Systematic identification of intron retention associated variants from massive publicly available transcriptome sequencing data

Yuichi Shiraishi\textsuperscript{1,†,*}, Ai Okada\textsuperscript{1,†}, Kenichi Chiba\textsuperscript{1}, Ikuko Omori\textsuperscript{2}, Raúl Nicolás Mateos\textsuperscript{1}, Naoko Iida\textsuperscript{1}, Hirofumi Yamauchi\textsuperscript{2}, Kenjiro Kosaki\textsuperscript{3}, Akihide Yoshimi\textsuperscript{2}

\textsuperscript{1} Division of Genome Analysis Platform Development, National Cancer Center Research Institute, Tokyo, Japan
\textsuperscript{2} Cancer RNA Research Unit, National Cancer Center Research Institute, Tokyo, Japan
\textsuperscript{3} Center for Medical Genetics, Keio University School of Medicine, Tokyo, Japan

\textsuperscript{†} These authors equally contributed to this work
\textsuperscript{*} Corresponding author: Yuichi Shiraishi, Ph.D.
5-1-1, Tsukiji, Chuo-ku, Tokyo, 104-0045, Japan
E-mail: yuishira@ncc.go.jp

Abstract

Many disease-associated genomic variants disrupt gene function through abnormal splicing. With the advancement of genomic medicine, identifying disease-associated splicing associated variants has become more important than ever. Most bioinformatics approaches to detect splicing associated variants require both genome and transcriptomic data. However, there are not many datasets where both of them are available. In this study, we developed a methodology to detect genomic variants that cause splicing changes (more specifically, intron retention), using transcriptome sequencing data alone. After demonstrating its high sensitivity and precision, we have applied it to 230,988 transcriptome sequencing data from the publicly available repository and identified 27,937 intron retention associated variants (IRAVs). In addition, by exploring positional relationships with variants registered in existing disease databases, we extracted 3,077 putative disease-associated IRAVs, which range from cancer drivers to variants linked with autosomal recessive disorders. The new in-silico screening framework proposed here provides a foundation for a platform that can automatically acquire medical knowledge making the most of massively accumulated publicly available sequencing data. Collections of IRAVs identified in this study are available through IRAVDB (https://iravdb.io/).

Introduction

During the continued innovation in sequencing technology, the effectiveness of large-scale genome analysis has been thoroughly validated and widely recognized. Nowadays, national-scale genome projects have been moving forward worldwide, and genome analysis is expected to revolutionize the medical system. It is more important than ever to identify disease-associated variants from a vast list of mutations obtained by sequencing patients. However, there still remain many
challenges for interpreting the effect of each genomic variant, especially those that contribute to
disease by a different mechanism than amino-acid substitutions.

One important class of pathogenic variants is those causing abnormal splicing changes,
typically by damaging existing splicing motifs or creating novel splicing motifs. They may comprise
15%-60% of human disease variants\textsuperscript{1,2}. There have been many attempts to catalog these splicing
associated variants. One approach is to resort to machine learning-based methods\textsuperscript{3}. However, such
approaches are still in their infancy in terms of precision and recall. In addition, predicting the
consequence of splicing changes, which is often vital for evaluating the pathogenicity of variants\textsuperscript{4}, is
even more challenging. Another approach is to perform genome and transcriptome analysis and
identify pairs of genome variants and the corresponding splicing changes through some statistical
methods\textsuperscript{5–8}. However, this approach needs a dataset where both genome and transcriptome data are
provided, which is not very common.

Here, to make the most of massive collections of transcriptome data available in public
sequence repositories\textsuperscript{9,10}, we provide a novel approach, IRAVNet (https://github.com/friend1ws/iravnet),
which can identify genomic variants that cause specific types of splicing aberrations, intron retention, by
using transcriptome sequencing data alone. We confirmed that intron retention associated variants
(IRAVs) can be detected with high sensitivity and accuracy based on transcriptome sequencing data\textsuperscript{6}. Using this method, we performed a comprehensive screening of intron retention variants mainly using
massive public transcriptome sequencing data registered in Sequence Read Archive (SRA). Also, we
have prepared a web-based portal site, IRAVDB (https://iravdb.io/), where users can view various
information about the IRAVs of interest, such as positional relationships with registered pathogenic
variants and a list of the SRA samples with those variants.

Result

Method overview

When a variant is a direct cause of intron retention by disrupting existing splicing donor/acceptor
motifs, we can observe mismatched bases in most of the retained short-reads at the exon-intron
boundary (Figure 1a, Supplementary Figure 1). We exploited this phenomenon to develop an algorithm
that can precisely identify IRAVs. Briefly, IRAVNet first lists up putative variants supported by three or
more short reads around the exon-intron boundaries (three exonic and six intronic bases for splice
donor sites, and six intronic and one exonic base for splice acceptor sites). Next, IRAVNet only keeps
the candidates whose variants are specifically supported by the intron-retention supporting reads and
not by normally spliced reads. Finally, IRAVNet removes the potential artifacts such as those
presumably produced by alignment errors around exon-intron boundaries. We also removed common
variants (those whose allele frequencies are greater than 0.01 by gnomAD database\textsuperscript{11}) to focus on
variants having a significant effect on disease while keeping the false positive rate low. See the Method
section for a more detailed description.
Application on TCGA dataset

