Article:
Chaurasia, A., Shukla, A., Pande, S. et al. (2026) Systematic analysis of homozygous autosomal copy number losses in exomes improves diagnostic yield and uncovers ultra-rare recessive disorders. European Journal of Human Genetics, 34. pp. 1059-1069. ISSN: 1018-4813

https://doi.org/10.1038/s41431-026-02158-y

Systematic analysis of homozygous autosomal copy number losses in exomes improves diagnostic yield and uncovers ultra-rare recessive disorders

Ankur Chaurasia (1, 2), Anju Shukla (1), Shruti Pande (1), Greeshma Purushothama (1), Akhil Kanathay Ashokan (1), Purvi Majethia (1), Namanpreet Kaur (1), Priyanka Upadhyai (1), Neha Quadri (1), Gandham SriLakshmi Bhavani (1), Dhanya Lakshmi Narayanan (1), Shalini S Nayak (1), Sheela Nampoothiri (3), Ataf H Sabir (4, 5), Alaa A Mohammed (6), Sophie Shaw (7), Verity L. Hartill (8, 9), Christopher M Watson (8, 10), Colin A Johnson (8), Afrah Alshammari (8, 11), Andrew E. Fry (7, 12), James A. Poulter (8), William G. Newman (1, 2, 13), Paul R. Kasher (14, 15), Siddharth Banka (1, 2, 13)*, Katta M. Girisha (1,16)*

(1) Department of Medical Genetics, Kasturba Medical College, Manipal, Manipal Academy of Higher Education, Manipal, India
(2) Division of Evolution, Infection and Genomics, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, M13 9PL, UK
(3) Department of Paediatric Genetics, Amrita Institute of Medical Sciences & Research Centre, Cochin, India
(4) West Midlands Clinical Genetics Unit, Birmingham Women's and Children's NHS Foundation Trust, Birmingham
(5) Department of Cancer and Genomic Sciences, College of Medicine and Health, University of Birmingham
(6) West Midlands Genomics Laboratory, Birmingham Women's and Children's NHS Foundation Trust, Birmingham
(7) All Wales Medical Genomics Service, Wales Genomic Health Centre, Cardiff Edge Business Park, Longwood Drive, Whitchurch, Cardiff CF14 7YU, UK
(8) Division of Molecular Medicine, Leeds Institute of Medical Research, University of Leeds, Leeds, LS9 7TF, UK
(9) Leeds Clinical Genomics Service, Chapel Allerton Hospital, Leeds, LS7 4SA, UK

(10) Yorkshire and North East Genomic Laboratory Hub, Central Lab, St. James's University Hospital, Leeds, LS9 7TF, UK
(11) Department of Biology, College of Science, University of Hafr Al-Batin, Hafr Al-Batin, Saudi Arabia
(12) Division of Cancer and Genetics, Cardiff University, Cardiff, CF14 4XN, UK
(13) Manchester Centre for Genomic Medicine, St Mary's Hospital, Health Innovation Manchester, Manchester University NHS Foundation Trust, Manchester, M13 9WL, UK
(14) Division of Neuroscience, School of Biological Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Oxford Road, Manchester M13 9PT, UK
(15) Geoffrey Jefferson Brain Research Centre, Manchester Academic Health Science Centre, Northern Care Alliance NHS Foundation Trust and University of Manchester, Manchester, UK
(16) Department of Genetics, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat, Oman

*Corresponding authors - g.kumar@squ.edu.om and Siddharth.Banka@manchester.ac.uk

Abstract

Systematic analysis of copy number variants (CNVs) in large datasets is challenging and there are limited studies of homozygous copy number losses in rare disease exomes. Here we leveraged the genomic uniqueness and relative under-representation of the Indian population in the current public genomic databases and identified 42,386 possible homozygous losses (median 20 per individual) in a heterogeneous cohort of 2,021 individuals with suspected Mendelian disorders, who had undergone exome sequencing using 12 different capture kits in a resource-limited setting. Employing a genomic position loss-count based approach, we filtered 1,224 rare homozygous loss calls in 718 individuals (median 1 per individual) for further analysis, thus significantly reducing the analysis burden. Clinical correlation and validation of these rare calls enabled 10 new diagnoses in 240 unsolved individuals. This led to a two-fold increase in diagnosis owing to homozygous deletions. Further analysis of the data and identification of additional affected individuals through collaboration led to identification of biallelic *FILIP1* and *FAM177A1* variants as causes of a syndromic arthrogryposis and a neuromuscular disorder respectively. Both conditions were recently reported as ultra-rare recessive disorders, thus validating our approach. We also show that biallelic loss-of-function *TFCP2L1* variants cause chronic kidney disease and *VPS36* variants cause a severe recessive neurodevelopmental disorder characterised by microcephaly, motor delay, agenesis of the corpus callosum, cerebellar atrophy, seizures, hypotonia, spasticity and early death. Overall, these results demonstrate a scalable approach to screen homozygous losses for improving diagnostic yield and discovering disease-genes in large exome cohorts.

Keywords

Exome, copy number variants, homozygous deletions, large-scale CNV analysis, consanguinity, rare disease diagnosis

Main Manuscript

Introduction

Exome sequencing is standard practice, particularly in resource-limited diagnostic settings. Incorporating copy number variants (CNVs) detection into exome analysis improves diagnostic yield[1–7] but significant obstacles remain in reliable detection, filtering, and interpretation of CNVs in biobank-level short-read exome sequencing datasets[8,9]. Furthermore, systematic studies investigating homozygous CNVs in rare diseases are rare[10,11].

Endogamy is common among Indian populations, contributing to higher inbreeding rates and unique genetic profiles with significant implications for genetic disease risk[12,13]. We previously generated a dataset of point variants in the Indian population using exome sequencing data from families with rare Mendelian disorders which facilitated identification of novel disease genes and human knockouts[14]. However, there has been no systematic study of CNVs in the Indian population despite its genomic uniqueness and relative underrepresentation in current public genomic databases.

Here, we present a CNV calling and genomic position-wise loss-count based filtering and prioritization strategy that significantly reduces the analysis burden for clinical interpretation and increases homozygous loss associated diagnostic yield in existing exome data of individuals from a consanguinity-enriched population in India. We also provide evidence and independent replication to identify pathogenic variants in *VPS36*, *TFCP2L1*, *FILIP1*, and *FAM177A1* as causative for recessive diseases.

Materials and methods

Exomes of individuals affected by rare diseases were collected and CNVs were called in 12 independent batches based on the capture kits used **(Supplementary Methods).** Variants with zero copies (homozygous loss calls) were selected and mapped to the GRCh38 Human Reference Assembly. The count of mapped loss calls at each genomic position was then determined. Genomic positions with loss-call count of fewer than six were identified and rare homozygous loss calls at these positions were filtered and annotated. The calls that overlapped at least one known OMIM autosomal recessive (AR) disease gene and the calls that did not overlap any known OMIM AR disease genes

were prioritised separately using participants' demographic data, Integrative Genomics Viewer (IGV) inspection, internal datasets, and control databases as described in the supplementary data. Prioritized calls were orthogonally tested using quantitative PCR (qPCR). Clinical correlations were performed to identify and report diagnoses in known disease associated genes. Candidate recessive disease genes were also identified, their functional relevance to the individuals' phenotypes studied, and additional similarly affected individuals with biallelic variants in the candidates were searched as detailed in the supplementary data. cDNA and protein studies were also performed using cultured fibroblast cells as described in the supplementary data.

Results

Curation of rare homozygous loss calls from whole exome data

We assembled a cohort of 2,021 individuals, including affected probands and their relatives, from a population enriched for consanguinity and endogamy[14] who had undergone exome sequencing at our centre from 2011 to 2022 using 12 different capture kits **(Figure 1a)**. We developed and applied a CNV calling pipeline to exomes in separate capture kits based batches **(Supplementary Table S1)** and identified 42,386 possible homozygous losses (median 20 per individual, range 0 – 55; median size 2.95 kb, range 99 bp – 4.76 Mb) **(Figure 1b and Supplementary Figure S2)**.

We defined rare calls as those with a relative semi-stringent internal frequency of five or less in our cohort. Usually in large datasets, the frequencies of similar overlapping CNVs are estimated by clustering them without considering the differences in their size and breakpoints[15]. However, we decided to take a genomic position-based approach to reduce the imprecise frequency estimation (i.e. due to dissimilar CNVs being mistakenly interpreted as being similar due to large overlap). For this, we mapped the 42,386 loss calls to GRCh38 human reference assembly, counted loss calls at each base position, and determined positions that overlapped with five or fewer homozygous losses (16.3 million base positions). These positions were then used as anchors to filter the corresponding 1,224 rare homozygous loss calls in 718 individuals (median 1 per individual, range 1 – 22; median size 3.49 kb, range 121 bp – 4.76 Mb) **(Figure 1b)**. 955 (78%) calls overlapped a single gene, 158 (13%) overlapped two genes, and 111 (9%) overlapped three or more genes **(Supplementary Figure S3)**. Only 97 (8%)

loss calls overlapped a known OMIM AR disease gene. Analysis of the demographic data of 718 individuals with rare homozygous loss calls showed that 390 individuals (54%) were male and 328 (46%) were female; 518 (72%) were probands, 166 (23%) were parents (81 fathers and 85 mothers), 21 (3%) were siblings, and 13 (2%) were other relatives; 178 (25%) were unaffected, 300 (42%) had already received a molecular diagnosis, and 240 (33%) were affected and undiagnosed at the time of our analysis **(Figure 1c-d)**.

Overall, our filtering strategy significantly reduced the analysis burden (from median 20 to median 1) and led to the identification of rare homozygous loss calls.

Prioritization of losses with morbid genes improves diagnostic yield

To prioritize potential diagnostic calls from the 1,224 rare homozygous loss calls, we selected those which overlapped one or more OMIM morbid AR genes. This revealed 97 calls in 74 individuals **(Figure 2a)**. Following manual IGV inspection of all 97 calls we identified 25 possibly true calls in 24 individuals. Two calls detected in unaffected individuals were excluded from the analysis **(Supplementary Note 1).** Trio validation of the remaining calls performed by qPCR confirmed 19 out of the remaining 23 to be homozygous in affected individuals and heterozygous in parents **(Supplementary Figure S9)**. Three calls could not be assessed due to sample unavailability, and one was determined to be a false call.

