Increasing diagnostic yield by RNA-Sequencing in rare disease—bypass hurdles of interpreting intronic or splice-altering variants

Dong Li[1], **Lifeng Tian**[1], **Hakon Hakonarson**[1,2,3]

[1]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA; [2]Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA, USA; [3]Divisions of Human Genetics and Pulmonary Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

Correspondence to: Hakon Hakonarson, MD, PhD. Center for Applied Genomics, The Children's Hospital of Philadelphia, Abramson Research Building, Suite 1216B, 3615 Civic Center Boulevard, Philadelphia, PA 19104-4318, USA. Email: hakonarson@chop.edu.

Provenance: This is a Guest Editorial commissioned by Section Editor Wan Wang, PhD (Medical Technology School, Xuzhou Medical University, Xuzhou, China).

Comment on: Cummings BB, Marshall JL, Tukiainen T, *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med 2017;9.

Submitted Oct 10, 2017. Accepted for publication Oct 13, 2017.
doi: 10.21037/atm.2018.01.14
View this article at: http://dx.doi.org/10.21037/atm.2018.01.14

Whole exome sequencing (WES) has proven to be a powerful tool for the diagnosis of Mendelian disorders. Most studies on the application of WES have reported a diagnostic yield of 25–50% (1-4), leaving a significant number of undiagnosed cases. As a result, various strategies to improve the genetic diagnosis have emerged. In this regard, reanalyzing exome data on a regular basis has been shown to identify additional pathogenic variants (5) increasing the diagnostic yield of undiagnosed cases by about 10% (6). Whole genome sequencing (WGS) has the potential to identify relatively small (<10 Kb) copy number variations (CNVs) and complex genomic rearrangements that are missed by genotyping arrays (7,8). The 10× Genomics platform which allows barcoded sequence reads to be assembled into long-range DNA fragments (50 kb in size or greater) also facilitates the identification of CNVs and complex rearrangements (9). While technologies to produce large amounts of sequence information have grown exponentially in the past decades, the ability to fully interpret the resulting variants has lagged behind (3). Recently, two studies reported on utilization of RNA-Sequencing (RNA-Seq) to help interpret variants of unknown significance (VUS) identified through WES/WGS in rare diseases, providing new insights and improving molecular diagnosis yield (10,11).

RNA-Seq (i.e., transcriptome sequencing) is routinely applied in gene expression analysis and pathogen detection of infectious diseases. This technology can also theoretically guide the prioritization in searching for causal variants by recognizing the pattern of allele specific expression (ASE), aberrant splicing (such as exon skipping or extension), and up- or down-regulated expression levels of any given transcripts. Multiple mechanisms have been shown to be associated with ASE (12-14), e.g., genetic imprinting, X-inactivation, truncating mutation, alternative splicing, and allele specific transcription induced by genetic changes. Some notable examples of aberrant splicing causing diseases are: (I) aberrant splicing of *SMN1* exon 7 caused by a variant in intron 7 resulting in spinal muscular atrophy (15); (II) multiple exons deletion (exons 73–76) of *DMD* caused by a point mutation resulting in Duchenne muscular dystrophy (16). Similarly, variants in promotor, enhancer, intronic or coding regions can also lead to aberrant expression, as documented in several expression quantitative trait loci (eQTL) studies (17-19) and in a RNA-Seq study (20). MacArthur's group integrated RNA-Seq data from affected muscle tissues and WES/WGS data to identify genetic causes in primary muscle disorders (10), demonstrating how large RNA-Seq dataset can be utilized to explore pathogenic relevance of both non-coding and coding variants. This initial success showed that RNA-Seq has substantive potential in reliably and affordably identifying pathogenic variants in both known and new disease genes.

In a recent study, Cummings *et al.* performed RNA-Seq on affected muscle tissues from 63 individuals with suspected primary muscle disorders, 13 of whom had genetic causes uncovered that were expected to affected transcription (positive controls) and 50 of whom didn't have a specific molecular diagnosis (10). The authors pointed out the reasons for sequencing the disease relevant tissues are twofold: ability to evaluate tissue-dependent expression and splicing profiles; and to overcome the issue of muscle disease genes not being well expressed in more easily-accessible tissues. The authors sequenced diseased muscle samples obtained from biopsies that were part of standard clinical protocol for undiagnosed patients using the same protocol employed by the Genotype-Tissue Expression (GTEx) Consortium project. They subsequently analyzed the data using identical parameters/pipelines looking for aberrant splicing events, aberrant gene expression levels, and ASE that were unique to patients when compared to 184 quality metrics matched skeletal muscle RNA-Seq samples from the GTEx project, used as a reference panel. In 13 previously diagnosed patients, the authors could manually identify the transcript level changes caused by the genetic variants previously reported in genomic testing and were subsequently able to define the parameters for subsequent analysis of the undiagnosed cases. This led to an overall diagnostic yield of 35% and delivered a molecular diagnosis for 17 patients in whom the genetic diagnosis was unrevealing before.

