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

Ernest Turro[1,2,3] ✉, William J Astle[3,4], Karyn Megy[1,2], Stefan Gräf[1,2,5], Daniel Greene[1,2,3], Olga Shamardina[1,2], Hana Lango Allen[1,2], Alba Sanchis-Juan[1,2], Mattia Frontini[1,2,6], Chantal Thys[7], Jonathan Stephens[1,2], Rutendo Mapeta[1,2], Oliver S Burren[5,8], Kate Downes[1,2], Matthias Haimel[1,2,5], Salih Tuna[1,2], Timothy J Aitman[9,10], David L Bennett[11], Paul Calleja[12], Keren Carss[1,2], Mark J Caulfield[13,14], Patrick F Chinnery[2,15,16], Peter H Dixon[17], Daniel P Gale[18,19], Roger James[1,2], Ania Koziell[20,21], Michael A Laffan[22,23], Adam P Levine[18], Eamonn R Maher[24,25], Hugh S Markus[26], Joannella Morales[27], Nicholas W Morrell[2,5], Andrew D Mumford[28,29], Elizabeth Ormondroyd[30,31], Stuart Rankin[12], Augusto Rendon[1,13], Sylvia Richardson[3], Irene Roberts[32,33,34], Noemi B Roy[32,33,34], Moin A Saleem[35,36], Kenneth G C Smith[5,8], Hannah Stark[2,37], Rhea Y Y Tan[26], Andreas C Themistocleous[11], Adrian J Thrasher[38], Hugh Watkins[30,31,39], Andrew R Webster[40,41], Martin R Wilkins[42], Catherine Williamson[17,43], James Whitworth[24,25,44], Sean Humphray[45], David R Bentley[45], NIHR BioResource for the 100 000 Genomes Project*, Nathalie Kingston[2], Neil Walker[1,2], John R Bradley[2,5,25,46,47], Sofie Ashford[2,37], Christopher J Penkett[1,2], Kathleen Freson[7], Kathleen E Stirrups[1,2], F Lucy Raymond[2,24] ✉, Willem H Ouwehand[1,2,4,6,48] ✉

Most patients with rare diseases do not receive a molecular diagnosis and the aetiological variants and mediating genes for more than half such disorders remain to be discovered[1]. We implemented whole-genome sequencing (WGS) in a national health 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 1,138 of the 7,065 patients with detailed phenotypic data. We identified 95 Mendelian associations between genes and rare diseases, of which 11 have been discovered since 2015 and at least 79 are confirmed aetiological. Using WGS of UK Biobank[2], we showed that rare alleles can explain the presence of some individuals in the tails of a quantitative red blood cell (RBC) trait. Finally, we identified 4 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.

Rare diseases affect approximately 1 in 20 people, but only a minority of patients receive a genetic diagnosis[3]. Approximately 10,000 rare diseases are known, but fewer than half have a resolved genetic aetiology[1]. Even when the aetiology is known, the prospects for diagnosis are severely diminished by fragmentary phenotyping and the restriction of testing to disease-specific panels of genes. On average, a molecular cause is determined after three misdiagnoses and 16 physician visits over several years[4]. Recent development of WGS technology allows systematic, comprehensive genetic testing in integrated health systems concurrent with aetiological discovery in the coding and non-coding genome.

We 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). The participants were distributed approximately uniformly across the sexes (Supplementary Table 1) and approximately according to the UK census across ethnic groups (Fig. 1c; https://www.ons.gov.uk/census/2011census). 9,802 of the participants (75%) had a rare disease or an extreme measurement of a quantitative trait, of whom 9,024 were probands and 778 were affected relatives. Each participant was assigned to one of 18 domains with pre-specified enrolment criteria (Supplementary Table 1): 7,388 to one of 15 rare disease domains, 50 to a control domain, 4,835 to a Genomics England Limited (GEL) domain and 764 to a domain comprising UK Biobank participants with extreme red blood cell indices (Supplementary Information, Supplementary Table 1, Extended Data Fig. 1b). Sample sizes varied across domains, primarily due to differences in recruitment rates, limiting the efficiency of the experimental design. The patients presented with pathologies of many organ systems, which we phenotyped using Human Phenotype Ontology (HPO) terms for all the rare disease domains except Leber Hereditary Optic Neuropathy (LHON) and Ehler-Danlos/Ehler-Danlos-like Syndromes (EDS) (Fig. 2a, Extended Data Fig. 1c). The GEL domain released only a binary affection phenotype for these pilot analyses. In total, 19,605 HPO terms were assigned to patients.

Following bioinformatic analysis (Extended Data Fig. 2–4), 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 ≤50bp (Extended Data Fig. 5). 48.6% and 40.8% of the SNVs and indels respectively were absent from variant databases (Fig. 1d). 54.8% of the variants had a minor allele count (MAC) of 1 in a maximal set of 10,259 unrelated participants (MSUP). 82.6% of these variants were novel. Only 9.08% of novel variants had a MAC>1 in the MSUP, in which cases the minor allele was typically carried exclusively by individuals with similar population ancestry (Fig. 1e). SNVs and indels were well represented in genetic databases if they were common in our dataset but, consistent with theory, most variants were very rare and uncatalogued. We called 24,436 distinct large deletions (>50bp) by synthesising inferences from two algorithms. We also called more complicated types of structural variant, such as inversions, but this was unreliable and we could not reconcile the calls across individuals (Supplementary Information). Only 13 (0.1%) individuals had non-standard WGS-determined sex chromosomal karyotypes (Extended Data Fig. 3e–g). We inferred familial relationships from the genetic data (Supplementary Information). Due to the enrolment strategies, most families were singletons (Fig. 1f).

Clinical reporting

For each of the 15 rare disease domains, we reviewed the scientific literature to establish a list of diagnostic-grade genes (DGGs) and identify the corresponding transcripts (Supplementary Information). The lists ranged in length from two for Intrahepatic Cholestasis of Pregnancy (ICP) to 1,423 for Neurological and Developmental Disorders (NDD). The lists were not mutually exclusive because mutations in some genes cause pathologies compatible with the enrolment criteria of multiple domains (Fig. 2b). Twelve multidisciplinary teams (MDTs) with domain-specific expertise examined the rare variants observed in DGGs in the context of the HPO phenotypes. 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 variant's contribution was assessed to be at least partial if, given all other known variants in the case, it was considered to be disease determining. Clinical reports containing molecular diagnoses featuring 1,103 distinct causal variants (731 SNVs, 264 indels, 102 large deletions, 6 complex structural variants) affecting 329 genes (Extended Data Fig. 5, Supplementary Table 2), were issued for 1,138 of the 7,065 (16.1%) patients reviewed. 266 of the 995 SNVs and indels (26.7%) were classed as novel because they were absent from the Human Gene Mutation Database (HGMD) and were not among the variants in ClinVar with at least one pathogenic or likely pathogenic interpretation and no benign interpretation. We ranked the 329 DGGs by the number of clinical reports in which they featured. The top three DGGs (*BMPR2*, *ABCA4* and *TNFRSF13B*) featured in a quarter of all reports. The subsequent 19 DGGs featured in a further quarter of reports. The remaining 307 DGGs mostly featured in a single report (Fig. 2c, Extended Data Fig. 6). The diagnostic yield by domain ranged from 0% (0/184) of patients for Primary Membranoproliferative Glomerulonephritis (PMG) to 53.9% (391/725) of patients for Inherited Retinal Disease (IRD). The variability of diagnostic yield is attributable to heterogeneity in: phenotypic and genetic pre-screening before enrolment, the genetic architecture of the diseases and prior knowledge of genetic aetiologies.

Clinical reporting was enhanced by the use of PCR-free WGS with a mean autosomal depth >35X instead of whole-exome sequencing (WES). For example, we identified a causal SNV encoding a start loss of *HPS6* in a case with Hermansky-Pudlak syndrome previously missed by WES. We compared the read coverage of WGS to that of research WES of participants in UK Biobank (https://www.biorxiv.org/content/10.1101/572347v1), INTERVAL[6] and the Columbia University exome sequencing study for chronic kidney disease (Supplementary Information). Although less costly to generate per sample, the variation in coverage within and between genomic sites harbouring known pathogenic SNVs or indels was much greater for WES than WGS (Extended Data Fig. 7). Of the 938 distinct autosomal SNVs reported in this study, the number with insufficient coverage in WES for reliable genotyping ranged between 25 and 99 (2.67%–10.5%) across WES datasets (Extended Data Fig. 7). Moreover, deletions spanning only a few short exons or part of a single exon are not reliably called by WES[7,8]. Of the 102 distinct large deletions that we reported (length range 203bp–16.80Mb, mean 786.33Kb, median 15.91Kb), 22 (21.6%) overlapped only one exon. Clinical and research WES may have different coverage characteristics, but we were unable to obtain an example clinical dataset for comparison.

Genetic discoveries arising from this study have informed treatment decisions: 27 patients with *KMT2B* mediated early-onset dystonia can be treated by deep brain stimulation[9]; cases with *DIAPH1*-related macrothrombocytopenia and deafness[10] can be treated for their thrombocytopenia in a preoperative setting with Eltrombopag[11]; and a case of severe thrombocytopenia, myelofibrosis and bleeding due to a gain-of-function mutation in *SRC*[12] was cured by an allogeneic haematopoietic stem cell transplant. Our diagnoses have stratified 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 thrombocytopenia due to variants in *ACTN1*, *CYCS* or *TUBB1*. Our discoveries have also improved the accuracy of prognosis: we found that mutations in *BMPR2*[17] and *EIF2AK4*[18] carry a poorer-than-average prognosis in Pulmonary Arterial Hypertension (PAH) and we plan prognostication studies of four genes (*ATP13A3*, *AQP1*, *GDF2* and *SOX17*) we recently reported as aetiological[19].

Measurement of quantitative intermediate phenotypes can elucidate genetic aetiology in difficult-to-diagnose patients. We considered patients with a clinically determined absence of a protein encoded by a DGG in whom we had called only one explanatory allele and examined the corresponding WGS read alignments for evidence of a variant in compound heterozygosity. Two patients with a severe unexplained bleeding disorder due to the absence of αIIbβ3 integrin on their platelet membranes carried complex variants in intron 9 of *ITGB3*: one carried a tandem repeat and the other an SVA retrotransposon, not called by structural variant callers, but which generated an excess of improperly mapped reads and was confirmed by long-read sequencing (Extended Data Fig. 8a–e). A third patient had a severe haemolytic anemia due to absence of the RhD and RhCE proteins on her red cell membranes, caused by a large tandem repeat in *RHAG* (Extended Data Fig. 8f).

