A variant in *TAF1* is associated with a new syndrome with severe intellectual disability and characteristic dysmorphic features

JASON A. O'RAWE[1,2,#], YIYANG WU[1,2,#], ALAN ROPE[3], LAURA T. JIMENEZ BARRÓN[1,4], JEFFREY SWENSEN[5], HAN FANG[1], DAVID MITTELMAN[6,7], GARETH HIGHNAM[6], REID ROBISON[7,8], EDWARD YANG[9], KAI WANG[7,8,10], AND GHOLSON J. LYON[1,2,8]

[1] *Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY, USA*
[2] *Stony Brook University, Stony Brook, NY, USA*
[3] *Department of Medical Genetics, Northwest Kaiser Permanente, Portland, OR, USA*
[4] *Centro de Ciencias Genomicas, Universidad Nacional Autonoma de Mexico, Cuernavaca, Morelos, MX*
[5] *Caris Life Sciences, Phoenix, Arizona, USA*
[6] *Gene by Gene, Ltd., Houston, TX, USA*
[7] *Tute Genomics, Ltd., Provo, UT, USA*
[8] *Utah Foundation for Biomedical Research, Salt Lake City, UT, USA*
[9] *Department of Radiology, Boston Children's Hospital, Boston, MA, USA*
[10] *Zilkha Neurogenetic Institute, Department of Psychiatry and Preventive Medicine, University of Southern California, Los Angeles, CA, USA*

[#] *Co-first authors*
[*] *Corresponding author: glyon@cshl.edu*

Compiled January 20, 2015

We describe the discovery of a new genetic syndrome, RykDax syndrome, driven by a whole genome sequencing (WGS) study of one family from Utah with two affected male brothers, presenting with severe intellectual disability (ID), a characteristic intergluteal crease, and very distinctive facial features including a broad, upturned nose, sagging cheeks, downward sloping palpebral fissures, prominent periorbital ridges, deep-set eyes, relative hypertelorism, thin upper lip, a high-arched palate, prominent ears with thickened helices, and a pointed chin. This Caucasian family was recruited from Utah, USA. Illumina-based WGS was performed on 10 members of this family, with additional Complete Genomics-based WGS performed on the nuclear portion of the family (mother, father and the two affected males). Using WGS datasets from 10 members of this family, we can increase the reliability of the biological inferences with an integrative bioinformatic pipeline. In combination with insights from clinical evaluations and medical diagnostic analyses, these DNA sequencing data were used in the study of three plausible genetic disease models that might uncover genetic contribution to the syndrome. We found a 2 to 5-fold difference in the number of variants detected as being relevant for various disease models when using different sets of sequencing data and analysis pipelines. We derived greater accuracy when more pipelines were used in conjunction with data encompassing a larger portion of the family, with the number of putative de-novo mutations being reduced by 80%, due to false negative calls in the parents. The boys carry a maternally inherited missense variant in a X-chromosomal gene *TAF1*, which we consider as disease relevant. TAF1 is the largest subunit of the general transcription factor IID (TFIID) multi-protein complex, and our results implicate mutations in *TAF1* as playing a critical role in the development of this new intellectual disability syndrome.

1. INTRODUCTION

Dramatic cost reductions and rapid advancements in the development of efficient sequencing technologies [1–3] have led to widespread use of exome and whole genome sequencing (WGS) [4–6] in a variety of research and clinical settings [7]. Computational tools for processing and analyzing these sequence data have been developed in parallel, and many are now freely available and straightforward to implement [4]. In this context, biomedical exome sequencing and WGS has led to the discov-

ery of the genetic basis for many conditions, including Miller Syndrome [8] and others [9].

The intellectual impetus for this study included how to identify the major genetic contribution to a particular syndrome in only one proband or two affected siblings, in the absence of other affected people with the same known syndrome. This task is much easier in the presence of multiple affected people spread out over two or more generations [10, 11], but this is usually not the case for most biomedical presentations of idiopathic disorders. Genetic discovery can also be easier in consanguineous pedigrees with autosomal recessive conditions [12, 13], but such consanguinity is not widespread in many parts of the world [14]. We instead focused on an analysis that initially included one family with only two siblings that are both afflicted by an idiopathic syndrome, and so we utilized WGS to cover as much of the genome as we could, particularly as any number of mutated nucleotides in the genome might influence some phenotype during embryogenesis or postnatal life [15–33].

This Caucasian family was recruited from Utah, USA. Illumina-based WGS was performed on 10 members of this family, with additional Complete Genomics-based WGS performed on the nuclear portion of the family (mother, father and the two affected males). Sequence data were processed with a number of bioinformatics pipelines. Using comprehensive datasets generated by an aggregation of results stemming from these pipelines, we can increase the reliability of the biological inferences stemming from these data. In combination with insights from clinical evaluations and medical diagnostic analyses, these DNA sequencing data were used in the study of three plausible genetic disease models that might uncover genetic contribution to the syndrome.

2. MATERIALS AND METHODS

This study consists of two methodological components, clinical and genomic sequencing/analysis. The clinical component includes research participant enrollment, clinical evaluation, diagnostic analyses and a detailed clinical report. The genomic sequencing component includes whole genome sequencing as well as downstream analyses aimed at annotating and assigning biological function to detected genomic variants.

A. Clinical methods

A.1. Enrollment of research participants

The collection and the analysis of DNA was conducted by the Utah Foundation for Biomedical Research, as approved by the Institutional Review Board (IRB) (Plantation, Florida). Written informed consent was also obtained from all study participants, and research was carried out in compliance with the Helsinki Declaration.

A.2. Clinical evaluation/diagnostics

A broad range of clinical diagnostic testing was performed on both affected male siblings, including karyotyping, a high resolution X-chromosome CGH array (720K Chromosome X Specific Array from Roche NimbleGen, Inc. USA), subtelomeric FISH study, methylation study for Angelman syndrome, *XNP* sequencing for ATRX, and fragile X DNA testing. In addition, we performed diagnostics on serum amino acid levels, urine organic acids levels, sweat chloride levels, plasma carnitine profile, and immunoglobulin levels. We also performed urine mucopolysaccharidosis (MPS) screening and examined thyroid profiles. Cranial ultrasound was performed and brain imagery was obtained using magnetic resonance imaging (MRI) and computed tomography (CT) scanning techniques. Images of the spine were also obtained using MRI. Moreover, cerebrospinal fluid (CSF) was collected from the elder sibling, and neurotransmitter metabolites, tetrahydrobiopterin (BH4) and neopterin (N) profile were screened.

A.3. Custom X CGH array and X-chromosome skewing assay

500 ng of each research participant's DNA was labeled with 5′-Cy3 tagged nanomers (NimbleGen) while a female control was labeled with Cy5 nonamers. After purification by isopropanol precipitation, 31 ug each of labeled research participant and reference DNA were combined. The mixture was hybridized to a custom NimbleGen 720K Chromosome X Specific Array for 42 hours at 42C in a MAUI Hybridization System (BioMicro Systems). The array was then washed according to the manufacturer's recommendation (Nimblegen) and immediately scanned. After scanning, fluorescence intensity raw data was extracted from the scanned images of the array using NimbleScan v2.6 software. For each of the spots on the array, normalized log2 ratios of the Cy3-labeled research participant sample vs the Cy5-reference sample were generated using the SegMNT program. The data was visualized with Nexus 6.1 software (Biodiscovery).

X-chromosome skewing assay analyses were performed using an adaptation of the technique described by Allen et al. (1992) [34].

