Article

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

https://doi.org/10.1038/s41586-020-2434-2

Received: 23 December 2018

Accepted: 5 May 2020

Published online: 24 June 2020

Check for updates

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

Most patients with rare diseases do not receive a molecular diagnosis and the aetiological variants and causative genes for more than half such disorders remain to be discovered[1]. Here we used 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. 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 extensively phenotyped participants. We identified 95 Mendelian associations between genes and rare diseases, of which 11 have been discovered since 2015 and at least 79 are confirmed to be aetiological. By generating WGS data of UK Biobank participants[2], we found that rare alleles can explain the presence of some individuals in the tails of a quantitative trait for red blood cells. Finally, we identified four novel non-coding variants that 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 for diseases with a resolved aetiology, the prospects for diagnosis are severely diminished by fragmentary phenotyping and the restriction of testing to disease-specific panels of genes. It may require more than 20 physician visits over several years to determine a molecular cause[4]. Recent development of WGS technology enables systematic, comprehensive genetic testing in integrated health systems, together with aetiological discovery in the coding and non-coding genome.

We performed WGS for 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 and 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 distribution reported by the UK census across ethnic groups (Fig. 1c; https://www.ons.gov.uk/census/2011census). Each participant was assigned to one of 18 domains with pre-specified enrolment criteria (Supplementary Table 1): 7,388 individuals were assigned to one of 15 rare disease domains, 50 individuals to a control domain, 4,835 individuals to a domain called the Rare Diseases Pilot of Genomics England Ltd (GEL) and 764 individuals to a domain comprising UK Biobank participants with extreme red blood cell indices (Extended Data Fig. 1b, Supplementary Information and Supplementary Table 1). Sample sizes varied across domains, primarily owing to differences in recruitment rates, limiting the efficiency of the study design. In total, 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. The patients presented with pathologies of many organ systems, which we phenotyped using Human Phenotype Ontology (HPO) terms for all of the rare disease domains except the domain comprising Leber's hereditary optic neuropathy and the domain comprising Ehler–Danlos/Ehler–Danlos-like syndromes (Fig. 2a and Extended Data

A list of affiliations appears at the end of the paper.

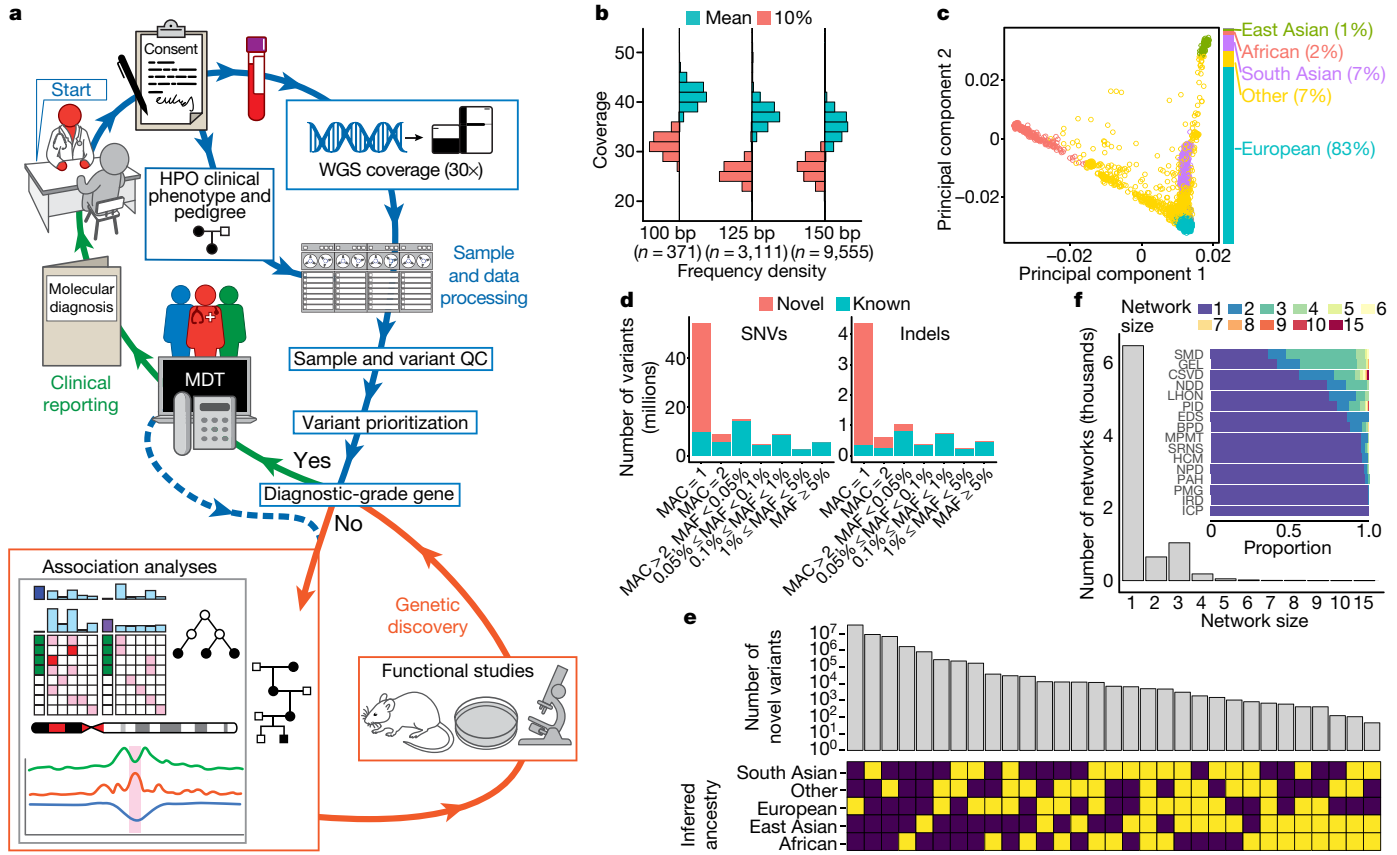


Fig. 1 | Study overview. a, Schematic of the diagnostic and research processes. Blue, patients are recruited, HPO and pedigree data are collected, DNA is extracted and sequenced and WGS data are transferred for quality control and variant prioritization. Green, variants are assessed and diagnoses are 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 (100 bp, 125 bp and 150 bp). **c**, Projection of genetic data of the 13,037 participants onto the first two principal components of variation in the 1000 Genomes Project and the distribution of participant ancestry. **d**, Histograms illustrating the observed distribution of the minor allele frequency (MAF) of variants called in the MSUP (n = 10,259), stratified by type (SNV or indel). Variants are labelled novel if they were uncatalogued in the 1000 Genomes, UK10K, TOPMed, gnomAD and HGMD Pro databases. MAC, minor allele count. **e**, The number of novel variants stratified by the ancestry groups in which they were observed (yellow, present; navy, absent). **f**, 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. BPD, bleeding, thrombotic and platelet disorders; CSVD, cerebral small vessel disease; EDS, Ehler–Danlos and Ehler–Danlos-like syndromes; HCM, hypertrophic cardiomyopathy; ICP, intrahepatic cholestasis of pregnancy; IRD, inherited retinal disorders; LHON, Leber's hereditary optic neuropathy; MPMT, multiple primary malignant tumours; NDD, neurological and developmental disorders; NPD, neuropathic pain disorders; PAH, pulmonary arterial hypertension; PID, primary immune disorders; PMG: primary membranoproliferative glomerulonephritis; SMD, stem cell and myeloid disorders; SRNS, steroid-resistant nephrotic syndrome.

Fig. 1c). The GEL domain released only a binary affection phenotype for these analyses. In total, 19,605 HPO terms were assigned to patients.

Following bioinformatic analysis (Extended Data Fig. 2–4), we considered a maximal set of 10,259 unrelated participants (MSUP), in which we identified 172,005,610 short variants. These variants comprised 157,411,228 (91.5%) single-nucleotide variants (SNVs) and 14,594,382 (8.5%) small insertions or deletions (indels) of ≤50 base pairs (bp) (Extended Data Fig. 5). Of these SNVs and indels, 48.6% and 40.8%, respectively, were absent from major public variant databases (Fig. 1d) and 54.8% had a minor-allele count of 1. Of these singleton variants, 82.6% were novel. Only 9.08% of the novel variants had a minor-allele count > 1; in these cases, the minor allele was typically carried exclusively by individuals with similar population ancestry (Fig. 1e). SNVs and indels were well represented in major variant databases if they were common in our dataset; however, consistent with theory, most variants were very rare and, of these, most were uncatalogued. We called 177,550 distinct large deletions (>50 bp) across the 13,037 participants by synthesizing inferences from two algorithms. We also called more complicated types of structural variant, such as inversions; however, 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). Owing 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 to identify the corresponding transcripts (Supplementary Information). The lists ranged in length from two for the intrahepatic cholestasis of pregnancy domain to 1,423 for the neurological and developmental disorders domain. The lists were not mutually exclusive because mutations in some genes cause pathologies that were 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

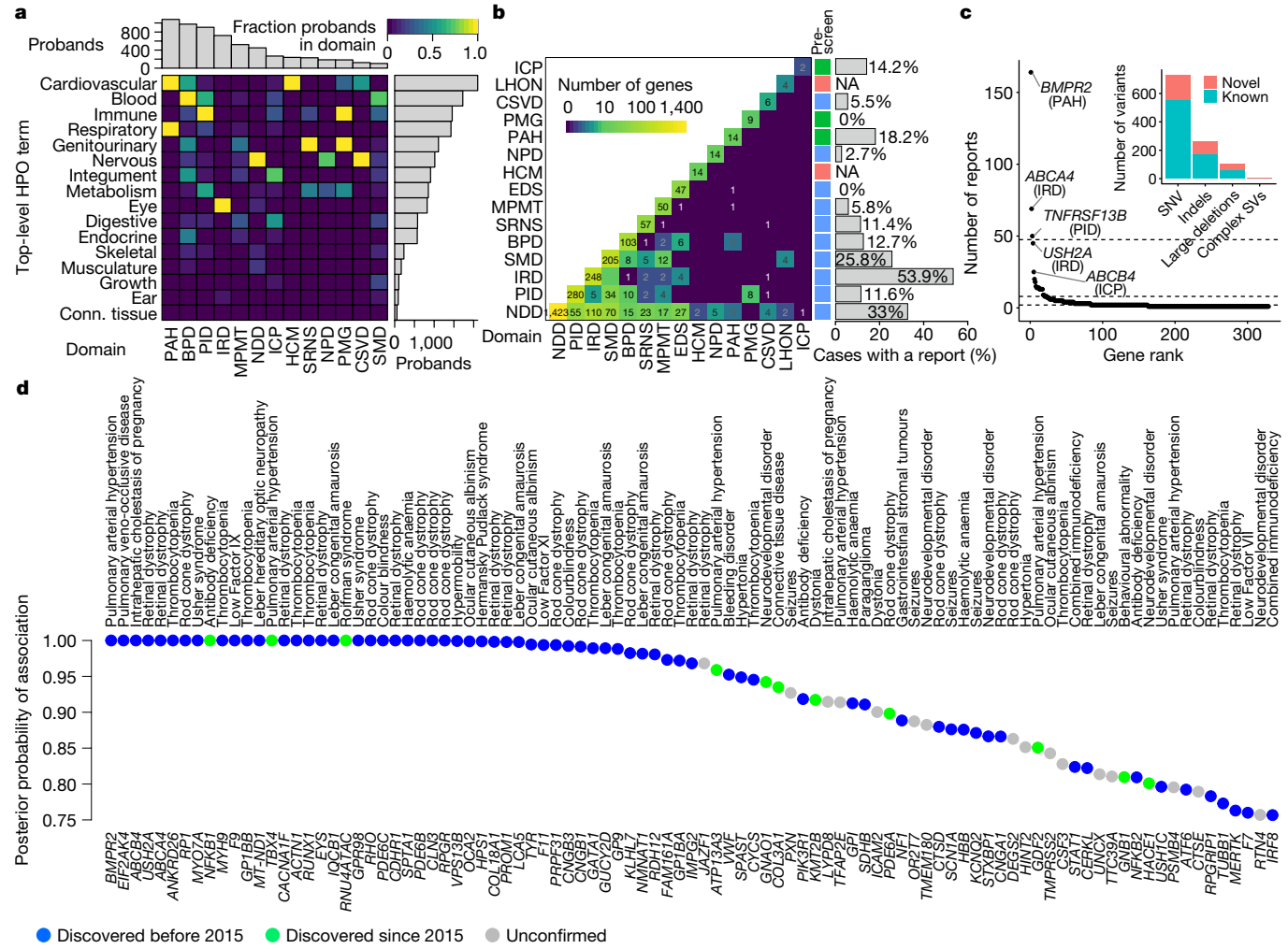


Fig. 2 | Variant reporting and genetic associations with rare diseases. a, The frequency of probands by domain (top) and by top-level HPO-phenotype abnormality term (right). The heat map shows the proportion of probands in each domain that were assigned a particular top-level HPO term (shown abbreviated). **b**, 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). The proportion of cases for which a clinical report was issued (right). **c**, The number of reports issued by DGG ordered inversely by count. Dashed lines indicate quartiles of the count distribution. Inset, the number of distinct clinically reported variants stratified by variant type. The colours in each bar indicate the proportion of variants that are known or novel (as defined in the main text). **d**, BeviMed posterior probabilities for genetic association >0.75. The colours indicate whether the associations were established in the scientific literature before 2015, since 2015 or remain unconfirmed. *GPR98* is also known as *ADGRV1*; *TMEM180* is also known as *MFSD13A*.

categorized 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'. The contribution of a variant was assessed to be full if it was considered to be the only variant for which an isolated reduction in copy number from conception would have eliminated the disease phenotype, otherwise the contribution was assessed to be partial. Clinical reports—which contained molecular diagnoses comprising 1,103 distinct causal variants (731 SNVs, 264 indels, 102 large deletions and 6 complex structural variants) that affected 329 DGGs (Supplementary Table 2)—were issued for 1,138 of the 7,065 (16.1%) patients reviewed. We classed 266 of the 995 SNVs and indels (26.7%) 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 and Extended Data Fig. 6). The diagnostic yield by domain ranged from 0% (0 out of 184) of patients for the primary membranoproliferative glomerulonephritis domain to 53.9% (391 out of 725) of patients for the inherited retinal disease domain (Fig. 2b). The variability in 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 greater than 35× 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 that was previously missed by WES. We compared the read coverage of WGS to that of research WES of participants in the UK Biobank[6], INTERVAL[7] 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 that contain 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 of autosomal SNVs with

