The American Journal of Human Genetics

Supplemental Data

TAF1 Variants Are Associated with Dysmorphic Features, Intellectual Disability, and Neurological Manifestations

Jason A. O'Rawe, Yiyang Wu, Max J. Dörfel, Alan F. Rope, P.Y. Billie Au, Jillian S. Parboosingh, Sungjin Moon, Maria Kousi, Konstantina Kosma, Christopher S. Smith, Maria Tzetis, Jane L. Schuette, Robert B. Hufnagel, Carlos E. Prada, Francisco Martinez, Carmen Orellana, Jonathan Crain, Alfonso Caro-Llopis, Silvestre Oltra, Sandra Monfort, Laura T. Jiménez-Barrón, Jeffrey Swensen, Sara Ellingwood, Rosemarie Smith, Han Fang, Sandra Ospina, Sander Stegmann, Nicolette Den Hollander, David Mittelman, Gareth Highnam, Reid Robison, Edward Yang, Laurence Faivre, Agathe Roubertie, Jean-Baptiste Rivière, Kristin G. Monaghan, Kai Wang, Erica E. Davis, Nicholas Katsanis, Vera M. Kalscheuer, Edith H. Wang, Kay Metcalfe, Tjitske Kleefstra, A. Micheil Innes, Sophia Kitsiou-Tzeli, Monica Rosello, Catherine E. Keegan, and Gholson J. Lyon

Supplemental Material and Methods

In general, testing and diagnostic procedures included karyotyping as well as disease-specific gene sequencing in some families to rule out known diseases with related phenotypes, such as methylation analysis for Angelman syndrome [MIM 105830], sequencing *ATRX* [MIM 300032] for X-linked mental retardation, *PTPN11* [MIM 176876], *SOS1* [MIM 182530], *KRAS* [MIM 190070], and *RAF1* [MIM 164760] for Noonan syndrome, *CREBBP* [MIM 600140] and *EP300* [MIM 602700] for Rubinstein-Taybi syndrome [MIM 180849, 613684], *CFTR* [MIM 602421] sequencing and deltaF508 testing for cystic fibrosis [MIM 219700], Cornelia de Lange syndrome-specific [MIM 122470, 300590, 300882, 614701, and 610759] gene sequencing, pyruvate dehydrogenase sequencing, as well as methylation studies for Prader-Willi syndrome [MIM 176270], DNA testing for myotonic dystrophy, DNA testing for fragile X syndrome [MIM 300624] and subtelomeric FISH studies. X-chromosome skewing assays[1] were also performed in Family 1.

In addition to this initial genotyping, echocardiography, brain MRI and CT imaging was performed in some families. Images of the spine were also obtained using MRI in those families indicated in the text. Metabolic and amino acid levels were investigated, which included plasma amino acids, mucopolysacharides, lysosomal enzymes, long fatty acids, A-glycosidase activity, serum amino acid levels, urine organic acid levels, sweat chloride levels, plasma acylcarnitine profile, and immunoglobulin levels. Urine mucopolysaccharidosis (MPS) screening and thyroid profiling was performed. Cerebrospinal fluid (CSF) was collected, and neurotransmitter metabolites, tetrahydrobiopterin (BH4) and neopterin (N) profiles were screened.

Whole genome sequencing

Whole genome sequencing was used to genotype 10 members from Family 1. The parents and two affected children were sequenced with two sequencing platforms, Complete Genomics (CG) and Illumina HiSeq 2000, and 6 other related members were sequenced using the Illumina HiSeq 2000 platform. CG WGS was optimized to cover 90% of the exome with 20 or more reads and 85% of the genome with 20 or more reads. Reads were mapped to the Genome Reference Consortium assembly GRCh37. Due to the proprietary data formats, all the sequencing data QC, alignment and variant calling were performed by CG as part of their sequencing service, using their version 2.0 pipeline. Illumina sequencing libraries were generated from 100 ng of genomic DNA using the Illumina TruSeq Nano LT kit, according to manufacturer recommendations. Illumina WGS (100 bp paired-end reads) resulted in an average mapped read depth coverage of 37.8X (SD=1.3X). >90% of the genome was covered by 30 reads or more and >80% of the bases had a quality score of >30. llumina reads were mapped to the hg19 reference genome using BWA v0.6.2-r126, and variant detection was performed using the GATK v. 2.8-1-g932cd3a. Illumina WGS resulted in an average mapped read depth coverage of 37.8X (SD=1.3X). A second analytical pipeline was used to map the Illumina reads and detect variants using novoalign v3.00.04 and the FreeBayes caller v9.9.2-43-ga97dbf8. Additional variant discovery procedures included Scalpel v0.1.1 for insertion or deletion (INDEL) detection[2; 3], RepeatSeq v0.8.2 for variant detection in short tandem repeat regions, and the ERDS (estimation by read depth) method v1.06.04 and PennCNV (2011Jun16 version) for detecting larger copy number variants (CNVs)[4]. Several methods were used to prioritize and identify possible disease-

contributory germ-line variants, including VAAST [5-8], Golden Helix SVS v8.1.4[9], ANNOVAR (2013Aug23 version)[10], and GEMINI v0.9.1[11].

Whole exome sequencing

Clinical whole exome sequencing (WES) was used for families 2 and 3, and WES sequencing was performed on a research basis for families 5, 6, 7, and 8. Clinical whole exome sequencing (XomeDx) of 3 members of Family 2 (a "trio" pedigree structure) was performed at GeneDx (Gaithersburg, MD). Exonic regions of genomic DNA from the affected individual and both parents were targeted using the Agilent SureSelect XT2 All Exon V4 kit, and sequenced using the Illumina HiSeq2000 sequencing system with 100bp paired-end reads. Raw sequence data was mapped to and analyzed in comparison with the published human genome build UCSC hg19 reference sequence. The targeted coding exons and splice junctions of the known protein-coding RefSeq genes were assessed for an average depth of coverage of 184X and a quality threshold of 99%. Family-based WES was used to genotype Family 3 in a diagnostic setting, using previously described techniques [12]. Briefly, WES was performed using a SOLiD 5500XL machine (Life Technologies) after enrichment with the Agilent SureSelectXT Human All Exon 50Mb Kit (Agilent, Santa Clara, CA, USA). The data were analyzed using LifeScope (Applied Biosystems, Life Technologies, Paisley, UK) software. All clinically relevant candidate *de novo* mutations were validated using Sanger sequencing, and subsequently tested for absence in parental DNA samples. WES was performed on the family 6 trio as part of an exome sequencing project. WES was used to genotype families 7 and 8 on a research basis using previously described methods [13 14]. WES was also used on a research basis to genotype Family 5. Genomic DNA from the affected individual and his parents was enriched for exonic sequences using the Agilent SureSelect V5 exon enrichment kit (Agilent technologies, Santa Clara, USA) prior to sequencing on a 5500xl SOLiD sequencer (Life Technologies, Carlsbad, USA). Sequencing achieved a median coverage of 102X in targeted regions. Bioinformatics was performed by the ACHRI genomics and bioinformatics facility using an in-house pipeline. In brief, reads were aligned to the human reference genome (GRCh37), PCR duplicate reads were removed using Picard and SAMTools and SNVs were called using FreeBayes and DiBayes. INDELs were called using Atlas2, and CNVs were called using Exome CNV.

Gene panel sequencing

Targeted gene enrichment and sequencing was used to genotype the parents and the affected proband from Family 9 using a custom SureSelect oligonucleotide probe library designed to capture 19,878 coding exons of 614 pathogenic and 642 candidate genes associated with intellectual disability. The design includes all the transcripts reported for each target gene in different databases (RefSeq, Ensembl, CCDS, Gencode, VEGA). The SureSelect DNA Standard Design Wizard (Agilent Technologies) was used for probe design with a 2X tiling density and a moderately stringent masking. A total of 71,994 probes, covering 5.073 Mbp (99.48% coverage of targets), were synthesized by Agilent Technologies (Santa Clara, CA, USA). Sequence capture, enrichment, and elution were performed according to the manufacturer's instructions. The libraries were sequenced on an IlluminaHiSeq 2000 platform with a paired-end run of 2 × 90 bp, following the manufacturer's protocol to generate at least a 100X effective mean depth. Variant calling was performed with the DNAnexus platform (DNAnexus, Mountain View, CA, USA) through the following pipeline: Fastq paired reads were aligned to the reference human genome UCSC hg19 using the BWA-MEM algorithm from the BWA software package. Duplicated reads were removed using Picard, realigned around sites of known indels using the GATK indel realigner, and their quality was recalibrated by looking at covariance in quality

metrics with frequently observed variation in the genome. After recalibration, variants were called with the GATK Unified Genotyper module. This pipeline follows the Broad Institute's recommendations for best practices in variant calling. Variants on regions with low mappability or variants in which there was not at least one sample with read depth ≥10 were filtered out. Annotation of nucleotide variants was performed by the Ion ReporterTM Software (Life Technologies). To evaluate the putative clinical impact of the variants, the following criteria were applied:1) an allele frequency <0.01 in the 1000G Phase 1 or EVS databases 2) stop gain, frameshift and splicing variants were a priori considered as most likely to pathogenic; 3) for missense mutations, amino acid conservation and prediction of pathogenicity (SIFT, Polyphen-2 and Grantham); 4) a de novo occurrence (dominant inheritance), the presence of two mutant alleles in the same gene, each from a different parent (recessive inheritance), or maternal inheritance of X-linked variants; 5) the absence of the variant in other samples (in-house database); 6) phenotypic consistency with clinical signs associated to mutations in the same gene when available. To evaluate the possible effect of synonymous or intronic variants in gene splicing we used the Human Splicer Finding web tool. Relevant variants were re-sequenced by Sanger sequencing.

Microarray genotyping

Microarray genotyping was also used to genotype Families 1, 9, 10, and 11. The mother, father and affected child from Family 9 was microarray genotyped using a custom array-CGH microarray (Agilent technology 8x60K) for the purposes of investigating large CNVs, including 453 candidate and pathological genes involved in neurodevelopmental disorders. The average distance between genomic probes was 160Kb. Microarray genotyping for the affected patients from families 10 and 11 was performed using the Agilent Human Genome CGH 4x180K (Sureprint G3 arrays) microarrays (Agilent Technologies, Santa Clara, CA, www.agilent.com). The average spatial resolution for the 180K platform is 13-25Kb. Labeling, hybridization and data processing was carried out according to manufacturer's recommendations and, as previously described[15]. The available parental DNA samples were processed in the same manner for both families. Subsequently, due to the X-chromosome finding, we designed a custom X-chromosome array (60K) using the Agilent SureDesign e-array platform and re-tested the affected individual from Family 10 as well as other members of his extended family (including his healthy brother, mother, father, maternal grandmother and affected maternal cousin) following the same methodology as previously described[22]. The custom X-chromosome array has median probe spacing for the X-chromosome 6.14Kb. Array genotyping was also performed in parallel as part of Illumina WGS and analysis for Family 1 on a research basis, with all 10 members genotyped using Illumina Omni 2.5 DNA microarrays (which contain approximately 2.5 million markers.

Supplemental Note: Case Reports

Family 1

The probands include two affected brothers, ages 12- and 14-years-old respectively, with severe ID, autistic behaviors, anxiety, attention deficit hyperactivity disorder, and very distinctive facial features (**Figure 1-1A/1B, S5, S6, Video S1-S5**). Among the facial features are a broad nasal bridge, sagging cheeks, downward sloping palpebral fissures, prominent supraorbital ridges, deep-set eyes, relative ocular hypertelorism, thin upper lip, a high-arched palate, prominent low-set ears with thickened helices, and a pointed chin. Other shared phenotypic symptoms include strabismus (exotropia), blocked tear ducts, microcephaly, oculomotor dysfunction, frequent otitis media with effusion, hearing impairments (mixed conductive/sensorineural), oral motor

dysphagia, kyphosis, a peculiar gluteal crease with sacral caudal remnant (without any spinal abnormalities), dysplastic toenails, hyperextensible joints (especially fingers and wrists), spasticity, ataxia, gait abnormalities, growth retardation and global developmental delays, especially in the areas of gross motor and verbal expression. The younger of the two affected siblings also suffers from orthopedic problems. He is reported to have scoliosis, dysplasia of his hips, genu valgum ("knock knees"), and overlapping toes. He also has frequent episodes of contact dermatitis, and eczema, sleep-wake dysregulation, and mild asthma documented in his record. The elder brother has spastic diplegia, and has received botulinum toxin (Botox) therapy for his lower-extremity hypertonicity for six years. A review of systems (ROS) questionnaire revealed no other obvious, shared or otherwise, symptoms or malformations.

The parents of the two affected siblings are non-consanguineous and healthy. The mother has been evaluated for PKU and has normal plasma amino acid levels. The family history does not reveal any members, living or deceased, with phenotypic or syndromic characteristics that resemble the described syndrome, and there is a male cousin who is unaffected. An assay performed on leukocyte DNA revealed that the mother has significantly skewed (99:1) X-chromosome inactivation. The maternal grandmother and aunt of the affected boys did not show any appreciable X-chromosome skewing (**Figure S5**), which suggested the possibility of a *de novo* deleterious X-chromosome mutation in the mother. However, skewed X-inactivation can also result from stochastic processes[16].

Both pregnancies with these male fetuses were complicated by placenta deterioration, and both affected siblings were diagnosed with intra-uterine growth retardation (IUGR) and were eventually delivered through Caesarean section (C-section). The mother denied alcohol, drug use, and exposure to environmental toxins during the course of either pregnancy. The elder boy was born in the 40th gestational week with a birth weight of 2.21 kg and a notable birth defect of aplasia cutis congenita, which was surgically corrected at the age of 4 days. The younger boy was born in the 37th gestational week with a birth weight of 1.76 kg. A heart murmur was noticed at birth, but echocardiography confirmed the absence of additional cardiovascular abnormalities. He was treated with light for neonatal jaundice, and required a feeding tube during the first few days of life due to difficulties swallowing and digesting food. During the most recent examinations, the younger boy (aged $10_{11/12}$ years) had a height of 129.7 cm (2% tile), a weight of 30.8 kg (19% tile, BMI 18.3 kg/m^2), and his occipital frontal circumference (OFC) was 51 cm (4.5th percentile); while his elder brother (aged $11_{11/12}$ years) had a height of 136.8 cm (5% tile), a weight of 26.3 kg (0% tile, BMI 14.1 kg/m^2), and his OFC was 49.5 cm (0.2th percentile) at the time.

Brain MRIs of the two brothers (**Figure S7**) demonstrated a remarkably similar constellation of abnormalities. In both subjects, there was hypoplasia of the isthmus and splenium of the corpus callosum with thickness falling below the third percentile reported for individuals of the same age[17]. There was mild lateral ventriculomegaly implying low cerebral volume given the microcephaly. There was also deficiency of the septum pellucidum in both brothers, with the older brother having absence of the posterior two-thirds of the septum pellucidum and the younger brother having complete absence of the septal leaflets. Other findings associated with septooptic dysplasia included underdeveloped pituitary glands for age, deficiency of the anterior falx with mild hemispheric interdigitation, and question of small olfactory bulbs despite fully formed olfactory sulci. However, the optic nerves appeared grossly normal in size. Finally, there was subjective vermian hypoplasia with the inferior vermis resting at the level of the pontomedullary junction rather than a more typical lower half of the medulla. Pertinent negatives included absence of a malformation of cortical development, evidence of prior injury, or conventional imaging evidence of a metabolic/neurodegenerative process.

Other clinical diagnostic testing performed on both affected siblings did not reveal any known disorders. Although chromosomal analysis revealed that both boys have the karyotype of 46,X,inv(Y)(p11.2q11.2), this is known to be a normal population variant.

A comprehensive set of putative variants was established from among three disease models; X-linked, autosomal recessive, and autosomal dominant inheritance. We found seven potentially important variants in non-coding regions of the genome and seven in coding regions (**Table S1**). These variants fall within coding or non-coding regions of eight known genes: *TAF1* [MIM 313650], *FAM47B*, *SLC28A1* [MIM 606207], *FRG2B*, *DMRTB1* [MIM 614805], *ZNF423* [MIM 604557), *SNAPC5* [MIM 605979] and *KCHN1*. Our analysis of genotyping arrays and the WGS data also led to the identification of a number of CNVs that are not known to be associated with any phenotype (**Table S3**). Only one of the above variants was robustly identified between the different prioritization analysis schemes applied to the WGS data, namely a non-synonymous change in *TAF1* that resulted in an isoleucine (hydrophobic) to threonine (polar) change, p.Ile1337Thr (NP_001273003.1). This variant was ranked highest among those tested with VAAST using an X-linked model (p-value of 0.00184; rank of 14.59), and it was ranked by CADD as being within the top 1% most deleterious variants in the human genome with a score of 27. In addition, it was categorized by PolyPhen-2[18] as being "Probably Damaging" with a score of 0.996, by SIFT[19] as "Damaging" with a score of 0.003, by GERP++ [20] with a score of 5.6, by phyloP[21] as "Deleterious" with a score of 7.695.

Additional sequencing information for family 1

Complete Genomics whole genome sequencing and variant detection for Family 1

After quality control to ensure lack of genomic degradation, we sent 10 μg DNA of each sample to Complete Genomics (CG) at Mountain View, California for sequencing. The whole-genome DNA was sequenced with a nanoarray-based short-read sequencing-by-ligation technology, including an adaptation of the pairwise end-sequencing strategy. Reads were mapped to the Genome Reference Consortium assembly GRCh37. Due to the proprietary data formats, all the sequencing data QC, alignment and variant calling were performed by CG as part of their sequencing service, using their version 2.0 pipeline. Complete Genomics WGS was optimized to cover 90% of the exome with 20 or more reads and 85% of the genome with 20 or more reads.

Illumina HiSeq 2000 whole genome sequencing and variant detection for Family 1

After the samples were quantified using Qubit® dsDNA BR Assay Kit (Invitrogen), 1 μg of each sample was sent out for whole genome sequencing using the Illumina® Hiseq 2000 platform. Sequencing libraries were generated from 100 ng of genomic DNA using the Illumina TruSeq Nano LT kit, according to manufacturer recommendations. The quality of each library was evaluated with the Agilent bioanalyzer high sensitivity assay (less than 5% primer dimers), and quantified by qPCR (Kappa Biosystem, CT). The pooled library was sequenced in three lanes of a HiSeq2000 paired end 100 bp flow cell. The number of clusters passing initial filtering was above 80%, and the number of bases at or above Q30 was above 85%. Illumina reads were mapped to the hg19 reference genome using BWA v0.6.2-r126, and variant detection was performed using the GATK v. 2.8-1-g932cd3a. Illumina WGS resulted in an average mapped read depth coverage of 37.8X (SD=1.3X). >90% of the genome was covered by 30 reads or more and >80% of the bases had a quality score of >30. A second analytical pipeline was used to map the Illumina reads and detect variants

using novoalign v3.00.04 and the FreeBayes caller v9.9.2-43-ga97dbf8. Additional variant discovery procedures included Scalpel v0.1.1 for insertion or deletion (INDEL) detection, RepeatSeq v0.8.2 for variant detection in short tandem repeat regions, and the ERDS (estimation by read depth) method v1.06.04 and PennCNV (2011Jun16 version) for detecting larger copy number variants (CNVs).

Variant prioritization for Family 1

We used several methods to prioritize and identify possible disease-contributory germ-line variants, including VAAST [5-8], Golden Helix SVS v8.1.4[9], ANNOVAR (2013Aug23 version)[10], and GEMINI v0.9.1[11]. VAAST employs a likelihood-based statistical framework for identifying the most likely disease-contributory variants given genomic makeup and population specific genomic information. SVS, ANNOVAR and GEMINI employ more traditional annotation and filtering-based techniques that leverage data stored in public genomic databases (i.e., dbSNP 137, 1000 Genomes phase 1 data, NHLBI 6500 exomes, etc.).

