Genomic diagnosis for children with intellectual disability and/or developmental delay

Kevin M. Bowling, PhD[1], Michelle L. Thompson, PhD[1], Michelle D. Amaral, PhD[1], Candice R. Finnila, PhD[1], Susan M. Hiatt, PhD[1], Krysta L. Engel, PhD[1], J. Nicholas Cochran, PhD[1], Kyle B. Brothers, MD, PhD[3], Kelly M. East, MS, CGC[1], David E. Gray, MS[1], Whitley V. Kelley, MS, CGC[1], Neil E. Lamb, PhD[1], Edward J. Lose, MD[2], Carla A. Rich, MA[3], Shirley Simmons, RN[2], Jana S. Whittle, BS[1], Benjamin T. Weaver, BS[1], Richard M. Myers, PhD[1], Gregory S. Barsh, MD, PhD[1], E. Martina Bebin, MD, MPA[2], Gregory M. Cooper, PhD[1*]

[1]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA
[2]University of Alabama at Birmingham, Birmingham, AL, USA
[3]University of Louisville, Louisville, KY, USA

*Corresponding Author: 601 Genome Way, Huntsville, AL 35806; gcooper@hudsonalpha.org

Abstract

Developmental disabilities have diverse genetic causes that must be identified to facilitate precise diagnoses. We here describe genomic data from 371 affected individuals, 310 of which were sequenced as proband-parent trios. Exomes were generated for 127 probands (365 individuals) and genomes were generated for 244 probands (612 individuals). Diagnostic variants were found in 102 individuals (27%), with variants of uncertain significance (VUS) in an additional 44 (11.8%). We found that a family history of neurocognitive disease, especially the presence of an affected first-degree relative, strongly reduces the diagnosis rate, reflecting both the disease relevance and relative ease of interpretation of *de novo* variants. We also found that improvements to genetic knowledge facilitated interpretation changes in many cases. Through systematic reanalyses we have thus far reclassified 15 variants, with 10.8% of families who initially received a VUS, and 4.7% of families who received no variant, subsequently given a diagnosis. To further such progress, the data described here are being shared through ClinVar, GeneMatcher, and dbGAP. Our results strongly support the value of genome sequencing as a first-choice diagnostic tool and means to continually advance, especially when coupled to rapid and free data sharing, clinical and research progress related to developmental disabilities.

Key Words

Whole genome sequencing, Whole exome sequencing, Pathogenic, Likely pathogenic, Reanalysis, Developmental Delay, Intellectual Disability, *De novo,* Trio sequencing, CSER

Abbreviations

DD/ID: Developmental delay and/or intellectual disability; VUS: Variant of uncertain significance; WGS: Whole genome sequencing; WES: Whole exome sequencing; CSER: Clinical Sequencing Exploratory Research; ACMG: American College of Medical Genetics and Genomics

INTRODUCTION

Developmental delay, intellectual disability, and related congenital defects (DD/ID) affect approximately 1-2% of children in the U.S and pose major medical, financial, and psychological challenges to both the affected children and their families [1]. While many are likely to be genetic in origin, a large fraction of DD/ID cases are either not diagnosed or are misdiagnosed, with many families undergoing a "diagnostic odyssey" involving numerous tests over many years with no effective result. A lack of diagnoses undermines counseling, education, and medical management and slows research towards improving educational or therapeutic options.

Standard clinical genetic testing for DD/ID-affected individuals often includes evaluation with karyotype, microarray, Fragile X, single gene, gene panel, and/or mitochondrial DNA testing [2]. The first two tests examine the whole genome with low resolution, while the latter offer higher resolution but examine only a small fraction of the genome. Whole exome or genome sequencing (WES or WGS) can provide both broad and high-resolution identification of genetic contributors to DD/ID, and thus hold great promise as a far more effective single assay to address this major clinical need [3].

As part of the NHGRI Clinical Sequencing Exploratory Research (CSER) project [4], we have sequenced a cohort of 371 individuals exhibiting DD/ID phenotypes, including, but not limited to, global developmental delay, intellectual disability, speech delay, seizures, autism, and motor movement disorder. One hundred and two affected individuals (27%) received a molecular diagnosis, with the majority of diagnosed individuals harboring a *de novo* variant in a gene previously associated with DD/ID. Fifteen percent of diagnoses were made after initial assessment in timescales ranging from several weeks to several years, with our data and experiences strongly supporting the importance of systematic reanalysis of exome/genome data. We describe 21 variants of uncertain significance (VUS) in 19 genes not currently associated with disease but which are intriguing candidates. The genomic data we describe, which are shared through dbGAP [5], ClinVar [6], and GeneMatcher [7], may prove useful to other clinical genetics labs and researchers interested in the genetics of DD/ID. Our experiences strongly support the value of large-scale sequencing as both a potent first-choice diagnostic tool and means to advance clinical and research progress related to pediatric neurological disease.

MATERIALS AND METHODS

IRB protocol, consent process

This study was approved and monitored by Western Institutional Review Board (20130675) and University of Alabama at Birmingham (UAB) Institutional Review Board (X130201001). All patient identifiers remained at the clinic where they are regularly seen by their physician and medical staff. All information provided to researchers was de-identified in order to minimize the risks to study participants.

Study Participant Population

There was no public recruitment for this study. Primary criteria for study inclusion was that the participants have a clinical relationship with the pediatric neurologist or medical geneticist listed as co-investigators for the study and that the participants be enrolled at North Alabama Children's Specialists in Huntsville, AL. The study physicians offered inclusion into the study to individuals with moderate to severe DD/ID (corresponding to milestones and/or cognitive performance more than two standard deviations below the age-adjusted means) which could not be accounted for by known causes (such as inborn errors of metabolism, lysosomal storage or mitochondrial disorders, Fragile X-associated mental retardation, Rett syndrome or other neurodegenerative conditions, Prader-Willi syndrome, or severe and documented birth asphyxia). Some of the participants exhibited autistic features and other behavioral abnormalities, but these characteristics had to be found in conjunction with other discrete DD/ID phenotypes prior to enrollment into the study. Initially, it was required that all families be a trio consisting of the DD/ID affected proband child and both unaffected (no DD/ID), biological parents. This was requested in order to confirm inheritance of identified variants. As the study progressed, this criterion was modified to allow for children without both available biological parents to enroll in the study.

In addition to having a documented diagnosis of a significant developmental delay and/or intellectual disability, with or without congenital structural or functional anomalies, child participants were required to be at least two years of age and weigh at least 9 kilos (19.8lbs). A parent or legal guardian was also required to give consent for the child to participate. Assent was obtained for children willing and able to do so. Biological parent participants were required to be at least 19 years old, willing to offer consent for themselves and their child, and able to attend two clinic visits for enrollment and return of results. Blood samples were collected from each of the enrolled study participants to be used for exome or genome sequencing at HudsonAlpha Institute for Biotechnology in Huntsville, AL. Independent variant validation by Sanger sequencing was conducted at Emory University in Atlanta, GA.

Return of results

After the results were analyzed, participants were scheduled for a return of results conversation with a medical geneticist and/or pediatric neurologist and a certified genetic

counselor. During that visit, variants of interest were discussed with each of the participant families in a private setting. Clinical significance of the findings were addressed along with any questions posed by the participant family. Genetic counselors provided each family with a document detailing information about variants identified in their sample(s). These letters also contained information about support groups, specialty clinics, or other resources relevant to their specific finding(s).

Exome/Genome sequencing

Genomic DNA was isolated from peripheral blood leukocytes and whole exome or genome sequencing was conducted to a mean depth of 65X or 35X, respectively, with at least 80% of targeted (exome) or hg19 (genome) bases covered at 20X for both sequencing assays. Exome capture was completed using Nimblegen SeqCap EZ Exome version 3 and sequencing was conducted on the Illumina HiSeq 2000 or 2500. Whole genome sequencing was done on the Illumina HiSeq X. Reads were aligned to reference hg19 using bwa (0.6.2) [8]. GATK best practice methods [9] were used to identify variants, with samples called jointly in batches of 10-20 trios. Maternity and paternity of each parent was confirmed by whole exome/genome kinship coefficient estimation using KING [10].

Whole Genome CNV calling

CNVs were called from whole genome bam files using two callers, ERDS [11] and read Depth [12]. Overlapping calls within the output of each caller were merged, then the two call sets were combined, keeping unique calls from each call set and merging any calls common to both callers that had 90% reciprocal overlap. Calls were retained when less than 50% of the call was made up of segmental duplications and when similar calls were observed in five or less unaffected parents. Of the calls meeting that criteria, any calls within 5kb of a known DD/ID gene, or within 5kb of an OMIM morbidity gene, or intersecting one or more exons of any gene, were manually inspected. Call inspection consisted of graphing the read depth of the call and flanking regions, and graphing the allele ratio of common SNPs in and around the call, and using the graphs to assess the quality of the calls. Calls with read depths and allele ratios consistent with a copy number change were considered accurate.

