Michael J. Field, Raman Kumar, Anna Hackett, Sayaka Kayumi, Cheryl A. Shoubridge, Lisa J. Ewans, Atma M. Ivancevic, Tracy Dudding, Byth, Renée Carroll, Thessa Kroes, Alison E. Gardner, Patricia Sullivan, Thuong T. Ha, Charles E. Schwartz, Mark J. Cowley, Marcel E. Dinger, Elizabeth E. Palmer, Louise Christie, Marie Shaw, Tony Roscioli, Jozef Gecz, Mark A. Corbett

Different types of disease-causing non-coding variants revealed by genomic and gene expression analyses in families with X-linked intellectual disability

Human Mutation, 2021; 42(7):835-847

Different types of disease-causing non-coding variants revealed by genomic and gene expression analyses in families with X-linked intellectual disability

Michael J. Field[1], Raman Kumar[2], Anna Hackett[1,3], Sayaka Kayumi[2], Cheryl A. Shoubridge[2], Lisa J. Ewans[4,5], Atma M. Ivancevic[6], Tracy Dudding-Byth[1,3], Renée Carroll[2], Thessa Kroes [2], Alison E. Gardner [2], Patricia Sullivan[7], Thuong T. Ha[8], Charles E. Schwartz[9], Mark J. Cowley[1,5,7], Marcel E. Dinger[10], Elizabeth E. Palmer[1,11], Louise Christie[1], Marie Shaw[2], Tony Roscioli[12,13], Jozef Gecz[2,14] and Mark A. Corbett[2]*

1. NSW Genetics of Learning Disability Service, Newcastle, NSW, Australia.
2. Adelaide Medical School and Robinson Research Institute, University of Adelaide, Adelaide, SA, Australia.
3. University of Newcastle, Newcastle, NSW, Australia.
4. St Vincent's Clinical School, University of New South Wales, Darlinghurst, Australia.
5. Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia.
6. University of Colorado, Boulder, CO, USA.
7. Children's Cancer Institute, University of New South Wales, Kensington, NSW, Australia.
8. Molecular Pathology Department, Centre for Cancer Biology, SA Pathology, Adelaide, SA, Australia.
9. Greenwood Genetics Centre, Greenwood, SC, USA.

10. School of Biotechnology and Biomolecular Sciences, University of New South Wales, Kensington, NSW, Australia.

11. School of Women's and Children's Health, University of New South Wales, Kensington, Sydney, NSW, Australia.

12. NeuRA, University of New South Wales, Sydney, NSW, Australia.

13. Centre for Clinical Genetics, Sydney Children's Hospital, Randwick, Sydney, NSW, Australia.

14. South Australian Health and Medical Research Institute, Adelaide, SA, Australia.

* For correspondence:

Mark Corbett, Ph.D.

Australian Collaborative Cerebral Palsy Research Group and Neurogenetics Research Program, Adelaide Medical School,

University of Adelaide, Adelaide,

South Australia, 5000, Australia.

Phone: +61 8 83137938

e-mail: mark.corbett@adelaide.edu.au

Key words: whole genome sequencing, transcriptome, splicing, non-coding, gene regulation, intellectual disability, X-linked, RNA-Seq

Abstract

The pioneering discovery research of X-linked intellectual disability (XLID) genes has benefitted thousands of individuals worldwide however, approximately 30% of XLID families still remain unresolved. We postulated that non-coding variants that affect gene regulation or splicing may account for the lack of a genetic diagnosis in some cases. Detecting pathogenic, gene-regulatory variants with the same sensitivity and specificity as structural and coding variants is a major challenge for Mendelian disorders. Here, we describe three pedigrees with suggestive XLID where distinctive phenotypes associated with known genes guided the identification of three different non-coding variants. We used comprehensive structural, single nucleotide and repeat expansion analyses of genome sequencing. RNA-Seq from patient-derived cell lines, RT-PCRs, western blots and reporter gene assays were used to confirm the functional effect of three fundamentally different classes of pathogenic non-coding variants: a retrotransposon insertion, a novel intronic splice donor and a canonical splice variant of an untranslated exon. In one family, we excluded a rare coding variant in *ARX,* a known XLID gene, in favour of a regulatory non-coding variant in *OFD1* that correlated with the clinical phenotype. Our results underscore the value of genomic research on unresolved XLID families to aid novel, pathogenic non-coding variant discovery.

Introduction

Massively parallel sequencing has led to an explosion in our knowledge of the genetics of monogenic disorders (Bamshad et al., 2019). Multiple, large clinical genomics studies report diagnostic rates between 40-60% (Liu et al., 2019; Wright et al., 2018). However, these genetic diagnoses are heavily biased towards the detection of *de novo* protein-coding or disrupting variants.

Genetic studies of families living with X-linked intellectual disability (XLID) have implicated over 140 genes with a diverse range of molecular functions (Neri et al., 2018). One of the earliest and most significant discoveries was the triplet repeat expansion in *FMR1* that causes fragile X syndrome (FRAXA; MIM# 309550). The expanded CGG repeat in the 5' untranslated region (UTR) of *FMR1* becomes hypermethylated, leading to silencing of transcription; a gene-regulatory disease mechanism (Chiurazzi et al., 1998; Oberlé et al., 1991). The high rate of XLID gene discovery has continued with 69 new genes reported between 2007 and 2017 (Neri et al., 2018). Many of these discoveries were achieved through systematic X-chromosome gene resequencing studies in large cohorts (Hu et al., 2016; P.S. Tarpey et al., 2009). Despite access to high-quality sequencing with near complete coverage of protein-coding regions, up to 30% of the large XLID pedigrees (traditionally coded with "MRX" or "MRXS" numbers) are yet to be explained (Neri et al., 2018).

Combining RNA-Seq with exome or genome sequencing (GS) data is a highly effective method for detecting gene regulatory variants (Cummings et al., 2017; Frésard et al., 2019; Kremer et al., 2017). Using this strategy on a broadly selected cohort of individuals, predominantly with rare neurodevelopmental disorders, a diagnostic rate of 7.5% - 10% was achieved (Frésard et al., 2019;

Kremer et al., 2017). A diagnostic rate as high as 35% was achieved using disease target tissue in a selected cohort of individuals with specific muscle disorders (Cummings et al., 2017).

In the case of XLID, we previously discovered causative non-coding variants in two large pedigrees that remained unresolved after X-chromosome exome sequencing (Huang et al., 2012; Kumar et al., 2016). In the first family, a variant in one of the YY1 transcriptional repressor binding motifs of the *HCFC1* promoter blocked YY1 binding and upregulated *HCFC1* expression (Huang et al., 2012). In the second family, a single base duplication in the 5' UTR of *DLG3* caused attenuation of mRNA translation (Kumar et al., 2016).

The lack of a genetic diagnosis in some XLID families, particularly those with a clinically recognisable phenotype, led us to explore the possibility of non-coding variation as the cause. Here, we report three different causative regulatory variants in three families. We show that GS and analysis of the effects of phenotype-driven candidate non-coding variants on transcription, even within non-neuronal tissue, has the power to deliver genetic diagnosis.

Methods

Ethics statement

Genetic studies were approved by the Women's and Children's Health Network human research ethics committee, Adelaide. Written informed consent was obtained for molecular genetic analysis, and written permission was obtained before the publication of clinical data from all participants or their legal guardians.

Family recruitment

Five families that were unresolved following research exome and in two cases genome sequencing were initially selected for non-coding (GS and RNA-Seq) analysis based on a high probability of being X-linked based on a multi-generational pedigree with inheritance through less severely affected or normal females. These included two large mapped but unresolved MRX pedigrees, two smaller pedigrees with a strong clinical suspicion of a specific X-linked phenotype without resolution on targeted and exome testing (Families 1 & 2) as well as a family of multiple affected males from a mother with different partners. A further six families with single generation male only, familial intellectual disability that were genetically unresolved by exome sequencing were re-analysed by whole genome sequencing as part of a cost utility study (Ewans et al., 2018). One of these, (Family 3) was included in this study.

Genomic analysis pipeline

All the families underwent GS of two or more distantly related affected males on the Illumina HiSeq X Ten platform at the Kinghorn Centre for Clinical Genomics, Sydney. Short read alignment to hg19 build of the human genome with the Burrows-Wheeler aligner (BWA MEM) (H. Li & Durbin, 2009), single nucleotide variant (SNV) and INDEL identification with the genome analysis toolkit haplotype caller (v3.7) (Van der Auwera et al., 2013) and annotation with ANNOVAR (Wang et al., 2010) was performed as previously described (Corbett et al., 2016).

Structural variant analysis (copy number variants [CNV], translocations, insertions and inversions) was performed using DELLY v0.7.8, Manta v-1.1.1 and Lumpy v-0.2.13 for detection of deletions, duplications, translocations, insertions and inversions (Chen et al., 2016; Layer et al., 2014; Rausch et al., 2012) with results being genotyped in combination with 150 in-house control genomes and the 1000 genomes CNV reference dataset. Novel sequence insertions were detected with the RetroSeq v1.5 pipeline using default parameters (Keane et al., 2013).

To identify short tandem repeat expansions we used ExpansionHunter (Dolzhenko et al., 2017) and exSTRa (Tankard et al., 2018). We created a custom target location JSON file or exSTRa database respectively, that included all recorded short tandem repeats with sequence unit lengths between 2 and 7 bp on the X chromosome extracted from the tandem repeat database (Gelfand et al., 2007). We used TRhist (Doi et al., 2014) to look for novel repeated sequences filling individual reads uniformly trimmed to 90 bp. Repeat reads and their pairs were extracted from the fastq file in samples with 20 or more reads that were enriched (Z-score > 2) with a specific repeat sequence compared to a population of 50 in-house control genomes of similar genetic background. These reads were assembled into contigs using the DNASTAR Lasergene v16 SeqMan Pro module with subsequent contigs matched to the NCBI non-redundant sequence database with BLAST (Altschul et al., 1990).

RNA-Seq

Total RNA was extracted from patient-derived lymphoblastoid cell lines (LCL) as described previously (Froyen et al., 2008). TruSeq stranded cDNA libraries were generated according to the manufacturer's protocols (Illumina). RNA sequencing was performed on the NovaSeq 6000 (Illumina) to yield a minimum of 7.7 x 10^7 100 bp paired reads per sample. Reads were mapped to GRCh38 build of the human genome using HISAT2 and read counts generated for known and novel transcripts using StringTie (Pertea et al., 2016). Outlier gene expression was tested from normalised read count data using the OUTRIDER package (Brechtmann et al., 2018). Significantly differentially spliced isoforms (FDR < 0.05) generated from known and novel splice junctions were detected and quantified with Leafcutter using default settings (Y. I. Li et al., 2018).

Detection of candidate disease-causing variants

Family 1 and Family 2: All variants were first filtered for those shared between the related individuals under an X-linked inheritance model. We removed SNV and INDELS that were frequent in population databases greater than the levels indicated in the following: gnomAD (v2.1.1) (Karczewski et al., 2020) or ExAC (v3) (Lek et al., 2016) to >0.0001, UK10K control data (Walter et al., 2015) or 1000 genomes project phase 3 (1000 Genomes Project Consortium, 2010) to >0.005. Structural variants on the X chromosome shared between affected family members were retained except those with greater than 80% overlap with CNV with minor allele frequencies > 0.01 in the DECIPHER (v9.25) common database.

Family 3: Variants were filtered using the web platform SEAVE (https://www.seave.bio/) that utilises GEMINI (Paila et al., 2013). SNVs and INDELs with a predicted impact severity of "high" or "medium" shared between both affected males were retained whose zygosity was consistent with X-linked, autosomal recessive (AR) or autosomal dominant (AD) inheritance. Population databases from the 1000 genomes project phase 3, ExAC or the exome variant server were utilised to remove variants with a minor allele frequency (MAF) of greater than 2% (X-linked/AR) or 0.1% (AD). Remaining gene variants underwent further prioritisation and manual interpretation.

Cloning of mutant full-length *ARX* constructs

Full-length human *ARX* cDNA construct in pCMV-Myc vector (pCMV-Myc-ARX WT) (C. Shoubridge et al., 2007) was used to generate pCMV-Myc-ARX c.1204G>A (p.Gly402Arg) using site-directed mutagenesis (QuikChange Multi Site-Directed Mutagenesis Kit, Agilent Technologies). The primer sequence is available upon request. The entire open reading frame was verified by Sanger sequencing to ensure no other mutation was introduced.

Luciferase reporter assays

HEK293T cells were maintained in Dulbecco's modified Eagle's medium supplemented with 10% (v/v) fetal bovine serum, 100 U ml^{-1} sodium penicillin and 100 μg ml^{-1} of streptomycin sulfate in 5% CO_2 at 37 °C. Cells were plated at 4x10^5 per well in 12 well plates without antibiotics and 24 hours later were transfected with 200 ng luciferase reporter plasmid DNA, 10 ng pGL4.74[hRluc/TK] plasmid DNA (Promega) and 500 ng of pCMV-Myc, pCMV-Myc-ARX-WT, pCMV-Myc-ARX-p.Gly402Arg, pCMV-Myc-ARX-p.Thr333Asn or pCMV-Myc-ARX-p.Pro353Leu plasmid DNAs using Lipofectamine 2000 (Invitrogen). Cells were lysed 24 hours post-transfection, and both Firefly and *Renilla* luciferase activity was quantified using Dual-Glo Luciferase Assay system (Promega) on the LUMIstar Optima (BMG Labtech), as previously described (Mattiske et al., 2018). In at least three independent transfections, each sample was measured in replicate, with triplicates of each replicate measured in the reporter assay. The Firefly luciferase activity was normalised to the corresponding *Renilla* luciferase activity, and each sample was reported relative to the pCMV-Myc empty vector.

cDNA, RT-PCR and qPCR protocols

Total RNA (1 μg) extracted from cell lines as previously described (Froyen et al., 2008), was reverse transcribed to cDNA using the iScript reverse transcription kit (Bio-Rad, Gladesville, NSW, Australia; cat# 1708891), according to the manufacturer's protocol. RT-PCR using primers and conditions were performed as indicated in Supp. Table S1.

Quantitative RT-PCR (qPCR) was performed using the relative standard curve method. PCR products were amplified with iTaq Universal Supermix (Bio-Rad; cat# 1725121) and primers as indicated in Supp. Table S1 in a StepOnePlus real-time PCR system (Applied Biosystems). Experiments were performed in duplicate with three technical replicates of each sample for each primer pair in each case. Product specificity was determined by melt-curve analysis at the end of each run.

Genomic PCR and Sanger Sequencing

Specific variants were validated and segregated through each family using dye terminator chemistry v3.1. Primer sequences and cycling conditions for all PCRs are recorded in Supp. Table S1.

Western blotting

Proteins from patient-derived or control cell lines were extracted with lysis buffer 50 mM Tris-HCl pH 7.5, 250 mM NaCl, 0.1% Triton X-100, 1 mM EDTA, 50 mM NaF and 0.1 mM Na_3VO_4 and 1x Protease inhibitor, no EDTA for OFD1 or 50 mM Tris-HCl pH 7.5, 50 mM KCl, 0.1% NP40, 5 mM EDTA, 50 mM

NaF, 0.1 mM Na_3VO_4 and 1x Protease inhibitor, no EDTA for AP1S2. Extracts were resolved by 7% denaturing polyacrylamide gel (SDS-PAGE) and transferred to nitrocellulose membrane by electroblotting. Primary antibodies for detection were rabbit polyclonal anti-AP1S2 antibody (Abcam cat# ab97590), rabbit anti-OFD1 (Sigma cat# SAB2702042) and rabbit anti-β-tubulin (Abcam cat# ab6046) antibodies. Secondary antibody was anti-rabbit IgG conjugated to horseradish peroxidase (HRP), (Dako cat# P0448). Enhanced chemiluminescent signal (Bio-Rad cat# 1705061) was visualised with the chemidoc detection system (Bio-Rad).

