Rare disease gene association discovery from burden analysis of the 100,000 Genomes Project data

Dr. Valentina Cipriani PhD[1,2,3,§], Dr. Letizia Vestito PhD[1,§], Dr. Emma F Magavern MD, MSc, MRCP[1], Dr. Julius OB Jacobsen PhD[1], Dr. Gavin Arno PhD[4,2], Professor. Elijah R Behr MA, MBBS, MD, FRCP, FESC[5,6], Dr. Katherine A Benson PhD[7], Dr. Marta Bertoli MD[8], Prof. Detlef Bockenhauer PhD[9,10], Dr. Michael R Bowl DPhil[11], Dr. Kate Burley PhD[12], Prof. Li F Chan MB, BChir, PhD[13], Prof. Patrick Chinnery[14], Prof. Peter Conlon MB, DMed[15], Mr. Marcos Costa MSc[2], Dr. Alice E Davidson PhD[2], Prof. Sally J Dawson PhD[11], Dr. Elhussein Elhassan MBBS[15], Prof. Sarah E Flanagan PhD[16], Dr. Marta Futema BSc, PhD[5,17], Prof. Daniel P Gale PhD FRCP[18], Dr. Sonia García-Ruiz BSc, MSc, PhD[19,20,21], Prof. Cecilia Gonzalez Corcia MD, PhD[22,23], Dr. Helen R Griffin PhD[24], Prof. Sophie Hambleton FRCPCH, DPhil[24,25], Ms. Amy R Hicks BSc[19,20,21], Prof. Henry Houlden MD, PhD[26,27], Prof. Richard S Houlston MD, PhD[28], Dr. Sarah A Howles DPhil, FRCS[29], Prof. Robert Kleta MD, PhD[18], Iris Lekkerkerker MD[30], Dr. Siying Lin MBBS, PhD[4,2], Prof. Petra Liskova MD, PhD[31,32], Prof. Hannah Mitchison PhD[20], Dr. Heba Morsy MB ChB, PhD[33,34], Prof. Andrew D Mumford PhD, MB ChB[12], Prof. William G Newman MB ChB, PhD[35,36], Miss. Ruxandra Neatu MSc[37], Prof. Edel A O'Toole MB, BCh, PhD, FRCP[38], Prof. Albert CM Ong DM, FRCP[39,40], Dr. Alistair T Pagnamenta PhD[41,42], Prof. Shamima Rahman FRCP, FRCPCH, PhD[20], Prof. Neil Rajan MBBS, PhD[24,43], Prof. Peter N Robinson MD, PhD[44], Prof. Mina Ryten MBBS, PhD, FRCP[20,45,21], Dr. Omid Sadeghi-Alavijeh MBBS[18], Prof. John A Sayer MB ChB, PhD[46,47,48], Prof. Claire L Shovlin PhD FRCP[49], Prof. Jenny C Taylor PhD[41,42], Dr. Omri Teltsh PhD[7], Prof. Ian Tomlinson FRS, FMedSci[50], Dr. Arianna Tucci MD, PhD[1], Prof. Clare Turnbull PhD, FRCP, FRCPath[51], Dr. Albertien M van Eerde MD, PhD[30], Prof. James S Ware PhD, MRCP[52,53], Dr. Laura M Watts PhD BMBCh[41,54], Prof. Andrew R Webster FRCOphth[2,55], Dr. Sarah K Westbury PhD, FRCP[56], Dr. Sean L Zheng MSc, MRCP[52,53], Prof. Mark Caulfield FRCP, FMedSci[1], Prof. Damian Smedley PhD[1,*]

[1]William Harvey Research Institute, Clinical Pharmacology and Precision Medicine, Queen Mary University of London, London, EC1M 6BQ, United Kingdom, [2]UCL Institute of Ophthalmology, University College London, London, EC1V 9EL, United Kingdom, [3]UCL Genetics Institute, University College London, London, WC1E 6BT, United Kingdom, [4]National Institute of Health Research Biomedical Research Centre at Moorfields Eye Hospital, London, EC1V 2PD, United Kingdom, [5]Cardiology Section, Molecular and Clinical Science Institute, St George's, University of London, London, SW17 0RE, United Kingdom, [6]Cardiology Department, St George's University Hospitals NHS Foundation Trust, London, SW17 0QT, United Kingdom, [7]School of Pharmacy and Biomolecular Sciences, Royal College of Surgeons in Ireland, Dublin, D02 YN77, Republic of Ireland, [8]Northern Genetics Centre, The Newcastle upon Tyne NHS Foundation Trust, Newcastle upon Tyne, NE1 3BZ, United Kingdom, [9]Paediatric Nephrology, University Hospital and Katholic University Leuven, Leuven, 3000, Belgium, [10]Deptartment of Renal Medicine, University College London, London, NW3 2PF, United Kingdom, [11]UCL Ear Institute, University College London, London, WC1X 8EE, United Kingdom, [12]School of Cellular and Molecular Medicine, University of Bristol, Bristol, BS8 1TD, United Kingdom, [13]William Harvey Research Institute, Centre for Endocrinology, Queen Mary University of London, London, EC1M 6BQ, United Kingdom, [14]Medical Research Council Mitochondrial Biology Unit, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB2 0XY, United Kingdom, [15]Department of Nephrology, Beaumont Hospital, Dublin, D02 YN77, Republic of Ireland, [16]Department of Clinical and Biomedical Science, University of Exeter Medical School, Exeter, EX2 5DW, United Kingdom, [17]Institute of Cardiovascular Science, University College London, London, WC1E 6DD, United Kingdom, [18]Department of Renal Medicine, University College London, London, NW3 2PF,

United Kingdom, [19]Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology, University College London, London, WC1N 3BG, United Kingdom, [20]Genetics and Genomic Medicine, UCL Great Ormond Street Institute of Child Health, University College London, London, WC1N 1EH, United Kingdom, [21]Aligning Science Across Parkinson's (ASAP) Collaborative Research Network, Chevy Chase, MD, 20815, USA, [22]Pediatric Cardiology, CHU Sainte Justine, University of Montreal, Montreal, H3T 1C5, Canada, [23]Mc Gill University, Montreal, H3A 0G4, Canada, [24]Newcastle University Translational and Clinical Research Institute, Newcastle upon Tyne, NE2 4HH, United Kingdom, [25]Great North Children's Hospital, Newcastle upon Tyne, NE1 4LP, United Kingdom, [26]UCL Institute of Neurology, London, WC1N 3BG, United Kingdom, [27]The National Hospital for Neurology and Neurosurgery, Queen Square, London, WC1N 3BG, United Kingdom, [28]Division of Genetics and Epidemiology, The Institute of Cancer Research, London, SW3 6JB, United Kingdom, [29]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, OX3 7LE, United Kingdom, [30]UMC Utrecht, Department of Genetics, Utrecht, Netherlands, [31]Department of Ophthalmology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague, Czech Republic, [32]Department of Paediatrics and Inherited Metabolic Disorders, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague, Czech Republic, [33]Department of Neuromuscular Disorders, UCLInstitute of Neurology, London, WC1N 3BG, United Kingdom, [34]London, WC1N 3BG, United Kingdom, [35]Division of Evolution, Infection & Genomics, University of Manchester, Manchester, M13 9PL, United Kingdom, [36]Manchester Centre for Genomic Medicine, Manchester University NHS Foundation Trust, Manchester, M13 9WL, United Kingdom, [37]Institute of Translational and Clinical Research, Newcastle University, Newcastle upon Tyne, NE1 3BZ, United Kingdom, [38]Centre for Cell Biology and Cutaneous Research, Blizard Institute, QMUL, London, E1 2AT, United Kingdom, [39]Kidney Genetics Group, Division of Clinical Medicine, School of Medicine and Population Health, University of Sheffield, Sheffield, United Kingdom, [40]Sheffield Kidney Institute, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, United Kingdom, [41]Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, United Kingdom, [42]Oxford NIHR Biomedical Research Centre, Oxford, OX3 9DU, United Kingdom, [43]Department of Dermatology and NIHR Biomedical Research Centre, Royal Victoria Infirmary, Newcastle upon Tyne, NE1 4LP, United Kingdom, [44]The Jackson Laboratory for Genomic Medicine, Farmington, CT, 6032, USA, [45]NIHR GOSH Biomedical Research Centre, Great Ormond Street Institute of Child Health, London, WC1N 1EH, United Kingdom, [46]Biosciences Institute, Newcastle University, Newcastle upon Tyne, NE1 3BZ, United Kingdom, [47]Renal Services, The Newcastle upon Tyne NHS Foundation Trust Hospitals, Newcastle upon Tyne, NE7 7DN, United Kingdom, [48]NIHR Biomedical Research Centre, Newcastle University, Newcastle upon Tyne, NE4 5PL, United Kingdom, [49]National Heart and Lung Institute, Imperial College London, London, W12 ONN, United Kingdom, [50]Dept of Oncology, University of Oxford, Old Road Campus Research Building, Roosevelt Drive, Oxford, OX3 7DQ, United Kingdom, [51]Division of Genetics and Epidemiology, Institute of Cancer Research, [52]National Heart and Lung Institute, Imperial College London, London, W12 0NN, United Kingdom, [53]MRC Laboratory of Medical Sciences, Imperial College London, London, W12 0HS, United Kingdom, [54]Oxford Centre for Genomic Medicine, Oxford University Foundation Trust, Oxford, OX3 7HE, United Kingdom, [55]National Institute of Health Research Biomedical Research Centre at Moorfields Eye Hospital, London, EC1V 9EL, United Kingdom, [56]Bristol Medical School, University of Bristol, Bristol, BS8 1UD, United Kingdom

§ Jointly contributed

* Corresponding author: d.smedley@qmul.ac.uk

Abstract

To discover rare disease-gene associations, we developed a gene burden analytical framework and applied it to rare, protein-coding variants from whole genome sequencing of 35,008 cases with rare diseases and their family members recruited to the 100,000 Genomes Project (100KGP). Following *in silico* triaging of the results, 88 novel associations were identified including 38 with existing experimental evidence. We have published the confirmation of one of these associations, hereditary ataxia with *UCHL1,* and independent confirmatory evidence has recently been published for four more. We highlight a further seven compelling associations: hypertrophic cardiomyopathy with *DYSF* and *SLC4A3* where both genes show high/specific heart expression and existing associations to skeletal dystrophies or short QT syndrome respectively; monogenic diabetes with *UNC13A* with a known role in the regulation of β cells and a mouse model with impaired glucose tolerance; epilepsy with *KCNQ1* where a mouse model shows seizures and the existing long QT syndrome association may be linked; early onset Parkinson's disease with *RYR1* with existing links to tremor pathophysiology and a mouse model with neurological phenotypes; anterior segment ocular abnormalities associated with *POMK* showing expression in corneal cells and with a zebrafish model with developmental ocular abnormalities; and cystic kidney disease with *COL4A3* showing high renal expression and prior evidence for a digenic or modifying role in renal disease. Confirmation of all 88 associations would lead to potential diagnoses in 456 molecularly undiagnosed cases within the 100KGP, as well as other rare disease patients worldwide, highlighting the clinical impact of a large-scale statistical approach to rare disease gene discovery.