B. Whole Genome Sequencing and analysis methods

Two different sequencing strategies were employed. Initial sequencing efforts focused on whole genome sequencing using the Complete Genomics (CG) sequencing and analysis pipeline v2.0 for the mother, father and two affected boys. Additional whole genome sequencing was performed subsequent to this initial effort, using the Illumina HiSeq 2000 sequencing platform. The mother, father, two affected boys and six other immediate family members were sequenced using the Illumina HiSeq 2000. Raw sequencing data stemming from Illumina sequencing was processed by a variety of analysis pipelines and subsequently pooled with variants detected by the CG sequencing an analysis pipeline. Downstream and functional annotation tools were used to evaluate variants detected by the various methods.

B.1. Complete Genomics whole genome sequencing and variant detection

After quality control to ensure lack of genomic degradation, we sent 10ug DNA samples 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.

B.2. Illumina HiSeq 2000 whole genome sequencing and variant detection

After the samples were quantified using Qubit dsDNA BR Assay Kit (Invitrogen), 1ug of each sample was sent out for whole genome sequencing using the Illumina® Hiseq 2000 platform. Sequencing libraries were generated from 100ng 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 100bp flow cell. The number of clusters passing initial filtering was above 80%, and the number of bases at or above Q30 was above 85%.

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

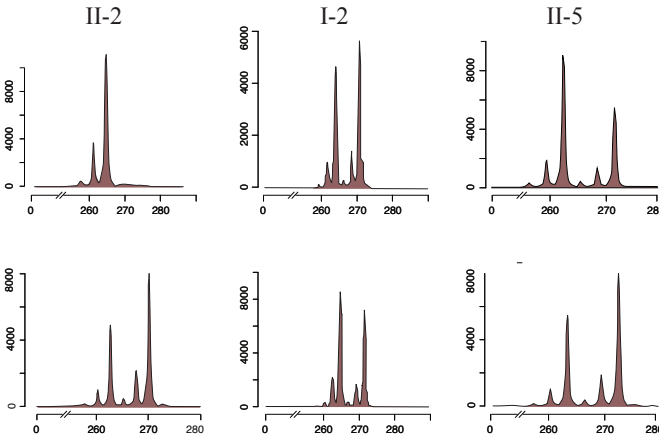

Figure 1. X-chromosome inactivation assays were performed on the mother (**II-2**), aunt (**II-5**), and maternal grandmother (**I-2**) of the two affected siblings. The assay reveals normal X-chromosome inactivation ratios in the aunt and maternal grandmother, however the mother of the affected siblings exhibits a 99:1 X-chromosome inactivation ratio.

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. 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).

We used several methods to prioritize and identify possible disease-contributory germ-line variants, including VAAST [10, 35], Golden Helix SVS v8.1.4 [36], ANNOVAR (2013Aug23 version) [37], and GEMINI v0.9.1 [38]. 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.).

We used two distinct variant prioritization schemes. The first scheme, which we will refer to as the 'coding' scheme, requires all variants to be within a coding region of the genome. Splice site variants are also included. The second scheme, which we will refer to as the 'CADD' scheme, requires all variants to have a CADD score of >20. CADD scores do not preclude non-coding genetic variants from the resulting list of potentially deleterious variants. Both schemes required each variant to have a low population frequency (MAF < 1%). Prioritized variants were manually verified by inspecting sequence alignments using Golden Helix GenomeBrowse v2.0.3 [http://www.goldenhelix.com/GenomeBrowse/index.html]. A signal was considered plausible if 4 or more reads supported the alternative allele in more than one family member. Population frequency information was corroborated using the NCBI Variation viewer [http://www.ncbi.nlm.nih.gov/variation/view/]. See the Supplementary Information for details about the variant calling and prioritization analyses.

3. RESULTS

A. Clinical evaluations and phenotypic presentation

The initial probands selected for study by the corresponding author (GJL) were 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. Among the facial features are a broad, upturned nose, sagging cheeks, downward sloping palpebral fissures, prominent periorbital ridges, deep-set eyes, relative hypertelorism, thin upper lip, a high-arched palate, prominent ears with thickened helices, and a pointed chin. Other shared phenotypic symptoms include strabismus (exotropia), blocked tear ducts, microcephaly, mild ventriculomegaly, deficiency of the septum pellucidum, hypoplasia of the corpus callosum, low cerebral white matter volume, 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 frequent episodes of contact dermatitis and eczema, scoliosis, sleep-wake dysregulation, as well as asthma, although he no longer requires medication for the latter. The elder brother, on the other hand, has diplegia, and has received Botulinum Toxin (Botox) therapy for his lower-extremity spasticity for six years. A review of systems (ROS) questionnaire revealed no other obvious, shared or otherwise, symptoms or malformations. See **Table 1** for a summary of the clinical features of the affected male siblings.

The parents of the two affected siblings are non-consanguineous and are both healthy. The mother has been evaluated for PKU and had 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 X-chromosome skewing assay revealed that the mother of the two affected boys has skewed, 99:1, X-chromosome inactivation. The grandmother, as well as the aunt of the affected boys, does not show any appreciable X-chromosome skewing (**Figure 1**), which suggested the possibility of a newly arising deleterious X-chromosome variant, although it is well established that this also could be non-specific [39].

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 any alcohol or drug use, nor any exposure to environmental toxins during the course of both pregnancies. 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 old. The younger boy was born in the 37th gestational week with a birth weight of 1.76 kg. A heart murmur was noticed at his birth, but echocardiography confirmed the absence of any further or more serious cardiovascular abnormalities. He was treated with light for neonatal jaundice, and required a feeding tube during the first few days of his life due to difficulties swallowing and digesting food. During the most recent examinations, the younger boy (aged 1011/12 years) had a height of 129.7 cm (2% tile), a weight of 30.8 kg (19% tile, BMI 18.3 kg/m2), and his occipital

