

Occurrence of splice-site-creating mutations in the human genome and their involvement in genetic disorders

坂口, 愛美

<https://hdl.handle.net/2324/4784434>

出版情報 : Kyushu University, 2021, 博士 (理学), 課程博士
バージョン :
権利関係 :



Doctoral Dissertation

Occurrence of splice-site-creating mutations in the human
genome and their involvement in genetic disorders

BY

Narumi Sakaguchi

A dissertation submitted in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy (Ph.D.) in Science

February 2022

Contents

Acknowledgments	1
Publications	2
Chapter 1. Introduction.....	3
1.1 Genome structure.....	3
1.2 RNA splicing mechanism.....	4
1.3 Alternative splicing.....	8
1.3.1 Alternative splicing and evolution	11
1.3.2 Alternative splicing and disease	12
1.4 Current status in the identification of mutations that cause genetic disorders	17
1.5 Thesis goals and outline.....	19
Chapter 2. Estimating the general frequency of splice-site-creating mutations	20
2.1 Introduction.....	20
2.2 Materials and methods	21
2.2.1 Dataset.....	21
2.2.2 Quality check of RNA-seq data	22
2.2.3 Personal genome sequence construction and RNA-seq read mapping	22
2.2.4 Splice site scoring	22
2.2.5 Analysis of protein domain architecture	23
2.2.6 Calculation of exon inclusion ratio.....	24
2.2.7 Analysis of splicing regulatory elements in the flanking regions of the identified pseudo-exons.....	24
2.2.8 Enrichment analysis of functional categories of genes	25
2.2.9 Evaluation of identified SCMs by SpliceAI	25
2.2.10 Data visualization and manipulation.....	25
2.3 Results	26

2.3.1 Identification of pseudo-exon activation events	26
2.3.2 Examples of pseudo-exon activation events.....	29
2.3.3 Systematic analysis of the identified pseudo-exon activation events.....	33
2.3.4 Frequency and spectrum of SCMs for each site	39
2.4 Discussion	40
Chapter 3. Comprehensive identification for splice-site-creating mutations capable of causing genetic disorders.....	44
3.1 Introduction.....	44
3.2 Materials and methods	45
3.2.1 Genomic variants	45
3.2.2 Creation of a list of causative genes for genetic disorders.....	45
3.2.3 Splice site scoring	46
3.2.4 Sequence-based prediction of SCMs.....	46
3.2.5 Validation using RNA-seq data	46
3.2.6 Calculation of junction allele fraction.....	47
3.2.7 Identification of SCMs in WES data of ciliopathy patients.....	47
3.2.8 Analysis of protein domain architecture	48
3.2.9 Modeling a three-dimensional protein structure	48
3.2.10 Data visualization and manipulation.....	49
3.3 Results	49
3.3.1 Identification of splice-site-creating mutations in the causative genes of genetic disorders	50
3.3.2 Estimation of sensitivity, specificity, and positive predictive value in identifying SCMs	52
3.3.3 Examples of novel exons induced by SCMs.....	53
3.3.4 Structural characteristics of the identified SCMs.....	57
3.3.5 Functional consequences of the identified SCMs and their possible contribution to pathogenesis.....	59
3.3.6 Identification of SCMs possibly involved in the pathogenesis of ciliopathies from WES data	62
3.4 Discussion	67
Chapter 4. Conclusion.....	70
4.1 Summary.....	70

4.2 Future tasks.....	70
Abbreviations used in this thesis.....	72
References	75
Supplementary materials	89

Acknowledgments

I would like to express my sincere gratitude to Professor Mikita Suyama for his appropriate advice and careful guidance in conducting this study. I would also like to express my heartfelt gratitude to assistant professors, Dr. Chie Kikutake and Dr. Naoto Kubota, for their guidance and encouragement from the beginning of my research.

I am also grateful to the JST SPRING (Grant Number: JPMJSP2136) for promoting my Ph.D. study by the financial support.

Finally, I would like to express my gratitude to my family who respected my opinions and warmly watched over me, my friends who supported me mentally, and all the people who have been involved in my study.

Publications

1. Narumi Sakaguchi and Mikita Suyama
In silico identification of pseudo-exon activation events in personal genome and transcriptome data
RNA Biology, 2021
<https://doi.org/10.1080/15476286.2020.1809195>
2. Narumi Sakaguchi and Mikita Suyama
Pervasive occurrence of splice-site-creating mutations and their possible involvement in genetic disorder
npj Genomic Medicine, *in press*

Chapter 1. Introduction

1.1 Genome structure

Genome is the complete set of genetic information in an organism. The instructions for the genome consist of deoxyribonucleic acid (DNA), and DNA is found in the cell nucleus and mitochondria. DNA molecules consist of four basic units: adenine (A), cytosine (C), guanine (G), and thymine (T). DNA is responsible for inheriting genetic information, which is a blueprint for animals, including humans. Genomic DNA contains protein-coding regions and non-coding regions that do not encode protein sequences. Proteins are produced through transcription (from DNA to messenger ribonucleic acid (mRNA)) and translation (from mRNA to protein). This process is called central dogma [1]. In transcription, firstly, a precursor mRNA (pre-mRNA) is synthesized from DNA. To form a mature mRNA from pre-mRNA, it is necessary to go through RNA processing, such as 5'-end capping, an intron removal mechanism called splicing, and 3'-end polyadenylation called tailing (Fig. 1-1). In RNA, uracil (U) replaces thymine. Processed mRNA is then transported from the nucleus to the cytoplasm, and the information in mRNA is finally converted into protein. In this study, I will focus on splicing, which is one of the RNA processing steps.

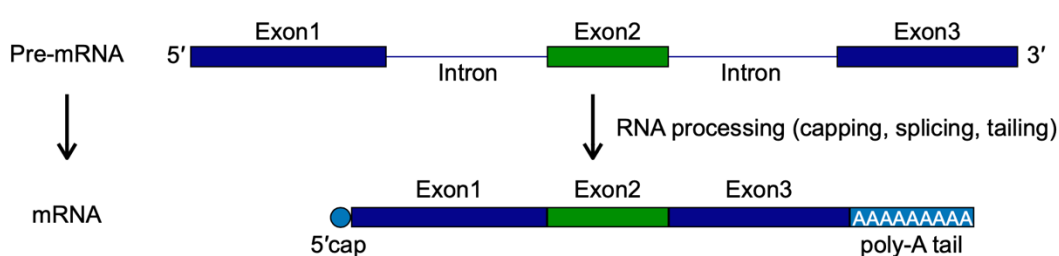


Fig. 1-1 Schematics of RNA processing.

Schematics of RNA processing to form mature mRNA from pre-mRNA are shown. Blue and green boxes indicate exons, and blue lines indicate introns. Pre-mRNA becomes mRNA after RNA processing, such as capping, splicing, and tailing.

1.2 RNA splicing mechanism

RNA splicing is a mechanism for generating mature mRNA from pre-mRNA by removing intron sequences that do not carry the information for amino acids that make up proteins [2]. After splicing, the remaining exons are joined together to form mature mRNA. Splicing involves RNA splice sites and branch point (BP) sequences (Fig. 1-2). There are two types of splice sites: a 5' splice site (5'ss) and a 3' splice site (3'ss). The 5'ss and 3'ss have consensus sequences with different lengths of 9-bp and 21-bp, respectively. Particularly, canonical splice sites, dinucleotide GU at the 5'-end of the intron for 5'ss and dinucleotide AG at the 3'-end of the intron for 3'ss [3], are highly conserved among species. Additionally, the 3'ss has a C or U repeat sequence called polypyrimidine tract (PPT) [4], which is longer than the 5'ss. The BP sequence is located further upstream from PPT and includes an adenine nucleotide, which is involved in the lariat formation of cleaved introns [5]. Splicing occurs when components of the spliceosome recognize and bind to RNA splice sites and BP sequence, respectively. Splicing is classified into two types, U2-type (major) and U12-type (minor), depending on the type of spliceosome involved [6]. In this thesis, I focused on the splicing mechanism of the U2-type, which accounts for 99% of all splicing [7].

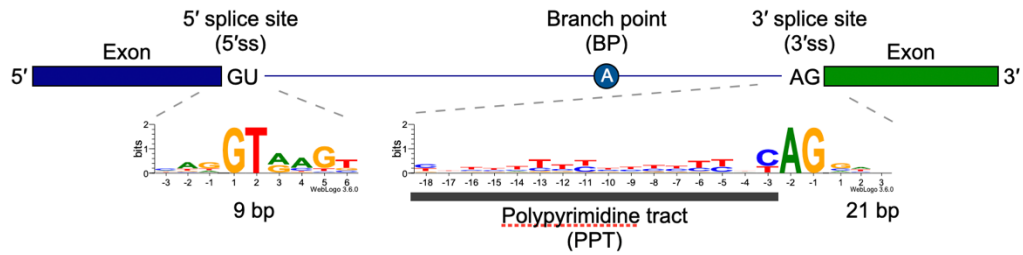


Fig. 1-2 Consensus sequences involved in RNA splicing.

Conserved sequences involved in splicing are shown. Blue and green boxes indicate exons, and a blue line indicates an intron. In 5'ss, dinucleotide GU at the 5'-end of the intron and, in 3'ss, dinucleotide AG at the 3'-end of the intron, are known as canonical splice sites. The 5'ss consensus sequence length is 9-bp (six bases in the intronic regions and three bases in the exonic regions), and 3'ss is 21-bp (18 bases in the intronic regions and three bases in the exonic regions) containing polypyrimidine tract (PPT). The height of the alphabetic characters represents the frequency of base occurrence at each splice site. The data are generated using WebLogo 3 [8]. Branch point (BP) was 18–40 nucleotides upstream of the 3'ss.

The splicing reaction is catalyzed by the spliceosome, which consists of five small nuclear ribonucleoproteins (snRNPs) (U1, U2, U4, U5, and U6 snRNPs in major splicing) and approximately 200 proteins, including splicing factor 1 (SF1) and U2 snRNP auxiliary factor (U2AF) [9,10](Fig. 1-3). Firstly, U1 snRNP recognizes the consensus sequence at 5'ss, SF1 recognizes BP sequence, and U2AF recognizes PPT at 3'ss. The U2AF helps the SF1 bind to the BP sequence. These spliceosomes bind to the pre-mRNA sequence, forming a first complex, complex E. Next, with the aid of the U2AF, SF1 bound to BP sequence is replaced by U2 snRNP, forming a pre-spliceosome (complex A). The U5 and the U4/U6 snRNPs are then assembled into a complex and bound to complex A to form complex B. The U5 snRNP binds exons at the 5'ss, and the U6 snRNP binds to the U2 snRNP. At this point, U1 and U2 snRNPs move closer to each other by interaction, and the end positions of each exon are closer to each other. The U1 snRNP and the U4 snRNP are released and the U6 snRNP binds at the 5'ss, forming a complex C. The U5 snRNP binds exon at

3'ss, then, the 5'-end of the intron is cleaved. In the intron, the floating GU sequence binds to the BP, forming a lariat structure. Finally, the 3'-end of the intron is cleaved, the intron is removed, and the exons are combined to form mature mRNA [11,12]. The introns are converted to linear forms by RNA lariat debranching enzyme Dbr1, and the linear introns are further degraded by exoribonucleases [13,14].

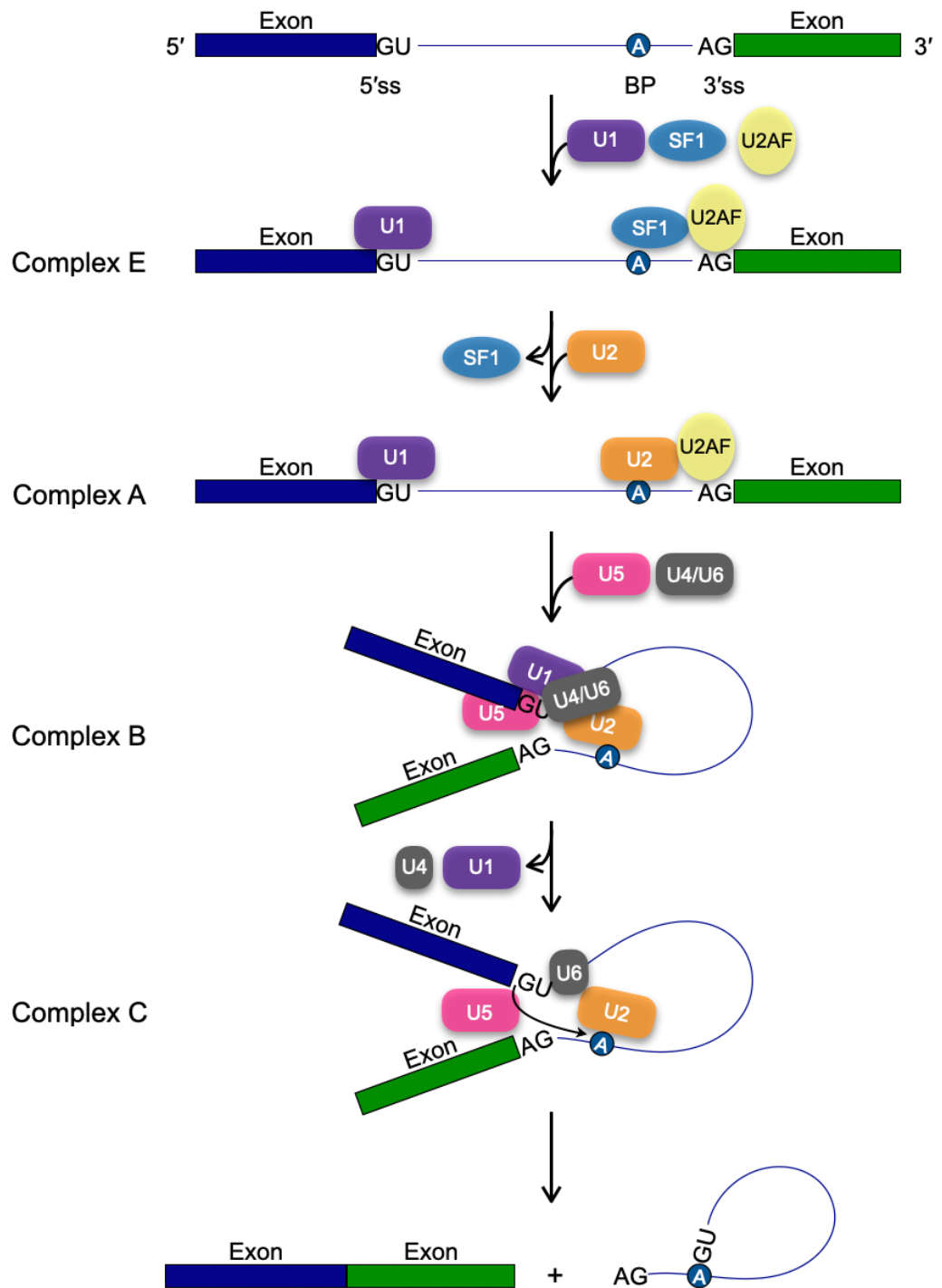


Fig. 1-3 Schematics of splicing reaction.

Assembly state of small nuclear ribonucleoproteins (snRNPs) at each complex. Blue and green boxes indicate exons, and blue lines indicate introns. Other colored boxes are snRNPs or other proteins. Five snRNPs and approximately 200 proteins recognize the consensus sequences and proceed with splicing while forming various complexes.

1.3 Alternative splicing

There are two types of splicing: constitutive splicing and alternative splicing. Constitutive splicing uses all exons in a gene to produce a single mRNA isoform. In contrast, alternative splicing can produce various mRNA isoforms by altering combinations of multiple exons. It was estimated that 95% of the 20,000–30,000 human genes are alternatively spliced, and the total number of mRNA produced is approximately 120,000 [15,16]. Furthermore, a high proportion of alternatively spliced genes are also found in eukaryotes other than humans. Thus, alternative splicing contributes to the diversity of proteins in various eukaryotes and is an essential process for gaining advantageous proteins for survival and adaptation to the environment.

Alternative splicing is commonly regulated by changes in the binding efficiency of the spliceosome to the splice site or by binding of one or more trans-acting factors to the cis-acting elements, namely, splicing regulatory elements (SREs) [12,17–19]. The SREs are located in exons or introns close to the splice sites, and they enhance or silence splicing (Fig. 1-4). Splicing enhancers, including exonic splicing enhancer (ESE) [20] and intronic splicing enhancer (ISE) [21], activates splicing. The ESEs promotes exon inclusion by binding serine/arginine-rich (SR) proteins and facilitating the binding of U2AFs and U1 snRNPs to the splice sites [22–24]. Whereas splicing silencers, including exonic splicing silencer (ESS) [25] and intronic splicing silencer (ISS), inhibit splicing. The ESSs promote exon skipping by binding heterogeneous nuclear ribonucleoproteins (hnRNPs) and inhibiting the binding of U2AFs and U1 snRNPs to the splice sites [26]. In addition to the above factors, recent studies have shown that many other factors are involved in the

splicing regulation, including RNA secondary structure [27–29], three-dimensional (3D) organization of chromatin [30], DNA methylation [31], and transcription rate [32].

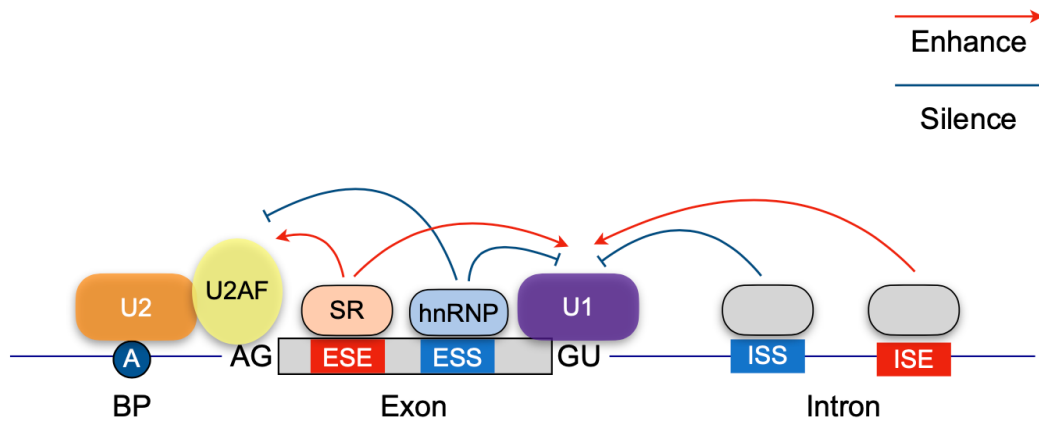


Fig. 1-4 Schematic of regulated splicing by trans-acting factors and cis-acting elements.

Splicing is regulated by the binding of trans-acting factors to cis-acting elements. Gray box indicates exon, and the blue line indicates intron. Binding of serine/arginine-rich (SR) protein to the exonic splicing enhancer (ESE) promotes 3'ss recognition by U2AF or 5'ss recognition by U1 snRNP. The binding of heterogeneous nuclear ribonucleoprotein (hnRNP) to exonic splicing silencer (ESS) inhibits 3'ss recognition by U2AF or 5'ss recognition by U1 snRNP. Gray ovals show other proteins that regulate splicing.

Following the state of the exons, alternative splicing can be classified into five categories: cassette exon, alternative 5'ss, alternative 3'ss, intron retention, and mutually exclusive exon [33] (Fig. 1-5). Cassette exon is a form of exon skipping or insertion in a transcript. Alternative 5' splice site and alternative 3' splice site are forms of shifting exon–intron boundaries. Intron retention is a state where the full length of an intron is included in transcripts. Mutually exclusive exons are forms in which only one of the two exons is retained, but not both at the same time. The abundance ratio of each splicing event to all alternative splicing events is reported to be approximately 38% for cassette exons, 26% for alternative 5' ss and alternative

3' ss, 3% for intron retention [34]. In this thesis, I focused on cassette exon form and alternative 3' or 5' ss forms, which account for most of the alternative splicing events.

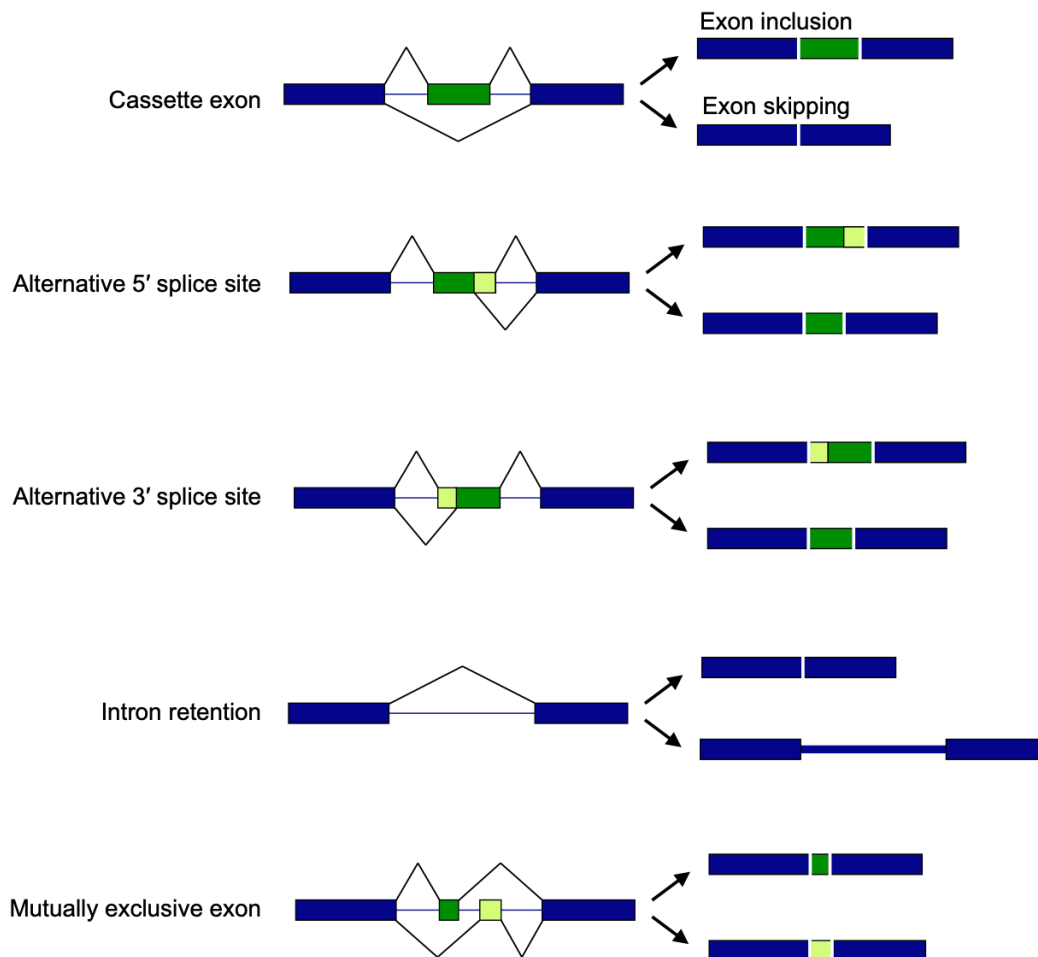


Fig. 1-5 Five main categories of alternative splicing.

According to the state of exons, alternative splicing is classified into five categories. Blue boxes indicate constitutive exons, and green boxes indicate alternative exons. Blue lines indicate introns. From top to bottom, schemes of a cassette exon, alternative 5' splice site, alternative 3' splice site, intron retention, and mutually exclusive exons are shown.

1.3.1 Alternative splicing and evolution

The human genome contains about 200,000 exons [35], close to the number found in mice and other mammals [36]. The number of exons is constantly fluctuating during the evolution process, and in metazoan, generally, the organism complexity increases with an increase in the number of exons [37,38]. Splicing pattern, which is the combination of exons that make up a gene, varies between species even for a given orthologous gene, and even if the splicing pattern is identical between species, the inclusion ratio of the alternative exons may vary. Therefore, alternative splicing is thought to result from the genetic diversity of species. For example, a human-specific exon loss event was found in the *SLC7A6* gene, which encodes an amino acid transporter [39]. Splicing regulates the combination of exons to produce a suitable protein for the survival of species.

Recently, comparative genomic analysis has revealed several cases in which species-specific exons were acquired in orthologous intronic regions. For example, Singer and coauthors [40] found a novel alternative 5' exon generated in the human tumor necrosis factor receptor (*p75TNFR*) gene by comparative genomic analysis of 13 species. The alternative exon is activated in humans, apes, and cercopithecoids, but not in New World monkeys because they lack the 5'ss sequence [40]. Exonization, the biological process of the generation of new exons, is most likely due to gene duplication, exon duplication, and insertion of Alu elements, which are transposable elements [41–44]. Alternative splicing is an essential regulatory mechanism for species diversity, but newly acquired exons do not necessarily serve a positive function in species. Newly inserted exons into transcripts often shift the original reading frames, introducing premature termination codons (PTCs) at a position upstream of the normal termination codons. Translated mRNA with PTC is often harmful to the cell because it produces a shorter protein than normal protein and it may not function properly; hence, newly created exons are skipped much more frequently than older exons [45]. In other words, the number of harmful transcripts is reduced by alternative splicing, and the survival of the species is maintained by keeping the normal transcripts at a certain level. During evolution, novel exons with

positive functions for survival are fixed in the species, and their content in transcripts gradually increases [45].

1.3.2 Alternative splicing and disease

Splicing is intimately involved in diseases. It has been reported that approximately 60% of human diseases are associated with splicing abnormalities [46,47]. Particularly, mutations that disrupt splice sites inhibit their recognition by the spliceosome, resulting in splicing abnormalities that can lead to disease [48]. It is known that when splice sites are disrupted, splicing abnormalities are likely to occur in the order of exon skipping, activation of cryptic splice site, and intron retention [49]. The highest prevalence of exon skipping can be attributed to the exon definition model which is common in organisms with short exons and long introns (for example, vertebrates) [50]. In the exon definition model, it is reported that the interaction between 5'ss and 3'ss spliceosomes is involved in splicing. If the spliceosome does not recognize splice site, the spliceosome interaction may be disrupted and exon skipping may occur. Additionally, mutations in the spliceosome components or in SREs can also cause splicing abnormalities [51,52]. Further, acquiring a new exon may be the cause of human diseases [53]. In patients with X-linked ocular albinism, a deficiency of melanin pigment in the eye, a novel 5'ss is formed within the deep intronic region of *GPR143*, inserting an aberrant exon into the coding sequence and causing the disease [54]. This new formation of the 5'ss is due to a mutation. This phenomenon, in which a part of the intron region is activated and used as an exon, is called pseudo-exon activation. In the next section, I will discuss two main reasons why splicing abnormalities may be responsible for genetic disorders.

1.3.2.1 Nonsense-mediated mRNA decay

A reason why splicing abnormalities may be responsible for genetic disorders is decrease in the amount of proteins due to nonsense-mediated mRNA decay (NMD) [55]. NMD is one of the mRNA quality control mechanisms in eukaryotes. Usually, the genetic information in mRNA is transcribed by an enzyme protein known as RNA

polymerase II (RNA Pol II) then, translated into an amino acid sequence from the start (initiation) codon to the stop (termination) codon. However, if a splicing error occurs, the reading frame of the codon sequence may shift, creating PTC before the position of an original termination codon. In this case, the protein becomes shorter and incomplete compared to the normal protein. Incomplete protein may not be able to conduct the original protein function and may cause diseases. Additionally, the NMD machinery recognizes PTCs and degrades the abnormal mRNA before they are translated to reduce the amount of these abnormal protein. As a result, the amount of a required protein is insufficient and causes diseases. However, NMD cannot completely degrade abnormal mRNAs, and their degradation rates vary for each gene and transcript. For example, PTC-containing transcript expression of Desmoglein 2 (*DSG2*) is reduced by NMD to approximately ~30%–40% in blood, whereas PTC-containing transcript expression of Pejvakin (*PJVK*) is reduced to less than 10% in the same tissue [56]. Additionally, according to the NMDetective [57], a tool for predicting the impact of PTCs in mRNA, only 51% of PTCs that may occur in humans undergo complete degradation by NMD, and the remaining 49% undergo partial or no degradation. Therefore, even if abnormal transcripts including PTCs are produced, they can still be detected by RNA-seq if these are not completely degraded by NMD [58,59].

How does NMD machinery distinguish between PTCs and normal termination codons? In mammals, PTC recognition involves the existence of an exon junction complex (EJC) [60,61] (Fig. 1-6). EJC is normally composed of four core proteins: eukaryotic translation initiation factor 4A3 (eIF4A3), cancer susceptibility candidate 3 (CASC3), RNA-binding motif protein 8A (RBM8A or Y14), and mago-nashi homolog (MAGOH). It is outfitted with up-frameshift suppressor 3 homolog B (UPF3B or UPF3X) and up-frameshift suppressor 2 homolog (UPF2) proteins. When two exons are connected by splicing, the EJC is bound to their exon-exon junction. EJC is then removed from the mRNA by the ribosome during translation. If PTC presents in a transcript, up-frameshift suppressor 1 homolog (UPF1) and its associated kinase, serine/threonine protein kinase SMG1, bind to the eukaryotic

release factor 1 (eRF1) and eukaryotic release factor 3 (eRF3) to form a SMG1-UPF1-eRF1/3 (SURF) complex. The SURF complex binds to the ribosome via eRF1 and eRF3, and interacts with EJC to form a complex called a decay-inducing complex (DECID). This inhibits the movement of the ribosome and terminates translation. In other words, the EJC serves as a signal to initiate NMD [62]. If PTC exists in the last exon where EJC does not bind, DECID cannot be formed, and NMD is escaped. Additionally, if PTC presents within 50–55-nt upstream of the last exon-exon junction, the DECID cannot be formed owing to the close distance between PTC and EJC, and NMD is escaped [61]. Therefore, NMD is controlled based on the positional relationship between PTC and EJC.

