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

Running title: Clinical sequencing of DD/ID-affected children

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[1,4], Benjamin T. Weaver, BS[1,2], Amy S. Nesmith, 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
[4]University of Alabama Huntsville, Huntsville, AL, USA

*Corresponding Author: 601 Genome Way, Huntsville, AL 35806; 256-327-9490; gcooper@hudsonalpha.org

ABSTRACT

Purpose: Developmental disabilities have diverse genetic causes that must be identified to facilitate precise diagnoses. We describe genomic data from 371 affected individuals, 310 of which were sequenced as proband-parent trios.

Methods: Exomes were generated for 365 individuals (127 affected) and genomes were generated for 612 individuals (244 affected).

Results: Diagnostic variants were found in 102 individuals (27%), with variants of uncertain significance in an additional 44 (11.8%). We found that a family history of neurological disease, especially the presence of an affected 1st degree relative, reduces the diagnostic rate, reflecting both the disease relevance and 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 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.

Conclusion: Our results strongly support the value of genome sequencing as a first-choice diagnostic tool and means to continually advance clinical and research progress related to developmental disabilities, especially when coupled to rapid and free data sharing.

KEYWORDS

Developmental Delay, Intellectual Disability, *De novo,* clinical sequencing, CSER

INTRODUCTION

Developmental delay, intellectual disability, and related phenotypes (DD/ID) affect 1-2% of children and pose medical, financial, and psychological challenges[1]. While many are genetic in origin, a large fraction of cases are not diagnosed, with many families undergoing a "diagnostic odyssey" involving numerous ineffective tests over many years. A lack of diagnoses undermines counseling and medical management and slows research towards improving educational or therapeutic options.

Standard clinical genetic testing for DD/ID includes 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 over a small fraction of the genome. Whole exome or genome sequencing (WES or WGS) can provide both broad and high-resolution identification of genetic variants, and hold great promise as effective diagnostic assays[3].

As part of the Clinical Sequencing Exploratory Research (CSER) consortium[4], we have sequenced 371 individuals with one or more DD/ID-related phenotypes. One hundred and two affected individuals (27%) received a diagnosis, most from a *de novo* variant in a gene known to associate with disease. Fifteen percent of diagnoses were made after initial assessment and results return, supporting the importance of systematic reanalysis of variant data to maximize clinical effectiveness. We also describe 21 variants of uncertain significance (VUS) in 19 genes not currently associated with disease but which are intriguing candidates. The genomic data we generated and shared through dbGAP[5], ClinVar[6], and GeneMatcher[7] may prove useful to other clinical genetics labs and researchers. Our experiences strongly support the value of large-scale sequencing, especially WGS, as both an effective first-choice diagnostic tool and means to advance clinical and research progress related to pediatric neurological disease.

MATERIALS AND METHODS

IRB approval and monitoring

Review boards at Western (20130675) and the University of Alabama at Birmingham (X130201001) approved and monitored this study.

Study participant population

Participants were enrolled at North Alabama Children's Specialists in Huntsville, AL. Affected individuals were required to be at least two years of age and weigh at least 9 kilos (19.8 lbs). A parent or legal guardian was required to give consent, and assent was obtained from those children who were capable. Blood samples were sent for sequencing at the HudsonAlpha Genomic Services Laboratory. The Emory Genetics Laboratory conducted variant validation by Sanger sequencing.

Exome/genome sequencing

Genomic DNA was isolated from peripheral blood and WES (Nimblegen v3) or WGS was conducted to a mean depth of 71X or 35X, respectively, with 80% of bases covered at 20X. WES was conducted on Illumina HiSeq 2000 or 2500 machines; WGS was done on Illumina HiSeq Xs. Reads were aligned and variants called according to standard protocols[8,9].

WGS CNV calling

CNVs were called from WGS bam files using ERDS[10] and read Depth[11]. Overlapping calls with at least 90% reciprocity, less than 50% segmental duplications, and that were observed in five or fewer unaffected parents were retained. Calls were manually inspected if they were within 5 kb of a known DD/ID gene, within 5 kb of an OMIM disease gene[12], or intersecting one or more exons of any gene.

Filtering and reanalysis of exomes

Using filters related to call quality, allele frequency, and impact predictions, we searched for rare, damaging *de novo* variation, or inherited X-linked, recessive, or compound heterozygous variation in affected probands, with modifications for probands with only one or neither biological parent available for sequencing.

