1. Extended Data

Figure or Table # Please group Extended Data items by type, in sequential order. Total number of items (Figs. + Tables) must not exceed 10.	Figure/Table title One sentence only	Filename Whole original file name including extension. i.e.: Smith_ED_Fig1.jpg	Figure/Table Legend If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.
Extended Data Fig. 1	Reduction in the number of genotypes stored per sample.	ext-data-fig1.eps	For 100 randomly chosen 100KGP participants belonging to each ancestry group (taken from amongst those with an inferred probability >0.9 of belonging): **a,** boxplots showing the distribution of the number of non-homozygous reference PASSing genotypes for variants on chromosomes 1–22 and X which meet the default Rareservoir MAF filtering criteria (i.e. a PMAF score >0 using gnomAD v3.0 and internal MAF <0.002); **b,** boxplots showing the distribution of the proportion of all PASSing non-homozygous reference genotypes that meet the default Rareservoir MAF filtering criteria. In both plots, the lower, centre and upper lines respectively indicate the lower quartile, median and upper quartile. Whiskers are drawn up to the most extreme points that are less than 1.5× the interquartile range away from the nearest quartile.
Extended Data Fig. 2	General schematic of the database build procedure and contents.	ext-data-fig2.eps	Variants are extracted from VCF files, filtered on internal cohort allele frequency, encoded as 64-bit RSVR IDs and loaded into a table containing the corresponding genotypes. The variants are annotated with scores reflecting their predicted deleteriousness (in this case, CADD scores) and probabilistic minor allele frequency scores (PMAF) from gnomAD. The consequences of each variant with respect to a reference set of transcripts are generated and loaded into a table. Sample information including pedigree membership

			and membership of an MSUP is loaded into a table. The case groupings for case/control association analyses are stored in a table.
Extended Data Fig. 3	Detailed schematic of the database build procedure.	ext-data-fig3.eps	Variants may be imported to a Rareservoir from either single gVCF files or a merged VCF file, following the procedures indicated by red and blue arrows respectively.
Extended Data Fig. 4	Schematic showing the variant data in the 100KGP Main Programme Rareservoir.	ext-data-fig4.eps	The number of variant/transcript pairs, the distribution of CADD scores and a breakdown of gnomAD frequency classes is shown for each annotated SO term in the context of the structure of the ontology.
Extended Data Fig. 5	The 269 case sets, Disease Groups A–I.	ext-data-fig5.eps	The names and sizes of the case sets used for the genetic association analyses, grouped by Disease Group and coloured by type (Disease Sub Group or Specific Disease). Disease Sub Groups with only one Specific Disease were excluded to avoid repeating identical analyses. Case sets smaller than 5 are labelled '<5' and shown as having size 4 to comply with 100KGP policy on limiting participant identifiability. For legibility, only Disease Groups starting with the letters A–I are shown here.
Extended Data Fig. 6	The 269 case sets, Disease Groups M–Z.	ext-data-fig6.eps	An extension of **Extended Data Fig. 5** showing the case sets in Disease Groups starting with the letters M–Z.
Extended Data Fig. 7	Breakdown of cases attributable to associations with 'Posterior segment abnormalities' by Specific Disease.	ext-data-fig7.eps	For each gene associated with the Disease Sub Group 'Posterior segment abnormalities', a bar plot showing the number of cases having each of the different Specific Diseases who have an inferred pathogenic configuration of alleles in the gene. This example illustrates that sets of cases with the same aetiological gene may be assigned different Specific Diseases. Consequently, pooling cases within Disease Sub Group can boost power.
Extended Data Fig. 8	Microscopy images of	ext-data-fig8.eps	Exemplar immunofluorescence microscopy images of HEK293 cells overexpressing wild

	HEK293 cells overexpressing ERG.		type ERG (from 20 replicates) and each of the p.S182Afs*22, p.T224Rfs*15 and p.A447Cfs*19 variants of ERG (each from 17 replicates). Cells were stained for ERG (green) and nuclear marker DAPI (blue). Scale bar, 20μm.
Extended Data Fig. 9	Illustrative audiograms for GPR156 cases.	ext-data-fig9.eps	Air and bone conduction audiograms for the two affected daughters of the family with compound heterozygous *GPR156* truncating alleles.

2. Supplementary Information:

A. PDF Files

Item	Present?	Filename Whole original file name including extension. i.e.: Smith_SI.pdf. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	No		
Reporting Summary	Yes.	nr-reporting-summary-comments-addressed.pdf	
Peer Review Information	No	*OFFICE USE ONLY*	

3. Source Data

Parent Figure or Table	Filename Whole original file name including extension. i.e.: *Smith_SourceData_Fig1.xls,* or *Smith_ Unmodified_Gels_Fig1.pdf*	Data description i.e.: Unprocessed western Blots and/or gels, Statistical Source Data, etc.
Source Data Fig. 1	source-data-fig1b.xlsx	**Sheet 1,** Table of associations shown in **Fig. 1b** annotated with BeviMed PPAs (PPA), the level of the case set in the disease label hierarchy (Level), the inferred variant class and MOI for the association, the matched

		PanelApp panel for the association, the method that was used to find the match (Match method, either 'Automatic' or 'Manual'), the associated evidence level for the match, and the notes on the consistency between the MOI listed by PanelApp for the association and the inferred MOI (MOI match comment). **Sheet 2,** Table of variants having a probability of pathogenicity >0.8 conditional on the modal model and forming a pathogenic configuration of alleles in at least one case. While these variants contributed to the reported statistical associations, they have not been individually scrutinised according to ACMG guidelines.
Source Data Fig. 2	source-data-fig2e.jpg	Uncropped western blot images corresponding to **Fig. 2e**.
Source Data Fig. 3	source-data-fig4e.jpg	Uncropped western blot images corresponding to **Fig. 4e**.

Genetic association analysis of 77,539 genomes reveals rare disease etiologies

Daniel Greene[1,2]**, Genomics England Research Consortium, Daniela Pirri**[3]**, Karen Frudd**[3,4]**, Ege Sackey**[5]**, Mohammed Al-Owain**[6]**, Arnaud P.J. Giese**[7]**, Khushnooda Ramzan**[8]**, Sehar Riaz**[7,9]**, Itaru Yamanaka**[10]**, Nele Boeckx**[11]**, Chantal Thys**[12]**, Bruce D. Gelb**[2,13,14]**, Paul Brennan**[15]**, Verity Hartill**[16,17]**, Julie Harvengt**[18]**, Tomoki Kosho**[19,20]**, Sahar Mansour**[5,21]**, Mitsuo Masuno**[22]**, Takako Ohata**[23]**, Helen Stewart**[24]**, Khalid Taibah**[25]**, Claire L.S. Turner**[26]**, Faiqa Imtiaz**[8]**, Saima Riazuddin**[7,9]**, Takayuki Morisaki**[10,27]**, Pia Ostergaard**[5]**, Bart L. Loeys**[11,28]**, Hiroko Morisaki**[10,29]**, Zubair M. Ahmed**[7,9]**, Graeme M. Birdsey**[3]**, Kathleen Freson**[12]**, Andrew Mumford**[30,31] **& Ernest Turro**[2,14,32,33*]

*Corresponding author. Contact: ernest.turro@mssm.edu

[1]Department of Medicine, University of Cambridge, Cambridge, UK. [2]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]National Heart and Lung Institute, Imperial College London, London, UK. [4]UCL Institute of Ophthalmology, University College London, London, UK. [5]Molecular and Clinical Sciences Institute, St George's University of London, London, UK. [6]Department of Medical Genomics, Centre for Genomic Medicine, King Faisal Specialist Hospital & Research Centre, Riyadh, Saudi Arabia. [7]Department of Otorhinolaryngology Head and Neck Surgery, School of Medicine, University of Maryland Baltimore, Baltimore, MD, USA. [8]Department of Clinical Genomics, Centre for Genomic Medicine, King Faisal Specialist Hospital & Research Centre, Riyadh, Saudi Arabia. [9]Department of Biochemistry and Molecular Biology, School of Medicine, University of Maryland, Baltimore, MD, USA. [10]Department of Bioscience and Genetics, National Cerebral and Cardiovascular Center, Osaka, Japan. [11]Center for Medical Genetics, Antwerp University Hospital/University of Antwerp, Antwerp, Belgium. [12]Department of Cardiovascular Sciences, Center for Molecular and Vascular Biology, KU Leuven, Leuven, Belgium. [13]Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY,

USA. [14]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [15]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Trust, International Centre for Life, Newcastle upon Tyne, UK. [16]Department of Clinical Genetics, Chapel Allerton Hospital, Leeds Teaching Hospitals NHS Trust, Leeds, UK. [17]Leeds Institute of Medical Research, University of Leeds, Leeds, UK. [18]Center for Medical Genetics, CHU Liege, Belgium. [19]Department of Medical Genetics, Shinshu University School of Medicine, Matsumoto, Japan. [20]Center for Medical Genetics, Shinshu University Hospital. [21]SW Thames Regional Genetics Service, St George's University Hospitals NHS Foundation Trust, London, UK. [22]Department of Medical Genetics, Kawasaki Medical School Hospital, Okayama, Japan. [23]Okinawa Chubu Hospital, Okinawa, Japan. [24]Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [25]ENT Medical Centre, Riyadh, Saudi Arabia. [26]Peninsula Clinical Genetics Service, Royal Devon & Exeter Hospital, Exeter, UK. [27]Division of Molecular Pathology and Department of Internal Medicine, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [28]Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands. [29]Department of Medical Genetics, Sakakibara Heart Institute, Tokyo, Japan. [30]School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK. [31]South West NHS Genomic Medicine Service Alliance, Bristol, UK. [32]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [33]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

Abstract

The genetic aetiologies of more than half of rare diseases remain unknown. Standardised genome sequencing (GS) and phenotyping of large patient cohorts provides an opportunity for discovering the unknown aetiologies, but this depends on efficient and powerful analytical methods. We built a compact database, the 'Rareservoir,' containing the rare variant genotypes and phenotypes of 77,539 participants sequenced by the 100,000 Genomes Project (100KGP). We then used the Bayesian genetic association method BeviMed to infer associations between genes and each of 269 rare disease classes assigned by clinicians to the participants. We identified 241 known and 19 previously unidentified associations. We validated associations with *ERG*, *PMEPA1* and *GPR156* by searching for pedigrees in other cohorts and using bioinformatic and experimental approaches. We provide evidence that (1) loss-of-function variants in the ETS-family transcription factor encoding gene *ERG* lead to primary lymphoedema, (2) truncating variants in the last exon of TGFβ regulator *PMEPA1* result in Loeys-Dietz syndrome, and (3) loss-of-function variants in *GPR156* give rise to recessive congenital hearing impairment. The Rareservoir provides a lightweight, flexible and portable system for synthesising the genetic and phenotypic data required to study rare disease cohorts with tens of thousands of participants.

Introduction

Collectively, rare diseases affect 1 in 20 people[1], but fewer than half of the approximately 10,000 catalogued rare diseases have a resolved genetic aetiology[2]. Standardised GS of large, phenotypically diverse collections of rare disease patients enables aetiological discovery across a wide range of pathologies[3,4,5] while boosting genetic diagnostic rates for patients. The 100KGP, the largest GS study of rare disease patients to date, sequenced 34,523 United Kingdom National Health Service patients with rare diseases and 43,016 of their unaffected relatives. The linked genetic and phenotypic data of 100KGP participants were then made available to researchers through a web portal called the Genomics England Research Environment. The scale and complexity of such large GS datasets and the hierarchical nature of patient phenotype coding[6] induce numerous bioinformatic and statistical challenges. Most importantly, the full genotype data from GS studies of tens of thousands of individuals are typically stored in unmodifiable files many terabytes in size, leading to high storage and processing costs. Recently developed frameworks such as Hail[7] and OpenCGA[8] afford greater flexibility. However, they are designed to capture genotypes for variants across the full minor allele frequency (MAF)

spectrum, from rare (MAF<0.1%) to common (MAF>5%) variants. To accommodate large numbers of genotypes, they depend on distributed storage systems and require numerous software packages, hindering deployment. We developed a database schema, the 'Rareservoir,' for working with rare variant genotypes and patient phenotypes flexibly and efficiently. We deployed a Rareservoir only 5.5GB in size of the 100KGP data and applied the Bayesian statistical method BeviMed[9] to identify genetic associations between coding genes and each of the 269 rare disease classes assigned to patients by clinicians. Out of the previously unknown associations that we identified, we followed up the most plausible subset in confirmatory analytical and experimental work.

Results

The 'Rareservoir'

Relational databases (RDBs) provide a unified, centralised structure for storing, querying and modifying data of multiple underlying types. In principle, an RDB could provide a convenient foundation for the analysis of genotypes, variants, genes, participants and statistical results, but they cannot accommodate tables of the scale required to store exome or genome-wide genotypes in a moderately sized cohort. An RDB can, however, accommodate a sparse representation of genotypes corresponding to rare variants only, which encompass almost all variants having a large effect on rare disease risk. We developed an RDB schema, the 'Rareservoir', and build procedure for the analysis of rare diseases, which, by default, stores genotypes corresponding to variants for which all population-specific MAFs are likely to be <0.1%. This reduces the number of stored genotypes in a large study by about 99% (**Extended Data Fig. 1**). The Rareservoir encodes variants as 64-bit integers ('RSVR IDs', **Extended Data Fig. 2**), which can represent 99.3% of variants encountered in practice without loss of information. RSVR IDs occupy a single column and increase numerically with respect to genomic position, allowing fast location-based queries within a simple database structure. To support the build process of a Rareservoir, we developed a complementary software package called 'rsvr' (**Extended Data Figs. 2–3**). The package includes tools to annotate variants with MAF information from control databases (e.g. gnomAD[10]), pathogenicity scores (e.g. combined annotation dependent depletion (CADD) scores[11]) and predicted Sequence Ontology (SO)[12] consequences with respect to a set of transcripts. We use a 64-bit integer ('CSQ ID') to record the consequences for interacting variant/transcript pairs, where each bit encodes one of the possible consequences, ordered by severity. Encoding the consequences in this way is efficient and enables succinct queries that threshold or sort based on severity of impact. The Rareservoir also contains a table with genetically derived data for each sample (including ancestry, sex and membership of a maximal set of unrelated participants (MSUP)), and a table of 'case sets' storing the rare disease classes assigned to each participant.

BeviMed infers 241 known and 19 unknown genetic associations

We built a Rareservoir, 5.5GB in size, containing 11.9 million rare exonic and splicing single nucleotide variants (SNVs) and short insertions or deletions (indels) affecting canonical transcripts of protein-coding genes in Ensembl v104[13] from a merged variant call format file (VCF) containing genotype calls for 77,539 participants, including 29,741 probands, in the Rare Diseases Main Programme of the 100KGP (Data Release Version 13) (**Extended Data Fig. 4**). During enrolment to the 100KGP, expert clinicians used the clinical characteristics of each affected participant to assign them to one or more of 220 *Specific Diseases*. The Specific Diseases are hierarchically arranged into 88 *Disease Sub Groups*, each of which belongs to one of 20 *Disease Groups*. Whereas the eligibility criteria for many specific diseases aligned

to the same or closely related rare diseases, for others such as ‘intellectual disability,’ the criteria were broader and encompassed diverse genetic aetiologies. We generated 269 analytical case sets corresponding to all distinct Specific Diseases and Disease Sub Groups, ranging in size from 5,809 to one proband, and stored them in the Rareservoir (**Fig. 1a, Extended Data Figs. 5–6**). We included these two levels of the phenotyping hierarchy to account for heterogeneity in presentation or diagnosis among cases sharing the same genetic aetiology, with the aim of boosting power to identify statistical genetic associations.

Using the Bayesian statistical method BeviMed[9], we obtained a posterior probability of association (PPA) between each of the 19,663 protein-coding genes and each of the 269 rare disease classes. BeviMed computes posterior probabilities over a baseline model of no association and six competing association models, each of which assumes a particular mode of inheritance (MOI, dominant or recessive) and consequence class of aetiological variant (high-impact, moderate-impact or 5' UTR). The PPA is obtained by summing the posterior probabilities over all six association models. The association model with the greatest posterior probability (the modal model) determines the inferred MOI and class of aetiological variant. Conditional on an association model, BeviMed models the pathogenicity of each included rare variant. In the model, participants with at least one pathogenic allele (under a dominant MOI), or at least as many pathogenic alleles as the ploidy (under a recessive MOI), have a *pathogenic configuration of alleles*, which determines their risk of case status. For each rare disease class, we selected a set of unrelated cases based on pedigree information provided by the 100KGP and compared them to participants not in the case set who belonged to different pedigrees and to an MSUP, also provided by the 100KGP. To account for correlation between case sets, we only recorded the association for each gene having the highest PPA within a given Disease Group. Using a significance threshold of PPA>0.95, we identified 260 significant associations, 241 of which were documented by the PanelApp gene panel database,[14] an expert-curated and annotated resource containing gene lists with high, medium or low levels of prior supporting evidence of causality for rare diseases (**Fig. 1b**). Out of the 241 known associations that we identified, 43 (17.8%) were with Disease Sub Groups. For example, within each of the nine known genes associated with the Disease Sub Group 'Posterior segment abnormalities,' the set of cases explained by variants with a posterior probability of pathogenicity >0.8 comprised a mixture of participants with five different Specific Diseases (**Extended Data Fig. 7**). This demonstrates that participants with different Specific Diseases belonging to the same Disease Sub Group sometimes share defects in the same gene, which confirms that treating Disease Sub Groups, not just Specific Diseases, as case sets, boosts statistical power.

Out of the 241 associations identified as previously known according to PanelApp, 237 (98.3%) had an inferred MOI that was consistent with the MOIs listed for the relevant gene. Of these, the consistent MOI was found in the matched panel (223 associations), in the notes for the matched panel (five associations) or in the MOIs listed for an alternative relevant panel (nine associations) in PanelApp (**Source data for Fig. 1**). This provided independent evidence that the genetic associations we labelled as known (without reference to MOI information) are genuinely supported by evidence in the literature, further demonstrating the accuracy of BeviMed's inference. Of the four known associations with an inferred MOI that was incongruous with PanelApp, two had supporting evidence for the inferred MOI in the literature that was absent from PanelApp: *EDA* with dominant 'Ectodermal dysplasia without a known gene mutation'[15] and *AICDA* with dominant 'Primary immunodeficiency'[16]. The two associations with an MOI

that was unsupported in the literature were between *UCHL1* and dominant 'Inherited optic neuropathies' and between *SLC39A8* and dominant 'Intellectual disability'.

Among 5,253 of the probands included in our analysis, the table of clinically reported variants available from the 100KGP Rare Diseases Main Programme at the time of this study comprised 4,907 distinct variants that had been classified as pathogenic or likely pathogenic in 1,863 genes. For 855 of these genes, aetiological variants had been reported for only one family, suggesting that many genes which are aetiological in the 100KGP are not identifiable by statistical association. Nevertheless, across the 260 associations identified by BeviMed, 2,536 distinct rare variants had a posterior probability of pathogenicity >0.8 conditional on the modal model and were observed as part of a pathogenic configuration of alleles in a case (**Source data for Fig. 1**). Interestingly, among the subset of 2,485 variants contributing to the 241 known associations, only 1,604 featured in the table of clinically reported variants.

