Category: News & Views

Title: Deciphering DNA variant-associated aberrant splicing with the aid of RNA sequencing

Bin Zhang[1,2] & Xin Gao[1,2]

1Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

2KAUST Computational Bioscience Research Center, King Abdullah University of Science and Technology

* Correspondence should be addressed to Xin Gao

Phone: +966-12-8080323

Email: Xin Gao, xin.gao@kaust.edu.sa

Aberrant RNA splicing resulting from DNA variations are common causes of genetic disorders. Two studies published in *Nature Genetics* independently contribute to decipher the DNA variant-associated aberrant splicing with high-throughput RNA sequencing data.

Main Text

Genetic disorders arise from variants of DNA that impair functions of single or multiple genes. In addition to the DNA variants directly impact coding sequence, a large proportion of variants alter RNA isoform structure with changes in exons or introns (Fig. 1), by affecting RNA splicing[1]. These aberrant splicing can lead to open reading frame changes or even shifts, resulting in loss of functional protein production or alterations of functional domains. Diagnosis of genetic disorders usually depend on the identification of mutations in genomic DNA, while aberrant RNA products caused by the mutations can also serve as additional supporting evidence[2]. However, RNA-based diagnostics remain limited in use for a variety of reasons. One major challenge is the tissue specificity of RNA expression, for which pathogenic RNA might be absent or too lowly expressed to be detected in clinical accessible tissues (CAT). Usually, tissues that can be sampled for clinical diagnostics are limited to the easily accessible ones, most commonly skin fibroblast and lymphoid cells from whole body. Another factor limiting the RNA diagnostics is the complex transcript structures in terms of different exon combinations and the varied location of exon/intron boundaries. To predict outcomes of DNA variants on RNA splicing, the current state-of-the-art methods are based on deep-learning models trained by a large set of primary DNA/RNA sequences[3]. For instance,

SpliceAI could predict scores for activating/deactivating splice sites, and loss/gain of exons with the pre-mRNA sequence, whereas MMSplice combines scores of exons, introns and splice sites based on neural networks to predict impacts of mutations on splicing, such as exon skipping, splice site choice, splicing efficiency, and pathogenicity[4,5]. However, their performance is far from satisfactory due to that the training dataset do not fully represent the complex, functional consequences of DNA variation on RNA splicing. Now two independent papers in *Nature Genetics* provide solutions to address these limitations[6,7], which developed much accurate methods for predicting impacts of DNA variants on RNA splicing with the incorporation of high-throughput RNA sequencing (RNA-seq) data.

There are over 20,000 protein-coding genes in the human genome but only a proportion of them are expressed in each tissue[8]. In addition to the expression, pervasive variations of RNA splicing, including switch-like alternative spliced exons, have also been observed across different human tissues[9]. To address this limitation, Wagner et al first comprehensively identified unannotated splicing events utilizing FRASER[10], an aberrant splicing caller, on 16,213 RNA-seq samples of 49 human tissues or cell lines in the Genotype-Tissue expression (GTEx) project[11]. By integrating with the matched genomes, they further characterized the aberrant splicing events likely arise from DNA variants (rare variants within 250 base pairs away from intron/exon boundaries) and created SpliceMap, a map of tissue-specific aberrant splice sites. By leveraging predictions from SpliceAI and MMSplice, and incorporating with SpliceMap, they developed Absplice-DNA, which is a generalized additive model utilizing information from these three sources. Comparing to SpliceAI and MMSplice with the capped 12% precision at 20% recall, Absplice-DNA achieved 2.5-3-folds improvement in precision. Moreover, with the available RNA-seq data from CAT, they further trained models by integrating Absplice-DNA features together with features from RNA-seq data, including the gene-level p-values from FRASER and the estimated differential splicing amplitudes. They named these models as Absplice-RNA. Taking fibroblast as an example, with RNA-seq data from an independent dataset, they demonstrated that Absplice-RNA has the highest precision as 60% at 20% recall.

Using a complementary approach to the machine-learning-based methods, Dawes *et al* proposed a data-derived evidence-based methods according to their empirical observation that aberrant splicing junctions caused by genetic mutations could also be detected in control RNA-seq samples from healthy individuals. With the annotation from RefSeq and Ensembl, the authors collated all unannotated splicing events surrounding each potential splice site by extracting splicing junctions from 335,301 published RNA-seq samples (300K-RNA) from GEO and GTEx using Monorail analysis pipeline from recount3[12]. For each splice site, they ranked all unannotated splicing events based on the number of samples, in which the event is detected, and further used the four most common ones (300K-RNA Top4) to nominate potential

aberrant splicing caused by genetic variants. To illustrate the power of this evidence-based heuristic, they could successfully identify 96% of exon-skipping events and 82% of cryptic splice-sites induced by 76 clinically variants reported in a standardized diagnostic study[13]. They found that 300K-RNA Top4 outperforms SpliceAI (10% increase in sensitivity) to correctly identify aberrant splicing events for these 76 variants. Finally, they developed SpliceVault, a web portal to retrieve unannotated splicing events in 300K-RNA Top4, which covered each splice site from protein-coding genes in the annotation. They proposed that the SpliceVault could aid in interpreting clinical RNA diagnostic tests, as will be prospectively trialed by the Australasian Consortium for RNA Diagnostics (SpliceACORD)[13].