More detailed methodology for the analysis of Family 1 is available below. The Illumina data were also re-analyzed in the recent development of SeqHBase, a big data-based toolset for analyzing family based sequencing data to detect de novo, inherited homozygous, or compound heterozygous mutations that may contribute to disease manifestations[22].

Sanger Sequencing

PCR primers were designed using Primer 3 (http://primer3.sourceforge.net) to produce amplicons of around 700 bp in size, with variants of interest located approximately in the center of each amplicon. Primers were obtained from Sigma-Aldrich®. Upon arrival, all primers were tested for PCR efficiency using a HAPMAP DNA sample (Catalog ID NA12864, Coriell Institute for Medical Research, USA) and LongAmp® Taq DNA Polymerase (New England Biolabs, USA). PCR products were visually inspected for amplification efficiency using agarose gel electrophoresis. PCR products were further purified using QIAquick PCR Purification Kit (QIAGEN Inc., USA), quantified by Qubit® dsDNA BR Assay Kit (Invitrogen Corp., USA), and diluted to 5 - 10 ng/µl in water for Sanger sequencing using the ABI 3700 sequencer. The resulting *.ab1 files were loaded into the CodonCode Aligner V4.0.4 for analysis. All sequence traces were manually reviewed to ensure the reliability of the genotype calls (see **Figure S1**).

DNA Microarrays

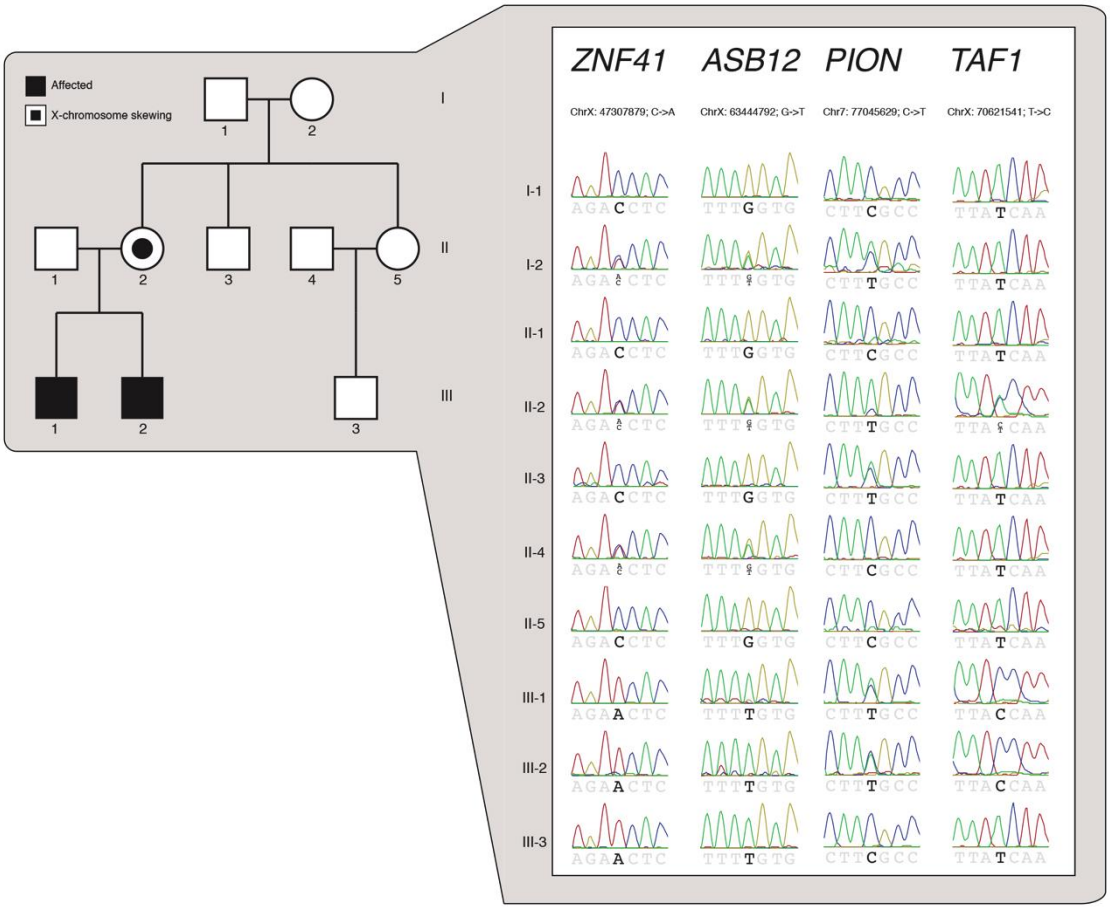


Figure S1. Sanger sequencing traces for all 10 family members for variants found in *ZNF41*, *ASB12*, *PION* and *TAF1*.

DNA samples were genotyped on Illumina Omni 2.5 DNA microarrays (which contain approximately 2.5 million markers). Total genomic DNA extracted from whole blood was used in the experiments. Standard data-normalization procedures and canonical genotype-clustering files provided by Illumina were used to process the genotyping signals.

Variant detection

Human sequence variation ranges in manifestation from differences that can be detected at the single nucleotide level, to whole chromosome differences. In our study, we used a number of bioinformatics software packages to extract signals for differences seen at the levels of single nucleotide variants (SNV), small insertions/deletions (INDELs), variants in short tandem repeat structure (STRs), and variants in copy number (CNVs) (see **Figure S2 and S3** for a general map of the analyses performed). When possible, we used more than a single bioinformatics software package to detect different classes of genetic variants, so as to arrive at a comprehensive and high-quality set of variants for each person sequenced. Standard data quality filtering approaches were used for all genetic variants detected by the various different methods. This includes, when appropriate, requiring sequencing to be at a depth of 10 or more reads at the location of a sequence variant, and a variant phred quality score of 30 or above. Specific variant detection parameters, which themselves detail internal or pipeline specific variant detection thresholds, are described below, in the Supplemental document or in the documentation of the software which has been described in detail elsewhere. We expect variants where all pipelines agree to be more reliable in

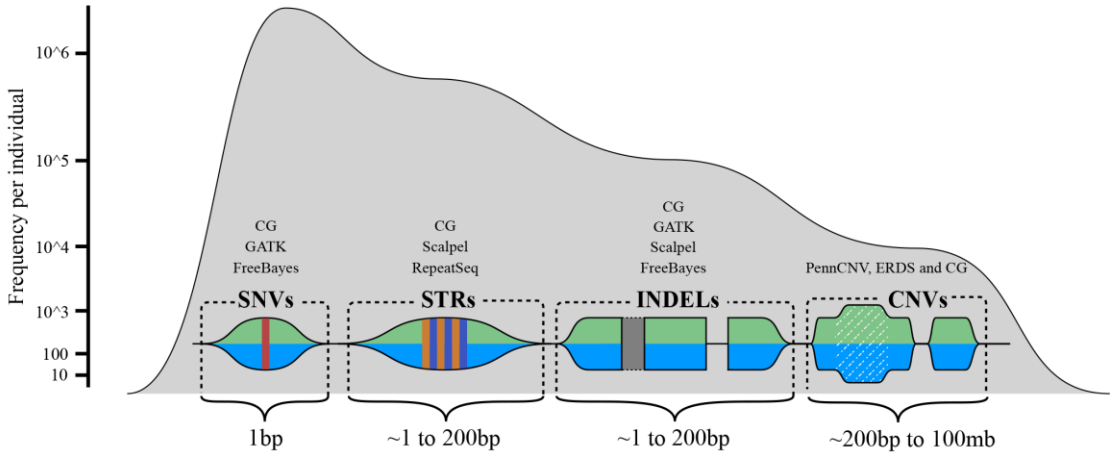


Figure S2. A conceptual map of human sequence variation. Here, we show approximate sizes, as well as the associated signature, of the various different types of human sequence variation that can be currently detected with the WGS and informatics technologies employed in this work. The frequency axis shows the approximate frequency of the various genetic variation types that currently detectable via germline WGS. Above the visual signatures of the different types of human sequence variation, the general names of the different informatics software tools for detecting the variation are noted which include, the Genome Analysis Took Kit (GATK), Scalpel, RepeatSeq PennCNV, the estimation by read depth with single-nucleotide variants (ERDS) CNV caller and the FreeBayes caller. We do not differentiate here by raw sequencing data generated by different sequencing technologies, but its important to note that Complete Genomics (CG) is listed here as a software tool but in actuality what we are referring to is the CG sequencing technology as well as its own proprietary sequence analysis.

terms of their validation rate, whereas those variants that were unique to single pipelines will likely have lower yet potentially vastly different validation rates. This expectation has been shown to be true in our previous studies that have used high-throughput MiSeq validation methods[23]. This information was carried through to the various stages of the variant prioritization and functional annotation stages of the study. If a variant was annotated as being highly deleterious by functional annotation or by frequency inference, sequence error was more easily identified by first checking how many detection pipelines found it. In contrast, if a variant was detected by one sequencing *platform* and not the other, sequence depth and quality variation between platforms contributed to these instances and were not as easily dismissed as errors.

SNVs and INDELs

bwa-GATK

Illumina reads were mapped to the hg19 reference genome using BWA v. 0.7.5a using default 'mem' parameters. BWA was directed to mark shorter split hits as secondary, so as to make the output compatible with Picard and the Genome Analysis Took Kit (GATK). BWA sequence alignments were converted into binary format using SAMtools v0.1.19-44428cd, and duplicate reads were marked using Picard tools v1.84. GATK 2.8-1-g932cd3a was used to realign the reads around putative INDELs, and base quality scores were then recalibrated. Variants were detected using the GATK HaplotypeCaller, and variant quality scores were then recalibrated using the GATK variant quality score recalibration (VQSR) protocol. The GATK HaplotypeCaller[24] works by generating a reference graph assembly, which starts out as a directed DeBruijn graph. The GATK HaplotypeCaller then tries to match each sequence read to a path in the reference graph, this is called the 'read-threading' graph. The graph is then pruned by removing sections of the graph that are supported by

fewer than 2 reads, which are considered to be the result of stochastic errors. Haplotype sequences are then constructed using likelihood scores for each path in the graph. A Smith-Waterman alignment of each haplotype to the original reference sequence is used to generate potential variant calls, which are then modeled using a genotype likelihood framework.

novoalign-FreeBayes

SNP and INDELs were also detected with a novoalign-FreeBayes pipeline, using novoalign v3.00.04 to map reads to the hg19 reference genome, and the FreeBayes caller v9.9.2-43-ga97dbf8 to detect variants. Novoalign was used to map the first 50,000 reads to the reference sequence in order to determine the empirical insert size for the Illumina paired end reads. Once the insert size was determined, novoalign was then used to map all of the Illumina reads to the reference sequence using default parameters. Sequence alignment output from novoalign was used to generate variant calls using FreeBayes with default parameters. FreeBayes uses a Bayesian genotype likelihood approach, but generalizes its use to perform over haplotype sequences, which is in contrast to precise alignment based implementations.

bwa-Scalpel

Sequence alignments obtained from the above 'bwa-GATK' pipeline were used in conjunction with Scalpel v0.1.1[25] to extract INDELS from the WGS data. Scalpel was run with near-default parameterizations in 'single' mode. The minimum coverage threshold and minimum coverage ratio for emitting a variant was set to 3 and 0.1 respectively, and the threshold at which low coverage nodes were removed was set to 1. Scalpel uses both sequence mapping and assembly to detect INDELs. First, Scalpel extracts aligned sequence reads to construct a de Bruijn graph. Low coverage nodes and sequencing errors are then removed and a repeat analysis of the each region is performed to tune the k-mer size. Assembled sequences are then aligned back to the reference genome where a standard Smith-Waterman-Gotoh alignment algorithm with affine gap penalties is used to detect candidate variants.

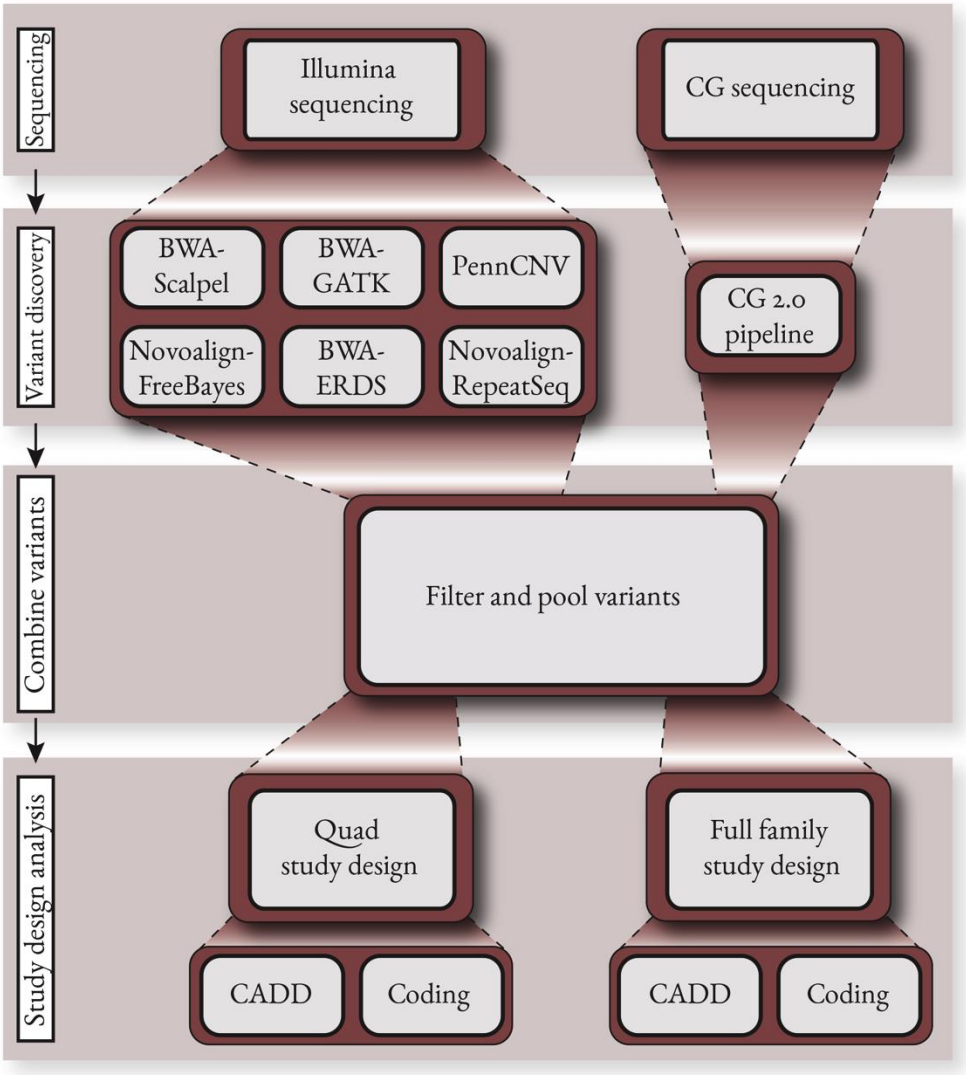


Figure S3. A generalized map of the flow of work performed during the course of the study for Family 1. Briefly, the family was sequenced using two different sequencing technologies, the Illumina (which includes WGS on the HiSeq 2000 and genotyping array data from Illumina Omni 2.5 microarrays) and the Complete Genomics (CG) sequencing platforms. Raw sequencing data resulting from the CG sequencing was processed using the internal CG informatics pipeline v. 2.0. Six variant discovery pipelines analyzed raw sequencing data resulting from the Illumina-based sequencing. Variants resulting from all of the post-sequencing data analysis and variant discovery pipelines were filtered using standard filtering methods/thresholds (see Methods section) and pooled for further post-variant discovery analyses. From the pooled variant data set, two family-based study designs were performed, a quad based study design and a study design that incorporates data from all of the sequenced members of the family. For these two studies, two variant prioritization strategies were employed, 'CADD' and 'Coding'. Both prioritization schemes required variants to be in low population frequencies (MAF < 1%), but the CADD strategy further required variants to have a CADD score of greater than 20 whereas the coding scheme required variants to be coding and non-synonymous.

STR profiling

RepeatSeq and Scalpel

We used RepeatSeq[26] v0.8.2 to extract variants in short tandem repeats across the genome

using default settings. RepeatSeq uses a Bayesian model selection approach to assign the most probable genotype using information about the full length of the sequence repeat, the repetitive unit size and the average base quality of mapped reads. Scalpel has been shown to perform well in terms of detecting variants in short tandem repeat regions. For this reason, we also consider Scalpel to be a good informatics pipeline for use in profiling STR regions. Scalpel was used to detect sequence variants in STRs using the same methods described in the section detailing pipelines on SNVs and INDELs (above).

CNVs

Bedtools v2.17.0[27] was used to compare CNVs . CNVs were required to overlap reciprocally by 90%. Hypervariable and invariant CG CNVs were excluded from the analysis, and CG CNVs were required to have a 'CNVTypeScore' of greater than 30.

PennCNV

The PennCNV[28] software package (2011Jun16 version) was used to perform Copy Number Variant CNV calling using the Illumina Omni 2.5 microarray data for all the samples. For kilobase-resolution detection of CNVs, PennCNV uses an algorithm that implements a hidden Markov model, which integrates multiple signal patterns across the genome and uses the distance between neighboring SNPs and the allele frequency of SNPs. The two signal patterns that it uses are the Log R Ratio (LRR), which is a normalized measure of the total signal intensity for two alleles of the SNP and the B Allele Frequency (BAF), a normalized measure of the allelic intensity ratio of two alleles. The combination of both signal patterns is then used to infer copy number changes in the genome.

Microarrays often show variation in hybridization intensity (genomic waves), that is related to the genomic position of the clones, and that correlates to GC content among the genomic features considered. For adjustment of such genomic waves in signal intensities, the cal_gc_snp.pl PennCNV program was used to generate a GC model that considered the GC content surrounding each Illumina Omni2.5 marker within 500kb on each side (1Mb total). The detect_cnv.pl program in the mode --test for individual CNV calling was used, the Hidden Markov Model used is contained in the hhall.hmm file provided by the latest PennCNV package, and custom Population Frequency of B allele (PFB) file for all the SNPs in the Illumina Omni2.5 array was generated from 600 controls (which consists of 600 unaffected parents from the Simons Simplex Collection), the GC model described above was also used during CNV calling. Chromosome X CNVs were called separately using the -- test mode with the --chrx option. We excluded CNVs with an inter-marker distance of >50kb and required each CNV to be supported by at least 10 markers.

ERDS

To detect CNVs from the Illumina WGS data, the Estimation by Read Depth with SNVs (ERDS) v1.06.04[29] method was employed, using default pipeline parameterization. ERDS uses WGS read depth information contained within sequence alignment files, along with soft-clip signatures, to detect CNVs. CNVs that were detected by the ERDS method were filtered to include CNVs that were greater than 200 kilobases in scale and CNVs called with a confident score of greater than 300.

Disease variant prioritization, post variant discovery analyses

We performed analyses to prioritize sequence variants conforming to three disease model pathways: de-novo, autosomal recessive and x-linked models of transmission. X-chromosome skewing in the mother of the two affected boys suggests that genetic components of the disease phenotype are most likely segregating and following an X-linked mode of inheritance. Recent work illustrates the existence of a substantial amount of complexity in elucidating genetic factors of human disease, with many syndromes likely being the result of an array of different genetic aberrations in conjunction with environmental effects and modification of gene/variant function by ancestral background[30-32]. There are also many still uncharacterized noncoding regions of the genome [33], along with continuous re-annotation of protein coding portions [34]. In light of this complexity, we sought to identify variants following de-novo, autosomal recessive and x-linked models of transmission that may be contributing, together or alone, to the disease phenotype. It is possible that a disease-contributory variant in the germline of a somatically mosaic parent could pass on to both children, appearing as "de-novo" when compared to DNA from the blood of the parents [35-38]. Similarly, variants benign in the heterozygous state might prove deleterious if present in the homozygous state in the two children, so we sought to identify these autosomal recessive variants as well.