For reanalysis, variants were reannotated with additional data (updated ClinVar[6], ExAC[13], DDG2P[14], and several publications[15-17]) and refiltered as described above and in Supplemental Materials and Methods. Genes harboring variation but not known to associate with disease were submitted to GeneMatcher (https://genematcher.org/).

Secondary variants in parents were also reviewed, including those within ACMG genes[18], those associated with recessive disease in OMIM[12], and carrier status for *CFTR*, *HBB,* and *HEXA*.

We also searched for variants listed as pathogenic or likely pathogenic in ClinVar[6], regardless of inheritance or affected status. Further details for variant annotation and filtration are supplied in Supplemental Materials and Methods.

Analysis of trios as singletons

For probands subjected to WGS as part of trios, we removed parental genotype information from their associated VCFs and subsequently filtered for rare "*de novo*-like" (i.e., extremely rare) variants. The CADD-based ranks of variants[19] returned to each family were then calculated. See Supplemental Materials and Methods for details.

Functional Assays

RNA isolation, cDNA synthesis, RTqPCR and western blotting were conducted according to standard protocols. Details are provided in Supplemental Materials and Methods.

RESULTS

Demographics of study population

We enrolled 339 families (977 individuals total) with at least one proband with an unexplained diagnosis of a DD/ID-related phenotype. Most participating families were parent-proband trios, with a subset being either parent-proband duos or proband-only singletons. Twenty-eight enrolled families had more than one affected child, resulting in a total of 371 affected individuals. WES was performed on 365 individuals (127 affected) and 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 both experiment types. DNA from probands subjected to WES was also analyzed via a SNP array to detect copy-number variants (CNV) if clinical array testing had not been previously performed.

The study population had a mean age of 11 years and was 58% male. Affected individuals displayed symptoms described by 333 unique HPO[20] terms with over 90% of individuals displaying intellectual disability, 69% with speech delay, 45% with seizures, 18% with an abnormal brain magnetic resonance imaging (MRI) result, and 20% with microcephaly or macrocephaly. Eighty-one percent of individuals had been previously subjected to genetic testing (Table 1).

DD/ID-associated genetic variation

WES and WGS data were processed with standard protocols to produce variant lists in each family that were subsequently annotated and filtered, followed by manual review of short candidate lists (see Materials and Methods). ACMG-guided designations of pathogenicity[21] were used to classify variants according to their disease relevance. All variants described here and returned to patients were confirmed by Sanger sequencing in probands and available family members.

One hundred and two (27%) of the 371 probands had pathogenic or likely pathogenic variants, while an additional 44 (11.8%) harbored a variant of uncertain significance (VUS, Table 2). We hereafter refer to pathogenic or likely pathogenic variants, but not VUSs, as “diagnostic”. Given that most probands had been previously tested via microarray prior to their enrollment in this study, diagnostic or VUS large CNVs were detected in only 11 affected individuals (Table 2).

Most (74%) diagnostic variants occurred *de novo*, while 12% of individuals inherited diagnostic variants as compound heterozygotes or homozygotes (Figure S1A). An additional 7% were males with an X-linked maternally inherited diagnosis. Finally, 7% of diagnosed participants were sequenced with one or no parent and thus have ambiguous inheritance (Figure S1A). Most diagnostic variants were missense mutations (53%), while 38% were nonsense or frameshift, 7% disrupted splicing, and 2% led to inframe deletion (Figure S1B). Diagnostic variants or VUSs were identified in 99 genes, excluding large CNVs, with variants in 24 (24%) of these genes observed in two or more unrelated individuals (Tables S1 and S2). *SCN1A* (Dravet syndrome MIM:607208) and *MTOR* (Smith-Kingsmore syndrome MIM:601231) were the most frequent genes identified, each affecting four unrelated families.

Diagnostic rates across families of varying structure and phenotypic complexity

Affected individuals were categorized as one of three structures based on the number of parents that were sequenced along with the proband(s): proband-parent trios (310); duos with one parent (41); and proband-only singletons (20). A diagnostic result was found in 30% of trio individuals, 17% of duo individuals, and 15% of singletons (Table 1).

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 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 (n=123) (Table S3). Thirty-eight probands with limited or no family history information were excluded from this analysis.