Table 1. Summary of the Clinical Features of RykDax Syndrome

Systems	Features	Utah family	
		III-1 (elder)	III-2 (younger)
Genetic Studies	Karyotype	46,X,inv(Y)(p11.2q11.2)	
	TAF1 Mutation	chrX:70621541 T>C, p.I1337T	
Gestation	Complications	placenta deterioration	
	Term (Weeks)	40	37
Birth	C-section	+	+
	Weight (Centile)	2.21 kg	1.76 kg
	Length (Centile)	NK	NK
	Head Circumference (Centile)	NK	NK
	Apgar Scores	NK	NK
Perinatal Course	Complications	NK	neonatal jaundice, poor feeding
Growth	Prenatal Onset Growth Retardation	+	
	Postnatal Growth Retardation	+	
Neurobehavioral/Development	Gross Motor Delay	+	
	Verbal Expression Delay	+	
	Oral Motor Dysphagia	+	
Craniofacial	Prominent Periorbital Ridges	+	
	Downslanted Palpebral Fissures	+	
	Deep-set Eyes	+	
	Prominent Forehead	+	
	Sagging Cheeks	+	
	Long Philtrum	+	
	Prominent Low-set Ears	+	
	Thickened Helices	+	
	Long Face	+	
	High Arched Palate	+	
	Thin Upper Lip	+	
	Pointed Chin	+	
	Broad Upturned Nose	+	
	Hypertelorism	+	
Skin	Aplasia Cutis Congenita	+ (on the scalp)	-
	Sacral Dimple	+	
	Hirsutism	-	
	Frequent Dermatitis & Eczema	-	+
	Dysplastic Toenails	+	
Ear, Nose, Mouth, and Throat (ENMT)	hearing Impairments	+	
	Chromic Otitis Media with Effusion	+	
Eyes	Strabismus (Exotropia)	+	
	Blocked Tear Ducts	+	
	Oculomotor Dysfuntion	+	
Gastrointestinal	Constipation	-	+
	Gastroesophageal Reflux	-	+
Neurological	Microcephaly	+	
	Ventriculomegaly	+	
	Low Cerebral White Matter Volume	+	
	Seizures	-	
	Hypotonia	+	
	Deficient Septum Pellucidum	+	
	Hypoplasia of the Corpus Callosum	+	
	Gait Abnormalities	+	
	Balance Problem	-	
	Diplegia	+	-
	Ataxia	+	
	Sleep-wake Dysregulation	-	+
Musculoskeletal	Osteopenia	+	NK
	Unusual Gluteal Crease with Sacral Caudal Remnant	+	
	Hyperextensible Joints	+ (metacarpophalangeal joints & wrists)	+ (fingers & wrists)
	Spasticity	+ (lower extremity)	
	Kyphosis	+	+ (thoracic)
	Scoliosis	-	+
	Short Neck	+	
Respitatory		-	+ (H/O asthma)
Cardiovascular	Structural Defects at Birth	-	- (H/O murmur without cardiovascular abnormalities)
Genital		-	
Hematologic/Lymphatic/Immunologic		-	
Psychiatric	Autistic Behaviors	+	
	Attention Deficit Hyperactivity Disorder	+	
	Anxiety	+	
	Intellectual Disability	+	
Other	Age of Death (Years)	N/A	
	Other Features	-	

"+": feature present; "-": feature absent; N/A: not applicable; NK: not known.

Table 1. This table illustrates known clinical features across the affected individuals, as well as other noted clinical characteristics on these individuals.

Table 2. A table of prioritized genetic variations in RykDax Syndrome

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 2. 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.

frontal circumference (OFC) was 51 cm (4.5th percentile); while his elder brother (aged 1111/12 years) had a height of 136.8 cm (5% tile), a weight of 26.3 kg (0% tile, BMI 14.1 kg/m2), and his OFC was 49.5 cm (0.2th percentile) at the time.

Brain MRIs of the two brothers 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 [40]. 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, 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. Findings associated with septooptic dysplasia included underdeveloped pituitary glands for age, deficiency of the anterior falx with mild hemispheric interdigitiation, 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 (see Clinical Methods section) 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.

B. Whole genome sequencing and bioinformatics analysis

B.1. Whole genome sequencing

With our effort to move beyond exome sequencing and into whole genome analyses, over the past few years we have developed and published comprehensive whole genome analysis pipelines, including for finding insertions and deletions (INDELs) [7, 41–46]. We and others have shown that the two dominant WGS platforms (Illumina and Complete Genomics) are complementary, as both miss variants [42, 47], and so both platforms were used in this study. Initial sequencing efforts focused on whole genome sequencing using the Complete Genomics (CG) sequencing and analysis pipeline v2.0 for the mother, father and two affected boys. Additional whole genome sequencing was performed subsequent to this initial effort, using the Illumina HiSeq 2000 sequencing platform on the mother, father, two affected boys and six other immediate family members. Raw sequencing data stemming from Illumina sequencing was processed by a variety of analysis pipelines and subsequently pooled with variants detected by the CG sequencing and analysis pipeline (see Whole Genome Sequencing and Analysis Methods section). Downstream and functional annotation tools were used to evaluate variants detected by the various methods. 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 Supplementary Table 1 for more details about the sequencing data.

B.2. Bioinformatics analyses and variant calling

The mean number of variants per individual that were detected using the Illumina sequencing data across all of the detection methods was 3,583,905.1 (SD=192,317.5) SNPs, 650,708.2 (SD=

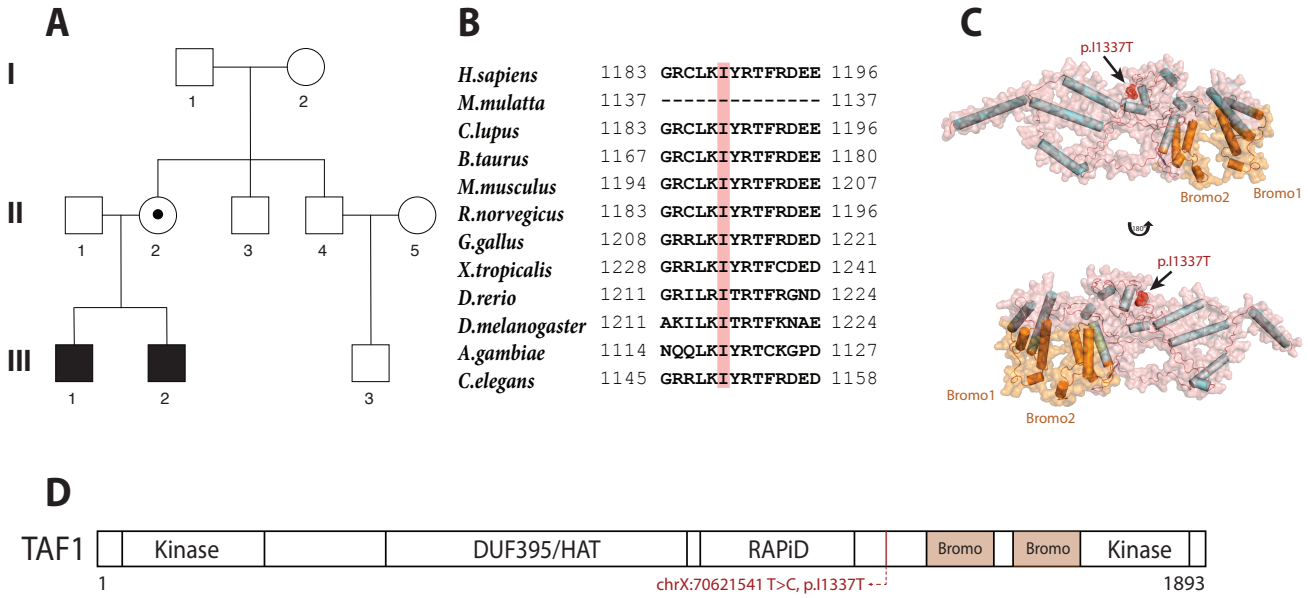

Figure 2. (**A**) Pedigree drawings of the Utah family, (**B**) A protein sequence alignment of TAF1 between *H.sapiens, M.mulatta, C.lupus, B.taurus, M.musculus, R.norvegicus, G.gallus, X.tropicalis, D.rerio, D.melanogaster, A.gambiae, and C.elegans,* listed from top to bottom. This alignment was generated using the MUSCLE [48] software in the HomoloGene [http://www.ncbi.nlm.nih.gov/ homologene] website. The TAF1 I1337T location is highlighted in red, which shows a high degree of protein sequence conservation, (**C**) computational modeling of the TAF1 protein structure from residues 1080 to 1579 using I-TASSER and (**D**) known *TAF1* domains, with the *TAF1* variant indicated.