In general, to identify de-novo variants, we isolated genetic variants shared by both affected boys. Variants in common with all other healthy people in the family were then filtered out. For X-linked variants, all X-chromosome variants shared by the two affected boys as well as their mother were identified. Then, variants in common with all other healthy males were removed. Autosomal recessive variants were identified by first selecting all heterozygous variants shared by the mother and father of the two affected children. Then, homozygous variants shared by the two affected boys where both parents were heterozygous were selected. Finally, homozygous variants that were also present in any other healthy family member were removed.

SVS, ANNOVAR and GEMINI

Recent work has identified differences between results generated by available annotation software packages, which can, in part, be the result of differences in choice of transcript by the user[39]. To capture and analyze this variability, we used three annotation and filtering software packages with similar databases to filter and prioritize disease variants. For variants conforming to each disease model, ANNOVAR, SVS and GEMINI were used to filter and annotate them.

To be consistent among the different software tools, we used the same filtering strategy for each of the three software packages. Depending on the analysis, filtering criteria required each variant to be characterized by a population frequency of less than 1 percent in the available variant frequency databases (this includes genotype frequencies derived from the 1000 genomes project as well as genotype frequencies reported in dbSNP 138 and the Exome Sequence Project, which includes genotype frequencies, among other information, on 6500 individuals of various recently derived ancestral lineages), a CADD (Combined Annotation Dependent Depletion) score of greater than 20 (top 1% of all possible human genomic variants in terms of deleteriousness) or be either a non-synonymous or a splice site variant. Variants that passed these filters were then annotated by gene and variant type using UCSC's Known Genes table for annotation.

VAAST

VAAST v2.0 was used to identify variants that are likely contributing to disease[40; 41]. SNPs and INDELs were converted into the GVF file format using the vast_converter tool, annotated using the the VAAST annotation tool (VST) and then converted into a condenser file (*.cdr). VAAST was run in CLRT mode without grouping variants when they are located within the same feature. Amino acid substation frequencies were included in the likelihood ratio test when scoring variants, and the maximum expected frequency of the 'causal' allele in the background population was set to 0.01. 10,000 permutations were performed. The VAAST background file that was used contains 1057 "1000 genome project" genomes, 54 Complete Genomics genomes, 184 genomes from Danish exomes, and 9 genomes from 10 Gen data[42]. Variants from dbSNP and NHLBI ESP that have a sample size >= 100 were randomly spiked into the dataset based on their allele frequencies. Coding variants only within CDS regions of the RefSeq gene set with 10 nts around each exon splice regions". The background file used in this study is public and freely available for download on the VAAST website (http://www.yandell-lab.org/software/VAAST).

Cluster analysis

Nonsynonymous *TAF1* hemi or homozygote variants from European and Latino populations were taken from the ExAC database. Unique genomic positions were collected. Following Cucala 2008[43], these locations were set such that

$$0 = X_{(0)} \leq X_{(1)} \leq \ldots \leq X_n \leq X_{(n+1)} = 1$$

We then compute all $(j - i)$ ordered spacing's such that

$$D_{i,j} = X_{(j)} - X_{(i)} \sum_{k=i+1}^{j} D_k, \qquad 1 \leq i < j \leq n$$

We also let

$$U_{i,} = B_{inc}\left(D_{i,j}, j - i, n + 1 - j + i\right), \qquad 1 \leq i < j \leq n,$$

And compute the hypothesis-free scan statistic, which is

$$\Lambda_{HF} = \sup_{1 \leq i < j \leq n} 1/U_{i,j}$$

p-values are computed via a Monte Carlo procedure. To identify more than one cluster, a multiple procedure is introduced. If there exists an interval that represents a significant cluster in the initial search, which Cacala 2008 and we note here as$[X_{(i^*)}, X_{(j^*)}]$, then let $T^* = 1 - X_{(j^*)} + X_{(i^*)}$. We will then transform the data such that

$$X_k^{(2)} = \begin{cases} \dfrac{X_{(k)}}{T^*} & if \qquad 1 \leq k \leq i^*, \\ \dfrac{X_{(k+j^*-i^*)} - X_{(j^*)} + X_{(i^*)}}{T^*} & if \qquad i^* + 1 \leq k \leq n - j^* + i^* \end{cases}$$

and test for clusters as described above.

RNA sequencing and analysis for family 1

We conducted RNA sequencing with RNA isolated from blood from Family 1. Blood was collected from the two probands, their parents and the maternal grandparents. Except for the single blood draws for the grandparents, two blood draws were taken on separate days for each subject. The blood was collected in PAXgene Blood RNA tubes and the RNA was isolated with the PAXgene Blood RNA kit (QIAGEN) according to the manufacturer's recommendations. The final pooled library was measured by qPCR using the KAPA SYBR® Fast Universal qPCR kit (Kapa Biosystem, Wilmington, MA) and sequenced on a HiSeq 2000 across three lanes (paired-end 100bp). A mean of 49,792,652 (sd = 11,666,119) properly paired reads were generated and a mean spliced mapping percentage of 85.41 (sd = 7.1) per sample was observed (**Table S6**). HISAT[44] was used for spliced alignment to the UCSC human reference sequence hg19 using 10 alignment threads and the --rna-strandness RF flag for stranded libraries. Stringtie[45] was used to quantify transcripts using the UCSC hg19 transcript annotations, with estimates of abundances being restricted to these annotated transcripts. Cuffdiff[46] was used to perform differential expression analysis using 32 threads with the --library-type fr-firststrand flag for our stranded libraries. The R package CummeRbund[47] was used to analyze, filter and visualize the results from the Cuffdiff differential expression analysis. WebGestalt[48] was used to perform gene set enrichment analyses using Molecular Signatures Database (MSigDB) for Transcription Factor Targets[49], KEGG[50], GO[51], and HPO[52] databases for gene annotations and set inclusion information. We used various analysis tools in the CummeRbund package[47] to evaluate the quality of our RNA sequencing data/analysis for Family 1 (**Figure S10**).

Family 2

A second family was recruited from Michigan, USA, after referral from GeneDx, which as of August 10, 2015, had reported out 4864 clinical exomes, of which 2706 of these had been referred with phenotypes involving development delay and/or intellectual disability. This family consists of one affected male child born to healthy non-consanguineous parents of mixed European descent. The family history was significant only for a paternal first cousin with a seizure disorder and mild developmental delay. The family has since had a second unaffected male child. The affected individual was initially evaluated by the pediatric genetics service as an infant following transfer to University of Michigan Health System (UMHS) for evaluation and management of multiple congenital anomalies and surgical repair of his congenital heart defect. The pregnancy was generally uncomplicated with the exception of a possible AVSD identified by ultrasound prenatally. Amniocentesis revealed a 46, XY karyotype. The proband was born by Cesarean section due to failure to progress. The affected individual is now 5 years old and shares remarkable similarity to the two brothers in Family 1 (Utah family) (**Figure 1-2A**), including dysmorphic facial features, such as a flat occiput, short palpebral fissures, low-set and posteriorly rotated ears, bulbous nasal tip, a highly arched palate, and mildly short digits, and a peculiar gluteal crease. He also presents with microcephaly, ventricular septal defect, coarctation of the aorta, cryptorchidism, vesicoureteral reflux, bilateral sensorineural hearing loss, nystagmus, myopia, talipes equinovarus and developmental delay. At the age of 5, he walks with assistance, is not toilet trained, and does not have any expressive language. A spinal ultrasound ordered to investigate any tethering of the spinal cord showed that the tip of the conus medullaris appeared normal and was in the normal position in the upper lumbar canal. The visualized low thoracic and lumbar spinal cord and roots of the cauda equina oscillated normally.

Brain MRI showed diffuse thinning (hypoplasia) of the corpus callosum; this hypoplasia was more severe anteriorly in contrast to the Family 1 brothers, where findings were more severe

posteriorly (**Figure S7**). The cerebellar vermis was hypoplastic and to a more severe degree than seen in the Family 1 brothers. There was anterior falx cerebri deficiency, as was seen in the Family 1 brothers. There was top normal size of the lateral ventricles (implying low cerebral volume given microcephaly). However, there was no deficiency of the septum pellucidum or obvious pituitary abnormality like in the Family 1 brothers. Unlike any of the other subjects in this study, there were germinolytic cysts at the caudothalamic grooves which can be seen with germinal matrix hemorrhages, in utero infection like CMV, and rarely with some inherited conditions (e.g. peroxisomal disorders).

Previous genetic testing included normal oligonucleotide (EMArray Cyto6000) and SNP (Illumina HumanCyto SNP-12) chromosomal microarrays, normal 7-dehydrocholesterol, normal *PTPN11*, *SOS1*, *KRAS*, and *RAF1* sequencing results for Noonan syndrome, and normal sequencing and deletion duplication studies of *CREBBP* and *EP300*, which are associated with Rubinstein-Taybi syndrome. Newborn screening revealed that he was a carrier for the cystic fibrosis mutation deltaF508. A sweat chloride test was normal, and full *CFTR* sequencing did not identify a second mutation.

Through exome sequencing at GeneDx, a *de novo* mutation, p.Cys807Arg (c.2419T>C) in exon 15 of *TAF1* was identified in the affected individual, and this was the only variant reported by GeneDx. The p.Cys807Arg variant was not observed in approximately 6500 individuals of European and African American ancestry in the NHLBI Exome Sequencing Project. The substitution is within a well-conserved central domain of the TAF1 protein. In silico analysis predicts that this variant is probably damaging to the protein structure/function (PPH2 0.999, MutTaster D, SIFTnew 1, Phylo P 1.66, Alamut Score 0).

Sequencing analysis for Family 2

Clinical whole exome sequencing (XomeDx) of the family trio was performed at GeneDx (Gaithersburg, MD). Exonic regions of genomic DNA from the affected individual and both parents were targeted using the Agilent SureSelect XT2 All Exon V4 kit. The targeted regions were sequenced using the Illumina HiSeq2000 sequencing system with 100bp paired-end reads. The DNA sequence was mapped to and analyzed in comparison with the published human genome build UCSC hg19 reference sequence. The targeted coding exons and splice junctions of the known protein-coding RefSeq genes were assessed for an average depth of coverage of 184X and a quality threshold of 99%.

Family 3

A third family from the Netherlands was identified through the "genotype-first" approach when pictures of an affected individual with a mutation in *TAF1* from an ongoing large sequencing study [12] were compared to those seen in families 1 and 2 (see **Table 1-3A**). This is a *de novo* missense change in *TAF1* in a boy (currently 6 years old) born after a normal pregnancy and delivery, with severe intellectual disability (ID), syndromic facial features that are very similar to that seen in Family 1 and 2, heart defects (AVSD, hypoplastic aortic bow), feeding difficulties in the first year, hearing loss (40 decibels), and frequent infections. He was referred for clinical exome sequencing at the age of 6 years with severe ID, hypotonia and hyperlaxity of joints, and frequent pneumonia. He has a healthy brother, and other maternal male relatives are all healthy. The family did not agree to publish photographs of the face, which did demonstrate remarkable similarity to the probands in Family 1 and 2. The family did allow publication of a photograph of an intergluteal crease also found in this affected individual (**Figure S9A**).

MRI of the brain was notable for a small anterior pituitary gland and findings consistent with an ectopic posterior pituitary (**Figure S7; O**). While posterior pituitary ectopia can be associated with septooptic dysplasia, no other findings are present to support septooptic dysplasia radiographically apart from questionably underdeveloped olfactory sulci: there is no septum pellucidum deficiency or falcine deficiency with hemispheric interdigitation like some of the other subjects, and the optic nerves appear grossly normal. The ventricles are slightly dysmorphic but still normal in absolute size, reflecting some prominence of the lateral ventricular bodies near mineralization suggestive of remote germinal matrix hemorrhage (**Figure S7; P)**. The pons is hypoplastic in appearance. The corpus callosum and cerebellar vermis appear normally formed. No other structural or myelination abnormality is present in the brain. Regarding the abnormal gluteal crease, spine MRI suggests an underlying coccygeal sinus connecting the gluteal crease deviated to the right and the coccyx.

Analysis for family 3

This individual was ascertained through family-based WES in a diagnostic setting using previously described techniques [12]. WES was performed using a SOLiD 5500XL machine (Life Technologies) after enrichment with the Agilent SureSelectXT Human All Exon 50Mb Kit (Agilent, Santa Clara, CA, USA). The data were analyzed using LifeScope (Applied Biosystems, Life Technologies, Paisley, UK) software. All clinically relevant candidate *de novo* mutations were validated using Sanger sequencing, and subsequently tested for absence in parental DNA samples.

Family 4

A fourth family from Maine, USA, was identified, after referral from GeneDx upon clinical exome sequencing. The affected is a male proband (**Figure 1-4A**) that was the product of a 42-week gestation to a 38-year-old, G3, P0-1, Caucasian female. Paternal age was 35. The pregnancy was complicated by some placental bleeding at 6 weeks' gestation. An amniocentesis at 16 weeks' gestation revealed normal chromosomes. There were no reported exposures to alcohol, tobacco, medications, illnesses, or drugs during the pregnancy. The proband weighed 6 pounds 4½ ounces and was 20 inches long at birth. Apgar scores were 1 at one minute, 8 at five minutes, and 9 at 10 minutes. Brain MRI and CT were read as normal on 2/7/2007.

The probands initial genetics evaluation was performed on 1/29/07. At the time, he had a history of developmental delay, truncal hypotonia with potential hypertonia of the extremities. He had a history of a right clubfoot, bifid uvula, small stature, and microcephaly with significant left parietal plagiocephaly. His features were thought to be consistent with possible Cornelia de Lange syndrome, although no mutations in any gene known to be associated with CdLS were found. The proband has been diagnosed with spastic diplegia; he has contractures of his lower extremities. These have been treated with Botox and surgery (hamstring lengthening). The proband had a secundum ASD that closed on its own. He has mild dilation of the ascending aorta and MPA, recurrent otitis media and has had several sets of tubes. He has mild astigmatism that does not required correction at this time.

Prior to XomeDx exome sequence at GeneDx, clinical testing included: normal *NIPBL* sequencing, 500k SNP array, a karyotype, and subtelomeric FISH. Urine MPS screen was normal. XomeDx sequencing resulted in 2 variants of uncertain significance, both in X-linked genes: *POLA1* p.His707Arg (maternal), *TAF1* p.IIe505Asn (*de novo*).

Family 5

A fifth family, recruited from Canada, consists of a single affected 3 year-old male (**Figure 1-5A**), originally born in Ecuador. His mother was 26 years old and G2P1. Ultrasounds at 12 weeks

and 20 weeks were done in Ecuador and reportedly normal. He was delivered at term by Ceasarean due to nuchal cord. He presented with hypotonia and feeding difficulties after birth, and was hospitalized for feeding and oxygen support in the NICU for one month. The family subsequently moved to Canada, where the boy was referred to Medical Genetics at 2 years of age for evaluation of developmental delay and hypotonia. At 2 years of age, growth parameters consisted of head circumference 50.5cm, (at <50th percentile), with a weight of 9.0kg (< 3rd percentile) and height of 81.5cm (3rd percentile). He was dysmorphic and in addition to global developmental delay, growth retardation and hypotonia, shared facial similarities with the other affected boys discussed above in Family 1, 2, 3 and 4. These features include sagging cheeks, downslanting palpebral fissures, broad nasal bridge and tip, long smooth philtrum, and a high palate. He also has other overlapping and unique phenotypic features, including hypermobility, bilateral vertical talus, cryptorchidism, and the characteristic intergluteal crease seen in affected males from the preceding four families.

Clinical investigations included karyotype, array CGH (60K), methylation studies for Prader-Willi syndrome, and DNA testing for myotonic dystrophy, which were all normal. Brain MRI demonstrated mild ventriculomegaly, mild hypoplasia of the corpus callosum splenium, and deficiency of the falx cerebri reminiscent of the Family 1 probands (**Figure S7**). However, there was no septum pellucidum deficiency or vermian hypoplasia. Echocardiogram was normal. Given that all previous tests were normal, he and his parents underwent trio-based whole exome sequencing as an investigation of the causes of sporadic ID at the Alberta Children's Hospital Research Institute (ACHRI) genomics and bioinformatics facility (University of Calgary, Calgary, Alberta, Canada). A total of 18819 variants were identified in protein coding genes. 732 such variants were rare (found at a frequency of less than 2% in internal databases and less than 1% in external databases such as ESP6500 and the 1000 genomes project). From this pool of rare variants, we identified no *de novo* variants, no biallelic putatively pathogenic variant pairs and six putatively pathogenic X-linked variants. All but one of the six X-linked variants were eliminated as potential candidates based upon a literature review of gene function. The remaining variant, *TAF1* chrX:70618449A>G, had not been previously reported, was maternally inherited, and was not predicted to alter the protein-coding sequence (synonymous) however, splice prediction algorithms (MaxEntScan, GeneSplicer, and Human Splicing Finder) suggested the introduction of a splice acceptor site. RNA studies were undertaken. PCR amplification and gel electrophoresis of cDNA from the affected individual and his mother revealed two transcripts of different sizes in both samples. Sanger sequencing of the PCR products revealed that the smaller fragment contained a 28 bp deletion, resulting in a frameshift and a premature stop codon (p.(Arg1228Ilefs*16)) (data not shown). The unaffected mother of the affected boy carries the variant too, and the alternative splice site is also used. The presence of two transcripts (r.[3708a>g, 3681_3708del28]) in the proband suggests that *TAF1* levels are critical for normal development.

Sequencing analysis for family 5

Genomic DNA from the affected individual and his parents was enriched for exonic sequences using the Agilent SureSelect V5 exon enrichment kit (Agilent technologies, Santa Clara, USA) prior to sequencing on a 5500xl SOLiD sequencer (Life Technologies, Carlsbad, USA). Sequencing achieved a median coverage of 102X in targeted regions. Bioinformatics was performed by the ACHRI genomics and bioinformatics facility using an in-house pipeline. In brief, reads were aligned to the human reference genome (GRCh37), PCR duplicate reads were removed using Picard and SAMTools. Reads were realigned using GATK. SNVs were called using FreeBayes and DiBayes, INDELs were called using Atlas2, and CNVs were called using Exome CNV.

Family 6

A sixth family, recruited from France contained one affected male (6A), 3 unaffected brothers and an unaffected sister. The male proband was born at term with intrauterine growth retardation (length 45 cm, weight 2500 g, OFC 33.5 cm). He had congenital torticollis. He walked independently at 24 months and later developed regression of his walking abilities. He could say a few words but did not develop sentences. He had epilepsy and psychotic behavior. He was enrolled in a school for special needs for severe ID. He had major scoliosis (80°) that required thoracic and lumbar arthodesis. He had Klippel-Feil cervical malformations. He had long and thin habitus but could not be measured, arachnodactyly with positive thumb and wrist signs and long feet. He had facial asymmetry, long face with pointed chin, prominent supraorbital ridges, malar hypoplasia and prominent ears. His cerebral MRI revealed corpus callosum agenesis and posterior periventricular heterotopia of the grey matter. He had unilateral strabismus and normal heart survey. He progressively developed tracheal spasm that necessitated tracheotomy. He had walking difficulties with severe hypokinesia. He was evaluated by a specialist in dystonia, who did not diagnose the proband with dystonia but instead with hypokinesia.

