Genetics inMedicine | BRIEF COMMUNICATION

Pitfalls of clinical exome and gene panel testing: alternative transcripts

Dale L. Bodian, PhD[1], Prachi Kothiyal, PhD[1] and Natalie S. Hauser, MD[1]

Purpose: Clinical exome and gene panel testing can provide molecular diagnoses for patients with rare Mendelian disorders, but for many patients these tests are nonexplanatory. We investigated whether interrogation of alternative transcripts in known disease genes could provide answers for additional patients.

Methods: We integrated alternative transcripts for known neonatal epilepsy genes with RNA-Seq data to identify brain-expressed coding regions that are not evaluated by popular neonatal epilepsy clinical gene panel and exome tests.

Results: We found brain-expressed alternative coding regions in 89 (30%) of 292 neonatal epilepsy genes. The 147 regions encompass 15,713 bases that are noncoding in the primary transcripts analyzed by the clinical tests. Alternative coding regions from at least 5 genes carry reported pathogenic variants. Three candidate variants in these regions were identified in public exome data from 337 epilepsy patients. Incorporating alternative transcripts into the analysis of neonatal epilepsy genes in 44 patient genomes identified the pathogenic variant for the epilepsy case and 2 variants of uncertain significance (VUS) among the 43 control cases.

Conclusion: Assessment of alternative transcripts in exon-based clinical genetic tests, including gene panel, exome, and genome sequencing, may provide diagnoses for patients for whom standard testing is unrevealing, without introducing many VUS.

Genetics in Medicine (2019) 21:1240–1245; https://doi.org/10.1038/s41436-018-0319-7

Keywords: gene panel testing; exome sequencing; alternative transcripts; epilepsy; diagnostic yield

INTRODUCTION

Clinical gene panel and exome sequencing have transformed the diagnosis of rare Mendelian disorders, and greatly reduce repeated blood samples, cost, and time. Despite the successes, many cases remain unresolved, with reported diagnostic yields ranging from ~15% to ~60%. Negative results have been attributed to factors including incomplete knowledge of disease architecture, a focus on exonic variation, challenges in variant pathogenicity interpretation, and technical limitations influencing variant calling. Assumptions incorporated into test design and analysis pipelines can also contribute to missed diagnoses. For example, assumptions about inheritance patterns led to overlooked variants in the imprinted genes *CDKN1C*[1] and *MAGEL2* (refs. [2,3]).

Incomplete consideration of alternative transcripts can also cause pathogenic variants to be missed. We recently reported a patient with epileptic encephalopathy for whom clinical gene panel testing was unrevealing.[4] Research-based genome sequencing identified a de novo variant in an alternative transcript of *CDKL5*, a gene targeted by the clinical panel. Similarly, in a reanalysis of previously undiagnosed epilepsy patients, the Epilepsy Genetics Initiative identified three cases with de novo variants in an alternative transcript of *SCN8A*, an isoform that had only recently been added to the set of transcripts evaluated.[5]

These variants demonstrate that alternative transcripts can be disease-relevant. Here, we investigate whether these examples are isolated cases, or whether alternative isoforms may be more widely relevant to clinical sequencing. Using neonatal epilepsy as an example, we found that clinically relevant alternative transcripts are common in known disease genes. The results suggest that alternative isoforms should be assessed more routinely in assays dependent on a set of reference transcripts, including gene panel, exome, and genome sequencing, and that reanalysis or resequencing incorporating alternative transcripts should be considered for patients with negative test results.

MATERIALS AND METHODS

Genes and transcripts

Gene symbols and RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) identifiers for the primary transcripts assessed by neonatal epilepsy clinical gene panels as of December 2017 were provided by the genetic testing companies. Genes were limited to those strongly associated with neonatal epilepsy, defined as a primary seizure condition starting in the first

[1]Inova Translational Medicine Institute, Inova Health System, Falls Church, VA, USA. Correspondence: Dale L. Bodian (dale.bodian@inova.org)

Submitted 29 April 2018; accepted: 14 September 2018
Published online: 8 October 2018

months of life. Genomic coordinates (hg19) of the panel transcripts and alternative transcripts associated with the neonatal epilepsy genes were extracted from RefSeq and the GENCODE v27 comprehensive data set (hg19.wgEncodeGencodeCompV27lift37), downloaded from the University of California–Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/). The GENCODE data were filtered for transcripts annotated as protein-coding and with complete coding regions.