Firstly, we sought to examine the effectiveness of our filtering, prioritization and validation strategy. For this we compiled from our cohort a list of all homozygous deletions in AR genes that had been previously determined as causative variants through selective exome CNV analysis using ExomeDepth[16]. This search revealed nine such instances **(Supplementary Table S4)**. These included seven partial single gene exonic deletions encompassing *CLCNKB* (OMIM #607364), *CYP21A2* (OMIM #201910), *DOCK8* (OMIM #243700), *FUCA1* (OMIM #230000), *GM2A* (OMIM #272750), *NPC2* (OMIM #607625), and *OCLN* (OMIM #251290), and two contiguous multi-gene deletions, one in the 6p21.33 locus encompassing *PSORS1C1*, *C6orf15*, and the morbid *CDSN* (OMIM #270300) and the other in 16p13.3 region removing *HBZ*, *HBM*, *HBA2*, *HBA1*, and *HBQ1* causing hemoglobin Bart hydrops fetalis syndrome (OMIM #604131). Importantly, the 19 prioritized validated homozygous deletions included all nine previous diagnostic findings, thus validating our strategy.

The remaining 10 homozygous deletions were predicted to result in partial or complete loss-of-function of an OMIM-morbid gene **(Figure 2a) (Supplementary Table S4)**. Reverse phenotyping revealed clinical features consistent with known disease descriptions, thus establishing new diagnoses in all 10 individuals **(Figure 2b) (detailed case reports in supplementary data)**.[17] In brief, seven were partial single-gene exonic deletions, one each in *ADPRS* (OMIM #618170), *CNTNAP2* (OMIM #610042), *DGAT1* (OMIM #615863), *FANCA* (OMIM #227650), and *PRKN* (OMIM #600116), and two identical losses in *SLC13A5* (OMIM #615905) in a sib-pair. The remaining three losses resulted in contiguous multigene deletions, one at the 1p36.33 locus affecting *ATAD3B* and *ATAD3A* (OMIM #618810), and two identical losses in the 2q13 locus affecting *MALL, NPHP1* (OMIM #256100), and *MTLN* in two unrelated individuals. Our investigation of HGMD database and the literature revealed that homozygous exonic deletions involving *ADPRS*, and *DGAT1* have not been previously reported **(Supplementary Note 2)**.

Together, these results show that our focused analysis of homozygous CNV losses enabled 10 new diagnoses in 240 unsolved individuals (4.2%) with at least one filtered rare homozygous loss call. This led to a 2.1-fold (from 9 to 19) increase in diagnosis owing to homozygous deletions. Notably, these losses had remained undetected previously, as systematic CNV analysis with the sensitivity to detect single-exon–level changes had not been applied across all the exomes.

Systematic analysis identifies candidate recessive disease genes

Next, we hypothesized that a systematic analysis of our data could lead to the identification of previously unrecognised recessive disorders. We, therefore, analysed the remaining 1,127 filtered rare homozygous loss calls from 681 individuals that did not overlap with any known AR disease genes **(Figure 3a)**. Exclusion of calls based on individuals' attributes, followed by IGV inspection, led to the identification of 96 possibly true calls in affected undiagnosed individuals. Phenotypic assessment-based prioritisation using our internal dataset and evaluation of affected genes in control databases (gnomAD[18], human gene knockout dataset[19], and our local unaffected individuals' database) resulted in selection of 12 relevant loss calls that had at least one gene without homozygous pLoF variant in controls. Orthogonal trio validation performed using qPCR confirmed homozygous losses in the affected individuals and heterozygous carrier state of the respective parents in five families **(Supplementary Figure S9)**. These included seven partial single gene exonic deletions (1-4 deleted exons), with two

identical losses in both *VPS36* and *TFCP2L1* in siblings of two families and one loss each in *FILIP1, FAM177A1*, and *PWWP3A* **(Figure 3b) (Supplementary Table S4)**. To study the possible gene-disease associations, we sought additional similarly affected individuals who had biallelic deleterious variants in the five candidate genes. Individuals were searched via GeneMatcher[20], 100,000 Genomes Project (100KGP)[21] tiering data, international collaborations, and literature **(Supplementary Table S5)**. We also studied the genes' function, expression, and available animal models to understand their role in the conditions.

In Family 21 we identified a homozygous deletion encompassing exons 2-3 of *VPS36* (NM_016075.4) **(Figures 4a)**. Through a GeneMatcher search we identified three more families with homozygous *VPS36* variants and overlapping clinical presentations, along with a fourth family discovered through a literature search. Details of our *VPS36*-related work are provided in the next section.

In Family 22, two siblings with homozygous deletions of exons 1-2 of the *TFCP2L1* gene (NM_014553.3) were identified to have chronic kidney disease, microcephaly, sensorineural hearing loss, and bilateral destruction of the femoral head, likely due to renal osteodystrophy **(Figure 3b)**. Another sibling with similar presentations and two spontaneous abortions of unknown cause were reported by the family **(Supplemental Data)**. The homozygous loss identified in the two siblings is likely to result in a null allele due to the loss of transcription start site **(Supplementary Figure S4)**. Through literature search, we identified another individual with a homozygous frameshift *TFCP2L1* variant (NM_014553.3:c.869del p.(Ser290Phefs*50)), end-stage renal disease, congenital cataracts, hypotonia, seizures, and global developmental delay[22]. The variant was heterozygous in the parents and a non-identical unaffected twin and was predicted to result in premature termination of the protein prior to the sterile alpha motif (SAM)-like domain likely causing nonsense-mediated mRNA decay. *TFCP2L1* ortholog is expressed in mouse kidneys at both embryonic and adult stages and has been shown to induce growth and maturation of renal tubules in rat mesenchymal cells[22]. The orthologous transcription factor *Cp2l1* in mouse kidneys is also necessary for duct maturation[23], indicating an important role of this gene in normal kidney function. Collectively, phenotypic convergence among the unrelated individuals identified in literature and our patients with chronic kidney disease, the deleterious effect of the variants on gene **(Supplementary Figure S4)**, and insights from animal studies together

suggest that biallelic loss-of-function variants in *TFCP2L1* are associated with chronic kidney disease. The cause of other presentations in the two families is unclear and emphasizes on the need for a larger patient cohort to understand the full spectrum of phenotypes that may occur.

In Family 23, we identified an individual with homozygous *FILIP1* exon 3-6 deletion (NM_015687.5), multiple joint contractures, scoliosis, microcephaly, and facial dysmorphism **(Figure 3b)**. Our search identified seven individuals from four families with biallelic truncating variants in *FILIP1*. Two of these families, identified via GeneMatcher, had two distinct homozygous stop gains in three siblings (NM_015687.5:c.463G>T p.(Glu155*)) and one child (NM_015687.5:c.2665C>T p.(Arg889*)), respectively. A study detailing the genetic and clinical findings in these families, and in the individual we identified, was recently published seperately[24]. Another non-consanguineous British family was detected in the 100KGP with an individual with homozygous *FILIP1* stop gain variant (NM_015687.5:exon 5:c.2152C>T p.(Arg718*)), presenting with joint contractures and dysmorphic features[25]. We also found a family in the literature with stop gain variants (NM_015687.5:exon 2:c.169C>T p.(Arg57*)) for whom the two affected female siblings, homozygous for the variant, presented with hypotonia, joint contractures, motor delay, and delayed speech development[26]. *FILIP1* regulates filamin homeostasis through filamin A (FLNA) degradation[27]. It is required for the differentiation of cross-striated muscle cells during myogenesis[28] and shows robust gene expression in both skeletal and smooth muscle[29]. It is also implicated in the regulation of neuronal cell positioning in the cortex[30]. Taken together, the phenotypic concurrence in the affected individuals, the predicted deleteriousness of all the identified variants **(Supplementary Figure S4)**, and the functional relevance of *FILIP1* in myogenesis, prove that biallelic loss-of-function variants in *FILIP1* are associated with a syndromic arthrogryposis.

In Family 24, we identified an individual with homozygous *FAM177A1* exon 5 loss (NM_173607.5) and spastic diplegia **(Figure 3b)**. We also found a 43-year-old female with progressive neurological deterioration (detailed phenotype not available) and homozygous *FAM177A1* frameshift variant (NM_173607.5:c.218del p.(Lys73Argfs*93)) in the 100KGP. Survey of the literature revealed nine additional individuals from four families with biallelic truncating *FAM177A1* variants and neurodevelopmental phenotypes[31,32]. Variants in the identified individuals and the variant in our

individual with spastic diplegia were predicted to cause a loss of function as a result of protein truncation **(Supplementary Figure S4)**. *FAM177A1* is a ubiquitously expressed gene that encodes a Golgi complex protein of unknown function. It is thought that the gene is involved in intracellular transport of lipids between Golgi sub-compartments by assisting VPS13B as a functional partner[33]. Collectively, the implicated functional role of *FAM177A1* in lipid transportation pathway, the disruption of which leads to similar phenotypes, the predicted deleteriousness of the identified variants, and the phenotypic concordance observed in individuals from unrelated families provide compelling evidence to associate biallelic loss-of-function variants in *FAM177A1* with neurodevelopmental disorder.

Collectively, our systematic approach to prioritize rare homozygous loss calls, which did not overlap with any known AR disease genes, led to the identification of candidate autosomal recessive diseases which include *TFCP2L1*-linked renal tubulopathy, *FILIP1*-related arthrogryposis with microcephaly, and *FAM177A1*-associated motor dysfunction. The contribution of *PWWP3A* to the observed clinical features in Family 25 remains undetermined as described in the supplementary data **(Supplementary Note 3)**.