We found 19 previously unidentified genetic associations. To select a shortlist for further investigation, we assigned a plausibility score (range 0–3) based on three sources of additional evidence (**Table 1**). Firstly, we considered evidence of purifying selection from gnomAD v2.1.1. Any dominant associations with high-impact variants in a gene having a probability of loss-of-function intolerance (pLI) >0.9 or with moderate-impact variants in a gene having a *Z*-score >2 were considered to be supported by population genetic metrics of purifying selection. To avoid disadvantaging recessive associations, which are unlikely to leave a detectable signature of purifying selection in gnomAD even if genuine, they were considered to be supported by default. Secondly, we considered co-segregation data: any association for which variants having a posterior probability of pathogenicity conditional on the modal model >0.8 tracked with case status in at least three additional family members and for which no affected relatives lacked the pertinent variants were considered to be supported by co-segregation. Thirdly, we performed a comprehensive review of the literature for each gene and made a subjective assessment of whether an association was supported by biological function or previously known disease associations for related genes. In total, three genetic associations had a plausibility score of 3 and were therefore investigated further by gathering additional experimental evidence and looking for replication in other sequenced rare disease collections.

Variants in *ERG* are responsible for primary lymphoedema

BeviMed identified a dominant genetic association between high-impact variants in *ERG* and the Specific Disease 'Primary lymphoedema,' a group of genetic conditions caused by abnormal development of lymphatic vessels or failure of lymphatic function[17,18]. Three such variants were responsible for the high PPA, with locations ranging from codon 182 to 463 on the canonical Ensembl transcript ENST00000288319.12. One of the probands had two unaffected parents without the variant allele—one sequenced by the 100KGP and the other by Sanger sequencing—suggesting the truncating heterozygous variant had appeared *de novo*. A participant in a fourth family who had been enrolled to the 100KGP for an unrelated condition also carried a predicted loss-of-function variant in *ERG*. Upon manual chart review, this participant had features associated with this unrelated condition, but additional features consistent with primary lymphoedema, providing internal replication within the discovery cohort (**Fig. 2a**).

The affected father of the proband with the variant encoding p.S182Afs*22 was called homozygous for the reference allele, initially suggesting a lack of co-segregation of the variant with the disease in that pedigree. However, a review of the GS read alignments for the father revealed that two out of the 48 reads overlapping that position supported the alternative allele. Specifically, these reads contained a deletion of a single G within the central poly-G tract of the motif "AGCTGGGGGTGAG." To assess whether this could be the result of erroneous sequencing, we counted the number of such reads in the 77,539 genomes in the 100KGP and found that the proband and the father were the only two with more than one such read. This indicated that these reads in the father were unlikely to be erroneous but instead that he was mosaic (**Fig. 2b**), consistent with the observation that his lymphoedema became clinically apparent over two decades later than his daughter, indicating milder disease. A further 130 samples collected through the 100KGP had a single read containing the deletion. This number was consistent with observations in the 80 other exonic loci that contain the same 13bp motif (mean: 99.67 samples, range: 4 to 149 samples), suggesting that, rather than being mosaic, the 130 samples contained individual sequencing errors. Furthermore, none of the participants who gave these samples had been assigned the Specific Disease 'Primary lymphoedema.'

ERG encodes a critical transcriptional regulator of blood vessel endothelial cell (EC) gene expression[19] that is essential for normal vascular development[20]. However, little is known about the contribution of ERG to lymphatic development or how primary lymphoedema could arise from loss-of-function *ERG* variants which affect different parts of the ERG protein (**Fig. 2c**). Total cellular expression of ERG detected by real-time quantitative polymerase chain reaction (PCR) in purified RNA and by immunoblotting of protein extracts was the same in primary human dermal lymphatic EC (HDLEC) as human umbilical vein EC (HUVEC) (**Fig. 2d** and **Fig. 2e** respectively). Moreover, immunofluorescence microscopy of cultured HDLEC showed that ERG expression co-localised with the lymphatic EC nuclear marker PROX1 (**Fig. 2f**) a finding confirmed *in vivo* by immunostaining whole mounts of ear skin from mice at three weeks after birth (**Fig. 2g)**. The positions of the p.S182Afs*22 and p.T224Rfs*15 variants suggest nonsense mediated decay and haploinsufficiency as a possible disease mechanism. The other two variants, however, are located in the final exon of ERG and may therefore evade nonsense mediated decay. We studied both types of variant in more detail to explore potential disease mechanisms. In HEK293 cells, which do not express endogenous ERG, overexpression of wild type ERG cDNA recapitulated the nuclear expression pattern observed in the HDLEC and mouse ear skin models. However, overexpression of each of the ERG mutant cDNAs resulted in mislocalisation of ERG outside of the nucleus, in the cytosol (**Fig. 2h–i**, **Extended Data Fig. 8**), preventing it from binding to DNA and exerting its function as a transcription factor[21]. Together, these data confirm high levels of ERG expression within the nuclei of the lymphatic endothelium consistent with a transcription regulatory function during lymphangiogenesis. They also suggest that in the primary lymphoedema cases, defective lymphangeogenesis may result from reduced ERG availability in the nucleus either because of haploinsufficiency resulting from nonsense mediated decay or mislocalisation.

Variants in *PMEPA1* result in Loeys-Dietz syndrome

BeviMed identified a dominant genetic association between high-impact variants in *PMEPA1* and the Specific Disease 'Familial Thoracic Aortic Aneurysm Disease' (FTAAD). The variant with the highest conditional probability of pathogenicity was an insertion of one cytosine within a seven-cytosine stretch in the last exon of the canonical Ensembl transcript ENST00000341744.8. This variant, which is predicted

to induce a p.S209Qfs*3 frameshift, was observed in three FTAAD pedigrees of European ancestry in the 100KGP discovery cohort. We replicated the association in three additional collections of cases. Firstly, the same variant was identified independently in eight affected members of three pedigrees of Japanese ancestry in a separate Japanese patient group. Secondly, a single-cytosine deletion within the same poly-cytosine stretch as the previous variant, and encoding p.S209Afs*61, was found in an FTAAD case enrolled in a separate collection of 2,793 participants in the 100KGP Pilot Programme. Lastly, we identified a family in Belgium wherein the affected members carried a five base-pair deletion in the same stretch of poly-cytosines inducing a frameshift two residues upstream of the other two variants (p.P207Qfs*3).

All pedigrees exhibited dominant inheritance of aortic aneurysm disease with incomplete penetrance and skeletal features including pectus deformity, scoliosis and arachnodactyly with complete penetrance, which co-segregated with the respective variants in genotyped participants (**Fig. 3a**). To assess whether *PMEPA1* families affected by FTAAD form a phenotypically distinct subgroup, we analysed the HPO terms assigned to the 593 FTAAD families in both programmes of the 100KGP. Using a permutation-based method[22,23] based on Resnik's semantic similarity measure[24], we found that the four 100KGP *PMEPA1* families were significantly more similar to each other than to other FTAAD families chosen at random (p=5.7x10^{-3}). To characterise the *PMEPA1* phenotype in greater detail, we compared the prevalence of each of the HPO terms in the minimal set of terms present in at least three of the four families with the prevalence in the other FTAAD families. We identified four HPO terms related to the musculoskeletal system that were significantly enriched (**Fig. 3b**), echoing the phenotypic characteristics of the syndromic aortopathy Loeys-Dietz syndrome[25,26].

To understand the molecular mechanisms underlying this defect, we examined the protein-protein interactions[27] for *PMEPA1* and the complete set of high-confidence genes in the 'Thoracic aortic aneurysm or dissection' PanelApp panel. *PMEPA1* encodes a negative regulator of Transforming Growth Factor β (TGFβ) signalling[28], a pathway previously implicated in multiple aortopathies, including Loeys-Dietz syndrome[29]. The genes underlying known forms of Loeys-Dietz syndrome encode part of a tightly interacting subgroup of proteins in the TGFβ pathway, in which there is a direct interaction between the proteins encoded by *SMAD2*, *SMAD3* and *PMEPA1* (**Fig. 3c**). As the two candidate variants occur in the last exon of the transcript, they are likely to evade nonsense-mediated decay[30]. However, their truncating effects are predicted to remove a PPxY interaction motif, while leaving the SMAD interaction motif intact (**Fig. 3d**), possibly affecting binding between PMEPA1 and SMAD2/3, and altering TGFβ signalling through a gain-of-function mechanism.

Variants in *GPR156* lead to recessive congenital hearing loss

BeviMed identified a recessive genetic association between high-impact variants in *GPR156* and the Specific Disease 'Congenital hearing impairment'. Two high-impact variants in *GPR156* were responsible for the strong evidence of association: a one base pair deletion predicting p.S207Vfs*113 and a one base pair insertion predicting p.P718Lfs*86 with respect to the canonical Ensembl transcript ENST00000464295.6. One family contained two affected siblings who were both homozygous for the p.S207Vfs*113 variant inherited from heterozygous parents. In a second family, there were also two affected siblings, in this case compound heterozygous for the same p.S207Vfs*113 variant that was maternally inherited and a different p.P718Lfs*86 variant that was paternal. Using GeneMatcher[31], we

identified a third pedigree from Saudi Arabia with biallelic truncating variants in *GPR156*. This consanguineous pedigree contained four siblings with hearing impairment, all of whom were homozygous for a variant predicting p.S642Afs*162 (**Fig. 4a**). The eight affected individuals in these three families all had congenital non-syndromic bilateral sensorineural hearing loss (see **Extended Data Fig. 9** for illustrative audiograms).

GPR156 encodes probable G-protein coupled receptor 156, which has sequence homology to the class C GABAB receptors[32]. Although previously designated as an orphan receptor, *GPR156* has recently been identified as a critical regulator of stereocilia orientation on hair cells of the auditory epithelium and other mechanosensory tissues[33]. Its expression is highly restricted to hair cells in the inner ear[34]. Disruption of stereocilia is a common pathogenic mechanism underlying many human Mendelian hearing loss disorders[35] and the over-expression of *GPR156* in hair cells relative to surrounding cells was commensurate with the over-expression of the 21 genes currently implicated in hearing impairment having a Gene Ontology (GO) term relating to stereocilia (**Fig. 4b**). By immunostaining of the Corti and vestibular system from wild type mice, we found that GPR156 strongly co-localises with actin at the apical surface of the outer and inner hair cells of the organ of Corti (**Fig. 4c**).

The p.S207Vfs*113 variant is located in the sixth of 10 exons of *GPR156* and therefore predicts absent expression through nonsense mediated decay of the *GPR156* mRNA. In contrast, the p.S642Afs*162 and p.P718Lfs*86 variants both occur within the final *GPR156* exon and likely result in expression of abnormal GPR156 with an altered amino acid sequence and premature truncation of the cytoplasmic tail (**Fig. 4d**). To determine the effect of the variants on protein expression, we transfected Cos7 cells, which do not express *GPR156* endogenously, with constructs containing cDNAs for wild type *GPR156* or *GPR156* containing each of the three mutant alleles, tagged with a green fluorescent protein (GFP) reporter. While cells transfected with wild type sequence expressed GPR156-GFP fusion protein robustly, cells transfected with the mutant constructs either did not express the protein appreciably or exhibited markedly reduced expression, suggesting that all three of the truncated proteins are degraded (**Fig. 4e**). These data suggest that the biallelic chain truncating variants in *GPR156* cause a congenital hearing loss by preventing expression of GPR156 protein, thereby disrupting stereocilia formation in the auditory epithelium.

Discussion

The standardisation of GS within a healthcare system, together with powerful frameworks for genetic and phenotypic data processing and statistical analysis, promises to advance the resolution of the remaining unknown aetiologies of rare diseases. We have developed a lightweight and easily deployable relational database, the Rareservoir, for genetic analysis of rare diseases using approaches such as BeviMed. In one unified analysis, we identified 260 associations, of which 241 had been published previously in a body of work spanning several decades of genetics research. Our results give an upper bound on the false discovery rate (FDR) of 7.3%. In contrast, a recent analysis of 57,000 samples in the 100KGP reported 249 known and 579 previously unidentified associations[36], giving an upper bound on the FDR of 70%, which suggests that our analytical approach has a greater specificity for a given sensitivity. The associations spanned 86 disease classes across a wide range of organ systems. Interestingly, only 64% of the variants contributing substantially to the known associations were present in the table of clinically reported variants available at the time of this study. This suggests that, as cohorts grow larger, the

results of statistical inference could help guide the clinical reporting process. The case sets we used in our genetic association analysis were based on the formal disease classifications used by the 100KGP. Some of the case sets, such as 'Intellectual disability' (5,529 probands), are particularly large and likely to be highly genetically heterogeneous, potentially limiting the power of our analyses. Careful partitioning of heterogeneous case sets using individual-level HPO terms[6] has the potential to boost power. Of the 19 previously unidentified associations, we shortlisted, replicated and validated three. These three aetiologies involve genes that had not previously been implicated in any of these human diseases. The remaining 16 associations include further plausible hypotheses. For example, *LRRC7*, which we identified to be associated with intellectual disability, encodes a brain-specific protein in post-synaptic densities[37], and Lrrc7-deficient mice exhibit a neuro-behavioural phenotype[38]. *USP33*, which we found to be associated with early-onset hypertension, encodes a deubiquitinating enzyme implicated in regulating expression of the β2-adrenergic receptor regulation[39]. These and other candidates will require replication and validation before they can be considered causative genes.

The present study has several limitations. Firstly, approximately 82% of the participants in the 100KGP are of European ancestry. While this percentage is in line with the proportion of residents in England and Wales reporting their ethnic group as White in the 2011 UK census (86%), its large magnitude constrains power to identify causative variants specific to other ancestry groups. Secondly, of the 260 case sets analysed, 28 contained fewer than 5 probands, limiting power to identify the causes of the corresponding disease classes and highlighting the need for continued enrolment of patients with ultra-rare disorders. Thirdly, we have only considered SNVs and indels in coding genes. The exploration of rare variation in non-coding genes and in regulatory elements of the genome may help identify further etiologies. Lastly, we focused our attention on monogenic models of rare disorders, even though the genetic etiologies of certain rare diseases may be polygenic. In addition, important variation in clinical presentation of monogenic disorders may be explained by polygenic effects. These limitations point towards multiple promising avenues of future research to uncover the remaining unknown genetic determinants of rare diseases.

Acknowledgements

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. GS was performed by Illumina at Illumina Laboratory Services and was overseen by Genomics England. We thank all NHS clinicians who have contributed clinical phenotype data to the 100,000 Genomes rare diseases programme, and all staff at Genomics England who have contributed to the sequencing, maintenance of the research environment and assembly of the standard bioinformatic files that were required for our analyses. We thank the participants of the rare diseases program who made this research possible. We are grateful to Vaughan Keeley for providing access to paternal DNA (*ERG*), Frances Elmslie for inviting a patient to the clinic (*ERG*), and Thomas Jaworek for technical assistance (*GPR156*). D. Greene was supported by the Cambridge BHF Centre of Research Excellence [RE/18/1/34212] and Wellcome Collaborative Award 219506/Z/19/Z. V. Hartill was supported by

MRC/NIHR Clinical Academic Research Partnership MR/V037617/1. G. Birdsey and K. Frudd were funded by BHF project grant PG/17/33/32990. G. Birdsey and D. Pirri were funded by BHF project grant PG/20/16/35047. E. Sackey was supported by Swiss Federal National Fund for Scientific Research n°CRSII5_177191/1. P. Ostergaard and S. Mansour were supported by Medical Research Council grant MR/P011543/1 and BHF grant RG/17/7/33217. K. Freson was supported by KU Leuven BOF grant C14/19/096 and FWO grant G072921N. Work at the University of Maryland Baltimore was supported by the NIDCD/NIH grant R01DC016295 to Z. Ahmed. M. Al-Owain, K. Ramzan and F. Imtiaz were supported by King Salman Center for Disability Research grant 85722. E. Turro was supported by the Mindich Child Health and Development Institute, the Charles Bronfman Institute for Personalized Medicine and the Lowy Foundation USA.

Author contributions

D. Greene developed software, conducted analyses and co-wrote the paper. G.E.R.C. provided genetic and phenotypic data and access to the Genomics England Research Environment. C. Thys performed experiments and interpreted results. B. D. Gelb provided biological interpretation and feedback on the manuscript. K. Freson designed and supervised experiments, provided biological interpretation and contributed to writing the paper. A. Mumford provided clinical oversight, provided biological interpretation and contributed to writing the paper. E. Turro oversaw the study and co-wrote the paper. The following contributions relate to the three gene-specific vignettes. *ERG*: D. Pirri, K. Frudd and E. Sackey performed experiments and interpreted results. S. Mansour and C. L. S. Turner provided additional clinical information. P. Ostergaard coordinated validation and contributed to writing the paper. G. Birdsey designed and supervised experiments and contributed to writing the paper. *PMEPA1*: I. Yamanaka and N. Boeckx conducted experiments and interpreted results. P. Brennan, V. Hartill, J. Harvengt, T. Kosho, M. Masuno and T. Ohata provided clinical information. T. Morisaki and B. Loeys oversaw clinical and experimental studies. H. Morisaki recruited the Japanese cases, conducted experiments, interpreted and analysed results, and oversaw genetic studies. *GPR156*: H. Stewart provided additional clinical information for the compound heterozygous family. K. Taibah clinically evaluated and recruited the p.S642Afs*162 family. A. Giese,K. Ramzan and S. Riaz conducted experiments and interpreted results. M. Al-Owain assisted with experiments, interpreted results and contributed clinical information. S. Riazuddin, F. Imtiaz and Z. M. Ahmed designed and supervised experiments, analysed results, and provided reagents and tools.

Competing interests No authors have competing interests.