Rare diseases collectively affect one in seventeen individuals in the United Kingdom [1]. Despite advances in genomic sequencing, molecular diagnosis continues to elude 50% to 80% of patients presenting to genetics clinics [2]. Furthermore, less than half of the 10,000 rare Mendelian diseases in the Online Mendelian Inheritance in Man (OMIM) database [3] have an established genetic basis. Diagnostic failure may arise due to a lack of routine screening for non-coding [1] or structural variants [4]. However, it is likely that a substantial proportion of the pathogenic variants responsible for undiagnosed cases reside in those yet to be discovered (possibly very rare) disease-associated genes. The scale of rare disease sequencing studies such as the Undiagnosed Disease Network [5], Centers for Mendelian Genomics [6], Deciphering Developmental Disorders [7] and the 100,000 Genomes Project (100KGP) [8], offers expanded opportunities to provide insight into pathogenic mechanisms of inherited disease, including the possibility of establishing disease-gene associations through case-control analyses, akin to methods previously used to identify common genetic variants influencing the risk of complex disorders. Such an approach provides much-needed power to identify genes harbouring rare pathogenic variants.

To identify disease-associated genes, we recently developed a framework that analyses rare protein coding variants identified by the Exomiser prioritisation tool [9], within a preliminary version of the 100KGP data [4], to conduct gene-based burden testing of single probands and family members relative to control families. *In silico* triage highlighted 22 novel disease-gene associations, three of which have been also reported in independent studies [10–12].

We have now extended our gene burden analytical framework for generic application to any large-scale, rare disease sequencing cohorts and complemented it with visualisation scripts. In addition, we report on the application of the approach to a larger cohort from the final 100KGP data including 35,008 families, 226 rare diseases and a pool of 4,676,866 rare candidate variants with improved *in silico* and added clinical expert triage of 88 probable new disease-gene associations.

Results

Gene burden analytical framework for disease gene discovery

We have developed an open-source R framework (https://github.com/whri-phenogenomics/geneBurdenRD) allowing users to perform gene burden testing of variants in user-defined cases versus controls from rare disease sequencing cohorts. The input to the framework is a file obtained from processing Exomiser output files for each of the cohort samples, a file containing a label for each case-control association analysis to perform within the cohort and a (set of) corresponding file(s) with user-defined identifiers and case/control assignment per each sample. Cases and controls in a cohort could be defined in many ways, for example, by recruited disease category as we have done for the 100KGP analysis below, by specific phenotypic annotations or phenotypic clustering. The framework will then assess false discovery rate (FDR)-adjusted disease-gene associations where genes are tested for an enrichment in cases vs controls of rare, protein-coding, segregating variants that are either (i) predicted loss-of-function (LoF), (ii) highly predicted pathogenic (Exomiser variant score >= 0.8), (iii) highly predicted pathogenic and present in a constrained coding region (CCR; [13]) or (iv) *de novo* (restricted to only trios or larger families where *de novo* calling was possible and provided by the user) (**Methods**). As well as various output files annotating these case-control association tests, Manhattan and volcano plots are generated summarising the FDR-adjusted p-values of all the gene-based tests for each case-control association analysis, along with lollipop plots of the relevant variants in cases and controls and plots of the hierarchical distribution of the Human Phenotype Ontology (HPO) case annotations for individual disease-gene associations.

Application to the 100,000 Genomes Project data identifies 88 novel disease-gene associations

A rare variant gene-burden analysis was performed on a cohort of 35,008 families (73,018 genomes) from the 100KGP rare disease pilot and main programme (Data Release v.11)

(**Fig. 1**, **Supplementary Table 1**). A pool of 4,676,866 rare, protein-coding, segregating and most predicted pathogenic (per gene) variants for the analysis was derived by running Exomiser for each family. Our pipeline was then used to detect statistically significant gene-based enrichment in relevant variant categories (predicted LoF, highly predicted pathogenic, highly predicted pathogenic in CCR regions, *de novo*) for *cases* in each of 226 'specific diseases' used for patient recruitment by the 100KGP [4] versus *controls* who were defined as probands from any other 20 broad 'disease groups' in the project (e.g. intellectual disability cases were compared to all non-neurological probands as controls). Applied cutoffs required at least 5 case probands per 'specific disease' and at least 4 probands (of which, at least one *case*) with a relevant rare variant per disease-gene burden test (**Methods**).

We identified 190 previously known and 316 novel potential disease-gene associations (**Fig. 2**), imposing a 5% FDR. Not previously known (novel) signals were initially defined, at the first round of analysis in March 2021, as having no documented evidence for an association within OMIM and absence from the 'specific disease' curated panel of high confidence (green) genes in PanelApp [14]. Enrichment of predicted LoF, highly predicted pathogenic, also in CCR and *de novo* variants was observed in 53%, 28%, 6% and 13% of known and 19%, 80%, 1%, and 0% of novel associations respectively, revealing discovery was mostly driven by predicted pathogenic, missense variants.

Potential associations were further filtered to 88/316 (28%) by removing those in which (i) the gene was a non-protein coding RNA one (31/316); (ii) for signals driven by dominant, LoF variants where the Genome Aggregation Database (gnomAD) [15] suggests there is no evidence for haploinsufficiency (i.e. gnomAD oe_lof >= 0.5) (44/316); (iii) > 33.3% of the variants driving the signal in the cases were present in controls (179/316) to exclude false positives unless penetrance is very low; or (iv) > 50% of the cases driving the signals had already received an alternative genetic diagnosis (36/316) (**Fig. 1** and **Supplementary Table 2**). Molecularly diagnosed individuals with a report as part of the 100KGP routine diagnostic pipeline were included in the experimental pipeline to evaluate the approach via detection of known associations and to allow for the possibility of misdiagnosis. Relatively relaxed thresholds were chosen for the last two criteria (iii and

iv) to avoid restricting associations based on two or three cases, but most were far below the relevant thresholds, e.g. 48/88 had no occurrence of putative disease-causing variants in controls and in 50/87 less than 5% of cases were otherwise assigned a molecular diagnosis. However, for all 12 cystic kidney disease associations 20-50% of cases had received a diagnosis involving well known genes such as *PKD1*, raising the possibility of digenic inheritance or variants in other genes modifying penetrance as a common theme (discussed further for the *COL4A3* example below). Similarly, most of the independent Irish renal cohort cases described in **Supplementary Table 2** already had an established molecular diagnosis. In comparison to the novel association signals, 174/190 (92%) of signals from known disease genes passed criteria from (i) to (iii). Given the relatively mixed ethnic diversity of the UK population and our cohort, we investigated if any of the 88 associations in **Supplementary Table 2** were linked to particular ancestries. For most, no significant enrichments were detected, but the *DYSF* association had 25% Asian/Asian British and 12.5% Black/Black British cases, the *DSG3* association had 27% Black/Black British cases, and both *IFN10* associated cases were Black/Black British. Different variants were observed in all these examples.

An extensive review of the literature as well as the phenotype evidence from Exomiser, in collaboration with members of the Genomics England Clinical Interpretation Partnerships (GeCIPs), was performed to identify supporting evidence from each of the 88 genes' biological function, known disease associations or the phenotypes of the gene-deficient mouse and/or other animal models. *In silico* analyses were also undertaken to identify high quality StringDB [16] protein-protein associations between the gene signal and any other genes known to be associated with the disease, or with highly specific expression in the most relevant tissue for the disease. This combined curation highlighted 38 associations supported by experimental evidence: 29 based on literature curation of a gene function fitting the likely disease mechanism with additional lines of evidence for many, seven based on mouse models and other evidence for some, and two based on protein-protein evidence only (**Fig. 1** and highlighted in bold in the summary column in **Supplementary Table 2**).

ClinGen [17] has developed a robust set of criteria to assess the evidence for disease-gene associations and we applied these to our 88 associations. Evidence of causality was strong for four associations, moderate for 34, and limited for the remainder (**Fig. 1 and Supplementary Table 2**). The four associations with strong evidence were familial thoracic aortic aneurysm disease and *UGGT2*; cystic kidney disease with *DNAH8* and *COL4A3*; hereditary ataxia and *UCHL1*. The latter two are described below but the first two are driven almost purely by the genetic evidence score due to many case variants, with little or no experimental evidence.

Four of our 88 novel signals have had interim independent confirmatory evidence emerge since our initial assessment was performed in 2021: hereditary ataxia associated with *TUBA4A* [18]; mitochondrial disorders with *MORC2* [19]; renal tract calcification with *SLC34A3* [20]; and epidermolysis bullosa with *TUFT1* [21]. In an interim published collaborative work [22], we were also able to confirm the association between heterozygous variants in *UCHL1* and a specific form of hereditary ataxia associated with neuropathy and optic atrophy, which is distinct from the spastic paraplegia associated with recessive variants in the same gene. The signal was driven by rare LoF variants in five cases and the gnomAD haploinsufficiency evidence is particularly compelling with no observed LoF variants (o/e = 0 (0-0.21) and pLI = 0.99). A heterozygous KO mouse model with neurological phenotypes also exists [23]. Through a network of collaborators in the UK and Germany, we were able to identify a total of 34 cases from 18 unrelated families and confirmed a 50% reduction in expression through mass spectrometry-based proteomics on patient fibroblasts.

Of the remaining 83 novel associations, nine had prior functional data fitting the likely disease mechanism and a high ClinGen classification score (>= 8) (**Supplementary Table 2**). Further investigation of the association between congenital heart disease and *DCP1A* revealed that the case variants are likely to be false positive calls resulting from two neighbouring inframe deletions. Investigation of the recessive association between *HBB* variants and renal tract calcification highlighted that nearly all the identified variants are reported to be known pathogenic/likely pathogenic in ClinVar and linked to beta-thalassemia; the patients included in our analyses were also clinically described with

thalassemia. Therefore, despite previous descriptions of the association of *HBB* with kidney stones [24] and nephropathy of unknown cause [25], we propose that our analyses have detected an association of thalassemia with renal tract calcification, appearing as a new association possibly more due to a treatment effect [26]. The remaining seven novel compelling candidates are described in the following sections.

Hypertrophic cardiomyopathy associated with *DYSF*

We identified a dominant association between variants in the muscle fiber repair gene Dysferlin (*DYSF)* and 'specific disease' *hypertrophic cardiomyopathy* (HCM). The association is driven by rare, predicted pathogenic variants throughout the gene (22 missense, 1 LoF, 1 intronic) in 24 cases (**Fig. 3a,** odds ratio (OR) = 2.8 [95% confidence interval (CI): 1.8 – 4.3], **Supplementary Table 2**). To assess if participants with predicted pathogenic mutations in this gene had a phenotypically distinct sub-endotype, we performed pairwise similarity by HPO term calculations of all the HCM cases as described in the **Methods**. The 24 cases associated with the *DYSF* signal shared a similar clinical syndrome (mean PhenoDigm score from pairwise, reciprocal, non-self hits was 0.67) with respiratory and metabolic phenotypes in addition to the cardiomyopathy. However, 301/1,000 randomly sampled HCM sets of the same size also achieved the same mean score or higher, suggesting the *DYSF* families are not a phenotypically distinct sub-group of HCM based on the available HPO annotations. Co-segregation was apparent in two trios where (i) a heterozygous, ClinVar known pathogenic intronic c.5667-824C>T variant [27] is inherited from the affected father in the male proband, (ii) a heterozygous p.Pro810Ser variant, classified as a variant of uncertain significance (VUS) in ClinVar, is inherited from the affected father in the male proband (**Fig. 3a**). *DYSF* shows strongest expression in skeletal muscle, but with significant levels of expression in heart muscle as well (**Fig. 3a**). In addition, co-expression network analysis of *DYSF* and known HCM genes from PanelApp shows the most significant enrichment in the heartAtrial Genotype Tissue Expression (GTEx) v6 module (FDR adjusted p-value of 3.6×10^{-13}), with *DYSF* co-expressed with 10/22 of the known genes (**Fig. 3a**). LoF in *DYSF is a* known cause of autosomal recessive limb-girdle muscular dystrophy (LGMD) 2 (OMIM:253601), although there is evidence for a cardiomyopathic component and a milder, later onset form for

heterozygous carriers [28]. The association between *DYSF* and HCM driven by heterozygous missense, rather than LoF, variants suggests that a milder, cardiomyopathy-only phenotype, rather than LGMD, is possible.