Bi-allelic loss of function *VPS36* variants are associated with a neurodevelopmental syndrome

In Family 21, in three siblings with microcephaly, motor delay, agenesis of the corpus callosum, cerebellar atrophy, seizures, and spasticity, we identified a homozygous frameshift deletion encompassing exons 2-3 of *VPS36* (NM_016075.4) **(Figures 4a-c) (Supplementary Figure S4) (Supplementary Table S6)**. *VPS36* encodes EAP45, a component of the heterotetrameric endosomal sorting required for transport (ESCRT)-II subcomplex. Importantly, this complex includes *SNF8* encoded EAP30/VPS22[34] **(Supplementary Figure S7)** and the clinical features of affected individuals in family 21 showed a remarkable overlap with the known phenotype of disorders resulting from bi-allelic *SNF8* variants (Developmental and epileptic encephalopathy 115, OMIM #620783 and neurodevelopmental disorder plus optic atrophy, OMIM # 620784)[35].

Literature search identified a homozygous *VPS36* likely splice-disrupting variant (NM_016075.4: intron 13: c.1068-2A>T; SpliceAI[36] acceptor loss delta score of 0.97) reported in a study of diagnostic panels and exomes from Saudi Arabia[37] **(Figure 4b)**. A gene-driven phenotype-agnostic GeneMatcher search

revealed an identical homozygous *VPS36* missense variant (NM_016075.4:c.1103G>A p.(Arg368His)) in two Pakistani and one Saudi Arabian family. The variant is rare in the heterozygous (in eight of 806,918 individuals) and absent in the homozygous state in gnomAD. The variant has a high Combined Annotation Dependent Depletion (CADD) score of 33[38] and predicted to disrupt native hydrogen bonding of residue R368 with E361, thereby affecting the structure and stability of EAP45 **(Supplementary Figure S7)**. We found evidence of recent common ancestry between the two F51 and F53 Pakistani families, but this haplotype was not shared with the Saudi Arabian family **(Supplementary Table S7)**.

The individual identified via literature search was documented to have speech delay and intellectual disability, but no other phenotype details were available[37]. Family 51 consisted of two deceased male siblings, born to consanguineous parents of Pakistani origin, with microcephaly, corpus callosal agenesis, motor delay, intellectual disability, seizures, and hypotonia. Family 52 included a male child, born to consanguineous parents from Saudi Arabia, with microcephaly, motor delay, dystonia, cerebellar atrophy, agenesis of the corpus callosum, and ventriculomegaly **(Figure 4a-d and Supplementary Table S6)**. Notably, in Family 52, an older similarly affected female sibling had died at 21 months of age. The *VPS36* variant in her could not be confirmed due to sample unavailability. However, previous SNP array data from her had shown a region of homozygosity, on chromosome 13, which includes *VPS36*. Family 53 consisted of two deceased males and one deceased female born to consanguineous parents of Pakistani origin. The female and one male had developmental delay, corpus callosum agenesis, ventriculomegaly, hypotonia and seizures (clinical in the brother and sub-clinical in the sister). The other male had corpus callosum abnormalities and ventriculomegaly but an additional skeletal dysplasia picture. Genetic testing confirmed a homozygous *VPS36* variant in the female. Testing of the males was not yet possible due to sample unavailability.

Complementary DNA (cDNA) from fibroblasts of individual F21:II.3 demonstrated loss of the two exons in the *VPS36* transcript **(Supplementary Figure S5)**. Western blot analysis revealed a marked reduction in VPS36 protein levels in the fibroblast sample as compared to the control samples **(Supplementary Figure S6)**. It has been previously shown that presence of all ESCRT-II subunits is essential for its stability and if any one subunit is missing, the levels of the other subunits are reduced[35].

Collectively, the rarity and deleterious nature of the *VPS36* variants, the striking phenotypic convergence observed in multiple affected individuals from unrelated families, and the functional relevance of the gene, and the significant clinical overlap with the known *SNF8*-related disorders suggest that bi-allelic *VPS36* variants cause a neurodevelopmental disorder.

Discussion

Most large scale CNV studies from WES or WGS utilise overlaps of CNVs to classify them as similar and to determine their allele frequency[6,15]. However, this approach overlooks the fact that CNVs sharing a common overlap can still vary in size and breakpoints, potentially disrupting distinct genomic regions. Consequently, this may erroneously exclude CNVs affecting pathogenic loci if their allele frequency is inflated due to overlapping yet non-identical CNVs. To address this limitation, we prioritised rare CNVs based on their frequency at the individual nucleotide level, using a genomic position-wise loss-count based filtering strategy which reduced the analysis burden without missing the previously known homozygous loss-associated diagnoses. Importantly, this approach also revealed new diagnoses and led to the identification of new candidate disease genes. Overall, homozygous losses accounted for phenotypes in 26 of the 240 (10.8%) unsolved individuals in our cohort who had at least one rare homozygous loss call.

However, there are several limitations of the approach. Firstly, we experienced a high false detection rate **(Supplementary Table S8)**. This is possibly due to the highly heterogeneous sequencing data collected over a period of 12 years from various service providers. Although we excluded uncertain calls from our analysis, they can still represent valid biological losses. Furthermore, in our analysis, we focused only on MANE Select canonical transcripts[39]. While MANE transcripts offer consistency, analysing alternative isoforms would be crucial for comprehending the complete functional consequences of the losses and the underlying disease mechanisms. Moreover, exact breakpoints were not assessed in most cases, and the possibility of the loss being part of complex rearrangements cannot be ruled out. To prioritize loss calls in our analysis of new candidate recessive disease genes, we excluded calls from previously diagnosed individuals. Although these individuals have a confirmed diagnosis, they may still carry a pathogenic loss in a gene not yet linked to disease. This could contribute

to a blended phenotype or a presentation closely resembling the known disease associated with the diagnosed gene, raising the possibility of a dual diagnosis. Moreover, we did not explore the functional consequences of the losses identified in unaffected individuals. Such losses could represent natural human knockouts that do not lead to a clinical phenotype, offering broader implications for understanding the functional roles of such genes[19]. Potentially, supplementing our approach with paired-end and split read-based structural variant (SV) detection tools [7,40] and long-read sequencing technologies may help to address some of these issues at scale. While we used cn.MOPS for detecting homozygous deletions, a key strength of our method lies in the downstream filtering and prioritisation framework which is largely tool-agnostic and can be applied alongside any robust homozygous loss detection algorithm to improve the discovery of clinically relevant deletions.

Our work provides evidence to support the existence of *FILIP1*-related arthrogryposis with microcephaly, *TFCP2L1*-related renal tubulopathy, *FAM177A1*-related motor dysfunction and *VPS36*-related neurodevelopmental disorder. *FILIP1*-related congenital neuromuscular disorder with dysmorphic facies (OMIM #620775) is now indexed in OMIM and has been shown to be characterized by impaired skeletal muscle development, usually resulting in hypotonia and secondary joint contractures, and dysmorphic facial features. Overall, this validates our approach. With regards to *TFCP2L1*-related renal tubulopathy, it is noteworthy that the affected individuals also exhibit other phenotypes (microcephaly, sensorineural hearing loss, congenital cataracts, hypotonia, seizures, and global developmental delay) that are not directly attributable to the renal dysfunction. This could be due to the known role of *TFCP2L1* beyond kidney development and function[41–43]. *FAM177A1*-related motor dysfunction seems to be characterized by global developmental delay, seizures, behavioral abnormalities, hypotonia, and gait disturbance. Although the cellular function of FAM177A1 is not fully understood, it has been observed to localize in the Golgi subcompartments and may be involved in the intracellular lipid and protein trafficking[33]. Identification of more individuals with these conditions will be important for better characterisation of the disorders.

We also provide genetic, clinical and functional evidence to support a relationship between homozygous *VPS36* variants and a neurodevelopmental disorder characterised by microcephaly, motor delay, agenesis of the corpus callosum, cerebellar atrophy, seizures, hypotonia, spasticity, and early death.

Although the precise mechanism by which this deletion causes disease is unclear, it is likely that the deletion impacts the stability of the ESCRT-II complex, of which VPS36 is a subunit, thereby disrupting the function of this highly conserved assembly. This ESCRT system comprises three distinct multi-subunit complexes, ESCRT-I, ESCRT-II, and ESCRT-III, and plays a critical role in various cellular processes, including cytokinesis, endosome maturation, autophagy, and membrane repair[44]. In humans, the ESCRT-II complex includes one subunit each of VPS36 (EAP45) and SNF8 (EAP30/VPS22) and two subunits of VPS25 (EAP20)[34]. Mouse knockout models of *Vps36* and other ESCRT subunits result in early embryonic or preweaning lethality[35,45]. Another subunit of the ESCRT-II complex, SNF8, has been associated with remarkably similar phenotypes, and biallelic variants in *SNF8* have been shown to reduce the protein levels of other ESCRT-II subunits (VPS36 and VPS25)[35]. Furthermore, several other genes encoding subunits of the ESCRT (consisting of ESCRT I to III subcomplexes) machinery have been implicated in a range of neurodevelopmental disorders in humans, such as AR spastic paraplegia 53 (*VPS37A*; OMIM #614898), AD spastic paraplegia 80 (*UBAP1*; OMIM #618418), AR pontocerebellar hypoplasia type 8 (*CHMP1A*; OMIM #614961), AD frontotemporal dementia and/or amyotrophic lateral sclerosis 7 (*CHMP2B*; OMIM #600795), and AD CIMDAG syndrome (*VPS4A*; OMIM #619273). This suggests that the ESCRT machinery plays an important role in normal development and its impairment can lead to a spectrum of neurodevelopmental conditions in humans.

In summary, employing filtering strategies for CNV screening can streamline the analysis process. By integrating these techniques with *in silico* phenotype matching, it is possible to scale up the analysis and make it feasible for large cohorts. Although we focussed on identification of disease-causing homozygous deletions, it is possible to customise this approach to identify heterozygous CNVs or to analyse copy number gains from large genomic datasets. To conclude, we developed a CNV calling pipeline to identify homozygous losses from exomes and applied filtering and prioritisation strategies to facilitate diagnostic screening and rare disease discovery in a cohort of 2,021 individuals. Our methods increased homozygous loss associated diagnoses by more than two-fold and defined four new recessive diseases caused by biallelic variants in *VPS36*, *TFCP2L1*, *FILIP1,* and *FAM177A1*.

Data and code availability

The data and code used in this study are available from the corresponding author upon reasonable request.

