RNA BIOLOGY
2021, VOL. 18, NO. 3, 382–390
https://doi.org/10.1080/15476286.2020.1809195

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS

Check for updates

In silico identification of pseudo-exon activation events in personal genome and transcriptome data

Narumi Sakaguchi and Mikita Suyama

Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan

ABSTRACT

Causative mutations for human genetic disorders have mainly been identified in exonic regions that code for amino acid sequences. Recently, however, it has been reported that mutations in deep intronic regions can also cause certain human genetic disorders by creating novel splice sites, leading to pseudo-exon activation. To investigate how frequently pseudo-exon activation events occur in normal individuals, we conducted in silico identification of such events using personal genome data and corresponding high-quality transcriptome data. With rather stringent conditions, on average, 2.6 pseudo-exon activation events per individual were identified. More pseudo-exon activation events were found in 5′ donor splice sites than in 3′ acceptor splice sites. Although pseudo-exon activation events have sporadically been reported as causative mutations in genetic disorders, it is revealed in this study that such events can be observed in normal individuals at a certain frequency. We estimate that human genomes typically contain on average at least 10 pseudo-exon activation events. The actual number should be higher than this, because we used stringent criteria to identify pseudo-exon activation events. This suggests that it is worth considering the possibility of pseudo-exon activation when searching for causative mutations of genetic disorders if candidate mutations are not identified in coding regions or RNA splice sites.

ARTICLE HISTORY
Received 11 June 2020
Revised 3 August 2020
Accepted 8 August 2020

KEYWORDS
Pseudo-exon activation; RNA splicing; mutation; RNA-Seq; personal genome

Introduction

Causative mutations for genetic disorders have mainly been identified in exons, especially in coding sequences, and in RNA splice sites at both ends of introns. This is because these regions are functionally important and, hence, conserved over the course of evolution. If a mutation occurs in these regions, the transcript or its protein product may be disrupted, often leading to a disease phenotype. Exome sequencing is a method that efficiently detects mutations in exonic regions and their flanking RNA splice sites [1], and this has been successfully applied to identify causative mutations for a variety of genetic disorders (for a review, see [2]).

Compared with mutations in coding regions and splice sites, mutations in deep intronic regions have not been a target for thorough analysis of disease-causing mutations, mainly for two reasons. First, such regions are usually not conserved among species and hence not thought to be functionally important. Second, conventional exome sequencing can only identify mutations in exons and flanking intronic sequences, which contain splice sites. It is thus intrinsically not possible to identify mutations that occur in deep intronic regions. However, growing evidence shows that mutations in deep intronic regions often create novel RNA splice sites that can trigger pseudo-exon activation, which in turn leads to aberrant transcripts containing an extra exon (e.g. [3–6],). Such mutations may disrupt reading frames, often introducing premature termination codons (PTCs) and, therefore, can be causative of a disease phenotype (for a review, see [7]).

As described above, since pseudo-exon activation events are only sporadically reported as causative mutations in genetic disorders, the genome-wide frequency and characteristics of such events in the human genome have yet to be analysed. With the advancement of personal genome sequencing [8] and transcriptome analysis for the corresponding individuals [9], it is now possible to identify pseudo-exon activation events on a genome-wide scale in individuals. In this study, to determine how frequently pseudo-exon activation events occur in normal individuals, we conducted in silico identification of such events in normal individuals by using publicly available personal genome data and transcriptome data for the corresponding individuals, especially focusing on those single nucleotide variants (SNVs) that create novel splice sites.

Materials and methods

Genomic variants and transcriptome data

We used genomic variant data of individuals as determined by the 1000 Genomes Project [8] in variant call format (VCF). They have already been registered as SNPs in dbSNP. Transcriptome data of lymphoblastoid cell lines for the

CONTACT Mikita Suyama mikita@bioreg.kyushu-u.ac.jp Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, Fukuoka 812-8582, Japan

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

corresponding individuals were obtained from the GEUVADIS project [9]. Initially, we downloaded the variant data for 462 individuals and their corresponding transcriptome data. We then evaluated the quality of the sequencing reads using the FastQC program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc); only those high-quality RNA-Seq data with a sequencing quality score greater than 30 were used. After this quality filtering, transcriptome data for 235 individuals remained.

Personal genome sequence construction and RNA-Seq read mapping

The reference genome sequence (hg19) was downloaded from the UCSC Genome Browser [10]. BCFtools (version 1.9) [11] was applied to the reference genome sequence and to the variant information for each individual in VCF format to construct individual-specific reference genome sequences. The HISAT2 program (version 2.1.0) [12] was used to map the RNA-Seq data onto the genomic sequences that reflect the variant information for the individual. The gene structures registered in the RefSeq data [13] in GTF format were used as reference transcriptome data. RefSeq data were downloaded from the UCSC Genome Browser [10]. The total numbers of reference transcripts and genes were 50,643 and 26,242, respectively. Default parameters were used for mapping. Gene expression levels were quantified using the StringTie program (version 1.3.5) [14].

Data visualization and manipulation

The mapping data, together with the data on genomic variants, were visualized using the Integrative Genomics Viewer (IGV) software (version 2.4.16) [15]. SAMtools (version 1.5) [16] and BEDTools (version 2.26.0) [17] were used for data manipulation.

Splice site scoring

The strength of splice sites was evaluated using the MaxEntScan program [18]. For 5′ss, genome sequence segments corresponding to the three bases at the end of pseudo-exons and the six bases at the start of introns were subjected to MaxEntScan. For 3′ss, genome sequence segments corresponding to the 20 bases at the end of introns and the 3 bases at the start of pseudo-exons were also subjected to MaxEntScan.

Analysis of protein domain architecture

For the cases in which the inclusion of a pseudo-exon introduce neither a frameshift nor an in-frame stop codon, the protein sequences of the wild-type gene were examined to determine whether pseudo-exon inclusion can disrupt any existing protein domains. For this, we used the SMART database [19].