Genetic associations with rare diseases

Several cases with similar aetiologies are typically needed for discovery in rare disease genetics. Cases can be aggregated across siloed studies using Matchmaker Exchange (MME)[20]. We identified novel aetiologies for *SLC18A2*[21] and *WASF1*[22] using MME (Supplementary Information). However, in a study of a large unified health system, it is possible to make discoveries by statistical analyses of patient collections.

We applied BeviMed[23] to identify associations between genes and rare diseases under various modes of inheritance (Supplementary Information). We labelled groups of cases with a common tag if their phenotypes were *a priori* judged compatible with a shared aetiology (Supplementary Table 3). The number of unrelated cases with each tag ranged from three for Roifman syndrome to 1,101 for PAH. We analysed each gene–tag pair independently and considered a posterior probability (PP) of association >0.75 to be strong evidence supporting a genetic aetiology. To account for correlation between tags, we recorded only one association per gene, corresponding to the tag for which the highest PP of association was obtained. Conditional on gene causality for

a tag, BeviMed reported PPs over the mode of inheritance, the molecular consequence class of variants mediating disease risk (e.g. 5' UTR variants or predicted loss-of-function variants) and the pathogenicity of each specific variant.

We recorded strong evidence for association between 95 genes and 29 tags. The distribution of PPs implied a posterior estimate for the positive predictive value (PPV) of 93%. The 95 genes included 68 established DGGs, 11 DGGs discovered since 2015[13,24,19,25,26,9,27,28,29] and 16 candidates requiring further investigation (Fig. 2d; Supplementary Table 3). Thus, 79 of the 95 associations are confirmed, setting a lower bound on the true PPV of 83%, which is broadly in line with an ancestry-controlled statistical estimate of the study-wide PPV of 79% (Supplementary Information). We estimated that 611.3 cases can be explained by rare variants in the 79 confirmed genes, 115.6 of which are explained by the association between *BMPR2* and PAH. Associations with 51 of the 95 genes relied solely on evidence from singleton variants, showing the power of joint statistical modelling of rare variants. Only three of the unconfirmed associations relied on evidence from alleles carried by more than one case, demonstrating the robustness of the results to cryptic relatedness. For one gene (*GP1BB*), the mode of inheritance inferred by BeviMed differed from that established in the literature, challenging long-held assumptions[30]. These results and other findings from this project[19,22,13,31,32,8,22,9,33,10,34] show that a unified analysis of homogeneously collected genetic and phenotypic data from a large phenotypically heterogeneous rare disease cohort is a powerful approach for genetic discovery.

Polygenic and rare variant associations with the tails of a quantitative trait

Several heritable 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. Alleles with large effects on a quantitative trait predispose carriers to lie in the extreme tails and hence to negative selection pressure. Consequently, such alleles are rare. We sought to identify genes likely to mediate RBC pathologies by WGS of UK Biobank participants in the tails of a univariate quantitative phenotype, computed from RBC full blood count (FBC) traits to optimise rare variant heritability. We derived the univariate phenotype from the joint distribution of estimated effect sizes from GWAS associations between variants with MAF <1% and four RBC FBC traits[6] (Fig. 4a). We sequenced 764 participants, 383 of which were in the left tail of the phenotype, corresponding to a low RBC count (RBC#) and a high mean cell volume (MCV), and 381 of which in the right tail of the phenotype, corresponding to a high RBC# and a low MCV (Fig. 3b–c).

The distribution of an RBC FBC GWAS derived polygenic predictor of the phenotype exhibited left and right shifts from the population distribution in the respectively named tails (Fig. 3d). However, these shifts were less strong than predicted by Gaussian variance components modelling, a discrepancy which might be partly explained by rare alleles generating excess density in the tails (phenotype kurtosis=6.9). A WGS GWAS of an ordinal outcome (left tail, unselected, right tail) did not yield novel associations. Consequently, we treated each of the tail groups as a set of cases in a BeviMed analysis, identifying 12 genes with PP of association greater than 0.4, a liberal threshold (Fig. 3e). *HBB* and *TFRC* can be considered causal, as known mutations cause microcytic anaemias. Other genes, including *CUX1* and *ALG1*, are plausible

candidates. These results (Supplementary Table 3) indicate that the analysis of quantitative extremes in apparently healthy population samples may identify medically relevant loci[6,35].

Aetiological variants in regulatory elements (REs)

Rare variants in REs can cause disease by disrupting transcription or translation[36,37]. Recent work suggests that, at least in neurodevelopmental disorders, a few percent of cases are attributable to *de novo* non-coding SNVs in REs active in relevant tissues[38]. Larger variants may be more disruptive to REs than SNVs. We searched for aetiological variants, including large deletions, in the REs of 246 DGGs implicated in recessive haematopoiesis-related disorders (Supplementary Information). Firstly, we defined a set of active REs, a 'regulome', for each of six haematological cell types, by merging transcription factor binding sites identified by ChIP-seq with genomic regions called by RedPop, a new detection method exploiting the negative covariance of ATAC-seq and H3K27ac ChIP-seq coverage in REs (Supplementary Information). We linked the REs to genes using genomic proximity and promoter capture Hi-C[39]. Secondly, we assigned each regulome to one or more of the BPD, PID and Stem Cell and Myeloid Disorders (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 linked to a DGG of the domain of the case and active in a relevant cell type. We also searched for deletions meeting these criteria that were in compound heterozygosity with a rare coding variant in a DGG linked to the deleted element. These approaches 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*, *Nature* 2020), a boy with autism spectrum disorder and thrombocytopenia carrying a hemizygous deletion of a *GATA1* enhancer, and a patient with several autoimmune-mediated cytopenias carrying a homozygous deletion of intronic CTCF binding sites[40] of *LRBA*.

The X-linked variant carried by the boy with autism deleted a *GATA1* enhancer and exons 1-4 of *HDAC6* (Extended Data Fig. 9). He had a persistently low platelet count ($52x10^9$/l), an elevated mean platelet volume (15.1fl) and normal RBC parameters except for mild dyserythropoiesis. Electron microscopy showed lower than usual platelet α-granule content. Stem cell culture recapitulated poor platelet formation by megakaryocytes. These symptoms are typical of patients carrying a pathogenic coding *GATA1* allele[41]. His platelets contained abnormally low GATA1 (Fig. 4g), consistent with weak transcription due to deletion of the enhancer[42]. HDAC6 deacetylases Lys40 of α-tubulin, which localises in polymerized microtubules[43]. The absence of HDAC6 was accompanied by elevated acetylated α-tubulin in platelets (Fig. 4e), concordant with *Hdac6* knockout mice[44], which have aberrant acetylation leading to bleeding44 and altered emotional behaviour[45]. 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* presented with an autoantibody-mediated pancytopenia due to a loss of tolerance for multiple autoantigens, which is characteristic of impaired *LRBA* function[46].

We adapted the approach to search for pathogenic non-coding SNVs. We focussed on SNVs with a CADD[47] score >20 in compound heterozygosity with a high-impact coding variant in the

assigned DGG. This approach identified two potentially aetiological SNVs in elements assigned to *AP3B1* and *MPL*. We studied the latter mutation (chr1:43803414 G>A), carried by a 10 year old boy, in more detail (Extended Data Fig. 10). *MPL* encodes the receptor for the megakaryocyte growth factor thrombopoietin[48]. Loss of *MPL* causes chronic amegakaryocytic thrombocytopenia[49]. The SNV was in a megakaryocyte-specific RedPop-identified RE (Extended Data Fig. 10). It had CADD=21.8, was absent from gnomAD and was in compound heterozygosity with a deletion of exon 10 of *MPL*. The mutant allele was associated with 50% reduced promoter activity, leading to significant reduction in platelet MPL levels. In contrast to *MPL*-null patients[50], who are severely thrombocytopenic because their bone marrow is almost devoid of megakaryocytes, the patient had platelet counts of $45x10^9$/l and a marrow that was only moderately depleted of megakaryocytes. As the regulatory SNV does not abolish *MPL* transcription completely, the boy has a milder clinical phenotype than *MPL*-null cases.

Discussion

The resolution of unknown rare disease aetiologies will be hastened by the standardisation and integration of clinical testing and research on a national scale. The NHS in England plans to increase provision of WGS-based diagnostics from 8,000 to 30,000 samples per month. To achieve this, it has reduced the number of clinical genomics laboratories to seven and introduced unified staff training in WGS, informatics and genomics. The development of statistical methodology to interpret the new data and participant consent to recall for follow up experiments will be of critical importance. Additionally, long-read sequencing may be needed to overcome the difficulty of calling complex structural variants by WGS. We have initiated WGS of UK Biobank participants to identify rare variant associations with participants in the extreme tails of a quantitative phenotype who are typically excluded from GWAS studies. These associations can identify genes mediating Mendelian pathologies. We have also shown that epigenetic data on cell types mediating aetiology, combined with WGS, can identify REs harbouring pathogenic non-coding mutations. The exploration of regulatory variation is a promising focus for future research and clinical intervention.