Gene	Case set	Level	Cases	Controls	Variant class	MOI	pLI	Z	Co-segregation evidence	Biological function and existing disease associations	Score	Replication
ERG	Primary lymphoedema	SD	94	55,400	High	Dom	**0.96**	2.53	**Co-segregation in 2 affected and 1 unaffected relatives (mosaicism in one affected parent).**	**ETS family transcription factor ERG is a critical regulator of endothelial lineage specification, vascular development, angiogenesis, and endothelial homeostasis.[40,20].**	3	Internal (case enrolled for a different Specific Disease)

GPR156	Congenital hearing impairment	SD	510	54,739	High	**Rec**	0	1.04	**Co-segregation in 2 affected and 4 unaffected relatives.**	**G protein-coupled receptor that regulates hair cell orientation in mechanosensory epithelia including in murine auditory epithelium[33].**	3	Riyadh
PMEPA1	Familial Thoracic Aortic Aneurysm Disease	SD	574	54,858	High	Dom	**0.94**	1.21	**Co-segregation in 3 affected relatives and distinctive phenotypic features**.	**Negative regulator of Transforming Growth Factor β (TGFβ) signalling[28]. Aberrant TGFβ signalling is implicated in multiple Mendelian aortopathies[29].**	3	100KGP pilot, Antwerp, Tokyo
LRRC7	Intellectual disability	SD	5,529	46,401	High	Dom	**1**	3.6		**Brain-specific scaffold protein in post-synaptic densities[37]. LRRC7-inactivated mice have neuro-behavioural phenotype[38]**	2	
USP33	Extreme early-onset hypertension	SD	182	55,305	High	Dom	0.86	**2.1**		**Deubiquitinating enzyme implicated in multiple cellular processes, including regulation of expression of the β2-adrenergic receptor[39], a critical regulator of circulatory function and blood pressure[41].**	2	
ARPC3	Charcot-Marie-Tooth disease	SD	549	54,856	Moderate	Dom	0.22	0.39		**Component of the Arp2/3 complex that regulates polymerisation of F-actin, abundant in axonal neurofilaments. Multiple Mendelian axonal filamentopathies manifest as Charcot-Marie-Tooth disease[42]. ArpC3-inactivation in mice causes axon dysfunction[43].**	1	
FMN1	Congenital hearing impairment	SD	510	54,738	High	**Rec**	0	-1.53	Co-segregation in 2 unaffected relatives.	**Formin family protein involved in linear actin and microtubule polymerisation[44]. Pathogenic variants in the formin DIAPH1 cause hearing loss via cytoskeletal disruption in auditory stereocilia[45].**	1	
RAB35	Familial Hypercholesterolaemia	SD	469	55,033	High	Dom	**0.98**	2.36	Co-segregation in 1 affected relative.	Small GTP-binding proteins that are a regulator of endosomal transport and function.	1	
RAB3A	Hereditary ataxia	SD	905	54,504	Moderate	Dom	0.95	**2.32**		Small GTP-binding proteins that regulate exocytosis and secretion. Although abundant in brain synaptic vesicles, rab3A-inactivated mice have no neuromuscular phenotype[46].	1	
TUFT1	Epidermolysis bullosa	SD	32	55,459	High	**Rec**	0	0.9	**Co-segregation in 1 affected and 4 unaffected relatives.**	Acidic protein that mediates dental enamel mineralisation.	1	
FAM222B	Ultra-rare undescribed monogenic disorders	SD	1,205	53,681	Moderate	Dom	0.29	0.42		Uncharacterised nucleosomal protein.	0	
INSL4	Rod Dysfunction Syndrome	SD	58	55,425	Moderate	Dom	0	-1.43		Insulin-like growth factor implicated in trophoblast and bone development.	0	
KRT14	Young onset tumour syndromes	DSG	256	55,207	Moderate	**Rec**	0	0.84	Co-segregation in 2 unaffected relatives.	Component of keratin intermediate filaments in epithelial cells. Pathogenic variants cause Epidermolysis	0	

										bullosa simplex !A-D(AD/AR); Dermatopathia pigmentosa reticularis (AD); Naegeli-Franceschetti-Jadassohn syndrome (AD).		
MPPE1	Primary ciliary dyskinesia	SD	105	55,360	High	**Rec**	0	0.35	Co-segregation in 2 unaffected relatives.	Metallophosphoesterase required for transport of GPI-anchor proteins from the endoplasmic reticulum to the Golgi.	0	
PKMYT1	Single autosomal recessive mutation in rare disease	SD	51	55,429	Moderate	**Rec**	0.22	0.07	Co-segregation in 2 unaffected relatives.	Serine/threonine protein kinase that is a negative regulator of cell entry into mitosis.	0	
RPL10A	Milroy disease	SD	20	55,470	High	**Dom**	0.85	2.06		Component of the large ribosomal subunit that mediates protein translation.	0	
SERPINB3	Autosomal recessive congenital ichthyosis	SD	46	55,437	Moderate	**Rec**	0	-1.66	Co-segregation in 2 unaffected relatives.	Cysteine endopeptidase inhibitor implicated in autocrine/paracrine signalling and cell protein metabolism	0	
SRP9	Ductal plate malformation	SD	54	55,445	High	**Dom**	0.42	1.13		Component of the signal recognition particle that targets secretory proteins to the endoplasmic reticulum.	0	
WWOX	Gastrointestinal disorders	DSG	59	55,413	Moderate	**Rec**	0	-4.44	Co-segregation in 1 unaffected relative.	Short-chain dehydrogenase/reductase that acts as a tumour suppressor and apoptosis regulator. Pathogenic variants cause developmental and epileptic encephalopathy 28 and spinocerebellar ataxia 12.	0	

Table 1 | Plausibility scoring of the 21 genetic associations identified by BeviMed. Each row corresponds to a genetic association between a gene and a case set in the 100KGP Main Programme without prior supporting evidence in PanelApp. Each column gives additional information for each association. Cells contributing to the final score are shown in bold. Rows are sorted by score in descending order and the genes corresponding to associations with a score of three are underlined. The level of the case set in the disease label hierarchy (DSG: Disease Sub Group, SD: Specific Disease), the class of variants and the MOI corresponding to the model with the greatest posterior probability are shown (Dom: dominant; Rec: recessive). A recessive association contributes one point to the score. A pLI >0.9 contributes one point to the score providing the inferred class of aetiological variants is high-impact variants. A *Z*-score >2 contributes one point to the score providing the inferred class of aetiological variants is moderate-impact variants. Evidence of co-segregation in ≥3 relatives in the 100KGP data contributes one point to the score (including mosaicism supported by ≥2 reads containing the alternate allele). Prior evidence of a relevant biological function or disease association contributes one point to the score. The 'Replication' column specifies cohorts in which additional cases were confirmed.

FIGURE LEGENDS

Fig. 1 | BeviMed analysis of the 100KGP. a, Bars showing the size of each case set used for the genetic association analyses, grouped by Disease Group and coloured by type (Disease Sub Group or Specific Disease). Case sets smaller than 5 are shown as having size 4 to comply with 100KGP policy on limiting participant identifiability. Below, the names and sizes of the case sets for an exemplar Disease Sub Group, 'Cardiovascular disorders', is shown. **b,** BeviMed PPAs >0.95 arranged by Disease

Group. Only the strongest association for each gene within a Disease Group is shown. Associations are coloured by their PanelApp evidence level (green, amber or red). Associations that were mapped to PanelApp by manual review, rather than using our automatic matching algorithm, are marked with an asterisk (**Source data for Fig. 1**). Previously unidentified associations are shown in grey. The shape of the points shows whether the association was with a Disease Sub Group (square), or Specific Disease (circle).

Fig. 2 | Loss-of-function variants in *ERG* are responsible for primary lymphoedema. a, Pedigrees for the four probands with loss-of-function variants in the canonical transcript of *ERG*, ENST00000288319.12. **b,** Truncated barchart showing the distribution of the number of reads supporting the p.S182Afs*22 alternate allele in the 100KGP. The embedded windows show the read pileups at this position in the two affected members of the family with the variant encoding p.S182Afs*22. The reads supporting the reference allele are in blue and those supporting the variant allele are in red. **c**, Schematic showing the effects of each variant at the cDNA and amino acid level, and on the protein product with respect to the canonical transcript. **d,** Reverse transcription-PCR amplification of ERG mRNA in HDLEC relative to HUVEC. Data are normalised to GAPDH. Statistical significance was assessed using a two-sided Student's *t*-test, n.s.: not significant (p=0.39). **e,** Immunoblot (representing two replicates) of HUVEC and HDLEC protein lysates identified several bands corresponding to ERG isoforms expressed at similar intensities in both cell types. **f,** Immunofluorescence microscopy (representing three replicates) of HDLEC shows ERG (green) nuclear co-localisation with lymphatic endothelial cell nuclear marker PROX1 (violet) and DAPI (blue). HDLEC junctions are shown using an antibody to VE-cadherin (yellow). Scale bar, 50μm. **g,** *En face* immunofluorescence confocal microscopy (representing five replicates) of mouse ear skin. Vessels are stained with antibodies to the lymphatic marker PROX1 (violet) and ERG (green). Scale bar, 100μm. **h,** Exemplar immunofluorescence microscopy image of HEK293 cells overexpressing wild type *ERG* and the p.S182Afs*22 variant *ERG*. Cells were stained for ERG (green) and nuclear marker DAPI (blue). Scale bar, 20μm. The brightness was optimised for print. **i,** Dot plot of estimated proportion of ERG not overlapping the nuclear marker DAPI in each of a set of immunofluorescence microscopy images of HEK293 cells overexpressing different *ERG* cDNAs (20 replicates for wild type (WT), 17 replicates per tested mutant). The estimated proportions were significantly higher in each of the variants compared to wild type: p=1.52x10^{-11}, 4.10x10^{-13} and 3.03x10^{-5} for each of p.S182Afs*22, p.T224Rfs*15 and p.A447Cfs*19, respectively (two-sided Student's *t*-test).

Fig. 3 | Truncating variants in *PMEPA1* result in Loeys-Dietz syndrome. a, Pedigrees for the three probands in the 100KGP (discovery cohort) heterozygous for the frameshift insertion predicting p.S209Qfs*3 and probands from replication cohorts, including: one from the 100KGP pilot programme heterozygous for the frameshift deletion predicting p.S209Afs*61, three of Japanese ancestry heterozygous for p.S209Qfs*3 and one Belgian pedigree heterozygous for a frameshift deletion encoding p.P207Qfs*3. All variant consequences are shown with respect to the canonical transcript of *PMEPA1*, ENST00000341744.8. **b,** HPO terms present in at least three of the four *PMEPA1* FTAAD families, excluding redundant terms within each level of frequency, alongside their frequency in four *PMEPA1* FTAAD families and the other 589 unexplained FTAAD families. Terms are ordered by p-value obtained by a Fisher's exact test of association between the term's presence in an FTAAD family and whether the family is one of the four *PMEPA1* families. Terms were declared significant (indicated by an asterisk), or not significant (n.s.) by comparing their Fisher test p-values and rank to a null distribution of equivalent

pairs obtained by permutation (10,000 replicates). For each rank, the *p*-value of the term on the 5th percentile was used as an upper bound for declaring an association significant, provided all terms at higher ranks were also significant. The *p*-values for each term were as follows. Dolichocephaly: $p=2.9\times10^{-4}$, Abnormal axial skeleton morphology: $p=6.7\times10^{-3}$, Striae distensae: $p=0.013$, Pes planus: $p=0.014$, Ascending tubular aorta aneurysm: $p=0.62$. **c**, Graph showing *PMEPA1* and genes with high evidence (green) of association with FTAAD in PanelApp. Edges connect genes where the string-db v11.5[27] confidence score for physical interactions between corresponding proteins was >0.6. Genes known to be associated with Loeys-Dietz syndrome are highlighted in blue. *PMEPA1* is highlighted yellow. **d**, Schematic showing the effects of each variant at the cDNA and amino acid level, and on the protein product.

Fig. 4 | Loss-of-function variants in *GPR156* give rise to recessive congenital hearing loss. a, Schematic of the three pedigrees with cases homozygous or compound heterozygous for loss-of-function variants in the canonical transcript of *GPR156*, ENST00000464295.6. Blank symbols indicate individuals with an unknown genotype. **b,** Histograms of expression log fold changes for different sets of genes in mouse hair cells compared to surrounding cells: all genes (left) and genes homologous to the human counterparts in the 'Hearing loss' PanelApp panel with and without a stereocilia-related GO term (i.e. a term whose name contained ‘stereocilia’ or ‘stereocilium’, or the descendant of such a term) (right). The log fold change for *Gpr156* is shown as a horizontal line. **c,** Maximum intensity projections of confocal Z-stacks in the organ of Corti and vestibular system of a P20 wild type mouse immunostained with GPR156 antibody (green) and counterstained with phalloidin (red). Top row: overview of the organ of Conti and vestibular system. Middle and bottom rows: magnified images of outer hair cells (OHC) and inner hell cells (IHC), respectively. No stereociliary bundle staining was observed. The punctate staining observed in the organ of Corti was absent or significantly decreased in the utricle of the vestibular system. Scale bars: 10 μm (each image representative of three replicates). **d,** Schematic showing the effects of each variant at the cDNA and amino acid level, and on the protein product. **e,** Exemplar western blot taken from three replicates of GFP-GPR156 using anti-GPR156 antibody in untransfected Cos7 cells (Cos7), Cos7 cells transfected with the wild type construct (W) and Cos7 cells transfected with the constructs containing each of the mutant alleles p.S642Afs*162 (S642), p.P718Lfs*86 (P718) and p.S207Vfs*113 (S207).

References

1. Boycott KM et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. Am J Hum Genet. 2017; 100(5):695--705.
2. Ferreira CR. The burden of rare diseases. Am J Med Genet A. 2019; 179(6):885--892.
3. Turro E et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020; 583(7814):96--102.
4. Wang Q et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature. 2021; 597(7877):527--532.
5. Kaplanis J et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature. 2020; 586(7831):757--762.

6. Greene D, Richardson S, Turro E. Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases. Am J Hum Genet. 2016; 98(3):490--499.
7. Hail Team. (2022). Hail 0.2. https://github.com/hail-is/hail.
8. Lopez J et al. HGVA: the Human Genome Variation Archive. Nucleic Acids Res. 2017; 45(W1):W189-W194.
9. Greene D, Richardson S, Turro E. A Fast Association Test for Identifying Pathogenic Variants Involved in Rare Diseases. Am J Hum Genet. 2017; 101(1):104--114.
10. Karczewski KJ et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581(7809):434--443.
11. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021; 13(1):31.
12. Eilbeck K et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2005; 6(5):R44.
13. Howe KL et al. Ensembl 2021. Nucleic Acids Res. 2021; 49(D1):D884-D891.
14. Martin AR et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019; 51(11):1560--1565.
15. Korber L, Schneider H, Fleischer N, Maier-Wohlfart S. No evidence for preferential X-chromosome inactivation as the main cause of divergent phenotypes in sisters with X-linked hypohidrotic ectodermal dysplasia. Orphanet J Rare Dis. 2021; 16(1):98.
16. Kasahara Y et al. Hyper-IgM syndrome with putative dominant negative mutation in activation-induced cytidine deaminase. J Allergy Clin Immunol. 2003; 112(4):755--760.
17. Martin-Almedina S, Mortimer PS, Ostergaard P. Development and physiological functions of the lymphatic system: insights from human genetic studies of primary lymphedema. Physiol Rev. 2021; 101(4):1809--1871.
18. Gordon K et al. Update and audit of the St George's classification algorithm of primary lymphatic anomalies: a clinical and molecular approach to diagnosis. J Med Genet. 2020; 57(10):653--659.
19. Kalna V et al. The Transcription Factor ERG Regulates Super-Enhancers Associated With an Endothelial-Specific Gene Expression Program. Circ Res. 2019; 124(9):1337--1349.
20. Shah AV, Birdsey GM, Randi AM. Regulation of endothelial homeostasis, vascular development and angiogenesis by the transcription factor ERG. Vascul Pharmacol. 2016; 86:3--13.
21. Hoesel B et al. Sequence-function correlations and dynamics of ERG isoforms. ERG8 is the black sheep of the family. Biochim Biophys Acta. 2016; 1863(2):205--218.
22. Westbury SK et al. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. Genome Med. 2015; 7(1):36.
23. Greene D, Richardson S, Turro E. ontologyX: a suite of R packages for working with ontological data. Bioinformatics. 2017; 33(7):1104--1106.
24. Resnik P, Others. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res.(JAIR). 1999; 11:95--130.
25. Ciurica S et al. Arterial Tortuosity. Hypertension. 2019; 73(5):951--960.
26. Loeys BL et al. A syndrome of altered cardiovascular, craniofacial, neurocognitive and skeletal development caused by mutations in TGFBR1 or TGFBR2. Nat Genet. 2005; 37(3):275--281.
27. Szklarczyk D et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; 47(D1):D607-D613.

28. Watanabe Y et al. TMEPAI, a transmembrane TGF-beta-inducible protein, sequesters Smad proteins from active participation in TGF-beta signaling. Mol Cell. 2010; 37(1):123--134.
29. Creamer TJ, Bramel EE, MacFarlane EG. Insights on the Pathogenesis of Aneurysm through the Study of Hereditary Aortopathies. Genes (Basel). 2021; 12(2).
30. Thermann R et al. Binary specification of nonsense codons by splicing and cytoplasmic translation. EMBO J. 1998; 17(12):3484--3494.
31. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat. 2015; 36(10):928--930.
32. Ellaithy A, Gonzalez-Maeso J, Logothetis DA, Levitz J. Structural and Biophysical Mechanisms of Class C G Protein-Coupled Receptor Function. Trends Biochem Sci. 2020; 45(12):1049--1064.
33. Kindt KS et al. EMX2-GPR156-Gai reverses hair cell orientation in mechanosensory epithelia. Nat Commun. 2021; 12(1):2861.
34. Scheffer DI, Shen J, Corey DP, Chen ZY. Gene Expression by Mouse Inner Ear Hair Cells during Development. J Neurosci. 2015; 35(16):6366--6380.
35. Miyoshi T et al. Human deafness-associated variants alter the dynamics of key molecules in hair cell stereocilia F-actin cores. Hum Genet. 2022; 141(3-4):363--382.
36. Smedley D et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med. 2021; 385(20):1868--1880.
37. Thalhammer A, Trinidad JC, Burlingame AL, Schoepfer R. Densin-180: revised membrane topology, domain structure and phosphorylation status. J Neurochem. 2009; 109(2):297--302.
38. Chong CH et al. Lrrc7 mutant mice model developmental emotional dysregulation that can be alleviated by mGluR5 allosteric modulation. Transl Psychiatry. 2019; 9(1):244.
39. Berthouze M, Venkataramanan V, Li Y, Shenoy SK. The deubiquitinases USP33 and USP20 coordinate beta2 adrenergic receptor recycling and resensitization. EMBO J. 2009; 28(12):1684--1696.
40. Birdsey GM et al. The endothelial transcription factor ERG promotes vascular stability and growth through Wnt/Beta-catenin signaling. Dev Cell. 2015; 32(1):82--96.
41. Motiejunaite J, Amar L, Vidal-Petiot E. Adrenergic receptors and cardiovascular effects of catecholamines. Ann Endocrinol (Paris). 2021; 82(3-4):193--197.
42. Munoz-Lasso DC, Roma-Mateo C, Pallardo FV, Gonzalez-Cabo P. Much More Than a Scaffold: Cytoskeletal Proteins in Neurological Disorders. Cells. 2020; 9(2).
43. Zuchero JB et al. CNS myelin wrapping is driven by actin disassembly. Dev Cell. 2015; 34(2):152--167.
44. DeWard AD, Eisenmann KM, Matheson SF, Alberts AS. The role of formins in human disease. Biochim Biophys Acta. 2010; 1803(2):226--233.
45. Ninoyu Y et al. The integrity of cochlear hair cells is established and maintained through the localization of Dia1 at apical junctional complexes and stereocilia. Cell Death Dis. 2020; 11(7):536.
46. Geppert M et al. The role of Rab3A in neurotransmitter release. Nature. 1994; 369(6480):493--497.

METHODS

Ethics

The 100,000 Genomes project was approved by East of England–Cambridge Central REC REF 20/EE/0035. Only participants who provided written informed consent for their data to be used for research were included in the analyses. The study at the University of Maryland was approved by the

Institutional Review Board (RAC#2100001) and written informed consent was obtained by clinicians at King Faisal Hospital in Saudi Arabia from the participating individuals. The study of the Japanese ancestry pedigrees bearing *PMEPA1* truncating alleles was approved by the Institutional Review Board of the National Cerebral and Cardiovascular Centre (M14-020) and Sakakibara Heart Institute (16-035), and written informed consent was obtained from the participating individuals.