Author contribution statement

Funding was acquired by SB, CAJ and KMG. The study was conceptualized by AC, SB, and KMG. Methodology was developed by AC, AS, NK, PU, SSN, SS, VLH, CMW, EH, AEF, JAP, PRK, SB, and KMG. Investigations were conducted by all authors. Data curation and analyses were performed by AC, AS, SP, GSLB, PU, NQ, PM, AA, SB, and KMG. Project administration and supervision were provided by PRK, WGN, CAJ, SB, and KMG. The original manuscript draft was prepared by AC, SB and KMG which was reviewed and approved by all authors.

Competing interests

Authors declare no competing interests.

Acknowledgements

The authors express gratitude to all the patients and their family members who consented to sequencing their exomes at participating centers.

Ethical approval

This study was approved by the Kasturba Medical College and Kasturba Hospital Institutional Ethics Committee (IEC:856/2020). Informed consent was obtained from all participants for exome sequencing and for publication of clinical details (including pedigree charts).

Funding information

1. This work was supported by the Global Challenges Research Fund Institutional Support Grant to the University of Manchester 2018 to 2021.
2. DBT/Wellcome Trust India Alliance for funding the study, “Centre for Rare Disease Diagnosis, Research and Training” (IA/CRC/20/1/600002).

3. National Institutes of Health, United States, for funding the study, “Genetic Diagnosis of Neurodevelopmental Disorders in India” (1R01HD093570-01A1).
4. Department of Biotechnology, Government of India, for funding the study, “Deep phenotyping, comprehensive genomic studies and investigations into pathomechanisms of congenital heart defects” (BT/PR40007/MED/97/490/2020)
5. DBT/Wellcome Trust India Alliance for funding the study, “Understanding autoinflammatory disorders through clinical, genomic and functional approaches” (IA/CPHE/20/1/505226)
6. Indian Council of Medical Research, for funding the study, “Genetic evaluation of inborn errors of immunity in Indian population” (5/7/1740/CH/Adhoc/2021-RBMCH)
7. Department of Health Research, Government of India. Grant numbers: V.25011/379/2015-GIA/HR and R.11012/02/2018-HR
8. Indian Council of Medical Research, Government of India. Grant numbers: 54/2/2013-HUM-BMS, No.4/13/58/2015/NCD-III, 5/7/1508/2016-CH, ISRM/12(40)/2019, 63/8/2010-BMS, and F. No. 63/01/2019-Genomics/BMS
9. Science and Engineering Research Board, Government of India. Grant numbers: SB/SO/HS/005/2014, YSS/2015/002009, and YSS/2015/001681
10. Department of Biotechnology, Government of India. Grant numbers: BT/Bio-CARe/07/9889/2013-14, BT/PR9635/MED/97/198/2013, BT/PR13921/MED/12/704/2015, and BT/PR3193/MED/12/521/2011
11. Medical Research Council/National Institute for Health and Care Research (MRC/NIHR) Clinical Academic Research Partnership award (CARP) fellowship MR/V037617/1 (V.H.)
12. Sir Jules Thorn Award “An international resource for the autozygosity mapping and identification of recessive disease genes in consanguineous families” (09/JTA)
13. Medical Research Council "Functional genomics identification and characterization of novel disease genes, mechanisms and pathways of ciliogenesis (MR/M000532/1)
14. This study has been delivered through the National Institute for Health and Care Research (NIHR) Manchester Biomedical Research Centre (BRC) (NIHR203308). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Web resources

OMIM, https://www.omim.org/

gnomAD, https://gnomad.broadinstitute.org/

Genotype-Tissue Expression (GTEx) Portal, https://www.gtexportal.org/home/

GeneMatcher, https://genematcher.org/

The Human Gene Mutation Database (HGMD®), https://www.hgmd.cf.ac.uk/ac/index.php

The International Mouse Phenotyping Consortium (IMPC), https://www.mousephenotype.org/

The Zebrafish Information Network (ZFIN), https://zfin.org/

References

1 Pfundt R, Rosario M del, Vissers LELM, Kwint MP, Janssen IM, Leeuw N de *et al.* Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genetics in Medicine* 2017; **19**: 667–675.

2 Marchuk DS, Crooks K, Strande N, Kaiser-Rogers K, Milko LV, Brandt A *et al.* Increasing the diagnostic yield of exome sequencing by copy number variant analysis. *PLoS One* 2018; **13**: e0209185.

3 Bergant G, Maver A, Lovrecic L, Čuturilo G, Hodzic A, Peterlin B. Comprehensive use of extended exome analysis improves diagnostic yield in rare disease: a retrospective survey in 1,059 cases. *Genetics in Medicine* 2018; **20**: 303–312.

4 Testard Q, Vanhoye X, Yauy K, Naud M-E, Vieville G, Rousseau F *et al.* Exome sequencing as a first-tier test for copy number variant detection: retrospective evaluation and prospective screening in 2418 cases. *Journal of Medical Genetics* 2022; **59**: 1234–1240.

5 Atik T, Durmusalioglu EA, Isik E, Kose M, Kanmaz S, Aykut A *et al.* Diagnostic yield of exome sequencing-based copy number variation analysis in Mendelian disorders: a clinical application. *BMC Medical Genomics* 2024; **17**: 239.

6 Lemire G, Sanchis-Juan A, Russell K, Baxter S, Chao KR, Singer-Berk M *et al.* Exome copy number variant detection, analysis, and classification in a large cohort of families with undiagnosed rare genetic disease. *The American Journal of Human Genetics* 2024; **111**: 863–876.

7 Demidov G, Laurie S, Torella A, Piluso G, Scala M, Morleo M *et al.* Structural variant calling and clinical interpretation in 6224 unsolved rare disease exomes. 2023; : 2023.10.28.23297720.

8 Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology* 2015; **3**: 92.

9 Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Human Mutation* 2015; **36**: 815–822.

10 Gambin T, Akdemir ZC, Yuan B, Gu S, Chiang T, Carvalho CMB *et al.* Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Res* 2017; **45**: 1633–1648.

11 Olinger E, Wilson IJ, Orr S, Barroso-Gil M, Neatu R, Ambrose JC *et al.* Copy-number analysis from genome sequencing data of 11,754 rare-disease parent-child trios: A model for identifying autosomal recessive human gene knockouts including a novel gene for autosomal recessive retinopathy. *Genetics in Medicine Open* 2024; **2**. doi:10.1016/j.gimo.2024.101834.

12 Machha P, Gopalan A, Elangovan Y, Veeravalli SCM, Sowpati DT, Thangaraj K. Endogamy and high prevalence of deleterious mutations in India: evidence from strong founder events. *Journal of Genetics and Genomics* 2025. doi:10.1016/j.jgg.2025.02.001.

13 Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N *et al.* The promise of discovering population-specific disease-associated genes in South Asia. *Nat Genet* 2017; **49**: 1403–1407.

14 Kausthubham N, Shukla A, Gupta N, Bhavani GS, Kulshrestha S, Das Bhowmik A *et al.* A data set of variants derived from 1455 clinical and research exomes is efficient in variant prioritization for early-onset monogenic disorders in Indians. *Human Mutation* 2021; **42**. doi:10.1002/humu.24172.

15 Babadi M, Fu JM, Lee SK, Smirnov AN, Gauthier LD, Walker M *et al.* GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nature Genetics* 2023; **55**: 1589–1597.

16 Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 2012; **28**: 2747–2754.

17 de Goede C, Yue WW, Yan G, Ariyaratnam S, Chandler KE, Downes L *et al.* Role of reverse phenotyping in interpretation of next generation sequencing data and a review of INPP5E related disorders. *European Journal of Paediatric Neurology* 2016; **20**: 286–295.

18 Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020; **581**: 434–443.

19 Narasimhan VM, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science* 2016; **352**: 474–477.

20 Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. *Human Mutation* 2015; **36**: 928–930.

21 Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB *et al.* The 100 000 Genomes Project: Bringing whole genome sequencing to the NHS. *BMJ (Online)* 2018; **361**. doi:10.1136/bmj.k1687.

22 Klämbt V, Werth M, Onuchic-Whitford AC, Getwan M, Kitzler TM, Buerger F *et al.* Mutations in transcription factor CP2-like 1 may cause a novel syndrome with distal renal tubulopathy in humans. *Nephrology Dialysis Transplantation* 2021; **36**: 237–246.

23 Yamaguchi Y, Yonemura S, Takada S. Grainyhead-related transcription factor is required for duct maturation in the salivary gland and the kidney of the mouse. *Development* 2006; **133**: 4737–4748.

24 Schnabel F, Schuler E, Al-Maawali A, Chaurasia A, Syrbe S, Al-Kindi A *et al.* Homozygous loss-of-function variants in FILIP1 cause autosomal recessive arthrogryposis multiplex congenita with microcephaly. *Human Genetics* 2023; **142**: 543–552.

25 Watts LM, Bunyan DJ, Giacopuzzi E, Walker S, Gazdagh G, Thomas NS *et al. FILIP1* -associated neuromuscular disorder and phenotypic blending due to paternal UPD6. *Brain Communications* 2024; **6**: fcae330.

26 Roos A, van der Ven PFM, Alrohaif H, Kölbel H, Heil L, Della Marina A *et al.* Bi-allelic variants of FILIP1 cause congenital myopathy, dysmorphism and neurological defects. *Brain* 2023; **146**: 4200–4216.

27 Nagano T, Yoneda T, Hatanaka Y, Kubota C, Murakami F, Sato M. Filamin A-interacting protein (FILIP) regulates cortical cell migration out of the ventricular zone. *Nat Cell Biol* 2002; **4**: 495–501.

28 Mao Z, Nakamura F. Structure and Function of Filamin C in the Muscle Z-Disc. *International Journal of Molecular Sciences* 2020; **21**: 2696.

29 Sato M, Nagano T. Involvement of Filamin A and Filamin A-interacting protein (FILIP) in controlling the start and cell shape of radially migrating cortical neurons. *Anatomical Science International* 2005; **80**: 19–29.

30 Yagi H, Oka Y, Komada M, Xie M-J, Noguchi K, Sato M. Filamin A interacting protein plays a role in proper positioning of callosal projection neurons in the cortex. *Neuroscience Letters* 2016; **612**: 18–24.