Clinical descriptions

Family 1

Family 1 had a putative X-linked ciliopathy in the three affected males examined (Fig. 1a). All had a mild cognitive delay in adulthood. Two males (IV-1 and IV-2) had progressive suppurative lung disease and retinal coloboma. IV-1 had severe early language delay, intermittent generalized tonic-clonic seizures from the age of 10 that were initially controlled with sodium valproate, but became drug resistant in mid adolescence and conductive hearing loss. There was evidence of cerebellar dysfunction on clinical examination with minor cerebellar vermis hypoplasia in IV-1 in infancy on MRI. All males had macrocephaly with head circumference in IV-1 and IV-2 in the 97th centile and II-6 in the 75th centile. The combination of suppurative lung disease, retinal and cerebellar changes made us consider a ciliopathy and a pathogenic variant in *OFD1* had been considered likely, but was not identified on an extensive ciliopathy panel including *OFD1*

(Vilboux et al., 2017). Assumed obligate female carriers had normal intellect and III-2 showed highly skewed X-inactivation (90:10).

Family 2

Family 2 had a clinical and biochemical diagnosis of Alan-Herndon-Dudley syndrome (AHDS; MIM# 300523). The family consisted of two affected males (the proband and his maternal second cousin) (Fig. 2a). The phenotype was severe early hypotonia with feeding difficulties, which evolved to a progressive spasticity resulting in contractures, scoliosis and a severe reduction in mobility. The affected individuals also had cognitive impairment, seizures and were non-verbal. Thyroid function studies were consistent with a diagnosis of AHDS. For III-4 at 53 years of age and IV-6 at 23 years of age, TSH and free T4 were in the normal range (NR) while free T3 was elevated, III-4 was 7.5 pmolL^{-1} (NR 3.3 – 6.2 pmolL^{-1}) and IV-6 was 8.4 pmolL^{-1} (NR 3.1 – 7.6 pmolL^{-1}) . However, Sanger sequencing of the exons of *SLC16A2* (a.k.a. *MCT8*) did not identify a pathogenic variant. Obligate female carriers were of normal intellect and there was no evidence of abnormal skewing of X-chromosome inactivation in III-2 (74:26).

Family 3

Family 3 consisted of two brothers with mild-severe ID, autistic spectrum disorder, microcephaly, hypotonia, abnormal gait and hyperextensible joints (Fig. 3a). Neither male has had seizures. One was non-verbal as an adolescent and the other had functional speech and basic literacy skills. Dysmorphic facial features included a depressed nasal bridge, peg teeth and prominent jaw. A brain MRI performed on II-3 at age 10, showed abnormal signal in the globus pallidus and caudate, compatible with intracerebral calcification (Supp. Fig. S1). Both parents were unaffected. Due to the relatively non-specific phenotypic features, no specific diagnosis was suspected clinically. Female carriers in the family have been of normal intellect. X-chromosome inactivation testing was uninformative in I-2.

Results

We performed GS of two affected males from each of the families in this study. Mapping to the hg19 build of the human genome achieved a median read depth of 38x in all samples with no significant mapping bias between coding and non-coding regions of the genome (Supp. Fig. S2). Initial filtering of variants in these three families failed to identify plausible disease-causing coding missense, truncating variants or copy number variants previously identified as pathogenic in ClinVar or DECIPHER databases.

Family 1: Ciliopathy caused by a deep intronic variant in OFD1

Given the apparent X-linked pattern of inheritance and the distinct ciliopathy, we made a targeted investigation of all coding and non-coding variants in *OFD1*, a known X-linked ciliopathy gene we had experience with (Field et al., 2012). We analysed GS data within the boundaries of the first and last exons of the OFD1 gene and a region 2kb upstream of the transcriptional start site. A novel variant of uncertain significance on chrX, NC_000023.10:g.13775457G>A (hg19), (ClinVar: VCV000929433.1) was identified within intron 13 of *OFD1* (NM_003611.2:c.1412-322G>A) that segregated with the affected males and obligate carriers in the family (Fig. 1a and Supp. Fig. S3). Comparing outlier transcripts from RNA-Seq from patient derived LCLs revealed a novel splicing event involving a cryptic splice donor site

3 bp upstream of the NC_000023.10:g.13775457G>A variant and two cryptic splice acceptors at positions chrX:g.13775250 and chrX:g.13775347 to create two novel *OFD1* transcripts (Fig. 1b). The SpliceAI program predicted the novel splice donor site and the most proximal splice acceptor site (chrX:g.13775347, 110 bp upstream of the variant site), with delta scores of 0.64 and 0.61 respectively. SpliceAI delta scores range between 0 and 1 and are an approximate measure of the probability that the variant will alter splicing (Jaganathan et al., 2019). Both novel transcripts were predicted to create truncated protein products NP_003602.1:p.(Leu472ProfsTer26) and NP_003602.1:p.(Leu472PhefsTer37) due to frameshifts (Fig. 1b and Supp. Data). Western blotting showed reduced OFD1 protein abundance in an available LCL from IV-2 compared to LCLs from unaffected males (Fig. 1c). The epitope for the antibody targets a region in the protein prior to p.Arg471, however bands corresponding to the predicted novel truncated protein products were not detected. Reduced OFD1 protein expression combined with the ciliopathy segregating in an X-linked pattern in this family strongly suggested this novel deep intronic splice variant was pathogenic.

Functional assessment of a predicted damaging coding ARX variant

We also identified a unique coding variant in *ARX* NM_139058.3:c.1204G>A:p.Gly402Arg (ClinVar: VCV000929432.1) that segregated with the affected individuals in Family 1 (Fig. 1a and Supp. Fig. S3) and was absent in all public variant databases used for filtering. The variant had a phred scaled CADD score of 27, and was not covered in previous exome sequencing (Supp. Fig. S4). The variant was located C-terminally and outside of the homeodomain, in a region with paucity of known *ARX* pathogenic variants (Cheryl Shoubridge et al., 2010). The phenotype in the patients with infratentorial changes without corpus callosum or cortical abnormalities was not typical for an *ARX* point mutation as was the absence of severe epilepsy, dystonia or genital abnormalities. The relative proximity to the homeodomain (ending at p.387) and high CADD score prompted us to investigate this variant using an ARX-responsive luciferase reporter assay (Mattiske et al., 2018). The p.Gly402Arg variant displayed levels of repression (45%) similar to ARX-WT when compared to the pCMV-Myc vector control, indicating the variant did not change the transcriptional activity of the ARX protein, in the context of this *in vitro* assay (Fig. 1d). Known pathogenic variants of *ARX* were also tested either within the nuclear localisation sequence (NLS) or homeodomain itself, and all abolished repression of *luciferase* expression. Furthermore, we did not observe any disruptions to subcellular localisation of the p.Gly402Arg variant compared to over-expression of the wild-type protein HEK293T cells (data not shown). These results suggested the ARX p.Gly402Arg variant was benign.

Family 2: Novel intronic mobile element insertion in SLC16A2

GS on two affected males from family 2 (III-4 and IV-6; Fig. 2a) revealed no shared, rare coding variants on the X chromosome. Given the clinical diagnosis of AHDS in this family, we targeted variants called in coding and non-coding regions of *SLC16A2* for further analysis. A SINE-VNTR-Alu (SVA_E) retrotransposon insertion, called by RetroSeq, was found in the fifth intron of *SLC16A2* (ClinVar: VCV000929441.1) in both affected males (Fig. 2b and Supp. Fig. S5) but was not observed in our in-house control GS data of 207 individuals.

We measured the expression of *SLC16A2* by qPCR using primers specific for cDNA of exons 5 and 6 and showed almost complete loss of gene expression relative to *GAPDH* in a fibroblast cell line from individual IV-6 compared to control fibroblasts (Fig. 2c). Further investigation by RT-PCR revealed all exon boundaries of *SLC16A2* that we tested except those involving exon 6 were correctly spliced (Fig. 2d and Supp. Fig. S6). Qualitative examination of RNA-Seq data from IV-6 showed the creation of at least one novel splice donor site in intron 5, just prior to the site of the SVA_E insertion and subsequent retention of the remainder of intron 5 (Supp. Fig. S7 and Supp. Data). The loss of the final exon in *SLC16A2* was predicted to be sufficient to account for the metabolic findings in this family and suggested that this retrotransposon insertion was pathogenic.

Family 3: Canonical splice site variant in a non-coding exon of AP1S2

The first pass analysis of coding and splicing variants in GS data from individuals II-1 and II-3 of Family 3 (Fig 3a) identified a shared variant NC_000023.10:g.15872810C>T (NM_003916.3:c.-1+1G>A) in *AP1S2* (ClinVar: VCV000929434.1), (Fig. 3b) that was predicted to affect splicing with a SpliceAI donor loss delta score = 0.98. RNA-Seq analysis showed retention of intron 1 and a significant down regulation of *AP1S2* expression (Fig. 3c). A comprehensive analysis of the exon boundaries of *AP1S2* by RT-PCR using cDNA from LCL of II-1, II-3 and an unrelated control showed splicing between the untranslated exon one and translated exon two was completely abolished, while transcripts containing exons three, four and five were spliced normally (Fig. 3d & e and Supp. Fig. S8a-d). We also detected aberrant *AP1S2* transcripts using primers specific for intron one and exon five and confirmed that intron one was retained in these transcripts by Sanger sequencing (Fig. 3f). Western blotting of whole-cell protein extracts of LCL from II-1 and II-3 showed absence of AP1S2 protein compared to control LCLs (Fig. 3g).

Discussion

We have shown that utilisation of GS and gene expression analysis in families unresolved by exome analyses can detect functionally significant non-coding variations that explain a specific phenotype. The range of non-coding variants we have detected in large X linked families to date, includes a transcription factor binding site (Huang et al., 2012), a 5'UTR insertion that impedes translation (Kumar et al., 2016), and now, genesis of a deep intronic splice donor site, a retrotransposon insertion with mRNA processing effect and destruction of the canonical splice donor site of a non-coding exon. Each of these variants required a combination of approaches for detection and subsequent variant-focused molecular assays to confirm their pathogenicity.

The intronic variant in *OFD1* created a novel splice donor site and activated novel usage of two cryptic splice acceptor sites within the same intron. Traditional splicing prediction tools failed to predict this outcome, however, the recently developed SpliceAI program (Jaganathan et al., 2019) was able to predict this event for one of the two upstream cryptic splice acceptor sites which were validated by our RNA-Seq data. Machine learning approaches like that taken by SpliceAI show promising results for discovery of pathogenic non-coding variants.

Classically, pathological variants in *OFD1* were associated with a female limited phenotype with polydactyly and midline clefting with male lethality. Hypomorphic or loss of function variants in the terminal exon of *OFD1* have been associated with a variable range of phenotypes from a Simpson-Golabi-Behmel like disorder, with chronic suppurative lung disease (SGBS2; MIM# 300209) (Budny et al., 2006), to X-linked Joubert syndrome (JBTS10; MIM# 300804)(Coene et al., 2009). Joubert syndrome is defined by a specific radiological sign (molar tooth sign) that was not seen in Family 1. X-linked retinal dystrophy has been described due to a deep intronic variant in *OFD1*, NM_003611.2:c.935+706A>G (ClinVar: VCV000101499.5). This variant caused abnormal splicing, thus introducing a novel exon with a predicted frameshift and reduced *OFD1* expression (Webb et al., 2012). The respiratory, retinal, cerebellar and cognitive features seen in the three affected males in our family fit well with the broader phenotype associated with *OFD1* variants in males (Supp. Table S2) (Sakakibara et al., 2019). This distinctive phenotype and the data from the luciferase assays was critical in confirming the *ARX* p.Gly402Arg variant was likely benign.

Reports of retrotransposon insertions causing Mendelian disease are extremely rare (Hancks & Kazazian, 2016). A *de novo* L1 insertion into intron 3 of *RPS6KA3* which caused skipping of exon 4 in a male diagnosed with Coffin-Lowry syndrome and the SVA insertion into the 3'UTR of *FKTN* that causes Fukuyama congenital muscular dystrophy are to our knowledge, the only previous reports of such an event linked to ID (Kobayashi et al., 1998; Martínez-Garay et al., 2003; Taniguchi-Ikeda et al., 2011). Sine-VNTR-Alu (SVA) retro-transposed elements are one of the youngest and most mobile elements in the genome. Transposon insertion is not random, but relies on the presence of specific target sequences, and therefore sites prone to rearrangement can be predicted to some degree. In singular cases of Fukuyama muscular dystrophy and Bruton agammaglobulinaemia, the disease-causing mechanism involved novel exonisation of the inserted SVA sequences within the respective target genes (Conley et al., 2005; Taniguchi-Ikeda et al., 2011). A polymorphic SVA insertion that is implicated in X-linked dystonia Parkinsonism drove retention of intron 32 of *TAF1* transcripts, which was more pronounced in patient-derived neuronal stem cells than fibroblasts from the same individual (Aneichyk et al., 2018). In Family 2 of this study, the SVA_E was inserted in the same sense as *SLC16A2* and based on our RT-PCR and RNA-Seq data, potentially leads to novel exonisation of the 3' end of the *SLC16A2* transcript and a predicted protein that lacks the most C-terminal of the 12 transmembrane domains.

AHDS is caused by pathogenic variants in the thyroid hormone (TH) transporter *SLC16A2* and is characterised by severe ID and altered TH serum levels. Other features include early hypotonia, which evolves to spastic paraplegia within the first few years of life, low muscle mass with generalised weakness, speech difficulties that range from dysarthria to completely absent speech, variable ataxia and occasional dystonia and/or athetoid movements as well as seizures (Remerand et al., 2019). Penetrance is complete, although the severity is variable. Both affected males from Family 2 had the typical clinical features of AHDS, and their biochemical profile of high serum T3, low-normal T4, low rT3 and normal-elevated TSH levels was consistent with the disorder.

Approximately 6.5% of pathogenic variants recorded in the Human Gene Mutation Database (HMGD) are splice variants (Stenson et al., 2017). Clinically relevant variants at both canonical and especially at non-canonical positions are under ascertained in clinical exome sequencing studies to date (Lord et al., 2019). Characterisation of the effects of splicing variants are most efficiently performed with RNA-Seq, however in low throughput situations, targeted analysis of the effects of specific variants by RT-PCR is a viable alternative. The first exon of *AP1S2* is not translated thus making interpretation of the functional consequence of the variant we detected that affects the splice donor site difficult by computational predictions alone. Examination of the GENCODEv35 build of gene annotations identified 4,646 transcripts within 1,318 genes that have a start codon within 5bp of a splice acceptor site. There were 190 Pathogenic or Likely Pathogenic variants in the ClinVar database within the introns upstream of these start codons with a Kozak sequence interrupted by an intron. The clinical features displayed by the brothers in Family 3 were in hindsight consistent with those described in other affected individuals with causative *AP1S2* variants (Huo et al., 2019; Patrick S. Tarpey et al., 2006). Individuals with *AP1S2* variants display highly variable degrees of ID, even between affected males in the same family. The history is often characterised by early hypotonia and significant speech delay. Aggressive symptoms are reported. In some individuals, there is borderline microcephaly, which was also seen in both affected males. Brain MRI results II-3 from Family 3 were consistent with studies in individuals with *AP1S2* variants which showed basal ganglia calcification. The clinical presentation of the affected males in Family 3, in combination with our functional characterization of the splicing defects in *AP1S2* were essential in reaching a diagnosis.