84,125.7) INDELs (6,310.2 mean INDELs from Scalpel with a SD of 74.3, as it only detects signals in exon regions), 1,338,503 (SD=7,622.2) variants observed in STR regions, and 327.4 CNVs with a SD of 8.2 (and a mean of 49.1 and a SD of 15.8 for CNVs from PennCNV, which detects signals over a smaller search space than ERDS). For the CG sequence data, the mean number variants per individual detected are 3,457,584 (SD=51,665.8), 565,691.5 (SD=16,247.7), and 175.3 (SD=12.8) for SNPs INDELs and CNVs respectively. 14 unique INDELS and SNVs were discovered using two different prioritization schemes, with only a single coding SNV being reliably identified by both schemes. 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 Supplementary Table 2 for the list of CNVs).

B.3. Variant prioritization

Using the two variant prioritization schemes described in the Materials and Methods section, we discovered a set of putative variants from among the three disease models tested here. We found 7 potentially important variants in non-coding regions of the genome and 7 in coding regions across the two prioritization schemes (**Table 2**). These variants fall within coding and non-coding regions of a total of 8 known genes, *TAF1*, *FAM47B*, *SLC28A1*, *FRG2B*, *DMRTB1*, *ZNF423*, *SNAPC5* and *KCHN1*. We found a number of CNVs that were not known to be associated with any deleterious phenotype that we could find (Supplementary Table 2). As stated above, only one variant was shared between the two different schemes employed here, namely a non-synonymous change in *TAF1* that resulted in an isoleucine (hydrophobic) to threonine (polar) change on the 1337th residue. The protein change occurs within a linker region before two bromodomains and after a co-factor (TAF7) interacting domain. This linker region is highly conserved in multicellular eukaryotes (**Figure 2B**), but is not present in S. cerevisiae TAF1. Therefore, this linker region is not located in a recently reported crystal structure for the S. cerevisiae TAF1-TAF7 complex [49].

B.4. Variant scores

The *TAF1* variant found in this family was ranked highest among the variants being tested with VAAST using an X-linked model (with a p-value of 0.00184 and a rank of 14.59) and it is ranked highly in terms of its potential functional significance. This variant is ranked by CADD as being within the top 1% most deleterious variants in the human genome, it is scored by PolyPhen-2 [50] as being "Probably Damaging" with a score of 0.996, by SIFT [51] as "Damaging" with a score of 0.003, and also by PROVEAN [52] as "Deleterious" with a score of -3.51. This variant in *TAF1* is novel, as it is not found in public databases (i.e., dbSNP 137, 1000 Genomes phase 1 data, NHLBI 6500 exomes, or ExAC version 0.2).

B.5. Protein modeling

To investigate how the *TAF1* variant may influence protein structure and packaging, we built a structure model for the region of residues 1080 to 1569 using I-TASSER (**Figure 2C**). Residues spanning 1120 to 1270 contain the RAP74 interacting do-main (RAPiD), which has been shown to be important for the interactions between TAF1 and RAP74 (GTF2F1) and TAF1-TAF7 [53]. Residues spanning 1373 to 1590 contain two Bromo domains (Bromo1: 1397-1467 and Bromo2: 1520-1590, **Figure 2C** and **2D**). These Bromo do-mains consist of a bundle of four alpha helices that form a hydrophobic pocket to recognize an acetyl lysine, such as those on the N-terminal tails of histones. The variant found in this family, I337T, is located between the RAPiD and Bromo domains (shown as red spheres). Although this variant is not within any known protein domain, we speculate that it may affect domain packing of TAF1, which may interfere with the TAF1-TAF7 interacting surface [49, 53] or mark the protein for proteolytic degradation.

4. DISCUSSION

In general, we found benefits in using multiple informatics pipelines with WGS across a multi-generational pedigree for the identification and prioritization of human genetic variation

potentially important in the disease phenotype discussed. We highlight here a variant found on the X-chromosome of the two affected male boys from Utah, USA, which was transmitted to them from their mother who acquired this variant spontaneously (de novo) and who is herself affected by extreme X-chromosome skewing. We were able to prioritize variants that were not observed in other members of the sequenced family, including not being present in an unaffected male cousin. The only variant found in both of our two variant prioritization schemes is in a highly conserved region of TAF1, which is the largest subunit of the TFIID multi-protein complex involved in transcription initiation [49, 54–57].

A. *TAF1* variant implicated in disease phenotype

TAF1 (TATA-box-binding protein associated factor 1) is part of the TFIID multi-protein complex, which consists of the TATA binding protein (TBP) and 12 additional TBP associated factors (TAFs). TFIID has been implicated in promoting transcription initiation by recognizing promoter DNA and facilitating the nucleation of other factors to aid in the assembly of the pre-initiation complex [55]. TFIID also interacts with transcriptional activators as a co-activator [55]. TAF1 is the largest known TFIID associated TAF, and it binds directly to the TBP via a conserved N-terminal domain. Through its binding to TBP, TAF1 is thought to influence some control over the activation of genes promoted by TATA or other DNA motifs by inhibiting the TBP subunit from binding to these regions [54]. More recent work has reported aberrant expression affecting hundreds of *D. melanogaster* genes as a result of *TAF1* transcript depletion, with many more of these genes being expressed more vigorously while a smaller number are expressed less vigorously than under wild-type conditions [58]. In addition and consistent with the notion that TAF1 is important in controlling the binding patterns of TFIID to specific promoter regions, this study showed that the set of genes conferring increased expression were enriched with genes containing TATA-motif promoters, suggesting an association between the depletion of TAF1 and increased expression of genes with TATA-based promoters. Recent structural work in yeast points to an epigenetic role of the TAF1-TAF7 complex in general TFIID function and/or pre-initiation complex (PIC) assembly, and that TAF1 is likely unstable in the absence of a binding partner, such as TAF7 [49].

There is evidence relating various TAFs and the TBP to important functional roles in human neuronal tissue. Indeed, mutations observed in *TAF1-2* and the TBP have been implicated in playing an important role in human neurodegenerative disorders such as X-linked dystonia-parkinsonism (XDP) [59, 60] as well as in intellectual disability and developmental delay [61–63], respectively. Both XDP studies demonstrated aberrant neuron-specific *TAF1* isoform expression levels in neuronal tissue containing *TAF1* mutations. Herzfeld et. al (2013) corroborated previous reports which suggested that a reduction in *TAF1* expression is associated with large-scale expression differences across hundreds of genes. Studies in rat and mice brain also corroborate the importance and relevance of *TAF1* expression patterns specific to neuronal tissues [64, 65]. In corroboration with biochemical and functional studies, a recent population-scale study reported *TAF1* as being ranked 53rd among the top 1,003 constrained human genes [66]. Using allele frequencies reported by the Blood Institute (NHLBI) Exome Sequencing Project (ESP), the authors of this study reasoned that the number of observed missense variants is lower than one would expect by chance (a signed Z score of 5.1779 ranks *TAF1* the 53rd most constrained among the 1,003 most constrained human genes) when compared to the expected probability of missense variants occurring in this gene (5.61E-05).

FRG2B and *FAM47B* are not known to be involved in the pathogenesis of human disease, although the detailed molecular function of these genes has been largely unexplored. *FRG2B* is homologous to *FRG2* located on chromosome 4, which 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 [67–69]. However, reductions in the homologous 10qt26 microsatellite repeat, proximal to *FRG2B*, have *not* been associated with FSHD.

ZNF423 acts as a transcriptional regulator and mutations in *ZNF423* coding regions have been implicated in the pathogenesis of Joubert syndrome [70, 71]. The mutation 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 [72] whereas SNAPC5, also known as SNAP19, plays a scaffolding role in the forming the complete snRNA-activating protein (SNAP) complex, which is required for the transcription of snRNA genes [73]. The molecular functions of KCHN1 and DMRTB1 are not well understood or studied.