Calculation of exon inclusion ratio

Exon inclusion ratio, or 'percent spliced in' (PSI), can be calculated using the following equation [20]:

$$\mathrm{PSI} = (Ji/2)/(Ji/2 + Js)$$

where *J*i is the number of inclusion junction reads consisting of reads mapped to upstream and downstream splice junctions of the pseudo-exon and *J*s is the number of skipping junction reads mapped to the junction that skips the pseudo-exon. The reason why *J*i is divided by 2 is that there are two junctions for an exon inclusion. Values of PSI can range from 0 (completely skipped) to 1 (complete inclusion).

Analysis of splicing regulatory elements in the flanking regions of the identified pseudo-exons

We used the Human Splicing Finder [21] to examine whether there are other SNVs affecting the splicing regulatory elements within 1000 bp of the identified pseudo-exons with lower MaxEntScan scores. For the input, the reference genome sequence (hg19) downloaded from the UCSC Genome Browser [10] were used as the 'reference sequence' and the personal genome data as the 'mutant sequence.'

Enrichment analysis of functional categories of genes

Enrichment analysis of functional categories of genes was performed using Metascape [22]. The following settings were used: 'Input as species' was set to '*H. sapiens*,' and 'Analysis as species' was set to '*H. sapiens*.'

Results

Identification of pseudo-exon activation events

Before mapping the RNA-Seq data of each individual onto the genome sequence, we prepared a reference genome that reflected the variant information for the corresponding individual using BCFtools (version 1.9) (Fig. 1A). By creating such a reference genome, variants that affect splicing can be identified [23]. This process is necessary because pseudo-exons are thought to be activated by SNVs at splice sites, which is not observed in the commonly used reference genome. If the reference genome does not contain SNVs at splice sites, the mapping program for RNA-Seq data might fail to correctly map junction reads because canonical sequence motifs for splice sites are not detected in the reference genome sequence without reflecting the information concerning variants for that individual.

Pseudo-exon activation events were identified through the following steps (Fig. 1B). First, we collected junction reads that were mapped onto annotated exons in RefSeq transcripts on one side, and the remainder were mapped directly onto the region in the reference genome that were not covered by any annotated exons. For each junction identified in the above step, only those junctions that were covered by two or more junction reads were further selected as pseudo-exon junction candidates. Among these pseudo-exon junction candidates, those pairs of junction candidates that matched their order

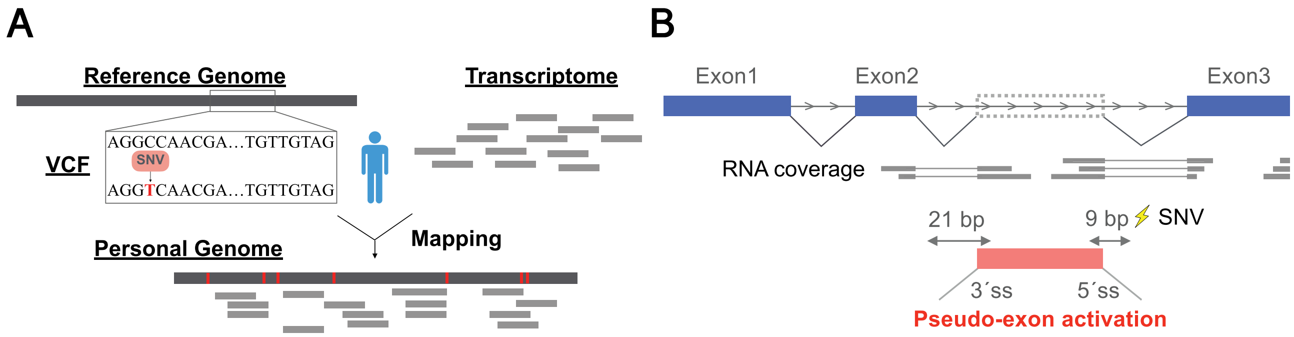

Figure 1. Workflow for the identification of pseudo-exon activation events. (A) RNA-Seq data for an individual is mapped against the reference genome that reflects variant information for the corresponding individual. (B) Identification of pseudo-exon activation events based on junction reads that are not mapped to annotated exons (see main text for further details).

to the direction of transcription and with pseudo-exon lengths of ≤1,000 bp were retained as pseudo-exon candidates, having novel splice sites at both 5′- and 3′-ends. Next, we selected only those pseudo-exon candidates that had an SNV either at the flanking region of the 5′ donor splice site (5′ss) or at the flanking region of the 3′ acceptor splice site (3′ss) as pseudo-exon activation events. Here, we considered three bases at the end of the pseudo-exon and six bases at the start of the intron as the flanking region of 5′ss, and 18 bases at the end of the intron and three bases at the start of the pseudo-exon as the flanking region of 3′ss. For 3′ss, we accepted longer intronic segments than those of 5′ss because 3′ss has a characteristic