References

1. Ferreira CR. The burden of rare diseases. Am J Med Genet A. 2019 Jun;179(6):885-892
2. 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
3. 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
4. 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
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. 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
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. 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
18. 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
19. Graf S, Haimel M, Bleda M, Hadinnapola C, Southgate L, Li W, et al. Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. Nat Commun. 2018 Apr 12;9(1):1416

20. 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
21. Padmakumar M, Jaeken J, Ramaekers V, Lagae L, Greene D, Thys C, et al. A novel missense variant in SLC18A2 causes recessive brain monoamine vesicular transport disease and absent serotonin in platelets. JIMD Rep. 2019 May;47(1):9-16
22. 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
23. 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
24. Merico D, Roifman M, Braunschweig U, Yuen RK, Alexandrova R, Bates A, et al. Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. Nat Commun. 2015 Nov 2;6:8718
25. Ananth AL, Robichaux-Viehoever A, Kim YM, Hanson-Kahn A, Cox R, Enns GM, et al. Clinical Course of Six Children With GNAO1 Mutations Causing a Severe and Distinctive Movement Disorder. Pediatr Neurol. 2016 Jun;59:81-4
26. Horn D, Siebert E, Seidel U, Rost I, Mayer K, Abou Jamra R, et al. Biallelic COL3A1 mutations result in a clinical spectrum of specific structural brain anomalies and connective tissue abnormalities. Am J Med Genet A. 2017 Sep;173(9):2534-2538
27. Khan SY, Ali S, Naeem MA, Khan SN, Husnain T, Butt NH, et al. Splice-site mutations identified in PDE6A responsible for retinitis pigmentosa in consanguineous Pakistani families. Mol Vis. 2015;21:871-82
28. Petrovski S, Kury S, Myers CT, Anyane-Yeboa K, Cogne B, Bialer M, et al. Germline De Novo Mutations in GNB1 Cause Severe Neurodevelopmental Disability, Hypotonia, and Seizures. Am J Hum Genet. 2016 May 5;98(5):1001-1010
29. Akawi N, McRae J, Ansari M, Balasubramanian M, Blyth M, Brady AF, et al. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. Nat Genet. 2015 Nov;47(11):1363-9
30. 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
31. 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
32. 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
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. Lentaigne C, Greene D, Sivapalaratnam S, Favier R, Seyres D, Thys C, et al. Germline mutations in the transcription factor IKZF5 cause thrombocytopenia. Blood. 2019 Dec 5;134(23):2070-2081

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. 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
37. 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
38. 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
39. 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
40. Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet. 2014 Apr;15(4):234-46
41. 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
42. 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
43. 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
44. 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
45. 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
46. 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
47. 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
48. 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
49. 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
50. Ballmaier M, Germeshausen M. Congenital amegakaryocytic thrombocytopenia: clinical presentation, diagnosis, and treatment. Semin Thromb Hemost. 2011 Sep;37(6):673-81

Main Figure Legends

Fig. 1. Study overview. a, Schematic of the diagnostic and research processes. Blue: patients are recruited, HPO and pedigree data are collected, DNA is extracted and sequenced, WGS data are transferred for QC and variant prioritisation. Green: variants are assessed and diagnoses are returned. Orange: the complete data are analysed by association and co-segregation to identify aetiological variants, disease-mediating genes and regulatory regions; functional studies, and model systems are used to study disease mechanisms. **b**, Histograms of read coverage across the 13,037 participants, stratified by WGS read length (100bp,125bp, 150bp). **c**, Projection of genetic data of the 13,037 participants onto the first two principal components of variation in the 1000 Genomes Project and barplot of the distribution of participant ancestry. **d**, Histograms illustrating the observed minor allele frequency (MAF) distribution of variants called in the MSUP (n=10,259), stratified by type (SNV or indel). Variants are labelled novel if they were uncatalogued in 1000 Genomes, UK10K, TOPMed, gnomAD and HGMD Pro. MAC: minor allele count. **e**, The frequency of novel variants stratified by the ancestry groups in which they were observed (yellow: present, navy: absent). **f**, Barplot of the sizes of genetically determined networks of closely related individuals across the 13,037 participants. Inset: distributions of network sizes for each rare disease domain. Here and throughout, SMD: Stem cell and Myeloid Disorders; GEL: 100,000 Genomes Project–Rare Diseases Pilot; CSVD: Cerebral Small Vessel Disease; NDD: Neurological and Developmental Disorders; LHON: Leber Hereditary Optic Neuropathy; PID: Primary Immune Disorders; EDS: Ehler-Danlos and Ehler-Danlos-like Syndromes; BPD: Bleeding, Thrombotic and Platelet Disorders; MPMT: Multiple Primary Malignant Tumours; SRNS: Steroid Resistant Nephrotic Syndrome; HCM: Hypertrophic Cardiomyopathy; NPD: Neuropathic Pain Disorders; PAH: Pulmonary Arterial Hypertension; PMG: Primary Membranoproliferative Glomerulonephritis; IRD: Inherited Retinal Disorders; ICP: Intrahepatic Cholestasis of Pregnancy.

Fig. 2. MDT reporting and genetic associations with rare diseases. a, Barplot of the frequency of probands by domain (top); barplot of the frequency of probands with each top-level HPO phenotype abnormality term (right). The heat map shows the proportion of probands in each domain assigned a particular top-level HPO term (shown abbreviated). **b**, Heat map of the number of DGGs shared by pairs of domains (left). Pre-screening level for each domain indicated in red (full), blue (partial) or green (none). Barplot of the proportion of cases for which a clinical report was issued (right). **c**, Frequency of reports issued by DGG ordered inversely by count. Dashed lines indicate quartiles of the count distribution. Inset: barplot of the frequency of distinct clinically reported variants stratified by variant type. The colours in each bar indicate the proportion of variants which are known/novel (as defined in the main text). **d,** Bevimed PPs for genetic association >0.75. The colours indicate whether the associations were established in the scientific literature prior to 2015, since 2015, or remain unconfirmed.

Fig 3. Genetic associations with the tails of an RBC trait. a, Histograms/scatterplots summarising the distribution of the additive effects of 65 red cell GWAS variants (MAF <1%) on four RBC traits (acronyms in Supplementary Information). The red square shows the bivariate distribution used to develop the selection phenotype. The red line was estimated by Deming regression. **b**, The (standardised) distribution of the selection phenotype (panels showing different y-axis ranges) in post-menopausal female and male European ancestry UK Biobank participants without record of illness/treatment known to perturb RBC indices (grey) and selected for WGS (turquoise/salmon). The area of the histogram represents the number of

contributing individuals in thousands, *N*=316,739. Many participants in the tails were unselected (see Supplementary Information). **c**, Scatterplots showing the distribution of RBC# and MCV in UK Biobank post-menopausal females (left) and males (right). The ellipsoids are contours of kernel density estimates. Open circles: participants ineligible for selection. Non-European ancestry thalassemias may explain the concentration with high RBC#/low MCV. Coloured circles: WGS'd participants. **d**, The boxplots summarise the distribution of a polygenic score for the selection phenotype in the 383/381 individuals selected from the left/right tails and in 508 European participants in domains other than UKB with pathology explained by rare variants (Unselected). The centre mark and lower and upper hinges of the boxplots respectively indicate the median, 25th and 75th percentiles. Outliers beyond 1.5 times the interquartile range from each hinge are shown. The violin plots show the expected distribution of the polygenic score under a Gaussian variance components model, conditional on the proportion of phenotypic variance explained by the score and the tail selection thresholds. **e**, BeviMed PPs for genetic association of each tail (distinguished by colour), for genes with PPs >0.4. Boldface indicates strong concordant biological evidence.

Fig. 4. Causal variants in regulatory elements. a, Top to bottom: X chromosome ideogram; read coverage of H3K27ac ChIP-seq (green) and ATAC-seq (orange) in MKs; the smoothed covariance (Cov) between MK H3K27ac ChIP-seq and MK ATAC-seq coverages, used to call regulatory elements (overlying coral rectangles); pink segments indicate regions in which the locally normalised ATAC-seq coverage exceeds the locally normalised H3K27ac ChIP-seq coverage (Supplementary Information); the corresponding three tracks and overlays for EBs; gene exons in orange; the *GATA1* enhancer and the large deletion in the proband as horizontal bars, respectively. A regulatory element overlapping the enhancer was identified by RedPop in MKs and EBs but not in the other four cell types (tracks for which not shown). The deleted element binds transcription factors characteristic of the MK lineage: FLI1, GATA1/2, MEIS1, RUNX1 and TAL1 (binding not shown). **b–d**, P: propositus, M: mother, F: father; C1, C2 and C3 are controls. **b–c**, m: marker. **b**, Representative immunoblots for total platelet lysates for the indicated proteins and individuals (n=2). **c**, Representative example of n=3 replicate immunoblots of total platelet lysates using two GATA1 antibodies. **d**, Dot plots of GATA1 protein quantifications (as in **c**). The underlying violin plots show posterior predictive densities for the distribution of standardised GATA1 expression. The 90% credible intervals for the ratio of N6-measured expression in F, M, P to the geometric mean in controls were (0.86, 1.45), (0.35, 0.59) and (0.37, 0.62) respectively; correspondingly, for NF-measured expression (0.80, 1.05), (0.51, 0.67) and (0.45, 0.60).

Methods

Enrolment, research ethics and consent Study participants were enrolled by one of three mechanisms between December 2012 and March 2017 under the overall coordination of the National Institute for Health Research BioResource (NBR) at the University of Cambridge. Patients with rare diseases and their close relatives were enrolled into 15 rare disease domains approved by the Sequencing and Informatics Committee of the NBR. Enrolment of controls was coordinated by the University of Cambridge. Enrolment in the GEL domain was coordinated by Genomics England Ltd. Enrolment in the UK Biobank domain was jointly coordinated by the NBR and UK Biobank Ltd[2]. Rare disease domain participants were recruited mainly at NHS Hospitals in the UK, but also at hospitals overseas (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 ethics for other REC-approved studies. Obtaining consent for overseas samples was the

responsibility of the respective principal investigators at the hospitals where enrolment took place. 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 in the domain descriptions (Supplementary Information). The clinical and laboratory phenotype data were captured through case report forms by paper questionnaires or by online data capture applications and deposited in the NBR study database. Online data capture allowed for the free entry of HPO terms[51] 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 NBR rare disease domains is depicted in Extended Data Fig. 1c.

DNA sequencing Pre-extracted DNA samples or EDTA treated whole blood samples were delivered to the NBR laboratory at Cambridge, where DNA was extracted from the whole blood. Samples were tested for adequate concentration (Picogreen), quality controlled (QC'd) for DNA degradation (gel electrophoresis) and purity (OD 260/280; 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 to ensure that the concentration of the DNA was >30 ng/μl 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 prior to data delivery. 0.5 μg of each DNA sample was fragmented using Covaris LE220 (Covaris Inc., Woburn, MA, USA) to obtain an average size of 450bp 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, 125bp and 150bp (377, 3,154 and 9,656 samples, respectively). Samples sequenced with 100bp and 125bp reads utilised three and two lanes of an Illumina HiSeq 2500 instrument, respectively, while 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 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 an MSUP of 10,259 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 the 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 MSUP, ethnicity and sex karyotype. Variants were annotated with 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 list of DGGs 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 DGGs 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[52] or designation as canonical in Ensembl. Variants (SNVs, indels) were shortlisted if (i) their MAF in control populations[53] 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[54] 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 DGG relevant to the patient’s disease. Variants with more than three alleles or a MAF ≥10% in the cohort were discarded respectively, to guard against errors in repetitive regions and to 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 prevented erroneous filtering of causal variants carried at elevated frequencies by 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 per person, while retaining almost all 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 gender and genetically inferred sex, ancestry, relatedness, and consanguinity level) were transferred to Congenica Inc (Cambridge, United Kingdom) for visualisation in the Sapientia™ web application during MDT meetings. MDTs comprised 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 uncertain significance (VUS). Only pathogenic and likely pathogenic variants 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[23] 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 aetiological mechanism. The BeviMed computed PPs can be used to estimate the number of cases attributable to variants in each gene, conditional on gene causality. The methodology is described in further detail in the Supplementary Information and in the original BeviMed publication23. BeviMed was applied gene-wise to infer associations between the genotypes of filtered rare variants and various case/control groupings (tags). For a given gene, only the maximum PP over tags was recorded, to account for correlation between tags.