Filtering

De novo variants were identified as heterozygous calls in the proband, and we required that the proband and their parents each had a read depth of 10X, at least 20% of proband reads contained the alternate allele, less than 5% of parental reads contained the alternate allele, and a minor allele frequency (MAF) ≤1% in 1000 Genomes, EVS and ExAC. Variants were also restricted to those where two or less alternate alleles were identified, the batch allele count was one, there were less than 40 counts in an internal allele frequency database, and the VQSR Filter class was not “Low Quality”. All candidate *de novo* variants that affected a protein, were within 10 nt of a splice site, or had a scaled CADD score ≥10 [13] were manually reviewed. Variants were similarly identified using X-linked, compound heterozygote, and recessive

inheritance filters. No specific gene lists were used for identifying DD/ID variants in the proband.

In cases where one or both parents were unavailable, rare, potentially damaging variants were identified in the proband as heterozygous calls with a read depth of 10X, at least 20% of proband reads contained the alternate allele, and a minor allele frequency ≤0.1% in 1000 Genomes, EVS and ExAC. Variants were further restricted to those that affected a protein, were within 10 nt of a splice site, or had a scaled CADD score ≥20 [13]. When available, parent read depth was required to be 10 reads, with less than 5% of reads representing the alternate allele. Variants were similarly identified using X-linked, potential compound heterozygote and recessive inheritance filters in probands with one or no parents available.

Three gene lists were employed for secondary variant identification in parent participants. Variants meeting the above guidelines in genes included on the ACMG gene list [14] were reviewed. Variants with recessive inheritance in genes that are associated with disease in OMIM were also reviewed. Lastly, carrier status was assessed for three genes: *CFTR*, *HBB*, and *HEXA*.

Secondary variants were identified in parents as non-reference calls in which there were two or less alternate alleles, a batch allele frequency of ≤10, <40 counts in an internal allele frequency database, and the VQSR Filter class was "PASS". Variants were also restricted to those that either affected a protein, a splice site, or had a scaled CADD score ≥15 [13]. Note that for the ACMG and OMIM gene lists, a minor allele frequency of ≤1% in 1000 Genomes, EVS and ExAC was used, while for the carrier gene list, a minor allele frequency ≤25% in 1000 Genomes and ExAC was used.

Rare variants (minor allele frequency of ≤1% in 1000 Genomes, EVS and ExAC) that have been submitted to the ClinVar database [6] as pathogenic or likely pathogenic (or conflicting reports of pathogenicity, but at least one report as pathogenic or likely pathogenic) were also manually reviewed in each family.

Analysis of Trios as Singletons

For all families that underwent whole genome sequencing, we used family-specific VCFs that were pre-filtered for rare variants (≤ 2 alternate alleles, ≤ 1% MAF in EVS, ExAC, 1000G, batch allele count ≤10, ≤40 counts in an internal allele frequency database, CADD ≥10 or protein-altering, or splice region, and exclude VQSR Low Quality) and then systematically decremented the batch allele count (AC) to reduce the allele count based on alleles contributed by parents. Using these AC-decremented files, we then filtered each proband (individually) for rare DeNovo-Like variants or compound heterozygous candidate variants.

Rare DeNovo-Like variants were defined as those where the proband was not homozygous reference, restricting to those with two or less alternate alleles, batch allele counts of one, ≤ four counts in an internal allele frequency database, EVS, ExAC Global allele frequencies of ≤

0.01% (≤0.2% each subpopulations), 1000G Global allele frequency of ≤0.2% (≤0.5% in AA and EUR subpopulations), and LowQuality VQSR Filtered variants were excluded. Additional variations included restrictions based on CADD scaled scores of ≥10 or 15 (unless protein-altering, or within 10 nt of a canonical splice site).

Rare potential compound heterozygous variants were defined as those where the proband had two heterozygous calls in the same gene, restricting to those with two or less alternate alleles, batch allele counts of ≤ 5, ≤ 40 counts in an internal allele frequency database, EVS, ExAC Global allele frequencies of ≤ 0.01% (≤0.2% each subpopulations), 1000G Global allele frequency of ≤0.2% (≤0.5% in AA and EUR subpopulations), and LowQuality VQSR Filtered variants were excluded.

Additional variations included restrictions to genes associated with disease in OMIM or DDG2P, or restrictions based on RVIS score (top 10% intolerant genes, top 20%, etc.).

Variants from these filter results were then ranked by CADD score. In cases where we identified a returned variant (VUS, likely athogenic or pathogenic), we calculated the CADD-based rank of that variant.

Reanalysis of Exomes

In December 2015, variants from all 120 exome families were reannotated and refiltered. Updated annotations included: ClinVar (accessed 12/3/2015, clinvar_20151201.vcf), ExAC release r0.2, data from DDG2P (https://decipher.sanger.ac.uk/ddd#ddgenes, accessed November 18, 2014 and December 3, 2015; ddg2p_20141118 and ddg2p_20150701), and data from several publications [15-18].

All 120 exome families were trios, and filtering for primary variants was performed as described above. One additional, less-stringent *de novo* filter was also employed, which removed the depth requirement for all members of the trio. Filters were also utilized to identify rare variants submitted to ClinVar as pathogenic or likely pathogenic.

Genes with no disease association that harbored suspicious VUSs were submitted to GeneMatcher (https://genematcher.org/).

RNA Isolation

2.5 mL of blood was collected in PAXgene RNA tubes (PreAnalytiX #762165) according to the manufacturer's instructions and stored short-term at -20°C. RNA was isolated using a the PAX gene Blood RNA Kit (Qiagen #762164) according to the manufacturer's instructions. Isolated RNA was quantified by Qubit® (Thermo Fisher #Q32855).

cDNA Synthesis

First strand synthesis of cDNA was performed from 250-500 ng of RNA using either Superscript™ III (Thermo Fisher #18080044) or Superscript™ IV VILO™ (Thermo Fisher #11766050) according to manufacturer's instructions using either random hexamers or a mix of random hexamers and oligodT, with the exception that reverse transcription was carried out at 55°C for 20 minutes for Superscript™ IV VILO™.

mTOR PCR

Amplicons were obtained by amplifying 3 ng of template cDNA using Phusion polymerase in HF buffer (NEB #M0531L) with 500 nM forward and reverse primers (IDT, Table S1). Cycling conditions were as follows: (98°C, 30s), (98°C, 10s; 61°C, 30s; 72°C, 90s)x35, (72°C, 10m), (4°C, ∞).

ALG1 PCR

Amplicons were obtained by amplifying 30 ng of template cDNA using Phusion polymerase (NEB #M0531L) with 1 μM forward and reverse primers (IDT, Table S1) and 3% DMSO. Cycling conditions were as follows: (98°C, 30s), (98°C, 10s; 58°C, 30s; 72°C, 90s)x7, (98°C, 10s; 60°C, 30s; 72°C, 80s)x7, (98°C, 10s; 62°C, 30s; 72°C, 70s)x7, (98°C, 10s; 64°C, 30s; 72°C, 60s)x7, (98°C, 10s; 66°C, 30s; 72°C, 50s)x7, (72°C, 10m), (4°C, ∞).

ROCK2 Western Blot

Live cells were collected using cell processing tubes (BD #362760), isolated according to the manufacturer's instructions, and stored in liquid nitrogen in CTS™ Synth-a-Freeze® Medium (Thermo Fisher # A13713-01) until use. Cell pellets were homogenized in RIPA buffer (1x PBS, 0.5% Sodium Deoxycholate, 0.1% SDS, 1% NP-40) supplemented with 1x cOmplete™ protease inhibitor cocktail (Roche #11697498001). Lysates were cleared at 500 x g for 5 minutes, then protein concentration was measured by Qubit® (Thermo Fisher # Q33212) and lysates were equalized in concentration by dilution with lysis buffer. 100 μg of protein was loaded per lane, blots were blocked for 1 hour at room temperature in 5% milk in 0.05% PBS-T, then blots were probed with 1:250 rabbit anti-ROCK2 (N-terminal, Sigma-Aldrich HPA007459) or 2 μg/mL mouse anti-ROCK2 (C-terminal, Abcam ab56661) overnight at 4°C in 0.05% PBS-T. Blots were probed with β-actin as a loading control at 1:10,000 (Cell Signaling #8H10D10) for 1 hour at room temperature. Secondary antibodies were HRP-conjugated goat anti-rabbit IgG (Thermo Fisher #31460) and HRP-conjugated goat anti-mouse IgG (Thermo Fisher #31430). Signal was detected using an enhanced chemiluminescent substrate (Thermo Fisher #34095).

qPCR

Quantitative PCR was performed on cDNA synthesized as described above after being diluted 5 fold. 10 μl reactions were composed of 2 μl diluted cDNA, 5 μl Power SYBR Green 2X master mix

(Applied Biosystems ref#4367659), 1.25 µl forward oligo (10 µM), 1.25 µl reverse oligo (10 µM), and 0.5 µl H_2O. The oligos were designed to amplify a 50-70 bp fragment and the melting temperature of the product was used to verify the correct specificity of the oligos during each qPCR. At least two independent cDNA synthesis reactions were performed for each biological sample and the reactions were performed in quadruplicate technical replicates for each cDNA synthesis. The reactions were performed using Applied Biosystems QuantStudio 6 Flex. C_T values were obtained and used to calculate the $\Delta\Delta C_T$ values as a percentage of the affected individuals.

