Published in final edited form as:

Genet Med. 2020 July ; 22(7): 1269–1275. doi:10.1038/s41436-020-0781-x.

Alternative transcripts in variant interpretation: the potential for missed diagnoses and misdiagnoses

Kelly Schoch, MS, CGC[1], **Queenie K.-G. Tan, MD, PhD**[1], **Nicholas Stong, PhD**[2], **Kristen L. Deak, PhD**[3], **Allyn McConkie-Rosell, PhD, CGC**[1], **Marie T. McDonald, MD**[1], **Undiagnosed Diseases Network**, **David B. Goldstein, PhD**[2], **Yong-hui Jiang, MD, PhD**[1], **Vandana Shashi, MD**[1]

[1]Division of Medical Genetics, Department of Pediatrics, Duke University Medical Center

[2]Institute of Genomic Medicine, Columbia University, New York, N.Y.

[3]Department of Pathology, Duke University Medical Center

Corresponding author: Vandana Shashi, Professor, Division of Medical Genetics, Department of Pediatrics, Duke University Medical Center, Vandana.shashi@duke.edu, (919) 681-2772.

Undiagnosed Diseases Network Member list

Maria T. Acosta, Margaret Adam, David R. Adams, Pankaj B. Agrawal, Mercedes E. Alejandro, Justin Alvey, Laura Amendola, Ashley Andrews, Euan A. Ashley, Mahshid S. Azamian, Carlos A. Bacino, Guney Bademci, Eva Baker, Ashok Balasubramanyam, Dustin Baldridge, Jim Bale, Michael Bamshad, Deborah Barbouth, Gabriel F. Batzli, Pinar Bayrak-Toydemir, Anita Beck, Alan H. Beggs, Edward Behrens, Gill Bejerano, Jimmy Bennet, Beverly Berg-Rood, Raphael Bernier, Jonathan A. Bernstein, Gerard T. Berry, Anna Bican, Stephanie Bivona, Elizabeth Blue, John Bohnsack, Carsten Bonnenmann, Devon Bonner, Lorenzo Botto, Brenna Boyd, Lauren C. Briere, Elly Brokamp, Gabrielle Brown, Elizabeth A. Burke, Lindsay C. Burrage, Manish J. Butte, Peter Byers, William E. Byrd, John Carey, Olveen Carrasquillo, Ta Chen Peter Chang, Sirisak Chanprasert, Hsiao-Tuan Chao, Gary D. Clark, Terra R. Coakley, Laurel A. Cobban, Joy D. Cogan, F. Sessions Cole, Heather A. Colley, Cynthia M. Cooper, Heidi Cope, William J. Craigen, Andrew B. Crouse, Michael Cunningham, Precilla D’Souza, Hongzheng Dai, Surendra Dasari, Mariska Davids, Jyoti G. Dayal, Matthew Deardorff, Esteban C. Dell’Angelica, Shweta U. Dhar, Katrina Dipple, Daniel Doherty, Naghmeh Dorrani, Emilie D. Douine, David D. Draper, Laura Duncan, Dawn Earl, David J. Eckstein, Lisa T. Emrick, Christine M. Eng, Cecilia Esteves, Tyra Estwick, Marni Falk, Liliana Fernandez, Carlos Ferreira, Elizabeth L. Fieg, Paul G. Fisher, Brent L. Fogel, Irman Forghani, Laure Fresard, William A. Gahl, Ian Glass, Rena A. Godfrey, Katie Golden-Grant, Alica M. Goldman, David B. Goldstein, Alana Grajewski, Catherine A. Groden, Andrea L. Gropman, Irma Gutierrez, Sihoun Hahn, Rizwan Hamid, Neil A. Hanchard, Kelly Hassey, Nichole Hayes, Frances High, Anne Hing, Fuki M. Hisama, Ingrid A. Holm, Jason Hom, Martha Horike-Pyne, Alden Huang, Yong Huang, Rosario Isasi, Fariha Jamal, Gail P. Jarvik, Jeffrey Jarvik, Suman Jayadev, Jean M. Johnston, Lefkothea Karaviti, Emily G. Kelley, Jennifer Kennedy, Dana Kiley, Isaac S. Kohane, Jennefer N. Kohler, Deborah Krakow, Donna M. Krasnewich, Elijah Kravets, Susan Korrick, Mary Koziura, Joel B. Krier, Seema R. Lalani, Byron Lam, Christina Lam, Brendan C. Lanpher, Ian R. Lanza, C. Christopher Lau, Kimberly LeBlanc, Brendan H. Lee, Hane Lee, Roy Levitt, Richard A. Lewis, Sharyn A. Lincoln, Pengfei Liu, Xue Zhong Liu, Nicola Longo, Sandra K. Loo, Joseph Loscalzo, Richard L. Maas, Ellen F. Macnamara, Calum A. MacRae, Valerie V. Maduro, Marta M. Majcherska, May Christine V. Malicdan, Laura A. Mamounas, Teri A. Manolio, Rong Mao, Kenneth Maravilla, Thomas C. Markello, Ronit Marom, Gabor Marth, Beth A. Martin, Martin G. Martin, Julian A. Martínez-Agosto, Shruti Marwaha, Jacob McCauley, Allyn McConkie-Rosell, Colleen E. McCormack, Alexa T. McCray, Elisabeth McGee, Heather Mefford, J. Lawrence Merritt, Matthew Might, Ghayda Mirzaa, Eva Morava-Kozicz, Paolo M. Moretti, Marie Morimoto, John J. Mulvihill, David R. Murdock, Mariko Nakano-Okuno, Avi Nath, Stan F. Nelson, John H. Newman, Sarah K. Nicholas, Deborah Nickerson, Donna Novacic, Devin Oglesbee, James P. Orengo, Laura Pace, Stephen Pak, J. Carl Pallais, Christina GS. Palmer, Jeanette C. Papp, Neil H. Parker, John A. Phillips III, Jennifer E. Posey, Lorraine Potocki, Barbara N. Pusey, Aaron Quinlan, Wendy Raskind, Archana N. Raja, Genecee Renteria, Chloe M. Reuter, Lynette Rives, Amy K. Robertson, Lance H. Rodan, Jill A. Rosenfeld, Natalie Rosenwasser, Robb K. Rowley, Maura Ruzhnikov, Ralph Sacco, Jacinda B. Sampson, Susan L. Samson, Mario Saporta, C. Ron Scott, Judy Schaechter, Timothy Schedl, Kelly Schoch, Daryl A. Scott, Prashant Sharma, Vandana Shashi, Jimann Shin, Rebecca Signer, Catherine H. Sillari, Edwin K. Silverman, Janet S. Sinsheimer, Kathy Sisco, Edward C. Smith, Kevin S. Smith, Emily Solem, Lilianna Solnica-Krezel, Rebecca C. Spillmann, Joan M. Stoler, Nicholas Stong, Jennifer A. Sullivan, Kathleen Sullivan, Angela Sun, Shirley Sutton, David A. Sweetser, Virginia Sybert, Holly K. Tabor, Cecelia P. Tamburro, Queenie K.-G. Tan, Mustafa Tekin, Fred Telischi, Willa Thorson, Cynthia J. Tifft, Camilo Toro, Alyssa A. Tran, Brianna M. Tucker, Tiina K. Urv, Adeline Vanderver, Matt Velinder, Dave Viskochil, Tiphanie P. Vogel, Colleen E. Wahl, Stephanie Wallace, Nicole M. Walley, Chris A. Walsh, Melissa Walker, Jennifer Wambach, Jijun Wan, Lee-kai Wang, Michael F. Wangler, Patricia A. Ward, Daniel Wegner, Mark Wener, Tara Wenger, Katherine Wesseling Perry, Monte Westerfield, Matthew T. Wheeler, Jordan Whitlock, Lynne A. Wolfe, Jeremy D. Woods, Shinya Yamamoto, John Yang, Guoyun Yu, Diane B. Zastrow, Chunli Zhao, Stephan Zuchner