To test the effectiveness of this approach, we conducted a preliminary analysis using 11,312 transcriptome sequencing data from The Cancer Genome Atlas (TCGA). In total, 2,767 IRAVs were identified after merging the variants with the same position and substitution found in multiple samples (Supplementary Data 1). To investigate whether the predicted IRAVs identified from transcriptome by IRAVNet are truly genomic variants or not, we developed a framework for validating the genomic mutation status of IRAVs using paired exome sequencing data. We classified the IRAVs into somatic, germline, somatic or germline, ambiguous, or false positive by exploring the number of supporting reads and sequencing depths at the positions of IRAVs for the 3,144 corresponding pairs of tumor and matched control exome sequencing data (Figure 2a). This revealed that the ratio of false positives was estimated to be as low as 0.93% (Figure 2b). Next, we evaluated the amount of intron retention for the samples with IRAVs compared to others. We confirmed that, in most cases, significantly higher ratios of intron retention were observed specifically in the samples having IRAVs (Figure 2c). To evaluate the sensitivity of the proposed approach, we performed a comparison with our previous approach, SAVNet\(^5,12\) which collects somatic splicing associated variants making use of paired genome and transcriptome data. IRAVNet, which just utilizes transcriptome data, detected about 43.7% (1,045 / 2,393) of intron retention causing somatic variants identified by SAVNet. In addition, the proposed approach identified 348 novel variants out of the 1,321 variants classified as “somatic” by the above...
Figure 2: Validation of the IRAVNet approach using TCGA transcriptome and exome sequencing data.
(a) The flowchart showing the classification and validation of genome level mutation status for IRAVs. Here, according to sequencing depths, the variant counts (the numbers of intron retention reads), and variant allele frequencies (VAF, the ratios of intron retention reads to the total number of sequencing reads covering the corresponding exon-intron boundaries), the status are classified into “Germline,” “Somatic,” “Somatic or germline,” “Ambiguous” and “False positive.” (b) The number of IRAVs categorized by the inferred mutation status for the identified IRAVs determined by the above procedure.
procedure (Supplementary Figure 2). Therefore, IRAVNet achieves a certain level of sensitivity and a high rate of precision, even though it only uses transcriptome data. Furthermore, most of the IRAVs detected by this approach are thought to actually cause intron retention. IRAVNet was able to identify 326 distinct IRAVs that affect well-known cancer genes such as TP53, STK11, SMARCA4, FAT1, SETD2 (Figure 2d, Supplementary Figure 3, Supplementary Data 2). In addition, strong enrichment of several genes on specific cancer types reflected well on the previous findings (e.g., high concentration of STK11 and KEAP1 in lung adenocarcinomas). These results indicate that this approach can effectively catalog disease-associated variants.

A computational framework for applying Sequence Read Archive

Next, to obtain a more comprehensive list of IRAVs, we applied this approach to publicly available transcriptome sequencing data from SRA (Figure 1b). For this purpose, we developed a cloud-based platform utilizing Amazon Web Service (Supplementary Figure 4) as well as on-premise computational clusters (Supplementary Figure 5). For the cloud-based platform, we automated the entire process of analysis using serverless architecture; download the raw sequence data, alignment to the human reference genome, and perform IRAVNet to identify IRAVs. To make the analysis reproducible, we utilized a container orchestration framework. We made various efforts to keep the cloud usage cost down by choosing the region where the data is located (to reduce downloading time), selecting the optimal instance type for each procedure, setting the right amount of block storage to be reserved, and usage of spot instances (Supplementary Figure 6).

Screening of intron retention associated variants using Sequencing Read Archive

We have analyzed 219,615 transcriptome sequencing data (counted based on run IDs) and integrated the result into that of TCGA data (Figure 3a). In the SRA, there were many sequencing data with different run IDs that were actually derived from the same individuals as exemplified by frequently used cell lines with multiple experimental conditions. Therefore, to avoid double counting of IRAVs, we basically focused on “distinct IRAVs” in the following, in which those with the same genomic position and substitution patterns identified from multiple sequence data are counted as one.

We detected 27,936 distinct IRAVs in total. Of which, 26,581 were on coding genes defined by RefSeq, and 145, 24,566 and 1,870 distinct IRAVs involved splices sites at 5'UTR, coding, and 3'UTR regions, respectively. Most of the IRAVs associated with coding regions were predicted to create premature termination codons (24,061, 97.9%), 23,849 (99.1%) of which were inferred to be nonsense-
mediated decay (NMD) sensitive by the 50nt rules\textsuperscript{15} (premature termination codons are located before the 50bp upstream of the last exon-exon junction). In the following, we focused on these NMD-sensitive IRAVs (22,060 SNVs and 1,789 indels in 7,913 RefSeq coding genes), because they are plausible to be associated with loss-of-function\textsuperscript{11}. 17,369 IRAVs were located at splice donor sites, whereas 6,480 IRAVs disrupted splice acceptor sites. 10,647 (44.6\%) IRAVs were those that did not involve GT-AT essential splice sites (Figure 3b). The distribution of substitution patterns of identified IRAVs at each exonic and intronic position of splice donor and acceptor sites show similar patterns for splicing associated variants with the previous study\textsuperscript{5}, where many were found at GT-AG essential splices sites followed by the last exonic and the 5th intronic bases of splice donor sites (Figure 3c). Furthermore, splicing donor motifs grouped by the relative positions of IRAVs were separated into left-handed and right-handed motifs, where higher entropy masses were placed in the exon and in the intron, respectively\textsuperscript{16} (Figure 3d).