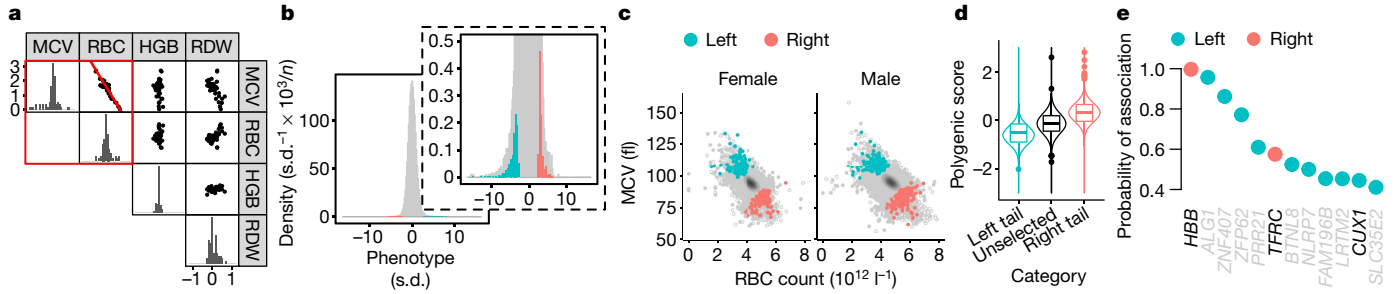


Fig. 3 | Genetic associations with the tails of an RBC trait. a, The distribution of the additive effects of 65 RBC GWAS variants (MAF < 1%) on four RBC traits (acronyms are defined in the Supplementary Information). The red square indicates the bivariate distribution used to develop the selection phenotype. The red line was estimated by Deming regression. **b**, The (standardized) distribution of the selection phenotype (panels showing different *y*-axis ranges) in post-menopausal female and male participants of European ancestry in the UK Biobank without record of illness or treatment that is known to perturb RBC indices (grey) and selected for WGS (turquoise and salmon). The scale of the *x*-axis shows the standard deviations (s.d.) of the phenotype and the scale of the *y*-axis is such that when the units of the axes are disregarded the area of the histogram represents the number of contributing individuals in thousands, where $n = 316{,}739$. Many participants in the tails were unselected (Supplementary Information). **c**, The distribution of RBC count and mean cell volume (MCV) in post-menopausal female (left) and male (right) participants in the UK Biobank. The ellipsoids are contours of kernel density estimates. Open circles, participants ineligible for selection. Non-European ancestry thalassaemias may explain the concentration with high RBC count/low mean cell volume. Coloured circles, participants who have WGS data. **d**, The distribution of a polygenic score for the selection phenotype in the 382 and 368 individuals of European ancestry selected from the left and right tails, respectively, and in 522 European participants in domains other than the UK Biobank (extreme red blood cell traits) domain with pathology explained by rare variants (unselected). The centre mark, lower and upper hinges of the boxplots, respectively, indicate the median, 25th and 75th percentiles. Outliers beyond 1.5× 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 posterior probabilities for genetic association of each tail (distinguished by colour), for genes with posterior probabilities >0.4. Indicated genes (black font) have strong concordant biological evidence.

insufficient coverage in WES analyses for reliable genotyping ranged between 25 and 99 (2.67–10.5%) across WES datasets (Extended Data Fig. 7). Moreover, deletions that span only a few short exons or part of a single exon are not reliably called by WES[8,9]. Of the 102 distinct large deletions that we reported (length range, 203 bp–16.80 Mb; mean, 786.33 kb; median, 15.91 kb), 22 (21.6%) overlapped only one exon. Although clinical and research WES may have different coverage characteristics, we were unable to obtain an example clinical dataset for comparison.

Measurement of quantitative intermediate phenotypes can elucidate the genetic aetiology in difficult-to-diagnose patients. We considered patients with a clinically determined absence of a protein encoded by a DGG for 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 owing 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 a SINE-VNTR-Alu (SVA) retrotransposon that was 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 severe haemolytic anaemia owing to absence of the RhD and RhCE proteins in the membranes of her red blood cells, which was caused by a large tandem repeat in *RHAG* (Extended Data Fig. 8f).

Research findings from this study have informed treatment decisions: patients with *KMT2B*-mediated early-onset dystonia were treated by deep brain stimulation[10]; individuals with *DIAPH1*-related macrothrombocytopenia and deafness[11] were treated for their thrombocytopenia in a preoperative setting with eltrombopag[12]; and a case of severe thrombocytopenia, myelofibrosis and bleeding owing to a gain-of-function mutation in *SRC*[13] was cured by an allogeneic haematopoietic stem cell transplant. Our diagnoses have stratified patient care: patients with primary immune disorders owing to variants in *NFKB1*, which we have shown are the most common monogenic cause of combined variable immunodeficiency[14], have unexplained splenomegaly and an increased risk of cancer; 27 cases with isolated thrombocytopenia caused by variants in *ANKRD26*, *ETV6* or *RUNX1* have an increased risk of malignancy[15–17] compared with 19 cases with thrombocytopenia caused by variants in *ACTN1*, *CYCS* or *TUBB1*. Our discoveries have also improved the accuracy of prognosis: we found that mutations in *BMPR2*[18] and *EIF2AK4*[19] carry a poorer-than-average prognosis in pulmonary arterial hypertension and we plan prognostication studies of four genes (*ATP13A3*, *AQP1*, *GDF2* and *SOX17*) that we recently reported are aetiological[20].

Genetic associations with rare diseases

Several cases with similar aetiologies are typically needed for discovery in rare disease genetics. Cases can be aggregated across distinct studies using Matchmaker Exchange[21]. We identified novel aetiologies for *SLC18A2*[22] and *WASF1*[23] using Matchmaker Exchange (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[24] 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 to be 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 pulmonary arterial hypertension. We analysed each gene–tag pair independently and considered a posterior probability of association of greater than 0.75 to be strong evidence of a genetic aetiology. To account for correlation between tags, we recorded only one association per gene, corresponding to the tag for which the highest posterior probability of association was obtained. Conditional on gene causality for a tag, BeviMed reported posterior probabilities over the mode of inheritance, the molecular consequence class of variants that mediate disease risk (for example, variants in the 5′ untranslated region 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 posterior probabilities implied a posterior estimate for the positive predictive value of 93%. The 95 genes included

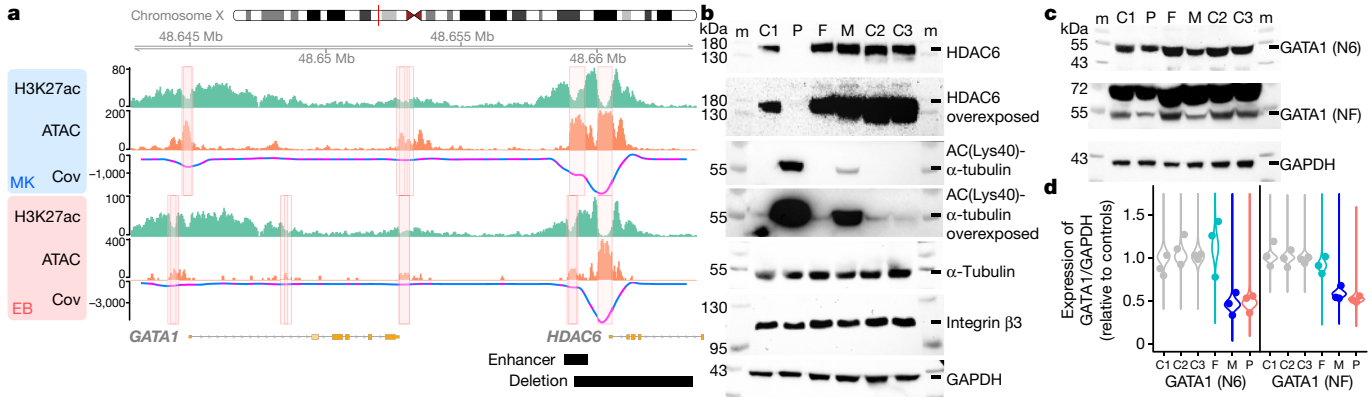


Fig. 4 | Causal variants in regulatory elements. a, From top to bottom: X chromosome ideogram; read coverage of H3K27ac ChIP–seq (green) and ATAC-seq (orange) in megakaryocytes (MK); the smoothed covariance (Cov) between H3K27ac ChIP–seq and ATAC-seq coverages for megakaryocytes, which were used to call regulatory elements (overlying coral rectangles); pink segments indicate regions in which the locally normalized ATAC-seq coverage exceeds the locally normalized H3K27ac ChIP–seq coverage (Supplementary Information); the corresponding three tracks and overlays for erythroblasts (EB); gene exons are shown in orange; the *GATA1* enhancer and the large deletion in the proband are shown as horizontal bars. A regulatory element overlapping the enhancer was identified by RedPop in megakaryocytes and erythroblasts but not in the other four cell types (tracks for these cell types are not shown). The deleted element binds to transcription factors that are characteristic of the megakaryocyte lineage: FLI1, GATA1/2, MEIS1, RUNX1 and TAL1 (binding not shown). **b**–**d**, P, proband; 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 (N6 and NF). **d**, Dot plots of GATA1 protein quantifications (as in **c**). The underlying violin plots show posterior predictive densities for the distribution of standardized GATA1 expression. The 90% credible intervals for the ratio of expression using the N6 antibody in father, mother, proband to the geometric mean in controls were 0.86–1.45, 0.35–0.59 and 0.37–0.62, respectively; similarly, for the expression using the NF antibody, the 95% credible intervals were 0.80–1.05, 0.51–0.67 and 0.45–0.60, respectively.

68 established DGGs, 11 DGGs that were discovered since 2015[10,14,20,25–30] and 16 candidates that require further investigation (Fig. 2d and Supplementary Table 3). Therefore, 79 of the 95 associations are confirmed, setting a lower bound on the true positive predictive value of 83%, which is broadly in line with an ancestry-controlled statistical estimate of the study-wide positive predictive value 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 pulmonary arterial hypertension. 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[31]. These results and other findings from this project[9–11,14,20,23,32–36], 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.

Genetics of the tails of a quantitative trait

Several heritable rare diseases (for example, familial hypercholesterolaemia, combined variable immunodeficiency, thrombocytopenia and von Willebrand disease) are diagnosed and clinically characterized 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 that were likely to mediate red blood cell (RBC)-associated pathologies by WGS of UK Biobank participants in the tails of a univariate quantitative phenotype, computed to optimize rare variant heritability. We derived the univariate phenotype by considering the joint distribution of estimated effect sizes from GWAS associations between variants with minor allele frequencies of <1% and four RBC full blood count traits[7] (Fig. 3a). We sequenced 764 participants, 383 of whom were in the left tail of the phenotype, corresponding to a low RBC count and a high mean cell volume, and 381 of whom were in the right tail of the phenotype, corresponding to a high RBC count and a low mean cell volume (Fig. 3b, c).

The distribution of a polygenic predictor of the phenotype derived from an RBC full blood count GWAS exhibited left and right shifts from the population distribution in the respective tails (Fig. 3d). However, these shifts were less strong than predicted by Gaussian variance components modelling, a discrepancy that 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. Therefore, we treated each of the tail groups as a set of cases in a BeviMed analysis and identified 12 genes with a posterior probability of association >0.4, which is 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[7,37].

Aetiological variants in regulatory elements

Rare variants in regulatory elements can cause disease by disrupting transcription or translation[38,39]. Recent studies have suggested that—at least in neurodevelopmental disorders—a small percentage of cases are attributable to de novo non-coding SNVs in regulatory elements that are active in relevant tissues[40]. Larger variants may be more disruptive to regulatory elements than SNVs. We searched for aetiological variants, including large deletions, in the regulatory elements of 246 DGGs implicated in recessive haematopoiesis-related disorders (Supplementary Information). First, we defined a set of active regulatory elements—a 'regulome'—for each of six haematological cell types, by merging transcription-factor-binding sites identified by chromatin immunoprecipitation followed by sequencing (ChIP–seq)

with genomic regions called by RedPop. RedPop is a detection method that uses the negative covariance between data from the assay for transposase-accessible chromatin using sequencing (ATAC-seq) and ChIP–seq coverage of histone H3 K27 acetylation (H3K27ac) in regulatory elements (Supplementary Information). We linked the regulatory elements to genes using genomic proximity and promoter capture chromosome conformation capture (pcHi-C)[41]. Second, we assigned each regulome to one or more of three rare disease domains—bleeding, thrombotic and platelet disorders, primary immune disorders and stem cell and myeloid disorders—according to the relevance of the corresponding cell types to the domains (Supplementary Table 3). Last, we searched for cases with a rare homozygous or hemizygous deletion of a regulatory element active in a relevant cell type and linked to a DGG of the domain of the case. We also searched for deletions that met 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 patient with a primary immune disorder who carried a deletion overlapping the 5′ untranslated region of *ARPC1B* in compound heterozygosity with a frameshift variant in the same gene[36], a boy with autism spectrum disorder and thrombocytopenia who carried a hemizygous deletion of a *GATA1* enhancer and a patient with several autoimmune-mediated cytopenias who carried a homozygous deletion of an intronic CTCF-binding site[42] of *LRBA*.

The X-linked variant carried by the boy with autism spectrum disorder deleted a *GATA1* enhancer and exons 1–4 of *HDAC6* (Fig. 4 and Extended Data Fig. 9). He had a persistently low platelet count (52×10^9 l^{-1}), an elevated mean platelet volume (15.1 fl) and normal RBC parameters except for mild dyserythropoiesis. Electron microscopy analyses showed lower than usual platelet α-granule content. Stem cell culture recapitulated poor platelet formation by megakaryocytes. These symptoms are typical of patients with a pathogenic coding *GATA1* allele[43]. His platelets contained abnormally low GATA1, consistent with weak transcription due to the deletion of the enhancer[44]. HDAC6 deacetylases Lys40 of α-tubulin, which localizes in polymerized microtubules[45]. The absence of HDAC6 was accompanied by an increase in acetylated α-tubulin in platelets. Knockout of the mouse homologue, *Hdac6*, causes aberrant acetylation of α-tubulin, which leads to bleeding[46] and abnormal behaviour[47]. Thus, the reduced expression of *GATA1* and the absence of HDAC6 jointly caused a previously undescribed syndrome of macrothrombocytopenia that is accompanied by neurodevelopmental problems. The patient with a homozygous deletion of a CTCF-binding site in the first intron of *LRBA* presented with autoantibody-mediated pancytopenia due to a loss of tolerance for multiple autoantigens, which is characteristic of impaired *LRBA* function[48].

We adapted our approaches for identifying pathogenic deletions in regulatory elements to identify pathogenic non-coding SNVs. We focused on SNVs with a combined annotation-dependent depletion (CADD)[49] 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 (chromosome 1: 43803414G>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[50]. Loss of *MPL* causes chronic amegakaryocytic thrombocytopenia[51]. The SNV was in a megakaryocyte-specific RedPop-identified regulatory element. 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 a significant reduction in platelet MPL levels. In contrast to *MPL*-null patients[52], who are severely thrombocytopenic because their bone marrow is almost devoid of megakaryocytes, the patient had platelet counts of 45×10^9 l^{-1} and a bone 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 individuals.

Discussion

The resolution of unknown rare disease aetiologies will be hastened by the standardization 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. These associations can identify genes that mediate Mendelian pathologies. We have also shown that epigenetic data for cell types that mediate aetiology, combined with WGS, can identify regulatory elements that contain pathogenic non-coding mutations. The exploration of regulatory variation is a promising focus for future research and clinical intervention.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2434-2.