There are no conflicts of interest.

Abstract

Purpose—Guidelines by professional organizations for assessing variant pathogenicity include the recommendation to utilize biologically relevant transcripts, however there is variability in transcript selection by laboratories.

Methods—We describe three patients whose genomic results were incorrect, because alternative transcripts and tissue expression patterns were not considered by the commercial laboratories.

Results—In individual 1, a pathogenic coding variant in a brain-expressed isoform of *CKDL5* was missed twice on sequencing, because the variant was intronic in the transcripts considered in analysis. In individual 2, a microdeletion affecting *KMT2C* was not reported on microarray, since deletions of proximal exons in this gene are seen in healthy individuals; however this individual had a more distal deletion involving the brain-expressed *KMT2C* isoform, giving her a diagnosis of Kleefstra syndrome. Individual 3 was reported to have a pathogenic variant in exon 10 of *OFD1* on exome, but had no typical features of the *OFD1*-related disorders. Since exon 10 is spliced from the more biologically relevant transcripts of *OFD1*, it was determined that he did not have an *OFD1* disorder.

Conclusion—These examples illustrate the importance of considering alternative transcripts as a potential confounder when genetic results are negative or discordant with the phenotype.

Keywords

alternative splicing; exome/genome sequencing; chromosomal microarray; transcript expression; isoforms

Introduction

Genomic sequencing (exome/genome/targeted gene sequencing) and allied techniques such as chromosomal microarray (CMA) are widely used for the diagnosis of genetic diseases and while these have revolutionized genomic medicine, determining variant pathogenicity remains a challenge in diagnostic decision-making. The ACMG/AMP standards and guidelines for the interpretation of sequence variants have stated that in addition to the reference transcript (i.e. canonical), alternate clinically relevant transcripts (e.g with additional exons, or expressed in a tissue of interest) should be evaluated in assessing variant pathogenicity[1]. For analysis of copy number variants (CNVs) on CMA alternate transcripts are less likely to be a confounding factor, but the *interpretation* of CNVs can be influenced by tissue specific transcript expression, a fact that is seldom discussed in the existing literature.

The importance of alternate transcripts in interpreting genetic testing is underscored by the finding that ~95% of multiexon genes undergo alternative splicing, with an average of seven transcripts per gene; furthermore, these can be differentially expressed across tissues and developmental timespans[2,3]. Interpretation of genomic variation may thus differ according to transcript selection and tissue expression[4]. Several transcript databases are available for variant annotation, including GENCODE (https://www.gencodegenes.org/), RefSeq (https://www.ncbi.nlm.nih.gov/refseq/), Ensembl ((https://www.ensembl.org); and Consensus

Coding Sequencing (CCDS; https://www.ncbi.nlm.nih.gov/projects/CCDS). Each laboratory selects its own reference transcript and as a result, there is great variability in the transcripts utilized by sequencing laboratories.[5].