**Pathogenicity of intron retention associated variants obtained from Sequence Read Archive**

Next, to search for IRAVs that are considered to be linked to disease, we took an approach of extending existing knowledge related to disease and we divided variants into 5 tiers based on the
positional relationships with the known pathogenic variants registered in ClinVar. 361 IRAVs sharing the same positions and substitution patterns with already registered pathogenic/likely pathogenic (P/LP) splicing variants were defined as Tier 1. 720 IRAVs

**Figure 4:** The landscape of putative pathogenic IRAVs. (a) A conceptual diagram for the classification of disease-related IRAVs in four tiers. The ranks of putative pathogenic IRAVs are determined by comparison with the positional relationships with the registered pathogenic/likely pathogenic (P/LP) variants in ClinVar. (b) The saturation analysis of putative pathogenic IRAVs via using Sequencing Read Archive. Transcriptome data were sorted by registration date, and the cumulative number of putative pathogenic IRAVs against the number of the analyzed transcriptome is shown. Gray vertical lines indicate the registration year changes. (c) The landscape of putative pathogenic IRAVs for frequently altered genes (total number ≥10). Genes were sorted by the total number of distinct variants, and known cancer-related genes are shown in red. (d) The number of distinct putative pathogenic IRAVs affecting cancer-related genes for each rank identified from Sequencing Read Archive, TCGA, and both.
 sharing the same splice sites with registered P/LP splice-site variants were classified as Tier 2. 887 IRAVs for which P/LP truncating variants are seen within 30bp from the exon-intron boundary of the IRAVs were categorized as Tier 3. 1,107 NMD-sensitive IRAVs that affect genes with at least one registered P/LP truncating variant were designated as Tier4, and other variants were set to Tier 5. Here, Tier 1 to Tier 4 IRAVs were considered to be putative pathogenic IRAVs (ppIRAVs) and will be discussed in the following. The number of ppIRAVs identified showed an almost linear increase with the number of transcriptome sequence data analyzed and has not yet reached saturation. As the data accumulate at an accelerated rate, much more ppIRAVs are expected to be found via further analysis (Figure 4b).

Putative pathogenic intron retention associated variants affecting disease-related genes

The transcriptome sequencing data from Sequencing Read Archive includes a number of cancer cells. In total with TCGA RNA sequencing data, 645 ppIRAVs (Tier1: 123, Tier2: 209, Tier3: 145, Tier4: 168) were those affecting cancer-related genes (Figure 4c). The fact that most of them (418, 81.3%) were only identified from SRA (Figure 4d) indicates the usefulness of SRA, which is overwhelmingly superior in number, compared to TCGA data. Most of the cancer-related IRAVs identified from SRA were considered to come from cancer transcriptome data and to be somatic variants from the fact that their population allele frequencies (via gnomAD) were generally very low. Genes with >= 15 novel ppIRAVs were TP53 (53 ppIRAVs), MYH9 (26), NF1 (24), SMARCA4 (23), STK11 (20), RB1 (19), FLNA (17), TSC2 (17), BCOR (15) (Figure 5a,c,d, Supplementary Figure 8). Generally, cancer-related ppIRAVs tended to concentrate on specific splice-sites, and 36 splice-sites had multiple ppIRAVs. In extreme cases, 19 ppIRAVs were concentrated at the 4th exon splice donor site of TP53, of which nine ppIRAVs have not been reported in ClinVar (Figure 5b). Intron retention caused by the IRAVs at this site was shown to be associated with overexpression of ∆133p53 transcript, which may promote cancer cell invasion. Six ppIRAVs were identified in the splice donor site of the 6th exon in YY1AP1 gene, whose relationships with Grange syndrome and hepatocellular carcinoma have been reported. Other hotspot splice-sites included the splice donor site of TP53 exon 9 (6 ppIRAVs), the splice donor site of SMARCA4 exon 34 (5), and the splice acceptor site of STK11 exon 7 (5).

Other recurrent ppIRAVs were detected in a number of genes known to be associated with disease such as MYH9 (26 ppIRAVs), PHKB (16), COL7A1 (14), and so on (Figure 4c, Supplementary Figure 9). MYH9 has been known to cause an autosomal-dominant disease called MYH9-related disorder (MYH9-RD), characterized by large platelets and thrombocytopenia as well as increased risk of progressive nephropathy, sensorineural deafness, pre-senile cataract, and aberration of liver enzymes. Moreover, several studies using animal models suggested that MYH9 may act as a tumor suppressor and inactivation of MYH9 was related to the development of squamous cell carcinomas and invasive lobular breast carcinomas. PHKB is associated with the autosomal recessive type of glycogen storage disease type IX. Many truncating and splicing variants have been reported in ClinVar and we have identified IRAVs near them as well as other locations. Mutations in the COL7A1 gene cause dystrophic epidermolysis bullosa, leading to subepidermal blistering and mucocutaneous fragility. The severity of the disease varies depending on the type of mutation, and the degree to which the ppIRAVs cause intron retention may have a different impact on this disorder. In addition, 688
were those affecting haploinsufficient genes, and 251 ppIRAVs (Tier1: 79, Tier2: 108, Tier3: 34, Tier4: 30) were identified in ACMG genes list version 3.0 (Supplementary Figure 7).

Collectively, our approach of screening ppIRAVs with massive transcriptome data can still greatly improve the current knowledge of variants causing splicing changes even for well-known disease-related genes.

Putative intron retention associated variants related to drug response

Loss of function mutations exemplified by variants causing splicing changes occasionally provides important information on the potential safety and effect of drugs. We also identified IRAVs related to drug response in the same manner and identified 14 IRAVs (Tier1: 2, Tier2: 2, Tier4: 10)
predicted to be related to drug response (Supplementary Data 4). Eleven distinct IRAVs were identified in \textit{DPYD}, mutations of which are implicated to increased risk of toxicity in cancer patients receiving 5-fluorouracil chemotherapy\textsuperscript{31}. One variant, c.1905+1G>A, is relatively frequent (allele frequency: 0.0047), well characterized\textsuperscript{32}, and “reviewed by expert panel” status in ClinVar. Although the other nine IRAVs were very rare (≤0.0001 allele frequency), they are, in aggregate, potentially very important when administering medication. Other genes included \textit{CYP2C19} (2 IRAVs), which is known to influence the enzyme activity in the metabolic pathway of drugs\textsuperscript{33}.