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 (Supplementary Information). Additionally, for MK and EB, we had access to the following

transcription factor (TF) ChIP-seq data, which were used to call peaks 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)[39] 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 the target gene TSS, and the 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 of platelets was performed as described in ref.[41]. Immunostaining of resting and fibrinogen spread platelets was performed as described in ref.[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 in ref.[55]. 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, Agilent 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[56]), rabbit anti-GAPDH (14C10, Cell Signaling) and anti-β3 integrin (sc-14009; Santa Cruz Biotechnology). The statistical analysis of the GATA1 data is described in the Supplementary Information.

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[57] with the aim of resolving 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 from the 4,835 participants enrolled in the National Institute for Health Research (NIHR) BioResource for the 100,000 Genomes Project–Rare Diseases Pilot can be accessed by application to Genomics England Limited following the procedure outlined at: https://www.genomicsengland.co.uk/about-gecip/joining-researchcommunity/. The genotype data for the 764 UK Biobank samples will be made available through a data release process which is being overseen by UK Biobank (https://www.ukbiobank.ac.uk/). The phenotype data from UK Biobank participants are available from UK Biobank using their access procedures.
The WGS and detailed phenotype data of the remaining 7,348 NIHR BioResource participants can be accessed by application to the NIHR BioResource Data Access Committee at dac@bioresource.nihr.ac.uk. Subject to ethical consent, the genotype data of 6,939 NIHR BioResource participants are also available from the European Genome-phenome Archive (EGA) at the EMBL European Bioinformatics Institute under access procedures managed by EGA. The domain specific accessions are as follows:
BPD: EGAD00001004519, CSVD: EGAD00001004513, EDS (EGAD00001005123), HCM: EGAD00001004514, ICP: EGAD00001004515, IRD: EGAD00001004520, LHON: EGAD00001005122, MPMT: EGAD00001004521, NDD: EGAD00001004522, NPD: EGAD00001004516, PAH: EGAD00001004525, PID: EGAD00001004523, PMG:
EGAD00001004517, SMD: EGAD00001004524, SRNS: EGAD00001004518. 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[58], B (ERR1043004, ERR1043129, ERR928206, ERR769436), EB (EGAD00001002377), MK (EGAD00001002362), MON (ERR829362 (ERS257420), ERR829412 (ERS222466), ERR493634 (ERS214696)), rCD458. ATAC-seq: aCD4 (GSE124867), B (SRR2126769 (GSE71338)), EB (SRR5489430 (GSM2594182)), MK (EGAD00001001871), MON (EGAD00001006065), rCD4 (GSE124867). Reported alleles and their clinical interpretation have been deposited with ClinVar under the study names "NIHR_Bioresource_Rare_Diseases_13k", "NIHR_Bioresource_Rare_Diseases_Retinal_Dystrophy",
"NIHR_Bioresource_Rare_Diseases_MYH9" and "NIHR_Bioresource_Rare_Diseases_PID". MDT-reported alleles and their clinical interpretation have been deposited in ClinVar (under the name “NIHR Bioresource Rare Diseases”) and DECIPHER.

References (continued)

51. 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
52. 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

53. 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
54. 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
55. 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
56. 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
57. 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
58. 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
59. Wijgaerts A, Wittevrongel C, Thys C, Devos T, Peerlinck K, Tijssen MR, et al. The transcription factor GATA1 regulates NBEAL2 expression through a long-distance enhancer. Haematologica. 2017 Apr;102(4):695-706

Extended Data Figure Legends

Extended Data Fig. 1: Demographic and phenotypic characteristics.

a, Barplot of the number of enrolments at the 40 hospitals with at least 20 enrolled participants. The heat map shows the proportion of enrolments per domain at each of the 40 hospitals. Hospital IDs are detailed in Supplementary Table 1. **b**, Top: boxplot of age at recruitment for all probands in the 15 rare disease domains, GEL and UK Biobank; Bottom: stacked barplot of the counts of probands in each domain with and without an available age at recruitment. **c**, Histograms of the number of HPO terms appended to affected probands for 13 of the rare disease domains.

Extended Data Fig. 2: Flowchart of the bioinformatic data processing.

Flowchart describing the processing of samples and variants. Beginning at the top left, all samples were checked for data quality (see **Extended Data Fig. 3**). Quick kinship and sex checks were regularly performed to ensure consistency with reported sex and family information. Samples failing QC, samples with clearly discordant sex data and the sub-optimal replicates of repeated samples were removed before further analysis (pink boxes). Sex chromosome karyotypes, ethnicities, and relatedness/family trees were computed on these filtered samples (orange boxes) and variants were recalled for those samples with X/Y-chromosome ploidies different to those automatically predicted by the quick checks. After variant normalisation, variant calls were loaded into HBase and merged, and summary statistics were calculated, stratified by technical factors (100, 125, and 150bp) and ancestry (e.g., unrelated African) (green boxes). Variant-specific minimum OPRs were calculated and used to filter inaccurately genotyped variants (see **Extended Data Fig. 4**). Finally, variants were

annotated in HBase with predicted consequence information and information from external databases, including allele frequencies (e.g., gnomAD) (blue box).

Extended Data Fig. 3: Sample QC, sex chromosome karyotyping and ancestry inference. **a**, Boxplot of the percentage of QC-passing autosomal bases (n=13,137; 4 exclusions highlighted). **b**, Boxplot of the percentage of common SNVs that failed QC (n=13,137; 2 exclusions highlighted). **c**, Batch-specific boxplots of Ts/Tv ratios (n=377 for 100bp samples; n=3,154 for 125bp samples; n=9,656 for 150bp samples; 3 exclusions highlighted). **d**, Boxplot of FREEMIX values representing sample contamination (n=13,137; 8 exclusions highlighted). **a–d**, Excluded samples are marked in red and labelled with an integer. Three samples were excluded due to failing more than one of the four QC checks (samples 5,12 and 14). The centre line of each boxplot indicates the median and the lower and upper hinges indicate the 25th and 75th percentiles respectively. The vertical line of each boxplot extends to 1.5 times the interquartile range from each hinge. **e**, H-ratios for 13,037 samples and predicted initial sexes. **f**, Scatterplot of ratios of X/Auto and Y/Auto coloured by the initial sex calls and showing the five sex karyotyping gates. **g**, Scatterplot of ratios of X/Auto and Y/Auto coloured by the final sex chromosome karyotype. Circles indicate samples falling within a sex karyotyping gate and triangles indicate samples falling outside all sex karyotyping gates. 1: confirmed XYY case; 2–4: confirmed XY female cases; 5, 6: confirmed XO cases; 7: confirmed XO case, this sample has some part of the second X chromosome present; 8–10: samples with large part of the X chromosome missing; 11–12: samples with multiple deletions on the X chromosome; 13: sample with two almost identical X chromosomes (normal karyotype); 14: confirmed XXY case. **h**, Projection of the 13,037 samples, shown as round circles, onto the 1000 Genomes derived PCAs. The 1000 Genomes samples are shown as diffuse points underneath in colour. **i**, Projection of the 13,037 samples, shown as round circles, coloured by assigned population. **j**, Barplot showing the number of individuals assigned to each population. The percentages are shown above each bar. NFE: Non-Finnish European; SAS: South Asian; AFR: African; EAS: East Asian; FIN: Finnish. **k**–**m**, Distribution of the sizes of small insertions (indel size >0) and small deletions (indel size <0) in coding regions, non-coding regions and non-coding regions excluding repetitive regions, specifically, the RepeatMasker track from the UCSC table browser and the Tandem Repeats Finder locations from the UCSC hg19 full data set download. In coding regions, natural selection against frameshift variants results in a systematic depletion of indel sizes that are not a multiple of 3bp. In non-coding regions, there is a slight excess of indel sizes that are a multiple of 2bp, but this pattern is almost indiscernible if repetitive regions are excluded.

Extended Data Fig. 4: Variant QC.
a–**c**, The proportion of HWE *P*-values <0.05 amongst 8,510 unrelated Europeans across different AF bins is shown for SNVs, small deletions and small insertions. Boxplots of the number of variants in each OPR and AF bin are shown in the bottom sub-panels. **d**, Table showing the possible combinations of genotypes in a pair of samples. The variables in the cells represent numbers of variants (see Supplementary Information for use). **e–g**, Three measures of genotype concordance (Supplementary Information) for pairs of duplicates and twins with results from 100, 125 and 150bp reads shown from left to right. **e**, Distribution of mutual non-

reference concordance in pairs of duplicates and twins. **f**, Probability of having a heterozygous genotype in a sample, given its duplicate/twin has this heterozygous genotype. **g**, Probability of having a non-reference homozygous genotype in a sample, given its duplicate/twin has this homozygous genotype. In panels **e**–**g**, the mean number of variants of each type used to compute concordance is shown in brackets after the variant type label. In panels **f**–**g**, red and blue colours represent distribution of the lowest and highest of the two probabilities (sample 1 compared to sample 2 and sample 2 compared to sample 1) in a pair of duplicates/twins.

Extended Data Fig. 5: Breakdown of genetic variants by their predicted primary consequence.
a, Counts of SNVs and indels in various Variant Effect Predictor consequence classes shown on logarithmic scales with exact numbers above each bar. Variants in the green bars are subdivided into more granular regions of genome space in the following panel in a recursive manner from left to right. Categories have been chosen to represent the most severe transcriptional consequences at each stage: i.e., from left, overall genome space, within genes, exonic parts of genes, and protein coding regions. **b**, Count of MDT SNVs and indels in various consequence classes with exact numbers above each bar. A star denotes a super-category with missense_variant including missense_variant or missense_variant&splice_region_variant; splice including splice_acceptor_variant, splice_donor_variant, splice_donor_variant&coding_sequence_variant or splice_region_variant or splice_region_variant&intron_variant; stop_gained including stop_gained, stop_gained&splice_region_variant or stop_gained&splice; frameshift variant including frameshift_variant, frameshift_variant&splice_region_variant or retained_intron; inframe indel including inframe_deletion or inframe_insertion.