Highlighting the impact of this variability in transcript selection, an annotation comparison of 80 million genome sequencing (GS) variants using two different transcript sets (RefSeq and Ensembl) produced agreement in annotation in only 44% of putative loss of function (LoF) variants[6]. Another study demonstrated that for 292 genes included on three neonatal epilepsy panels, only one transcript was considered by the commercial laboratories for 96% of the genes, although 30% of these genes had alternative coding regions expressed in fetal/neonatal brain tissue[5]. Four missed pathogenic variants were found when variants were reannotated as LoF in alternate transcripts[5]. The opposite could also occur, with a putative pathogenic variant being reannotated as not disease associated, in the context of alternative isoforms. Thus, errors in considering alternative transcripts can result in both missed and incorrect diagnoses.

Here we present three individuals, in whom genomic results were either negative or discordant with the clinical phenotype, and subsequent evaluation of alternate transcripts and their expression in the tissues of interest provided diagnostic clarity.

Patients and Methods

All evaluations were performed as part of the NIH Undiagnosed Diseases Network (UDN) (https://undiagnosed.hms.harvard.edu/) under an IRB-approved protocol (NHGRI 15-HG-0130), and informed consent was obtained from all subjects. Further consent for photographs to be used in a publication was provided for Individual 2 (Supplementary Material). Trio exome sequencing (ES) had previously been performed at a commercial lab for all three individuals, and FASTQ files were requested and reanalyzed in the UDN as described previously[7]. Trio GS was performed for Individuals 1 and 2 through the UDN[8]. Review of the literature, transcript databases and CMA data was conducted for variant reannotation. Clinically relevant results were confirmed by an orthogonal method before communication to the individuals/parents.

Results

Individual 1,is a 3 year, 4 month-old Caucasian male, with refractory infantile spasms, atonic seizures, developmental regression with onset of seizures and current skills of a 9–12 month old (details in Supplementary Material). The family history was unremarkable. Pre-UDN testing included a normal CMA, normal comprehensive epilepsy panel (which included *CDKL5*) and non-diagnostic trio ES. The latter two tests had been performed through the same commercial laboratory

Reanalysis of the ES data in the UDN was non-diagnostic, but GS revealed a novel *de novo* hemizygous *CDKL5* c.2842C>T; p.(Arg948*) variant (NM_001323289.1) in exon 17, interpreted as pathogenic by the UDN laboratory. The alternative NM_001323289.1 isoform chosen by this laboratory for analysis is more biologically relevant than the canonical transcript, because it is the most abundant isoform expressed within the central nervous system[9]. The patient's clinical phenotype was consistent with *CDKL5*- associated epileptic encephalopathy (MIM#300672). Since LoF is the disease mechanism for this disorder we evaluated why this variant had not been reported previously on the epilepsy panel, ES or the UDN reanalysis of exome data.

CDKL5 has multiple isoforms due to alternative splicing, three of which are included in the NCBI RefSeq Database (Figure 1A). The commercial laboratory had used the canonical (longest) transcript NM_003159 for ES analysis in October 2016 (and for the prior epilepsy panel in 2015), and this variant had not been reported because it appeared in an intronic region just past exon 18. However, in the alternative RefSeq transcript NM_001323289.1 that was considered by the UDN genomic sequencing laboratory, this variant is within a coding region designated as exon 17. This exon extends further in the 3' direction in this transcript, into what is an intronic region in the canonical transcript. The variant identified in Individual 1 is located in this "extra" region of exon 17, indicated by the red arrow in Figure 1A. In ClinVar there are two other individuals with different pathogenic variants within this "extra" region of exon 17 (these variants have been previously reported[10,11]), and there are no loss-of-function variants within this region in gnomAD (https://gnomad.broadinstitute.org). Interestingly, the research-based pipeline used for UDN exome reanalysis also did not report the variant as the corresponding Ensembl transcript had not yet been added to the annotation build used.

The alternative transcript NM_001323289.1 was added to the NCBI RefSeq database in April 2016, but had not yet been added to the commercial laboratory's annotation pipeline at the time of this individual's epilepsy panel and ES analyses (personal communication with commercial laboratory). The commercial laboratory listed two other reasons for this variant not being reported, including (1) the presence of structural errors in the hg19 genome assembly making it difficult to accurately assess variants in this region and (2) prior literature indicating that truncating variants in *CDKL5* past amino acid 938 in the canonical transcript are not given "very strong" pathogenic designation[12]. However, the latter reason is irrelevant for the variant identified in Individual 1, as amino acid 938 in the canonical transcript is further downstream. Therefore, we asserted that this *de novo* CDKL5 variant is pathogenic and clinically relevant because it is a loss-of-function coding variant in the brain-expressed transcript.

Individual 2, a 6 year, 2 month-old Caucasian female had developmental delays, borderline intelligence, macrocephaly, mild dysmorphic features and a paternal family history of intellectual disability (details in Supplementary Material). A pre-UDN CMA had revealed a 142kb deletion of uncertain significance on chromosome 7q31.1 in 2013, interpreted as a risk factor for autism, Tourette syndrome and ADHD with reduced penetrance. Pre-UDN commercial trio ES and a reanalysis were non-diagnostic.