Hypertrophic cardiomyopathy associated with *SLC4A3*

We identified another dominant association between variants in *SLC4A3* and 'specific disease' HCM. The association is driven by rare, predicted pathogenic variants (8 missense, 1 splice region) in 9 cases located throughout the gene (**Fig. 3b,** OR = 5.9 [95% CI: 2.9 – 12.0], **Supplementary Table 2**). One family contains the same p.Gly728Glu variant in the proband and her affected maternal aunt. A p.Arg938His variant is seen in two of the other cases. The gene is depleted for rare missense variants in gnomAD (o/e = 0.73 (0.68-0.78) and Z = 2.74). *SLC4A3*, a plasma membrane anion exchange protein, has recently been associated with short QT syndrome [29] and exhibits highly specific expression in heart muscle. None of our cases had recorded short QT phenotypes, but the underlying ECG data was not available for closer evaluation. The 9 cases show strong phenotypic similarity to each other (mean PhenoDigm score from pairwise, reciprocal, non-self hits was 0.61) with palpitations/arrhythmia phenotypes observed in 6 cases in addition to the HCM. However, 596/1000 randomly sampled case sets of the same size achieved the same mean score or higher, suggesting the *SLC4A3* families are not phenotypically distinct from the other HCM cases based on the available HPO terms.

Monogenic diabetes associated with *UNC13A*

We identified a dominant association between variants in *UNC13A* and 'specific disease' *diabetes with additional phenotypes suggestive of a monogenic aetiology.* The association is driven by rare predicted LoF variants in 2 singleton cases with the only recorded phenotypes being diabetes mellitus in both and one further phenotype in one: p.Ala53Serfs*50 and p.Gly44* (**Fig. 3c,** OR = 355.1 [95% CI: 71.7 – 1759.5], **Supplementary Table 2**). The gene is depleted for rare LoF variants in gnomAD (o/e = 0.09 (0.05-0.16) and pLI = 1). *UNC13A* is a diacylgycerol and phorbol ester receptor with evidence for a role in regulation of β cells [30]. Neonatal pancreatic β cells extracted from

UNC13A-knockout mice and knock-in mice that lacked the DAG binding domain showed impaired second phase of insulin secretion in response to glucose stimulation [31] and the heterozygous mouse knockout model shows impaired glucose tolerance [32]. In addition, co-expression network analysis of *UNC13A* and known monogenic diabetes genes from PanelApp shows the most significant enrichment in the pancreas GTEx v6 module (FDR adjusted p-value of 0.01), with *UNC13A* co-expressed with 6/43 of the known genes. However, predicted LoF variants (3 splice site, 5 stop gain or frameshift) were also seen in controls with no apparent history of diabetes suggesting either incomplete penetrance, later onset (year of birth of the two was 1973 and 1974 compared to 1956-2007 (mean 1980) for controls), or that the variants in controls are not genuinely LoF.

Epilepsy associated with *KCNQ1*

We identified a dominant association between variants in *KNCQ1* and 'specific disease' *epilepsy plus other features*. The association is driven by rare, predicted pathogenic variants (20 missense, 1 LoF) in 21 cases, located throughout the gene (**Fig. 4a,** OR = 3.0 [95% CI: 1.9 – 4.8], **Supplementary Table 2**). *KCNQ1* is depleted for rare missense variants in gnomAD (o/e = 0.82 (0.77-0.87) and Z = 1.99). The 21 cases show strong phenotypic similarity to each other (mean PhenoDigm score from pairwise, reciprocal, non-self hits was 0.63). Only 62/1000 randomly sampled epilepsy sets of the same size achieved the same mean score or higher, suggesting the *KCNQ1* families with common features of seizures, deterioration of higher mental function and cerebral morphology abnormalities are phenotypically distinct from the other epilepsy cases. The only co-segregation evidence comes from a duo where the p.His509Gln variant in the proband is inherited from an affected mother. *KNCQ1* is a voltage-gated potassium channel required for the repolarization phase of the cardiac action potential and associated with autosomal dominant long-QT syndrome sub-type 1 (OMIM:192500), while a mouse model supports the potential for associated epileptic seizures [33]. Other voltage-gated potassium channel family members, *KCNQ2* (OMIM:613720), *KCNQ3 (*OMIM:121201) and *KCNQ5 (*OMIM:617601), have been associated with epilepsy and are on the virtual panel for this specific disease in PanelApp. However, distinguishing primary seizures from seizures secondary to cerebral hypoperfusion caused by ventricular arrhythmia is challenging [34].

Furthermore, repeated ischaemic insults could explain cerebral abnormalities. Thus, our association may reflect, at least in part, detection of concealed long QT syndrome and may account for the aforementioned, distinct phenotype. Supporting this, two of the case variants (p.Ala302Val and p.Arg594*) are classified as known/likely pathogenic for long-QT syndrome in Clinvar. Nonetheless, another gene associated with long-QT syndrome, *KCNH2* has been more strongly associated with primary epilepsy [35] but was not detected in our analysis. Investigation of the cardiac phenotype in these cases of epilepsy will be required but these data are not available currently.

Early onset Parkinson's disease associated with *RYR1*

We identified a dominant association between variants in *RYR1* and 'specific disease' *early onset and familial Parkinson's disease.* The association is driven by rare, predicted pathogenic variants (35 missense, 1 LoF) in 36 cases, located throughout the gene (**Fig. 4b,** OR = 2.1 [95% CI: 1.5 – 2.9], **Supplementary Table 2**). The gene is depleted for rare missense variants in gnomAD (o/e = 0.9 (0.87-0.93) and Z = 1.92). The 36 cases show strong phenotypic similarity to each other (mean PhenoDigm score from pairwise, reciprocal, non-self hits was 0.76). However, 407/1000 randomly sampled case sets of the same size achieved the same mean score or higher, suggesting the *RYR1* families do not have any phenotypic distinction from the other Parkinson's disease (PD) cases based on the available HPO terms. *RYR1* functions as a calcium release channel and is associated with malignant hyperthermia, congenital myopathies and tremor pathophysiology [36]. A mouse knockout shows various neurological phenotypes such as being unresponsive to tactile stimuli, abnormal posture, and paralysis [37]. The GTEx substantia nigra gene co-expression network was examined for enrichment of 45 genes related to Mendelian forms of PD and complex Parkinsonism (green genes in PanelApp) as well as 151 genes linked to sporadic PD through Mendelian randomisation (Nalls et al., 2019). *RYR1* was contained within the substantia nigra 'darkmagenta' module. Although this module was not significantly enriched for PD-associated genes, it is interesting to note that the module contained multiple genes causally implicated in PD, including *PINK1* and *PARK2* amongst others (FDR adjusted p-value = 0.146). Furthermore, the module in which *RYR1* is located is highly enriched for gene ontology

terms associated with mitochondrial function, a key process in PD pathophysiology (**Supplementary Fig. 1**).

Anterior segment ocular abnormalities associated with *POMK*

We identified a dominant association between variants in *POMK* and 'specific disease' *corneal abnormalities.* The association is driven by rare, predicted pathogenic variants (2 LoF, 1 missense) in 3 cases with recorded phenotypes collectively suggestive of Anterior Segment Dysgenesis (ASD) (**Fig. 4c,** OR = 92.5 [95% CI: 26.8 – 319.6], **Supplementary Table 2**). ASD is a spectrum of developmental disorders affecting the anterior segment of the eye often with incomplete penetrance and/or variable expressivity [38]. Co-segregation was apparent in two trios where a heterozygous, splice acceptor variant c.-21-1G>A (gnomAD v4.0.0 allele frequency (AF) = 0.000011) and a heterozygous, frameshift stop gain variant p.Arg339* (gnomAD v4.0.0 AF = 0.000001) are inherited from the affected mothers in the female probands (**Fig. 4c**). In the independent cohort described in **Supplementary Table 2**, one rare (gnomAD v4.0.0 AF = 0.000038), heterozygous missense variant p.Arg86His was identified in a mother and a son of a Czech family diagnosed with ASD, comprising 4 affected individuals across 3 generations. *POMK* is involved in the presentation of the laminin-binding O-linked carbohydrate chain of alpha-dystroglycan (a-DG), which forms transmembrane linkages between the extracellular matrix and the exoskeleton. Given the absence of corneal specific expression data in the GTEx Project, we interrogated publicly available bulk RNA-seq datasets [39,40] which showed expression across all corneal cell types analysed, with highest levels detected within the corneal epithelium (**Fig. 4c**). Bi-allelic (predicted LoF) mutations in *POMK* are associated with autosomal recessive muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 12 (OMIM:615249), a disease that also includes several ocular abnormalities (microphthalmia, buphthalmos, coloboma, retinal degeneration and cataract), suggesting *POMK* plays a crucial role in ocular development [41]. Morpholino knockdown of the *pomk* gene in zebrafish has been reported to show multiple defects, including developmental ocular abnormalities [41]. Whether the identified rare variants here could induce ASD via *POMK* haploinsufficiency, or by exerting dominant gain-of-function effects, deserves

future investigation. With *POMK* apparently LoF-tolerant (pLI = 0) and ASD not been reported in carriers of muscular dystrophy-dystroglycanopathy type A, 12-associated variants, the latter seems more likely.

Cystic kidney disease associated with *COL4A3*

We identified a dominant association between variants in *COL4A3* and 'specific disease' *cystic kidney disease* and another component of Type IV collagen, family member *COL4A1,* is already known to be associated with cystic kidney disease (Plaisier et al. 2007). The association was driven by rare, predicted pathogenic variants (27 missense, 2 LoF) in 29 cases, located throughout this kidney expressed gene (**Fig. 4d,** OR = 4.4 [95% CI: 2.9 – 6.6], **Supplementary Table 2**). *COL4A3* is constrained for rare missense variants in gnomAD (o/e = 0.77 (0.67-0.82) and Z = 1.83). The 29 cases show strong phenotypic similarity to each other (mean PhenoDigm score from pairwise, reciprocal, non-self-hits was 0.81) with cardiovascular phenotypes observed in some cases in addition to kidney cysts. 224/1000 randomly sampled case sets of the same size achieved the same mean score or higher, suggesting the *COL4A3* families do not form a phenotypically distinct group from the other cases based on HPO terms. Three of the cases involved duos where the variant was also observed in the other affected family member: (i) a ClinVar known pathogenic variant, p.Gly395Glu [42], in a brother and sister, (ii) a ClinVar VUS variant, p.Ser1600del, in a father and son, (iii) a ClinVar pathogenic variant, p.Arg1481*, in a mother and son. *COL4A3* is associated with dominant (van der Loop et al. 2000) and recessive (Mochizuki et al. 1994) forms of Alport syndrome with renal and hearing phenotypes, as well as a milder familial hematuria (Badenas et al. 2002). Kidney cysts in people with Alport syndrome/thin basement membrane nephropathy have previously been observed (Savige, Mack, et al. 2022). However, only one of our cases, with a p.Asn1508Ser variant reported with conflicting evidence in Clinvar, had any hearing impairment suggestive of classical Alport syndrome. Eleven of the cases have already been issued with molecular diagnoses in other cystic kidney genes (9 with *PKD1* and 2 with *PKD2*) within the 100KGP, raising the possibility of digenic inheritance or modification of incompletely penetrant variants. In the independent Irish cohort described in **Supplementary Table 2**, two related patients with chronic kidney

disease are already diagnosed with rare, heterozygous *COL4A3* variants and two have *COL4A3* and missense *PKD1* VUS variants. A digenic form of Alport syndrome has been described with pathogenic *COL4A4*/*COL4A5* and milder *COL4A3* variants [43] . This more severe, form of Alport syndrome is associated with increased incidence of proteinuria, hypertension and kidney failure is occasionally observed and the recommendation is to look for second variants in *COL4A3*, *COL4A4* and *COL4A5* in at-risk individuals (Savige, Lipska-Zietkiewicz, et al. 2022). In our cases with an existing cystic kidney disease molecular diagnosis, the renal phenotype is not as severe and pathogenic *COL4A4* and *COL4A5* variants are not observed suggesting the *COL4A3* variants are having a milder digenic effect or acting as modifiers with other cystic kidney genes.