The *TAF1* variant found in this Utah family was the only variant identified as important by the two prioritization schemes that we used. The variant arose in this family as a de novo variant on the X-chromosome of the mother (**II-2**) of the two affected children (as it is not found in any of the other members of the family) and was then transmitted to both of them. The mother also exhibits extreme X-chromosome skewing, whereas her mother and her sister do not. The protein change occurs within a linker region N-terminal to two bromo domains and C-terminal of a co-factor (TAF7) interacting domain. This linker region is highly conserved in multicellular eukaryotes, but is not present in *S. cerevisiae TAF1*. This linker region is not located in a recently reported crystal structure for the *S. cerevisiae* TAF1-TAF7 complex [49]. The crystal structure of a human TAF1-TAF7 complex was also reported [53], although the characterized region, again, does not include the linker region where our variant lies. This variant represented the only variation that we were able to identify with comprehensive WGS that has a clear molecular function in neuronal tissue and functions as part of TFIID, a larger multi-protein complex involved in general transcription regulation that has been suggested to play a possible role in neurodegenerative diseases and developmental delay when its constituents are disrupted [13, 59–62, 64, 65].

Taken together, this evidence suggests that the variant in *TAF1* may be playing an important role in this newly identified syndrome, whereas other variants found in this family (**Table 2**) are not as strongly implicated by previous work exploring their putative function(s).

5. CONCLUSIONS

Our work demonstrates the value of performing more comprehensive genomic analyses when confronted with an undescribed and undiagnosed syndrome or disease affliction, particularly with only one or two affected probands in one family. WGS led to the identification of genomic aberrations that were not tested for by more traditional, clinical assays. Among the multiple rare and potentially disease-contributory variants discovered this this family, the variant in *TAF1* is likely contributing to the phenotype, in concert with the environment and other genetic aberrations to contribute to the sum total of this disease phenotype in the two brothers in Utah. Of course, the differences in genetic background and the environment can certainly account for the phenotypic differences between the two brothers. This phenomenon has been well known for many years in genetics [74, 75], but seems to be more recently appreciated and has become a current active research topic [76–80]. Other work

also suggests a physiological link between developmental delay and the TFIID multi-protein complex [13, 61, 62], although the phenotypic variability and expression of other variants in *TAF1* remains to be determined.

ACKNOWLEDGMENTS

G.J.L. is supported by funds from the Stanley Institute for Cognitive Genomics at Cold Spring Harbor Laboratory. K.W. is supported by NIH grant number HG006465. The sequencing by Complete Genomics was provided by a CG data analysis grant to K.W. We would like to thank Max Doerfel for helpful discussions. The authors would like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found at http://exac.broadinstitute.org/about.

REFERENCES

1. G. F. Schneider and C. Dekker, "Dna sequencing with nanopores," Nature biotechnology **30**, 326–328 (2012).
2. J. Shendure and E. Lieberman Aiden, "The expanding scope of dna sequencing," Nature biotechnology **30**, 1084–1094 (2012).
3. D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, E. C. M. Chiara, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith, "Accurate whole human genome sequencing using reversible terminator chemistry," Nature **456**, 53–59 (2008).
4. G. J. Lyon and K. Wang, "Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress," Genome Med **4**, 58 (2012).
5. S. H. Katsanis and N. Katsanis, "Molecular genetic testing and the future of clinical genomics," Nature reviews. Genetics **14**, 415–426 (2013).
6. G. J. Lyon, "Personal account of the discovery of a new disease using next-generation sequencing. interview by natalie harrison," Pharmacogenomics **12**, 1519–1523 (2011).
7. C. A. Brownstein, A. H. Beggs, N. Homer, B. Merriman, T. W. Yu, K. C. Flannery, E. T. DeChene, M. C. Towne, S. K. Savage, E. N. Price, I. A. Holm, L. J. Luquette, E. Lyon, J. Majzoub, P. Neupert, D. McCallie Jr., P. Szolovits, H. F. Willard, N. J. Mendelsohn, R. Temme, R. S. Finkel, S. W. Yum, L. Medne, S. R. Sunyaev, I. Adzhubey, C. A. Cassa, P. I. de Bakker, H. Duzkale, P. Dworzynski, W. Fairbrother, L. Francioli, B. H. Funke, M. A. Giovanni, R. E. Handsaker, K. Lage, M. S. Lebo, M. Lek, I. Leshchiner, D. G. MacArthur, H. M. McLaughlin, M. F. Murray, T. H. Pers, P. P. Polak, S. Raychaudhuri, H. L. Rehm, R. Soemedi, N. O. Stitziel, S. Vestecka, J. Supper, C. Gugenmus, B. Klocke, A. Hahn, M. Schubach, M. Menzel, S. Biskup, P. Freisinger, M. Deng, M. Braun, S. Perner, R. J. Smith, J. L. Andorf, J. Huang, K. Ryckman, V. C. Sheffield, E. M. Stone, T. Bair, E. A. Black-Ziegelbein, T. A. Braun, B. Darbro, A. P. DeLuca, D. L. Kolbe, T. E. Scheetz, A. E. Shearer, R. Sompallae, K. Wang, A. G. Bassuk, E. Edens, K. Mathews, S. A. Moore, O. A. Shchelochkov, P. Trapane, A. Bossler, C. A. Campbell, J. W. Heusel, A. Kwitek, T. Maga, K. Panzer, T. Wassink, D. Van Daele, H. Azaiez, K. Booth, N. Meyer, M. M. Segal, M. S. Williams, G. Tromp, P. White, D. Corsmeier, S. Fitzgerald-Butt, G. Herman, D. Lamb-Thrush, K. L. McBride, D. Newsom, C. R. Pierson, A. T. Rakowsky, A. Maver, L. Lovrecic, A. Palandacic, B. Peterlin, A. Torkamani, A. Wedell, M. Huss, A. Alexeyenko, J. M. Lindvall, M. Magnusson, D. Nilsson, H. Stranneheim, F. Taylan, C. Gilissen, A. Hoischen, B. van Bon, H. Yntema, M. Nelen, W. Zhang, J. Sager, L. Zhang, K. Blair, D. Kural, M. Cariaso, G. G. Lennon, A. Javed, S. Agrawal, P. C. Ng, K. S. Sandhu, S. Krishna, V. Veeramachaneni, O. Isakov, E. Halperin, E. Friedman, N. Shomron, G. Glusman, J. C. Roach, J. Caballero, H. C. Cox, D. Mauldin, S. A. Ament, L. Rowen, D. R. Richards, F. A. San Lucas, M. L. Gonzalez-Garay, C. T. Caskey, Y. Bai, Y. Huang, F. Fang, Y. Zhang, Z. Wang, J. Barrera, J. M. Garcia-Lobo, D. Gonzalez-Lamuno, J. Llorca, M. C. Rodriguez, I. Varela, M. G. Reese, F. M. De La Vega, E. Kiruluta, M. Cargill, R. K. Hart, J. M. Sorenson, G. J. Lyon, D. A. Stevenson, B. E. Bray, B. M. Moore, K. Eilbeck, M. Yandell, H. Zhao, L. Hou, X. Chen, X. Yan, M. Chen, C. Li, C. Yang, M. Gunel, P. Li, Y. Kong, A. C. Alexander, Z. I. Albertyn, K. M. Boycott, D. E. Bulman, P. M. Gordon, A. M. Innes, B. M. Knoppers, J. Majewski, C. R. Marshall, J. S. Parboosingh, S. L. Sawyer, M. E. Samuels, J. Schwartzentruber, I. S. Kohane, and D. M. Margulies, "An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the clarity challenge," Genome Biol **15**, R53 (2014).
8. J. C. Roach, G. Glusman, A. F. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas, "Analysis of genetic inheritance in a family