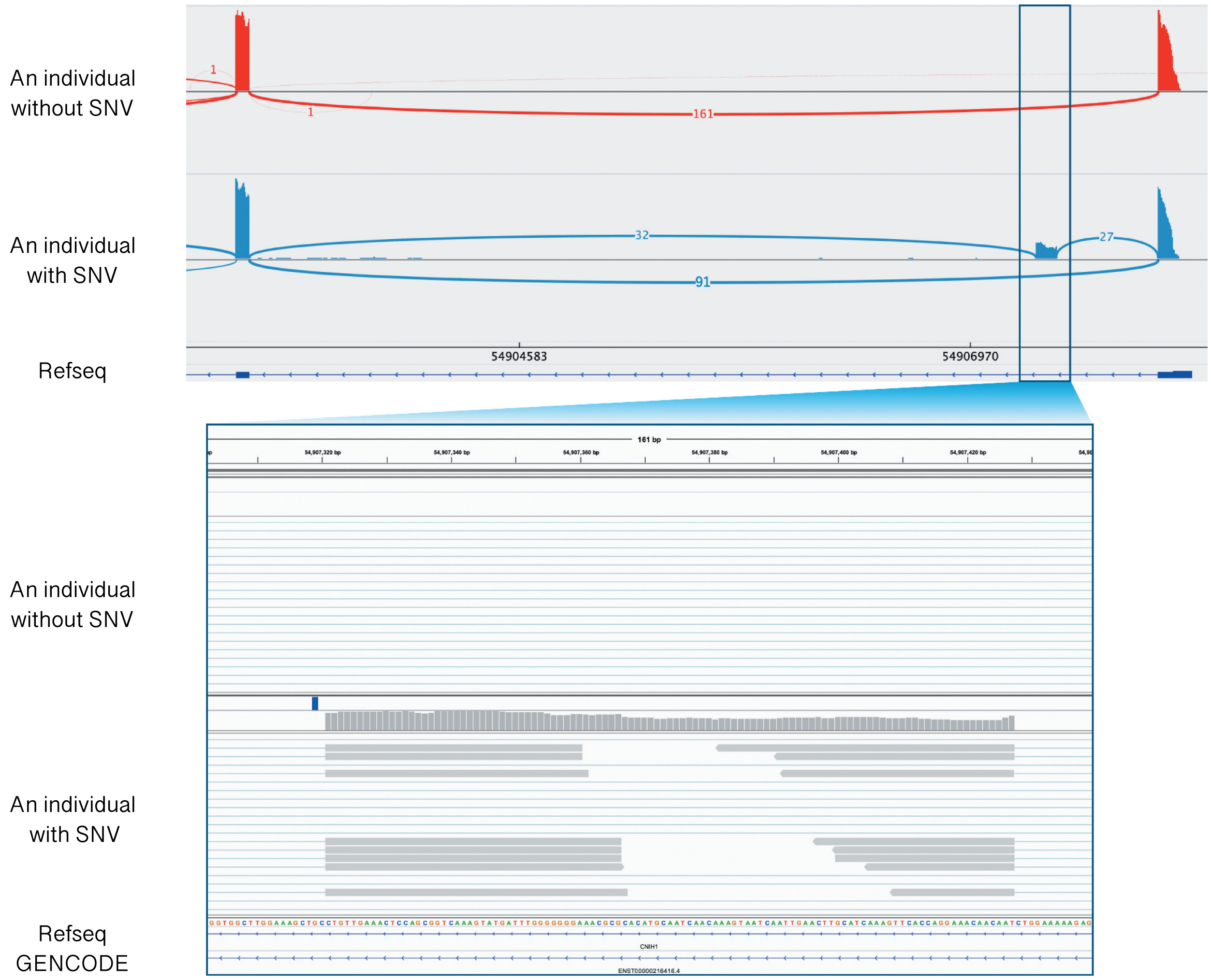

Figure 2. An example of a pseudo-exon activation event identified in the first intron of cornichon family AMPA receptor auxiliary protein 1 (*CNIH1*) by an SNV that creates a novel 5′ss. The upper panel illustrates a Sashimi plot [25] of the activated pseudo-exon and neighbouring exons. Each number represents the number of exon-exon junction reads. The bottom half shows an individual (HG00238) with the SNV and activated pseudo-exon. The top half shows an individual (HG00116) without the SNV for comparison. The lower panel shows a close-up view of the activated pseudo-exon.

polypyrimidine tract at the end of the intron. Finally, we examined the correspondence between the presence of SNVs and pseudo-exons for all of the individuals analysed in this study. We conducted this analysis because cases can occur in which individuals harbour the same candidate SNV without pseudo-exon activation and vice versa. This happens because, for some pseudo-exons, the expression levels are around the borderline of the aforementioned detection criteria, and we are not able to detect pseudo-exons for some individuals with low or no expression. To deal with this possibility, we introduced a condition that, in the final set, a candidate pseudo-exon was kept only if the number of individuals having both the SNV and the pseudo-exon was larger than the number of those having the same SNV but not having the pseudo-exon. On the other hand, a candidate pseudo-exon was rejected from the final set if there was an individual without the SNV having the candidate. This is because, in such a case, it is clear that the pseudo-exon activation is not caused by that SNV.

From the 235 individuals analysed in this study, we identified 116 distinct pseudo-exons. On average, there were 2.6 pseudo-exons per individual.

Examples of pseudo-exon activation events

As an example, we here show a pseudo-exon identified in the first intron of cornichon family AMPA receptor auxiliary protein 1 (*CNIH1*) (Fig. 2). The length of the wild-type intron was measured as 4,807 bases, and the pseudo-exon was found 537 bases from the 3′ terminal of the upstream exon. The suspected causative SNV is a C-to-T transition at the second base of the canonical dinucleotide at the 5′ss. In the mutated sequence with the canonical dinucleotide, MaxEntScan score, which calculates the strength of a potential splice site based on the sequence, was determined to be 6.52 at the 5′ss, whereas the score for the reference genome sequence at the corresponding regions was −1.23, suggesting that the mutated sequence has much higher potential to be a 5′ss than the reference sequence at the corresponding region. At the 3′ss of the pseudo-exon, the MaxEntScan value was rather high, 12.02, indicating that the sequence pattern can be taken as a cryptic splice site even without SNVs. The flanking exons of the pseudo-exon are coding exons; therefore, the newly incorporated exonic sequence might be translated. The length of the pseudo-exon is 107 bases, which is not a multiple of 3, indicating that the inclusion of the pseudo-exon disrupts the coding potential of the downstream exons. Moreover, the pseudo-exon itself introduces a premature termination codon (PTC). The PSI value, which represents the exon inclusion ratio (see Materials and Methods), for the pseudo-exon was calculated to be 0.24. Since the maximum PSI value for the heterozygous SNV is assumed to be 0.5, the deviation from that value can be attributed either to the strength of the splice sites of the pseudo-exon or to nonsense-mediated mRNA decay (NMD), which selectively degrades transcripts having a PTC [24], or both. In addition, although the pseudo-exon contains ATG, it is out-of-frame and is unlikely to be a start codon for alternative translation initiation.