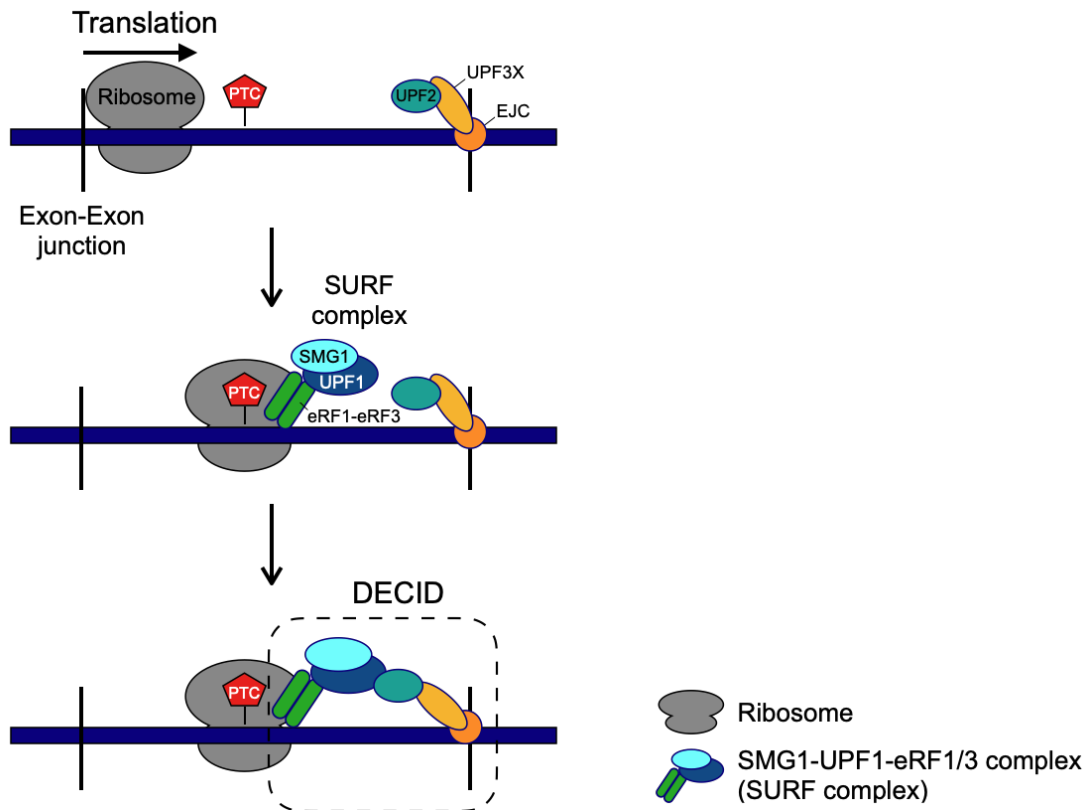


Fig. 1-6 Formation of decay-inducing complex.

A schematic of premature termination codon (PTC) containing transcript leads to nonsense-mediated mRNA decay (NMD). Blue box indicates exons, and vertical lines indicate exon-exon junction. A EJC (shown as orange circle) is outfitted UPF3X (yellow oval) and UPF2 (green oval). A SURF complex consists of eRF1-eRF3 (light green ovals), UPF1 (blue oval), and SMG1 (light blue oval). A ribosome is shown as a combination of gray ovals.

1.3.2.2 Disruption of protein domain structure

The second reason why splicing abnormalities may be responsible for genetic disorders is the disruption of protein domain structure. In the human growth hormone 1 (*GH1*) gene, the causative gene of isolated growth hormone deficiency type II (IGHD II), splicing variants within exon 3 are involved in the pathogenesis of the disease [63–65]. Particularly, mutations in the 5'ss of exon 3 causes exon skipping, which results in the generation of aberrant *GH1* isoforms that causes the disease [66–68]. These mutations do not change the downstream reading frame and do not generate PTC. Therefore, the aberrant *GH1* isoform is not a target of NMD, but the 40 amino acids coded by exon 3 are deleted from the protein (Fig. 1-7a). The deleted fragment is located in the central part of the Hormone_1 domain (Fig. 1-7b), where it affects the domain structure and alters its protein function. Thus, splicing abnormalities may cause disease by disrupting the structure of protein domains, even if PTC does not generate.

a

```

1      11      21      31      41      51      61      71      81      91      101      111      121
MATGSR TSL LAFGL LCLPW LQEGSA FPTI PLSRL FDNAM LRAHRL HQLA FDTYQ EF EEAYIPKEOKYSF LQNPQ TSLCF SESIPT PSNR EETQ QKS NLELL RLSLL LIQSW LEPVQ FLRSVF
131      141      151      161      171      181      191      201      211
ANSLVYG ASDSNVY DLLKDLE EGIQTL MGRLED GSPRTG QIFKQ TYSKFDT NSHND DALLKNY GLLYCF RKDMD RVE TFLRIV QCRS VEGSCGF

```

b

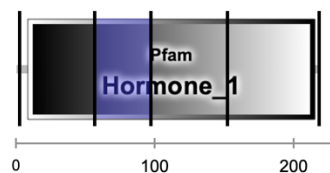


Fig. 1-7 Example of disrupted protein domain by splicing abnormality.

Disruption of domain structure in *GH1* gene. (a) Amino acid sequence of *GH1* gene. Deleted sequence of 40 amino acids is shown in blue. (b) Effect of splicing abnormality on the Hormone_1 domain structure. The protein domain architecture is taken from Simple Modular Architecture Research Tool (SMART) [69]. The blue rectangle indicates the 40-amino-acid-long fragment coded by the exon 3. The vertical line indicates the position of the intron.

1.4 Current status in the identification of mutations that cause genetic disorders

There are two types of human diseases: multifactorial diseases, caused by a complex combination of several genetic and environmental factors, and monogenic diseases, caused by an abnormality in one specific gene. Particularly, in monogenic diseases, splicing abnormalities may be critically involved in the pathogenesis of the disease [70]. Therefore, in this thesis, I mainly focused on monogenic diseases. There are many varieties of monogenic diseases, and the total number of patients with all monogenic diseases is estimated to be approximately 350 million worldwide [71]. Many monogenic diseases and their causative genes have been identified, but many rare and intractable diseases remain unidentified. Rare and intractable diseases are often difficult to properly diagnose because their pathogenesis is unknown. It has also been reported that there are no therapeutic drugs for more than 95% of the rare diseases [72].

It is necessary to identify causative mutations of genetic disorders for the correct diagnosis and development of therapeutic drugs. Whole exome sequencing (WES) has served mainly as an effective method to search for causative mutations [73]. WES can be used to detect missense mutations, nonsense mutations, frameshift mutations, and mutations around splice site regions; it has contributed to the identification of many disease-associated mutations. However, it has been reported that the probability of successfully identifying the causative mutation using WES is only about 50% of all patients [74]. The reason is that WES can usually only detect mutations in exonic regions or in intronic regions within approximately 50-bp exon–intron boundaries [75]. If a causative mutation is located in a deep intronic region more than 50-bp from an exon–intron boundary, it cannot be detected by WES. Furthermore, although it is possible to detect mutations in deep intronic regions using whole genome sequencing (WGS), the sheer number of mutations detected makes it difficult to identify the actual causative mutations if the pathogenic mechanism is unknown. Therefore, it is necessary to clarify how mutations contribute to disease pathogenicity. Since the discovery of new causative mutations possibly leads to the diagnosis of patients and the development of therapeutic drugs based

on the pathogenic mechanism, it is essential to improve the identification rate of disease-causing mutations.

1.5 Thesis goals and outline

In this thesis, I aimed to identify novel mutations noteworthy in identifying causative mutations of genetic disorders to improve the identification rate. I focused on splice-site-creating mutations (SCMs), which have been sporadically reported to cause various genetic disorders and cancers in recent years [53,76,77]. To achieve this goal, I had to resolve the two issues as shown below.

1. Estimating the general frequency of splice-site-creating mutations
2. Comprehensive identification for splice-site-creating mutations capable of causing genetic disorders

Based on the above, I have conducted two studies. In the study described in Chapter 2, I analyzed the frequency and characteristics of pseudo-exon activation events that occur when SCMs exist in deep intronic regions. From the obtained results, pseudo-exon activation is estimated to occur at least 10 loci per normal individual. This suggests that SCM is not a rare mutation. Based on the study described in Chapter 2, I thought that if I actively searched for SCMs, I could identify causative mutations even in patients in whom they had not been identified. Therefore, in the study described in Chapter 3, I conducted a cross-disease analysis for SCMs that may be involved in the pathogenesis of genetic disorders. From the analysis, I identified 5,656 SCMs in 4,054 genes that may cause genetic disorders, and 3,942 SCMs of these were expected to be involved in the pathogenesis by triggering NMD or by disrupting protein domain structures. Additionally, the number of potentially pathogenic SCMs identified in this study was compared to the number of pathogenic mutations that disrupt protein-coding regions or splice sites, which are commonly focused on in the search for causative mutations. It was estimated that the identification rate of causative mutations could be improved by approximately 5.9%–8.5% by considering the possibility that SCM is attributable to genetic disorders. This suggests that SCMs are mutations noteworthy of attention in the search for causative mutations of genetic disorders.

Chapter 2. Estimating the general frequency of splice-site-creating mutations

2.1 Introduction

Causative mutations for genetic disorders have mainly been identified in exons, especially in the coding sequences, and in RNA splice sites at both ends of the introns. This is because these regions are functionally important and, hence, conserved throughout evolution. If a mutation occurs in these regions, the transcript or its protein product may be disrupted, often leading to a disease phenotype. Whole exome sequencing (WES) is a method that efficiently detects mutations in exonic regions and their flanking RNA splice sites [78]. WES has been successfully applied to identify causative mutations in various genetic disorders (for a review, see [79]).

Compared with mutations in coding regions and splice sites, mutations in deep intronic regions have not been a target for a thorough analysis of disease-causing mutations, mainly for two reasons. Firstly, such regions are usually not conserved among species and, hence, not thought to be functionally important. Secondly, conventional WES can only identify mutations in exons and flanking intronic sequences, which contain splice sites. Thus, it is impossible to identify mutations that occur in deep intronic regions. However, growing evidence has shown that mutations in deep intronic regions often create novel RNA splice sites capable of triggering pseudo-exon activation, which in turn leads to aberrant transcripts containing an extra exon (e.g. [54,80–82]). 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 [53]).

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 is yet to be analyzed. With the advancement of personal genome sequencing [83] and transcriptome analysis for the corresponding individuals [59], 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, I conducted an *in silico* identification of such events in normal individuals using publicly available personal genome and transcriptome data for the corresponding individuals, especially focusing on those single-nucleotide variants (SNVs) that create novel splice sites.

2.2 Materials and methods

2.2.1 Dataset

I downloaded the variant data for 462 individuals as determined by the 1000 Genomes Project [81] in the Variant Call Format (VCF) and their corresponding RNA-seq data registered in the Geuvadis project [59] in the fastq format. The RNA-seq data are derived from lymphoblastoid cell lines for the corresponding individuals. The 1000 Genomes Project is an international project that creates a huge public catalog of human variation and genotype data ran from 2008 to 2015. These data were used because they are a large dataset with both the human genome and the corresponding RNA-seq data. The reference genome sequence (GENCODE V19) in the FASTA format and the gene annotation data in the General Transfer Format (GTF) were downloaded from the University of California, Santa Cruz (UCSC) Genome Browser [84]. The public data used in this study are listed (Table 2-1).

Table 2-1. List of data used in the study of Chapter 2.

Data	Number of samples	Format	Database	URL
Variant data	462	VCF	1000 Genomes Project	https://www.internationalgenome.org/home
RNA-seq	462	fastq	Geuvadis	https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-3/
Reference genome data (GRCh37.p13)	1	FASTA	GENCODE	https://genome.ucsc.edu/cgi-bin/hgTables
Gene annotation data (V19)	1	GTF	GENCODE	https://genome.ucsc.edu/cgi-bin/hgTables
Gene annotation data	1	GTF	RefSeq	https://genome.ucsc.edu/cgi-bin/hgTables

2.2.2 Quality check of RNA-seq data

To evaluate the quality of the sequencing reads obtained from the Geuvadis project, I performed a quality check using the FastQC program (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) [85]. In this study, only high quality RNA-seq data with a sequencing quality score greater than 30 were used. After this quality filtering, the RNA-seq data for 235 individuals remained. The individual IDs used in this analysis and the number of SNVs are shown separately for the exonic and intronic regions (Supplementary Table 2-1).

2.2.3 Personal genome sequence construction and RNA-seq read mapping

BCFtools (version 1.9) [86] was applied to the reference genome sequence (GENCODE V19) and to the variant information for each individual in the VCF to construct individual-specific reference genome sequences. The hierarchical indexing for spliced alignment of transcripts (HISAT) 2 program (version 2.1.0) [87] was used to map the RNA-seq data onto the genomic sequences that reflect the variant information for the individual. Default parameters were used for mapping. Gene expression levels were quantified using the StringTie program (version 1.3.5) [88]. The gene structures registered in the RefSeq data [89] in the GTF were used as reference transcriptome data. RefSeq data were downloaded from the UCSC Genome Browser. The total number of reference transcripts and genes were 50,643 and 26,242, respectively.

2.2.4 Splice site scoring

The strength of splice sites was evaluated using the MaxEntScan program [90]. The MaxEntScan is a bioinformatics tool for scoring how submitted genome sequence segments resemble the sequence of splice site. A Perl script of the tool was downloaded from <http://hollywood.mit.edu/burgelab/maxent/download/> and used. 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 (Fig. 2-1). For 3'ss, genome sequence segments corresponding to the 20 bases at the end of introns and the three bases at the start of pseudo-exons were also subjected to MaxEntScan (Fig. 2-1).

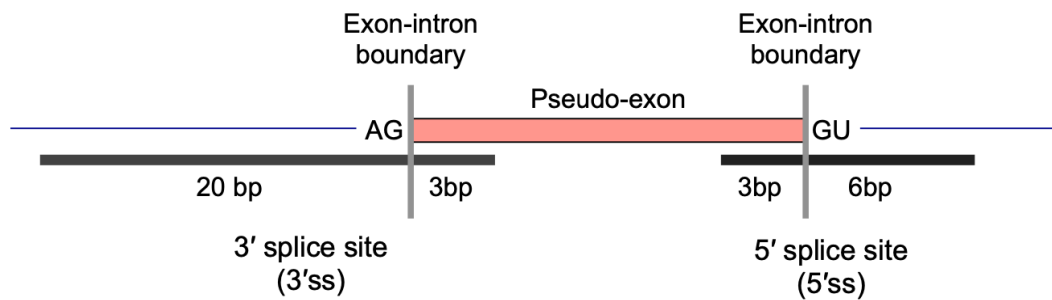


Fig. 2-1 A schematic of genome sequence segments targeted for MaxEntScan.

The red box indicates an exon, while blue lines indicate introns. The vertical lines indicate exon–intron boundaries, while the bold black line indicates the range of each splice site used for MaxEntScan.

2.2.5 Analysis of protein domain architecture

For cases where the inclusion of a pseudo-exon introduces 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, I used the Simple Modular Architecture Research Tool (SMART) database [69]. The SMART is a tool that enables the identification and annotation of domains for more than 500 domain families. For the input data, I used the protein sequence, which was created using gffread (<https://github.com/gperte/gffread>) based on gene annotation data (GENCODE V19).

2.2.6 Calculation of exon inclusion ratio

Exon inclusion ratio, or “percent spliced in” (PSI), can be calculated using the following equation [91] (Fig. 2-2):

$$PSI = (J_i / 2) / (J_i / 2 + J_s)$$

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 two is because there are two junctions for an exon inclusion. PSI values can range from zero (completely skipped) to one (complete inclusion).

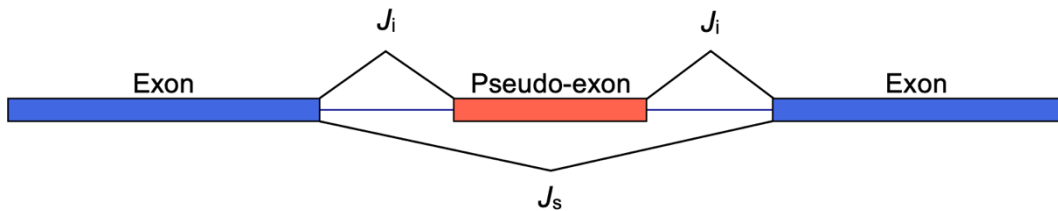


Fig.2-2 Calculation of PSI by mapped junction reads around pseudo-exons.

Junction reads that span exon (Blue boxes) and pseudo-exon (Red box) are denoted as J_i , and junction reads that span between exons by skipping pseudo-exon is denoted as J_s .

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

The pseudo-exons with identified SCMs that did not increase in the MaxEntScan scores, which is a measure of splice site strength, may have other SNVs that affect splicing within or flanking the pseudo-exons. The Human Splicing Finder [92] was used to examine whether there are other SNVs affecting the SREs within 1000 bp of identified pseudo-exons with lower MaxEntScan scores. For the input, the reference genome sequence (GRCh37.p13) downloaded from the UCSC Genome Browser was used as the “reference sequence” and the personal genome data as the “mutant sequence.”

2.2.8 Enrichment analysis of functional categories of genes

Enrichment analysis of functional categories of genes that included pseudo-exons was conducted using Metascape [93]. The following settings were used: “Input as species” was set to “*H. sapiens*,” and “Analysis as species” was set to “*H. sapiens*.”

2.2.9 Evaluation of identified SCMs by SpliceAI

To evaluate the SCMs that induce pseudo-exons, SpliceAI was used [94]. The SpliceAI is a tool for predicting genetic variants that affect splicing. As input data, I used SCMs that induced pseudo-exon activation and gene annotation data (GENCODE V19) obtained from the UCSC Genome Browser.

2.2.10 Data visualization and manipulation

The mapping data, with the data on genomic variants, were visualized using the Integrative Genomics Viewer (IGV) software (version 2.4.16) [95]. SAMtools (version 1.5) [96] and BEDTools (version 2.26.0) [97] were used for data manipulation. All existing tools used in my analysis were listed (Table 2-2). All analyses in this paper were conducted using Python (version 2.7.16), Java (10.0.2), Perl (v5.32.0), and GNU Bash (version 4.2.26) environments.

Table 2-2. List of tools used in the study of Chapter 2.

Tools	Version
FastQC	0.11.9
BCFtools	1.9
HISAT2	2.1.0
StringTie	1.3.5
IGV	2.4.16
SAMtools	1.5
BEDTools	2.26.0
MaxEntScan	-
SMART	-
Human Splicing Finder	3.1
Matascope	-
SpliceAI	1.3.1

2.3 Results

2.3.1 Identification of pseudo-exon activation events

For this study, genomic variant data and high quality RNA-seq data were used for 235 individuals. Before mapping the RNA-seq data of each individual onto the genome sequence, a reference genome that reflected the variant information for the corresponding individual using BCFtools (version 1.9) was prepared (Fig. 2-3). By creating such a personal reference genome, variants affecting splicing can be identified [98]. 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. Next, the best mapping program for pseudo-exon detection was selected. This analysis requires a mapping program that does not depend on gene annotation data. TopHat2 [99] and STAR [100] are mapping programs that are generally used with the input of gene annotation data. If the annotation data are not given, the mapping will take longer, and the accuracy of the mapping will be reduced. Therefore, I thought that these mapping programs were not suitable for pseudo-exon detection. I used HISAT2 (version 2.1.0) [87] because it can accurately and rapidly map RNA-seq data without using gene annotation data.

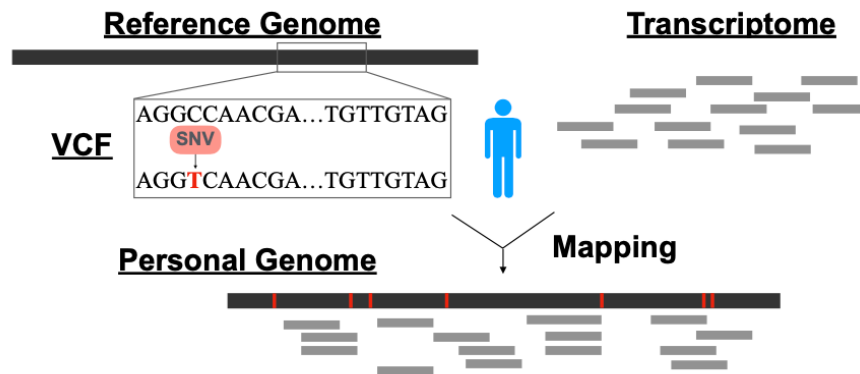


Fig. 2-3 Construction of a personal genome.

Schematic of the construction of personal genome and RNA-seq mapping. The RNA-seq data for an individual are mapped against the reference genome reflecting variant information for the corresponding individual.

Pseudo-exon activation events were identified through the following steps (Fig. 2-4). Firstly, I collected junction reads that were mapped onto annotated exons in RefSeq transcripts on one side, and the remainder side 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 to the direction of transcription and with pseudo-exon lengths of $\leq 1,000$ bp were retained as pseudo-exon candidates, having novel splice sites at both 5'- and 3'-ends. Next, only those pseudo-exon candidates that had an SNV either at the flanking region of the 5' splice site (5'ss) or at the flanking region of the 3' splice site (3'ss) as pseudo-exon activation events were selected. Here, I 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, I accepted longer intronic segments than those of 5'ss because 3'ss has a characteristic polypyrimidine tract at the end of the intron. Finally, I examined the correspondence between the presence of SNVs and pseudo-exons for all the

individuals analyzed in this study. I conducted this analysis because cases can occur in which individuals harbor 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 I was not able to detect pseudo-exons for some individuals with low or no expression. To deal with this possibility, I 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 cases, the pseudo-exon activation is not caused by that SNV but may be due to splicing noise or pre-mRNA contamination. After these filtering, the remaining mutations were defined as splice-site-creating mutations (SCMs) because they are probably involved in pseudo-exon activation.

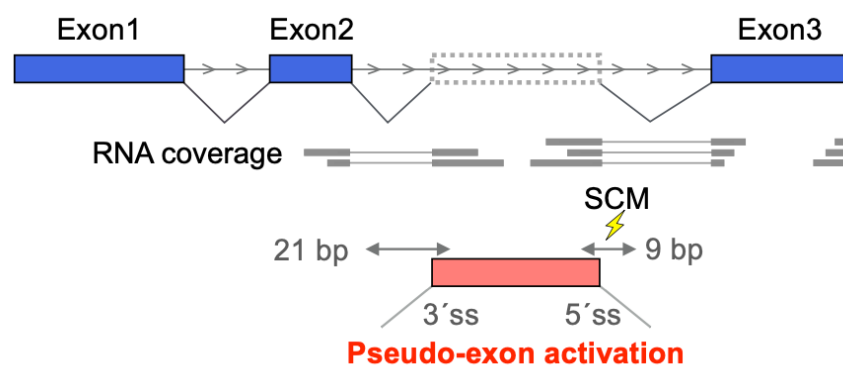


Fig. 2-4 A schematic of the identification of pseudo-exon activation events.

Identification of pseudo-exon activation events based on junction reads that are not mapped to annotated exons (see main text for further details).

From the 235 individuals analyzed in this study, 116 distinct pseudo-exons (Supplementary Table 2-2), which were created by 111 unique SCMs were identified. On average, there were 2.6 pseudo-exons per individual.

2.3.2 Examples of pseudo-exon activation events

As an example, here, I showed a pseudo-exon identified in the first intron of cornichon family AMPA receptor auxiliary protein 1 (*CNIH1*) (Fig. 2-5). 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 SCM 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, the 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, while the score for the reference genome sequence at the corresponding regions was -1.23, suggesting that the mutated sequence has a 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 relatively high, 12.02, indicating that the sequence pattern can be taken as a cryptic splice site even without SCMs. 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 three, 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 SCM 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 [55], or both. Additionally, although the pseudo-exon contains ATG, it is out-of-frame and is unlikely to be a start codon for alternative translation initiation. I also used the SpliceAI program [94] to evaluate whether the identified SCM has a potential to create the splice site. The calculated SpliceAI score was 0.80, suggesting a very strong possibility of creating a splice site by the SCM.

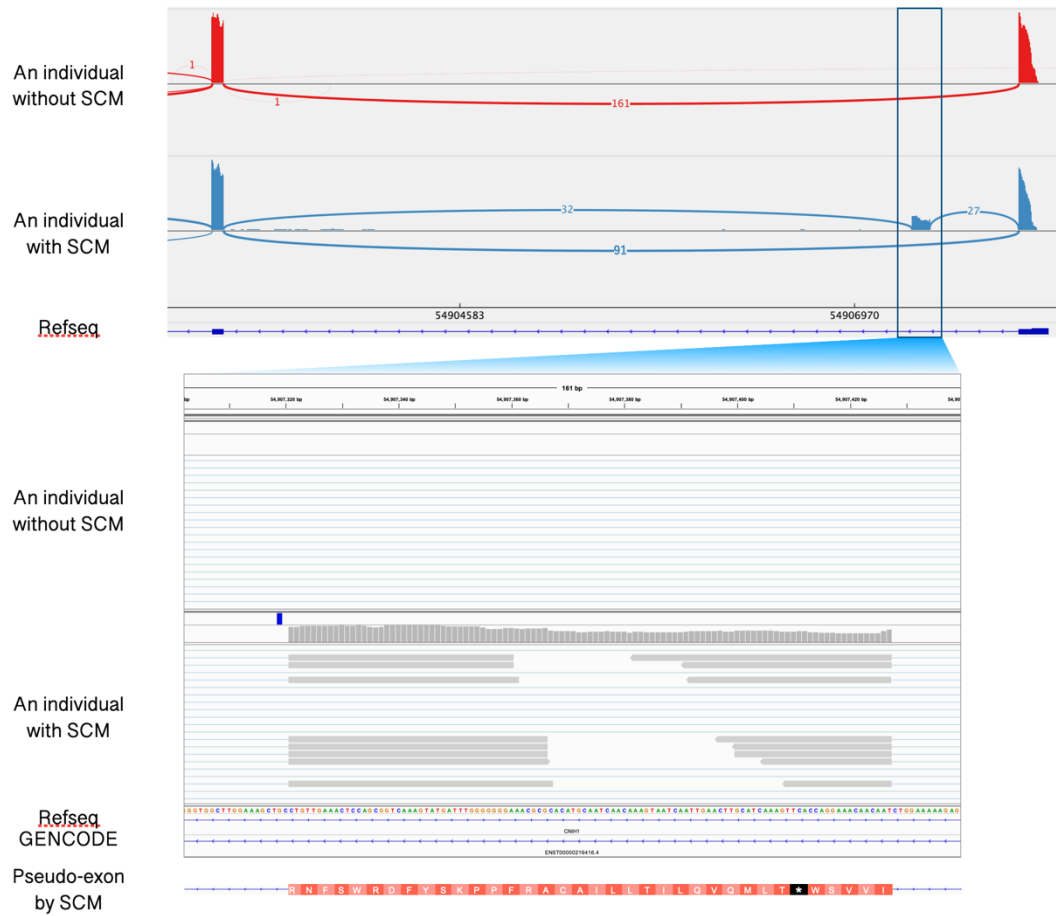


Fig. 2-5 An example of a pseudo-exon activation event identified in the first intron of the cornichon family AMPA receptor auxiliary protein 1 (*CNIH1*) by an SCM that creates a novel 5'ss.

The upper panel illustrates a Sashimi plot [101] of the activated pseudo-exon and neighboring exons. Each number represents the number of exon-exon junction reads. The bottom half shows an individual (HG00238) with SCM and activated pseudo-exon. The top half shows an individual (HG00116) without SCM for comparison. The lower panel shows a close-up view of the activated pseudo-exon.

Another example is a pseudo-exon identified in the seventh intron of cysteinyl-tRNA synthetase 2 (*CARS2*) (Fig. 2-6). In this case, the suspected causative SCM 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 three, 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 [69] shows that the pseudo-exon is inserted 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 SCM can be attributed to the strength of the splice sites of the pseudo-exon. Out of the 235 individuals analyzed, there were two individuals who shared the same SNV and activation of the pseudo-exon. The calculated SpliceAI score was 0.13, which is lower than the commonly used SpliceAI score cutoff of 0.20.

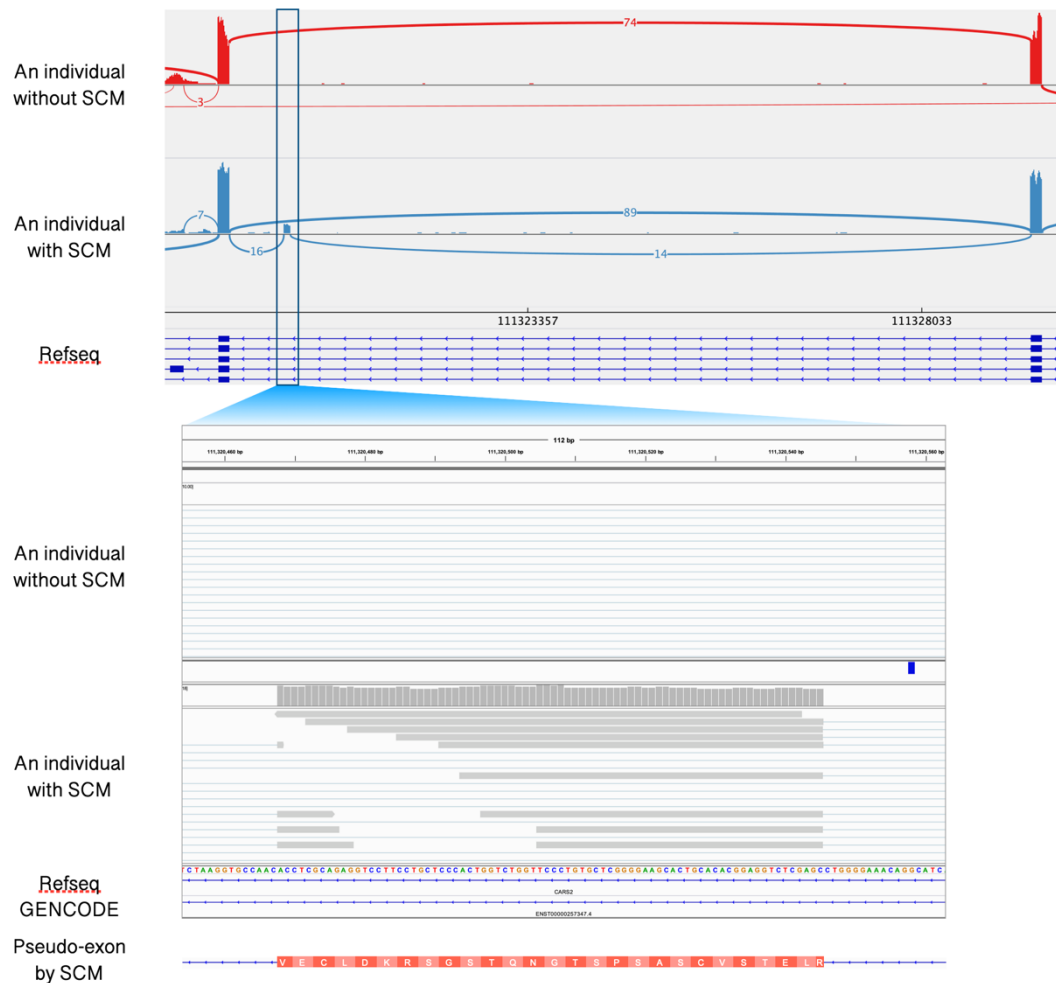


Fig. 2-6 An example of a pseudo-exon activation event identified in the seventh intron of cysteinyl-tRNA synthetase 2 (*CARS2*) by an SCM that creates a novel 3'ss.

