ARTICLE | Genetics inMedicine

Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq

Joseph K. Aicher, MA[1,2], Paul Jewell, BS[1,3], Jorge Vaquero-Garcia, MSc[1,3], Yoseph Barash, PhD[1,3] and Elizabeth J. Bhoj, MD, PhD[1,2,4]

Purpose: RNA-seq is a promising approach to improve diagnoses by detecting pathogenic aberrations in RNA splicing that are missed by DNA sequencing. RNA-seq is typically performed on clinically accessible tissues (CATs) from blood and skin. RNA tissue specificity makes it difficult to identify aberrations in relevant but nonaccessible tissues (non-CATs). We determined how RNA-seq from CATs represent splicing in and across genes and non-CATs.

Methods: We quantified RNA splicing in 801 RNA-seq samples from 56 different adult and fetal tissues from Genotype-Tissue Expression Project (GTEx) and ArrayExpress. We identified genes and splicing events in each non-CAT and determined when RNA-seq in each CAT would inadequately represent them. We developed an online resource, MAJIQ-CAT, for exploring our analysis for specific genes and tissues.

Results: In non-CATs, 40.2% of genes have splicing that is inadequately represented by at least one CAT; 6.3% of genes have splicing inadequately represented by all CATs. A majority (52.1%) of inadequately represented genes are lowly expressed in CATs (transcripts per million (TPM) < 1), but 5.8% are inadequately represented despite being well expressed (TPM > 10).

Conclusion: Many splicing events in non-CATs are inadequately evaluated using RNA-seq from CATs. MAJIQ-CAT allows users to explore which accessible tissues, if any, best represent splicing in genes and tissues of interest.

Genetics in Medicine (2020) 22:1181–1190; https://doi.org/10.1038/s41436-020-0780-y

Keywords: clinical genetics; medical genetics; alternative splicing; diagnostic markers; RNA-seq

INTRODUCTION

Exome sequencing is the most advanced standard-of-care genetic test for patients with suspected Mendelian disorders. Yet, the diagnostic rate of exome sequencing is around 31%.[1–4] Genome sequencing, where it has begun being implemented, has been reported to improve upon this diagnostic rate by around 10–15%.[1,5,6] As a result, we are unable to provide a molecular diagnosis for the majority of patients tested with either exome or genome sequencing.

Of many factors hypothesized to contribute to this diagnostic gap, one particularly significant challenge is our inability to adequately interpret the numerous noncoding and synonymous variants these tests produce[7,8]. These variants can cause disease through various well-described mechanisms but are currently challenging to predict. These difficulties have led to most current clinical pipelines largely ignoring these variants.

One such mechanism by which these variants can cause disease is by altering RNA splicing[9,10]. RNA splicing is the process by which different segments of precursor messenger RNA (pre-mRNA) are selectively included or excluded and removed as exons and introns to create a mature mRNA (from which proteins are translated) (Fig. 1a). This process is highly regulated across developmental stages and tissues and is mediated by the spliceosome and numerous RNA-binding proteins (RBPs) that recognize different conserved sequence elements. Variants in these splice factors can thus alter splicing in *trans*, and intronic and exonic variants can alter splicing in *cis* by changing the strength of existing sequence elements or introducing cryptic ones. These different changes in splicing can alter protein function by inserting or removing part of the mRNA transcript (Fig. 1a). Furthermore, they can sometimes cause a frameshift and/or insertion of a premature termination codon, leading to loss of function. Such splicing-altering variants are known to cause Mendelian disorders (e.g., familial dysautonomia, Crouzon syndrome, etc.) and are associated with complex diseases such as Alzheimer disease and cancer[9,11,12].