31 Alazami AM, Patel N, Shamseldin HE, Anazi S, Al-Dosari MS, Alzahrani F *et al.* Accelerating Novel Candidate Gene Discovery in Neurogenetic Disorders via Whole-Exome Sequencing of Prescreened Multiplex Consanguineous Families. *Cell Reports* 2015; **10**: 148–161.

32 Kohler JN, Legro NR, Baldridge D, Shin J, Bowman A, Ugur B *et al.* Loss of function of *FAM177A1*, a Golgi complex localized protein, causes a novel neurodevelopmental disorder. *Genetics in Medicine* 2024; **26**: 101166.

33 Ugur B, Schueder F, Shin J, Hanna MG, Wu Y, Leonzino M *et al.* VPS13B is localized at the cis-trans Golgi complex interface and is a functional partner of FAM177A1. *bioRxiv : the preprint server for biology* 2023. doi:10.1101/2023.12.18.572081.

34 Hierro A, Sun J, Rusnak AS, Kim J, Prag G, Emr SD *et al.* Structure of the ESCRT-II endosomal trafficking complex. *Nature* 2004; **431**: 221–225.

35 Brugger M, Lauri A, Zhen Y, Gramegna LL, Zott B, Sekulić N *et al.* Bi-allelic variants in SNF8 cause a disease spectrum ranging from severe developmental and epileptic encephalopathy to syndromic optic atrophy. *The American Journal of Human Genetics* 2024; **111**: 594–613.

36 Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 2019; **176**: 535-548.e24.

37 Monies D, Abouelhoda M, AlSayed M, Alhassnan Z, Alotaibi M, Kayyali H *et al.* The landscape of genetic diseases in Saudi Arabia based on the first 1000 diagnostic panels and exomes. *Hum Genet* 2017; **136**: 921–939.

562 38 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for
563 estimating the relative pathogenicity of human genetic variants. *Nature Genetics* 2014; **46**: 310–
564 315.

565 39 Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A *et al.* A joint NCBI and EMBL-EBI
566 transcript set for clinical genomics and research. *Nature* 2022; **604**: 310–315.

567 40 Gardner EJ, Sifrim A, Lindsay SJ, Prigmore E, Rajan D, Danecek P *et al.* Detecting cryptic clinically
568 relevant structural variation in exome-sequencing data increases diagnostic yield for
569 developmental disorders. *Am J Hum Genet* 2021; **108**: 2186–2194.

570 41 Wang X, Wang X, Zhang S, Sun H, Li S, Ding H *et al.* The transcription factor TFCP2L1 induces
571 expression of distinct target genes and promotes self-renewal of mouse and human embryonic
572 stem cells. *Journal of Biological Chemistry* 2019; **294**: 6007–6016.

573 42 Ye S, Li P, Tong C, Ying Q-L. Embryonic stem cell self-renewal pathways converge on the
574 transcription factor Tfcp2l1. *EMBO J* 2013; **32**: 2548–2560.

575 43 Qiu D, Ye S, Ruiz B, Zhou X, Liu D, Zhang Q *et al.* Klf2 and Tfcp2l1, Two Wnt/β-Catenin Targets,
576 Act Synergistically to Induce and Maintain Naive Pluripotency. *Stem Cell Reports* 2015; **5**: 314–
577 322.

578 44 Vietri M, Radulovic M, Stenmark H. The many functions of ESCRTs. *Nat Rev Mol Cell Biol* 2020;
579 **21**: 25–42.

580 45 Brown SDM, Moore MW. The International Mouse Phenotyping Consortium: past and future
581 perspectives on mouse phenotyping. *Mamm Genome* 2012; **23**: 632–640.

582

583

Figure legends

Figure 1: Homozygous loss calling, filtering, and annotation. (**a**) Analysis workflow. In total, 2,021 exomes from affected individuals and their relatives were collected for CNV calling. Exomes were processed in 12 separate batches based on the capture kits used. A total of 42,386 homozygous losses were called from these exomes. Calls were then mapped to the GRCh38 human reference assembly, and a position-wise count of the overlapping loss calls was determined. Positions with count of fewer than six were identified and calls at these positions were selected. The resulting 1,224 filtered rare homozygous loss calls from 718 individuals were annotated with genes, associated diseases, and inheritance patterns. Loss calls with or without known OMIM morbid autosomal recessive (AR) genes were prioritized separately using the attributes of individuals. Prioritized calls were then validated and clinically interpreted. (**b**) Per individual homozygous loss call counts before and after filtering. After CNV calling, a median of 20 homozygous losses was identified per individual (range, 0–55). Filtering reduced this number to a median of 1 filtered homozygous loss call per individual (range, 1–22). (**c**) relationship and (**d**) disease and diagnostic status of individuals with filtered rare homozygous loss calls.

Figure 2: Prioritization, validation, and interpretation of filtered rare homozygous loss calls with a morbid AR gene and the resulting new diagnoses. (**a**) Results of prioritization, validation, and clinical interpretation. Calls overlapping known AR disease genes were inspected using IGV. Possibly true calls in the affected individuals were assessed for validity using qPCR. Confirmed homozygous losses were interpreted to establish diagnoses. (**b**) New diagnoses identified using this method. The figure shows pedigree of families with confirmed losses and position of the loss in morbid gene. Each family pedigree illustrates the carriers, affected individuals (solid symbols), and test status of the identified deletion (+/−, heterozygous; −/−, homozygous; NA, sample not available). Gene diagrams represent the affected exons and protein domains. The displayed size corresponds to the represented genomic segment, and not to the size of the deletions. Previously undescribed homozygous losses are marked with ☆.

Figure 3: Prioritization of filtered rare homozygous loss calls without morbid AR genes and association study of the resulting candidate recessive disease genes. (**a**) Results of prioritization, validation, and gene-disease association studies. Calls from affected undiagnosed individuals were inspected using IGV. Possibly true calls with no similar calls in unaffected or phenotypically non-

matching individuals were selected. From these, calls overlapping genes with no homozygous pLoF variants in control databases were validated. Candidate recessive genes in confirmed homozygous losses were searched for additional similarly affected individuals with biallelic variants, and were studied for their function, expression in relevant tissues, and available animal models recapitulating the patient's phenotype. (**b**) Pedigrees of the families with candidate recessive disease genes. Figure shows pedigrees of the families with confirmed losses and the genomic position of each loss. Each family pedigree depicts carriers, affected individuals (solid symbols), and the test status of the identified deletion (+/−, heterozygous; −/−, homozygous; NA, sample not available).

Figure 4: Genotypes, phenotypes, and brain MRI scans of individuals with bi-allelic VPS36 variants. (**a**) Pedigrees of the four families with VPS36 variants, illustrating carriers, affected individuals (solid symbols), and the test status of the identified variants (+/−, heterozygous; −/−, homozygous; NA, sample not available). (**b**) Schematic representation of the positions of identified variants across the VPS36 gene. (**c**) Key phenotypic characteristics of the seven individuals with homozygous VPS36 variants. (**d**) T1 and T2-weighted MRI scans of affected individuals' brains, demonstrating key abnormalities. Axial and sagittal slices of the following individuals: F21 II.2 (first row; panels i–vi), F21 II.3 (second row; vii–xii), F51 II.1 (third row; xiii–xviii), F51 II.2 (fourth row; xix–xxiv), and F52 II.2 (fifth row; xxv–xxx) reveal pontocerebellar hypoplasia, posterior predominant ventriculomegaly due to agenesis of corpus callosum and white matter volume loss, and cerebral atrophy.

Supplementary Information

Supplementary Methods

Supplementary Notes

Supplementary Note 1: Homozygous loss calls with OMIM AR disease genes detected in unaffected relatives

Supplementary Note 2: Known homozygous losses in diagnosed disease-causing genes

Supplementary Tables

Supplementary Table S1: Capture kits used for enrichment

Supplementary Table S2: IGV inspection framework

Supplementary Table S3: Primer pairs for qPCR testing of DNA copy number

Supplementary Table S4: Valid diagnoses and candidate recessive disease genes

Supplementary Table S5: Additional individuals with biallelic variants in candidate recessive genes

Supplementary Table S6: Clinical features of individuals with homozygous VPS36 variants

Supplementary Table S7: Homozygous variants flanking the VPS36 c.1103G>A variant in three families

Supplementary Table S8: Breakdown of possibly true, possibly false, and uncertain homozygous loss calls identified through IGV inspection

Supplementary Figures

Supplementary Figure S1: Example IGV screenshots of possibly true, possibly false, and uncertain homozygous loss calls

Supplementary Figure S2: Capture kit-wise breakdown of homozygous loss calls

Supplementary Figure S3: Gene annotations of the filtered rare homozygous loss calls

Supplementary Figure S4: Predicted effect of variants in candidate recessive disease genes

Supplementary Figure S5: Sanger sequencing of cDNA derived from patient fibroblast showing VPS36 exon 2 and 3 deletion

Supplementary Figure S6: Immunoblot analyses shows reduced expression of VPS36 protein in affected individual's fibroblast

Supplementary Figure S7: Structure of the ESCRT-II complex and the predicted impact of p.Arg368His substitution on VPS36 protein structure.