The upper panel illustrates a Sashimi plot [101] of the activated pseudo-exon and neighboring exons. Each number represents the number of exon-exon junction reads. The bottom half shows an individual (HG00336) with SCM and activated pseudo-exon. The top half shows an individual (HG00178) without the SCM for comparison. The lower panel shows a close-up view of the activated pseudo-exon.

2.3.3 Systematic analysis of the identified pseudo-exon activation events

Of the 235 individuals analyzed in this study, the maximum number of events found in an individual was seven, and this was observed in two individuals (Fig. 2-7a). Pseudo-exon activation event was not identified 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. The maximum number of individuals that shared the same pseudo-exon was 92 as determined although most of the pseudo-exons were observed in only a single individual (Fig. 2-7b). 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 (Supplementary Table 2-2). For each of these pseudo-exons, I calculated the PSI value and counted the number of individuals in terms of whether they had homozygous or heterozygous SCMs. 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.

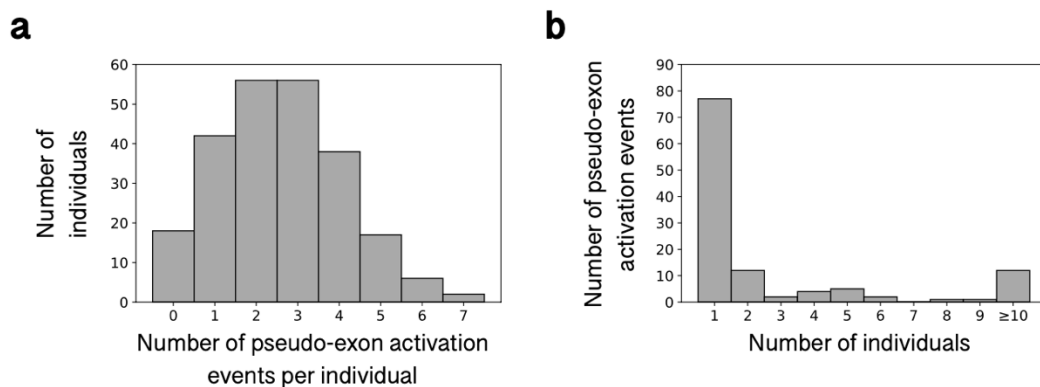


Fig. 2-7 Frequencies of the identified pseudo-exon activation events.

(a) Histogram of the number of pseudo-exon activation events per individual. (b) Histogram of the number of pseudo-exons shared among multiple individuals.

I then analyzed the length distribution of the pseudo-exons and compared this with that of exons (Fig. 2-8). The modal value was the same between pseudo-exons and exons (81–120 bases). However, the median length was found to be longer for exons, that is, 105.0 bases (pseudo-exons) vs. 130.0 bases (exons), mainly because a certain fraction of exons are much longer than the other exons, and because I set the cutoff for pseudo-exon length at $\leq 1,000$ bp. This trend is consistent with a previous report, which compiled approximately 81 cases of disease-causing pseudo-exon activation events from the literature [53].

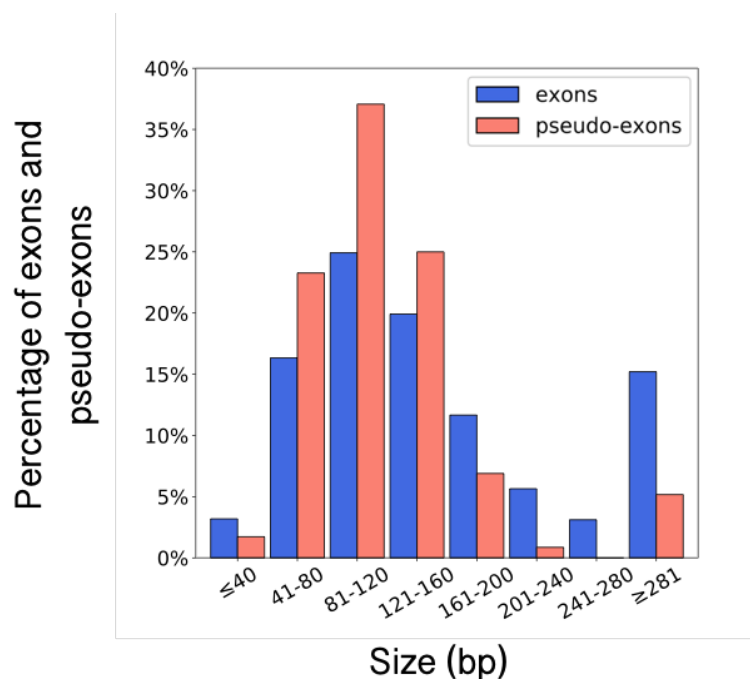


Fig. 2-8 Length distribution of the pseudo-exons.

Length distribution of exons (blue) and pseudo-exons (red).

To evaluate the effects of the SCMs that are thought to be a cause of pseudo-exon activation on the strength of the splice site, I used the MaxEntScan program [90], which quantitatively assesses whether a local sequence segment has the potential to be a splice site, to the wild-type sequence segment and the mutated one. Among the 116 distinct pseudo-exon activation events, 110 (94.8%) showed gains

in scores of more than 1.0 in the sequences with SCMs (Fig. 2-9). However, six instances showed little or no gains in scores in the sequences with SCMs. I found that these cases already have relatively high MaxEntScan scores, even in wild-type sequences. For these six instances, the Human Splicing Finder [92] was further applied 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 SCMs 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.

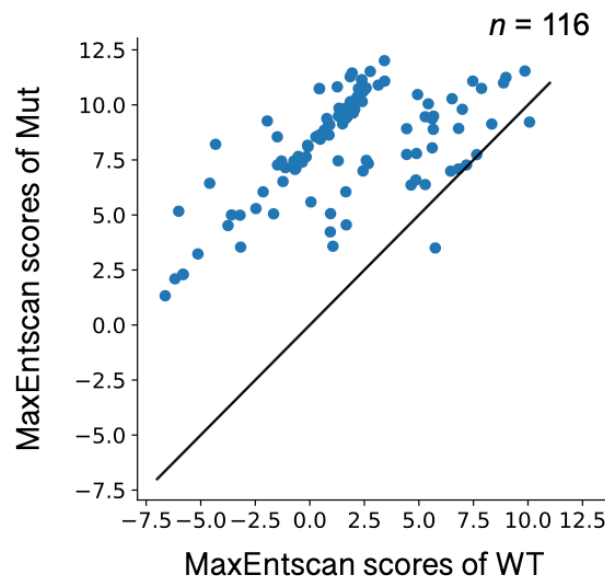


Fig. 2-9 Scatter plot of the strength of the splice site before and after the SCM for each of the 116 pseudo-exons.

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.

To evaluate the effect of the pseudo-exons inclusion on the coding potential, whether each identified pseudo-exon might disrupt the original reading frame and introduce PTC that seem to trigger NMD was analyzed. Among the 116 identified pseudo-exons, 83 were located in the coding regions. Of these, 65 pseudo-exons

were expected to introduce PTCs either in the pseudo-exons themselves (58 cases) or in their downstream region by frameshifts (7 cases). The remaining 18 pseudo-exon activation events may maintain the open reading frame and did not seem to trigger the NMD (Fig. 2-10 and Supplementary Table 2-2). For these 18 pseudo-exon activation events, their protein sequences were analyzed using the SMART database [69] to check the insert positions of the pseudo-exons in their protein domain architecture. I discovered that, in three instances, the insertions of the pseudo-exons occur within protein domains, indicating severe disruption of the domains (Supplementary Table 2-2 and Supplementary Figure 2-1).

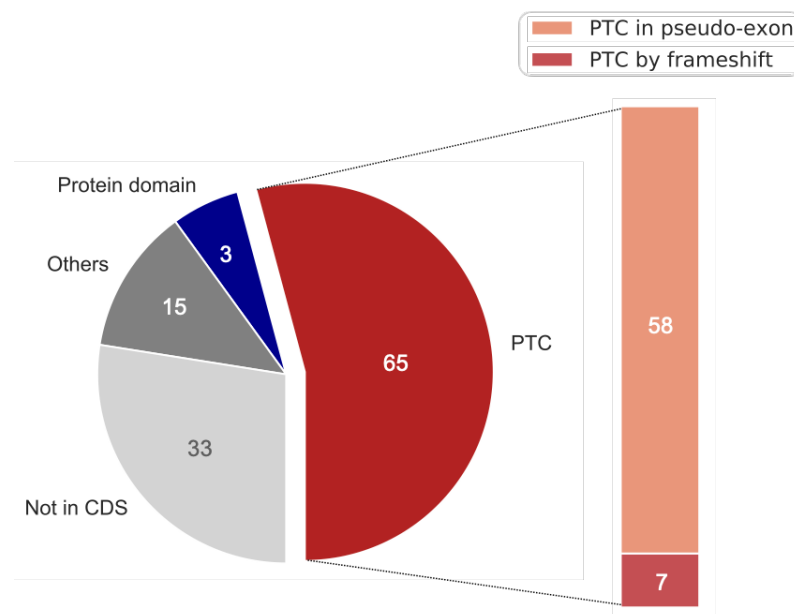


Fig. 2-10 Effect of pseudo-exons on the transcripts and proteins.

The left pie chart shows the functional consequences of the 116 pseudo-exons identified. “PTC” indicates those that seem to induce NMD by creating PTC. These were further divided into those that create PTCs by frameshifts and those that introduce in-frame PTC in the pseudo-exon. “Protein domain” indicates that the insertion of the pseudo-exon disrupts the protein domain structure. “Not in CDS” indicates that the pseudo-exons are located outside the protein-coding regions. “Others” indicates pseudo-exons that do not seem to trigger NMD or those that reside outside the protein domain structures.

Additionally, for the 116 identified pseudo-exon activation events, Gene Ontology analysis was conducted using Metascape [93]. According to the results, there were no specifically enriched functional categories for the genes with pseudo-exons. I also individually examined the function of genes with pseudo-exons that occurred in more than 10 people. For example, the pseudo-exon found in the *FANCA* gene occurred in 87 of 235 individuals. *FANCA* genes belong to the Fanconi anemia complementation group (FANC) gene family and are thought to function as a post-replication repair or a cell cycle checkpoint [102]. FANCA proteins repair interstrand DNA crosslinks and regulate the differentiation of hematopoietic stem cells into mature blood cells. The SCM found in the *FANCA* gene have some advantageous function for human and may be starting to be fixed.

The distribution of allele frequencies of the SCMs was further examined. The allele frequencies were referred to the gnomAD database of human mutation information [103]. There were 88 SCMs with allele frequencies on this database, of which 69 SCMs were allele frequencies ≤ 0.01 (Fig. 2-11 and Supplementary Table 2-2). In contrast, the maximum allele frequency among the 88 SCMs was 0.415.

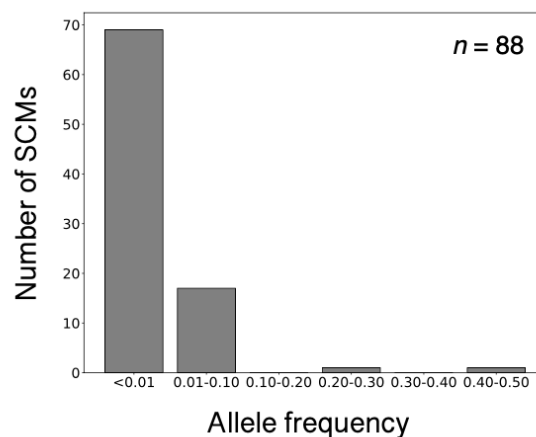


Fig. 2-11 Distribution of allele frequencies of SCMs.

Distribution of allele frequencies for the 88 SCMs obtained from gnomAD.

Additionally, I used SpliceAI [94] to see if I could assess the impact of identified SCMs on splicing. The SpliceAI score (Δ score) ranges from 0 to 1 and the closer the score is to 1, the higher the probability that the SNV will alter splicing. The SpliceAI was applied to the 111 SCMs I identified and to 500 randomly selected SNVs (Fig. 2-12). Of the 111 SCMs, 36 SCMs had Δ score ≥ 0.50 , which is a recommended cutoff of SpliceAI, and only 10 SCMs had Δ score ≥ 0.80 , which is a high precision cutoff. The median Δ score was 0.335, which was significantly higher than the score for randomly selected SNVs (Mann-Whitney U-test, $p = 4.068e^{-101}$). Next, I examined the distance between the pseudo-exon and an annotated exon. The SpliceAI is known to be difficult to predict the effect of SNVs against splicing in the introns more than 50-bp away from the exon–intron boundaries [94]. Of the 116 pseudo-exons identified in this study, the minimum distance from the annotated exon was 84-bp, the maximum distance was 132,342 bp, and the median distance was 1,792 bp. In my laboratory analysis, the median Δ score for SCMs occurring relatively close to the exon was 0.700, which was higher than the score for deep intronic SCMs identified in this study (manuscript in preparation). Therefore, this result, and the result of a previous study [94], indicate that spliceAI is not suitable for predicting the effects of splicing by deep intronic variants.

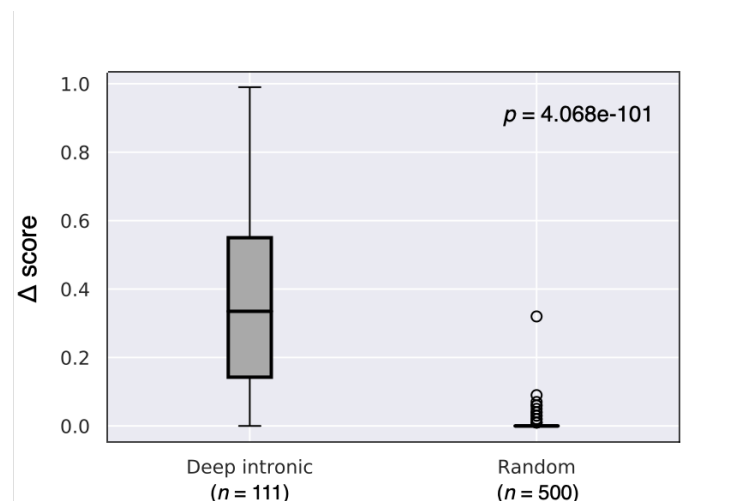


Fig. 2-12 Evaluation of SCMs that induce pseudo-exon using SpliceAI.

The SpliceAI scores (Δ score) for SCMs and randomly selected SNVs. The box plot on the left is the Δ score for the identified SCMs, and the box plot on the right is the Δ score for random SNVs.

2.3.4 Frequency and spectrum of SCMs for each site

For the 116 distinct pseudo-exon activation events, I summarized the type and frequency of SCMs (Fig. 2-13). The total number of SCMs that I identified was approximately 2.7 times higher in 5'ss (85 SCMs) than in 3'ss (31 SCMs), even though the lengths that I analyzed were longer in 3'ss because of the existence of polypyrimidine tracts. As expected, most of the SCMs were observed in the canonical dinucleotides in both 5'ss (Fig. 2-13a) and 3'ss (Fig. 2-13b). More specifically, in 5'ss, 62.4% of the SCMs were observed in the canonical GU dinucleotides, and all of these SCMs created the dinucleotide. In 3'ss, 41.9% of the SCMs were observed in the canonical AG dinucleotides, and all of these SCMs created the dinucleotide. Most of the SCMs were substitutions toward splice site motifs. For example, in dinucleotides at both 5'ss and 3'ss, all the SCMs were changed toward the bases in the canonical dinucleotides. Even when SCMs formed splice sites other than canonical splice sites, the dinucleotides in the canonical splice sites were originally GU or AG dinucleotides.

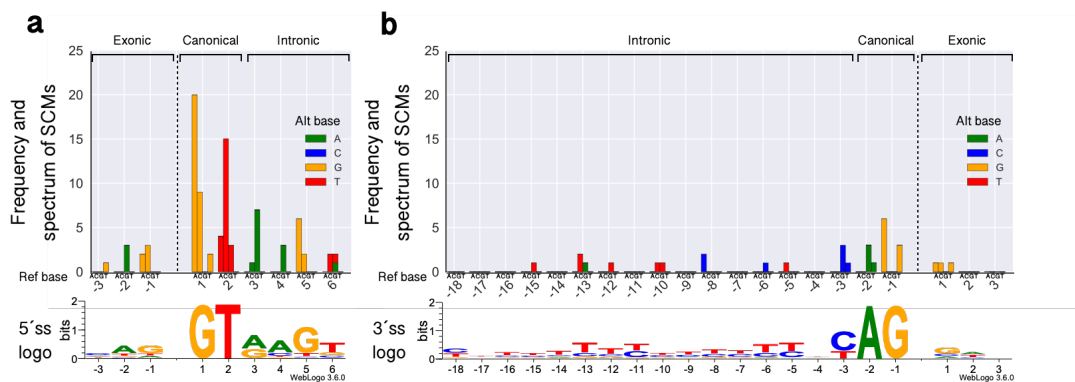


Fig. 2-13 Frequency and spectrum of SCMs involved in pseudo-exon activation events.

(a) 5'ss. (b) 3'ss. The dotted vertical line represents the exon–intron boundary. The color 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, was taken from WebLogo 3 [8].

2.4 Discussion

Pseudo-exon activation events have been mainly reported disease-causing mutations [53]. In this study, by using personal genome data [83] together with the RNA-seq data for the peripheral blood samples of the corresponding individuals [59], I successfully identified pseudo-exon activation events in normal individuals. Such events often introduce PTCs in the downstream region of the mRNA, which trigger NMD [104]. Even if such a surveillance mechanism of irregular mRNA exists, I could detect transcripts with pseudo-exons because NMD does not usually degrade all transcripts with a PTC. Indeed, it has been demonstrated that a substantial number of transcripts supposed to trigger NMD could still be detected in transcriptome data [58,59].

Although the involvement of the identified SNVs in pseudo-exon activation events is yet to be directly demonstrated, they seem to be causative SCMs because the mutation spectrum correlated well with the sequence motifs of splice sites (Fig. 2-13). Interestingly, the number of SCMs at 5'ss was found to be about 2.7 times higher than those at the 3'ss. This trend corresponds well with the known cases of pseudo-exon activation events reported so far [53], which also shows that more pseudo-exons were created by SCMs at the 5'ss than at the 3'ss; this further confirms that most pseudo-exon activation events that I identified in this study are authentic. Two possible explanations for the excess of SCMs at 5'ss can be provided. Firstly, the canonical AG dinucleotides at the 3'ss can be created by SNV of CG dinucleotides, which are under-represented in vertebrate genomes [105]. Indeed, I did not observe any SNVs of CG to AG to create 3'ss (Fig. 2-13). Secondly, it was indicated that the recognition of 5'ss is a key step in RNA splicing [106], and the number of SNVs that disrupt splicing is also high at 5'ss compared with that at 3'ss [106,107].

If the causative SCMs for pseudo-exon activation reside in exonic regions, the events and the associated SCMs can possibly be identified solely by transcriptome analysis, namely, RNA-seq, without conducted WGS. The proportion of such SCMs, 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, 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 SCM 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 [108]. From 1,188 individuals, with an average number of somatic SNVs was 22,144, analyzed in their study, they identified 46 distinct pseudo-exons. This corresponds to one pseudo-exon activation event in approximately 0.6 million SNVs. This frequency is similar to what I found in my 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). Additionally, the fact that the SNVs that cause pseudo-exon activation in cancer were enriched in the canonical splice sites is well consistent with my results.

The actual number of pseudo-exon activation events might be higher than the number that I identified in this study for the following four reasons. Firstly, I adopted rather stringent criteria for the identification of pseudo-exon activation events. For example, I considered only those cases where both sides of the pseudo-exon were covered by at least two junction reads. The correspondence between the presence of SCMs and pseudo-exons for all of the individuals was examined. Secondly, I only considered SNVs but not insertions/deletions (indels). Thirdly, I did not take the SNVs creating splicing regulatory elements (SREs), 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 SREs

[109–111], I did not consider them in this study because those elements are not always located close to the pseudo-exons, also, it is difficult to determine the correspondence between an SNV at a regulatory element and a pseudo-exon. Additionally, I focused only on pseudo-exon activation events caused by SCMs in intronic regions, although SCMs 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 [76]. The fourth reason is that I only used transcriptome data for peripheral blood samples. Because of this, what I identified as pseudo-exon activation events is limited to those genes actively transcribed in peripheral blood, and those genes unexpressed in the peripheral blood were not covered in the present study. The lowest expression level of the gene for which I identified a pseudo-exon activation event was fragments per kilobase of exon per million reads mapped (FPKM) value of 0.53. This value corresponds to approximately the 43rd percentile from the top of all genes in the genome following their expression levels. For the remaining genes (57% of all genes in the genome), I could not identify pseudo-exon activation events because of the low expression levels in peripheral blood cells. From this proportion, I can estimate that at least two times more pseudo-exon activation events (i.e., on average $5.2 = 2.6 \times 2$ instances) might exist in normal individuals. Moreover, given that there are on average five SREs per exon [112], the number of pseudo-exon activation events can further be estimated as at least twice as much, that is, more than 10 per individual. Because I have adopted rather stringent criteria in identifying pseudo-exon activation events, the actual number of pseudo-exon activation events in an individual is probably 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 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 the capture target for exome sequencing. It has been reported that the success rate of exome sequencing is approximately 25%–40% [74,113,114]. For the remaining cases, in which causative mutations have not yet been identified, some might be caused by deep intronic mutations that trigger pseudo-exon activation. In fact, of the 116 pseudo-exons identified in the study, 25 occurred within the gene responsible for genetic disorders. These 25 pseudo-exons were caused by heterozygous mutations and did not exhibit a disease phenotype, but if these SCMs are homozygous or compound heterozygous due to another harmful mutation, they are probably disease-causing mutations. My 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.

Chapter 3. Comprehensive identification for splice-site-creating mutations capable of causing genetic disorders

3.1 Introduction

Whole exome sequencing (WES) has been widely used to detect mutations causing genetic disorders in humans. However, WES identifies causative mutations in less than 50% of patients with genetic disorders [74,113,114], and many causative mutations remain unidentified. Causative mutations are detected at low rates because research to date has focused mainly on mutations that disrupt either protein-coding regions or canonical splice sites. Identifying novel causative mutations in genetic disorders will lead to better diagnosis and treatment of patients.

Based on the results of Chapter 2, I thought that it would be possible to identify mutations that may be responsible for genetic disorders by actively searching for splice-site-creating mutations (SCMs). However, as mentioned above, in the search for causative mutations of genetic disorders, WES does not identify potentially causative SCMs that occur in deep intronic regions. Therefore, in this study, I focused on SCMs occurring around exonic regions that can be detected by WES.

Recently, it was reported that mutations in intronic or exonic regions can create novel splice sites, resulting in the inclusion of abnormal exons in transcripts [53,76,77,115]. This modification often generates premature termination codons (PTCs), giving rise to truncated protein products and often leading to nonsense-mediated mRNA decay (NMD) [55]. Even without the generation of PTCs, the inclusion of abnormal exons may disrupt protein domain structures, producing diseases. However, such SCMs have only been sporadically reported as being involved in genetic disorders to date. Hence, the overall picture of the extent to which SCMs are involved as a cause of genetic disorders is still unclear.

In this study, to address the above question, I sought to identify SCMs that could be involved in various genetic disorders. I used SpliceAI, a recently developed highly accurate tool for evaluating the effect of mutations on splice site function [94].

The input data were variants registered in the gnomAD database [103]. A total of 3,942 candidate causative SCMs were identified in 4,054 genes listed in the Online Mendelian Inheritance in Man (OMIM) database [116] (<https://www.omim.org/>) as being responsible for genetic disorders. I also analyzed existing WES data from ciliopathy patients and identified 38 novel candidate causative SCMs. I further confirmed that one of the SCMs is exclusively homozygotic in multiple patients. These results indicate that it is worth considering the possibility that SCMs may be involved in genetic disorders when causative mutations, such as the disruption of coding regions or splice sites, were not found.

3.2 Materials and methods

3.2.1 Genomic variants

The genomic variant data for 71,702 individuals from the Genome Aggregation Database (gnomAD release 3.0) [103] in Variant Call Format (VCF) was downloaded. The gnomAD aggregates and harmonizes exome and genome sequencing data from various large-scale sequencing projects and makes the data widely available to the scientific community. The genomic variant data from the gnomAD were used for my analysis because it is one of the datasets that contains genomic variants, and I thought it would be serve as an adaptable data for searching for many causative mutations. A total of 564 million rare SNVs (allele frequency ≤ 0.01) were used in identifying splice-site-creating mutations (SCMs) caused by genetic disorders.

3.2.2 Creation of a list of causative genes for genetic disorders

I downloaded the Online Mendelian Inheritance in Man (OMIM) (<https://www.omim.org/>) data [116], including names of genes causing human genetic disorders. OMIM is an organization that compiles a comprehensive set of human genes and genetic phenotypes, and the data was freely available. I used gene structures registered in GENCODE [117] (V 29) in GTF. From the GTF file, I extracted transcripts corresponding to the 4,054 causative genes for genetic disorders in OMIM. If there were two or more transcript isoforms for a gene, I selected the longest transcript isoform as the representative gene structure.

3.2.3 Splice site scoring

The strength of the splice site was calculated using the MaxEntScan program [90] (see Materials and Methods of Chapter 2 for further details). I used the genome sequence segments around the SNV that might form a canonical splice site.

3.2.4 Sequence-based prediction of SCMs

SpliceAI [94] was used to detect SCMs. Each SNV in the exonic regions or in introns within 50 bp of an exon–intron boundary was evaluated. The SNVs in VCF, reference genome data (GRCh38.p12), and the gene annotation data from GENCODE (V 29) (https://www.gencodegenes.org/human/release_29.html) were used for the input of SpliceAI. The SNVs with SpliceAI scores (Δ score) ≥ 0.80 were considered SCMs.

3.2.5 Validation using RNA-seq data

To validate whether a candidate SCM indeed functioned as a novel splice site, the RNA-seq data obtained from individuals who carried the candidate SCM were used. For this validation, variant data of 670 individuals in the VCF format and RNA-seq data in the Binary Alignment Map (BAM) format from the Genotype-Tissue Expression (GTEx) project were downloaded [118]. These are all whole blood RNA-seq data. The GTEx project, launched by the National Institutes of Health (NIH) in September 2010, aims to create a resource for studying how inherited changes in genes lead to common diseases. As of December 2021, the GTEx project provides RNA-seq data from 54 non-diseased tissues and approximately 1000 individuals. The GTEx data were used because I needed a larger number of RNA-seq data to validate candidate SCMs. I mapped the RNA-seq data onto the personal genome reflecting mutation data as with Chapter 2 (see Materials and Methods in Chapter 2 for details). The RNA-seq data in the BAM format for the corresponding individual were converted to FASTQ format using Picard tools (<http://broadinstitute.github.io/picard/>).

3.2.6 Calculation of junction allele fraction

Junction allele fraction (JAF), a measure of the relative usage of the novel exon–intron junction created by an SCM compared to the annotated junction, was calculated using the following equation [76]:

$$\text{JAF} = J_n / (J_n + J_a)$$

where J_n is the number of junction reads supporting the novel splice site and J_a is the number of junction reads supporting the annotated splice site. JAF values range from 0 to 1, and the closer the value is to 1, the higher the usage of the novel junction (Figure 3-1).

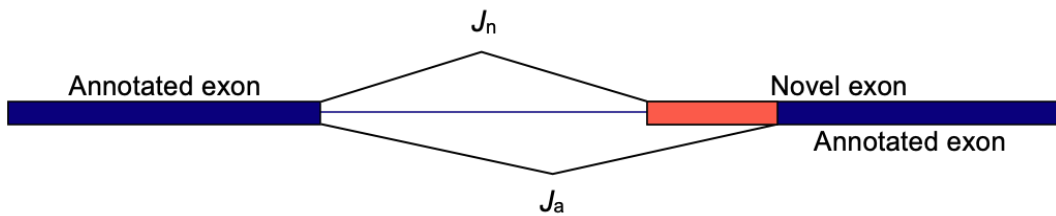


Fig.3-1 Calculation of JAF using mapped junction reads.

Junction reads that span annotated exon (Blue box) and novel exon (Red box) are denoted as J_n , and junction reads that span between annotated exons are denoted as J_a .

3.2.7 Identification of SCMs in WES data of ciliopathy patients

To analyze ciliopathies, I obtained whole exome sequencing (WES) data of 2,569 patients from the Database of Genotypes and Phenotypes (dbGaP) [119] (dbGaP Study Accession: [phs000288.v2.p2](#)). These data were obtained from inbred families where the parents were the first- or second-degree cousins, and there were two or more affected family members. The data included 1,773 affected individuals and 796 individuals who were not affected but had two or more affected individuals in their family members. Of the variants obtained from these patients, 1,401,863 SNVs located on autosomes and the X chromosome were used in this study. A list of 434 ciliopathy-related genes [120–134] was generated to identify causative SCMs in

these genes (Supplementary Table 3-1). According to a previous study [120], these genes are classified into two groups: the genes labeled as “established” are those for which the causative mutations have been identified, and the genes labeled as “candidate” are those for which the causative mutations have not been identified but which have been implicated in ciliopathies in other species or at the cellular level. I evaluated all SNVs identified by WES for their potential to be SCMs using SpliceAI with a Δ score cutoff of 0.80. The SCMs with allele frequencies greater than 0.01 or those homozygous in normal individuals in gnomAD were excluded. To analyze the ciliopathies, GRCh37.p13 was used as the reference genome and GENCODE (V24) as the gene annotation data (https://www.gencodegenes.org/human/release_24lift37.html). For each gene, all transcript isoforms were analyzed for possible pathogenicity of the SCMs.

3.2.8 Analysis of protein domain architecture

For the 2,167 novel exons that do not seem to be the target of NMD—those that do not disrupt the amino acid sequences either by PTCs or by frameshifts—I examined whether the extension or shrinkage of the sequence by SCMs should affect the protein domains using the hmmscan program [135]. The hmmscan program is a tool for searching protein sequences against a protein profile database and can be run from the command line. An E-value cutoff of 0.001 was adopted. Protein domain architecture was visualized using the SMART database [69].