Clinical examination at 22 years of age disclosed prominent motor symptoms. The patient exhibited bradykinesia with a lack of spontaneous facial expressions and decreased blink rate. Slowness in initiating movement, especially concerning the upper limbs was obvious. This slowness was associated to reduced speed and amplitude of the movements, and interfered with daily-living motor activities. Gait was clumsy with a stiff posture, a truncal retropulsion and reduced arm swing. Mild dystonic symptoms were also noticed, with permanent right-sided laterocollis associated to elevation of the right shoulder, and equinovarus posture of the left feet when walking. Clinical examination did not disclose rigidity, tremor, pyramidal tract signs, cerebellar signs, or balance troubles.

Review of MRI imaging of the brain revealed that there is posterior corpus callosum thinning (hypoplasia) similar to some of the other subjects. The septum pellucidum, olfactory apparatus, pituitary, and cerebellar vermis all appear grossly intact. There is no convincing evidence of volume loss. There is periventricular gray matter heterotopia in the frontal white matter, which is the first malformation of cortical development seen in this cohort.

Regarding prioritization of variants for proband 6A, the patient was tested using an exome sequencing trio analysis and the *TAF1* variant was the only de novo variant found for this patient.

Family 7

Another affected individual was identified through the Deciphering Developmental Disorders study team who reported recently a *TAF1* c.4355G>A *de novo* change in one patient [13]. This patient has very distinctive dysmorphology that is very similar to the probands reported in this study, except his palpebral fissures are upslanting, instead of downslanting (**Figure 1-7A**). This individual is now almost 11 years old and is in special education. He has some basic reading skills and knows some multiplication tables. He can ride a bicycle. He has many sensory issues and some anger issues. He is in generally good health, but did have previous atrial septal defect, floppy larynx, narrow ear canals and chronic otitis media. His height is 50th centile, weight 91-98th centile and OFC 51.5cm. No spasticity has been noted.

Sequencing for Family 7

The methodology for the identification of this mutation was recently reported [13].

Family 8

An additional family from Colombia was identified with three affected brothers, ages 9, 4 and 1 years-old **(Figure 1- 8A, 8B** and **8C**, respectively), who presented with early onset epilepsy, postnatal microcephaly, severe intellectual disability, and dysmorphic features. All patients presented with seizures during the first 24 hours post delivery and severe lactic acidosis. Metabolic testing during the neonatal period included normal urine organic acids, plasma amino acids, and plasma acylcarnitine profile. Additionally, pyruvate dehydrogenase sequencing was performed with no mutations identified. Parents reported that the fontanelle closed during the first postnatal month in all three siblings. The parents were consanguineous (2nd degree cousins) and the mother had a personal history of keratoconus treated with corneal transplantation. There is no family history of similar microcephaly, seizures, or developmental delays in other members of the family.

Physical examination was remarkable for severe microcephaly, long palpebral fissures, prominent low-set ears with thickened helices, bulbous nasal tip, long philtrum, thin upper lip, high palate, and pointed chin. During infantile period, facies were round with sagging cheeks (**Figure 1-8B**) and as patients got older the face was long (**Figure 1-8A**). An intergluteal crease was also more prominent during early infancy (see **Figure S9,A-3** for an image of the intergluteal crease of **8C**). Thoracic deformities, kyphosis and scoliosis worsened over time.

Neurodevelopment was poor with no language development, inability to walk, bilateral sensorineural hearing loss, and refractory seizures. Brain CT showed calcifications (pending imaging), and brain MRI showed decreased white matter volume, and corpus callosum hypoplasia. All siblings exhibited autistic behavior and the older sibling had repetitive dystonic movements of hands and upper extremities.

Additional genetic testing included karyotyping analysis, performed with the Spectral Genomics Chip™ 1Mb which includes approximately 2500 clones (1/Mb over the length of the chromosome). The microarray detected no chromosomal imbalances and there was no detectable large scale duplications or deletions of chromosomal material within the genome for the specific regions tested. Given previously normal evaluations a familial exome study was performed on the three affected male siblings, at which point a mutation in *TAF1* was identified (c.1786C>T;p.Pro596Ser).

Sequencing analysis for family 8

Whole exome sequencing and variant calling was performed as previously described[14].

Family 9

A 10-month-old male affected individual from a Spanish family was found with multiple congenital anomalies and psychomotor delay (**Figure 1-9A**). He was the fourth child of non-consanguineous parents. He was born at 37th weeks of gestation by spontaneous vaginal delivery, with a weight of 1.69 kg (<3rd centile), length of 45 cm (10-25 percentile) and OFC of 28 cm (<3rd centile). During pregnancy, increased nuchal translucency and intrauterine growth retardation were noted with normal fetal karyotype. Since birth, he had particular facial features and multiple congenital anomalies. He had left-eye pseudo albinism, strabismus, microretrognathia, low set ears and very sparse, fine and depigmented hair. His physical examination revealed plagiocephaly and microcephaly, short neck, congenital heart disease (ostium secundum, atrial septal defect, mild aortic coarctation and pulmonary hypertension), scoliosis, pectus excavatum, sacral dimple and bilateral cryptorchidism. On clinical examination at 9 months of age, his length was 62 cm (<3rd centile), weight was 4.84 kg (<3rd centile), and

OFC was 39 cm (<3rd centile). Neurological examination showed psychomotor delay with axial hypotonia and spasticity of the lower limbs. The child has global developmental delay, being unable to sit at 10 months. Unfortunately, the patient died at three years of age by respiratory failure caused by pulmonary infection.

MRI imaging of the brain at 5 months of age showed diffuse corpus callosum hypoplasia relative to age-matched controls (**Figure S7**) [17]. There was also presence of anterior falcine deficiency like the Utah and Michigan subjects. The ventricles were slightly prominent. There was no vermian hypoplasia or septum pellucidum deficiency like the Utah probands. The spine MRI from the same date demonstrated abnormal spinal curvature, compatible with clinically observed scoliosis. No sacrococcygeal hypoplasia or other abnormality was found.

The prenatal and postnatal karyotypes of this affected individual were both evaluated and showed a normal result. A custom array-CGH was also performed to rule out large CNVs and revealed a normal result. Molecular genetic testing by next-generation sequencing was performed for 1274 pathogenic and candidate neurodevelopmental genes. After filtering for *de novo* variants, the missense c.2926G>C (p.Asp976His) transition in exon 19 of the *TAF1* gene was detected as the most potentially relevant variant (NM_004606.4). *In silico* predictions of functional relevance suggested a pathogenic effect in the conserved central domain of TAF1.

Capture array design, NGS and analysis pipeline for family 9

A custom SureSelect oligonucleotide probe library was designed to capture 19,878 coding exons of 614 pathogenic and 642 candidate genes associated with intellectual disability (manuscript in preparation). The design includes all the transcripts reported for each target gene in different databases (RefSeq, Ensembl, CCDS, Gencode, VEGA). The SureSelect DNA Standard Design Wizard (Agilent Technologies) was used for probe design with a 2X tiling density and a moderately stringent masking. A total of 71,994 probes, covering 5.073 Mbp (99.48% coverage of targets), were synthesized by Agilent Technologies (Santa Clara, CA, USA). Sequence capture, enrichment, and elution were performed according to the manufacturer's instructions. The libraries were sequenced on an IlluminaHiSeq 2000 platform with a paired-end run of 2 × 90 bp, following the manufacturer's protocol to generate at least a 100X effective mean depth.

Variant calling was performed with the DNAnexus platform (DNAnexus, Mountain View, CA, USA) through the following pipeline: Fasq paired reads were aligned to the reference human genome UCSC hg19 using the BWA-MEM algorithm from the BWA software package. Duplicated reads were removed using Picard, realigned around sites of known indels, and their quality was recalibrated by looking at covariance in quality metrics with frequently observed variation in the genome. After recalibration, variants were called with the GATK Unified Genotyper module. This pipeline follows the Broad Institute's recommendations for best practices in variant calling. Variants on regions with low mappability or variants in which there was not at least one sample with read depth ≥10 were filtered out. Annotation of nucleotide variants was performed by the Ion ReporterTM Software (Life Technologies).

The custom array-CGH microarray (Agilent technology 8x60K) to investigate large CNVs includes 453 candidate and pathological genes involved in neurodevelopmental disorders, with an average distance between genomic probes of 160Kb.

Prioritization of variants for family 9

To evaluate the putative clinical impact of the variants, the following criteria were applied:1) an allele frequency <0.01 in the 1000 g or EVS databases 2) stop gain, frameshift and splicing variants were a priori considered as most likely to pathogenic; 3) for missense mutations, amino acid conservation and prediction of pathogenicity (SIFT, Polyphen-2 and Grantham); 4) a de novo occurrence (dominant inheritance), the presence of two mutant alleles in the same gene, each from a different parent (recessive inheritance), or maternal inheritance of X-linked variants; 5) the absence of the variant in other samples (in-house database); 6) phenotypic consistency with clinical signs associated to mutations in the same gene when available. To evaluate the possible effect of synonymous or intronic variants in gene splicing we used the Human Splicer Finding web tool.

Validation by Sanger sequencing for family 9

Relevant variants were re-sequenced by Sanger sequencing. After PCR amplification from DNA of the patient and his parents using specific primers, bidirectional sequencing was performed following the BigDye Terminator kit and an ABI PRISM 3500 automated sequencer (Life Technologies, Carlsbad, CA, USA). All primers for amplification and sequencing were selected with exon-primer (primers and PCR conditions are available on request).

Family 10

A 16 year-old male proband, the second child of non-consanguineous parents of Albanian descent, was also identified (**Figure 3B**). His mother had a brother (**Figure S9B; III-1**) who was wheel chair bound and unable to speak since the age of 22 years old and who died at the age of 37 years old. The sister (**Figure S9B; II-8**) of the affected individual's maternal grandmother had 2 sons (**Figure S9B; III-8 and III-9**) who progressively were unable to walk or speak and both died at the age of 26 years old without a diagnosis. Another sister of the affected individual's maternal grandmother (**Figure S9B; II-2**) has a 21 year old son (**Figure S9B; III-2**) who developed neurological symptoms, similar to the affected individual, at the age of 14 years old and has been wheel chair bound for the last five years.

The affected individual was born after full-term pregnancy with a normal delivery. Birth weight was 2.85 kg (<10th centile). The pregnancy was uneventful and the prenatal serial ultrasound examinations were reported as normal. Perinatal period was reported as uncomplicated, but at the age of 2 years he was significantly delayed in supporting his head with other psychomotor milestones within normal limits. At that age he started to present an ataxic walk and stereotypic movements of the hands. At the age of 16 years he was referred for comprehensive genetic testing, due to progressive difficulties in walking, tremors in his movements and learning difficulties since the age of 14 years old. His speech is very poor and slow with dysarthria. He developed swallowing problems at the age of 18 for which he was treated with botulinum toxin. Since the age of 18, he is wheel chair bound, unable to properly use his hands (see **Videos S6 and S7**). Routine blood and biochemical plasma and urine investigation were normal, while the visual examination revealed myopia and strabismus.

Brain MRI revealed significant degree of cerebellar atrophy and a smaller pons with an enlargement of the 4th ventricle. The cerebral hemispheres presented a mild atrophy with an enlargement of the lateral and 3rd ventricles. The flair and T2 signals were increased in the posterior periventricular and triangular white matter.

Microarray screening revealed a 0.423 Mb duplication at Xq13.1 in the proband [arr Xq13.1(70,370,794-70,794,385)x2 (NCBI build 37/ hg19)] that included *NLGN3, GJB1, ZMYM3, NONO*, and *TAF1* genes. This duplication is illustrated in **Figure 3**. This affected individual additionally carried a de novo deletion on 17q21.31 (0.63 Mb) [arr 17q21.31(44,159,803-44,787,924)x1 (NCBI build 37/hg19)] containing *KANSL1, LRRC37A, ARL17B,* and *NSFP1*. The 17q21.31 (KANSL gene deletion syndrome) presents with moderate to severe ID associated with highly distinctive facial features, but not with brain anomalies or progressive inability to walk or to speak and certainly not death[53-55].

The affected individual, his healthy brother (**Figure S9B; IV-4**), parents (**Figure S9B; III-15, III-16**), maternal grandmother (**Figure S9B; II-12**) and an affected maternal cousin (**Figure S9B; III-2**) were tested with the custom X-chromosome array platform. The Xq13.1 duplication was found to be inherited from his mother [arrXq13.1(70,359,782-70,766,824) (NCBI build 37/ hg19)]. Regarding other family members tested, the maternal grandmother and the maternal male cousin (exhibiting similar phenotype according to his aunt but not seen by our group) were positive for the same duplication while the healthy sibling (**Figure S9B; IV-4**) was negative.

Family 11

Another affected individual was ascertained from a family in Greece. The boy was the first child of non-consanguineous parents, both of them healthy and 30 years old at the time of the child's birth. This individual was referred to the Department of Medical Genetics at the age of 6 years, due to global developmental delay, hypotonia, inability to speak, walk and control his head. During the clinical evaluation, dysmorphic facial features were noted, which included a broad upturned nose, sagging cheeks, prominent peripheral ridges, deep-set eyes, high arched palate and prominent ears. There was no family history of any related symptoms. The family refused to allow pictures to be published.

The medical history revealed that the boy was delivered after an uneventful pregnancy with C-section because of umbilical cord entrapment at 38 weeks of gestation. The prenatal serial ultrasound examinations were reported as normal. Birth weight was 2.83 kg (10th centile), length was 49 cm (25th centile), and the OFC was 34 cm (25th centile). Perinatal period was uncomplicated except from jaundice that required phototherapy.

At the age of 5 months, he initially presented to medical attention with motor regression, severe hypotonia and very poor head control. The patient was reevaluated at the age of 10 months due to a light head tremor. At that time, the neurologic examination revealed truncal hypotonia, low deep tendon reflexes and proximal weakness with inability to raise the neck, arm and leg against gravity, but there were no signs of fasciculation or atrophy. At the age of 13 months there was no motor progress and neurological evaluation was unchanged except for the loss of deep tendon reflexes and hand tremor. EMG showed no abnormalities, while the brain MRI showed increase flair and T2 signal in the periventricular white matter. Routine blood and biochemical plasma and urine investigation, mucopolysacharides, lysosomal enzymes, acylcarnitine, very long fatty acids, transferring isoelectric activity, and A-glycosidase activity (on fibroblasts) were normal. Bone X rays did not reveal dramatic issues at that time. In addition, testing for mitochondrial disorders was negative.

During the following two years, he developed severe thoracic cage deformities and joint contractures (arm and legs). At the age of 8 years old he was admitted to the hospital due to severe cardiopulmonary insufficiency attributed to an infection, and he died from this illness. Microarray screening revealed a 0.42 Mb duplication [arr Xq13.1(70,287,519-70,711,110)x2 (NCBI build 37/ hg19)] that included *NLGN3, GJB1, SNX12, FOXO4, MED12, ZMY3, NONO,*

and *TAF1*. Both parental DNAs tested by the same methodology, as well as a new pregnancy proved negative for the affected individual's duplication.

Microarray analysis for family 10 and family 11

Genomic DNA was obtained from 3 ml of peripheral blood using the BioRobot® M48 System (Qiagen, Hilden, Germany) and the commercially available kit MagAttract® DNA Blood Midi M48 Kit (Qiagen). The quality and quantity of the DNA samples was determined using a NanoDrop ND-1000 UV-VIS spectrophotometer. Agilent Human Genome CGH 4x180K (Sureprint G3 arrays) were used (Agilent Technologies, Santa Clara, CA, www.agilent.com) for the initial discovery of the CNVs in the two affected individuals. The average spatial resolution for the 180K platform is 13-25Kb. Labeling, hybridization and data processing was carried out according to manufacturer's recommendations and as previously published[15]. The available parental DNA samples were processed in the same manner. Subsequently due to the X-chromosome finding we designed a custom X-chromosome (60K) using the Agilent SureDesign e-array platform and re-tested the affected individual of family 10 and other members of his extended family (healthy brother, mother, father, maternal grandmother, affected maternal cousin) following the same methodology as previously described[22]. The custom X-chromosome array has median probe spacing for the X-chromosome of 6.14Kb.

Additional results for family 1-based analyses

Whole genome sequencing

Complete Genomics WGS was optimized to cover 90% of the exome with 20 or more reads and 85% of the genome with 20 or more reads. Illumina WGS resulted in an average mapped read depth coverage of 37.8X (SD=1.3X). >90% of the genome was covered by 30 reads or more and >80% of the bases had a quality score of >30. See **Table S1** for more details about the sequencing data.

Concordance among variant detection pipelines

SNP and INDEL concordance across SNP and INDEL detecting pipelines applied to Illumina raw data was computed. In agreement with various other studies that have focused on computing SNP and INDEL concordance across pipelines, the mean concordance for SNPs across the two SNP detecting pipelines (GATK and FreeBayes) among the 10 sequenced individuals was 81.8%, whereas the mean concordance for INDELs between GATK and FreeBayes was 62.2% (with a mean of 80.3% of Scalpel calls being detected by the other two pipelines). Agreement between CNV detecting pipelines was low; with 6.3% percent of PennCNV found by ERDS and 0.9% percent of ERDS CNVs found by PennCNV. No known disease-contributory CNVs were discovered, but we archive in our study 8 de-novo CNVs that are not currently associated with any biological phenotype (see **Table S3** for the list of CNVs).

Between Illumina and CG sequencing platforms, the SNP concordance was 77.1% whereas the INDEL concordance was 44.8%. CNV concordance between the two sequencing platforms was 5.7%. To make these cross platform comparisons, variants generated from the different informatics pipelines applied to the Illumina raw data were combined into a larger set that only included unique calls from each caller.

Study design comparisons

We explored differences between study-design scenarios in their output in terms of variants conforming to the disease models (de-novo, autosomal recessive and X-linked, see **Figure S4**). Results were compared between two distinct study designs: a quad study design and a full family study, which integrates data from all of the sequenced family members as well as all of the variant detecting pipelines previously described. We found that there was a mean fold difference of 2.4 to 14.0 in the number of variants that were segregating in terms of the three different disease models (a mean fold difference of 2.4 was observed for the autosomal recessive disease model, 3.1 for the de-novo disease model and 14.0 for the X-linked disease model). For each disease model, simple python set operations, SVS operations and GEMINI operations were used to divide variants segregating according to each model. For the de-novo disease model, python set operations, SVS and GEMINI identified 52,360, 40,440 and 42,625 variants respectively for a quad-based study design and 17,908, 12,298 and 14,249 variants for a study design incorporating all of the family members recruited into the study. Similarly for the autosomal recessive disease model, 59,111, 57,614 and 64,366 variants were found using a quad study design and 37,678, 20,579 and 22,302 variants were found when using all of the family members. Lastly, for the x-linked model, 26,228, 27,121 and 39,316 variants were found under a quad study design whereas 2,322, 2,538 and 1,958 variants were found when all family members were incorporated into the analysis.