Supplementary Figure S8: IGV screenshots of homozygous loss calls

Supplementary Figure S9: qPCR validation of homozygous loss calls

Supplementary Clinical Reports

a

2,021 exomes

SUPRATECH

FOCUSED

MGI v5

CEv5

CEv3

TWIST

TruSeq

CEv4

V6

CREv3

NE

CREv2

12 batches based on capture kits

CNV calling

42,386 homozygous loss calls in 2007 individuals

Filtering

1,224 filtered homozygous loss calls in 718 individuals

Annotation

Annotated loss calls

Prioritization

Calls with at lease one **OMIM morbid AR** gene

Calls **without any OMIM morbid AR** genes

Diagnoses involving known disease genes

Previously undelineated disorders

Reference genome

Homozygous loss call

Genomic positions with no homozygous loss call

Genomic positions with less than six homozygous loss calls

Genomic positions with six or more homozygous loss calls

* Filtered homozygous loss calls (at position with less than six loss calls)

99%

2.9 billion bases

0.84%

24.2 million bases

33%

67%

b

60

50

40

30

20

10

0

Per individual homozygous loss calls

Per individual filtered homozygous loss calls

c

Sibling, 21, 3%

Mother, 85, 12%

Father, 81, 11%

n = 718

Other relatives, 13, 2%

Proband, 518, 72%

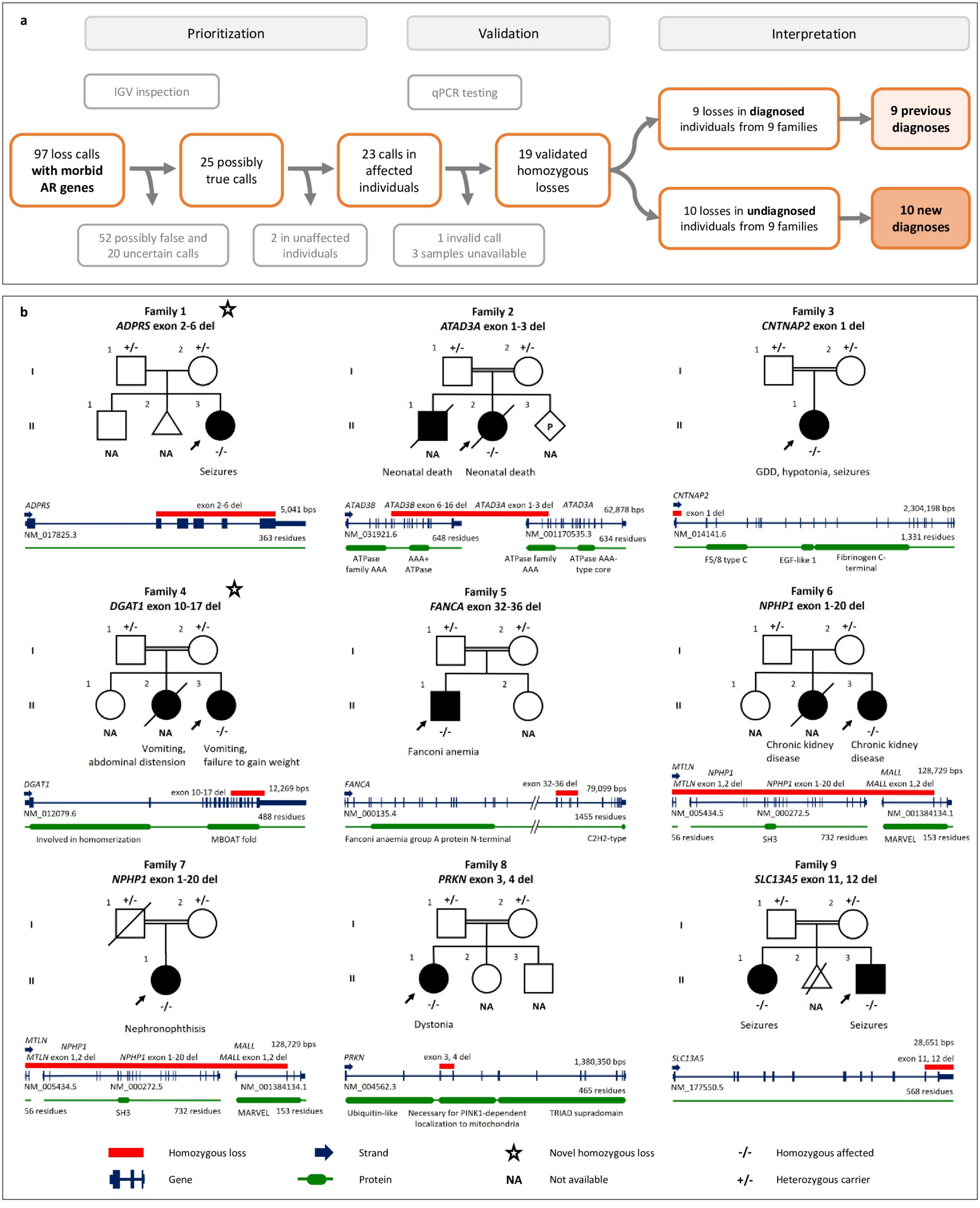
d

Unaffected, 178, 25%

n = 718

Undiagnosed, 240, 33%

Diagnosed, 300, 42%


a
Prioritization
Validation
Interpretation
IGV inspection
qPCR testing
97 loss calls with morbid AR genes
25 possibly true calls
23 calls in affected individuals
19 validated homozygous losses
9 losses in diagnosed individuals from 9 families
9 previous diagnoses
10 losses in undiagnosed individuals from 9 families
10 new diagnoses
52 possibly false and 20 uncertain calls
2 in unaffected individuals
1 invalid call 3 samples unavailable
b
Family 1
ADPRS exon 2-6 del
Seizures
ADPRS
exon 2-6 del
5,041 bps
NM_017825.3
363 residues
Family 2
ATAD3A exon 1-3 del
Neonatal death
Neonatal death
ATAD3B
ATAD3B exon 6-16 del
ATAD3A exon 1-3 del
ATAD3A
62,878 bps
NM_031921.6
648 residues
NM_001170535.3
634 residues
ATPase family AAA
AAA+ ATPase
ATPase family AAA
ATPase AAA-type core
Family 3
CNTNAP2 exon 1 del
GDD, hypotonia, seizures
CNTNAP2
exon 1 del
2,304,198 bps
NM_014141.6
1,331 residues
F5/8 type C
EGF-like 1
Fibrinogen C-terminal
Family 4
DGAT1 exon 10-17 del
Vomiting, abdominal distension
Vomiting, failure to gain weight
DGAT1
exon 10-17 del
12,269 bps
NM_012079.6
488 residues
Involved in homomerization
MBOAT fold
Family 5
FANCA exon 32-36 del
Fanconi anemia
FANCA
exon 32-36 del
79,099 bps
NM_000135.4
1455 residues
Fanconi anaemia group A protein N-terminal
C2H2-type
Family 6
NPHP1 exon 1-20 del
Chronic kidney disease
Chronic kidney disease
MTLN
NPHP1
MALL
128,729 bps
MTLN exon 1,2 del
NPHP1 exon 1-20 del
MALL exon 1,2 del
NM_005434.5
NM_000272.5
NM_001384134.1
56 residues
SH3
732 residues
MARVEL
153 residues
Family 7
NPHP1 exon 1-20 del
Nephronophthisis
MTLN
NPHP1
MALL
128,729 bps
MTLN exon 1,2 del
NPHP1 exon 1-20 del
MALL exon 1,2 del
NM_005434.5
NM_000272.5
NM_001384134.1
56 residues
SH3
732 residues
MARVEL
153 residues
Family 8
PRKN exon 3, 4 del
Dystonia
PRKN
exon 3, 4 del
1,380,350 bps
NM_004562.3
465 residues
Ubiquitin-like
Necessary for PINK1-dependent localization to mitochondria
TRIAD supradomain
Family 9
SLC13A5 exon 11, 12 del
Seizures
Seizures
28,651 bps
SLC13A5
exon 11, 12 del
NM_177550.5
568 residues
Homozygous loss
Strand
Novel homozygous loss
-/-
Homozygous affected
Gene
Protein
NA
Not available
+/-
Heterozygous carrier

a

Prioritization

1,127 loss calls **without any morbid AR** genes

276 in unaffected individuals

851 calls in affected individuals

457 in diagnosed individuals

394 calls in undiagnosed individuals

143 possibly false and 155 uncertain calls

96 possibly true calls

68 calls overlapping calls in unaffected/phenotypically non-matching individuals

28 possibly true calls with candidate gene(s)

16 calls where gene(s) had homozygous pLoF variants in control databases

qPCR validation

12 calls with candidate recessive disease gene(s)