Clinical RNA-seq is one approach by which laboratories can identify splicing aberrations among other transcriptomic

[1]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA; [2]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA; [3]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA; [4]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA. Correspondence: Yoseph Barash (yosephb@upenn.edu) or Elizabeth J. Bhoj (bhoje@email.chop.edu)

Submitted 12 August 2019; revised 3 March 2020; accepted: 5 March 2020
Published online: 30 March 2020

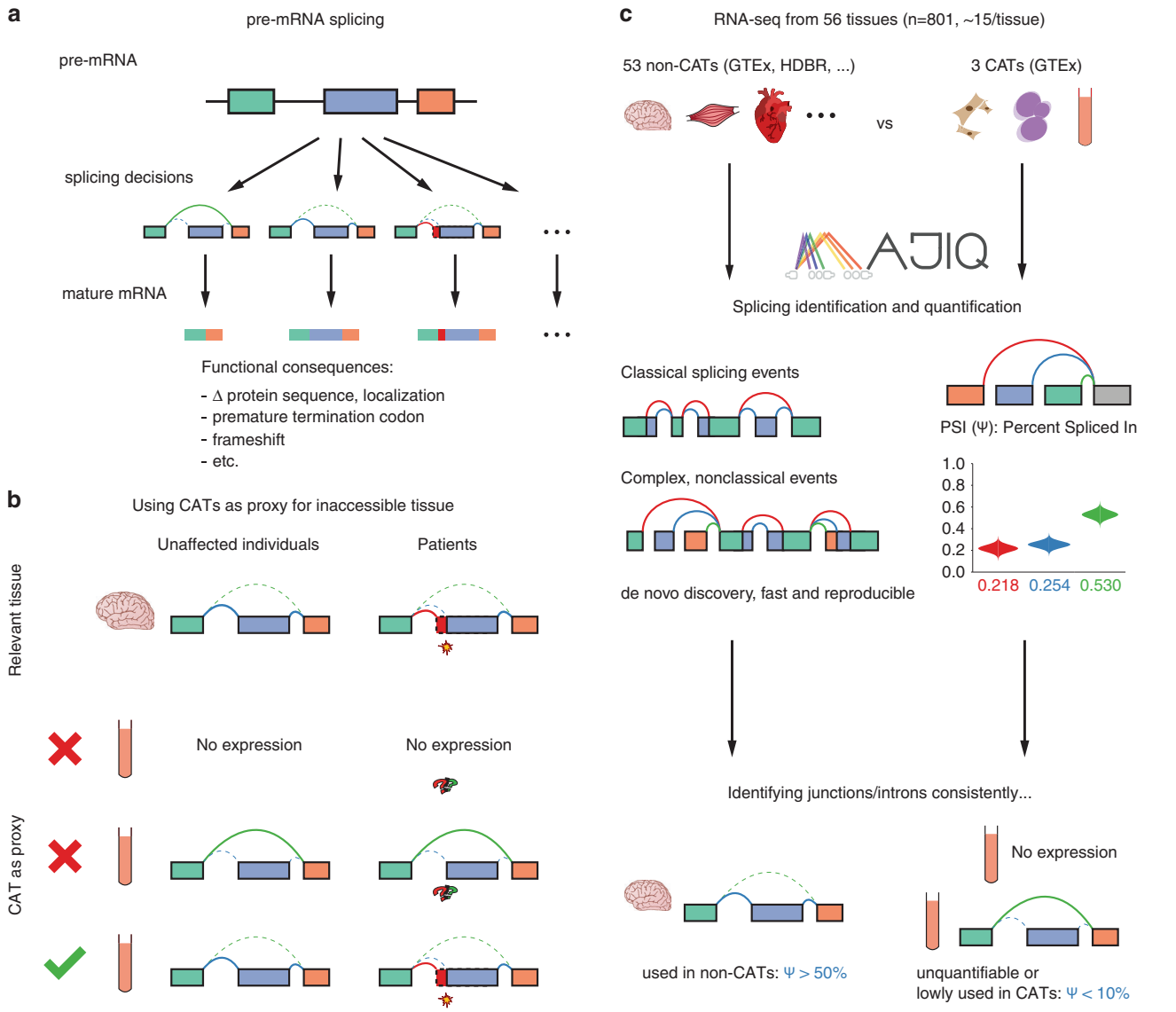

Fig. 1 Identification of splicing events inadequately represented by clinically accessible tissues (CATs). (**a**) Different precursor mRNA (pre-mRNA) splicing decisions can have significant, potentially pathogenic, functional consequences. (**b**) Splicing events in inaccessible tissues (non-CATs) can only be adequately represented by accessible tissues as a proxy if the gene is both expressed and similarly spliced. (**c**) We used MAJIQ on RNA-seq samples from 56 different tissues to define and identify inadequately represented splicing events between inaccessible and accessible tissues.

variations such as gene expression outliers, allele-specific expression, and gene fusions. Indeed, previous work in several labs have demonstrated that RNA-seq can enable genetic diagnosis in patients previously unsolved by exome or genome sequencing[13–18].

As laboratories move to measuring the transcriptome directly with RNA-seq, one challenge they face is tissue specificity. Tissue-specific expression is the most discussed complicating factor of RNA-based analysis, as a gene must be expressed in the tissue to be studied[13,17]. Alternative splicing between tissues is less often addressed, and further complicates analysis. If a tissue other than the tissue of clinical interest is tested, a gene that is expressed in both tissues can still be spliced differently. Thus, splicing defects affecting the tissue of clinical interest might not be realized in the tested tissue despite the gene being expressed in both. Therefore, one tissue can be an adequate proxy for a gene's splicing in a different tissue only if it is both expressed and spliced similarly (Fig. 1b).

Clinicians and researchers can only perform RNA-seq on tissues they have access to. In the clinical setting, these tissues are typically limited to those from blood or skin biopsies: whole blood, Epstein–Barr virus (EBV)-transformed lymphocytes, and fibroblasts. We refer to these three as clinically accessible tissues (CATs). At the same time, laboratories are often interested in pathology occurring in inaccessible tissues (non-CATs, e.g., brain, heart, etc.).

Several recent studies consider limitations of using RNA-seq from CATs for clinical diagnosis. Frésard et al. demonstrate that RNA-seq in whole blood can make some diagnoses in