3.2.9 Modeling a three-dimensional protein structure

To investigate the effect of shrinkage of the exon, which codes a part of the differentially expressed in normal and neoplastic cells (DENN) domain of DENN domain containing 4A (*DENND4A*), on a three-dimensional (3D) protein structure, a 3D model structure of the DENN domain was constructed using SWISS-MODEL [136] and AlphaFold2 [137]. The domain sequence obtained from the SMART database was used as the input data. The 3D model structure was visualized using PyMOL [138] (version 2.1).

3.2.10 Data visualization and manipulation

The IGV [95] was used to visualize the RNA-seq read mapping and the variant data. The UCSC Genome Browser [84] was used to visualize the gene annotation data. Gene expression levels were quantified using the StringTie program (version 1.3.5) [88]. SAMtools (version 1.5) [96] and BEDTools (version 2.26.0) [97] were used for data manipulation. All used public data and tools in my analysis were listed (Table 3-1 and Table 3-2). All analyses were conducted under Python (version 3.7.8), Java (10.0.2), Perl (v5.32.0), and GNU Bash (version 4.2.26) environments.

Table 3-1. List of data used in the study of Chapter 3.

Data	Number of samples	Format	Database	URL
Variant data	71,102	VCF	gnomAD	https://gnomad.broadinstitute.org/
OMIM data	1	txt	OMIM	https://www.omim.org/
Gene annotation data (V29)	1	GTF	GENCODE	https://www.encodegenes.org/human/release_29.html
Reference genome data (GRCh38.p12)	1	FASTA	GENCODE	https://www.encodegenes.org/human/release_29.html
RNA-seq	670	BAM	GTEx project (via dbGaP)	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2
Variant data (Ciliopathy)	2,569	VCF	dbGaP	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000288.v2.p2
Gene annotation data (V24)	1	GTF	GENCODE	https://www.encodegenes.org/human/release_24lift37.html
Reference genome data (GRCh37.p13)	1	FASTA	GENCODE	https://www.encodegenes.org/human/release_19.html

Table 3-2. List of tools used in the study of Chapter 3.

Tools	Version
MaxEntScan	-
SpliceAI	1.3.1
BCFtools	1.9
Picard	
HISAT2	2.1.0
IGV	2.4.16
UCSC Genome Browser	-
hmmscan	3.3
SWISS-MODEL	-
SMART	-
PyMOL	2.1
StringTie	1.3.5
SAMtools	1.5
BEDTools	2.26.0

3.3 Results

3.3.1 Identification of splice-site-creating mutations in the causative genes of genetic disorders

To understand the extent to which splice-site-creating mutations (SCMs) affect the pathogenesis of genetic disorders, I sought to identify mutations that might shrink or extend an exon in 4,054 genes known to be responsible for genetic disorders (Fig. 3-2a). SCMs were identified using the following steps (Fig. 3-2b). Firstly, a list of genes involved in genetic disorders was created. From the single-nucleotide variants (SNVs) registered in the gnomAD database (v3.0), only those located in the 4,054 genes were selected. The gnomAD database contains more than 800 million SNVs with various allele frequencies and is one of the most comprehensive databases of genetic variation in humans. This step yielded 72,567,596 SNVs. I further selected the SNVs located in intronic regions within 50 bp of an annotated exon–intron boundary or in exonic regions. I introduced this positional cutoff for the intronic SNVs because these genomic intervals are the limits using conventional WES procedures for identifying mutations [75]. This filtering produced 4,215,150 SNVs. Next, I selected those SNVs that create canonical splice sites, namely, GT or AG, followed by further selection based on the MaxEntScan scores [90]. The MaxEntScan is a computational method for evaluating the strength of a splice site. Then, I selected those SNVs that had a MaxEntScan score ≥ 0 , which is a rather relaxed threshold. This cutoff was adopted to reduce the number of false negatives. Although the cutoff also increases the number of false positives, such cases can be filtered out in the following steps. I then selected only those SNVs with allele frequencies ≤ 0.01 in the gnomAD database. I set this condition to obtain SNVs that could serve as causative variants for genetic disorders. After filtering, the number of SNVs was reduced to 268,836. These SNVs were further evaluated for their potential to be SCMs using the SpliceAI program, which is a highly accurate splice site prediction method. From the distribution of SpliceAI score (Δ score) for the 268,836 SNVs, I found that the recommended cutoff of 0.80 for accurate prediction of splice sites [94] was located almost at the minimum count of SNVs (Fig. 3-2c). From this observation, I adopted

this cutoff and obtained 5,656 SNVs that were potential SCMs (see paper published in npj Genomic Medicine).

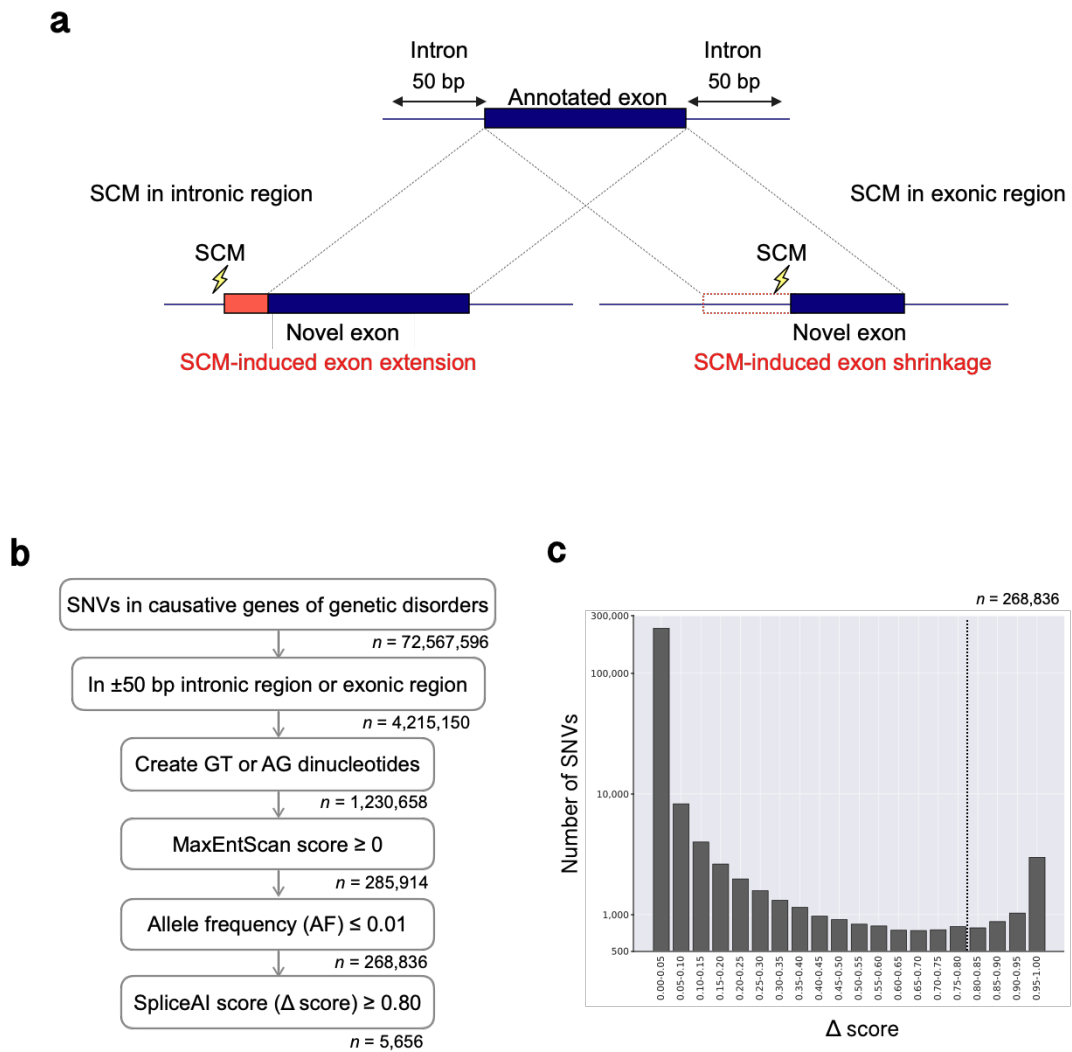


Fig. 3-2 Identification of splice-site-creating mutations in 4,054 genes responsible for genetic disorders.

(a) Schematic of novel exons by splice-site-creating mutations (SCMs). An exon extension by an SCM in the intronic region (left) and an exon shrinkage by an SCM in the exonic region (right). (b) Workflow for the identification of SCMs. Each number represents the number of SNVs at each step. (c) Distribution of SpliceAI scores (Δ scores) for the 268,836 SNVs. The dotted line indicates the cutoff score (0.80) that I adopted for the potential SCMs.

3.3.2 Estimation of sensitivity, specificity, and positive predictive value in identifying SCMs

To evaluate the quality of the potential SCMs identified as described above, the potential SCMs were compared to those already identified. To create a set of known SCMs, the ClinVar database was used [139]. Focus was placed on intronic and synonymous exonic SNVs that do not disrupt an annotated splice site and those that create canonical splice sites because these SNVs are more likely to be SCMs if they are pathogenic. Precisely, I selected those SNVs that (1) were classified as “pathogenic” or “likely pathogenic” in ClinVar, (2) and were located in intronic regions within 50 bp of annotated exon–intron boundaries or synonymous sites of exonic regions, and (3) created either GT or AG and did not disrupt the annotated splice sites. There were 775 potential SCMs in ClinVar that met these criteria. Among them, 72 SNVs were within the gene bodies of the 4,054 genes responsible for genetic disorders and overlapped with the variants registered in gnomAD. Hence, these 72 SNVs should be known SCMs that my pipeline should detect. Of these 72 SNVs, my pipeline was capable of identifying 49 SNVs as SCMs. This means that the sensitivity of my pipeline is 68.1% (49 / 72).

To calculate the specificity, firstly, I artificially introduced the 5,656 potential SCMs to the reference genome sequences. Then, RNA-seq data from individuals without these SCMs were mapped onto the artificially mutated reference genome. For this, I used data of 10 individuals that have relatively high expression of the genes containing the potential SCMs. This analysis found only three false positives out of the 5,656 potential SCMs, indicating the specificity to be 99.9% (5,653 / 5,656).

I also measured the positive predictive value (PPV), which is, in this case, the value representing how many of the identified potential SCMs forms novel splice sites. I used whole blood RNA-seq data of 670 individuals obtained from Genotype-Tissue Expression (GTEx) [118] (<https://gtexportal.org/home/>). Although mRNAs are thought to be degraded by NMD if an alteration of an exon by an SCM generates premature termination codons (PTCs), it was reported that a substantial number of transcripts supposed to trigger NMD could be detected in RNA-seq data [58,59].

Among the 5,656 potential SCMs, one or more individuals had the mutant allele for 114 SCMs. Of these SCMs, gene expression was confirmed for 53 genes in the RNA-seq data from individuals with the mutant allele. I considered a locus to be expressed if there were at least five junction reads that covered either the novel junction created by the SCM or the annotated junction at that locus. The cutoff was adopted from previous studies [140,141]. Of these 53 loci, 40 had junction reads that supported the novel junctions created by the SCMs. For these 40 loci, I further confirmed that these are true positives using the RNA-seq data of all individuals without the SCMs as the negative control. Accordingly, the PPV was calculated to be 75.4% (40 / 53). I also compared the strength of the annotated and newly created splice sites in these 40 loci and the newly splice sites had a higher MaxEntScan score than the annotated splice sites in 21 loci. The remaining 19 loci were confirmed to be used as splice sites, although the MES scores of the newly splice sites were not higher than the annotated splice sites.

3.3.3 Examples of novel exons induced by SCMs

One of the potential SCMs was found in the intronic region of the pericentrin (*PCNT*) gene, which causes microcephalic osteodysplastic primordial dwarfism type II (MOPD2) [142]. MOPD2 is a rare disease and is characterized by small head size, short stature, and skeletal abnormalities. No cure has been established. There was one individual who was heterozygous for this variant in GTEx (GTEx-14BMV). From the RNA-seq data from this individual, it was confirmed that the variant was indeed an SCM that induces an exon extension (Fig. 3-3). As described in the previous section, such an exon extension was not observed in all the individuals without the SCM. The SCM is an A-to-G transition, which forms the first base of the canonical dinucleotide at the 5' splice site (5'ss). The novel 5'ss is formed 41 bp downstream of the annotated original 5'ss. Although this locus contains two annotated isoforms, the SCM does not seem to affect the splicing of the shorter isoform because its exon–intron boundary is located 278 bases apart from the SCM. Additionally, only a few junction reads that support this isoform in both individuals with and without the

SCM exists. According to gnomAD, the allele frequency of this SNV is $1.39505e-05$. The JAF value, which measures the relative usage of the novel splice site based on RNA-seq reads [76], was calculated to be 0.53. The extended exonic region introduces an in-frame PTC, hence, it may trigger NMD. To check if the gene expression decreased as a result of degradation of the transcripts by NMD, the normalized expression levels were compared between the individual with the SCM and 100 randomly selected individuals without the SCM and found that the expression level did not reduce from those without the SCM. Therefore, the SCM, which introduce a PTC in the transcript, might be detrimental due to the production of truncated proteins by the PTC rather than the reduction of the expression by NMD. The SNV has not been reported as a causative factor for MOPD2 so far. My results suggest that the SNV might be a novel causative variant for MOPD2, an autosomal recessive monogenic disorder, through exon extension by the SCM, creating a novel 5'ss.

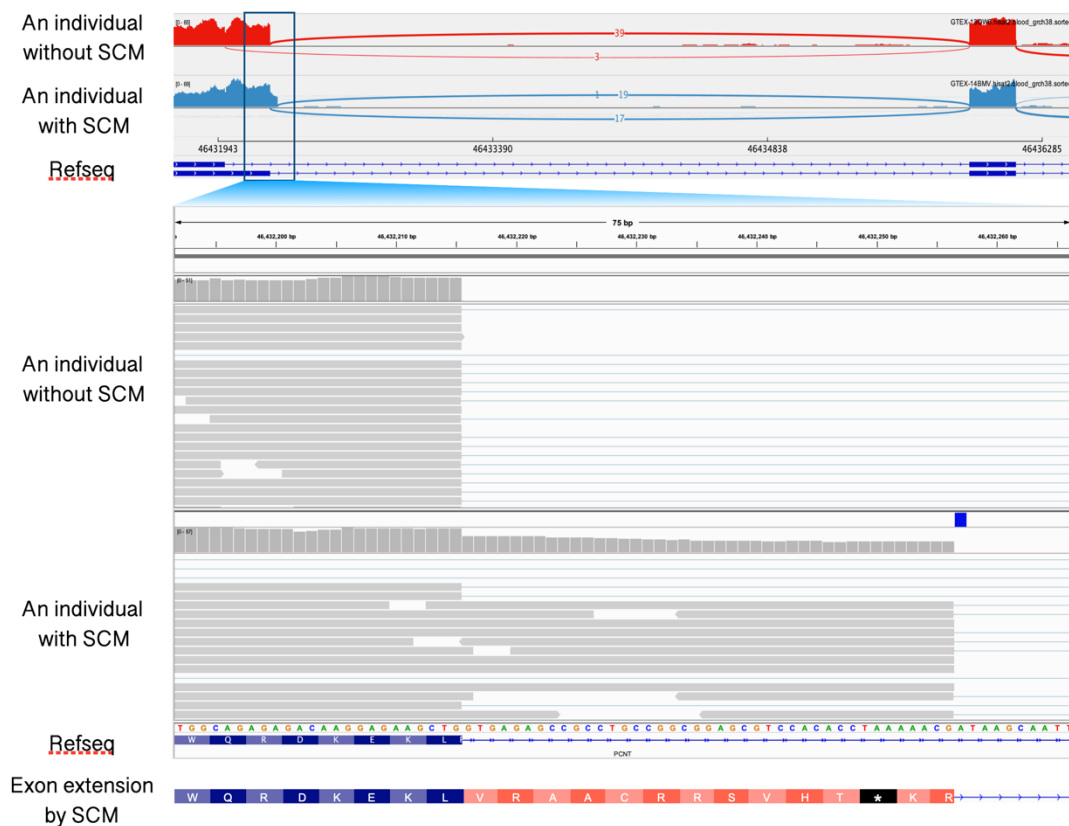


Fig. 3-3 Exon extension by an SCM in intron 38 of pericentrin (*PCNT*).

The upper panel illustrates Sashimi plots [101] of the extended exon and the downstream exon. The individuals without and with the SCM are shown in red and blue, respectively. Each number represents the number of the junction reads. The lower part of this panel is a close-up view of the genomic region around the SCM. The SCM is indicated at the top of the individual in a blue rectangle.

The next example is an SCM found in the exonic region of the acid alpha-glucosidase (*GAA*) gene, which was reported to be a causative gene of Pompe disease, an autosomal recessive disorder caused by an abnormal accumulation of glycogen in the lysosomes [143]. Symptoms of the disease include muscle weakness, cardiac dysfunction, and developmental delays. In GTEx, one individual (GTEx-1QCLY) who was heterozygous for this variant was found. From the RNA-seq data for this individual, it was also confirmed that the variant works as an SCM that induces exon shrinkage (Fig. 3-4). As described in the previous section, such an exon shrinkage was not observed in all the individuals without the SCM. The SCM is a G-to-A transition, which forms the first base of the canonical dinucleotide at the 3' splice site (3'ss). The novel 3'ss is formed 35 bp downstream of the annotated original 3'ss, leading to a frameshift in the downstream exons. The allele frequency reported in gnomAD was 0.00099162. The JAF value was calculated as 0.25. Although the variant was annotated as synonymous using Annotate Variation (ANNOVAR) [144] and was reported to be "benign/likely benign" or "uncertain significance" in ClinVar, it might be highly damaging because of the frameshift due to the SCM-induced exon shrinkage, which introduce a PTC. In this example, there was also no reduction of the expression in the individual with the SCM, suggesting that the transcript with the PTC will lead to the production of truncated proteins. A pathogenic variant has been reported in ClinVar within the region of the exon shrinkage (Supplementary Figure 3-1), and there are also other pathogenic mutations in its downstream exons, further supporting the possible involvement of the SCM in the pathogenesis of Pompe disease. The earlier the onset of the disease,

the more severe the symptoms and the more rapidly the disease progresses. However, there is a treatment for Pompe disease called enzyme replacement therapy, which replenishes the missing enzyme by administering the enzyme preparation through intravenous infusion, and this improves symptoms. This therapy must be initiated before or immediately after onset. Thus, requiring prompt diagnosis and identification of the causative mutation.

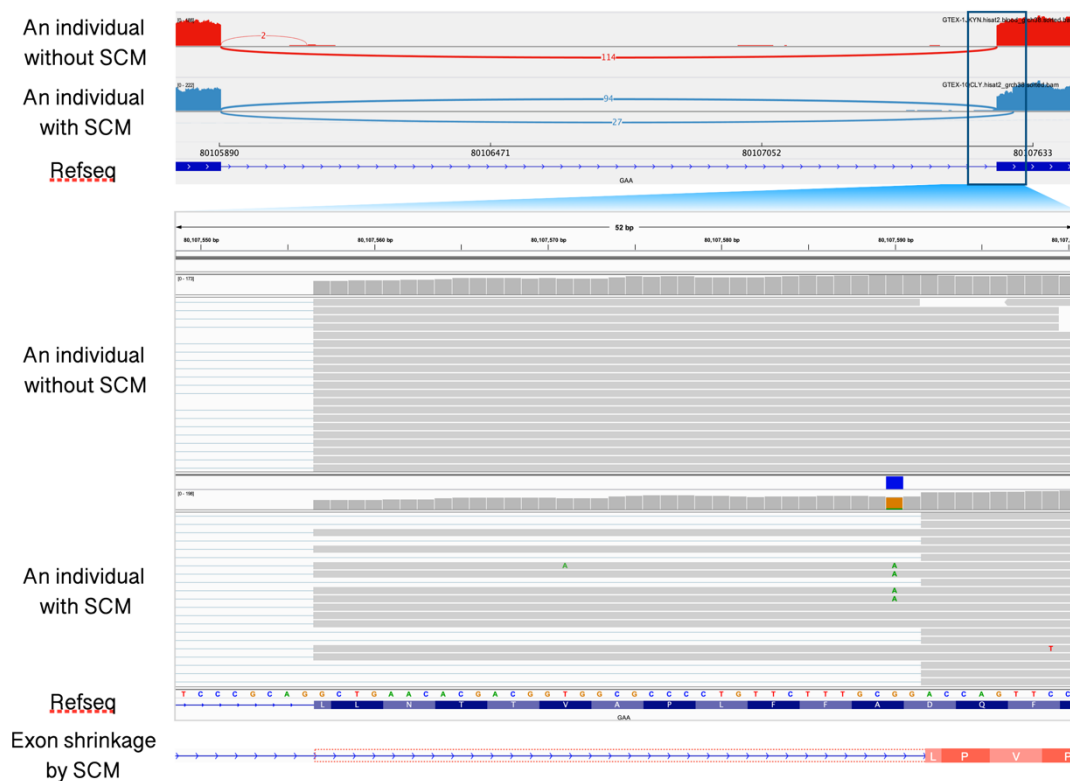


Fig. 3-4 Exon shrinkage by an SCM in exon 5 of acid alpha-glucosidase (GAA).

The upper panel illustrates Sashimi plots [101] of the shrunken exon and the upstream exon. The individuals without and with the SCM are shown in red and blue, respectively. Each number represents the number of junction reads. The lower part of this panel is a close-up view of the genomic region around the SCM. The SCM is indicated at the top of the individual in a blue rectangle and indicated in the sequencing reads as mismatched residues.

3.3.4 Structural characteristics of the identified SCMs

The structural characteristics of the potential SCMs identified in this study, such as how they were distributed in terms of gene structures were analyzed. Of the 5,656 SCMs, 2,218 (39.2%) created 5'ss and 3,438 (60.8%) created 3'ss. In intronic regions, 2,937 (51.9%) potential SCMs were found, and 2,719 (48.1%) were found in exonic regions. For each of the 5'ss and 3'ss, the positional frequencies of the SCMs relative to the annotated splice junctions and the mutational spectrum (Fig. 3-5) were further analyzed. In the 5'ss, SCMs were concentrated in the exonic regions at -3 and -2 (Fig. 3-5a). The dominant alternate bases were G and T at these sites, respectively, indicating that these substitutions created GYNGYN motifs [145], often observed in 5'ss. A comparatively smaller number of SCMs were observed in the intronic region next to the annotated exon–intron boundaries, in which the sequence motif for the 5'ss is located. In the rest of the regions, SCMs are fairly and uniformly distributed without positional preferences. In the 3'ss, the most frequent position for SCMs was position -1 in the intronic region, where the second base of the canonical dinucleotide is located (Fig. 3-5b). Some degree of concentration of SCMs were also observed at positions -5 and -4, with the dominant alternate base being A and G, respectively, creating NAGNAG motifs [146]. In contrast to the 5'ss, SCMs were rarely observed in exonic regions in 3'ss except for the positions closer to exon–intron boundaries. This might be because both a canonical dinucleotide and a polypyrimidine tract are required to function as a novel 3'ss. In the case of SCMs in the intronic regions of 3'ss, an already existing polypyrimidine tract can be utilized to function as a novel 3'ss. This requirement for polypyrimidine tract in 3'ss might explain why the number of SCMs gradually decreased with distance from the original 3'ss.



Fig. 3-5 Positional frequency and spectrum of SCMs.

(a) 5'ss. (b) 3'ss. The positions of SCMs within 20-bp upstream and downstream of annotated splice sites, respectively, are shown in each panel. The dotted vertical line indicates the annotated original exon–intron boundary. The color codes for alternative bases are shown on the right side of each panel. Note that in each panel, there are only two alternative bases because only variants that create GT (for 5'ss) or AG (for 3'ss) were considered. The SCMs that are more than 20-bp distant from the annotated original exon–intron boundary are not shown. The data for the information content at each splice site were taken from WebLogo 3 [8].

3.3.5 Functional consequences of the identified SCMs and their possible contribution to pathogenesis

The possible involvement in the pathogenesis of the potential SCMs identified in the 4,054 genes known to be responsible for genetic disorders was systematically analyzed. Firstly, the identified SCMs were annotated in terms of genic regions, such as intronic and nonsynonymous, using ANNOVAR [144] (Table 3-3). Then, they were further divided into seven subcategories based on the classification in ClinVar, which reports the associated phenotypes for the variants [139], to see whether the SNVs identified as SCMs are already classified as pathogenic (Table 3-3). Only 149 SNVs (2.6%) identified as SCMs are classified as either “pathogenic” (104 SNVs) or “likely pathogenic” (45 SNVs) in ClinVar, and for the rest of the SNVs, most (4,986 SNVs) are even unregistered in ClinVar and have not yet been reported as pathogenic.

Table 3-3. Classification of the 5,656 SCMs using ANNOVAR and ClinVar.

ClinVar	In intronic	Nonsynonymous	Synonymous	Stopgain	ncRNA	5'UTR	3'UTR	Total
Not in ClinVar	2,640	1,207	709	160	1	222	47	4,986
Benign	12	5	18	0	0	0	0	35
Likely Benign	26	9	17	0	0	0	1	53
Uncertain Significance	142	141	56	1	0	3	1	344
Pathogenic	58	23	5	18	0	0	0	104
Likely Pathogenic	31	6	2	6	0	0	0	45
Others	35	27	26	1	0	0	0	89
Total	2,944	1,418	833	186	1	225	49	5,656

As the pathogenicity of most of the SCMs is unclear, to investigate the functional consequences of the mutations, I examined whether the length alteration of exons induced by the SCMs might introduce PTCs that seem to trigger NMD or disrupt protein domain structures (Fig. 3-6). Among the 5,656 exons, whose lengths were altered by the SCMs, 466 had altered portions outside the protein-coding regions. In total, there were 3,023 SCMs (53.4%) that introduce PTCs either by frameshift (1,385 cases) or in the extended exon (1,638 cases). Using hmmscan, a tool for assigning annotated protein domains in amino acid sequences [135], also, 919 exons (16.2%) in which the altered portions were located within annotated protein domains were identified. Therefore, 3,942 SCMs (69.7%) seemed to be associated with disorders, i.e., potentially pathogenic, either by introducing PTC or disrupting protein domains.

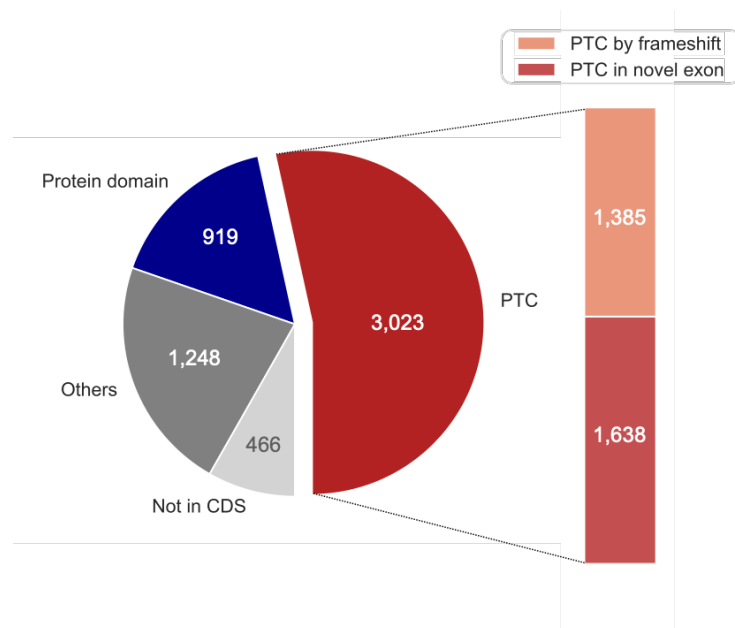


Fig. 3-6 Effect of SCM-induced length alterations on the transcripts and the proteins.

The left pie chart shows the functional consequences of the 5,656 SCMs identified. “PTC” indicates those that seem to induce NMD by creating PTC. These are further divided into those that create PTCs by frameshifts and those that introduce in-frame PTC in the extended part of the affected exon. “Protein domain” indicates that the coding alteration by the SCM disrupts the protein domain structure. “Not in CDS” indicates that the SCMs are located outside of the protein-coding regions. “Others” indicates SCMs that do not seem to trigger NMD or those that reside outside the protein domain structures.

I also conducted an analysis of the effect of SCMs on the transcripts using RNA-seq data that were obtained from the SCM containing samples in GTEx. For this, I focused on the 53 loci that I used to calculate PPV in the earlier section. Of these 53 loci, 29 seem to trigger NMD. For each of the 29 loci, I compared the normalized gene expression levels between the individual with the SCM and 100 randomly selected individuals without the SCM to see if the gene expression is reduced as a result of NMD. The result shows that 13 (44.8%) of them have lower expression levels compared to the median values of the individuals without the

SCMs, although I could not assess their statistical significance because of the limited number of individuals for each SCM. Even if NMD is not efficiently working on the transcript with a PTC, it will lead to the production of truncated proteins by the PTC, which could have functional consequences.

To further evaluate the contribution of SCMs to the pathogenesis of genetic disorders, the number of potentially pathogenic SCMs was compared with the number of known pathogenic mutations, most of which were identified as nonsense or missense mutations, or those that disrupt existing splice sites. There are 45,505 mutations in ClinVar labeled as “pathogenic” in the 4,054 genes responsible for genetic disorders. Among the 3,942 potentially pathogenic SCMs, 60 overlapped with pathogenic mutations registered in ClinVar. Although these overlapping mutations are annotated as “pathogenic” in ClinVar because they either are nonsynonymous or disrupt existing splice sites, they might be pathogenic because of their splice-site-creating capability. The rest of the SCMs are not registered in ClinVar, indicating that 3,882 more mutations can be added, corresponding to 8.5% ($3,882/45,505$) of the known pathogenic mutations, as novel potential causes of genetic disorders. Even if the “likely pathogenic” mutations were included in this calculation, the total number will be 65,570 together with “pathogenic” mutations, the number of SCMs corresponds to 5.9% of these mutations.