Motivation for developing a sparse relational database

Computational approaches for discovering the aetiologies of rare diseases typically depend on the analysis of a heterogeneous set of files, each of which can be very large and follow a distinct convention. Genotypes, for example, are ordinarily stored in VCFs containing data for one sample or for multiple samples. In the latter case, the data are usually distributed in files covering many different "chunks" of the reference genome. Variant-level information, such as consequence predictions or pathogenicity scores, are typically encoded in strings that require extensive parsing to decode, either from within the VCFs containing the genotypes, or in separate files. Modifying genotype or annotation files, for example in order to incorporate newly generated data, requires rewriting files in their entirety. Phenotype data, pedigree data and the results of statistical inference are stored in a further set of files. Consequently, analyses are often burdensome to conduct and prone to error. Frameworks such as Hail[7] and OpenCGA[8] afford greater flexibility but they depend on the centrally organised deployment of a distributed storage system, hindering usability and portability.

Relational databases are widely used, mature technologies, well known for their speed, reliability, flexibility, structure and extensibility. In the context of rare diseases, a relational database can in principle render the modification, combination and addition of data on samples, variants, genes and other entities efficient, reliable and straightforward to implement using a single query language. Unfortunately, the performance of relational databases degrades substantially when the number of records in a table reaches several billion, and the number of genotypes in a cohort the size of the 100KGP easily surpasses this threshold. However, the MAFs of pathogenic variants with strong effects on rare disease risk are typically kept below 1/1,000 by negative selection, and the proportion of non-homozygous reference genotypes for variants within that MAF stratum is only about 1% of the total (**Extended Data Fig. 1**). Consequently, it is possible to construct a compact relational database that includes virtually all the pathogenic variants even in a large cohort such as the 100KGP. This provides an opportunity for exploiting the benefits of a single unified relational database containing the non-homozygous genotypes of rare variants upon which to conduct the entirety of the aetiological discovery process. Furthermore, it provides a natural foundation for developing web applications for the multidisciplinary review of genetic, phenotypic, statistical and other data.

Rareservoir

The Rareservoir is a relational database schema and a complementary software package ‘rsvr’ for working with rare disease data. The database stores data including rare variant genotypes, variant annotations, phenotypes, sample information and pedigrees (**Extended Data Fig. 1**) but it can be extended arbitrarily. A Rareservoir is built through a series of steps from a set of input data and parameters (**Extended Data Fig. 3**). The 'bcftools' program[47] extracts ('bcftools view') and normalises ('bcftools norm') variants from either a set of single sample genome VCF files (gVCFs) or from a merged VCF. In all steps of the procedure, variants are encoded as RSVR IDs using the 'rsvr enc' tool (see

Encoding RSVR IDs). Merged VCFs typically contain cohort-wide variant quality information in the FILTER column, which can be used to select variants for processing. However, this is not readily obtained from single gVCFs. To address this, we developed the ‘rsvr depth’ tool, which computes variant quality pass rates at all positions in the genome based on a random subsample of gVCFs. If the input is a merged VCF, an internal (i.e. within-VCF) allele frequency threshold is applied with bcftools to filter out internally common variants. If the input is a set of single-sample gVCFs, internally common variants are filtered out in two steps, for computational efficiency. Firstly, a set of variants that are statistically almost certain to be common based on a random sample of gVCFs is identified—by default, the variants for which a one-sided binomial test under the null hypothesis that the MAF=0.01 is rejected at a significance level of 10^{-6} (done using the ‘rsvr tabulate’ tool). Secondly, all gVCFs are read sequentially, filtering out the variants identified in the previous step (using the ‘rsvr mix’ tool) and those for which the pass rates identified with 'rsvr depth' do not meet the threshold. Retained genotypes are then loaded into a temporary genotype table in the database in order to apply the final internal allele frequency filter by executing an SQL 'DELETE' statement. These variants are then annotated with gnomAD 'probabilistic minor allele frequency' (PMAF) scores[3] using the ‘rsvr pmaf’ tool. The PMAF score is calculated with respect to a given allele frequency threshold *t*, by evaluating a binomial test (at a significance threshold of 0.05) on the observed frequency of the variant under the null hypothesis that the variant has an allele frequency of *t*. If, in any gnomAD population, the null is rejected for *t*=0.001 and the allele count is at least 2, the score is set to 0. If the null is rejected for *t*=0.0001, the score is set to 1. If the null is not rejected, the score is set to 2. Finally, if the variant is absent from gnomAD, the score is set to 3. For the non-pseudo autosomal dominant regions of chromosome X, only allele counts for males are used in calculations. Variants are then additionally annotated with their CADD phred scores using the 'rsvr ann' program, and loaded into the VARIANT table. At this point, variants in the VARIANT and GENOTYPE table which have a PMAF score of 0 may be deleted because they are unlikely to be involved in rare diseases. We then annotate the retained variants with predicted transcript consequences for a given set of transcripts specified in a Gene Transfer Format (GTF) file. The 100KGP Rareservoir, uses Ensembl v104 canonical transcripts with a protein-coding biotype, of which >90% are Matched Annotation from NCBI and EMBL-EBI (MANE)[48] transcripts. The 'rsvr seqfx' program determines a set of SO terms for each interacting transcript–variant pair and encodes them as a CSQ ID, which is added to the CONSEQUENCE table. This table can also hold LOFTEE scores corresponding to a transcript-variant pair. Note that, as LOFTEE scores on the Genomics England Research Environment correspond to Ensembl v99 transcripts, we mapped Ensembl v104 canonical transcripts to the most similar v99 transcripts having an identical CDS in order to obtain the LOFTEE scores for the 100KGP Rareservoir, finding a match for >98% of transcripts. The contents of the GTF file are also imported into the database to create tables of transcript features (FEATURE), transcripts (TX) and genes (GENE). Optionally, VARIANT, GENOTYPE and CONSEQUENCE may be filtered for RSVR IDs that have CSQ IDs meeting particular criteria, for instance, in order to retain only variants with protein-coding consequences. The SAMPLE table of metadata and genetic statistics for each sample represented in the input VCF(s) must then be added to the database, including mandatory columns containing the ID, sex, family, and an indicator of inclusion in the maximal unrelated set of samples in the database. The VARIANT, GENOTYPE and CONSEQUENCE tables are indexed by RSVR ID, in order to support fast lookups by genomic location. The SAMPLE table and GENOTYPE table are indexed by sample ID allowing fast lookups by sample. The CONSEQUENCE, TX and GENE tables are indexed by transcript and gene ID, allowing fast lookups of variants based on gene/transcript specific consequences. If sample phenotypes

have been encoded using phenotypic terms (e.g. ICD10 codes or HPO terms), terms from the relevant coding systems can be added to a generic PHENOTYPE table mapping code IDs to descriptions, and codes assigned to samples can be added to the SAMPLE_PHENOTYPE table. Disease labels may be added to the CASE_SET table. The majority of the compute time required for building the database is taken by reading the genotype data from the input VCF, which may be executed in parallel over separate regions against a merged VCF or over single gVCFs. The rsvr tool, implemented in C++, executes rapidly, with 'rsvr seqfx' capable of assigning CSQ IDs for all Ensembl v104 canonical transcripts to all variants (over 685M) in gnomAD v3.0 in under 40 minutes on a single core. The 100KGP Rareservoir, which is stored in a SQLite database, returns complex gene-specific queries in under one second. For example: (1) a table with 628 rows containing the moderate and high-impact variants with a PMAF score ≥1 in *TTN*, along with the corresponding SeqFx consequence predictions and CADD scores takes 0.57 seconds; (2) a table with 1,498 rows containing, for each variant, the samples and genotypes for individuals who carry an alternate allele takes 0.61 seconds; and (3) a classification for each of the 77,539 participants into proband with Dilated Cardiomyopathy, relative of such a proband, unrelated control, or relative of a control takes 0.65 seconds. Specific details on implementation of the workflow, code for encoding data as SQL statements compatible with Rareservoir and the mapping between bits in the 64 bit CSQ ID and each SO term assigned by 'rsvr seqfx' can be found in the rsvr software package (see **Code availability**). Software packages rsvr 1.0, bcftools 1.9 and perl 5 were used to build the 100KGP Rareservoir.

Encoding RSVR IDs

SNVs and indels may be encoded as 64-bit integers called RSVR IDs. In order to compute an RSVR ID for a given variant, the following expression is evaluated:

$$c \times 2^{58} + p \times 2^{30} + |r| \times 2^{24} + |a| \times 2^{18} + \sum_{i=1}^{|A|} A_i \times 4^{i-1},$$

where *c* is the chromosome number (using 23, 24 and 25 respectively to represent X, Y and MT), *p* is the position, and |*r*| and |*a*| are the lengths of the reference and alternate alleles, respectively. *A* is a sequence identical to the alternate allele, *a*, when its length is less than 10, and otherwise equal to the first five followed by the last four elements of *a*. In the summation, nucleotides are assigned values: `A`= 0, `C` = 1, `G` = 2 and `T` = 3. The expression evaluates to integers that can be represented using 63 bits, setting the most significant bit to 0 when encoding as 64-bit integers. The chromosome, position, reference and alternate allele lengths and alternate allele bases are thereby encoded respectively by the subsequent 5, 28, 6, 6 and 18 bits (with two bits per base for the alternate allele). This procedure and its inverse are implemented in the 'rsvr enc' and 'rsvr dec' programs respectively. The reference and alternate alleles of input variants are first normalised by removing any redundant identical sequence from the starts and then the ends. The proportion of variants in gnomAD 3.0 weighted by allele count that can be encoded losslessly is 99.3%, while 99.8%can be represented by a distinct RSVR ID. The full variant information corresponding to any encountered ambiguous RSVR ID may be stored in full in a dedicated table. Structural variants that can be represented by a position and length may also be encoded using distinct 64-bit RSVR IDs alongside SNVs and indels by setting the most significant bit to 1, and subsequently encoding the type of structural variant using 2 bits (Deletion 0, Duplication 1, Inversion 2, Insertion 3), the chromosome using 5 bits (as done for SNVs and indels), and the start and length consecutively using 28 bits.

Genetic association analysis of 100KGP data

We constructed a Rareservoir in the Genomics England Research Environment containing the PASSing[49] variants in the merged VCF of 77,539 consented participants in the 100KGP rare diseases programme. This Rareservoir only included variants with a PMAF >0 according to GnomAD v3.0, an internal MAF <0.002 and at least one predicted consequence on a canonical transcript in Ensembl v104. Variants with a greater MAF are unlikely to be highly penetrant for diseases eligible for inclusion in the 100KGP and are likely to have, at most, small effects on risk, making them challenging to validate. Variants with a median genotype quality <35 and SNVs with a CADD Phred score <10 were also excluded from the analyses.

For each of the 269 rare disease classes (**Extended Data Figs. 5–6**), we applied the BeviMed[9] association test to rare variants extracted from the Rareservoir database in each of the 19,663 canonical transcripts belonging to a gene with a 'protein_coding' biotype. The analysis was carried out using R 3.6.2, making use of functionality from packages: Matrix 1.2-18, dplyr 0.8.5, bit64 0.9-7, bit 1.1-14, DBI 1.1.0, RSQLite 2.1.4 and BeviMed 5.7. The case set for a given disease class and gene was constructed by selecting one case from each pedigree containing at least one person affected with the disease class. For the purposes of the association analysis, participants were labelled 'explained' by a given gene if they had variants in that gene classified as 'pathogenic_variant' or 'likely_pathogenic_variant' in the 'gmc_exit_questionnaire' table in the Genomics England Research Environment. To boost power, we used this information to reassign cases who were explained by variants in a different gene to the control group.

Using BeviMed, we performed a Bayesian comparison of a baseline model of no association and each of six association models defined by a mode of inheritance and a class of aetiological variant:

1. No association (prior probability: 0.99),
2. Dominant association with "high"-impact variants having a PMAF ≥2 (i.e., corresponding to a target MAF <0.01%) (prior probability: 0.002475),
3. Dominant association with "moderate"-impact variants having a PMAF ≥2 (prior probability: 0.002475),
4. Dominant association with "5' UTR" variants having a PMAF ≥2 (prior probability: 0.00005),
5. Recessive association with "high"-impact variants having a PMAF ≥1 (i.e., corresponding to a target MAF <0.1%) (prior probability: 0.002475),
6. Recessive association with "moderate"-impact variants having a PMAF ≥1 (prior probability: 0.002475),
7. Recessive association with "5' UTR" variants having a PMAF ≥1 (prior probability: 0.00005).

Thus the overall prior probability of association was 0.01 and there was an equal prior probability of dominant and recessive inheritance. The PPA was the sum of the posterior probabilities of models 2 through 7. We imposed a stricter PMAF threshold under a dominant MOI than under a recessive MOI because, ceteris paribus, dominant variants are under stronger negative selection than recessive variants. The three groups of variants were selected as follows:

- 5' UTR variants: those with a 5_prime_UTR_variant consequence,
- High-impact variants: those with any consequence amongst start_lost, stop_lost, frameshift_variant, stop_gained, splice_donor_variant or splice_acceptor_variant, excluding variants with a "low-confidence" LOFTEE score[10],

- Moderate-impact variants: those with any consequence amongst start_lost, stop_lost, frameshift_variant, stop_gained, splice_donor_variant or splice_acceptor_variant, missense_variant or inframe_deletion.

The rationale for embedding variants from the high-impact class in the moderate-impact class is that both types of variant are capable of inducing a loss of function. The prior on the probability that a modelled rare variant is pathogenic, conditional on either the association model mediated by 5' UTR variants or the association model mediated by moderate-impact variants, was set to Beta(2,8). This encodes a prior conditional expectation that 20% of rare variants are pathogenic, which is well suited to missense and 5' UTR variants. However, we specified a distribution with a greater mean for the high-impact models. Specifically, the prior on the probability that a modelled high-impact variant is pathogenic was set to Beta(3,1), which reflects a prior conditional expectation that 75% of rare variants are pathogenic because loss-of-function variants tend to be functionally equivalent to each other. BeviMed reports the posterior probability that each variant is pathogenic conditional on the mode of inheritance and the class of aetiological variant. The methodology is described in further detail in the original BeviMed publication[9].

We applied the following post-processing of BeviMed results with a PPA >0.95:

- We re-ran BeviMed including all samples (i.e. with relatives of cases and controls). Associations for which the analysis with all samples caused the PPA to fall below 0.9 were filtered out due to conflicting evidence for the association within families.
- We re-ran BeviMed after removing variants absent from affected relatives of the cases. Associations for which this removal caused the PPA to drop below 0.25 were filtered out because they depended on variants that were not shared by affected cases within families.
- To guard against false positives due to incorrect pedigree data, population structure or cryptic relatedness, we applied the following algorithm. We obtained the distribution of the number of rare variants in the Rareservoir shared by pairs of individuals within each assigned ancestry in the 100KGP. The top percentile in each of these distributions was used to indicate potential relatedness between participants in the same population. We re-ran BeviMed after removing cases so as to ensure that no more than one case from any set of potentially related cases sharing a variant were included in the analysis. . Associations for which this analysis caused the PPA to fall below 0.25 were filtered out.

To account for correlation between case sets, for each gene, we removed all but the most strongly associated disease class within each Disease Group before reporting the 260 associations remaining (**Source data for Fig. 1**). Without the post-processing, the number of reported associations would have been 302. Conditional on the modal model underlying each of the 260 associations, we recorded the variants with a posterior probability of pathogenicity >0.8 accounting for at least one case in the 100KGP (**Source data for Fig. 2**).

PanelApp annotation

Significant associations were coloured according to PanelApp[14] (**Fig. 1b**) evidence levels for panel–gene relations (green for high evidence, amber for moderate evidence, and red for low evidence) for panels of type ‘Rare Disease 100K’, which are organised hierarchically by Disease Sub Group and Disease Group, or of type ‘GMS Rare Disease’. Given an association between a gene and a case set (corresponding either to a Specific Disease or a Disease Sub Group), we searched for panels which contained the gene

and had the same name as the case set (ignoring case). If such a match was not found, we searched for panels which contained the gene and which belonged to a Disease Sub Group with the same name as the Disease Sub Group of the case set. When this matching rule generated multiple matches, we selected the panel(s) with the highest evidence. If multiple panels still remained, we selected the panel with the smallest number of genes. Associations for which no matching panel in PanelApp could be found were inspected manually to assess whether PanelApp contained an alternative suitable panel (marked with an asterisk in **Fig. 1b**).

Shortlisting previously unidentified genetic associations for validation

Several sources of independent evidence were used to shortlist significant associations for validation. For each source, a score of one was awarded if the evidence was supportive, and zero otherwise. Scores were then added over the different sources and used to rank the associations. Associations for which at least three sources of evidence were supportive were taken forward for further investigation. The sources of evidence and qualifying criteria for being considered supportive are listed below. Note that here we refer to variants which had a probability of pathogenicity >0.8 conditional on the modal model as 'probably pathogenic'.

- *Counting co-segregating pedigree members.* The pedigrees harbouring pathogenic configurations of probably pathogenic alleles were checked for co-segregation between genotype and affection status. This evidence counted as supportive for associations for which all such pedigrees demonstrated co-segregation, and there were at least three additional relatives who had not been included in the association analysis but for whom there was co-segregation. Note that BAM files for the affected members of pedigrees who were called homozygous reference for probably pathogenic variants were checked for evidence of mosaicism to guard against the possibility that they were falsely portraying a lack of co-segregation.
- *pLI and Z-scores.* pLI and *Z*-scores for depletion of missense variants were obtained from the gnomAD v2.2.1 browser[10]. pLI >0.9 for associations in which high impact variants were most strongly associated were counted as supportive, whilst *Z*-scores greater than 2 for associations in which moderate impact variants were most strongly associated were counted as supportive.
- *Recessive association.* Population genetic metrics of purifying selection (pLI scores and *Z*-scores) are sensitive to depletion of high-impact variants and missense variants, respectively. They are therefore useful measures to corroborate dominant associations. However, these metrics have low sensitivity to identify the signatures of selection against recessive diseases because isolated pathogenic variants in heterozygous form do not lead to a reduction in reproductive fitness. To avoid disadvantaging recessive associations identified by BeviMed, they were assigned a contribution of one point to the score.
- *Literature review.* A comprehensive literature review, assessing the gene's role (if any) in biological processes relevant to the disease, other diseases, and a survey of model organisms was undertaken, and determined to be either supportive or not.

***ERG*: Primary endothelial cell culture**

Single donor primary human dermal lymphatic endothelial cells (HDLEC) (Promocell, Heidelberg) were cultured in Endothelial Cell Growth Medium MV2 (Promocell). Pooled donor human umbilical vein endothelial cells (HUVEC) (Lonza, Slough) were grown in Endothelial Cell Growth Media-2 (EGM-2) (Lonza). HUVEC and HDLEC were grown on 1% (v/v) gelatin and used between passages 3-5.