Extended Data Fig. 6: Breakdown of diagnostic reports by domain.
a, Number of reports issued for the 11 rare disease domains that issued clinical reports. Each panel corresponds to a domain, the title denotes the domain acronym and number of reports issued. PMG and EDS are not shown because no reports were issued for cases in these domains. The panels are arranged in decreasing order of the maximum number of within domain reports issued for a single DGG. Each point represents a gene featuring in at least one report for a case in the domain. The genes with the most reports issued for each domain are labelled. Full details of all the reports issued are given in Supplementary Table 2. **b**, Barplots of the number of distinct reported autosomal short variants (SNVs and indels) for each domain in different gnomAD/TOPMed allele frequency/count bins in samples of European ancestry, broken down by rare disease domain (left) and by mode of inheritance (right). MAC: Minor allele count. MAF: Minor allele frequency. The domain acronyms are defined in Supplementary Table 1. MOI: Mode of inheritance. AD: Autosomal dominant. AR: Autosomal recessive. For a given position and minor allele, the combined MAF was defined as the sum of allele counts divided by the sum of allele numbers over gnomAD and TOPMed. The first bin in the plots (MAC=0) corresponds to variants not observed in either gnomAD or TOPMed. **c**, Some genes featured in reports for cases in more than one domain. The heatmap shows the number of reports featuring these genes, broken down by domain.

Extended Data Fig. 7: Comparison of WGS and WES for genetic testing.
For each of four WES datasets – UK Biobank, INTERVAL, Columbia (IDTERv1) and Columbia (Roche) – four groups of panels (labelled **a–d)** are shown, each of which corresponds to a different comparison of coverage characteristics, as follows. **a**, Scatterplot of WGS vs WES mean coverage at 116,449 sites of diagnostic importance (Supplementary Information). The red axes show the threshold for clinical reporting and the numbers of variants in each quadrant are indicated. **b**, Scatterplot of WGS vs WES coverage of the MDT-reported known (turquoise) and novel (salmon) SNVs and indels in autosomal diagnostic-grade genes. **c**, Barplots of the percentage of samples with coverage below the threshold for clinical reporting, with variants ranked on the x-axis by their corresponding values on the y-axis within the WGS and WES datasets. The barplots corresponding to WGS are superimposed on those corresponding to WES. The inset panel shows the mean percentage of individuals covered below 20X by WGS and WES in a zoomed-in view. **d**, Vertical bars indicate the 1%–99% coverage range in WGS (turquoise) and WES (salmon), with variants ranked by the mean coverage values within the WGS and WES datasets.

Extended Data Fig. 8: Cases with protein-null phenotypes.
a, Alignments in the *ITGB3* locus for a Glanzmann's thrombasthenia case with a premature stop (blue bar) and a tandem repeat revealed by improperly mapped read pairs. **b**, Number of improperly mapped read pairs in the 9th intron of *ITGB3* in 6,656 samples sequenced by 150bp reads before (light grey dots) or after (dark grey squares) the data freeze. The Glanzmann's thrombasthenia cases with the tandem repeat and with the SVA insertion, and the carrier mother of the latter, are highlighted. **c–d,** Alignments in the *ITGB3* locus for the Glanzmann's thrombasthenia proband (**c**) and his mother (**d**) with a p.T456P variant for the proband (blue bar) and an insertion revealed by an excess of mapped reads for the 9th intron for the proband and his mother. **e,** Top: long-read alignments for the PCR-amplified *ITGB3* DNA from the Glanzmann's thrombasthenia proband covering the element with excess reads. Downstream Read Elements (DRE) starts are represented in the histogram. Bottom (from left to right): the Glanzmann's thrombasthenia pedigree (A: proband, B: mother, C: grandmother) with the flow cytometric measurements of platelet GPIIbIIIa expression indicated as percentage of normal levels and genotypes; confirmation of the insertion by gel electrophoresis of PCR products covering the insertion; diagram of the inserted SVA (Alu, SINE-VNTR-Alu) retrotransposon element (insSVA). **f**, Alignments in the *RHAG* locus of the Rhesus-null case with a splice donor variant (blue bar) and a tandem duplication revealed by improperly mapped read pairs.

Extended Data Fig. 9: Deletion of a *GATA1* enhancer and part of the *HDAC6* open reading frame and its effects.
a, WGS reads show a hemizygous 4108 bp deletion (X:48,659,245-48,663,353) in the proband. **b–k**, P: proband, F: father, M: mother, C: control. **b**, Pedigree of the proband with thrombocytopenia and autism. PLT: platelet count, MPV: mean platelet volume, PDW: platelet distribution width, ASD: autism spectrum disorder, ID: Intellectual disability. **c**, Left: representative image of n=2 rounds of gel electrophoresis showing presence and absence of short PCR amplicons using primers flanking the deletion. Right: control PCR. '-': no DNA added. **d**, Sanger sequencing of PCR fragments (shown in panel **c**) with primers flanking the 4801 bp

deletion. The red arrow points to the position of the fusion between bp 48,659,245 and bp 48,663,353. **e**, Electron microscopy images (n=1 sample preparation per subject) show that platelets of P were larger and rounder than those of C (unrelated healthy control), and in some instances had abnormal semi-circular empty vacuoles (*) and a depletion of alpha granules. Marker is 1.5 μM. **f–g**, Analysis of electron microscopy images (n=21, 14, 21, 20 and 20 platelets in samples E1, E2, E3, C and P respectively); E1, E2, E3 and C are controls; the data for E1, E2 and E3 were obtained from ref.[59]. Dot plots of platelet area (μm^2) and the alpha granule count per unit area ($1/(\mu m^2)$), computed using ImageJ. The underlying violin plots show posterior predictive densities for the mean platelet area/granule density in controls and in P under a mixed model accounting for intra-individual correlation. The 90% credible intervals for the ratio of the mean in P to the mean in controls were (1.38, 2.03) and (0.15, 0.87) for area and granule density respectively. The abnormalities of platelet area and alpha granule density in the proband are very similar to the defects described in *GATA1* deficiency59. **h**, Platelet spreading analysis using SIM (Z-stacks) and staining for F-actin (red) and acetylated α-tubulin (green). Washed platelets were spread on fibrinogen for 0 (basal condition), 30 and 60 minutes for control, father, mother and proband. This experiment was performed once and representative images are shown. Marker is 1.5 μM. **i**, Platelet analysis using structured illumination microscopy (SIM) and staining for acetylated α-tubulin (green) before spreading (time point 0). The microtubule marginal bands are clearly disturbed and hyper-acetylated for non-activated platelets of the proband while being normal for the father and mother. This experiment was performed once. Marker is 1.5 μM. **j**, Dot plots of the mean ImageJ-quantified platelet area in groups of n=5 images of F-actin stained platelets at three time points (0, 30 and 60 minutes after spreading on fibrinogen) for C, F, M and P. There was no evidence of a difference in the mean of the mean platelet area in F and M compared to C within time points (P >0.12 for all six two-sided Welch t-tests), so F and M were treated as controls in subsequent modelling. The underlying violin plots show posterior predictive densities for the mean platelet area at time points 30 and 60 under a mixed model accounting for intra-individual correlation. The 90% credible intervals for the ratio of the mean in P to the mean in controls were (1.87, 4.56) and (2.07, 3.61) at time points 30 and 60 respectively. **k**, The upper sub-panels show representative images from the control and the proband. In the latter, large MKs are present but proplatelet formation is strongly reduced. The lower sub-panel shows the quantification of proplatelet formation by MKs at day 12 of differentiation from cultures performed in duplicate for each individual. 10 images per culture were used to compute the % proplatelet-forming MKs per individual, shown as dot plots. There was no evidence of a difference in the mean of the percentage between F and C (P=0.90, two-sided Welch t-test), so F was treated as a control in subsequent modelling. The underlying violin plots show posterior predictive densities for the % proplatelet-forming MKs in controls, in M and in P under a mixed model accounting for intra-individual correlation. The 90% credible intervals for the odds ratio of the mean in M and P to the mean in controls were (0.32, 0.46) and (0.18, 0.28) respectively. **l**, Day 12 differentiated MKs for the indicated individuals were stained for F-actin (red) and HDAC6 (green). Upper two panels: HDAC6 is expressed in the cytosol and is trafficked to proplatelets as shown in MKs for the control and father (bold arrows). Middle two panels: MKs for the proband show no HDAC6 expression while cultures from the mother contain a mixture of MKs that are positive and negative (15 of the 45 MKs) for HDAC6 expression. Lower two panels show only the HDAC6

staining. This experiment was performed once. **m**, Day 12 differentiated MKs for the indicated individuals were stained for acetylated α-tubulin (green). Highly organised tubulin structures are present in all MKs from the control and father while the patient (47 of the 57 MKs) and mother (16 of the 46 MKs) contain MKs that show signs of tubulin depolymerisation (*). This experiment was performed once.

Extended Data Fig. 10: Thrombocytopenia due to compound regulatory and coding rare variants in *MPL*.

a, 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. **b**, Pedigree for the proband (P) with thrombocytopenia due to a 454bp deletion encompassing exon 10 of *MPL*, which was inherited from the mother (M), and an SNV just upstream of the 5' UTR of *MPL*. **c**, Sanger sequencing traces confirming the presence of the heterozygous SNV in P and its absence in M. **d**, Gel electrophoresis of PCR amplicons covering the deletion confirming presence of the deletion in P and M. The PCR was conducted on two independent samples in P and once in M and the control (wt). **e**, Mean fluorescence intensities (MFI) on the y-axis obtained by the flow cytometric measurement of MPL abundance (CD110) on the membrane of platelets from five unrelated healthy controls (C), M and P. The MFI was normalised with unstained platelets. We fitted a linear regression model with an intercept term representing the mean in C, a coefficient representing the difference in means between M and C (P=0.1828) and a coefficient representing the difference in means between P and C (P=0.0086). Distribution summaries show mean ± s.e.m. where multiple observations are available. **f**, Results of luciferase reporter assays in K562 cells expressed with empty pGL3 vector or after cloning with an MPL promoter fragment containing the wild type G allele (MPL-SNV-G) or the variant A allele (MPL-SNV-A). The measurements were derived from n=4 independent transfection experiments. The P-values were obtained by one-way ANOVA and adjusted for multiple comparisons using Tukey's method. Distribution summaries show mean ± s.e.m.