RESULTS

Demographics of study population

We enrolled 339 families (977 individuals total) that included at least one child with a documented diagnosis of developmental delay and/or intellectual disability (DD/ID), with or without structural or functional anomalies. Participating families consisted mostly of parent-proband trios, with a smaller group representing one parent-proband duos and child-only singletons. Twenty-nine of the enrolled families had more than one affected child, resulting in a total of 371 affected individuals. Whole exome sequencing (WES) was performed on 365 individuals (127 affected) and whole genome sequencing (WGS) was performed on 612 individuals (244 affected). Exomes and genomes were sequenced to an average depth of 71X and 35X, respectively, with >80% of bases covered ≥ 20X in each experiment type. DNA from affected participants that received exome sequencing was also analyzed via a SNP genotyping array if not previously done so in a clinical setting.

The study population had a mean age of 11 years and was 58% male. Affected individuals displayed a wide array of symptoms described by 333 unique HPO terms with over 90% of individuals displaying mild to severe intellectual disability, 69% with speech delay, 45% with seizures, and nearly 10% with microcephaly or macrocephaly. Eighteen percent of individuals had an abnormal brain magnetic resonance imaging (MRI) result. Nearly 81% of individuals had some form of previous genetic testing that failed to give a diagnosis prior to enrollment into this study (Table 1).

DD/ID-associated genetic variation

WES and WGS data were processed with standard alignment and genotyping protocols to produce variant lists in each family that were subsequently subject to annotation and filtering, followed by manual review of short candidate lists (see Methods). We used ACMG-guided designations of pathogenicity [19] to classify variants according to their potential relevance to phenotype within each proband. All variants described here and returned to patients were confirmed by Sanger sequencing in affected individuals, and when available, their parents or other enrolled family members.

Among the 371 DD/ID-affected individuals included in our study, 146 (39%) had DD/ID-associated or potentially DD/ID-associated genetic variation deemed to be returnable (pathogenic, likely pathogenic, VUS) following thorough variant review (Table 2). More specifically, 102 (27%) affected individuals harbored a likely pathogenic or pathogenic variant, while 44 (11.8%) harbored a VUS. We hereafter refer to pathogenic or likely pathogenic variants, but not VUSs, as “diagnostic”. Given that the majority of probands had been tested via microarray and/or karyotype prior to their enrollment in this study, diagnostic or VUSs were large copy-number variants (CNVs) were detected in 11 affected individuals (Table 2).

Most diagnostic variants occurred *de novo* (74%), while 12% of individuals inherited diagnostic variants as either compound heterozygotes or homozygotes (Figure 1A). An additional 7% were male individuals who inherited an X-linked variant maternally. A smaller percentage (7%) of diagnosed participants were sequenced with either one or no first-degree relative, therefore mode of inheritance was indeterminable (Figure 1A). Most diagnostic variants were missense mutations (53%), while another 38% were nonsense or frameshift mutations. A small number of variants led to altered splicing (7%) and inframe deletion of amino acids (2%; Figure 1B).

Mutations deemed to be returnable to participants and their families were identified in 99 different genes, excluding large CNVs, with variants in 24 (24%) of these genes observed in two or more unrelated individuals (Table 3; Table S2). This includes a number of well-characterized genes that have been previously implicated in DD/ID including *SCN1A* (Dravet syndrome MIM:607208), *MECP2* (Rett syndrome, MIM:312750), *SLC2A1* (GLUT1 deficiency, MIM:612126), *ANKRD11* (KBG syndrome, MIM:148050), *CREBBP* (Rubinstein-Taybi syndrome, MIM:180849), *SATB2* (Glass syndrome, MIM:612313), *TCF4* (Pitt-Hopkins syndrome, MIM:610954), and *WDR45* (neurodegeneration with brain iron accumulation, MIM:300894). *SCN1A* and *MTOR* were the most frequent genes in which diagnostic variants were identified, with diagnoses made in four unrelated families in each gene.

Diagnostic rates across families of varying structure and phenotypic complexity

Affected individuals were categorized into one of three family structures based on the number of biological parents that were sequenced along with the DD/ID-affected proband(s): trios consisted of the affected individual and both parents; duos consisted of the affected individual and one parent; and singletons were comprised of only the affected individual. In total, we enrolled 310 affected individuals as part of a trio, 41 affected individuals as part of a duo, and 20 affected individuals as singletons. A diagnostic result was found in 30% of trio individuals, 17% of individuals who were part of a duo, and 15% of singletons (Table 1). The higher diagnostic rate in proband-parent trios reflects the benefit of using parental genotypes to identify *de novo*, causal variation.

Given that one or both biological parents were unavailable or unwilling to participate in duo or singleton analyses, the diagnostic rate comparisons among trios/duos/singletons may be confounded by other disease-associated factors (depression, schizophrenia, ADHD, etc.). For example, a number of the singleton probands were adopted owing to death or disability associated with neurological disease in their biological parents. To further assess the relationship between diagnostic rate and family history, we separated probands into three types: simplex families in which there was only one affected proband and no 1st to 3rd degree relatives reported to be affected with any neurological condition (n=93); families in which the enrolled proband had no affected 1st degree relatives but with one or more reported 2nd or 3rd degree relatives who were affected with a neurological condition (n=85); and multiplex families in which the proband had at least one first degree relative affected with a neurological condition, including in some cases one or more siblings with DD/ID that were enrolled into this

study (n=123) (Table 4). For 38 probands, there was limited or no family history information and they were excluded from this analysis.

In total, a diagnosis was made in 24 (~20%) of the 123 multiplex families (20 out of 97 when considering only trios), in contrast with 36 (39%) of 93 simplex families (32 out of 80 when considering only trios), suggesting a diagnostic success rate that is 2X higher for probands in simplex, relative to multiplex, families. While larger sample sizes are needed to confirm this effect, the diagnostic rate difference is significant whether or not all enrolled families (p=0.017) or only those sequenced as trios (p=0.033) are considered. Rates in families that were neither simplex nor multiplex (i.e., proband lacks an affected 1st degree relative but has one or more affected 2nd or 3rd degree relatives) were intermediate, with 26% of all such families diagnosed (28% for those sequenced as part of trios). Of relevance to the trio/duo/singleton comparison described above, 11 of 13 (85%) singletons for whom we had some information on family history had an affected first-degree relative, in contrast with 41% for duos and 39% for trios (Table 4). This enrichment for affected first-degree relatives likely contributed to the generally reduced diagnostic rate in singletons in this study.

Multiplex family diagnoses include examples of both expected and unexpected inheritance patterns. For example, two, affected male siblings from family 315 were both found to be hemizygous for a nonsense mutation in *PHF6* (Börjeson-Forssman-Lehmann syndrome MIM:301900) inherited from their unaffected mother. In family 180, we found the affected individual to be compound heterozygous for two different variants in *GRIK4* (subunit of the inotropic glutamate receptor), with one allele inherited from each parent. Interestingly, both the mother and father of the affected-individual report psychiatric illness, and extended family history of psychiatric phenotypes is noted. In contrast, we observed multiple independent *de novo* causal variants within two families. Affected siblings in family 135 each harbored a *de novo* variant deemed to be returnable in a different gene, one in *SPR* and one in *RIT1*, while a pair of probands (75 and 78) who were second degree relatives to one another also harbored independent, *de novo* diagnostic events, one each in *DDX3X* and *TCF20* (Table S2).

Alternative mechanisms of disease

While the majority of DD/ID-associated genetic variations detected in our study has been missense, frameshift, or nonsense mutations (Figure 1B), a subset of sequenced affected individuals harbor variants leading to altered splicing, and in some cases, potentially alternative mechanisms of disease. As an example, we sequenced an affected 14-year-old girl (00003-C, Table S2) who presented with severe intellectual disability, seizures, speech delay, autism and stereotypic behaviors. Exome sequencing revealed a point mutation near the canonical acceptor splice site of intron 2 in *MECP2* (c.27-6C>G, MIM:312750), identical to a previously observed *de novo* variant in a 5-year-old female patient who presented with several features of Rett syndrome, but lacked deceleration of head growth and exhibited typical growth development [20]. Laccone, et al. showed by RT-PCR that the variant produces a cryptic splice acceptor site that introduces five additional nucleotides into the mRNA sequence between exons 2 and 3, ultimately resulting in a transcript with a frameshift (R9fs24X) [20]. It is likely that

both the canonical and cryptic splice sites exhibit functionality, allowing for a majority of *MECP2* transcripts to be wild type, resulting in the milder Rett phenotype observed in both the individual described here and the girl described by Laccone and colleagues [20].