The variants we found in *OFD1* and *SLC16A2* were both hypomorphic and some normal splicing occurred. The residual level of normal transcript expression and mild reduction in protein abundance may explain why the individuals in Family 1 fit the milder end of the *OFD1* disease spectrum. Most cases of ID are in the mild rather than moderate to severe spectrum, and this is an area where it has been less tractable so far to reach a genetic diagnosis. A proportion of these cases may be due to as yet unrecognised non-coding variants reducing the expression of known ID genes. An excellent example of this is the association of X-linked dystonia with an anti-sense inserted SVA_E transposon (Bragg et al., 2017) as opposed to a severe neurocognitive disability caused by coding variants in *TAF1* (O'Rawe et al., 2015, p. 1). Similarly, the mild neurocognitive features associated with the YY1 binding site variant regulating *HCFC1* we previously described (Huang et al., 2012), compared to the cobalamin deficiency and severe phenotype associated with loss of function variants within *HCFC1* (Yu et al., 2013). Each individual class of regulatory variant may only contribute modestly to the diagnostic rate in a cohort. For example, *de novo* variants in ultra-conserved, brain-active regulatory elements were estimated to be causative in 1-3% of cases (Short et al., 2018). Taken together, however, non-coding variants may account for as much as 50% of unresolved cases depending on the cohort selection (Burdick et al., 2020; Cummings et al., 2017).

We have shown three examples where combined analysis of clinical, genetic and molecular data was used to reach a genetic diagnosis involving a non-coding variant. RNA-Seq or hypothesis-driven RT-PCR analyses were necessary to reveal the effects of the candidate variants on transcription. Western blotting where a suitable antibody was available to show the effect on protein abundance in patient cell lines was highly informative in determining variant pathogenicity. There are important lessons to be learned from our study that will help to improve diagnostic yield in the currently 50-60% of individuals with a strongly suspected monogenic disorder who remain unresolved on current diagnostic testing (Hartley et al., 2020). Firstly, we have demonstrated that it is possible to make use of non-neuronal, patient-derived cell lines to genetically resolve non-coding variants of uncertain significance in patients with a primary neurodevelopmental disorder. Secondly, relatively simple molecular techniques that are tractable for molecular genetics laboratories can be powerful tools for functionally validating the effects of such variants and consequently, confirm their pathogenicity. Finally, we show that using multiple strategies for analysis of genome sequencing data including coding, non-coding, structural variation and repeat expansion detection is advisable in light of the heterogeneity of non-coding variants we have observed in this study and our previous investigations (Huang et al., 2012; Kumar et al., 2016). A comprehensive analysis of GS data, phenotype-driven, candidate-gene identification combined with gene expression analysis can successfully locate the most elusive causative non-coding variants and enable a confident genetic diagnosis.

Acknowledgements

We wish to thank the families involved in this project. This work was funded by grants from the Women's and Children's Hospital Foundation (MAC, JG, RK), the Tenix foundation (MAC, JG) and MJC was supported by NSW Health and a Cancer

Australia Project Grant. We wish to thank Velimir Gayevskiy for the use of SEAVE and Bree Hodgson for technical assistance.

Web Resources

ClinVar: https://www.ncbi.nlm.nih.gov/clinvar/

GENCODE: https://www.gencodegenes.org/

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

SEAVE: https://www.seave.bio/

Conflict of Interest Statement

The authors declare they have no conflicts of interest relevant to this work.

Data Availability Statement

Data not provided with this manuscript are available from the authors on reasonable request subject to the limitations of initial patient consent and approval by human research ethics committee. Links to variants mentioned in this manuscript are as follows:

OFD1: https://www.ncbi.nlm.nih.gov/clinvar/variation/929433/

ARX: https://www.ncbi.nlm.nih.gov/clinvar/variation/929432/

SLC16A2: https://www.ncbi.nlm.nih.gov/clinvar/variation/929441/

AP1S2: https://www.ncbi.nlm.nih.gov/clinvar/variation/929434/

Authors' Contributions

MAC, RK, JG and MF designed the study; RK, CSh, SK, RC, TH, MS and MAC designed and performed different aspects of the molecular and cell biology experiments; AH, ST, T.D-B, LC, EP, CES, TR and MF were clinicians involved with the families; SK, LE, AI, TH, MED, MF, MJC and MAC performed the bioinformatics and genomic analyses; MAC, AH and MF wrote the manuscript; all authors critically discussed results, revised and approved the manuscript.

References

1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/s0022-2836(05)80360-2

Aneichyk, T., Hendriks, W. T., Yadav, R., Shin, D., Gao, D., Vaine, C. A., Collins, R. L., Domingo, A., Currall, B., Stortchevoi, A., Multhaupt-Buell, T., Penney, E. B., Cruz, L., Dhakal, J., Brand, H., Hanscom, C., Antolik, C., Dy, M., Ragavendran, A., … Talkowski, M. E. (2018). Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell*, *172*(5), 897-909.e21. https://doi.org/10.1016/j.cell.2018.02.011

Bamshad, M. J., Nickerson, D. A., & Chong, J. X. (2019). Mendelian Gene Discovery: Fast and Furious with No End in Sight. *American Journal of Human Genetics*, *105*(3), 448–455. https://doi.org/10.1016/j.ajhg.2019.07.011

Bragg, D. C., Mangkalaphiban, K., Vaine, C. A., Kulkarni, N. J., Shin, D., Yadav, R., Dhakal, J., Ton, M.-L., Cheng, A., Russo, C. T., Ang, M., Acuña, P., Go, C., Franceour, T. N., Multhaupt-Buell, T., Ito, N., Müller, U., Hendriks, W. T., Breakefield, X. O., … Ozelius, L. J. (2017). Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(51), E11020–E11028. https://doi.org/10.1073/pnas.1712526114

Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V. A., Avsec, Ž., Herzog, M., Bader, D. M., Prokisch, H., & Gagneur, J. (2018). OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *American Journal of Human Genetics*, *103*(6), 907–917. https://doi.org/10.1016/j.ajhg.2018.10.025

Budny, B., Chen, W., Omran, H., Fliegauf, M., Tzschach, A., Wisniewska, M., Jensen, L. R., Raynaud, M., Shoichet, S. A., Badura, M., Lenzner, S., Latos-Bielenska, A., & Ropers, H.-H. (2006). A novel X-linked recessive mental retardation syndrome comprising macrocephaly and ciliary dysfunction is allelic to oral-facial-digital type I syndrome. *Human Genetics*, *120*(2), 171–178. https://doi.org/10.1007/s00439-006-0210-5

Burdick, K. J., Cogan, J. D., Rives, L. C., Robertson, A. K., Koziura, M. E., Brokamp, E., Duncan, L., Hannig, V., Pfotenhauer, J., Vanzo, R., Paul, M. S., Bican, A., Morgan, T., Duis, J., Newman, J. H., Hamid, R., Phillips, J. A., & Undiagnosed Diseases Network. (2020). Limitations of exome sequencing in detecting rare and undiagnosed diseases. *American Journal of Medical Genetics. Part A*, *182*(6), 1400–1406. https://doi.org/10.1002/ajmg.a.61558

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, *32*(8), 1220–1222. https://doi.org/10.1093/bioinformatics/btv710

Chiurazzi, P., Pomponi, M. G., Willemsen, R., Oostra, B. A., & Neri, G. (1998). In Vitro Reactivation of the FMR1 Gene Involved in Fragile X Syndrome. *Human Molecular Genetics*, *7*(1), 109–113. https://doi.org/10.1093/hmg/7.1.109

Coene, K. L. M., Roepman, R., Doherty, D., Afroze, B., Kroes, H. Y., Letteboer, S. J. F., Ngu, L. H., Budny, B., van Wijk, E., Gorden, N. T., Azhimi, M., Thauvin-Robinet, C., Veltman, J. A., Boink, M., Kleefstra, T., Cremers, F. P. M., van Bokhoven, H., & de Brouwer, A. P. M. (2009). OFD1 is mutated in X-linked Joubert syndrome and interacts with LCA5-encoded lebercilin. *American Journal of Human Genetics*, *85*(4), 465–481. https://doi.org/10.1016/j.ajhg.2009.09.002

Conley, M. E., Partain, J. D., Norland, S. M., Shurtleff, S. A., & Kazazian, H. H. (2005). Two independent retrotransposon insertions at the same site within the coding region of BTK. *Human Mutation*, *25*(3), 324–325. https://doi.org/10.1002/humu.9321

Corbett, M. A., Bellows, S. T., Li, M., Carroll, R., Micallef, S., Carvill, G. L., Myers, C. T., Howell, K. B., Maljevic, S., Lerche, H., Gazina, E. V., Mefford, H. C., Bahlo, M., Berkovic, S. F., Petrou, S., Scheffer, I. E., & Gecz, J. (2016). Dominant KCNA2 mutation causes episodic ataxia and pharmacoresponsive epilepsy. *Neurology*, *87*(19), 1975–1984. https://doi.org/10.1212/WNL.0000000000003309

Cummings, B. B., Marshall, J. L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A. R., Bolduc, V., Waddell, L. B., Sandaradura, S. A., O'Grady, G. L., Estrella, E., Reddy, H. M., Zhao, F., Weisburd, B., Karczewski, K. J., O'Donnell-Luria, A. H., Birnbaum, D., Sarkozy, A., Hu, Y., … MacArthur, D. G. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine*, *9*(386), eaal5209. https://doi.org/10.1126/scitranslmed.aal5209

Doi, K., Monjo, T., Hoang, P. H., Yoshimura, J., Yurino, H., Mitsui, J., Ishiura, H., Takahashi, Y., Ichikawa, Y., Goto, J., Tsuji, S., & Morishita, S. (2014). Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. *Bioinformatics*, *30*(6), 815–822. https://doi.org/10.1093/bioinformatics/btt647

Dolzhenko, E., van Vugt, J. J. F. A., Shaw, R. J., Bekritsky, M. A., van Blitterswijk, M., Narzisi, G., Ajay, S. S., Rajan, V., Lajoie, B. R., Johnson, N. H., Kingsbury, Z., Humphray, S. J., Schellevis, R. D., Brands, W. J., Baker, M., Rademakers, R., Kooyman, M., Tazelaar, G. H. P., van Es, M. A., … Eberle, M. A. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Research*, *27*(11), 1895–1903. https://doi.org/10.1101/gr.225672.117

Ewans, L. J., Schofield, D., Shrestha, R., Zhu, Y., Gayevskiy, V., Ying, K., Walsh, C., Lee, E., Kirk, E. P., Colley, A., Ellaway, C., Turner, A., Mowat, D., Worgan, L., Freckmann, M.-L., Lipke, M., Sachdev, R., Miller, D., Field, M., … Roscioli, T. (2018). Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. *Genetics in Medicine*, *20*(12), 1564–1574. https://doi.org/10.1038/gim.2018.39

Field, M., Scheffer, I. E., Gill, D., Wilson, M., Christie, L., Shaw, M., Gardner, A., Glubb, G., Hobson, L., Corbett, M., Friend, K., Willis-Owen, S., & Gecz, J. (2012). Expanding the molecular basis and phenotypic spectrum of X-linked Joubert syndrome associated with OFD1 mutations. *European Journal of Human Genetics*, *20*(7), 806–809. https://doi.org/10.1038/ejhg.2012.9

Frésard, L., Smail, C., Ferraro, N. M., Teran, N. A., Li, X., Smith, K. S., Bonner, D., Kernohan, K. D., Marwaha, S., Zappala, Z., Balliu, B., Davis, J. R., Liu, B., Prybol, C. J., Kohler, J. N., Zastrow, D. B., Reuter, C. M., Fisk, D. G., Grove, M. E., … Montgomery, S. B. (2019). Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nature Medicine*, *25*(6), 911–919. https://doi.org/10.1038/s41591-019-0457-8

Froyen, G., Corbett, M., Vandewalle, J., Jarvela, I., Lawrence, O., Meldrum, C., Bauters, M., Govaerts, K., Vandeleur, L., Van Esch, H., Chelly, J., Sanlaville, D., van Bokhoven, H., Ropers, H. H., Laumonnier, F., Ranieri, E., Schwartz, C. E., Abidi, F., Tarpey, P. S., … Gecz, J. (2008). Submicroscopic duplications of the hydroxysteroid dehydrogenase HSD17B10 and the E3 ubiquitin ligase HUWE1 are associated with mental retardation. *American Journal of Human Genetics*, *82*(2), 432–443. https://doi.org/10.1016/j.ajhg.2007.11.002

Gelfand, Y., Rodriguez, A., & Benson, G. (2007). TRDB—The Tandem Repeats Database. *Nucleic Acids Research*, *35*(Database issue), D80–D87. https://doi.org/10.1093/nar/gkl1013

Hancks, D. C., & Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA*, *7*, 9. https://doi.org/10.1186/s13100-016-0065-9

Hartley, T., Lemire, G., Kernohan, K. D., Howley, H. E., Adams, D. R., & Boycott, K. M. (2020). New Diagnostic Approaches for Undiagnosed Rare Genetic Diseases. *Annual Review of Genomics and Human Genetics*, *21*(1), 351–372. https://doi.org/10.1146/annurev-genom-083118-015345

Hu, H., Haas, S. A., Chelly, J., Van Esch, H., Raynaud, M., de Brouwer, A. P. M., Weinert, S., Froyen, G., Frints, S. G. M., Laumonnier, F., Zemojtel, T., Love, M. I., Richard, H., Emde, A.-K., Bienek, M., Jensen, C., Hambrock, M., Fischer, U., Langnick, C., … Kalscheuer, V. M. (2016). X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. *Molecular Psychiatry*, *21*(1), 133–148. https://doi.org/10.1038/mp.2014.193

Huang, L., Jolly, L. A., Willis-Owen, S., Gardner, A., Kumar, R., Douglas, E., Shoubridge, C., Wieczorek, D., Tzschach, A., Cohen, M., Hackett, A., Field, M., Froyen, G., Hu, H., Haas, S. A., Ropers, H.-H., Kalscheuer, V. M., Corbett, M. A., & Gecz, J. (2012). A noncoding, regulatory mutation implicates HCFC1 in nonsyndromic intellectual disability. *American Journal of Human Genetics*, *91*(4), 694–702. https://doi.org/10.1016/j.ajhg.2012.08.011

Huo, L., Teng, Z., Wang, H., & Liu, X. (2019). A novel splice site mutation in AP1S2 gene for X-linked mental retardation in a Chinese pedigree and literature review. *Brain and Behavior*, *9*(3), e01221. https://doi.org/10.1002/brb3.1221

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., & Farh, K. K.-H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, *176*(3), 535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Keane, T. M., Wong, K., & Adams, D. J. (2013). RetroSeq: Transposable element discovery from next-generation sequencing data. *Bioinformatics*, *29*(3), 389–390. https://doi.org/10.1093/bioinformatics/bts697

Kobayashi, K., Nakahori, Y., Miyake, M., Matsumura, K., Kondo-Iida, E., Nomura, Y., Segawa, M., Yoshioka, M., Saito, K., Osawa, M., Hamano, K., Sakakihara, Y., Nonaka, I., Nakagome, Y., Kanazawa, I., Nakamura, Y., Tokunaga, K., & Toda, T. (1998). An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature*, *394*(6691), 388–392. https://doi.org/10.1038/28653

Kremer, L. S., Bader, D. M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T. B., Graf, E., Schwarzmayr, T., Terrile, C., Koňaříková, E., Repp, B., Kastenmüller, G., Adamski, J., Lichtner, P., Leonhardt, C., Funalot, B., Donati, A., Tiranti, V., … Prokisch, H. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature Communications*, *8*, 15824. https://doi.org/10.1038/ncomms15824

Kumar, R., Ha, T., Pham, D., Shaw, M., Mangelsdorf, M., Friend, K. L., Hobson, L., Turner, G., Boyle, J., Field, M., Hackett, A., Corbett, M., & Gecz, J. (2016). A non-coding variant in the 5' UTR of DLG3 attenuates protein translation to cause non-syndromic intellectual disability. *European Journal of Human Genetics*, *24*(11), 1612–1616. https://doi.org/10.1038/ejhg.2016.46

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), R84. https://doi.org/10.1186/gb-2014-15-6-r84

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., & Pritchard, J. K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, *50*(1), 151–158. https://doi.org/10.1038/s41588-017-0004-9

