Genetic association analysis of 269 rare diseases reveals novel aetiologies

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

[1]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, 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]Molecular and Clinical Sciences Institute, St George's University of London, London, UK. [5]Department of Medical Genomics, Centre for Genomics Medicine, King Faisal Specialist Hospital & Research Centre, Riyadh, Saudi Arabia. [6]Department of Otorhinolaryngology Head and Neck Surgery, School of Medicine, University of Maryland Baltimore, Baltimore, MD, USA. [7]Department of Clinical Genomics, Centre for Genomics Medicine, King Faisal Specialist Hospital & Research Centre, Riyadh, Saudi Arabia. [8]Department of Bioscience and Genetics, National Cerebral and Cardiovascular Center, Osaka, Japan. [9]Center for Medical Genetics, Antwerp University Hospital/University of Antwerp, Antwerp, Belgium. [10]Department of Cardiovascular Sciences, Center for Molecular and Vascular Biology, KU Leuven, Leuven, Belgium. [11]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [12]Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [13]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Trust, International Centre for Life, Newcastle upon Tyne, UK. [14]Department of Clinical Genetics, Chapel Allerton Hospital, Leeds, UK. [15]Center for Medical Genetics, CHU Liege, Belgium. [16]Department of Medical Genetics, Shinshu University School of Medicine, Matsumoto, Japan. [17]SW Thames Regional Genetics Service, St George's University Hospitals NHS Foundation Trust, London, UK. [18]Department of Medical Genetics, Kawasaki Medical School Hospital, Okayama, Japan. [19]Okinawa Chubu Hospital, Okinawa, Japan. [20]Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [21]ENT Medical Centre, Riyadh, Saudi Arabia. [22]Peninsula Clinical Genetics Service, Royal Devon & Exeter Hospital, Exeter, UK. [23]Department of Biochemistry and Molecular Biology, School of Medicine, University of Maryland, Baltimore, MD, USA. [24]Division of Molecular Pathology and Department of Internal Medicine, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [25]Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands. [26]Department of Medical Genetics, Sakakibara Heart Institute, Tokyo, Japan. [27]School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK. [28]University Hospitals Bristol NHS Foundation Trust, Bristol, UK.

The genetic aetiologies of more than half of rare diseases remain unknown[1]. Standardised genome sequencing (GS) and phenotyping of large patient cohorts provides an opportunity for discovering the unknown aetiologies[2], but this depends on efficient and powerful analytical methods[3]. We have developed a portable computational and statistical framework for inferring genetic associations with rare diseases. At its core lies the 'Rareservoir', a compact database of rare variant genotypes and phenotypes. We built a Rareservoir of 77,539 genomes sequenced by the 100,000 Genomes Project (100KGP)[4]. We then applied the Bayesian association method, BeviMed[3], across 269 rare diseases assigned to participants in the project, identifying 238 known[5] and 21 novel associations. The novel results included three which we selected for validation. We provide compelling 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[6], and (3) loss-of-function mutations in *GPR156* gives rise to recessive congenital hearing impairment. These novel findings confirm the power of our analytical approach for the aetiological discovery of rare diseases.

Collectively, rare diseases affect 1 in 20 people[7], but fewer than half of the approximately 10,000 catalogued rare diseases have a resolved genetic aetiology[1]. Standardised GS of large, phenotypically diverse collections of rare disease patients enables aetiological discovery across a wide range of pathologies[2]. However, the scale and complexity of large GS datasets and the hierarchical nature of patient phenotype coding[8] induces 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 variant call format (VCF) files many terabytes in size, leading to high storage and processing costs. Recently developed frameworks such as Hail[9] and OpenCGA[10] 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.

The 'Rareservoir'

We developed a compact and portable relational database schema, the 'Rareservoir', which enables efficient statistical analysis of rare variant genotypes and clinical phenotypes. A Rareservoir stores genotypes sparsely and for rare variants only—by default, those for which all population-specific MAFs are likely to be <0.1%. This greatly reduces the number of stored genotypes in a large study (**Extended Data Fig. 1**), allowing the data to reside in a relational database on centralised storage. Variants in this MAF stratum encompass the vast majority of those having a large effect on rare disease risk. The Rareservoir encodes variants as 64-bit integers ('RSVR IDs', **Fig. 1a**), which can encode 99.2% 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' (**Fig. 1a, Extended Data Fig. 2**). The package includes tools to annotate variants with MAF information from control databases (e.g. gnomAD[11]), pathogenicity scores (e.g. CADD scores[12]) and predicted Sequence Ontology (SO)[13] consequences with respect to a set of known 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 diseases assigned to each participant.

BeviMed identifies 238 known and 21 novel genetic associations

We built a Rareservoir of 11.9 million rare exonic and splicing single nucleotide variants (SNVs) and short insertions or deletions (indels) in canonical transcripts of protein-coding genes in Ensembl v104[14] from a merged VCF containing genotype calls for 77,539 participants in the Rare Diseases Main Programme of the 100KGP (Data Release Version 13) (**Fig. 1b**). During enrolment, expert clinicians assigned eligible cases to one or more of 220 Specific Diseases, arranged into 88 Disease Sub Groups, each of which belonged to one of 20 Disease Groups. We generated 269 case sets corresponding to the distinct Specific Diseases and Disease Sub Groups, ranging in size from 5,809 to one proband, and encoded them in the Rareservoir (**Fig. 1c, Extended Data Fig. 3**). 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 BeviMed, we obtained a posterior probability of association (PPA) between each of the 19,663 protein-coding genes and each of the 269 disease case sets. We selected probands in a given case set on the basis of 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 guard against false positive associations due to errors in the pedigree data and cryptic relatedness, we ensured the PPAs were robust to removal of cases sharing an excess of rare variants in the Rareservoir relative to ancestry-specific distributions. To account for correlation between case sets, we only recorded the association for each gene having the highest PPA within a given Disease Group. Conditional on a gene being causal for a particular disease, we recorded posterior probabilities over the mode of inheritance (MOI), the consequence class of variants that mediate disease risk (e.g. predicted loss-of-function variants or variants in the 5′ UTR) and the pathogenicity of each specific variant. Using a significance threshold of PPA>0.95, we identified 259 significant associations, 238 of which were documented by the PanelApp gene panel database[5] with high, medium or low levels of prior supporting evidence in the literature (**Fig. 2, Supplementary Table 1**). These results give an upper bound on the false discovery rate (FDR) of 8%. In contrast, a recent analysis of 57,000 samples in the 100KGP reported 249 known and 579 novel associations[15], giving an upper bound on the FDR of 70%, which suggests that our analytical approach has a greater specificity for a given sensitivity. Out of the 238 known associations that we identified, 40 (16.8%) were with Disease Sub Groups. 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 238 associations identified as known according to PanelApp data, 234 (98.3%) had an inferred MOI that was consistent with the MOIs listed for the relevant gene in the matched panel (220 associations), in the notes for the matched panel (five associations) or in the MOIs listed for an alternative relevant panel (nine associations) in PanelApp (**Supplementary Table 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'[16] and *AICDA* with dominant 'Primary immunodeficiency'[17]. 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'.

We found 21 novel genetic associations. To select a shortlist for further investigation, we developed 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 not disadvantage 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 conditional posterior probability of pathogenicity >0.5 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 on 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 considering additional experimental evidence or whether there was replication in other sequenced rare disease collections.

Loss-of-function mutations 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'. Primary lymphoedema is a group of genetic conditions caused by abnormal development of lymphatic vessels or failure of lymphatic function[18,19]. Three variants affecting different parts of the ERG protein 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 homozygous for the reference 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 also carried a predicted loss-of-function variant in *ERG* but had been enrolled to the 100KGP for an unrelated condition. 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. 3a**).

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 two independent reads overlapping that position which supported the alternative allele. 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. 3b**). This might explain why the father had a milder form of the disease, with an onset of lymphoedema over two decades later than his daughter.

ERG is a critical transcriptional regulator of blood vessel endothelial cell (EC) gene expression[20] and is essential for normal vascular development[21]. However, little is known about the contribution of ERG to lymphatic development or how the loss-of-function variants identified in our analysis could result in primary lymphoedema. To compare the expression of ERG in lymphatic EC versus blood vessel EC, we carried out real-time quantitative PCR for *ERG* in RNA obtained from primary cultures of human umbilical vein EC (HUVEC) and human dermal lymphatic EC (HLDEC). No significant difference in *ERG* transcript levels were observed between HUVEC and HDLEC (**Fig. 3c**). Likewise, similar ERG expression was detected by immunoblotting of HUVEC and HLDEC protein lysates (**Fig. 3d**). Immunofluorescence

microscopy was carried out to study the cellular localization of ERG in cultured HDLEC. These results confirmed nuclear co-localization of ERG with the lymphatic EC marker PROX1 (**Fig. 3e**). To understand whether ERG is present within the lymphatic vasculature *in vivo*, we analysed whole mount immunostaining of ERG and PROX1 in the ear skin of mice at three weeks after birth. At this time point, the lymphatic vessels have formed into a network of blind-ended lymphatic capillaries and valve-containing collecting vessels. We showed that ERG co-localizes with PROX1 in discrete lymphatic vascular structures that are distinct from the ERG signal present throughout the blood vessels in the mouse ear (**Fig. 3f**). Thus, we confirmed high levels of ERG expression in the lymphatic endothelium. In separate work, we have shown that deletion of the murine *ERG* homolog *Erg* in lymphatic endothelial cells results in defective lymphangiogenesis *in vivo*, confirming a role for *ERG* in lymphatic vessel remodelling and maintenance (Pirri et al., in preparation).

Truncating mutations 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. 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 as above, 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 Marfanoid skeletal features with complete penetrance, which co-segregated with the respective variants in genotyped participants (**Fig. 4a**). 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. 4b**), echoing the phenotypic characteristics of the syndromic aortopathy Loeys-Dietz syndrome[25]. The cases in the Japanese and Belgian families also exhibited a constellation of skeletal phenotypes, including pectus deformity, scoliosis and arachnodactyly, consistent with observations in the UK families.

To understand the molecular mechanisms underlying this novel defect, we examined the protein-protein interactions[26] among the complete set of high-confidence genes in the 'Thoracic aortic aneurysm or dissection' PanelApp panel and *PMEPA1*. PMEPA1 is a negative regulator of Transforming Growth Factor β (TGFβ) signalling[27], which has previously been implicated in multiple aortopathies, including

Loeys-Dietz syndrome[28]. The genes implicated in Loeys-Dietz syndrome form part of a tightly interacting subgroup of proteins in the TGFβ pathway, and PMEPA1 interacts with two of these: SMAD2 and SMAD3 (**Fig. 4c**). As the two candidate variants occur in the last exon of the transcript, they are likely to evade nonsense-mediated decay[29]. Their truncating effects are predicted to remove a PPxY interaction motif, while leaving the SMAD interaction motif intact (**Fig. 4d**), possibly affecting binding between PMEPA1 and SMAD2/3, and altering TGFβ signalling through a gain-of-function mechanism.