In another DD/ID-affected child (00126-C), we identified compound heterozygous variants in *ALG1* (Table S2). This child presents clinically with phenotypes consistent with those previously described for ALG1-CDG (congenital disorder of glycosylation) including severe intellectual disability, hypotonia, growth retardation, microcephaly and seizures [21]. The paternally inherited missense mutation (c.773C>T, S258L) has been previously shown to be pathogenic [22]. The maternally inherited variant, previously unreported (c.1187+3A>G), is three bases downstream of an exon/intron junction (Figure 2A). To confirm that this D-3 splice results in altered splicing, we performed RTqPCR across exon-intron junctions from patient blood RNA, and found that intron 11 of *ALG1* is completely retained in both the affected proband and the mother, from whom the variant was transmitted (Figure 2A-D). The retention of intron 11 results in stop-gain after addition of 84 nucleotides (28 codons) to the transcript.

In a separate family consisting of affected maternal half siblings (00218-C and 00218-S, Table S2, Figure 2E) we observed variation in a canonical splice acceptor site (c.505-2A>G) of *MTOR* intron 4. The half siblings described here both have intellectual disability; the younger sibling has no seizures but also presents with facial dysmorphism, speech delay, and autism, while his older sister exhibits seizures. It is presumed that the maternal half siblings inherited the splice variant from their mother, for whom DNA was not available, but who was reported to exhibit seizures as well. To determine the consequences of the *MTOR* canonical splice variant on mRNA synthesis in the half siblings, we conducted RTqPCR and Sanger sequencing using blood RNA. Analysis of the transcripts from the affected siblings revealed inclusion of 134 nucleotides from the 3' end of intron 4, ultimately leading to the addition of 20 additional amino acids after exon 4, with the 21st additional codon resulting in stop-gain (Figure 2F-H, Figure S1). Because the stop-gain occurs early on in protein translation, the splice variant likely results in loss-of-function. Previous studies have associated mutations in *MTOR* with a broad spectrum of phenotypes including epilepsy, focal epilepsy without brain abnormalities and normal development, hemimegalencephaly, and intellectual disability [23-29]. However, previously reported pathogenic variants in *MTOR* are all missense and suspected to result in gain-of-function [30]. Owing to this mechanistic uncertainty, we have classified this variant as a VUS. However, given the overlap between phenotypes observed in this family and previously reported families, we find this variant to be highly intriguing and suggestive that *MTOR* loss-of-function variation may also give rise to a phenotype. *MTOR* is highly intolerant of mutations in the general population (RVIS score of 0.09%), further supporting the hypothesis that loss-of-function is deleterious and likely to lead to disease consequences.

Proband-only versus trio sequencing

Our trio-based study design allows rapid identification of *de novo* variants, which are enriched among variants that are causally related to deleterious, pediatric phenotypes [31]. However, we were also interested in assessing to what extent our diagnostic rate would differ if we had only

enrolled probands. Thus, and to avoid the confounding of family history differences among trios, duos, and singletons (see above), we subjected variants within all trio-based probands to various filtering scenarios blinded to parental status and assessed the CADD score [13] ranks of *de novo* variants we previously found to be diagnostic (Figure 3). While parentally informed *de novo* filters proved most effective at highly ranking causal variants (e.g., >60% of diagnostic variants were the top-ranked variant among the list of all *de novo* events in each given patient), filters defined without parental information were also highly effective. For example, among all rare, protein-altering mutations found in genes associated with Mendelian disease via OMIM or associated with DD/ID via the DECIPHER project [32], nearly 20% of diagnostic variants were the top-ranked variant in the given patient, most ranked among the top 5, and >80% ranked among the top 25, a number of variants that can be readily manually assessed.

We found VUSs to be more difficult to identify without parental information (Figure S2), owing largely to the fact that many VUSs do not affect genes listed in OMIM or DDG2P as being associated with disease, resulting in lower ranks (i.e., the more obviously deleterious variants in these genes are more likely to be diagnostic). Thus, while most currently diagnostic variants could be found (with additional curation time) without parental sequence, such data is tremendously valuable for the discovery of potential novel disease associations.

Secondary findings in DD/ID-affected individuals and participating parents

During analysis of sequencing data, we often encountered variation in genes not related to the indication for testing (DD/ID; Table S3). Overall, secondary findings were observed in 8% of parental DNA samples. One and a half percent of these corresponded to secondary diagnoses, such as the identification of pathogenic variants in *SLC22A5* that explain one parent's self-reported primary carnitine deficiency (MIM:212140). We also examined the 56 genes named by the American College of Medical Genetics (ACMG) as potentially harboring actionable secondary findings that should be considered by diagnostic labs [14]. Thus far, we have returned pathogenic/likely pathogenic variants that reside in ACMG genes to 13 study parents (2.1%), a rate similar to that observed in previously described cohorts [14,33]. Finally, we performed a limited carrier screening assessment, identifying 27 (4.5%) parent participants as carriers of pathogenic/likely pathogenic variation in genes associated with recessive disease. All of these were variants in *HBB* (MIM:603903), *HEXA* (MIM:272800), or *CFTR* (MIM:219700). However, we also assessed parents as mate pairs and searched for genes in which both are heterozygous for a diagnostic recessive allele. We have to date, only identified one parental pair as carriers for pathogenic variation in *ATP7B*, associated with Wilson disease (MIM:277900).

Reanalysis of exomes and genomes

Due to the steady and progressive increase in available information on genomic variation in the scientific community, we have, and continue, to perform regular reanalysis of patient data to reduce the high rate of unsolved cases. We approached reanalysis in three ways: 1) systematic reanalysis of old data, with the goal of reassessing each dataset 12 months after initial analysis; 2) continual mining of all variant data based on new publications related to DD/ID gene

discovery; and 3) use of GeneMatcher [7] to aid in the interpretation of variants in genes thought to be of uncertain disease significance.

As shown in Table 5, these combined efforts led to a change in pathogenicity score for 15 variants in 17 individuals. For 9 of the 15 variants, a new publication became available that allowed a variant that had not been previously reported, or that was previously reported as a VUS, to be reclassified as diagnostic. Three additional upgrades were a result of discussions facilitated by GeneMatcher [7], while the remaining upgrades resulted from reduction in filter stringency (changes to read depth and batch allele frequency) or clarification of clinical phenotype. Of all VUSs identified through our study, 5 (10.8%) have subsequently been upgraded to diagnostic. The most rapid of these, a *de novo* variant in *DDX3X*, was upgraded approximately 1 month after initial assessment, while a variant in *EBF3* was upgraded almost 2.5 years post initial assessment. These data clearly indicate that VUSs, especially when identified via parent-proband trio sequencing, have considerable diagnostic potential. Moreover, of the 211 families who originally received no DD/ID-associated findings, diagnostic variation was identified for 10 (4.7%) through reanalysis. For the 15 variants that were ultimately upgraded to diagnostic, time from initial review to upgrade ranged from 1-29 months, with upgrade occurring within 1 year of primary assessment for several variants. Taken together, this data suggests that regular reanalysis of both uncertain and negative sequencing data is an effective mechanism to improve diagnostic yield over both short and long-term periods.

Identification of novel candidate DD/ID genes

We have also identified 21 variants harbored within 19 genes with no known disease association but which are interesting candidates. For example, in one DD/ID-affected child we identified a variant in ROCK2 leading to early protein termination starting at codon 714, with absence of mutant ROCK2 protein confirmed by western blot (Figure S3). ROCK2 is a conserved Rho-associated serine/threonine kinase involved in a number of different cellular processes including actin cytoskeleton organization, proliferation, apoptosis, extracellular matrix remodeling and smooth muscle cell contraction, and has an RVIS score placing it among the top 17.93% most intolerant genes [34-36]. Additionally, in two affected individuals, we identified *de novo* variation in *NBEA*, a nonsense variant at codon 2213 of 2946, and a missense, H946Y. NBEA is a kinase anchoring protein with roles in the recruitment of cAMP dependent protein kinase A to endomembranes near the trans-Golgi network [37]. It is unclear whether alterations in this gene are linked to DD/ID; although its RVIS score of 0.75% suggests that it does not tolerate genetic variation. While all these variants remain VUSs, the evidence for selective conservation suggests these may be candidate genes that ultimately hold up as being disease associated, similar to VUSs in other genes like *EBF3*.

DISCUSSION

As part of the CSER study performed at HudsonAlpha, we have sequenced 371 individuals that presented clinically with a variety of different phenotypes, all affected with DD/ID related

conditions. Twenty-seven percent of sequenced individuals received a molecular diagnosis, with most of these resulting from a *de novo* variant in a known DD/ID gene. This diagnostic rate is likely to be an underestimate for what might be observed in less screened populations; 81% of our enrolled participants had undergone previous genetic testing and remained undiagnosed by standard clinical diagnostic methods. We found that the diagnostic yield is also impacted by presence of disease in related family members, as our success rate drops from 39% for probands without any affected relatives to 20% for probands with one or more affected first degree relatives. These data are consistent with the observation of higher causal variant yields in simplex families relative to multiplex families affected with autism, [38] and in part, reflects the eased interpretation of *de novo* causal variation relative to inherited, and likely in many cases variably expressive or incompletely penetrant, causal variation (e.g., 16p12) [39,40].

Of the 371 DD/ID-affected individuals that were sequenced, 127 underwent exome sequencing, while 244 were genome sequenced. The diagnostic rate was not significantly different between the two assays when considering only SNVs or small indels. However, genome sequencing is a better assay for detection of CNVs, [41] and while our patient population is heavily depleted for large causal CNVs owing to prior array or karyotype testing, we have identified diagnostic CNVs in 8 individuals.

