institute. Deposition of genotype at EGA is grouped by rare disease domain: EGA accession codes: BPD: EGAD00001004519, CSVD: EGAD00001004513, HCM: EGAD00001004514, ICP: EGAD00001004515, IRD: EGAD00001004520, MPMT: EGAD00001004521, NDD: EGAD00001004522, NPD: EGAD00001004516, PAH: EGAD00001004525, PID: EGAD00001004523, PMG: EGAD00001004517, SMD: EGAD00001004524, SRNS: EGAD00001004518. Genotype data will be available at the time of publication from EGA under the principles of obtaining access to data, which are controlled by a Data Access Committee (DAC) (https://www.ebi.ac.uk/ega/). Access to more detailed phenotypic datasets on the vast majority the 7,438 NIHR BioResource participants can be requested by completing the NIHR BioResource Data Access Agreement application (dac@bioresource.nihr.ac.uk). Decisions about granting access are controlled by a DAC. A DAC has the full right to not approve requests for access to data and is under no obligation to provide reasons for a refusal. The DAC will in reaching its decisions about granting right of access to data respect the agreement with the study participants as set out in their signed consent.

The ATAC-seq and H3K27ac ChIP-seq data to support the generation of the regulomes are available from GEO or EGA, or referenced to their publication as follows. H3K27ac ChIP-seq: aCD4 [64], B (ERR1043004, ERR1043129, ERR928206, ERR769436), EB (EGAD00001002377), MK (EGAD00001002362), MON (ERR829362 (ERS257420), ERR829412 (ERS222466), ERR493634 (ERS214696)), rCD4 [64]. ATAC-seq: aCD4 (accession will be available before publication), B (SRR2126769 (GSE71338)), EB (SRR5489430 (GSM2594182)), MK (EGAD00001001871), MON (accession number requested), rCD4 (GEO accession will be available before publication).

MDT-reported alleles and their clinical interpretation have been deposited with ClinVar (accession number will be available before publication) and also with DECIPHER (accession will be available before publication).

References (continued)

57. Conlon M, Murphy RF. The interaction of immobilized transition-metal ions with some gastrointestinal polypeptides. Biochem Soc Trans. 1976;4(5):860-1
58. 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
59. 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
60. 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
61. 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
62. 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
63. 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
64. Burren OS, Rubio Garcia A, Javierre BM, Rainbow DB, Cairns J, Cooper NJ, et al. Chromosome contacts in activated T cells identify autoimmune disease candidate genes. Genome Biol. 2017 Sep 4;18(1):165

Author Contributions

Corresponding author: Willem H Ouwehand[1,2,3,4,5]

Writing Group: William J Astle[1,2,4,6,7], Kathleen Freson[8], Karyn Megy[1,2], Willem H Ouwehand[1,2,3,4,5], F Lucy Raymond[2,9], Kathleen E Stirrups[1,2], Ernest Turro[1,2,6]

NIHR BioResource Principal Investigators: Timothy J Aitman[10,11], David L Bennett[12], Mark J Caulfield[13,14], Patrick F Chinnery[2,15,16], Peter H Dixon[17], Kathleen Freson[8], Daniel P Gale[18], Ania Koziell[19,20], Taco W Kuijpers[21,22], Michael A Laffan[23,24], Eamonn R Maher[9,25], Hugh S Markus[26], Nicholas Morrell[2,27], Irene Roberts[28,29,30], Kenneth G C Smith[27], Adrian J Thrasher[31], Hugh Watkins[32,33,34], Catherine Williamson[17,35], Christopher Geoffrey Woods[9,36], F Lucy Raymond[2,9], Willem H Ouwehand[1,2,3,4,5]

Ethics, Governance, Recruitment Coordination and Clinical Bioinformatics: Matthew Brown[1,2], Naomi Clements Brod[1,2], John Davis[1,2], Eleanor F Dewhurst[1,2], Marie Erwood[1,2], Amy J Frary[1,2], Rachel Linger[2,37], Jennifer Martin[2,27,37], Sofia Papadia[2,37], Crina Samarghitean[1,2], Emily

Staples[27], Catherine Titterton[1,2], Julie von Ziegenweidt[1,2], Katherine Yates[1,2,27], Ping Yu[1,2], Hannah Stark[2,37], Roger James[1,2], Sofie Ashford[2,37]

Sample and Data processing: Congenica Eugene Bragin[38], Calvin Cheah[38], Radhika Prathalingam[38], Anthony Rogers[38], Charles Steward[38], Katie Tate[38], Nick Lench[38]; **EMBL-European Bioinformatics Institute** Jeff Almeida-King[39], Aoife McMahon[39], Joannella Morales[39]; **GENALICE** Jack Findhammer[40], Tim Karten[40], Bas Tolhuis[40], Maarten Vandekuilen[40], Johannes Karten[40]; **High Performance Computing Facility, University of Cambridge** Robert Klima[41], Ignacio Medina Castello[41], Stuart Rankin[41], Wojciech Turek[41], Paul Calleja[41]; **Illumina** Christian J Bourne[42], Camilla Colombo[42], Claire Geoghegan[42], Terence S A Gerighty[42], Russell J Grocock[42], Joseph Hughes[42], Sarah Hunter[2,42], John Peden[42], Christine Rees[42], Sean Humphray[42], David R Bentley[42]; **University of Cambridge** Anthony Attwood[1,2], Abigail Crisp-Hihn[1,2], Sri V V Deevi[1,2], Karen Edwards[1,2], James Fox[1,2], Fengyuan Hu[1,2], Jennifer Jolley[1,2], Rutendo Mapeta[1,2], Stuart Meacham[1,2], Paula J Rayner-Matthews[1,2], Olga Shamardina[1,2], Ilenia Simeoni[1,2], Simon Staines[1,2], Jonathan Stephens[1,2], Salih Tuna[1,2], Christopher Watt[1,2], Deborah Whitehorn[1,2], Yvette Wood[1,2], Christopher J Penkett[1,2], Kathleen E Stirrups[1,2]

Software Development: High Performance Computing Facility, University of Cambridge Stuart Rankin[41]; **Medical Research Council (MRC) Biostatistics Unit** Sylvia Richardson[6]; **University of Cambridge** Keren Carss[1,2], Daniel Greene[1,2,6], Matthias Haimel[1,2,27], Tobias Tilley[1,2], Eliska Zlamalova[1], Ernest Turro[1,2,6], Stefan Gräf[1,2,27]

Data Analysis: William J Astle[1,2,4,6,7], Christian Babbs[28,30], Agnieszka Bierzynska[43], Marta Bleda[27], Oliver S Burren[27], Peter H Dixon[17], Courtney E French[44], Daniel Greene[1,2,6], Charaka Hadinnapola[27], Matthias Haimel[1,2,27], Adam P Levine[18], Eleni Louka[28,30], Adam J Mead[28], Karyn Megy[1,2], Monika Mozere[18], Jennifer O'Sullivan[45], Steven Okoli[28,30], David Parry[11], Beth Psaila[28,30,46], Anupama Rao[47], Omid Sadeghi-Alavijeh[18], Alba Sanchis-Juan[1,2], Katherine R Smith[13], Emilia Swietlik[27], Rhea Y Y Tan[26], Natalie van Zuydam[12], Wei Wei[15,16], James Whitworth[9,25,48], Eliska Zlamalova[1], Augusto Rendon[1,13], Keren Carss[1,2], Stefan Gräf[1,2,27], Hana Lango Allen[1,2], Ernest Turro[1,2,6]

Clinical Interpretation and Multi-Disciplinary Teams: Stephen Abbs[49], Timothy J Aitman[10,11], Philip Ancliff[47], Gavin Arno[50,51], Chiara Bacchelli[31], David L Bennett[12], Agnieszka Bierzynska[43], Siobhan O Burns[52,53], Keren Carss[1,2], Louise C Daugherty[1,2,13], Sri V V Deevi[1,2], Peter H Dixon[17], Kate Downes[1,2], Anna M Drazyk[26], Courtney E French[44], Kathleen Freson[8], Daniel P Gale[18], Kimberly C Gilmour[31,47], Keith Gomez[54,55], Detelina Grozeva[9], Charaka Hadinnapola[27], Simon Holden[56], Ania Koziell[19,20], Taco W Kuijpers[21,22], Michael A Laffan[23,24], Hana Lango Allen[1,2], D Mark Layton[23,24], Adam P Levine[18], Eleni Louka[28,30], Eamonn R Maher[9,25], Jesmeen Maimaris[31], Rutendo Mapeta[1,2], Hugh S Markus[26], Jennifer Martin[2,27,37], Sarju Mehta[56], Nicholas Morrell[2,27], Andrew D Mumford[57,58], David Parry[11], Irene Roberts[28,29,30], Noemi B Roy[28,29,30], Moin Saleem[43,59], Alba Sanchis-Juan[1,2], Sinisa Savic[60,61], Ilenia Simeoni[1,2], Emilia Swietlik[27], Rhea Y Y Tan[26], James E Thaventhiran[62], Andreas C Themistocleous[12], David Thomas[27], Marc Tischkowitz[49,63], Matthew Traylor[26], Ernest Turro[1,2,6], Natalie van Zuydam[12], Anthony M Vandersteen[64], Andrew R

Webster[50,51], James Whitworth[9,25,48], Catherine Williamson[17,35], Christopher Geoffrey Woods[9,36], Willem H Ouwehand[1,2,3,4,5], F Lucy Raymond[2,9], Stefan Gräf[1,2,27], Karyn Megy[1,2]