4 invalid calls, 1 sample unavailable

7 valid homozygous losses in 7 individuals from 5 families

5 candidate recessive disease genes

Additional patients search

Functional characterization

1 gene of unknown significance

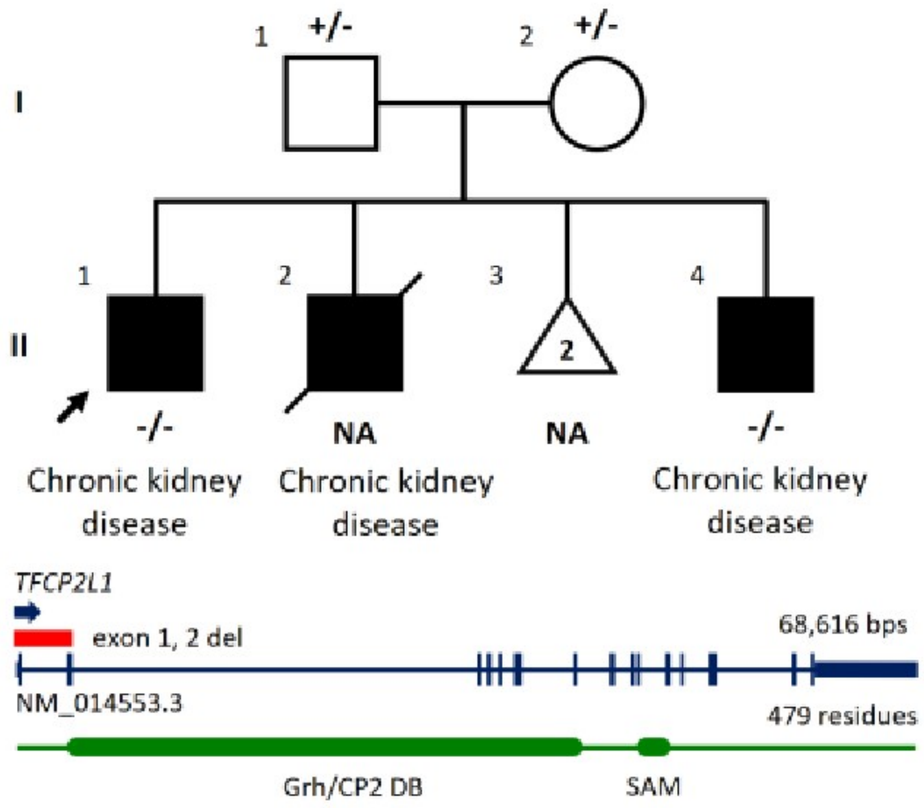
PWWP3A

4 previously undelineated disorders

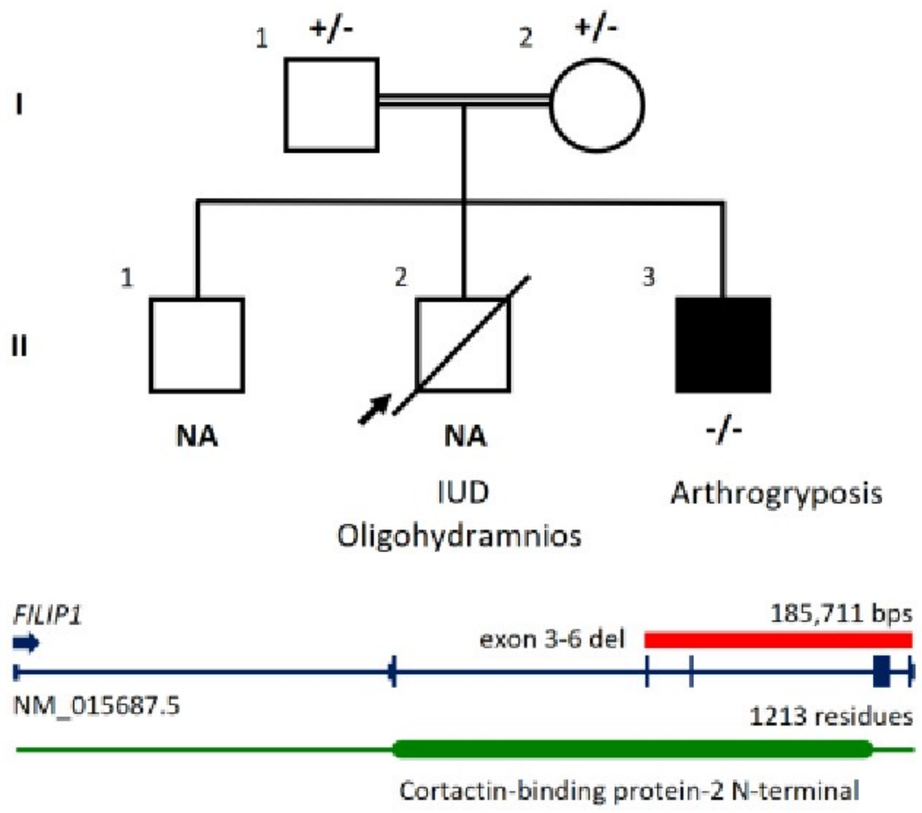
VPS36 ***TFCP2L1*** ***FILIP1*** ***FAM177A1***