On GS through the UDN, an interstitial deletion of ~135kb in 7q36.1 involving exons 8 through 55 of *KMT2C* (NM_170606.3) was reported. *KMT2C* haploinsufficiency causes Kleefstra syndrome type 2 (MIM #617768), and there was clinical overlap between this disease entity and features described in Individual 2. The deletion was inherited from her father who had similar clinical characteristics. The cytogenetics lab had not reported it on initial microarray in March of 2013 but upon request they re-evaluated the data and confirmed the interstitial deletion “arr[hg19] 7q36.1(151,839,151–151,965,981)x1”, estimating it to be approximately 127kb in size. The reason for the cytogenetics laboratory not reporting this deletion previously was that the size was below their threshold of 300 kb and additionally deletions within *KMT2C* are found in healthy individuals in the Database of Genomic Variants (DGV, http://dgv.tcag.ca/dgv/app/home). However, upon further scrutiny, it is evident that these are restricted to the 5’ region (exons 2–6) of the gene (Figure 1B). In contrast, the deletion identified in Individual 2 and her father includes exons 8 through 55, extending into the 3’ region of the NM_170606.3 transcript. Deletions overlapping the exons deleted in our individual have been associated with Kleefstra syndrome 2, including a deletion of exons 2–43 and a stop gain variant in exon 12 that resulted in a truncated protein (Figure 1B)[13].

We examined tissue-specific differences in expression of the isoforms of *KMT2C*. While NCBI RefSeq lists only one transcript for *KMT2C* (NM_170606.3, which corresponds to ENST00000262189.11), there are multiple splice variants in Ensembl and two (ENST00000424877.5 and ENST00000360104.7) are most highly expressed in the brain according to the GTEx database (V8). These shorter brain-expressed transcripts do not include the proximal exons that are deleted in healthy individuals, but do contain the more distal exons deleted in our individual, as well as those previously reported as pathogenic[13]. With this information, the clinical laboratory agreed that the deletion in our individual and their father is likely pathogenic [personal communication] and she was given a diagnosis of Kleefstra syndrome 2.

Individual 3, an 8 year, 5 month-old African American male, had autistic features, and mild developmental delays, (details in Supplementary Material). On pre-UDN commercial trio ES, a novel maternally inherited pathogenic variant (c.967delA, p.(Ser323Alafs*2)) in exon 10 of X-linked *OFD1* (MIM#300170) had been reported (NM_003611.2). Further evaluations were performed for specific features of an *OFD-1* related disorder (MIM#30084, MIM#311200, MIM#300209), but he was found to have normal cognition and no oral, digital, facial, renal or brain abnormalities. The mother’s clinical evaluation and kidney ultrasound were normal. Since the laboratory’s report of a pathogenic variant in *OFD1* was discordant with the individual’s phenotype, we examined this variant further, including the transcript that had been selected for annotation and tissue specific expression.

The canonical transcript NM_003611.2 (ENST00000340096.11) of *OFD1* is used most widely for variant annotation by commercial laboratories (personal communication). However, there are two additional NCBI RefSeq transcripts including NM_001330209.1 and NM_001330210.1 (Figure 1C). The NM_001330209.1 transcript undergoes alternative splicing for exon 10, thus encoding a protein which lacks the corresponding 40 amino acids encoded by this exon[14]. To our knowledge this transcript is not used by commercial labs in their pipelines (personal communication with 4 commercial labs).

We then interrogated control and disease databases to determine if there were differences in variants reported in exon 10 compared to the other exons of *OFD1.* In gnomAD there are 9 high confidence *OFD1* LoF variants in the canonical transcript NM_003611.2. Four of these are located within exon 10 (2 frameshift variants including the variant identified in Individual 3, 1 nonsense variant, and 1 splice acceptor variant), and two of these four are seen in the hemizygous state (Figure 1C). Furthermore, these two hemizygous variants are seen in a total of 15 male individuals in gnomAD. Presumably none of these individuals have *OFD1*-related disorders since individuals with severe pediatric disease are excluded from this database.

In HGMD (Human Gene Mutation Database, http://www.hgmd.cf.ac.uk/ac/index.php) there are 40 LoF variants reported in *OFD1*, but none are in exon 10. However, in ClinVar there are four pathogenic or likely pathogenic variants reported in exon 10 with gender not specified. Two of these are among the frameshift variants reported in gnomAD, one of which is the variant also identified in Individual 3 reported here. The other two variants are nonsense, of which one is reportedly associated with clinical features of an OFD syndrome, and the other without clinical details provided. We contacted three other commercial labs to determine if they had reported pathogenic/likely pathogenic variants in exon 10 of *OFD1*, and only one had - this was a female patient with a frameshift variant with clinical features including bifid tongue, ankyloglossia, alveolar ridging and clinodactyly (personal communication with lab). Attempts to contact the referring provider for further details were unsuccessful. We also contacted an OFD research team with a large cohort, who reported that they had no pathogenic/likely pathogenic variants in exon 10 of *OFD1* in their cohort (personal communication).

Interestingly, another male individual in our clinic was found to have a *de novo* frameshift variant (c.1007dupA p.(Ser337Glufs*3)) in exon 10 of *OFD1* (NM_003611.2), reported as pathogenic on ES by a commercial laboratory. This patient had severe epilepsy, but like Individual 3 he had no other findings suggestive of *OFD1*-associated disorders and upon further examination, there were 6 male individuals in gnomAD with the exact variant. His unaffected brother was found to have the *OFD1* variant as well. He was subsequently found to have a likely pathogenic variant in *ATP1A3* that was consistent with the phenotype.