Loss-of-function mutations in *GPR156* give rise 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 encoding p.S207Vfs*113 and a one base pair insertion encoding p.P718Lfs*86 with respect to the canonical Ensembl transcript ENST00000464295.6. In one family, each of the two parents heterozygous for p.S207Vfs*113 transmitted the frameshift allele to each of their two affected children, who were therefore homozygous for that allele. In a second family, the mother was heterozygous for the same p.S207Vfs*113 allele, which she transmitted to her two affected daughters. The father, however, was heterozygous for a different frameshift allele, encoding p.P718Lfs*86. He transmitted that allele to the same two affected daughters, who were thus compound heterozygous for truncating alleles. Using GeneMatcher[30], we identified a third pedigree with biallelic truncating mutations in *GPR156*. This pedigree concerned a consanguineous family with four siblings affected by hearing impairment, all of whom were homozygous for a variant encoding p.S642Afs*162 (**Fig. 5a**). The eight affected individuals in these three families all had congenital non-syndromic bilateral sensorineural hearing loss (see **Extended Data Fig. 4** for illustrative audiograms).

GPR156 encodes a class C G protein-coupled receptor (GPCR) with sequence homology to the GABAB receptors[31]. Although previously designated as an orphan GPCR, *GPR156* has recently been identified as a critical regulator of stereocilia orientation on hair cells of the auditory epithelium and other mechanosensory tissues[32]. Its expression is highly restricted to hair cells in the inner ear[33]. Disruption of stereocilia is a common pathogenic mechanism underlying many human Mendelian hearing loss disorders[34] 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. 5b**). By immunostaining of the Corti and vestibular system from P10 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. 5c**).

Although one of the variants interrupted the sequence of seven transmembrane receptors of GPR156, two of the variants were located in the region encoding the GPCR's cytoplasmic tail, within the gene's last exon (**Fig. 5d**). To determine the effect of the truncating mutations on GPR156 protein, we transfected Cos7 cells, which do not express human *GPR156* endogenously, with GFP-coupled constructs of wild type *GPR156* sequence and mutant *GPR156* sequence containing the p.S642Afs*162 allele. While cells transfected with wild type sequence expressed GFP-GPR156 protein robustly, cells transfected with the mutant construct did not express the protein appreciably, suggesting that the mutant protein is degraded and pointing to a loss-of-function mechanism (**Fig. 5e**). These data suggest that the

biallelic chain truncating variants in *GPR156* cause a congenital hearing loss by preventing expression of GPR156 protein and 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 association analysis of rare diseases using approaches such as BeviMed. Of the 21 novel associations, we shortlisted, replicated and validated three. These three novel aetiologies, which we identified in a single unified analysis of one large cohort, involve genes that had not previously been implicated in any of these human diseases. The remaining 18 novel associations include further tantalising hypotheses. For example, *LRRC7*, which we identified to be associated with intellectual disability, encodes a brain-specific protein in post-synaptic densities[35], and Lrrc7-deficient mice exhibit a neuro-behavioural phenotype[36]. *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[37]. These and other candidates will require replication and validation before they can be considered causative genes. In addition, the exploration of rare variation in other regions of the genome using similar analytical approaches are a promising focus of future research.

Gene	Case set	Level	Variant class	MOI	pLI	Z	Co-segregation evidence	Biological function and existing disease associations	Score	Replication
ERG	Primary lymphoedema	SD	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.[38,21].**	3	Internal (case enrolled for a different Specific Disease)
GPR156	Congenital hearing impairment	SD	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[32].**	3	Maryland
PMEPA1	Familial Thoracic Aortic Aneurysm Disease	SD	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[27]. Aberrant TGFβ signalling is implicated in multiple Mendelian aortopathies[28].**	3	100KGP pilot, Antwerp, Tokyo
LRRC7	Intellectual disability	SD	High	Dom	**1**	3.6		**Brain-specific scaffold protein in post-synaptic densities[35]. LRRC7-inactivated mice have neuro-behavioural phenotype[36]**	2	
USP33	Extreme early-onset hypertension	SD	High	Dom	0.86	**2.1**		**Deubiquitinating enzyme implicated in multiple cellular processes, including regulation of expression of the β2-adrenergic receptor[37], a critical regulator of circulatory function and blood pressure[39].**	2	
ARPC3	Charcot-Marie-Tooth disease	SD	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[40]. ArpC3-inactivation in mice causes axon dysfunction[41].**	1	

ATF7	Arthrogryposis	SD	High	Dom	**1**	2.28		Chromatin regulator implicated in innate immune response, adipocyte differentiation and telomerase regulation.	1	
FMN1	Congenital hearing impairment	SD	High	**Rec**	0	-1.53	Co-segregation in 2 unaffected relatives.	**Formin family protein involved in linear actin and microtubule polymerisation[42]. Pathogenic variants in the formin DIAPH1 cause hearing loss via cytoskeletal disruption in auditory stereocilia[43].**	1	
RAB35	Familial Hypercholesterolaemia	SD	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	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[44].	1	
TUFT1	Epidermolysis bullosa	SD	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	Moderate	Dom	0.29	0.42		Uncharacterised nucleosomal protein.	0	
INSL4	Rod Dysfunction Syndrome	SD	Moderate	Dom	0	-1.43		Insulin-like growth factor implicated in trophoblast and bone development.	0	
KRT14	Young onset tumour syndromes	DSG	Moderate	**Rec**	0	0.84	Co-segregation in 2 unaffected relatives.	Component of keratin intermediate filaments in epithelial cells. Pathogenic variants cause Epidermolysis bullosa simplex !A-D(AD/AR); Dermatopathia pigmentosa reticularis (AD); Naegeli-Franceschetti-Jadassohn syndrome (AD).	0	
LSM6	Familial congenital heart disease	SD	High	Dom	0.87	1.54		Sm-like protein involved in spliceosome assembly and pre-mRNA splicing function. Putative regulator of mRNA degradation.	0	
MPPE1	Primary ciliary dyskinesia	SD	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	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	High	Dom	0.85	2.06		Component of the large ribosomal subunit that mediates protein translation.	0	
SERPINB3	Autosomal recessive congenital ichthyosis	SD	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	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	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. Associations with a score of three are

shaded. 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 identified as aetiological by BeviMed 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 | The Rareservoir in the 100KGP. a, Schematic of the database build procedure and contents. 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 (e.g. CADD score) and their gnomAD probabilistic minor allele frequency scores[2]. 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. **b,** Schematic showing the variant data in the 100KGP Main Programme Rareservoir. 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. **c,** 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.

Fig. 2 | BeviMed genetic association results. 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 with substantial supporting evidence from the literature, but which were not automatically mapped to PanelApp, are marked with an asterisk (**Supplementary Table 1**). Novel 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. 3 | Loss-of-function mutations in ERG are responsible for primary lymphoedema. a, Pedigrees for the four probands with loss-of-function variants in *ERG*: three were assigned the Specific Disease 'Primary lymphoedema' (left) and one was assigned to an unrelated disorder but later determined by chart review to also have lymphoedema (right-most pedigree). **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 pileups of reads 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,** RT-PCR amplification of ERG mRNA in HDLEC relative to HUVEC. Data are normalised to GAPDH. Statistical significance was assessed using a Student's *t*-test, n.s.: not significant. **d,** Representative immunoblot of HUVEC and HDLEC protein lysates identified several bands corresponding to ERG isoforms expressed at similar intensities in both cell types. **e,** Immunofluorescence microscopy of HDLEC shows ERG nuclear co-localisation with PROX1 and DAPI.

HDLEC junctions are shown using an antibody to VE-cadherin. Scale bar, 50μm. **f,** *En face* immunofluorescence confocal microscopy of mouse ear skin. Vessels are stained with antibodies to the lymphatic marker PROX1 (green) and ERG (red). Scale bar, 100μm.

Fig. 4 | Truncating mutations in *PMEPA1* result in Loeys-Dietz syndrome. a, Pedigrees for the three probands in the 100KGP (discovery cohort) heterozygous for the frameshift insertion encoding p.S209Qfs*3 and probands from replication cohorts, including: one from the 100KGP pilot programme heterozygous for the frameshift deletion encoding 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. **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 a 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 a threshold for declaring an association significant, provided all terms at higher ranks were also significant. **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[26] 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. 5 | Loss-of-function mutations 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 mutations in *GPR156.* 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. **d,** Schematic showing the effects of each variant at the cDNA and amino acid level, and on the protein product. **e,** Representative western blot of GFP-GPR156 in untransfected Cos7 cells (C), Cos7 cells transfected with the wild type construct (W) and Cos7 cells transfected with the construct containing the mutant p.S642Afs*162 allele (M) (left). The ratios of GFP/GAPDH intensities in the three different conditions across technical replicates. The ratios corresponding to cells that had been transfected with the mutant construct were significantly lower than those corresponding to cells that had been transfected with the wild type construct (p=0.0017, *t*-test), but not significantly different from those corresponding to untransfected cells (p>0.05, *t*-test).

METHODS

Ethics

The 100,000 Genomes project was approved by East of England–Cambridge Central REC REF 20/EE/0035. Only participants who consented 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 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.

Building a Rareservoir

A Rareservoir is a relational database built through a series of steps from a set of input data and parameters (**Fig. 1a**, **Extended Data Fig. 2**). The 'bcftools' program[45] extracts ('bcftools view') and normalises ('bcftools norm') variants from either a set of single sample 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[2] 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 '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. 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 files, which may be executed in parallel over separate regions against a merged VCF or over single gVCF files. 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. 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**).

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 proportion of variants in gnomAD 3.0 weighted by allele count that can be encoded losslessly is 99.2%, while 99.7% of all variants in gnomAD 3.0 can be represented by a distinct RSVR ID. 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[46] variants in the merged VCF file of 77,539 consented participants in the 100KGP rare diseases programme. Variants were filtered according to the default criteria described above using CADD v1.5 and gnomAD v3.0. Additionally, we imposed a minimum genotype quality of 35 and at least one predicted consequence on a canonical transcript in Ensembl v104. For each of the 269 case sets (**Extended Data Fig. 3**), we applied the BeviMed 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. When we applied BeviMed, we averaged over models including the following classes of variant:

- 5' UTR variants: those with a 5_prime_UTR_variant consequence (prior probability of association 0.0001),
- High impact variants: those with any consequences amongst start_lost, stop_lost, frameshift_variant, stop_gained, splice_donor_variant or splice_acceptor_variant (prior 0.00495),
- Moderate or high impact variants: those with any high impact consequence or missense_variant or inframe_deletion consequences (prior 0.00495).

Thus the prior probability of association was set to 0.01. To select the samples and set the case status used for each application of BeviMed to a disease–gene pair, one affected person per pedigree was selected, whilst the controls not from selected pedigrees were chosen from the MSUP. 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 Research Environment. Case status was then set to one for affected individuals who were either not explained or explained by the target gene and zero for everyone else. Analyses yielding a BeviMed PPA >0.95 were re-run including all samples (i.e. with relatives). Associations for which the analysis with all samples caused the PPA to fall below 0.9 were filtered out, as the drop indicated a lack of co-segregation. To guard against false positives due to incorrect pedigree data, population structure or cryptic relatedness, we applied the following algorithm. Firstly, we obtained the distribution of the number of rare variants in the Rareservoir shared by pairs of individuals with a given ancestry in the 100KGP. The top percentile in each of these distributions was used to indicate potential relatedness between two cases in a given population. Analyses yielding a BeviMed PPA >0.95 were re-run after removing cases so as to ensure that any remaining set of cases carrying the same rare variant in the gene were not potentially related according to the criterion above. Associations for which this analysis caused the PPA to fall below 0.25 were filtered out.

PanelApp annotation

Significant associations were coloured according to PanelApp[5] (**Fig. 2**) 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 in PanelApp could be found were inspected manually to assess whether PanelApp contained a suitable panel that could not be matched by the algorithm (marked with an asterisk in **Fig. 2**).

Shortlisting novel 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 >0.5 conditional on association as 'probably pathogenic'.