***ERG*: Real-time polymerase chain reaction**

HUVEC and HDLEC were grown to confluency in a pre-gelatinised 6-well dish. Total RNA was isolated using the RNeasy Mini Kit (Qiagen) and 1 µg of total RNA was transcribed into cDNA using Superscript III Reverse Transcriptase (Thermo Fisher Scientific). Quantitative real-time PCR was performed using PerfCTa SYBR Green FastMix (Quanta Biosciences) on a Bio-Rad CFX96 System. Gene expression values of ERG in HUVEC and HDLEC were normalised to GAPDH expression and compared using the ΔΔCt method. The following oligonucleotides were used: ERG, 5'-GGAGTGGGCGGTGAAAGA-3' and 5'-AAGGATGTCGGCGTTGTAGC-3'; GAPDH, 5'-CAAGGTCATCCATGACAACTTTG-3' and 5'-GGGCCATCCACAGTCTTCTG-3'.

***ERG*: Immunoblotting analysis**

Immunoblotting was performed according to standard conditions. Proteins were labelled with the following primary antibodies: rabbit anti-human ERG antibody (1:1000; ab133264, Abcam) and mouse anti-human GAPDH (1:10000; MAB374, Millipore). Primary antibodies were detected using fluorescently labelled secondary antibodies: goat anti-rabbit IgG DyLight 680 and goat anti-mouse IgG Dylight 800 (Thermo Scientific). Detection of fluorescence intensity was performed using an Odyssey CLx imaging system (Li-COR Biosciences, Lincoln) and Odyssey version 4 software.

***ERG*: Immunofluorescence analysis of endothelial cells and mouse tissues**

Confluent cultures of HUVEC and HDLEC were fixed with 4% (w/v) paraformaldehyde for 15 minutes and permeabilised with 0.5% (v/v) Triton-X100, before incubation with 3% BSA (w/v) in PBS containing the following primary antibodies: goat anti-human PROX1 antibody (1:100; AF2727, R&D Systems), rabbit anti-human ERG antibody (1:100; ab92513, Abcam), mouse anti-human VE-cadherin (1:100; 555661, BD Biosciences). Secondary antibody incubation was carried out in 3% BSA (w/v) in PBS, using the following antibodies: donkey anti-goat IgG Alexa Fluor-488 (1:1000; A-11055), donkey anti-rabbit IgG Alexa Fluor-555 (1:1000; A-31572), donkey anti-mouse Alexa Fluor-594 (1:1000; A-21203). All secondary antibodies from Thermo Fisher Scientific. Nuclei were visualised using DAPI (4′,6-diamidino-2-phenylindole) (Sigma-Aldrich). Confocal microscopy was carried out on a Carl Zeiss LSM780 confocal laser scanning microscope with Zen 3.2 software. All animal experiments were conducted with ethical approval from Imperial College London under UK Home Office Project Licence number PEDBB1586 in compliance with the UK Animals (Scientific Procedures) Act of 1986. Ear tissue was collected from euthanised 3-week old male and female C57BL/6J mice and fixed in 4% (w/v) paraformaldehyde at room temperature for 2h. Tissue was then washed with PBS followed by a blocking and permeabilization step using 3% (w/v) milk in PBST (containing 0.3% (v/v) Triton X-100 in PBS) for 1h at room temperature. The following primary antibodies were used for immunofluorescence staining: goat anti-human PROX1 antibody (1:100; AF2727, R&D Systems) and rabbit anti-human ERG antibody (1:100; ab92513, Abcam). Primary antibodies were incubated at 4°C overnight in 3% (w/v) milk in PBST. The following day, tissues were washed three times with PBST over the course of 2h at room temperature. Tissues were incubated with secondary antibodies at room temperature for 2h in 3% milk (w/v) in PBST. Primary antibodies were detected using fluorescently labelled secondary antibodies: donkey anti-goat IgG Alexa Fluor-488 (1:400; A-11055, Thermo Fisher Scientific) and donkey anti-rabbit IgG Alexa Fluor-555; A-31572, Thermo Fisher Scientific). Stained samples were mounted onto glass slides using Fluoromount G (Thermo Fisher

Scientific). Images were acquired using Zeiss LSM-780 confocal laser scanning microscope with Zen 3.2 software. All confocal images represent maximum intensity projection of Z-stacks of single tiles.

ERG: Subcloning and overexpression in HEK293 cells

We subcloned *ERG* (ENST00000288319.12) from HUVEC into the mammalian expression vector pcDNA3.1 (Thermo Fisher). *ERG* variants were generated by site-directed mutagenesis using the Quikchange Lightning kit (Agilent, Stockport, Cheshire) using the wild type *ERG* cDNA as template. Expression of wild type and mutant *ERG* was carried out using Polyethylenimine (PEI; Sigma-Aldrich) transfection reagent in HEK293 cells grown in DMEM (Thermo Fisher) with 10% (v/v) fetal bovine serum. After 24 hr, cells were fixed with 4% (w/v) paraformaldehyde for 15 minutes and permeabilised with 0.5% (v/v) Triton-X100, before incubation with 3% BSA (w/v) in PBS containing mouse monoclonal anti-ERG antibody (1:100; sc-376293, Santa Cruz Biotechnology). Secondary antibody incubation was carried out in 3% BSA (w/v) in PBS, using donkey anti-mouse Alexa Fluor-488 (1:1000; A-21202, Thermo Fisher). Nuclei were visualised using DAPI (4′,6-diamidino-2-phenylindole) (Sigma-Aldrich). Confocal microscopy was carried out on a Carl Zeiss LSM780 confocal laser scanning microscope with Zen 3.2 software.

ERG: Estimation of nuclear and non-nuclear ERG in HEK293 cells

Each image was read into a pair of channel-specific 1,024 x 1,024 matrices in R 4.2.1 using the readCzi function from the readCzi R package v0.2.0. A pixel was declared to contain a nuclear region if the intensity in the blue channel exceeded 60% of the 95th percentile of blue intensities across all pixels above background (identified as exceeding 1.35×10^{-2} by visual inspection of bimodal intensity histograms). A pixel was declared to contain ERG if the intensity in the green channel exceeded 30% of the 95th percentile of the green intensities within the pixels previously declared to be nuclear. To fill in intranuclear gaps, any non-nuclear pixels adjacent to at least 5 nuclear pixels were declared nuclear. The estimated proportion of ERG that was cytosolic in an image was set to the number of ERG pixels that did not overlap nuclear pixels divided by the number of ERG pixels.

GPR156: Western blots

We subcloned *GPR156* from human brain cDNA, into EGFP-N2 vector. The three mutant *GPR156* constructs were generated by mutagenesis using the QuickChange kit (Stratagene, La Jolla, CA) and a wild type GPR156-GFP as a template. For expression analysis, the WT and mutants were transfected in COS7 cells grown in DMEM (Gibco, Gaithersburg, MD, USA) with 10% fetal bovine serum. Transfections were performed with Lipofectamine 2000 reagent (Life Technologies). Cells were harvested 48hr after transfection, lysed in buffer containing 1% CHAPS, 100mM NaCl, and 25mM HEPES, pH 7.4 and clarified by centrifugation at 18,407 x g. Lysates (20µg) were run on a 4–20% SDS-PAGE gel. Membrane was blocked with 5% milk then incubated with anti-GPR156 (1:200) and immunoblots developed with HRP conjugated secondary (sheep anti-rabbit) antibody (1:1,000). Comparable loading was checked by stripping and reprobing the blots with anti-GAPDH (1:500) antibodies (Santa Cruz Biotechnology, Heidelberg, Germany).

GPR156: Whole mount immunostaining of GPR156 in mouse inner ears

All the animal work was approved by the University of Maryland, Baltimore Institutional Animal Care and use Committee (IACUC 420002). Inner ears were dissected from C57BL/6J mice with a postnatal age of 10 days and fixed in 4% paraformaldehyde (PFA) in phosphate buffered saline (PBS) overnight. For

whole mount immunostaining, the cochleae were micro-dissected and were subjected to blocking for 1 hour with 10% normal goat serum in PBS containing 0.25% tritonX100, followed by overnight incubation at 4°C with anti-GPR156 antibodies (1:200; Cat#PA5-23857; Thermo Fisher) in 3% normal goat serum with PBS. F-Actin was decorated using Phalloidin (1:300). Confocal images were acquired from Zeiss LSM710 confocal microscope and images were processed using ImageJ 1.53t software.

Data Availability

Genetic and phenotypic data for the 100KGP study participants are available through the Genomics England Research Environment via application at https://www.genomicsengland.co.uk/join-a-gecip-domain. PanelApp gene panels and evidence of associations were obtained using the PanelApp application programming interface (https://panelapp.genomicsengland.co.uk/api/docs/) on the 20th October 2021. CADD v1.5 (https://cadd.gs.washington.edu/), gnomAD v3.0 (https://cadd.gs.washington.edu/) and Ensembl v104 (http://may2021.archive.ensembl.org/index.html) were used for variant annotation.

Code Availability

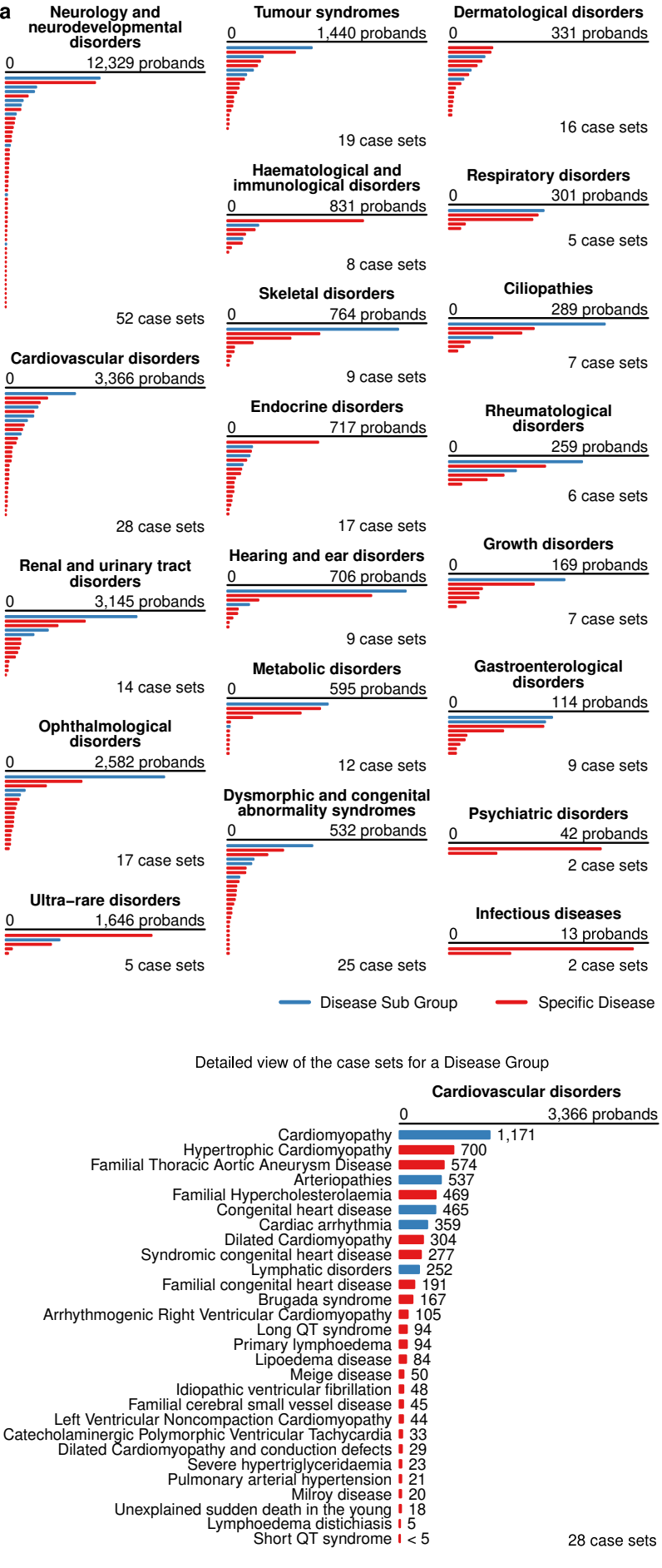
The rsvr tool and Rareservoir schema are available from https://github.com/turrogroup/rsvr.

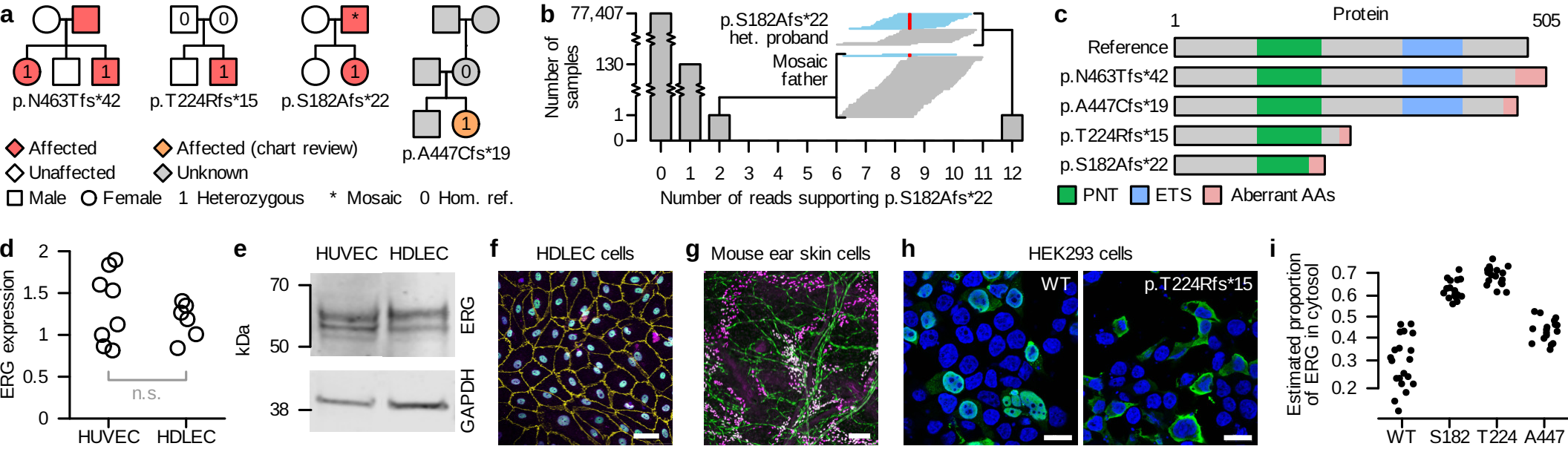
Methods-only references

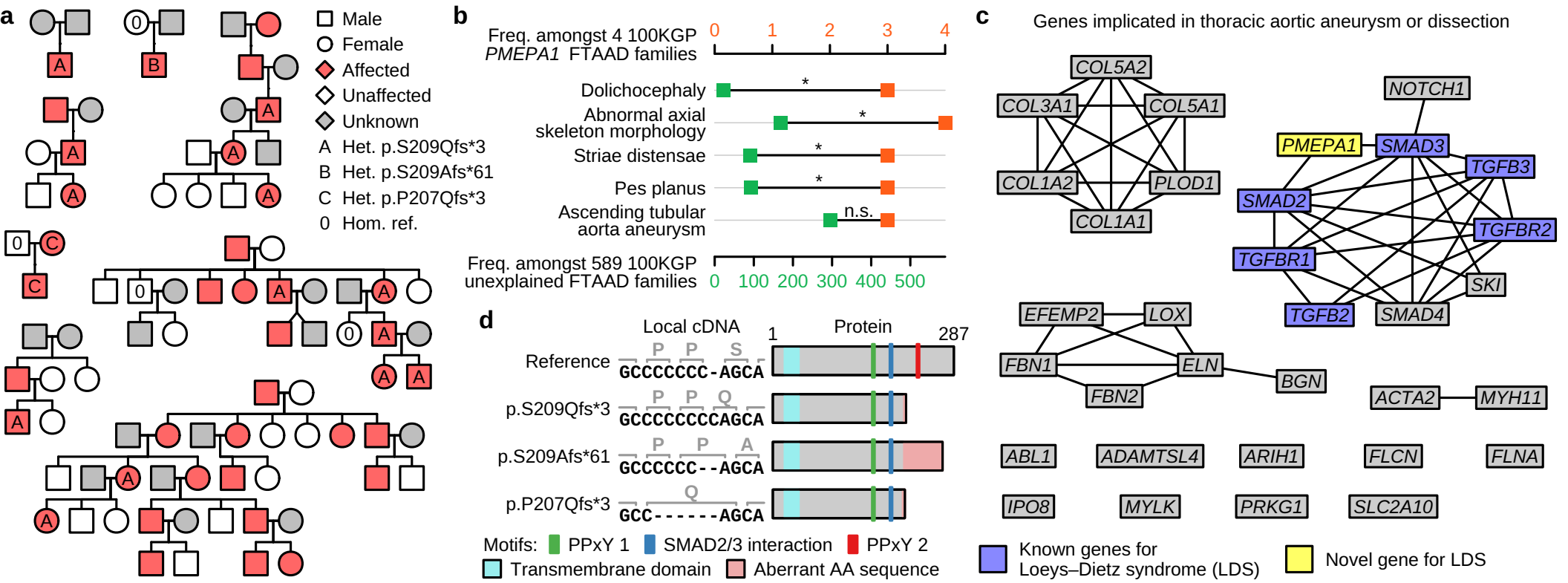
47. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078--2079.

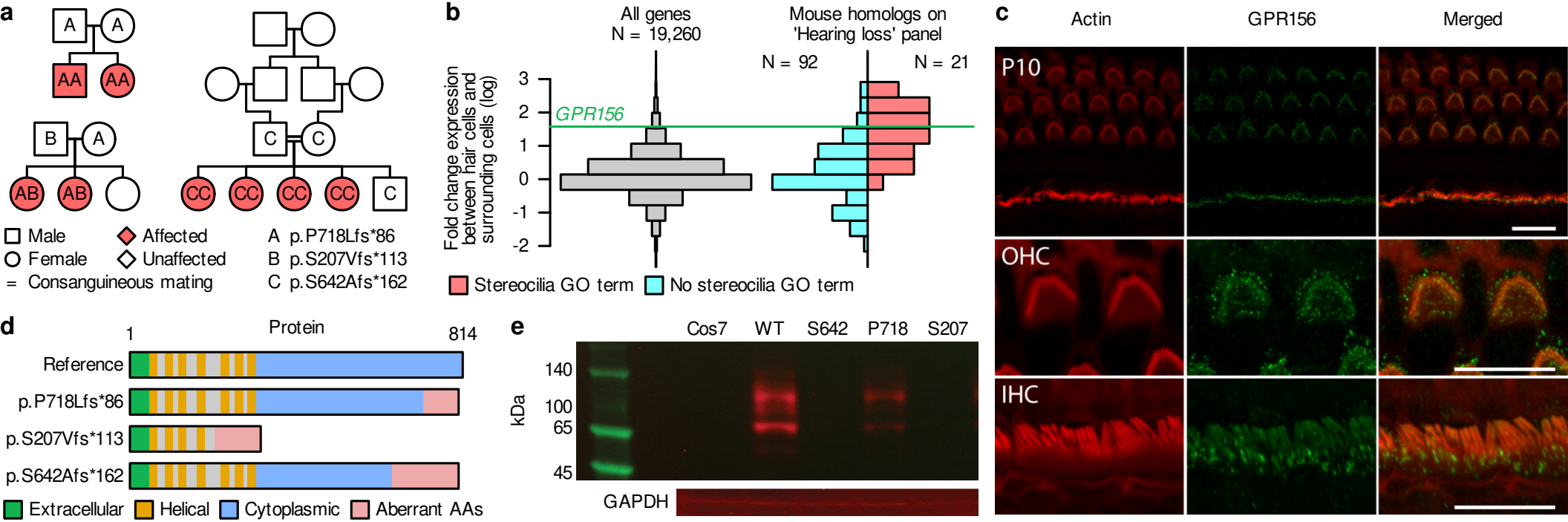
48. Morales J et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022; 604(7905):310--315.