\textbf{Relatively common pathogenic intron retention associated variants implicated in genetic disorders}

Even though we focused on rare variants (≤0.01 allele frequency), 180 ppIRAVs (Tier1: 54, Tier2: 36, Tier3: 35, Tier4: 55) were relatively common (≥ 0.0001 allele frequency). Among them, 18 ppIRAVs (Tier1: 8, Tier2: 5, Tier3: 5, Tier4: 0) affected cancer predisposition genes\textsuperscript{34} and 12 ppIRAVs (Tier1: 2, Tier2: 8, Tier3: 0, Tier4: 2) were located at ACMG genes list version 3.0\textsuperscript{29}. They are considered to be germline variants that work as factors for genetic diseases and might be a good candidate for a drug target. Many of them have been registered in ClinVar as “Benign,” “Likely benign,” “Conflicting interpretations of pathogenicity,” or “Uncertain significance” partly because the effects of splicing are unclear. Using transcriptome sequence data collected from multiple tissues (Genotype-Tissue Expression project\textsuperscript{35}), we have confirmed that individuals with ppIRAVs tend to show specifically high intron retention ratios at corresponding exon-intron boundaries in multiple tissues (Figure 6a, Supplementary Figure 10, 11). Examples of relatively frequent ppIRAVs included the c.1473+5G>T variant at the 9th exon splice donor site of \textit{P3H1}, linked with osteogenesis imperfecta. This variant was observed in 405/152,156 (0.2662%) alleles (392/41,440 (0.9459%) for the African ancestry) in gnomAD database v3.1.1 and has not been reported as pathogenic. We confirmed by mini-gene assays that this variant induces strong intron 9 retention (Figure 6b). Given that this variant may cause intron 9 retention in multiple tissues and that a pathogenic variant was reported at the same splice site (c.1473+1G>T)\textsuperscript{36}, this variant may have some pathogenicity properties. The next example is the c.424+5G>A variant at the 8th exon splice donor site of \textit{SMAD4}. Although this variant has an even smaller allele frequency (19/151,834, (0.0125%)), it has been reported in several cancer-cohort studies\textsuperscript{37–39}. Previous studies have described this variant as a variant of uncertainty significance even though they pointed out the possibility that it causes some kind of abnormal splicing. However, intron retention verified by our mini-gene assay implied some kind of pathogenicity of this variant (Figure 6b). Overall, our approach provides an effective screening method for cataloging major genomic variants that are responsible for genetic disease through abnormal splicing.

\textbf{Discussion}

We have proposed a new framework for screening pathogenic variants through abnormal splicing, making the most of the vast amount of transcriptome data available in public repositories. Our methodology identifies not only many previously detected pathogenic variants but also a vast amount of novel ones. In addition, our saturation analysis demonstrated that more and more mutations would be identifiable by keeping applying this methodology as the sequence data accumulate in the repository.
It is possible to build a system that automatically stores IRAVs by running IRAVNet in conjunction with increasing importance to acquire relevant knowledge for medical treatment and prevention. In principle, the regulation mechanism of NMD and the 50nt rule may not be sufficient typically did improved. We utilized the 50nt rule for the prediction of NMD sensitivity as other genomic information such as race, tissue, and cancer status. Third, the prediction on the pathogenicity could be not well organized in terms of metadata, making it difficult to perform association studies with some sequencing data alone. Second, the Sequence Read Archive, which we mainly used in this paper, is other types of splicing associated variants than those causing intron retention. Furthermore, the correspondence table between the variants and sample IDs will enable researchers to download the set of sequencing data with the variants of interest for further detailed analyses. There are three possible challenges in line with this approach. First, we may be able to extract other types of splicing associated variants than those causing intron retention by using transcriptome sequencing data alone. Second, the Sequence Read Archive, which we mainly used in this paper, is not well organized in terms of metadata, making it difficult to perform association studies with some phenotypes. It may be useful to devise and apply computational prediction methods for basic information such as race, tissue, and cancer status. Third, the prediction on the pathogenicity could be improved. We utilized the 50nt rule for the prediction of NMD sensitivity as other genomic studies typically did. However, recent advances in this field have indicated the complex physiological regulation mechanism of NMD and the 50nt rule may not be sufficient.

With the accumulation of genome data via the implementation of genomic medicine, it will be increasingly important to acquire relevant knowledge for medical treatment and prevention. In principle, it is possible to build a system that automatically stores IRAVs by running IRAVNet in conjunction with

![Figure 6: Validation of relatively common IRAVs implicated in genetic disorders.](image-url)

(a) Relative ratios of corresponding intron retention for samples with (orange) and without (gray) the IRAVs across tissues measured using GTEx transcriptome data. (b) In vitro splicing analyses using wildtype (WT) or mutant (Mut) minigene constructs (left) showing intron retention introduced by IRAVs at P3H1 intron 9 and SMAD4 intron 3.
data registration and upload. Still, there are many challenges such as reanalysis after updating the software, data harmonization, and storage issues. However, while providing an integrative solution to them, we believe that platforms that can efficiently and autonomously perform knowledge discovery will revolutionize the research and healthcare system.