3.3.6 Identification of SCMs possibly involved in the pathogenesis of ciliopathies from WES data

The above analyses demonstrate the importance of SCMs as a cause of genetic disorders. Although WES analysis has been widely applied to identify causative mutations in genetic disorders, it is often difficult to identify mutations. In some cases, SCMs, which have not been actively considered causes of disorders, can be involved in pathogenesis. Therefore, I searched for SCMs in WES data in which causative mutations are yet to be identified. I focused on ciliopathies because WES data from a large cohort of patients and their unaffected family members are available in dbGaP [119] (dbGaP Study Accession: [phs000288.v2.p2](#)). Ciliopathies are human disorders in which genes related to primary cilia or motile cilia are mutated. Various phenotypes associated with ciliopathies have been reported throughout the body [120].

I identified potential SCMs by applying SpliceAI with a cutoff of 0.80 to the variant data obtained from the WES analysis of patients with ciliopathies and their unaffected family members downloaded from dbGaP. I restricted my search space for the SCMs to 434 genes for which involvement in ciliopathies was already suggested [120–134] (Supplementary Table 3-1). From these potential SCMs, I selected only those that introduced PTC or disrupted domain structures and then excluded those mutations with $AF \geq 0.01$ or homozygous in normal individuals. I finally obtained 38 potential causative SCMs (Table 3-4).

Table 3-4. List of 38 potential causative SCMs that may be responsible for ciliopathies.

Nucleotide change	Amino acid change	AF in gnomAD	ss	PTC	Protein domain disruption	Homozygous	Heterozygous	CH variants	Gene	Inheritance	Ciliopathy gene
						No. individuals (Affected)	No. individuals (Affected)				
ENST00000378888.9:c.408C>T	ENSP00000368166.5:p.Gly136Gly	-	5'ss	No PTC	Dishevelled	0 (0)	1 (1)	-	<i>DVL1</i>	AD	Established ^b
ENST00000366637.7:c.1269-2A>G	-	-	3'ss	PTC in novel exon	-	0 (0)	1 (0)	-	<i>DISC1</i>	Unknown	Candidate
ENST00000272321.11:c.2079-20A>G	-	-	3'ss	PTC in novel exon	-	0 (0)	1 (0)	-	<i>WDPCP</i>	AR	Established
ENST00000331683.9:c.1527+36A>G	-	0.0007296	5'ss	PTC by frameshift	-	0 (0)	8 (6)	-	<i>SPAG16</i>	AD	Established ^b
ENST00000295709.7:c.3058-16C>G	-	-	3'ss	No PTC	HEAT,HEAT_2,HEAT_EZ,Cnd1	0 (0)	1 (1)	-	<i>STK36</i>	AR	Established ^b
ENST00000440121.1:c.2197-2A>G	-	0.00006569	3'ss	PTC in novel exon	-	0 (0)	1 (1)	-	<i>TTC21A</i>	Unknown	Candidate
ENST00000301831.8:c.1349-20T>G	-	0.000486	3'ss	PTC in novel exon	-	0 (0)	3 (1)	-	<i>ULK4</i>	Unknown	Candidate
ENST00000420323.6:c.4086+27G>T	-	0.00001314	5'ss	PTC by frameshift	-	0 (0)	1 (1)	-	<i>DNAH1</i>	AR	Established
ENST00000420323.6:c.9585G>T	ENSP00000401514.2:p.Gly3195Gly	-	5'ss	PTC by frameshift	-	0 (0)	1 (1)	-	<i>DNAH1</i>	AR	Established
ENST00000330852.9:c.76-9A>G	-	0.00001977	3'ss	PTC in novel exon	-	0 (0)	2 (1)	-	<i>RAB28</i>	AR	Established
ENST00000274192.6:c.654C>T	ENSP00000274192.5:p.Gly218Gly	0.00007229	5'ss	Do not trigger NMD	Steroid_dh, DUF1295	0 (0)	2 (1)	-	<i>SRD5A1</i>	Unknown	Candidate
ENST00000265104.4:c.10593T>A	ENSP00000265104.4:p.Ser3531Ser	0.00007882	3'ss	No PTC	AAA_9	0 (0)	1 (1)	-	<i>DNAH5</i>	AR	Established
ENST00000356031.7:c.3621-3A>G	-	0.0000657	3'ss	PTC in novel exon	-	0 (0)	2 (2)	-	<i>SPEF2</i>	AR	Established ^b
ENST00000356971.3:c.473A>G	ENSP00000349458.3:p.Asp158Gly	-	5'ss	PTC by frameshift	-	0 (0)	1 (0)	-	<i>ICK</i>	AR	Established
ENST00000440747.5:c.1336-8G>A	-	-	3'ss	PTC by frameshift	-	0 (0)	2 (1)	-	<i>DNAAF5</i>	AR	Established
ENST00000404984.5:c.1591-16T>A	-	-	3'ss	No PTC	IQ	1 (1)	0 (0)	-	<i>IQCE</i>	AR	Established ^b
ENST00000409508.7:c.4255-5A>G	-	0.00001315	3'ss	PTC in novel exon	-	0 (0)	1 (0)	-	<i>DNAH11</i>	AR	Established
ENST00000262210.9:c.3206-13G>A	-	-	3'ss	PTC in novel exon	-	3 (3)	1 (0)	-	<i>CSPP1</i>	AR	Established
ENST00000242317.8:c.1309C>T	ENSP00000242317.4:p.Gln437Ter	-	5'ss	PTC by frameshift	-	0 (0)	1 (1)	-	<i>DNAI1</i>	AR	Established
ENST00000297814.6:c.2240-7G>A	-	0.000711	3'ss	PTC in novel exon	-	0 (0)	3 (2)	-	<i>KIF27</i>	Unknown	Candidate
ENST00000305242.9:c.820-1G>A	-	0.0000987	3'ss	PTC in novel exon	-	0 (0)	2 (1)	-	<i>ARMC4</i>	AR	Established
ENST00000534099.5:c.1407G>A	ENSP00000434400.1:p.Pro469Pro	0.00009203	3'ss	No PTC	Tub	0 (0)	2 (1)	-	<i>TUB</i>	AR	Established
ENST00000546141.5:c.101-9G>A	-	0.00001316	3'ss	PTC by frameshift	-	0 (0)	1 (1)	-	<i>GLI1</i>	Unknown	Candidate
ENST00000411698.6:c.1504+9C>T	-	-	5'ss	PTC by frameshift	-	0 (0)	1 (1)	-	<i>MDM1</i>	Unknown	Candidate
ENST00000553106.5:c.1066-11G>A	-	0.0003024	3'ss	No PTC	Biopterin_H	0 (0)	4 (2)	-	<i>PAH</i>	Unknown	Candidate
ENST00000409039.7:c.10213A>G	ENSP00000386770.3:p.Ile3405Val	-	5'ss	PTC by frameshift	-	0 (0)	1 (1)	-	<i>DNAH10</i>	Unknown	Candidate
ENST00000319980.10:c.1710-4T>G	-	-	3'ss	No PTC	TPR_8,TPR_1	0 (0)	2 (2)	-	<i>IFT88</i>	AR	Established ^b
ENST00000536576.5:c.656A>G	ENSP00000445067.2:p.Asp219Gly	-	5'ss	PTC by frameshift	-	0 (0)	3 (3)	c.192G>C ^a (p.Glu64Asp)	<i>TTC8</i>	AR	Established
ENST00000559838.5:c.438G>A	ENSP00000453449.1:p.Val146Val	-	3'ss	PTC in novel exon	-	0 (0)	1 (1)	-	<i>MOK</i>	Unknown	Candidate
ENST00000443035.7:c.1361T>A	ENSP00000391167.3:p.Leu454Gln	-	3'ss	No PTC	DENN	1 (1)	0 (0)	-	<i>DENND4A</i>	Unknown	Candidate
ENST00000379925.7:c.1877A>G	ENSP00000369257.3:p.Asp626Gly	-	5'ss	No PTC	C2-C2_1	0 (0)	1 (1)	-	<i>RPGRIP1L</i>	AR	Established
ENST00000338694.6:c.357-1G>A	-	0.0006964	3'ss	PTC in novel exon	-	0 (0)	2 (2)	-	<i>TEKT1</i>	AR	Established ^b
ENST00000570817.5:c.617C>T	ENSP00000461374.1:p.Ala206Val	0.00003942	5'ss	PTC by frameshift	-	0 (0)	3 (1)	-	<i>CEP131</i>	Unknown	Candidate
ENST00000269392.8:c.516-10G>A	-	-	3'ss	PTC by frameshift	-	0 (0)	2 (2)	-	<i>CEP131</i>	Unknown	Candidate
ENST00000315396.7:c.1391+7C>T	-	0.00001974	5'ss	PTC in novel exon	-	0 (0)	2 (1)	-	<i>CCDC114</i>	AR	Established
ENST00000278886.10:c.2125A>G	ENSP00000278886.6:p.Met709Val	0.00001971	5'ss	PTC by frameshift	-	0 (0)	1 (1)	-	<i>NINL</i>	AR	Candidate
ENST00000359568.9:c.1872C>G	ENSP00000352572.5:p.Cys624Trp	-	5'ss	PTC by frameshift	-	0 (0)	1 (1)	-	<i>PCNT</i>	AR	Candidate
ENST00000357137.8:c.590-16G>A	-	-	3'ss	PTC in novel exon	-	0 (0)	1 (1)	-	<i>XPNPEP3</i>	AR	Established

a: All affected individuals have this mutation. This mutation and the SCM seem to be in trans because, for one patient, there is a sibling who has this mutation but not the SCM. The other two patients are siblings in another family.

b: In a previous study [120], these genes were labeled as "candidate," but I changed them to "established" because causative mutations have been identified in these genes in recent reports [123–125,127,129–131,133].

AF: allele frequency; ss: splice site; CH: compound heterozygous.

One of the potential causative SCMs was found in the intronic region of the centrosome and spindle pole-associated protein 1 (*CSPP1*) (Fig. 3-7). The SCM is a G-to-A transition, which forms the first base of the canonical dinucleotide at the 3'ss. The novel 3'ss is formed 11-bp upstream of the annotated original 3'ss (Fig. 3-7). Due to the SCM, exon 27 (E27) is extended by 11-bp, generating an in-frame stop codon (Fig. 3-7). This mutation seems to be a rare variant as it is not registered in gnomAD. However, three ciliopathy patients were homozygous for the variant, one unaffected family member was heterozygous, and none of these patients had any known causative mutations, strongly suggesting that the SCM is a causative mutation. Although the SCM was found close to the 3'-end of the gene, exon 27, located downstream of the SCM, has a known pathogenic nonsense mutation (c.3212dup [p.Tyr1071Ter]) in Joubert syndrome, one of the disorders classified as ciliopathies, providing further evidence that the SCM may be involved in pathogenesis.

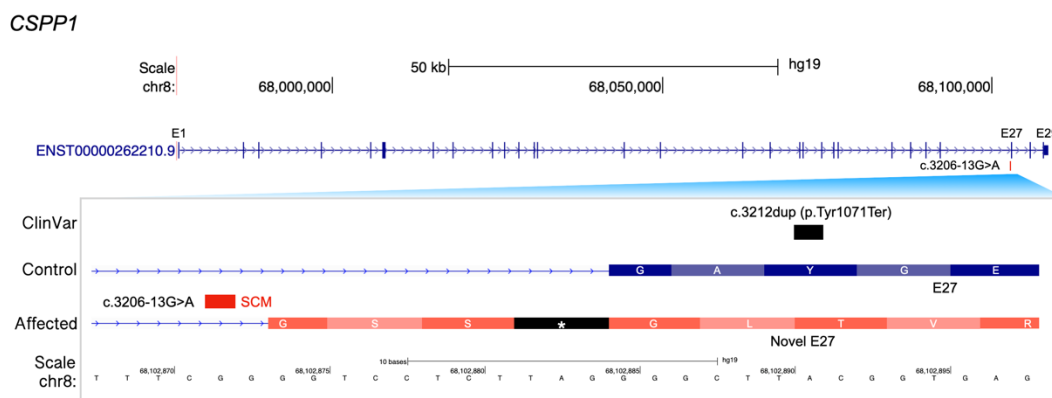


Fig. 3-7 Schematics of exon extension by SCM found in *CSPP1*.

The upper panel shows the gene structure of *CSPP1* obtained from the gene annotation data of GENCODE [117] V24. The lower panel is a close-up view of the SCM and exon 27 (E27). The ClinVar track shows a known pathogenic variant, c.3212dup, for ciliopathy. The gene structure for a control individual is followed by that for an affected individual with the SCM (c.3206-13G>A) which induces an exon extension. The extended exon contains an in-frame stop codon, shown as an asterisk with a black background.

Another example of an SCM was found in the exonic region of the DENN domain containing 4A (*DENND4A*) gene (Fig. 3-8). The potential causative SCM is a T-to-A transversion, which forms the first base of the canonical dinucleotide at the 3'ss. This mutation is a nonsynonymous substitution that replaces leucine with glutamine and is reported to be “probably damaging” by PolyPhen2 [147]. The novel 3'ss is formed 51-bp downstream of the annotated original 3'ss, which results in a 17 amino acid deletion without changing the downstream reading frame (Fig. 3-8a). Because the deleted segment is located within the DENN domain (Fig. 3-8b) and spans the three-dimensional protein structure (Fig. 3-8c and Supplementary Figure 3-2), the deletion may affect the domain structure, hence, its function. The variant is not registered in gnomAD. One ciliopathy patient who does not have any known causative mutations for ciliopathy was homozygous for this variant. No pathogenic mutations have been reported for *DENND4A* in ClinVar. Although the involvement of this gene in ciliogenesis has been predicted by high-throughput genome-wide RNAi screens [134], direct evidence of its involvement in ciliopathy is yet to be reported. My finding that the ciliopathy patient has a homozygous potential SCM in this gene provides further evidence supporting its possible involvement in the pathogenesis of ciliopathy.

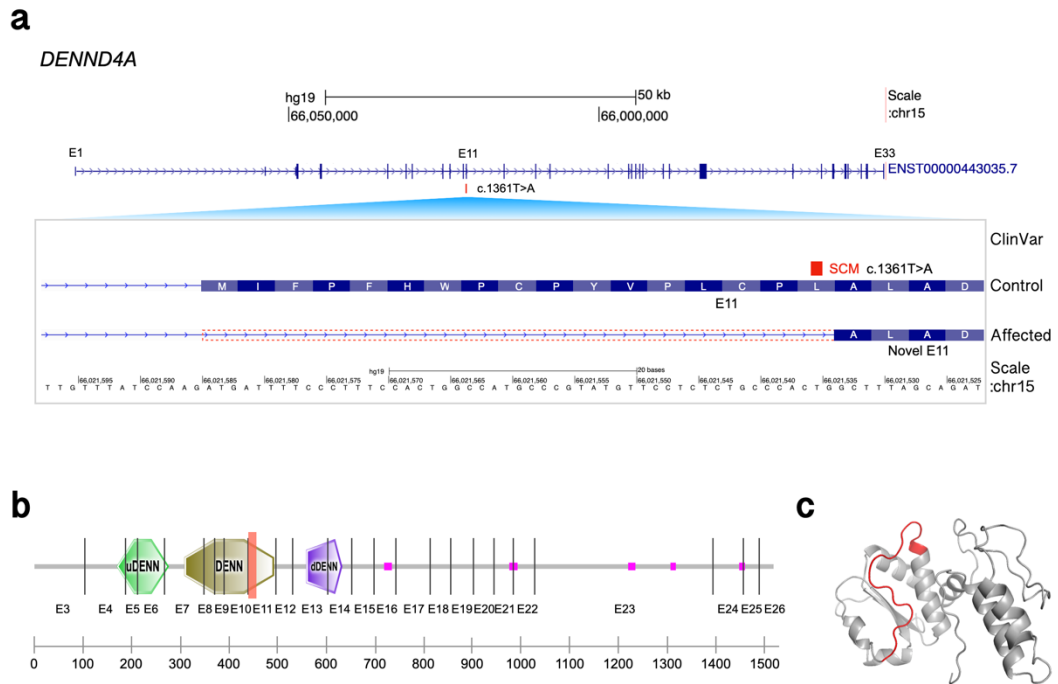


Fig. 3-8 Schematics of exon shrinkage by SCM found in *DENND4A*.

(a) The upper panel shows the gene structure of *DENND4A* obtained from gene annotation data of the GENCODE [117] V24. The genomic coordinate is reversed so that the direction of transcription is from left to right. The lower panel is a close-up view of the SCM and exon 11 (E11). The gene structure for a control individual is followed by that for an affected individual with the SCM (c.1361T>A), which induces exon shrinkage. The dotted box indicates the region of the shrinkage. (b) Effect of the SCM on the protein domain structure. The protein domain architecture is taken from SMART [69]. The red rectangle indicates the 17-amino-acid-long fragment that is coded by the shrunken part of the exon. The vertical line indicates the position of the intron. The pink box indicates the low-complexity region. (c) A three-dimensional protein structural model of the DENN domain constructed by SWISS-MODEL [136]. The 17-amino-acid-long fragment supposed to be deleted by the SCM is shown in red.

3.4 Discussion

In this Chapter, to clarify the extent to which SCMs serve as a cause of genetic disorders, a systematic search for SCMs in genes known to be responsible for genetic disorders was conducted. By applying SpliceAI to variant data from gnomAD and evaluating their ability to form novel splice sites, I comprehensively searched for potentially disease-causing SCMs in various genetic diseases without using patient samples. I identified 3,942 potentially pathogenic SCMs in the 4,054 genes known to be responsible for genetic disorders. Because I adopted relatively simple criteria to assess their functional consequences, some of them may not have functional effects, hence, the number of potentially pathogenic SCMs may have been overestimated. For example, some cases in which the small insertion of peptides by SCMs are located on the surface loop region without any functions of the proteins might exist. Conversely, even among the remaining 1,248 SCMs, there can be some damaging mutations caused by other mechanisms than by introducing PTCs or disrupting protein domains, such as those altering disordered regions with a certain function like protein–protein interaction [148,149]. Using the publicly available RNA-seq data obtained from individuals with the potential SCMs, I confirmed that the SCMs function as novel splice sites, as shown in the examples of the *PCNT* (Fig. 3-3) and *GAA* loci (Fig. 3-4). By reanalyzing WES data of ciliopathies, with a focus on SCMs, I identified 38 novel candidate causative mutations.

SCMs have been largely overlooked because they usually occur in non-conserved regions, where it is difficult to evaluate their functional importance. With the advent of the highly accurate splice site prediction tool SpliceAI, it has become possible to evaluate the mutations occurring in such regions in terms of their potential as splice sites. Additionally, by combining large-scale RNA-seq data with corresponding individual genotype data [118], I confirmed the functional impact of some potential SCMs without conducting *in vitro* molecular biological experiments. However, for most SCMs, their function has not been proven yet because of the lack of RNA-seq data with those mutations. Nevertheless, I believe that my data could be a valuable resource in identifying causative mutations of genetic disorders even in

such a situation. For example, suppose a mutation corresponds to one of the potential SCMs, but no other obvious damaging mutations are identified in a disease sample. In that case, the potential SCM is a plausible candidate for a causative mutation. As the possible involvement of such mutations in abnormal splicing can be confirmed by analyzing the structure of transcripts, conducting RNA-seq along with genotyping can help improve diagnostic rate [74,150–152].

The number of SCMs associated with the genetic disorders would be a lower limit for two main reasons. The first reason is that I focused only on single-nucleotide substitutions that create either GT or AG dinucleotides in the canonical splice sites. There must be other mutations that strengthen existing cryptic splice sites [76,77,115] or create novel splicing regulatory elements (SREs), such as exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and intronic splicing silencers (ISSs) [53]. I also excluded mutations that occur in deep intronic regions more than 50 bp from an annotated exon–intron boundary from my analysis. I did not take these mutations into account because either the computational methods currently available to assess potential splice sites [90,94] were not designed to handle these sites or the accuracy of the predictions is known to be rather low. For example, if I apply SpliceAI to 111 deep intronic SCMs that I identified in Chapter 2 (Fig.2-12), only 10 had SpliceAI scores ≥ 0.80 , indicating that SpliceAI is not so suitable in detecting deep intronic SCMs. I adopted this rather stringent cutoff in this study to identify potential SCMs to exclude as many false positives as possible. However, this can be a bias in our list of potential SCMs. The second reason is that, in the search for SCMs in the genes involved in genetic disorders, I focused only on the variants registered in the gnomAD database. As I show in the reanalysis of the exome data for ciliopathies, in which I used all the variants identified for each sample, 52.6% (20/38) of the potential SCMs are not registered in gnomAD (Table 3-4). This suggests that the list of 5,656 potential SCMs has another bias relying on the gnomAD database. There could be approximately the same number of potential SCMs that are not registered in the current version of gnomAD.

The SCMs identified in this study can be applied in several ways. One such application is the diagnosis of genetic disorders. Diagnoses of genetic disorders may often be delayed because of their rarity in populations and of sometimes weak or no symptoms in early life. If, however, patients with genetic disorders, especially some metabolic disorders, are correctly diagnosed and properly treated in their early life, symptoms may be diminished or almost completely suppressed, leading to improved quality of life in the future [153]. For this purpose, neonatal screening has been developed and is being applied clinically. Although the actual involvement of the potentially pathogenic SCMs has yet to be experimentally proven, adding those SCMs that are confirmed to be pathogenic ones as part of neonatal screening will directly contribute to the improvement of the diagnostic rate of genetic disorders. Another application of the identified SCMs might be as therapeutic targets. Unlike mutations in coding regions and those that disrupt existing splice sites, SCMs can be direct therapeutic targets and treated, for example, by applying antisense oligonucleotides [154]. In the case of disorders caused by SCMs, the underlying pathogenic mechanism is the formation of novel splice sites by mutation. Thus, a well-designed antisense oligonucleotide that could bind and suppress the activity of the novel splice site created by the SCM could be used as a therapeutic agent.

In conclusion, I showed that SCMs might be a common type of mutations causing genetic disorders as the number of SCMs is approximately 5.9%–8.5% of all currently known pathogenic mutations. I demonstrated that, by focusing on SCMs, the reanalysis of existing WES data for which the causative mutations have not yet been identified, can lead to the successful identification of candidate novel causative mutations, suggesting that SCMs should be considered a cause of genetic disorders.

Chapter 4. Conclusion

4.1 Summary

The purpose of this thesis is to identify novel mutations that are noteworthy in identifying causative mutations of genetic disorders and to improve the current low identification rate of causative mutations. I focused on splice-site-creating mutations (SCMs), which have been overlooked because they usually occur in non-conserved regions, and succeeded in identifying 3,942 novel candidate causative mutations. This number is equivalent to 5.9%–8.5% of all currently known causative mutations. Using whole exome sequencing (WES) data from ciliopathy patients available in public databases, I also successfully identified 38 novel candidate causative mutations. The results of this thesis demonstrate that by focusing on SCM, it is possible to identify novel causative mutations even in patients in whom causative mutations had not been identified. Therefore, I also strongly hope that researchers, doctors, and genetic counselors worldwide will consider the possibility that SCMs may be involved in genetic disorders and search for causative mutations in the future.

4.2 Future tasks

In Chapter 2, whole genome sequencing (WGS) and RNA-seq data was used to clarify the general frequency of SCMs in normal individuals. However, Chapter 2 focused on pseudo-exons caused by SCMs in deep intronic regions and not on exon extension or exon shrinkage caused by SCMs around exonic regions. Therefore, I will analyze exon extension and exon shrinkage in normal individuals and aim to clarify the frequency and characteristics of their occurrence.

The analysis pipeline for identifying candidate causative SCMs constructed in Chapter 3 targets mutations around exonic regions that can be detected by WES. This means that candidate causative SCMs in deep intronic regions cannot be analyzed with the pipeline. This is because it is difficult to identify SCMs in deep intronic regions by SpliceAI [94] used in Chapter 3. However, as of 2021, SpliceAI is the most accurate and powerful tool for predicting genetic variants that affect splicing [94,155,156], and there is no tool that can accurately assess variants in deep intronic

regions. Therefore, as soon as a tool that can accurately predict mutations in deep intronic regions is developed, I would like to identify candidate causative SCMs in those regions comprehensively.

Additionally, insertions/deletions (indels) were excluded from the analysis in this study. Indels are known to occur at a frequency of 20% of single-nucleotide variants (SNVs) [157] and are essential mutations that should be considered for possible involvement in genetic disorders. Indeed, indels have been reported as causative SCMs [158]. However, SpliceAI cannot be used to analyze indels. Therefore, I would like to identify the SCMs of indels that are responsible for genetic disorders using WES and RNA-seq data obtained from patients. Recently, these data have become available from public databases. For example, there are data on patients with rare diseases that have not been diagnosed yet in Database of Genotypes and Phenotypes (dbGaP) [119] (for example, dbGaP Study Accession: [phs001232.v4.p2](#)), and I am planning to use these data to identify cases where SCMs of indels are involved in genetic disorders. I aim to further improve the identification rate of causative mutations by conducting these analyses.

Abbreviations used in this thesis

3D	three-dimensional
3' ss	3' splice site
5' ss	5' splice site
ANNOVAR	Annotate Variation
BAM	Binary Alignment Map
BP	branch point
<i>CARS2</i>	cysteinyl-tRNA synthetase 2
<i>CASC3</i>	cancer susceptibility candidate 3
<i>CNIH1</i>	cornichon family AMPA receptor auxiliary protein 1
<i>CSPP1</i>	centrosome and spindle pole-associated protein 1
dbGaP	Database of Genotypes and Phenotypes
DECID	decay-inducing complex
DENN	differentially expressed in normal and neoplastic cells
<i>DENND4A</i>	DENN domain containing 4A
DNA	deoxyribonucleic acid
eIF4A3	eukaryotic translation initiation factor 4A3
EJC	exon junction complex
eRF1	eukaryotic release factor 1
eRF3	eukaryotic release factor 3
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
FPKM	fragments per kilobase of exon per million reads mapped
GAA	acid alpha-glucosidase
<i>GH1</i>	growth hormone 1
gnomAD	Genome Aggregation Database
GTEx	Genotype-Tissue Expression
GTF	General Transfer Format
HISAT	hierarchical indexing for spliced alignment of transcripts
hnRNP	heterogeneous nuclear ribonucleoprotein

IGHD II	isolated growth hormone deficiency type II
IGV	Integrative Genomics Viewer
Indel	insertion/deletion
ISE	intronic splicing enhancer
ISS	intronic splicing silencer
JAF	junction allele fraction
MAGOH	mago-nashi homolog
MOPD2	microcephalic osteodysplastic primordial dwarfism type II
NIH	National Institutes of Health
NMD	nonsense-mediated mRNA decay
OMIM	Online Mendelian Inheritance in Man
<i>p75TNFR</i>	human tumor necrosis factor receptor
<i>PCNT</i>	pericentrin
RNA Pol II	RNA polymerase II
PPT	polypyrimidine tract
PPV	positive predictive value
pre-mRNA	precursor mRNA
PSI	percent spliced in
PTC	premature termination codon
RBM8A	RNA-binding motif protein 8A
RNA	ribonucleic acid
RNA Pol II	RNA polymerase II
SCM	splice-site-creating mutation
SF1	splicing factor 1
SMART	Simple Modular Architecture Research Tool
snRNP	small nuclear ribonucleoprotein
SNV	single-nucleotide variant
SR protein	serine/arginine-rich protein
SRE	splicing regulatory element
SURF	SMG1-UPF1-eRF1/3
U2AF	U2 snRNP auxiliary factor

UCSC	University of California, Santa Cruz
UPF1	up-frameshift suppressor 1 homolog
UPF2	up-frameshift suppressor 2 homolog
UPF3B	up-frameshift suppressor 3 homolog B
VCF	Variant Call Format
WES	whole exome sequencing
WGS	whole genome sequencing

References

1. Crick F. Central dogma of molecular biology. *Nature*. 1970;227: 561–563.
2. Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A*. 1977;74: 3171–3175.
3. Mount SM. A catalogue of splice junction sequences. *Nucleic Acids Res*. 1982;10: 459–472.
4. Coolidge CJ, Seely RJ, Patton JG. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res*. 1997;25: 888–896.
5. Hornig H, Aepli M, Weissmann C. Effect of mutations at the lariat branch acceptor site on beta-globin pre-mRNA splicing in vitro. *Nature*. 1986;324: 589–591.
6. Sharp PA, Burge CB. Classification of introns: U2-type or U12-type. *Cell*. 1997;91: 875–879.
7. Turunen JJ, Niemelä EH, Verma B, Frilander MJ. The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA*. 2013;4: 61–76.
8. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14: 1188–1190.
9. Konarska MM, Grabowski PJ, Padgett RA, Sharp PA. Characterization of the branch site in lariat RNAs produced by splicing of mRNA precursors. *Nature*. 1985;313: 552–557.
10. Wahl MC, Will CL, Lührmann R. The spliceosome: design principles of a dynamic RNP machine. *Cell*. 2009;136: 701–718.
11. Green MR. Pre-mRNA splicing. *Annu Rev Genet*. 1986;20: 671–708.
12. Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*. 2003;72: 291–336.
13. Ruskin B, Green MR. An RNA processing activity that debranches RNA lariats. *Science*. 1985;229: 135–140.

14. Hesselberth JR. Lives that introns lead after splicing. *Wiley Interdiscip Rev RNA*. 2013;4: 677–691.
15. Wang ET, Sandberg R, Luo S, Khrebukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456: 470–476.
16. Pertea M. The human transcriptome: an unfinished story. *Genes* . 2012;3: 344–360.
17. Matlin AJ, Clark F, Smith CWJ. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*. 2005;6: 386–398.
18. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*. 2008;14: 802–813.
19. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature*. 2010;463: 457–463.
20. Blencowe BJ. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci*. 2000;25: 106–110.
21. Wang Y, Ma M, Xiao X, Wang Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat Struct Mol Biol*. 2012;19: 1044–1052.
22. Graveley BR. Sorting out the complexity of SR protein functions. *RNA*. 2000;6: 1197–1211.
23. Black DL. A simple answer for a splicing conundrum. *Proceedings of the National Academy of Sciences of the United States of America*. 2005. pp. 4927–4928.
24. Matera AG, Wang Z. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol*. 2014;15: 108–121.
25. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. Systematic identification and analysis of exonic splicing silencers. *Cell*. 2004;119: 831–845.
26. Wagner EJ, Garcia-Blanco MA. Polypyrimidine tract binding protein antagonizes exon definition. *Mol Cell Biol*. 2001;21: 3281–3288.
27. Hiller M, Zhang Z, Backofen R, Stamm S. Pre-mRNA secondary structures influence exon recognition. *PLoS Genet*. 2007;3: e204.