Another example is a pseudo-exon identified in the seventh intron of cysteinyl-tRNA synthetase 2 (*CARS2*) (Fig. 3). In this case, the suspected causative SNV was a C-to-T transition in the polypyrimidine tract at position −13 of the 3′ ss. The length of the wild-type intron is 9,500 bases, and the pseudo-exon was found 8,775 bases from the upstream exon. The MaxEntScan scores for the mutated sequence and the reference sequence at the corresponding regions were 9.13 and 8.34, respectively. The length of the pseudo-exon is 78, which is a multiple of 3, and it does not disrupt the original reading frame. Moreover, the sequence of the pseudo-exon itself does not contain any PTCs, suggesting that this would not trigger NMD. The analysis of protein domain architecture using the SMART database [19] shows that the pseudo-exon is inserted in between the annotated protein domains. The PSI value for the pseudo-exon was calculated to be 0.14. In this case, the deviation from the maximum PSI value for the heterozygous SNV can be attributed to the strength of the splice sites of the pseudo-exon. Out of the 235 individuals analysed, there were 2 individuals who shared the same SNV and activation of the pseudo-exon.

Systematic analysis of the identified pseudo-exon activation events

Of the 235 individuals analysed in this study, the maximum number of events found in an individual was 7; this was observed in two individuals (Fig. 4A). We did not identify any pseudo-exon activation events in 18 individuals. There were pseudo-exons shared among multiple individuals. In such cases, they also shared the same variations that are thought to be causative ones, as we applied such a condition in identifying pseudo-exons. The maximum number of individuals that shared the same pseudo-exon was determined to be 92, although most of the pseudo-exons were observed in only a single individual (Fig. 4B). By counting these pseudo-exons that are shared by multiple individuals as a single case, the number of distinct pseudo-exons in terms of genomic loci was 116 (Supplemental Table S1). For each of these pseudo-exons, we calculated the PSI value and also counted the number of individuals in terms of whether they had homozygous or heterozygous SNVs. Although the PSI value can be a quantitative index to evaluate the degree of the inclusion of the pseudo-exon, it may also deviate markedly from the true expression ratio of the transcript isoforms, if the read depth, that is, the expression level, is rather low.

We then analysed the length distribution of the pseudo-exons and compared this with that of exons (Fig. 4C). The modal value was the same between pseudo-exons and exons (81–120 bases); however, the average length was found to be significantly longer for exons, that is, 132.3 bases (pseudo-exons) vs. 258.2 bases (exons) (Student's t-test, $p < 0.001$), mainly because a certain fraction of exons are much longer than the other exons, and also because we set our cut-off for pseudo-exon length at ≤1,000 bp. This trend is consistent with a previous report, which compiled approximately 81 cases of disease-causing pseudo-exon activation events from the literature [7].