- *Counting co-segregating pedigree members.* The pedigrees harbouring probably pathogenic alleles (1 for dominant associations and 2 for recessive) 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[11]. 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 not disadvantage 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 ver.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 permeabilized 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 visualized 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.

GPR156: Western blot

We subcloned *GPR156* from human brain cDNA, into EGFP-C1 vector. The mutant *GPR156* construct was generated by mutagenesis using the QuickChange kit (Stratagene, La Jolla, CA) and a wild type GFP-GPR156 as a template. For expression analysis, the WT and mutant GPR156 experiments 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 24hr after transfection, lysed in buffer containing 1% CHAPS, 100mM NaCl, and 25mM HEPES, pH 7.4 and clarified by centrifugation at 14,000 rpm. Lysates (20µg) were run on a 15% SDS-PAGE gel. Membrane was blocked with 5% milk then incubated with anti-GFP antibody and immunoblots developed with HRP conjugated secondary (sheep anti-rabbit) antibody. Comparable loading was checked by stripping and reprobing the blots with anti-GAPDH 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 P10 mice 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 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.

Code Availability

The rsvr tool and Rareservoir schema is available from https://github.com/turrogroup/rsvr. The BeviMed software package is available from https://cran.r-project.org/web/packages/BeviMed/.

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 is supported by Wellcome Collaborative Award 219506/Z/19/Z. V. Hartill is supported by MRC/NIHR Clinical Academic Research Partnership MR/V037617/1 (VH). K. Frudd was funded by PG/17/33/32990. D. Pirri is funded by PG/20/16/35047. E. Sackey is supported by Swiss Federal National Fund for Scientific Research n°CRSII5_177191/1. K. Freson is supported by KU Leuven BOF grant C14/19/096. Work at the University of Maryland Baltimore was supported by the NIDCD/NIH grant R01DC016295 to Z. Ahmed.

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 and K. Ramzan 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.

Additional information

Supplementary information is available for this paper online.

Correspondence and requests for materials should be addressed to E. Turro.

EXTENDED DATA FIGURE LEGENDS

Extended Data Fig. 1 | Reduction in the number of genotypes stored per sample. For each of five sets of 100 randomly chosen 100KGP participants with a probability >0.9 of belonging to a specific ancestry group[47]: **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 | Detailed schematic of the database build procedure. 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. 3 | The 269 case sets. 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.

Extended Data Fig. 4 | Illustrative audiograms for *GPR156* cases. Air and bone conduction audiograms for the two affected daughters of the family with compound heterozygous *GPR156* truncating alleles.

SUPPLEMENTARY TABLE LEGENDS

Supplementary Table 1 | Detailed annotated results of the genetic association analyses. Table of associations shown in Fig. 2 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 and associated evidence level, notes on the consistency between the MOI listed by PanelApp for the association and the inferred MOI (MOI match comment), and the method that was used to find the match (Match method, either ‘Automatic’ or ‘Manual’).

References

1. Ferreira CR. The burden of rare diseases. Am J Med Genet A. 2019; 179(6):885--892.
2. Turro E et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020; 583(7814):96--102.
3. 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.
4. The National Genomics Research and Healthcare Knowledgebase v5, Genomics England. doi:10.6084/m9.figshare.4530893.v5. 2019.
5. Martin AR et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019; 51(11):1560--1565.
6. 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.
7. Boycott KM et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. Am J Hum Genet. 2017; 100(5):695--705.

8. 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.
9. Hail Team. (2022). Hail 0.2. https://github.com/hail-is/hail.
10. Lopez J et al. HGVA: the Human Genome Variation Archive. Nucleic Acids Res. 2017; 45(W1):W189-W194.
11. Karczewski KJ et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581(7809):434--443.
12. 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.
13. Eilbeck K et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2005; 6(5):R44.
14. Howe KL et al. Ensembl 2021. Nucleic Acids Res. 2021; 49(D1):D884-D891.
15. 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.
16. Körber 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.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
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. 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.
27. 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.
28. Creamer TJ, Bramel EE, MacFarlane EG. Insights on the Pathogenesis of Aneurysm through the Study of Hereditary Aortopathies. Genes (Basel). 2021; 12(2).
29. Thermann R et al. Binary specification of nonsense codons by splicing and cytoplasmic translation. EMBO J. 1998; 17(12):3484--3494.

30. 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.
31. 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.
32. Kindt KS et al. EMX2-GPR156-Gai reverses hair cell orientation in mechanosensory epithelia. Nat Commun. 2021; 12(1):2861.
33. 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.
34. 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.
35. 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.
36. 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.
37. 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.
38. Birdsey GM et al. The endothelial transcription factor ERG promotes vascular stability and growth through Wnt/ß-catenin signaling. Dev Cell. 2015; 32(1):82--96.
39. Motiejunaite J, Amar L, Vidal-Petiot E. Adrenergic receptors and cardiovascular effects of catecholamines. Ann Endocrinol (Paris). 2021; 82(3-4):193--197.
40. Muñoz-Lasso DC, Romá-Mateo C, Pallardó FV, Gonzalez-Cabo P. Much More Than a Scaffold: Cytoskeletal Proteins in Neurological Disorders. Cells. 2020; 9(2).
41. Zuchero JB et al. CNS myelin wrapping is driven by actin disassembly. Dev Cell. 2015; 34(2):152--167.
42. DeWard AD, Eisenmann KM, Matheson SF, Alberts AS. The role of formins in human disease. Biochim Biophys Acta. 2010; 1803(2):226--233.
43. 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.
44. Geppert M et al. The role of Rab3A in neurotransmitter release. Nature. 1994; 369(6480):493--497.
45. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078--2079.
46. 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
47. Ancestry inference for the 100,000 Genomes Project participants (2022). https://research-help.genomicsengland.co.uk/display/GERE/Ancestry+inference

Fig. 1

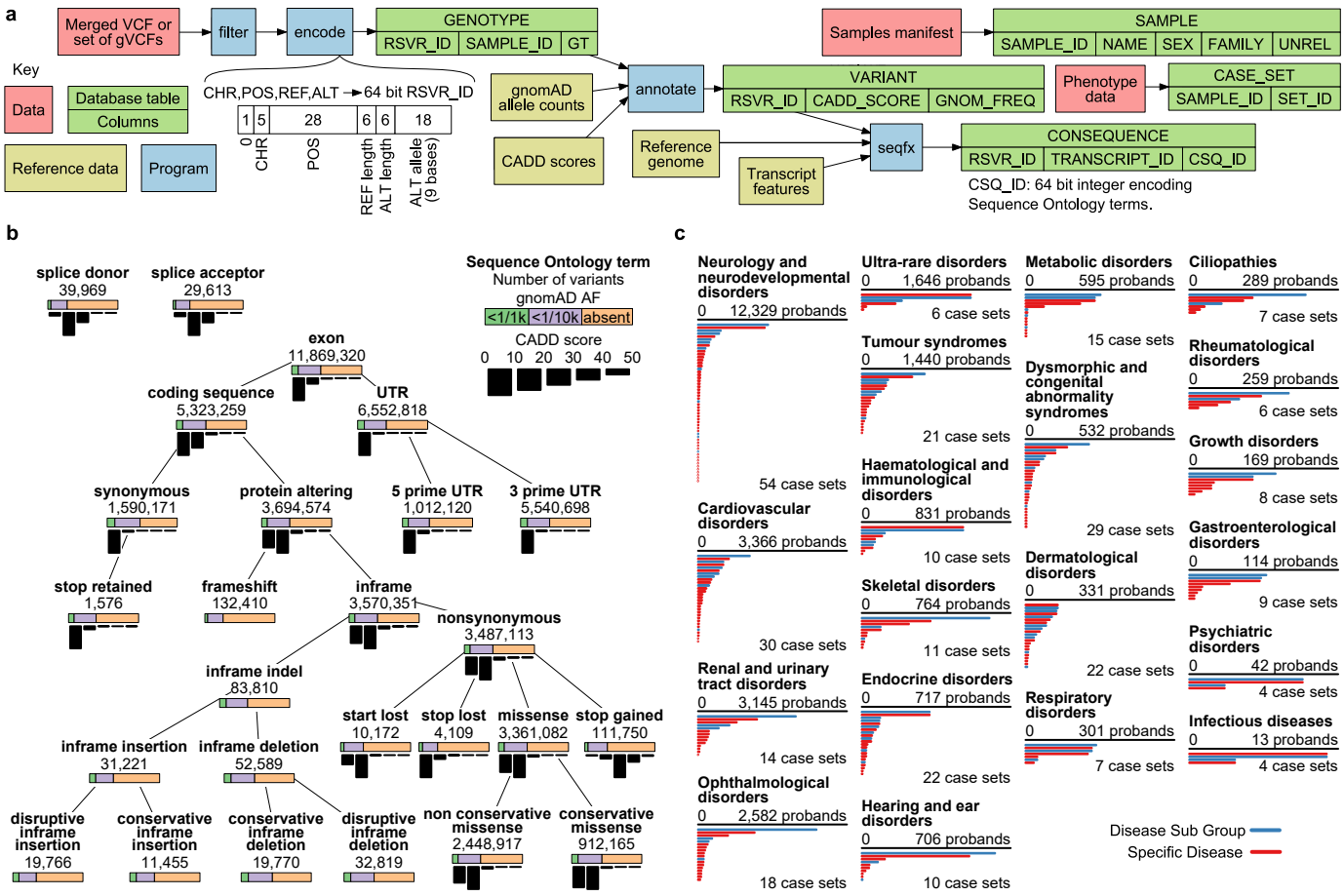

a
Merged VCF or set of gVCFs
filter
encode
GENOTYPE
RSVR_ID
SAMPLE_ID
GT
Samples manifest
SAMPLE
SAMPLE_ID
NAME
SEX
FAMILY
UNREL
Key
Data
Database table
Columns
Reference data
Program
CHR,POS,REF,ALT → 64 bit RSVR_ID
1
5
28
6
6
18
0
CHR
POS
REF length
ALT length
ALT allele (9 bases)
gnomAD allele counts
CADD scores
annotate
VARIANT
RSVR_ID
CADD_SCORE
GNOM_FREQ
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.
b
splice donor 39,969
splice acceptor 29,613
Sequence Ontology term
Number of variants
gnomAD AF
<1/1k
<1/10k
absent
CADD score
0 10 20 30 40 50
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
c
Neurology and neurodevelopmental disorders
0 12,329 probands
54 case sets
Cardiovascular disorders
0 3,366 probands
30 case sets
Renal and urinary tract disorders
0 3,145 probands
14 case sets
Ophthalmological disorders
0 2,582 probands
18 case sets
Ultra-rare disorders
0 1,646 probands
6 case sets
Tumour syndromes
0 1,440 probands
21 case sets
Haematological and immunological disorders
0 831 probands
10 case sets
Skeletal disorders
0 764 probands
11 case sets
Endocrine disorders
0 717 probands
22 case sets
Hearing and ear disorders
0 706 probands
10 case sets
Metabolic disorders
0 595 probands
15 case sets
Dysmorphic and congenital abnormality syndromes
0 532 probands
29 case sets
Dermatological disorders
0 331 probands
22 case sets
Respiratory disorders
0 301 probands
7 case sets
Ciliopathies
0 289 probands
7 case sets
Rheumatological disorders
0 259 probands
6 case sets
Growth disorders
0 169 probands
8 case sets
Gastroenterological disorders
0 114 probands
9 case sets
Psychiatric disorders
0 42 probands
4 case sets
Infectious diseases
0 13 probands
4 case sets
Disease Sub Group
Specific Disease