There do not appear to be clear tissue specific differences among *OFD1* isoforms (https://www.gtexportal.org). However, our findings of LoF variants in exon 10 in healthy individuals, and a general lack of LoF variants in exon 10 in affected individuals, and the fact that exon 10 is spliced out in an alternative transcript, suggest that all putative damaging variants in exon 10 of *OFD1* may not cause disease.

Discussion

Negative results from genetic tests may be due to pathogenic variants being overlooked; less frequently variants may be declared pathogenic erroneously, resulting in a misdiagnosis[11]. Many factors may account for such erroneous results, and here we describe three individuals in which molecular diagnoses were missed, or erroneously assigned, because alternative transcripts and their tissue specific expression patterns were not considered in variant annotation.

A major factor that leads to inconsistencies in how transcripts are applied between laboratories is that there is no single, standardized transcript database utilized, although the ACMG/AMP guidelines for variant interpretation emphasize the importance of understanding the transcript architecture of genes and information about alternative splicing of genes[1]. In Individual 1 an exonic nonsense *CDKL5* variant had been overlooked by the commercial lab because this variant was inferred as being within a non-coding region of the canonical transcript used for annotation, although the sequencing software had identified the variant. Interestingly, another individual with epileptic encephalopathy has recently been reported with a pathogenic variant within this same region of *CDKL5*, and similar cases with pathogenic variants being overlooked because they are in the non-coding region of the canonical transcript have been reported in *SCN8A* and *MITF*[11,15,16]. Additionally, for Individual 1, the guidelines used by the commercial lab for not reporting variants in the distal 3'-end of the canonical transcript were erroneously applied to the alternative isoform; the *CDKL5* variant in Individual 1 was in the longer exon 17 within the alternative transcript NM_001323289.1 and not distal to amino acid 938 in the canonical transcript (since pathogenicity of variants distal to this is thought to be unlikely[12]). Thus, for genes with multiple isoforms it is critical for sequencing laboratories to consider regions in which alternative splicing can occur as well as tissue expression of the various transcripts when variants are prioritized and assessed for pathogenicity.

Similarly, the frameshift *OFD1* variant in Individual 3 was reported as pathogenic because loss-of-function variants throughout *OFD1* are known to cause a spectrum of *OFD1*-related disorders. However, the laboratory did not take into account the presence of an alternative transcript NM_001330209.1 in which exon 10 is spliced out. Although we do not have tissue specific expression differences among the transcripts, the presence of 15 presumably healthy male individuals in gnomAD with hemizygous *OFD1* nonsense variants supports the notion that the transcript containing this exon may not be highly expressed in biologically relevant tissues. Two of the four variants reported as pathogenic or likely pathogenic in ClinVar have also been found in presumably healthy individuals in gnomAD (1 in a female, and the second in 6 males). It is unclear whether the two other individuals reported in ClinVar have an OFD1-related disorder or another condition with overlapping features. However, clearly most individuals with pLoF variants in exon 10 of the canonical *OFD1* transcript do not have an OFD1 disorder. Individual 3 and the individual from our clinical cohort (both with *OFD1* variants reported as pathogenic) reinforce the fact that not all LoF variants in a gene wherein LoF is the mechanism of disease, are indeed associated with disease. Databases such as ClinVar can have erroneous data for disease associations, and when there is remarkable phenotypic discordance with the genetic results there must be a reexamination of the pathogenicity of the variants (including alternative transcripts) by the clinicians and the laboratory to avoid conferring erroneous diagnoses[17]. Incorporation of next generation phenotyping data may provide a more objective determination of whether the phenotype is consistent with the associated disease[18].

When genes known to have multiple transcripts are strong candidates in the differential diagnosis, incorporating tissue specific and temporal expression data using resources such as GTEx may help identify transcripts that are most biologically relevant[5,16]. For individual 1, the canonical transcript NM_003159.2 is mostly expressed in the fetal brain and adult testes, while the alternative transcript NM_001323289.1 is the most abundant isoform expressed in the central nervous system[9]. Pertinent to individual 2, the shorter alternative transcripts (that include the distal exons that were deleted in her), are the isoforms most highly expressed in the brain. The original CMA report did not consider the tissue specific expression of the isoforms of *KMT2C*, or location of Individual 2's deletion in relation to the other small deletions in DGV in the context of alternative transcripts, resulting in an erroneous interpretation that the deleted exons were of uncertain significance. Brain expression data from the GTEx database (V8) was helpful in this case, since the phenotype of Individual 2 and her father was primarily neurologic. In both of these cases, the canonical transcript was not the most biologically relevant isoform, underscoring the importance of laboratories considering tissue specific expression data in variant interpretation. Differential expression patterns of isoforms by tissue system continue to be updated, and continued curation of transcripts is important for accurately annotating variants[16,19,20].

Reanalysis of ES or GS data is a powerful tool and should also include evaluation of alternative transcripts[5]. Considering alternate transcripts allows for the selection of the most biologically relevant transcript based on tissue expression, analysis of exons that may be spliced from the canonical transcript, and detection of variants that may be intronic in the canonical transcript but protein coding in the alternate transcript. Individuals 1 and 2 had negative initial testing (panel, ES or CMA) and then received molecular diagnoses with GS, utilizing updated variant annotation and information about transcript expression. This emphasizes the importance of allowing adequate time to elapse before further genetic testing or reanalysis of prior data so that new information may be included. Similarly, utilizing a different laboratory may overcome limitations specific to an analytic pipeline or reporting protocol[7]. Our experience with Individual 2 highlights the importance of applying a similar reanalysis process to existing CMA data, considering whether genes have multiple transcripts along with transcript-specific expression data when interpreting CMA results.