Diagnoses were found in 24 (20%) of the 123 multiplex families (20 out of 97 trios), in contrast with 36 (39%) of 93 simplex families (32 out of 80 trios), suggesting a diagnostic rate that is twice as high for 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=0.002) or only those sequenced as trios (p=0.008) 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% of trios). Of relevance to the trio/duo/singleton comparison described above, 11 of 13 (85%) singletons for which we had family history information had an affected 1st degree relative, in contrast with 41% for duos and 39% for trios (Table S3). This enrichment for affected 1st degree relatives likely contributed to the generally reduced diagnostic rate in singletons observed here.

Multiplex family diagnoses include examples of both expected and unexpected inheritance patterns. For example, two affected male siblings were found to be hemizygous for a nonsense mutation in *PHF6* (Börjeson-Forssman-Lehmann syndrome MIM:301900) inherited from their unaffected mother. In another family, we found the proband to be compound heterozygous for two variants in *GRIK4*, with one allele inherited from each parent. Interestingly, both the mother and father of this proband report psychiatric illness, and extended family history of psychiatric phenotypes is notable. We also observed independent *de novo* causal variants within two families. Affected siblings in family 00135 each harbored a returnable *de novo* variant in a different gene, one in *SPR* (Dystonia MIM:612716) and one in *RIT1* (Noonan syndrome MIM:615355), while two probands (00075 and 00078) who were second degree relatives to one another harbored independent *de novo* variants, one each in *DDX3X* (X-linked ID MIM:300958) and *TCF20* (Table S2).

Alternative mechanisms of disease

While the majority of DD/ID-associated genetic variation detected in our study has been missense, frameshift, or nonsense mutations (Figure S1B), 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 ID, seizures, speech delay, autism and stereotypic behaviors. WES revealed an SNV near the splice acceptor 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 with several features of Rett syndrome, but who lacked deceleration of head growth and exhibited typical growth development[22]. Laccone, et al. showed by RTqPCR that the variant produces a cryptic splice acceptor site that adds five nucleotides to the mRNA resulting in a frameshift (R9fs24X)[22]. It is likely that both the canonical and cryptic splice sites function, allowing for most *MECP2* transcripts to produce full-length protein, resulting in the milder Rett phenotype observed in the individual described here and the girl described by Laccone et al.

In another affected proband, we identified compound heterozygous variants in *ALG1* (Table S2). This proband has phenotypes consistent with ALG1-CDG (congenital disorder of glycosylation MIM:608540) including severe ID, hypotonia, growth retardation, microcephaly, and seizures[23]. The paternally inherited missense mutation (c.773C>T, S258L) was known to be pathogenic[24], while the maternally inherited variant, previously unreported (c.1187+3A>G), is three bases downstream of an exon/intron junction (Figure 1A). We performed RTqPCR from patient blood RNA and found that intron 11 of *ALG1* is completely retained in both the proband and the mother (Figure 1A-D). The retention of intron 11 results in a stop-gain after adding 84 nucleotides (28 codons).

In a separate family consisting of affected maternal half siblings (00218-C and 00218-S, Table S2, Figure 1E) we observed variation in a canonical splice acceptor site (c.505-2A>G) of *MTOR* intron 4. The half siblings described here both have ID; the younger sibling has no seizures but has facial dysmorphism, speech delay, and autism, while his older sister exhibits seizures. We presume that the maternal half siblings inherited the splice variant from their mother, for whom DNA was not available, who was reported to exhibit seizures. We conducted RTqPCR and Sanger sequencing using blood-derived RNA from both siblings, finding transcripts that included an additional 134 nucleotides from the 3' end of intron 4, ultimately leading to the addition of 20 amino acids before a stop-gain (Figure 1F-H, Figure S2). Because the stop-gain occurs early in protein translation, this splice variant likely leads to *MTOR* loss-of-function. Mutations in *MTOR* associate with a broad spectrum of phenotypes including epilepsy, hemimegalencephaly, and intellectual disability[25]. However, previously reported pathogenic variants in *MTOR* are all missense and suspected to result in gain-of-function[26]. Owing to this mechanistic uncertainty, we have classified this splice 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 disease. *MTOR* is highly intolerant of mutations in the general population (RVIS score of 0.09%) supporting the hypothesis that loss-of-function is deleterious and likely leads 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[27]. However, we also assessed 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[19] ranks of *de novo* variants previously found to be diagnostic (Figure 2). While parentally informed filters were the most effective (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[12] or associated with DD/ID via DECIPHER[14], 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 that VUSs would have been more difficult to identify without parental sequencing (Figure S3), owing to the fact that many VUSs do not affect genes known to associate with disease. Also, those VUSs that do affect genes known to associate with disease tended to have lesser computationally estimated effects, and therefore lower CADD ranks[19]; if they were more overtly deleterious, they would likely have been found 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 participating parents