Fig. 2

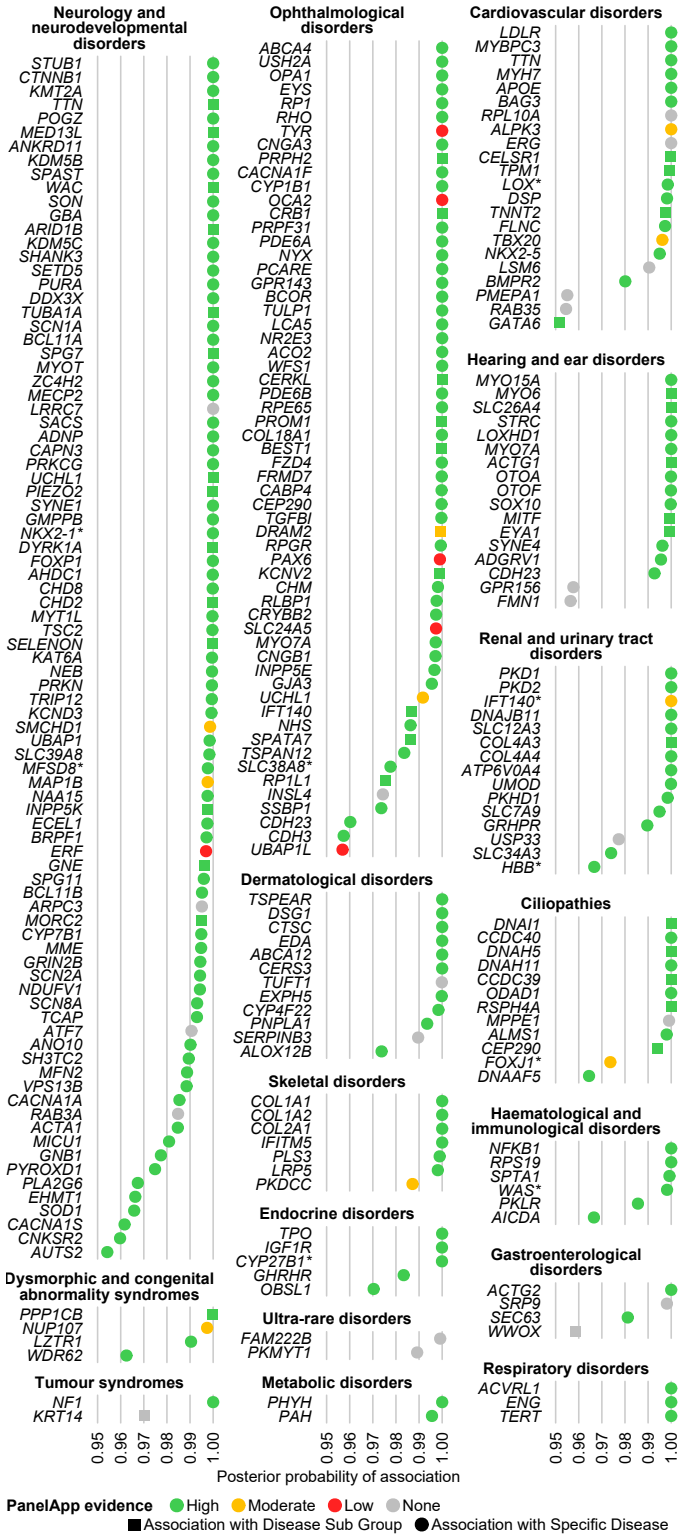

Neurology and neurodevelopmental disorders
STUB1
CTNNB1
KMT2A
TTN
POGZ
MED13L
ANKRD11
KDM5B
SPAST
WAC
SON
GBA
ARID1B
KDM5C
SHANK3
SETD5
PURA
DDX3X
TUBA1A
SCN1A
BCL11A
SPG7
MYOT
ZC4H2
MECP2
LRRC7
SACS
ADNP
CAPN3
PRKCG
UCHL1
PIEZO2
SYNE1
GMPPB
NKX2-1*
DYRK1A
FOXP1
AHDC1
CHD8
CHD2
MYT1L
TSC2
SELENON
KAT6A
NEB
PRKN
TRIP12
KCND3
SMCHD1
UBAP1
SLC39A8
MFSD8*
MAP1B
NAA15
INPP5K
ECEL1
BRPF1
ERF
GNE
SPG11
BCL11B
ARPC3
MORC2
CYP7B1
MME
GRIN2B
SCN2A
NDUFV1
SCN8A
TCAP
ATF7
ANO10
SH3TC2
MFN2
VPS13B
CACNA1A
RAB3A
ACTA1
MICU1
GNB1
PYROXD1
PLA2G6
EHMT1
SOD1
CACNA1S
CNKSR2
AUTS2
Dysmorphic and congenital abnormality syndromes
PPP1CB
NUP107
LZTR1
WDR62
Tumour syndromes
NF1
KRT14
0.95
0.96
0.97
0.98
0.99
1.00
Ophthalmological disorders
ABCA4
USH2A
OPA1
EYS
RP1
RHO
TYR
CNGA3
PRPH2
CACNA1F
CYP1B1
OCA2
CRB1
PRPF31
PDE6A
NYX
PCARE
GPR143
BCOR
TULP1
LCA5
NR2E3
ACO2
WFS1
CERKL
PDE6B
RPE65
PROM1
COL18A1
BEST1
FZD4
FRMD7
CABP4
CEP290
TGFBI
DRAM2
RPGR
PAX6
KCNV2
CHM
RLBP1
CRYBB2
SLC24A5
MYO7A
CNGB1
INPP5E
GJA3
UCHL1
IFT140
NHS
SPATA7
TSPAN12
SLC38A8*
RP1L1
INSL4
SSBP1
CDH23
CDH3
UBAP1L
Dermatological disorders
TSPEAR
DSG1
CTSC
EDA
ABCA12
CERS3
TUFT1
EXPH5
CYP4F22
PNPLA1
SERPINB3
ALOX12B
Skeletal disorders
COL1A1
COL1A2
COL2A1
IFITM5
PLS3
LRP5
PKDCC
Endocrine disorders
TPO
IGF1R
CYP27B1*
GHRHR
OBSL1
Ultra-rare disorders
FAM222B
PKMYT1
Metabolic disorders
PHYH
PAH
0.95
0.96
0.97
0.98
0.99
1.00
Cardiovascular disorders
LDLR
MYBPC3
TTN
MYH7
APOE
BAG3
RPL10A
ALPK3
ERG
CELSR1
TPM1
LOX*
DSP
TNNT2
FLNC
TBX20
NKX2-5
LSM6
BMPR2
PMEPA1
RAB35
GATA6
Hearing and ear disorders
MYO15A
MYO6
SLC26A4
STRC
LOXHD1
MYO7A
ACTG1
OTOA
OTOF
SOX10
MITF
EYA1
SYNE4
ADGRV1
CDH23
GPR156
FMN1
Renal and urinary tract disorders
PKD1
PKD2
IFT140*
DNAJB11
SLC12A3
COL4A3
COL4A4
ATP6V0A4
UMOD
PKHD1
SLC7A9
GRHPR
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
Respiratory disorders
ACVRL1
ENG
TERT
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