28. Warf MB, Berglund JA. Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci.* 2010;35: 169–178.
29. Kalmykova S, Kalinina M, Denisov S, Mironov A, Skvortsov D, Guigó R, et al. Conserved long-range base pairings are associated with pre-mRNA processing of human genes. *Nat Commun.* 2021;12: 2300.
30. Ruiz-Velasco M, Kumar M, Lai MC, Bhat P, Solis-Pinson AB, Reyes A, et al. CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals. *Cell Syst.* 2017;5: 628-637.e6.
31. Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell.* 2014;26: 577–590.
32. Saldi T, Riemony K, Erickson B, Bentley DL. Alternative RNA structures formed during transcription depend on elongation rate and modify RNA processing. *Mol Cell.* 2021. doi:10.1016/j.molcel.2021.01.040
33. Ast G. How did alternative splicing evolve? *Nat Rev Genet.* 2004;5: 773–782.
34. Sugnet CW, Kent WJ, Ares M, Haussler D. TRANSCRIPTOME AND GENOME CONSERVATION OF ALTERNATIVE SPLICING EVENTS IN HUMANS AND MICE. *Biocomputing 2004. WORLD SCIENTIFIC; 2003.* pp. 66–77.
35. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409: 860–921.
36. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature.* 2002;420: 520–562.
37. Sorek R. The birth of new exons: mechanisms and evolutionary consequences. *RNA.* 2007;13: 1603–1608.
38. Koralewski TE, Krutovsky KV. Evolution of exon-intron structure and alternative splicing. *PLoS One.* 2011;6: e18055.
39. Wang J, Lu Z-X, Tokheim CJ, Miller SE, Xing Y. Species-specific exon loss in human transcriptomes. *Mol Biol Evol.* 2015;32: 481–494.

40. Singer SS, Männel DN, Hehlhans T, Brosius J, Schmitz J. From “junk” to gene: curriculum vitae of a primate receptor isoform gene. *J Mol Biol.* 2004;341: 883–886.
41. Makałowski W, Mitchell GA, Labuda D. Alu sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet.* 1994;10: 188–193.
42. Sorek R, Ast G, Graur D. Alu-containing exons are alternatively spliced. *Genome Res.* 2002;12: 1060–1067.
43. Kondrashov FA, Koonin EV. Origin of alternative splicing by tandem exon duplication. *Hum Mol Genet.* 2001;10: 2661–2669.
44. Häsler J, Strub K. Alu elements as regulators of gene expression. *Nucleic Acids Res.* 2006;34: 5491–5497.
45. Zhang XH-F, Chasin LA. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc Natl Acad Sci U S A.* 2006;103: 13427–13432.
46. López-Bigas N, Audit B, Ouzounis C, Parra G, Guigó R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.* 2005;579: 1900–1903.
47. Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A.* 2011;108: 11093–11098.
48. Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nat Rev Genet.* 2016;17: 19–32.
49. Nakai K, Sakamoto H. Construction of a novel database containing aberrant splicing mutations of mammalian genes. *Gene.* 1994;141: 171–177.
50. De Conti L, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA.* 2013;4: 49–60.
51. Růžicková Š, Staněk D. Mutations in spliceosomal proteins and retina degeneration. *RNA Biol.* 2017;14: 544–552.
52. Grodecká L, Buratti E, Freiburger T. Mutations of Pre-mRNA Splicing Regulatory Elements: Are Predictions Moving Forward to Clinical Diagnostics? *Int J Mol Sci.* 2017;18. doi:10.3390/ijms18081668

53. Vaz-Drago R, Custódio N, Carmo-Fonseca M. Deep intronic mutations and human disease. *Hum Genet.* 2017;136: 1093–1111.
54. Naruto T, Okamoto N, Masuda K, Endo T, Hatsukawa Y, Kohmoto T, et al. Deep intronic GPR143 mutation in a Japanese family with ocular albinism. *Sci Rep.* 2015;5: 11334.
55. Lykke-Andersen S, Jensen TH. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol.* 2015;16: 665–677.
56. Magyar I, Colman D, Arnold E, Baumgartner D, Bottani A, Fokstuen S, et al. Quantitative sequence analysis of FBN1 premature termination codons provides evidence for incomplete NMD in leukocytes. *Hum Mutat.* 2009;30: 1355–1364.
57. Lindeboom RGH, Vermeulen M, Lehner B, Supek F. The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. *Nat Genet.* 2019;51: 1645–1651.
58. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science.* 2012;335: 823–828.
59. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature.* 2013;501: 506–511.
60. Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res.* 2016;44: 1483–1495.
61. Popp MW, Maquat LE. Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine. *Cell.* 2016;165: 1319–1322.
62. Zhang J, Sun X, Qian Y, LaDuca JP, Maquat LE. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol Cell Biol.* 1998;18: 5272–5283.
63. DeNoto FM, Moore DD, Goodman HM. Human growth hormone DNA sequence and mRNA structure: possible alternative splicing. *Nucleic Acids Res.* 1981;9: 3719–3730.

64. Masuda N, Watahiki M, Tanaka M, Yamakawa M, Shimizu K, Nagai J, et al. Molecular cloning of cDNA encoding 20 kDa variant human growth hormone and the alternative splicing mechanism. *Biochim Biophys Acta*. 1988;949: 125–131.
65. Asada N, Takahashi Y, Wada M, Naito N, Uchida H, Ikeda M, et al. GH induced lipolysis stimulation in 3T3-L1 adipocytes stably expressing hGHR: analysis on signaling pathway and activity of 20K hGH. *Mol Cell Endocrinol*. 2000;162: 121–129.
66. Binder G, Ranke MB. Screening for growth hormone (GH) gene splice-site mutations in sporadic cases with severe isolated GH deficiency using ectopic transcript analysis. *J Clin Endocrinol Metab*. 1995;80: 1247–1252.
67. Cogan JD, Ramel B, Lehto M, Phillips J 3rd, Prince M, Blizzard RM, et al. A recurring dominant negative mutation causes autosomal dominant growth hormone deficiency--a clinical research center study. *J Clin Endocrinol Metab*. 1995;80: 3591–3595.
68. Fofanova OV, Evgrafov OV, Polyakov AV, Poltaraus AB, Peterkova VA, Dedov II. A novel IVS2 -2A>T splicing mutation in the GH-1 gene in familial isolated growth hormone deficiency type II in the spectrum of other splicing mutations in the Russian population. *J Clin Endocrinol Metab*. 2003;88: 820–826.
69. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A*. 1998;95: 5857–5864.
70. Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet*. 2018;59: 253–268.
71. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet*. 2020;28: 165–173.
72. The Lancet Diabetes Endocrinology. Spotlight on rare diseases. *Lancet Diabetes Endocrinol*. 2019;7: 75.
73. Akter H, Hossain MS, Dity NJ, Rahaman MA, Furkan Uddin KM, Nassir N, et al. Whole exome sequencing uncovered highly penetrant recessive mutations for a spectrum of rare genetic pediatric diseases in Bangladesh. *NPJ Genom Med*. 2021;6: 14.

74. Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, et al. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am J Hum Genet.* 2019;104: 466–483.
75. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A.* 2015;112: 5473–5478.
76. Jayasinghe RG, Cao S, Gao Q, Wendl MC, Vo NS, Reynolds SM, et al. Systematic Analysis of Splice-Site-Creating Mutations in Cancer. *Cell Rep.* 2018;23: 270-281.e3.
77. Cao S, Zhou DC, Oh C, Jayasinghe RG, Zhao Y, Yoon CJ, et al. Discovery of driver non-coding splice-site-creating mutations in cancer. *Nat Commun.* 2020;11: 5573.
78. Hodges E, Xuan Z, Baliya V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. *Nat Genet.* 2007;39: 1522–1527.
79. Rabbani B, Tekin M, Mahdiah N. The promise of whole-exome sequencing in medical genetics. *J Hum Genet.* 2014;59: 5–15.
80. Blázquez L, Aiastui A, Goicoechea M, Martins de Araujo M, Avril A, Beley C, 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.
81. Flanagan SE, Xie W, Caswell R, Damhuis A, Vianey-Saban C, Akcay T, 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.
82. Chmel N, Danescu S, Gruler A, Kiritsi D, Bruckner-Tuderman L, Kreuter A, et al. A Deep-Intronic FERMT1 Mutation Causes Kindler Syndrome: An Explanation for Genetically Unsolved Cases. *J Invest Dermatol.* 2015;135: 2876–2879.
83. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012;491: 56–65.
84. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res.* 2002;12: 996–1006.

85. Andrews S, Others. FastQC: a quality control tool for high throughput sequence data. 2010.
86. 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.
87. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12: 357–360.
88. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015;33: 290–295.
89. 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.
90. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004;11: 377–394.
91. Zhao K, Lu Z-X, Park JW, Zhou Q, Xing Y. GLIMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biol*. 2013;14: R74.
92. Desmet F-O, Hamroun D, Lalande M, Collod-Bérout G, Claustres M, Bérout C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*. 2009;37: e67.
93. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*. 2019;10: 1523.
94. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*. 2019;176: 535-548.e24.
95. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29: 24–26.

96. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25: 2078–2079.
97. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26: 841–842.
98. Stein S, Lu Z-X, Bahrami-Samani E, Park JW, Xing Y. Discover hidden splicing variations by mapping personal transcriptomes to personal genomes. *Nucleic Acids Res*. 2015;43: 10612–10622.
99. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14: R36.
100. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29: 15–21.
101. Katz Y, Wang ET, Silterra J, Schwartz S, Wong B, Thorvaldsdóttir H, et al. Sashimi plots: Quantitative visualization of alternative isoform expression from RNA-seq data. *bioRxiv*. 2014. p. 002576. doi:10.1101/002576
102. Abdul-Sater Z, Cerabona D, Potchanant ES, Sun Z, Enzor R, He Y, et al. FANCA safeguards interphase and mitosis during hematopoiesis in vivo. *Exp Hematol*. 2015;43: 1031-1046.e12.
103. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581: 434–443.
104. 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.
105. Nakashima H, Nishikawa K, Ooi T. Differences in dinucleotide frequencies of human, yeast, and *Escherichia coli* genes. *DNA Res*. 1997;4: 185–192.
106. Krawczak M, Thomas NST, Hundrieser B, Mort M, Wittig M, Hampe J, 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.

107. Pros E, Gómez C, Martín T, Fábregas P, Serra E, Lázaro C. Nature and mRNA effect of 282 different NF1 point mutations: focus on splicing alterations. *Hum Mutat.* 2008;29: E173-93.
108. PCAWG Transcriptome Core Group, Calabrese C, Davidson NR, Demircioğlu D, Fonseca NA, He Y, et al. Genomic basis for RNA alterations in cancer. *Nature.* 2020;578: 129–136.
109. Faà V, Incani F, Meloni A, Corda D, Masala M, Baffico AM, 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.
110. Homolova K, Zavadakova P, Doktor TK, Schroeder LD, Kozich V, Andresen BS. 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.
111. Käsäkoski J, Jääskeläinen J, Jääskeläinen T, Tommiska J, Saarinen L, Lehtonen R, et al. Complete androgen insensitivity syndrome caused by a deep intronic pseudoexon-activating mutation in the androgen receptor gene. *Sci Rep.* 2016;6: 32819.
112. Fairbrother WG, Yeh R-F, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. *Science.* 2002;297: 1007–1013.
113. Sawyer SL, Hartley T, Dymant DA, Beaulieu CL, Schwartzentruber J, Smith A, 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.
114. Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: a practical guide to its clinical application. *Brief Funct Genomics.* 2016;15: 374–384.
115. Sakaguchi N, Suyama M. In silico identification of pseudo-exon activation events in personal genome and transcriptome data. *RNA Biol.* 2021;18: 382–390.
116. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 2015;43: D789-98.

117. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 2019;47: D766–D773.
118. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.* 2013;45: 580–585.
119. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet.* 2007;39: 1181–1186.
120. Reiter JF, Leroux MR. Genes and molecular pathways underpinning ciliopathies. *Nat Rev Mol Cell Biol.* 2017;18: 533–547.
121. Bayne CW, Donnelly F, Chapman K, Bollina P, Buck C, Habib F. A novel coculture model for benign prostatic hyperplasia expressing both isoforms of 5 alpha-reductase. *J Clin Endocrinol Metab.* 1998;83: 206–213.
122. Inglis PN, Boroevich KA, Leroux MR. Piecing together a ciliome. *Trends Genet.* 2006;22: 491–500.
123. Zariwala MA, Knowles MR, Omran H. Genetic defects in ciliary structure and function. *Annu Rev Physiol.* 2007;69: 423–450.
124. Zhang Z, Zariwala MA, Mahadevan MM, Caballero-Campo P, Shen X, Escudier E, et al. A heterozygous mutation disrupting the SPAG16 gene results in biochemical instability of central apparatus components of the human sperm axoneme. *Biol Reprod.* 2007;77: 864–871.
125. White J, Mazzeu JF, Hoischen A, Jhangiani SN, Gambin T, Alcino MC, et al. DVL1 frameshift mutations clustering in the penultimate exon cause autosomal-dominant Robinow syndrome. *Am J Hum Genet.* 2015;96: 612–622.
126. Shaheen R, Szymanska K, Basu B, Patel N, Ewida N, Faqeih E, et al. Characterizing the morbid genome of ciliopathies. *Genome Biol.* 2016;17: 242.
127. Edelbusch C, Cindrić S, Dougherty GW, Loges NT, Olbrich H, Rivlin J, et al. Mutation of serine/threonine protein kinase 36 (STK36) causes primary ciliary dyskinesia with a central pair defect. *Hum Mutat.* 2017;38: 964–969.

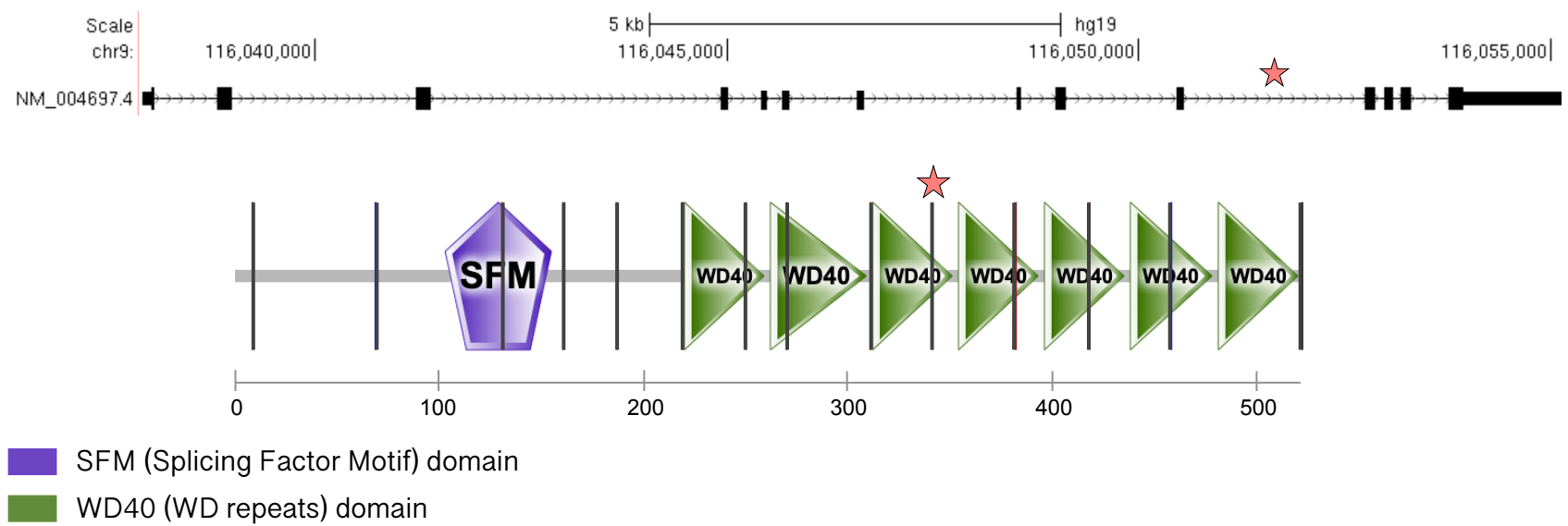
128. Quigley IK, Kintner C. Rfx2 Stabilizes Foxj1 Binding at Chromatin Loops to Enable Multiciliated Cell Gene Expression. *PLoS Genet.* 2017;13: e1006538.
129. Umair M, Shah K, Alhaddad B, Haack TB, Graf E, Strom TM, et al. Exome sequencing revealed a splice site variant in the IQCE gene underlying post-axial polydactyly type A restricted to lower limb. *Eur J Hum Genet.* 2017;25: 960–965.
130. Chekuri A, Guru AA, Biswas P, Branham K, Borooah S, Soto-Hermida A, et al. IFT88 mutations identified in individuals with non-syndromic recessive retinal degeneration result in abnormal ciliogenesis. *Hum Genet.* 2018;137: 447–458.
131. Ryan R, Failler M, Reilly ML, Garfa-Traore M, Delous M, Filhol E, et al. Functional characterization of tektin-1 in motile cilia and evidence for TEK1 as a new candidate gene for motile ciliopathies. *Hum Mol Genet.* 2018;27: 266–282.
132. Acevedo Luna NI. In-silico guided identification of ciliogenesis candidate genes in a non-conventional animal model. Iowa State University. 2019. Available: <https://lib.dr.iastate.edu/etd/17382>
133. Cindrić S, Dougherty GW, Olbrich H, Hjejij R, Loges NT, Amirav I, et al. SPEF2- and HYDIN-Mutant Cilia Lack the Central Pair-associated Protein SPEF2, Aiding Primary Ciliary Dyskinesia Diagnostics. *Am J Respir Cell Mol Biol.* 2020;62: 382–396.
134. Failler M, Giro-Perafita A, Owa M, Srivastava S, Yun C, Kahler DJ, et al. Whole-genome screen identifies diverse pathways that negatively regulate ciliogenesis. *Mol Biol Cell.* 2021;32: 169–185.
135. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 2011;39: W29-37.
136. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 2018;46: W296–W303.
137. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596: 583–589.
138. DELANO, W. L. The PyMOL Molecular Graphics System. <http://www.pymol.org>. 2002 [cited 4 Jun 2021]. Available: <https://ci.nii.ac.jp/naid/10020095229/>

139. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 2016;44: D862-8.
140. Zhang X-O, Wang H-B, Zhang Y, Lu X, Chen L-L, Yang L. Complementary sequence-mediated exon circularization. *Cell.* 2014;159: 134–147.
141. Demirdjian L, Xu Y, Bahrami-Samani E, Pan Y, Stein S, Xie Z, et al. Detecting Allele-Specific Alternative Splicing from Population-Scale RNA-Seq Data. *Am J Hum Genet.* 2020. doi:10.1016/j.ajhg.2020.07.005
142. Rauch A, Thiel CT, Schindler D, Wick U, Crow YJ, Ekici AB, et al. Mutations in the pericentrin (PCNT) gene cause primordial dwarfism. *Science.* 2008;319: 816–819.
143. Lim J-A, Kakhlon O, Li L, Myerowitz R, Raben N. Pompe disease: Shared and unshared features of lysosomal storage disorders. *Rare Dis.* 2015;3: e1068978.
144. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38: e164.
145. Hiller M, Huse K, Szafranski K, Rosenstiel P, Schreiber S, Backofen R, et al. Phylogenetically widespread alternative splicing at unusual GYNGYN donors. *Genome Biol.* 2006;7: R65.
146. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, et al. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet.* 2004;36: 1255–1257.
147. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7: 248–249.
148. Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell.* 2012;46: 884–892.
149. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science.* 2012;338: 1587–1593.

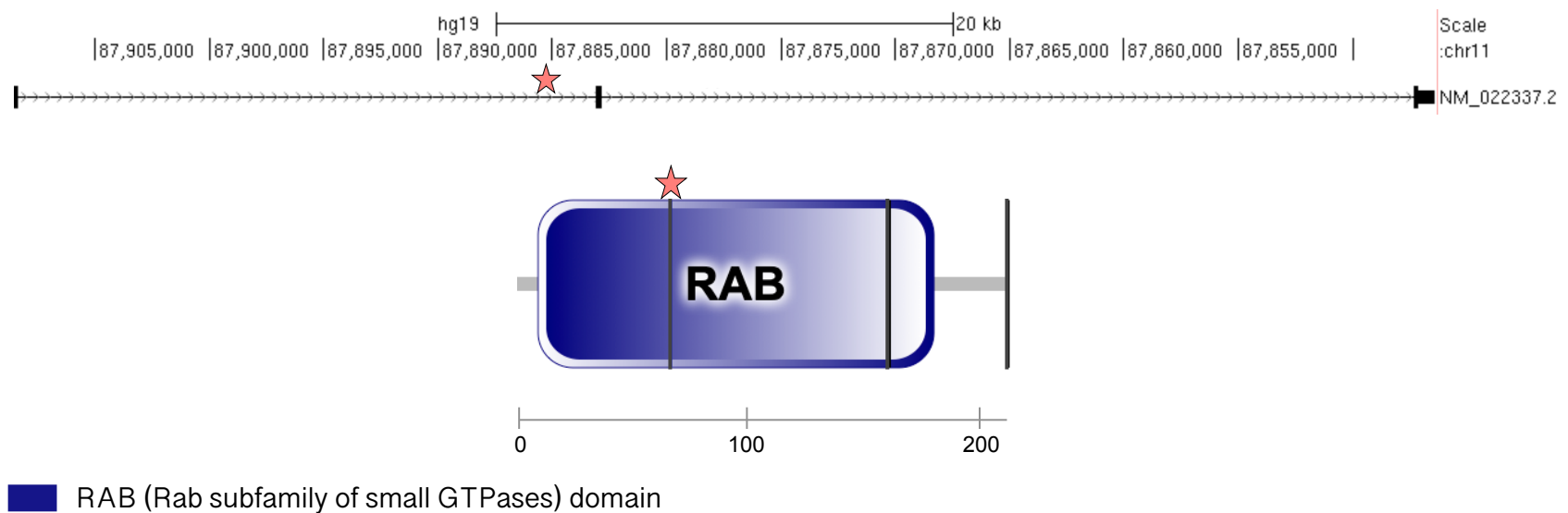
150. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med*. 2017;9. doi:10.1126/scitranslmed.aal5209
151. Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun*. 2017;8: 15824.
152. Lee H, Huang AY, Wang L-K, Yoon AJ, Renteria G, Eskin A, et al. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genet Med*. 2020;22: 490–499.
153. Morteza Pourfarzam FZ. Newborn Screening for inherited metabolic disorders; news and views. *J Res Med Sci*. 2013;18: 801.
154. Sazani P, Kole R. Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing. *J Clin Invest*. 2003;112: 481–486.
155. Chen J-M, Lin J-H, Masson E, Liao Z, Férec C, Cooper DN, et al. The Experimentally Obtained Functional Impact Assessments of 5' Splice Site GT'GC Variants Differ Markedly from Those Predicted. *Curr Genomics*. 2020;21: 56–66.
156. Ha C, Kim J-W, Jang J-H. Performance Evaluation of SpliceAI for the Prediction of Splicing of NF1 Variants. *Genes* . 2021;12. doi:10.3390/genes12091308
157. Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, et al. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res*. 2011;21: 830–839.
158. Chen LL, Sabripour M, Wu EF, Prieto VG, Fuller GN, Frazier ML. A mutation-created novel intra-exonic pre-mRNA splice site causes constitutive activation of KIT in human gastrointestinal stromal tumors. *Oncogene*. 2005;24: 4271–4280.

Supplementary materials

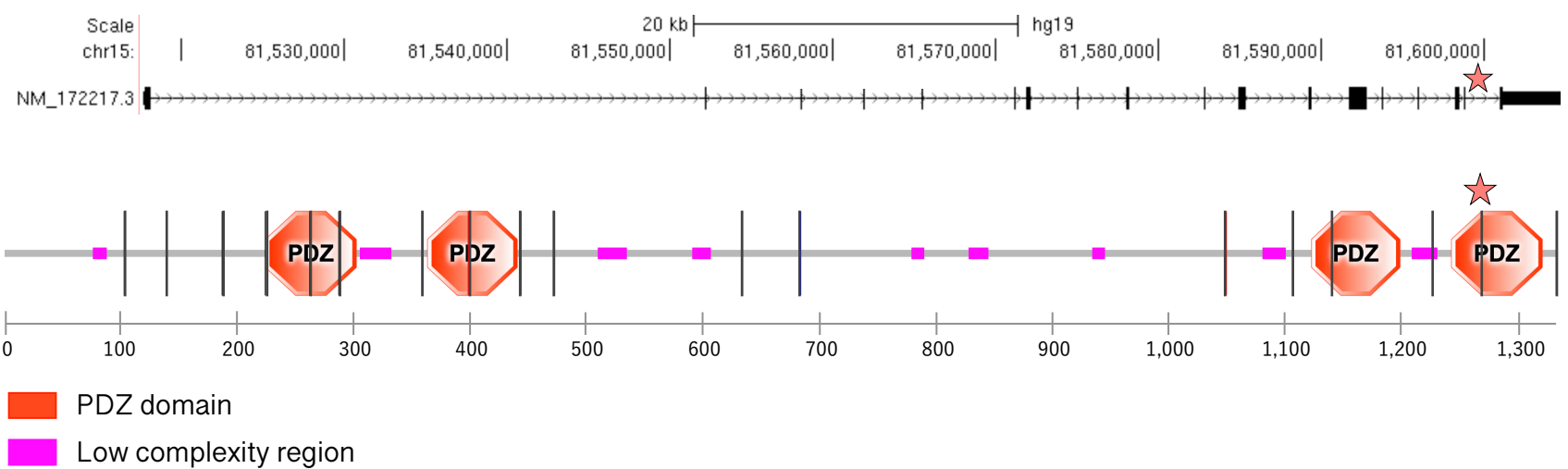
a



b

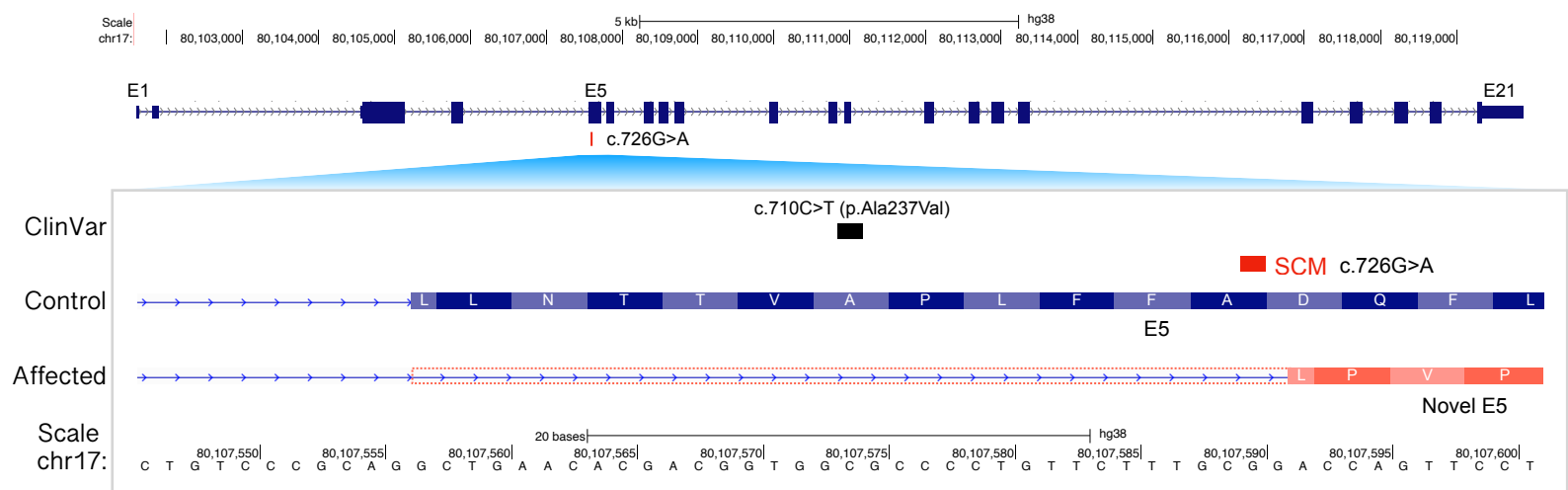


c



Supplementary Figure 2-1. Pseudo-exon activation events that disrupt known protein domains. These pseudo-exons do not disrupt the reading frame of the original genes neither by frameshifts nor by introducing premature termination codons. For each panel, the gene structure on the genome is shown at the top, followed by the protein domain architecture. In the gene structure, the position of the pseudo-exon is indicated by a star in orange. Vertical lines in the protein domain architecture indicate the positions of introns. The star in orange in the protein domain architecture indicates the position where the pseudo-exon is inserted. (A) *PRPF4*, (B) *RAB38*, and (C) *IL16*.

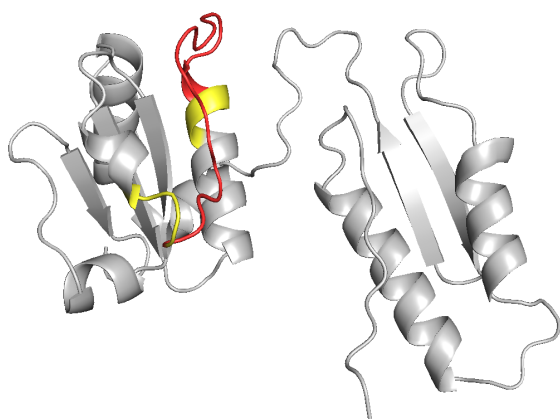
GAA



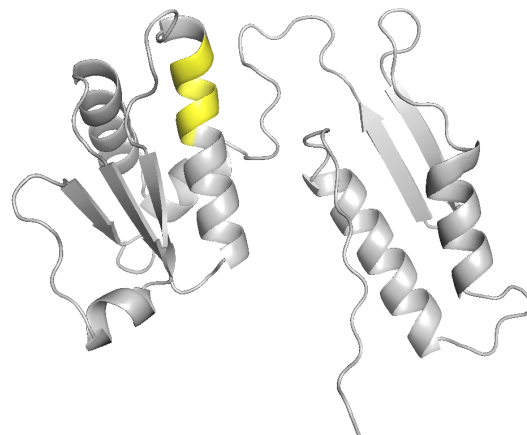
Supplementary Figure 3-1. Schematics of exon shrinkage by SCM found in GAA.

The upper panel shows the gene structure of GAA obtained from the gene annotation data of GENCODE V29. The lower panel is a close-up view of the SCM and exon 5 (E5). The ClinVar track shows a known pathogenic variant, c.710C>T, for Pompe disease. The gene structure for a control individual is followed by that for an affected individual with the SCM (c.726G>A) which induces an exon shrinkage. The dotted box indicates the region of the shrinkage. The shrunken region changes the downstream reading frame.

a



b



Supplementary Figure 3-2. A three-dimensional protein structural model of the DENN domain.