We found genetic variation unrelated to DD/ID, i.e., secondary findings, in 8% of parents. One and a half percent of parents were given a secondary diagnosis, such as variants in *SLC22A5* that explain one parent's self-reported primary carnitine deficiency (MIM:212140). We also examined the 56 genes named by the ACMG as potentially harboring actionable secondary findings[18], revealing pathogenic/likely pathogenic variants in 13 parents (2.1%), a rate similar to that observed in other cohorts[18,28]. Finally, we performed a limited carrier screening assessment, identifying 27 (4.5%) parents as carriers of pathogenic/likely pathogenic variation in *HBB* (Sickle cell anemia MIM:603903), *HEXA* (Tay-Sachs disease MIM:272800), or *CFTR* (Cystic fibrosis MIM:219700). We also assessed parents as mate pairs and searched for genes in which both are heterozygous for a pathogenic/likely pathogenic recessive allele, resulting in one parental pair (among 285 total pairs) identified as carriers for variants in *ATP7B*, associated with Wilson disease (MIM:277900).

Reanalysis of WES and WGS data

To exploit steady increases in human genetic knowledge, we performed systematic reanalyses of WES/WGS data. We approached reanalysis in three ways: 1) systematic reanalysis of old data, with the goal of reassessing each dataset every 12 months after initial analysis; 2) continual mining of all variant data based on new DD/ID genetic publications; and 3) use of GeneMatcher[7] to aid in the interpretation of variants in genes of uncertain disease significance.

As shown in Table 3, these combined efforts led to a change in pathogenicity score for 15 variants in 17 individuals. In nine cases, 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 changes were a result of discussions facilitated by GeneMatcher[7], while the remaining upgrades resulted from reductions in filter stringency (changes to read depth and batch allele frequency) or clarification of the clinical phenotype. Among all 46 VUSs thus far identified, five (10.8%) have been upgraded to diagnostic. The most rapid change affected a *de novo* variant in *DDX3X*, which was upgraded from VUS to pathogenic 1 month after initial assessment, while a *de novo* disruption of *EBF3* was upgraded from VUS to pathogenic 2.5 years after initial assessment. These data indicate that VUSs, especially when identified via parent-proband trio sequencing, have considerable diagnostic potential. Additionally, of the 211 families who originally received a negative result, diagnostic variation was identified for 10 (4.7%) through reanalysis. These data show that regular reanalysis of both uncertain and negative results is an effective mechanism to improve diagnostic yield.

Identification of novel candidate genes

We have identified 21 variants within 19 genes with no known disease association but which are interesting candidates. For example, in one proband we identified an early nonsense variant (c.2140C>T, R714X, CADD score 44) in *ROCK2*, with reduction of ROCK2 protein confirmed by western blot (Figure S4). *ROCK2* is a conserved Rho-associated serine/threonine kinase involved in a number of 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[29]. As a second example, in two unrelated probands, we identified *de novo* variation in *NBEA*, a nonsense variant at codon 2213 (of 2946, c.6637C>T, R2213X, CADD score 52), and a missense at codon 946 (c.2836C>T, H946Y, CADD score 25.6). *NBEA* is a kinase anchoring protein with roles in the recruitment of cAMP dependent protein kinase A to endomembranes near the trans-Golgi network[30]. The RVIS score of *NBEA* is 0.75%. While these variants remain VUSs, the fact that they are *de novo*, predicted to be deleterious, and affect genes under strong selective conservation in human populations, suggests they have a good chance to be disease-associated.

DISCUSSION

We have sequenced 371 individuals with various DD/ID-related phenotypes. Twenty-seven percent of these individuals received a molecular diagnosis, mostly as a result of *de novo* protein-altering variants. We found that the diagnostic yield is impacted by presence of disease in family members, as our success rate drops from 39% for probands without any affected relatives to 20% for probands with one or more affected 1st relatives. These data are consistent with the observation of higher causal variant yields in simplex families relative to multiplex families affected with autism[31]. It 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)[32].