Discussion

In this study we have described a gene burden analysis of a large cohort of rare disease cases and identified 88 novel disease-gene associations after triaging of the statistically significant signals. Recent publications describe confirmation of five of these and we highlight a further seven with strong genetic and experimental evidence: hypertrophic cardiomyopathy associated with *DYSF* and *SLC4A3*; monogenic diabetes with *UNC13A*; epilepsy with *KCNQ1*; early onset Parkinson's disease with RYR1; anterior segment ocular abnormalities with *POMK*; and cystic kidney disease with *COL4A3;* and. Further evidence is necessary before all the novel associations described here can be used clinically for diagnostics, counselling, and management. For example, addition of further functional study evidence would increase the score of 47/50 of the limited ClinGen classification evidence candidates shown in **Supplementary Table 2,** such that they would be re-classified as moderate. We are also pursuing the identification of variants in further independent cases through GeneMatcher to boost the ClinGen evidence category. Collection of additional affected family members for cases in collaboration with the original recruiting clinicians could also raise the category although this is likely to be more difficult,

e.g., for a dominant association at least two large families with five affected members are needed to add one point of evidence.

Although our approach applied to a large, rare disease cohort has successfully highlighted numerous known associations and suggested many previously unreported associations, it is not without limitations. There is an assumption that variant calls are accurate and that cases are always recruited to the correct 'specific disease' category within the 100KGP, which is not always true as shown for *DCP1A* with heart disease and *HBB* with kidney stones respectively. Probands are assumed to be unrelated, but when we further investigated the association of *ADAMTS4* with hearing impairment (**Supplementary Table 2**), this was not the case with 3 of the 4 probands belonging to the same family. The *GATA2* signal was similarly diluted on further investigation with 4 of the 6 probands linked to the association being closely related (mother and 3 siblings). For maximal performance of a gene burden testing approach, there should be no pathogenic variants associated with the disease cases in the control samples but mis-recruitment or incomplete penetrance may mean this is not always true, otherwise we could have added this absolute condition prior to the statistical testing. Given the multi-disease nature of our analysis, we opted for a non-disease specific, less strict AF filtering strategy so as not to incur inflated false negatives and afterwards discarded, in our *in silico* triage, signals where > 33.3% of the contributing case variants were present in controls (179/316, 57%). Those many discarded signals could be the focus of a future clinical expert revised triage using a *disease-specific* AF filtering strategy such as the *maximum credible population AF* [44] where disease-specific prevalence, genetic and allelic heterogeneity, inheritance mode, penetrance, and sampling variance in reference datasets are considered to identify less arbitrary and/or unnecessarily lenient AF cutoffs.

The relatively small ORs for some of the associations in **Supplementary Table 2** suggest that some of the 88 associations may involve incomplete penetrance and that is known to be the case for kidney stones and *SLC34A3*. Incomplete penetrance will lead to the same/similar variants in controls and lower ORs and less significant signals, as well as possible loss of signal in multi-sample cases under Exomiser's assumed full penetrance

model. Finally, associations linked to gain-of-function variants may have been difficult to detect with our methodology, unless they clustered in the CCR regions we analysed.

A recent analysis of approximately the same cohort of rare disease patients using BeviMed found 241 known but only 19 novel associations [24]. 115/190 of our known and only 5/88 of the novel signals were also detected by BeviMed: *UCHL1* associated with hereditary ataxia; *SLC34A3* and *HBB* associated with renal tract calcification; *TUFT1* associated with epidermolysis bullosa; *SRP9* with ductal plate malformations. The relatively small overlap in signals highlights a possible complementarity of the two methods to discover new disease-gene associations from the same cohort. As well as differences in the statistical approach, different case-control selection strategies were used.

There are 606 cases with no molecular diagnosis but with variant(s) contributing to one of the known disease-gene association signals, that had not already been considered and classified as VUS or benign in the diagnostic report, giving an upper bound on the increase in diagnostic yield from review of these variants of 1.7% (606 of 35,008 cases analysed). Furthermore, 456 molecularly unsolved cases had a variant contributing to one of the 88 novel associations and could potentially increase the diagnostic yield by 1.3% (456 of 35,008 cases analysed), if all genes were confirmed and the variants considered penetrant enough to be deemed pathogenic rather than just predictive. By making our analytical framework openly available for wider application to similar cohort data globally, we hope to substantially aid disease-gene discovery and new molecular diagnoses in rare Mendelian diseases in numerous other cohorts.

Online Methods

Rare disease genomes from the 100,000 Genomes Project

Patients with rare diseases and affected and unaffected family members were enrolled to the 100KGP through one of the 13 NHS Genomic Medicine Centres (GMCs) across England, Northern Ireland, Scotland and Wales [8]. The recruiting clinicians assigned each proband to a specific disease (according to a hierarchical disease classification available within the project which is described below) and provided patient's phenotypic data based on the HPO [45]. An initial cohort of 74,061 genomes (35,548 single probands and larger families) from the rare disease pilot and main programme of the 100KGP (Data Release v.11) was available for analysis (March 2021). Genome sequencing was performed with the use of the TruSeq DNA polymerase-chain-reaction (PCR)–free sample preparation kit (Illumina) on a HiSeq 2500 sequencer, which generates a mean depth of 32× (range, 27 to 54) and a depth greater than 15× for at least 95% of the reference human genome. Whole-genome sequencing reads were aligned to the Genome Reference Consortium human genome build 37 (GRCh37) with the use of Isaac Genome Alignment Software. Family-based variant calling of single-nucleotide variants (SNVs) and insertion or deletions (indels) for chromosomes 1 to 22, the X chromosome, and the mitochondrial genome (mean coverage, 2814×; range, 142 to 16,581) was performed with the use of the Platypus variant caller [46]. Quality control performed by Genomics England highlighted that 81 of the probands had been recruited and sequenced twice and these duplicates were removed from our cohort. In addition, the required data for our Exomiser-based gene burden analysis, e.g. recruited disease category and phenotypic terms, was not available for 16 families and these were also excluded from our cohort.

Pool of rare, putative disease-causing variants for gene burden testing

The variant prioritisation tool Exomiser [9] (version 12.1.0 with default settings and latest 2007* (July2020) databases) was then run on all available 35,451 single proband and family-based variant call format (VCF) files to obtain a pool of rare, protein-coding, segregating and most predicted pathogenic (per each gene) variants to use in a rare-

variant gene-based burden testing analysis for the discovery of novel rare Mendelian disease-gene associations as described below. Per each proband/family and each gene, Exomiser selected a single configuration of *contributing* variants, i.e. the most predicted (i.e. REVEL and MVP) pathogenic, rare (< 0.1% autosomal/X-linked dominant or homozygous recessive, < 2% autosomal/X-linked compound heterozygous recessive; using publicly available sequencing datasets including gnomAD) protein-coding homozygous/heterozygous variant or compound-heterozygote variants that segregated with disease for each possible mode of inheritance. Coding variants were selected by Exomiser by removing all those classified as FIVE_PRIME_UTR_EXON_VARIANT, FIVE_PRIME_UTR_INTRON_VARIANT, THREE_PRIME_UTR_EXON_VARIANT, THREE_PRIME_UTR_INTRON_VARIANT, NON_CODING_TRANSCRIPT_EXON_VARIANT, UPSTREAM_GENE_VARIANT, INTERGENIC_VARIANT, REGULATORY_REGION_VARIANT, CODING_TRANSCRIPT_INTRON_VARIANT, NON_CODING_TRANSCRIPT_INTRON_VARIANT, DOWNSTREAM_GENE_VARIANT. The Exomiser analysis did not return any candidate variants for 29 families, generally for larger families with multiple affected individuals where no rare, putative disease-causing variants remained after filtering, leading to an interim dataset size of 35,422 single probands and larger families and 5,733,899 Exomiser-based candidate variants (*Exomiser master dataset*). To control for false positive variant calls and/or relatively common variants within the project itself, we further discarded variants based on how often they were observed within the Exomiser master dataset itself (frequency > 2% for variants in a compound-heterozygote genotype, > 0.2% for mtDNA genome variants, > 0.1% for heterozygote/homozygote variants). This led us to discard data from 41 additional families. Finally, potentially digenic probands with more than one recruited disease category were discarded from the analysis, leading to a final analysis dataset of 35,008 families (40,584 probands and affected family members and 32,434 unaffected family members) and 4,676,866 Exomiser-based candidate heterozygote/homozygote variants and compound-heterozygote genotypes (**Supplementary Table 1 and Fig. 1**).

Exomiser-based rare variant gene burden testing

A rare variant gene-based burden case-control analytical framework which exploits rare, putative disease-causing variants as annotated, filtered and scored by the variant prioritisation tool Exomiser has been used to identify novel rare Mendelian disease-gene associations. The framework has been described previously [4]. Briefly, as to the application of the analytical framework to the rare disease component of the 100KGP, *cases* and *controls* were defined exploiting the hierarchical disease classification within the project itself where the recruiting clinicians assigned each proband to any of 226 'specific diseases' (level 4); the 'specific diseases' are in turn grouped into less specific 91 'disease sub groups' (level 3), each of which corresponds to one of 20 broad 'disease groups' (level 2) (**Supplementary Table 1**). A *case* set was then defined as all probands recruited under each of the 226 level 4 disease categories and its corresponding *control* set as all recruited probands except those under the level 2 category containing the specific level 4 disease, e.g. 'hypertrophic cardiomyopathy' cases were compared to all non-'cardiovascular disorders' probands as controls. As to the gene burden testing, right-tailed Fisher's exact tests were performed to assess the gene-based enrichment of variants in cases versus controls under four proband's genotype scenarios (irrespective of the mode of inheritance): (i) presence of at least one rare, predicted LoF variant, (ii) presence of at least one rare, highly predicted pathogenic variant (Exomiser variant score >= 0.8, i.e. either LoF or missense variants predicted to be pathogenic) (iii) presence of at least one rare, highly predicted pathogenic variant in a constrained coding region (CCR) and (iv) presence of a rare, *de novo* variant (restricted to only trios or larger families where *de novo* calling was possible). To maintain statistical validity and power, the analysis was limited to those disease-gene associations where at least five cases exist for the specific disease tested and, per each of the four gene-based proband's genotype scenarios above, where relevant variants in the gene were seen in at least four probands, of which at least one was a case. The Benjamini and Hochberg method [47] was used to correct for multiple testing; an overall false discovery rate (FDR) adjusted p-value (q-value) threshold of 0.05 was used for claiming statistically significant disease-gene associations to pursue for further triaging.

Triaging

First in silico triage

The statistically significant associations were further filtered for those where (i) the gene was protein-coding as the Exomiser coding variant filtering settings also identified variants disrupting non-protein coding RNA genes, (ii) less than 33.3% of the variants driving the signal in the cases were present in the controls, (iii) for dominant, LoF signals there was gnomAD evidence for haploinsufficiency (gnomAD LOEUF < 0.5), and (iv) <= 50% of the cases driving the signal were already assigned a molecular diagnosis in other genes as part of the 100KGP routine diagnostic pipeline.