Finally, these cases illustrate the complexities of having multiple public databases using different languages to describe gene information that is similar but not always identical. The CCDS project is a step toward consistently annotating coding regions of human genes, but this is an imperfect process and these cases illustrate that the CCDS transcript is not always the most biologically relevant. It would be helpful for the genomics community to develop a single resource for describing all known isoforms for each gene and their biological relevance. The Clinical Genome Resource (https://clinicalgenome.org/) curation teams have made strides toward determining the significance of variation for specific genes[21], and including relevant transcript and/or expression data as part of this process would be beneficial to the genetics community.

We understand that our approach to resolving missed diagnoses and misdiagnoses may not be generalizable to all clinic workflows, since it involves resources that may not be readily available, such as reanalysis of sequencing data by an in-house bioinformatician. In addition, our approach involved testing of family members and frequent communication with the sequencing or cytogenetics laboratory directors, which may be challenging in many clinical settings, due to time constraints. However, we wish to highlight the need for clinicians to carefully consider CMA and ES or GS results, or the lack thereof, in situations where there is phenotypic mismatch and understand that these results may change over time, with our understanding of transcript expression and alternative splicing. The three examples in our study highlight how the choice of selecting only the reference/canonical transcript in variant annotation can lead to both the failure to detect pathogenic variants as well as false attribution of pathogenicity to variants in exons that are not biologically relevant.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We wish to thank the families who participated in this study. Research reported in this manuscript was supported by the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under Award Number U01HG007672. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. 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. [PubMed: 25741868]
2. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40(12):1413–1415. [PubMed: 18978789]
3. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol. 2017;18(7):437–451. [PubMed: 28488700]
4. Zimmermann MT. The Importance of Biologic Knowledge and Gene Expression Context for Genomic Data Interpretation. Front Genet. 2018;9:670. [PubMed: 30619486]
5. Bodian DL, Kothiyal P, Hauser NS. Pitfalls of clinical exome and gene panel testing: alternative transcripts. Genet Med. 2019;21(5):1240–1245. [PubMed: 30293991]
6. McCarthy DJ, Humburg P, Kanapin A, et al. Choice of transcripts and software has a large effect on variant annotation. Genome Med. 2014;6(3):26. [PubMed: 24944579]
7. Shashi V, Schoch K, Spillmann R, et al. A comprehensive iterative approach is highly effective in diagnosing individuals who are exome negative. Genet Med. 2019;21(1):161–172. [PubMed: 29907797]
8. 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. [PubMed: 21478889]
9. Hector RD, Dando O, Landsberger N, et al. Characterisation of CDKL5 Transcript Isoforms in Human and Mouse. PLoS One. 2016;11(6):e0157758. [PubMed: 27315173]
10. Archer HL, Evans J, Edwards S, et al. CDKL5 mutations cause infantile spasms, early onset seizures, and severe mental retardation in female patients. J Med Genet. 2006;43(9):729–734. [PubMed: 16611748]
11. Bodian DL, Schreiber JM, Vilboux T, Khromykh A, Hauser NS. Mutation in an alternative transcript of CDKL5 in a boy with early-onset seizures. Cold Spring Harb Mol Case Stud. 2018;4(3).
12. Diebold B, Delepine C, Gataullina S, Delahaye A, Nectoux J, Bienvenu T. Mutations in the C-terminus of CDKL5: proceed with caution. Eur J Hum Genet. 2014;22(2):270–272. [PubMed: 23756444]
13. Koemans TS, Kleefstra T, Chubak MC, et al. Functional convergence of histone methyltransferases EHMT1 and KMT2C involved in intellectual disability and autism spectrum disorder. PLoS Genet. 2017;13(10):e1006864. [PubMed: 29069077]

14. Coene KL, Roepman R, Doherty D, et al. OFD1 is mutated in X-linked Joubert syndrome and interacts with LCA5-encoded lebercilin. Am J Hum Genet. 2009;85(4):465–481. [PubMed: 19800048]

15. Epilepsy Genetics I De novo variants in the alternative exon 5 of SCN8A cause epileptic encephalopathy. Genet Med. 2018;20(2):275–281. [PubMed: 29121005]

16. DiStefano MT, Hemphill SE, Cushman BJ, et al. Curating Clinically Relevant Transcripts for the Interpretation of Sequence Variants. J Mol Diagn. 2018;20(6):789–801. [PubMed: 30096381]