Further, WGS in principle allows for discovery of pathogenic genetic variation outside of coding exons. For example, we identified a *de novo* variant deep within intron 20 of *SCN1A* in an individual who presented clinically with Dravet syndrome (MIM:607208). This individual had previously undergone *SCN1A* single-gene testing, and lacked coding or splice variation. The intronic variant identified in this study has a CADD scaled score of 19.28 (higher than ~99% of all possible hg19 SNVs and near the median of pathogenic variants reported to ClinVar), and affects a highly conserved nucleotide within a highly conserved region of hg19. While this remains a VUS, we believe it has a reasonable chance of being pathogenic and is anecdotally supportive of the value of WGS as a benefit to longer-term research discovery.

We have also demonstrated the value and effectiveness of reanalysis, having diagnosed an additional 17 DD/ID-affected individuals (16.7% of total diagnoses, 4.6% of total proband participants). Given the rates of progress in Mendelian disease genetic discovery [42] and the development of new genomic annotations, we believe that systematic reanalysis of genomic data should become standard practice given the considerable benefits to both research and clinical goals. While non-trivial, reanalysis requires relatively modest, especially in proportion to the initial sequencing and analysis, investments of time and cost. Further, as more pathogenic coding and non-coding variants are found, the reanalysis benefit potential is largest for WGS relative to WES; the former typically has slightly better coverage of coding exons in both our data (Table S4) and previous studies [41], and reanalysis of non-coding variants is precluded entirely by WES.

Although sequencing parent-proband trios is the most powerful way of identifying disease causal genetic variation in this population, we are cognizant of the fact that proband-only sequencing would allow for sequencing more affected individuals, with the potential of making

more diagnoses per dollar spent. To evaluate this possibility, we conducted an analysis to ask how efficiently we can detect diagnostic variation when sequencing affected individuals only, in the absence of any first-degree relatives. While our analysis suggests that sequencing only the proband certainly leads to an increase in the number of variants that have to be manually curated and interpreted, it also suggests that focusing on rare coding variants in disease-associated genes ranked by CADD scores can lead to efficient diagnosis. However, VUSs are more difficult to identify without parental sequence data, and thus proband-only approaches ultimately confer less benefit in terms of discovery of new disease associations.

Variation detected through our studies has already helped lead to the discovery of a novel DD/ID-associated gene. We identified two patients that harbored variants in *EBF3* (one variant results in altered splicing; the other a missense mutation), which is a highly conserved transcription factor involved in neurodevelopment that is relatively intolerant to mutations in the general population (RVIS: 6.78%). Through collaboration with other researchers via GeneMatcher, we were able to identify a total of 10 DD/ID-affected individuals who also harbor *EBF3* variants, supporting that *de novo* disruption of *EBF3* function leads to neurodevelopmental phenotypes [43].

We have demonstrated the benefits of utilizing genomic sequencing to detect causal variation in children exhibiting DD/ID phenotypes who have been clinically and/or genetically undiagnosed. Through our study, we have diagnosed 102 DD/ID-affected individuals who may not have otherwise received a specific diagnosis. It is important to note that the 27% of patients who received a diagnosis in our study are additive to diagnostic rates determined by karyotyping, microarrays and single-gene testing as most of the affected individuals that we sequenced had previous genetic testing. Moreover, we believe our data demonstrate that large-scale sequencing is the best tool for DD/ID-affected individuals given the extreme genetic heterogeneity observed, as we identified variants in 71 different genes deemed to be diagnostic across 102 affected individuals. In general, our data and experience point to WGS as a highly effective choice as the first diagnostic test for DD/ID, with its benefits and effectiveness likely to grow over time both by accelerating research and by facilitating effective longer-term reanalysis.

Acknowledgements

We are grateful to the patients and their families who contributed to this study. We thank the HudsonAlpha Software Development and Informatics team and the Genome Sequencing Center who contributed to data collection and analysis. We would also like to thank Jeremy Herskowitz for discussions about ROCK2. This work was supported by grant from the US National Human Genome Research Institute (NHGRI; UM1HG007301).

References

1. Boyle CA, Boulet S, Schieve LA, et al. Trends in the prevalence of developmental disabilities in US children, 1997-2008. *Pediatrics.* 2011;127(6):1034-1042.
2. Sun F, Oristaglio J, Levy SE, et al. Genetic Testing for Developmental Disabilities, Intellectual Disability, and Autism Spectrum Disorder. *Rockville (MD): Agency for Healthcare Research and Quality (US).* 2015:1-55.
3. Stavropoulos D, Merico D, Jobling R, Bowdin S. Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. *npj Genomic Medicine.* 2016;1:1-9.
4. Green RC, Goddard KA, Jarvik GP, et al. Clinical Sequencing Exploratory Research Consortium: Accelerating Evidence-Based Practice of Genomic Medicine. *Am J Hum Genet.* 2016;98(6):1051-1066.
5. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet.* 2007;39(10):1181-1186.
6. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 2016;44(D1):D862-868.
7. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat.* 2015;36(10):928-930.
8. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754-1760.
9. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43(5):491-498.
10. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics.* 2010;26(22):2867-2873.
11. Zhu M, Need AC, Han Y, et al. Using ERDS to infer copy-number variants in high-coverage genomes. *Am J Hum Genet.* 2012;91(3):408-421.
12. Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One.* 2011;6(1):e16327.
13. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014;46(3):310-315.
14. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med.* 2013;15(7):565-574.
15. Consortium EK, Project EPG, Allen AS, et al. De novo mutations in epileptic encephalopathies. *Nature.* 2013;501(7466):217-221.
16. Karaca E, Harel T, Pehlivan D, et al. Genes that Affect Brain Structure and Function Identified by Rare Variant Analyses of Mendelian Neurologic Disease. *Neuron.* 2015;88(3):499-513.
17. Krumm N, Turner TN, Baker C, et al. Excess of rare, inherited truncating mutations in autism. *Nat Genet.* 2015;47(6):582-588.
18. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. *Nat Genet.* 2014;46(9):944-950.
19. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* 2015;17(5):405-424.

20. Laccone F, Huppke P, Hanefeld F, Meins M. Mutation spectrum in patients with Rett syndrome in the German population: Evidence of hot spot regions. *Hum Mutat.* 2001;17(3):183-190.
21. Ng BG, Shiryaev SA, Rymen D, et al. ALG1-CDG: Clinical and Molecular Characterization of 39 Unreported Patients. *Hum Mutat.* 2016;37(7):653-660.
22. Dupre T, Vuillaumier-Barrot S, Chantret I, et al. Guanosine diphosphate-mannose:GlcNAc2-PP-dolichol mannosyltransferase deficiency (congenital disorders of glycosylation type Ik): five new patients and seven novel mutations. *J Med Genet.* 2010;47(11):729-735.
23. Mirzaa GM, Campbell CD, Solovieff N, et al. Wide spectrum of developmental brain disorders from megalencephaly to focal cortical dysplasia and pigmentary mosaicism caused by mutations of MTOR. *JAMA Neurol.* 2016;73(7):836-845.
24. D'Gama AM, Geng Y, Couto JA, et al. Mammalian target of rapamycin pathway mutations cause hemimegalencephaly and focal cortical dysplasia. *Ann Neurol.* 2015;77(4):720-725.
25. Lim JS, Kim WI, Kang HC, et al. Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. *Nat Med.* 2015;21(4):395-400.
26. Nakashima M, Saitsu H, Takei N, et al. Somatic Mutations in the MTOR gene cause focal cortical dysplasia type IIb. *Ann Neurol.* 2015;78(3):375-386.
27. Mroske C, Rasmussen K, Shinde DN, et al. Germline activating MTOR mutation arising through gonadal mosaicism in two brothers with megalencephaly and neurodevelopmental abnormalities. *BMC Med Genet.* 2015;16:102.
28. Baynam G, Overkov A, Davis M, et al. A germline MTOR mutation in Aboriginal Australian siblings with intellectual disability, dysmorphism, macrocephaly, and small thoraces. *Am J Med Genet A.* 2015;167(7):1659-1667.
29. Smith L, Saunders C, Dinwiddie D. Exome Sequencing reveals de novo germline mutation of hte mammalian target of rapamycin (MTOR) in a patient with megalencephaly and intractable seizures. *Journal of Genomes and Exomes.* 2013:63-72.
30. Moller R. Germline and somatic mutations in MTOR gene in focal epilepsies with and without brain malformations. *Neurology: Genetics.* In press.
31. Vissers LE, de Ligt J, Gilissen C, et al. A de novo paradigm for mental retardation. *Nat Genet.* 2010;42(12):1109-1112.
32. McRae J, Clayton S, Fitzgerald T, et al. Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. *BioRxiv.* 2016.
33. Johnston JJ, Rubinstein WS, Facio FM, et al. Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. *Am J Hum Genet.* 2012;91(1):97-108.
34. Amano M, Nakayama M, Kaibuchi K. Rho-kinase/ROCK: A key regulator of the cytoskeleton and cell polarity. *Cytoskeleton (Hoboken).* 2010;67(9):545-554.
35. Shi J, Wei L. Rho kinase in the regulation of cell death and survival. *Arch Immunol Ther Exp (Warsz).* 2007;55(2):61-75.
36. Khalil RA. *Regulation of Vascular Smooth Muscle Function.* San Rafael (CA): Morgan & Claypool Life Sciences; 2010.
37. Castermans D, Wilquet V, Parthoens E, et al. The neurobeachin gene is disrupted by a translocation in a patient with idiopathic autism. *J Med Genet.* 2003;40(5):352-356.
38. Klei L, Sanders SJ, Murtha MT, et al. Common genetic variants, acting additively, are a major source of risk for autism. *Mol Autism.* 2012;3(1):9.
39. Girirajan S, Moeschler J, Rosenfeld J. 16p12.2 Microdeletion. In: Pagon RA, Adam MP, Ardinger HH, et al., eds. *GeneReviews(R)*. Seattle (WA)1993.