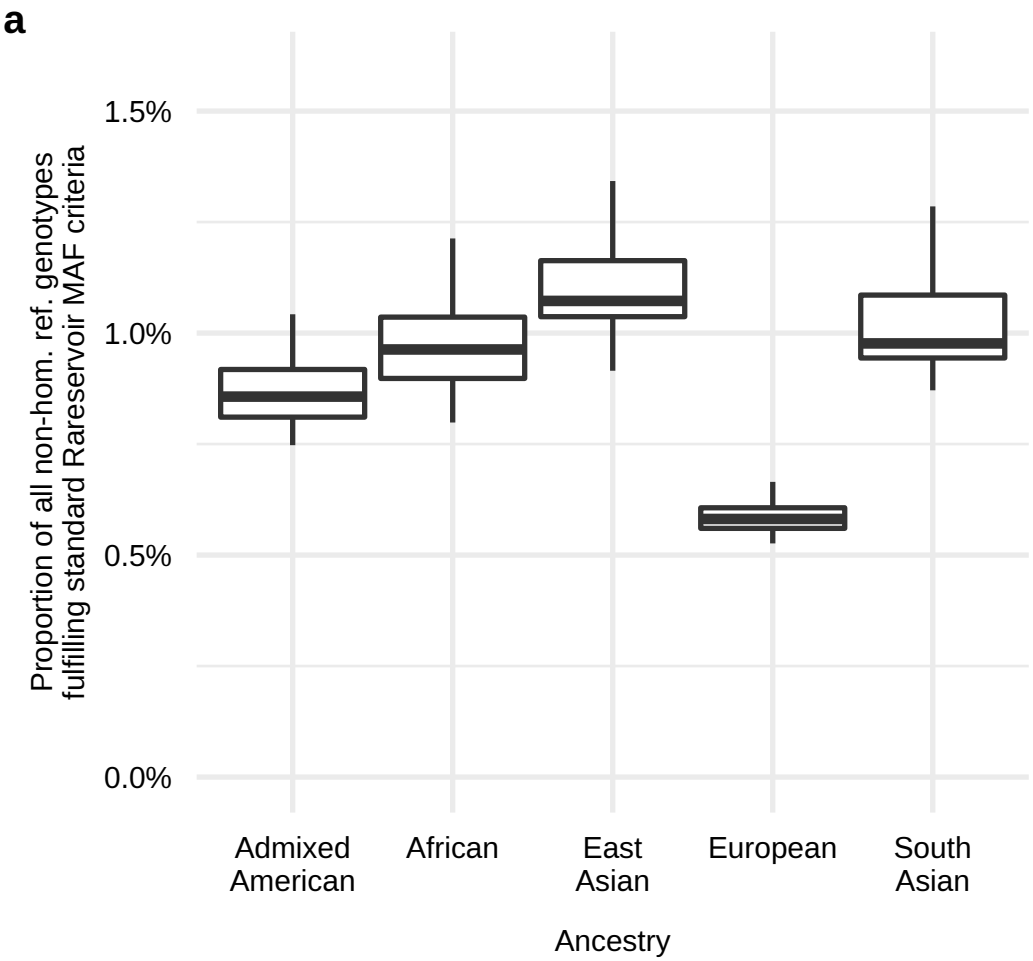
49. Variant QC for 100,000 Genomes Project merged VCF files (2022). https://research-help.genomicsengland.co.uk/display/GERE/Site+QC%2C+FILTER+and+INFO+Fields

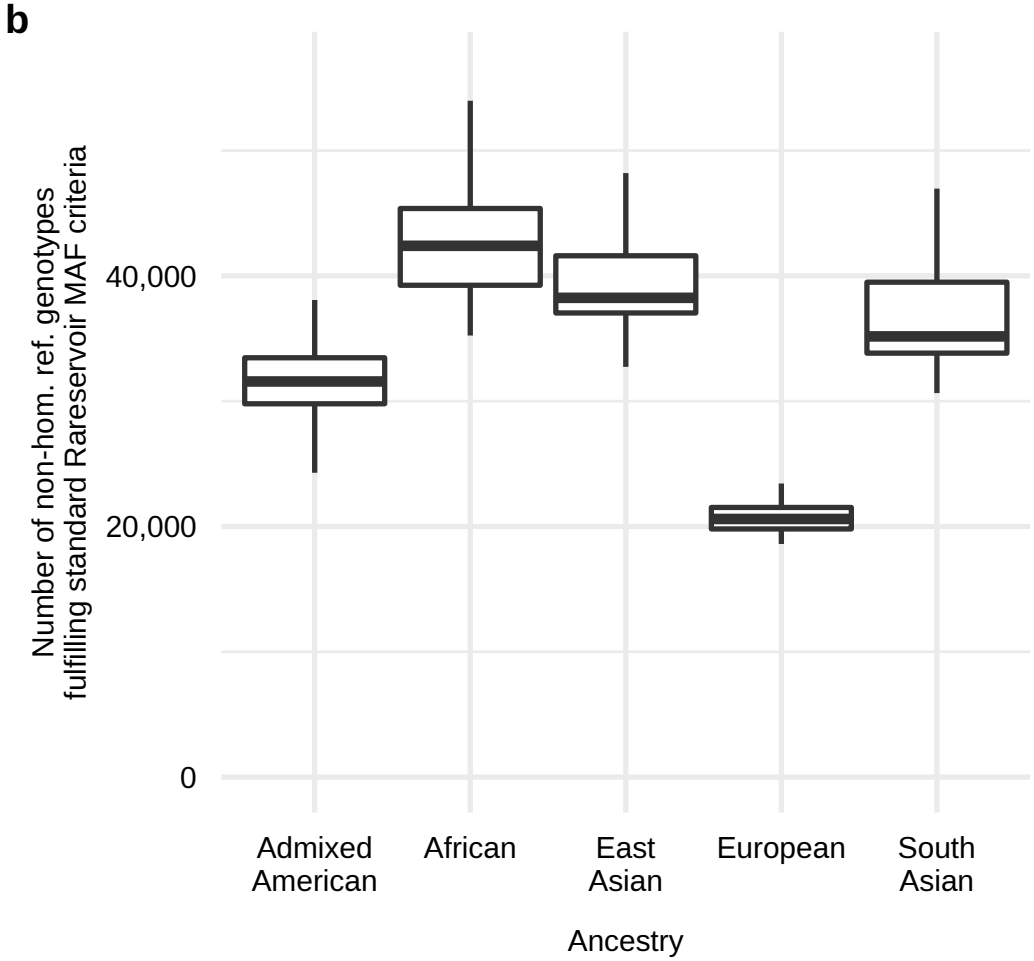

a
Neurology and neurodevelopmental disorders
0
12,329 probands
52 case sets
Cardiovascular disorders
0
3,366 probands
28 case sets
Renal and urinary tract disorders
0
3,145 probands
14 case sets
Ophthalmological disorders
0
2,582 probands
17 case sets
Ultra−rare disorders
0
1,646 probands
5 case sets
Tumour syndromes
0
1,440 probands
19 case sets
Haematological and immunological disorders
0
831 probands
8 case sets
Skeletal disorders
0
764 probands
9 case sets
Endocrine disorders
0
717 probands
17 case sets
Hearing and ear disorders
0
706 probands
9 case sets
Metabolic disorders
0
595 probands
12 case sets
Dysmorphic and congenital abnormality syndromes
0
532 probands
25 case sets
Dermatological disorders
0
331 probands
16 case sets
Respiratory disorders
0
301 probands
5 case sets
Ciliopathies
0
289 probands
7 case sets
Rheumatological disorders
0
259 probands
6 case sets
Growth disorders
0
169 probands
7 case sets
Gastroenterological disorders
0
114 probands
9 case sets
Psychiatric disorders
0
42 probands
2 case sets
Infectious diseases
0
13 probands
2 case sets
Disease Sub Group
Specific Disease
Detailed view of the case sets for a Disease Group
Cardiovascular disorders
0
3,366 probands
Cardiomyopathy 1,171
Hypertrophic Cardiomyopathy 700
Familial Thoracic Aortic Aneurysm Disease 574
Arteriopathies 537
Familial Hypercholesterolaemia 469
Congenital heart disease 465
Cardiac arrhythmia 359
Dilated Cardiomyopathy 304
Syndromic congenital heart disease 277
Lymphatic disorders 252
Familial congenital heart disease 191
Brugada syndrome 167
Arrhythmogenic Right Ventricular Cardiomyopathy 105
Long QT syndrome 94
Primary lymphoedema 94
Lipoedema disease 84
Meige disease 50
Idiopathic ventricular fibrillation 48
Familial cerebral small vessel disease 45
Left Ventricular Noncompaction Cardiomyopathy 44
Catecholaminergic Polymorphic Ventricular Tachycardia 33
Dilated Cardiomyopathy and conduction defects 29
Severe hypertriglyceridaemia 23
Pulmonary arterial hypertension 21
Milroy disease 20
Unexplained sudden death in the young 18
Lymphoedema distichiasis 5
Short QT syndrome < 5
28 case sets
b
Neurology and neurodevelopmental disorders
STUB1 CTNNB1 KMT2A POGZ TTN ANKRD11 MED13L KDM5B SPAST WAC SON GBA ARID1B KDM5C SHANK3 PURA DDX3X TUBA1A SCN1A BCL11A SPG7 SETD5 MYOT ZC4H2 ADNP MECP2 LRRC7 SACS CAPN3 PRKCG UCHL1 CHD8 MYT1L PIEZO2 SYNE1 GMPPB NKX2−1* DYRK1A AHDC1 CHD2 UBAP1 SELENON NEB KAT6A TSC2 PRKN TRIP12 KCND3 FOXP1 SMCHD1 SLC39A8 MFSD8* MAP1B NAA15 INPP5K BRPF1 ECEL1 ERF GNE SPG11 BCL11B ARPC3 MORC2 MME CYP7B1 GRIN2B SCN2A NDUFV1 SETD1A CACNA1A SCN8A TCAP ANO10 SH3TC2 MFN2 VPS13B RAB3A ACTA1 MICU1 CNOT3 GNB1 PYROXD1 PLA2G6 EHMT1 SOD1 DEPDC5 CACNA1S CNKSR2 AUTS2
Dysmorphic and congenital abnormality syndromes
PPP1CB NUP107 LZTR1 WDR62
Tumour syndromes
NF1 KRT14
Ophthalmological disorders
ABCA4 USH2A OPA1 EYS RHO TYR RP1 CNGA3 PRPH2 CACNA1F CYP1B1 OCA2 CRB1 PRPF31 PDE6A NYX PCARE GPR143 BCOR TULP1 LCA5 NR2E3 ACO2 WFS1 CERKL PDE6B RPE65 BEST1 PROM1 COL18A1 FZD4 FRMD7 CABP4 CEP290 TGFBI PAX6 DRAM2 RPGR KCNV2 CHM RLBP1 CRYBB2 SLC24A5 MYO7A CNGB1 INPP5E GJA3 UCHL1 IFT140 SPATA7 NHS TSPAN12 INSL4 SLC38A8* SSBP1 CDH3 UBAP1L CDH23
Dermatological disorders
TSPEAR DSG1 CTSC EDA ABCA12 CERS3 TUFT1 CYP4F22 PNPLA1 EXPH5 SERPINB3 ALOX12B
Skeletal disorders
COL1A1 COL1A2 COL2A1 IFITM5 PLS3 LRP5 PKDCC
Endocrine disorders
TPO IGF1R CYP27B1* GHRHR OBSL1
Respiratory disorders
ACVRL1 ENG TERT
Cardiovascular disorders
LDLR MYBPC3 TTN MYH7 APOE BAG3 RPL10A ALPK3 ERG CELSR1 TPM1 LOX* DSP TNNT2 FLNC TBX20 NKX2−5 BMPR2 PMEPA1 RAB35 GATA6
Hearing and ear disorders
MYO15A MYO6 SLC26A4 STRC LOXHD1 MYO7A ACTG1 SOX10 OTOA OTOF MITF EYA1 SYNE4 ADGRV1 CDH23 GPR156 FMN1
Renal and urinary tract disorders
PKD1 PKD2 IFT140* DNAJB11 SLC12A3 COL4A3 COL4A4 ATP6V0A4 UMOD PKHD1 SLC7A9 GRHPR COL4A5 USP33 SLC34A3 HBB*
Ciliopathies
DNAI1 CCDC40 DNAH5 DNAH11 CCDC39 ODAD1 RSPH4A MPPE1 ALMS1 CEP290 FOXJ1* DNAAF5
Haematological and immunological disorders
NFKB1 RPS19 SPTA1 WAS* PKLR AICDA
Gastroenterological disorders
ACTG2 SRP9 SEC63 WWOX
Ultra−rare disorders
FAM222B PKMYT1
Metabolic disorders
PHYH PAH
0.95 0.96 0.97 0.98 0.99 1.00
Posterior probability of association
PanelApp evidence
High
Moderate
Low
None
Association with Disease Sub Group
Association with Specific Disease

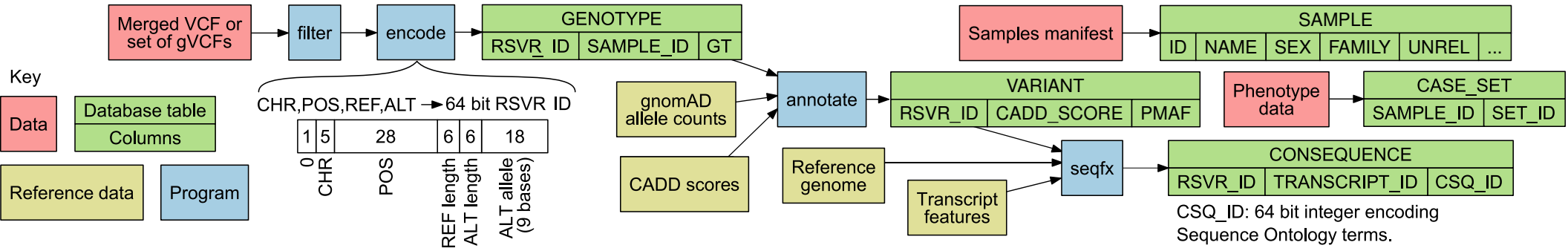

a
p.N463Tfs*42
p.T224Rfs*15
p.S182Afs*22
p.A447Cfs*19
Affected
Affected (chart review)
Unaffected
Unknown
Male
Female
1 Heterozygous
* Mosaic
0 Hom. ref.
b
77,407
130
1
0
Number of samples
p.S182Afs*22 het. proband
Mosaic father
0 1 2 3 4 5 6 7 8 9 10 11 12
Number of reads supporting p.S182Afs*22
c
Protein
1
505
Reference
p.N463Tfs*42
p.A447Cfs*19
p.T224Rfs*15
p.S182Afs*22
PNT
ETS
Aberrant AAs
d
ERG expression
2
1.5
1
0.5
0
n.s.
HUVEC
HDLEC
e
HUVEC
HDLEC
kDa
70
50
38
ERG
GAPDH
f
HDLEC cells
g
Mouse ear skin cells
h
HEK293 cells
WT
p.T224Rfs*15
i
Estimated proportion of ERG in cytosol
0.7
0.6
0.5
0.4
0.3
0.2
WT
S182
T224
A447

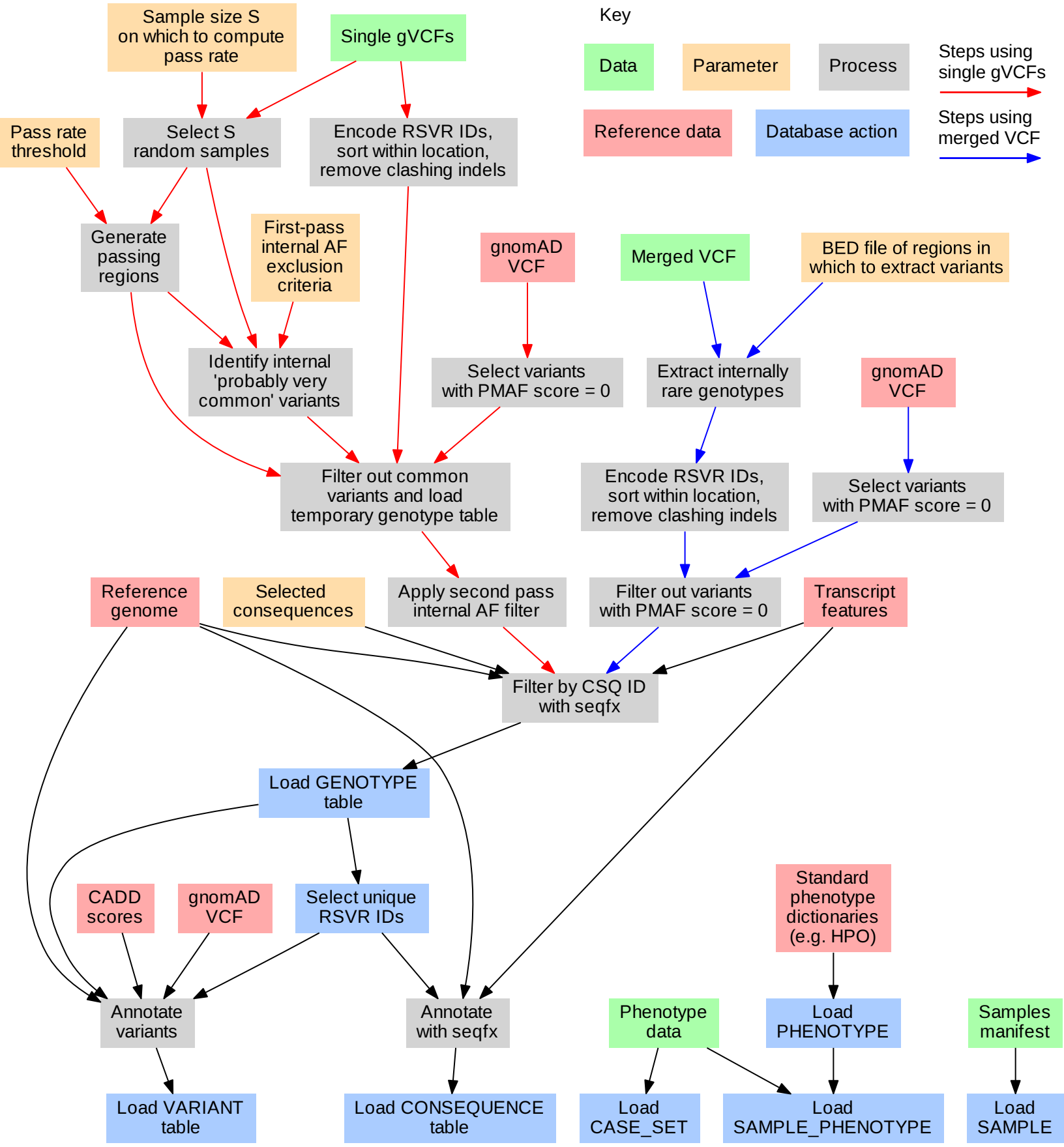

a
Male
Female
Affected
Unaffected
Unknown
A Het. p.S209Qfs*3
B Het. p.S209Afs*61
C Het. p.P207Qfs*3
0 Hom. ref.
b
Freq. amongst 4 100KGP PMEPA1 FTAAD families
0 1 2 3 4
Dolichocephaly
Abnormal axial skeleton morphology
Striae distensae
Pes planus
Ascending tubular aorta aneurysm
n.s.
Freq. amongst 589 100KGP unexplained FTAAD families
0 100 200 300 400 500
d
Local cDNA
Protein
1
287
Reference
GCCCCCCC-AGCA
p.S209Qfs*3
GCCCCCCCCAGCA
p.S209Afs*61
GCCCCCC--AGCA
p.P207Qfs*3
GCC------AGCA
Motifs: PPxY 1 SMAD2/3 interaction PPxY 2
Transmembrane domain Aberrant AA sequence
c
Genes implicated in thoracic aortic aneurysm or dissection
COL5A2
COL3A1
COL5A1
COL1A2
PLOD1
COL1A1
NOTCH1
PMEPA1
SMAD3
TGFB3
SMAD2
TGFBR2
TGFBR1
SKI
TGFB2
SMAD4
EFEMP2
LOX
FBN1
ELN
BGN
FBN2
ACTA2
MYH11
ABL1
ADAMTSL4
ARIH1
FLCN
FLNA
IPO8
MYLK
PRKG1
SLC2A10
Known genes for Loeys–Dietz syndrome (LDS)
Novel gene for LDS