quartet by whole-genome sequencing," Science **328**, 636–639 (2010).
9. M. J. Bamshad, J. A. Shendure, D. Valle, A. Hamosh, J. R. Lupski, R. A. Gibbs, E. Boerwinkle, R. P. Lifton, M. Gerstein, M. Gunel, S. Mane, D. A. Nickerson, and G. Centers for Mendelian, "The centers for mendelian genomics: a new large-scale initiative to identify the genes underlying rare mendelian conditions," Am J Med Genet A **158A**, 1523–1525 (2012).
10. A. F. Rope, K. Wang, R. Evjenth, J. Xing, J. J. Johnston, J. J. Swensen, W. E. Johnson, B. Moore, C. D. Huff, L. M. Bird, J. C. Carey, J. M. Opitz, C. A. Stevens, T. Jiang, C. Schank, H. D. Fain, R. Robison, B. Dalley, S. Chin, S. T. South, T. J. Pysher, L. B. Jorde, H. Hakonarson, J. R. Lillehaug, L. G. Biesecker, M. Yandell, T. Arnesen, and G. J. Lyon, "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 (2011).
11. H. Hirata, I. Nanda, A. van Riesen, G. McMichael, H. Hu, M. Hambrock, M. A. Papon, U. Fischer, S. Marouillat, C. Ding, S. Alirol, M. Bienek, S. Preisler-Adams, A. Grimme, D. Seelow, R. Webster, E. Haan, A. MacLennan, W. Stenzel, T. Y. Yap, A. Gardner, L. S. Nguyen, M. Shaw, N. Lebrun, S. A. Haas, W. Kress, T. Haaf, E. Schellenberger, J. Chelly, G. Viot, L. G. Shaffer, J. A. Rosenfeld, N. Kramer, R. Falk, D. El-Khechen, L. F. Escobar, R. Hennekam, P. Wieacker, C. Hubner, H. H. Ropers, J. Gecz, M. Schuelke, F. Laumonnier, and V. M. Kalscheuer, "Zc4h2 mutations are associated with arthrogryposis multiplex congenita and intellectual disability through impairment of central and peripheral synaptic plasticity," American journal of human genetics **92**, 681–695 (2013).
12. A. B. Alsalem, A. S. Halees, S. Anazi, S. Alshamekh, and F. S. Alkuraya, "Autozygome sequencing expands the horizon of human knockout research and provides novel insights into human phenotypic variation," PLoS Genet **9**, e1004030 (2013).
13. H. Najmabadi, H. Hu, M. Garshasbi, T. Zemojtel, S. S. Abedini, W. Chen, M. Hosseini, F. Behjati, S. Haas, P. Jamali, A. Zecha, M. Mohseni, L. Puttmann, L. N. Vahid, C. Jensen, L. A. Moheb, M. Bienek, F. Larti, I. Mueller, R. Weissmann, H. Darvish, K. Wrogemann, V. Hadavi, B. Lipkowitz, S. Esmaeeli-Nieh, D. Wieczorek, R. Kariminejad, S. G. Firouzabadi, M. Cohen, Z. Fattahi, I. Rost, F. Mojahedi, C. Hertzberg, A. Dehghan, A. Rajab, M. J. Banavandi, J. Hoffer, M. Falah, L. Musante, V. Kalscheuer, R. Ullmann, A. W. Kuss, A. Tzschach, K. Kahrizi, and H. H. Ropers, "Deep sequencing reveals 50 novel genes for recessive cognitive disorders," Nature **478**, 57–63 (2011).
14. G. Romeo and A. H. Bittles, "Consanguinity in the contemporary world," Hum Hered **77**, 6–9 (2014).
15. E. F. Keller, *The mirage of a space between nature and nurture* (2010).
16. L. A. Pennacchio, W. Bickmore, A. Dean, M. A. Nobrega, and G. Bejerano, "Enhancers: five essential questions," Nature reviews. Genetics **14**, 288–295 (2013).
17. D. E. Dickel, A. Visel, and L. A. Pennacchio, "Functional anatomy of distant-acting mammalian enhancers," Philosophical transactions of the Royal Society of London. Series B, Biological sciences **368**, 20120359 (2013).
18. F. Cartault, P. Munier, E. Benko, I. Desguerre, S. Hanein, N. Boddaert, S. Bandiera, J. Vellayoudom, P. Krejbich-Trotot, M. Bintner, J. J. Hoarau, M. Girard, E. Genin, P. de Lonlay, A. Fourmaintraux, M. Naville, D. Rodriguez, J. Feingold, M. Renouil, A. Munnich, E. Westhof, M. Fahling, S. Lyonnet, and A. Henrion-Caude, "Mutation in a primate-conserved retrotransposon reveals a noncoding rna as a mediator of infantile encephalopathy," Proceedings of the National Academy of Sciences of the United States of America **109**, 4980–4985 (2012).
19. J. Salzman, C. Gawad, P. L. Wang, N. Lacayo, and P. O. Brown, "Circular rnas are the predominant transcript isoform from hundreds of human genes in diverse cell types," PloS one **7**, e30733 (2012).
20. P. J. Batista and H. Y. Chang, "Long noncoding rnas: cellular address codes in development and disease," Cell **152**, 1298–1307 (2013).
21. T. B. Hansen, T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard, and J. Kjems, "Natural rna circles function as efficient microrna sponges," Nature **495**, 384–388 (2013).
22. A. Kapusta, Z. Kronenberg, V. J. Lynch, X. Zhuo, L. Ramsay, G. Bourque, M. Yandell, and C. Feschotte, "Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding rnas," PLoS Genet **9**, e1003470 (2013).
23. V. Khoddami and B. R. Cairns, "Identification of direct targets and modified bases of rna cytosine methyltransferases," Nature biotechnology (2013).
24. H. Ledford, "Circular rnas throw genetics for a loop," Nature **494**, 415 (2013).
25. A. Maxmen, "Rna: The genome's rising stars," Nature **496**, 127–129 (2013).
26. S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. le Noble, and N. Rajewsky, "Circular rnas are a large class of animal rnas with regulatory potency," Nature **495**, 333–338 (2013).
27. T. R. Mercer and J. S. Mattick, "Structure and function of long noncoding rnas in epigenetic regulation," Nature structural & molecular biology **20**, 300–307 (2013).
28. P. Miura, S. Shenker, C. Andreu-Agullo, J. O. Westholm, and E. C. Lai, "Widespread and extensive lengthening of 3′ utrs in the mammalian brain," Genome Res **23**, 812–825 (2013).
29. M. P. Moreau, S. E. Bruse, R. Jornsten, Y. Liu, and L. M. Brzustowicz, "Chronological changes in microrna expression in the developing human brain," PloS one **8**, e60480 (2013).
30. S. Ning, P. Wang, J. Ye, X. Li, R. Li, Z. Zhao, X. Huo, L. Wang, and F. Li, "A global map for dissecting phenotypic variants in human lincrnas," European journal of human genetics : EJHG (2013).
31. P. N. Perrat, S. DasGupta, J. Wang, W. Theurkauf, Z. Weng, M. Rosbash, and S. Waddell, "Transposition-driven genomic heterogeneity in the drosophila brain," Science **340**, 91–95 (2013).
32. L. R. Sabin, M. J. Delas, and G. J. Hannon, "Dogma derailed: the many influences of rna on the genome," Molecular cell **49**, 783–794 (2013).
33. J. E. Wilusz and P. A. Sharp, "Molecular biology. a circuitous route to noncoding rna," Science **340**, 440–441 (2013).
34. R. C. Allen, H. Y. Zoghbi, A. B. Moseley, H. M. Rosenblatt, and J. W. Belmont, "Methylation of hpaii and hhai sites near the polymorphic cag repeat in the human androgen-receptor gene correlates with x chromosome inactivation," American journal of human genetics **51**, 1229 (1992).
35. M. Yandell, C. Huff, H. Hu, M. Singleton, B. Moore, J. Xing, L. B. Jorde, and M. G. Reese, "A probabilistic disease-gene finder for personal genomes," Genome research **21**, 1529–1542 (2011).
36. G. B. Christensen and C. G. Lambert, "Search for compound