Liu, P., Meng, L., Normand, E. A., Xia, F., Song, X., Ghazi, A., Rosenfeld, J., Magoulas, P. L., Braxton, A., Ward, P., Dai, H., Yuan, B., Bi, W., Xiao, R., Wang, X., Chiang, T., Vetrini, F., He, W., Cheng, H., … Yang, Y. (2019). Reanalysis of Clinical Exome Sequencing Data. *New England Journal of Medicine*, *380*(25), 2478–2480. https://doi.org/10.1056/NEJMc1812033

Lord, J., Gallone, G., Short, P. J., McRae, J. F., Ironfield, H., Wynn, E. H., Gerety, S. S., He, L., Kerr, B., Johnson, D. S., McCann, E., Kinning, E., Flinter, F., Temple, I. K., Clayton-Smith, J., McEntagart, M., Lynch, S. A., Joss, S., Douzgou, S., … Study, on behalf of the D. D. D. (2019). Pathogenicity and selective constraint on variation near splice sites. *Genome Research*, *29*(2), 159–170. https://doi.org/10.1101/gr.238444.118

Martínez-Garay, I., Ballesta, M. J., Oltra, S., Orellana, C., Palomeque, A., Moltó, M. D., Prieto, F., & Martínez, F. (2003). Intronic L1 insertion and F268S, novel mutations in RPS6KA3 (RSK2) causing Coffin-Lowry syndrome. *Clinical Genetics*, *64*(6), 491–496. https://doi.org/10.1046/j.1399-0004.2003.00166.x

Mattiske, T., Tan, M. H., Dearsley, O., Cloosterman, D., Hii, C. S., Gécz, J., & Shoubridge, C. (2018). Regulating transcriptional activity by phosphorylation: A new mechanism for the ARX homeodomain transcription factor. *PloS One*, *13*(11), e0206914. https://doi.org/10.1371/journal.pone.0206914

Neri, G., Schwartz, C. E., Lubs, H. A., & Stevenson, R. E. (2018). X-linked intellectual disability update 2017. *American Journal of Medical Genetics. Part A*, *176*(6), 1375–1388. https://doi.org/10.1002/ajmg.a.38710

Oberlé, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., Boué, J., Bertheas, M. F., & Mandel, J. L. (1991). Instability of a 550-Base Pair DNA Segment and Abnormal Methylation in Fragile X Syndrome. *Science*, *252*(5009), 1097–1102. https://doi.org/10.1126/science.252.5009.1097

O'Rawe, J. A., Wu, Y., Dörfel, M. J., Rope, A. F., Au, P. Y. B., Parboosingh, J. S., Moon, S., Kousi, M., Kosma, K., Smith, C. S., Tzetis, M., Schuette, J. L., Hufnagel, R. B., Prada, C. E., Martinez, F., Orellana, C., Crain, J., Caro-Llopis, A., Oltra, S., … Lyon, G. J. (2015). TAF1 Variants Are Associated with Dysmorphic Features, Intellectual Disability, and Neurological Manifestations. *American Journal of Human Genetics*, *97*(6), 922–932. https://doi.org/10.1016/j.ajhg.2015.11.005

Paila, U., Chapman, B. A., Kirchner, R., & Quinlan, A. R. (2013). GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLOS Computational Biology*, *9*(7), e1003153. https://doi.org/10.1371/journal.pcbi.1003153

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, *11*(9), 1650–1667. https://doi.org/10.1038/nprot.2016.095

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Remerand, G., Boespflug-Tanguy, O., Tonduti, D., Touraine, R., Rodriguez, D., Curie, A., Perreton, N., Des Portes, V., Sarret, C., & RMLX/AHDS Study

Group. (2019). Expanding the phenotypic spectrum of Allan-Herndon-Dudley syndrome in patients with SLC16A2 mutations. *Developmental Medicine and Child Neurology*, *61*(12), 1439–1447. https://doi.org/10.1111/dmcn.14332

Sakakibara, N., Morisada, N., Nozu, K., Nagatani, K., Ohta, T., Shimizu, J., Wada, T., Shima, Y., Yamamura, T., Minamikawa, S., Fujimura, J., Horinouchi, T., Nagano, C., Shono, A., Ye, M. J., Nozu, Y., Nakanishi, K., & Iijima, K. (2019). Clinical spectrum of male patients with OFD1 mutations. *Journal of Human Genetics*, *64*(1), 3–9. https://doi.org/10.1038/s10038-018-0532-x

Short, P. J., McRae, J. F., Gallone, G., Sifrim, A., Won, H., Geschwind, D. H., Wright, C. F., Firth, H. V., FitzPatrick, D. R., Barrett, J. C., & Hurles, M. E. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature*, *555*(7698), 611–616. https://doi.org/10.1038/nature25983

Shoubridge, C., Cloosterman, D., Parkinson-Lawerence, E., Brooks, D., & Gecz, J. (2007). Molecular pathology of expanded polyalanine tract mutations in the Aristaless-related homeobox gene. *Genomics*, *90*(1), 59–71. https://doi.org/10.1016/j.ygeno.2007.03.005

Shoubridge, Cheryl, Tan, M. H., Fullston, T., Cloosterman, D., Coman, D., McGillivray, G., Mancini, G. M., Kleefstra, T., & Gécz, J. (2010). Mutations in the nuclear localization sequence of the Aristaless related homeobox; sequestration of mutant ARX with IPO13 disrupts normal subcellular distribution of the transcription factor and retards cell division. *PathoGenetics*, *3*, 1. https://doi.org/10.1186/1755-8417-3-1

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A. D., & Cooper, D. N. (2017). The Human Gene Mutation

Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, *136*(6), 665–677. https://doi.org/10.1007/s00439-017-1779-6

Taniguchi-Ikeda, M., Kobayashi, K., Kanagawa, M., Yu, C., Mori, K., Oda, T., Kuga, A., Kurahashi, H., Akman, H. O., DiMauro, S., Kaji, R., Yokota, T., Takeda, S., & Toda, T. (2011). Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. *Nature*, *478*(7367), 127–131. https://doi.org/10.1038/nature10456

Tankard, R. M., Bennett, M. F., Degorski, P., Delatycki, M. B., Lockhart, P. J., & Bahlo, M. (2018). Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. *American Journal of Human Genetics*, *103*(6), 858–873. https://doi.org/10.1016/j.ajhg.2018.10.015

Tarpey, Patrick S., Stevens, C., Teague, J., Edkins, S., O'Meara, S., Avis, T., Barthorpe, S., Buck, G., Butler, A., Cole, J., Dicks, E., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jones, D., Menzies, A., Mironenko, T., … Raymond, F. L. (2006). Mutations in the gene encoding the Sigma 2 subunit of the adaptor protein 1 complex, AP1S2, cause X-linked mental retardation. *American Journal of Human Genetics*, *79*(6), 1119–1124. https://doi.org/10.1086/510137

Tarpey, P.S., Smith, R., Pleasance, E., Whibley, A., Edkins, S., Hardy, C., O'Meara, S., Latimer, C., Dicks, E., Menzies, A., Stephens, P., Blow, M., Greenman, C., Xue, Y., Tyler-Smith, C., Thompson, D., Gray, K., Andrews, J., Barthorpe, S., … Stratton, M. R. (2009). A systematic, large-scale resequencing screen of

X-chromosome coding exons in mental retardation. *Nature Genetics*, *41*(5), 535–543. https://doi.org/10.1038/ng.367

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *11*(1110), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Vilboux, T., Doherty, D. A., Glass, I. A., Parisi, M. A., Phelps, I. G., Cullinane, A. R., Zein, W., Brooks, B. P., Heller, T., Soldatos, A., Oden, N. L., Yildirimli, D., Vemulapalli, M., Mullikin, J. C., Nisc Comparative Sequencing Program, null, Malicdan, M. C. V., Gahl, W. A., & Gunay-Aygun, M. (2017). Molecular genetic findings and clinical correlations in 100 patients with Joubert syndrome and related disorders prospectively evaluated at a single center. *Genetics in Medicine*, *19*(8), 875–882. https://doi.org/10.1038/gim.2016.204

Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Danecek, P., Li, R., Floyd, J., Wain, L. V., Barroso, I., Humphries, S. E., … Management committee. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*(7571), 82–90. https://doi.org/10.1038/nature14962

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Webb, T. R., Parfitt, D. A., Gardner, J. C., Martinez, A., Bevilacqua, D., Davidson, A. E., Zito, I., Thiselton, D. L., Ressa, J. H. C., Apergi, M., Schwarz, N., Kanuga, N., Michaelides, M., Cheetham, M. E., Gorin, M. B., & Hardcastle, A. J. (2012). Deep intronic mutation in OFD1, identified by targeted genomic next-generation sequencing, causes a severe form of X-linked retinitis pigmentosa (RP23). *Human Molecular Genetics*, *21*(16), 3647–3654. https://doi.org/10.1093/hmg/dds194

Wright, C. F., McRae, J. F., Clayton, S., Gallone, G., Aitken, S., FitzGerald, T. W., Jones, P., Prigmore, E., Rajan, D., Lord, J., Sifrim, A., Kelsell, R., Parker, M. J., Barrett, J. C., Hurles, M. E., FitzPatrick, D. R., & Firth, H. V. (2018). Making new genetic diagnoses with old data: Iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genetics in Medicine*, *20*(10), 1216–1223. https://doi.org/10.1038/gim.2017.246

Yu, H.-C., Sloan, J. L., Scharer, G., Brebner, A., Quintana, A. M., Achilly, N. P., Manoli, I., Coughlin, C. R., Geiger, E. A., Schneck, U., Watkins, D., Suormala, T., Van Hove, J. L. K., Fowler, B., Baumgartner, M. R., Rosenblatt, D. S., Venditti, C. P., & Shaikh, T. H. (2013). An X-linked cobalamin disorder caused by mutations in transcriptional coregulator HCFC1. *American Journal of Human Genetics*, *93*(3), 506–514. https://doi.org/10.1016/j.ajhg.2013.07.022

Figure Legends

Figure 1. Functional genomic assessment of co-segregating *OFD1* and *ARX* variants in a family affected by X-linked ciliopathy. **a**. Family pedigree showing probable X-linked inheritance of ciliopathy (black symbols). DNA from individuals marked by (*) was analysed by GS. Genotypes for wild type (wt) and variant (mt) alleles of *OFD1* (O) and *ARX* (A) are shown for family members analysed by Sanger sequencing. **b**. Sashimi plot of RNA-Seq data from LCL of individual IV-2 (red) and a representative control LCL (blue) for OFD1 exons 13, 14 and 15. The percentage of reads supporting each intron from the total number or reads supporting all splice junctions using the exon 13 splice donor site are shown for both samples. The predicted outcomes for protein translation caused by the novel exon are shown below the plot (the predicted translated sequences are in Supp. Data). **c**. Western blot of protein extracts from a LCL from IV-2 (first two lanes are extracts from cell pellets from independent cultures) compared to extracts from three unrelated male control LCLs and adult mouse cortex. Blots were probed with anti-OFD1 (Sigma cat# SAB2702042) and rabbit anti-β-tubulin (Abcam cat# ab6046) antibodies. **d.** Luciferase reporter activity normalised to *Renilla* reporter activity and expressed as a percentage relative to empty Myc-vector transfected cells (dark grey). Full-length Myc-tagged constructs; ARX WT (white), a nuclear localisation sequence (NLS) variant T333N (black), a variant in the homeodomain but outside the NLS regions P353L (diagonal lines) and the novel missense variant G402R (light grey). Error bars show standard deviations of three independent transfections carried out in triplicate.

Figure 2. Retrotransposon insertion of an SVA_E attenuates *SLC16A2* expression. **a**. Pedigree shows two affected males with phenotypes characteristic of AHDS potentially linked through unaffected obligate carrier females. DNA from individuals indicated by (*) was analysed by GS. Genotypes of individuals with either the reference (wt) or SVA_E inserted allele (i) are shown where tested (see also Supp. Fig. S5c). **b**. IGV screen shot showing a cluster of discordantly mapped reads in III-4 and IV-6 but not in an unrelated control genome alignment. The different colours correspond to the identity of the chromosome to which the other end of the read-pair is mapped as indicated by the key on the right of the image. Below the alignment is a schematic of the *SLC16A2* gene structure and the orientation of the SVA_E transposon inserted into intron 5. **c**. Quantified expression of *SLC16A2* expression relative to *GAPDH* in three unrelated control fibroblast cell lines compared with a fibroblast line derived from IV-6. The PCR product crosses the boundary between exon 5 and 6. Error bars show standard deviations between biological replicate samples averaged from two experiments done in triplicate. **d**. PCR products of *SLC16A2* from a fibroblast line derived from IV-6 and three control fibroblast cell lines, size separated on 1% agarose gel and stained with ethidium bromide. PCRs were run for 30 cycles for all amplicons. Note that all products that cross the exon boundaries over intron 5 are substantially reduced in IV-6. The uncropped gels are shown in Supp. Fig. S6.

Figure 3. A canonical splice site variant in *AP1S2* leads to aberrant splicing and reduced protein expression. **a**. Pedigree showing two affected brothers whose genomic DNA was analysed by GS (*). **b**. IGV alignment of GS data from II-1 and

II-3 showing the chrX:g.15872810C>T transition in *AP1S2*. Colours indicate mapping orientation of the reads. **c**. Stylised representation of the *AP1S2* gene (not to scale). Exons are indicated by boxes and within these, the open reading frame (grey shading) and untranslated regions (white shading) are shown. Positions of primers used to evaluate *AP1S2* expression and splicing are shown on the image as numbered half-arrows. Below the gene model is a sashimi plot of RNA-Seq data from LCL of individual II-1 (red, maximum read depth 39) and a representative control LCL (blue, maximum read depth 515) for *AP1S2* exons 1 and 2. Note that intron 1 is retained in the affected male. The peak that appears relatively prominently upstream of *AP1S2* Exon 1 in II-1 is an antisense transcript that is present in all samples. **d-f**. RT-PCR analyses of cDNA reactions carried out in the presence (+) or absence (-) of reverse transcriptase (RT) using RNA extracted from affected individuals II-1 and II-3 of Family 3 compared to an unrelated male control LCL and human fetal brain. Genomic DNA (gDNA) from II-1 and II-3 are included as controls to show when the primer pairs also amplified the closely related *AP1S2P1* pseudogene sequence. **d**. Primer pair P356 and P344 amplify a 405 bp band in control LCL and fetal brain but not II-1 and II-3, suggesting splicing of *AP1S2* is impaired between exon 1 and exon 2. **e**. Primer pair P354 and P351 show potentially reduced amplification of the 178 bp band corresponding to *AP1S2* transcript in II-1 and II-3. Note that an identical 178 bp band corresponding to the *AP1S2P1* pseudogene amplifies in the genomic DNA samples in addition to the 564bp band spanning intron 3 of *AP1S2*. **f**. Primer pair P343 and P359 generate a 1182 bp product in II-1 and II-3 that suggests retention of intron1 in a transcript that is otherwise correctly spliced for exons 2 – 5, (ns; non-specific). **g**. Short and long exposures of the same western blot detecting AP1S2 (with a primary antibody from Abcam cat# ab97590) in protein extracts from II-1, II-3 and four unrelated male control LCLs compared to β-III tubulin (as a loading control). Blot shows absence of AP1S2 in both II-1 and II-3. The uncropped gel is shown in Supp. Fig. S8e.