The two papers demonstrate the feasibility of utilizing aberrant RNA to facilitate genetic diagnostics for disorders caused by splicing defects. Together with several previous studies[2,13,14], these results implicate a new direction of genetic diagnostics, which is known as RNA-based diagnostics. In addition, these two studies could be also helpful for functional interpretation of DNA variants in non-coding regions identified by genome-wide association studies (GWAS) as it can predict the outcome of these variants on splicing. Compared with the current state-of-the-art methods, these two exhibit higher precision and sensitivity. Notably, unlike the deep-learning based methods that requires significant computational capacities to implement, SpliceVault is very easy and convenient to use. Finally, the approach for the detection of aberrant RNA in these two papers is conventional RNA-seq, which is the second generation high-throughput sequencing technology that captures short fragments of RNA molecules transcriptome-wide. It can effectively identify short sequences around aberrant splicing junctions (Fig. 1), but it remains challenging to reconstruct the full-length of aberrant RNA transcripts, which might lead to misinterpretation of the outcome. On the other hand, the third-generation sequencing, such as Oxford Nanopore Technology (ONT), can sequence long fragments or even the full-length DNA or RNA molecules with a much faster speed. Just recently, ONT data has been applied to investigate splicing variations associated with different DNA alleles in human tissues from the GTEx[15]. This indicate that long read sequencing can be a promising approach to predict the full-length of RNA isoform induced by DNA variants and might facilitate RNA-based diagnostics in the future (Fig. 1).

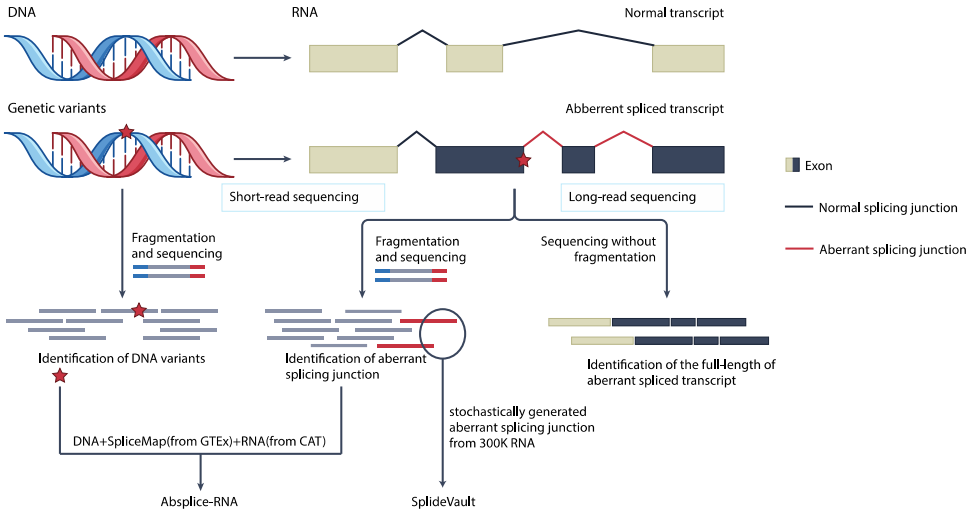

Figure 1: A schematic illustration of the genetic diagnostics based on the identification of DNA variants, aberrant splicing junctions and the full-length of aberrant spliced transcripts. The stars indicate the genetic variant that cause alteration of RNA isoforms by affecting splicing. Absplice-RNA is addictive model integrating DNA variants, reference SpliceMap derived from multiple tissues of the GTEx RNA-seq data and RNA-seq data from CAT. SpliceVault is an evidence-based methods by collecting non-variants-induced, stochastically generated aberrant splicing junctions from 300K-RNA of healthy individuals to predict the most likely splicing outcome caused by the variants.

Competing interests

The authors declare no competing interests.

Reference

1 Baralle, D. & Buratti, E. RNA splicing in human disease and in the clinic. *Clin Sci (Lond)* **131**, 355-368, doi:10.1042/CS20160211 (2017).

2 Maddirevula, S. *et al.* Analysis of transcript-deleterious variants in Mendelian disorders: implications for RNA-based diagnostics. *Genome biology* **21**, 1-21 (2020).

3 Li, Z. *et al.* Applications of deep learning in understanding gene regulation. *Cell Reports Methods* **3**, 100384, doi:https://doi.org/10.1016/j.crmeth.2022.100384 (2023).

4 Cheng, J. *et al.* MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol* **20**, 48, doi:10.1186/s13059-019-1653-z (2019).

5 Jaganathan, K. *et al.* Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535-548. e524 (2019).
6 Celik, M. H. *et al.* Aberrant splicing prediction across human tissues. *Nat Genet* (2023).
7 Dawes, R. *et al.* SpliceVault: predicting the precise nature of variant-associated mis-splicing. *Nat Genet* (2023).
8 Mele, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660-665, doi:10.1126/science.aaa0355 (2015).
9 Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476, doi:10.1038/nature07509 (2008).
10 Mertes, C. *et al.* Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat Commun* **12**, 529, doi:10.1038/s41467-020-20573-7 (2021).
11 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013).
12 Wilks, C. *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol* **22**, 323, doi:10.1186/s13059-021-02533-6 (2021).
13 Bournazos, A. M. *et al.* Standardized practices for RNA diagnostics using clinically accessible specimens reclassifies 75% of putative splicing variants. *Genet Med* **24**, 130-145, doi:10.1016/j.gim.2021.09.001 (2022).
14 Fresard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med* **25**, 911-919, doi:10.1038/s41591-019-0457-8 (2019).
15 Glinos, D. A. *et al.* Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**, 353-359, doi:10.1038/s41586-022-05035-y (2022).