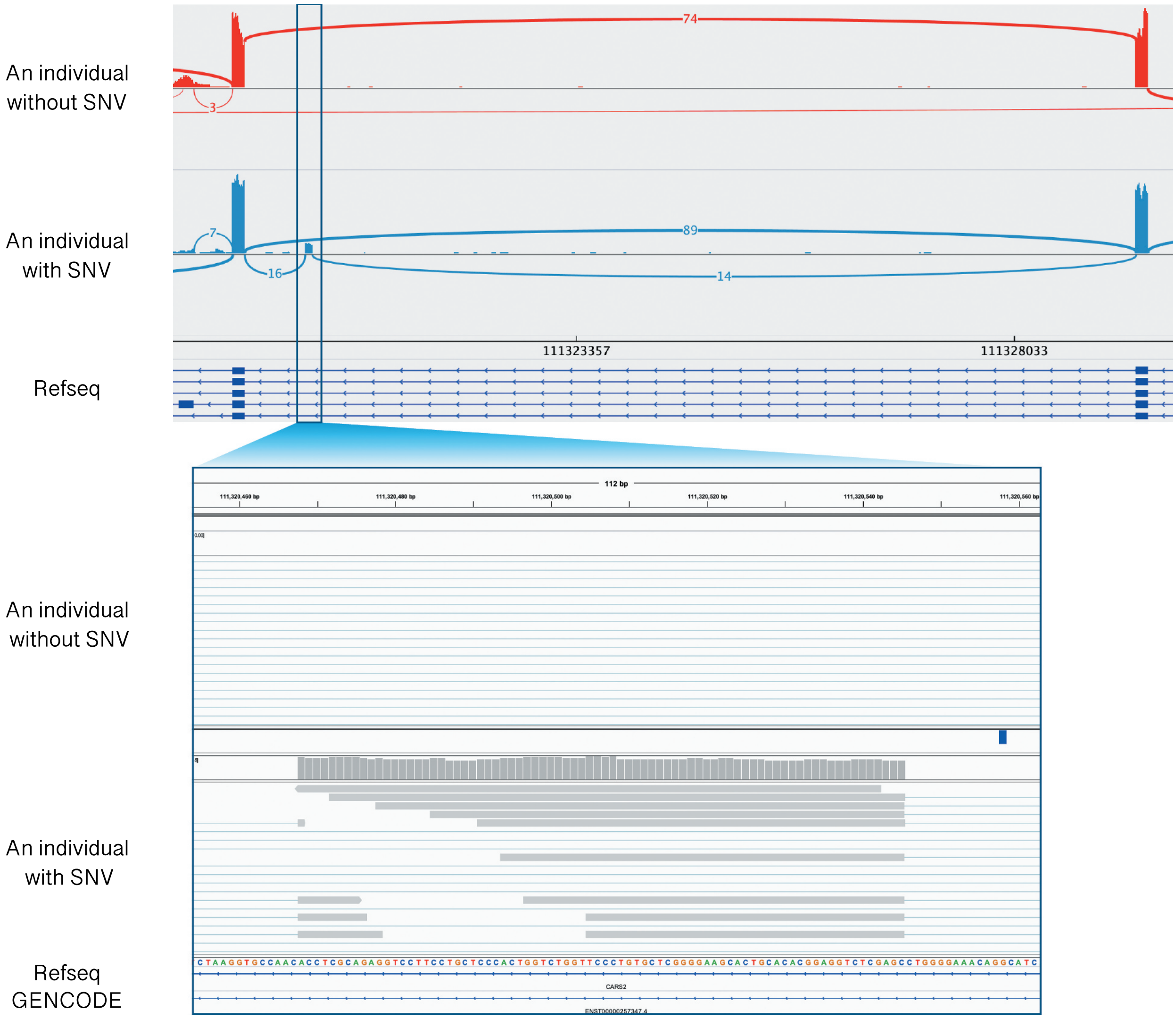

Figure 3. An example of a pseudo-exon activation event identified in the seventh intron of cysteinyl-tRNA synthetase 2 (*CARS2*) by an SNV that creates a novel 3′ss. The upper panel illustrates a Sashimi plot [25] of the activated pseudo-exon and neighbouring exons. Each number represents the number of exon-exon junction reads. The bottom half shows an individual (HG00336) with the SNV and activated pseudo-exon. The top half shows an individual (HG00178) without the SNV for comparison. The lower panel shows a close-up view of the activated pseudo-exon.

To evaluate the effects of the SNVs that are thought to be a cause of pseudo-exon activation on the strength of the splice site, we used the MaxEntScan program [18], which quantitatively assesses whether a local sequence segment has the potential to be a splice site, to the wild-type sequence segment and to the mutated one. Among the 116 distinct pseudo-exon activation events, 110 (94.8%) of them showed gains in scores of more than 1.0 in the sequences with SNVs (Fig. 4D). On the other hand, there were six instances that showed little or no gains in scores in the sequences with SNVs. We found that these cases already have rather high MaxEntScan scores even in the wild-type sequences. For these six instances, we further applied the Human Splicing Finder [21] to identify other possible SNVs that might affect the splicing regulatory element within the pseudo-exons or in the regions flanking them. However, there were no significant SNVs in those regions, indicating that the SNVs identified at the splice sites of the pseudo-exons would be a cause of the pseudo-exon activation events even though there were little or no gains in the MaxEntScan scores.

To evaluate the effect of the inclusion of the pseudo-exons on the transcripts in terms of the coding potential, we analysed whether each identified pseudo-exon might disrupt the original reading frame and induce NMD. Among the 116 pseudo-exons that we identified, 83 were located in the coding regions. Of these, 65 pseudo-exons were expected to induce NMD either by in-frame termination codons in the pseudo-exons themselves (58 cases) or by frameshifts that create PTCs in their downstream region (7 cases). The remaining 18 pseudo-exon activation events may maintain the open reading frame and do not seem to trigger NMD (Supplemental Table S1). For these 18 pseudo-exon activation events, their protein sequences using the SMART database were analysed [19] to check the insert positions of the pseudo-exons in their protein domain architecture. We found that, in three instances, the insertions of the pseudo-exons occur within protein domains, indicating severe disruption of the domains (Supplemental Table S1 and Supplemental Fig. S1).