(a) The three-dimensional structure of the DENN domain was predicted using AlfaFold2. The 17-amino-acid-long fragment supposed to be deleted by the SCM is shown in red. The 3-amino-acids around deleted 17-amino-acid-long fragment is shown in yellow. (b) The three-dimensional structure of the DENN domain upon deletion of 17-amino-acid-long fragment was predicted using AlfaFold2. The 3-amino-acids around deleted 17-amino-acid-long fragment is shown in yellow.

Supplementary Table 2-1. Number of SNVs per sample used in this study.

Individual IDs	Number of SNVs	Number of SNVs
	(In exonic)	(In intronic)
HG00096	84,807	1,654,349
HG00097	84,437	1,661,294
HG00099	84,678	1,650,875
HG00100	85,903	1,654,194
HG00102	83,860	1,640,779
HG00105	84,447	1,651,108
HG00106	85,123	1,666,431
HG00109	85,123	1,649,263
HG00110	84,878	1,645,359
HG00111	84,722	1,639,194
HG00112	84,392	1,645,480
HG00114	83,752	1,641,813
HG00116	84,165	1,646,082
HG00119	85,382	1,657,672
HG00120	84,513	1,652,054
HG00129	84,044	1,636,635
HG00132	83,911	1,645,430
HG00133	84,994	1,664,582
HG00137	84,446	1,646,310
HG00138	84,751	1,652,593
HG00146	84,553	1,643,219
HG00148	84,421	1,656,367
HG00151	84,240	1,638,560
HG00177	83,998	1,634,551
HG00178	84,318	1,648,009
HG00181	85,701	1,666,234
HG00183	85,470	1,650,441
HG00187	84,012	1,650,898
HG00189	83,977	1,647,036
HG00231	84,454	1,642,625
HG00233	84,494	1,649,325
HG00234	84,988	1,652,023
HG00236	85,257	1,659,568

HG00238	85,286	1,653,450
HG00239	84,899	1,654,820
HG00245	85,531	1,657,607
HG00251	84,973	1,651,470
HG00255	85,220	1,655,835
HG00261	83,871	1,649,893
HG00262	84,430	1,649,749
HG00265	84,210	1,644,372
HG00267	84,971	1,645,628
HG00269	83,551	1,642,857
HG00271	85,226	1,655,200
HG00272	83,814	1,642,051
HG00275	84,035	1,646,184
HG00276	85,176	1,652,691
HG00277	83,659	1,633,271
HG00278	85,257	1,651,722
HG00282	84,562	1,653,239
HG00309	84,955	1,653,546
HG00311	84,102	1,635,234
HG00313	84,738	1,644,691
HG00315	85,015	1,644,300
HG00321	83,841	1,640,221
HG00325	84,780	1,644,358
HG00326	84,871	1,649,595
HG00327	84,520	1,637,534
HG00328	84,583	1,644,102
HG00330	85,173	1,649,774
HG00332	85,015	1,654,933
HG00334	84,962	1,658,633
HG00335	84,932	1,658,977
HG00336	84,624	1,652,399
HG00337	84,740	1,654,394
HG00338	85,132	1,648,289
HG00341	85,810	1,663,706
HG00345	84,864	1,656,748
HG00346	84,300	1,639,322

HG00349	84,051	1,644,513
HG00355	84,750	1,645,026
HG00356	84,921	1,651,451
HG00358	84,746	1,646,726
HG00364	84,094	1,664,035
HG00367	85,284	1,654,794
HG00369	84,686	1,648,685
HG00373	84,904	1,640,369
HG00376	84,404	1,645,551
HG00378	84,381	1,648,745
HG00382	83,803	1,646,513
HG00384	85,098	1,649,038
HG01791	84,907	1,647,637
NA06985	83,899	1,658,904
NA06986	84,674	1,651,490
NA07037	85,385	1,665,455
NA07048	84,592	1,656,373
NA07051	83,805	1,635,812
NA07357	85,494	1,643,686
NA10847	85,085	1,647,934
NA10851	84,282	1,651,771
NA11829	85,081	1,650,702
NA11830	83,840	1,641,578
NA11832	84,611	1,663,987
NA11843	85,047	1,658,739
NA11894	84,889	1,660,717
NA11920	85,805	1,657,866
NA11931	84,386	1,651,019
NA11994	84,717	1,654,256
NA12005	83,641	1,641,624
NA12006	85,023	1,659,045
NA12043	84,123	1,640,826
NA12044	83,533	1,653,770
NA12154	84,506	1,657,581
NA12155	85,152	1,654,953
NA12156	85,125	1,655,001

NA12249	84,316	1,635,872
NA12283	84,727	1,657,547
NA12286	84,667	1,660,565
NA12341	84,790	1,662,946
NA12342	84,277	1,643,610
NA12347	84,747	1,653,636
NA12348	85,436	1,664,076
NA12399	85,579	1,651,855
NA12400	84,749	1,647,595
NA12413	83,406	1,645,350
NA12546	85,006	1,662,261
NA12718	84,963	1,641,939
NA12750	84,580	1,652,698
NA12751	84,616	1,647,875
NA12761	83,693	1,647,275
NA12763	84,858	1,656,396
NA12776	84,505	1,655,702
NA12777	84,852	1,650,948
NA12812	84,279	1,651,412
NA12813	84,893	1,648,142
NA12829	84,248	1,646,868
NA12842	85,377	1,655,304
NA12843	84,358	1,659,486
NA12872	84,710	1,661,904
NA12889	84,843	1,648,241
NA18858	104,734	2,033,794
NA18861	105,423	2,041,043
NA18868	105,025	2,036,500
NA18870	105,690	2,041,623
NA18907	104,469	2,036,812
NA18908	104,399	2,032,602
NA18909	104,508	2,020,617
NA18916	105,201	2,037,634
NA18923	104,933	2,029,081
NA18934	104,264	2,023,067
NA19093	104,700	2,016,628

NA19095	104,808	2,031,059
NA19096	104,941	2,036,754
NA19098	105,186	2,034,196
NA19107	104,258	2,029,643
NA19108	104,666	2,032,627
NA19114	105,398	2,033,407
NA19116	104,948	2,027,086
NA19117	103,589	2,023,405
NA19119	104,826	2,023,214
NA19121	103,322	2,014,147
NA19129	104,535	2,029,605
NA19130	104,963	2,035,639
NA19131	105,019	2,036,398
NA19137	103,871	2,015,373
NA19138	103,622	2,022,352
NA19141	103,872	2,024,076
NA19143	104,454	2,027,288
NA19144	105,564	2,038,903
NA19146	104,688	2,035,345
NA19147	105,466	2,035,562
NA19152	104,266	2,030,778
NA19153	104,371	2,024,412
NA19171	104,437	2,024,050
NA19175	103,599	2,027,835
NA19189	104,172	2,019,442
NA19197	104,896	2,033,364
NA19198	104,703	2,023,341
NA19200	104,018	2,024,070
NA19201	104,390	2,018,551
NA19206	103,964	2,019,439
NA19209	104,154	2,021,124
NA19210	104,924	2,032,632
NA19213	104,777	2,030,351
NA19214	104,412	2,037,254
NA19222	104,799	2,023,795
NA19225	104,774	2,032,245

NA19247	103,953	2,018,218
NA19248	104,160	2,030,504
NA19256	104,605	2,025,495
NA19257	104,582	2,030,927
NA20502	83,959	1,639,744
NA20503	85,427	1,665,891
NA20504	84,988	1,663,401
NA20506	85,594	1,666,787
NA20509	86,322	1,670,333
NA20512	85,212	1,661,825
NA20513	85,630	1,663,639
NA20514	84,632	1,646,703
NA20515	84,561	1,659,903
NA20516	85,403	1,651,224
NA20517	85,823	1,668,090
NA20521	85,975	1,660,738
NA20524	85,106	1,657,620
NA20527	85,068	1,670,587
NA20528	85,302	1,662,176
NA20529	85,451	1,652,799
NA20530	84,976	1,657,024
NA20531	85,507	1,660,933
NA20535	84,818	1,658,575
NA20538	85,635	1,664,724
NA20541	85,372	1,658,896
NA20543	85,777	1,673,619
NA20581	85,480	1,665,255
NA20585	83,271	1,635,781
NA20586	84,770	1,653,462
NA20588	85,322	1,659,632
NA20760	85,871	1,661,855
NA20761	84,811	1,662,765
NA20765	85,935	1,665,110
NA20766	85,909	1,665,766
NA20768	85,364	1,662,049
NA20771	85,563	1,657,664

NA20772	85,836	1,660,727
NA20778	84,196	1,660,549
NA20783	85,239	1,663,446
NA20786	86,281	1,669,706
NA20787	85,855	1,661,808
NA20790	85,965	1,669,821
NA20792	85,210	1,655,634
NA20796	84,325	1,645,297
NA20797	85,217	1,663,356
NA20798	85,307	1,654,331
NA20800	84,973	1,653,449
NA20802	86,381	1,662,610
NA20806	85,154	1,668,485
NA20807	84,744	1,660,134
NA20808	85,073	1,665,899
NA20809	85,690	1,659,032
NA20810	85,397	1,669,856
NA20811	85,494	1,669,742
NA20812	85,558	1,661,593
NA20813	85,250	1,665,392
NA20814	84,997	1,667,814
NA20828	84,998	1,665,969

Supplementary Table 2-2. List of pseudo-exons identified in this study.

Chromosome	Start	End	Direction	SNV position	Ref	Alt	Alternate allele frequency	ss	Gene	PTC	Protein domain disruption ^a	PSI value ^b	No. individuals	No. individuals with homozygous mutations
chr1	882,137	882,244	-	882,254	G	A	0.000732764	3'ss	NOC2L	No PTC	Inter domain region	0.03±0.01	5	0
chr1	6,251,934	6,251,981	-	6,251,932	G	A	-	5'ss	RPL22	No PTC	Inter domain region	<0.01	1	0
chr1	6,251,934	6,252,049	-	6,251,932	G	A	-	5'ss	RPL22	PTC by frameshift	-	<0.01	1	0
chr1	32,610,948	32,611,068	+	32,611,070	C	T	0.000160803	5'ss	KPNA6	PTC in pseudo-exon	-	0.04	1	0
chr1	33,313,272	33,313,323	+	33,313,271	A	G	0.000191083	3'ss	S100PBP	PTC in pseudo-exon	-	0.92	1	0
chr1	95,405,349	95,405,481	+	95,405,348	T	G	0.000637186	3'ss	LOC729970	Not in CDS	-	0.83	1	0
chr1	95,702,899	95,703,016	+	95,702,898	A	G	0.0521736	3'ss	RWDD3	PTC in pseudo-exon	-	0.63±0.10	19	1
chr1	110,946,107	110,946,227	-	110,946,104	C	T	-	5'ss	LAMTOR5	PTC in pseudo-exon	-	0.01	1	0
chr1	173,694,857	173,694,997	+	173,694,842	G	T	0.00321574	3'ss	KLHL20	PTC in pseudo-exon	-	0.22	1	0
chr1	181,549,278	181,549,460	+	181,549,466	G	A	3.18573E-05	5'ss	CACNA1E	PTC in pseudo-exon	-	0.15	1	0
chr1	200,553,866	200,554,015	-	200,554,015	A	C	0.00165584	3'ss	KIF14	No PTC	Inter domain region	0.18	1	0
chr2	27,011,589	27,011,686	+	27,011,688	C	T	-	5'ss	CENPA	PTC in pseudo-exon	-	0.69	1	0
chr2	47,225,808	47,226,796	+	47,226,801	A	G	-	5'ss	TTC7A	PTC in pseudo-exon	-	0.1	1	0
chr2	47,226,626	47,226,796	+	47,226,801	A	G	-	5'ss	TTC7A	PTC in pseudo-exon	-	0.1	1	0
chr2	74,190,779	74,190,843	-	74,190,779	G	C	0.00272418	5'ss	DGUK-AS1	Not in CDS	-	0.7	2	0
chr2	127,438,330	127,438,380	+	127,438,329	T	G	0.00433038	3'ss	GYPC	Not in CDS	-	0.06	1	0
chr2	176,020,248	176,020,299	-	176,020,301	C	T	3.18674E-05	3'ss	ATF2	Not in CDS	-	0.15	1	0
chr3	16,488,265	16,488,355	-	16,488,266	C	T	-	5'ss	RFTN1	PTC in pseudo-exon	-	0.08	1	0
chr3	123,649,509	123,649,726	-	123,649,732	C	G	-	3'ss	CCDC14	PTC in pseudo-exon	-	0.13	1	0
chr4	17,580,691	17,580,803	+	17,580,681	G	T	0.00340764	3'ss	LAP3	PTC in pseudo-exon	-	0.01	1	0
chr4	17,600,401	17,600,502	+	17,600,399	G	A	0.000300581	3'ss	LAP3	No PTC	Inter domain region	<0.01	1	0
chr4	100,009,173	100,009,286	-	100,009,170	G	T	0.00191083	5'ss	ADH5	No PTC	Inter domain region	0.03	1	0
chr4	113,483,924	113,484,034	-	113,483,926	A	C	0.000031839	5'ss	ZGRF1	PTC in pseudo-exon	-	0.19	1	0
chr4	186,276,303	186,276,363	+	186,276,364	A	G	-	5'ss	SNX25	PTC in pseudo-exon	-	0.88	1	0
chr5	72,850,950	72,851,088	-	72,851,090	C	T	0.00124228	3'ss	ANKRA2	PTC in pseudo-exon	-	0.19	1	0
chr5	77,026,223	77,026,280	-	77,026,221	C	A	0.0717929	5'ss	TBCA	PTC in pseudo-exon	-	0.01±0.00	30	0
chr5	81,572,895	81,572,990	-	81,572,890	G	C	0.00202271	5'ss	RPS23	No PTC	Inter domain region	0.06	1	0
chr5	134,042,392	134,042,543	+	134,042,544	C	G	9.55536E-05	5'ss	SEC24A	PTC in pseudo-exon	-	0.03	1	0
chr5	139,932,741	139,932,889	-	139,932,740	T	C	0.0163203	5'ss	SRA1	PTC in pseudo-exon	-	0.03±0.01	4	0
chr6	5,325,369	5,325,539	+	5,325,357	G	T	0.0023893	3'ss	FARS2	Not in CDS	-	0.37	1	0
chr6	30,623,468	30,623,633	-	30,623,466	T	A	-	5'ss	DHX16	PTC by frameshift	-	0.08	1	0
chr6	43,015,325	43,015,513	-	43,015,321	C	T	0.000127453	5'ss	CUL7	No PTC	Inter domain region	0.5	1	0
chr6	43,022,414	43,022,508	-	43,022,408	G	A	-	5'ss	MRPL2	PTC in pseudo-exon	-	0.06	1	0
chr6	53,194,275	53,194,735	-	53,194,274	T	C	0.00359964	5'ss	ELOVL5	Not in CDS	-	0.04	2	0
chr6	88,294,961	88,295,044	-	88,294,960	T	C	9.55171E-05	5'ss	RARS2	PTC in pseudo-exon	-	0.52	1	0
chr6	111,214,505	111,214,549	+	111,214,550	A	G	0.00464909	5'ss	AMD1	No PTC	Inter domain region	0.13	1	0
chr6	139,354,886	139,354,992	+	139,354,992	A	G	0.0609304	5'ss	ABRACL	Not in CDS	-	0.04±0.01	29	0
chr6	143,507,663	143,507,763	+	143,507,763	C	G	0.00101898	5'ss	AIG1	PTC by frameshift	-	0.57	1	0
chr7	5,266,632	5,266,732	+	5,266,629	G	C	0.00624761	3'ss	WIP1	PTC in pseudo-exon	-	0.03	1	0
chr7	29,964,230	29,964,314	-	29,964,316	A	T	0.000159571	3'ss	SCRN1	PTC by frameshift	-	0.06	2	0
chr7	43,695,632	43,695,752	-	43,695,628	C	T	0.0984259	5'ss	COA1	Not in CDS	-	0.83±0.08	41	6
chr7	76,634,115	76,634,264	+	76,634,267	G	A	0.0155617	5'ss	DTX2P1-UPK3BP1-PMS2P11	Not in CDS	-	0.34±0.12	5	0
chr7	94,265,107	94,265,389	-	94,265,102	T	C	0.00375724	5'ss	SGCE	PTC in pseudo-exon	-	0.08	1	0
chr7	99,955,063	99,955,158	+	99,955,060	G	C	6.37308E-05	3'ss	STAG3LSP-PVRIG2P-PILRB	Not in CDS	-	0.49	1	0
chr7	99,964,625	99,964,749	+	99,964,750	C	G	0.0017226	5'ss	PILRB	PTC in pseudo-exon	-	0.26	1	0
chr7	99,964,721	99,964,749	+	99,964,750	C	G	0.0017226	5'ss	PILRB	PTC in pseudo-exon	-	0.26	1	0
chr7	135,351,122	135,351,259	+	135,351,260	C	G	0.00535339	5'ss	C7orf73	No PTC	Inter domain region	0.06±0.01	5	0
chr8	99,227,961	99,228,061	-	99,227,960	G	C	0.000127364	5'ss	NIPAL2	PTC in pseudo-exon	-	0.12	1	0
chr9	86,255,902	86,255,974	+	86,255,979	A	G	0.0150637	5'ss	IDNK	PTC in pseudo-exon	-	0.28±0.03	3	0
chr9	92,035,276	92,035,336	-	92,035,275	T	C	0.00563515	5'ss	SEMA4D	Not in CDS	-	0.28±0.06	4	0
chr9	111,637,915	111,638,015	-	111,637,912	C	T	0.00168768	5'ss	IKBKAP	PTC in pseudo-exon	-	0.08	1	0
chr9	116,051,691	116,051,795	+	116,051,798	G	A	-	5'ss	PRPF4	No PTC	Within domain	0.02	1	0
chr9	135,519,082	135,519,210	-	135,519,076	G	A	0.00095645	5'ss	DDX31	PTC in pseudo-exon	-	0.21	1	0
chr9	136,239,462	136,239,529	-	136,239,462	T	C	0.000063678	5'ss	SRF4	PTC in pseudo-exon	-	0.04	1	0
chr10	7,837,521	7,837,579	+	7,837,584	A	G	0.000987198	5'ss	ATP5C1	PTC in pseudo-exon	-	0.03	1	0
chr10	7,839,772	7,839,839	+	7,839,840	A	G	0.000318471	5'ss	ATP5C1	PTC in pseudo-exon	-	<0.01	1	0
chr10	32,635,026	32,635,076	-	32,635,025	T	C	0.00101911	5'ss	EPC1	PTC in pseudo-exon	-	0.93	1	0
chr10	38,654,838	38,654,939	+	38,654,940	T	G	0.216145	5'ss	HSD17B7P2	Not in CDS	-	0.49±0.15	92	11
chr10	127,485,385	127,485,430	-	127,485,384	T	C	6.36902E-05	5'ss	UROS	PTC in pseudo-exon	-	0.15	1	0
chr10	127,502,483	127,502,635	-	127,502,635	T	C	0.00146515	3'ss	UROS	PTC in pseudo-exon	-	0.02	1	0
chr11	4,414,142	4,414,337	-	4,414,137	T	C	-	5'ss	TRIM21	Not in CDS	-	0.33	1	0
chr11	69,487,151	69,487,271	-	69,487,150	T	C	-	5'ss	ORAOV1	PTC in pseudo-exon	-	0.31	1	0
chr11	71,157,568	71,157,656	-	71,157,567	G	C	0.0271355	5'ss	DHCR7	Not in CDS	-	0.02±0.01	9	0
chr11	72,610,673	72,610,777	-	72,610,668	G	C	0.00180111	5'ss	FCHSD2	No PTC	Inter domain region	0.05	2	0
chr11	74,663,547	74,663,624	+	74,663,544	T	C	0.000732484	3'ss	SPCS2	No PTC	Inter domain region	0.03	2	0
chr11	87,884,977	87,885,042	-	87,884,976	G	C	9.55231E-05	5'ss	RAB38	No PTC	Within domain	0.14	1	0
chr12	4,646,529	4,646,680	-	4,646,528	T	C	-	5'ss	C12orf4	Not in CDS	-	0.64	1	0
chr12	16,503,692	16,503,788	+	16,503,789	T	G	0.00915821	5'ss	MGST1	Not in CDS	-	0.31±0.12	6	0
chr12	32,835,808	32,835,889	+	32,835,891	C	T	0.00168811	5'ss	DNM1L	PTC in pseudo-exon	-	0.05	2	0
chr12	113,348,549	113,348,652	+	113,348,654	C	T	0.0295243	5'ss	OAS1	PTC by frameshift	-	0.12±0.25	8	1
chr12	120,883,446	120,883,539	-	120,883,445	T	C	6.36983E-05	5'ss	TRIAP1	PTC in pseudo-exon	-	0.94	1	0
chr12	123,038,820	123,038,906	+	123,038,909	G	A	0.00324196	5'ss	KNTC1	PTC in pseudo-exon	-	0.19	1	0
chr12	133,312,785	133,312,935	-	133,312,783	G	A	0.00442647	5'ss	ANKLE2	PTC in pseudo-exon	-	0.14±0.02	4	0
chr13	42,349,639	42,350,609	-	42,349,637	G	A	0.000573431	5'ss	VWA8	PTC in pseudo-exon	-	0.12	1	0
chr13	50,007,912	50,007,982	-	50,007,910	T	A	-	5'ss	CAB39L	Not in CDS	-	0.45	1	0
chr13	111,320,468	111,320,545	-	111,320,558	G	A	0.0030249	3'ss	CARS2	No PTC	Inter domain region	0.14	2	0
chr14	23,251,147	23,251,268	-	23,251,146	G	C	0.00506402	5'ss	SLC7A7	PTC in pseudo-exon	-	0.11	1	0
chr14	54,907,321	54,907,427	-	54,907,319	G	A	-	5'ss	CNIH1	PTC in pseudo-exon	-	0.24	1	0
chr14	105,954,227	105,954,364	+	105,954,368	G	A	0.00885689	5'ss	CRIP1	PTC in pseudo-exon	-	0.43±0.41	5	0
chr14	106,327,524	106,327,578	-	106,327,581	C	G	-	3'ss	IGHE	Not in CDS	-	0.03	1	0
chr15	23,311,070	23,311,188	-	23,311,201	C	T	0.014111	3'ss	HERC2P2	Not in CDS	-	0.32±0.13	5	1
chr15	28,900,641	28,900,819	+	28,900,821	G	T	0.00169912	5'ss	HERC2P9	Not in CDS	-	N/A ^c	1	1

chr15	31,516,048	31,516,188	-	31,516,046	C	A	6.36983E-05	5'ss	LOC283710	PTC in pseudo-exon	-	0.67	1	0
chr15	77,542,038	77,542,165	-	77,542,037	T	C	0.00200675	5'ss	PEAK1	Not in CDS	-	0.43	1	0
chr15	81,600,452	81,600,478	+	81,600,451	T	G	0.0594889	3'ss	IL16	No PTC	Within domain	0.10±0.03	23	0
chr16	4,457,198	4,457,255	-	4,457,256	T	C	3.22269E-05	3'ss	CORO7	PTC in pseudo-exon	-	0.08	1	0
chr16	57,199,458	57,199,564	-	57,199,456	G	A	0.000382312	5'ss	FAM192A	PTC in pseudo-exon	-	0.03	1	0
chr16	58,662,843	58,663,002	-	58,662,841	G	A	0.0459236	5'ss	CNOT1	Not in CDS	-	0.09±0.04	10	3
chr16	70,318,544	70,318,636	-	70,318,544	G	C	0.00114694	5'ss	AARS	Not in CDS	-	0.35±0.04	5	0
chr16	88,739,004	88,739,579	+	88,738,999	C	T	0.000350162	3'ss	SNAI3-AS1	Not in CDS	-	0.38	1	0
chr16	89,829,046	89,829,201	-	89,829,201	G	C	0.414918	3'ss	FANCA	PTC in pseudo-exon	-	0.12±0.04	83	37
chr16	90,020,442	90,020,517	+	90,020,523	G	T	-	5'ss	DEF8	PTC by frameshift	-	0.39	1	0
chr17	4,211,715	4,211,816	-	4,211,714	G	C	-	5'ss	UBE2G1	PTC in pseudo-exon	-	0.08	1	0
chr17	28,253,297	28,253,389	-	28,253,295	G	A	-	5'ss	SSH2	No PTC	Inter domain region	0.34	1	0
chr17	40,440,948	40,441,015	+	40,441,014	G	A	0.0230534	5'ss	STAT5A	Not in CDS	-	0.55±0.05	6	0
chr17	43,712,667	43,712,764	+	43,712,766	C	T	0.00404613	5'ss	MGC57346	Not in CDS	-	0.62	2	0
chr18	2,608,230	2,608,347	+	2,608,350	G	A	0.00227652	5'ss	NDC80	PTC in pseudo-exon	-	0.04	1	0
chr18	23,641,965	23,642,060	-	23,641,963	G	A	0.00499936	5'ss	SS18	No PTC	Inter domain region	0.37±0.30	4	0
chr19	8,312,329	8,312,446	+	8,312,447	A	G	0.033161	5'ss	CERS4	Not in CDS	-	0.74±0.10	18	2
chr19	13,886,291	13,886,376	+	13,886,283	A	C	0.00322087	3'ss	C19orf53	PTC in pseudo-exon	-	0.01	2	0
chr19	13,886,291	13,886,427	+	13,886,283	A	C	0.00322087	3'ss	C19orf53	PTC in pseudo-exon	-	0.03±0.02	3	0
chr19	19,151,482	19,151,601	+	19,151,481	A	G	3.36225E-05	3'ss	ARMC6	PTC in pseudo-exon	-	0.54	1	0
chr19	19,151,482	19,151,640	+	19,151,481	A	G	3.36225E-05	3'ss	ARMC6	PTC in pseudo-exon	-	0.33	1	0
chr19	58,776,501	58,776,589	-	58,776,500	T	C	-	5'ss	ZNF544	Not in CDS	-	0.31	1	0
chr20	252,598	252,655	-	252,596	T	A	0.00197503	5'ss	C20orf96	PTC in pseudo-exon	-	0.3	2	0
chr20	13,604,563	13,604,608	-	13,604,561	T	A	-	5'ss	TASP1	PTC in pseudo-exon	-	0.43	1	0
chr20	35,236,988	35,237,109	+	35,237,112	G	A	0.007424	5'ss	C20orf24	PTC by frameshift	-	0.03	2	0
chr21	17,240,254	17,240,381	+	17,240,253	A	G	0.00194391	3'ss	USP25	Not in CDS	-	0.06	1	0
chr21	26,758,995	26,759,072	-	26,758,994	T	C	0.0993755	5'ss	LINC00158	Not in CDS	-	0.17±0.06	37	2
chr21	33,980,707	33,980,799	-	33,980,705	G	A	0.0394239	5'ss	C21orf59	No PTC	Inter domain region	0.04±0.01	10	0
chr21	34,789,253	34,789,394	+	34,789,395	A	G	-	5'ss	IFNGR2	Not in CDS	-	0.4	1	0
chr21	44,326,950	44,327,012	+	44,327,013	A	G	0.0610665	5'ss	NDUFV3	PTC in pseudo-exon	-	0.29±0.26	34	1
chr21	45,759,596	45,759,682	+	45,759,683	A	G	0.00226086	5'ss	AP001062.9	Not in CDS	-	0.83	2	0
chr22	39,419,690	39,419,852	+	39,419,677	C	T	0.00188498	3'ss	APOBEC3D	PTC in pseudo-exon	-	0.16	1	0
chr22	45,717,923	45,718,049	+	45,718,050	A	G	-	5'ss	FAM118A	Not in CDS	-	0.82	1	0
chr22	50,695,706	50,696,089	-	50,695,707	C	T	0.0025834	5'ss	MAPK12	PTC in pseudo-exon	-	0.14	1	0

a: For pseudo-exons that do not disrupt the reading frame of the original genes, I determined the position of the pseudo-exon insertion in the protein domain architecture.

b: PSI values were calculated only for the samples heterozygous for the causative SCM. A standard deviation value is provided for the pseudo-exons identified in more than two individuals.

c: PSI value was not calculated because the sample was homozygous for the causative SCM.

Supplementary Table 3-1. List of the 434 ciliopathy-related genes.