One hundred and twenty-seven probands were subject to WES and 244 were subject to WGS. The diagnostic rate was not significantly different between the two assays when considering only SNVs or small indels (p=0.30). However, WGS is a better assay for detection of CNVs[33] and, while our patient population is depleted for large causal CNVs owing to prior array or karyotype testing, we have identified diagnostic CNVs in eight individuals.

We have also demonstrated the value of systematic reanalysis, which has thus far yielded diagnoses for an additional 17 individuals (16.7% of total diagnoses, 4.6% of total probands). Given the rates of progress in Mendelian disease genetics[34] and the development of new genomic annotations, we believe that systematic reanalysis of genomic data should become standard practice. While non-trivial, reanalysis requires relatively modest investments of time and cost, especially in proportion to the initial sequencing and analysis. Furthermore, 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 S5) and previous studies[33], and re-analysis of pathogenic non-coding variation is impossible with 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. Our analyses shows that proband-only sequencing can lead to effective diagnosis, particularly via combinations of disease gene databases like OMIM[12] and variant annotations like CADD[19], with modest increases in the number of variants that must be manually curated, relative to that achieved from trio sequencing. However, VUSs are more difficult to identify without parental sequence data, and 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 at least one new disease association, as we identified two patients that harbor *de novo* variants in *EBF3*, 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[7], we were able to identify a total of 10 DD/ID-affected individuals who harbor *EBF3* variants, supporting that *de novo* disruption of *EBF3* function leads to neurodevelopmental phenotypes[35]. It is our hope that the other VUSs described here, shared via ClinVar[6] and GeneMatcher[7], will also help to facilitate new associations.

We have demonstrated the benefits of genomic sequencing to identify diagnostic variation in children with developmental disabilities who are otherwise lacking a precise diagnosis. Indeed, by combining genomic breadth with resolution capable of detecting SNVs, indels, and CNVs in a single assay, WGS is a highly effective choice as the first diagnostic test, rather than last resort, for unexplained developmental disabilities. The ability for WGS to serve as a single-assay replacement for WES and microarrays underscores its value as a frontline test. Furthermore, the benefits and effectiveness of WGS testing is likely to grow over time both by accelerating research (for example into the discovery of smaller pathogenic CNVs and pathogenic SNVs outside of coding exons), and by facilitating more effective reanalysis, a process which we show to be an essential component to maximize diagnostic yield.

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 acquisition and analysis. We would also like to thank Dr. Jeremy Herskowitz in the Department of Neurology at University of Alabama at Birmingham for discussions about ROCK2. This work was supported by grants from the US National Human Genome Research Institute (NHGRI; UM1HG007301) and the National Cancer Institute (NCI; R01CA197139).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

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. 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.
11. 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.
12. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 2005;33(Database issue):D514-517.
13. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536(7616):285-291.
14. McRae J, Clayton S, Fitzgerald T, et al. Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. *BioRxiv.* 2016.
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. 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.
18. 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.

19. 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.
20. Kohler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* 2014;42(Database issue):D966-974.
21. 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.
22. 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.
23. 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.
24. 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.
25. Baulac S. mTOR signaling pathway genes in focal epilepsies. *Prog Brain Res.* 2016;226:61-79.
26. Moller R, Wechkuysen S, Chipaux M, et al. Germline and somatic mutations in MTOR gene in focal cortical dysplasia and epilepsy. *Neurology: Genetics.* In press;2(6):e118.
27. Vissers LE, de Ligt J, Gilissen C, et al. A de novo paradigm for mental retardation. *Nat Genet.* 2010;42(12):1109-1112.
28. 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.
29. Loirand G. Rho Kinases in Health and Disease: From Basic Science to Translational Research. *Pharmacol Rev.* 2015;67(4):1074-1095.
30. 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.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. Deciphering Developmental Disorders S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature.* 2015;519(7542):223-228.

37. Farach LS, Northrup H. KIAA2022 nonsense mutation in a symptomatic female. *Am J Med Genet A.* 2016;170(3):703-706.
38. Shang L, Cho MT, Retterer K, et al. Mutations in ARID2 are associated with intellectual disabilities. *Neurogenetics.* 2015;16(4):307-314.
39. Wortmann SB, Zietkiewicz S, Kousi M, et al. CLPB mutations cause 3-methylglutaconic aciduria, progressive brain atrophy, intellectual disability, congenital neutropenia, cataracts, movement disorder. *Am J Hum Genet.* 2015;96(2):245-257.
40. Siekierska A, Isrie M, Liu Y, et al. Gain-of-function FHF1 mutation causes early-onset epileptic encephalopathy with cerebellar atrophy. *Neurology.* 2016;86(23):2162-2170.