1. Ferreira, C. R. The burden of rare diseases. *Am. J. Med. Genet. A.* **179**, 885–892 (2019).
2. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
3. Boycott, K. M. et al. International cooperation to enable the diagnosis of all rare genetic diseases. *Am. J. Hum. Genet.* **100**, 695–705 (2017).
4. Vissers, L. E. L. M. et al. A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. *Genet. Med.* **19**, 1055–1063 (2017).
5. Richards, S. 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.* **17**, 405–423 (2015).
6. Van Houten, C. V. et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. Preprint at *bioRxiv* https://doi.org/10.1101/572347 (2019).
7. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
8. Belkadi, A. et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl Acad. Sci. USA* **112**, 5473–5478 (2015).
9. Carss, K. J. et al. Comprehensive rare variant analysis via whole-genome sequencing to determine the molecular pathology of inherited retinal disease. *Am. J. Hum. Genet.* **100**, 75–90 (2017).
10. Meyer, E. et al. Mutations in the histone methyltransferase gene *KMT2B* cause complex early-onset dystonia. *Nat. Genet.* **49**, 223–237 (2017).
11. Stritt, S. et al. A gain-of-function variant in *DIAPH1* causes dominant macrothrombocytopenia and hearing loss. *Blood* **127**, 2903–2914 (2016).
12. Westbury, S. K. et al. Phenotype description and response to thrombopoietin receptor agonist in *DIAPH1*-related disorder. *Blood Adv.* **2**, 2341–2346 (2018).
13. Turro, E. et al. A dominant gain-of-function mutation in universal tyrosine kinase SRC causes thrombocytopenia, myelofibrosis, bleeding, and bone pathologies. *Sci. Transl. Med.* **8**, 328ra30 (2016).
14. Tuijnenburg, P. et al. Loss-of-function nuclear factor κB subunit 1 (*NFKB1*) variants are the most common monogenic cause of common variable immunodeficiency in Europeans. *J. Allergy Clin. Immunol.* **142**, 1285–1296 (2018).
15. Noris, P. et al. *ANKRD26*-related thrombocytopenia and myeloid malignancies. *Blood* **122**, 1987–1989 (2013).
16. Noetzli, L. et al. Germline mutations in *ETV6* are associated with thrombocytopenia, red cell macrocytosis and predisposition to lymphoblastic leukemia. *Nat. Genet.* **47**, 535–538 (2015).
17. Song, W. J. et al. Haploinsufficiency of *CBFA2* causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. *Nat. Genet.* **23**, 166–175 (1999).
18. Evans, J. D. et al. *BMPR2* mutations and survival in pulmonary arterial hypertension: an individual participant data meta-analysis. *Lancet Respir. Med.* **4**, 129–137 (2016).
19. Hadinnapola, C. et al. Phenotypic characterization of *EIF2AK4* mutation carriers in a large cohort of patients diagnosed clinically with pulmonary arterial hypertension. *Circulation* **136**, 2022–2033 (2017).

20. Gräf, S. et al. Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. *Nat. Commun.* **9**, 1416 (2018).
21. Philippakis, A. A. et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum. Mutat.* **36**, 915–921 (2015).
22. Padmakumar, M. et al. A novel missense variant in *SLC18A2* causes recessive brain monoamine vesicular transport disease and absent serotonin in platelets. *JIMD Rep.* **47**, 9–16 (2019).
23. Ito, Y. et al. De novo truncating mutations in *WASF1* cause intellectual disability with seizures. *Am. J. Hum. Genet.* **103**, 144–153 (2018).
24. Greene, D., Richardson, S. & Turro, E. A fast association test for identifying pathogenic variants involved in rare diseases. *Am. J. Hum. Genet.* **101**, 104–114 (2017).
25. Merico, D. et al. Compound heterozygous mutations in the noncoding *RNU4ATAC* cause Roifman syndrome by disrupting minor intron splicing. *Nat. Commun.* **6**, 8718 (2015).
26. Ananth, A. L. et al. Clinical course of six children with *GNAO1* mutations causing a severe and distinctive movement disorder. *Pediatr. Neurol.* **59**, 81–84 (2016).
27. Horn, D. et al. Biallelic *COL3A1* mutations result in a clinical spectrum of specific structural brain anomalies and connective tissue abnormalities. Am. *J. Med. Genet. A.* **173**, 2534–2538 (2017).
28. Khan, S. Y. et al. Splice-site mutations identified in *PDE6A* responsible for retinitis pigmentosa in consanguineous Pakistani families. *Mol. Vis.* **21**, 871–882 (2015).
29. Petrovski, S. et al. Germline de novo mutations in GNB1 cause severe neurodevelopmental disability, hypotonia, and seizures. *Am. J. Hum. Genet.* **98**, 1001–1010 (2016).
30. Akawi, N. et al. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat. Genet.* **47**, 1363–1369 (2015).
31. Sivapalaratnam, S. et al. Rare variants in *GP1BB* are responsible for autosomal dominant macrothrombocytopenia. *Blood* **129**, 520–524 (2017).
32. Westbury, S. K. et al. Expanded repertoire of *RASGRP2* variants responsible for platelet dysfunction and severe bleeding. *Blood* **130**, 1026–1030 (2017).
33. Pleines, I. et al. Mutations in tropomyosin 4 underlie a rare form of human macrothrombocytopenia. *J. Clin. Invest.* **127**, 814–829 (2017).
34. Heremans, J. et al. Abnormal differentiation of B cells and megakaryocytes in patients with Roifman syndrome. *J. Allergy Clin. Immunol.* **142**, 630–646 (2018).
35. Lentaigne, C. et al. Germline mutations in the transcription factor IKZF5 cause thrombocytopenia. *Blood* **134**, 2070–2081 (2019).
36. Thaventhiran, J. E. D. et al. Whole-genome sequencing of a sporadic primary immunodeficiency cohort. *Nature* (2020). https://doi.org/10.1038/s41586-020-2265-1
37. Natarajan, P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun.* **9**, 3391 (2018).
38. Giardine, B. et al. Updates of the HbVar database of human hemoglobin variants and thalassemia mutations. *Nucleic Acids Res.* **42**, D1063–D1069 (2014).
39. Albers, C. A. 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.* **44**, 435–439 (2012).
40. Short, P. J. et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).
41. Javierre, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384 (2016).
42. Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **15**, 234–246 (2014).
43. Freson, K. et al. Platelet characteristics in patients with X-linked macrothrombocytopenia because of a novel GATA1 mutation. *Blood* **98**, 85–92 (2001).
44. Fulco, C. P. et al. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
45. Skultetyova, L. et al. Human histone deacetylase 6 shows strong preference for tubulin dimers over assembled microtubules. *Sci. Rep.* **7**, 11547 (2017).
46. Sadoul, K. et al. HDAC6 controls the kinetics of platelet activation. *Blood* **120**, 4215–4218 (2012).
47. Fukada, M. et al. Loss of deacetylation activity of Hdac6 affects emotional behavior in mice. *PLoS One* **7**, e30924 (2012).
48. Lopez-Herrera, G. et al. Deleterious mutations in *LRBA* are associated with a syndrome of immune deficiency and autoimmunity. *Am. J. Hum. Genet.* **90**, 986–1001 (2012).
49. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
50. Wendling, F. et al. cMpl ligand is a humoral regulator of megakaryocytopoiesis. *Nature* **369**, 571–574 (1994).
51. Tijssen, M. R. et al. Functional analysis of single amino-acid mutations in the thrombopoietin-receptor Mpl underlying congenital amegakaryocytic thrombocytopenia. *Br. J. Haematol.* **141**, 808–813 (2008).
52. Ballmaier, M. & Germeshausen, M. Congenital amegakaryocytic thrombocytopenia: clinical presentation, diagnosis, and treatment. *Semin. Thromb. Hemost.* **37**, 673–681 (2011).