Fig. 3

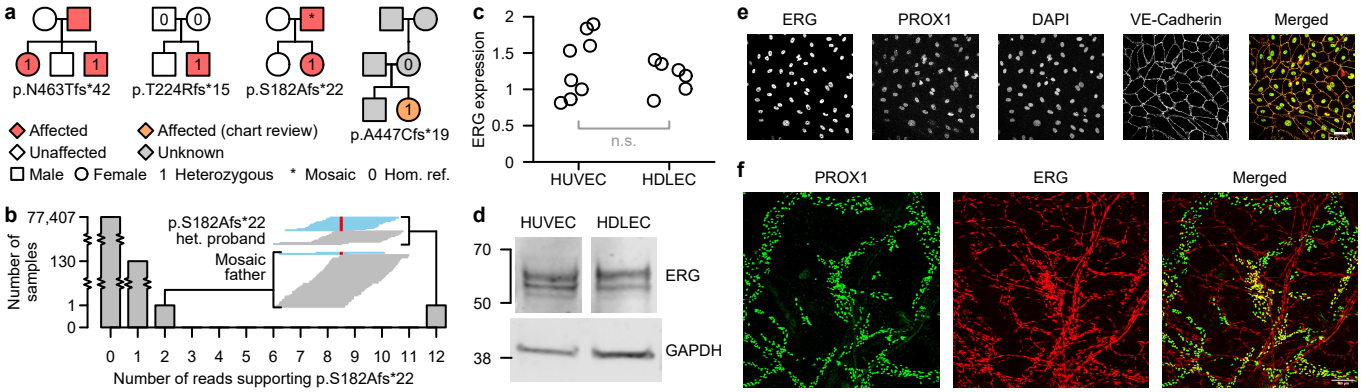


Fig. 4

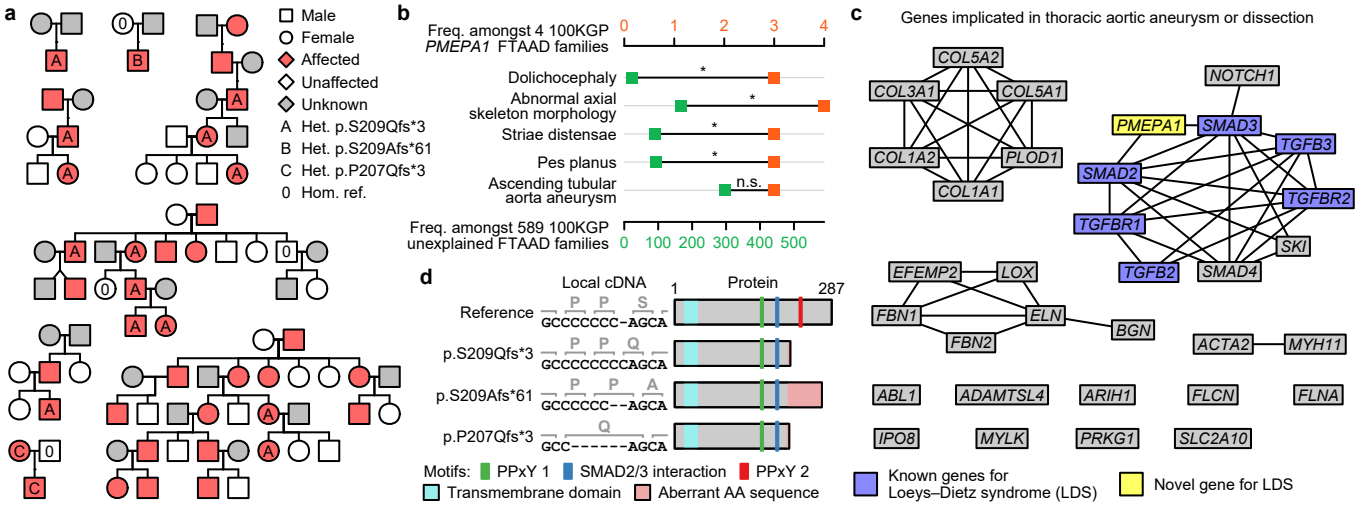
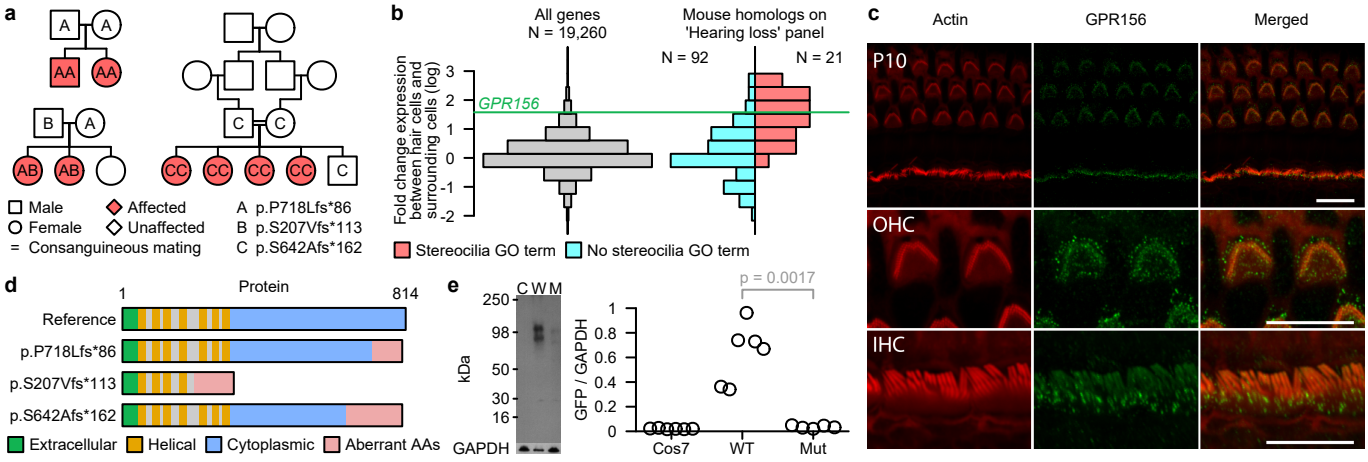
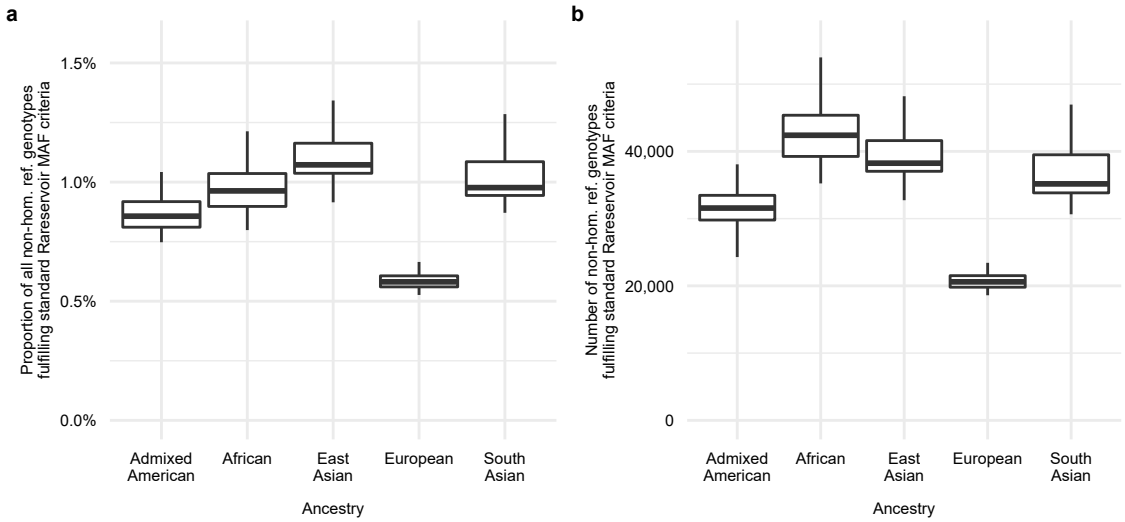