40. Quintans B, Ordonez-Ugalde A, Cacheiro P, Carracedo A, Sobrido MJ. Medical genomics: The intricate path from genetic variant identification to clinical interpretation. *Appl Transl Genom.* 2014;3(3):60-67.
41. Belkadi A, Bolze A, Itan Y, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A.* 2015;112(17):5473-5478.
42. Chong JX, Buckingham KJ, Jhangiani SN, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet.* 2015;97(2):199-215.
43. Harms FL, K.M. G, Hardigan AA, et al. Mutations in EBF3 disturb transcriptional profiles and underlie a novel syndrome of intellectual disability, ataxia and facial dysmorphism. *BioRxiv.* 2016.

Figure and Table Legends

Table 1: Diagnostic rates by age, sex, clinical specifics, previous genetic testing, and family structure among 371 DD/ID-affected individuals

Table 2: Diagnostic yield by exome and/or genome sequencing for 371 DD/ID-affected individuals

Table 3: Recurrent gene findings across affected 371 DD/ID-affected individuals

Table 4: Diagnostic rates across families of varying structure and phenotypic complexity

Table 5: Variants with a change in pathogenicity score due to reanalysis

Figure 1: Inheritance mechanism and molecular consequence of diagnostic variants returned to DD/ID-affected individuals. (A) The inheritance pattern of variants returned to DD/ID-affected individuals who received a likely pathogenic or pathogenic finding as a result of WES or WGS is expressed as a percentage of affected probands who received a diagnostic result. In the event that both parents were not available, inheritance could not be determined and this is represented by the term "unknown". (B) The mutation type of identified diagnostic variants is expressed as a percentage of affected individuals that received a diagnostic finding. n=94 affected individuals.

Figure 2: Intronic variants in *ALG1* and *MTOR* disrupt splicing and introduce early stop codons. (A) Diagram showing the region of *ALG1* surrounding the variant found in the proband and mother, an A>G transition 3 nucleotides downstream from the splicing donor site of intron 11. E=exon. (B) PCR F and PCR R indicate the position of the oligos used to amplify the region from patient derived cDNA. The control sample is from the RNA extracted from blood of an unrelated individual that did not harbor the variant. Control reactions lacking RT were also performed and did not show the PCR product containing the fully retained intron (data not shown). (C and D) RTqPCR analysis shows that the variant leads to inclusion of the entire intron 11. Controls are two unrelated individuals and the father of the proband. The affected individuals are the proband and mother. (E) Diagram showing the region of mTOR surrounding the variant, an A>G transition 2 nucleotides upstream of the splicing acceptor site. E=exon. (F) The region surrounding intron 4 was amplified using PCR F and PCR R (position indicated in E), and shows partial retention of the intron. The retained partial intron was not detected in control reactions lacking RT (data not shown). (G and H) RTqPCR from blood RNA shows that the 5' splice site is not affected by the variant, but that the 3' acceptor site is, leading to partial retention (134bp) of intron 4. Controls included unrelated individuals and the maternal half aunt of the proband. Affected individuals are the proband and half-sibling. For all RTqPCR analyses RNA was extracted from blood and $\Delta\Delta C_T$ values were calculated as a percent of affected individuals and normalized to GAPDH. The sequences of all oligos used are found in Table S4.

Figure 3: Cumulative fractions (y-axis) of CADD-based ranks of diagnostic variants filtered without parental data relative to *de novo* events ("DeNovo") defined with parental data. Most diagnostic variants, even under models that only consider population frequencies (e.g., "Rare"), rank among the top 25 hits in a patient, and many rank as the top hit. Restrictions to rare coding variants and/or those affecting OMIM/DDG2P genes further enrich for causal variants among top candidates, making diagnosis feasible without parents.

Supplemental Figure and Table Legends

Table S1: Oligos for quantitative PCR, PCR, and sequencing of ALG1, mTOR, and ROCK2 cDNA.

Table S2: Primary results in DD/ID-affected individuals. All returned variants identified in the study are listed. For each observed variant, we list the individual ID, sequencing method (exome/genome), gene, variant info, functional category, inheritance pattern, and pathogenicity score (VUS, likely pathogenic, pathogenic).

Table S3: Secondary findings in study participants. All secondary findings identified in the study are listed. For each observed variant, we list the individual ID, sequencing method (exome/genome), gene, variant info, functional category, Gene List (ACMG, carrier, secondary carrier, secondary diagnosis), associated phenotype, and pathogenicity score (likely pathogenic, pathogenic).

Table S4: Coverage metrics across 365 exomes and 612 genomes. Exome and genome coverage metrics across CCDS, Nimblegen exome v3 targets, 56 ACMG genes and genes included as part of GeneTests (www.genetests.org; February 2015). For exomes, n=365; for genomes, n=612.

Figure S1: The splice variant identified in mTOR leads to retention of 134 nucleotides of the 3' end of intron 4 in the mRNA transcript. Sanger sequencing confirmed the partial retention of intron 4 in cDNA synthesized from the maternal half-siblings.

Figure S2: Cumulative fractions (y-axis) of CADD-based ranks of variants of uncertain significance (VUSs) filtered without parental data relative to *de novo* events ("DeNovo") defined with parental data. VUSs are more difficult to identify without parental information, and restrictions to those variants affecting OMIM/DDG2P genes further limits the identification of these variants.

Figure S3: mRNA is produced from the variant *ROCK2* allele, but truncated protein is not detected in the proband's blood. (A) Sanger sequencing from cDNA showed that the variant allele is transcribed and produces detectable mRNA. (B) Western blots were performed using antibodies directed against the N-terminus (Sigma-Aldrich HPA007459) and C-terminus (Abcam ab56661) from protein extracted from the proband, father, and mother. β-actin was used as a control (Cell Signaling #8H10D10).

Table 1. Diagnostic rates by age, sex, clinical specifics, previous genetic testing, and family structure among the 371 DD/ID-affected individuals

CHARACTERISTIC	% Individuals (No. of individuals)	% individuals with diagnostic result[a] (No. of individuals)
AGE OF INDIVIDUAL		
2-5 years	25.8% (96)	27.1% (26/96)
6-12 years	44.5% (165)	25.4% (42/165)
13-18 years	16.5% (61)	32.8% (20/61)
19-40 years	12.7% (47)	32.0% (15/47)
>40 years	0.54% (2)	0.00% (0/2)
Average age of subject (age range)	10.56 (2 to 54 years)	
SEX OF INDIVIDUAL		
Male	57.7% (214)	24.3% (52/214)
Female	42.3% (157)	32.5% (51/157)
CLINICAL SPECIFICS		
Intellectual disability (moderate, severe) (HP:0002342, HP:0010864)	92.2% (342)	28.7% (98/342)
Speech delay (HP:0000750)	68.7% (255)	27.1% (69/255)
Seizures (HP:0001250)	45.3% (168)	30.9% (52/168)
Facial dysmorphism (HP:0001999)	30.2% (112)	29.5% (33/112)
Autism spectrum disorder (HP:000729)	25.6% (95)	18.9% (18/95)
Hypotonia (HP:0001252)	20.2% (75)	34.6% (26/75)
Positive Brain MRI	17.5% (65)	28.1% (18/64)
Macrocephaly (HP:0000256)	9.70% (36)	25.0% (9/36)
Microcephaly (HP:0000252)	9.16% (34)	47.0% (16/34)
ADHD (HP:0007018)	7.28% (27)	25.9% (7/27)
Failure to thrive (HP:0001508)	5.90% (22)	27.3% (6/22)
Short stature (HP:0004322)	4.85% (18)	44.4% (8/18)
PREVIOUS GENETIC TESTING		
Microarray	59.8% (222)	27.5% (61/222)
Single Gene/Gene Panel	38.3% (142)	30.3% (43/142)
Karyotype	29.1% (108)	36.1% (39/108)
Fragile-X	27.2% (101)	27.7% (28/101)
Mito DNA Screen	7.55% (28)	25.0% (7/28)
FAMILY STRUCTURE		
Trio	83.6% (310)	30.0% (93/310)
Duo	11.1% (41)	17.1% (7/41)
Singleton	5.39% (20)	15.0% (3/20)