[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, KU Leuven, 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]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, UK. [13]High Performance Computing Service, University of Cambridge, Cambridge, UK. [14]Genomics England Ltd, London, UK. [15]William Harvey Research Institute, NIHR Biomedical Research Centre at Barts, Queen Mary University of London, London, UK. [16]Department of Clinical Neurosciences, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [17]Medical Research Council Mitochondrial Biology Unit, Cambridge Biomedical Campus, Cambridge, UK. [18]Women and Children's Health, School of Life Course Sciences, King's College London, London, UK. [19]Department of Renal Medicine, University College London, London, UK. [20]Rare Renal Disease Registry, UK Renal Registry, Bristol, UK. [21]King's College London, London, UK. [22]Department of Paediatric Nephrology, Evelina London Children's Hospital, Guy's & St Thomas' NHS Foundation Trust, London, UK. [23]Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [24]Centre for Haematology, Imperial College London, London, UK. [25]Department of Medical Genetics, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [26]NIHR Cambridge Biomedical Research Centre, Cambridge Biomedical Campus, Cambridge, UK. [27]Cancer Research UK Cambridge Centre, Cambridge Biomedical Campus, Cambridge, UK. [28]Stroke Research Group, Department of Clinical Neurosciences, University of Cambridge, Cambridge Biomedical Campus, Bristol, UK. [29]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [30]School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK. [31]University Hospitals Bristol NHS Foundation Trust, Bristol, UK. [32]Department of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [33]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [34]Department of Paediatrics, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [35]Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [36]Bristol Renal and Children's Renal Unit, Bristol Medical School, University of Bristol, Bristol, UK. [37]Bristol Royal Hospital for Children, University Hospitals Bristol NHS Foundation Trust, Bristol, UK. [38]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [39]UCL Great Ormond Street Institute of Child Health, London, UK. [40]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [41]Moorfields Eye Hospital NHS Trust, London, UK. [42]UCL Institute of Opthalmology, University College London, London, UK. [43]Department of Medicine, Imperial College London, London, UK. [44]Institute of Reproductive and Developmental Biology, Department of Surgery and Cancer, Faculty of Medicine, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [45]Illumina Cambridge, Little Chesterford, 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, 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[29], Philip Ancliff[50], Richard Antrobus[56], Ruth Armstrong[25,26,27], Gavin Arno[41,42], Sofie Ashford[2,38], William J. Astle[3,4], Anthony Attwood[1,2], Paul Aurora[50], Christian Babbs[12,33], Chiara Bacchelli[12,39], 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[11,12], David R. Bentley[45], Agnieszka Bierzynska[36], 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[2,5,26,46,47], 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[21], John Burn[76], Siobhan O. Burns[85,86], Oliver S. Burren[5,8], Nigel Burrows[46], Paul Calleja[13], Carolyn Campbell[87], Gerald Carr-White[88], Keren Carss[1,2], Ruth Casey[25,26,27], Mark J. Caulfield[14,15], Jenny Chambers[18,89], John Chambers[90,91,92,93,94], Melanie M. Y. Chan[19], Calvin Cheah[71], Floria Cheng[89], Patrick F. Chinnery[2,16,17], 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[36], 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[43], Paul A. Corris[105,106], Abigail Furnell[1,2], Fiona Cunningham[29], Nicola S. Curry[107], Antony J. Cutler[108], Matthew J. Daniels[32,35,110], Mehul Dattani[67,111], Louise C. Daugherty[1,2,14], John Davis[1,2], Anthony De Soyza[75,106,112], Sri V. V. Deevi[1,2], Timothy Dent[35], Charu Deshpande[83], Eleanor F. Dewhurst[1,2], Peter H. Dixon[18], Sofia Douzgou[58,59], Kate Downes[1,2], Anna M. Drazyk[28], 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[126,128], Hiva Fassihi[127], Remi Favier[129], Jack Findhammer[130], Debra Fletcher[1,2], Frances A. Flinter[83], R. Andres Floto[5,46,66], Tom Fowler[14,15], James Fox[1,2], Amy J. Frary[1,2], Courtney E. French[64], Kathleen Freson[7], Mattia Frontini[1,4,6], Daniel P. Gale[19,20], 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[43,131], J. Simon R. Gibbs[80], Kate Gibson[87], Kimberly C. Gilmour[39,50], 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], Lynn Greenhalgh[141], Andreas Greinacher[142], Paolo Gresele[143], Philip Griffiths[144,145], Sofia Grigoriadou[146], Russell J. Grocock[45], Detelina Grozeva[25], 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[32,40], Claire L. Harris[106,109], Daniel Hart[98], Ahamad Hassan[152], Grant Hayman[153], Alex Henderson[76], Archana Herwadkar[53], Jonathan Hoffman[97], Simon Holden[154], Rita Horvath[144,155], Henry Houlden[156], Arjan C. Houweling[223], Luke S. Howard[80,157], Fengyuan Hu[1,2], Gavin Hudson[145], 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[106,109,158], Stephen Jolles[159], Jennifer Jolley[1,2], Dragana Josifova[83], Neringa Jurkute[41,61], Tim Karten[130], Johannes Karten[130], Mary A. Kasanicki[46], Hanadi Kazkaz[160], Rashid Kazmi[70], Peter Kelleher[161,162], Anne M. Kelly[46], Wilf Kelsall[46], Carly Kempster[1,2], David G. Kiely[100], Nathalie Kingston[1,2], Robert Klima[13], Nils Koelling[163], Myrto Kostadima[1], Gabor Kovacs[164,165], Ania Koziell[21,22], Roman Kreuzhuber[1,2], Taco W. Kuijpers[166,167], Ajith Kumar[52], Dinakantha Kumararatne[168], Manju A. Kurian[169,170], Michael A. Laffan[23,24], Fiona Lalloo[59], Michele Lambert[171,172], Hana Lango Allen[1,2], Allan Lawrie[173], D. Mark Layton[23,24], Nick Lench[71], Claire Lentaigne[23,24], Tracy Lester[87], Adam P. Levine[19], Rachel Linger[2,38], Hilary Longhurst[174], Lorena E. Lorenzo[146], Eleni Louka[12,33], Paul A. Lyons[5,8], Rajiv D. Machado[175,176], Robert V. MacKenzie Ross[177], Bella Madan[178], Eamonn R. Maher[25,26,27], Jesmeen Maimaris[39], Samantha Malka[41,42], Sarah Mangles[125], Rutendo Mapeta[1,2], Kevin J. Marchbank[106,179], Stephen Marks[50], Hugh S. Markus[28], Hanns-Ulrich Marschall[180], Andrew Marshall[181,182,183], Jennifer Martin[2,5,38], Mary Mathias[184], Emma Matthews[185,186], Heather Maxwell[149], Paul McAlinden[106], Mark I. McCarthy[12,40,187], Harriet McKinney[1,2], Aoife McMahon[29], Stuart Meacham[1,2], Adam J. Mead[33], Ignacio Medina Castello[13], Karyn Megy[1,2], Sarju G. Mehta[154], Michel Michaelides[41,42], Carolyn Millar[23,24], Shehla N. Mohammed[83], Shahin Moledina[50], David Montani[134,135,136], Anthony T. Moore[41,42,188], Joannella Morales[29], Nicholas W. Morrell[2,5], Monika Mozere[19], Keith W. Muir[189], Andrew D. Mumford[30,31], Andrea H. Nemeth[11,190], William G. Newman[58,59], Michael Newnham[5,66], Sadia Noorani[191], Paquita Nurden[192], Jennifer O'Sullivan[178], Samya Obaji[193], Chris Odhams[14], Steven Okoli[12,33], Andrea Olschewski[164], Horst Olschewski[164,165], Kai Ren Ong[97], S. Helen Oram[194], Elizabeth Ormondroyd[12,32], Willem H. Ouwehand[1,2,4,6,48], Claire Palles[195], Sofia Papadia[2,38], Soo-Mi Park[26,27,49], David Parry[10], Smita Patel[196], Joan Paterson[25,26,27], Andrew Peacock[96], Simon H. 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], Kenneth E. S. Poole[5,46], Radhika Prathalingam[71], Bethan Psaila[12,33], Angela Pyle[145], Richard Quinton[75,145], Shamima Rahman[50,67], Stuart Rankin[13], Anupama Rao[50], F. Lucy Raymond[2,25], Paula J. Rayner-Matthews[1,2], Christine Rees[45], Augusto Rendon[1,14], Tara Renton[198], Christopher J. Rhodes[43], Andrew S. C. Rice[199,200], Sylvia Richardson[3], Alex Richter[56], Leema Robert[83], Irene Roberts[12,33,34], Anthony Rogers[71], Sarah J. Rose[83], Robert Ross-Russell[46], Catherine Roughley[123], Noemi B. A. Roy[12,33,35], Deborah M. Ruddy[83], Omid Sadeghi-Alavijeh[19], Moin A. Saleem[36,37], Nilesh Samani[202], Crina Samarghitean[1,2], Alba Sanchis-Juan[1,2], Ravishankar B. Sargur[117], Robert N. Sarkany[127], Simon Satchell[36,203], Sinisa Savic[204,205,206], John A. Sayer[75,145], Genevieve Sayer[83], Laura Scelsi[133], Andrew M. Schaefer[75,144], Sol Schulman[207], Richard Scott[14,50], Marie Scully[160], Claire Searle[208], Werner Seeger[131], Arjune Sen[12,209,210], W. A. Carrock Sewell[211], Denis Seyres[1,2], Neil Shah[39,50], Olga Shamardina[1,2], Susan E. Shapiro[107], Adam C. Shaw[83], Patrick J. Short[48], Keith Sibson[184], Lucy Side[212], Ilenia Simeoni[1,2], Michael A. Simpson[213], Matthew C. Sims[1,214], Suthesh Sivapalaratnam[4,98,215,216], Damian Smedley[14,15], Katherine R. Smith[14], Kenneth G. C. Smith[5,8], Katie Snape[150], Nicole Soranzo[1,48], Florent Soubrier[126], Laura Southgate[176,217], Olivera Spasic-Boskovic[49], Simon Staines[1,2], Emily Staples[5], Hannah Stark[2,38], Jonathan Stephens[1,2], Charles Steward[71], Kathleen E. Stirrups[1,2], Alex Stuckey[14], Jay Suntharalingam[177], Emilia M. Swietlik[5], Petros Syrris[118], R. Campbell Tait[218], Kate Talks[68], Rhea Y. Y. Tan[28], Katie Tate[71], John M. Taylor[87], Jenny C. Taylor[12,40], James E. Thaventhiran[8,219], Andreas C. Themistocleous[11], Ellen Thomas[14,15,88], David Thomas[5], Moira J. Thomas[220,221], Patrick Thomas[1,2], Kate Thomson[32,35], Adrian J. Thrasher[39], Glen Threadgold[29], Chantal Thys[7], Tobias Tilly[1,2], Marc Tischkowitz[25,26,49], Catherine Titterton[1,2], John A. Todd[108], Cheng-Hock Toh[114], Bas Tolhuis[130], Ian P. Tomlinson[195], Mark Toshner[5,66], Matthew Traylor[28], Carmen Treacy[5,66], Paul Treadaway[1,2], Richard Trembath[21], Salih Tuna[1,2], Wojciech Turek[13], Ernest Turro[1,2,3], Philip Twiss[49], Tom Vale[11], Chris Van Geet[7], Natalie van Zuydam[40,197], Maarten Vandekuilen[130], Anthony M. Vandersteen[222], Marta Vazquez-Lopez[89], Julie von Ziegenweidt[1,2], Anton Vonk Noordegraaf[69], Annette Wagner[46], Quinten Waisfisz[223], Suellen M. Walker[39,50], Neil Walker[1,2], Klaudia Walter[48], James S. Ware[80,81,101], Hugh Watkins[32,35,40], Christopher Watt[1,2], Andrew R. Webster[41,42], Lucy Wedderburn[39,224,225], Wei Wei[16,17], Steven B. Welch[226], Julie Wessels[137], Sarah K. Westbury[30,31], John-Paul Westwood[160], John Wharton[43], Deborah Whitehorn[1,2], James Whitworth[25,26,27], Andrew O. M. Wilkie[163], Martin R. Wilkins[43], Catherine Williamson[18,44], Brian T. Wilson[52,75,145], Edwin K. S. Wong[75,106], Nicholas Wood[156,227], Yvette Wood[1,2], Christopher Geoffrey Woods[25,46], Emma R. Woodward[59], Stephen J. Wort[81,228], Austen Worth[50], Michael Wright[76], Katherine Yates[1,2,5], Patrick F. K. Yong[229], Timothy Young[1,2], Ping Yu[1,2], Patrick Yu-Wai-Man[16,17,230] & 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, Tubingen, 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, Amesterdam University Medical Centres, VU University Medical Centre, Amsterdam, The Netherlands. [70]Southampton General Hospital, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [71]Congenica, Biodata Innovation Centre, 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, Department of 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, London, 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 and Metabolic Disorders, Duke-National University of Singapore, Singapore, Singapore. [104]Department of Immunology and Inflammation, Faculty of Medicine, Imperial College London, London, UK. [105]National Pulmonary Hypertension Service (Newcastle), The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [106]Translational and Clinical Research Institute, 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]The National Renal Complement Therapeutics Centre, Royal Victoria Infirmary, Newcastle upon Tyne, UK. [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, and PathWest Laboratory Medicine, Crawley, Western Australia, 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, Guy's and St Thomas' NHS Foundation Trust, London, UK. [128]UMRS 1166-ICAN, INSERM, UPMC, Sorbonne Universités, Paris, France. [129]Service d'Hematologie biologique, Centre de Reference des Pathologies Plaquettaires, Hopital Armand Trousseau, Assistance Publique-Hopitaux de Paris, Paris, France. [130]GENALICE, 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, NY, USA. [133]Division of Cardiology, Fondazione IRCCS Policlinico S. Matteo, Pavia, Italy. [134]Université 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, NY, 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, South Australia, 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]John Walton Muscular Dystrophy Research Centre, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [156]Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK. [157]National Pulmonary Hypertension Service, Imperial College Healthcare NHS Trust, London, UK. [158]Department of Paediatric Nephrology, Great North Children's Hospital, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [159]Immunodeficiency Centre for Wales, University Hospital of Wales, Cardiff, UK. [160]University College London Hospitals NHS Foundation Trust, London, UK. [161]Centre for Immunology & Vaccinology, Department of Medicine, Chelsea & Westminster Hospital, Imperial College London, London, UK. [162]Department of Respiratory Medicine, Royal Brompton & Harefield NHS Foundation Trust, London, UK. [163]MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK. [164]Ludwig Boltzmann Institute for Lung Vascular Research, Graz, Austria. [165]Department of Internal Medicine, Division of Pulmonology, Medical University of Graz, Graz, Austria. [166]Department of Pediatric Hematology, Immunology, Rheumatology and Infectious Diseases, Emma Children's Hospital, Academic Medical Center (AMC), University of Amsterdam, Amsterdam, The Netherlands. [167]Department of Blood Cell Research, Sanquin, Amsterdam, The Netherlands. [168]Department of Clinical Immunology, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [169]Developmental Neurosciences, UCL Great Ormond Street Institute of Child Health, London, UK. [170]Department of Neurology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [171]Division of Hematology, The Children's Hospital of Philadelphia, Philadelphia, PA, USA. [172]Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [173]Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, UK. [174]Department of Specialist Allergy and Clinical Immunology, University College Hospital, University College London Hospitals NHS Foundation Trust, London, UK. [175]School of Life Sciences, University of Lincoln, Lincoln, UK. [176]Molecular and Clinical Sciences Research Institute, St George's University of London, London, UK. [177]Royal United Hospitals Bath NHS Foundation Trust, Bath, UK. [178]Department of Haematology, Guy's and St Thomas' NHS Foundation Trust, London, UK. [179]The National Renal Complement Therapeutics Centre, Royal Victoria Infirmary, Newcastle upon Tyne, UK. [180]Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [181]Faculty of Biology, Medicine and Health, School of Biological Sciences, Division of Neuroscience and Experimental Psychology, University of Manchester, Manchester, UK. [182]Department of Clinical Neurophysiology, Manchester University NHS Foundation Trust, Manchester, UK. [183]National Institute for Health Research/Wellcome Trust Clinical Research Facility, Manchester, UK. [184]Department of Haematology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [185]The National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK. [186]MRC Centre for Neuromuscular Diseases, Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK. [187]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK. [188]Ophthalmology Department, UCSF School of Medicine, San Francisco, CA, USA. [189]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. [190]Department of Clinical Genetics, Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK. [191]Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK. [192]Institut Hospitalo-Universitaire de Rythmologie et de Modelisation Cardiaque, Plateforme Technologique d'Innovation Biomedicale, Hopital Xavier Arnozan, Pessac, France. [193]The Arthur Bloom Haemophilia Centre, University Hospital of Wales, Cardiff, UK. [194]Department of Paediatric Haematology, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [195]Institute of Cancer and Genomic Sciences, Institute of Biomedical Research, University of Birmingham, Birmingham, UK. [196]Department of Clinical Immunology, John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [197]Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [198]King's College Hospital NHS Foundation Trust, London, UK. [199]Pain Research, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK. [200]Pain Medicine, Chelsea and Westminster Hospital NHS Foundation Trust, London, UK. [201]Department of Haematology, Oxford University Hospital Foundation Trust, Oxford, UK. [202]Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, UK. [203]North Bristol NHS Trust, Bristol, UK. [204]Department of Clinical Immunology and Allergy, St James's University Hospital, Leeds, UK. [205]The NIHR Leeds Biomedical Research Centre, Leeds, UK. [206]Leeds Institute of Rheumatic and Musculoskeletal Medicine, Leeds, UK. [207]Beth Israel Deaconess Medical Centre and Harvard Medical School, Boston, MA, USA. [208]Department of Clinical Genetics, Nottingham University Hospitals NHS Trust, Nottingham, UK. [209]Oxford Epilepsy Research Group, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. [210]Department of Neurology, John Radcliffe Hospital, Oxford, UK. [211]Scunthorpe General Hospital, Northern Lincolnshire and Goole NHS Foundation Trust, Scunthorpe, UK. [212]Wessex Clinical Genetics Service, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [213]Genetics and Molecular Medicine, King's College London, London, UK. [214]Oxford Haemophilia and Thrombosis Centre, Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK. [215]Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [216]Queen Mary University of London, London, UK. [217]Faculty of Life Sciences and Medicine, King's College London, London, UK. [218]Glasgow Royal Infirmary, NHS Greater Glasgow and Clyde, Glasgow, UK. [219]MRC Toxicology Unit, School of Biological Sciences, University of Cambridge, Cambridge, UK. [220]Gartnavel General Hospital, NHS Greater Glasgow and Clyde, Glasgow, UK. [221]Queen Elizabeth University Hospital, Glasgow, UK. [222]Division of Medical Genetics, IWK Health Centre, Dalhousie University, Halifax, Nova Scotia, Canada. [223]Department of Clinical Genetics, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [224]NIHR Great Ormond Street Biomedical Research Centre, London, UK. [225]Arthritis Research UK Centre for Adolescent Rheumatology, University College London, London, UK. [226]Birmingham Chest Clinic and Heartlands Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [227]UCL Genetics Institute, UCL Division of Biosciences, University College London, London, UK. [228]Imperial College London, London, UK. [229]Frimley Park Hospital, NHS Frimley Health Foundation Trust, Camberley, UK. [230]NIHR Biomedical Research Centre at Moorfields Eye Hospital, UCL Institute of Ophthalmology, London, UK.
*A full list of members and their affiliations appears in the Supplementary Information.

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 Cambridge University Hospitals. 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 (extreme red blood cell traits) (UKB) domain was jointly coordinated by the NBR and UK Biobank[2]. Participants in the rare disease domains were recruited mainly at NHS Hospitals in the United Kingdom, but also at hospitals overseas (Extended Data Fig. 1a and Supplementary Table 1). All 13,187 participants provided written informed consent, either under the East of England Cambridge South national research ethics committee (REC) reference no. 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 at which 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[53] 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 for DNA degradation (gel electrophoresis) and purity (using ratio of the absorbance at 260 and 280 nm ($A_{260/280}$); Trinean) before selection for WGS. DNA samples were prepared at a minimum concentration of 30 ng μl^{-1} in 110 µl, visually inspected for degradation and had to have an $A_{260/280}$ between 1.75 and 2.04. They were then prepared in batches of 96 and shipped on dry ice to the sequencing provider (Illumina). Further sample quality control was performed by Illumina to ensure that the concentration of the DNA was >30 ng μl^{-1} and that every sample generated high-quality microarray genotyping data (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 before data delivery. For each sample, 0.5 µg of DNA was fragmented using Covaris LE220 (Covaris) to obtain an average size of 450-bp DNA fragments. DNA samples were processed using the Illumina TruSeq DNA PCR-Free Sample Preparation kit (Illumina) on the Hamilton Microlab Star (Hamilton Robotics). The final libraries were checked using the Roche LightCycler 480 II (Roche Diagnostics) with KAPA Library Quantification Kit (Kapa Biosystems) for concentration. From February 2014 to June 2017, three read lengths were used: 100 bp, 125 bp and 150 bp (377, 3,154 and 9,656 samples, respectively). Samples sequenced with 100-bp and 125-bp reads used three and two lanes of an Illumina HiSeq 2500 instrument, respectively, while samples sequenced with 150-bp reads used a single lane of a HiSeq X instrument. At least 95% of the autosomal genome had to be covered at 15× and a maximum of 5% of insert sizes had to be less than twice the read length. Following sample and data quality control at Illumina, 13,187 sets of WGS data files were received by the University of Cambridge High Performance Computing Service (HPC) for further quality control.