Non-coding Space Analysis Group: University of Cambridge Oliver S Burren[27], Luigi Grassi[1,2], Daniel Greene[1,2,6], Myrto Kostadima[1], Roman Kreuzhuber[1,2], Hana Lango Allen[1,2], Romina Petersen[1,2], Denis Seyres[1,2], James E Thaventhiran[62], Mattia Frontini[1,2,5], Ernest Turro[1,2,6]; **University of Oxford** Anthony J Cutler[65], John A Todd[65]; **Wellcome Sanger Institute** Patrick J Short[3], Matthew Hurles[3]

Functional Analysis Group: Nichola Cooper[66], Nicholas S Gleadall[1,2], Andrew D Mumford[57,58], Helen Oram[67], Alba Sanchis-Juan[1,2], Olga Shamardina[1,2], Jonathan Stephens[1,2], Patrick Thomas[1,2], Chantal Thys[8], Sarah K Westbury[57,58], Suthesh Sivapalaratnam[4,68,69,70], Kate Downes[1,2], Kathleen Freson[8]

Data Visualisation: Salih Tuna[1,2], William J Astle[1,2,4,6,7], Sri V V Deevi[1,2], Daniel Greene[1,2,6], Matthias Haimel[1,2,27], Christopher J Penkett[1,2], Alba Sanchis-Juan[1,2], Olga Shamardina[1,2], Ernest Turro[1,2,6]

Steering groups: NIHR BioResource Sequencing and Informatics Committee (SIC) Gerome Breen[71,72], John Chambers[73,74,75,76,77], Matthew Hurles[3], Nathalie Kingston[2], Mark I McCarthy[30,33,78], Nilesh Samani[79], Michael Simpson[80], Nicholas Wood[81,82], Willem H Ouwehand[1,2,3,4,5], F Lucy Raymond[2,9]; **NIHR BioResource – Rare Diseases Senior Management Team (SMT)** Sofie Ashford[2,37], Debra Fletcher[1,2], Mary A Kasanicki[36], Nathalie Kingston[2], Christopher J Penkett[1,2], Hannah Stark[2,37], Kathleen E Stirrups[1,2], Timothy Young[1,2], Roger James[1,2], F Lucy Raymond[2,9], John R Bradley[2,25,27,36,83], Willem H Ouwehand[1,2,3,4,5]

NIHR BioResource - Rare Diseases Study teams: Bleeding, thrombotic and Platelet Disorders (BPD) Tadbir Bariana[54,55], Claire Lentaigne[23,24], Suthesh Sivapalaratnam[4,68,69,70], Sarah K Westbury[57,58], David J Allsup[84,85], Tamam Bakchoul[86], Tina Biss[87], Sara Boyce[88], Janine Collins[1,68], Peter W Collins[89], Nicola S Curry[90], Kate Downes[1,2], Tina Dutt[91], Wendy N Erber[92], Gillian Evans[93], Tamara Everington[94,95], Remi Favier[96,97], Keith Gomez[54,55], Daniel Greene[1,2,6], Andreas Greinacher[98], Paolo Gresele[99], Daniel Hart[68], Rashid Kazmi[88], Anne M Kelly[36], Michele Lambert[100,101], Bella Madan[45], Sarah Mangles[95], Mary Mathias[102], Carolyn Millar[23,24], Paquita Nurden[103], Samya Obaji[104], Kathelijne Peerlinck[8], Catherine Roughley[93], Sol Schulman[105], Marie Scully[106], Susan E Shapiro[90], Keith Sibson[102], Ilenia Simeoni[1,2], Matthew C Sims[1,107], R Campbell Tait[108], Kate Talks[87], Chantal Thys[8], Cheng-Hock Toh[91], Chris Van Geet[8], John-Paul Westwood[106], Sofia Papadia[2,37], Ernest Turro[1,2,6], Andrew D Mumford[57,58], Willem H Ouwehand[1,2,3,4,5], Kathleen Freson[8], Michael A Laffan[23,24]; **Cerebral Small Vessel Disease (CSVD)** Rhea Y Y Tan[26], Julian Barwell[109,110], Kate Downes[1,2], Kirsty Harkness[111], Sarju Mehta[56], Keith W Muir[112], Ahamad Hassan[113], Matthew Traylor[26], Anna M Drazyk[26], Stefan Gräf[1,2,27], Hugh S Markus[26]; **Ehlers Danlos Syndrome (EDS)** David Parry[11], Munaza Ahmed[114], Alex Henderson[115], Hanadi Kazkaz[106], Anthony M Vandersteen[64], Timothy J Aitman[10,11]; **Hypertrophic Cardiomyopathy (HCM)** Elizabeth Ormondroyd[32,34], Kate Thomson[32,34], Timothy Dent[34], Paul Brennan[115,116,117], Rachel J Buchan[118,119], Teofila Bueser[19,120,121], Gerald Carr-White[122], Stuart

Cook[118,123,124,125], Matthew J Daniels[32,34,126], Andrew R Harper[32,33,118], Alex Henderson[115], James S Ware[118,119,123], Hugh Watkins[32,33,34]; **Intrahepatic Cholestasis of Pregnancy (ICP)** Peter H Dixon[17], Jenny Chambers[17,127], Floria Cheng[127], Maria C Estiu[128], William M Hague[129], Hanns-Ulrich Marschall[130], Marta Vazquez-Lopez[127], Catherine Williamson[17,35]; **Inherited Retinal Disorders (IRD)** Gavin Arno[50,51], Eleanor F Dewhurst[1,2], Marie Erwood[1,2], Courtney E French[44], Michel Michaelides[50,51], Anthony T Moore[50,51,131], Alba Sanchis-Juan[1,2], Keren Carss[1,2], Andrew R Webster[50,51], F Lucy Raymond[2,9]; **Leber Hereditary Optic Neuropathy (LHON)** Patrick F Chinnery[2,15,16], Philip Griffiths[132,133], Rita Horvath[134,135], Gavin Hudson[134], Neringa Jurkute[50,55], Angela Pyle[134], Wei Wei[15,16], Patrick Yu-Wai-Man[15,16,136]; **Multiple Primary Malignant Tumours (MPMT)** James Whitworth[9,25,48], Julian Adlard[137], Munaza Ahmed[114], Ruth Armstrong[9,25,48], Julian Barwell[109,110], Carole Brewer[138], Ruth Casey[9,25,48], Trevor R P Cole[139], Dafydd Gareth Evans[140], Lynn Greenhalgh[141], Helen L Hanson[142], Alex Henderson[115], Jonathan Hoffman[139], Louise Izatt[143], Ajith Kumar[114], Fiona Lalloo[144], Kai Ren Ong[139], Soo-Mi Park[25,48,49], Joan Paterson[9,25,48], Claire Searle[145], Lucy Side[146], Katie Snape[142], Emma Woodward[144], Marc Tischkowitz[49,63], Eamonn R Maher[9,25]; **Neurological and Developmental Disorders (NDD)** Keren Carss[1,2], Eleanor F Dewhurst[1,2], Marie Erwood[1,2], Courtney E French[44], Detelina Grozeva[9], Alba Sanchis-Juan[1,2], Manju A Kurian[147,148], F Lucy Raymond[2,9]; **Neuropathic Pain Disorders (NPD)** Andreas C Themistocleous[12], Iulia Blesneac[12], David Gosal[149], Rita Horvath[134,135], Andrew Marshall[150,151,152], Emma Matthews[153,154], Mark I McCarthy[30,33,78], Tara Renton[121], Andrew S C Rice[155,156], Tom Vale[12], Natalie van Zuydam[12], Suellen M Walker[31,47], Christopher Geoffrey Woods[9,36], David L Bennett[12]; **Primary Immune Disorders (PID)** James E Thaventhiran[62], Hana Lango Allen[1,2], Siobhan O Burns[52,53], Sinisa Savic[60,61], Oliver S Burren[27], Hana Alachkar[149], Richard Antrobus[157], Helen E Baxendale[27,44,158,159], Michael J Browning[160], Matthew S Buckland[161], Nichola Cooper[66], Elizabeth Drewe[162], J David M Edgar[163,164], William Egner[165], Kimberly C Gilmour[31,47], Sarah Goddard[166], Pavels Gordins[167], Sofia Grigoriadou[168], Scott Hackett[169], Rosie Hague[170], Grant Hayman[171], Archana Herwadkar[149], Aarnoud P Huissoon[169], Stephen Jolles[172], Peter Kelleher[173,174], Dinakantha Kumararatne[175], Hilary Longhurst[168], Lorena E Lorenzo[168], Paul A Lyons[27], Jesmeen Maimaris[31], Sadia Noorani[176], Alex Richter[157], Crina Samarghitean[1,2], Ravishankar B Sargur[165], W A Carrock Sewell[177], Ilenia Simeoni[1,2], Emily Staples[27], David Thomas[27], Moira J Thomas[178,179], Steven B Welch[180], Austen Worth[47], Patrick F K Yong[181], Taco W Kuijpers[21,22], Adrian J Thrasher[31], Kenneth G C Smith[27]; **Primary Membranoproliferative Glomerulonephritis (PMG)** Adam P Levine[18], Omid Sadeghi-Alavijeh[18], Edwin K S Wong[117,182], H Terence Cook[183], Melanie M Y Chan[18], Martin T Christian[184], Matthew Hall[162], Claire Harris[182], Paul McAlinden[182], Kevin J Marchbank[182,185], Stephen Marks[47], Heather Maxwell[170], Monika Mozere[18], Julie Wessels[166], MPGN/C3 Glomerulopathy Rare Renal Disease group[186], Sally A Johnson[182,187], Daniel P Gale[18]; **Pulmonary Arterial Hypertension (PAH)** Marta Bleda[27], Charaka Hadinnapola[27], Matthias Haimel[1,2,27], Emilia Swietlik[27], Harm Bogaard[188], Colin Church[189], Gerry Coghlan[161], Robin Condliffe[190], Paul Corris[182,191], Cesare Danesino[192], Mélanie Eyries[193], Henning Gall[194], Stefano Ghio[195], Hossein-Ardeschir Ghofrani[66,194], J Simon R Gibbs[118], Barbara Girerd[196,197,198], Simon Holden[56], Arjan Houweling[188], Luke S Howard[118,199], Marc Humbert[196,197,198], David G Kiely[190], Gabor Kovacs[200,201], Allan Lawrie[202], Robert V MacKenzie Ross[203], Jennifer Martin[2,27,37], Shahin Moledina[47], David Montani[196,197,198], Michael Newnham[27,159], Andrea Olschewski[200], Horst Olschewski[200,201], Andrew Peacock[189], Joanna Pepke-Zaba[159], Laura Scelsi[195], Werner Seeger[194], Florent Soubrier[193], Jay Suntharalingam[203], Mark Toshner[27,159],