We also explored differences in disease variation discovery due to varying prioritization schemes. We looked at these differences in combination with applying a quad or full family study design. In general, 40 variants were identified using a quad based study design and using the CADD scheme, whereas 15 variants were found using the same study design but instead using the Coding prioritization scheme, with only two variants being identified by both; a non-frameshift substitution in *NLGN4X* and a nonsynonymous variant in *TAF1* conferring a p.Ile1337Thr change (see **Table S3**). 8 variants were identified using the full-family based study design in combination with the CADD prioritization scheme whereas 7

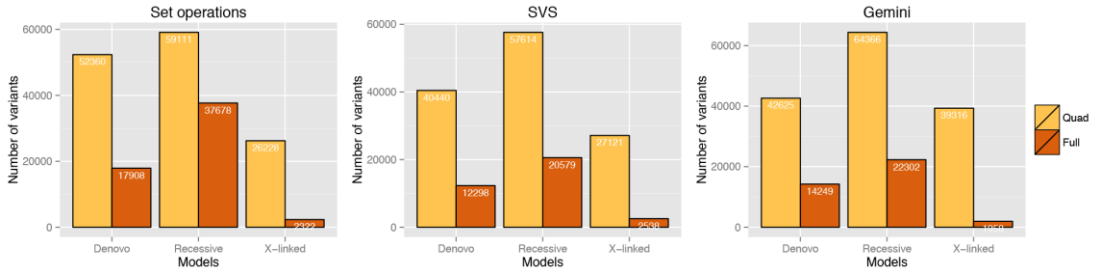


Figure S4. Bar plots showing differences in the number of variants conforming to de-novo, autosomal recessive and X-linked disease models using python set operations, SVS and Gemini software between quad and full family-based study designs in the Family 1 study.

variants were found using this same study design and the coding scheme, only one of which was found using both schemes; a nonsynonymous variant in *TAF1* conferring a p.Ile1337Thr change.

Multi-generational pedigrees reduce erroneous findings

More variants are reliably eliminated when a greater portion of the family is incorporated into the analysis. This is likely due to varying false positive and false negative rates across

sequencing and informatics platforms due in part to variation in data quality across the sequenced portion of each genome in each individual. Trio and quad-based study designs are prevalent in the literature[56-60], and many human genetics studies using high-throughput sequencing technologies only employ the use of a single, or a limited number, of variant detection pipelines. Our findings highlight the need for more comprehensive family-based study designs, and we demonstrate benefits in focusing high-throughput sequencing efforts on studying large related cohorts, where intra-familial relationships allow for more rigorous variant filtering and identification of true positive alleles that might be contributing to a disease phenotype.

We were able to minimize false negative variant detections by using many orthogonal informatics pipelines, as each alone miss some true and possibly functional sequence variants but together capture a greater portion of the true call set. The multi-generational pedigree structure allowed us to minimize false positive findings by using expanded disease model operations that included three generations, effectively reducing false positive findings by corroborating genotypic evidence across the familial generations. In general, reductions in false negative and false positive calls should increase the efficacy of prioritization strategies, and reduce the number of candidate variants to a manageable and robust number in terms of performing validation and functional follow-up studies. In our study, we found this reduction to vary across disease models, with autosomal recessive, de-novo and X-linked models having candidate variant reductions of 2.4, 3.1 and 14.0 fold respectively. Further, the number of final prioritized variants was reduced by a factor of 3.8 across both of the prioritization schemes that were employed (53 unique variants were identified through prioritization using a quad study design and 14 were identified using the full-family study design).

Before WGS was performed, a SNV of unknown significance was detected by clinical gene-panel sequencing: *ZNF41*; *p.Asp397Glu*. This variant was determined to be a variant of unknown significance due to the clinical ambiguity of the variant as well as the limited scope of the gene panel. There is some previous work implicating other variants in this gene as contributing to X-linked mental retardation[61], although during the course of this study, the significance of this finding was challenged[62]. When the study was expanded to include WGS data generated by CG on the two affected children and their parents, this variant was still identified as a putative disease-contributory variant. Only when a larger portion of the family was recruited for genotyping and additional Illumina-based WGS performed were we able to show that this variant was observed in other, unaffected, family members, including a male cousin. We found this to be the case for other variants detected under a quad-based study design. For example, functional prediction algorithms (Polyphen and SIFT) indicated that another variant, located in *ASB12*, was deleterious and thus suspected as a potential disease-contributory variant. This inference was found to be invalid due to its presence in other unaffected members of their family (see **Figure S1** for Sanger sequence traces which show *ZNF41* and *ASB12* variants to be present in other members of the family, despite being identified as important in disease using a quad-based study design). In another instance, a variant in *PION* was thought to be de-novo in the children, but was found to be the result of poor sequencing coverage at that position, as this variant was indeed present in the mother, hence not de-novo in the children (**Figure S1**). We have observed that some studies use trio or quad based designs and assert genetic "causality" when there is very little evidence supporting their case. This runs the risk of polluting the literature further with many false positive findings [63].

An extensive literature review was conducted in pursuit of genotype-phenotype correlations with the above variants. *FRG2B* and *FAM47B* are not known to be involved in the pathogenesis of any human disease, although the detailed molecular function of these genes has been largely unexplored. *FRG2B* is homologous to *FRG2*, which locates on chromosome 4 and has been implicated in playing a role in the pathogenesis of facioscapulohumeral muscular dystrophy (FSHD) in patients with substantial reductions in a 11-150 unit 4q35 microsatellite repeat[64-66]. However, reductions in the homologous 10qt26 microsatellite repeat, proximal to *FRG2B*, have not been associated with FSHD. ZNF423 acts as a transcriptional regulator, and variants in ZNF423 coding regions have been implicated in the pathogenesis of Joubert syndrome[67; 68]. The variant that we have identified in *ZNF423* is located within an intron, and its molecular function is unknown. SLC28A1 is thought to mediate sodium-depedent fluxes of uridine, adenosine and azidodeoxythymidine[69], whereas SNAPC5, also known as SNAP19, plays a scaffolding role in the forming the complete SNAP complex, which is required for the transcription of snRNA genes[70]. The molecular functions of KCHN1 and DMRTB1 are not well understood or studied.

X-chromosome Skewing in Family 1

The X-chromosome skewing assay revealed that the mother of the two affected boys has skewed, 99:1, X-chromosome inactivation (**Figure S5**). The grandmother, as well as the aunt of the affected boys, does not show any appreciable X-chromosome skewing, suggesting the possibility of a newly arising deleterious X-chromosome variant.

Zebrafish studies

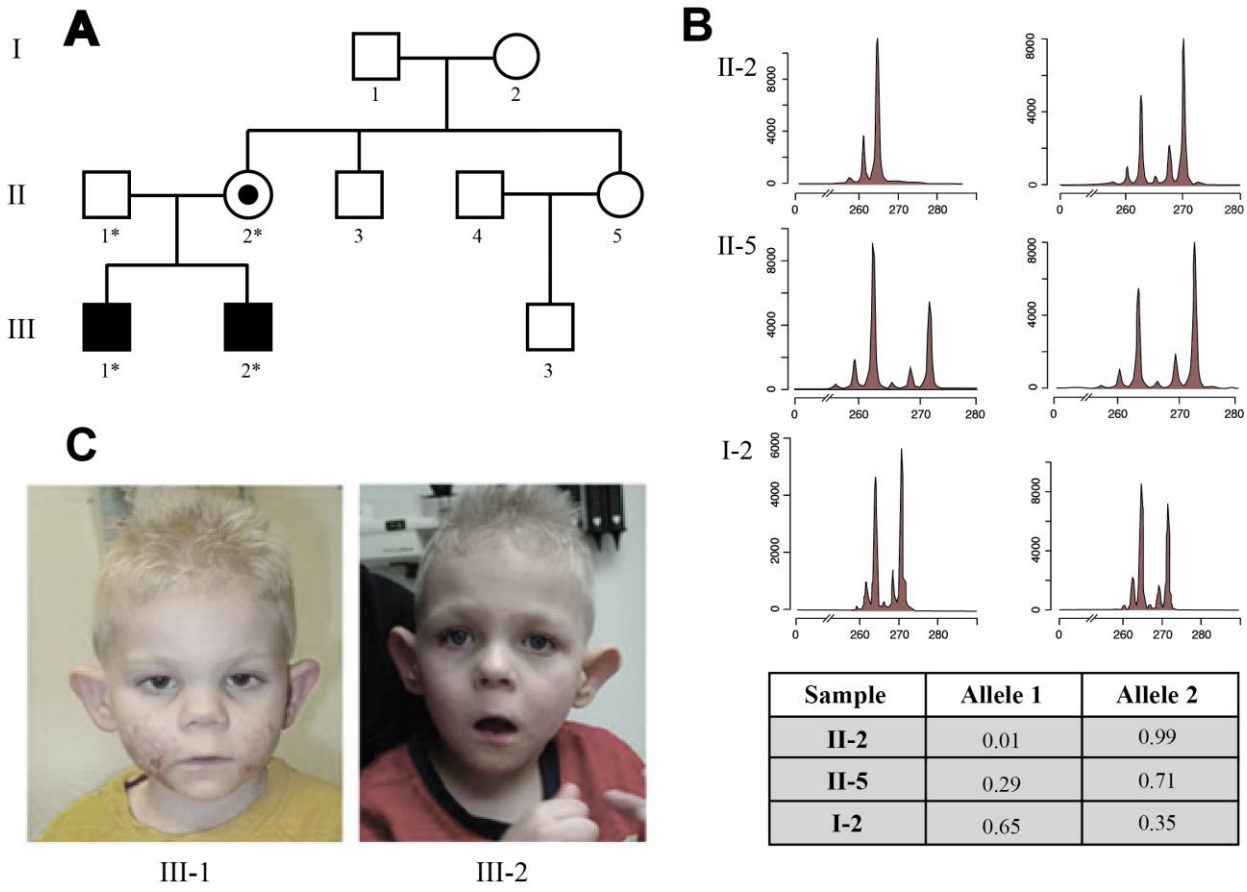
Zebrafish were maintained and mated as previously described[40] and all experiments were carried out with the approval of the Duke University Institutional Animal Care and Use Committee (IACUC).

To suppress the endogenous expression of *taf1I,* we injected 1nl of diluted MO (12 ng). For our rescue experiments (to ensure phenotypic specificity of the MO), we co-injected 1nl of MO (12 ng) and WT human *TAF1* RNA (50 pg) into wild-type zebrafish embryos at the one- to four-cell stage. To generate RNA for rescue and overexpression experiments, we first obtained a *TAF1* ORF clone from Open Biosystems (clone ID, 100069121) corresponding to the human wild type *TAF1* cDNA (ENST00000373790); this was then cloned into the pCS2+ vector using Stbl2 competent cells (Invitrogen) for transformation. Finally, we transcribed RNA *in vitro* using the SP6 Message Machine kit (Ambion).

For the cloning of the guide template RNA for the CRISPR experiments , we utilized the 5'-ATCGGCCCTTCCACTTGACA-3' oligonucleotide sequence, that was ligated into the pT7Cas9sgRNA vector (Addgene) using the *Bsm* BI sites. For the generation of the guide RNA, the template DNA was linearized with *Bam* HI, purified by phenol/chloroform extraction and *in vitro* transcribed using the MEGAshortscript T7 kit (Invitrogen). To generate F0 CRISPR mutants we injected 1nl containing 100pg *taf1* guide RNA and 200ng Cas9 protein (PNA bio, CP01) to one-cell stage embryos. To determine the efficiency of the guide RNA, embryos were allowed to grow to 3 days post fertilization (dpf), at which time they were euthanized and subjected to digestion with proteinase K (Life Technologies, AM2548) to extract genomic DNA. The targeted locus was PCR amplified using the drTAF1_g1test_1F 5'- TTACCGTATCCAGCACACTGTC-3' and drTAF1_g1test_1R 5'-TTTTACTGCAAATGTCGTGTCC-3' primer pair. The PCR amplicon was subject to digestion using T7 endonuclease I (New England Biolabs, M0302L) at 37 °C for 1 hr and was

visualized on a 1.5 % of agarose gel. For Sanger sequencing of individual products from the *taf1* locus, PCR fragments were cloned into the pCR4/TOPO TA cloning vector (Life technologies, 450030) and each clone was Sanger sequenced. We observed sequence aberrations in ~20% of the evaluated clones.

Injected embryos were fixed in Dent's fixative (80% methanol, 20% DMSO) overnight at 4°C. Embryos were permeabilized with proteinase K followed by post-fixation with 4% paraformaldehyde (PFA). PFA-fixed embryos were then washed, first in PBS and subsequently in IF buffer (0.1% Tween-20, 1% BSA in PBS) for 10' at room temperature. Embryos were incubated in blocking buffer (10% FBS, 1% BSA in PBS) for 1 hour at room temperature. After two washes in IF Buffer for 10' each, embryos were incubated in the primary antibody (anti-acetylated tubulin (T7451, mouse, Sigma-Aldrich), 1:1000) in blocking solution, overnight at 4°C. After two additional washes in IF Buffer for 10' each, embryos were incubated in the secondary antibody solution (Alexa Fluor goat anti-mouse IgG (A21207, Invitrogen), 1:1000) in blocking solution, for 1 hour at room temperature. All the experiments were repeated at least twice; statistical significance values were computed using student`s t-test.



Sample	Allele 1	Allele 2
II-2	0.01	0.99
II-5	0.29	0.71
I-2	0.65	0.35

Figure S5. (A) Pedigree structure of all individuals in Family 1 that were sequenced during the course of this study and images of the two affected siblings, who display strikingly similar facial dysmorphology. Affected brothers, III-1 III-2, are sons to mother II-2, who tested positive for extreme X-chromosome skewing **(B)**. Individuals with a star next to their number indicates that their whole genomes were sequenced with both the Complete Genomics sequencing and analysis pipeline as well as with Illumina sequencing technology and the various downstream analysis pipelines. All other numbered individuals had their whole genomes sequenced only with the Illumina WGS technology, followed by the downstream analysis pipelines described in the methods section. **(C)** The affected have distinctive shared facial features, including broad, upturned nose, sagging cheeks, downward sloping palpebral fissures, prominent periorbital ridges, deep-set eyes, relative hypertelorism, a high-arched palate, and prominent ears.

We show additional longitudinal data for Family 1 in **Figure S6**.

Figure S6. Additional images of the male siblings from the Utah Family 1 at different ages, and their distinctive facial features, aplasia cutis congenita and a unique standing posture. The younger brother at the ages of 19 months (**A**), 9 (**B**), 10 (**C**), and 12 (**D**) years old, and the elder brother at the ages of 3 (**F**), 11 (**G**), 12 (**H**), and 14 (**I**) years old share distinctive facial features, including a high-arched palate (**E**). The elder boy has a notable scar on his head from the surgery of treating his birth defect of aplasia cutis congenital (**J**). The younger boy at the age of 12 shows a unique standing posture with protruding abdomen and bent knees (**K**).

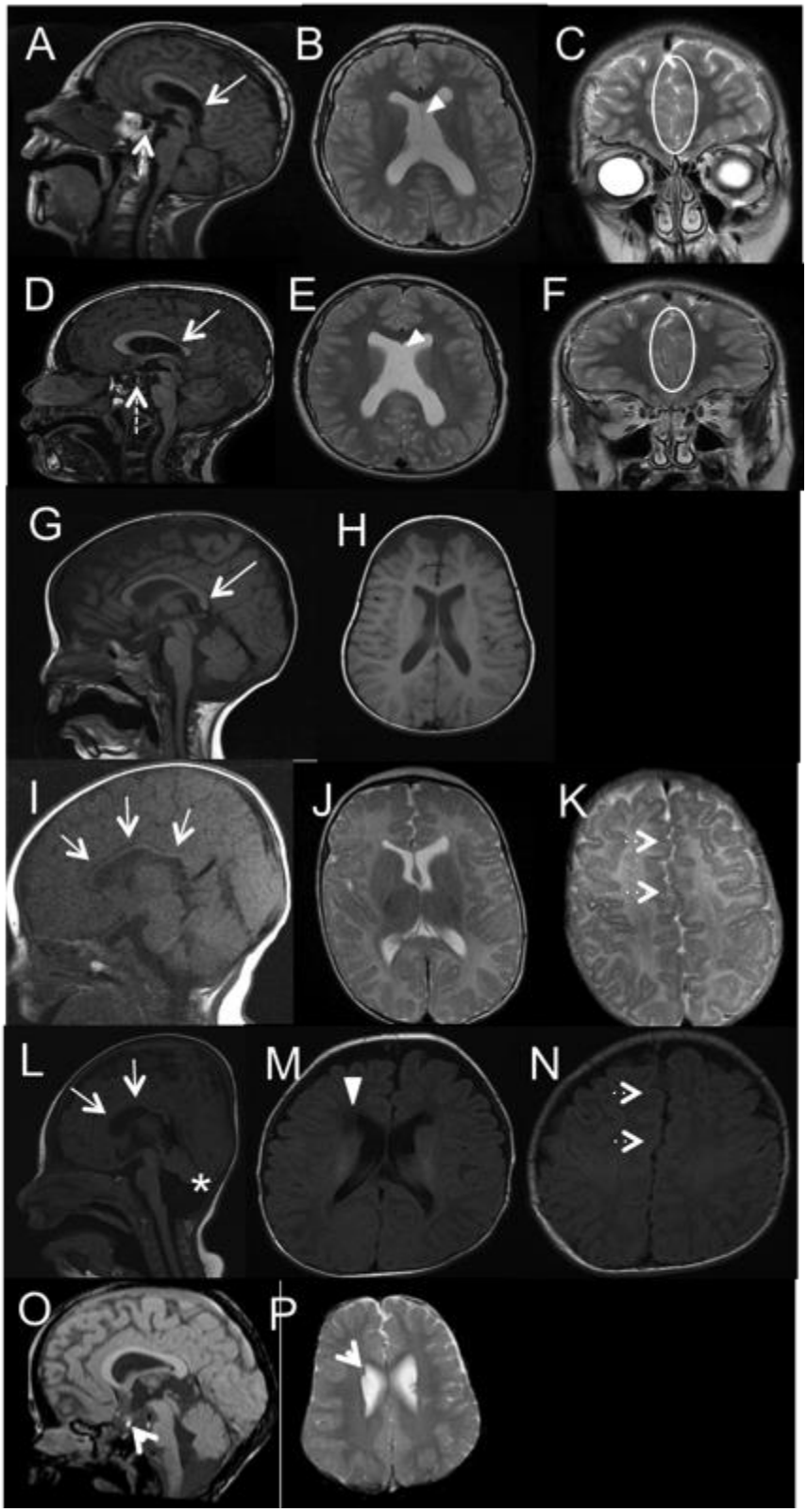


Figure S7. Sagittal T1 (**A,D**), axial T2 (**B, E**), and coronal T2 (**C,F**) weighted images from brain MRIs of the Utah brothers from family 1 (older brother at 13 years A-C, younger

brother at 11 years D-F) demonstrated a remarkably similar constellation of abnormalities. In both subjects, there was hypoplasia of the isthmus and splenium of the corpus callosum (arrows) with thickness falling below the third percentile reported for individuals of the same age (1). As is often the case with callosal hypoplasia, there was associated dysmorphic configuration of the lateral ventricles and mild lateral ventriculomegaly without positive findings of abnormal CSF dynamics (i.e. no imaging evidence of hydrocephalus). There was also deficiency of the septum pellucidum in both brothers (arrowheads), with the older brother having absence of the posterior two-thirds of the septum pellucidum and the younger brother having complete absence of the septal leaflets. Other findings associated with septooptic dysplasia included underdeveloped pituitary glands for age (dashed arrow), deficiency of the anterior falx with mild hemispheric interdigitiation (circles), and question of small olfactory bulbs despite fully formed olfactory sulci. However, the optic nerves appeared grossly normal in size. Finally, there was subjective vermian hypoplasia with the inferior vermis resting at the level of the pontomedullary junction rather than a more typical lower half of the medulla.

Sagittal T1 (**G**) and axial T1 (**H**) weighted brain MRI of subject 5A at 2 years of age demonstrated mild hypoplasia of the splenium of the corpus callosum (arrow, **G**) relative to age matched controls. There was also mild ventriculomegaly. However, there was no deficiency of the septum pellucidum, deficiency of falx cerebri, or grossly abnormal size of the pituitary gland for a 2 year old. The vermis was fully formed.