patients from diverse disease categories[16]. However, Cummings et al. studying a cohort of patients with neuromuscular disease, performed RNA-seq on skeletal muscle biopsies motivated by low gene expression of many known neuromuscular disease genes in whole blood and fibroblasts[13]. Gonorazky et al. further show that they can identify aberrant splicing in muscle that they would not detect from fibroblasts from the same patients[17]. While serving as an important proof of concept of the limitations of RNA-seq with CATs in neuromuscular disease genes, these studies raise the more general question: what are the limitations of RNA-seq in CATs across other non-CATs and genes in general, and how can one quantify them? An answer to this question could inform the selection of the best tissue to send for RNA-seq in clinical practice by evaluating the degree to which each CAT faithfully represents splicing in genes and tissues of interest for different patient phenotypes.

We address the question by considering splicing in nonaccessible versus accessible tissues in terms of splicing events. We model splicing events as local splicing choices either starting or ending at a single exon (the reference exon) in a given gene. These local choices involve splice junctions or retained introns that are included in the gene's transcripts. These splicing events include constitutive splice junctions and local splicing variations (LSVs)[19], which are splicing events where the reference exons can be spliced to multiple RNA segments, thus allowing variation.

We identify and quantify splicing events from RNA-seq data using the recently updated MAJIQ 2.1 toolkit for splicing detection, quantification, and visualization[19]. Previous studies have shown MAJIQ is able to accurately detect and quantify splicing variations, comparing favorably to other tools and offering unique advantages for the specific task we consider here[20]. Briefly, MAJIQ can robustly detect LSVs by combining given transcriptome annotations with de novo (unannotated) splice junctions and retained introns found in the input RNA-seq data. This allows us to assess splicing corresponding not only to known transcripts but for unannotated events (novel junctions, exons, and retained introns) identified in the input samples. For each LSV junction (or retained intron), MAJIQ quantifies its relative inclusion compared with the other LSV junctions, measured as percent splicing inclusion (PSI or $\Psi \in [0, 100]$), in any given sample, allowing us to quantify and compare splicing between samples and tissue types. Previous work has shown that MAJIQ highly correlates with reverse transcription polymerase chain reaction (RT-PCR) validation experiments ($r = 0.97$), the gold standard in the RNA field, and identifies differentially spliced events between RNA-seq samples with high reproducibility ($RR = 78$–86%) (Vaquero-Garcia et al., 2018 [bioRxiv], unpublished data).