Method

A workflow for the discovery of intron retention associated variants

Selection of Sequence Read Archive and TCGA samples

Public Sequence Read Archive (SRA) samples were selected in a manner similar to the previous study. We queried the SRA website (https://www.ncbi.nlm.nih.gov/sra) with the following search term: "platform illumina"[Properties] AND "strategy rna seq"[Properties] AND "human"[Organism] AND "cluster public"[Properties] AND "biomol ma"[Properties]. Then, we extracted samples whose base number is ≥ 1 billion bases, to secure sufficient sequence coverage for reliably detect mutations. There were a number of run data that could not be downloaded even after repeated trials probably due to some technical issue. Some of the downloaded sequence data had severe problems such as inconsistencies between two paired-end files, different lengths between sequence letters and base qualities. In addition, we removed run data whose number of IRAVs was extremely high due to potential DNA contamination.

Downloading Sequence Read Archive samples

We used SRA Toolkit version 2.10.0. First, we performed prefetch command with the "----max-size 100000000" option to download SRA format file. Then, we executed fasterq-dump command with the option "-v --split-files."

Alignment of RNA-seq data

We used the GRCh38 based reference genome provided from Genomic Data Commons (https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files). First, genome indexes were generated using STAR version 2.7.2b (Dobin et al. 2013) with that reference genome and the release 31 GTF file (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_31/gencode.v31.annotation.gtf.gz) and --sjdbOverhang 100 option. For each sample, alignment to the reference genomes was performed by the same version of STAR with the following options: --runThreadN 6 --outSAMtype BAM Unsorted - -outSAMstrandField intronMotif --outSAMunmapped Within --outSJfilterCountUniqueMin 1 1 1 1 -- outSJfilterCountTotalMin 1 1 1 1 --outSJfilterOverhangMin 12 12 12 --outSJfilterDistToOtherSJmin 0 0 0 --alignIntronMax 500000 --alignMatesGapMax 500000 --alignSJstitchMismatchNmax -1 -1 -1 -1 -- chimSegmentMin 12 --chimJunctionOverhangMin 12. After the alignment, BAM files were sorted, and converted into CRAM format, and indexed using SAMtools version 1.9 (https://www.htslib.org/).
Detection of intron retention associated variants from CRAM format files

We focused on the known splice-site regions (from the 3rd exonic base to the 6th intronic base for splice donor sites, and from the 6th intronic base to the 1st exonic base) registered in RefSeq genes (https://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/refGene.txt.gz) as previously described, and created the BED format file for the above splice-site regions. From these splice-site regions, we removed those completely included in the exonic regions of other genes or isoforms.

First, we piled up the CRAM file with the “samtools mpileup” command confining to the remaining 418,464 splice-site regions and identified putative variants with at least 3 variant supporting reads and ≥0.05 variant allele frequencies. Then, we checked the position of the variant in each supporting read and kept variants that are supported by at least 3 different positions (and at least one of which must be inside the 5 bases from the edges).

Next, for each variant remaining at this stage, we classified short reads around the variant by pair-wise alignment into:
- Splicing junction positive (negative): short reads that are normally spliced at corresponding splice-site and do (not) support the target variant.
- Intron retention positive (negative): short reads that are not spliced and retained around the splice-site and do (not) support the target variant.

Let us denote the number of splicing junction positive, splicing junction negative, intron retention positive, and intron retention negative short reads as #SJ_Pos, #SJ_Neg, #IR_Pos, and #IR_Neg, respectively. We requested the following conditions:
- #IR_Pos ≥ 3.
- #IR_Pos / (#IR_Pos + #IR_Neg) ≥ 0.9 (to confirm that most intron retention reads include the target variant).
- #IR_Pos / (#SJ_Neg + #IR_Pos) ≥ 0.1 (to remove the variants with too low variant allele frequencies or too weak penetrance on the intron retention effect).

Also, at least one intron retention read needs to cover 25 intronic bases from the exon-intron boundary. Furthermore, to restrict reliable IRAVs, we requested that the MaxEntScore be reduced by at least 2.5 due to the variants. Finally, using gnomAD version 3.0, we removed the variants with high allele frequencies (>0.01). The entire workflow is available at https://github.com/friend1ws/iravnet.

Quantification of intron retention

Intron retention was quantified by our in-house program (intron_retention_utils, https://github.com/friend1ws/intron_retention_utils) as in the previous study. For each exon-intron boundary registered in RefSeq database, the number of presumed intron retention reads (those spanning ≥10 bp of both sides of the boundary), as well as that of normally spliced reads covering the last exonic base of the boundary, was counted.
TCGA exome and transcriptome analysis

Classification of intron retention associated variants using tumor and matched-control exome sequencing data

For each transcriptome sequencing data having detected IRAVs, we first checked whether the corresponding tumor and matched control exome sequencing data is available in the Genomic Data Commons. Here, we focused on the sample type of “Primary Tumor,” and “Primary Blood Derived Cancer - Peripheral Blood” (as well as “Metastatic” in cases when the cancer type is TCGA-SKCM) for tumor and “Solid Tissue Normal” “Blood Derived Normal” for matched control. When both tumor and matched control were obtained, we downloaded part of the BAM file around the detected IRAVs using the BAM Slicing API. Then, we measured sequencing depths, the numbers of intron retention supporting reads, and the ratios of intron retention supporting reads for the IRAVs and classified them into “germline,” “somatic,” “somatic or germline,” “ambiguous,” and “false positive” according to the flowchart presented in Figure 2a.

Investigation of the specificity of the amount of intron retention

The amount of intron retention was quantified by our in-house program (intron_retention_utils simple_count command) as described previously5. Briefly, for each exon-intron boundary, the number of intron retention reads (those covering ≥10 bp of both sides of the exon-intron boundary) and that of normally spliced reads covering the last exonic base of the exon-intron boundary were counted. For each detected IRAV, we measured the ratios of intron retention for the IRAV having sample and other samples with the same cancer type and quantified the Z-value comparing them.