Sagittal T1 (**I**) and axial T2 (**J,K**) images of subject 9A at 5 months of age demonstrated diffuse hypoplasia of the corpus callosum (arrows, I). Like the family 1 subjects, there was some deficiency of the anterior falx cerebri with mild cerebral hemisphere interdigitation (dashed arrows, **K**). The ventricles were top normal for age. There was no deficiency of the septum pellucidum, overt abnormality of the pituitary gland for age, or vermian hypoplasia.

Sagittal T1 (**L**) and axial T1 (**M,N**) images of subject 2A at 3 months of age demonstrated anterior callosal hypoplasia (arrows, L) and moderate severity vermian hypoplasia (asterisk, **L**). There was also mild lateral ventriculomegaly greatest anteriorly (**M**) and moderate sized germinolytic cysts at the caudothalamic grooves. Nonspecific right frontal periventricular white matter gliosis was present (arrowhead, **M**). Like the family 1 and 9 subjects, there was deficiency of the anterior falx cerebri with some mild cerebral hemisphere interdigitation (dashed arrows, **N**). There was no discernible abnormality of the pituitary gland or deficiency of the septum pellucidum.

Sagittal T1 weighted image (**O**) of subject 3A demonstrates a small size of the anterior pituitary and ectopic positioning of the posterior pituitary gland (arrow). Axial T2* GRE images (**P**) demonstrate mild prominence of the bodies of the lateral ventricles with a focus of mineralization (arrow) suggestive of remote injury/hemorrhage.

Figure S8. Sagittal T2 weighted imaging of the spine for patient 5A. There is incomplete segmentation of the T12 and L1 vertebrae with partial fusion and wasting across the disk space with questionable similar changes at T11-T12. Assuming the lowest visualized rib corresponds to T12, the cord terminates low normal around L2-L3 and there is no tethering lesion. There were no segmentation anomalies in the only other subject whose spine MRI was reviewed (subject 9A).

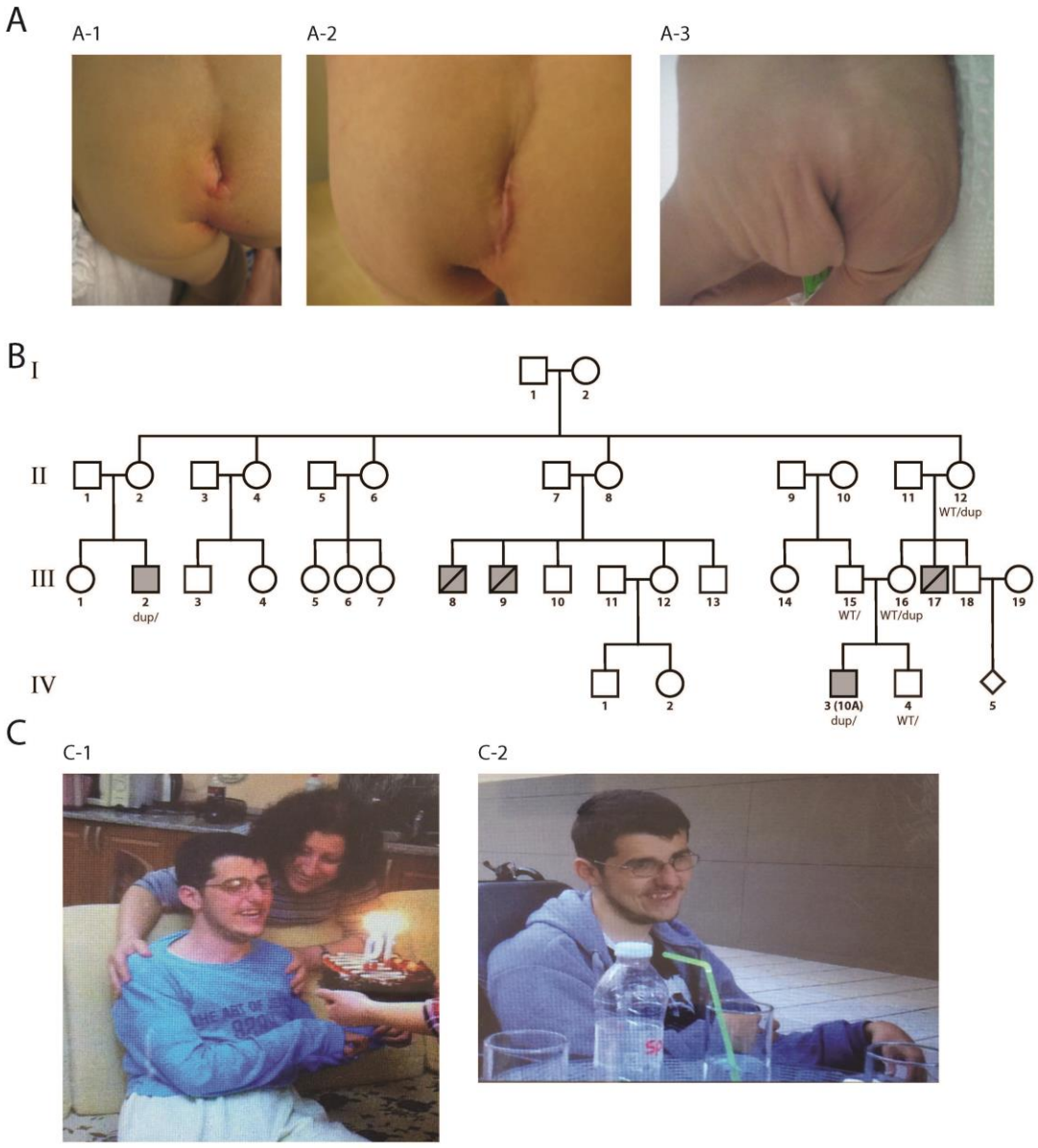


Figure S9. Intergluteal crease in the boy from Family 3. **A-1**) Before surgical repair. **A-2**) After surgical repair. **A-3**) The intergluteal crease from the affected proband 8C from Family 8. **B**) the pedigree structure of family 10, marking the segregation of the duplication CNV as hemizygous in males and heterozygous in female carriers. **C)** Pictures of the proband 10A shown in the pedigree for family 10, after his neurodegenerative course.

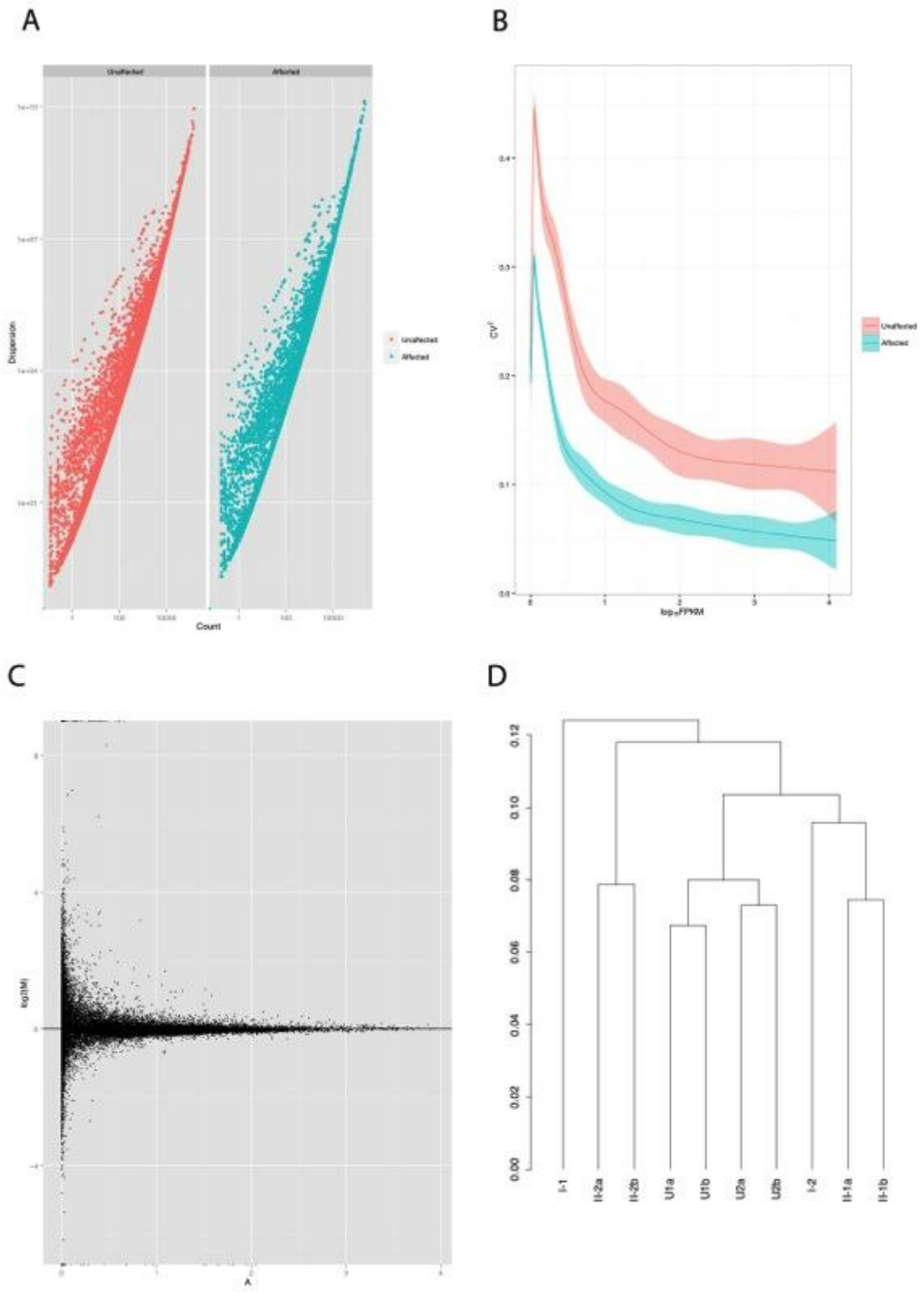


Figure S10. RNA sequencing data quality evaluations for Family 1. Various analysis tools in the CummeRbund R package[47] were used to evaluate the quality of our RNA

sequencing-based differential expression analysis for Family 1. (**A**) The degree of read count dispersion between the affected and unaffected groups was plotted; both groups appear to be quite similar in this regard. (**B**) The squared coefficient of variation in log base 10 FPKM values, used here as a measure of cross-replicate variability, was plotted. In general, we saw a higher degree of variability in the unaffected group than the affected group. (**C**) MA plot shows no obvious evidence of systematic bias between conditions. (**D**) The Jensen-Shannon distance (shown on the y-axis of **D**) was used to constructed Dendrograms between replicates (a and b) and groups (affected vs unaffected). The replicates are generally closer to each other than they are to other samples and their replicates. The replicates among the affected group (U1a-b and U2a-b) appear closer to each other than they are with any of the other replicate samples in the unaffected group (I-1, II-2a, II-2b, I-2, II-1a, II-1b).

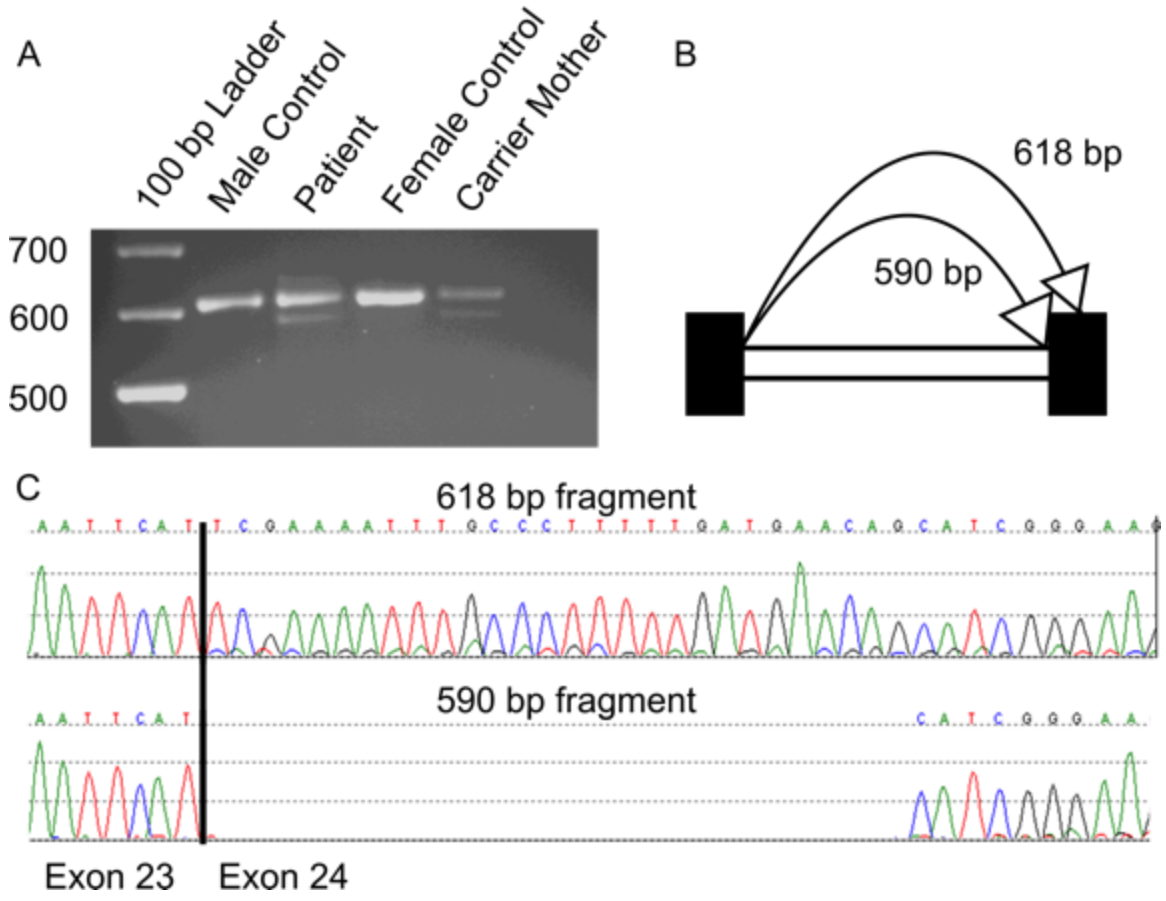


Figure S11. Splice variant in family 5. In family 5, exome sequencing revealed a synonymous variant in *TAF1*. Splice prediction algorithms Splice Site Finder-like[71], MaxEntScan[72], GeneSplicer[73] and Human splicing finder[74] predicted the introduction of a cryptic splice acceptor site with scores of 81.6/100, 5.4/16, 1.8/15 and 81.4/100 respectively. For reference, the canonical splice acceptor site scores are 74.9/100, 4.48/16, 4.5/15, and 81.2/100 respectively. (A) RNA studies were undertaken using peripheral blood samples. rtPCR amplification of *TAF1* cDNA shows a second splice isoform. The smaller band corresponds to a loss of 28 base pairs. The splice isoform can be detected in the proband and unaffected mother that also carries the variant. (B) a schematic shows how the use of a cryptic splice acceptor site results in the second splice isoform. (C) Gel extraction and Sanger sequencing of the smaller band confirms a 28 bp loss. This loss is out of frame and is predicted to result in a stop codon 16 amino acids later [p.Arg1228Ilefs*16].

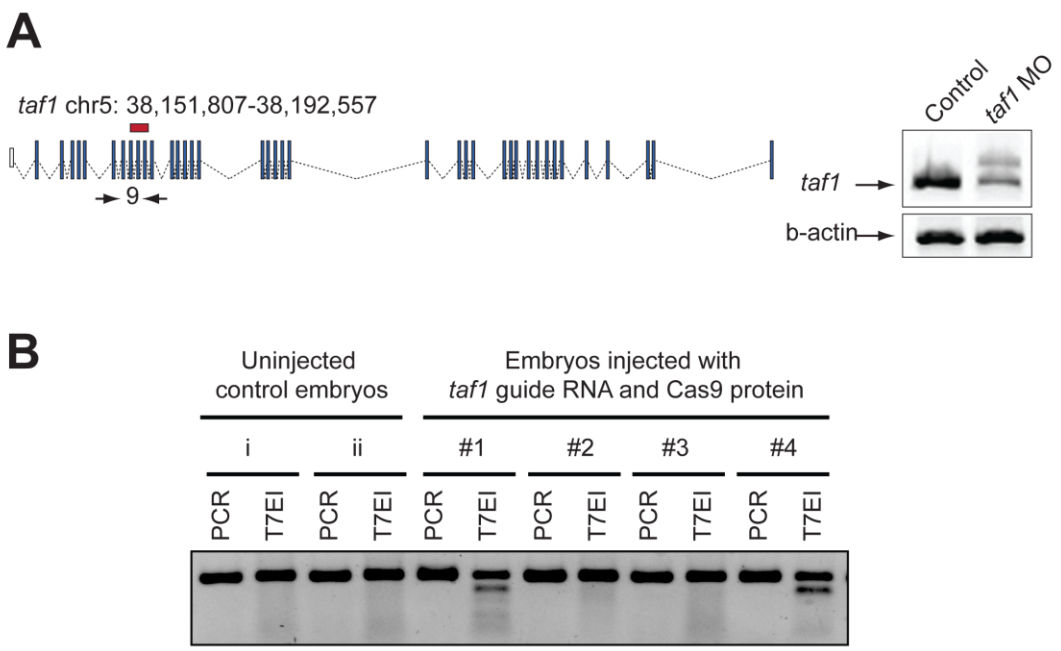


Figure S12. Assessment of the efficiency of the MO and CRISPR reagents used to suppress endogenous *taf1* in developing zebrafish embryos.
(A) Schematic of the zebrafish *taf1* ortholog, showing the site targeted by the MO. Exons are shown as blue boxes and introns as dashed lines. The MO is presented as a red box above exon 9. The arrows below the gene cartoon show the positions of the primers used to amplify the cDNA sequence flanking the MO site. The *in vivo* effect of the MO is shown in the gel image where a splice aberration of the wild-type *taf1* transcript is observed in the morphant embryos only. *B*-actin was used as a loading control. (B) Gel image showing the efficiency of the *taf1* guide RNA following T7 endonuclease assay evaluation. The first four lanes show amplicons from the locus flanking the targeted sequence, with no aberrations observed. Two of the embryos injected with *taf1* guide RNA and Cas9 protein (#1 and #4), show that the guide introduced sequence aberrations.