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. In brief, 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 quality control checks were applied to ensure the SNV and indel call data 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 to 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 normalized with bcftools, loaded into our HBase database and filtered on their overall pass rate, as 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 or 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 (that is, all domains except UKB, GEL and a domain comprising technical controls) 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 on the basis of (by order of priority) community input, presence in the Locus Reference Genomic resource[54] or designation as 'canonical' in Ensembl. Variants (SNVs, indels) were shortlisted if (1) their MAFs in control populations[55] were less than 1 in 1,000 for putative novel causal variants and less than 25 in 1,000 for variants listed as disease-causing in HGMD; (2) their predicted impact according to the Variant Effect Predictor[56] 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; (3) the variant affected a DGG relevant to the disease of the patient. 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 MDTs 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 prioritized 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 for visualization in the Sapientia web application during MDT meetings. MDTs comprised experts from different hospitals across the United Kingdom 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 were marked in Sapientia as pathogenic, likely pathogenic or of uncertain significance. Only pathogenic and likely pathogenic variants were systematically reported and variants of uncertain significance were reported at the discretion of the MDT. As per the REC-approved study protocol, secondary findings (for example, 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[24] 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 to 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 (for example, because the phenotypes are similar) and the rare variants should be chosen such that they potentially share a mechanism of action on the phenotype (for example, because they are predicted to have a similar effect on a particular gene product). BeviMed computes posterior probabilities of no association, dominant association and recessive association and, conditional on dominant or recessive association, it computes the posterior probability 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 posterior probabilities 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 publication[24]. 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 posterior probability 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, B cells, erythroblasts, megakaryocytes, monocytes and resting $CD4^+$ T cells. For each cell type, we used open chromatin data (ATAC-seq or DNase-seq) and histone-modification data (H3K27ac) to identify regulatory elements using the RedPop method (Supplementary Information). Additionally, for megakaryocytes and erythroblasts, we had access to the following transcription-factor ChIP–seq data, which were used to call peaks and supplement the regulomes: FLI1, GATA1, GATA2, MEIS1, RUNX1, TAL1 and CTCF for megakaryocytes; GATA1, KLF1, NFE2 and TAL1 for erythroblasts; and CTCF for monocytes and B cells. For each cell type, the regulome build process proceeded as follows: (1) call RedPop regions using ATAC-seq or DNase-seq and H3K27ac-seq data; (2) call transcription factor and CTCF-binding peaks using ChIP–seq data if available and obtain enrichment scores; (3) discard peaks with an enrichment score <10 unless they overlap at least two other peaks; (4) collapse overlapping features to obtain a single genomic track; (5) merge features within 100 bp of each other. Each regulome feature was assigned a gene label using either gene annotations from Ensembl (v.75) or a compendium of previously published pcHi-C[41] as follows: (1) assign to a gene if the feature overlaps the gene or the region up to 10 kb either side of the gene body; (2) assign to a gene if the feature overlaps the pcHi-C 'blind' spot of the gene (this region is defined by three HindIII restriction fragments, incorporating the capture fragment overlapping the transcription start site of the target gene, 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 and *HDAC6* deletion

The *GATA1* enhancer and *HDAC6* deletion was confirmed by PCR using primers HDAC6-F: 5′-CATCTTCAAGAGGATCAGAGG-3′ and HDAC6-R: 5′-CATAGCTAGACACTGGTT-3′. Electron microscopy analysis of platelets was performed as described previously[43]. Immunostaining of resting and fibrinogen spread platelets was performed as described previously[34] and analysed by structured illumination microscopy (SIM, Elyra S.1, Zeiss). Total protein lysates were obtained from platelets for immunoblot analysis as described previously[57]. The following antibodies were used for SIM and immunoblot analysis: rabbit anti-HDAC6 (clone D2E5, Cell Signaling Technology), mouse anti-acetylated tubulin antibody (clone 6-11B-1, Sigma), mouse anti-α-tubulin (A11126, Thermo Fisher Scientific), rabbit anti-VWF (DAKO), mouse anti-CD63 and rat anti-GATA1 N6 (Santa Cruz Biotechnology), rabbit anti-GATA1 (NF; the antibody was produced against the recombinant N-terminal zinc finger[58]), rabbit anti-GAPDH (14C10, Cell Signaling) and anti-β3 integrin (sc-14009, Santa Cruz Biotechnology). The statistical analysis of the GATA1 data are 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, 551061), PE-labelled IgG1 against CD110 (clone REA250, Miltenyi Biotec) and a PE-labelled isotype control (clone MOPC-21, BD Pharmingen, 555749). In brief, a sample of EDTA-anticoagulated blood was incubated with anti-CD110 (or control) and anti-CD42b antibodies for 30 min. Mean fluorescence intensity (MFI) produced by the anti-CD110 antibody 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[59] 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 h, and the mean coverage was 863,986×.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

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 Ltd 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

that is being overseen by the UK Biobank (https://www.ukbiobank.ac.uk/). The full blood count 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 (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 (refer to the legend of Fig. 1 for domain acronym definitions): 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 (https://www.ncbi.nlm.nih.gov/geo/), EGA (https://ega-archive.org), or referenced to their publication as follows. H3K27ac ChIP–seq: activated $CD4^+$ T cells[60], B cells (ERR1043004, ERR1043129, ERR928206, ERR769436), erythroblasts (EGAD00001002377), megakaryocytes (EGAD00001002362), monocytes (ERR829362 (ERS257420), ERR829412 (ERS222466), ERR493634 (ERS214696)), resting $CD4^+$ T cells[60]. ATAC-seq: activated $CD4^+$ T cells (GSE124867), B cells (SRR2126769 (GSE71338)), erythroblasts (SRR5489430 (GSM2594182)), megakaryocytes (EGAD00001001871), monocytes (EGAD00001006065), resting $CD4^+$ T cells (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.

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/.

53. Robinson, P. N. et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).
54. MacArthur, J. A. et al. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Res.* **42**, D873–D878 (2014).
55. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
56. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
57. Di Michele, M. et al. An integrated proteomics and genomics analysis to unravel a heterogeneous platelet secretion defect. *J. Proteomics* **74**, 902–913 (2011).
58. de Waele, L. et al. Severe gastrointestinal bleeding and thrombocytopenia in a child with an anti-GATA1 autoantibody. *Pediatr. Res.* **67**, 314–319 (2010).
59. Sanchis-Juan, A. et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med.* **10**, 95 (2018).
60. Burren, O. S. et al. Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol.* **18**, 165 (2017).
61. Wijgaerts, A. et al. The transcription factor GATA1 regulates NBEAL2 expression through a long-distance enhancer. *Haematologica* **102**, 695–706 (2017).

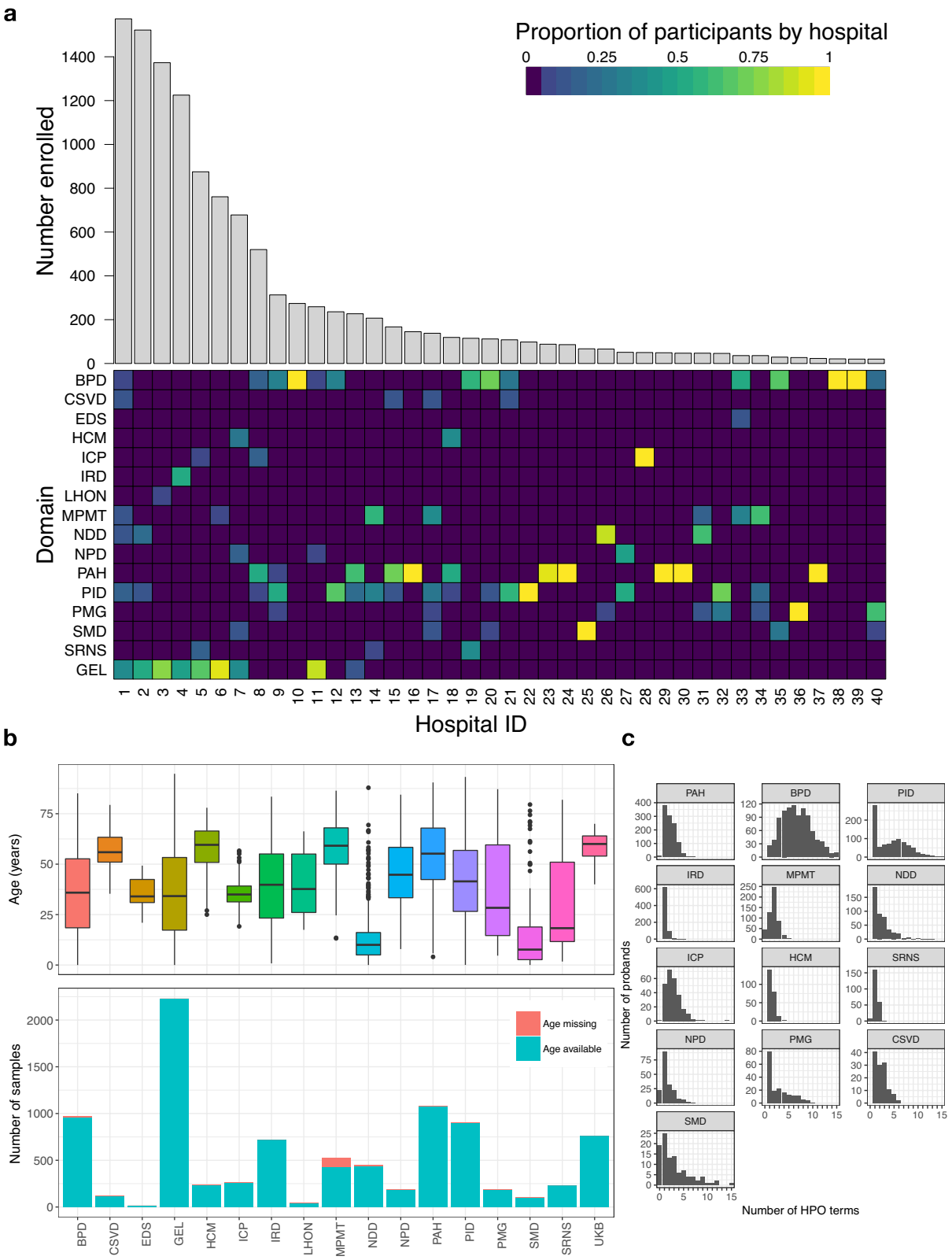
Acknowledgements This research was made possible through access to the data and findings generated by two pilot studies for the 100,000 Genomes Project. The enrolment was coordinated for one by the NIHR BioResource and for the other by Genomics England Ltd (GEL), a company wholly owned by the Department of Health in the United Kingdom. These pilot studies were mainly funded by grants from the NIHR in England to Cambridge University Hospitals and GEL, respectively. Additional funding was provided by the British Heart Foundation (BHF), MRC, NHS England, the Wellcome Trust and many other fund providers (see funding acknowledgements 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 the NIHR Biomedical Research Centres (BRC), NIHR BioResource Centres, NHS Trust Hospitals, NHS Blood and Transplant and their 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, BHF, Diabetes UK, Northwest Regional Development Agency, Scottish Government and Welsh Assembly Government. The MRC and Wellcome Trust had a key role in the decision to establish the UK Biobank. A. McMahon and J. Morales are funded by The Wellcome Trust (WT200990/Z/16/Z) and the European Molecular Biology Laboratory; K.G.C.S. holds a Wellcome Investigator Award, MRC Programme Grant (number MR/L019027/1); M.I.M. 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); R. Horvath 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); D.L.B. 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); J.S.W. 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; A.J.T. 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; L. Southgate is supported by the Wellcome Trust Institutional Strategic Support Fund (204809/Z/16/Z) awarded to St George's, University of London; M.J.D. receives funding from Wellcome Trust (WT098519MA); M.C.S. holds an MRC Clinical Research Training Fellowship (MR/R002363/1); J.A.S. is funded by MRC UK grant MR/M012212/1; A.J.M. received funding from an MRC Senior Clinical Fellowship (MR/L006340/1); C. Lentaigne received funding from an MRC Clinical Research Training Fellowship (MR/J011711/1); M.R.W. holds a NIHR award to the NIHR Imperial Clinical Research Facility at Imperial College Healthcare NHS Trust; C. Williamson holds a NIHR Senior Investigator Award; M. A. Kurian holds a NIHR Research Professorship (NIHR-RP-2016-07-019) and Wellcome Intermediate Fellowship (098524/Z/12/A); M.J.C. is an NIHR Senior Investigator and is funded by the NIHR Barts Biomedical Research Centre; N. Cooper is partially funded by NIHR Imperial College Biomedical Research Centre; C. Hadinnapola was funded through a PhD Fellowship by the NIHR Translational Research Collaboration - Rare Diseases; A.D.M. and S.K.W. were funded by the NIHR Bristol Biomedical Research Centre; E.L.M. received funding from the NIHR Biomedical Research Centre at University College London Hospitals; K.C.G. received funding from the NIHR Great Ormond Street Biomedical Research Centre; I.R. and E. Louka are supported by the NIHR Translational Research Collaboration - Rare Diseases; J.C.T., J.M.T. and S. Patel are funded by the NIHR Oxford Biomedical Research Centre; G. Arno is funded by a Fight for Sight (UK) Early Career Investigator Award (5045-5046); All authors affiliated with Moorfields Eye hospital and Institute of Ophthalmology are 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; A.T.M. is funded by Retinitis Pigmentosa Fighting Blindness, P.Y.-W.-M. is supported by grants from MRC UK (G1002570), Fight for Sight (1570/1571 and 24TP171), NIHR (IS-BRC-1215-20002); S.O.B. is supported by NIHR Translational Research Collaboration - Rare Diseases (01/04/15-30/04/2017); A.R.W. 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; A.C.T. is a member of the International Diabetic Neuropathy Consortium, the Novo Nordisk Foundation (NNF14SA0006) and is a member of the DOLORisk consortium funded by the European Commission Horizon 2020 (ID633491); J. Whitworth is a recipient of a Cancer Research UK Cambridge Cancer Centre Clinical Research Training Fellowship; S.A.J. is funded by Kids Kidney Research; D.P.G. 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 for the PMG domain; K.J.M. is supported by the Northern Counties Kidney Research Fund; P.H.D. receives funding from ICP Support; T.K.B. received a PhD fellowship from the NHSBT and British Society of Haematology; H.S.M. receives support from BHF Programme Grant RG/16/4/32218; A.L. is a BHF Senior Basic Science Research Fellow - FS/13/48/30453; K.F. and C.V.-G. are supported by the Research Council of the University of Leuven (BOF KU Leuven, Belgium; OT/14/098); H.J.B. works for the Netherlands CardioVascular Research Initiative (CVON); Fiona Cunningham, Aoife McMahon, Glen Threadgold, and Joannella Morales received funding from the Wellcome Trust (grant numbers WT108749/Z/15/Z and WT200990/Z/16/Z) and the European Molecular Biology Laboratory. 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 of England or any of the other funding agencies.

Author contributions Details of author contributions can be found in the Supplementary information, which contains the full list of consortium members and working groups.