Author Contributions Details of author contributions can be found in the supplementary file containing the full list of consortium members and working groups.

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 was coordinated for one by the NIHR BioResource and for 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 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 acknowledgement 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 (BRC), 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. 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; AJT is supported by the Wellcome Trust (104807/Z/14/Z) and the NIHR Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University 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 an MRC Clinical Research Training Fellowship (MR/R002363/1); JAS is funded by MRC UK grant MR/M012212/1; AJM received funding from an MRC Senior Clinical Fellowship (MR/L006340/1); CLe received funding from an 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; CWi holds a NIHR Senior Investigator Award; 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; The MPGN/DDD/C3 Glomerulopathy Rare Disease Group contributed to the recruitment and analysis of the PMG cohort; 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. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care or any of the 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.

[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]MRC Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge, UK. [4]NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, UK. [5]Department of Medicine, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [6]British Heart Foundation Cambridge Centre of Excellence, University of Cambridge, Cambridge, UK. [7]Department of Cardiovascular Sciences, Center for Molecular and Vascular Biology, KULeuven, Leuven, Belgium. [8]Cambridge Institute of Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, Cambridge, UK. [9]MRC Clinical Sciences Centre, Faculty of Medicine, Imperial College London, London, UK. [10]Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [11]The Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford, UK. [12]High Performance Computing Service, University of Cambridge, Cambridge, 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]Rare Renal Disease Registry, UK Renal Registry, Bristol, UK. [20]King's College London, London, UK. [21]Department of Paediatric Nephrology, Evelina London Children's Hospital, Guy's & St Thomas' NHS Foundation Trust, London, UK. [22]Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [23]Centre for Haematology, Imperial College London, London, UK. [24]Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge Biomedical Campus, Cambridge, 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]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, UK. [28]School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK. [29]University

Hospitals Bristol NHS Foundation Trust, Bristol, UK. [30]Department of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [31]Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [32]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [33]Department of Paediatrics, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [34]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, UK. [35]Bristol Renal and Children's Renal Unit, Bristol Medical School, University of Bristol, Bristol, UK. [36]Bristol Royal Hospital for Children, University Hospitals Bristol NHS Foundation Trust, Bristol, UK. [37]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [38]UCL Great Ormond Street Institute of Child Health, London, UK. [39]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [40]Moorfields Eye Hospital NHS Foundation Trust, London, UK. [41]UCL Institute of Ophthalmology, University College London, London, UK. [42]Department of Medicine, Imperial College London, London, UK. [43]Institute of Reproductive and Developmental Biology, Surgery and Cancer, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [44]Cancer Research UK Cambridge Centre, Cambridge Biomedical Campus, Cambridge, UK. [45]Illumina Cambridge Limited, Chesterford Research Park, Little Chesterford, Saffron Walden, Essex, UK. [46]Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [47]Department of Renal Medicine, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [48]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. *A list of authors and their affiliations appears in the online version of the paper. ✉e-mail: et341@cam.ac.uk; flr24@cam.ac.uk; who1000@cam.ac.uk.

NIHR BioResource for the 100 000 Genomes Project

Stephen Abbs[49], Lara Abulhoul[50], Julian Adlard[51], Munaza Ahmed[52], Timothy J Aitman[9,10], Hana Alachkar[53], David J Allsup[54,55], Jeff Almeida-King[27], Philip Ancliff[50], Richard Antrobus[56], Ruth Armstrong[24,25,44], Gavin Arno[40,41], Sofie Ashford[2,37], William J Astle[3,4], Anthony Attwood[1,2], Paul Aurora[50], Christian Babbs[32,34], Chiara Bacchelli[38], Tamam Bakchoul[57], Siddharth Banka[58,59], Tadbir Bariana[60,61], Julian Barwell[62,63], Joana Batista[1,2], Helen E Baxendale[5,64,65,66], Phil L Beales[50,67], David L Bennett[11], David R Bentley[45], Agnieszka Bierzynska[35], Tina Biss[68], Maria A K Bitner-Glindzicz[50,67], Graeme C Black[58,59], Marta Bleda[5], Iulia Blesneac[11], Detlef Bockenhauer[50], Harm Bogaard[69], Christian J Bourne[45], Sara Boyce[70], John R Bradley[2,5,25,46,47], Eugene Bragin[71], Gerome Breen[72,73], Paul Brennan[74,75,76], Carole Brewer[77], Matthew Brown[1,2], Andrew C Browning[78], Michael J Browning[79], Rachel J Buchan[80,81], Matthew S Buckland[82], Teofila Bueser[83,84], Carmen Bugarin Diz[20], John Burn[76], Siobhan O Burns[85,86], Oliver S Burren[5,8], Nigel Burrows[46], Paul Calleja[12], Carolyn Campbell[87], Gerald Carr-White[88], Keren Carss[1,2], Ruth Casey[24,25,44], Mark J Caulfield[13,14], Jenny Chambers[17,89], John Chambers[90,91,92,93,94], Melanie M Y Chan[18], Calvin Cheah[71], Floria Cheng[89], Patrick F Chinnery[2,15,16], Manali Chitre[64], Martin T Christian[95], Colin Church[96], Jill Clayton-Smith[58,59], Maureen Cleary[50], Naomi Clements Brod[1,2], Gerry Coghlan[82], Elizabeth Colby[35], Trevor R P Cole[97], Janine Collins[1,98], Peter W Collins[99], Camilla Colombo[45], Cecilia J Compton[83], Robin Condliffe[100], Stuart Cook[80,101,102,103], H Terence Cook[104], Nichola Cooper[42], Paul Corris[105,106], Abigail Crisp-Hihn[1,2], Fiona Cunningham[27], Nicola S Curry[107], Antony J Cutler[108], Cesare Danesino[109], Matthew J Daniels[30,31,110], Mehul Dattani[67,111], Louise C Daugherty[1,2,13], John Davis[1,2], Anthony De Soyza[75,106,112], Sri V V Deevi[1,2],

Timothy Dent[31], Charu Deshpande[83], Eleanor F Dewhurst[1,2], Peter H Dixon[17], Sofia Douzgou[58,59], Kate Downes[1,2], Anna M Drazyk[26], Elizabeth Drewe[113], Daniel Duarte[1,2], Tina Dutt[114], J David M Edgar[115,116], Karen Edwards[1,2], William Egner[117], Melanie N Ekani[88], Perry Elliott[118,119], Wendy N Erber[120], Marie Erwood[1,2], Maria C Estiu[121], Dafydd Gareth Evans[122], Gillian Evans[123], Tamara Everington[124,125], Mélanie Eyries[126], Hiva Fassihi[127], Remi Favier[128,129], Jack Findhammer[130], Debra Fletcher[1,2], Frances A Flinter[83], R Andres Floto[5,46,66], Tom Fowler[13], James Fox[1,2], Amy J Frary[1,2], Courtney E French[64], Kathleen Freson[7], Mattia Frontini[1,2,6], Daniel P Gale[18,19], Henning Gall[131], Vijeya Ganesan[50], Michael Gattens[46], Claire Geoghegan[45], Terence S A Gerighty[45], Ali G Gharavi[132], Stefano Ghio[133], Hossein-Ardeschir Ghofrani[42,131], J Simon R Gibbs[80], Kate Gibson[87], Kimberly C Gilmour[38,50], Barbara Girerd[134,135,136], Nicholas S Gleadall[1,2], Sarah Goddard[137], David B Goldstein[138], Keith Gomez[60,61], Pavels Gordins[139], David Gosal[53], Stefan Gräf[1,2,5], Jodie Graham[140], Luigi Grassi[1,2], Daniel Greene[1,2,3], Lynn Greenhalgh[141], Andreas Greinacher[142], Paolo Gresele[143], Philip Griffiths[144,145], Sofia Grigoriadou[146], Russell J Grocock[45], Detelina Grozeva[24], Mark Gurnell[5,46], Scott Hackett[147], Charaka Hadinnapola[5], William M Hague[148], Rosie Hague[149], Matthias Haimel[1,2,5], Matthew Hall[113], Helen L Hanson[150], Eshika Haque[83], Kirsty Harkness[151], Andrew R Harper[30,39], Claire Harris[106], Daniel Hart[98], Ahamad Hassan[152], Grant Hayman[153], Alex Henderson[76], Archana Herwadkar[53], Jonathan Hoffman[97], Simon Holden[154], Rita Horvath[144], Rita Horvath[155,156], Henry Houlden[157], Arjan Houweling[69], Luke S Howard[80,158], Fengyuan Hu[1,2], Gavin Hudson[155], Joseph Hughes[45], Aarnoud P Huissoon[147], Marc Humbert[134,135,136], Sean Humphray[45], Sarah Hunter[2,45], Matthew Hurles[48], Melita Irving[83], Louise Izatt[83], Roger James[1,2], Sally A Johnson[106,159], Stephen Jolles[160], Jennifer Jolley[1,2], Dragana Josifova[83], Neringa Jurkute[40,61], Tim Karten[130], Johannes Karten[130], Mary A Kasanicki[46], Hanadi Kazkaz[161], Rashid Kazmi[70], Peter Kelleher[162,163], Anne M Kelly[46], Wilf Kelsall[46], Carly Kempster[1,2], David G Kiely[100], Nathalie Kingston[2], Robert Klima[12], Nils Koelling[164], Myrto Kostadima[1], Gabor Kovacs[165,166], Ania Koziell[20,21], Roman Kreuzhuber[1,2], Taco W Kuijpers[167,168], Ajith Kumar[52], Dinakantha Kumararatne[169], Manju A Kurian[170,171], Michael A Laffan[22,23], Fiona Lalloo[59], Michele Lambert[172,173], Hana Lango Allen[1,2], Allan Lawrie[174], D Mark Layton[22,23], Nick Lench[71], Claire Lentaigne[22,23], Tracy Lester[87], Adam P Levine[18], Rachel Linger[2,37], Hilary Longhurst[175], Lorena E Lorenzo[146], Eleni Louka[32,34], Paul A Lyons[5,8], Rajiv D Machado[176,177], Robert V MacKenzie Ross[178], Bella Madan[179], Eamonn R Maher[24,25], Jesmeen Maimaris[38], Samantha Malka[40,41], Sarah Mangles[125], Rutendo Mapeta[1,2], Kevin J Marchbank[106,180], Stephen Marks[50], Hugh S Markus[26], Hanns-Ulrich Marschall[181], Andrew Marshall[182,183,184], Jennifer Martin[2,5,37], Mary Mathias[185], Emma Matthews[186,187], Heather Maxwell[149], Paul McAlinden[106], Mark I McCarthy[34,39,188], Harriet McKinney[1,2], Aoife McMahon[27], Stuart Meacham[1,2], Adam J Mead[32], Ignacio Medina Castello[12], Karyn Megy[1,2], Sarju Mehta[154], Michel Michaelides[40,41], Carolyn Millar[22,23], Shehla N Mohammed[83], Shahin Moledina[50], David Montani[134,135,136], Anthony T Moore[40,41,189], Joannella Morales[27], Nicholas W Morrell[2,5], Monika Mozere[18], Keith W Muir[190], Andrew D Mumford[28,29], Andrea H Nemeth[11,191], William G Newman[58,59], Michael Newnham[5,66], Sadia Noorani[192], Paquita Nurden[193], Jennifer O'Sullivan[179], Samya Obaji[194], Chris Odhams[13], Steven Okoli[32,34], Andrea Olschewski[165], Horst Olschewski[165,166], Kai Ren Ong[97], Helen Oram[195], Elizabeth Ormondroyd[30,31], Willem H Ouwehand[1,2,4,6,48], Claire Palles[196], Sofia Papadia[2,37], Soo-Mi Park[25,44,49], David Parry[10], Smita Patel[197], Joan Paterson[24,25,44], Andrew Peacock[96], Simon Pearce[75,140], John Peden[45], Kathelijne Peerlinck[7], Christopher J Penkett[1,2], Joanna Pepke-Zaba[66], Romina Petersen[1,2], Clarissa