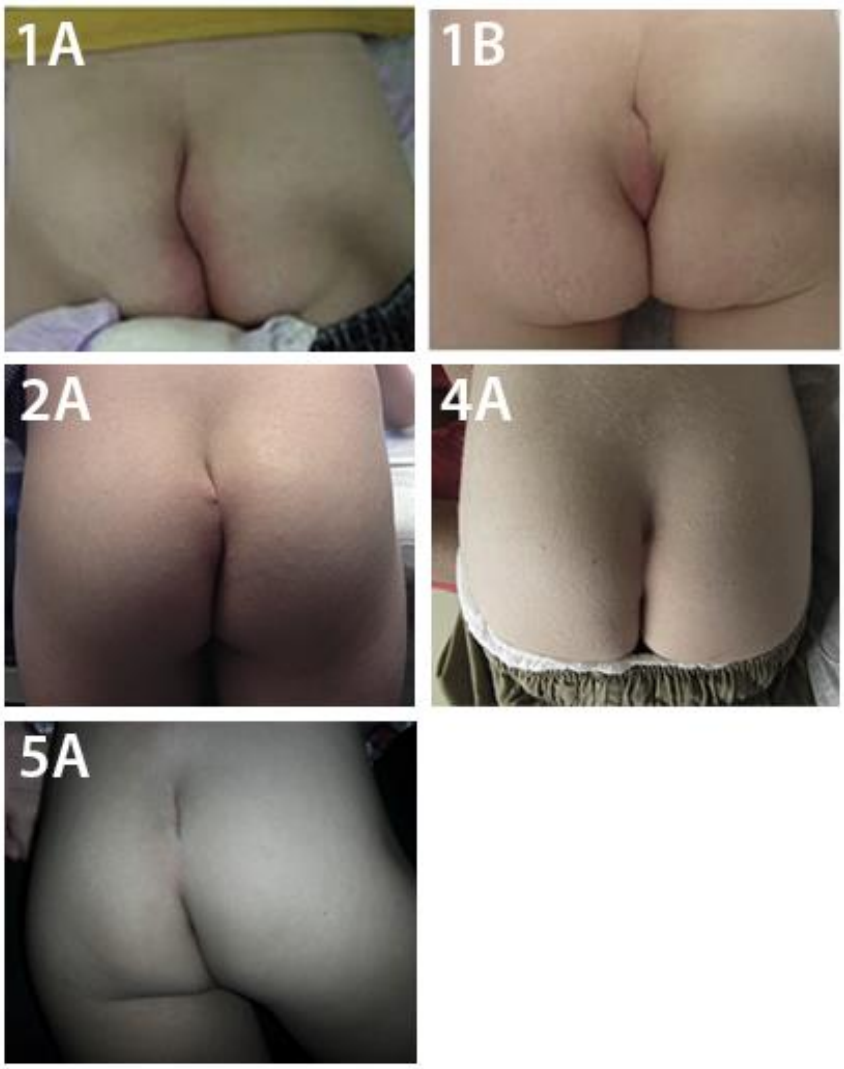


Figure S13. Some probands share a characteristic gluteal crease, with a sacral caudal remnant, shown for families 1, 2, 4 and 5 in **1A**, **1B**, **2A**, **4A** and **5A**.

Table S1. **A table summarizing the Illumina HiSeq 2000 sequencing WGS sequencing data for Family 1.** We report basic sequencing statistics generated by initial sequence data analysis performed by the sequencing facility, including estimated library size, percent PCR duplicates, genome coverage of mapped bases, percent bases with a base quality of above 30, percent reads which were passing filter aligning to genome, initial estimates of the number of SNPs and the number of novel SNPs along with initial transition to transversion ratios. We also report initial estimates of the percentage of heterozygous sites, the heterozygous to homozygous ratio, the percent of non-N ref bases covered, insert size and the tight insert size distribution.

Sample ID	Est library size	% PCR duplicates	Genome coverage of mapped bases	% bases Passing filter > Q30	% reads Passing filter aligning to genome	% of non-N ref bases covered	Ins size	Tight insert size distribution
II-3	20703336864	2.05	40	0.883566667	0.9435	0.949	261	81
II-5	21316370174	1.67	38.15	0.903266667	0.9447	0.962	268	86
III-1	21317347309	1.75	37.76	0.8971	0.9459	0.941	262	84
III-3	21209346469	1.71	37.2	0.895866667	0.9439	0.938	253	80
II-4	22830743528	1.73	39.4	0.82	0.94	0.952	261	79
I-2	21587072423	1.72	37.1	0.85	0.94	0.963	269	82
II-1	20937437047	1.74	35.4	0.87	0.94	0.936	265	86
III-2	22626476447	1.8	38.2	0.84	0.94	94.6	262	85
I-1	19959631060	1.9	38.14	-	93.99	94.2	257	86
II-2	20095126759	2	36.68	-	94.04	95.8	271	93

Table S2. **A table of prioritized genetic variations in TAF1 intellectual disability syndrome.** Variants conforming to the three disease models, de-novo, autosomal recessive and X-linked were identified. We show a list resulting from the CADD prioritization scheme as well as from the coding prioritization scheme. Both schemes required each variant to have a low population frequency (MAF < 1%). The coding scheme required all variants to also be within a coding region of the genome and to be a non-synonymous change. The CADD scheme requires all variants to have a CADD score of >20, along with the aforementioned population frequency. A variation in *TAF1* was the only variation to be reliably detected using both prioritization schemes.

Model	Location	Ref	Alt	Variant Caller	Function	Scheme
Recessive	chr1:210851705	TT	T	CG, GATK, FreeBayes, RepeatSeq	KCNH1:UTR3	CADD, score:27.5
Recessive	chr1:224772440	AATAATTTG	TA	CG, GATK, FreeBayes	intergeneic	CADD, score:22.1
Recessive	chr2:60537356	TTTTATTT	ATTATTA	CG, FreeBayes, GATK, RepeatSeq	intergeneic	CADD, score:22.3
Recessive	chr8:109098066	AT	A	CG, FreeBayes, GATK, RepeatSeq	intergeneic	CADD, score:24.6
Recessive	chr15:66786022		A	FreeBayes, GATK	SNAPC5:intronic	CADD, score:23.6
Recessive	chr16:49061346	TA	T	CG, FreeBayes, GATK	intergenic	CADD, score:25.3
Recessive	chr16:49612367		G	CG, FreeBayes, GATK	ZNF423:intronic	CADD, score:20.5
Recessive	chr10:135438929	T	G	CG, FreeBayes, GATK	I171L	Coding, gene:FRG2B
Recessive	chr10:135438951		AGCCT	FreeBayes, Scalpel	sub	Coding, gene:FRG2B
Recessive	chr10:135438967	C	T	GATK, FreeBayes	R158Q	Coding, gene:FRG2B
Recessive	chr15:85438314	C	CTTG	CG, FreeBayes, GATK, Scalpel	K141delinsIE	Coding, gene:SLC28A1
De-novo	chr1:53925373	G	GCCGCCC	FreeBayes, CG, Scalpel	A83delinsAAP	Coding, gene:DMRTB1
X-linked	chrX:34961492	T	C	CG, FreeBayes, GATK	Y182H	Coding, gene:FAM47B
X-linked	chrX:70621541	T	C	CG, FreeBayes, GATK	I1337T	Coding, gene:TAF1; CADD, score:22.9

Table S3. CNVs were detected with Illumina HiSeq 2000 WGS data using the Estimation by Read Depth with SNVs (ERDS) pipeline, Illumina Omni 2.5 microarray data using the PennCNV software package and Complete Genomics sequencing. We report 8 CNVs following in the de-novo disease model. No CNVs were found to be segregating in an autosomal recessive or X-linked fashion.

Disease model	Location	Ploidy	CNV-type	Software	Function
De-novo	chr2:177266000-177272000	0	DEL	CG	intergenic
De-novo	chr6:256000-292000	3	DUP	CG	intergenic
De-novo	chr6:62200000-62206000	0	DEL	CG	intergenic
De-novo	chr8:11895601-12091800	0	DEL	ERDS	DEFB130,FAM86B1,LOC100133267,USP17L2,USP17L7,ZNF705D
De-novo	chr11:50326000-50440000	3	DUP	CG	LOC646813:ncRNA
De-novo	chr16:33846000-33848000	3	DUP	CG	intergenic
De-novo	chr16:55796000-55822000	1	DEL	CG	CES1P1:ncRNA
De-novo	chr22:24274601-24398600	0	DEL	ERDS	DDT,DDTL,GSTT1,GSTT2,GSTT2B,LOC391322

Table S4. A table of prioritized genetic variations identified using sequencing data taken only from the two affected boys and their parents (a quad based study design). Variants conforming to the three disease models, de-novo, autosomal recessive and X-linked were identified. We show a list resulting from the CADD prioritization scheme as well as from the coding prioritization scheme. Both schemes required each variant to have a low population frequency (MAF < 1%). The coding scheme required all variants to also be within a coding region of the genome and to be a non-synonymous change. The CADD scheme requires all variants to have a CADD score of >20, along with the aforementioned population frequency.

CADD

Disease model	Location	Ref	Alt	CADD	Annotation software	Function
De-novo	15:38033077	T	-	25.6	GEMINI	intergenic
Autosomal recessive	1:10577762	AA	-	23.5	GEMINI	PEX14:intronic
Autosomal recessive	1:14257942	-	GT	21.7	ANNOVAR	intergenic
Autosomal recessive	1:25758431	TTTT	C	22.8	GEMINI	TMEM57:intronic
Autosomal recessive	1:70034794	ACTCA	C	20.4	GEMINI	intergenic
Autosomal recessive	1:72607649	-	A	20.6	ANNOVAR	NEGR1:intronic
Autosomal recessive	1:80084116	T	-	24	ANNOVAR, GEMINI	intergenic
Autosomal recessive	1:210851705	AC	-	27.5	ANNOVAR, GEMINI, SVS	KCNH1:UTR3
Autosomal recessive	1:224772442	ATTTG	-	22.1	GEMINI	intergenic
Autosomal recessive	2:60489199	A	-	21.5	ANNOVAR, SVS	intergenic
Autosomal recessive	2:60537356	TTTTATTT	ATTATTA	22.3	GEMINI	intergenic
Autosomal recessive	2:100765742	A	-	25.2	GEMINI	intergenic
Autosomal recessive	2:220966775	TTT	-	27.1	GEMINI	intergenic
Autosomal recessive	6:70553736	-	T	23.9	GEMINI	intergenic
Autosomal recessive	8:22568256	T	-	21.9	ANNOVAR, GEMINI, SVS	intergenic
Autosomal recessive	8:109098067	T	-	24.6	GEMINI	intergenic
Autosomal recessive	9:24823309	C	T	26.2	GEMINI, SVS	intergenic
Autosomal recessive	11:7183451	-	TCAAA	20.9	SVS	intergenic
Autosomal recessive	12:91451415	TT	-	21.1	GEMINI	KERA:intronic
Autosomal recessive	14:35346947	AATTAT	-	26.7	ANNOVAR, GEMINI, SVS	intergenic
Autosomal recessive	15:36869352	A	-	20.3	GEMINI	intergenic
Autosomal recessive	15:36884292	TT	-	20.4	GEMINI	C15orf41:intronic
Autosomal recessive	15:46849033	-	AA	21.1	ANNOVAR, GEMINI	intergenic
Autosomal recessive	15:60520690	ATAG	-	22.9	ANNOVAR, GEMINI, SVS	intergenic
Autosomal recessive	15:66786023	AAACA	-	23.6	GEMINI	SNAPC5:intronic
Autosomal recessive	15:86909147	CT	-	22.2	ANNOVAR, GEMINI, SVS	AGBL1:intronic
Autosomal recessive	16:49061338	T	-	25.3	ANNOVAR, GEMINI	intergenic
Autosomal recessive	16:49612366	AG	-	20.5	GEMINI, SVS	ZNF423:intronic
Autosomal recessive	16:54579809	A	-	25.3	GEMINI	intergenic
Autosomal recessive	17:55591582	-	T	23.3	GEMINI	MSI2:intronic
X-linked	X:5811532	GAG	CAA	21.2	GEMINI, SVS	nonframeshift substitution
X-linked	X:12410845	TATG	-	20.9	GEMINI	FRMPD4:intronic
X-linked	X:16023611	GT	-	20.3	GEMINI	intergenic

X-linked	X:16143148	TCTT	-	20.1	SVS	GRPR:intronic
X-linked	X:31200843	G	-	22.6	GEMINI	DMD:intronic
X-linked	X:38725226	-	CA	22.2	ANNOVAR, GEMINI, SVS	intergenic
X-linked	X:46410136	ATT	-	20.2	ANNOVAR, GEMINI, SVS	intergenic
X-linked	X:70621541	T	C	22.9	ANNOVAR, GEMINI, SVS	TAF1:NM_004606:I1337T
X-linked	X:71530110	TC	-	21.2	GEMINI	intergenic
X-linked	X:150248428	GT	-	23	SVS	intergenic

Coding

Disease model	**Location**	**Ref**	**Alt**	**Gene**	**Annotation software**	**Function**
De-novo	1:12887549	T	C	PRAMEF11	GEMINI	NM_001146344:E103G
De-novo	10:135438950	GGCCC	AGCCT	FRG2B	GEMINI, SVS	nonframeshift substitution
Autosomal recessive	1:53925370	-	CCCCGC	DMRTB1	ANNOVAR, GEMINI, SVS	nonframeshift insertion
Autosomal recessive	1:92944187	G	A	GFI1	GEMINI	NM_001127215:H350Y
Autosomal recessive	10:135438929	T	G	FRG2B	GEMINI	NM_001080998:I171L
Autosomal recessive	X:47307978	G	T	ZNF41	GEMINI, SVS	NM_153380:D397E
Autosomal recessive	X:63444792	C	A	ASB12	GEMINI	NM_130388:G247C
Autosomal recessive	X:70621541	T	C	TAF1	ANNOVAR, GEMINI, SVS	NM_004606:I1337T
X-linked	X:5811529	GAG	CAA	NLGN4X	GEMINI	nonframeshift substitution
X-linked	1:12885288	G	T	PRAMEF11	GEMINI	NM_001146344:S275T
X-linked	10:51859757	A	G	FAM21A	GEMINI	NM_001005751:K523R
X-linked	10:129901721	GGCAC	AGCAT	MKI67	ANNOVAR	nonframeshift substitution
X-linked	10:135438887	C	T	FRG2B	ANNOVAR, GEMINI, SVS	NM_001080998:A185T
X-linked	14:20666175	-	A	OR11G2	ANNOVAR, GEMINI, SVS	frameshift insertion
X-linked	X:34961491	T	C	FAM47B	GEMINI	NM_152631:K181N

Table S5. Summary of the clinical features of *TAF1* intellectual disability syndrome. This table demonstrates all shared clinical features across all affected individuals in the families, as well as other noted clinical characteristics on these individuals that are unique.

See Excel Spreadsheet.

Table S6. RNA sequencing library numbers, adapter sequences and general data quality statistics.

Sample	**I-1**	**I-2**	**II-2a**	**II-2b**	**II-1a**	**II-1b**	**U1a**	**U1b**	**U2a**	**U2b**
LID	295782	295783	295784	295785	295786	295787	295788	295789	295790	295791
Illumina index	AR001	AR002	AR003	AR004	AR005	AR006	AR007	AR008	AR009	AR010
Index sequence	ATCACG	CGATGT	TTAGGC	TGACCA	ACAGTG	GCCAAT	CAGATC	ACTTGA	GATCAG	TAGCTT
Mapping percent	66.44	88.43	87.88	91.73	89.62	84.76	83.92	84.92	88.01	88.37
Properly paired reads	34561120	50224027	50812379	66555506	60522016	35767249	45865914	60357217	35409852	57851237

Table S7. A list of differentially expressed genes from the RNA sequencing study using samples from family 1.

Gene	log2 fold change	p value	q value
ATAD3C	-2.87112	5.00E-05	0.00574912
TNFRSF8	-1.04439	5.00E-05	0.00574912
TMEM51	-1.84068	0.00015	0.0139211
PADI4	1.01336	8.00E-04	0.048346
C1QA	-2.67333	5.00E-05	0.00574912
C1QB	-1.92782	5.00E-05	0.00574912
NDUFS5	-1.05429	1.00E-04	0.0102307
C1orf228	-1.39134	5.00E-05	0.00574912
LOC729041	-1.94259	5.00E-05	0.00574912
FCRL6	-1.26542	5.00E-05	0.00574912
HSPA7	-1.51413	5.00E-05	0.00574912
XCL1	-1.53244	5.00E-05	0.00574912
SHISA4	-2.70934	5.00E-05	0.00574912
LGR6	-1.43213	5.00E-05	0.00574912
ADORA1	-1.91863	2.00E-04	0.0170961
SNRPE	-1.05709	0.00035	0.0261353
CR2	1.27743	2.00E-04	0.0170961
C1orf115	-1.40868	1.00E-04	0.0102307
TNFRSF4	-1.05997	1.00E-04	0.0102307
ID3	0.861095	3.00E-04	0.0233407
ZNF683	1.8759	5.00E-05	0.00574912
SDC3	-2.18599	5.00E-05	0.00574912
SLC1A7	-2.04795	5.00E-05	0.00574912
TACSTD2	-2.41214	5.00E-05	0.00574912
GBP5	-0.859719	0.00025	0.0203016
FCGR1B	-1.28223	0.00025	0.0203016
SELENBP1	1.12944	5.00E-04	0.0336606
XCL2	-1.0419	0.00055	0.0364599
PIGR	-4.61117	0.00025	0.0203016
RPS24	-2.15991	5.00E-05	0.00574912
DPYSL4	-4.83167	2.00E-04	0.0170961
DIP2C	-1.29248	0.00055	0.0364599
PCBD1	-1.17198	0.00025	0.0203016
DLG5	-1.40821	5.00E-05	0.00574912
TSPAN4	-1.05754	0.00035	0.0261353
PRRG4	0.965322	4.00E-04	0.0290346
SERPING1	-1.19021	5.00E-05	0.00574912
HBG1	-4.32488	5.00E-05	0.00574912
HBG2	-2.42133	5.00E-05	0.00574912

DKK3	-3.16059	5.00E-05	0.00574912
KCNJ11	-3.58184	5.00E-05	0.00574912
CD248	1.2272	5.00E-05	0.00574912
CASP5	-1.60506	5.00E-04	0.0336606
CARD16	-1.86252	5.00E-05	0.00574912
IL18	-1.22313	0.00045	0.0314347
GPR162	-1.24786	5.00E-05	0.00574912
LEPREL2	-1.71941	0.00015	0.0139211
GNB3	-1.91873	0.00035	0.0261353
CLEC4D	-1.52432	1.00E-04	0.0102307
CLEC9A	-1.92409	1.00E-04	0.0102307
PFDN5	-1.31396	5.00E-05	0.00574912
KLRB1	-1.25551	5.00E-05	0.00574912
CLECL1	-1.98607	0.00045	0.0314347
CLEC2B	-1.51225	5.00E-05	0.00574912
CLEC7A	-1.00512	4.00E-04	0.0290346
KRT72	-1.0832	0.00035	0.0261353
KRT1	-2.36912	5.00E-05	0.00574912
RPL21	-0.995223	5.00E-05	0.00574912
COMMD6	-1.07029	1.00E-04	0.0102307
RNASE3	-1.87241	5.00E-05	0.00574912
TMEM63C	-1.96086	5.00E-05	0.00574912
SLIRP	-1.32867	0.00025	0.0203016
IFI27	1.81763	0.00025	0.0203016
KIAA0125	1.19957	5.00E-05	0.00574912
RPPH1	-2.53142	5.00E-05	0.00574912
GZMH	-1.62045	5.00E-05	0.00574912
GZMB	-1.18755	5.00E-05	0.00574912
NDUFB1	-1.55007	5.00E-05	0.00574912
C14orf2	-1.22803	5.00E-05	0.00574912
NUDT14	-1.08596	2.00E-04	0.0170961
GOLGA8A	-1.10963	0.00015	0.0139211
RBPMS2	-2.28351	5.00E-05	0.00574912
BCL2A1	-1.9344	5.00E-05	0.00574912
HBQ1	-1.22025	5.00E-05	0.00574912
PRSS21	-2.65918	5.00E-05	0.00574912
ZG16B	-1.60549	8.00E-04	0.048346
CMTM2	-1.24795	5.00E-05	0.00574912
PRSS30P	-1.77163	5.00E-05	0.00574912
NPIPA5	-3.04186	5.00E-05	0.00574912
LOC643802	-1.62919	5.00E-05	0.00574912