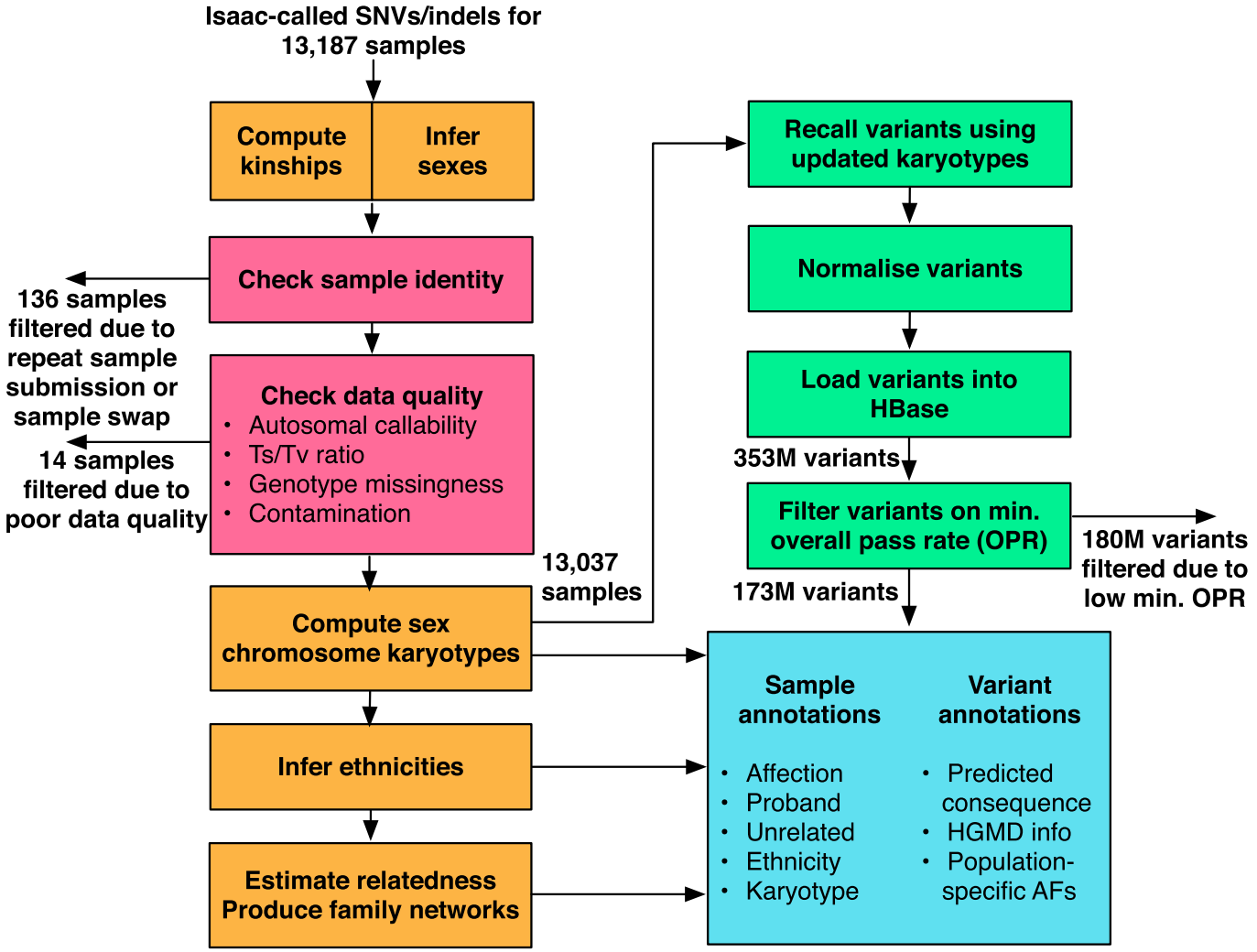
Competing interests A.M.K. had no competing interests at the time of the study, but after the study received an educational grant from CSL Behring to attend the ISTH meeting (2017); T.J.A. has received consultancy payments from AstraZeneca within the past 5 years and has received speaker honoraria from Illumina; A. Rogers, C. Cheah, C. Steward, E.B., K. Tate, N. Lench and R. Prathalingam are employees of Congenica; B. Tolhuis, J. Findhammer, J.K., M.V. and T. Karten

are employees of GENALICE; C. Colombo, C. Geoghegan, C.J.B., C. Rees, D.R.B., J.F.P., J. Hughes, R.J.G., S. Humphray, S. Hunter and T.S.A.G. are employees of Illumina Cambridge; C.V.-G. is the holder of the Bayer and Norbert Heimburger (CSL Behring) Chair; K.J.M. previously received funding for research and is currently on the scientific advisory board of Gemini Therapeutics; M.C.S. received travel and accommodation fees from NovoNordisk; D.M.L. serves on advisory boards for Agios, Novartis and Cerus; M.I.M. serves on advisory panels for Pfizer, NovoNordisk and Zoe Global, has received honoraria from Pfizer, NovoNordisk and Eli Lilly, has stock options in Zoe Global, and has received research funding from Abbvie, AstraZeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, NovoNordisk, Pfizer, Roche, Sanofi Aventis, Servier and Takeda; DLB has acted as a consultant on behalf of Oxford Innovation in the last 2 years for the following companies: Amgen, CODA therapeutic, Bristows, Lilly, Mundipharma, Regeneron and Theranexus, he holds an MRC Industrial Partnership grant with Astra Zeneca. The remaining authors declare no competing interests.

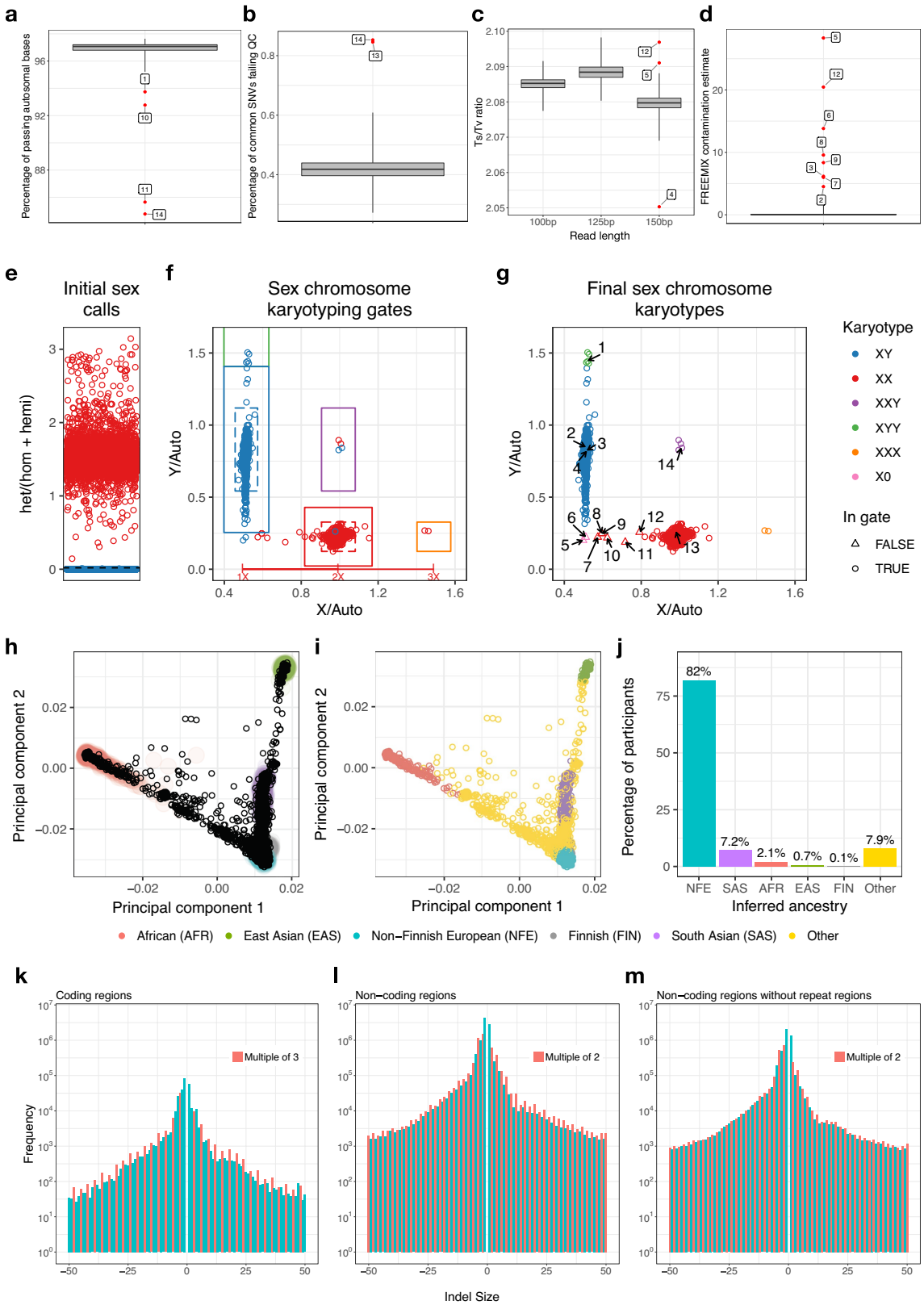
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2434-2.
Correspondence and requests for materials should be addressed to E.T., F.L.R. or W.H.O.
Peer review information *Nature* thanks Heidi L. Rehm, V. G. Sankaran, Shamil Sunyaev and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Demographic and phenotypic characteristics. a, The number of enrolments at the 40 hospitals with at least 20 enrolled participants. The heat map shows the distribution of enrolments over domains at each of the 40 hospitals. Hospital IDs are described in Supplementary Table 1. **b**, Top, age at recruitment for all probands in the 15 rare disease domains, GEL and UKB. Bottom, 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.

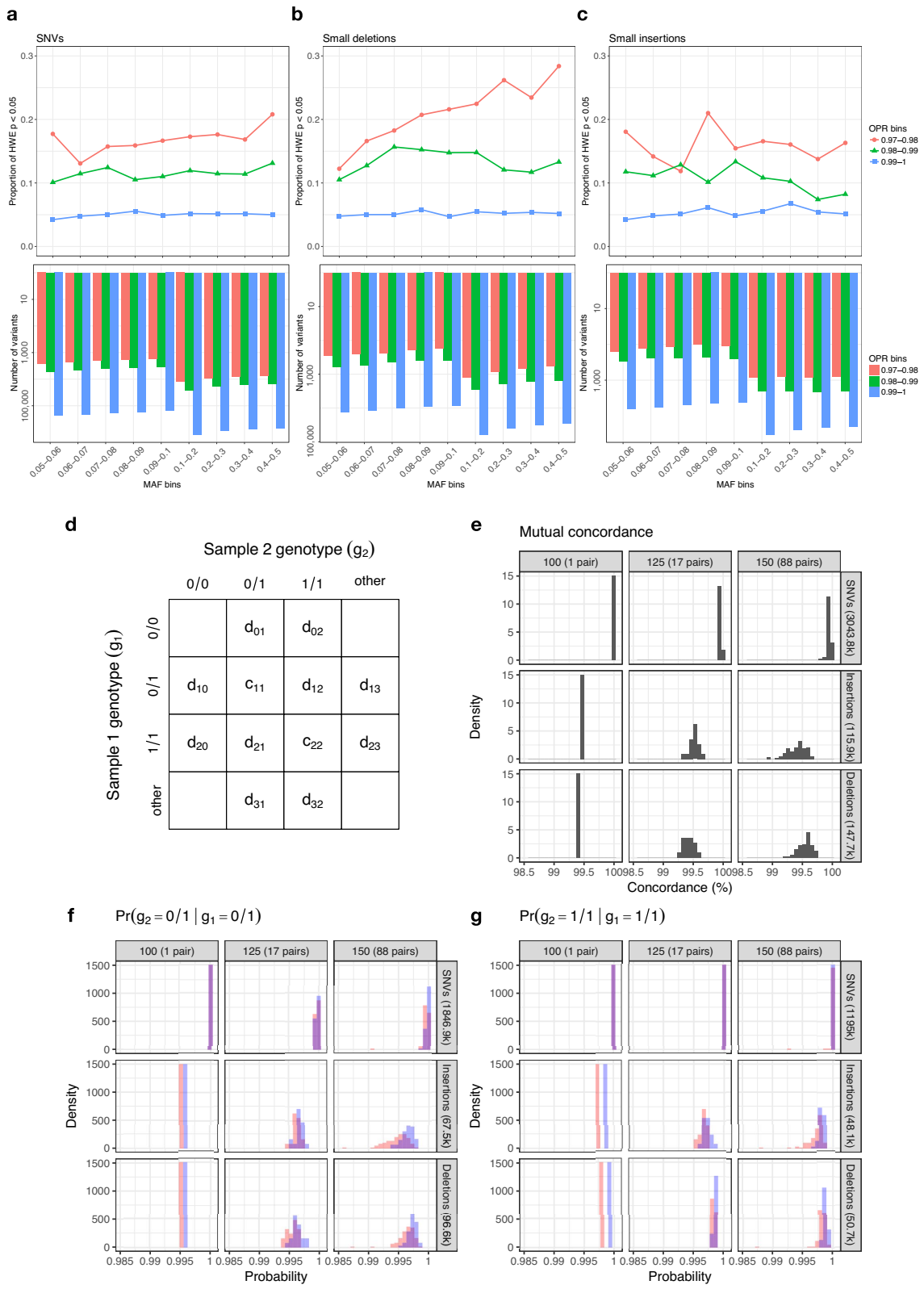


Extended Data Fig. 2 | Flowchart of the bioinformatic data processing. Flowchart describing the processing of samples and variants. Beginning at the top left, all samples were checked for data quality (Extended Data Fig. 3). Quick kinship and sex checks were regularly performed to ensure consistency with reported sex and family information. Samples that failed quality control, 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 normalization, variant calls were loaded into HBase and merged, and summary statistics were calculated, stratified by technical factors (100, 125 and 150 bp) and ancestry (for example, African) (green boxes). Variant-specific minimum overall pass rates were calculated and used to filter inaccurately genotyped variants (Extended Data Fig. 4). Finally, variants were annotated in HBase with predicted consequence information and information from external databases, including allele frequencies (AF) (for example, gnomAD) (blue box).



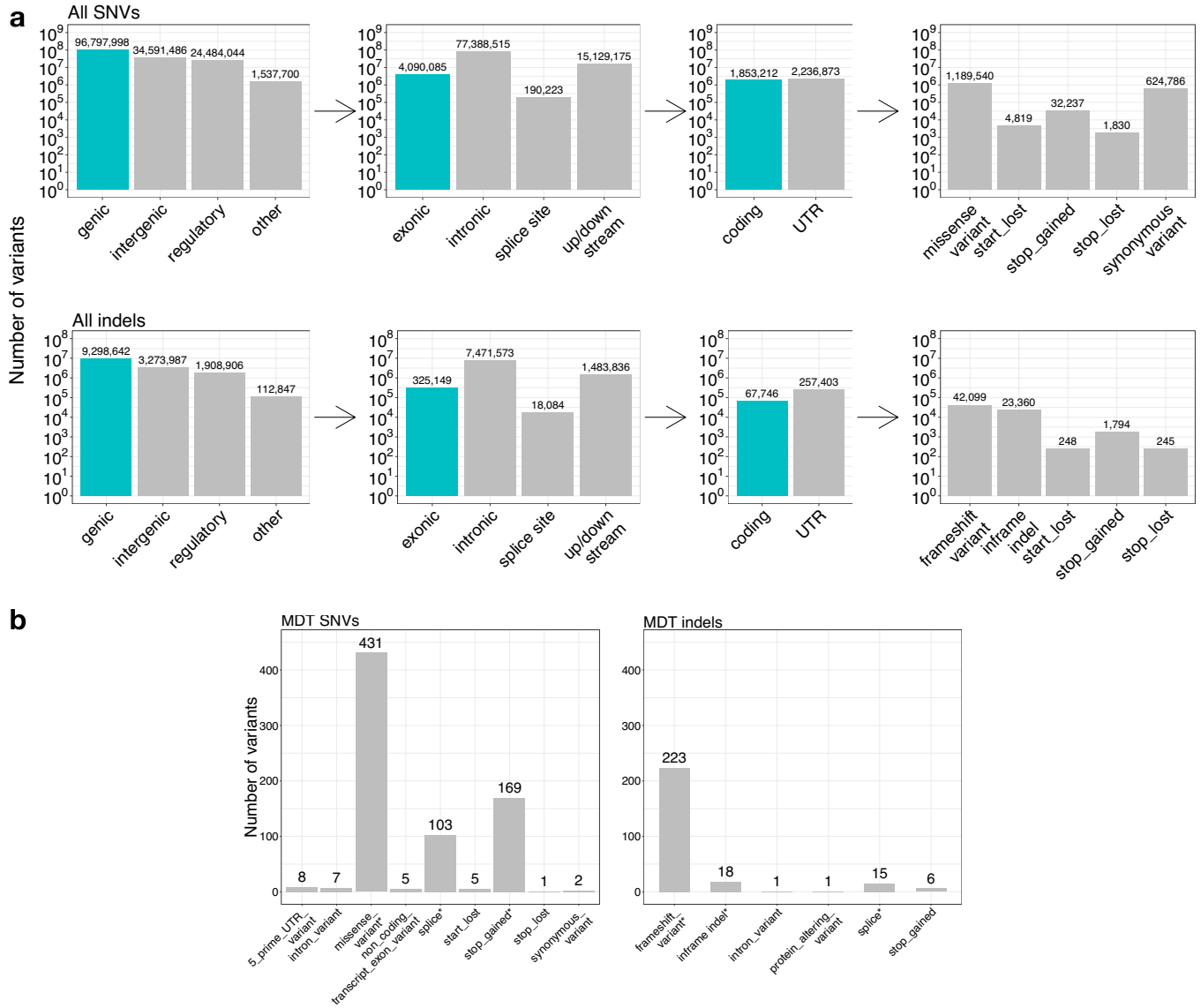
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Sample quality control, sex chromosome karyotyping and ancestry inference. a, The percentage of quality-control-passing autosomal bases (n = 13,187; 4 exclusions highlighted). **b**, The percentage of common SNVs that failed quality control (n = 13,187; 2 exclusions highlighted). **c**, Batch-specific box plots of Ts/Tv ratios (n = 377 for 100-bp samples; n = 3,154 for 125-bp samples; n = 9,656 for 150-bp samples; 3 exclusions highlighted). **d**, FREEMIX values representing sample contamination (n = 13,187; 8 exclusions highlighted). **a**–**d**, Excluded samples are marked in red and labelled with an integer. Three samples were excluded because they failed more than one of the four quality control checks (samples 5, 12 and 14). The centre line of each box plot 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× the interquartile range from each hinge. **e**, The number of heterozygous variants divided by the number of homozygous and hemizygous variants coloured by the initial predicted sexes for 13,037 samples. **f**, Scatter plot of ratios of X/Auto and Y/Auto coloured by the initial sex calls and showing the five sex karyotyping gates. **g**, Scatter plot 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 a 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 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 circles, coloured by assigned population. **j**, 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 (**k**), non-coding regions (**l**) 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 dataset download (**m**). In coding regions, natural selection against frameshift variants results in a systematic depletion of indel sizes that are not a multiple of 3 bp. In non-coding regions, there is a slight excess of indel sizes that are a multiple of 2 bp, but this pattern is almost indiscernible if repetitive regions are excluded.