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: Variants with a change in pathogenicity score due to reanalysis.

Figure 1: 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 S6.

Figure 2: 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: Recurrent gene findings across affected 371 DD/ID-affected individuals.

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: Diagnostic rates across families of varying structure and phenotypic complexity.

Table S4: 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 S5: 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.

Table S6: Oligos for quantitative PCR, PCR, and sequencing of ALG1, mTOR, and ROCK2 cDNA.

Figure S1: 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 S2: 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 S3: 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 S4: 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).

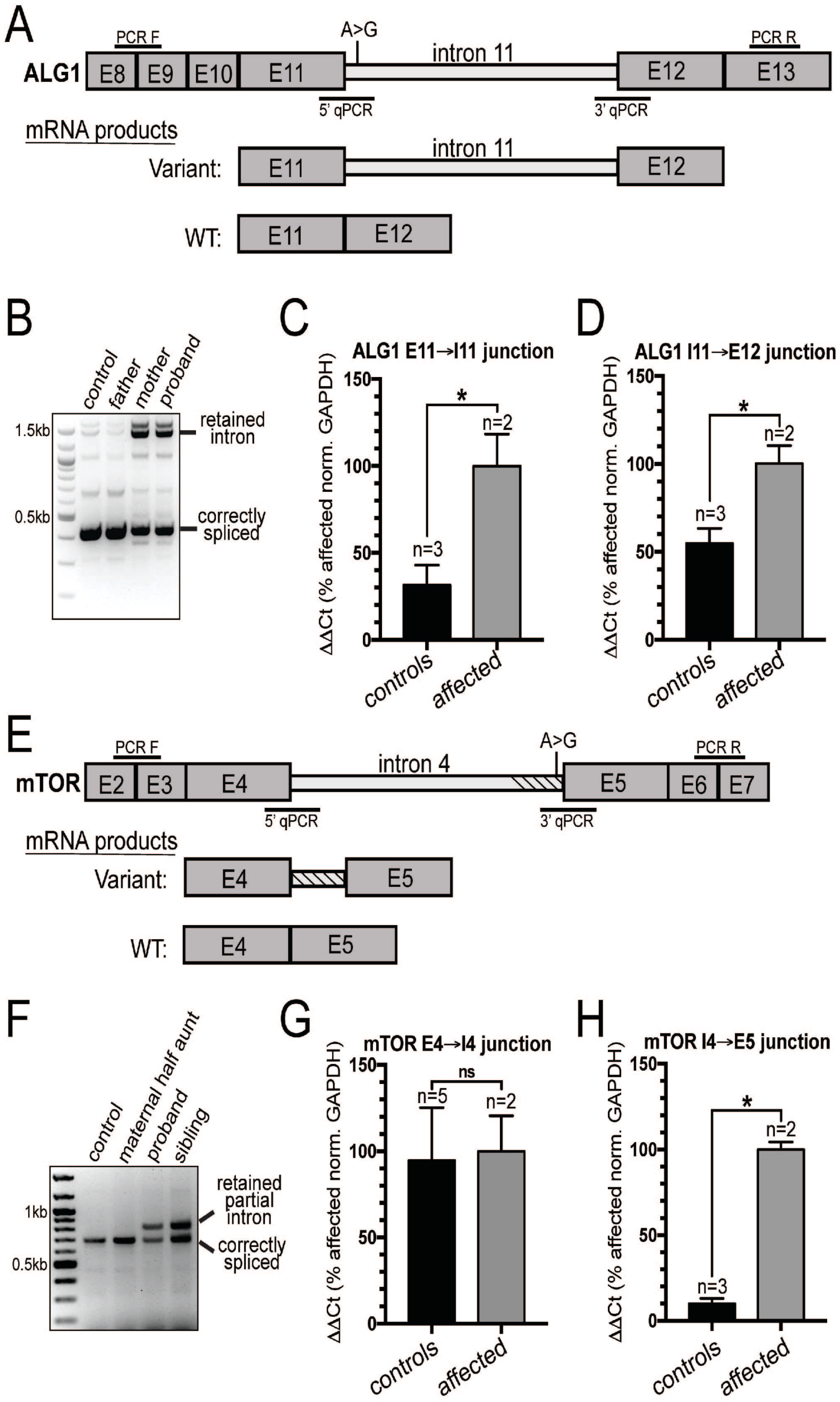
A
PCR F
A>G
PCR R
ALG1
E8
E9
E10
E11
intron 11
E12
E13
5' qPCR
3' qPCR
mRNA products
Variant:
E11
intron 11
E12
WT:
E11
E12
B
control
father
mother
proband
1.5kb
0.5kb
retained intron
correctly spliced
C
ALG1 E11→I11 junction
ΔΔCt (% affected norm. GAPDH)
150
100
50
0
*
n=3
n=2
controls
affected
D
ALG1 I11→E12 junction
ΔΔCt (% affected norm. GAPDH)
150
100
50
0
*
n=3
n=2
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=3
n=2
controls
affected