Second clinical expert (GeCIP) and in silico triage

An automated classification of the disease-gene associations according to ClinGen criteria (https://www.clinicalgenome.org/site/assets/files/9232/gene-disease validity standard operating procedures version 10.pdf) was applied. The case-level variant score was calculated from scoring and summing all case variants that support a particular mode of inheritance for a disease-gene association. LoF variants (stop gain, frameshift or splice acceptor/donor) scored 1.5 points or 2 if *de novo* whilst others scored 0.1 points or 0.5 if *de novo*. A case-control study score of 5 points for an OR > 5, 4 points for OR > 3 or 3 points for OR < 3 was assigned. The larger of the case-level variant score or case-control study score was used as the genetic evidence score, capped at a maximum of 12 for those associations that had many supporting case variants. Experimental evidence categories were calculated using a variety of sources. Existing evidence for a gene function fitting the likely disease mechanism was assessed via PubMed searches using the disease and gene name and the background knowledge of the experts in the various disease-specific GeCIPs. Scores of 0, 1 or 2 were awarded depending on whether there was no, some tenuous or lots of evidence. Gene expression was assessed using GTEx Project data through the web portal of the Human Protein Atlas (https://www.proteinatlas.org/; [48] and/or publicly available relevant RNA-seq datasets [39,40], and a score of 0, 1 or 2 assigned for no, widespread or solely specific expression in the relevant disease tissue. Defaults of 1 point for protein-protein association evidence

(high quality, direct experimental interactions (StringDB interactions with a score > 0.7 and experimental evidence) with genes on the disease panel from PanelApp [14] and 2 points for mouse/zebrafish evidence where there was some phenotypic similarity as calculated by Exomiser between the patient's phenotypes and the mouse/zebrafish phenotypes where the orthologous gene was disrupted. The rounded sum of genetic and experimental evidence points was used to assign the final ClinGen classification of limited (0.1-6 points), moderate (7-11 points) or strong (12-18 points). Definitive evidence for an association is considered a score of 12-18 as well as convincing replication of the result in more than 2 publications over more than 3 years, and therefore not applicable here.

Visual representation of variant location in lollipop plots

Visual representations of the variant locations within the protein were generated by extending the Mutplot software [49]. The x-axis represents the amino acid chain and their annotated protein domain from UniProt. Each lolly indicates a variant by its protein change annotated on one single transcript (specified in the plot) and the frequency is shown on the y-axis. Its shape indicates the genotype found in the proband. The colour indicates the type of variant and the variant's functional annotation. If the variant has both a p. change annotation and a number in parenthesis it means that the original p. change was annotated on a different transcript and the amino acid position in parenthesis indicates the re-annotation on the selected transcript. If the only annotation available indicates a number in parenthesis it means that the variant was in the non-coding region for that transcript, therefore the lolly was placed on the closest amino acid.

PhenoDigm patient similarity comparisons

During assessment of some disease-gene associations, the phenotypic similarity between the probands driving the signal was calculated using their HPO term annotations and the Exomiser API to give a PhenoDigm [50] score between 0 and 1. The mean of the pairwise, reciprocal, non-self hits was calculated and compared to those obtained from 1000 iterations when the same number of probands was selected at random from the set of cases with that disease.

Co-expression network analysis

Co-expression network analysis of our candidate genes and known genes linked to the potentially associated disease (green genes in PanelApp version 1.120) was performed using GTEx v6 tissue-specific modules and the CoExp tool accessible at https://rytenlab.com/coexp [51].

Peripheral blood mononuclear cell (PBMC) expression analysis

RNA-seq data from PBMC cells collected from three volunteer donors was analysed (poly A-selected libraries, mean of two replicates untreated and two replicates treated with cycloheximide for 1hr to inhibit protein translation and mimic integrated stress response). In-house, R was used for DeSeq2 normalisations per library and calculation of the mean values for each transcript for the 2 replicate libraries per donor per condition. For global evaluations, across all 3 donors, the mean base value, log2fold change post cycloheximide and Benjamini-Hochberg adjusted p-value were then calculated.

Gene and variant look up in independent rare disease cohorts

In a cohort of Irish renal patients (278 cystic kidney disease and 141 chronic kidney disease cases), rare (gnomAD MAF < 0.1%) LoF, missense, splicing or intronic variants were extracted for our novel renal disease-associated genes. A further cohort of over 3000 Dutch renal patients was queried for likely pathogenic/pathogenic variants in those genes using the Alissa bioinformatics pipeline. Similarly, a sequencing cohort of 212 participants with inherited corneal diseases, recruited in the UK and Czech Republic and pre-screened for known genetic causes, was interrogated for any rare variants in the candidate gene *POMK*.

Author contributions

M.C, A.E.D, P.L contributed to the analysis of the corneal patient cohort, evaluation of candidates and reviewing/editing manuscript. A.M.E, I.L contributed to the analysis of the Dutch cohort. S.E.F contributed to the analysis of the Exeter diabetes patient cohort, evaluation of candidates and reviewing/editing manuscript. P.C, E.E, O.T contributed to the analysis of the Irish cohort. K.A.B contributed to the analysis of the Irish cohort, evaluation of candidates and reviewing/editing manuscript. V.C, L.V, D.S developed the analysis pipeline, conducted analyses and cowrote the manuscript. M.C contributed to the development of analysis and reviewing/editing the manuscript. P.N.R contributed to the development of parts of the analysis pipeline. J.OB.J contributed to the development of parts of the analysis pipeline and reviewing/editing manuscript. H.R.G, H.H, S.L, R.N, A.T.P, N.R, S.K.W, S.L.Z contributed to the evaluation of candidates. E.F.M, G.A, E.R.B, D.B, M.R.B, K.B, L.F.C, P.C, S.J .D, M.F, D.P.G, R.S.H, S.A.H, H.M, H.M, A.D.M, W.G.N, E.A.OT, A.CM.O, S.R, O.SA, J.A.S, J.C.T, I.T, A.T, J.S.W, L.M.W contributed to the evaluation of candidates and reviewing/editing the manuscript. S.GR, A.R.H contributed to the expression analysis. M.R, C.L.S contributed to the expression analysis, evaluation of candidates and reviewing/editing manuscript. M.B, R.K, A.R.W, C.T contributed to the patient recruitment and phenotyping. C.GC, S.H contributed to reviewing/editing the manuscript.

Acknowledgements

This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library 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. PL was supported by GACR 24-10324S. This research was funded in part by Aligning Science Across Parkinson's [Grant numbers: ASAP-000478 and ASAP-000509] through the Michael J. Fox Foundation for Parkinson's Research (MJFF). This is part of the NIHR Barts Biomedical Research Centre (Caulfield, Jones) portfolio of research. The analysis was supported by a grant from the NIH, National Institute of Child Health and Human Development 1R01HD103805-01 and PhD funding from the UCLH NIHR Hearing Health BRC, and we would like to dedicate this work to the wonderful supervision of the late Professor Maria Bitner-Glindzicz.

Conflict of interests

The authors declare the following competing interests: D.S. and M.C. were seconded to and received salary from Genomics England, a wholly owned Department of Health and Social Care company, from 2016-2018 and 2013-2021 respectively. EOT has research funding from Kamari Pharma, Pavella Therapeutics, Unilever and the Leo Foundation unrelated to this work. She is CI for a trial for Kamari Pharma and performs consultancy for Kamari Pharma, Azitra and Palvella Therapeutics (all money goes to the university).

Figure legends

Figure 1. Rare variant gene burden analysis of the 100,000 Genomes Project data. Flowchart of the rare variant gene-based analytical framework, including triaging of the results.

Figure 2. Rare disease gene discoveries from gene burden analysis of the 100,000 Genomes Project data. The gene burden testing identified 506 disease-gene associations at 5% False Discovery Rate (FDR) including 316 potentially novel. An initial triage of the novel signals identified 88 signals for further investigation through *in silico* collection for additional evidence and clinical expert review. Statistical significance, expressed as -log10 of FDR adjusted p-value, is shown for each of the 506 gene-disease associations significant at 5% FDR, arranged by 'disease group'. The 190 known associations are in green, 88 triaged novel signals in blue and the discarded signals in grey.

Figure 3. Evidence for associations based on location of rare, predicted pathogenic variants in the 100,000 Genomes Project cases, Genotype Tissue Expression (GTEx) Project data and co-segregation data. (a) hypertrophic cardiomyopathy with *DYSF* including the heartAtrial GTEx co-expression module with known HCM genes in blue and *DYSF* in yellow, **(b)** hypertrophic cardiomyopathy with *SLC4A3*, **(c)** monogenic diabetes with *UNC13A*.

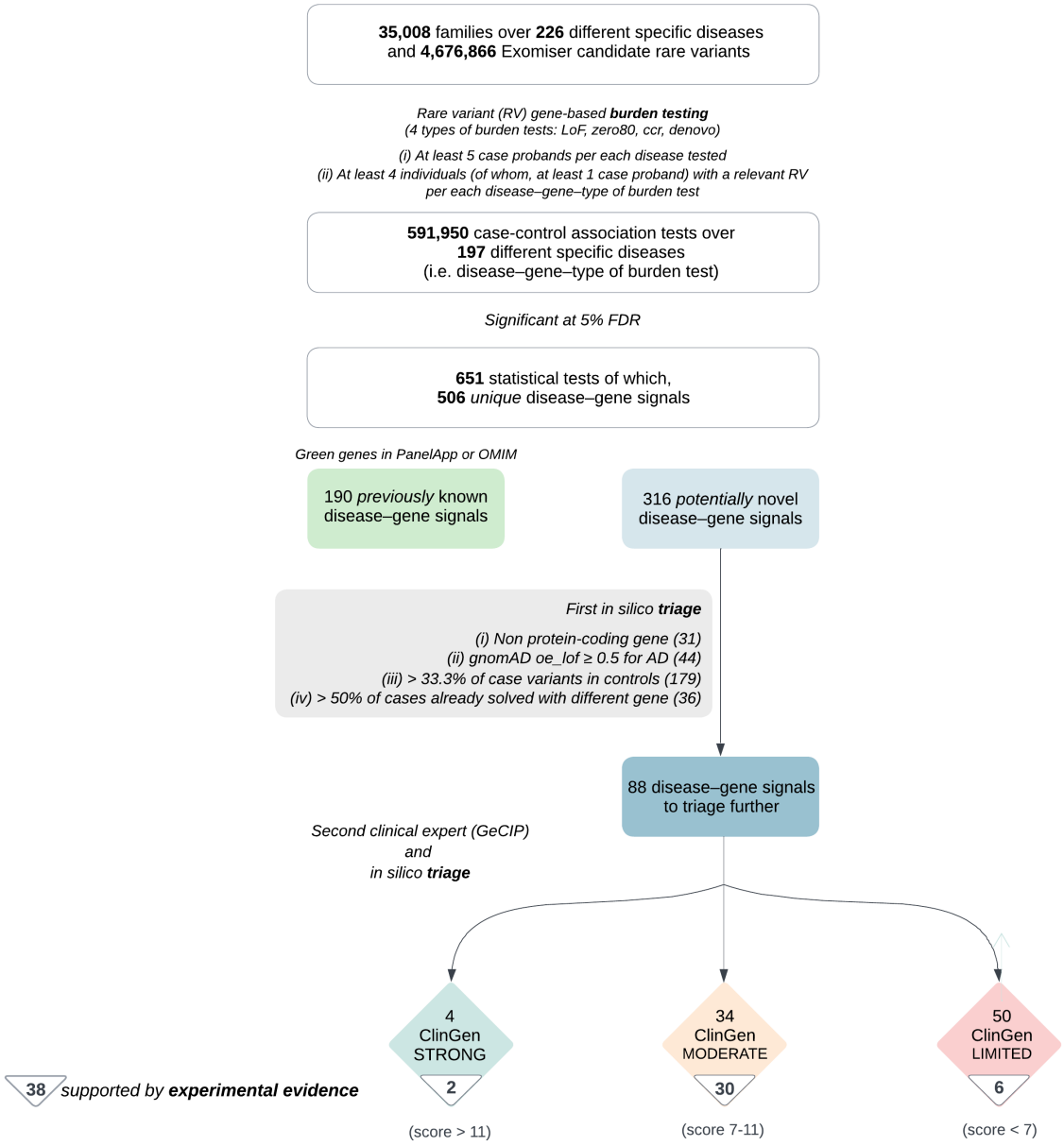
Figure 4. Evidence for associations based on location of rare, predicted pathogenic variants in the 100,000 Genomes Project cases, Genotype Tissue Expression (GTEx) Project or other RNA-seq relevant data, and co-segregation data. (a) epilepsy with *KCNQ1,* **(b)** early onset Parkinson's disease with *RYR1*, **(c)** anterior segment ocular abnormalities with *POMK* **(d)** cystic kidney disease with *COL4A3*.