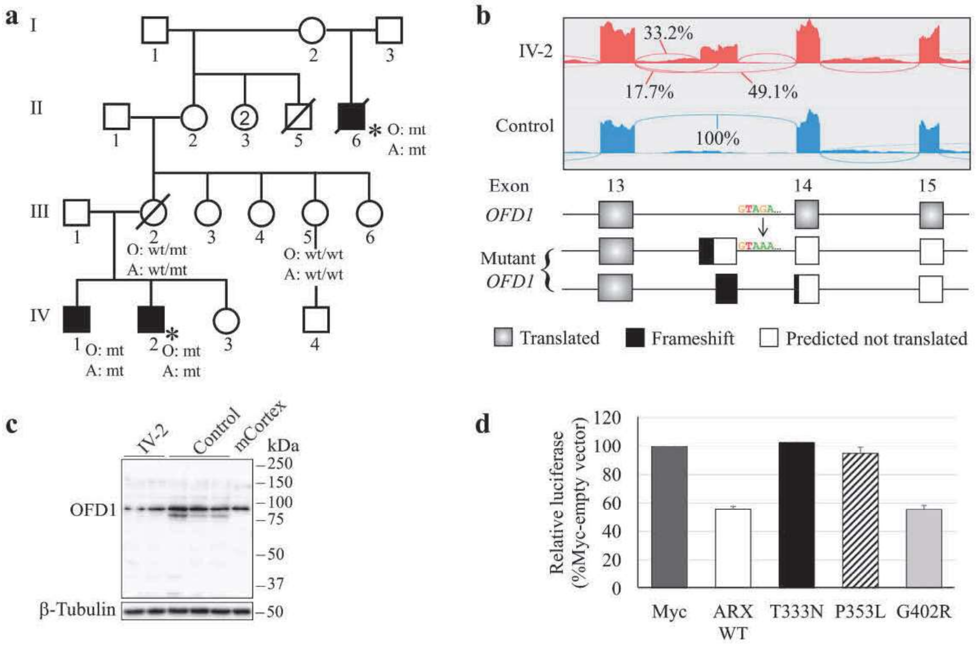


Figure 1. Functional genomic assessment of co-segregating OFD1 and ARX variants in a family affected by X-linked ciliopathy. a. Family pedigree showing probable X-linked inheritance of ciliopathy (black symbols). DNA from individuals marked by (*) was analysed by GS. Genotypes for wild type (wt) and variant (mt) alleles of OFD1 (O) and ARX (A) are shown for family members analysed by Sanger sequencing. b. Sashimi plot of RNA-Seq data from LCL of individual IV-2 (red) and a representative control LCL (blue) for OFD1 exons 13, 14 and 15. The percentage of reads supporting each intron from the total number or reads supporting all splice junctions using the exon 13 splice donor site are shown for both samples. The predicted outcomes for protein translation caused by the novel exon are shown below the plot (the predicted translated sequences are in Supplementary Data). c. Western blot of protein extracts from a LCL from IV-2 (first two lanes are extracts from cell pellets from independent cultures) compared to extracts from three unrelated male control LCLs and adult mouse cortex. Blots were probed with anti-OFD1 (Sigma cat# SAB2702042) and rabbit anti-β-tubulin (Abcam cat# ab6046) antibodies. d. Luciferase reporter activity normalised to Renilla reporter activity and expressed as a percentage relative to empty Myc-vector transfected cells (dark grey). Full-length Myc-tagged constructs; ARX WT (white), a nuclear localisation sequence (NLS) variant T333N (black), a variant in the homeodomain but outside the NLS regions P353L (diagonal lines) and the novel missense variant G402R (light grey). Error bars show standard deviations of three independent transfections carried out in triplicate.

180x118mm (600 x 600 DPI)

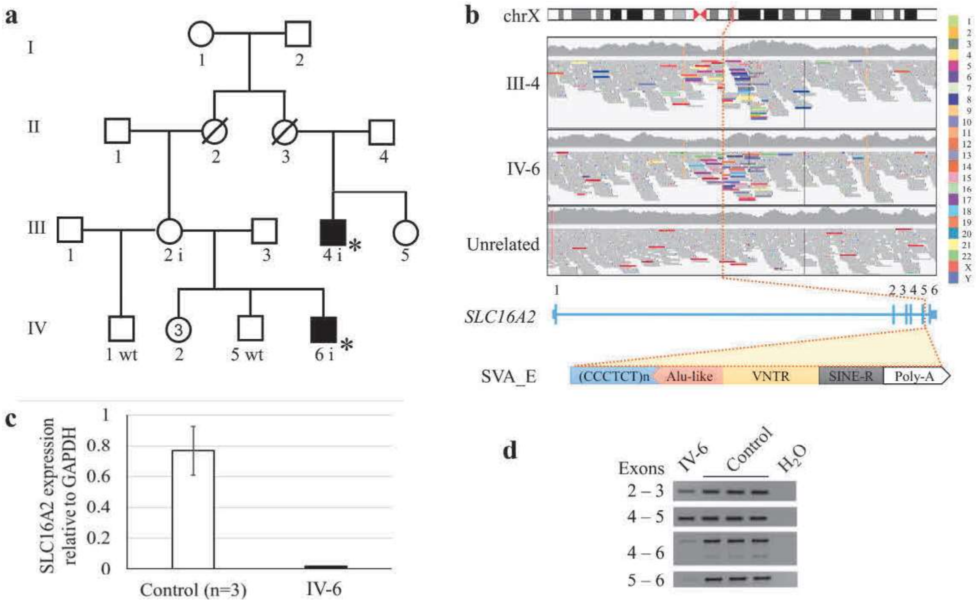


Figure 2. Retrotransposon insertion of an SVA_E attenuates SLC16A2 expression. a. Pedigree shows two affected males with phenotypes characteristic of AHDS potentially linked through unaffected obligate carrier females. DNA from individuals indicated by (*) was analysed by GS. Genotypes of individuals with either the reference (wt) or SVA_E inserted allele (i) are shown where tested (see also Supplementary Fig. 4c). b. IGV screen shot showing a cluster of discordantly mapped reads in III-4 and IV-6 but not in an unrelated control genome alignment. The different colours correspond to the identity of the chromosome to which the other end of the read-pair is mapped as indicated by the key on the right of the image. Below the alignment is a schematic of the SLC16A2 gene structure and the orientation of the SVA_E transposon inserted into intron 5. c. Quantified expression of SLC16A2 expression relative to GAPDH in three unrelated control fibroblast cell lines compared with a fibroblast line derived from IV-6. The PCR product crosses the boundary between exon 5 and 6. Error bars show standard deviations between biological replicate samples averaged from two experiments done in triplicate. d. PCR products of SLC16A2 from a fibroblast line derived from IV-6 and three control fibroblast cell lines, size separated on 1% agarose gel and stained with ethidium bromide. PCRs were run for 30 cycles for all amplicons. Note that all products that cross the exon boundaries over intron 5 are substantially reduced in IV-6. The uncropped gels are shown in supplementary Fig. 5.

180x110mm (600 x 600 DPI)

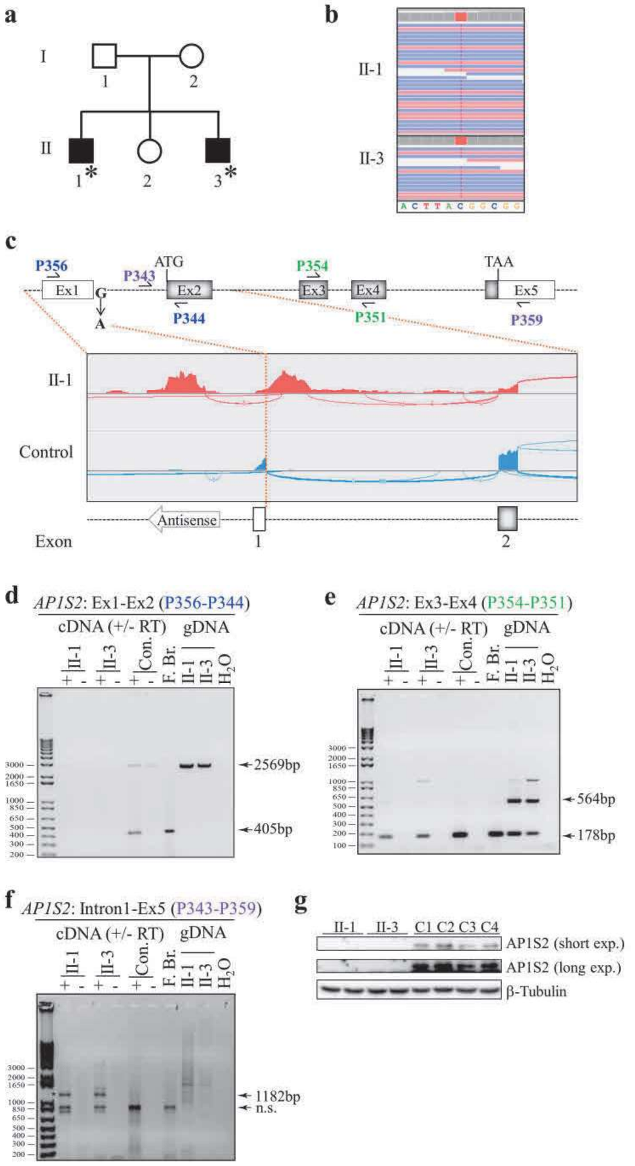


Figure 3. A canonical splice site variant in AP1S2 leads to aberrant splicing and reduced protein expression. a. Pedigree showing two affected brothers whose genomic DNA was analysed by GS (*). b. IGV alignment of GS data from II-1 and II-3 showing the chrX:g.15872810C>T transition in AP1S2. Colours indicate mapping orientation of the reads. c. Stylised representation of the AP1S2 gene (not to scale). Exons are indicated by boxes, the open reading frame (grey shading) and untranslated regions (white shading) are shown. Positions of primers used to evaluate AP1S2 expression and splicing are shown on the image as numbered half-arrows. Below the gene model is a sashimi plot of RNA-Seq data from LCL of individual II-1 (red, maximum read depth 39) and a representative control LCL (blue, maximum read depth 515) for AP1S2 exons 1 and 2. Note that intron 1 is retained in the affected male. The peak that appears relatively prominently upstream of AP1S2 Exon 1 in II-1 is an antisense transcript that is present in all samples. d-f. RT-PCR analyses of cDNA reactions carried out in the presence (+) or absence (-) of reverse transcriptase (RT) using RNA extracted from affected individuals II-1 and II-3 of Family 3 compared to an unrelated male control LCL and human fetal brain. Genomic DNA (gDNA) from II-1 and II-3 are included as controls to show when the primer pairs also amplified the closely related AP1S2P1 pseudogene sequence. d. Primer pair P356

and P344 amplify a 405 bp band in control LCL and fetal brain but not II-1 and II-3, suggesting splicing of AP1S2 is impaired between exon 1 and exon 2. e. Primer pair P354 and P351 show potentially reduced amplification of the 178 bp band corresponding to AP1S2 transcript in II-1 and II-3. Note that an identical 178 bp band corresponding to the AP1S2P1 pseudogene amplifies in the genomic DNA samples in addition to the 564bp band spanning intron 3 of AP1S2. f. Primer pair P343 and P359 generate a 1182 bp product in II-1 and II-3 that suggests retention of intron1 in a transcript that is otherwise correctly spliced for exons 2 – 5, (ns; non-specific). g. Short and long exposures of the same western blot detecting AP1S2 (with a primary antibody from Abcam cat# ab97590) in protein extracts from II-1, II-3 and four unrelated male control LCLs compared to □-III tubulin (as a loading control). Blot shows absence of AP1S2 in both II-1 and II-3. The uncropped gel is shown in Supp. Fig. S8e.

202x372mm (600 x 600 DPI)

Supplementary Material

Different types of disease-causing non-coding variants revealed by genomic and gene expression analyses in families with X-linked intellectual disability

Michael J. Field[1], Raman Kumar[2], Anna Hackett[1,3], Sayaka Kayumi[2], Cheryl A. Shoubridge[2], Lisa J. Ewans[4,5], Atma M. Ivancevic[6], Tracy Dudding-Byth[1,3], Renée Carroll[2], Thessa Kroes [2], Alison E. Gardner [2], Patricia Sullivan[7], Thuong T. Ha[8], Charles E. Schwartz[9], Mark J. Cowley[1,5,7], Marcel E. Dinger[10], Elizabeth E. Palmer[1,11], Louise Christie[1], Marie Shaw[2], Tony Roscioli[12,13], Jozef Gecz[2,14] and Mark A. Corbett[2]*

1. NSW Genetics of Learning Disability Service, Newcastle, NSW, Australia.
2. Adelaide Medical School and Robinson Research Institute, University of Adelaide, Adelaide, SA, Australia.
3. University of Newcastle, Newcastle, NSW, Australia.
4. St Vincent's Clinical School, University of New South Wales, Darlinghurst, Australia.
5. Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia.
6. University of Colorado, Boulder, CO, USA.
7. Children's Cancer Institute, UNSW Sydney, Kensington, NSW, Australia.
8. Molecular Pathology Department, Centre for Cancer Biology, SA Pathology, Adelaide, SA, Australia.
9. Greenwood Genetics Centre, Greenwood, SC, USA.
10. School of Biotechnology and Biomolecular Sciences, UNSW Sydney, Kensington NSW 2052, Australia.
11. School of Women's and Children's Health, UNSW Sydney, Kensington, Randwick, Sydney, NSW, Australia.

12. NeuRA, University of New South Wales, Sydney, NSW, Australia.

13. Centre for Clinical Genetics, Sydney Children's Hospital, Randwick, Sydney, NSW, Australia.

14. South Australian Health and Medical Research Institute, Adelaide, SA, Australia.

* For correspondence:

Mark Corbett, Ph.D.

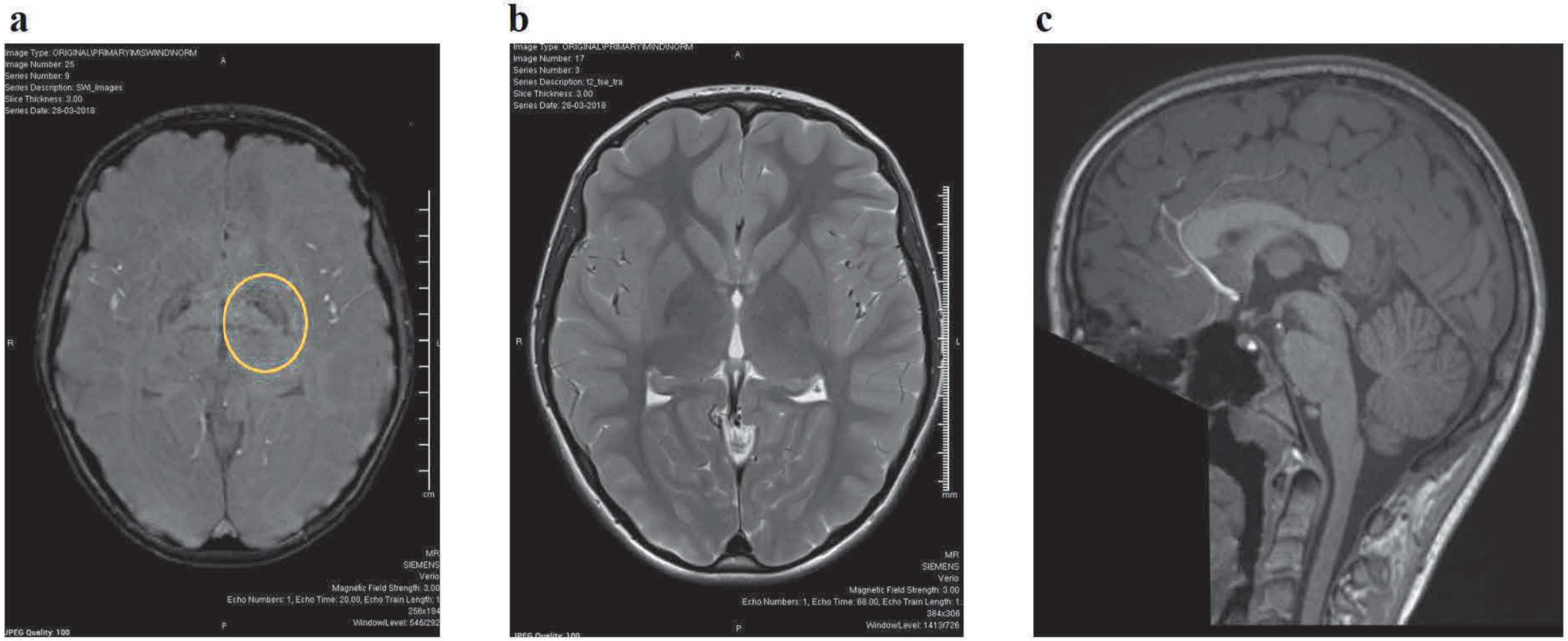
Australian Collaborative Cerebral Palsy Research Group and Neurogenetics Research Program, Adelaide Medical School,