Figure 1

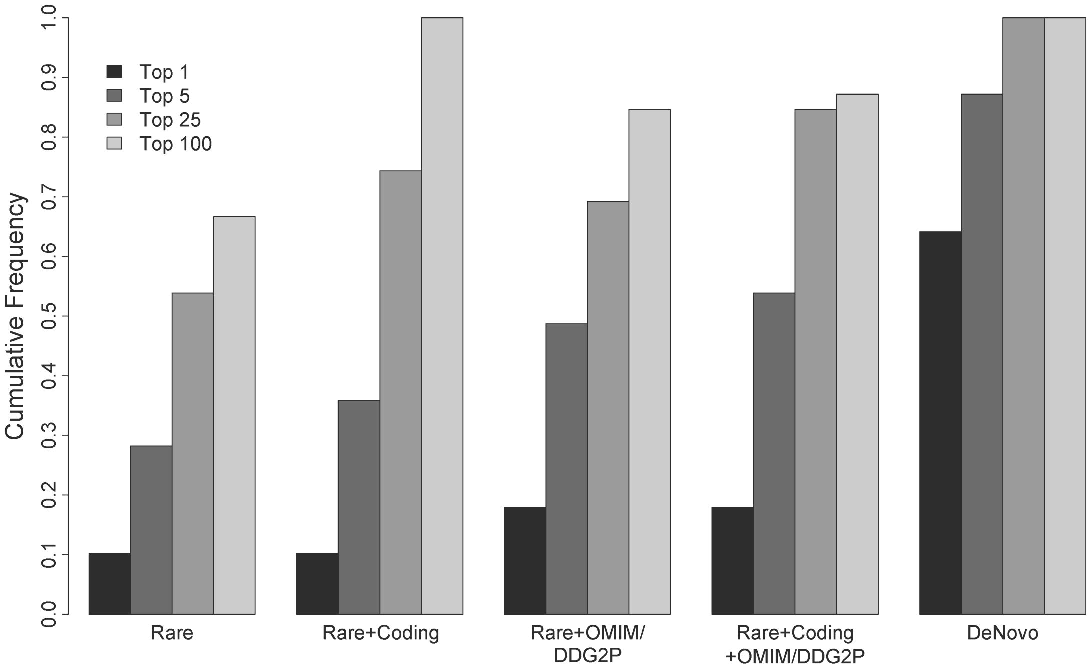
Cumulative Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
■ Top 1
■ Top 5
■ Top 25
□ Top 100
Rare
Rare+Coding
Rare+OMIM/
DDG2P
Rare+Coding
+OMIM/DDG2P
DeNovo

Figure 2

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)