Carmen Treacy[27,159], Richard Trembath[19], Anton Vonk Noordegraaf[188], Quinten Waisfisz[204], John Wharton[66], Martin R Wilkins[66], Stephen J Wort[119,205], Katherine Yates[1,2,27], Stefan Gräf[1,2,27], Nicholas Morrell[2,27]; **Stem cell and Myeloid Disorders (SMD)** Eleni Louka[28,30], Noemi B Roy[28,29,30], Anupama Rao[47], Philip Ancliff[47], Christian Babbs[28,30], D Mark Layton[23,24], Adam J Mead[28], Jennifer O'Sullivan[45], Steven Okoli[28,30], Irene Roberts[28,29,30]; **Steroid Resistant Nephrotic Syndrome (SRNS)** Moin Saleem[43,59], Agnieszka Bierzynska[43], Carmen Bugarin Diz[19], Elizabeth Colby[43], Melanie N Ekani[122], Simon Satchell[43,206], Ania Koziell[19,20]; **UK Biobank Extreme Red Blood Cell Traits (UKB)** William J Astle[1,2,4,6,7], Suthesh Sivapalaratnam[4,68,69,70], Noemi B Roy[28,29,30]

Genomics England Rare Diseases Pilot Study (GEL RD Pilot): Genomics England Core Teams Tom Fowler[13], Augusto Rendon[1,13], Richard Scott[13,47], Damian Smedley[13,14], Katherine R Smith[13], Ellen Thomas[13,122], Mark J Caulfield[13,14]; **Cambridge University Hospitals NHS Foundation Trust** Stephen Abbs[49], Nigel Burrows[36], Manali Chitre[44], Eleanor F Dewhurst[1,2], R Andres Floto[27,36,159], Michael Gattens[36], Mark Gurnell[27,36], Simon Holden[56], Wilf Kelsall[36], Sarju Mehta[56], Ken E S Poole[27,36], Robert Ross-Russell[36], Olivera Spasic-Boskovic[49], Philip Twiss[49], Annette Wagner[36], F Lucy Raymond[2,9]; **Central Manchester University Hospitals NHS Trust and Manchester University** Siddharth Banka[144,207], Graeme C Black[144,207], Jill Clayton-Smith[144,207], Sofia Douzgou[144,207], William G Newman[144,207]; **Great Ormond Street Hospital for Children NHS Foundation Trust and University College London** Lara Abulhoul[47], Paul Aurora[47], Detlef Bockenhauer[47], Maureen Cleary[47], Mehul Dattani[208,209], Vijeya Ganesan[47], Clarissa Pilkington[47], Shamima Rahman[47,208], Neil Shah[31,47], Lucy Wedderburn[31,210,211], Maria A K Bitner-Glindzicz[47,208]; **Guy's and St Thomas' Hospital NHS Foundation Trust and King's College London** Teofila Bueser[120,212], Cecilia J Compton[120], Charu Deshpande[120], Hiva Fassihi[213], Eshika Haque[120], Louise Izatt[120], Dragana Josifova[120], Shehla N Mohammed[120], Leema Robert[120], Sarah J Rose[120], Deborah M Ruddy[120], Robert N Sarkany[213], Genevieve Sayer[120], Adam C Shaw[120], Melita Irving[120], Frances A Flinter[120]; **Moorfields Eye Hospital NHS Trust and University College London** Gavin Arno[50,51], Samantha Malka[50,51], Michel Michaelides[50,51], Anthony T Moore[50,51,131], Andrew R Webster[50,51]; **Oxford University Hospitals NHS Trust and the University of Oxford** Carolyn Campbell[214], Kate Gibson[214], Nils Koelling[215], Tracy Lester[214], Andrea H Nemeth[12,216], Claire Palles[217], Smita Patel[218], Noemi B Roy[215,219], Arjune Sen[30,220,221], John M Taylor[214], Ian P Tomlinson[217], Jenny C Taylor[30,33], Andrew O Wilkie[215]; **Newcastle upon Tyne Hospitals NHS Foundation Trust and Newcastle University** Paul Brennan[115,116,117], Andrew C Browning[222], John Burn[115], Patrick F Chinnery[2,15,16], Anthony De Soyza[117,182,223], Jodie Graham[224], Rita Horvath[132], Simon Pearce[117,224], Richard Quinton[117,134], Andrew M Schaefer[117,132], Brian T Wilson[114,117,134], Michael Wright[115], Patrick Yu-Wai-Man[15,16,136], John A Sayer[117,134]; **University College London Hospitals NHS Trust and University College London** Michael Simpson[80], Petros Syrris[225], Perry Elliott[225,226], Henry Houlden[81], Phil L Beales[47,208]

Acknowledgements This research was made possible through access to the data and findings generated by two pilot studies for the 100,000 Genomes Project. The enrolment for one pilot study was coordinated by the NIHR BioResource and the other by Genomics England Limited (GEL), a wholly owned company of the Department of Health in the UK. These pilot studies were mainly funded by grants from the National Institute for Health Research (NIHR) in England to the

University of Cambridge and GEL, respectively. Additional funding was provided by the BHF, MRC, NHS England, the Wellcome Trust and many other fund providers (also see Funding acknowledgment for individual researchers). The pilot studies use data provided by patients and their close relatives and collected by the NHS and other healthcare providers as part of their care and support. The vast majority of participants in the two pilot studies have been enrolled in the NIHR BioResource. We thank all volunteers for their participation, and also gratefully acknowledge NIHR Biomedical Research Centres, NIHR BioResource Centres, NHS Trust Hospitals, NHS Blood and Transplant and staff for their contribution. This research has been conducted using the UK Biobank Resource under Application Number 9616, granting access to DNA samples and accompanying participant data. UK Biobank has received funding from the MRC, Wellcome Trust, Department of Health, British Heart Foundation (BHF), Diabetes UK, Northwest Regional Development Agency, Scottish Government, and Welsh Assembly Government. The MRC and Wellcome Trust played a key role in the decision to establish UK Biobank.

Funding acknowledgment for individual researchers AMM and JMo are funded by The Wellcome Trust (WT200990/Z/16/Z) and the European Molecular Biology Laboratory; KGCS holds a Wellcome Investigator Award, MRC Programme Grant (number MR/L019027/1); MIM is a Wellcome Senior Investigator and receives support from the Wellcome Trust (090532, 0938381) and is a member of the DOLORisk consortium funded by the European Commission Horizon 2020 (ID633491); RHo is a Wellcome Trust Investigator (109915/Z/15/Z), who receives support from the Wellcome Centre for Mitochondrial Research (203105/Z/16/Z), MRC (MR/N025431/1), the European Research Council (309548), the Wellcome Trust Pathfinder Scheme (201064/Z/16/Z), the Newton Fund (UK/Turkey, MR/N027302/1) and the European Union H2020 – Research and Innovation Actions (SC1-PM-03-2017, Solve-RD); DLB is a Wellcome clinical scientist (202747/Z/16/Z) and is a member of the DOLORisk consortium funded by the European Commission Horizon 2020 (ID633491); JSW is funded by Wellcome Trust [107469/Z/15/Z], NIHR Cardiovascular Biomedical Research Unit at Royal Brompton & Harefield NHS Foundation Trust and Imperial College London; LSo is supported by the Wellcome Trust Institutional Strategic Support Fund (204809/Z/16/Z) awarded to St. George's, University of London; MJD receives funding from Wellcome Trust (WT098519MA); MCS holds a MRC Clinical Research Training Fellowship (MR/R002363/1); JAS is funded by MRC UK grant MR/M012212/1; AJM received funding from a MRC Senior Clinical Fellowship (MR/L006340/1); CLe received funding from a MRC Clinical Research Training Fellowship (MR/J011711/1); MRW holds a NIHR award to the NIHR Imperial Clinical Research Facility at Imperial College Healthcare NHS Trust; DJW receives part of his salary from the NIHR University College London Hospitals Biomedical Research Centre; MAKu holds a NIHR Research Professorship (NIHR-RP-2016-07-019) and Wellcome Intermediate Fellowship (098524/Z/12/A); MJC is an NIHR Senior Investigator and is funded by the NIHR Barts Biomedical Research Centre; NCo is partially funded by NIHR Imperial College Biomedical Research Centre; CHad was funded through a PhD Fellowship by the NIHR Translational Research Collaboration - Rare Diseases; ADM and SKW were funded by the NIHR Bristol Biomedical Research Centre; ELM received funding from the NIHR Biomedical Research Centre at University College London Hospitals; KCG received funding from the NIHR Great Ormond Street Biomedical Research Centre; IR and ELo are supported by the NIHR Translational