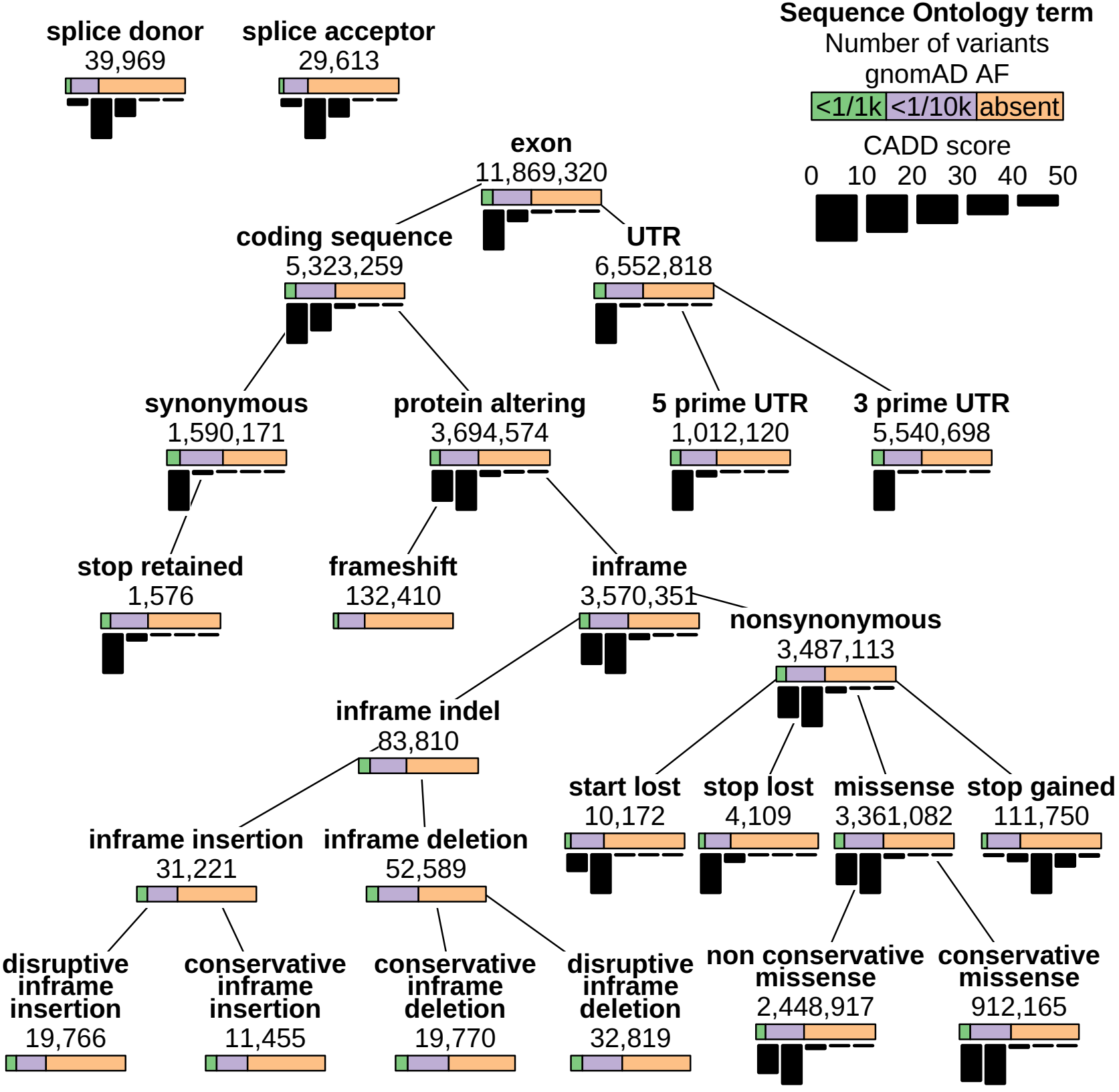

a
A
A
AA
AA
B
A
AB
AB
C
C
CC
CC
CC
CC
C
Male
Female
= Consanguineous mating
Affected
Unaffected
A p.P718Lfs*86
B p.S207Vfs*113
C p.S642Afs*162
b
All genes
N = 19,260
Mouse homologs on
'Hearing loss' panel
N = 92
N = 21
Fold change expression
between hair cells and
surrounding cells (log)
3
2
1
0
-1
-2
GPR156
Stereocilia GO term
No stereocilia GO term
c
Actin
GPR156
Merged
P10
OHC
IHC
d
1
Protein
814
Reference
p.P718Lfs*86
p.S207Vfs*113
p.S642Afs*162
Extracellular
Helical
Cytoplasmic
Aberrant AAs
e
Cos7
WT
S642
P718
S207
kDa
140
100
65
45
GAPDH

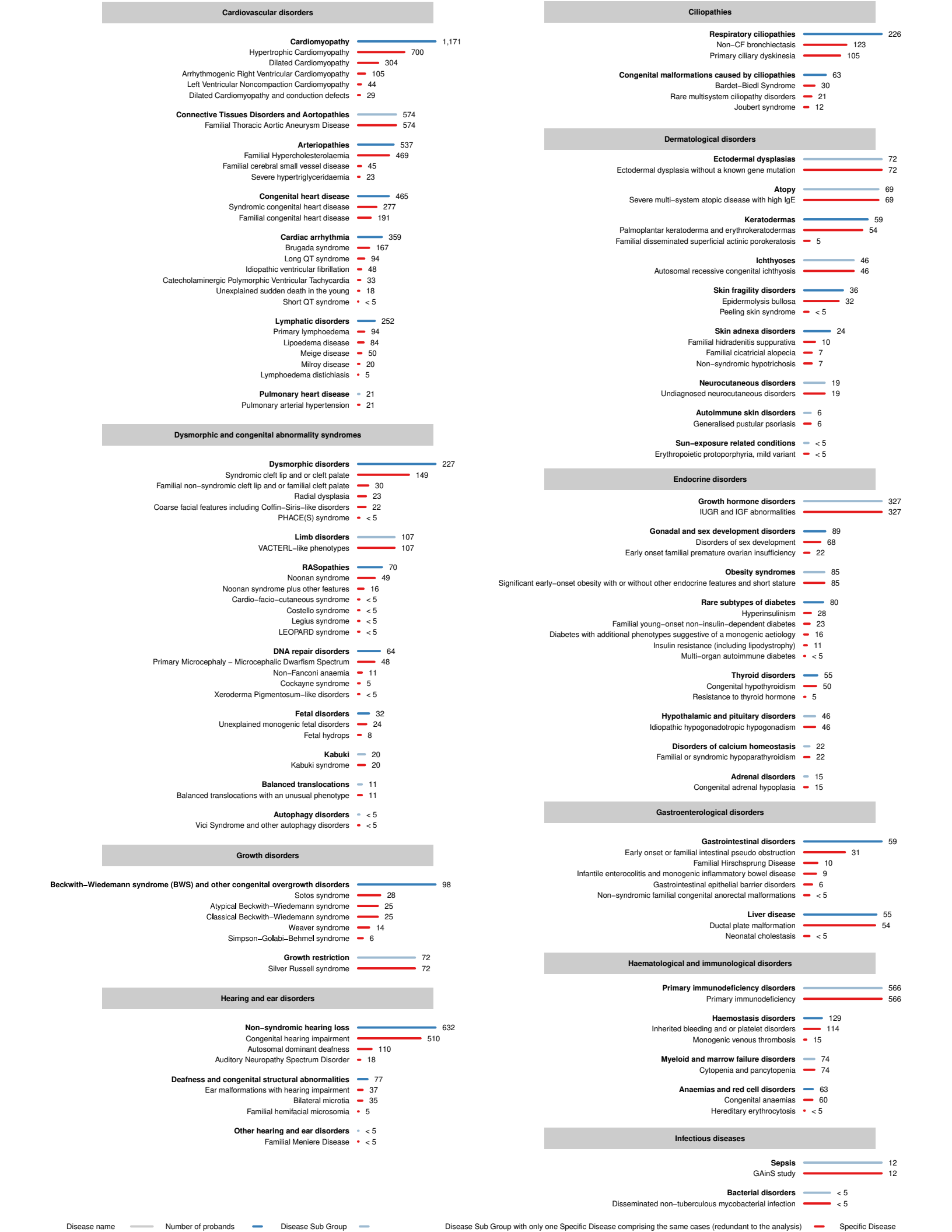

a
1.5%
1.0%
0.5%
0.0%
Proportion of all non-hom. ref. genotypes fulfilling standard Rareservoir MAF criteria
Admixed American
African
East Asian
European
South Asian
Ancestry

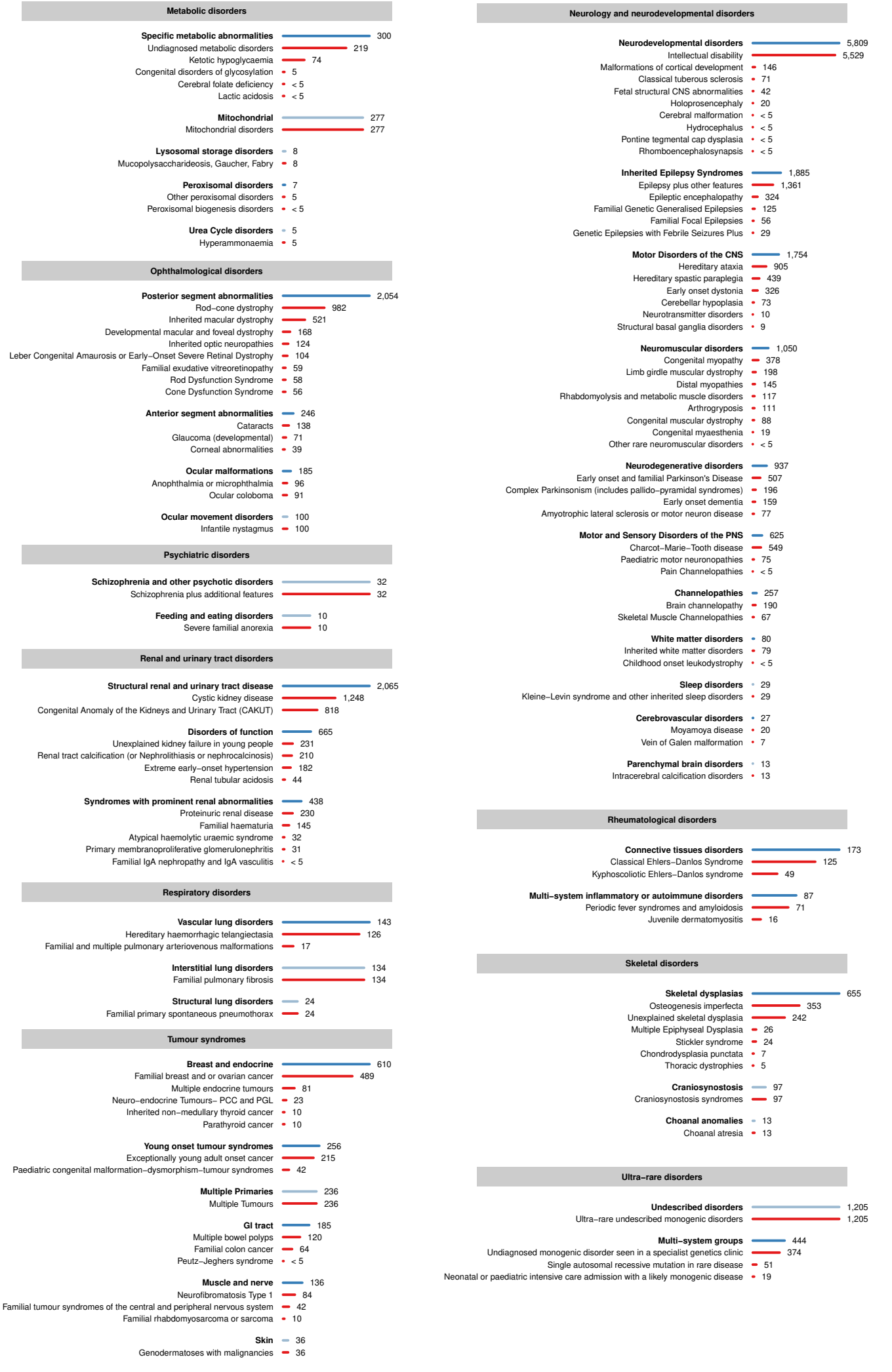

b
40,000
20,000
0
Number of non-hom. ref. genotypes fulfilling standard Rareservoir MAF criteria
Admixed American
African
East Asian
European
South Asian
Ancestry

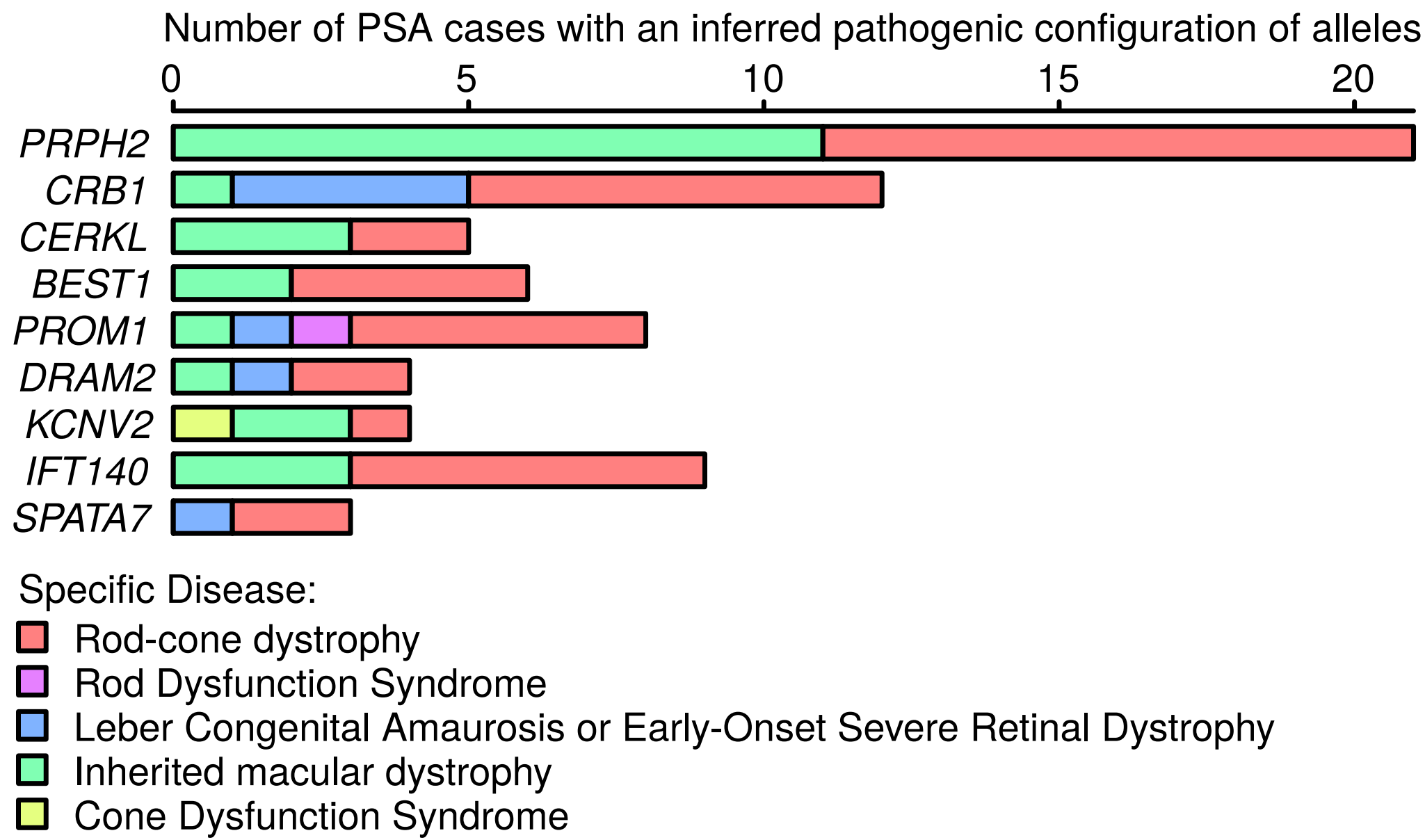

Merged VCF or set of gVCFs
filter
encode
GENOTYPE
RSVR_ID
SAMPLE_ID
GT
Samples manifest
SAMPLE
ID
NAME
SEX
FAMILY
UNREL
...
Key
Data
Database table
Columns
Reference data
Program
CHR,POS,REF,ALT → 64 bit RSVR_ID
1
5
28
6
6
18
0
CHR
POS
REF length
ALT length
ALT allele (9 bases)
gnomAD allele counts
CADD scores
annotate
VARIANT
RSVR_ID
CADD_SCORE
PMAF
Phenotype data
CASE_SET
SAMPLE_ID
SET_ID
Reference genome
Transcript features
seqfx
CONSEQUENCE
RSVR_ID
TRANSCRIPT_ID
CSQ_ID
CSQ_ID: 64 bit integer encoding Sequence Ontology terms.