b

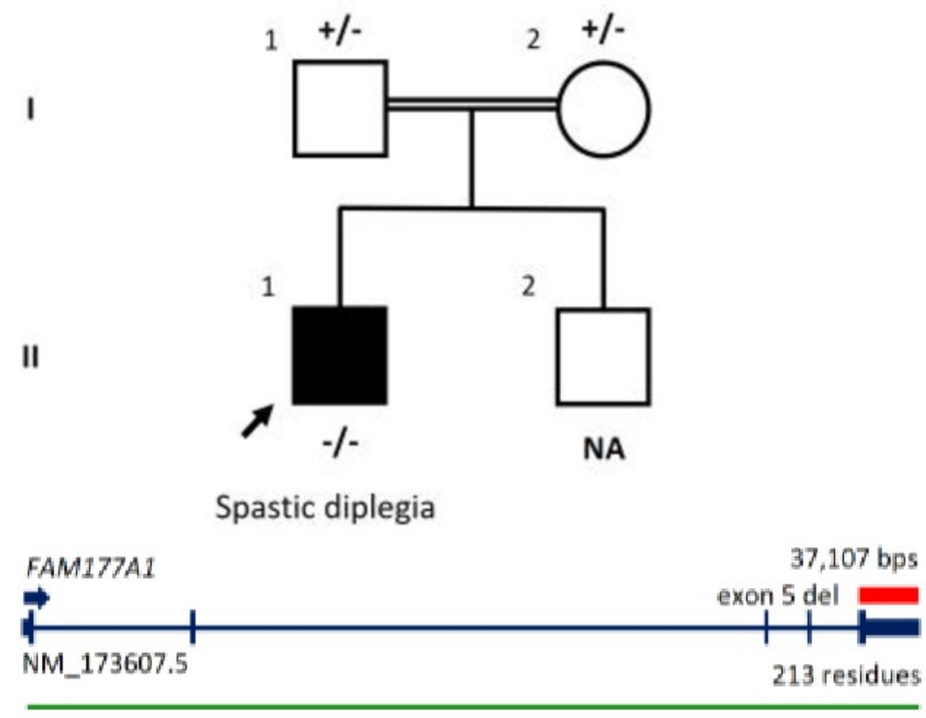
Family 22: *TFCP2L1* exon 1, 2 del



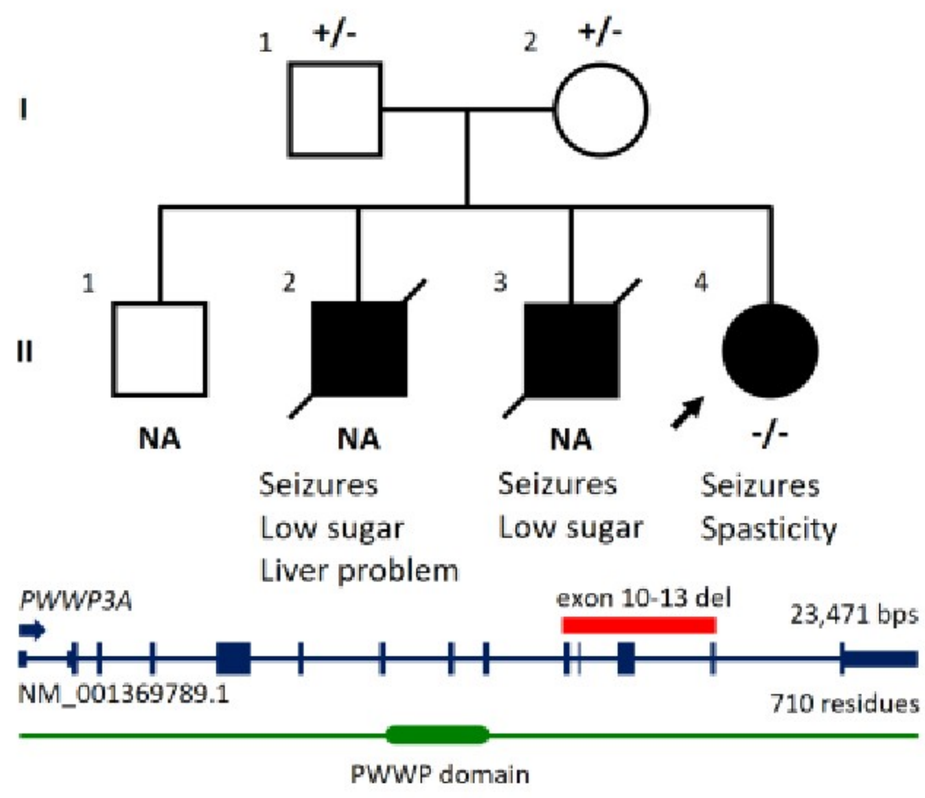
Family 23: *FILIP1* exon 3-6 del



Family 24: *FAM177A1* exon 5 del



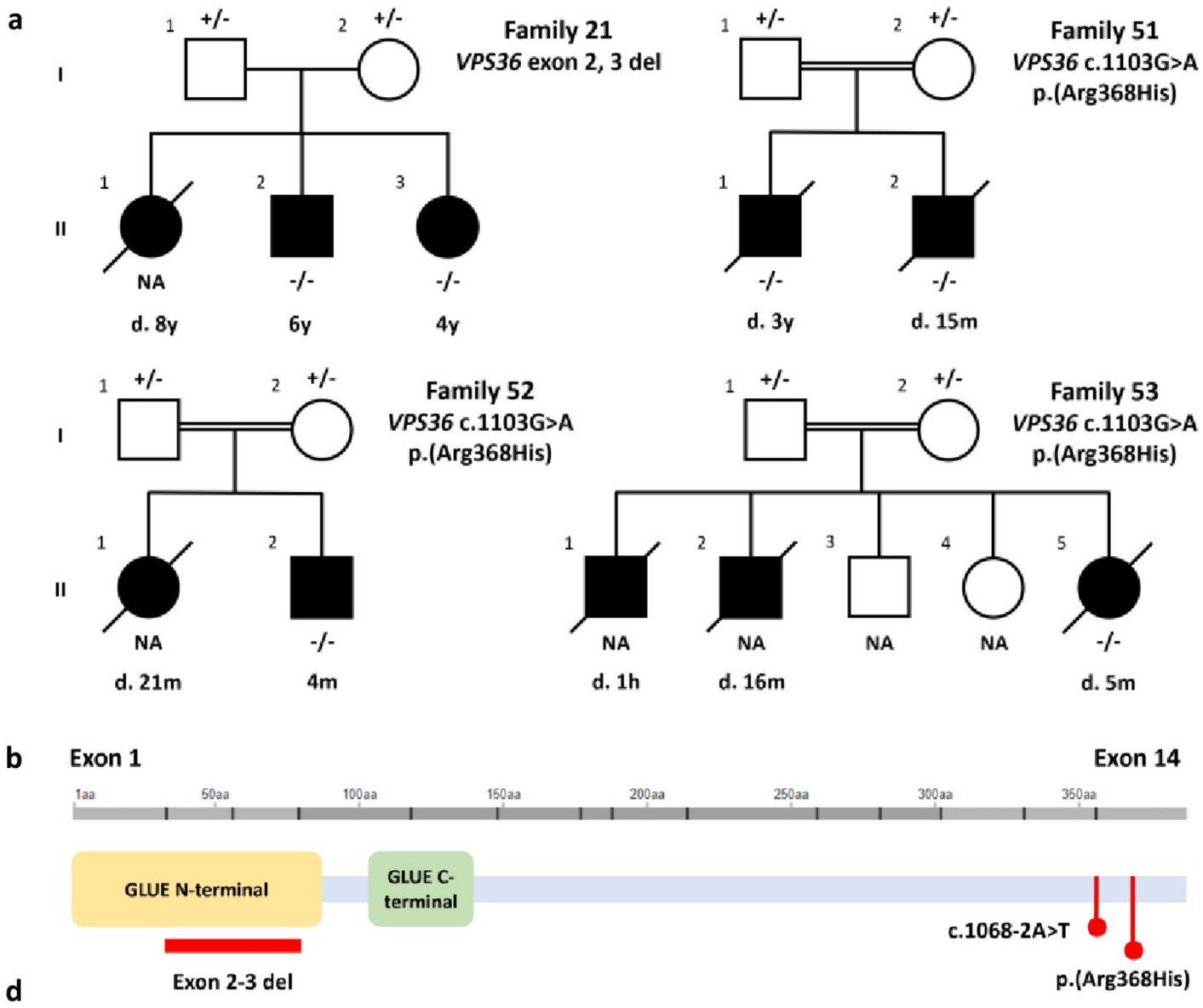
Family 25: *PWWP3A* exon 10-13 del



Homozygous loss -/- Homozygous affected

NA Not available +/- Heterozygous carrier

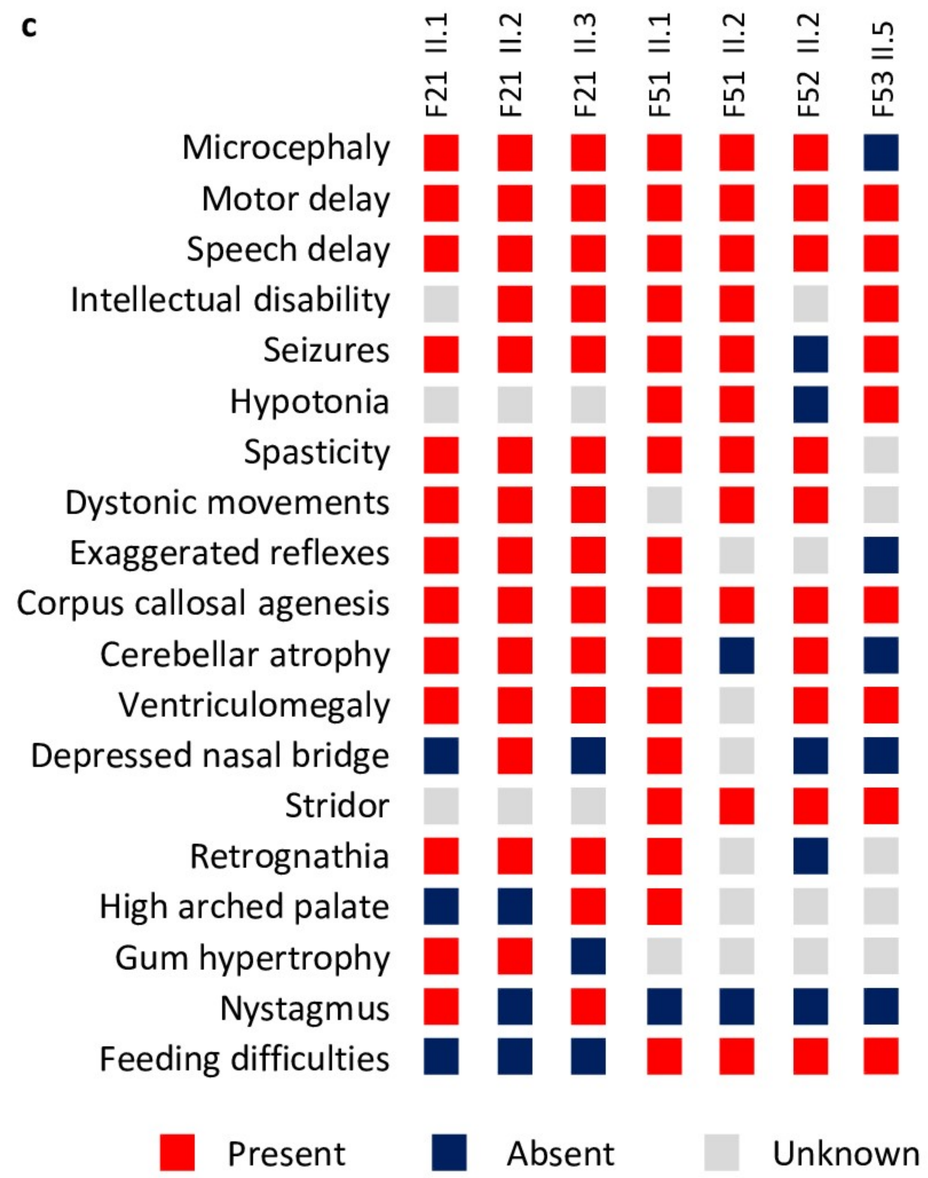
a


Family 21
VPS36 exon 2, 3 del
I
II
1 +/-
2 +/-
1 NA
2 -/-
3 -/-
d. 8y
6y
4y
Family 51
VPS36 c.1103G>A
p.(Arg368His)
1 -/-
2 -/-
d. 3y
d. 15m
Family 52
VPS36 c.1103G>A
p.(Arg368His)
1 NA
2 -/-
d. 21m
4m
Family 53
VPS36 c.1103G>A
p.(Arg368His)
1 NA
2 NA
3 NA
4 NA
5 -/-
d. 1h
d. 16m
d. 5m

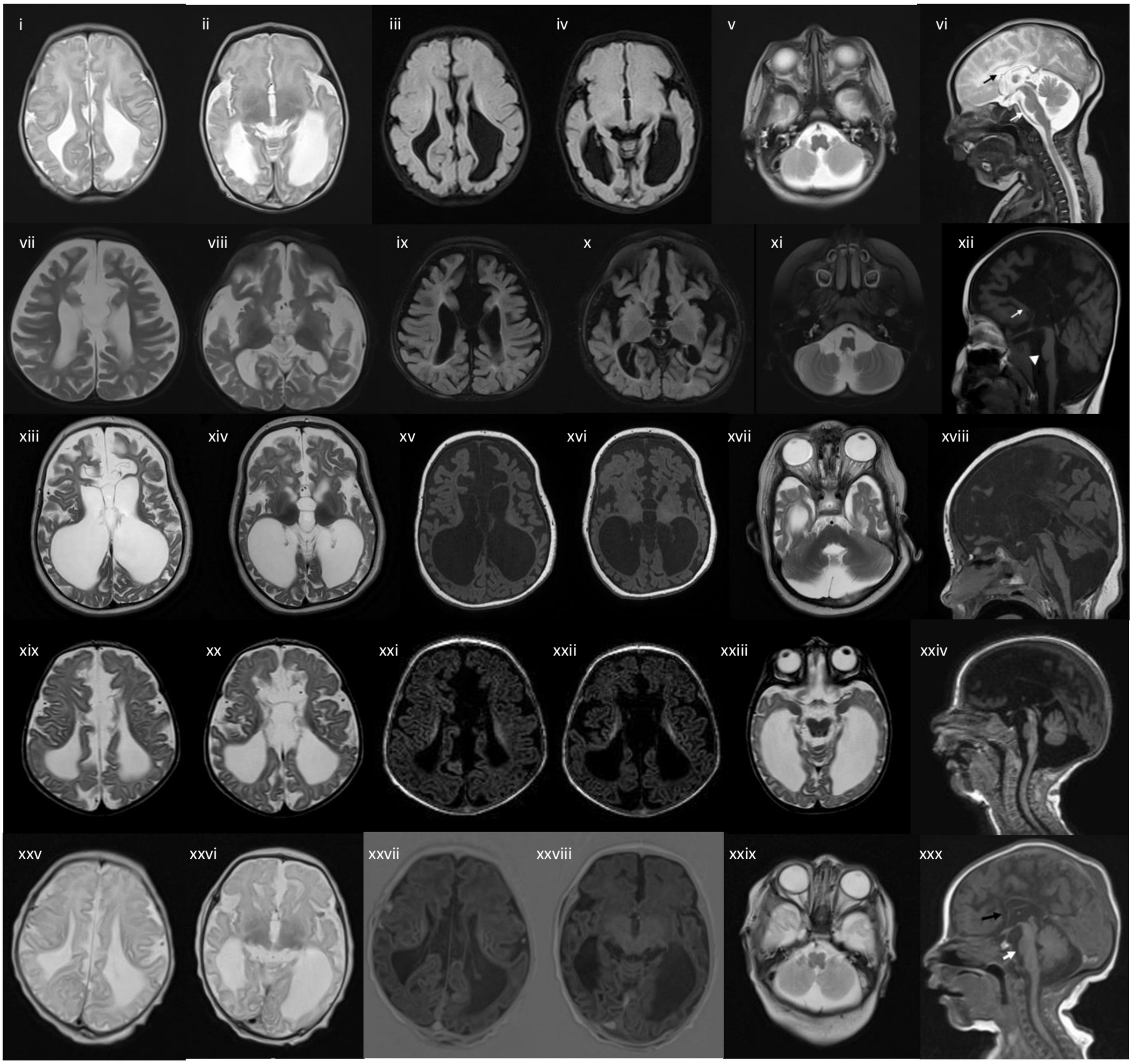
b

Exon 1
Exon 14
1aa
50aa
100aa
150aa
200aa
250aa
300aa
350aa
GLUE N-terminal
GLUE C-terminal
Exon 2-3 del
c.1068-2A>T
p.(Arg368His)

c


F21 II.1
F21 II.2
F21 II.3
F51 II.1
F51 II.2
F52 II.2
F53 II.5
Microcephaly
Motor delay
Speech delay
Intellectual disability
Seizures
Hypotonia
Spasticity
Dystonic movements
Exaggerated reflexes
Corpus callosal agenesis
Cerebellar atrophy
Ventriculomegaly
Depressed nasal bridge
Stridor
Retrognathia
High arched palate
Gum hypertrophy
Nystagmus
Feeding difficulties
Present
Absent
Unknown

d


F21 II.2
i
ii
iii
iv
v
vi
F21 II.3
vii
viii
ix
x
xi
xii
F51 II.1
xiii
xiv
xv
xvi
xvii
xviii
F51 II.2
xix
xx
xxi
xxii
xxiii
xxiv
F52 II.2
xxv
xxvi
xxvii
xxviii
xxix
xxx