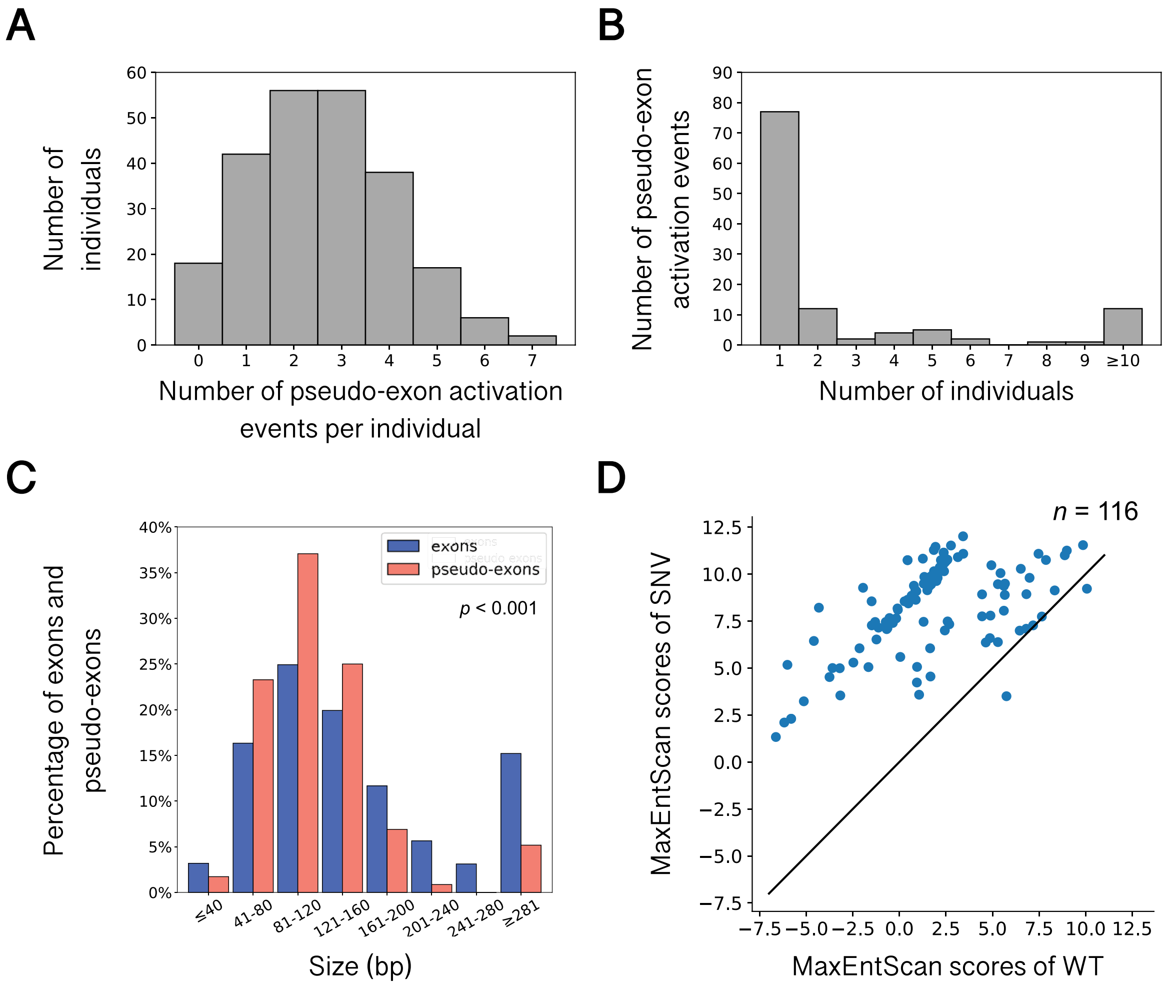

Figure 4. Basic characteristics of the identified pseudo-exon activation events. (A) Histogram of the number of pseudo-exon activation events per individual. (B) Histogram of the number of pseudo-exons shared among multiple individuals. (C) Length distribution of exons (blue) and pseudo-exons (red). (D) Scatter plot of the strength of the splice site before and after the SNV for each of the 116 pseudo-exons. The x- and y-axes indicate the MaxEntScan scores for the wild-type and mutated sequences, respectively. The diagonal line shows equal scores between the wild-type and mutated sequences.

Additionally, for the 116 identified pseudo-exon activation events, we also performed Gene Ontology analysis using Metascape [22]. The results showed that there were no specifically enriched functional categories for the genes having pseudo-exons.

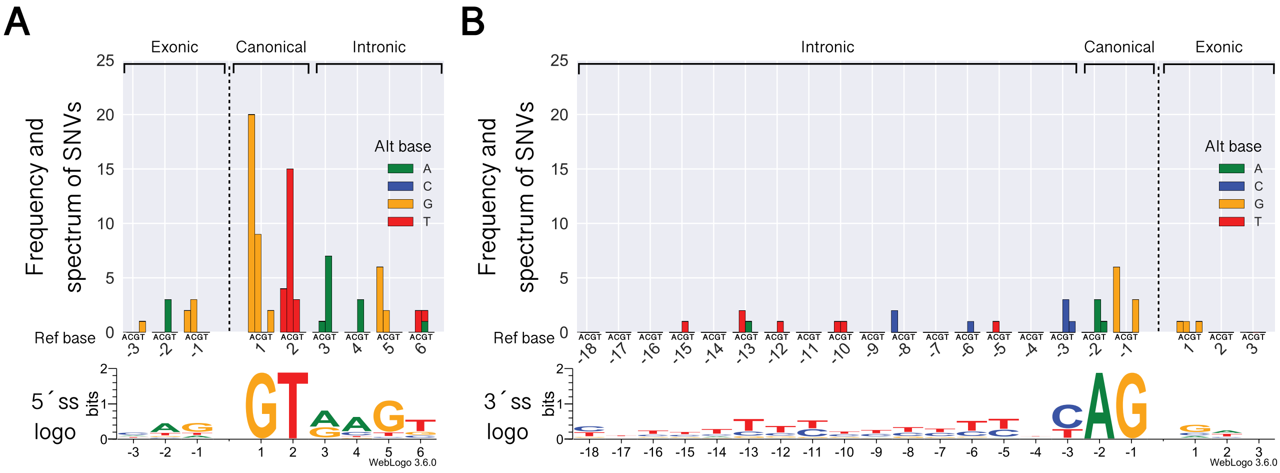

Figure 5. Frequency and spectrum of SNVs involved in pseudo-exon activation events. (A) 5′ss. (B) 3′ss. The dotted vertical line represents the intron-exon border. The colour codes for alternative bases are shown on the right side of each panel. The base frequency data for splice sites, which is represented as a sequence logo, is taken from WebLog 3 [26].

Frequency and spectrum of SNVs for each site

For the 116 distinct pseudo-exon activation events, we summarized the type and frequency of SNVs (Fig. 5). The total number of SNVs that we identified was approximately 2.7 times higher in 5′ss (85 SNVs) than in 3′ss (31 SNVs), even though the lengths that we analysed were longer in 3′ss because of the existence of polypyrimidine tracts. As expected, most of the SNVs were observed in the canonical dinucleotides in both 5′ss (Fig. 5A) and 3′ss (Fig. 5B). More specifically, in 5′ss, 62.4% of the SNVs were observed in the canonical GU dinucleotides, and all of these SNVs created the dinucleotide. In 3′ss, 41.9% of the SNVs were observed in the canonical AG dinucleotides, and all of these SNVs created the dinucleotide. Most of the SNVs were substitutions towards splice site motifs. For example, in dinucleotides at both 5′ss and 3′ss, all of the SNVs were changes towards the bases in the canonical dinucleotides.