heterozygous effects in exome sequence of unrelated subjects," BMC proceedings **5 Suppl 9**, S95 (2011).
37. K. Wang, M. Li, and H. Hakonarson, "Annovar: functional annotation of genetic variants from high-throughput sequencing data," Nucleic Acids Res **38**, e164 (2010).
38. U. Paila, B. A. Chapman, R. Kirchner, and A. R. Quinlan, "Gemini: Integrative exploration of genetic variation and genome annotations," PLoS Comput Biol **9**, e1003153 (2013).
39. J. M. Amos-Landgraf, A. Cottle, R. M. Plenge, M. Friez, C. E. Schwartz, J. Longshore, and H. F. Willard, "X chromosome-inactivation patterns of 1,005 phenotypically unaffected females," American journal of human genetics **79**, 493–499 (2006).
40. C. Garel, I. Cont, C. Alberti, E. Josserand, M. L. Moutard, and H. Ducou le Pointe, "Biometry of the corpus callosum in children: Mr imaging reference data," AJNR Am J Neuroradiol **32**, 1436–1443 (2011).
41. G. Narzisi, J. A. O. Rawe, I. Iossifov, Y.-h. Lee, Z. Wang, Y. Wu, G. J. Lyon, M. Wigler, and M. C. Schatz, "Accurate detection of de novo and transmitted indels within exome-capture data using micro-assembly," bioRxiv (2013).
42. J. O'Rawe, T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang, and G. J. Lyon, "Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing," Genome Med **5**, 28 (2013).
43. J. A. O'Rawe, H. Fang, S. Rynearson, R. Robison, E. S. Kiruluta, G. Higgins, K. Eilbeck, M. G. Reese, and G. J. Lyon, "Integrating precision medicine in the study and clinical treatment of a severely mentally ill person," PeerJ **1**, e177 (2013).
44. L. Shi, X. Zhang, R. Golhar, F. G. Otieno, M. He, C. Hou, C. Kim, B. Keating, G. J. Lyon, K. Wang, and H. Hakonarson, "Whole-genome sequencing in an autism multiplex family," Mol Autism **4**, 8 (2013).
45. K. Wang, C. Kim, J. Bradfield, Y. Guo, E. Toskala, F. G. Otieno, C. Hou, K. Thomas, C. Cardinale, G. J. Lyon, R. Golhar, and H. Hakonarson, "Whole-genome dna/rna sequencing identifies truncating mutations in rbck1 in a novel mendelian disease with neuromuscular and cardiac involvement," Genome medicine **5**, 67 (2013).
46. H. Fang, Y. Wu, G. Narzisi, J. A. O'Rawe, L. T. Jimenez Barrón, J. Rosenbaum, M. Ronemus, I. Iossifov, M. C. Schatz, and G. J. Lyon, *Reducing INDEL calling errors in whole-genome and exome sequencing data* (2014).
47. H. Y. Lam, M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. E. Dewey, L. Habegger, E. A. Ashley, M. B. Gerstein, A. J. Butte, H. P. Ji, and M. Snyder, "Performance comparison of whole-genome sequencing platforms," Nature biotechnology **30**, 78–82 (2012).
48. R. C. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," Nucleic Acids Research **32**, 1792–1797 (2004).
49. S. Bhattacharya, X. Lou, P. Hwang, K. R. Rajashankar, X. Wang, J. A. Gustafsson, R. J. Fletterick, R. H. Jacobson, and P. Webb, "Structural and functional insight into taf1-taf7, a subcomplex of transcription factor ii d," Proc Natl Acad Sci U S A **111**, 9103–9108 (2014).
50. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," Nat Methods **7**, 248–249 (2010).
51. P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm," Nat Protoc **4**, 1073–1081 (2009).
52. Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, "Predicting the functional effect of amino acid substitutions and indels," PLoS ONE **7**, e46688 (2012).
53. H. Wang, E. C. Curran, T. R. Hinds, E. H. Wang, and N. Zheng, "Crystal structure of a taf1-taf7 complex in human transcription factor iid reveals a promoter binding module," Cell Res **24**, 1433–1444 (2014).
54. T. Kotani, T. Miyake, Y. Tsukihashi, A. G. Hinnebusch, Y. Nakatani, M. Kawaichi, and T. Kokubo, "Identification of highly conserved amino-terminal segments of dtafii230 and ytafii145 that are functionally interchangeable for inhibiting tbp-dna interactions in vitro and in promoting yeast cell growth in vivo," Journal of Biological Chemistry **273**, 32254–32264 (1998).
55. G. Papai, P. A. Weil, and P. Schultz, "New insights into the function of transcription factor tfiid from recent structural studies," Curr Opin Genet Dev **21**, 219–224 (2011).
56. M. Cianfrocco, G. Kassavetis, P. Grob, J. Fang, T. Juven-Gershon, J. Kadonaga, and E. Nogales, "Human tfiid binds to core promoter dna in a reorganized structural state," Cell **152**, 120–131.
57. C. Bieniossek, G. Papai, C. Schaffitzel, F. Garzoni, M. Chaillet, E. Scheer, P. Papadopoulos, L. Tora, P. Schultz, and I. Berger, "The architecture of human general transcription factor tfiid core complex," Nature **493**, 699–702 (2013).
58. K. L. Pennington, S. K. Marr, G. W. Chirn, and M. T. Marr 2nd, "Holo-tfiid controls the magnitude of a transcription burst and fine-tuning of transcription," Proc Natl Acad Sci U S A **110**, 7678–7683 (2013).
59. T. Herzfeld, D. Nolte, M. Grznarova, A. Hofmann, J. L. Schultze, and U. Muller, "X-linked dystonia parkinsonism syndrome (xdp, lubag): disease-specific sequence change dsc3 in taf1/dyt3 affects genes in vesicular transport and dopamine metabolism," Hum Mol Genet **22**, 941–951 (2013).
60. S. Makino, R. Kaji, S. Ando, M. Tomizawa, K. Yasuno, S. Goto, S. Matsumoto, M. D. Tabuena, E. Maranon, M. Dantes, L. V. Lee, K. Ogasawara, I. Tooyama, H. Akatsu, M. Nishimura, and G. Tamiya, "Reduced neuron-specific expression of the taf1 gene is associated with x-linked dystonia-parkinsonism," American journal of human genetics **80**, 393–406 (2007).
61. L. Rooms, E. Reyniers, S. Scheers, R. van Luijk, J. Wauters, L. Van Aerschot, Z. Callaerts-Vegh, R. D'Hooge, G. Mengus, I. Davidson, W. Courtens, and R. F. Kooy, "Tbp as a candidate gene for mental retardation in patients with subtelomeric 6q deletions," Eur J Hum Genet **14**, 1090–1096 (2006).
62. S. Hellman-Aharony, P. Smirin-Yosef, A. Halevy, M. Pasmanik-Chor, A. Yeheskel, A. Har-Zahav, I. Maya, R. Straussberg, D. Dahary, A. Haviv, M. Shohat, and L. Basel-Vanagaite, "Microcephaly thin corpus callosum intellectual disability syndrome caused by mutated taf2," Pediatr Neurol **49**, 411–416 e1 (2013).
63. K. K. Abu-Amero, A. Hellani, M. A. Salih, A. Al Hussain, M. al Obailan, G. Zidan, I. A. Alorainy, and T. M. Bosley, "Ophthalmologic abnormalities in a de novo terminal 6q deletion," Ophthalmic Genet **31**, 1–11 (2010).
64. J. Jambaldorj, S. Makino, B. Munkhbat, and G. Tamiya, "Sustained expression of a neuron-specific isoform of the taf1 gene in development stages and aging in mice," Biochem Biophys Res Commun **425**, 273–277 (2012).
65. W. Sako, R. Morigaki, R. Kaji, I. Tooyama, S. Okita, K. Kitazato, S. Nagahiro, A. M. Graybiel, and S. Goto, "Identification and localization of a neuron-specific isoform of taf1 in rat brain: implications for neuropathology of dyt3 dystonia," Neuroscience **189**, 100–107 (2011).
66. K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens,