University of Adelaide, Adelaide,

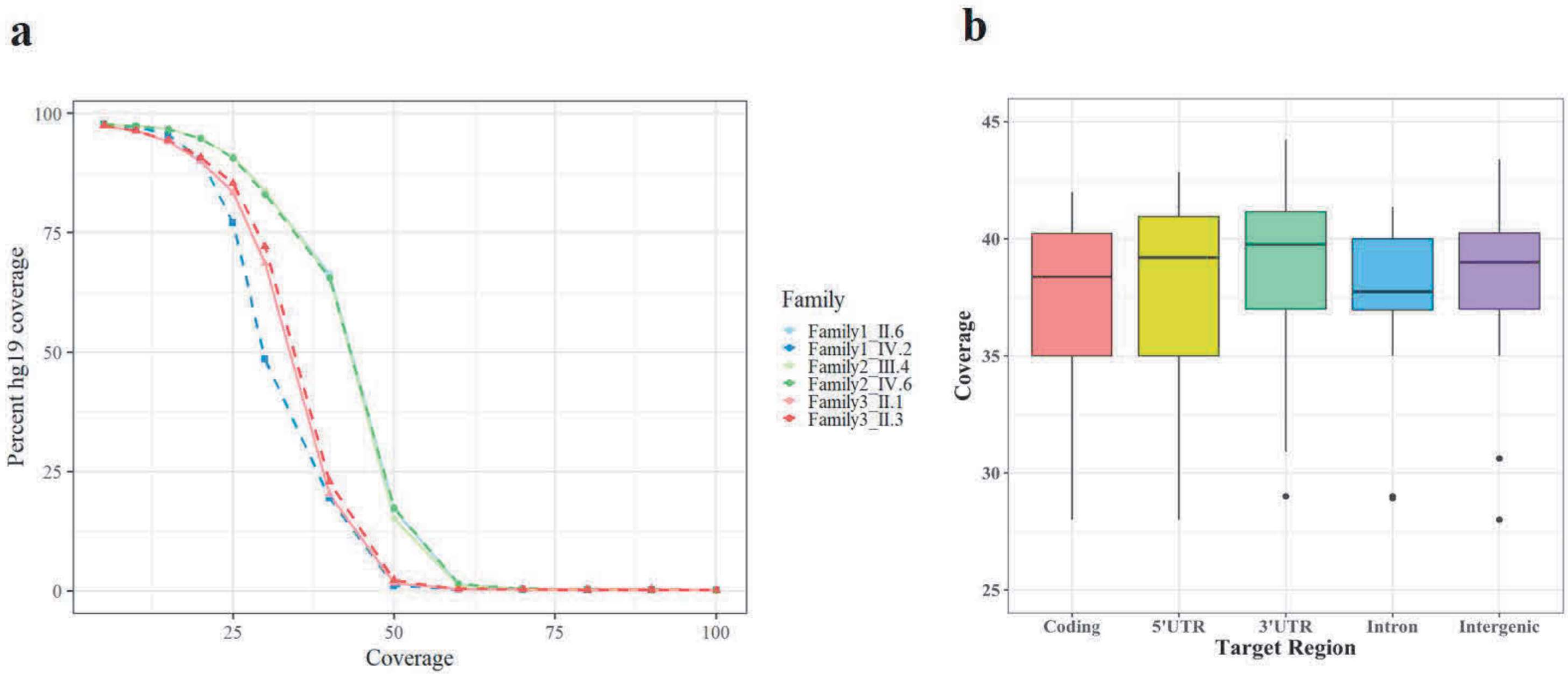
South Australia, 5000, Australia.

Phone: +61 8 83137938

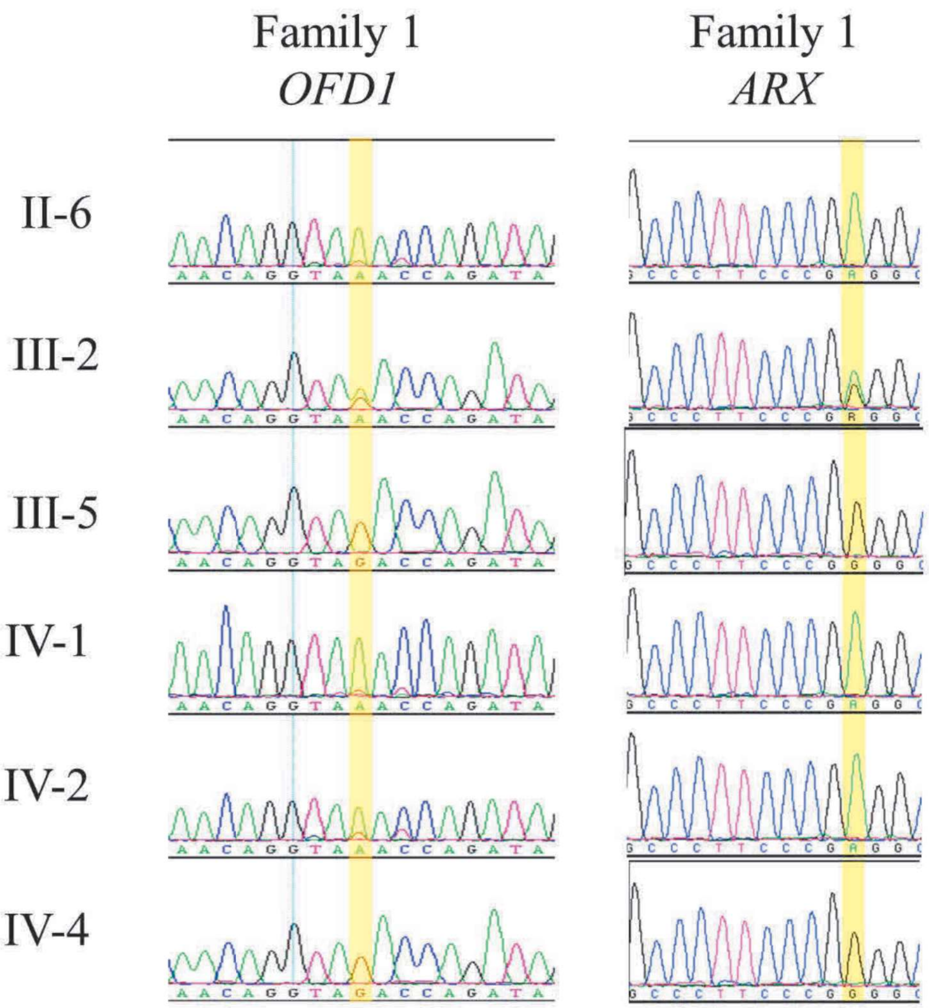
e-mail: mark.corbett@adelaide.edu.au


a
Image Type: ORIGINAL\PRIMARY\M\SW\ND\NORM
Image Number: 25
Series Number: 9
Series Description: SWI_Images
Slice Thickness: 3.00
Series Date: 28-03-2018
A
R
L
P
cm
MR
SIEMENS
Verio
Magnetic Field Strength: 3.00
Echo Numbers: 1, Echo Time: 20.00, Echo Train Length: 1
256x184
Window/Level: 546/292
JPEG Quality: 100
b
Image Type: ORIGINAL\PRIMARY\M\ND\NORM
Image Number: 17
Series Number: 3
Series Description: t2_tse_tra
Slice Thickness: 3.00
Series Date: 28-03-2018
A
R
L
P
mm
MR
SIEMENS
Verio
Magnetic Field Strength: 3.00
Echo Numbers: 1, Echo Time: 68.00, Echo Train Length: 1
384x306
Window/Level: 1413/726
JPEG Quality: 100
c

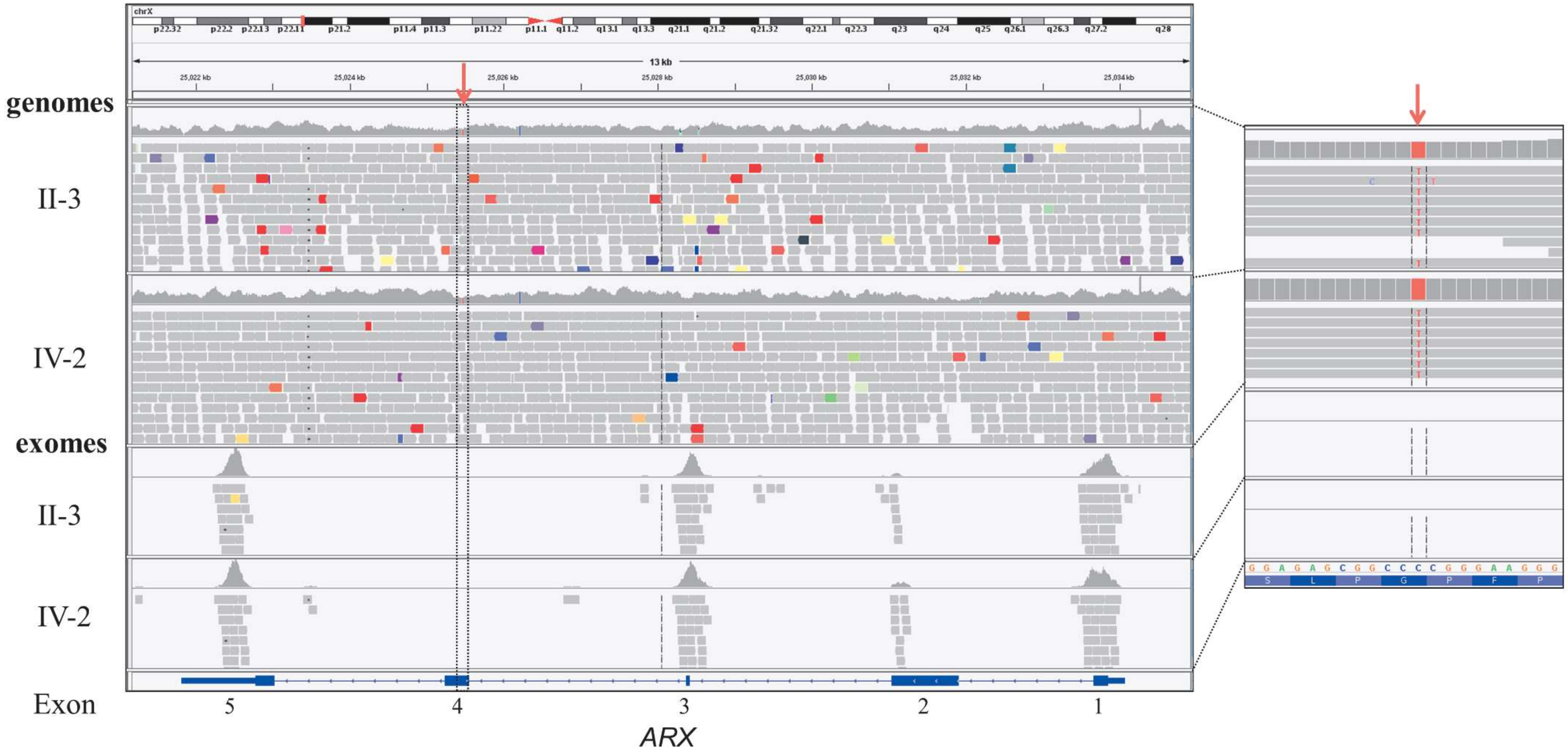
Supplementary Figure S1. MRI from affected male II-1 from Family 3. **a.** A transverse susceptibility weighted image showing bilateral lucent areas indicating calcification of the basal ganglia outlined by the yellow circle. **b.** A transverse T2 weighted image from the same individual. **c.** Sagittal T1 weighted image shows no distinct infratentorial abnormalities



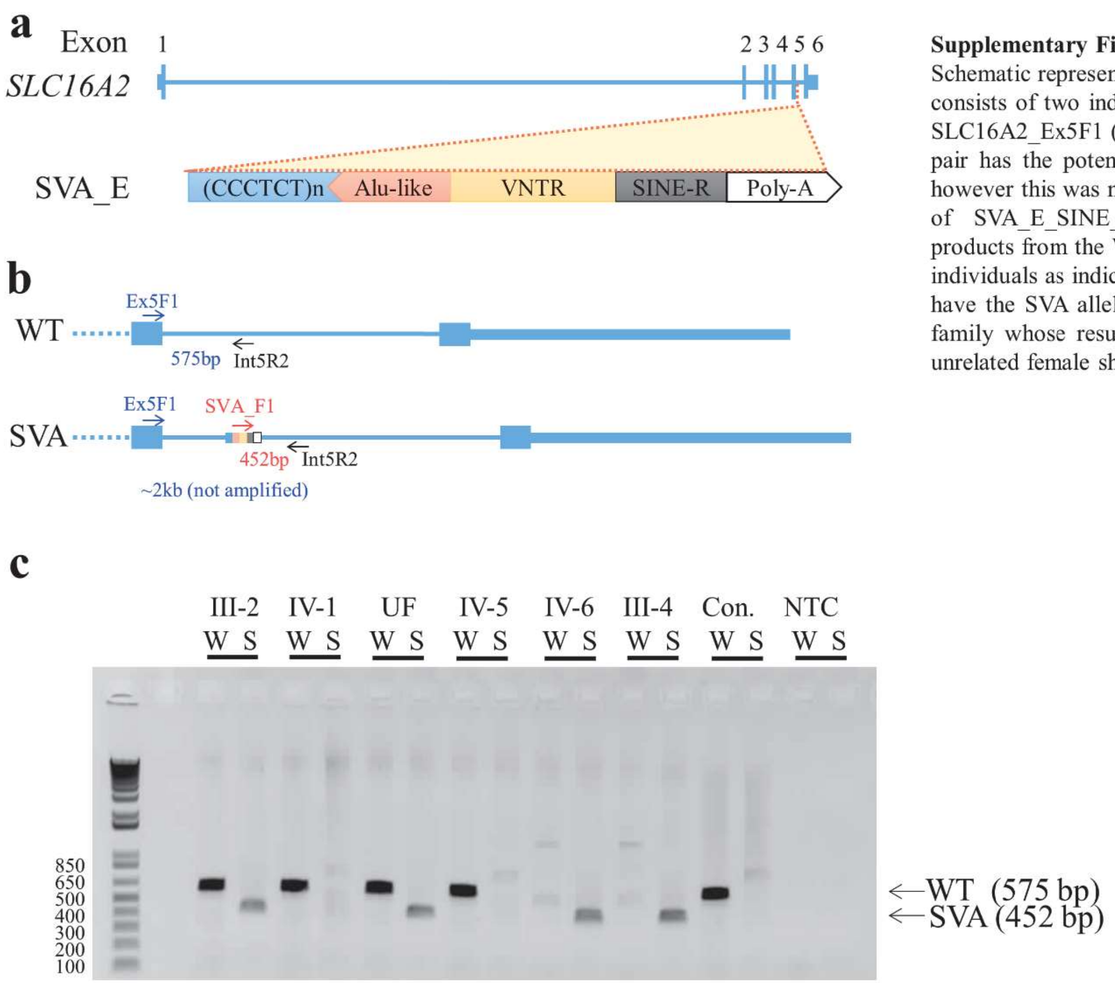
Supplementary Figure S2. Alignment statistics. **a**. Fraction of bases aligned to the hg19 build of the genome at specific read depths. All samples had more than 80% of target bases covered with at least 30 reads, except II-6 from Family 1 (68%). **b**. Box plots of GS data from family 1 and 2 as well as 6 additional GS from unrelated individuals sequenced at the same time shows the median coverage depths of different genomic regions are not significantly different from each other as indicated by the solid black horizontal line in each box. Outliers with values > 1.5x the interquartile range are shown as points.