Discussion

Pseudo-exon activation events have been mainly reported as a consequence of disease-causing mutations [7]. In this study, by using personal genome data [8] together with the RNA-Seq data for the peripheral blood samples of the corresponding individuals [9], we were able to successfully identify pseudo-exon activation events also in normal individuals. Such events often introduce PTCs in the downstream region of the mRNA, which trigger NMD [27]. Even if such a surveillance mechanism of irregular mRNA exists, we were still able to detect transcripts with pseudo-exons because NMD does not usually degrade all transcripts with a PTC. Indeed, it has been shown that a substantial number of the transcripts supposed to trigger NMD could still be detected in transcriptome data [9,28].

Although the involvement of the identified SNVs in pseudo-exon activation events is yet to be directly demonstrated, they seem to be causative SNVs because the mutation spectrum correlated well with the sequence motifs of splice sites (Fig. 5). It is interesting to note that the number of SNVs at 5′ss was found to be about 2.7 times higher than those at 3′ss. This trend corresponds well with the known cases of pseudo-exon activation events reported so far [7], which also shows that there are more pseudo-exons created by SNVs at 5′ss than those at 3′ss; this further confirms that most of the pseudo-exon activation events that we identified in this study are authentic. We can provide two possible explanations for the excess of SNVs at 5′ss. Firstly, the canonical AG dinucleotides at 3′ss can be created by SNV of CG dinucleotides, which are under-represented in vertebrate genomes [29]. Indeed, we did not observe any SNVs of CG to AG to create 3′ss (Fig. 5). Secondly, it is indicated that the recognition of 5′ss is a key step in RNA splicing [30] and the number of SNVs that disrupt splicing is also high at 5′ss compared with that at 3′ss [30,31].

If the causative SNVs for pseudo-exon activation reside in exonic regions, the events and the associated SNVs can possibly be identified solely by transcriptome analysis, namely, RNA-Seq, without carrying out WGS. The proportion of such SNVs, however, is rather low (10.3% of the total events identified in this study), mainly because the canonical dinucleotides at both ends of the splice sites are not in exonic regions but at the termini of introns. Moreover, without WGS, it is difficult to map RNA-Seq reads onto the reference genome; this is because the reference genome might not have the splice site for the pseudo-exon, and hence the mapping program of the RNA-Seq reads would fail to align the junction reads of the pseudo-exon. One possible solution for this difficulty might be to use *de novo* assembly of RNA-Seq reads to identify pseudo-exon activation solely from RNA-Seq data. Once the candidates of pseudo-exon activation are obtained, the causative SNV can be identified by PCR experiments at both sides of the exon-intron junctions of the pseudo-exon candidates.

Pseudo-exon activation caused by somatic variants in deep intronic regions was recently reported in cancer [32]. From 1,188 individuals, with an average number of somatic SNVs was 22,144, analysed in their study, they identified 46 distinct pseudo-exons. This corresponds to one pseudo-exon activation event in approximately 0.6 million SNVs. Interestingly, this frequency is similar to what we found in our study, that is one pseudo-exon activation event in approximately 1.4 million SNVs (average numbers of SNVs and pseudo-exons per individual are 3.7 million and 2.6, respectively). In addition, the fact that the SNVs that cause pseudo-exon activation in cancer were enriched in the canonical splice sites is well consistent with our results.

The actual number of pseudo-exon activation events might be higher than the number that we identified in this study for the following four reasons. First is that we adopted rather stringent criteria for the identification of pseudo-exon activation events. For example, we considered only those cases where both sides of the pseudo-exon were covered by at least two junction reads. We also examined the correspondence between the presence of the SNVs and pseudo-exons for all of the individuals. Second is that we only considered single-nucleotide variants but not insertions/deletions (indels). The third reason is that we did not take the SNVs creating splicing regulatory elements, such as exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs), into account. Although several papers have reported that pseudo-exons are activated by mutations in splicing regulatory elements [33–35], we did not consider them in the present study because those elements are not always located close to the pseudo-exons, and it is difficult to determine the correspondence between an SNV at a regulatory element and a pseudo-exon. In addition, we focused only on pseudo-exon activation events caused by SNVs in intronic regions, although SNVs in exonic regions can also cause pseudo-exon activation by creating not only an ESE and ESS but also a novel splice site, as has been reported in cancer [36]. The fourth reason is that we only used transcriptome data for peripheral blood samples. Because of this, what we identified as pseudo-exon activation events is limited to those genes actively transcribed in peripheral blood, and those genes that are not expressed in peripheral blood are not covered in the present study. The lowest expression level of the gene for which we identified a pseudo-exon activation event was an FPKM value of 0.53. This value corresponds to approximately the 43rd percentile from the top of all of the genes in the genome according to their expression levels. For the remaining genes (57% of all genes in the genome), we could not identify pseudo-exon activation events because of low expression levels in peripheral blood cells. From this proportion, we can estimate that at least two times more pseudo-exon activation events (i.e., on average 5.2 = 2.6 × 2 instances) might exist in normal individuals. Moreover, given that there are on average five splicing regulatory elements per exon [37], the number of pseudo-exon activation events can further be estimated as at least twice as much, that is, more than 10 per individual. Considering the fact that we have adopted rather stringent conditions in identifying pseudo-exon activation events, the actual number of pseudo-exon activation events in an individual is likely to be more than the above estimate.

Pseudo-exon activation events are thought to occur from variants in deep intronic regions. Such regions are often considered as usually under weak or no selective pressure because there may be no functional constraints. Variants in these regions are often overlooked as causative ones because they are thought to be benign and also because their number is relatively high compared with those under selective pressure. Moreover, variants in deep intronic regions cannot be detected by exome sequencing, which is often employed to identify causative mutations for genetic disorders because such regions are outside of the capture target for exome sequencing. It is reported that the success rate of exome sequencing is approximately 25%–40% [38,39]. For the remaining cases, in which causative mutations have not been identified yet, some might be caused by deep intronic mutations that trigger pseudo-exon activation. Our results suggest that it is worth considering the possible involvement of pseudo-exon activation events in identifying causative mutations of genetic disorders for which the responsible mutations have not yet been identified.