Alternative coding regions

Alternative coding regions were computed for each gene by subtracting the genomic positions of the coding exons and 20 flanking bases of the panel transcript(s) from the coding exons and 20 flanking bases of the filtered GENCODE transcript(s). Evidence of expression in neonatal brain for a region was defined as ≥50% of the coding bases supported by >20 normalized reads in the fetal or infant RNA-Seq data from polyA+ transcriptomes of human dorsolateral prefrontal cortex (DLPFC), downloaded from the Lieber Institute for Brain Development (LIBD) DLPFC Development UCSC custom track hub.[6] For alternative coding regions confined to intronic flanking sequence, expression was computed using the associated exonic bases.

Variants

Variants were obtained from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) version 201711 and Human Genome Mutation Database (HGMD) Professional (Qiagen) version 2013.2, preprocessed as described[7] and annotated with ANNOVAR (http://annovar.openbioinformatics.org/) version 2017-06-01. Variants were considered pathogenic if they were categorized as pathogenic or likely pathogenic in ClinVar or as disease-causing mutations (DM) in HGMD, limited to neonatal epilepsy-related disorders when the condition was provided. Relative variant deleteriousness was defined as putative loss of function (stopgain, stoploss, consensus splice site, frameshift, startloss) > nonsynonymous > synonymous. Variants were filtered for allele frequency <0.0001, using the maximum frequency from the gnomAD genomes and exomes,[8] where average coverage was ≥20 or ≥50, respectively. Annotations were computed from brain-expressed transcripts, defined as transcripts for which every exon has >50% of its coding bases supported by >20 normalized reads in the fetal or infant LIBD RNA-Seq data.[6]

Genome sequencing

This study was approved by the Inova Institutional Review Board (IRB 15–18196). Full written informed consent was obtained from the participants, with the parents providing consent for minors. Genome sequencing methods are described in the Supplementary Methods.

RESULTS

Neonatal epilepsy genes have brain-expressed coding regions that are not evaluated by clinical tests

The genomic positions sequenced by gene panel tests are limited to the exons and flanking sequences of a set of reference transcripts. We determined a set of "alternative coding regions," defined as genomic regions of a gene that would be newly sequenced by consideration of alternative transcripts, using neonatal epilepsy genes as an example. We first generated a list of transcripts sequenced by three representative clinical gene panel tests, the Invitae Epilepsy Panel (189 genes), the EpilepsyNext panel from Ambry Genetics (100 genes), and the Fulgent NeoNatal Epilepsy panel (276 genes), from data kindly provided by the companies. All three companies confirmed that the provided transcripts are the primary reference transcripts for these genes in both their gene panel and exome tests. The combined set of transcripts from the three panels has 292 genes and 305 transcripts (Supplementary Table S1). Most of the genes (96%) are represented by a single transcript, and 13 genes (4%) are represented by two transcripts.

To determine the alternative coding regions, we subtracted the genomic positions of the coding exons and flanking bases of the panel transcripts from those of transcripts from GENCODE[9] (Supplementary Figure S1). The GENCODE data set has 1372 quality-filtered transcripts for the 292 panel genes, with 1–39 transcripts per gene (median 3). Most of the genes (85%) have alternative transcripts, consistent with the ubiquity of alternative splicing.[10] The alternative coding regions were then limited to those transcribed in fetal or infant brain to prioritize sequences more likely to be relevant for neonatal epilepsy, resulting in 147 regions (Supplementary Table S2). Eighty-nine genes (30%) have at least one alternative coding region (range 1–6). The regions are 1–801 nucleotides long (median 74) and encompass a total of 15,713 genomic positions, of which 11,369 are exonic coding bases (72%) and 4,344 are in flanking sequences. The regions are distributed throughout the length of the encoded proteins: 19% encode alternate N-termini, 58% encode alternate C-termini, and a partially overlapping 50% are middle regions.

The set of alternative coding regions includes exons from transcripts previously shown to be expressed in brain, including alternative isoforms of *CACNA1A*, *CDKL5*, *DNM1*, *SCN2A*, and *SCN3A* (see Additional References). Alternative coding regions were also found for two bicistronic loci, *MOCS1* and *MOCS2*, each of which encodes two overlapping open reading frames, of which only one is in the set of transcripts assessed by the clinical panels. Alternative exon 5A from *SCN8A*, the location of recently identified pathogenic variants,[5] was excluded because both isoforms are assessed by the Invitae panel.