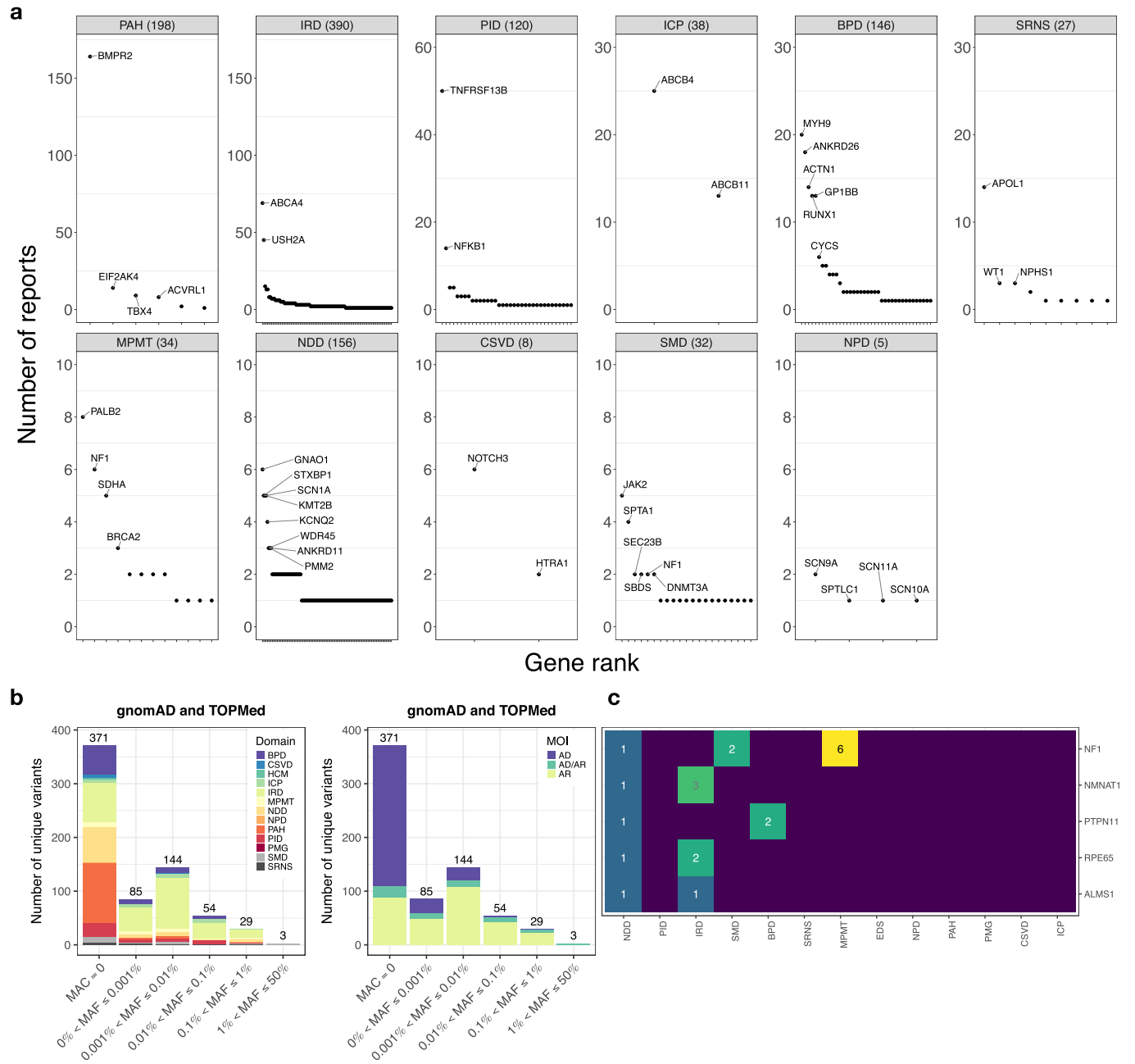


Extended Data Fig. 4 | See next page for caption.

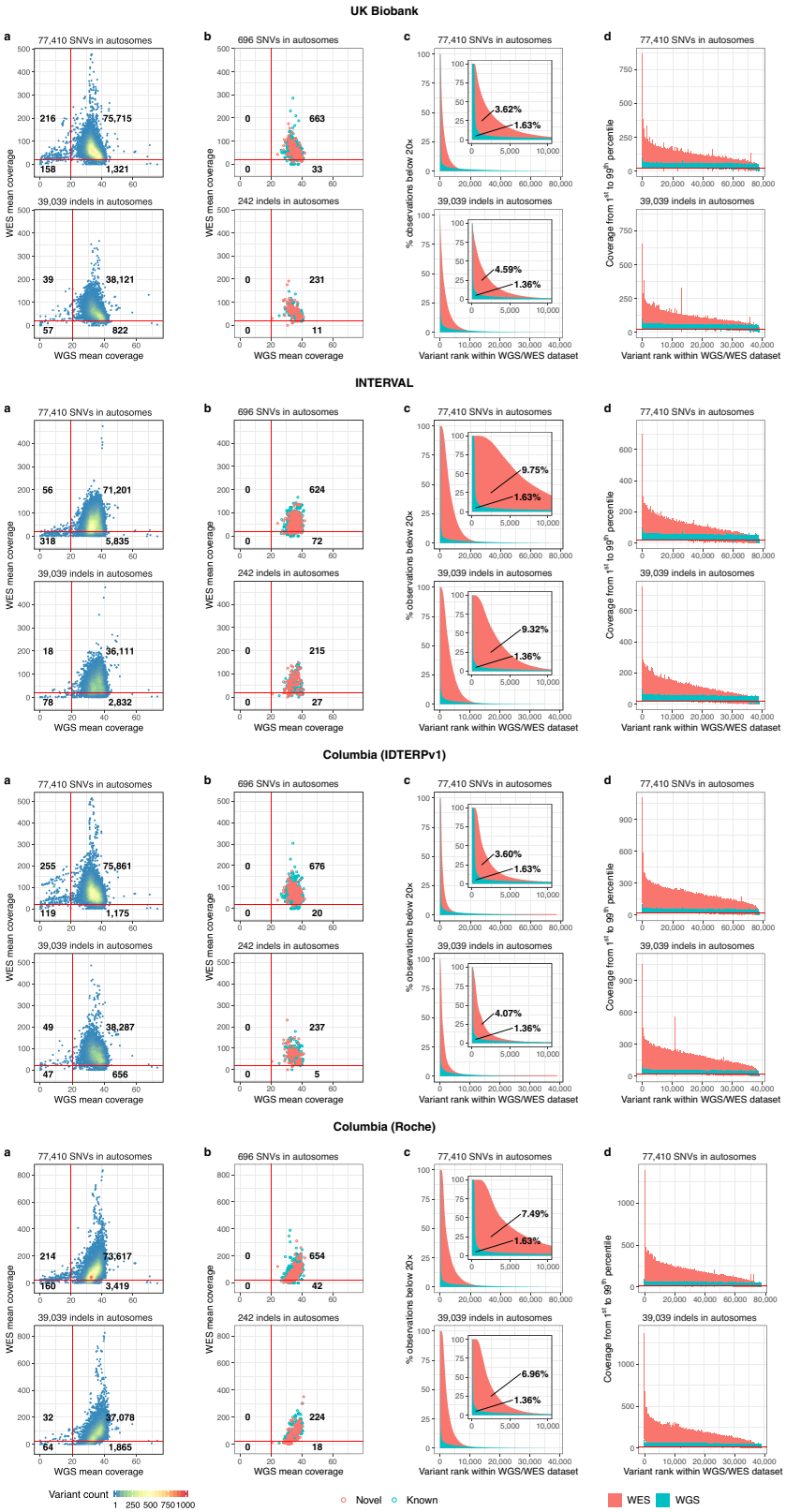
Extended Data Fig. 4 | Variant quality control. a–**c**, The proportion of *P* values computed to test the null hypothesis of Hardy–Weinberg equilibrium < 0.05 among 8,510 unrelated Europeans across different allele frequency (AF) bins for SNVs (**a**), small deletions (**b**) and small insertions (**c**). The number of variants in each overall pass rate (OPR) and allele frequency 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 150-bp 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 or twin has this heterozygous genotype. **g**, Probability of having a non-reference homozygous genotype in a sample, given its duplicate or twin has this homozygous genotype. In **e**–**g**, the mean number of variants of each type used to compute concordance is shown in brackets after the variant type label. In **f**, **g**, red and blue colours represent the 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 or 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 turquoise 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: that is, 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. An asterisk denotes a supercategory 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'.

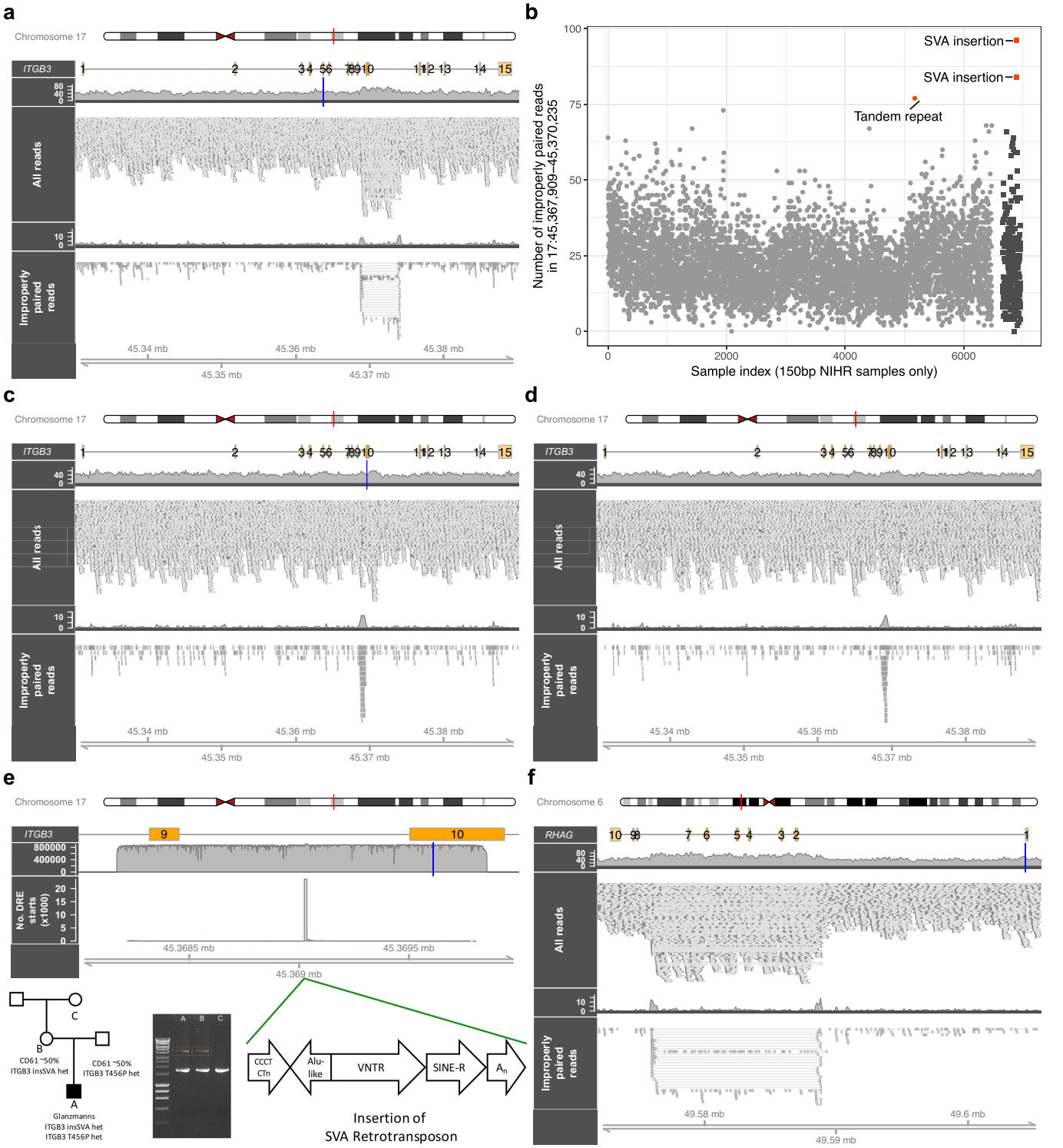


Extended Data Fig. 6 | Breakdown of diagnostic reports by domain. **a**, Number of reports issued for the 11 rare disease domains that issued clinical reports. Each panel corresponds to a domain, the title denotes the domain acronym and number of reports issued. PMG and EDS domains are not shown because 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**, The number of distinct reported autosomal short variants (SNVs and indels) for each domain in different gnomAD/TOPMed allele frequency bins in samples of European ancestry, broken down by rare disease domain (left) and by mode of inheritance (right). 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 heat map shows the number of reports featuring these genes, broken down by domain.