Table 1

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 2

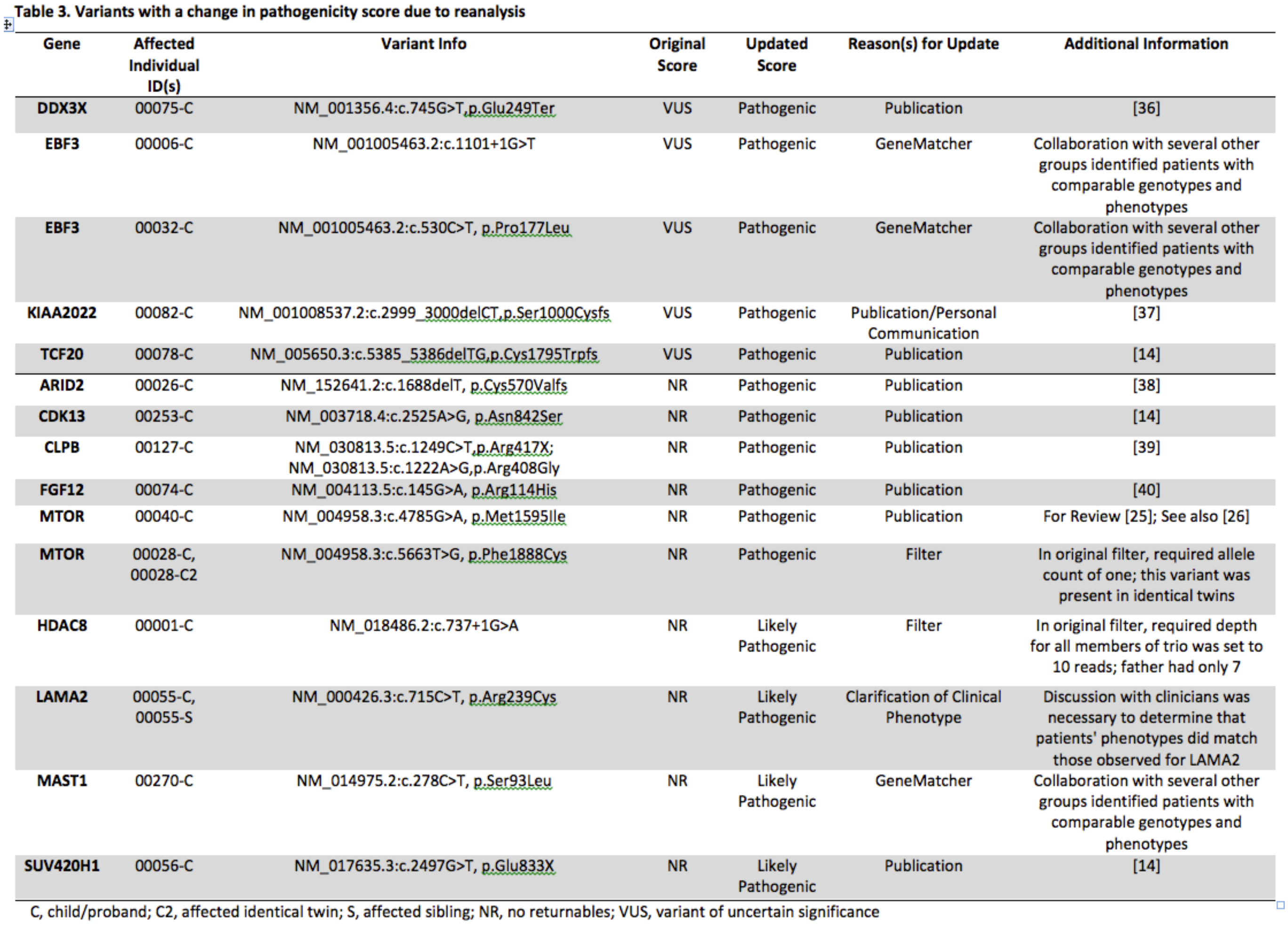

Table 3. Variants with a change in pathogenicity score due to reanalysis

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	[36]
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	[37]
TCF20	00078-C	NM_005650.3:c.5385_5386delTG,p.Cys1795Trpfs	VUS	Pathogenic	Publication	[14]
ARID2	00026-C	NM_152641.2:c.1688delT, p.Cys570Valfs	NR	Pathogenic	Publication	[38]
CDK13	00253-C	NM_003718.4:c.2525A>G, p.Asn842Ser	NR	Pathogenic	Publication	[14]
CLPB	00127-C	NM_030813.5:c.1249C>T,p.Arg417X; NM_030813.5:c.1222A>G,p.Arg408Gly	NR	Pathogenic	Publication	[39]
FGF12	00074-C	NM_004113.5:c.145G>A, p.Arg114His	NR	Pathogenic	Publication	[40]
MTOR	00040-C	NM_004958.3:c.4785G>A, p.Met1595Ile	NR	Pathogenic	Publication	For Review [25]; See also [26]
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	[14]

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

Table 3