SERPINF1	-1.1821	1.00E-04	0.0102307
LOC440461	-1.52764	3.00E-04	0.0233407
KCNJ2	-1.06415	0.00015	0.0139211
KIF19	-1.64362	7.00E-04	0.0437264
SCARF1	-1.03786	1.00E-04	0.0102307
RPL26	-1.9475	5.00E-05	0.00574912
RPL23	-1.29913	5.00E-05	0.00574912
LOC644172	1.44569	7.00E-04	0.0437264
COPZ2	-1.66441	0.00025	0.0203016
PTPRM	-1.55076	0.00015	0.0139211
RPL17	-1.25555	5.00E-05	0.00574912
MRPL54	-1.36795	5.00E-05	0.00574912
SHD	-2.29846	5.00E-05	0.00574912
ZNF155	-2.66016	5.00E-05	0.00574912
CEACAM19	-1.41753	1.00E-04	0.0102307
TRPM4	-2.16561	5.00E-05	0.00574912
FPR2	-1.08469	0.00035	0.0261353
CACNG6	-1.69296	0.00015	0.0139211
KIR2DL1	-1.62018	0.00045	0.0314347
KIR3DL1	-1.84708	5.00E-05	0.00574912
KIR2DS4	-1.68174	0.00015	0.0139211
PRSS57	-1.65305	5.00E-05	0.00574912
ADAMTSL5	-1.55167	3.00E-04	0.0233407
SMIM24	-2.02715	5.00E-05	0.00574912
SEMA6B	-2.65273	5.00E-05	0.00574912
PTPRS	-1.31403	0.00015	0.0139211
COL5A3	-1.78196	5.00E-05	0.00574912
PLA2G4C	1.19294	0.00015	0.0139211
HSD17B14	-2.18426	5.00E-04	0.0336606
LIM2	-1.93953	0.00045	0.0314347
LILRA4	-1.73447	5.00E-05	0.00574912
TNNT1	-3.23274	5.00E-05	0.00574912
RPS7	-1.2135	5.00E-05	0.00574912
SPTBN1	0.938105	2.00E-04	0.0170961
LOC100507006	-1.06325	0.00035	0.0261353
RPL31	-1.69122	5.00E-05	0.00574912
DBI	-1.07164	5.00E-05	0.00574912
NDUFB3	-1.27689	5.00E-05	0.00574912
STRADB	0.852128	0.00075	0.0459658
EEF1B2	-1.16347	5.00E-05	0.00574912
CYP27A1	-1.07854	5.00E-05	0.00574912

POMC	-1.46799	5.00E-05	0.00574912
SNRPG	-1.3855	5.00E-05	0.00574912
NMUR1	-1.20728	5.00E-05	0.00574912
TGM3	1.20048	5.00E-05	0.00574912
PI3	-1.18351	5.00E-05	0.00574912
OPRL1	0.858477	6.00E-04	0.0391749
CD93	-1.14257	5.00E-05	0.00574912
SLPI	-1.19749	1.00E-04	0.0102307
MAP3K7CL	-0.977451	0.00065	0.0413993
LINC00189	-4.15011	4.00E-04	0.0290346
COL6A1	-2.97605	5.00E-05	0.00574912
COL6A2	-1.3214	5.00E-05	0.00574912
TEKT4P2	-1.10974	2.00E-04	0.0170961
S100B	-1.82951	5.00E-05	0.00574912
BMS1P20	1.04278	1.00E-04	0.0102307
IGLL5	1.46423	5.00E-05	0.00574912
CACNA1I	-2.9258	5.00E-05	0.00574912
RBX1	-0.946039	4.00E-04	0.0290346
CECR6	-1.46779	5.00E-05	0.00574912
VPREB3	0.842423	7.00E-04	0.0437264
LARGE	-1.49021	2.00E-04	0.0170961
PVALB	-1.94211	2.00E-04	0.0170961
LGALS2	-1.20562	0.00015	0.0139211
SHISA8	-1.90778	0.00015	0.0139211
LSM3	-1.77289	5.00E-05	0.00574912
CCR2	-0.902303	2.00E-04	0.0170961
TMA7	-1.59142	5.00E-05	0.00574912
CACNA2D3	-1.17562	8.00E-04	0.048346
GRAMD1C	-1.86215	5.00E-05	0.00574912
CSTA	-0.979863	0.00035	0.0261353
LAMB2	-2.14303	5.00E-05	0.00574912
CXCL8	-2.35914	5.00E-05	0.00574912
RPL34	-2.00543	5.00E-05	0.00574912
SNHG8	-1.80959	5.00E-05	0.00574912
ATP5I	-1.28951	5.00E-05	0.00574912
SPON2	-0.873661	0.00045	0.0314347
FGFBP2	-0.841868	5.00E-04	0.0336606
TLR6	-0.902228	0.00045	0.0314347
HOPX	-0.99001	1.00E-04	0.0102307
IGJ	1.15773	5.00E-05	0.00574912
ROPN1L	0.991963	0.00065	0.0413993

SUB1	-0.901715	0.00065	0.0413993
NDUFS4	-1.21626	1.00E-04	0.0102307
GZMA	-0.848316	6.00E-04	0.0391749
F2RL1	-1.32069	5.00E-05	0.00574912
COX7C	-1.04086	5.00E-05	0.00574912
ERAP2	-0.997813	3.00E-04	0.0233407
SYNPO	1.4416	5.00E-05	0.00574912
GPX3	-1.34279	1.00E-04	0.0102307
HIGD2A	-0.84974	4.00E-04	0.0290346
HINT1	-0.904249	2.00E-04	0.0170961
MZB1	1.07242	5.00E-05	0.00574912
ZNF300	-2.37759	5.00E-05	0.00574912
HLA-G	1.17109	7.00E-04	0.0437264
DDX43	-2.23091	5.00E-04	0.0336606
FAM26F	-1.33926	5.00E-05	0.00574912
LINC01013	-3.92321	0.00015	0.0139211
MLLT4	-2.30667	5.00E-05	0.00574912
ETV7	-1.58053	2.00E-04	0.0170961
CPNE5	0.803557	6.00E-04	0.0391749
LRRN3	1.14407	0.00055	0.0364599
NDUFB2	-0.944644	0.00015	0.0139211
AOC1	-1.8003	5.00E-05	0.00574912
TOMM7	-1.26135	5.00E-05	0.00574912
HOXA10	1.72339	2.00E-04	0.0170961
IGFBP3	-2.5023	5.00E-05	0.00574912
KIAA1324L	1.89842	5.00E-05	0.00574912
SHFM1	-1.19882	0.00025	0.0203016
EPHX2	0.924312	5.00E-04	0.0336606
DNAJC5B	-3.16812	5.00E-05	0.00574912
NOV	1.02823	0.00075	0.0459658
RPL7	-1.11852	5.00E-05	0.00574912
MRPS28	-1.1122	0.00075	0.0459658
CA1	1.74612	5.00E-05	0.00574912
UQCRB	-1.91418	5.00E-05	0.00574912
COX6C	-1.32948	5.00E-05	0.00574912
MVB12B	-0.971247	0.00065	0.0413993
CERCAM	-2.06868	5.00E-05	0.00574912
PTGDS	-1.48965	5.00E-05	0.00574912
RMRP	-3.55787	5.00E-05	0.00574912
GOLM1	-1.77318	5.00E-05	0.00574912
GGTA1P	-1.3323	3.00E-04	0.0233407

LRRC26	-2.33032	0.00065	0.0413993
COX7B	-1.48867	5.00E-05	0.00574912
RPL36A	-1.6161	5.00E-05	0.00574912
NGFRAP1	0.969799	5.00E-04	0.0336606
LINC00892	-1.77887	0.00045	0.0314347
ALAS2	1.58355	5.00E-05	0.00574912
XIST	-10.1293	5.00E-05	0.00574912
RPL39	-0.957696	0.00075	0.0459658
PRKY	1.06368	3.00E-04	0.0233407
KDM5D	1.1741	3.00E-04	0.0233407

Supplemental References

1. Allen, R.C., Zoghbi, H., Moseley, A., Rosenblatt, H., and Belmont, J. (1992). Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation. Am J Hum Genet 51, 1229.
2. Fang, H., Wu, Y., Narzisi, G., O'Rawe, J.A., Barron, L.T., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M.C., and Lyon, G.J. (2014). Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med 6, 89.
3. Narzisi, G., O'Rawe, J.A., Iossifov, I., Fang, H., Lee, Y.H., Wang, Z., Wu, Y., Lyon, G.J., Wigler, M., and Schatz, M.C. (2014). Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nature methods 11, 1033-1036.
4. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17, 1665-1674.
5. Rope, A.F., Wang, K., Evjenth, R., Xing, J., Johnston, J.J., Swensen, J.J., Johnson, W.E., Moore, B., Huff, C.D., Bird, L.M., et al. (2011). Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. Am J Hum Genet 89, 28-43.
6. Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., and Reese, M.G. (2011). A probabilistic disease-gene finder for personal genomes. Genome research 21, 1529-1542.
7. Hu, H., Huff, C.D., Moore, B., Flygare, S., Reese, M.G., and Yandell, M. (2013). VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. Genet Epidemiol 37, 622-634.
8. Kennedy, B., Kronenberg, Z., Hu, H., Moore, B., Flygare, S., Reese, M.G., Jorde, L.B., Yandell, M., and Huff, C. (2014). Using VAAST to Identify Disease-Associated Variants in Next-Generation Sequencing Data. Curr Protoc Hum Genet 81, 6 14 11-16 14 25.
9. Christensen, G.B., and Lambert, C.G. (2011). Search for compound heterozygous effects in exome sequence of unrelated subjects. BMC Proc 5 Suppl 9, S95.
10. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164.
11. Paila, U., Chapman, B.A., Kirchner, R., and Quinlan, A.R. (2013). GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. PLoS Comput Biol 9, e1003153.
12. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367, 1921-1929.

13. Deciphering Developmental Disorders, S. (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature 519, 223-228.
14. Hufnagel, R.B., Arno, G., Hein, N.D., Hersheson, J., Prasad, M., Anderson, Y., Krueger, L.A., Gregory, L.C., Stoetzel, C., Jaworek, T.J., et al. (2015). Neuropathy target esterase impairments cause Oliver-McFarlane and Laurence-Moon syndromes. Journal of medical genetics 52, 85-94.
15. Tzetis, M., Kitsiou-Tzeli, S., Frysira, H., Xaidara, A., and Kanavakis, E. (2012). The clinical utility of molecular karyotyping using high-resolution array-comparative genomic hybridization. Expert Rev Mol Diagn 12, 449-457.
16. Amos-Landgraf, J.M., Cottle, A., Plenge, R.M., Friez, M., Schwartz, C.E., Longshore, J., and Willard, H.F. (2006). X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. Am J Hum Genet 79, 493-499.
17. Garel, C., Cont, I., Alberti, C., Josserand, E., Moutard, M.L., and Ducou le Pointe, H. (2011). Biometry of the corpus callosum in children: MR imaging reference data. AJNR American journal of neuroradiology 32, 1436-1443.
18. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nature methods 7, 248-249.
19. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 4, 1073-1081.
20. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6, e1001025.
21. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20, 110-121.
22. He, M., Person, T.N., Hebbring, S.J., Heinzen, E., Ye, Z., Schrodi, S.J., McPherson, E.W., Lin, S.M., Peissig, P.L., Brilliant, M.H., et al. (2015). SeqHBase: a big data toolset for family based sequencing data analysis. Journal of medical genetics 52, 282-288.
23. O'Rawe, J., Guangqing, S., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, E., Wei, Z., Jiang, T., et al. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5, 28.
24. (Broad Institute). Local re-assembly and haplotype determination by HaplotypeCaller. In. (
25. Narzisi, G., Rawe, J.A., Iossifov, I., Lee, Y.-h., Wang, Z., Wu, Y., Lyon, G.J., Wigler, M., and Schatz, M.C. (2013). Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly. bioRxiv.
26. Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A., and Mittelman, D. (2013). Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Research 41, e32.
27. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.

28. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Research 17, 1665-1674.
29. Zhu, M., Need, Anna C., Han, Y., Ge, D., Maia, Jessica M., Zhu, Q., Heinzen, Erin L., Cirulli, Elizabeth T., Pelak, K., He, M., et al. (2012). Using ERDS to Infer Copy-Number Variants in High-Coverage Genomes. The American Journal of Human Genetics 91, 408-421.
30. Lyon, G.J., and O'Rawe, J. (2014). Human genetics and clinical aspects of neurodevelopmental disorders.
31. Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421-427.
32. Gaugler, T., Klei, L., Sanders, S.J., Bodea, C.A., Goldberg, A.P., Lee, A.B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., et al. (2014). Most genetic risk for autism resides with common variation. Nat Genet 46, 881-885.
33. Cech, T.R., and Steitz, J.A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157, 77-94.
34. Ezkurdia, I., Vazquez, J., Valencia, A., and Tress, M. (2014). Analyzing the First Drafts of the Human Proteome. Journal of proteome research.
35. Shanske, A.L., Goodrich, J.T., Ala-Kokko, L., Baker, S., Frederick, B., and Levy, B. (2012). Germline mosacism in Shprintzen-Goldberg syndrome. Am J Med Genet A 158A, 1574-1578.
36. Slavin, T.P., Lazebnik, N., Clark, D.M., Vengoechea, J., Cohen, L., Kaur, M., Konczal, L., Crowe, C.A., Corteville, J.E., Nowaczyk, M.J., et al. (2012). Germline mosaicism in Cornelia de Lange syndrome. Am J Med Genet A 158A, 1481-1485.
37. Meyer, K.J., Axelsen, M.S., Sheffield, V.C., Patil, S.R., and Wassink, T.H. (2012). Germline mosaic transmission of a novel duplication of PXDN and MYT1L to two male half-siblings with autism. Psychiatr Genet 22, 137-140.
38. Campbell, I.M., Yuan, B., Robberecht, C., Pfundt, R., Szafranski, P., McEntagart, M.E., Nagamani, S.C., Erez, A., Bartnik, M., Wisniowiecka-Kowalnik, B., et al. (2014). Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. Am J Hum Genet 95, 173-182.
39. McCarthy, D., Humburg, P., Kanapin, A., Rivas, M., Gaulton, K., Consortium, T.W., Cazier, J.-B., and Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. Genome Med 6, 26.
40. Niederriter, A.R., Davis, E.E., Golzio, C., Oh, E.C., Tsai, I.C., and Katsanis, N. (2013). In Vivo Modeling of the Morbid Human Genome using Danio rerio. e50338.
41. Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., and Reese, M.G. (2011). A probabilistic disease-gene finder for personal genomes. Genome Res 21, 1529-1542.
42. Reese, M.G., Moore, B., Batchelor, C., Salas, F., Cunningham, F., Marth, G.T., Stein, L., Flicek, P., Yandell, M., and Eilbeck, K. (2010). A standard variation file format for human genome sequences. Genome biology 11, R88.

43. Cucala, L. (2008). A Hypothesis-Free Multiple Scan Statistic with Variable Window. Biometrical Journal 50, 299-310.
44. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat Meth 12, 357-360.
45. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotech 33, 290-295.
46. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protocols 7, 562-578.
47. L. Goff, C.T., D. Kelley. (2013). cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. R package version 2100.
48. Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Research 41, W77-W83.
49. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102, 15545-15550.
50. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research 42, D199-D205.
51. Consortium, T.G.O. (2015). Gene Ontology Consortium: going forward. Nucleic Acids Research 43, D1049-D1056.
52. Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C.M., Brown, D.L., Brudno, M., Campbell, J., et al. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Research 42, D966-D974.
53. Koolen, D.A., Kramer, J.M., Neveling, K., Nillesen, W.M., Moore-Barton, H.L., Elmslie, F.V., Toutain, A., Amiel, J., Malan, V., Tsai, A.C., et al. (2012). Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. Nat Genet 44, 639-641.
54. Zollino, M., Orteschi, D., Murdolo, M., Lattante, S., Battaglia, D., Stefanini, C., Mercuri, E., Chiurazzi, P., Neri, G., and Marangi, G. (2012). Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. Nat Genet 44, 636-638.
55. Kitsiou-Tzeli, S., Frysira, H., Giannikou, K., Syrmou, A., Kosma, K., Kakourou, G., Leze, E., Sofocleous, C., Kanavakis, E., and Tzetis, M. (2012). Microdeletion and microduplication 17q21.31 plus an additional CNV, in patients with intellectual disability, identified by array-CGH. Gene 492, 319-324.
56. O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., MacKenzie, A.P., Ng, S.B., Baker, C., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet 43, 585-589.

57. Neale, B.M., Kou, Y., Liu, L., Ma/'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485, 242-245.
58. O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485, 246-250.
59. Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A., et al. (2012). De novo gene disruptions in children on the autistic spectrum. Neuron 74, 285-299.
60. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature 511, 344-347.
61. Shoichet, S.A., Hoffmann, K., Menzel, C., Trautmann, U., Moser, B., Hoeltzenbein, M., Echenne, B., Partington, M., van Bokhoven, H., Moraine, C., et al. (2003). Mutations in the ZNF41 Gene Are Associated with Cognitive Deficits: Identification of a New Candidate for X-Linked Mental Retardation. The American Journal of Human Genetics 73, 1341-1354.
62. Piton, A., Redin, C., and Mandel, J.L. (2013). XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. Am J Hum Genet 93, 368-383.
63. Ioannidis, J.P., Greenland, S., Hlatky, M.A., Khoury, M.J., Macleod, M.R., Moher, D., Schulz, K.F., and Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. Lancet 383, 166-175.
64. Richards, M., Coppee, F., Thomas, N., Belayew, A., and Upadhyaya, M. (2012). Facioscapulohumeral muscular dystrophy (FSHD): an enigma unravelled? Human genetics 131, 325-340.
65. Rijkers, T., Deidda, G., van Koningsbruggen, S., van Geel, M., Lemmers, R.J., van Deutekom, J.C., Figlewicz, D., Hewitt, J.E., Padberg, G.W., Frants, R.R., et al. (2004). FRG2, an FSHD candidate gene, is transcriptionally upregulated in differentiating primary myoblast cultures of FSHD patients. Journal of medical genetics 41, 826-836.
66. Gabellini, D., Green, M.R., and Tupler, R. (2002). Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. Cell 110, 339-348.
67. Chaki, M., Airik, R., Ghosh, A.K., Giles, R.H., Chen, R., Slaats, G.G., Wang, H., Hurd, T.W., Zhou, W., Cluckey, A., et al. (2012). Exome capture reveals ZNF423 and CEP164 mutations, linking renal ciliopathies to DNA damage response signaling. Cell 150, 533-548.
68. Gupta, R.K., Arany, Z., Seale, P., Mepani, R.J., Ye, L., Conroe, H.M., Roby, Y.A., Kulaga, H., Reed, R.R., and Spiegelman, B.M. (2010). Transcriptional control of preadipocyte determination by Zfp423. Nature 464, 619-623.
69. Ritzel, M.W., Yao, S.Y., Huang, M.Y., Elliott, J.F., Cass, C.E., and Young, J.D. (1997). Molecular cloning and functional expression of cDNAs encoding a human Na+-nucleoside cotransporter (hCNT1). The American journal of physiology 272, C707-714.

70. Henry, R.W., Mittal, V., Ma, B., Kobayashi, R., and Hernandez, N. (1998). SNAP19 mediates the assembly of a functional core promoter complex (SNAPc) shared by RNA polymerases II and III. Genes & Development 12, 2664-2672.
71. Houdayer, C. (2011). In Silico Prediction of Splice-Affecting Nucleotide Variants. In In Silico Tools for Gene Discovery, B. Yu and M. Hinchcliffe, eds. (Humana Press), pp 269-281.
72. Yeo, G., and Burge, C.B. (2004). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. Journal of Computational Biology 11, 377-394.
73. Pertea, M., Lin, X., and Salzberg, S.L. (2001). GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Research 29, 1185-1190.
74. Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Research 37, e67.