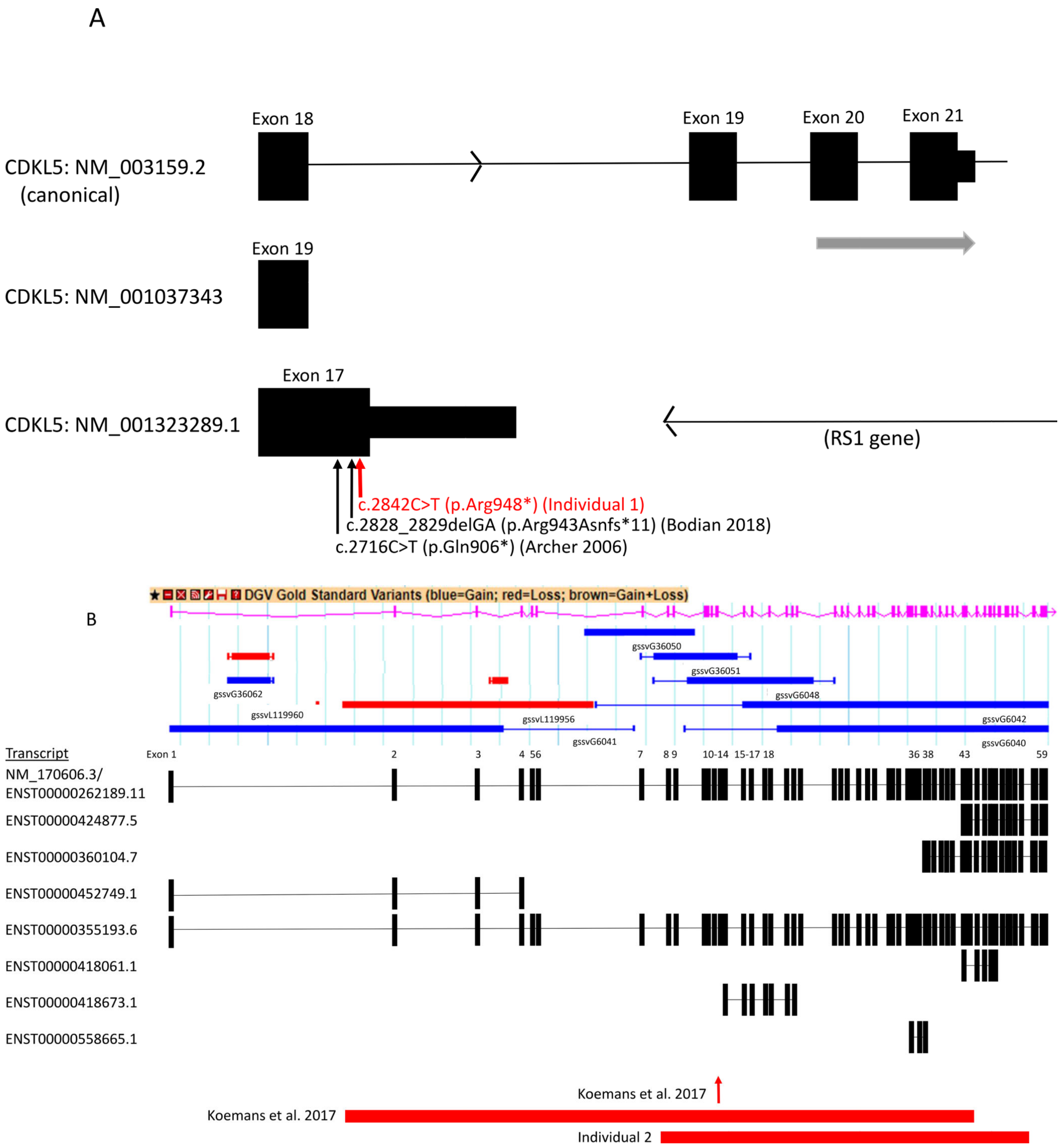
17. Shashi V, McConkie-Rosell A, Schoch K, et al. Practical considerations in the clinical application of whole-exome sequencing. Clin Genet. 2016;89(2):173–181. [PubMed: 25678066]

18. van der Donk R, Jansen S, Schuurs-Hoeijmakers JHM, et al. Next-generation phenotyping using computer vision algorithms in rare genomic neurodevelopmental disorders. Genet Med. 2019;21(8):1719–1725. [PubMed: 30568311]

19. Mele M, Ferreira PG, Reverter F, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348(6235):660–665. [PubMed: 25954002]

20. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648–660. [PubMed: 25954001]

21. DiStefano MT, Hemphill SE, Oza AM, et al. ClinGen expert clinical validity curation of 164 hearing loss gene-disease pairs. Genet Med. 2019;21(10):2239–2247. [PubMed: 30894701]


A
Exon 18
Exon 19
Exon 20
Exon 21
CDKL5: NM_003159.2
(canonical)
Exon 19
CDKL5: NM_001037343
Exon 17
CDKL5: NM_001323289.1
(RS1 gene)
c.2842C>T (p.Arg948*) (Individual 1)
c.2828_2829delGA (p.Arg943Asnfs*11) (Bodian 2018)
c.2716C>T (p.Gln906*) (Archer 2006)
B
DGV Gold Standard Variants (blue=Gain; red=Loss; brown=Gain+Loss)
gssvG36050
gssvG36051
gssvG36062
gssvG6048
gssvL119960
gssvL119956
gssvG6042
gssvG6041
gssvG6040
Transcript
Exon 1
2
3
4 56
7
8 9
10-14
15-17 18
36 38
43
59
NM_170606.3/
ENST00000262189.11
ENST00000424877.5
ENST00000360104.7
ENST00000452749.1
ENST00000355193.6
ENST00000418061.1
ENST00000418673.1
ENST00000558665.1
Koemans et al. 2017
Koemans et al. 2017
Individual 2