Gene	ENSEMBL ID	Ciliopathy gene	References
<i>ADCY3</i>	ENSG00000138031.14	Candidate	[120]
<i>ADCY6</i>	ENSG00000174233.11	Established	[120]
<i>AHI1</i>	ENSG00000135541.20	Established	[120]
<i>AK7</i>	ENSG00000140057.8	Established	[120]
<i>AKAP9</i>	ENSG00000127914.16	Candidate	[120]
<i>ALMS1</i>	ENSG00000116127.17	Established	[120]
<i>ANKRD26</i>	ENSG00000107890.16	Candidate	[120]
<i>ANKS3</i>	ENSG00000168096.14	Established	[120]
<i>ANKS6</i>	ENSG00000165138.17	Established	[120]
<i>ANO2</i>	ENSG00000047617.14	Candidate	[120]
<i>ARF4</i>	ENSG00000168374.10	Candidate	[120]
<i>ARL13B</i>	ENSG00000169379.15	Established	[120]
<i>ARL2</i>	ENSG00000213465.7	Candidate	[120]
<i>ARL2BP</i>	ENSG00000102931.7	Established	[120]
<i>ARL3</i>	ENSG00000138175.8	Established	[120]
<i>ARL6</i>	ENSG00000113966.9	Established	[120]
<i>ARMC4</i>	ENSG00000169126.15	Established	[120]
<i>ASAP1</i>	ENSG00000153317.14	Candidate	[120]
<i>ATXN10</i>	ENSG00000130638.15	Established	[120]
<i>AURKA</i>	ENSG00000087586.17	Candidate	[120]
<i>B9D1</i>	ENSG00000108641.14	Established	[120]
<i>B9D2</i>	ENSG00000123810.7	Established	[120]
<i>BBIP1</i>	ENSG00000214413.7	Established	[120]
<i>BBS1</i>	ENSG00000174483.19	Established	[120]
<i>BBS2</i>	ENSG00000125124.11	Established	[120]
<i>BBS4</i>	ENSG00000140463.13	Established	[120]
<i>BBS5</i>	ENSG00000163093.11	Established	[120]
<i>BBS7</i>	ENSG00000138686.9	Established	[120]
<i>BBS9</i>	ENSG00000122507.20	Established	[120]
<i>BBS10</i>	ENSG00000179941.6	Established	[120]
<i>BBS12</i>	ENSG00000181004.9	Established	[120]
<i>C2CD3</i>	ENSG00000168014.16	Established	[120]
<i>CACNA1I</i>	ENSG00000100346.17	Candidate	[122]
<i>CATIP</i>	ENSG00000158428.3	Candidate	[120]

<i>CBY1</i>	ENSG00000100211.10	Candidate	[120]
<i>CC2D2A</i>	ENSG00000048342.15	Established	[120]
<i>CCDC13</i>	ENSG00000244607.5	Candidate	[120]
<i>CCDC14</i>	ENSG00000175455.14	Candidate	[120]
<i>CCDC18</i>	ENSG00000122483.17	Candidate	[120]
<i>CCDC22</i>	ENSG00000101997.12	Candidate	[120]
<i>CCDC39</i>	ENSG00000145075.11	Established	[120]
<i>CCDC40</i>	ENSG00000141519.14	Established	[120]
<i>CCDC63</i>	ENSG00000173093.12	Candidate	[120]
<i>CCDC65</i>	ENSG00000139537.10	Established	[120]
<i>CCDC66</i>	ENSG00000180376.16	Candidate	[120]
<i>CCDC103</i>	ENSG00000167131.16	Established	[120]
<i>CCDC112</i>	ENSG00000164221.12	Candidate	[120]
<i>CCDC114</i>	ENSG00000105479.15	Established	[120]
<i>CCDC151</i>	ENSG00000198003.11	Established	[120]
<i>CCDC28B</i>	ENSG00000160050.14	Established	[120]
<i>CCDC88A</i>	ENSG00000115355.15	Candidate	[120]
<i>CCNO</i>	ENSG00000152669.8	Established	[120]
<i>CDK5RAP2</i>	ENSG00000136861.17	Candidate	[120]
<i>CDKL5</i>	ENSG00000008086.10	Candidate	[120]
<i>CELSR2</i>	ENSG00000143126.7	Established	[126]
<i>CENPF</i>	ENSG00000117724.12	Established	[120]
<i>CENPJ</i>	ENSG00000151849.14	Candidate	[120]
<i>CEP19</i>	ENSG00000174007.7	Established	[120]
<i>CEP41</i>	ENSG00000106477.18	Established	[120]
<i>CEP63</i>	ENSG00000182923.16	Candidate	[120]
<i>CEP72</i>	ENSG00000112877.7	Candidate	[120]
<i>CEP76</i>	ENSG00000101624.10	Candidate	[120]
<i>CEP78</i>	ENSG00000148019.12	Established	[120]
<i>CEP83</i>	ENSG00000173588.14	Established	[120]
<i>CEP89</i>	ENSG00000121289.17	Candidate	[120]
<i>CEP97</i>	ENSG00000182504.10	Candidate	[120]
<i>CEP104</i>	ENSG00000116198.12	Established	[120]
<i>CEP120</i>	ENSG00000168944.15	Established	[120]
<i>CEP128</i>	ENSG00000100629.16	Candidate	[120]
<i>CEP131</i>	ENSG00000141577.13	Candidate	[120]

<i>CEP135</i>	ENSG00000174799.10	Candidate	[120]
<i>CEP152</i>	ENSG00000103995.13	Candidate	[120]
<i>CEP162</i>	ENSG00000135315.11	Candidate	[120]
<i>CEP164</i>	ENSG00000110274.14	Established	[120]
<i>CEP170</i>	ENSG00000143702.15	Candidate	[120]
<i>CEP192</i>	ENSG00000101639.18	Candidate	[120]
<i>CEP250</i>	ENSG00000126001.15	Candidate	[120]
<i>CEP290</i>	ENSG00000198707.14	Established	[120]
<i>CEP350</i>	ENSG00000135837.15	Candidate	[120]
<i>CETN1</i>	ENSG00000177143.4	Candidate	[120]
<i>CETN2</i>	ENSG00000147400.8	Candidate	[120]
<i>CETN3</i>	ENSG00000153140.8	Candidate	[120]
<i>CFAP20</i>	ENSG00000070761.7	Candidate	[120]
<i>CFAP36</i>	ENSG00000163001.11	Candidate	[120]
<i>CFAP52</i>	ENSG00000166596.14	Established	[120]
<i>CFAP53</i>	ENSG00000172361.5	Established	[120]
<i>CFAP54</i>	ENSG00000188596.9	Candidate	[120]
<i>CFAP57</i>	ENSG00000243710.7	Candidate	[120]
<i>CFAP58</i>	ENSG00000120051.14	Candidate	[120]
<i>CFAP61</i>	ENSG00000089101.17	Candidate	[120]
<i>CFAP126</i>	ENSG00000188931.3	Candidate	[120]
<i>CFAP161</i>	ENSG00000156206.13	Candidate	[120]
<i>CLUAP1</i>	ENSG00000103351.12	Established	[120]
<i>CNGA1</i>	ENSG00000198515.13	Established	[120]
<i>CNGA2</i>	ENSG00000183862.5	Established	[120]
<i>CNGA4</i>	ENSG00000132259.12	Candidate	[120]
<i>CNGB1</i>	ENSG00000070729.13	Established	[120]
<i>CNTRL</i>	ENSG00000119397.16	Candidate	[120]
<i>CNTROB</i>	ENSG00000170037.13	Candidate	[120]
<i>CRB3</i>	ENSG00000130545.15	Candidate	[120]
<i>CROCC</i>	ENSG00000058453.16	Candidate	[120]
<i>CRX</i>	ENSG00000105392.15	Established	[120]
<i>CSNK1D</i>	ENSG00000141551.14	Candidate	[120]
<i>CSPP1</i>	ENSG00000104218.14	Established	[120]
<i>CTTN</i>	ENSG00000085733.15	Candidate	[120]
<i>CTU2</i>	ENSG00000174177.12	Candidate	[126]

<i>DCDC2</i>	ENSG00000146038.11	Established	[120]
<i>DCTN1</i>	ENSG00000204843.12	Candidate	[120]
<i>DDX59</i>	ENSG00000118197.13	Established	[120]
<i>DENND4A</i>	ENSG00000174485.15	Candidate	[134]
<i>DISC1</i>	ENSG00000162946.20	Candidate	[120]
<i>DNAAF1</i>	ENSG00000154099.17	Established	[120]
<i>DNAAF2</i>	ENSG00000165506.14	Established	[120]
<i>DNAAF3</i>	ENSG00000167646.13	Established	[120]
<i>DNAAF5</i>	ENSG00000164818.15	Established	[120]
<i>DNAH1</i>	ENSG00000114841.17	Established	[120]
<i>DNAH2</i>	ENSG00000183914.14	Candidate	[120]
<i>DNAH5</i>	ENSG00000039139.9	Established	[120]
<i>DNAH6</i>	ENSG00000115423.18	Established	[120]
<i>DNAH9</i>	ENSG00000007174.17	Candidate	[120]
<i>DNAH10</i>	ENSG00000197653.14	Candidate	[120]
<i>DNAH11</i>	ENSG00000105877.17	Established	[120]
<i>DNAI1</i>	ENSG00000122735.15	Established	[120]
<i>DNAI2</i>	ENSG00000171595.13	Established	[120]
<i>DNAJB13</i>	ENSG00000187726.8	Established	[120]
<i>DNAL1</i>	ENSG00000119661.14	Established	[120]
<i>DNAL4</i>	ENSG00000100246.12	Candidate	[120]
<i>DNALI1</i>	ENSG00000163879.10	Candidate	[120]
<i>DPCD</i>	ENSG00000166171.12	Candidate	[120]
<i>DRC1</i>	ENSG00000157856.10	Established	[120]
<i>DRC3</i>	ENSG00000171962.17	Candidate	[120]
<i>DVL1</i>	ENSG00000107404.17	Established*	[125]
<i>DYNC2H1</i>	ENSG00000187240.13	Established	[120]
<i>DYNC2LI1</i>	ENSG00000138036.18	Established	[120]
<i>DYNLT1</i>	ENSG00000146425.10	Candidate	[120]
<i>DYX1C1</i>	ENSG00000256061.7	Established	[120]
<i>DZIP1</i>	ENSG00000134874.17	Candidate	[120]
<i>EFCAB7</i>	ENSG00000203965.12	Candidate	[120]
<i>EFHC1</i>	ENSG00000096093.14	Established	[120]
<i>EFHC2</i>	ENSG00000183690.12	Candidate	[120]
<i>EVC</i>	ENSG00000072840.12	Established	[120]
<i>EVC2</i>	ENSG00000173040.12	Established	[120]

<i>EXOC3L2</i>	ENSG00000130201.7	Established	[126]
<i>EXOC4</i>	ENSG00000131558.14	Candidate	[120]
<i>FAM58A</i>	ENSG00000262919.7	Established	[120]
<i>FAM161A</i>	ENSG00000170264.12	Established	[120]
<i>FAM98C</i>	ENSG00000130244.12	Established	[126]
<i>FBF1</i>	ENSG00000188878.17	Candidate	[120]
<i>FGFR1OP</i>	ENSG00000213066.11	Candidate	[120]
<i>FIGNL1</i>	ENSG00000132436.11	Candidate	[120]
<i>FLCN</i>	ENSG00000154803.12	Established	[120]
<i>FOPNL</i>	ENSG00000133393.12	Candidate	[120]
<i>FOXJ1</i>	ENSG00000129654.7	Candidate	[120]
<i>FUZ</i>	ENSG00000010361.13	Candidate	[120]
<i>GALNT11</i>	ENSG00000178234.12	Established	[120]
<i>GAS8</i>	ENSG00000141013.14	Established	[120]
<i>GDI2</i>	ENSG00000057608.16	Candidate	[120]
<i>GLI1</i>	ENSG00000111087.9	Candidate	[120]
<i>GLI2</i>	ENSG00000074047.20	Established	[120]
<i>GLI3</i>	ENSG00000106571.12	Established	[120]
<i>GLIS2</i>	ENSG00000126603.8	Established	[120]
<i>GMPPB</i>	ENSG00000173540.12	Candidate	[122]
<i>GNAI2</i>	ENSG00000114353.16	Candidate	[120]
<i>GNAI3</i>	ENSG00000065135.8	Candidate	[120]
<i>GNAS</i>	ENSG00000087460.23	Candidate	[120]
<i>GPBAR1</i>	ENSG00000179921.14	Candidate	[120]
<i>GPR161</i>	ENSG00000143147.14	Established	[120]
<i>GSK3A</i>	ENSG00000105723.11	Candidate	[120]
<i>GSTP1</i>	ENSG00000084207.15	Candidate	[122]
<i>HNF1B</i>	ENSG00000275410.4	Candidate	[120]
<i>HSPB11</i>	ENSG00000081870.11	Candidate	[120]
<i>HTR6</i>	ENSG00000158748.3	Candidate	[120]
<i>HTT</i>	ENSG00000197386.10	Candidate	[120]
<i>HYDIN</i>	ENSG00000157423.17	Established	[120]
<i>HYLS1</i>	ENSG00000198331.10	Established	[120]
<i>ICK</i>	ENSG00000112144.15	Established	[120]
<i>IFT20</i>	ENSG00000109083.13	Candidate	[120]
<i>IFT27</i>	ENSG00000100360.14	Established	[120]

<i>IFT43</i>	ENSG00000119650.12	Established	[120]
<i>IFT46</i>	ENSG00000118096.7	Candidate	[120]
<i>IFT52</i>	ENSG00000101052.12	Established	[120]
<i>IFT57</i>	ENSG00000114446.4	Established	[120]
<i>IFT74</i>	ENSG00000096872.15	Candidate	[120]
<i>IFT80</i>	ENSG00000068885.14	Established	[120]
<i>IFT81</i>	ENSG00000122970.15	Established	[120]
<i>IFT88</i>	ENSG00000032742.17	Established*	[133]
<i>IFT122</i>	ENSG00000163913.11	Established	[120]
<i>IFT140</i>	ENSG00000187535.13	Established	[120]
<i>IFT172</i>	ENSG00000138002.14	Established	[120]
<i>INPP5B</i>	ENSG00000204084.12	Candidate	[120]
<i>INPP5E</i>	ENSG00000148384.11	Established	[120]
<i>INTU</i>	ENSG00000164066.12	Established	[120]
<i>INVS</i>	ENSG00000119509.12	Established	[120]
<i>IQCB1</i>	ENSG00000173226.16	Established	[120]
<i>IQCE</i>	ENSG00000106012.17	Established*	[126]
<i>IQCG</i>	ENSG00000114473.13	Candidate	[120]
<i>IQUB</i>	ENSG00000164675.10	Candidate	[120]
<i>KIAA0556</i>	ENSG00000047578.12	Established	[120]
<i>KIAA0586</i>	ENSG00000100578.14	Established	[120]
<i>KIAA0753</i>	ENSG00000198920.9	Established	[120]
<i>KIAA1217</i>	ENSG00000120549.16	Candidate	[120]
<i>KIF7</i>	ENSG00000166813.14	Established	[120]
<i>KIF17</i>	ENSG00000117245.12	Candidate	[120]
<i>KIF19</i>	ENSG00000196169.14	Candidate	[120]
<i>KIF24</i>	ENSG00000186638.15	Candidate	[120]
<i>KIF27</i>	ENSG00000165115.14	Candidate	[120]
<i>KIF2A</i>	ENSG00000068796.16	Candidate	[120]
<i>KIF3A</i>	ENSG00000131437.15	Candidate	[120]
<i>KIF3B</i>	ENSG00000101350.7	Candidate	[120]
<i>KIF3C</i>	ENSG00000084731.13	Candidate	[120]
<i>KIFAP3</i>	ENSG00000075945.12	Candidate	[120]
<i>KIZ</i>	ENSG00000088970.15	Established	[120]
<i>LCA5</i>	ENSG00000135338.13	Established	[120]
<i>LRRC6</i>	ENSG00000129295.8	Established	[120]

<i>LRRC49</i>	ENSG00000137821.11	Candidate	[120]
<i>LRRCC1</i>	ENSG00000133739.15	Established	[126]
<i>LUZP1</i>	ENSG00000169641.13	Candidate	[120]
<i>LZTFL1</i>	ENSG00000163818.16	Established	[120]
<i>MAK</i>	ENSG00000111837.11	Established	[120]
<i>MAL</i>	ENSG00000172005.10	Candidate	[120]
<i>MAPRE1</i>	ENSG00000101367.8	Candidate	[120]
<i>MARK4</i>	ENSG00000007047.14	Candidate	[120]
<i>MCHR1</i>	ENSG00000128285.4	Candidate	[120]
<i>MCIDAS</i>	ENSG00000234602.7	Established	[120]
<i>MCRS1</i>	ENSG00000187778.13	Candidate	[120]
<i>MDM1</i>	ENSG00000111554.14	Candidate	[120]
<i>MKKS</i>	ENSG00000125863.17	Established	[120]
<i>MKS1</i>	ENSG00000011143.16	Established	[120]
<i>MLF1</i>	ENSG00000178053.17	Candidate	[120]
<i>MNS1</i>	ENSG00000138587.5	Candidate	[120]
<i>MOK</i>	ENSG00000080823.21	Candidate	[120]
<i>MORN3</i>	ENSG00000139714.12	Candidate	[120]
<i>MST1</i>	ENSG00000173531.15	Candidate	[120]
<i>MXI1</i>	ENSG00000119950.20	Candidate	[120]
<i>MYB</i>	ENSG00000118513.18	Candidate	[120]
<i>MYCBP</i>	ENSG00000214114.8	Candidate	[122]
<i>NASP</i>	ENSG00000132780.16	Candidate	[122]
<i>NEK1</i>	ENSG00000137601.15	Established	[120]
<i>NEK2</i>	ENSG00000117650.12	Established	[120]
<i>NEK4</i>	ENSG00000114904.12	Candidate	[120]
<i>NEK8</i>	ENSG00000160602.13	Established	[120]
<i>NEK9</i>	ENSG00000119638.12	Established	[120]
<i>NIN</i>	ENSG00000100503.23	Candidate	[120]
<i>NINL</i>	ENSG00000101004.14	Candidate	[120]
<i>NME5</i>	ENSG00000112981.4	Candidate	[120]
<i>NME7</i>	ENSG00000143156.13	Established	[120]
<i>NME8</i>	ENSG00000086288.11	Established	[120]
<i>NOTO</i>	ENSG00000214513.3	Candidate	[120]
<i>NPHP1</i>	ENSG00000144061.12	Established	[120]
<i>NPHP3</i>	ENSG00000113971.18	Established	[120]

<i>NPHP4</i>	ENSG00000131697.17	Established	[120]
<i>NUBP1</i>	ENSG00000103274.10	Candidate	[120]
<i>NUBP2</i>	ENSG00000095906.16	Candidate	[120]
<i>OCRL</i>	ENSG00000122126.15	Candidate	[120]
<i>ODF2</i>	ENSG00000136811.16	Candidate	[120]
<i>OFD1</i>	ENSG00000046651.14	Established	[120]
<i>ORC1</i>	ENSG00000085840.12	Candidate	[120]
<i>OSR1</i>	ENSG00000143867.6	Candidate	[120]
<i>PACRG</i>	ENSG00000112530.11	Candidate	[120]
<i>PAFAH1B1</i>	ENSG00000007168.12	Candidate	[120]
<i>PAH</i>	ENSG00000171759.9	Candidate	[122]
<i>PARD3</i>	ENSG00000148498.15	Candidate	[120]
<i>PCM1</i>	ENSG00000078674.17	Candidate	[120]
<i>PCNT</i>	ENSG00000160299.16	Candidate	[120]
<i>PDE1C</i>	ENSG00000154678.16	Candidate	[122]
<i>PDE6D</i>	ENSG00000156973.13	Established	[120]
<i>PEBP1</i>	ENSG00000089220.4	Candidate	[120]
<i>PIBF1</i>	ENSG00000083535.15	Established	[120]
<i>PIFO</i>	ENSG00000173947.13	Candidate	[120]
<i>PIH1D3</i>	ENSG00000080572.12	Established	[120]
<i>PIK3R4</i>	ENSG00000196455.7	Established	[120]
<i>PKD1</i>	ENSG00000008710.17	Established	[120]
<i>PKD2</i>	ENSG00000118762.7	Established	[120]
<i>PKD1L1</i>	ENSG00000158683.7	Established	[120]
<i>PKD2L1</i>	ENSG00000107593.16	Candidate	[120]
<i>PKHD1</i>	ENSG00000170927.14	Established	[120]
<i>PLK1</i>	ENSG00000166851.14	Candidate	[120]
<i>PLK4</i>	ENSG00000142731.10	Candidate	[120]
<i>POC1A</i>	ENSG00000164087.7	Candidate	[120]
<i>POC1B</i>	ENSG00000139323.13	Established	[120]
<i>POC5</i>	ENSG00000152359.14	Candidate	[120]
<i>POMGNT1</i>	ENSG00000085998.13	Established	[120]
<i>POMT2</i>	ENSG00000009830.11	Candidate	[132]
<i>PPP1R32</i>	ENSG00000162148.10	Candidate	[120]
<i>PPP2CA</i>	ENSG00000113575.9	Candidate	[120]
<i>PPP2R1A</i>	ENSG00000105568.17	Candidate	[120]

<i>PRKACA</i>	ENSG00000072062.13	Candidate	[120]
<i>PRKACB</i>	ENSG00000142875.19	Candidate	[120]
<i>PRKAR1A</i>	ENSG00000108946.14	Candidate	[120]
<i>PRKAR2A</i>	ENSG00000114302.15	Candidate	[120]
<i>PRKAR2B</i>	ENSG00000005249.12	Candidate	[120]
<i>PROSER3</i>	ENSG00000167595.14	Candidate	[120]
<i>PTCH1</i>	ENSG00000185920.15	Established	[120]
<i>PTK7</i>	ENSG00000112655.15	Candidate	[120]
<i>PTPDC1</i>	ENSG00000158079.14	Candidate	[120]
<i>RAB11A</i>	ENSG00000103769.9	Candidate	[120]
<i>RAB11B</i>	ENSG00000185236.11	Candidate	[120]
<i>RAB11FIP3</i>	ENSG00000090565.15	Candidate	[120]
<i>RAB17</i>	ENSG00000124839.12	Candidate	[120]
<i>RAB23</i>	ENSG00000112210.11	Established	[120]
<i>RAB28</i>	ENSG00000157869.14	Established	[120]
<i>RAB29</i>	ENSG00000117280.12	Candidate	[120]
<i>RAB3IP</i>	ENSG00000127328.21	Candidate	[120]
<i>RAB8A</i>	ENSG00000167461.11	Candidate	[120]
<i>RFX2</i>	ENSG00000087903.12	Candidate	[120]
<i>RFX3</i>	ENSG00000080298.15	Candidate	[120]
<i>RILPL1</i>	ENSG00000188026.11	Candidate	[120]
<i>RILPL2</i>	ENSG00000150977.10	Candidate	[120]
<i>ROPN1</i>	ENSG00000065371.17	Candidate	[120]
<i>ROPN1L</i>	ENSG00000145491.11	Candidate	[120]
<i>RP1</i>	ENSG00000104237.6	Established	[120]
<i>RP2</i>	ENSG00000102218.5	Established	[120]
<i>RP1L1</i>	ENSG00000183638.5	Established	[120]
<i>RPGR</i>	ENSG00000156313.12	Established	[120]
<i>RPGRIP1</i>	ENSG00000092200.12	Established	[120]
<i>RPGRIP1L</i>	ENSG00000103494.12	Established	[120]
<i>RSG1</i>	ENSG00000132881.11	Candidate	[120]
<i>RSPH1</i>	ENSG00000160188.9	Established	[120]
<i>RSPH3</i>	ENSG00000130363.11	Established	[120]
<i>RSPH9</i>	ENSG00000172426.15	Established	[120]
<i>RSPH14</i>	ENSG00000100218.11	Candidate	[120]
<i>RSPH4A</i>	ENSG00000111834.12	Established	[120]

<i>RTTN</i>	ENSG00000176225.12	Candidate	[120]
<i>SASS6</i>	ENSG00000156876.9	Candidate	[120]
<i>SAV1</i>	ENSG00000151748.14	Candidate	[120]
<i>SAXO1</i>	ENSG00000155875.13	Candidate	[120]
<i>SCLT1</i>	ENSG00000151466.11	Established	[120]
<i>SDCCAG8</i>	ENSG00000054282.15	Established	[120]
<i>SEPT2</i>	ENSG00000168385.17	Candidate	[120]
<i>SEPT7</i>	ENSG00000122545.18	Candidate	[120]
<i>SIL1</i>	ENSG00000120725.12	Established	[126]
<i>SLC47A2</i>	ENSG00000180638.17	Candidate	[120]
<i>SLC4A1</i>	ENSG00000004939.13	Candidate	[120]
<i>SMO</i>	ENSG00000128602.9	Established	[120]
<i>SNAP25</i>	ENSG00000132639.12	Candidate	[120]
<i>SNX10</i>	ENSG00000086300.15	Candidate	[120]
<i>SPA17</i>	ENSG000000064199.6	Candidate	[120]
<i>SPAG1</i>	ENSG00000104450.12	Established	[120]
<i>SPAG6</i>	ENSG00000077327.15	Candidate	[120]
<i>SPAG16</i>	ENSG00000144451.18	Established*	[123],[127]
<i>SPAG17</i>	ENSG00000155761.13	Candidate	[120]
<i>SPATA7</i>	ENSG00000042317.16	Established	[120]
<i>SPEF1</i>	ENSG00000101222.12	Candidate	[120]
<i>SPEF2</i>	ENSG00000152582.12	Established*	[133]
<i>SPICE1</i>	ENSG00000163611.11	Candidate	[120]
<i>SRD5A1</i>	ENSG00000145545.11	Candidate	[121]
<i>SSNA1</i>	ENSG00000176101.11	Candidate	[120]
<i>SSTR3</i>	ENSG00000278195.1	Candidate	[120]
<i>SSX2IP</i>	ENSG00000117155.16	Candidate	[120]
<i>STIL</i>	ENSG00000123473.15	Candidate	[120]
<i>STK3</i>	ENSG00000104375.15	Candidate	[120]
<i>STK36</i>	ENSG00000163482.11	Established*	[127]
<i>STK39</i>	ENSG00000198648.10	Candidate	[120]
<i>STK38L</i>	ENSG00000211455.7	Candidate	[120]
<i>STOML3</i>	ENSG00000133115.11	Candidate	[120]
<i>STX3</i>	ENSG00000166900.14	Candidate	[120]
<i>SUFU</i>	ENSG00000107882.11	Established	[120]
<i>SYNE2</i>	ENSG00000054654.15	Candidate	[120]

<i>TAPT1</i>	ENSG00000169762.16	Established	[120]
<i>TBC1D7</i>	ENSG00000145979.17	Candidate	[120]
<i>TBC1D31</i>	ENSG00000156787.16	Candidate	[120]
<i>TBC1D32</i>	ENSG00000146350.13	Established	[120]
<i>TBX3</i>	ENSG00000135111.14	Candidate	[120]
<i>TCTEX1D2</i>	ENSG00000213123.10	Established	[120]
<i>TCTN1</i>	ENSG00000204852.15	Established	[120]
<i>TCTN2</i>	ENSG00000168778.11	Established	[120]
<i>TCTN3</i>	ENSG00000119977.20	Established	[120]
<i>TEKT1</i>	ENSG00000167858.12	Established*	[131]
<i>TEKT2</i>	ENSG00000092850.11	Candidate	[120]
<i>TEKT3</i>	ENSG00000125409.12	Candidate	[120]
<i>TEKT4</i>	ENSG00000163060.7	Candidate	[120]
<i>TEKT5</i>	ENSG00000153060.7	Candidate	[120]
<i>TEX9</i>	ENSG00000151575.14	Candidate	[120]
<i>TMEM17</i>	ENSG00000186889.9	Established	[120]
<i>TMEM67</i>	ENSG00000164953.15	Established	[120]
<i>TMEM80</i>	ENSG00000177042.14	Candidate	[120]
<i>TMEM107</i>	ENSG00000179029.14	Established	[120]
<i>TMEM138</i>	ENSG00000149483.11	Established	[120]
<i>TMEM216</i>	ENSG00000187049.9	Established	[120]
<i>TMEM231</i>	ENSG00000205084.10	Established	[120]
<i>TMEM237</i>	ENSG00000155755.18	Established	[120]
<i>TOPORS</i>	ENSG00000197579.7	Established	[120]
<i>TPRA1</i>	ENSG00000163870.14	Candidate	[120]
<i>TRAF3IP1</i>	ENSG00000204104.11	Established	[120]
<i>TRAPPC3</i>	ENSG00000054116.11	Established	[126]
<i>TRIM32</i>	ENSG00000119401.10	Established	[120]
<i>TRIP11</i>	ENSG00000100815.12	Candidate	[120]
<i>TTBK2</i>	ENSG00000128881.16	Established	[120]
<i>TTC8</i>	ENSG00000165533.18	Established	[120]
<i>TTC17</i>	ENSG00000052841.14	Candidate	[120]
<i>TTC25</i>	ENSG00000204815.8	Established	[120]
<i>TTC26</i>	ENSG00000105948.13	Candidate	[120]
<i>TTC29</i>	ENSG00000137473.17	Candidate	[120]
<i>TTC21A</i>	ENSG00000168026.16	Candidate	[120]

<i>TTC30A</i>	ENSG00000197557.6	Candidate	[120]
<i>TTC21B</i>	ENSG00000123607.14	Established	[120]
<i>TTC30B</i>	ENSG00000196659.9	Candidate	[120]
<i>TTLL3</i>	ENSG00000214021.15	Candidate	[120]
<i>TTLL5</i>	ENSG00000119685.19	Established	[120]
<i>TTLL6</i>	ENSG00000170703.15	Candidate	[120]
<i>TTLL9</i>	ENSG00000131044.16	Candidate	[120]
<i>TUB</i>	ENSG00000166402.8	Established	[120]
<i>TUBA1A</i>	ENSG00000167552.13	Candidate	[120]
<i>TUBA4A</i>	ENSG00000127824.13	Candidate	[120]
<i>TUBA1C</i>	ENSG00000167553.14	Candidate	[120]
<i>TUBB</i>	ENSG00000196230.12	Candidate	[120]
<i>TUBB2A</i>	ENSG00000137267.5	Candidate	[120]
<i>TUBB2B</i>	ENSG00000137285.9	Candidate	[120]
<i>TUBB3</i>	ENSG00000258947.6	Candidate	[120]
<i>TUBE1</i>	ENSG00000074935.13	Candidate	[120]
<i>TUBGCP6</i>	ENSG00000128159.11	Candidate	[120]
<i>TULP1</i>	ENSG00000112041.12	Established	[120]
<i>TULP3</i>	ENSG00000078246.15	Candidate	[120]
<i>ULK4</i>	ENSG00000168038.10	Candidate	[120]
<i>UNC119</i>	ENSG00000109103.11	Established	[120]
<i>USP33</i>	ENSG00000077254.14	Candidate	[120]
<i>USP9X</i>	ENSG00000124486.12	Established	[120]
<i>VHL</i>	ENSG00000134086.7	Candidate	[120]
<i>WDPCP</i>	ENSG00000143951.15	Established	[120]
<i>WDR3</i>	ENSG00000065183.15	Candidate	[122]
<i>WDR19</i>	ENSG00000157796.17	Established	[120]
<i>WDR34</i>	ENSG00000119333.11	Established	[120]
<i>WDR35</i>	ENSG00000118965.14	Established	[120]
<i>WDR60</i>	ENSG00000126870.15	Established	[120]
<i>WDR62</i>	ENSG00000075702.16	Candidate	[120]
<i>WDR78</i>	ENSG00000152763.16	Candidate	[120]
<i>WNK1</i>	ENSG00000060237.16	Candidate	[120]
<i>WNK4</i>	ENSG00000126562.16	Candidate	[120]
<i>WRAP73</i>	ENSG00000116213.15	Candidate	[120]
<i>XPNPEP3</i>	ENSG00000196236.12	Established	[120]

<i>ZIC2</i>	ENSG00000043355.10	Established	[120]
<i>ZMYND10</i>	ENSG00000004838.13	Established	[120]
<i>ZMYND12</i>	ENSG000000066185.12	Candidate	[122]
<i>ZNF423</i>	ENSG00000102935.11	Established	[120]

*: In a previous study [120], these genes were labeled as “candidate,” but I changed them to “established” because causative mutations have been identified in these genes in recent reports [123–125,127,129–131,133].