Supplementary table/figure legends

Supplementary Table 1. Demographics of the 100,000 Genomics Project cohort.

Supplementary Table 2. 88 novel, putative disease-gene associations discovered by gene burden testing of the 100,000 Genomes Project data. Bolding in cells is used to indicate where there is co-segregation evidence from multiple families, experimental evidence (literature, mouse models or protein-protein interactions to related disease-genes), or the mode of inheritance (MOI) was mixed, recessive or gnomAD oe_lof < 0.5 for dominant, LoF signals or gnomAD oe_missense < 1 for dominant, score >= 0.8 signals.

Supplementary Figure 1. Gene ontology enrichment analysis of the GTEx substantia nigra gene co-expression network showing highly enrichment for terms associated with mitochondrial function, a key process in Parkinson's disease pathophysiology.

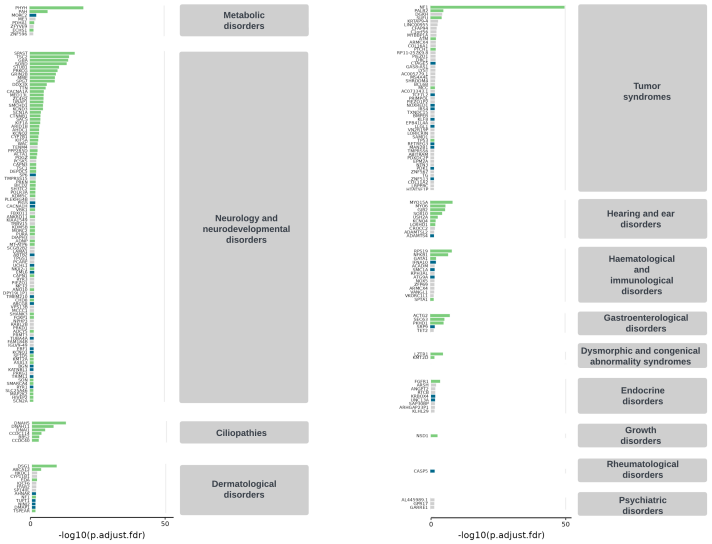


5 **interim confirmations** via collaboration or independent study

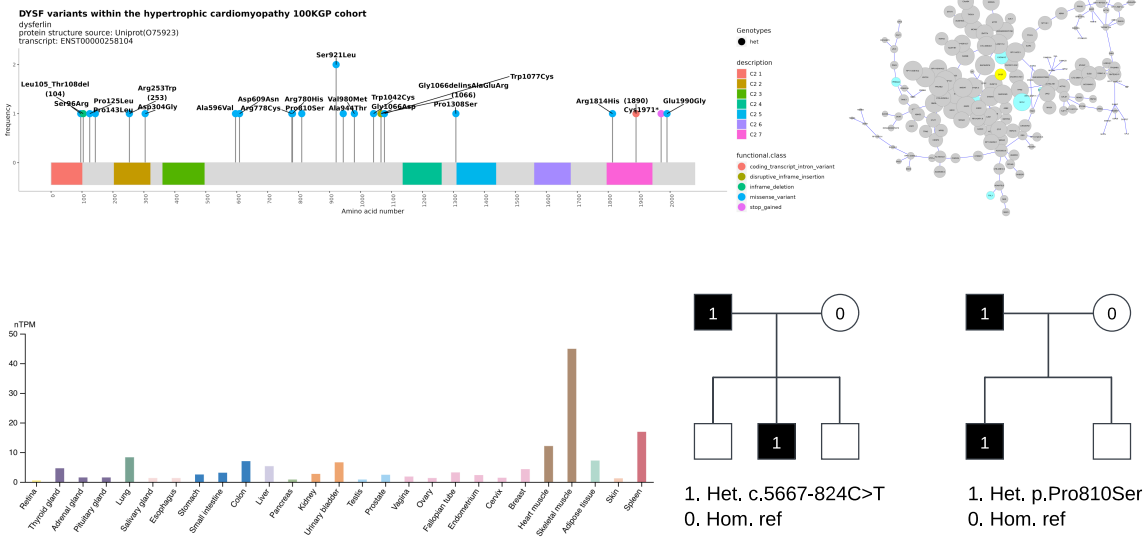
- Hereditary ataxia and **UCHL1** (Park et al., 2022)
- Mitochondrial disorders and **MORC2** (OMIM:619090)
- Renal tract calcification and **SLC34A3** (Sadeghi-Alavijeh et al., 2022)
- Hereditary ataxia and **TUBA4A** (Torella et al., 2023)
- Epidermolysis bullosa and **TUFT1** (Jackson et al. 2023)

7 **compelling candidate** disease-gene associations

- Cystic kidney disease and **COL4A3**
- Hypertrophic cardiomyopathy and **DYSF**
- Hypertrophic cardiomyopathy and **SLC4A3**
- Monogenic diabetes and **UNC13A**
- Epilepsy and **KCNQ1**
- Early onset Parkinson's disease and **RYR1**
- Anterior segment ocular abnormalities and **POMK**



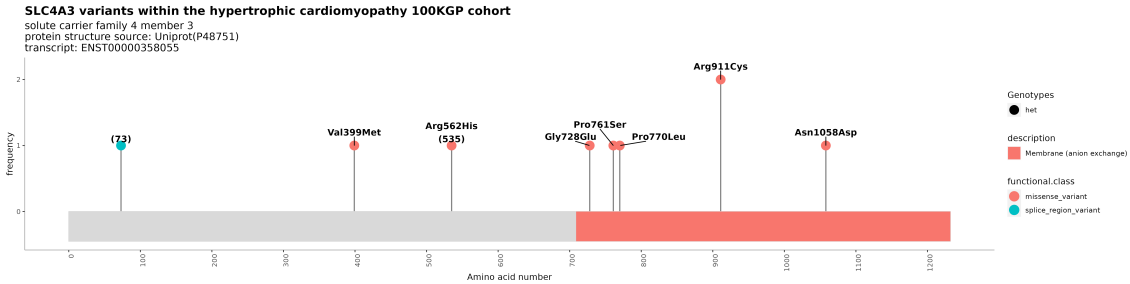
(a)

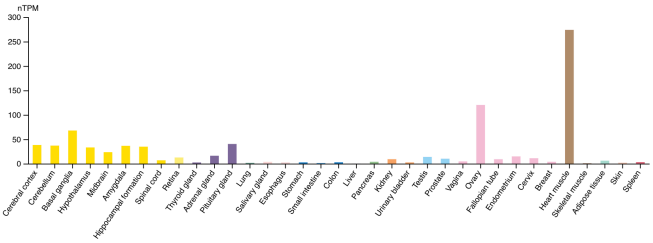


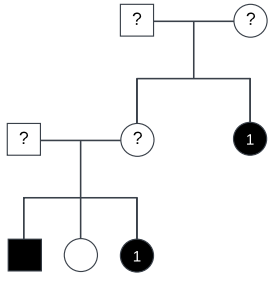
1. Het. c.5667-824C>T
0. Hom. ref

1. Het. p.Pro810Ser
0. Hom. ref

(b)

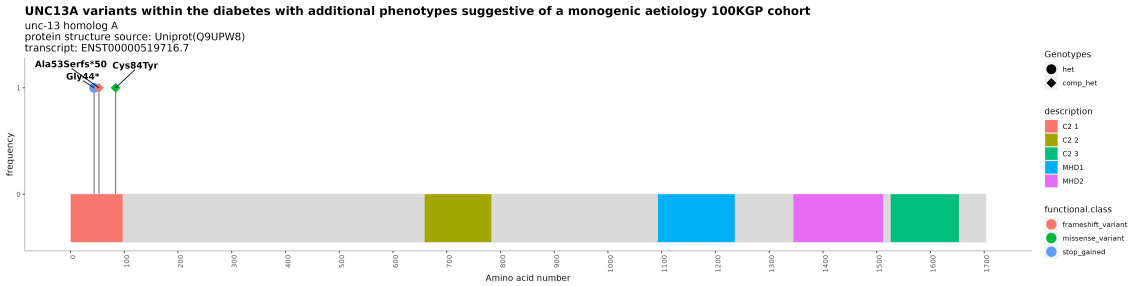




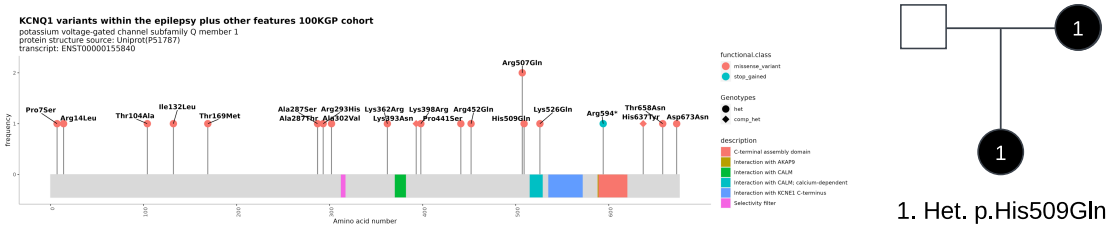


1. Het. p.Gly755Glu

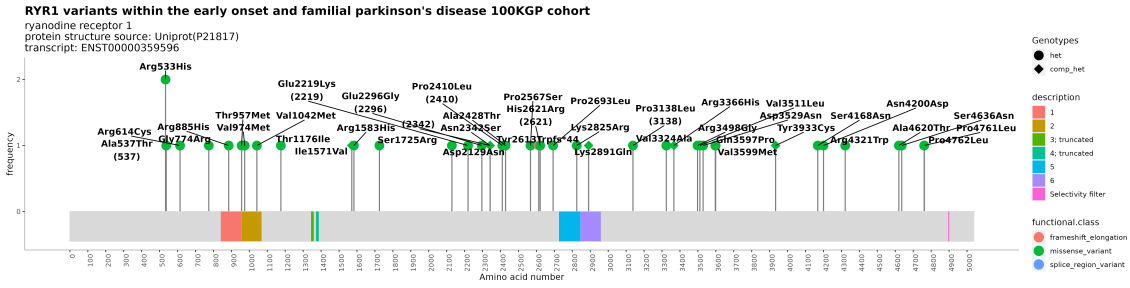
(c)