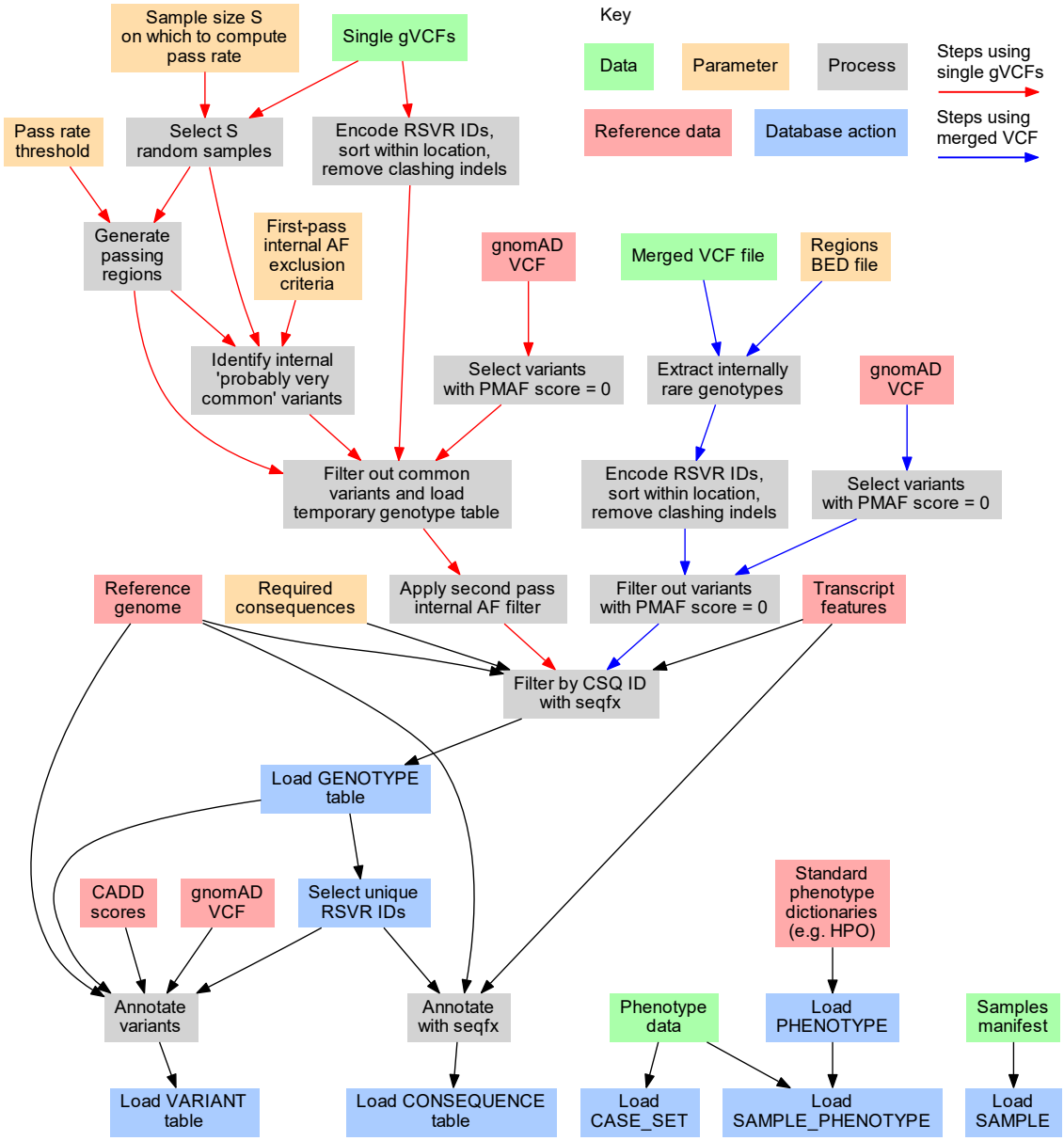
Fig. 5



Extended Data Fig. 1

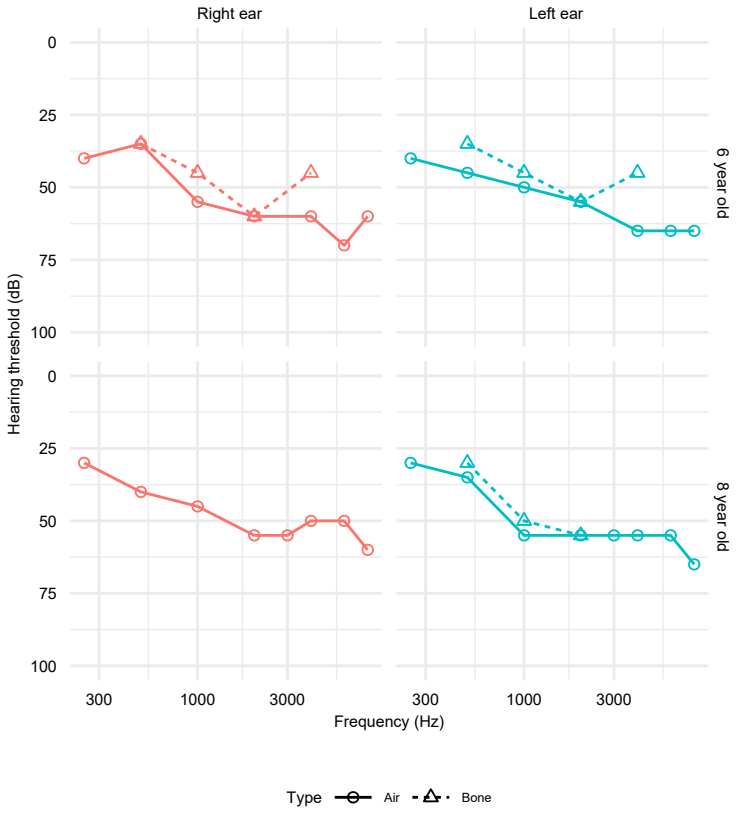


Extended Data Fig. 2



Extended Data Fig. 3

Extended Data Fig. 4



Supplementary Table 1

Gene	Case set	PPA	Level	Variant class	MOI	Matched panel	Evidence	MOI match comment	Match method
PKD1	Cystic kidney disease	1.000	Specific Disease	Moderate	Dominant	Cystic kidney disease	green	Consistent.	Automatic
PKD2	Cystic kidney disease	1.000	Specific Disease	High	Dominant	Cystic kidney disease	green	Consistent.	Automatic
COL1A1	Osteogenesis imperfecta	1.000	Specific Disease	High	Dominant	Osteogenesis imperfecta	green	Consistent.	Automatic
ABCA4	Inherited macular dystrophy	1.000	Specific Disease	Moderate	Dominant	Retinal disorders	green	Supporting evidence for the inferred MOI in the literature which is absent from PanelApp (PMID:11818392).	Automatic
ACVRL1	Hereditary haemorrhagic telangiectasia	1.000	Specific Disease	Moderate	Dominant	Hereditary haemorrhagic telangiectasia	green	Consistent.	Automatic
LDLR	Familial Hypercholesterolaemia	1.000	Specific Disease	Moderate	Dominant	Familial hypercholesterolaemia	green	Consistent.	Automatic
ENG	Hereditary haemorrhagic telangiectasia	1.000	Specific Disease	Moderate	Dominant	Hereditary haemorrhagic telangiectasia	green	Consistent.	Automatic
COL1A2	Osteogenesis imperfecta	1.000	Specific Disease	Moderate	Dominant	Osteogenesis imperfecta	green	Consistent.	Automatic
NF1	Neurofibromatosis Type 1	1.000	Specific Disease	High	Dominant	Neurofibromatosis Type 1	green	Consistent.	Automatic
USH2A	Rod-cone dystrophy	1.000	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
OPA1	Inherited optic neuropathies	1.000	Specific Disease	High	Dominant	Optic neuropathy	green	Consistent.	Automatic
EYS	Rod-cone dystrophy	1.000	Specific Disease	High	Recessive	Retinal disorders	green	Consistent.	Automatic
MYBPC3	Hypertrophic Cardiomyopathy	1.000	Specific Disease	Moderate	Dominant	Left Ventricular Noncompaction Cardiomyopathy	green	Consistent.	Automatic
RP1	Rod-cone dystrophy	1.000	Specific Disease	High	Dominant	Retinal disorders	green	Consistent.	Automatic
IFT140	Cystic kidney disease	1.000	Specific Disease	High	Dominant	Cystic renal disease	amber	Alternative compatible panel found matching MOI: "Cystic kidney disease" (not auto-matched as listed with "grey" evidence level).	Manual
RHO	Rod-cone dystrophy	1.000	Specific Disease	Moderate	Dominant	Retinal disorders	green	Consistent.	Automatic
TYR	Developmental macular and foveal dystrophy	1.000	Specific Disease	Moderate	Recessive	Retinal disorders	red	Alternative compatible panel found matching MOI: "Structural eye disease".	Automatic
TTN	Dilated Cardiomyopathy	1.000	Specific Disease	High	Dominant	Dilated cardiomyopathy - adult and teen	green	Consistent.	Automatic
STUB1	Hereditary ataxia	1.000	Specific Disease	Moderate	Dominant	Hereditary ataxia	green	Alternative compatible panel found matching MOI: "Severe Paediatric Disorders".	Automatic
TERT	Familial pulmonary fibrosis	1.000	Specific Disease	Moderate	Dominant	Familial pulmonary fibrosis	green	Consistent.	Automatic
DNAI1	Respiratory ciliopathies	1.000	Disease Sub Group	Moderate	Recessive	Primary ciliary disorders	green	Consistent.	Automatic
MYH7	Hypertrophic Cardiomyopathy	1.000	Specific Disease	Moderate	Dominant	Left Ventricular Noncompaction Cardiomyopathy	green	Consistent.	Automatic
CCDC40	Primary ciliary dyskinesia	1.000	Specific Disease	High	Recessive	Primary ciliary disorders	green	Consistent.	Automatic
CNGA3	Cone Dysfunction Syndrome	1.000	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
CTNNB1	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
DNAH5	Respiratory ciliopathies	1.000	Disease Sub Group	High	Recessive	Non-CF bronchiectasis	green	Consistent.	Automatic
PRPH2	Posterior segment abnormalities	1.000	Disease Sub Group	Moderate	Dominant	Retinal disorders	green	Consistent.	Automatic
KMT2A	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
CACNA1F	Rod Dysfunction Syndrome	1.000	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
CYP1B1	Glaucoma (developmental)	1.000	Specific Disease	Moderate	Recessive	Glaucoma (developmental)	green	Consistent.	Automatic
APOE	Familial Hypercholesterolaemia	1.000	Specific Disease	Moderate	Dominant	Familial hypercholesterolaemia	green	Consistent.	Automatic
DNAJB11	Cystic kidney disease	1.000	Specific Disease	High	Dominant	Cystic kidney disease	green	Consistent.	Automatic
OCA2	Developmental macular and foveal dystrophy	1.000	Specific Disease	Moderate	Recessive	Retinal disorders	red	Alternative compatible panel found matching MOI: "Structural eye disease".	Automatic
SLC12A3	Renal tubular acidosis	1.000	Specific Disease	Moderate	Recessive	Renal tubulopathies	green	Consistent.	Automatic
TTN	Neuromuscular disorders	1.000	Disease Sub Group	High	Recessive	Neuromuscular disorders	green	Consistent.	Automatic
TSPEAR	Ectodermal dysplasia without a known gene mutation	1.000	Specific Disease	Moderate	Recessive	Ectodermal dysplasia without a known gene mutation	green	Consistent.	Automatic
ACTG2	Early onset or familial intestinal pseudo obstruction	1.000	Specific Disease	Moderate	Dominant	Gastrointestinal neuromuscular disorders	green	Consistent.	Automatic
POGZ	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
MED13L	Neurodevelopmental disorders	1.000	Disease Sub Group	High	Dominant	Intellectual disability	green	Consistent.	Automatic
ANKRD11	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
CRB1	Posterior segment abnormalities	1.000	Disease Sub Group	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
KDM5B	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
MYO15A	Congenital hearing impairment	1.000	Specific Disease	Moderate	Recessive	Hearing loss	green	Consistent.	Automatic
PRPF31	Rod-cone dystrophy	1.000	Specific Disease	High	Dominant	Retinal disorders	green	Consistent.	Automatic
DSG1	Palmoplantar keratoderma and erythrokeratodermas	1.000	Specific Disease	High	Dominant	Palmoplantar keratoderma and erythrokeratodermas	green	Consistent.	Automatic
SPAST	Hereditary spastic paraplegia	1.000	Specific Disease	Moderate	Dominant	Hereditary spastic paraplegia	green	Consistent.	Automatic
BAG3	Dilated Cardiomyopathy	1.000	Specific Disease	High	Dominant	Dilated cardiomyopathy - adult and teen	green	Consistent.	Automatic
WAC	Neurodevelopmental disorders	1.000	Disease Sub Group	High	Dominant	Intellectual disability	green	Consistent.	Automatic
MYO6	Non-syndromic hearing loss	1.000	Disease Sub Group	High	Dominant	Hearing loss	green	Consistent.	Automatic
RPL10A	Milroy disease	1.000	Specific Disease	High	Dominant				
COL2A1	Stickler syndrome	1.000	Specific Disease	High	Dominant	Stickler syndrome	green	Consistent.	Automatic
SLC26A4	Non-syndromic hearing loss	1.000	Disease Sub Group	Moderate	Recessive	Hearing loss	green	Consistent.	Automatic
PDE6A	Rod-cone dystrophy	1.000	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
SON	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
GBA	Early onset and familial Parkinson's Disease	1.000	Specific Disease	Moderate	Dominant	Parkinson Disease and Complex Parkinsonism	green	Consistent.	Automatic
ARID1B	Neurodevelopmental disorders	1.000	Disease Sub Group	High	Dominant	Intellectual disability	green	Consistent.	Automatic
COL4A3	Syndromes with prominent renal abnormalities	1.000	Disease Sub Group	Moderate	Dominant	Haematuria	green	Consistent.	Automatic
CTSC	Palmoplantar keratoderma and erythrokeratodermas	1.000	Specific Disease	Moderate	Recessive	Palmoplantar keratoderma and erythrokeratodermas	green	Consistent.	Automatic
KDM5C	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Alternative compatible panel found matching MOI: "Hereditary spastic paraplegia" (associated with intellectual disability).	Automatic
SHANK3	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
STRC	Congenital hearing impairment	1.000	Specific Disease	High	Recessive	Hearing loss	green	Consistent.	Automatic
LOXHD1	Congenital hearing impairment	1.000	Specific Disease	Moderate	Recessive	Hearing loss	green	Consistent.	Automatic
MYO7A	Congenital hearing impairment	1.000	Specific Disease	Moderate	Recessive	Hearing loss	green	Consistent.	Automatic
ALPK3	Hypertrophic Cardiomyopathy	1.000	Specific Disease	High	Dominant	Hypertrophic cardiomyopathy - teen and adult	amber	Supporting evidence for the inferred MOI in the literature which is absent from PanelApp (PMID:34263907).	Automatic
DNAH11	Primary ciliary dyskinesia	1.000	Specific Disease	Moderate	Recessive	Primary ciliary disorders	green	Consistent.	Automatic
CCDC39	Respiratory ciliopathies	1.000	Disease Sub Group	High	Recessive	Primary ciliary disorders	green	Consistent.	Automatic
NYX	Rod Dysfunction Syndrome	1.000	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
SETD5	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
NFKB1	Primary immunodeficiency	1.000	Specific Disease	High	Dominant	Primary immunodeficiency	green	Consistent.	Automatic
PHYH	Other peroxisomal disorders	1.000	Specific Disease	Moderate	Recessive	Peroxisomal disorders	green	Consistent.	Automatic
PURA	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
RPS19	Congenital anaemias	1.000	Specific Disease	High	Dominant	Cytopenias and congenital anaemias	green	Consistent.	Automatic
PCARE	Rod-cone dystrophy	1.000	Specific Disease	High	Recessive	Retinal disorders	green	Consistent.	Automatic
DDX3X	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
TUBA1A	Neurodevelopmental disorders	1.000	Disease Sub Group	Moderate	Dominant	Malformations of cortical development	green	Consistent.	Automatic
SCN1A	Epilepsy plus other features	1.000	Specific Disease	High	Dominant	Genetic epilepsy syndromes	green	Consistent.	Automatic
BCL11A	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
EDA	Ectodermal dysplasia without a known gene mutation	1.000	Specific Disease	Moderate	Dominant	Ectodermal dysplasia without a known gene mutation	green	Supporting evidence for the inferred MOI in the literature which is absent from PanelApp (PMID:17256800).	Automatic
SPG7	Motor Disorders of the CNS	1.000	Disease Sub Group	Moderate	Recessive	Hereditary spastic paraplegia	green	Consistent.	Automatic
GPR143	Developmental macular and foveal dystrophy	1.000	Specific Disease	Moderate	Dominant	Retinal disorders	green	Consistent.	Automatic
BCOR	Cataracts	1.000	Specific Disease	High	Dominant	Cataracts	green	Consistent.	Automatic
COL4A4	Familial haematuria	1.000	Specific Disease	Moderate	Dominant	Haematuria	green	Consistent.	Automatic
TULP1	Rod-cone dystrophy	1.000	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic

LCA5	Leber Congenital Amaurosis or Early-Onset Severe Retinal Dystrophy	1.000	Specific Disease	High	Recessive	Retinal disorders	green	Consistent.	Automatic
ABCA12	Autosomal recessive congenital ichthyosis	1.000	Specific Disease	Moderate	Recessive	Autosomal recessive congenital ichthyosis	green	Consistent.	Automatic
IFITM5	Osteogenesis imperfecta	1.000	Specific Disease	5'UTR	Dominant	Osteogenesis imperfecta	green	Consistent.	Automatic
NR2E3	Rod-cone dystrophy	1.000	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
MYOT	Distal myopathies	1.000	Specific Disease	Moderate	Dominant	Distal myopathies	green	Consistent.	Automatic
ACO2	Inherited optic neuropathies	1.000	Specific Disease	Moderate	Dominant	Optic neuropathy	green	Supporting evidence for the inferred MOI in the literature which is absent from PanelApp (PMID:30689204).	Automatic
ZC4H2	Arthrogryposis	1.000	Specific Disease	High	Dominant	Arthrogryposis	green	Consistent.	Automatic
MECP2	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
LRRC7	Intellectual disability	1.000	Specific Disease	High	Dominant				
SACS	Hereditary ataxia	1.000	Specific Disease	High	Recessive	Hereditary ataxia	green	Consistent.	Automatic
WFS1	Inherited optic neuropathies	1.000	Specific Disease	Moderate	Dominant	Optic neuropathy	green	Consistent.	Automatic
TPO	Congenital hypothyroidism	1.000	Specific Disease	Moderate	Recessive	Congenital hypothyroidism	green	Consistent.	Automatic
CERS3	Autosomal recessive congenital ichthyosis	1.000	Specific Disease	Moderate	Recessive	Autosomal recessive congenital ichthyosis	green	Consistent.	Automatic
ADNP	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
ATP6V0A4	Renal tubular acidosis	1.000	Specific Disease	Moderate	Recessive	Nephrocalcinosis or nephrolithiasis	green	Consistent.	Automatic
ERG	Primary lymphoedema	1.000	Specific Disease	High	Dominant				
ODAD1	Primary ciliary dyskinesia	1.000	Specific Disease	High	Recessive	Primary ciliary disorders	green	Consistent.	Automatic
CERKL	Posterior segment abnormalities	1.000	Disease Sub Group	High	Recessive	Retinal disorders	green	Consistent.	Automatic
PDE6B	Rod-cone dystrophy	1.000	Specific Disease	High	Recessive	Retinal disorders	green	Consistent.	Automatic
CAPN3	Limb girdle muscular dystrophy	1.000	Specific Disease	Moderate	Recessive	Limb girdle muscular dystrophy	green	Consistent.	Automatic
RSPH4A	Respiratory ciliopathies	1.000	Disease Sub Group	High	Recessive	Primary ciliary disorders	green	Consistent.	Automatic
ACTG1	Non-syndromic hearing loss	1.000	Disease Sub Group	Moderate	Dominant	Hearing loss	green	Consistent.	Automatic
PRKCG	Hereditary ataxia	1.000	Specific Disease	Moderate	Dominant	Hereditary ataxia	green	Consistent.	Automatic
RPE65	Rod-cone dystrophy	1.000	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
UCHL1	Motor Disorders of the CNS	1.000	Disease Sub Group	High	Dominant	Hereditary ataxia	green	Alternative compatible panel found matching MOI: "Adult onset movement disorder".	Automatic
PIEZO2	Neuromuscular disorders	1.000	Disease Sub Group	High	Recessive	Neuromuscular disorders	green	Consistent.	Automatic
PROM1	Posterior segment abnormalities	1.000	Disease Sub Group	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
SYNE1	Hereditary ataxia	1.000	Specific Disease	High	Recessive	Hereditary ataxia	green	Consistent.	Automatic
IGF1R	IUGR and IGF abnormalities	1.000	Specific Disease	High	Dominant	IUGR and IGF abnormalities	green	Consistent.	Automatic
CYP27B1	Familial or syndromic hypoparathyroidism	1.000	Specific Disease	High	Recessive	Hypophosphataemia or rickets	green	Consistent.	Manual
PPP1CB	RASopathies	1.000	Disease Sub Group	Moderate	Dominant	RASopathies	green	Consistent.	Automatic
OTOA	Congenital hearing impairment	1.000	Specific Disease	Moderate	Recessive	Hearing loss	green	Consistent.	Automatic
COL18A1	Leber Congenital Amaurosis or Early-Onset Severe Retinal Dystrophy	1.000	Specific Disease	High	Recessive	Retinal disorders	green	Consistent.	Automatic
GMPPB	Limb girdle muscular dystrophy	1.000	Specific Disease	Moderate	Recessive	Limb girdle muscular dystrophy	green	Consistent.	Automatic
BEST1	Posterior segment abnormalities	1.000	Disease Sub Group	Moderate	Dominant	Retinal disorders	green	Consistent.	Automatic
UMOD	Unexplained kidney failure in young people	1.000	Specific Disease	Moderate	Dominant	Unexplained kidney failure in young people	green	Consistent.	Automatic
NKX2-1	Hereditary ataxia	1.000	Specific Disease	High	Dominant	Hereditary ataxia - adult onset	green	Consistent.	Manual
OTOF	Auditory Neuropathy Spectrum Disorder	1.000	Specific Disease	High	Recessive	Auditory Neuropathy Spectrum Disorde	green	Consistent.	Automatic
DYRK1A	Neurodevelopmental disorders	1.000	Disease Sub Group	High	Dominant	Intellectual disability	green	Consistent.	Automatic
FOXP1	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
FZD4	Familial exudative vitreoretinopathy	1.000	Specific Disease	High	Dominant	Retinal disorders	green	Consistent.	Automatic
TUFT1	Epidermolysis bullosa	1.000	Specific Disease	High	Recessive				
AHDC1	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
FRMD7	Infantile nystagmus	1.000	Specific Disease	Moderate	Recessive	Infantile nystagmus	green	Consistent.	Automatic
EXPH5	Epidermolysis bullosa	1.000	Specific Disease	High	Recessive	Epidermolysis bullosa	green	Consistent.	Automatic
CHD8	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
CELSR1	Lymphatic disorders	1.000	Disease Sub Group	High	Dominant	Primary lymphoedema	green	Consistent.	Automatic
SOX10	Congenital hearing impairment	1.000	Specific Disease	High	Dominant	Hearing loss	green	Consistent.	Automatic
CABP4	Cone Dysfunction Syndrome	1.000	Specific Disease	High	Recessive	Retinal disorders	green	Consistent.	Automatic
CHD2	Inherited Epilepsy Syndromes	1.000	Disease Sub Group	High	Dominant	Genetic epilepsy syndromes	green	Consistent.	Automatic
CEP290	Leber Congenital Amaurosis or Early-Onset Severe Retinal Dystrophy	1.000	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
MYT1L	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
TSC2	Classical tuberous sclerosis	1.000	Specific Disease	Moderate	Dominant	Classical tuberous sclerosis	green	Consistent.	Automatic
TGFBI	Corneal abnormalities	1.000	Specific Disease	Moderate	Dominant	Corneal abnormalities	green	Consistent.	Automatic
SELENON	Neuromuscular disorders	1.000	Disease Sub Group	Moderate	Recessive	Neuromuscular disorders	green	Consistent.	Automatic
KAT6A	Intellectual disability	1.000	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
NEB	Congenital myopathy	1.000	Specific Disease	High	Recessive	Congenital myopathy	green	Consistent.	Automatic
PRKN	Early onset and familial Parkinson's Disease	0.999	Specific Disease	High	Recessive	Parkinson Disease and Complex Parkinsonism	green	Consistent.	Automatic
DRAM2	Posterior segment abnormalities	0.999	Disease Sub Group	High	Recessive	Retinal disorders	amber	Consistent.	Automatic
TRIP12	Intellectual disability	0.999	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
RPGR	Rod-cone dystrophy	0.999	Specific Disease	Moderate	Dominant	Retinal disorders	green	Consistent.	Automatic
MITF	Non-syndromic hearing loss	0.999	Disease Sub Group	Moderate	Dominant	Hearing loss	green	Consistent.	Automatic
KCND3	Hereditary ataxia	0.999	Specific Disease	Moderate	Dominant	Hereditary ataxia	green	Consistent.	Automatic
SPTA1	Congenital anaemias	0.999	Specific Disease	High	Recessive	Cytopenias and congenital anaemias	green	Consistent.	Automatic
EYA1	Deafness and congenital structural abnormalities	0.999	Disease Sub Group	High	Dominant	Deafness and congenital structural abnormalities	green	Consistent.	Automatic
FAM222B	Ultra-rare undescribed monogenic disorders	0.999	Specific Disease	Moderate	Dominant				
TPM1	Cardiomyopathy	0.999	Disease Sub Group	Moderate	Dominant	Left Ventricular Noncompaction Cardiomyopathy	green	Consistent.	Automatic
MPPE1	Primary ciliary dyskinesia	0.999	Specific Disease	High	Recessive				
PAX6	Developmental macular and foveal dystrophy	0.999	Specific Disease	High	Dominant	Retinal disorders	red	Consistent.	Automatic
PLS3	Osteogenesis imperfecta	0.999	Specific Disease	High	Recessive	Osteogenesis imperfecta	green	Consistent.	Automatic
KCNV2	Posterior segment abnormalities	0.999	Disease Sub Group	High	Recessive	Retinal disorders	green	Consistent.	Automatic
SMCHD1	Limb girdle muscular dystrophy	0.999	Specific Disease	High	Dominant	Limb girdle muscular dystrophy	amber	Alternative compatible panel found matching MOI: "Severe Paediatric Disorders".	Automatic
LOX	Familial Thoracic Aortic Aneurysm Disease	0.999	Specific Disease	High	Dominant	Thoracic aortic aneurysm or dissection	green	Consistent.	Manual
PKHD1	Cystic kidney disease	0.998	Specific Disease	Moderate	Recessive	Cystic kidney disease	green	Consistent.	Automatic
UBAP1	Hereditary spastic paraplegia	0.998	Specific Disease	High	Dominant	Hereditary spastic paraplegia	green	Consistent.	Automatic
CYP4F22	Autosomal recessive congenital ichthyosis	0.998	Specific Disease	Moderate	Recessive	Autosomal recessive congenital ichthyosis	green	Consistent.	Automatic
SLC39A8	Intellectual disability	0.998	Specific Disease	Moderate	Dominant	Intellectual disability	green	No supporting evidence in the literature.	Automatic
WAS	Cytopenia and pancytopenia	0.998	Specific Disease	Moderate	Recessive	Cytopenia - NOT Fanconi anaemia	green	Consistent.	Manual
DSP	Dilated Cardiomyopathy	0.998	Specific Disease	High	Dominant	Arrhythmogenic cardiomyopathy	green	Consistent.	Automatic
CHM	Rod-cone dystrophy	0.998	Specific Disease	High	Dominant	Retinal disorders	green	Consistent.	Automatic
SRP9	Ductal plate malformation	0.998	Specific Disease	High	Dominant				
LRP5	Osteogenesis imperfecta	0.998	Specific Disease	Moderate	Dominant	Osteogenesis imperfecta	green	Consistent.	Automatic
ALMS1	Bardet-Biedl Syndrome	0.998	Specific Disease	High	Recessive	Renal ciliopathies	green	Consistent.	Automatic
TNNT2	Cardiomyopathy	0.998	Disease Sub Group	Moderate	Dominant	Left Ventricular Noncompaction Cardiomyopathy	green	Consistent.	Automatic
MFSD8	Hereditary ataxia	0.998	Specific Disease	Moderate	Recessive	Neuronal ceroid lipofuscinosis	green	Consistent.	Manual
MAP1B	Intellectual disability	0.998	Specific Disease	High	Dominant	Intellectual disability	amber	Consistent.	Automatic
RLBP1	Rod-cone dystrophy	0.998	Specific Disease	High	Recessive	Retinal disorders	green	Consistent.	Automatic
NAA15	Intellectual disability	0.998	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic

INPP5K	Neuromuscular disorders	0.998	Disease Sub Group	Moderate	Recessive	Neuromuscular disorders	green	Consistent.	Automatic
ECEL1	Arthrogryposis	0.997	Specific Disease	Moderate	Recessive	Arthrogryposis	green	Consistent.	Automatic
NUP107	Primary Microcephaly - Microcephalic Dwarfism Spectrum	0.997	Specific Disease	Moderate	Recessive	Severe microcephaly	amber	Consistent.	Automatic
CRYBB2	Cataracts	0.997	Specific Disease	Moderate	Dominant	Cataracts	green	Consistent.	Automatic
SLC24A5	Developmental macular and foveal dystrophy	0.997	Specific Disease	Moderate	Recessive	Retinal disorders	red	Alternative compatible panel found matching MOI: "Structural eye disease".	Automatic
FLNC	Dilated Cardiomyopathy	0.997	Specific Disease	High	Dominant	Arrhythmogenic cardiomyopathy	green	Consistent.	Automatic
MYO7A	Rod-cone dystrophy	0.997	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
BRPF1	Intellectual disability	0.997	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
CNGB1	Rod-cone dystrophy	0.997	Specific Disease	High	Recessive	Retinal disorders	green	Consistent.	Automatic
ERF	Intellectual disability	0.997	Specific Disease	High	Dominant	Intellectual disability	red	Consistent.	Automatic
INPP5E	Rod-cone dystrophy	0.997	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
GNE	Neuromuscular disorders	0.996	Disease Sub Group	Moderate	Recessive	Neuromuscular disorders	green	Consistent.	Automatic
SYNE4	Congenital hearing impairment	0.996	Specific Disease	Moderate	Recessive	Hearing loss	green	Consistent.	Automatic
TBX20	Left Ventricular Noncompaction Cardiomyopathy	0.996	Specific Disease	High	Dominant	Dilated cardiomyopathy - adult and teen	amber	Consistent.	Automatic
SPG11	Hereditary spastic paraplegia	0.996	Specific Disease	High	Recessive	Hereditary spastic paraplegia	green	Consistent.	Automatic
PAH	Undiagnosed metabolic disorders	0.996	Specific Disease	Moderate	Recessive	Undiagnosed metabolic disorders	green	Consistent.	Automatic
ADGRV1	Congenital hearing impairment	0.996	Specific Disease	High	Recessive	Hearing loss	green	Consistent.	Automatic
GJA3	Cataracts	0.996	Specific Disease	Moderate	Dominant	Cataracts	green	Consistent.	Automatic
BCL11B	Intellectual disability	0.995	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
ARPC3	Charcot-Marie-Tooth disease	0.995	Specific Disease	Moderate	Dominant				
MORC2	Motor and Sensory Disorders of the PNS	0.995	Disease Sub Group	Moderate	Dominant	Hereditary neuropathy	green	Consistent.	Automatic
NKX2-5	Left Ventricular Noncompaction Cardiomyopathy	0.995	Specific Disease	High	Dominant	Dilated cardiomyopathy - adult and teen	green	Consistent.	Automatic
SLC7A9	Renal tract calcification (or Nephrolithiasis or nephrocalcinosis)	0.995	Specific Disease	Moderate	Recessive	Nephrocalcinosis or nephrolithiasis	green	Consistent.	Automatic
CYP7B1	Hereditary spastic paraplegia	0.995	Specific Disease	Moderate	Recessive	Hereditary spastic paraplegia	green	Consistent.	Automatic
MME	Charcot-Marie-Tooth disease	0.995	Specific Disease	High	Dominant	Hereditary neuropathy	green	Alternative compatible panel found matching MOI: "Hereditary neuropathy NOT PMP22 copy number".	Automatic
GRIN2B	Intellectual disability	0.994	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
SCN2A	Intellectual disability	0.994	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
NDUFV1	Inherited white matter disorders	0.994	Specific Disease	Moderate	Recessive	Inherited white matter disorders	green	Consistent.	Automatic
CEP290	Congenital malformations caused by ciliopathies	0.994	Disease Sub Group	High	Recessive	Neurological ciliopathies	green	Consistent.	Automatic
PNPLA1	Autosomal recessive congenital ichthyosis	0.994	Specific Disease	Moderate	Recessive	Autosomal recessive congenital ichthyosis	green	Consistent.	Automatic
SCN8A	Intellectual disability	0.993	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
TCAP	Limb girdle muscular dystrophy	0.993	Specific Disease	High	Recessive	Limb girdle muscular dystrophy	green	Consistent.	Automatic
CDH23	Congenital hearing impairment	0.993	Specific Disease	Moderate	Recessive	Hearing loss	green	Consistent.	Automatic
UCHL1	Inherited optic neuropathies	0.992	Specific Disease	High	Dominant	Optic neuropathy	amber	Alternative compatible panel found matching MOI: "Neurodegenerative disorders - adult onset".	Automatic
ATF7	Arthrogryposis	0.991	Specific Disease	High	Dominant				
LSM6	Familial congenital heart disease	0.990	Specific Disease	High	Dominant				
LZTR1	Noonan syndrome	0.990	Specific Disease	Moderate	Dominant	RASopathies	green	Consistent.	Automatic
ANO10	Hereditary ataxia	0.990	Specific Disease	High	Recessive	Hereditary ataxia	green	Consistent.	Automatic
GRHPR	Renal tract calcification (or Nephrolithiasis or nephrocalcinosis)	0.990	Specific Disease	High	Recessive	Nephrocalcinosis or nephrolithiasis	green	Consistent.	Automatic
SERPINB3	Autosomal recessive congenital ichthyosis	0.990	Specific Disease	Moderate	Recessive				
SH3TC2	Charcot-Marie-Tooth disease	0.989	Specific Disease	Moderate	Recessive	Hereditary neuropathy	green	Consistent.	Automatic
PKMYT1	Single autosomal recessive mutation in rare disease	0.989	Specific Disease	Moderate	Recessive				
MFN2	Charcot-Marie-Tooth disease	0.989	Specific Disease	Moderate	Dominant	Hereditary neuropathy	green	Consistent.	Automatic
VPS13B	Intellectual disability	0.989	Specific Disease	High	Recessive	Intellectual disability	green	Consistent.	Automatic
PKDCC	Unexplained skeletal dysplasia	0.987	Specific Disease	High	Recessive	Skeletal dysplasia	amber	Consistent.	Automatic
IFT140	Posterior segment abnormalities	0.987	Disease Sub Group	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
NHS	Cataracts	0.986	Specific Disease	High	Recessive	Cataracts	green	Consistent.	Automatic
SPATA7	Posterior segment abnormalities	0.986	Disease Sub Group	High	Recessive	Retinal disorders	green	Consistent.	Automatic
PKLR	Congenital anaemias	0.986	Specific Disease	Moderate	Recessive	Cytopenias and congenital anaemias	green	Consistent.	Automatic
CACNA1A	Hereditary ataxia	0.985	Specific Disease	Moderate	Dominant	Hereditary ataxia	green	Consistent.	Automatic
RAB3A	Hereditary ataxia	0.985	Specific Disease	Moderate	Dominant				
ACTA1	Congenital myopathy	0.985	Specific Disease	Moderate	Dominant	Congenital myopathy	green	Consistent.	Automatic
TSPAN12	Familial exudative vitreoretinopathy	0.984	Specific Disease	High	Dominant	Retinal disorders	green	Consistent.	Automatic
GHRHR	IUGR and IGF abnormalities	0.983	Specific Disease	High	Recessive	IUGR and IGF abnormalities	green	Consistent.	Automatic
SEC63	Ductal plate malformation	0.981	Specific Disease	High	Dominant	Ductal plate malformation	green	Consistent.	Automatic
MICU1	Congenital myopathy	0.981	Specific Disease	High	Recessive	Congenital myopathy	green	Consistent.	Automatic
BMPR2	Pulmonary arterial hypertension	0.980	Specific Disease	High	Dominant	Pulmonary arterial hypertension	green	Consistent.	Automatic
SLC38A8	Infantile nystagmus	0.978	Specific Disease	High	Recessive	Albinism or congenital nystagmus	green	Consistent.	Manual
GNB1	Intellectual disability	0.977	Specific Disease	Moderate	Dominant	Intellectual disability	green	Consistent.	Automatic
USP33	Extreme early-onset hypertension	0.977	Specific Disease	High	Dominant				
RP1L1	Posterior segment abnormalities	0.975	Disease Sub Group	High	Recessive	Retinal disorders	green	Consistent.	Automatic
PYROXD1	Limb girdle muscular dystrophy	0.975	Specific Disease	Moderate	Recessive	Limb girdle muscular dystrophy	green	Consistent.	Automatic
INSL4	Rod Dysfunction Syndrome	0.974	Specific Disease	Moderate	Dominant				
SLC34A3	Renal tract calcification (or Nephrolithiasis or nephrocalcinosis)	0.974	Specific Disease	Moderate	Recessive	Nephrocalcinosis or nephrolithiasis	green	Consistent.	Automatic
ALOX12B	Autosomal recessive congenital ichthyosis	0.974	Specific Disease	Moderate	Recessive	Autosomal recessive congenital ichthyosis	green	Consistent.	Automatic
FOXJ1	Non-CF bronchiectasis	0.974	Specific Disease	High	Dominant	Respiratory ciliopathies including non-CF bronchiectasis.	amber	Consistent.	Manual
SSBP1	Inherited optic neuropathies	0.974	Specific Disease	Moderate	Dominant	Optic neuropathy	green	Consistent.	Automatic
OBSL1	IUGR and IGF abnormalities	0.970	Specific Disease	High	Recessive	IUGR and IGF abnormalities	green	Consistent.	Automatic
KRT14	Young onset tumour syndromes	0.970	Disease Sub Group	Moderate	Recessive				
PLA2G6	Cerebellar hypoplasia	0.967	Specific Disease	Moderate	Recessive	Structural basal ganglia disorders	green	Consistent.	Automatic
HBB	Renal tract calcification (or Nephrolithiasis or nephrocalcinosis)	0.967	Specific Disease	High	Recessive	Groopman et al 2019 - Genes with diagnostic variants	green	Consistent.	Manual
AICDA	Primary immunodeficiency	0.967	Specific Disease	High	Dominant	Primary immunodeficiency	green	Supporting evidence for the inferred MOI in the literature which is absent from PanelApp (PMID: 35246784).	Automatic
EHMT1	Intellectual disability	0.966	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
SOD1	Amyotrophic lateral sclerosis or motor neuron disease	0.966	Specific Disease	Moderate	Dominant	Amyotrophic lateral sclerosis/motor neuron disease	green	Consistent.	Automatic
DNAAF5	Primary ciliary dyskinesia	0.964	Specific Disease	Moderate	Recessive	Primary ciliary disorders	green	Consistent.	Automatic
WDR62	Primary Microcephaly - Microcephalic Dwarfism Spectrum	0.962	Specific Disease	Moderate	Recessive	Severe microcephaly	green	Consistent.	Automatic
CACNA1S	Congenital myopathy	0.962	Specific Disease	High	Recessive	Congenital myopathy	green	Consistent.	Automatic
CDH23	Rod-cone dystrophy	0.960	Specific Disease	Moderate	Recessive	Retinal disorders	green	Consistent.	Automatic
CNKSR2	Intellectual disability	0.960	Specific Disease	High	Dominant	Intellectual disability	green	Alternative compatible panel found matching MOI: "Severe Paediatric Disorders".	Automatic
WWOX	Gastrointestinal disorders	0.958	Disease Sub Group	Moderate	Recessive				
GPR156	Congenital hearing impairment	0.958	Specific Disease	High	Recessive				
CDH3	Inherited macular dystrophy	0.957	Specific Disease	High	Recessive	Retinal disorders	green	Consistent.	Automatic
UBAP1L	Inherited macular dystrophy	0.957	Specific Disease	High	Recessive	Retinal disorders	red	Supporting evidence for the inferred MOI in the literature which is absent from PanelApp (PMID:28041643).	Automatic
FMN1	Congenital hearing impairment	0.956	Specific Disease	High	Recessive				
PMEPA1	Familial Thoracic Aortic Aneurysm Disease	0.955	Specific Disease	High	Dominant				
RAB35	Familial Hypercholesterolaemia	0.955	Specific Disease	High	Dominant				
AUTS2	Intellectual disability	0.954	Specific Disease	High	Dominant	Intellectual disability	green	Consistent.	Automatic
GATA6	Congenital heart disease	0.952	Disease Sub Group	High	Dominant	Familial non syndromic congenital heart disease	green	Consistent.	Automatic