Pilkington[50], Ken E S Poole[5,46], Radhika Prathalingam[71], Beth Psaila[32,34,198], Angela Pyle[155], Richard Quinton[75,155], Shamima Rahman[50,67], Stuart Rankin[12], Anupama Rao[50], F Lucy Raymond[2,24], Paula J Rayner-Matthews[1,2], Christine Rees[45], Augusto Rendon[1,13], Tara Renton[199], Christopher J Rhodes[42], Andrew S C Rice[200,201], Sylvia Richardson[3], Alex Richter[56], Leema Robert[83], Irene Roberts[32,33,34], Anthony Rogers[71], Sarah J Rose[83], Robert Ross-Russell[46], Catherine Roughley[123], Noemi B Roy[164,202], Deborah M Ruddy[83], Omid Sadeghi-Alavijeh[18], Moin A Saleem[35,36], Nilesh Samani[203], Crina Samarghitean[1,2], Alba Sanchis-Juan[1,2], Ravishankar B Sargur[117], Robert N Sarkany[127], Simon Satchell[35,204], Sinisa Savic[205,206,207], John A Sayer[75,155], Genevieve Sayer[83], Laura Scelsi[133], Andrew M Schaefer[75,144], Sol Schulman[208], Richard Scott[13,50], Marie Scully[161], Claire Searle[209], Werner Seeger[131], Arjune Sen[34,210,211], W A Carrock Sewell[212], Denis Seyres[1,2], Neil Shah[38,50], Olga Shamardina[1,2], Susan E Shapiro[107], Adam C Shaw[83], Patrick J Short[48], Keith Sibson[185], Lucy Side[213], Ilenia Simeoni[1,2], Michael Simpson[214], Matthew C Sims[1,215], Suthesh Sivapalaratnam[4,98,216,217], Damian Smedley[13,14], Katherine R Smith[13], Kenneth G C Smith[5,8], Katie Snape[150], Nicole Soranzo[1,48], Florent Soubrier[126], Laura Southgate[177,218], Olivera Spasic-Boskovic[49], Simon Staines[1,2], Emily Staples[5], Hannah Stark[2,37], Jonathan Stephens[1,2], Charles Steward[71], Kathleen E Stirrups[1,2], Alex Stuckey[13], Jay Suntharalingam[178], Emilia M Swietlik[5], Petros Syrris[118], R Campbell Tait[219], Kate Talks[68], Rhea Y Y Tan[26], Katie Tate[71], John M Taylor[87], Jenny C Taylor[34,39], James E Thaventhiran[8,220], Andreas C Themistocleous[11], Ellen Thomas[13,88], David Thomas[5], Moira J Thomas[221,222], Patrick Thomas[1,2], Kate Thomson[30,31], Adrian J Thrasher[38], Glen Threadgold[27], Chantal Thys[7], Tobias Tilly[1,2], Marc Tischkowitz[49,223], Catherine Titterton[1,2], John A Todd[108], Cheng-Hock Toh[114], Bas Tolhuis[130], Ian P Tomlinson[196], Mark Toshner[5,66], Matthew Traylor[26], Carmen Treacy[5,66], Paul Treadaway[1,2], Richard Trembath[20], Salih Tuna[1,2], Wojciech Turek[12], Ernest Turro[1,2,3], Philip Twiss[49], Tom Vale[11], Chris Van Geet[7], Natalie van Zuydam[11], Maarten Vandekuilen[130], Anthony M Vandersteen[224], Marta Vazquez-Lopez[89], Julie von Ziegenweidt[1,2], Anton Vonk Noordegraaf[69], Annette Wagner[46], Quinten Waisfisz[225], Suellen M Walker[38,50], Neil Walker[1,2], Klaudia Walter[48], James S Ware[80,81,101], Hugh Watkins[30,31,39], Christopher Watt[1,2], Andrew R Webster[40,41], Lucy Wedderburn[38,226,227], Wei Wei[15,16], Steven B Welch[228], Julie Wessels[137], Sarah K Westbury[28,29], John-Paul Westwood[161], John Wharton[42], Deborah Whitehorn[1,2], James Whitworth[24,25,44], Andrew O Wilkie[164], Martin R Wilkins[42], Catherine Williamson[17,43], Brian T Wilson[52,75,155], Edwin K S Wong[75,106], Nicholas Wood[157,229], Yvette Wood[1,2], Christopher Geoffrey Woods[24,46], Emma Woodward[59], Stephen J Wort[81,230], Austen Worth[50], Michael Wright[76], Katherine Yates[1,2,5], Patrick F K Yong[231], Timothy Young[1,2], Ping Yu[1,2], Patrick Yu-Wai-Man[15,16,232], Eliska Zlamalova[1]

[49]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [50]Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [51]Yorkshire Regional Genetics Service, Chapel Allerton Hospital, Leeds Teaching Hospitals NHS Trust, Leeds, UK. [52]North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [53]Salford Royal NHS Foundation Trust, Salford, UK. [54]Queens Centre for Haematology and Oncology, Castle Hill Hospital, Hull and East Yorkshire NHS Trust, Cottingham, UK. [55]Hull York Medical School, University of Hull, Hull, UK. [56]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [57]Center for Clinical Transfusion Medicine, University Hospital of Tübingen,

Tübingen, Germany. [58]Evolution and Genomic Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. [59]Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester Universities Foundation NHS Trust, Manchester, UK. [60]The Katharine Dormandy Haemophilia Centre and Thrombosis Unit, Royal Free London NHS Foundation Trust, London, UK. [61]University College London, London, UK. [62]Department of Clinical Genetics, Leicester Royal Infirmary, University Hospitals of Leicester, Leicester, UK. [63]University of Leicester, Leicester, UK. [64]Department of Paediatrics, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [65]Division of Clinical Biochemistry and Immunology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [66]Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK. [67]Genetics and Genomic Medicine Programme, UCL Great Ormond Street Institute of Child Health, London, UK. [68]Haematology Department, Royal Victoria Infirmary, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [69]Department of Pulmonary Medicine, VU University Medical Centre, Amsterdam, The Netherlands. [70]Southampton General Hospital, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [71]Congenica, Biodata Innovation Centre, Wellcome Genome Campus, Hinxton, Cambridge, UK. [72]MRC Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK. [73]NIHR Biomedical Research Centre for Mental Health, Maudsley Hospital, London, UK. [74]Newcastle University, Newcastle upon Tyne, UK. [75]Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [76]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [77]Department of Clinical Genetics, Royal Devon & Exeter Hospital, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK. [78]Newcastle Eye Centre, Royal Victoria Infirmary, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [79]Department of Immunology, Leicester Royal Infirmary, Leicester, UK. [80]National Heart and Lung Institute, Imperial College London, London, UK. [81]Royal Brompton Hospital, Royal Brompton and Harefield NHS Foundation Trust, London, UK. [82]Royal Free London NHS Foundation Trust, London, UK. [83]Clinical Genetics Department, Guy's and St Thomas NHS Foundation Trust, London, UK. [84]Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, UK. [85]Institute of Immunity and Transplantation, University College London, London, UK. [86]Department of Immunology, Royal Free London NHS Foundation Trust, London, UK. [87]Oxford Medical Genetics Laboratories, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [88]Guy's and St Thomas' Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK. [89]Women’s Health Research Centre, Surgery and Cancer, Faculty of Medicine, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [90]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. [91]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [92]Department of Cardiology, Ealing Hospital, Middlesex, UK. [93]Imperial College Healthcare NHS Trust, London, UK. [94]MRC-PHE Centre for Environment and Health, Imperial College London, London, UK. [95]Children's Renal and Urology Unit, Nottingham Children's Hospital, QMC, Nottingham University Hospitals NHS Trust, Nottingham, UK. [96]Golden Jubilee National Hospital, Glasgow, UK. [97]West Midlands Regional Genetics Service, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. [98]The Royal London Hospital, Barts Health NHS Foundation Trust, London, UK. [99]Institute of Infection and