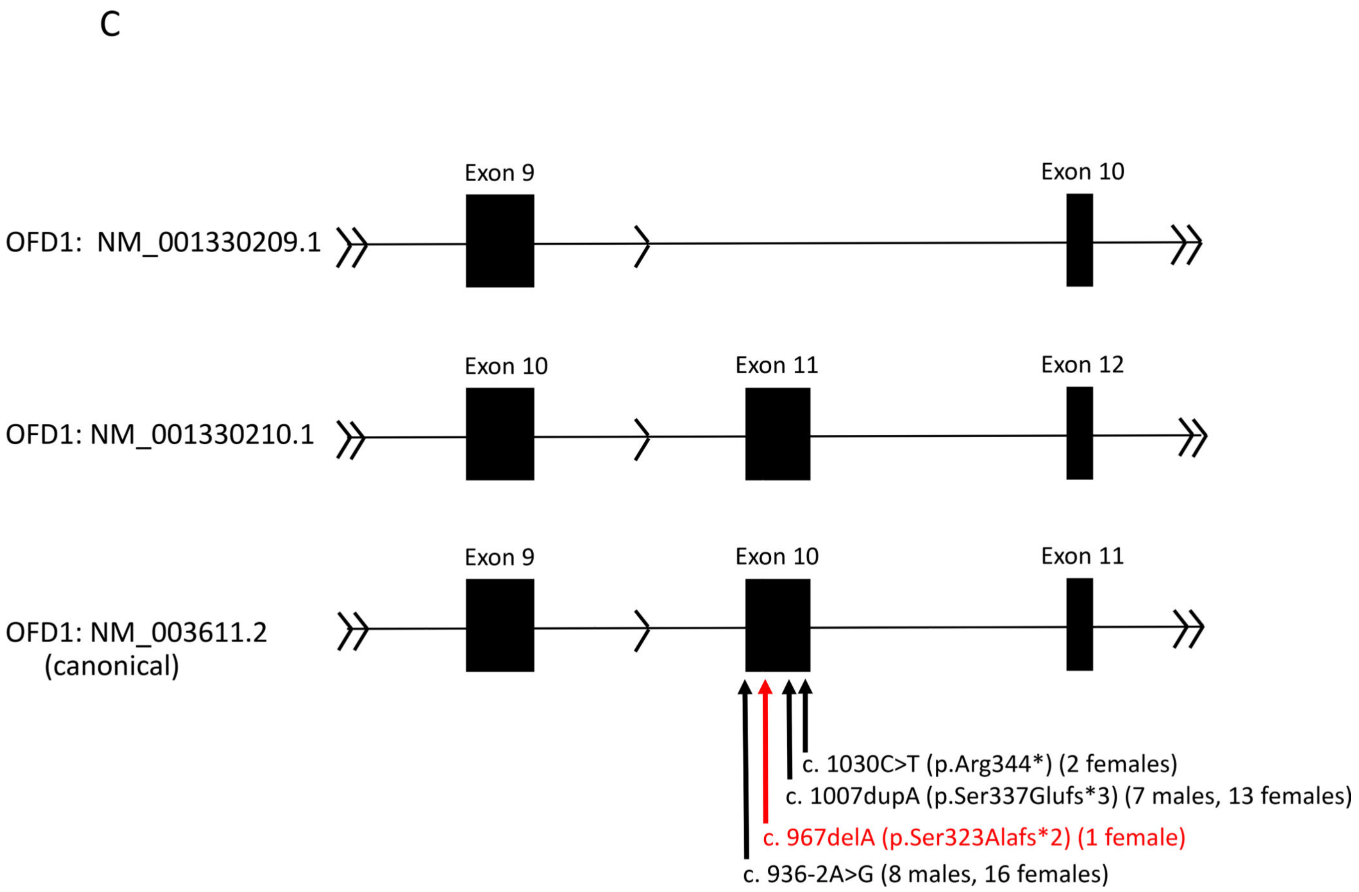

C
Exon 9
Exon 10
OFD1: NM_001330209.1
Exon 10
Exon 11
Exon 12
OFD1: NM_001330210.1
Exon 9
Exon 10
Exon 11
OFD1: NM_003611.2
(canonical)
c. 1030C>T (p.Arg344*) (2 females)
c. 1007dupA (p.Ser337Glufs*3) (7 males, 13 females)
c. 967delA (p.Ser323Alafs*2) (1 female)
c. 936-2A>G (8 males, 16 females)

Figure 1.

Black boxes represent predicted coding sequence, smaller black boxes represent untranslated regions (UTRs), and black lines represent intronic sequences with arrowheads indicating the direction of transcription. Exons are labeled.

Figure 1A. The three *CDKL5* RefSeq transcripts NM_003159.2 (canonical), NM_001037343, and NM_001323289.1 are shown. The variant identified in Individual 1 is designated with a red arrow. Pathogenic variants in this “extra” region of exon 17 previously reported by Bodian et al. (2018) and Archer et al. (2006) are designated with black arrows. The thick grey arrow indicates the region beyond codon 938 (NM_003159.2 only) for which Diebold et al. (2014) urges caution against over-assigning pathogenicity.

Figure 1B.The *KMT2C* RefSeq transcript NM_170606.3 (Ensembl transcript ENST00000262189) and seven additional protein coding Ensembl transcripts are shown. Deletions (red bars) and duplications (blue bars) in DGV are shown at the top of the figure. At the bottom, the p.(Lys564*) pathogenic variant reported by Koemans et al. (2017) is represented by a red arrow. The 127kb deletion identified in Individual 2 and the 203kb deletion previously reported by Koemans et al. (2017) are represented by horizontal red bars.

Figure 1C.The three OFD1 RefSeq transcripts NM_001330209.1, NM_001330210.1, and NM_003611.2 (canonical) are shown. The variant identified in Individual 3 is designated with a red arrow. The other 3 predicted LoF variants seen in gnomAD are labeled with black arrows, along with frequency and gender in which they were identified.