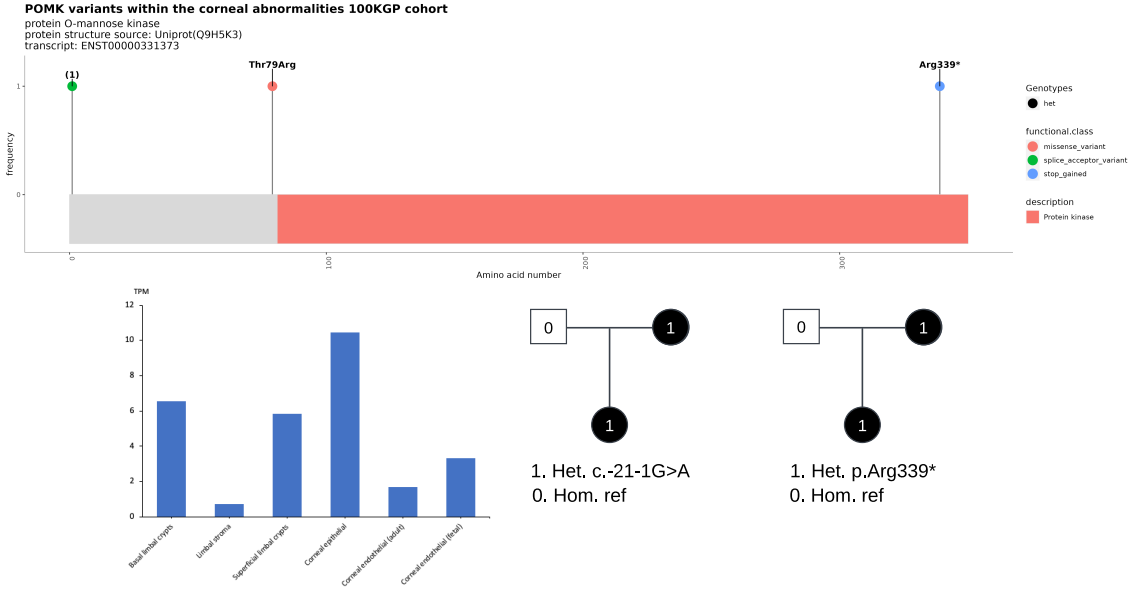
(a)



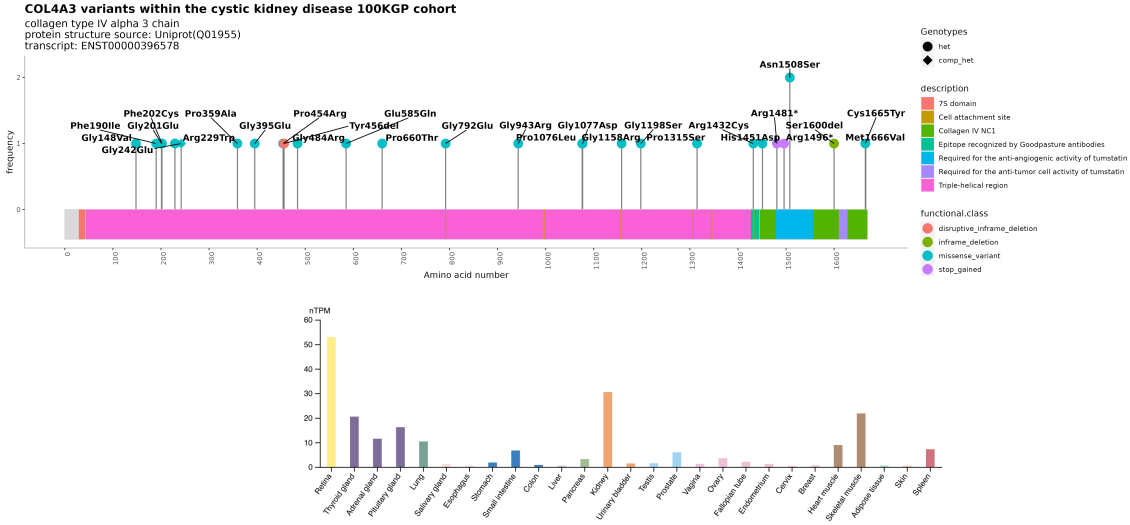
(b)

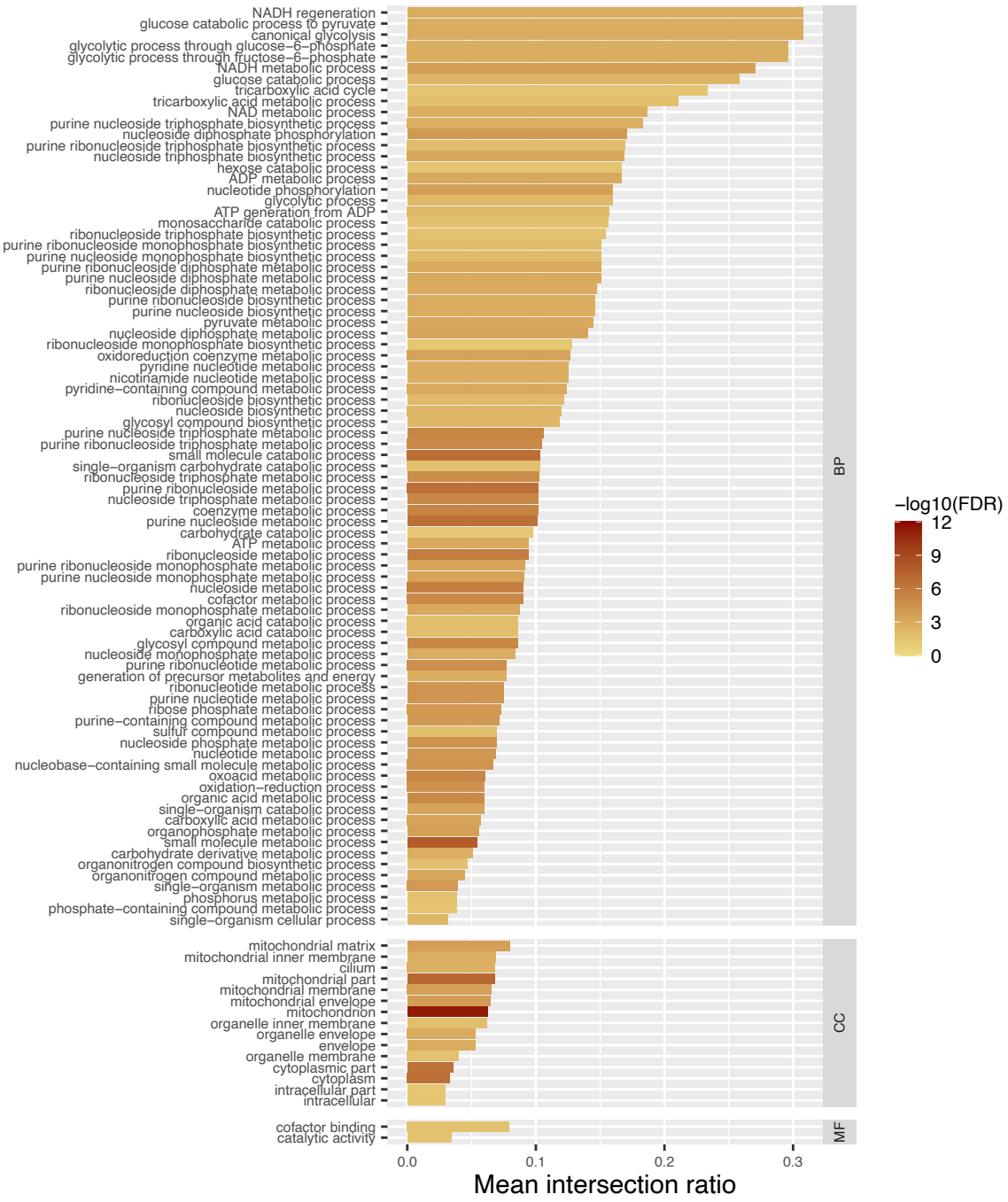


(c)



(d)





Supplementary Table 2 References

1. Chen, R. *et al.* Identification of biomarkers correlated with hypertrophic cardiomyopathy with co-expression analysis. *J. Cell. Physiol.* **234**, 21999–22008 (2019).
2. Hallmann, K. *et al.* A homozygous splice-site mutation in CARS2 is associated with progressive myoclonic epilepsy. *Neurology* **83**, 2183–2187 (2014).
3. Wiedmer, T. *et al.* Adiposity, dyslipidemia, and insulin resistance in mice with targeted deletion of phospholipid scramblase 3 (PLSCR3). *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13296–13301 (2004).
4. Whitfield, M. *et al.* Mutations in DNAH17, encoding a sperm-specific axonemal outer dynein arm heavy chain, cause isolated male infertility due to asthenozoospermia. *Am. J. Hum. Genet.* **105**, 198–212 (2019).
5. Tocchetti, A., Confalonieri, S., Scita, G., Di Fiore, P. P. & Betsholtz, C. In silico analysis of the EPS8 gene family: genomic organization, expression profile, and protein structure. *Genomics* **81**, 234–244 (2003).
6. Bai, R.-Y. *et al.* SMIF, a Smad4-interacting protein that functions as a co-activator in TGFbeta signalling. *Nat. Cell Biol.* **4**, 181–190 (2002).
7. Bashir, R. *et al.* A gene related to Caenorhabditis elegans spermatogenesis factor fer-1 is mutated in limb-girdle muscular dystrophy type 2B. *Nat. Genet.* **20**, 37–42 (1998).
8. Jackson, A. *et al.* Biallelic TUFT1 variants cause woolly hair, superficial skin fragility and desmosomal defects. *Br. J. Dermatol.* **188**, 75–83 (2023).
9. Cataldo, L. R. *et al.* MAFA and MAFB regulate exocytosis-related genes in human β-cells. *Acta Physiol. (Oxf.)* **234**, e13761 (2022).
10. Seist, R. *et al.* Cochlin deficiency protects against noise-induced hearing loss. *Front. Mol. Neurosci.* **14**, 670013 (2021).

11. Cavalleri, V. *et al.* Thrombocytopenia and Cornelia de Lange syndrome: Still an enigma? *Am. J. Med. Genet. A* **170A**, 130–134 (2016).
12. Downing, L. J. *et al.* IL-10 regulates thrombus-induced vein wall inflammation and thrombosis. *J. Immunol.* **161**, 1471–1476 (1998).
13. Guillen Sacoto, M. J. *et al.* De Novo Variants in the ATPase Module of MORC2 Cause a Neurodevelopmental Disorder with Growth Retardation and Variable Craniofacial Dysmorphism. *Am. J. Hum. Genet.* **107**, 352–363 (2020).
14. Goldman, A. M. *et al.* Arrhythmia in heart and brain: KCNQ1 mutations link epilepsy and sudden unexplained death. *Sci. Transl. Med.* **1**, 2ra6 (2009).
15. Torella, A. *et al.* A new genetic cause of spastic ataxia: the p.Glu415Lys variant in TUBA4A. *J. Neurol.* (2023) doi:10.1007/s00415-023-11816-w.
16. Park, J. *et al.* Heterozygous UCHL1 loss-of-function variants cause a neurodegenerative disorder with spasticity, ataxia, neuropathy, and optic atrophy. *Genet. Med.* **24**, 2079–2090 (2022).
17. Lory, P., Nicole, S. & Monteil, A. Neuronal Cav3 channelopathies: recent progress and perspectives. *Pflugers Arch.* **472**, 831–844 (2020).
18. Martuscello, R. T. *et al.* Defective cerebellar ryanodine receptor type 1 and endoplasmic reticulum calcium “leak” in tremor pathophysiology. *Acta Neuropathol.* **146**, 301–318 (2023).
19. Groopman, E. E. *et al.* Diagnostic Utility of Exome Sequencing for Kidney Disease. *N. Engl. J. Med.* **380**, 142–151 (2019).
20. Sadeghi-Alavijeh, O. *et al.* Rare variants in SLC34A3 explain missing heritability of urinary stone disease. *bioRxiv* (2022) doi:10.1101/2022.12.02.22283024.
21. Bartram, M. P. *et al.* Loss of Dgcr8-mediated microRNA expression in the kidney results in hydronephrosis and renal malformation. *BMC Nephrol.* **16**, 55 (2015).