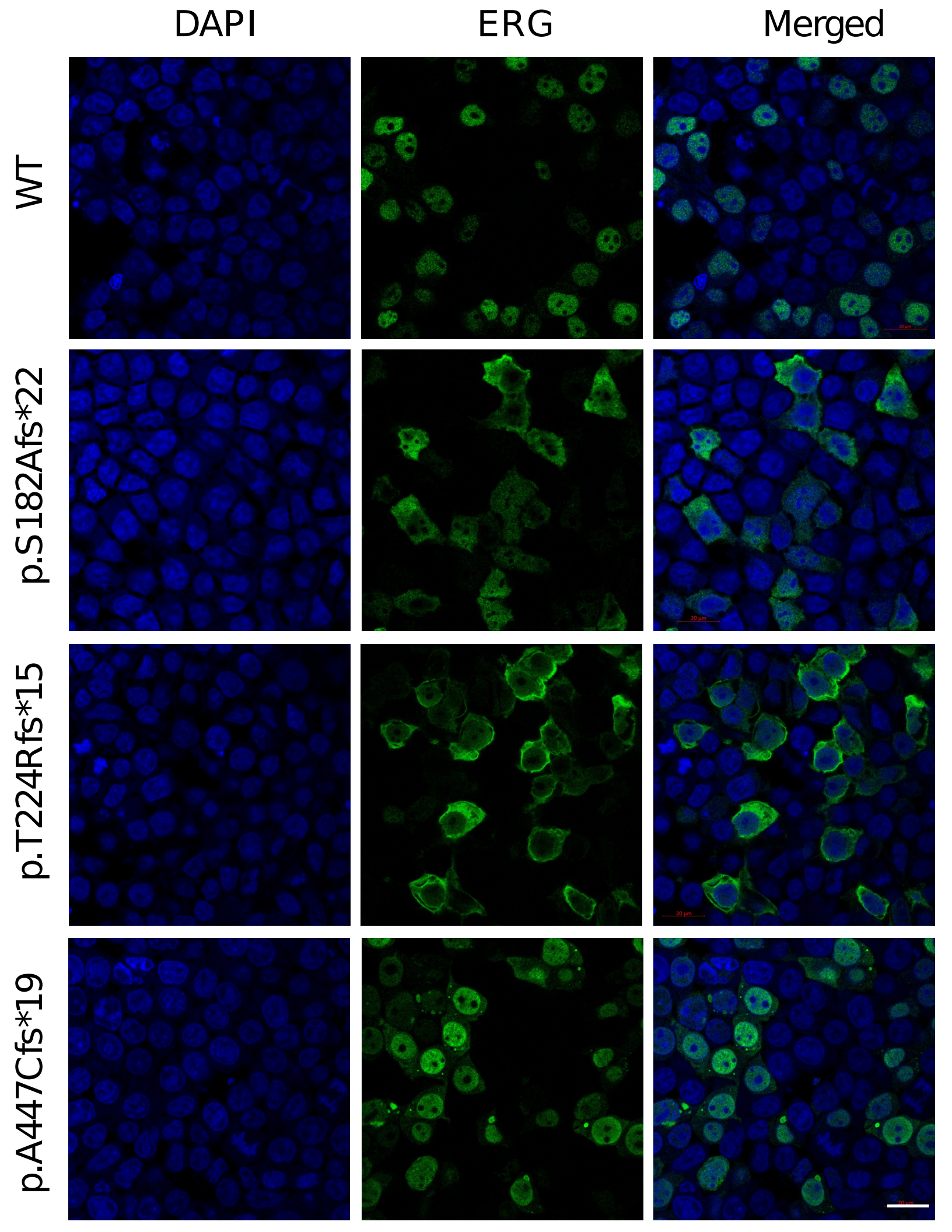

Key
Data
Parameter
Process
Steps using single gVCFs
Reference data
Database action
Steps using merged VCF
Sample size S on which to compute pass rate
Single gVCFs
Pass rate threshold
Select S random samples
Encode RSVR IDs, sort within location, remove clashing indels
Generate passing regions
First-pass internal AF exclusion criteria
gnomAD VCF
Merged VCF
BED file of regions in which to extract variants
Identify internal 'probably very common' variants
Select variants with PMAF score = 0
Extract internally rare genotypes
gnomAD VCF
Filter out common variants and load temporary genotype table
Encode RSVR IDs, sort within location, remove clashing indels
Select variants with PMAF score = 0
Reference genome
Selected consequences
Apply second pass internal AF filter
Filter out variants with PMAF score = 0
Transcript features
Filter by CSQ ID with seqfx
Load GENOTYPE table
CADD scores
gnomAD VCF
Select unique RSVR IDs
Standard phenotype dictionaries (e.g. HPO)
Annotate variants
Annotate with seqfx
Phenotype data
Load PHENOTYPE
Samples manifest
Load VARIANT table
Load CONSEQUENCE table
Load CASE_SET
Load SAMPLE_PHENOTYPE
Load SAMPLE


Sequence Ontology term
Number of variants
gnomAD AF
<1/1k
<1/10k
absent
CADD score
0 10 20 30 40 50
splice donor
39,969
splice acceptor
29,613
exon
11,869,320
coding sequence
5,323,259
UTR
6,552,818
synonymous
1,590,171
protein altering
3,694,574
5 prime UTR
1,012,120
3 prime UTR
5,540,698
stop retained
1,576
frameshift
132,410
inframe
3,570,351
nonsynonymous
3,487,113
inframe indel
83,810
start lost
10,172
stop lost
4,109
missense
3,361,082
stop gained
111,750
inframe insertion
31,221
inframe deletion
52,589
disruptive inframe insertion
19,766
conservative inframe insertion
11,455
conservative inframe deletion
19,770
disruptive inframe deletion
32,819
non conservative missense
2,448,917
conservative missense
912,165


Cardiovascular disorders
Cardiomyopathy 1,171
Hypertrophic Cardiomyopathy 700
Dilated Cardiomyopathy 304
Arrhythmogenic Right Ventricular Cardiomyopathy 105
Left Ventricular Noncompaction Cardiomyopathy 44
Dilated Cardiomyopathy and conduction defects 29
Connective Tissues Disorders and Aortopathies 574
Familial Thoracic Aortic Aneurysm Disease 574
Arteriopathies 537
Familial Hypercholesterolaemia 469
Familial cerebral small vessel disease 45
Severe hypertriglyceridaemia 23
Congenital heart disease 465
Syndromic congenital heart disease 277
Familial congenital heart disease 191
Cardiac arrhythmia 359
Brugada syndrome 167
Long QT syndrome 94
Idiopathic ventricular fibrillation 48
Catecholaminergic Polymorphic Ventricular Tachycardia 33
Unexplained sudden death in the young 18
Short QT syndrome < 5
Lymphatic disorders 252
Primary lymphoedema 94
Lipoedema disease 84
Meige disease 50
Milroy disease 20
Lymphoedema distichiasis 5
Pulmonary heart disease 21
Pulmonary arterial hypertension 21
Dysmorphic and congenital abnormality syndromes
Dysmorphic disorders 227
Syndromic cleft lip and or cleft palate 149
Familial non–syndromic cleft lip and or familial cleft palate 30
Radial dysplasia 23
Coarse facial features including Coffin–Siris–like disorders 22
PHACE(S) syndrome < 5
Limb disorders 107
VACTERL–like phenotypes 107
RASopathies 70
Noonan syndrome 49
Noonan syndrome plus other features 16
Cardio–facio–cutaneous syndrome < 5
Costello syndrome < 5
Legius syndrome < 5
LEOPARD syndrome < 5
DNA repair disorders 64
Primary Microcephaly – Microcephalic Dwarfism Spectrum 48
Non–Fanconi anaemia 11
Cockayne syndrome 5
Xeroderma Pigmentosum–like disorders < 5
Fetal disorders 32
Unexplained monogenic fetal disorders 24
Fetal hydrops 8
Kabuki 20
Kabuki syndrome 20
Balanced translocations 11
Balanced translocations with an unusual phenotype 11
Autophagy disorders < 5
Vici Syndrome and other autophagy disorders < 5
Growth disorders
Beckwith–Wiedemann syndrome (BWS) and other congenital overgrowth disorders 98
Sotos syndrome 28
Atypical Beckwith–Wiedemann syndrome 25
Classical Beckwith–Wiedemann syndrome 25
Weaver syndrome 14
Simpson–Golabi–Behmel syndrome 6
Growth restriction 72
Silver Russell syndrome 72
Hearing and ear disorders
Non–syndromic hearing loss 632
Congenital hearing impairment 510
Autosomal dominant deafness 110
Auditory Neuropathy Spectrum Disorder 18
Deafness and congenital structural abnormalities 77
Ear malformations with hearing impairment 37
Bilateral microtia 35
Familial hemifacial microsomia 5
Other hearing and ear disorders < 5
Familial Meniere Disease < 5
Ciliopathies
Respiratory ciliopathies 226
Non–CF bronchiectasis 123
Primary ciliary dyskinesia 105
Congenital malformations caused by ciliopathies 63
Bardet–Biedl Syndrome 30
Rare multisystem ciliopathy disorders 21
Joubert syndrome 12
Dermatological disorders
Ectodermal dysplasias 72
Ectodermal dysplasia without a known gene mutation 72
Atopy 69
Severe multi–system atopic disease with high IgE 69
Keratodermas 59
Palmoplantar keratoderma and erythrokeratodermas 54
Familial disseminated superficial actinic porokeratosis 5
Ichthyoses 46
Autosomal recessive congenital ichthyosis 46
Skin fragility disorders 36
Epidermolysis bullosa 32
Peeling skin syndrome < 5
Skin adnexa disorders 24
Familial hidradenitis suppurativa 10
Familial cicatricial alopecia 7
Non–syndromic hypotrichosis 7
Neurocutaneous disorders 19
Undiagnosed neurocutaneous disorders 19
Autoimmune skin disorders 6
Generalised pustular psoriasis 6
Sun–exposure related conditions < 5
Erythropoietic protoporphyria, mild variant < 5
Endocrine disorders
Growth hormone disorders 327
IUGR and IGF abnormalities 327
Gonadal and sex development disorders 89
Disorders of sex development 68
Early onset familial premature ovarian insufficiency 22
Obesity syndromes 85
Significant early–onset obesity with or without other endocrine features and short stature 85
Rare subtypes of diabetes 80
Hyperinsulinism 28
Familial young–onset non–insulin–dependent diabetes 23
Diabetes with additional phenotypes suggestive of a monogenic aetiology 16
Insulin resistance (including lipodystrophy) 11
Multi–organ autoimmune diabetes < 5
Thyroid disorders 55
Congenital hypothyroidism 50
Resistance to thyroid hormone 5
Hypothalamic and pituitary disorders 46
Idiopathic hypogonadotropic hypogonadism 46
Disorders of calcium homeostasis 22
Familial or syndromic hypoparathyroidism 22
Adrenal disorders 15
Congenital adrenal hypoplasia 15
Gastroenterological disorders
Gastrointestinal disorders 59
Early onset or familial intestinal pseudo obstruction 31
Familial Hirschsprung Disease 10
Infantile enterocolitis and monogenic inflammatory bowel disease 9
Gastrointestinal epithelial barrier disorders 6
Non–syndromic familial congenital anorectal malformations < 5
Liver disease 55
Ductal plate malformation 54
Neonatal cholestasis < 5
Haematological and immunological disorders
Primary immunodeficiency disorders 566
Primary immunodeficiency 566
Haemostasis disorders 129
Inherited bleeding and or platelet disorders 114
Monogenic venous thrombosis 15
Myeloid and marrow failure disorders 74
Cytopenia and pancytopenia 74
Anaemias and red cell disorders 63
Congenital anaemias 60
Hereditary erythrocytosis < 5
Infectious diseases
Sepsis 12
GAinS study 12
Bacterial disorders < 5
Disseminated non–tuberculous mycobacterial infection < 5
Disease name
Number of probands
Disease Sub Group
Disease Sub Group with only one Specific Disease comprising the same cases (redundant to the analysis)
Specific Disease


Metabolic disorders
Specific metabolic abnormalities 300
Undiagnosed metabolic disorders 219
Ketotic hypoglycaemia 74
Congenital disorders of glycosylation 5
Cerebral folate deficiency < 5
Lactic acidosis < 5
Mitochondrial 277
Mitochondrial disorders 277
Lysosomal storage disorders 8
Mucopolysaccharideosis, Gaucher, Fabry 8
Peroxisomal disorders 7
Other peroxisomal disorders 5
Peroxisomal biogenesis disorders < 5
Urea Cycle disorders 5
Hyperammonaemia 5
Ophthalmological disorders
Posterior segment abnormalities 2,054
Rod–cone dystrophy 982
Inherited macular dystrophy 521
Developmental macular and foveal dystrophy 168
Inherited optic neuropathies 124
Leber Congenital Amaurosis or Early–Onset Severe Retinal Dystrophy 104
Familial exudative vitreoretinopathy 59
Rod Dysfunction Syndrome 58
Cone Dysfunction Syndrome 56
Anterior segment abnormalities 246
Cataracts 138
Glaucoma (developmental) 71
Corneal abnormalities 39
Ocular malformations 185
Anophthalmia or microphthalmia 96
Ocular coloboma 91
Ocular movement disorders 100
Infantile nystagmus 100
Psychiatric disorders
Schizophrenia and other psychotic disorders 32
Schizophrenia plus additional features 32
Feeding and eating disorders 10
Severe familial anorexia 10
Renal and urinary tract disorders
Structural renal and urinary tract disease 2,065
Cystic kidney disease 1,248
Congenital Anomaly of the Kidneys and Urinary Tract (CAKUT) 818
Disorders of function 665
Unexplained kidney failure in young people 231
Renal tract calcification (or Nephrolithiasis or nephrocalcinosis) 210
Extreme early–onset hypertension 182
Renal tubular acidosis 44
Syndromes with prominent renal abnormalities 438
Proteinuric renal disease 230
Familial haematuria 145
Atypical haemolytic uraemic syndrome 32
Primary membranoproliferative glomerulonephritis 31
Familial IgA nephropathy and IgA vasculitis < 5
Respiratory disorders
Vascular lung disorders 143
Hereditary haemorrhagic telangiectasia 126
Familial and multiple pulmonary arteriovenous malformations 17
Interstitial lung disorders 134
Familial pulmonary fibrosis 134
Structural lung disorders 24
Familial primary spontaneous pneumothorax 24
Tumour syndromes
Breast and endocrine 610
Familial breast and or ovarian cancer 489
Multiple endocrine tumours 81
Neuro–endocrine Tumours– PCC and PGL 23
Inherited non–medullary thyroid cancer 10
Parathyroid cancer 10
Young onset tumour syndromes 256
Exceptionally young adult onset cancer 215
Paediatric congenital malformation–dysmorphism–tumour syndromes 42
Multiple Primaries 236
Multiple Tumours 236
GI tract 185
Multiple bowel polyps 120
Familial colon cancer 64
Peutz–Jeghers syndrome < 5
Muscle and nerve 136
Neurofibromatosis Type 1 84
Familial tumour syndromes of the central and peripheral nervous system 42
Familial rhabdomyosarcoma or sarcoma 10
Skin 36
Genodermatoses with malignancies 36
Neurology and neurodevelopmental disorders
Neurodevelopmental disorders 5,809
Intellectual disability 5,529
Malformations of cortical development 146
Classical tuberous sclerosis 71
Fetal structural CNS abnormalities 42
Holoprosencephaly 20
Cerebral malformation < 5
Hydrocephalus < 5
Pontine tegmental cap dysplasia < 5
Rhomboencephalosynapsis < 5
Inherited Epilepsy Syndromes 1,885
Epilepsy plus other features 1,361
Epileptic encephalopathy 324
Familial Genetic Generalised Epilepsies 125
Familial Focal Epilepsies 56
Genetic Epilepsies with Febrile Seizures Plus 29
Motor Disorders of the CNS 1,754
Hereditary ataxia 905
Hereditary spastic paraplegia 439
Early onset dystonia 326
Cerebellar hypoplasia 73
Neurotransmitter disorders 10
Structural basal ganglia disorders 9
Neuromuscular disorders 1,050
Congenital myopathy 378
Limb girdle muscular dystrophy 198
Distal myopathies 145
Rhabdomyolysis and metabolic muscle disorders 117
Arthrogryposis 111
Congenital muscular dystrophy 88
Congenital myaesthenia 19
Other rare neuromuscular disorders < 5
Neurodegenerative disorders 937
Early onset and familial Parkinson's Disease 507
Complex Parkinsonism (includes pallido–pyramidal syndromes) 196
Early onset dementia 159
Amyotrophic lateral sclerosis or motor neuron disease 77
Motor and Sensory Disorders of the PNS 625
Charcot–Marie–Tooth disease 549
Paediatric motor neuronopathies 75
Pain Channelopathies < 5
Channelopathies 257
Brain channelopathy 190
Skeletal Muscle Channelopathies 67
White matter disorders 80
Inherited white matter disorders 79
Childhood onset leukodystrophy < 5
Sleep disorders 29
Kleine–Levin syndrome and other inherited sleep disorders 29
Cerebrovascular disorders 27
Moyamoya disease 20
Vein of Galen malformation 7
Parenchymal brain disorders 13
Intracerebral calcification disorders 13
Rheumatological disorders
Connective tissues disorders 173
Classical Ehlers–Danlos Syndrome 125
Kyphoscoliotic Ehlers–Danlos syndrome 49
Multi–system inflammatory or autoimmune disorders 87
Periodic fever syndromes and amyloidosis 71
Juvenile dermatomyositis 16
Skeletal disorders
Skeletal dysplasias 655
Osteogenesis imperfecta 353
Unexplained skeletal dysplasia 242
Multiple Epiphyseal Dysplasia 26
Stickler syndrome 24
Chondrodysplasia punctata 7
Thoracic dystrophies 5
Craniosynostosis 97
Craniosynostosis syndromes 97
Choanal anomalies 13
Choanal atresia 13
Ultra–rare disorders
Undescribed disorders 1,205
Ultra–rare undescribed monogenic disorders 1,205
Multi–system groups 444
Undiagnosed monogenic disorder seen in a specialist genetics clinic 374
Single autosomal recessive mutation in rare disease 51
Neonatal or paediatric intensive care admission with a likely monogenic disease 19
Disease name
Number of probands
Disease Sub Group
Disease Sub Group with only one Specific Disease comprising the same cases (redundant to the analysis)
Specific Disease


Number of PSA cases with an inferred pathogenic configuration of alleles
0
5
10
15
20
PRPH2
CRB1
CERKL
BEST1
PROM1
DRAM2
KCNV2
IFT140
SPATA7
Specific Disease:
Rod-cone dystrophy
Rod Dysfunction Syndrome
Leber Congenital Amaurosis or Early-Onset Severe Retinal Dystrophy
Inherited macular dystrophy
Cone Dysfunction Syndrome


DAPI
ERG
Merged
WT
p.S182Afs*22
p.T224Rfs*15
p.A447Cfs*19

Right ear
Left ear
0
25
50
75
100
Hearing threshold (dB)
6 year old
8 year old
300
1000
3000
Frequency (Hz)
Type
Air
Bone