Comparison with SAVNet

First, we confined the transcriptome data to those used in the previous study5 for comparison. We collected the splicing associated variant identified as causing intron retention in that study and transformed the coordinates of these variants from those based on GRCh37 to GRCh38 using liftOver (https://genome-store.ucsc.edu/). Finally, we evaluated the overlap between the IRAVs identified in this study and them.

Cloud-based platform for detecting intron retention associated variants from Sequence Read Archive

We used our in-house batch job engine on Amazon Web Service, ecsub (https://github.com/aokad/ecsub). This software first launches a virtual machine of an instance type suited to a target task (e.g., transcriptome alignment, gene expression measurement, and so on). Then it performs a series of Extraction Transformation Load (ETL) procedures, in which it extracts input files (e.g. FASTQ files) stored in Amazon S3 to the virtual machines, transforms the input files to output files (e.g. BAM files converted from the FASTQ file by alignment software), and then load the output file to Amazon S3. Finally, ecsub removes the virtual machines used for the ETL procedure. This software basically wraps Amazon Elastic Container Service to implement the above approaches, with additional
functions such as selecting the instance type and availability zone of SpotInstance, automatically deleting the master instance using the serverless framework (Amazon Lambda), and so on. This software is inspired by dsub (https://github.com/DataBiosphere/dsub), which is a tool to submit batch jobs on Google Cloud Platform. In this study, the pipeline is composed of four steps; downloading SRA samples, Alignment of transcriptome data, detection of IRAVs using iravnet, and intron retention quantification.

Investigating generation of premature termination codon and sensitivity to nonsense mediated-decay sensitivity

To explore the generation of premature termination codon (PTC) for each IRAV, we chose the longest transcript (defined by RefSeq transcript annotation) harboring the corresponding exon-intron boundary affected by the IRAV. Then, we constructed the transcript with the nucleotide sequence of the retained intron and investigated whether a stop codon occurs before the original termination codon or not. For IRAVs with PTCs, if the PTCs were located before 50 bp upstream of the last exon-exon junction, then those IRAVs were classified into NMD-sensitive, whereas the remaining PTC harboring IRAVs were deemed as NMD-insensitive.

Definition and classification of putative pathogenic intron retention associated variants

We investigated the positional relationships of the IRAVs with pathogenic variants registered in ClinVar VCF file downloaded from ClinVar FTP site (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/) as of 4th, 2021. The “pathogenic” variant in ClinVar is defined as those whose CLNSIG INFO key is either of “Pathogenic,” “Likely_pathogenic,” or “Pathogenic/Likely_pathogenic” in this study. For each IRAV, we inspected the positional relationships with pathogenic variants in the ClinVar VCF file. First, when a pathogenic variant with the same genomic position and substitution patterns with the IRAV exist in the ClinVar VCF file, then the IRAVs were classified into Tier1. When a pathogenic splicing variant (MC INFO key is “SO:0001575|splice_donor_variant,” or “SO:0001574|splice_acceptor_variant”) sharing the same splicing site (3 exonic and 6 intronic bases for splice donor sites, and 6 intronic and 1 exonic base for splice acceptor sites) with the IRAV were observed, the IRAV was designated as Tier2. Next, we searched for a pathogenic truncating variant (MC INFO key is “SO:0001587|nonsense,” or “SO:0001589|frameshift_variant”) within 30 bp from the corresponding exon-intron boundary at the transcript level, and the pathogenicity rank of the IRAV was set to Tier3 if a variant satisfying the condition was found. Finally, we explored for a pathogenic variant in the same gene with the IRAV, and the IRAV was categorized into Tier4 if one could be found. In the case of Tier4 investigation, we imposed additional constraints, CLNSTAT INFO key is either of “criteria_provided,_multiple_submitters,_no_conflicts,” “reviewed_by_expert_panel,” “practice_guideline,” to focus on variants with solid evidence.
Measurement of the amount of intron retention across multiple tissues

We have downloaded the GTEx transcriptome sequence data from SRA and aligned them using STAR, and quantified the amount of intron retention using our in-house program (intron_retention_utils simple_count command) as described in the previous sections. For genotype data, we used exome genotype calls from the GTEx Analysis V7.

Validation using In vitro assay

The P3H1 and SMAD4 mini-gene constructs were generated by inserting the DNA fragments containing the P3H1 genomic sequence spanning exons 9 and 10 and intervening intron 9, and SMAD4 genomic sequence from exon 3 to exon 4 in between the BamHI and EcoRI restriction sites of the pcDNA3.0(-) plasmid. Mutagenesis was performed with the primeSTAR Mutagenesis Basal Kit (Takara) with specific primers according to the manufacturer’s instructions. For transient transfection experiments, 293T cells were seeded into a 6-well plate one day before transfection of P3H1 or SMAD4 minigene constructs in the presence of PEI MAX (polysciences). 36 hours after transfection, cells were collected and RNA was extracted with FavorPrep Mini Kit (FAVORGEN). Mini-gene-derived and endogenous transcripts of P3H1 and SMAD4 were analyzed by RT-PCR using specific primers. Primers and oligonucleotides used in RT-PCR reactions were: P3H1 fwd (mini-gene) CGCAAAATGGGCGGTAGGCGTG, P3H1 fwd (endogenous) AGTCACTGGATGTGAGCAGACTGAC and rev (common) TGAGGGCTTTGAAGA CAGTGACAC; SMAD4 fwd (mini-gene) CGCAAAATGGGCGGTAGGCGTG, SMAD4 fwd (endogenous) ATACAGAGAACATTGGATGGGAGGCTTC and rev (common) ATTACTCTGCAGTGTAAATCCTGAGAG.