MRI, magnetic resonance imaging

[a]Includes both pathogenic and likely pathogenic results

Table 2. Diagnostic yield by exome and/or genome sequencing for 371 DD/ID-affected individuals

	SNV/indel			CNV		
Assay (Affected individuals)	**Pathogenic**	**Likely pathogenic**	**VUS**	**Pathogenic**	**Likely pathogenic**	**VUS**
Exome (127)	20.4% (26)	11.0% (14)	11.8% (15)	1.6% (2)*	0% (0)	0% (0)
Genome (244)	18.0% (44)	4.1% (10)	10.7% (26)	2.0% (5)	0.4% (1)	1.2% (3)
Exome and genome (Total individuals: 371)	18.9% (70)	6.5% (24)	11.0% (41)	1.9% (7)	0.3% (1)	0.8% (3)

*Identified by microarray

Table 3. Recurrent gene findings across 371 DD/ID-affected individuals

Gene	Function	Associated clinical syndromes	No. of unrelated subjects	Mutation type	Inheritance pattern
MTOR	Protein kinase; mediates cell responses to stresses such as DNA damage and nutrient deprivation	Smith-Kingsmore	4	Missense (3); Splice	De novo (3); Unknown
SCN1A	Na channel; generation and propagation of action potential in neurons and muscle	Dravet syndrome	4	Missense; Frameshift; Nonsense; Intronic	Maternal Inherited; De novo (3)
FOXG1	Transcriptional repressor; involved in brain development	Rett Syndrome	3	Missense (2); Frameshift	De novo
MECP2	Transcriptional repressor; involved in embryonic development	Rett syndrome	3	Frameshift; Splice; Nonsense	De novo
SLC2A1	Glucose transporter in mammalian blood-brain barrier	GLUT1 deficiency	3	Splice; Missense; Frameshift	De novo
ANKRD11	Inhibits ligand-dependent activation of transcription	KGB syndrome	2	Nonsense; Frameshift	De novo
ARID1B	Cell-cycle activation; component of SWI/SNF chromatin remodeling complex	Mental retardation, AD 12	2	Nonsense	De novo
ARX	Transcriptional repressor activity; involved in CNS development	Partington syndrome; Proud syndrome	2	Frameshift; Missense	De novo; Inherited
ASXL3	Transcriptional regulator	Bainbridge-Ropers syndrome	2	Nonsense; Missense	Unknown
CHD2	Chromatin remodeling	Epileptic encephalopathy	2	Frameshift	De novo
CREBBP	Plays critical roles in embryonic development, growth control, homeostasis by coupling chromatin remodeling to TF recognition	Rubinstein-Taybi syndrome	2	Missense; Splice	De novo
DDX3X	ATP-dependent RNA helicase activity	Mental retardation	2	Nonsense	De novo
EBF3	Involved in B-cell differentiation, bone development, and neurogenesis	Epilepsy; Mental retardation	2	Splice; Missense	De novo
GRIA3	Glutamate receptor; Excitatory neurotransmitter receptor activated in normal neurophysiological processes	Mental retardation	2	Missense	Inherited
HCFC1	Control of the cell cycle and transcriptional regulation during herpes simplex virus infection	Mental retardation 3, X-linked	2	Missense	De novo; Inherited
KIF1A	Anterograde motor protein	Mental retardation	2	Missense	De novo
NBEA	Neuronal post-Golgi membrane traffic	Implicated in autism	2	Nonsense; Missense	De novo
MED13L	Transcriptional coactivator; involved in early development of heart and brain	Mental retardation	2	Frameshift; Missense	De novo
SATB2	Transcription regulation and chromatin remodeling	Glass syndrome	2	Frameshift; Nonsense	Unknown; De novo
SCN2A	Na channel; generation and propagation of action potential in neurons and muscle	Epileptic encephalopathy	2	Missense	De novo
SCN8A	Na channel; Membrane depolarization during action potentials in most electrically excitable cells	Epileptic encephalopathy	2	Missense	De novo
SYNGAP1	Ras GTPase activation protein; component of PSD associated with NMDA receptors at synapses	Mental retardation	2	Splice; Frameshift	De novo
TCF4	Transcription factor; may play a role in nervous system development	Pitt-Hopkins syndrome	2	Missense; Frameshift	Unknown; De novo
WDR45	Cell cycle progression, signal transduction, apoptosis, gene regulation	Neurodegeneration with brain iron accumulation 5	2	Missense; Splice	De novo

Table 4. Diagnostic rates across families of varying structure and phenotypic complexity

Family Structure	Trio (n=251)	Duo (n=37)	Singleton (n=13)	Total* (n=301)
Simplex (n=93)	40.0% (32/80)	30.7% (4/13)	0.00% (0/0)	38.7% (36/93)
Multiplex (n=123)	20.6% (20/97)	13.3% (2/15)	18.2% (2/11)	19.5% (24/123)
2nd & 3rd degree (n=85)	28.3% (21/74)	11.1% (1/9)	0.00% (0/2)	25.9% (22/85)

*38 families were excluded due to limited or no family history

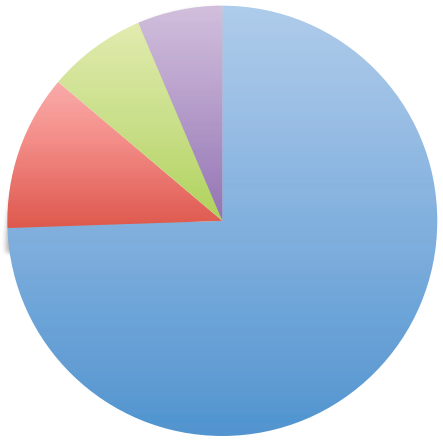
Table 5. Variants with a change in pathogenicity score due to reanalysis

Gene	Affected Individual ID(s)	Variant Info	Original Score	Updated Score	Reason(s) for Update	Additional Information
DDX3X	00075-C	NM_001356.4:c.745G>T,p.Glu249Ter	VUS	Pathogenic	Publication	[43]
EBF3	00006-C	NM_001005463.2:c.1101+1G>T	VUS	Pathogenic	GeneMatcher	Collaboration with several other groups identified patients with comparable genotypes and phenotypes
EBF3	00032-C	NM_001005463.2:c.530C>T, p.Pro177Leu	VUS	Pathogenic	GeneMatcher	Collaboration with several other groups identified patients with comparable genotypes and phenotypes
KIAA2022	00082-C	NM_001008537.2:c.2999_3000delCT,p.Ser1000Cysfs	VUS	Pathogenic	Publication/Personal Communication	[44]
TCF20	00078-C	NM_005650.3:c.5385_5386delTG,p.Cys1795Trpfs	VUS	Pathogenic	Publication	[21]
ARID2	00026-C	NM_152641.2:c.1688delT, p.Cys570Valfs	NR	Pathogenic	Publication	[45]
CDK13	00253-C	NM_003718.4:c.2525A>G, p.Asn842Ser	NR	Pathogenic	Publication	[21]
CLPB	00127-C	NM_030813.5:c.1249C>T,p.Arg417X; NM_030813.5:c.1222A>G,p.Arg408Gly	NR	Pathogenic	Publication	[46]
FGF12	00074-C	NM_004113.5:c.145G>A, p.Arg114His	NR	Pathogenic	Publication	[47]
MTOR	00040-C	NM_004958.3:c.4785G>A, p.Met1595Ile	NR	Pathogenic	Publication	For Review [48]; See also [15,49]
MTOR	00028-C, 00028-C2	NM_004958.3:c.5663T>G, p.Phe1888Cys	NR	Pathogenic	Filter	In original filter, required allele count of one; this variant was present in identical twins
HDAC8	00001-C	NM_018486.2:c.737+1G>A	NR	Likely Pathogenic	Filter	In original filter, required depth for all members of trio was set to 10 reads; father had only 7
LAMA2	00055-C, 00055-S	NM_000426.3:c.715C>T, p.Arg239Cys	NR	Likely Pathogenic	Clarification of Clinical Phenotype	Discussion with clinicians was necessary to determine that patients' phenotypes did match those observed for LAMA2
MAST1	00270-C	NM_014975.2:c.278C>T, p.Ser93Leu	NR	Likely Pathogenic	GeneMatcher	Collaboration with several other groups identified patients with comparable genotypes and phenotypes
SUV420H1	00056-C	NM_017635.3:c.2497G>T, p.Glu833X	NR	Likely Pathogenic	Publication	[21]

C, child/proband; C2, affected identical twin; S, affected sibling; NR, no returnables; VUS, variant of uncertain significance

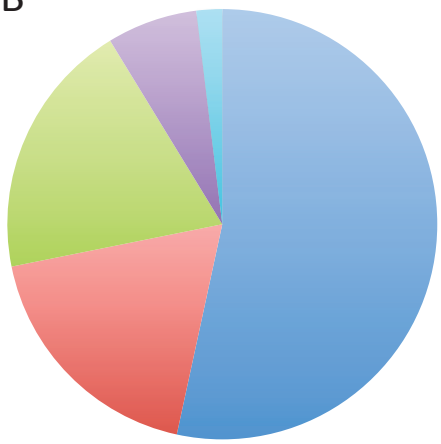
Figure 1

A



De Novo (74%)
Compound heterozygous/Recessive (12%)
X-linked Recessive (7%)
Unknown (7%)