22. Nagalakshmi, V. K. *et al.* Dicer regulates the development of nephrogenic and ureteric compartments in the mammalian kidney. *Kidney Int.* **79**, 317–330 (2011).

23. McGrath-Morrow, S. *et al.* VEGF receptor 2 blockade leads to renal cyst formation in mice. *Kidney Int.* **69**, 1741–1748 (2006).

24. Humbert, M. C. *et al.* ARL13B, PDE6D, and CEP164 form a functional network for INPP5E ciliary targeting. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19691–19696 (2012).

25. Bongers, E. M. H. F. *et al.* Genotype-phenotype studies in nail-patella syndrome show that LMX1B mutation location is involved in the risk of developing nephropathy. *Eur. J. Hum. Genet.* **13**, 935–946 (2005).

26. Le Minh, G. *et al.* Kruppel-like factor 8 regulates triple negative breast cancer stem cell-like activity. *Front. Oncol.* **13**, 1141834 (2023).

27. Bao, F., An, S., Yang, Y. & Xu, T.-R. SODD promotes lung cancer tumorigenesis by activating the PDK1/AKT and RAF/MEK/ERK signaling. *Genes (Basel)* **14**, (2023).

28. Gu, X. *et al.* Accumulated genetic mutations leading to accelerated initiation and progression of colorectal cancer in a patient with Gardner syndrome: A case report. *Medicine (Baltimore)* **100**, e25247 (2021).

29. Ortega, M. A. *et al.* Prognostic role of IRS-4 in the survival of patients with pancreatic cancer. *Histol. Histopathol.* **37**, 449–459 (2022).

30. Masunaga, T. *et al.* Desmoyokin/AHNAK protein localizes to the non-desmosomal keratinocyte cell surface of human epidermis. *J. Invest. Dermatol.* **104**, 941–945 (1995).

31. Kouno, M. *et al.* Ahnak/Desmoyokin is dispensable for proliferation, differentiation, and maintenance of integrity in mouse epidermis. *J. Invest. Dermatol.* **123**, 700–707 (2004).

32. Paragh, G. *et al.* Whole genome transcriptional profiling identifies novel differentiation regulated genes in keratinocytes. *Exp. Dermatol.* **19**, 297–301 (2010).

33. Araki, T. & Milbrandt, J. Ninjurin2, a novel homophilic adhesion molecule, is expressed in mature sensory and enteric neurons and promotes neurite outgrowth. *J. Neurosci.* **20**, 187–195 (2000).

34. Wiedemann, J. *et al.* Differential cell composition and split epidermal differentiation in human palm, sole, and hip skin. *Cell Rep.* **42**, 111994 (2023).

35. Silhavy, J. *et al.* Spontaneous nonsense mutation in Tuft1 (tuftelin 1) gene is associated with abnormal hair appearance and amelioration of glucose and lipid metabolism in the rat. *Physiol. Genomics* (2023) doi:10.1152/physiolgenomics.00084.2023.

36. Verkerk, A. J. M. H. *et al.* Disruption of TUFT1, a desmosome-associated protein, causes skin fragility, woolly hair, and palmoplantar keratoderma. *J. Invest. Dermatol.* (2023) doi:10.1016/j.jid.2023.02.044.

37. Py, B. F. *et al.* Cochlin produced by follicular dendritic cells promotes antibacterial innate immunity. *Immunity* **38**, 1063–1072 (2013).

38. Jung, J. *et al.* Cleaved cochlin sequesters Pseudomonas aeruginosa and activates innate immunity in the inner ear. *Cell Host Microbe* **25**, 513-525.e6 (2019).

39. Di Costanzo, S. *et al.* POMK mutations disrupt muscle development leading to a spectrum of neuromuscular presentations. *Hum. Mol. Genet.* **23**, 5781–5792 (2014).

40. Ealy, M. *et al.* Gene expression analysis of human otosclerotic stapedial footplates. *Hear. Res.* **240**, 80–86 (2008).

41. Højland, A. T. *et al.* A wide range of protective and predisposing variants in aggrecan influence the susceptibility for otosclerosis. *Hum. Genet.* **141**, 951–963 (2022).

References

1. Xiao, S. *et al.* Functional filter for whole-genome sequencing data identifies HHT and stress-associated non-coding SMAD4 polyadenylation site variants >5 kb from coding DNA. *Am. J. Hum. Genet.* **110**, 1903–1918 (2023).

2. Halley, M. C., Ashley, E. A. & Tabor, H. K. Supporting undiagnosed participants when clinical genomics studies end. *Nat. Genet.* **54**, 1063–1065 (2022).

3. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. Omim.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).

4. 100,000 Genomes Project Pilot Investigators *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).

5. Splinter, K. *et al.* Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. *N. Engl. J. Med.* **379**, 2131–2139 (2018).

6. Baxter, S. M. *et al.* Centers for Mendelian Genomics: A decade of facilitating gene discovery. *Genet. Med.* **24**, 784–797 (2022).

7. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).

8. Turnbull, C. *et al.* The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* **361**, k1687 (2018).

9. Bone, W. P. *et al.* Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genet. Med.* **18**, 608–617 (2016).

10. Farazi Fard, M. A. *et al.* Truncating Mutations in UBAP1 Cause Hereditary Spastic Paraplegia. *Am. J. Hum. Genet.* **104**, 767–773 (2019).

11. Wallmeier, J. *et al.* De Novo Mutations in FOXJ1 Result in a Motile Ciliopathy with Hydrocephalus and Randomization of Left/Right Body Asymmetry. *Am. J. Hum. Genet.* **105**,

1030–1039 (2019).

12. Cortese, A. *et al.* Biallelic mutations in SORD cause a common and potentially treatable hereditary neuropathy with implications for diabetes. *Nat. Genet.* **52**, 473–481 (2020).

13. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).

14. Martin, A. R. *et al.* PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* **51**, 1560–1565 (2019).

15. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

16. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).

17. Rehm, H. L. *et al.* ClinGen--the clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).

18. Torella, A. *et al.* A new genetic cause of spastic ataxia: the p.Glu415Lys variant in TUBA4A. *J. Neurol.* (2023) doi:10.1007/s00415-023-11816-w.

19. Guillen Sacoto, M. J. *et al.* De Novo Variants in the ATPase Module of MORC2 Cause a Neurodevelopmental Disorder with Growth Retardation and Variable Craniofacial Dysmorphism. *Am. J. Hum. Genet.* **107**, 352–363 (2020).

20. Sadeghi-Alavijeh, O. *et al.* Rare variants in SLC34A3 explain missing heritability of urinary stone disease. *bioRxiv* (2022) doi:10.1101/2022.12.02.22283024.

21. Jackson, A. *et al.* Biallelic TUFT1 variants cause woolly hair, superficial skin fragility and desmosomal defects. *Br. J. Dermatol.* **188**, 75–83 (2023).

22. Park, J. *et al.* Heterozygous UCHL1 loss-of-function variants cause a neurodegenerative disorder with spasticity, ataxia, neuropathy, and optic atrophy. *Genet. Med.* **24**, 2079–2090 (2022).

23. Groza, T. *et al.* The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res.* **51**, D1038–D1045 (2023).

24. Greene, D. *et al.* Genetic association analysis of 77,539 genomes reveals rare disease etiologies. *Nat. Med.* **29**, 679–688 (2023).

25. Groopman, E. E. *et al.* Diagnostic Utility of Exome Sequencing for Kidney Disease. *N. Engl. J. Med.* **380**, 142–151 (2019).

26. Aliberti, L. *et al.* Beta-thalassaemia major: Prevalence, risk factors and clinical consequences of hypercalciuria. *Br. J. Haematol.* **198**, 903–911 (2022).

27. Dominov, J. A. *et al.* Correction of pseudoexon splicing caused by a novel intronic dysferlin mutation. *Ann. Clin. Transl. Neurol.* **6**, 642–654 (2019).

28. Rosales, X. Q. *et al.* Cardiovascular magnetic resonance of cardiomyopathy in limb girdle muscular dystrophy 2B and 2I. *J. Cardiovasc. Magn. Reson.* **13**, 39 (2011).

29. Thorsen, K. *et al.* Loss-of-activity-mutation in the cardiac chloride-bicarbonate exchanger AE3 causes short QT syndrome. *Nat. Commun.* **8**, 1696 (2017).

30. Cataldo, L. R. *et al.* MAFA and MAFB regulate exocytosis-related genes in human β-cells.

Acta Physiol. (Oxf.) **234**, e13761 (2022).

31. Kang, L. *et al.* Munc13-1 is required for the sustained release of insulin from pancreatic beta cells. *Cell Metab.* **3**, 463–468 (2006).

32. Kwan, E. P. *et al.* Munc13-1 deficiency reduces insulin secretion and causes abnormal glucose tolerance. *Diabetes* **55**, 1421–1429 (2006).

33. Goldman, A. M. *et al.* Arrhythmia in heart and brain: KCNQ1 mutations link epilepsy and sudden unexplained death. *Sci. Transl. Med.* **1**, 2ra6 (2009).

34. Auerbach, D. S. *et al.* Genetic biomarkers for the risk of seizures in long QT syndrome. *Neurology* **87**, 1660–1668 (2016).

35. Johnson, J. N. *et al.* Identification of a possible pathogenic link between congenital long QT syndrome and epilepsy. *Neurology* **72**, 224–231 (2009).

36. Martuscello, R. T. *et al.* Defective cerebellar ryanodine receptor type 1 and endoplasmic reticulum calcium “leak” in tremor pathophysiology. *Acta Neuropathol.* **146**, 301–318 (2023).

37. Zvaritch, E. *et al.* An Ryr1I4895T mutation abolishes Ca2+ release channel function and delays development in homozygous offspring of a mutant mouse line. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 18537–18542 (2007).

38. Reis, L. M. *et al.* Comprehensive phenotypic and functional analysis of dominant and recessive FOXE3 alleles in ocular developmental disorders. *Hum. Mol. Genet.* **30**, 1591–1606 (2021).

39. Chen, Y. *et al.* Identification of novel molecular markers through transcriptomic analysis in human fetal and adult corneal endothelial cells. *Hum. Mol. Genet.* **22**, 1271–1279 (2013).

40. Bath, C. *et al.* Transcriptional dissection of human limbal niche compartments by massive parallel sequencing. *PLoS One* **8**, e64244 (2013).

41. Di Costanzo, S. *et al.* POMK mutations disrupt muscle development leading to a spectrum of neuromuscular presentations. *Hum. Mol. Genet.* **23**, 5781–5792 (2014).

42. Adam, J. *et al.* Genetic testing can resolve diagnostic confusion in Alport syndrome. *Clin. Kidney J.* **7**, 197–200 (2014).

43. Savige, J. *et al.* Digenic Alport syndrome. *Clin. J. Am. Soc. Nephrol.* **17**, 1697–1706 (2022).

44. Whiffin, N. *et al.* Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet. Med.* **19**, 1151–1158 (2017).

45. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).

46. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).

47. Benjamini, Y. & Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* **25**, 60 (2000).

48. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

49. Zhang, W., Wang, C. & Zhang, X. Mutplot: An easy-to-use online tool for plotting complex mutation data with flexibility. *PLoS One* **14**, e0215838 (2019).

50. Smedley, D. *et al.* PhenoDigm: analyzing curated annotations to associate animal models with human diseases. *Database* **2013**, bat025 (2013).

51. García-Ruiz, S. *et al.* CoExp: A web tool for the exploitation of co-expression networks. *Front. Genet.* **12**, 630187 (2021).