Table 1 Pathogenic variants

Variant	Gene	Panel transcript change		Alternative transcript change		AF
		Type	Details	Type	Details	
a) Variants in alternative coding regions[a]						
chr19:13318673 delCTGCTGCTGCTGCTGCTGCTG	*CACNA1A*	UTR3	NM_001127221.1:c.*187_*167delCAGCAGCAGCAGCAGCAGCAG	Polyglutamine repeat	ENST00000614285.4_1:c.6973_6993del:p.2325_2331del	na[b]
chr17:42987501 CAGCTAAC>GAT	*GFAP*	na	na	Frameshift	ENST00000435360.7_2:c.1292_1299ATC	0
chr17:42987511C>T	*GFAP*	na	na	nonsynonymous	ENST00000435360.7_2:c.G1289A:p.R430H	8e-6
chr6:39874535 delCT	*MOCS1*	UTR3	NM_005943.5:c.*366_*365delAG	Frameshift	ENST00000373195.7_1:c.1199_1200del:p.E400fs	2e-5
chr6:39902057 delCC	*MOCS1*	na	na	Frameshift	ENST00000373188.6_1:c.99_100del:p.G33fs	0
chr5:52404362G>A	*MOCS2*	na	NM_004531.4:c.-1358C>T	Stopgain	ENST00000361377.8_3:c.C130T:p.R44X	0
chr5:52404386G>A	*MOCS2*	na	NM_004531.4:c.-1382C>T	Stopgain	ENST00000361377.8_3:c.C106T:p.Q36X	0
chr5:52404404G>A	*MOCS2*	na	NM_004531.4:c.-1400C>T	Stopgain	ENST00000361377.8_3:c.C88T:p.Q30X	8e-6
chr5:52404447A>T	*MOCS2*	na	NM_004531.4:c.-1443T>A	Nonsynonymous	ENST00000361377.8_3:c.T45A:p.S15R	8e-6
chr5:52404459A>C	*MOCS2*	na	NM_004531.4:c.-1455T>G	Stopgain	ENST00000361377.8_3:c.T33G:p.Y11X	2e-5
chr5:52404473C>A	*MOCS2*	UTR5	NM_004531.4:c.-1469G>T	Nonsynonymous	ENST00000361377.8_3:c.G19T:p.V7F	4e-6
chr5:52405544G>A	*MOCS2*	UTR5	NM_004531.4:c.-2540C>T	Stopgain	ENST00000361377.8_3:c.C16T:p.Q6X	2e-5
chr5:52405545 delGCACAGCGGCACCATCCCGCCTA	*MOCS2*	UTR5	NM_004531.4:c.-2541_-2563 delTAGGCGGGATGGTGCCGCTGTGC	Startloss	ENST00000361377.8_3:c.-8_15del23	0
chr5:52405557C>T	*MOCS2*	na	NM_004531.4:c.-2553G>A	Startloss	ENST00000361377.8_3:c.G3A:p.M1I	0
chr5:52405559T>C	*MOCS2*	na	NM_004531.4:c.-2555A>G	Startloss	ENST00000361377.8_3:c.A1G:p.M1V	1e-5
chr9:130453077 delC	*STXBP1*	UTR3	NM_003165.3:c.*40delC	Frameshift	ENST00000373299.4_1:c.1726delC:p.Q576fs	0
b) Reannotated variants[a,c]						
chrX:18646710C>T	*CDKL5*	Splicing	NM_003159.2:c.2713+3C>T	Stopgain	ENST00000623535.1_2:c.C2716T:p.Q906X	0
chrX:76952065C>A	*ATRX*	Nonsynonymous	NM_000489.4:c.G370T:p.G124C	Stopgain	ENST00000395603.7_1:c.G370T:p.E124X	0
chr3:33106965A>T	*GLB1*	Nonsynonymous	NM_000404.3:c.T542A:p.I181K	Stopgain	ENST00000307377.12_1:c.T330A:p.Y110X	0
chr5:52402972G>A	*MOCS2*	Synonymous	NM_004531.4:c.C33T:p.F11F	Stopgain	ENST00000361377.8_3:c.C220T:p.Q74X	0

AF allele frequency in gnomAD, *UTR3* 3′ untranslated region, *UTR5* 5′ untranslated region.
[a]Variant sources and data supporting pathogenicity are provided in Supplementary Table S3.
[b]The variant is located in a low complexity region that may be difficult to sequence.
[c]Constraint metrics reflecting the probability of intolerance to loss-of-function variation are provided in Supplementary Table S4.