Research Collaboration - Rare Diseases; JCT, JMT and SPat are funded by the NIHR Oxford Biomedical Research Centre; GArn is funded by the NIHR Moorfields Biomedical Research Centre and UCL Institute of Ophthalmology, Fight for Sight (UK) Early Career Investigator Award, Moorfields Eye Hospital Special Trustees, Moorfields Eye Charity, Foundation Fighting Blindness (USA) and Retinitis Pigmentosa Fighting Blindness; ATM is funded by Retinitis Pigmentosa Fighting Blindness, PY-W-M is supported by grants from MRC UK (G1002570), Fight for Sight (1570/1571), Fight for Sight (24TP171), NIHR (IS-BRC-1215-20002); SOB is supported by NIHR Translational Research Collaboration - Rare Diseases (01/04/15-30/04/2017); ARW works for the NIHR Moorfields Biomedical Research Centre and the UCL Institute of Ophthalmology and Moorfields Eye Hospital; the following NIHR Biomedical Research Centres contributed to the enrolment for the ICP domain: Imperial College Healthcare NHS Trust, Guy's and St Thomas' NHS Foundation Trust and King's College London. All authors affiliated with Moorfields Eye hospital and Institute of Ophthalmology are funded by the NIHR Biomedical Resource Centre at UCL Institute of Ophthalmology and Moorfields; ACT is a member of the International Diabetic Neuropathy Consortium, the Novo Nordisk Foundation (Ref. NNF14SA0006) and is a member of the DOLORisk consortium funded by the European Commission Horizon 2020 (ID633491); JWhi is a recipient of a Cancer Research UK Cambridge Cancer Centre Clinical Research Training Fellowship; PSh holds a Henry Smith Charity and Department of Health (UK) Senior Fellowship; SAJ is funded by Kids Kidney Research; DPG is funded by the MRC, Kidney Research UK and St Peters Trust for Kidney, Bladder and Prostate Research; KJM is supported by the Northern Counties Kidney Research Fund; PHD receives funding from ICP Support; TKB received a PhD fellowship from the NHSBT and British Society of Haematology; HSM receives support from BHF Programme Grant no. RG/16/4/32218; AL is a BHF Senior Basic Science Research Fellow - FS/13/48/30453; KF and CVG are supported by the Research Council of the University of Leuven (BOF KU Leuven, Belgium; OT/14/098); HJB works for the Netherlands CardioVascular Research Initiative (CVON); GBa holds a WA Department of Health, Raine Clinician Research Fellowship 2015GB.

Disclaimer The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or of the any other funding agencies.

Competing Interests LHM acts as a consultant for Drayson Technologies; AMK had no competing interests at the time of the study, since the study has received an educational grant from CSL Behring to attend the ISTH meeting (2017); TJA has received consultancy payments from AstraZeneca within the last 5 years and has received speaker honoraria from Illumina Inc.; SW has received an educational grant from CSL Behring and an honorarium from Biotest, LFB; CLS has received educational grants to attend conferences from CSL Behring, Alk and Baxter; MJP has received support for attending educational events and speaker's fees from Biotest UK, Shire UK, and Baxter; TE-S has received support for attending educational events from Biotest UK, CSL and Shire UK; YMK holds a grant from Roche; ARo, CChe, CSt, EB, KTat, NLe, RPr are employees of Congenica Ltd; BTo, JFi, JK, MV, TKa are employees of GENALICE; CCol, CGe, CJBo, CRe, DRB, JFP, JHu, RJG, SHum, SHun, TSAG are employees of Illumina Cambridge Limited; CVG is holder of the Bayer and Norbert Heimburger (CSL Behring) Chair; KJM previously received funding for research and currently on the scientific advisory board of Gemini

Therapeutics, Boston, USA; YMCH received free IVD diagnostic tools and reagents from companies in laboratory haemostasis for studies and/or validations (Werfen, Roche, Siemens, Stage, Nodia); MCS received travel and accommodation fees from NovoNordisk; DML serves on advisory boards for Agios, Novartis and Cerus; MIM serves on advisory panels for Pfizer, NovoNordisk, Zoe Global, has received honoraria from Pfizer, NovoNordisk and Eli Lilly, has stock options in Zoe Global, has received research funding from Abbvie, AstraZeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, NovoNordisk, Pfizer, Roche, Sanofi Aventis, Servier, Takeda. The remaining authors declare no competing financial interests.

Additional information

Extended data is available for this paper at TBC

Supplementary information is available for this paper at TBC

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to who1000@cam.ac.uk.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

APPENDIX - Affiliations

[1]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [2]NIHR BioResource, Cambridge University Hospitals NHS Foundation, Cambridge Biomedical Campus, Cambridge, UK. [3]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [4]NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, UK. [5]British Heart Foundation Cambridge Centre of Excellence, University of Cambridge, Cambridge, UK. [6]MRC Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge, UK. [7]MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, Cambridge, UK. [8]Department of Cardiovascular Sciences, Center for Molecular and Vascular Biology, KULeuven, Leuven, Belgium. [9]Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [10]MRC Clinical Sciences Centre, Faculty of Medicine, Imperial College London, London, UK. [11]Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [12]The Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford, UK. [13]Genomics England, Charterhouse Square, London, UK. [14]William Harvey Research Institute, NIHR Biomedical Research Centre at Barts, Queen Mary University of London, London, UK. [15]Department of Clinical Neurosciences, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [16]Medical Research Council Mitochondrial Biology Unit, Cambridge Biomedical Campus, Cambridge, UK. [17]Women and Children's Health, School of Life Course Sciences, King's College London, London, UK. [18]UCL Centre for Nephrology, University College London, London, UK. [19]King's College London, London, UK. [20]Department of Paediatric Nephrology, Evelina London Children's Hospital, Guy's & St Thomas' NHS Foundation Trust, London, UK. [21]Department of Pediatric Hematology, Immunology, Rheumatology and Infectious Diseases, Emma Children's Hospital, Academic Medical Center (AMC), University of Amsterdam, Amsterdam, The Netherlands. [22]Department of Blood Cell Research, Sanquin, Amsterdam, The Netherlands. [23]Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [24]Centre for Haematology, Imperial College London, London, UK. [25]NIHR Cambridge Biomedical Research Centre, Cambridge Biomedical Campus, Cambridge, UK. [26]Stroke Research Group, Department of Clinical Neurosciences, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [27]Department of Medicine, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [28]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [29]Department of Paediatrics, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [30]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, UK. [31]UCL Great Ormond Street Institute of Child Health, London, UK. [32]Department of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [33]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [34]Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [35]Institute of Reproductive and Developmental Biology, Surgery and Cancer, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [36]Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [37]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [38]Congenica, Biodata Innovation Centre, Wellcome Genome Campus, Hinxton,

Cambridge, UK. [39]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, UK. [40]GENALICE BV, Harderwijk, The Netherlands. [41]High Performance Computing Service, University of Cambridge, Cambridge, UK. [42]Illumina Cambridge Limited, Chesterford Research Park, Little Chesterford, Saffron Walden, Essex, UK. [43]Bristol Renal and Children's Renal Unit, Bristol Medical School, University of Bristol, Bristol, UK. [44]Department of Paediatrics, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [45]Department of Haematology, Guy's and St Thomas' NHS Foundation Trust, London, UK. [46]Centre for Haematology, Department of Medicine, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [47]Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [48]Cancer Research UK Cambridge Centre, Cambridge Biomedical Campus, Cambridge, UK. [49]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [50]Moorfields Eye Hospital NHS Foundation Trust, London, UK. [51]UCL Institute of Ophthalmology, University College London, London, UK. [52]Institute of Immunity and Transplantation, University College London, London, UK. [53]Department of Immunology, Royal Free London NHS Foundation Trust, London, UK. [54]The Katharine Dormandy Haemophilia Centre and Thrombosis Unit, Royal Free London NHS Foundation Trust, London, UK. [55]University College London, London, UK. [56]Department of Clinical Genetics, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [57]School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK. [58]University Hospitals Bristol NHS Foundation Trust, Bristol, UK. [59]Bristol Royal Hospital for Children, University Hospitals Bristol NHS Foundation Trust, Bristol, UK. [60]The Department of Clinical Immunology and Allergy and The NIHR Leeds Biomedical Research Centre, Leeds, UK. [61]Leeds Institute of Rheumatic and Musculoskeletal Medicine, St James's University Hospital, Leeds, UK. [62]MRC Toxicology Unit, School of Biological Sciences, University of Cambridge, Cambridge, UK. [63]Department of Medical Genetics and NIHR Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK. [64]Division of Medical Genetics, IWK Health Centre, Dalhousie University, Halifax, Canada. [65]JDRF/Wellcome Diabetes and Inflammation Laboratory, Wellcome Centre for Human Genetics, Nuffield Department of Medicine, NIHR Oxford Biomedical Research Centre, University of Oxford, Oxford, UK. [66]Department of Medicine, Imperial College London, London, UK. [67]Department of Paediatric Haematology, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [68]The Royal London Hospital, Barts Health NHS Foundation Trust, London, UK. [69]Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [70]Queen Mary University of London, London, UK. [71]MRC Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK. [72]NIHR Biomedical Research Centre for Mental Health, Maudsley Hospital, London, UK. [73]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. [74]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [75]Department of Cardiology, Ealing Hospital, Middlesex, UK. [76]Imperial College Healthcare NHS Trust, London, UK. [77]MRC-PHE Centre for Environment and Health, Imperial College London, London, UK. [78]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. [79]Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Research Centre, University of Leicester, Leicester, UK. [80]Genetics and Molecular Medicine, King's College London, London, UK. [81]Department of