B



Missense (53%)
Nonsense (18%)
Frameshift (20%)
Splice (7%)
Inframe Deletion (2%)

Figure 2

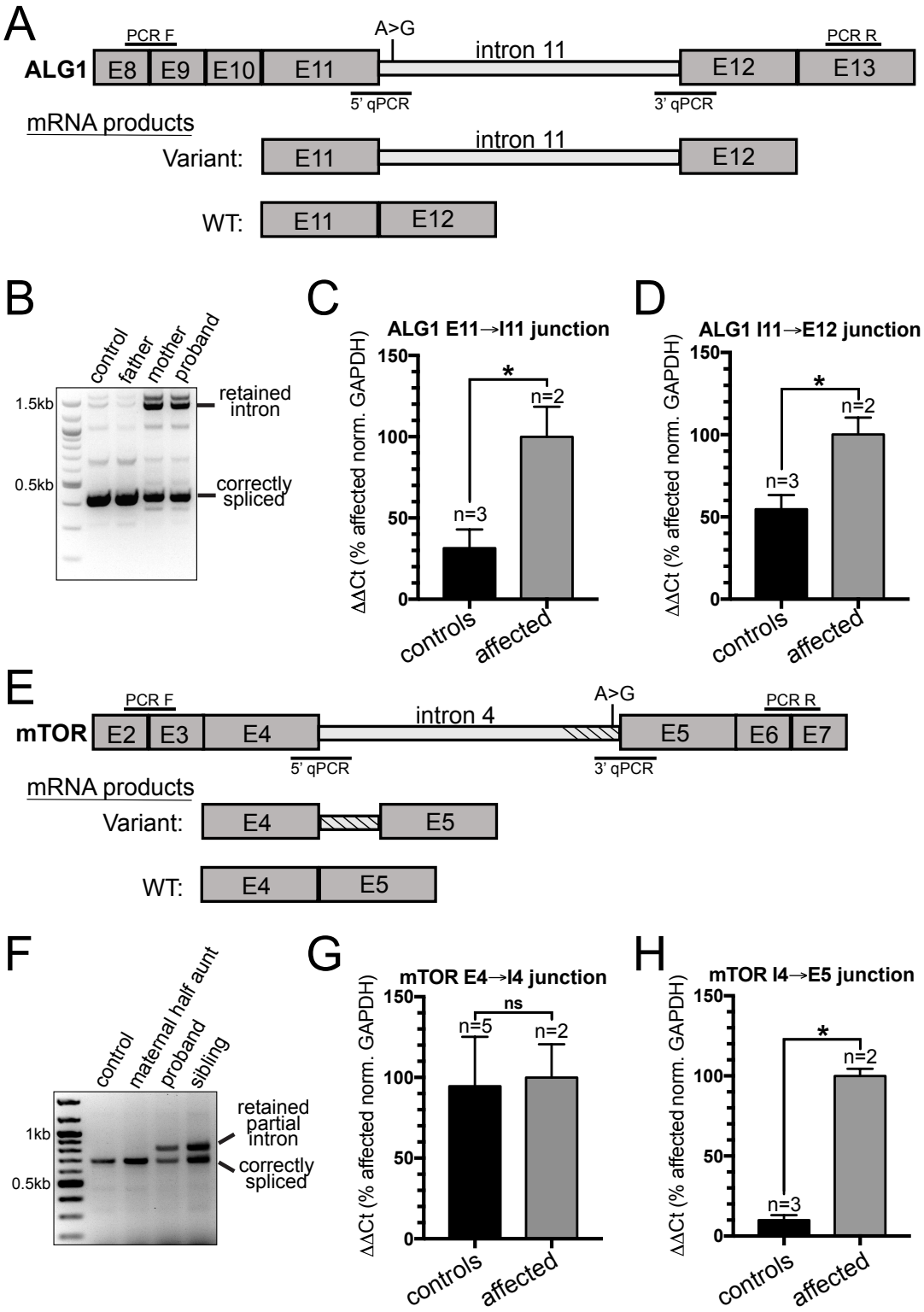
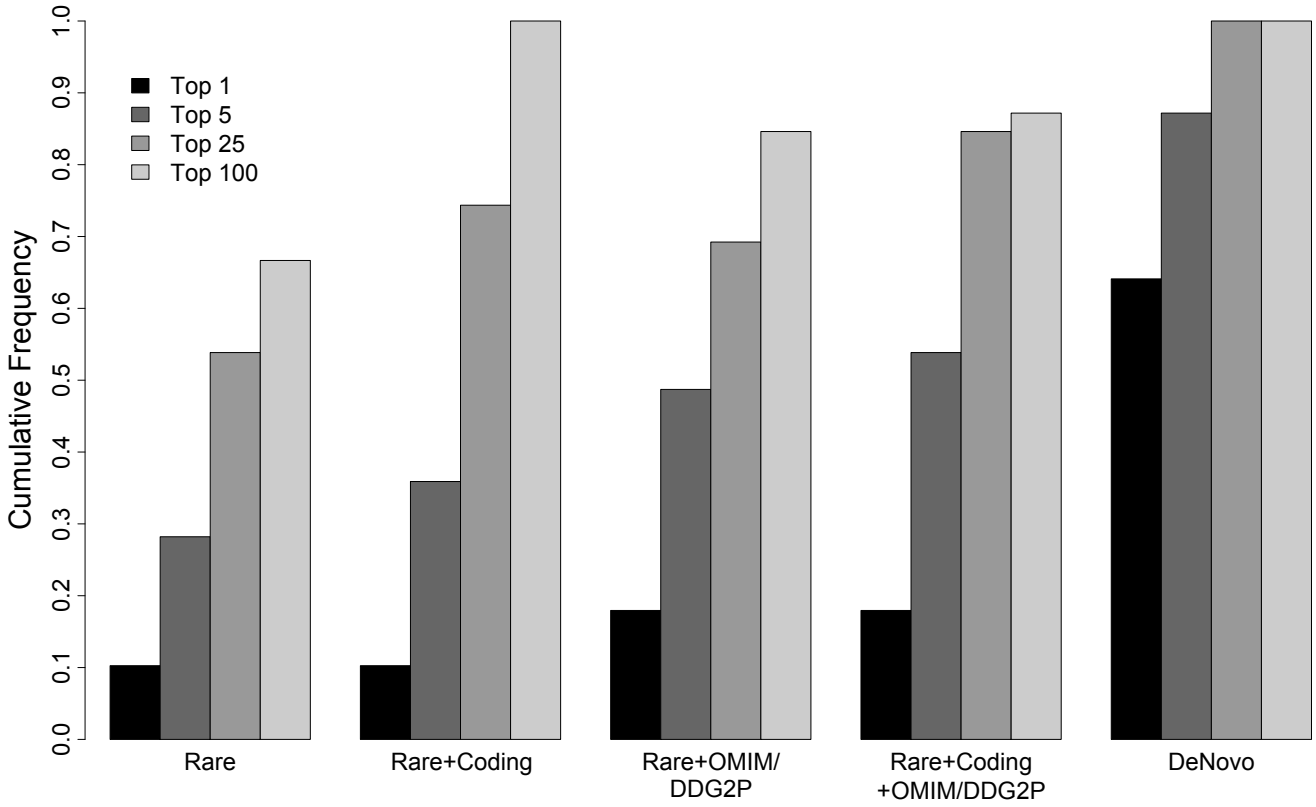

A
PCR F
A>G
PCR R
ALG1
E8
E9
E10
E11
intron 11
E12
E13
5' qPCR
3' qPCR
mRNA products
Variant:
E11
intron 11
E12
WT:
E11
E12
B
control
father
mother
proband
1.5kb
0.5kb
retained intron
correctly spliced
C
ALG1 E11→I11 junction
ΔΔCt (% affected norm. GAPDH)
150
100
50
0
*
n=2
n=3
controls
affected
D
ALG1 I11→E12 junction
ΔΔCt (% affected norm. GAPDH)
150
100
50
0
*
n=2
n=3
controls
affected
E
PCR F
A>G
PCR R
mTOR
E2
E3
E4
intron 4
E5
E6
E7
5' qPCR
3' qPCR
mRNA products
Variant:
E4
E5
WT:
E4
E5
F
control
maternal half aunt
proband
sibling
1kb
0.5kb
retained partial intron
correctly spliced
G
mTOR E4→I4 junction
ΔΔCt (% affected norm. GAPDH)
150
100
50
0
ns
n=5
n=2
controls
affected
H
mTOR I4→E5 junction
ΔΔCt (% affected norm. GAPDH)
150
100
50
0
*
n=2
n=3
controls
affected

Figure 3


Cumulative Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Top 1
Top 5
Top 25
Top 100
Rare
Rare+Coding
Rare+OMIM/
DDG2P
Rare+Coding
+OMIM/DDG2P
DeNovo