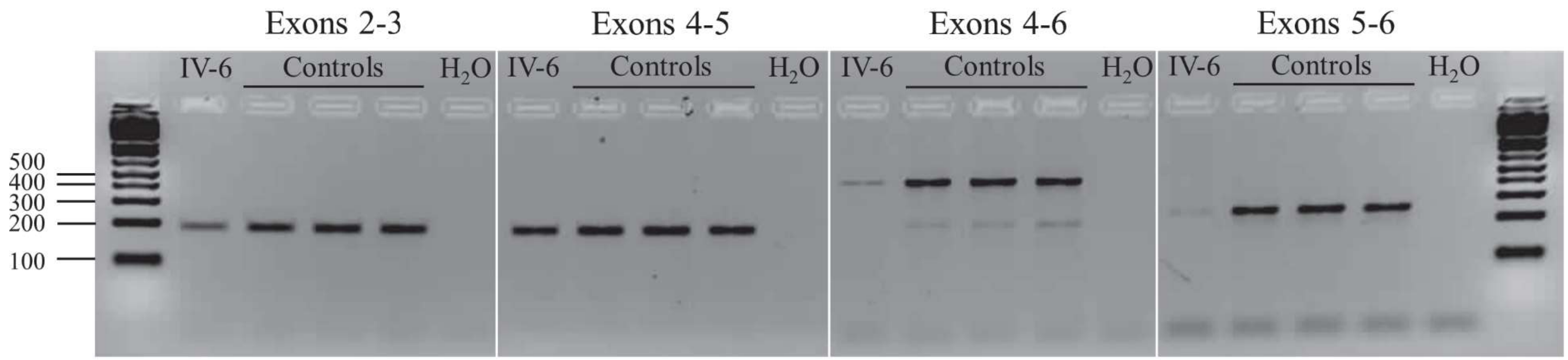
Supplementary Figure S3. Sanger sequence traces showing segregation of the *OFD1* NM_003611.2:c.1412-322G>A and *ARX* NM_139058.3:c.1204G>A variants in this family as indicated relative to members of the pedigree shown in Fig. 1a.



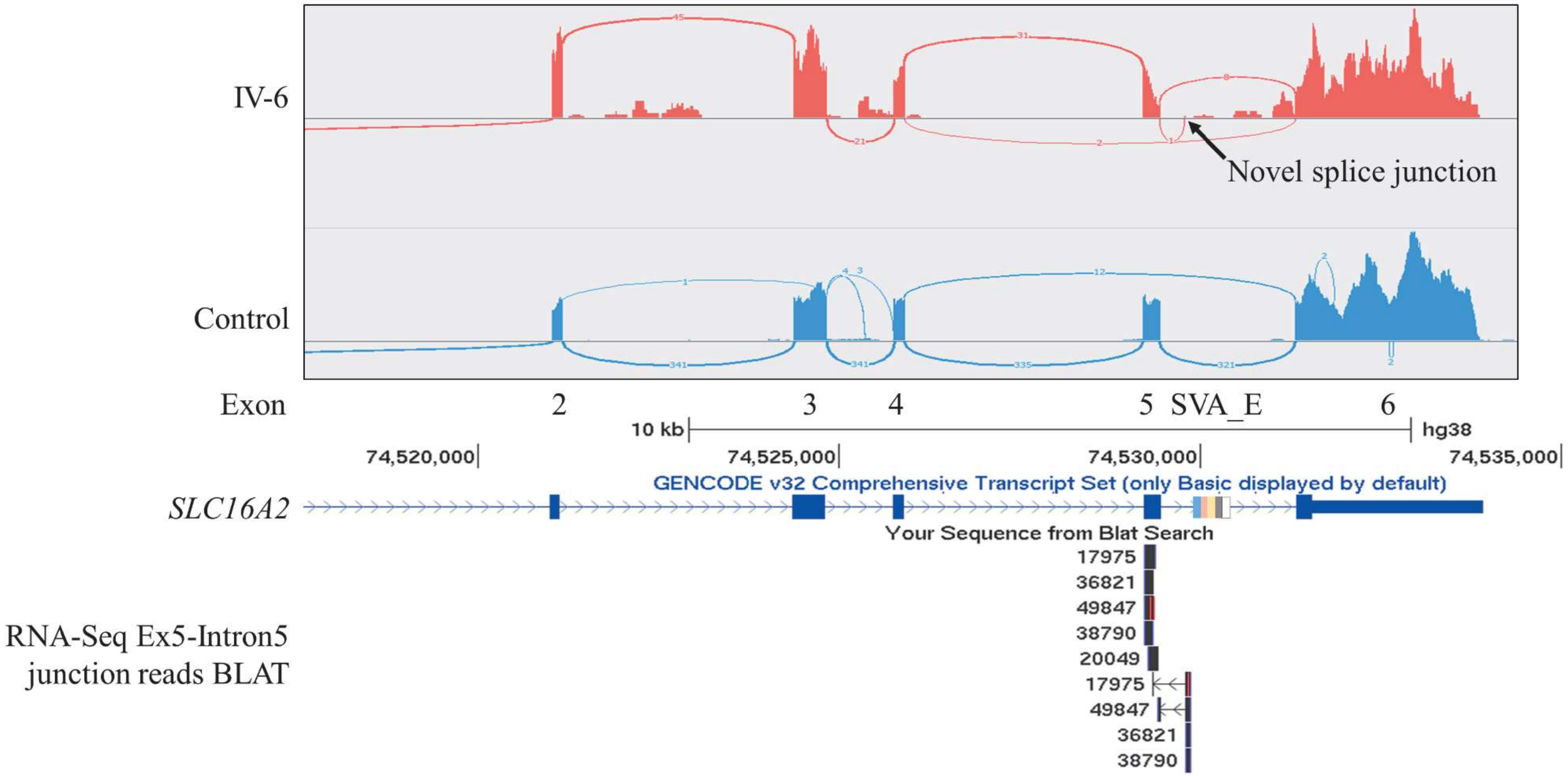
Supplementary Figure S4. Comparison of whole genome and whole exome sequencing alignments and coverage for *ARX* in Family 1. Integrative genome viewer alignment of the entire *ARX* gene (left) or zoomed in on the boxed region (right). The red arrows show the NM_139058.3:c.1204G>A (C>T in the genomic context) variant in exon 4 of *ARX*.



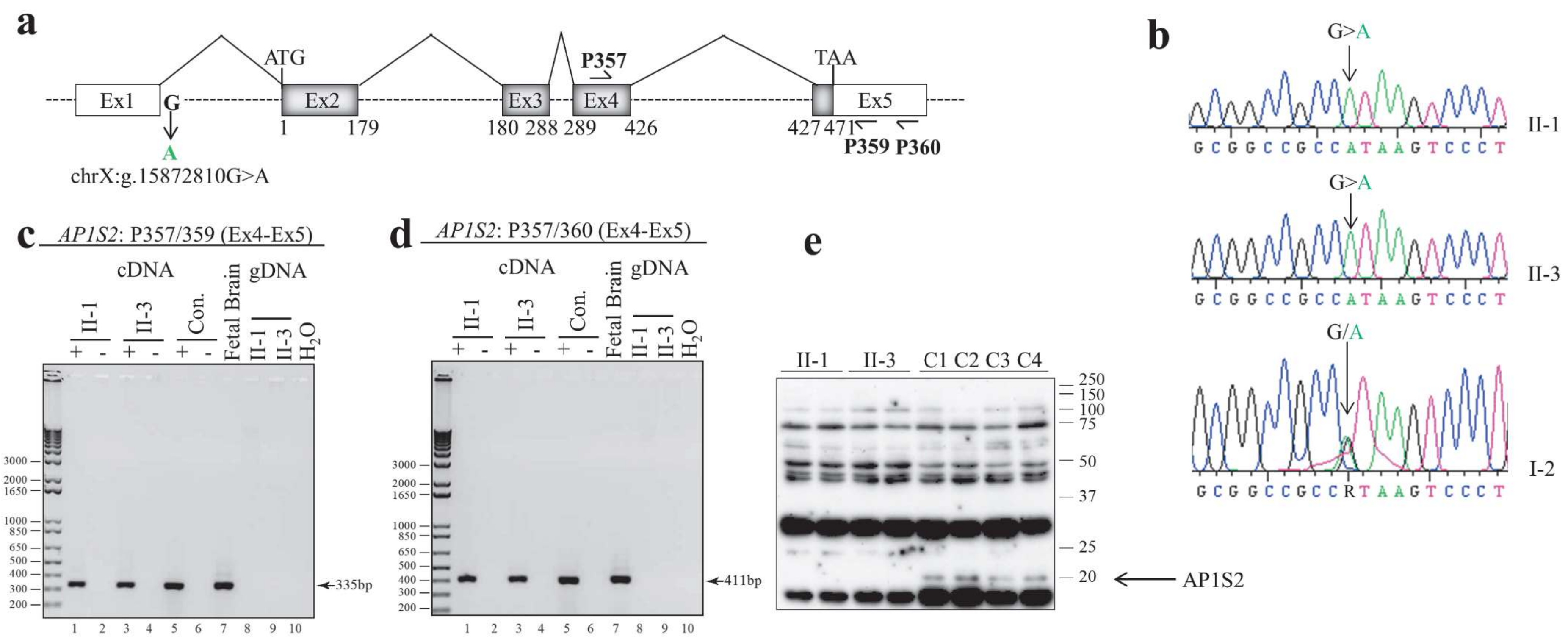
Supplementary Figure S5. Segregation of the SVA_E insertion into intron 5 of *SLC16A2*. **a**. Schematic representation of the SVA_E retrotransposon insertion. **b**. The PCR screening assay consists of two independent reactions: To detect the wild type allele (WT) a combination of SLC16A2_Ex5F1 (blue) and SLC16A2_Int5R2 (black) produce a 575bp product. This primer pair has the potential to also produce an estimated 1957 bp band from the mutant allele however this was never observed. To detect the SVA_E insertion allele (SVA) a combination of SVA_E_SINE_F (red) and SLC16A2_Int5R2 (black) produce a 452 bp band. **c**. PCR products from the WT (W) and SVA (S) alleles separated by agarose gel electrophoresis from individuals as indicated on the pedigree in Figure 2a. Note, affected males III-4 and IV-6 only have the SVA allele. UF is an unaffected female (not identified on the pedigree) from the family whose result indicates carrier status; Con. screening result from an unaffected and unrelated female shows amplification of the WT allele only, NTC no template control.



Supplementary Figure S6. Uncropped gels from Figure 2d. PCR products of *SLC16A2* from a fibroblast line derived from IV-6 and three male control fibroblast cell lines, size separated on 1% agarose gel and stained with ethidium bromide. PCRs were run for 30 cycles for all amplicons. Note that all products that cross the exon boundaries over intron 5 are substantially reduced in IV-6.



Supplementary Figure S7. Aberrant splicing of *SLC16A2* caused by insertion of an SVA_E retrotransposon. Upper panel shows a sashimi plot of RNA-Seq from data from RNA extracted from fibroblasts of IV-6 from family 2 (red) and a representative male control fibroblast (blue). Only a single read pair supported the novel splice junction indicated in intron 5 from the plot. Lower panel shows alignment of the read supporting the novel junction (49847) and additional read pairs where one read was mapped to exon 5 and the second read was not mapped by HISAT2 (17975, 36821 and 38790) using the BLAT program in UCSC genome browser. All of these reads support the novel splice junction at hg38 chrX:74529780, 76 base pairs upstream of the 5'end of the novel SVA_E insertion NC_000023.11:g.74529856_74529857 (ClinVar; VCV000929441.1).



Supplementary Figure S8. Splicing of *AP1S2* **a**. From Figure 3c, stylized representation of the *AP1S2* gene (not to scale). Exons are indicated by boxes, the open reading frame (grey shading) and untranslated regions (white shading) are shown respectively. Positions of primers used to evaluate *AP1S2* expression and splicing in the gels below are shown on the image as numbered arrows. **b** Sanger sequencing confirming segregation of the NM_003916.3:c.-1+1G>A variant in this family **c** and **d**. RT-PCR analyses of cDNA produced from RNA extracted from affected individuals II-11 and II-3 compared to a control LCL, human fetal brain. Genomic DNA from II-1 and II-3 are included as controls to show these primer pairs do not amplify the closely related *AP1S2P1* pseudogene sequence. Both primer pairs P357 and P359 (**b**) and P357 and P360 (**c**) amplify correctly spliced transcripts between exon 4 and 5. A slight reduction in the abundance of transcript in II-1 and II-3 is seen compared to the control. **e**. Full western blot from Figure 3g detecting AP1S2 in protein extracts from II-1, II-3 and four unrelated control LCL compared to β-III tubulin. Blot shows absence of a 20kDa band corresponding to AP1S2 in both II-1 and II-3.

Predicted translation of novel *OFD1* transcripts

>NP_003602.1 oral-facial-digital syndrome 1 protein isoform 1 [*Homo sapiens*]
MMAQSNMFTVADVLSQDELRKKLYQTFKDRGILDTLKTQLRNQLIHELMH
PVLSGELQPRSISVEGSSLLIGASNSLVADHLQRCGYEYSLSVFFPESGL
AKEKVFTMQDLLQLIKINPTSSLYKSLVSGSDKENQKGFLMHFLKELAEY
HQAKESCNMETQTSSTFNRDSLAEKLQLIDDQFADAYPQRIKFESLEIKL
NEYKREIEEQLRAEMCQKLKFFKDTEIAKIKMEAKKKYEKELTMFQNDFE
KACQAKSEALVLREKSTLERIHKHQEIETKEIYAQRQLLLKDMDLLRGRE
AELKQRVEAFELNQKLQEEKHKSITEALRRQEQNIKSFEETYDRKLKNEL
LKYQLELKDDYIIRTNRLIEDERKNKEKAVHLQEELIAINSKKEELNQSV
NRVKELELELESVKAQSLAITKQNHMLNEKVKEMSDYSLLKEEKLELLAQ
NKLLKQQLEESRNENLRLLN**R**LAQPAPELAVFQKELRKAEKAIVVEHEEF
ESCRQALHKQLQDEIEHSAQLKAQILGYKASVKSLTTQVADLKLQLKQTQ
TALENEVYCNPKQSVIDRSVNGLINGNVVPCNGEISGDFLNNPFKQENVL
ARMVASRITNYPTAWVEGSSPDSDLEFVANTKARVKELQQEAERLEKAFR
SYHRRVIKNSAKSPLAAKSPPSLHLLEAFKNITSSSPERHIFGEDRVVSE
QPQVGTLEERNDVVEALTGSAASRLRGGTSSRRLSSTPLPKAKRSLESEM
YLEGLGRSHIASPSPCPDRMPLPSPTESRHSLSIPPVSSPPEQKVGLYRR
QTELQDKSEFSDVDKLAFKDNEEFESSFESAGNMPRQLEMGGLSPAGDMS
HVDAAAAAVPLSYQHPSVDQKQIEEQKEEEKIREQQVKERRQREERRQSN
LQEVLERERRELEKLYQERKMIEESLKIKIKKELEMENELEMSNQEIKDK
SAHSENPLEKYMKIIQQEQDQESADKSSKKMVQEGSLVDTLQSSDKVESL
TGFSHEELDDSW
> NP_003602.1:p.Leu472ProfsTer26
MMAQSNMFTVADVLSQDELRKKLYQTFKDRGILDTLKTQLRNQLIHELMH
PVLSGELQPRSISVEGSSLLIGASNSLVADHLQRCGYEYSLSVFFPESGL
AKEKVFTMQDLLQLIKINPTSSLYKSLVSGSDKENQKGFLMHFLKELAEY
HQAKESCNMETQTSSTFNRDSLAEKLQLIDDQFADAYPQRIKFESLEIKL
NEYKREIEEQLRAEMCQKLKFFKDTEIAKIKMEAKKKYEKELTMFQNDFE
KACQAKSEALVLREKSTLERIHKHQEIETKEIYAQRQLLLKDMDLLRGRE
AELKQRVEAFELNQKLQEEKHKSITEALRRQEQNIKSFEETYDRKLKNEL
LKYQLELKDDYIIRTNRLIEDERKNKEKAVHLQEELIAINSKKEELNQSV
NRVKELELELESVKAQSLAITKQNHMLNEKVKEMSDYSLLKEEKLELLAQ
NKLLKQQLEESRNENLRLLNR**PRSANSMALLLAHPGNSTILCAYPE**
> NP_003602.1:p.Leu472PhefsTer37
MMAQSNMFTVADVLSQDELRKKLYQTFKDRGILDTLKTQLRNQLIHELMH
PVLSGELQPRSISVEGSSLLIGASNSLVADHLQRCGYEYSLSVFFPESGL
AKEKVFTMQDLLQLIKINPTSSLYKSLVSGSDKENQKGFLMHFLKELAEY
HQAKESCNMETQTSSTFNRDSLAEKLQLIDDQFADAYPQRIKFESLEIKL
NEYKREIEEQLRAEMCQKLKFFKDTEIAKIKMEAKKKYEKELTMFQNDFE
KACQAKSEALVLREKSTLERIHKHQEIETKEIYAQRQLLLKDMDLLRGRE
AELKQRVEAFELNQKLQEEKHKSITEALRRQEQNIKSFEETYDRKLKNEL
LKYQLELKDDYIIRTNRLIEDERKNKEKAVHLQEELIAINSKKEELNQSV
NRVKELELELESVKAQSLAITKQNHMLNEKVKEMSDYSLLKEEKLELLAQ
NKLLKQQLEESRNENLRLLNR**FLDDLDRESHLPSAWIPTAAVRCPDHIGS
QGCHQQA**