Focusing on known disease genes, the authors primarily investigated the splicing defects resulting from both coding and non-coding pathogenic variants, such as exon skipping, exon extension, and exonic and intronic splice gain, and subsequently confirmed the findings by reverse transcription polymerase chain reaction (RT-PCR) analysis. First, the authors highlighted the value of RNA-Seq in classification of VUS. In two cases, the previously identified splice-site variants were showed to disrupt the normal splicing, enabling the pathogenic classification. Second, the authors showed the advantage of RNA-Seq in finding the second allele in the recessive disorder when WES only returned one pathogenic variant due to shallow coverage on the certain exons. Despite the continuous improvement of exome capture which could potentially overcome such issue, RNA-Seq has the ability to detect and assess the deep intronic pathogenic variants that are beyond the resolution of WES. Even though detectable by WGS, it exceeds our ability to interpret. Indeed, the authors next addressed deep intronic variants in muscular dystrophy, including three patients with either Duchenne muscular dystrophy or milder Becker muscular dystrophy that resulted from a pseudo-exon and disrupting the reading frame of the gene. RNA-Seq also pinpointed three structural variants in *DMD*, which were confirmed by WGS. RNA-Seq also successfully assessed a missense in *TTN* and a synonymous variant in *RYR1* creating novel splice sites and one synonymous variant in *POMGNT1* causing exon skipping involving three patients. Finally, a recurrent *de novo* deep intronic pathogenic variant creating a novel splice-site was also revealed in *COL6A1* in four patients who were completely negative after extensive diagnostic work-up, including WGS and clinical WES. RT-PCR confirmed the variant resulted in a cryptic splice donor that retained a pseudo-exon of 72 bp and led to an in-frame insertion of 24 amino acids disrupting the conserved triple-helical collagenous G-X-Y repeats of COL6A1. Accordingly, this intronic variant explained 27 additional diagnoses from a larger cohort with genetically undiagnosed collagen VI–like dystrophy.

As discussed by the authors, pathogenic variants altering gene expression levels and/or resulting in ASE are of fundamental importance in disease pathogenesis and important to uncover. Whether or not RNA-Seq becomes a front-line approach in rare disease diagnosis remains unknown. However, the authors have presented a compelling case for the efficiency of RNA-Seq in identifying pathogenic variants that are disease causing and often missed by WES/WGS, using a large sample set (10).

Acknowledgements

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Ankala A, da Silva C, Gualandi F, et al. A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield. Ann Neurol 2015;77:206-14.
2. Chong JX, Buckingham KJ, Jhangiani SN, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet

2015;97:199-215.
3. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet 2015;47:717-26.
4. Yang Y, Muzny DM, Xia F, et al. Molecular findings among patients referred for clinical whole-exome sequencing. JAMA 2014;312:1870-9.
5. Bhoj EJ, Li D, Harr M, et al. Mutations in TBCK, Encoding TBC1-Domain-Containing Kinase, Lead to a Recognizable Syndrome of Intellectual Disability and Hypotonia. Am J Hum Genet 2016;98:782-8.
6. Wenger AM, Guturu H, Bernstein JA, et al. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med 2017;19:209-14.
7. Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. N Engl J Med 2014;370:2418-25.
8. Dewey FE, Grove ME, Pan C, et al. Clinical interpretation and implications of whole-genome sequencing. JAMA 2014;311:1035-45.
9. Spies N, Weng Z, Bishara A, et al. Genome-wide reconstruction of complex structural variants using read clouds. Nat Methods 2017;14:915-20.
10. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med 2017;9.
11. Kremer LS, Bader DM, Mertes C, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat Commun 2017;8:15824.
12. Li G, Bahn JH, Lee JH, et al. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. Nucleic Acids Res 2012;40:e104.
13. Massah S, Beischlag TV, Prefontaine GG. Epigenetic events regulating monoallelic gene expression. Crit Rev Biochem Mol Biol 2015;50:337-58.
14. Rivas MA, Pirinen M, Conrad DF, et al. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 2015;348:666-9.
15. Lorson CL, Hahnen E, Androphy EJ, et al. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. Proc Natl Acad Sci U S A 1999;96:6307-11.
16. Roberts RG, Bobrow M, Bentley DR. Point mutations in the dystrophin gene. Proc Natl Acad Sci U S A 1992;89:2331-5.
17. Zeng Y, Wang G, Yang E, et al. Aberrant gene expression in humans. PLoS Genet 2015;11:e1004942.
18. Zhao J, Akinsanmi I, Arafat D, et al. A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. Am J Hum Genet 2016;98:299-309.
19. Pickrell JK, Marioni JC, Pai AA, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 2010;464:768-72.
20. Jain M, Burrage LC, Rosenfeld JA, et al. The incorporation of whole blood and fibroblast RNAseq with whole exome sequencing implicates LZTR1 in a novel syndrome with features of rasopathy and mitochondrial dysfunction. Presented at the 66th Annual Meeting of The American Society of Human Genetics. October 20, 2016. Vancouver, BC, Canada.

Cite this article as: Li D, Tian L, Hakonarson H. Increasing diagnostic yield by RNA-Sequencing in rare disease—bypass hurdles of interpreting intronic or splice-altering variants. Ann Transl Med 2018;6(7):126. doi: 10.21037/atm.2018.01.14