Acknowledgments

We thank Naoto Kubota for discussion and critical reading of the manuscript. This work was supported by JSPS Grants-in-Aid for Scientific Research (Grant number: 17H03619 to M.S.) and the "Advanced Computational Scientific Program" of the Research Institute for Information Technology, Kyushu University. The results shown here are partly based on data generated by the 1000 Genomes Project and the GEUVADIS Project.

Disclosure of potential conflicts of interest

The authors report no conflicts of interest.

Funding

This work was supported by the JSPS Grants-in-Aid for Scientific Research [17H03619].

ORCID

Narumi Sakaguchi http://orcid.org/0000-0002-4077-0099
Mikita Suyama http://orcid.org/0000-0001-9526-3193

References

[1] Hodges E, Xuan Z, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007;39:1522–1527.

[2] Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014;59:5–15.

[3] Blázquez L, Aiastui A, Goicoechea M, et al. In vitro correction of a pseudoexon-generating deep intronic mutation in LGMD2A by antisense oligonucleotides and modified small nuclear RNAs. Hum Mutat. 2013;34:1387–1395.

[4] Flanagan SE, Xie W, Caswell R, et al. Next-generation sequencing reveals deep intronic cryptic ABCC8 and HADH splicing founder mutations causing hyperinsulinism by pseudoexon activation. Am J Hum Genet. 2013;92:131–136.

[5] Chmel N, Danescu S, Gruler A, et al. A deep-intronic FERMT1 mutation causes kindler syndrome: an explanation for genetically unsolved cases. J Invest Dermatol. 2015;135:2876–2879.

[6] Naruto T, Okamoto N, Masuda K, et al. Deep intronic GPR143 mutation in a Japanese family with ocular albinism. Sci Rep. 2015;5:11334.

[7] Vaz-Drago R, Custódio N, Carmo-Fonseca M. Deep intronic mutations and human disease. Hum Genet. 2017;136:1093–1111.

[8] Abecasis GR, Auton A, Brooks LD, et al.; 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.

[9] Lappalainen T, Sammeth M, Friedländer MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501:506–511.

[10] Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.

[11] Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–2993.

[12] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–360.

[13] Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005;33:D501–4.

[14] Pertea M, Pertea GM, Antonescu CM, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–295.

[15] Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–26.

[16] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–2079.

[17] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–842.

[18] Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004;11:377–394.

[19] Schultz J, Milpetz F, Bork P, et al. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A. 1998;95:5857–5864.

[20] Zhao K, Lu Z-X, Park JW, et al. GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. Genome Biol. 2013;14:R74.

[21] Desmet F-O, Hamroun D, Lalande M, et al. Human splicing finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009;37:e67.

[22] Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10:1523.

[23] Stein S, Lu Z-X, Bahrami-Samani E, et al. Discover hidden splicing variations by mapping personal transcriptomes to personal genomes. Nucleic Acids Res. 2015;43:10612–10622.

[24] Lykke-Andersen S, Jensen TH. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. Nat Rev Mol Cell Biol. 2015;16:665–677.

[25] Katz Y, Wang ET, Silterra J, et al. Sashimi plots: quantitative visualization of alternative isoform expression from RNA-seq data. bioRxiv. 2014;002576. DOI:10.1101/002576

[26] Crooks GE, Hon G, Chandonia J-M, et al. WebLogo: a sequence logo generator. Genome Res. 2004;14:1188–1190.

[27] Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A. 2003;100:189–192.

[28] MacArthur DG, Balasubramanian S, Frankish A, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335:823–828.

[29] Nakashima H, Nishikawa K, Ooi T. Differences in dinucleotide frequencies of human, yeast, and Escherichia coli genes. DNA Res. 1997;4:185–192.

[30] Krawczak M, Thomas NST, Hundrieser B, et al. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum Mutat. 2007;28:150–158.

[31] Pros E, Gómez C, Martín T, et al. Nature and mRNA effect of 282 different NF1 point mutations: focus on splicing alterations. Hum Mutat. 2008;29:E173–93.

[32] PCAWG Transcriptome Core Group, Calabrese C, Davidson NR, Demircioğlu D, et al. Genomic basis for RNA alterations in cancer. Nature. 2020;578:129–136.

[33] Faà V, Incani F, Meloni A, et al. Characterization of a disease-associated mutation affecting a putative splicing regulatory element in intron 6b of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. J Biol Chem. 2009;284:30024–30031.

[34] Homolova K, Zavadakova P, Doktor TK, et al. The deep intronic c.903+469T>C mutation in the MTRR gene creates an SF2/ASF binding exonic splicing enhancer, which leads to pseudoexon activation and causes the cblE type of homocystinuria. Hum Mutat. 2010;31:437–444.

[35] Känsäkoski J, Jääskeläinen J, Jääskeläinen T, et al. Complete androgen insensitivity syndrome caused by a deep intronic pseudoexon-activating mutation in the androgen receptor gene. Sci Rep. 2016;6:32819.

[36] Jayasinghe RG, Cao S, Gao Q, et al. Systematic analysis of splice-site-creating mutations in cancer. Cell Rep. 2018;23:270–281.e3.

[37] Fairbrother WG, Yeh R-F, Sharp PA, et al. Predictive identification of exonic splicing enhancers in human genes. Science. 2002;297:1007–1013.

[38] Sawyer SL, Hartley T, Dyment DA, et al. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. Clin Genet. 2016;89:275–284.

[39] Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: a practical guide to its clinical application. Brief Funct Genomics. 2016;15:374–384.