Table 2 Reportable variants in patient data

Variant	Clinical significance	Gene	Panel transcript change		Alternative transcript change		AF
			Type	Details	Type	Details	
a) Epilepsy patient exomes							
chr12:7343151G>C	Uncertain	*PEX5*	na	na	Nonsynonymous	ENST00000434354.6_1:c.G178C:p.A60P	0
chr17:42987518G>T	Uncertain	*GFAP*	na	na	Nonsynonymous	ENST00000435360.7_2:c.C1282A:p.P428T	1e-5
chr4:15480865C>T	Uncertain	*CC2D2A*	na	na	Stopgain	ENST00000438599.6_2:c.C142T:p.R48X	6e-5
b) Epilepsy patient genome							
chrX:18646821 delAG	Pathogenic	*CDKL5*	na	na	Frameshift	ENST00000623535.1_2:c.2827_2828del:p.R943fs	0
c) Control genomes							
chr2:86121052C>T	Uncertain	*ST3GAL5*	na	na	Splicing	ENST00000640418.1_1:c.19–1G>A	0
chrX:153296784G>A	Uncertain	*MECP2*	Synonymous	NM_004992.3: c.C495T:p.P165P	Nonsynonymous	ENST00000611468.1_1:c.C481T:p.L161F	4e-05

AF allele frequency, *na* not applicable.

Variants in the alternative coding regions can be disease-relevant

To determine whether the alternative coding regions may be clinically relevant, we asked whether any known pathogenic variants are located in these regions. Although these regions are not routinely examined by exon-based clinical tests, variants may have been identified using other methods. We found 16 pathogenic variants located in alternative coding regions of 5 genes, *CACNA1A*, *GFAP*, *MOCS1*, *MOCS2*, and *STXBP1* (Table **1a**). Of the 15 published variants, the reported impact is consistent with the alternative transcript annotation, and 6 variants have functional data supporting an effect on protein function or expression (Supplementary Table S3). These examples confirm that variants in alternative coding regions can be disease-relevant.

Alternative transcripts may alter reporting of variants detected by panel transcripts

In addition to identifying previously undetected variants, assessment of alternative transcripts could alter variant reporting by providing alternative annotations of the sequenced variants. Alternative annotations could impact interpretation of pathogenicity, variant prioritization, and hypothesized mechanisms of disease pathogenesis. To explore this effect, we searched for reported variants that are reannotated as loss-of-function variants based on alternative, brain-expressed, transcripts. We identified four pathogenic variants, in the genes *CDKL5*, *ATRX*, *GLB1*, and *MOCS2*, and no benign variants or variants of uncertain significance (VUS) (Table **1b**). The accuracy of the predicted protein changes is unknown, but the variant in *ATRX* was shown to impact splicing, an effect not predicted by either the panel or alternate transcript annotations. These results suggest that alternative transcripts could alter variant reporting, consistent with studies demonstrating dependence of annotations on the set of reference transcripts,[11] and that, like all variant annotations, the predicted impact should be interpreted cautiously.

Impact of alternative transcripts on patient data

To examine the potential impact of alternate transcripts on patient data, we reannotated publicly available exome results from 337 probands diagnosed with epileptic encephalopathy (epi4kdb.org). Although these data are themselves limited by a set of reference transcripts, we identified three rare protein-coding variants in alternative coding regions and no variants with potentially more deleterious annotations (Table **2a**). We also analyzed genomes from 44 probands with congenital disorders, including 1 patient with neonatal epilepsy,[4] and identified three rare protein-coding variants in alternative transcripts from the epilepsy panel genes, the pathogenic *CDKL5* variant in the epilepsy patient and two VUS in patients without epilepsy (Table **2b, c**). These results suggest that consideration of alternative transcripts can improve detection of pathogenic variants without introducing a large number of VUS.

DISCUSSION

Clinical gene panel and exome sequencing have provided molecular diagnoses for many rare disease patients, but for some patients these tests are nonexplanatory. Ongoing efforts to identify overlooked pathogenic variants include novel disease gene discovery and analysis of regulatory variants. The results presented here reaffirm that incomplete representation of alternative transcripts also causes pathogenic variants to be missed, and suggest that more complete evaluation of protein-coding regions in known disease genes will increase diagnostic yields.

Recent publications have highlighted the importance of reanalysis of sequencing data from initially uninformative exome tests.[12–15] Our study underscores this conclusion, and suggests that assessment of alternative transcripts should be part of the re-evaluation. Because many of the alternative coding regions identified here are fully captured by commonly used exome capture kits (48% by Illumina TruSeq and 67% by Agilent SureSelect), initial re-evaluation may require only computational reanalysis without additional sequencing.

Our analysis identified a set of alternative coding regions for 292 neonatal epilepsy genes, but this set is not expected to be comprehensive. Determination of the regions relied on a database of reference transcripts that is incomplete,[6] and regions may have been excluded due to low read counts in the expression data. The set of alternative coding regions is also likely to include false positives, including regions resulting from computational errors such as incorrect transcript mapping to the reference genome, and regions that are not relevant to seizure disorders. The regions may also include segments difficult to sequence by short-read technologies, such as the polyglutamine repeat region of *CACNA1A*.