A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnstrom, S. Mallick, A. Kirby, D. P. Wall, D. G. MacArthur, S. B. Gabriel, M. DePristo, S. M. Purcell, A. Palotie, E. Boerwinkle, J. D. Buxbaum, E. H. Cook Jr, R. A. Gibbs, G. D. Schellenberg, J. S. Sutcliffe, B. Devlin, K. Roeder, B. M. Neale, and M. J. Daly, "A framework for the interpretation of de novo mutation in human disease," Nat Genet **46**, 944–950 (2014).

67. M. Richards, F. Coppee, N. Thomas, A. Belayew, and M. Upadhyaya, "Facioscapulohumeral muscular dystrophy (fshd): an enigma unravelled?" Hum Genet **131**, 325–340 (2012).
68. T. Rijkers, G. Deidda, S. van Koningsbruggen, M. van Geel, R. J. Lemmers, J. C. van Deutekom, D. Figlewicz, J. E. Hewitt, G. W. Padberg, R. R. Frants, and S. M. van der Maarel, "Frg2, an fshd candidate gene, is transcriptionally upregulated in differentiating primary myoblast cultures of fshd patients," J Med Genet **41**, 826–836 (2004).
69. D. Gabellini, M. R. Green, and R. Tupler, "Inappropriate gene activation in fshd: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle," Cell **110**, 339–348 (2002).
70. M. Chaki, R. Airik, A. K. Ghosh, R. H. Giles, R. Chen, G. G. Slaats, H. Wang, T. W. Hurd, W. Zhou, A. Cluckey, H. Y. Gee, G. Ramaswami, C. J. Hong, B. A. Hamilton, I. Cervenka, R. S. Ganji, V. Bryja, H. H. Arts, J. van Reeuwijk, M. M. Oud, S. J. Letteboer, R. Roepman, H. Husson, O. Ibraghimov-Beskrovnaya, T. Yasunaga, G. Walz, L. Eley, J. A. Sayer, B. Schermer, M. C. Liebau, T. Benzing, S. Le Corre, I. Drummond, S. Janssen, S. J. Allen, S. Natarajan, J. F. O'Toole, M. Attanasio, S. Saunier, C. Antignac, R. K. Koenekoop, H. Ren, I. Lopez, A. Nayir, C. Stoetzel, H. Dollfus, R. Massoudi, J. G. Gleeson, S. P. Andreoli, D. G. Doherty, A. Lindstrad, C. Golzio, N. Katsanis, L. Pape, E. B. Abboud, A. A. Al-Rajhi, R. A. Lewis, H. Omran, E. Y. Lee, S. Wang, J. M. Sekiguchi, R. Saunders, C. A. Johnson, E. Garner, K. Vanselow, J. S. Andersen, J. Shlomai, G. Nurnberg, P. Nurnberg, S. Levy, A. Smogorzewska, E. A. Otto, and F. Hildebrandt, "Exome capture reveals znf423 and cep164 mutations, linking renal ciliopathies to dna damage response signaling," Cell **150**, 533–548 (2012).
71. R. K. Gupta, Z. Arany, P. Seale, R. J. Mepani, L. Ye, H. M. Conroe, Y. A. Roby, H. Kulaga, R. R. Reed, and B. M. Spiegelman, "Transcriptional control of preadipocyte determination by zfp423," Nature **464**, 619–623 (2010).
72. M. W. Ritzel, S. Y. Yao, M. Y. Huang, J. F. Elliott, C. E. Cass, and J. D. Young, "Molecular cloning and functional expression of cdnas encoding a human na+-nucleoside cotransporter (hcnt1)," Am J Physiol **272**, C707–14 (1997).
73. R. W. Henry, V. Mittal, B. Ma, R. Kobayashi, and N. Hernandez, "Snap19 mediates the assembly of a functional core promoter complex (snapc) shared by rna polymerases ii and iii," Genes & Development **12**, 2664–2672 (1998).
74. L. T. Hogben, *Nature and nurture* (1933).
75. W. F. R. Weldon, "Mendel's laws of alternative inheritance in peas," Biometrika **1**, 228–254 (1902).
76. J. Carayol, G. D. Schellenberg, B. Dombroski, C. Amiet, B. Genin, K. Fontaine, F. Rousseau, C. Vazart, D. Cohen, T. W. Frazier, A. Y. Hardan, G. Dawson, and T. Rio Frio, "A scoring strategy combining statistics and functional genomics supports a possible role for common polygenic variation in autism," Front Genet **5**, 33 (2014).
77. G. J. Lyon, J. O'Rawe, and Wiley, "Human genetics and clinical aspects of neurodevelopmental disorders," in "The Genetics of Neurodevelopmental Disorders.", , K. Mitchell, ed. (2015).
78. W. Huang, S. Richards, M. A. Carbone, D. Zhu, R. R. Anholt, J. F. Ayroles, L. Duncan, K. W. Jordan, F. Lawrence, M. M. Magwire, C. B. Warner, K. Blankenburg, Y. Han, M. Javaid, J. Jayaseelan, S. N. Jhangiani, D. Muzny, F. Ongeri, L. Perales, Y. Q. Wu, Y. Zhang, X. Zou, E. A. Stone, R. A. Gibbs, and T. F. Mackay, "Epistasis dominates the genetic architecture of drosophila quantitative traits," Proceedings of the National Academy of Sciences of the United States of America **109**, 15553–15559 (2012).
79. T. F. Mackay and J. H. Moore, "Why epistasis is important for tackling complex human disease genetics," Genome medicine **6**, 42 (2014).
80. J. S. Bloom, I. M. Ehrenreich, W. T. Loo, T. L. Lite, and L. Kruglyak, "Finding the sources of missing heritability in a yeast cross," Nature **494**, 234–237 (2013).