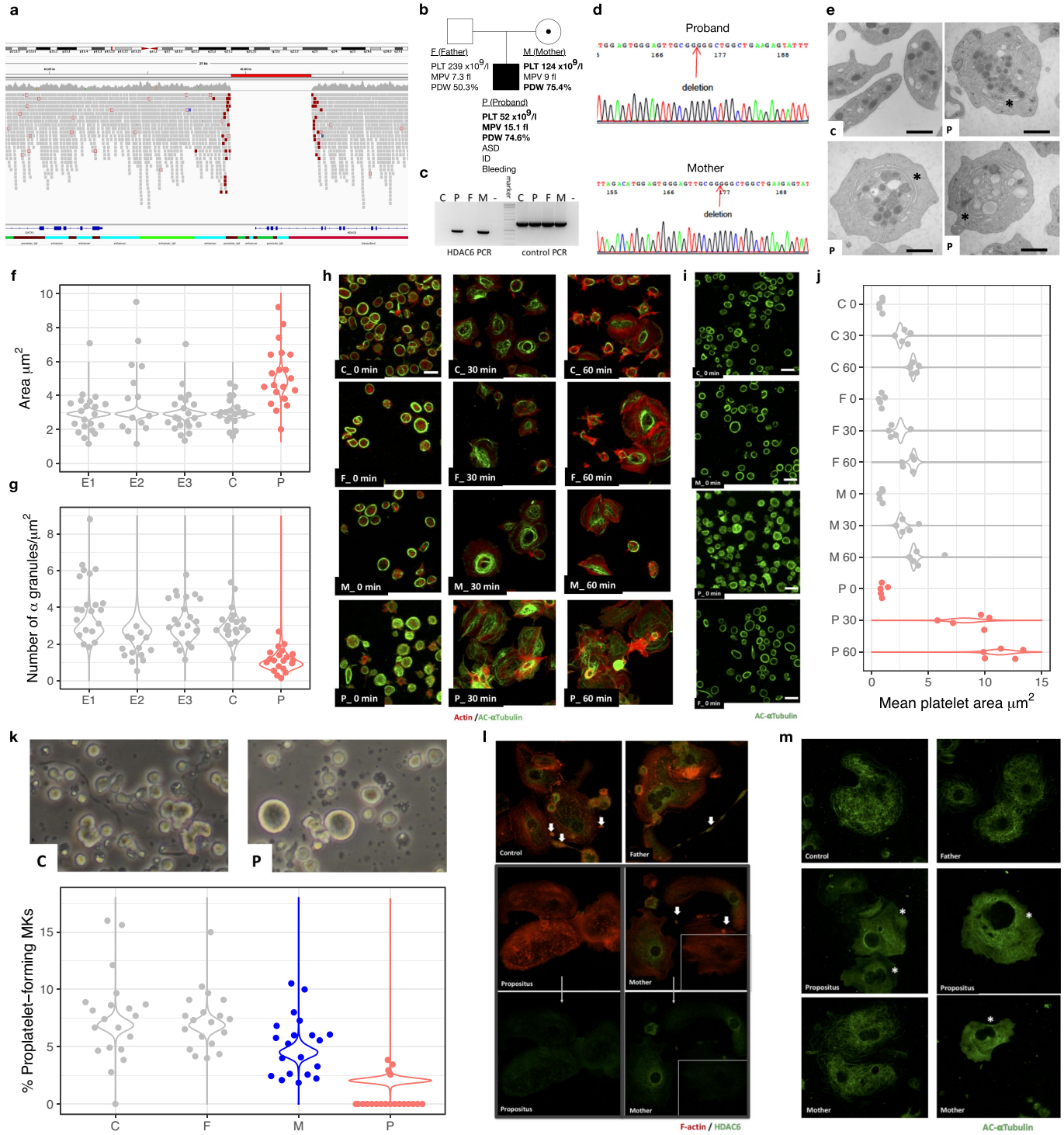


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Comparison of WGS and WES for genetic testing. **a**–**d**, For each of four WES datasets—'UK Biobank', 'INTERVAL', 'Columbia (IDTERPv1)' and 'Columbia (Roche)'—four groups of panels are shown, each of which corresponds to a different comparison of coverage characteristics, as follows. **a**, WGS versus 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**, WGS versus WES coverage of the MDT-reported known (turquoise) and novel (salmon) SNVs and indels in autosomal diagnostic-grade genes. **c**, 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 bar plots corresponding to WGS are superimposed on those corresponding to WES. The inset shows the mean percentage of individuals covered below 20× by WGS and WES in a magnified 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 data sets.

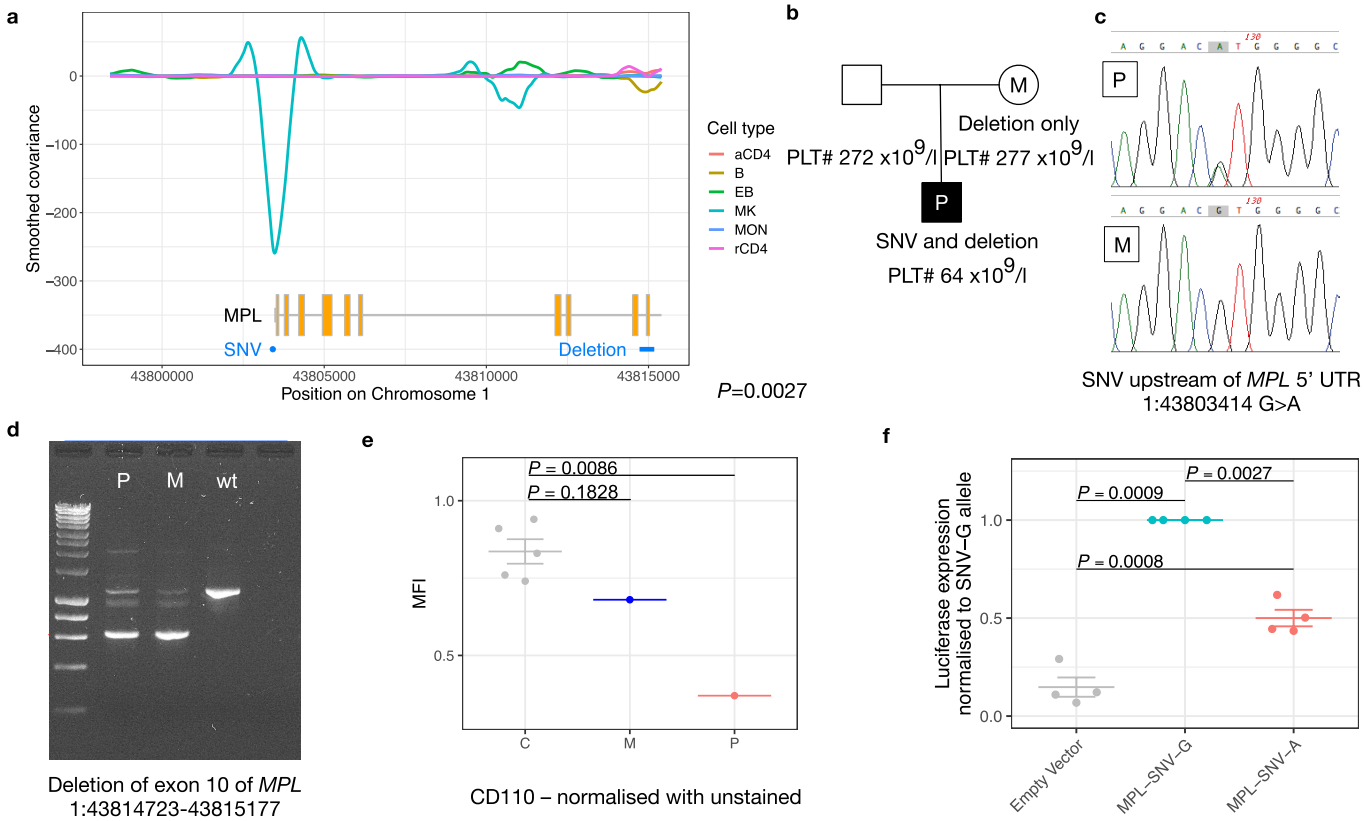


Extended Data Fig. 8 | Cases with protein-null phenotypes. a, Alignments in the *ITGB3* locus for an individual with Glanzmann's thrombasthenia with a premature stop (blue bar) and a tandem repeat revealed by improperly mapped read pairs. **b**, Number of improperly mapped read pairs in the ninth intron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



Extended Data Fig. 9 | See next page for caption.

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 4,108-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 **c**) with primers flanking the 4,801-bp deletion. The red arrow points to the position of the fusion between base pair 48,659,245 and base pair 48,663,353. **e**, Electron microscopy images ($n = 1$ sample preparation per subject) show that platelets of the proband were larger and rounder than those of the control (unrelated healthy control), and in some instances had abnormal semi-circular empty vacuoles (marked by an asterisk) and a depletion of alpha granules. Scale bars, 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 a previous study[61]. Dot plots of platelet area (μm^2) and the alpha granule count per unit area (μm^{-2}), computed using ImageJ. The underlying violin plots show posterior predictive densities for the mean platelet area or granule density in controls and in the proband under a mixed model accounting for intra-individual correlation. The 90% credible intervals for the ratio of the mean in the proband 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* deficiency[61]. **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 min for control, father, mother and proband. This experiment was performed once and representative images are shown. Scale bars, 1.5 µm. **i**, Platelet analysis using 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; whereas those of the father and mother are normal. This experiment was performed once. Scale bars, 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 min after spreading on fibrinogen) for the control, father, mother and proband. There was no evidence of a difference between the mean of the mean platelet area of either the father or the mother and the control within time points ($P > 0.12$ for all six two-sided Welch *t*-tests), so the father and mother 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 min under a mixed model accounting for intra-individual correlation. The 90% credible intervals for the ratio of the mean in the proband to the mean in controls were 1.87–4.56 and 2.07–3.61 at time points 30 and 60 min, respectively. **k**, Top, representative images from the control and the proband. In the latter, large megakaryocytes are present but proplatelet formation is strongly reduced. Bottom, the quantification of proplatelet formation by megakaryocytes at day 12 of differentiation from cultures performed in duplicate for each individual. Ten images per culture were used to compute the percentage proplatelet-forming megakaryocytes per individual, shown as dot plots. There was no evidence of a difference in the mean of the percentage between the father and the control ($P = 0.90$, two-sided Welch *t*-test), so the father was treated as a control in subsequent modelling. The underlying violin plots show posterior predictive densities for the percentage proplatelet-forming megakaryocytes in controls, in the mother and the proband under a mixed model accounting for intra-individual correlation. The 90% credible intervals for the odds ratio of the mean in the mother and the proband to the mean in controls were 0.32–0.46 and 0.18–0.28, respectively. **l**, Day-12 differentiated megakaryocytes for the indicated individuals were stained for F-actin (red) and HDAC6 (green). Top, HDAC6 is expressed in the cytosol and is trafficked to proplatelets as shown in megakaryocytes from the control and the father (bold arrows). Middle, megakaryocytes from the proband show no HDAC6 expression while cultures from the mother contain a mixture of megakaryocytes that are positive and negative (15 of the 45 megakaryocytes) for HDAC6 expression. Bottom, only the HDAC6 staining for the proband and mother. This experiment was performed once. **m**, Day-12 differentiated megakaryocytes for the indicated individuals were stained for acetylated α-tubulin (green). Highly organized tubulin structures are present in all megakaryocytes from the control and father while the patient (47 of the 57 megakaryocytes) and mother (16 of the 46 megakaryocytes) contain megakaryocytes that show signs of tubulin depolymerization (as indicated by an asterisk). 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 in Fig. 4a) and coverage tracks generated by RedPop for activated CD4$^+$ T cells (aCD4), B cells, erythroblasts, megakaryocytes, monocytes 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 with thrombocytopenia owing to a 454-bp deletion encompassing exon 10 of *MPL*, which was inherited from the mother, and an SNV just upstream of the 5′ untranslated region of *MPL*. **c**, Sanger sequencing traces confirming the presence of the heterozygous SNV in the proband and its absence in the mother. **d**, Gel electrophoresis of PCR amplicons covering the deletion confirming presence of the deletion in the proband and the mother. The PCR was conducted on two independent samples in the proband and once in the mother and the control (wt). **e**, MFI on the *y*-axis obtained by the flow cytometry measurement of MPL abundance (CD110) on the membrane of platelets from five unrelated healthy controls, the mother and the proband. The MFI was normalized to unstained platelets. We fitted a linear regression model with an intercept term representing the mean in the control, a coefficient representing the difference in means between the mother and control ($P = 0.1828$) and a coefficient representing the difference in means between the proband and control ($P = 0.0086$). Distribution summaries show mean ± s.e.m. where multiple observations are available. **f**, Results of luciferase reporter assays in K562 cells expressing 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.

nature research

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

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

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

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

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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

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

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

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

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

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used

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

Validation

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

Human research participants

Policy information about studies involving human research participants

Population characteristics

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

Recruitment

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

Ethics oversight

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

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

ChIP-seq

Data deposition

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

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

Data access links
May remain private before publication.

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

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

H3K27ac ChIP-seq:

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

Files in database submission: n/a

Genome browser session (e.g. UCSC): n/a

Methodology

Replicates: As in publications (MK - PMID:25258084 and PMID:21571218; EB - PMID:25521328; aCD4 and rCD4 - PMID:28870212) or the ENCODE website (MONO - https://www.encodeproject.org/experiments/ENCSR000ATN/; B - https://www.encodeproject.org/experiments/ENCSR000AUV).

Sequencing depth: As in publications (MK - PMID:25258084 and PMID:21571218; EB - PMID:25521328; aCD4 and rCD4 - PMID:28870212) or the ENCODE website (MONO - https://www.encodeproject.org/experiments/ENCSR000ATN/; B - https://www.encodeproject.org/experiments/ENCSR000AUV).

Antibodies: As in publications (MK - PMID:25258084 and PMID:21571218; EB - PMID:25521328; aCD4 and rCD4 - PMID:28870212) or the ENCODE website (MONO - https://www.encodeproject.org/experiments/ENCSR000ATN/; B - https://www.encodeproject.org/experiments/ENCSR000AUV).

Peak calling parameters: TFs and H3K27ac peaks were called with MACS2, significance threshold was set to qvalue<1e-5, narrow option was used.

Data quality: Low quality reads (-q 15), multi-mapped and duplicate reads were marked and removed with samtools and picard respectively. ChIP-seq efficiency was assessed with deepTools fingerPrint.

Software: BWA (v.0.7); picard (v.1 to v.2); deepTools plotFingerprint (v.2.3.5); MACS2 (v.2.1.1)

Flow Cytometry

Plots

Confirm that:

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

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

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

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

Methodology

Sample preparation: For MPL expression on platelets: the level of MPL protein on the platelet membrane was measured by flow cytometry (Beckman Coulter FC500) using the monoclonal antibodies: APC-labelled IgG1 against CD42b (clone HIP1, BD Pharmingen, cat. no 551061), PE-labelled IgG1 against CD110 (clone REA250, Miltenyi Biotec, cat. no 130-101-648) and a PE-labelled isotype control (clone MOPC-21, BD Pharmingen, cat. no 555749). In short, a sample of EDTA anticoagulated blood was incubated with anti-CD110 (or control) and anti-CD42b for 30 minutes.

Instrument: Beckman Coulter FC500

Software: Kaluza Analysis Software from Beckman (Version 2.1)

Cell population abundance: n/a

Gating strategy: Platelets were gated based on size using forward scatter and size scatter. The median fluorescent intensity of the CD110 PE-antibody was calculated for all CD42b positive platelets.

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