In this study, we focused on neonatal epilepsy but consideration of alternative transcripts is likely to benefit clinical testing for a broad range of diseases. Alternative splicing occurs in a wide variety of tissues and cell types[16] and affects ~95% of multiexon genes.[10] Pathogenic variants affecting alternative exons have been identified in the gene *ACTG2* for the smooth muscle disorder megacystis–microcolon–intestinal hypoperistalsis syndrome, in *SCN5A* for the cardiovascular disorder congenital long-QT syndrome, and in *ABCA4* for the retinal disorder Stargardt disease (see Additional References).

Currently there is no standardized method for selecting transcripts for clinical tests. Each company individually defines a set of primary transcripts based on sources such as HGMD (Qiagen), Alamut (Interactive Biosoftware), and literature review, or selects the longest transcript. Efforts underway to more fully characterize the human transcriptome across cell types and developmental stages[6,17,18] and to curate clinically relevant exons[19] will aid the detection and evaluation of variants in alternative transcripts. Variants presented here support the disease relevance of some alternative transcripts. Incorporating alternative transcripts into sequencing tests will likely yield data useful for determining additional disease-relevant regions.

This study has important implications for clinical practice. Although it is unknown how many attainable diagnoses are missed due to the nonassessment of alternative transcripts, our results indicate that clinicians should consider genetic tests that assess multiple isoforms, particularly for patients with negative test results for whom a positive result was expected. Adding additional transcripts may also introduce VUS requiring further evaluation, but ongoing efforts to fully characterize the transcriptome will help resolve the uncertain results, yielding additional pathogenic variants, increasing diagnosis rates, and ensuring a more complete genetic evaluation.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article (https://doi.org/10.1038/s41436-018-0319-7) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

We thank Callie Diamonstein, Ram Iyer, Rebecca Miller, and Thierry Vilboux for helpful discussions and comments on the manuscript. This work was funded by the Inova Health System.

DISCLOSURE

The authors declare no conflicts of interest.

REFERENCES

1. Bodian DL, Solomon BD, Khromykh A, et al. Diagnosis of an imprinted-gene syndrome by a novel bioinformatics analysis of whole-genome sequences from a family trio. Mol Genet Genomic Med. 2014;2:530–538.
2. Palomares-Bralo M, Vallespin E, Del Pozo A, et al. Pitfalls of trio-based exome sequencing: imprinted genes and parental mosaicism-MAGEL2 as an example. Genet Med. 2017;19:1285–1286.
3. Aten E, Fountain MD, van Haeringen A, Schaaf CP, Santen GW. Imprinting: the Achilles heel of trio-based exome sequencing. Genet Med. 2016;18:1163–1164.
4. 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:a002360.
5. Epilepsy Genetics Initiative. De novo variants in the alternative exon 5 of SCN8A cause epileptic encephalopathy. Genet Med. 2018;20:275–281.
6. Jaffe AE, Shin J, Collado-Torres L, et al. Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. Nat Neurosci. 2015;18:154–161.
7. Bodian DL, Vilboux T, Hourigan SK, et al. Genomic analysis of an infant with intractable diarrhea and dilated cardiomyopathy. Cold Spring Harb Mol Case Stud. 2017;3:a002055.
8. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–291.
9. Rosenbloom KR, Sloan CA, Malladi VS, et al. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. 2013;41:D56–63.
10. 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;40:1413–1415.
11. McCarthy DJ, Humburg P, Kanapin A, et al. Choice of transcripts and software has a large effect on variant annotation. Genome Med. 2014;6:26.
12. Wright CF, McRae JF, Clayton S, et al. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. Genet Med. 2018: https://doi.org/10.1038/gim.2017.246. Accessed 3 October 2018.
13. Nambot S, Thevenon J, Kuentz P, et al. Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. Genet Med. 2018;20:645–654.

14. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med. 2017;19:209–214.
15. Eldomery MK, Coban-Akdemir Z, Harel T, et al. Lessons learned from additional research analyses of unsolved clinical exome cases. Genome Med. 2017;9:26.
16. Mele M, Ferreira PG, Reverter F, et al. The human transcriptome across tissues and individuals. Science. 2015;348:660–665.
17. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
18. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–660.
19. DiStefano MT, Hemphill SE, Cushman BJ, et al. Curating clinically relevant transcripts for the interpretation of sequence variants. J Mol Diagn. 2018; https://doi.org/10.1016/j.jmoldx.2018.06.005. Accessed 3 October 2018.