Molecular Neuroscience, UCL Institute of Neurology, London, UK. [82]UCL Genetics Institute, London, UK. [83]Department of Renal Medicine, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [84]Queens Centre for Haematology and Oncology, Castle Hill Hospital, Hull and East Yorkshire NHS Trust, Cottingham, UK. [85]Hull York Medical School, University of Hull, Hull, UK. [86]Center for Clinical Transfusion Medicine, University Hospital of Tübingen, Tübingen, Germany. [87]Haematology Department, Royal Victoria Infirmary, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [88]Southampton General Hospital, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [89]Institute of Infection and Immunity, School of Medicine Cardiff University, Cardiff, UK. [90]Oxford Haemophilia and Thrombosis Centre, Oxford University Hospitals NHS Trust, Oxford Comprehensive Biomedical Research Centre, Oxford, UK. [91]The Roald Dahl Haemostasis and Thrombosis Centre, The Royal Liverpool University Hospital, Liverpool, UK. [92]Medical School and School of Biomedical Sciences, Faculty of Health and Medical Sciences, The University of Western Australia, Crawley, Australia. [93]Haemophilia Centre, Kent & Canterbury Hospital, East Kent Hospitals University Foundation Trust, Canterbury, UK. [94]Salisbury District Hospital, Salisbury NHS Foundation Trust, Salisbury, UK. [95]Haemophilia, Haemostasis and Thrombosis Centre, Hampshire Hospitals NHS Foundation Trust, Basingstoke, UK. [96]INSERM UMR 1170, Gustave Roussy Cancer Campus, Universite Paris-Saclay, Villejuif, France. [97]Service d'Hematologie biologique, Centre de Reference des Pathologies Plaquettaires, Hopital Armand Trousseau, Assistance Publique-Hopitaux de Paris, Paris, France. [98]Institute for Immunology and Transfusion Medicine, University Medicine Greifswald, Greifswald, Germany. [99]Section of Internal and Cardiovascular Medicine, University of Perugia, Perugia, Italy. [100]Division of Hematology, The Children's Hospital of Philadelphia, Philadelphia, USA. [101]Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA. [102]Department of Haematology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [103]Institut Hospitalo-Universitaire de Rythmologie et de Modelisation Cardiaque, Plateforme Technologique d'Innovation Biomedicale, Hopital Xavier Arnozan, Pessac, France. [104]The Arthur Bloom Haemophilia Centre, University Hospital of Wales, Cardiff, UK. [105]Beth Israel Deaconess Medical Centre and Harvard Medical School, Boston, USA. [106]University College London Hospitals NHS Foundation Trust, London, UK. [107]Oxford Haemophilia and Thrombosis Centre, The Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK. [108]Glasgow Royal Infirmary, NHS Greater Glasgow and Clyde, Glasgow, UK. [109]Department of Clinical Genetics, Leicester Royal Infirmary, University Hospitals of Leicester, Leicester, UK. [110]University of Leicester, Leicester, UK. [111]Department of Neurology, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [112]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. [113]Department of Neurology, Leeds Teaching Hospital NHS Trust, Leeds, UK. [114]North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [115]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [116]Newcastle University, Newcastle upon Tyne, UK. [117]Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [118]National Heart and Lung Institute, Imperial College London, London, UK. [119]Royal Brompton Hospital, Royal Brompton and Harefield NHS Foundation Trust, London, UK. [120]Clinical Genetics Department, Guy's and St Thomas NHS Foundation Trust, London, UK. [121]King's College Hospital NHS Foundation Trust, London, UK. [122]Guy's and St Thomas' Hospital, Guy's

and St Thomas' NHS Foundation Trust, London, UK. [123]MRC London Institute of Medical Sciences, Imperial College London, London, UK. [124]National Heart Research Institute Singapore, National Heart Centre Singapore, Singapore, Singapore. [125]Division of Cardiovascular & Metabolic Disorders, Duke-National University of Singapore, Singapore, Singapore. [126]Department of Biotechnology, Graduate School of Engineering, Osaka University, Suita, Osaka, Japan. [127]Women's Health Research Centre, Surgery and Cancer, Faculty of Medicine, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [128]Ramón Sardá Mother's and Children's Hospital, Buenos Aires, Argentina. [129]Robinson Research Institute, Discipline of Obstetrics and Gynaecology, The University of Adelaide, Women's and Children's Hospital, Adelaide, Australia. [130]Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [131]Ophthalmology Department, UCSF School of Medicine, San Francisco, USA. [132]Wellcome Centre for Mitochondrial Research, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [133]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [134]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [135]John Walton Muscular Dystrophy Research Centre, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [136]NIHR Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology, London, UK. [137]Yorkshire Regional Genetics Service, Chapel Allerton Hospital, Leeds Teaching Hospitals NHS Trust, Leeds, UK. [138]Department of Clinical Genetics, Royal Devon & Exeter Hospital, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK. [139]West Midlands Regional Genetics Service, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. [140]Manchester University NHS Foundation Trust, Manchester, UK. [141]Department of Clinical Genetics, Liverpool Women's NHS Foundation, Liverpool, UK. [142]Department of Clinical Genetics, St George's University Hospitals NHS Foundation Trust, London, UK. [143]Department of Clinical Genetics, Guy's and St Thomas' NHS Foundation Trust, London, UK. [144]Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester Universities Foundation NHS Trust, Manchester, UK. [145]Department of Clinical Genetics, Nottingham University Hospitals NHS Trust, Nottingham, UK. [146]Wessex Clinical Genetics Service, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [147]Developmental Neurosciences, UCL Great Ormond Street Institute of Child Health, London, UK. [148]Department of Neurology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [149]Salford Royal NHS Foundation Trust, Salford, UK. [150]Faculty of Medical and Human Sciences, Centre for Endocrinology and Diabetes, Institute of Human Development, University of Manchester, Manchester, UK. [151]Department of Clinical Neurophysiology, Manchester Royal Infirmary, Central Manchester University Hospitals National Health Service Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. [152]National Institute for Health Research/Wellcome Trust Clinical Research Facility, Manchester, UK. [153]The National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK. [154]MRC Centre for Neuromuscular Diseases, Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK. [155]Pain Research, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK. [156]Pain Medicine, Chelsea and Westminster Hospital NHS Foundation Trust, London, UK. [157]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [158]Division of Clinical Biochemistry and Immunology, Cambridge University Hospitals NHS Foundation Trust,

Cambridge, UK. [159]Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK. [160]Department of Immunology, Leicester Royal Infirmary, Leicester, UK. [161]Royal Free London NHS Foundation Trust, London, UK. [162]Nottingham University Hospitals NHS Trust, Nottingham, UK. [163]Regional Immunology Service, The Royal Hospitals, Belfast, UK. [164]Queen's University Belfast, Belfast, UK. [165]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [166]University Hospitals of North Midlands NHS Trust, Stoke-on-Trent, UK. [167]East Yorkshire Regional Adult Immunology and Allergy Unit, Hull Royal Infirmary, Hull and East Yorkshire Hospitals NHS Trust, Hull, UK. [168]Barts Health NHS Foundation Trust, London, UK. [169]Birmingham Heartlands Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [170]Royal Hospital for Children, NHS Greater Glasgow and Clyde, Glasgow, UK. [171]Epsom & St Helier University Hospitals NHS Trust, London, UK. [172]Immunodeficiency Centre for Wales, University Hospital of Wales, Cardiff, IUK. [173]Centre for Immunology & Vaccinology , Chelsea & Westminster Hospital , Department of Medicine, Imperial College London, London, UK. [174]Department of Respiratory Medicine Royal Brompton & Harefield NHS Foundation Trust, London, UK. [175]Department of Clinical Immunology, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [176]Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK. [177]Scunthorpe General Hospital, Northern Lincolnshire and Goole NHS Foundation Trust, Scunthorpe, UK. [178]Gartnavel General Hospital, NHS Greater Glasgow and Clyde, Glasgow, UK. [179]Queen Elizabeth University Hospital, Glasgow, UK. [180]Birmingham Chest Clinic and Heartlands Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [181]Frimley Park Hospital, NHS Frimley Health Foundation Trust, Camberley, UK. [182]Institute of Cellular Medicine, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK. [183]Imperial College Renal and Transplant Centre, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [184]Children's Renal and Urology Unit, Nottingham Children's Hospital, QMC, Nottingham University Hospitals NHS Trust, Nottingham, UK. [185]The National Renal Complement Therapeutics Centre, Royal Victoria Infirmary, Newcastle upon Tyne, UK. [186]MPGN/C3 Glomerulopathy Rare Renal Disease group, , UK. [187]Department of Paediatric Nephrology, Great North Children's Hospital, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [188]Department of Pulmonary Medicine, VU University Medical Centre, Amsterdam, The Netherlands. [189]Golden Jubilee National Hospital, Glasgow, UK. [190]Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital NHS Foundation Trust, Sheffield, UK. [191]National Pulmonary Hypertension Service (Newcastle), The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [192]Department of Molecular Medicine, General Biology, and Medical Genetics Unit, University of Pavia, Pavia, Italy. [193]Departement de Genetique & ICAN, Hopital Pitie-Salpetriere, Assistance Publique Hopitaux de Paris, Paris, France. [194]University of Giessen and Marburg Lung Center (UGMLC), Giessen, Germany. [195]Division of Cardiology, Fondazione IRCCS Policlinico S. Matteo, Pavia, Italy. [196]Univ. Paris-Sud, Faculty of Medicine, University Paris-Saclay, Le Kremlin Bicetre, France. [197]Service de Pneumologie, Centre de Reference de l'Hypertension Pulmonaire, Hopital Bicetre (Assistance Publique Hopitaux de Paris), Le Kremlin Bicetre, France. [198]INSERM U999, Hospital Marie Lannelongue, Le Plessis Robinson, France. [199]National Pulmonary Hypertension Service, Imperial College Healthcare NHS Trust, London, UK. [200]Ludwig Boltzmann Institute for Lung Vascular Research, Graz, Austria. [201]Dept of Internal Medicine, Division of Pulmonology, Medical University of Graz, Graz, Austria. [202]Department of

Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, UK. [203]Royal United Hospitals Bath NHS Foundation Trust, Bath, UK. [204]Department of Clinical Genetics, VU University Medical Centre, Amsterdam, The Netherlands. [205]Imperial College London, London, UK. [206]North Bristol NHS Trust, Bristol, UK. [207]Evolution and Genomic Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. [208]Genetics and Genomic Medicine Programme, UCL Great Ormond Street Institute of Child Health, London, UK. [209]London Centre for Paediatric Endocrinology and Diabetes, Great Ormond Street Hospital for Children, London, UK. [210]NIHR Great Ormond Street Biomedical Research Centre, London, UK. [211]Arthritis Research UK Centre for Adolescent Rheumatology at UCL UCLH and GOSH, London, UK. [212]Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, UK. [213]St Johns Institute of Dermatology, St Thomas' Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK. [214]Oxford Medical Genetics Laboratories, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [215]MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK. [216]Department of Clinical Genetics, Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK. [217]Institute of Cancer and Genomic Sciences, Institute of Biomedical Research, University of Birmingham, Birmingham, UK. [218]Department of Clinical Immunology, John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [219]Department of Haematology, Oxford University Hospital Foundation Trust, Oxford, UK. [220]Oxford Epilepsy Research Group, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. [221]Nuffield Department of Surgery, University of Oxford, Oxford, UK. [222]Newcastle Eye Centre, Royal Victoria Infirmary, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [223]NIHR Centre for Aging, Newcastle University, Newcastle upon Tyne, UK. [224]Newcastle BRC, Newcastle University, Newcastle upon Tyne, UK. [225]UCL Institute of Cardiovascular Science, University College London, London, UK. [226]Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, UK.

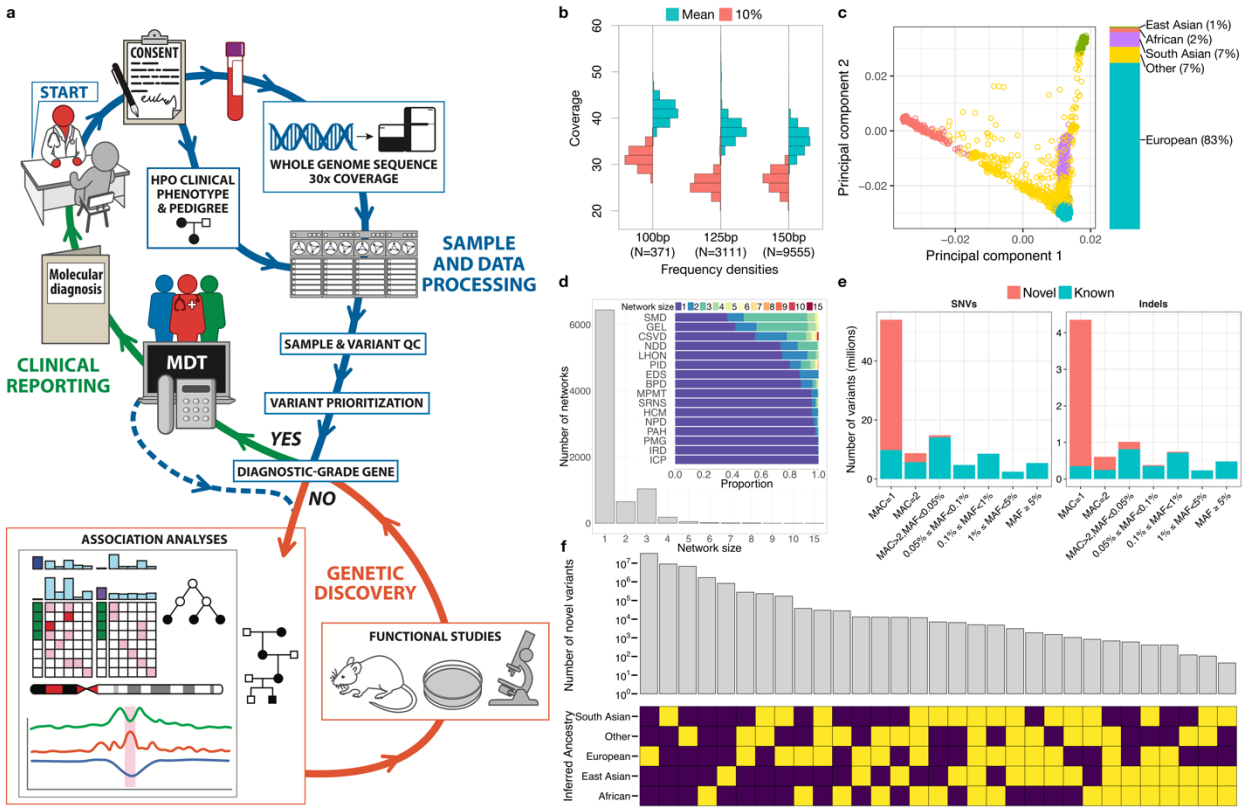

Fig. 1. Overview of the study and genetic data analysis. a, Schematic depicting the flow of information through the study and the synergy between diagnosis and discovery. In blue: an undiagnosed patient is recruited into the study by his/her clinician, informed consent is obtained and the clinician enters Human Phenotype Ontology (HPO) terms and pedigree information into the study database, biological samples are taken and DNA is sent to a single Illumina laboratory for WGS, sequencing data are transferred to a high performance computing cluster for bioinformatic QC and the prioritisation of variants in diagnostic-grade genes. In green: selected variants meeting predefined characteristics are presented to the multi-disciplinary teams (MDTs) using the Sapientia web application, variants are categorised as pathogenic or likely pathogenic, a molecular diagnosis may be returned to the referring clinician. In orange: statistical and bioinformatic analyses are applied to the genetic and phenotypic data to identify aetiological variants, disease-mediating genes and regulatory regions. Participants and close relatives are invited to participate in co-segregation and functional studies, and model systems are used to study disease mechanisms. **b**, Histograms showing the distribution of read coverage across 13,037 samples, stratified by sequencing read lengths of 100bp,125bp and 150bp. **c**, Projection of participants onto the first two principal components of genetic variation in the 1000 Genomes Project (left sub-panel), bar plot showing the percentage of participants whose ancestry was assigned to different 1000 Genomes populations (right sub-panel). **d**, Bar plot showing the size distribution of genetically determined networks of closely related individuals. Inset: Distributions of network sizes for each rare disease domain. **e**, Histograms illustrating the observed allele frequency distribution of variants measured in 10,259 unrelated samples, stratified by variant type (SNV or indel). Variants were labelled novel if they were uncatalogued in the the following databases: 1000 Genomes, UK10K, TOPMed, gnomAD, HGMD. MAC: minor allele count; MAF:

minor allele frequency. **f**, Histogram counting (log_{10} scale) the novel variants according to the ancestry groups in which they were observed (yellow: present, navy: absent).