Predicted translation of a novel *SLC16A2* transcript caused by aberrant splicing into intron 5 predicted from RNA-Seq data of IV-6 from Family 2. The first 11 transmembrane domains are highlighted in cyan while the twelfth domain, which is deleted in the predicted p.(Leu468LysfsTer1) mutant protein is highlighted in magenta.

```
>NP_006508.2 monocarboxylate transporter 8 [Homo sapiens]
MALQSQASEEAKGPWQEADQEQQEPVGSPEPESEPEPEPEPEPVPVPPPE
PQPEPQPLPDPAPLPELEFESERVHEPEPTPTVETRGTARGFQPPEGGFG
WVVVFAATWCNGSIFGIHNSVGILYSMLLEEEKEKNRQVEFQAAWVGALA
MGMIFFCSPIVSIFTDRLGCRITATAGAAVAFIGLHTSSFTSSLSLRYFT
YGILFGCGCSFAFQPSLVILGHYFQRRLGLANGVVSAGSSIFSMSFPFLI
RMLGDKIKLAQTFQVLSTFMFVLMLLSLTYRPLLPSSQDTPSKRGVRTLH
QRFLAQLRKYFNMRVFRQRTYRIWAFGIAAAALGYFVPYVHLMKYVEEEF
SEIKETWVLLVCIGATSGLGRLVSGHISDSIPGLKKIYLQVLSFLLLGLM
SMMIPLCRDFGGLIVVCLFLGLCDGFFITIMAPIAFELVGPMQASQAIGY
LLGMMALPMIAGPPIAGLLRNCFGDYHVAFYFAGVPPIIGAVILFFVPLM
HQRMFKKEQRDSSKDKMLAPDPDPNGELLPGSPNPEEPI

>NP_006508.2:p.(Leu468LysfsTer1)
MALQSQASEEAKGPWQEADQEQQEPVGSPEPESEPEPEPEPEPVPVPPPE
PQPEPQPLPDPAPLPELEFESERVHEPEPTPTVETRGTARGFQPPEGGFG
WVVVFAATWCNGSIFGIHNSVGILYSMLLEEEKEKNRQVEFQAAWVGALA
MGMIFFCSPIVSIFTDRLGCRITATAGAAVAFIGLHTSSFTSSLSLRYFT
YGILFGCGCSFAFQPSLVILGHYFQRRLGLANGVVSAGSSIFSMSFPFLI
RMLGDKIKLAQTFQVLSTFMFVLMLLSLTYRPLLPSSQDTPSKRGVRTLH
QRFLAQLRKYFNMRVFRQRTYRIWAFGIAAAALGYFVPYVHLMKYVEEEF
SEIKETWVLLVCIGATSGLGRLVSGHISDSIPGLKKIYLQVLSFLLLGLM
SMMIPLCRDFGGLIVVCLFLGLCDGFFITIMAPIAFELVGPMQASQAIGY
LLGMMALPMIAGPPIAGK
```

Supplementary Table 1: Primer Sequences and PCR conditions

Name	Sequence (5' → 3')	Size (bp)		Taq Polymerase and PCR conditions
		cDNA	gDNA	
P357 / AP1S2_Ex4_F1	TCAGGAAACATCCAAGAAAAATGTCC	335	Out of range	PS; 98°C-30 s, 35 cycles of 98°C-10s, 60°C-10s, 72°C-40s, incubation at 72°C-10 min
P359 / AP1S2_Ex5_R1	AAGGTATCTCTTTCTGCACCATTCTA			
P357 / AP1S2_Ex4_F1	TCAGGAAACATCCAAGAAAAATGTCC	411	Out of range	PS; 98°C-30 s, 35 cycles of 98°C-10s, 60°C-10s, 72°C-40s, incubation at 72°C-10 min
P360 / AP1S2_Ex5_R2	ATATGATGTGCCATTTTCATATGTGC			
P356 / APIS2_Ex1_2_F	CTCAGCGAAGAAACCTCCAATCGGCT	405	2569	κ; 95°C-3 min, 35 cycles of 98°C-10s, 59°C-10s, 72°C-2min 30s, incubation at 72°C-10 min
P344 / APIS2_Ex2_1_R	CTCTTTGTCTGATAGTGGGACATACCAT			
P354 / APIS2_Ex3_F2	TCATCGTTATGTGGAATTACTTGAC	178	178 & 564	PS; 98°C-30 s, 31 cycles of 98°C-10s, 60°C-10s, 72°C-40s, incubation at 72°C-10 min. Primers amplify a retrotransposed pseudogene in gDNA.
P351 / AP1S2_Ex4_R	CTCCTGCAGTAGATCAGCCTGCTC			
P343 / AP1S2_In_4_F	AGACATAAGCTACTGTCTGCAAGTA	1182	Out of range	κ; 95°C-3 min, 37 cycles of 98°C-10s, 59°C-10s, 72°C-1min 30s, incubation at 72°C-10 min
P359 / AP1S2_Ex5_R1	AAGGTATCTCTTTCTGCACCATTCTA			
P339 / APIS2_Ex1_1_F	ACAGCACCACGGCTTCTCTTCCTCA		509	κ; 95°C-3 min, 36 cycles of 98°C-10s, 62°C-10s, 72°C-1min, incubation at 72°C-10 min
P348 / AP1S2_In_1_R	TGGCCACACTCCATCACTGACCAA			
OFD1_Ex14_F1	AAACCTGCGTCTCCTAAACC	91	983	R; 95°C-3 min, 35 cycles of 95°C-30s, 60°C-15s, 72°C-1min, incubation at 72°C-7 min
OFD1_Ex15_R1	CACTATAGCCTTTTCGGCTTTC			
SLC16A2_Ex2_F	CGCGATGGGTATGATCTTCTTC	195	Out of range	R; 95°C-3 min, 30 cycles of 95°C-30s, 60°C-30s, 72°C-30s, incubation at 72°C-7 min
SLC16A2_Ex3_R	TGAAAGGCGAAGGAACAGCC			
SLC16A2_Ex4_F	GGGTGCTCTTGGTGTGTATTG	192	Out of range	R; 95°C-3 min, 30 cycles of 95°C-30s, 60°C-30s, 72°C-30s, incubation at 72°C-7 min
SLC16A2_Ex5_R	CCAGGAAAAGACAGACGACG			
SLC16A2_Ex4_F	GGGTGCTCTTGGTGTGTATTG	204 & 433	Out of range	R; 95°C-3 min, 30 cycles of 95°C-30s, 60°C-30s, 72°C-30s, incubation at 72°C-7 min
SLC16A2_Ex6_R	TGCATCAGAGGGACGAAGAAG			
SLC16A2_Ex5_F2	CCATTGCATTTGAGCTGGTG	246	Out of range	R; 95°C-3 min, 30 cycles of 95°C-30s, 60°C-30s, 72°C-30s, incubation at 72°C-7 min
SLC16A2_Ex6_R2	CCTTGCTGGAATCTCTCTGC			
SLC16A2_Ex5_F1	CTCACAGGCCATTGGCTACC		575	R; 95°C-3 min, 30 cycles of 95°C-30s, 60°C-30s, 72°C-30s, incubation at 72°C-3 min
SLC16A2_Int5_R2	CTGAAAGATGGCAAGTCAACAC			
SVA_E_SINE_F	TAAGTACCCAGGGACACAAACG		452	R; 95°C-3 min, 30 cycles of 95°C-30s, 60°C-30s, 72°C-30s, incubation at 72°C-3 min
SLC16A2_Int5_R2	CTGAAAGATGGCAAGTCAACAC			

Platinum SuperFi DNA Polymerase (PS), KAPA HiFi PCR Kit (κ), Roche Taq DNA polymerase (R)

Supplementary Table 2: Post-natal phenotypes in males with pathogenic *OFD1* variants.

Reference (individual ID)	Sakakibara et al. 2019 (2)	Zhang et al. 2021 (III-2)	Field et al. 2012	Webb et al. 2012	Wentzensen et al 2016
DNA variant*	c.539A>T	c.599T>C	c.689_706del	c.935+706A>G	c.1129+4A>T
Protein change or splicing effect	p.Asp180Val	p.Leu200Pro	p.Ile230_Lys235del	splicing	splicing
Location	Exon 7	Exon 7	Exon 8	Intron 9	Intron 11
# affected males	1	3	4	4	1
Age oldest male	6y	4y	7y	35y	17y
OFC	Macrocephaly	NA	>97th	NA	50th
Obesity	Yes	NA	No	NA	No
Speech delay	NA	NA	Yes >5y	No	Yes (Severe)
Ambulant	Motor developmental delay	Severe delay	>5y	NA	Severe delay
Recurrent infections	NA	NA	No	No	Yes
Polydactyly	No	No	No	No	Yes (all limbs)
Malrotation / situs inversus	No	No	No	No	Yes
CNS anomalies	ventricular dilation	MTS, hypoplastic vermis, macrogyria of right temporal lobe	PMG, MTS, hydrocephalus	No	ACC, MTS, ventriculomegaly
Retinal pathology	Optic nerve hypoplasia	No	No	Childhood RD	Optic atrophy; Severe RP
Nephrolithiasis	Yes	No	Yes (6y)	No	Yes (5y)
Other	Sz, ID	ID, apnea, feeding difficulties	Generalised Sz (4y)		Cleft tongue; Oral hamartoma; Sz

ACC: agenesis of the corpus callosum, ID: intellectual disability, m: months, MTS: Molar tooth sign, NA: Information

* All variant annotations are relative to NM_003611.2

Sharma et al. 2016	Budny et al. 2006	Sakakibara et al. 2019 (3 and 4)	Linpeng et al. 2018	Sakakibara et al. 2019 (1)
c.1654+833_2599+423del	c.2122dupAAGA	c.2260+2T>G	c.2488+2T>C	c.2600-18_2600delinsACCT
deletion exons 16-19		splicing	splicing	p.Ser867_Asp869delinsAsn
Intron 15 to Intron 20	Exon 16	Intron 17	Intron 19	Intron19 / Exon 20
1	11	3	4	1
9y	9	29y	13m	11y
NA	>97th	No	NA	No
Yes	Yes	Yes (1 of 2)	NA	No
Yes (mild)	Yes (severe)	Yes (1 of 2)	NA	No
Yes	Severe delay	NA	NA	Yes
Yes	Yes	NA	NA	NA
No	No	No	Yes (hands)	No
No	No	No	NA	No
No	Hydrocephalus	NA	hypoplastic cerebellum, absent vermis, enlarged ventricles	No
RD	NA	NA	NA	NA
Yes (4y)	No	Yes	NA	Yes
		ID		kidney transplant at 8y

not available, PMG: Polymicrogyria, y: years, RD: Retinal dystrophy, RP: Retinisis pigmentosa, Sz: seizures

Bukowy-Bieryllo, et al. 2019 (855)	**Bukowy-Bieryllo, et al. 2019 (343)**	**Coene et al. 2009 (UW87)**	**Coene et al. 2009 (W07-713)**	**Thauvin-Robinet et al. 2013 (1)**
c.2615_2619del	c.2746insT	c.2767del	c.2844_2850del	c. 2789_2793del
p.Gln872fs*26	p.Tyr916fs*7	p.Glu923Lysfs*4	p.Lys948Asnfs*9	p.Ile930Lysfs*8
Exon 20	Exon 20	Exon 21	Exon 21	Exon 21
1	1	11	8	1
16y	20y	1y	34y	13y
>97th	No	>97th	<3rd	NA
Yes	Yes	Yes	No	Yes
NA	Yes	Yes (severe)	Yes (absent)	NA
NA	Yes	No	NA	Yes
Yes	Yes	No	No	Yes
No	No	Yes (hands and feet)	Yes (all limbs)	Yes (left hand)
Yes	No	No	No	No
No	No	MTS, encephalocele, hydrocephalus	MTS, cerebral atrophy	MTS
No	No	Optic atrophy	Juvenile RD	Juvenile RD
NA	No	No	No	No
mild ID	mild ID			

Hannah et al. 2019 (1)	Zhang et al. 2017	Hannah et al. 2019 (2)	Hannah et al. 2019 (3)	Bukowy-Bieryllo, et al. 2019 (581)
c.2789_2793del	c.2843_2844del	c.2862dupT	c.2868del	c.2797G>T
p.Ile930Lysfs*8	p.Lys948Argfs*7	p.Glu995*	p.Pro957Leufs*2	p.Glu933*
Exon 21	Exon 21	Exon 21	Exon 21	Exon 21
1	1	1	1	1
33y	4m	16	32	16
Macrocephaly	No	Macrocephaly	NA	No
Yes	No	Yes	NA	Yes
NA	NA	Yes (severe)	NA	No
Yes	NA	Minimal	NA	Yes
Yes	Yes	Yes	Yes	Yes
Yes (post-axial)	All limbs	Yes (hands)	NA	No
No	Yes	No	NA	No
Enlarged ventricles, abnormal white matter	MTS	arachnoid cyst, enlarged ventricles, no MTS	NA	No
possible RD	Optic coloboma	NA	No	No
No	NA	Yes	No	No
mild ID, alopecia		Atrial septal defect; Sz		

Bukowy-Bieryllo, et al. 2019 (961)	**This study IV-2**	**This study IV-1**	**This Study II-6**
c.2815G>T	c.1412-322G>A	c.1412-322G>A	c.1412-322G>A
p.Glu939*	splicing	splicing	splicing
Exon 21	Intron 13	Intron 13	Intron 13
1	3	3	3
6	12y	20y	57y
No	97th	97th	75th
No	No	No	No
NA	Yes	NA	NA
NA	Yes >2y	NA	NA
Yes	Yes	Yes	Yes
Yes (all limbs)	No (minor syndactyly)	No	No
No	No	No	No
No	Cerebellar vermis, hypoplasia	Nil	NA
No	Optic coloboma	Optic coloboma	Optic coloboma
No	No	NA	NA
mild ID	Generalised Sz		