Data Availability

The list of IRAVs are available through IRAVDB (https://iravdb.io).

Acknowledgment

This work is supported by Grant-in-Aid for Scientific Research (KAKENHI 18H03327, 21H03549) and Grand-in-Aid from the Japan Agency for Medical Research and Development (Platform Program for Promotion of Genome Medicine: 20km0405207h9905, Program for an Integrated Database of Clinical and Genomic Information: 20kk0205014h0005, Practical Research Project for Rare / Intractable Diseases: 20ek0109485h0001, Practical Research for Innovative Cancer Control: 21ck0106641h0001). AY acknowledges support from the Japan Society for the Promotion of Science (JSPS) Home-Returning Researcher Development Research (grant number 19K24691), KAKENHI (grant number 21H04828), the Japan Agency for Medical Research and Development (grant number 21jm0210085h0002), National Cancer Center Research and Development Funds (2020-A-2). We used the super-computing resource provided by the Human Genome Center (The University of Tokyo) and ROIS National Institute of Genetics. The results shown here are partly based upon data generated by TCGA Research Network (https://cancergenome.nih.gov/) and the Genotype-Tissue Expression
Author Contributions

YS designed the study. YS developed the software for detecting intron retention associated variants. AO developed a cloud-based platform for analyzing massive transcriptome sequencing data deposited in Sequence Read Archive and a web-based database. AO developed the portal website for IRAVs with assistance from YS. IO and HY performed mini-gene assays under the supervision of AY. KC, RNM, and NI provided computational assistance. YS interpreted and organized the list of intron retention associated variants with assistance from KK and AY. YS and AO generated figures. YS wrote the manuscript. All authors participated in the discussion and interpretation of the data and results.

Supplementary Information

Supplementary Figure Legends

Supplementary Figure 1: Example of alignment status around intron retention associated variants by the depiction of Integrative Genomics Viewer. Here, the alignment views of the $TP53$ exon 4 donor site were shown for (a) The sample having the corresponding IRAVs (chr17:7675994 C>A) as well as (b) one with no IRAVs at that splice site.

Supplementary Figure 2: Venn diagrams showing the overlap of intron retention associated variants identified by the current (IRAVNet) and previous approach (SAVNet). (a) The overlaps with all the IRAVs detected by IRAVNet are depicted. (b) The overlap with IRAVs restricted to be those classified into “somatic” via tumor and matched control exome sequencing data are shown, since IRAVNet also detects germline variants whereas SAVNet only focuses on somatic variants. Splicing associated variants identified by SAVNet were restricted to those causing intron retention.

Supplementary Figure 3: Landscape of IRAVs in frequent cancer-related genes (total number ≥ 3) across cancer types. The font size indicates the number of affected samples. Genes are sorted by the total number of IRAVs in all cancer types. See also Figure 2d.

Supplementary Figure 4: A workflow for detecting IRAVs using Amazon Web Service. We performed three tasks for each transcriptome sequencing data; The tasks:fastq-dump performs downloading from Sequence Read Archive and converts the sequence data into FASTQ format. The task:star performs alignment to the reference genome using STAR. The task:irav performs IRAVNet to identify IRAVs. A serverless framework (Amazon Lambda) controls the job execution, which detects the placement meta-data describing the sample to analyze and launch the batch job using ecsub package (a wrapper for Amazon Elastic Container Service).
**Supplementary Figure 5:** A workflow for detecting IRAVs using Univa Grid Engine. We developed snakemake based pipeline, GCATWorkflow, to execute downloading FASTQ file, alignment to the reference genome, and IRAV detection in order. Also, two background jobs (uploader and notifier) upload the output to Amazon S3 and notify the success or failure to Slack, respectively.

**Supplementary Figure 6:** Evaluation of data size and computing resource. (a) The relationships between the number of bases extracted from meta-data file from Sequence Read Archive and actual FASTQ file size. The points that represent each sample line up almost exactly with the regression line for each project. (b) Relationships between FASTQ file size and disk usage when performing STAR alignment. The read lengths were 75bp and 100bp for DRP000987 and DRP001919, respectively.

**Supplementary Figure 7:** The landscape of putative pathogenic IRAVs in (a) haploinsufficient genes and (b) ACMG actionable genes. Genes are sorted by the total number of distinct variants.

**Supplementary Figure 8:** Distribution of IRAVs in genes with frequent IRAVs. (a-d) Frequencies of putative pathogenic IRAVs for each rank at each splice-site are shown in (a) RB1, (b) STK11, (c) TSC2, and (d) YY1AP1.

**Supplementary Figure 9:** Distribution of IRAVs in genes with frequent IRAVs. (a-d) Frequencies of putative pathogenic IRAVs for each rank at each splice-site are shown in (a) MYH9, (b) PHKB, (c) COL7A1, and (d) FLNA.

**Supplementary Figure 10:** (a, b) Relative ratios of corresponding intron retention for samples with and without the IRAVs of Tier1 (a), Tier2 (b) across tissues measured using GTEx transcriptome data. Each point shows an individual colored by IRAV mutation status. See also Figure 6a and Supplementary Figure 11.

**Supplementary Figure 11:** (a, b) Relative ratios of corresponding intron retention for samples with and without the IRAVs of Tier3 (a), Tier4 (b) across tissues measured using GTEx transcriptome data. Each point shows an individual colored by IRAV mutation status. See also Figure 6a and Supplementary Figure 10.