Immunity, School of Medicine Cardiff University, Cardiff, UK. [100]Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital NHS Foundation Trust, Sheffield, UK. [101]MRC London Institute of Medical Sciences, Imperial College London, London, UK. [102]National Heart Research Institute Singapore, National Heart Centre Singapore, Singapore, Singapore. [103]Division of Cardiovascular & Metabolic Disorders, Duke-National University of Singapore, Singapore, Singapore. [104]Imperial College Renal and Transplant Centre, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [105]National Pulmonary Hypertension Service (Newcastle), The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [106]Institute of Cellular Medicine, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK. [107]Oxford Haemophilia and Thrombosis Centre, Oxford University Hospitals NHS Trust, Oxford Comprehensive Biomedical Research Centre, Oxford, UK. [108]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. [109]Department of Molecular Medicine, General Biology, and Medical Genetics Unit, University of Pavia, Pavia, Italy. [110]Department of Biotechnology, Graduate School of Engineering, Osaka University, Suita, Osaka, Japan. [111]London Centre for Paediatric Endocrinology and Diabetes, Great Ormond Street Hospital for Children, London, UK. [112]NIHR Centre for Aging, Newcastle University, Newcastle upon Tyne, UK. [113]Nottingham University Hospitals NHS Trust, Nottingham, UK. [114]The Roald Dahl Haemostasis and Thrombosis Centre, The Royal Liverpool University Hospital, Liverpool, UK. [115]St James's Hospital, Dublin, Ireland. [116]Trinity College Dublin, Dublin, Ireland. [117]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [118]UCL Institute of Cardiovascular Science, University College London, London, UK. [119]Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, UK. [120]Medical School and School of Biomedical Sciences, Faculty of Health and Medical Sciences, The University of Western Australia, Crawley, Australia. [121]Ramón Sardá Mother's and Children's Hospital, Buenos Aires, Argentina. [122]Manchester University NHS Foundation Trust, Manchester, UK. [123]Haemophilia Centre, Kent & Canterbury Hospital, East Kent Hospitals University Foundation Trust, Canterbury, UK. [124]Salisbury District Hospital, Salisbury NHS Foundation Trust, Salisbury, UK. [125]Haemophilia, Haemostasis and Thrombosis Centre, Hampshire Hospitals NHS Foundation Trust, Basingstoke, UK. [126]Departement de Genetique & ICAN, Hopital Pitie-Salpetriere, Assistance Publique Hopitaux de Paris, Paris, France. [127]St Johns Institute of Dermatology, St Thomas' Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK. [128]INSERM UMR 1170, Gustave Roussy Cancer Campus, Universite Paris-Saclay, Villejuif, France. [129]Service d'Hematologie biologique, Centre de Reference des Pathologies Plaquettaires, Hopital Armand Trousseau, Assistance Publique-Hopitaux de Paris, Paris, France. [130]GENALICE BV, Harderwijk, The Netherlands. [131]University of Giessen and Marburg Lung Center (UGMLC), Giessen, Germany. [132]Division of Nephrology and Center for Precision Medicine and Genomics, Department of Medicine Columbia University Vagelos College of Physicians and Surgeons, New York, USA. [133]Division of Cardiology, Fondazione IRCCS Policlinico S. Matteo, Pavia, Italy. [134]Univ. Paris-Sud, Faculty of Medicine, University Paris-Saclay, Le Kremlin Bicetre, France. [135]Service de Pneumologie, Centre de Reference de l'Hypertension Pulmonaire, Hopital Bicetre (Assistance Publique Hopitaux de Paris), Le Kremlin Bicetre, France. [136]INSERM U999, Hospital Marie Lannelongue, Le Plessis Robinson, France. [137]University Hospitals of North

Midlands NHS Trust, Stoke-on-Trent, UK. [138]Institute of Genomic Medicine and the Department of Genetics and Development, Columbia University Vagelos College of Physicians and Surgeons, New York, USA. [139]East Yorkshire Regional Adult Immunology and Allergy Unit, Hull Royal Infirmary, Hull and East Yorkshire Hospitals NHS Trust, Hull, UK. [140]Newcastle BRC, Newcastle University, Newcastle upon Tyne, UK. [141]Department of Clinical Genetics, Liverpool Women's NHS Foundation, Liverpool, UK. [142]Institute for Immunology and Transfusion Medicine, University Medicine Greifswald, Greifswald, Germany. [143]Section of Internal and Cardiovascular Medicine, University of Perugia, Perugia, Italy. [144]Wellcome Centre for Mitochondrial Research, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [145]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [146]Barts Health NHS Foundation Trust, London, UK. [147]Birmingham Heartlands Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [148]Robinson Research Institute, Discipline of Obstetrics and Gynaecology, The University of Adelaide, Women's and Children's Hospital, Adelaide, Australia. [149]Royal Hospital for Children, NHS Greater Glasgow and Clyde, Glasgow, UK. [150]Department of Clinical Genetics, St George's University Hospitals NHS Foundation Trust, London, UK. [151]Department of Neurology, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [152]Department of Neurology, Leeds Teaching Hospital NHS Trust, Leeds, UK. [153]Epsom & St Helier University Hospitals NHS Trust, London, UK. [154]Department of Clinical Genetics, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [155]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [156]John Walton Muscular Dystrophy Research Centre, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [157]Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK. [158]National Pulmonary Hypertension Service, Imperial College Healthcare NHS Trust, London, UK. [159]Department of Paediatric Nephrology, Great North Children’s Hospital, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [160]Immunodeficiency Centre for Wales, University Hospital of Wales, Cardiff, IUK. [161]University College London Hospitals NHS Foundation Trust, London, UK. [162]Centre for Immunology & Vaccinology , Chelsea & Westminster Hospital , Department of Medicine, Imperial College London, London, UK. [163]Department of Respiratory Medicine Royal Brompton & Harefield NHS Foundation Trust, London, UK. [164]MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK. [165]Ludwig Boltzmann Institute for Lung Vascular Research, Graz, Austria. [166]Dept of Internal Medicine, Division of Pulmonology, Medical University of Graz, Graz, Austria. [167]Department of Pediatric Hematology, Immunology, Rheumatology and Infectious Diseases, Emma Children’s Hospital, Academic Medical Center (AMC), University of Amsterdam, Amsterdam, The Netherlands. [168]Department of Blood Cell Research, Sanquin, Amsterdam, The Netherlands. [169]Department of Clinical Immunology, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [170]Developmental Neurosciences, UCL Great Ormond Street Institute of Child Health, London, UK. [171]Department of Neurology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [172]Division of Hematology, The Children's Hospital of Philadelphia, Philadelphia, USA. [173]Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA. [174]Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, UK. [175]University College Hospital, University College London Hospitals NHS Foundation Trust,

London, UK. [176]School of Life Sciences, University of Lincoln, Lincoln, UK. [177]Molecular and Clinical Sciences Research Institute, St George's University of London, London, UK. [178]Royal United Hospitals Bath NHS Foundation Trust, Bath, UK. [179]Department of Haematology, Guy's and St Thomas' NHS Foundation Trust, London, UK. [180]The National Renal Complement Therapeutics Centre, Royal Victoria Infirmary, Newcastle upon Tyne, UK. [181]Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [182]Faculty of Biology, Medicine and Health, School of Biological Sciences, Division of Neuroscience and Experimental Psychology, University of Manchester, Manchester, UK. [183]Department of Clinical Neurophysiology, Manchester University NHS Foundation Trust, Manchester, Manchester Academic Health Science Centre,, Manchester, UK. [184]National Institute for Health Research/Wellcome Trust Clinical Research Facility, Manchester, UK. [185]Department of Haematology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [186]The National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK. [187]MRC Centre for Neuromuscular Diseases, Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK. [188]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. [189]Ophthalmology Department, UCSF School of Medicine, San Francisco, USA. [190]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. [191]Department of Clinical Genetics, Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK. [192]Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK. [193]Institut Hospitalo-Universitaire de Rythmologie et de Modelisation Cardiaque, Plateforme Technologique d'Innovation Biomedicale, Hopital Xavier Arnozan, Pessac, France. [194]The Arthur Bloom Haemophilia Centre, University Hospital of Wales, Cardiff, UK. [195]Department of Paediatric Haematology, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [196]Institute of Cancer and Genomic Sciences, Institute of Biomedical Research, University of Birmingham, Birmingham, UK. [197]Department of Clinical Immunology, John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [198]Centre for Haematology, Department of Medicine, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [199]King's College Hospital NHS Foundation Trust, London, UK. [200]Pain Research, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK. [201]Pain Medicine, Chelsea and Westminster Hospital NHS Foundation Trust, London, UK. [202]Department of Haematology, Oxford University Hospital Foundation Trust, Oxford, UK. [203]Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Research Centre, University of Leicester, Leicester, UK. [204]North Bristol NHS Trust, Bristol, UK. [205]Department of Clinical Immunology and Allergy, St James's University Hospital, Leeds, UK. [206]The NIHR Leeds Biomedical Research Centre, Leeds, UK. [207]Leeds Institute of Rheumatic and Musculoskeletal Medicine, Leeds, UK. [208]Beth Israel Deaconess Medical Centre and Harvard Medical School, Boston, USA. [209]Department of Clinical Genetics, Nottingham University Hospitals NHS Trust, Nottingham, UK. [210]Oxford Epilepsy Research Group, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. [211]Nuffield Department of Surgery, University of Oxford, Oxford, UK. [212]Scunthorpe General Hospital, Northern Lincolnshire and Goole NHS Foundation Trust, Scunthorpe, UK. [213]Wessex Clinical Genetics Service, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [214]Genetics and Molecular Medicine, King's College London, London, UK.

[215]Oxford Haemophilia and Thrombosis Centre, The Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK. [216]Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [217]Queen Mary University of London, London, UK. [218]Faculty of Life Sciences and Medicine, King's College London, London, UK. [219]Glasgow Royal Infirmary, NHS Greater Glasgow and Clyde, Glasgow, UK. [220]MRC Toxicology Unit, School of Biological Sciences, University of Cambridge, Cambridge, UK. [221]Gartnavel General Hospital, NHS Greater Glasgow and Clyde, Glasgow, UK. [222]Queen Elizabeth University Hospital, Glasgow, UK. [223]Department of Medical Genetics and NIHR Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK. [224]Division of Medical Genetics, IWK Health Centre, Dalhousie University, Halifax, Canada. [225]Department of Clinical Genetics, VU University Medical Centre, Amsterdam, The Netherlands. [226]NIHR Great Ormond Street Biomedical Research Centre, London, UK. [227]Arthritis Research UK Centre for Adolescent Rheumatology at UCL UCLH and GOSH, London, UK. [228]Birmingham Chest Clinic and Heartlands Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [229]UCL Genetics Institute, London, UK. [230]Imperial College London, London, UK. [231]Frimley Park Hospital, NHS Frimley Health Foundation Trust, Camberley, UK. [232]NIHR Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology, London, UK.