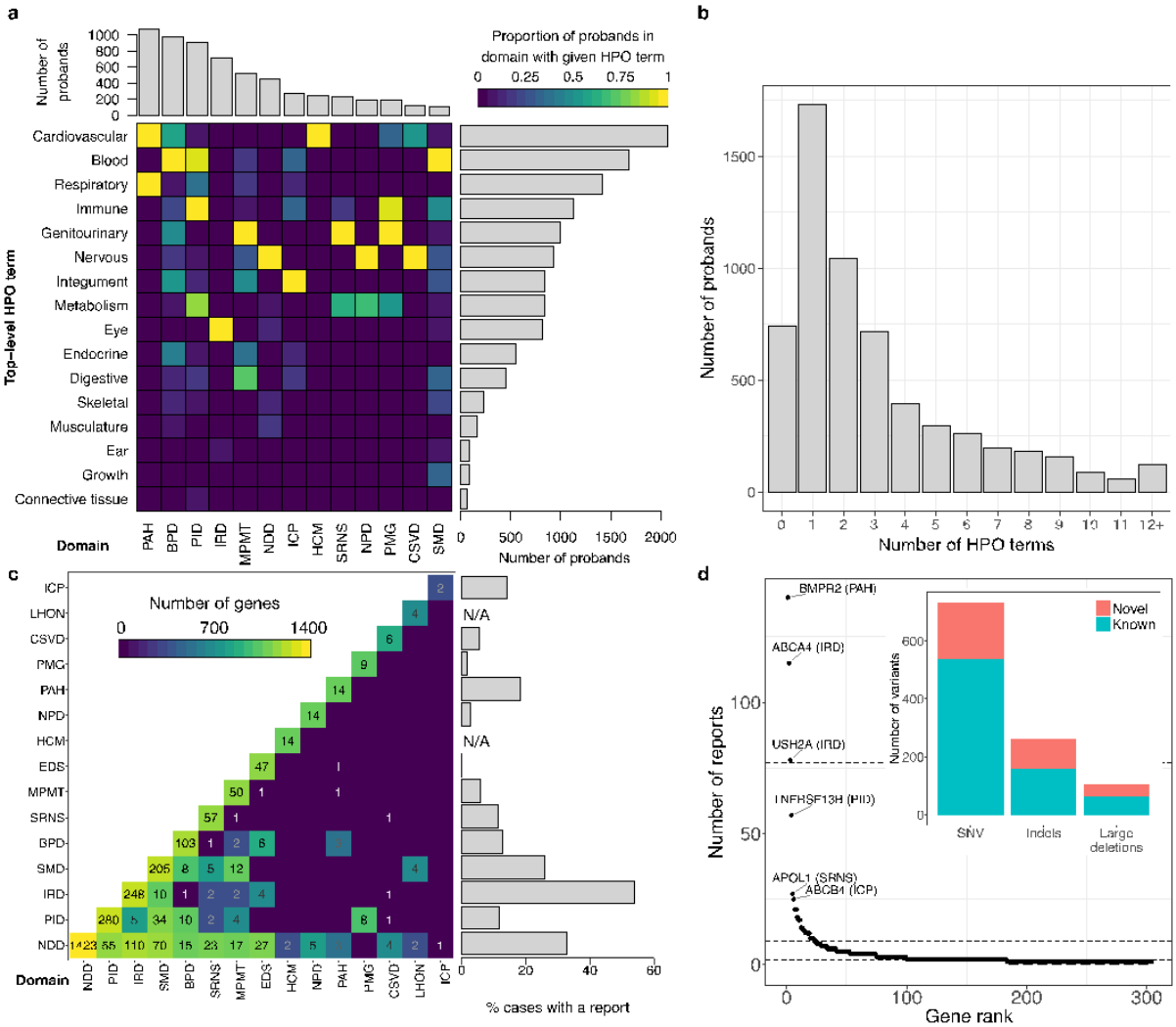

Fig. 2. Phenotyping data, diagnostic-grade genes and MDT-reported results. a, Bar plot showing the distribution of probands by domain (top); bar plot counting the number of probands with each top-level HPO term (right). The heat map shows the proportion of probands in each domain who have been assigned a particular top-level HPO term. Top-level HPO terms have been abbreviated. The full term names read 'Abnormality of,' followed by, from top to bottom: the cardiovascular system; blood and blood-forming tissues; the respiratory system; the immune system; the genitourinary system; the nervous system; integument; metabolism/homeostasis; the eye; the endocrine system; the digestive system; the skeletal system; the musculature; of the ear; growth; connective tissue. **b**, Bar plot showing the count distribution of the number of HPO terms assigned to probands. **c**, Heat map showing the number of DGGs shared by pairs of domains (left). Bar plot of the proportion of cases in each domain for which a clinical report was issued (right). **d**, Number of reports issued per DGG ordered inversely by count. Dashed lines indicate quartiles of the count distribution. Inset: bar plot showing the number of unique clinically reported variants stratified by variant type (SNVs, indels and large deletions). The proportion of each bar coloured iris blue/salmon indicates the proportion of variants which are known (present in the HGMD Pro database)/novel.

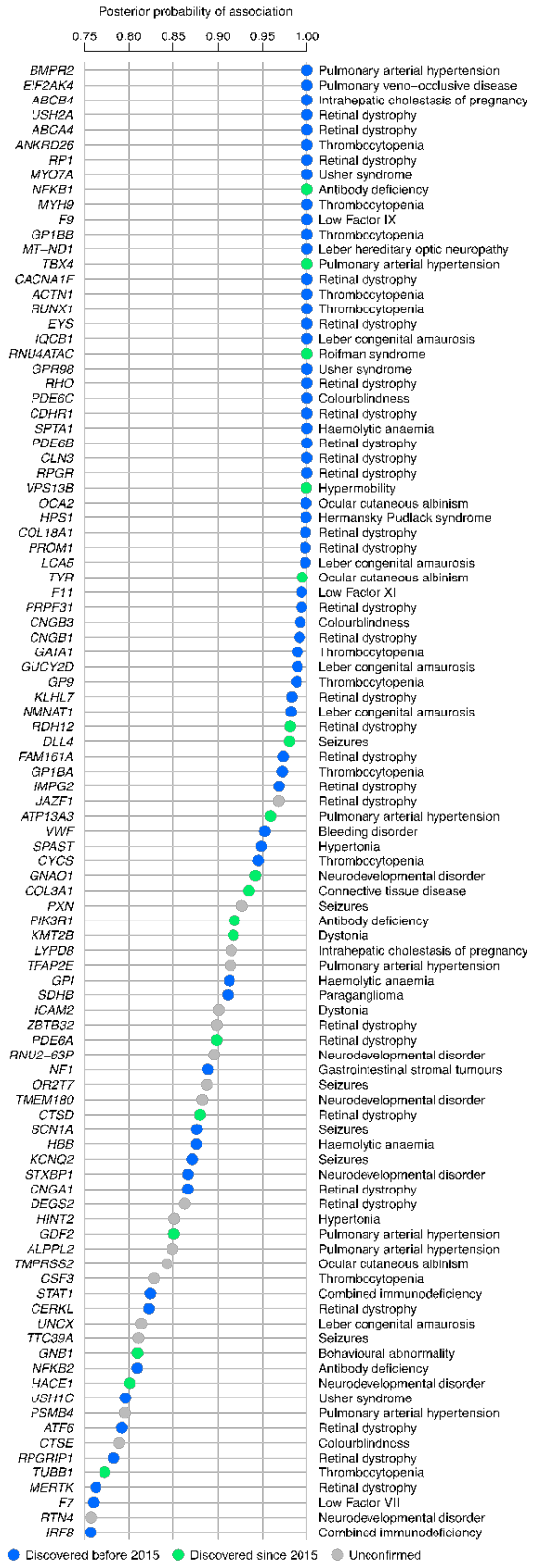

Fig. 3. BeviMed genetic association results for rare diseases. BeviMed was applied gene-wise to infer associations between the genotypes of filtered rare variants and various

case/control groupings (tags). The PPs for genetic association inferred by BeviMed exceeding 0.75 are shown. Gene names are given on the left and the corresponding tag names of the case groupings are given on the right. The green and blue colouring denotes genetic associations supported by original scientific publications since or before 2015, respectively, while grey denote associations that are currently unconfirmed in the literature.