**Supplementary Data**

**Supplementary Data 1:** List of TCGA samples used in this study and the frequencies of their IRAVs.

**Supplementary Data 2:** List of IRAVs affecting cancer-related genes (those registered in Cancer Gene Census) identified in TCGA samples.

**Supplementary Data 3:** List of putative pathogenic IRAVs.

**Supplementary Data 4:** List of IRAVs related to drug response.
Reference

1. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.* **102**, 11–26 (2018).

2. Wang, G.-S. & Cooper, T. A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* **8**, 749–761 (2007).

3. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).

4. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

5. Shiraishi, Y. *et al.* A comprehensive characterization of cis-acting splicing-associated variants in human cancer. *Genome Res.* **28**, 1111–1125 (2018).

6. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* **47**, 1242–1248 (2015).

7. Jayasinghe, R. G. *et al.* Systematic Analysis of Splice-Site-Creating Mutations in Cancer. *Cell Reports* **23**, 270-281.e3.

8. Cao, S. *et al.* Discovery of driver non-coding splice-site-creating mutations in cancer. *Nat. Commun.* **11**, 5573 (2020).

9. Collado-Torres, L. *et al.* Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319–321 (2017).

10. Wilks, C. *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing. *bioRxiv* 2021.05.21.445138 (2021) doi:10.1101/2021.05.21.445138.

11. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
12. PCAWG Transcriptome Core Group et al. Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).

13. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).

14. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19-21 (2011).

15. Supek, F., Lehner, B. & Lindeboom, R. G. H. To NMD or Not To NMD: Nonsense-Mediated mRNA Decay in Cancer and Other Genetic Diseases. *Trends Genet.* (2020) doi:10.1016/j.tig.2020.11.002.

16. Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. *Nat. Rev. Genet.* **17**, 407–421 (2016).

17. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

18. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).

19. Lasham, A., Knowlton, N., Mehta, S. Y., Braithwaite, A. W. & Print, C. G. Breast Cancer Patient Prognosis Is Determined by the Interplay between TP53 Mutation and Alternative Transcript Expression: Insights from TP53 Long Amplicon Digital PCR Assays. *Cancers* **13**, (2021).

20. Gadea, G. et al. TP53 drives invasion through expression of its Δ133p53β variant. *Elife* **5**, e14734 (2016).

21. Guo, D.-C. et al. Loss-of-Function Mutations in YY1AP1 Lead to Grange Syndrome and a Fibromuscular Dysplasia-Like Vascular Disease. *Am. J. Hum. Genet.* **100**, 21–30 (2017).

22. Zhao, X. et al. Integrative genomics identifies YY1AP1 as an oncogenic driver in EpCAM(+) AFP(+) hepatocellular carcinoma. *Oncogene* **34**, 5095–5104 (2015).

23. Kunishima, S. & Saito, H. Advances in the understanding of MYH9 disorders. *Curr. Opin. Hematol.* **17**, 405–410 (2010).

24. Bury, L. et al. Next-generation sequencing for the diagnosis of MYH9-RD: Predicting pathogenic
variants. *Hum. Mutat.* **41**, 277–290 (2020).

25. Schramek, D. *et al.* Direct in vivo RNAi screen unveils myosin IIa as a tumor suppressor of squamous cell carcinomas. *Science* **343**, 309–313 (2014).

26. Kas, S. M. *et al.* Insertional mutagenesis identifies drivers of a novel oncogenic pathway in invasive lobular breast carcinoma. *Nat. Genet.* **49**, 1219–1230 (2017).

27. Shinkuma, S. Dystrophic epidermolysis bullosa: a review. *Clin. Cosmet. Investig. Dermatol.* **8**, 275–284 (2015).

28. Rehm, H. L. *et al.* ClinGen--the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).

29. Miller, D. T. *et al.* ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* (2021) doi:10.1038/s41436-021-01172-3.

30. Ingelman-Sundberg, M., Mkrtchian, S., Zhou, Y. & Lauschke, V. M. Integrating rare genetic variants into pharmacogenetic drug response predictions. *Hum. Genomics* **12**, 26 (2018).

31. van Kuilenburg, A. B. P. Dihydropyrimidine dehydrogenase and the efficacy and toxicity of 5-fluorouracil. *Eur. J. Cancer* **40**, 939–950 (2004).

32. Del Re, M. *et al.* DPYD*6 plays an important role in fluoropyrimidine toxicity in addition to DPYD*2A and c.2846A>T: a comprehensive analysis in 1254 patients. *Pharmacogenomics J.* **19**, 556–563 (2019).

33. Fleming, I. The pharmacology of the cytochrome P450 epoxygenase/soluble epoxide hydrolase axis in the vasculature and cardiovascular disease. *Pharmacol. Rev.* **66**, 1106–1140 (2014).

34. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302–308 (2014).

35. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

36. Cabral, W. A. *et al.* Prolyl 3-hydroxylase 1 deficiency causes a recessive metabolic bone disorder resembling lethal/severe osteogenesis imperfecta. *Nat. Genet.* **39**, 359–365 (2007).
37. Yurgelun, M. B. et al. Identification of a Variety of Mutations in Cancer Predisposition Genes in Patients With Suspected Lynch Syndrome. *Gastroenterology* **149**, 604–13.e20 (2015).

38. Tung, N. *et al.* Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel. *Cancer* **121**, 25–33 (2015).

39. Jelsig, A. M. *et al.* Germline variants in Hamartomatous Polyposis Syndrome-associated genes from patients with one or few hamartomatous polyps. *Scand. J. Gastroenterol.* **51**, 1118–1125 (2016).

40. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).