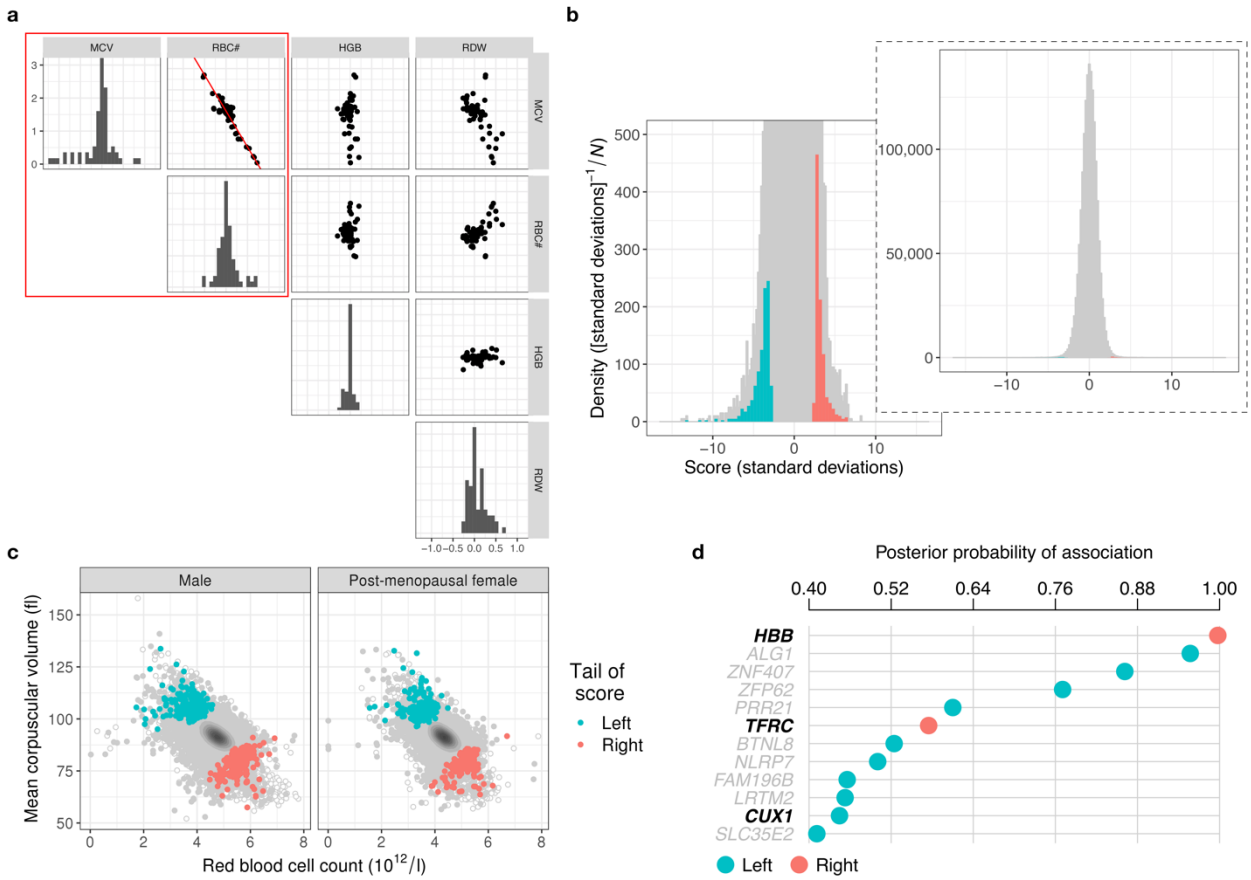

Fig 4. BeviMed genetic associations for extreme RBC traits in UK Biobank. a, Graphical summaries of the joint distribution of the estimated per allele additive effect of 65 variants with MAF<1% on the mean of four rank inverse normalised quantitative properties of red cells: mean corpuscular volume (MCV), RBC count (RBC#), haemoglobin concentration (HGB) and red cell distribution width (RDW). The 65 variants were chosen for being significantly associated with at least one of the 12 red cell traits in Astle *et al*. The on-diagonal panels depict the univariate distribution of the estimated effect sizes of each trait (measured in standard deviations per allele) and the off-diagonal panels depict the bivariate relationship between the estimated effect sizes. The red square highlights the bivariate marginal distribution used to develop the quantitative selection score. The red line in the (RBC#, MCV) panel was estimated by a Deming regression of the MCV effect sizes on the RBC# effect sizes. **b**, Both sub-panels show the distribution of the (centred and standardised) quantitative score in the European ancestry male and post-menopausal female UK Biobank participants without a baseline self report or medical history of an illness or treatment known to perturb RBC indices (grey density histograms) and the distribution of the score in individuals chosen for the left (iris blue density histogram) and right (salmon density histogram) selection tails. The density scale has been chosen so that the area under each histogram represents the respective number of contributing individuals, *N*=316,739. The lower left sub-panel is a vertical stretch of the bottom part of the overlying upper right sub-panel. Many participants in the extreme tails were not be selected because of poor quality DNA in the UK Biobank archive or because of recalibration of the score to ensure a

representative distribution of age and sex amongst those selected. **c**, Bivariate scatter showing the distribution of RBC# and MCV (both adjusted for technical but not biological noise) in UK Biobank males (left sub-panel) and post-menopausal females (right sub-panel). The overlaid ellipsoids are contours from kernel density estimates of the central parts of the distributions. The closed grey circles are due to the participants contributing to the grey density histograms in sub-figure 4b. The (underlying) open circles are due to participants excluded from selection for sequencing on the grounds of ancestry or medical history. The excess of open circles in the bottom right of each sub-panel is probably explained by the high prevalence of thalassemia in participants with African or Mediterranean ancestry. The coloured circles indicate the participants selected for sequencing from the two tails of the score. **d**, BeviMed computed PPs for genetic association with each of the tails of the score (distinguished by colour), for genes with probabilities greater than 0.4. The strength of concordant biological evidence for the genes given in boldface is such that they can be considered positive controls.

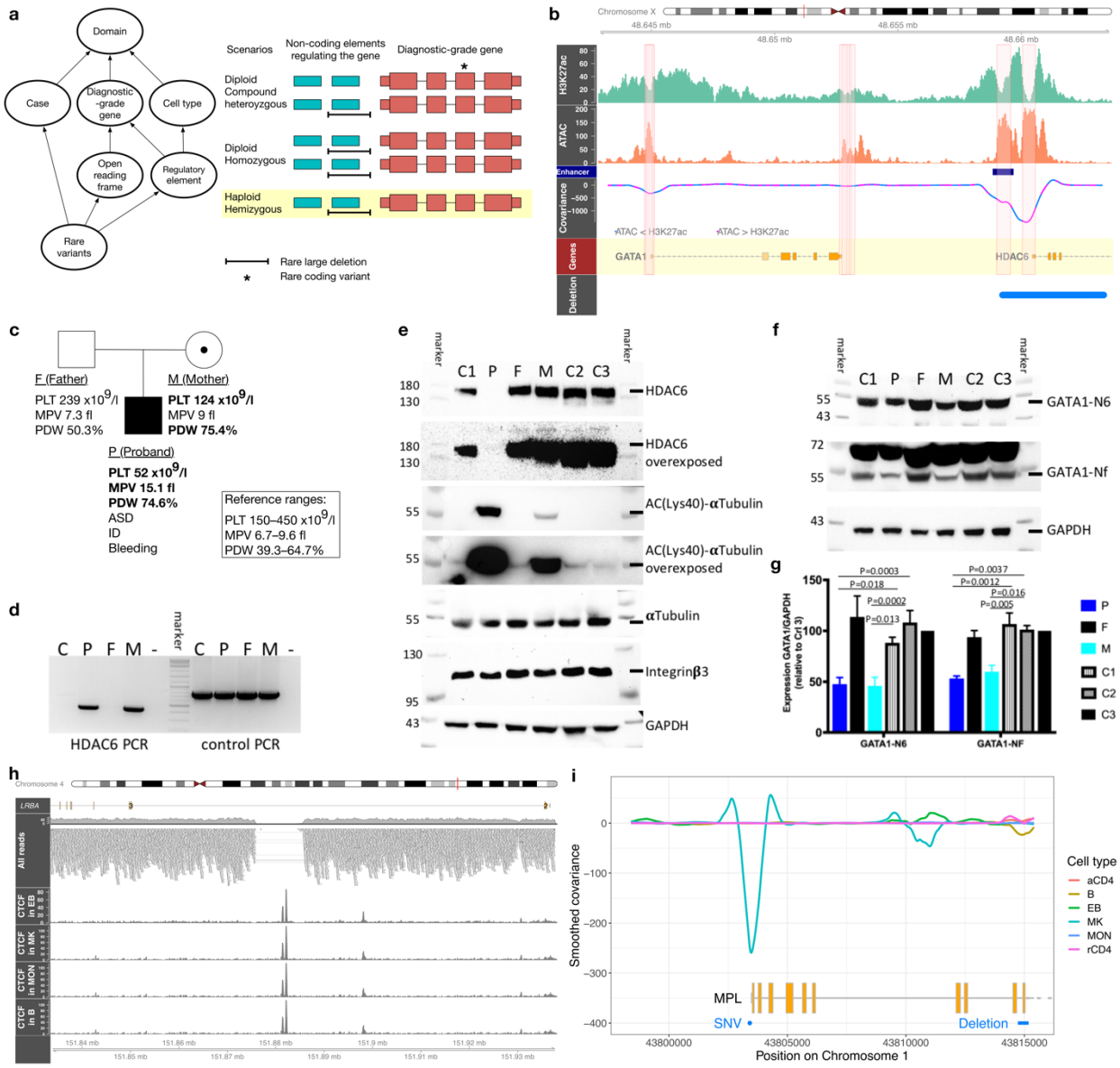

Fig. 5. Causal variants in regulatory elements. a, Left: schematic of the procedure for identifying causal deletions in regulatory elements of DGGs. Right: schematic showing models of expanded DGGs (in salmon) including regulatory elements (in iris blue) and three possible genetic architectures underlying a rare disease. **b**, From top to bottom: X chromosome, position on chromosome, genomic coverage of H3K27ac ChIP-seq (green) and ATAC-seq (orange) in megakaryocytes. The *GATA1* enhancer is shown as a dark blue horizontal bar. The smoothed covariance between H3K27ac ChIP-seq and ATAC-seq coverage was used to call the regulatory elements indicated by the shaded pink panels in overlay. Gene exons are shown in orange and the large deletion in the proband is shown as a light blue horizontal bar. The deleted element binds the transcription factors characteristic of the MK lineage: FLI1, GATA1/2, MEIS1, RUNX1 and TAL1 (not shown). **c**, Pedigree of proband (P) with thrombocytopenia and autism and his parents (F and M). PLT: platelet count, MPV: mean platelet volume, PDW: platelet distribution width, ASD: autism spectrum disorder, ID: Intellectual disability. **d**, Left: Gel

electrophoresis showing presence and absence of short PCR amplicons using primers flanking the deletion. Right: control PCR; no DNA added indicated by '-'. **e**, Representative immunoblots performed in duplicate for total platelet lysates for the indicated proteins and individuals. **f**, Representative immunoblots of total platelet lysates using two different GATA1 antibodies. **g**, Quantification of GATA1 protein levels obtained from three independent immunoblots (as per **f**) showing the mean and SEM and two-way ANOVA analysis *P* values (multiple comparisons). **h**, Depiction of sequencing reads in IGV showing a homozygous deletion in *LRBA* and CTCF ChIP-seq coverage in erythroblasts (EB), megakaryocytes (MK), monocytes (MON) and B cells (B). **i**, Top: Smoothed covariance between H3K27ac ChIP-seq and ATAC-seq (as per **b**) and coverage tracks generated by RedPop for activated CD4+ T-cells (aCD4), B, EB, MK MON and resting CD4+ T-cells (rCD4); Middle: *MPL* gene with exons in yellow; Bottom: positions of the deletion (blue bar) and SNV (blue dot) in the proband.