When using splicing in CATs as a proxy to splicing in some other tissue of interest, we consider three possible scenarios or splicing event categories (Fig. **1b**): (1) the event is unquantifiable in the CAT due to low gene expression and/or sequencing depth, (2) the event is quantifiable but spliced differently, and (3) the event is quantifiable and not spliced differently. We further focus on splicing events that are consistently included, meaning that they are similarly quantified in nearly all samples for a given tissue type. Naturally, categorizing events into these scenarios depends on the thresholds used to define them. Here, we define consistently spliced events in non-CATs to be events with a junction or retained intron with $\Psi > 50\%$ in more than 85% of samples (Fig. **1c**). We emphasize finding the subset of these events that correspond with either of the first two scenarios where splicing measured in a CAT inadequately represents splicing in the non-CAT. We define these events as those that are unquantifiable or have $\Psi < 10\%$ in more than 85% of a CAT's samples.

In this work, we analyze 53 adult tissues in the Genotype-Tissue Expression Project (GTEx)[21] and 3 fetal tissues from the Human Developmental Biology Resource (HDBR)[22] (cerebellum, cortex) and ArrayExpress accession E-MTAB-7031[23] (heart) (Fig. **1c**). We map all transcriptome variations across these data sets, contrasting splicing between CATs and non-CATs. We make our analyses accessible as an online resource, which we call MAJIQ-CAT (https://tools.biociphers.org/majiq-cat). This online resource has been designed for clinicians and researchers interested in obtaining patient RNA-seq in the context of Mendelian disease. With MAJIQ-CAT, these users can explore how faithfully different CATs represent splicing in their specific genes and tissues of interest, informing their choice of patient tissue to collect. Finally, we discuss implications for RNA-seq in clinical practice and the need for alternative solutions for the genes and tissues that are inadequately represented by CATs.

MATERIALS AND METHODS

Sample selection criteria

We used RNA-seq data for samples from 56 different tissue types: 53 adult tissues and 3 fetal tissues. We obtained samples for all 53 adult tissue types from GTEx (dbGaP accession phs000424). Meanwhile, we obtained samples for fetal cerebral cortex and cerebellum from HDBR (ArrayExpress accession E-MTAB-4840), and we obtained samples for fetal heart from ArrayExpress accession E-MTAB-7031.

All RNA-seq data have been previously described and were derived from tissues collected ethically. For GTEx, tissues from deceased donors are not legally classified as human subjects research under US Code of Federal Regulations Title 45, Part 46 (45 CFR 46) but were collected under written or recorded verbal authorization from next of kin, while tissues collected from living donors were only included after full, written consent was obtained[21]. HDBR is a tissue bank regulated by the UK Human Tissue Authority, and samples in HDBR were collected with appropriate maternal written consent and approval from a National Research Ethics Service (NRES) Committee[22]. The fetal heart samples were acquired with informed written parental consent obtained from all subjects under approval of NHS Lothian, the University of Edinburgh Research Governance Hope, and the University of Leeds Ethical Committee[23].

We restricted sample selection from each of these data sets/tissue types using available metadata as follows. We restricted selection to unique donors per tissue type. This restriction was relevant to both GTEx and HDBR, which include donors contributing multiple sample per tissue type. Available HDBR metadata did not suggest criteria for preferring one sample over another, so we restricted selection to the first available sample per donor. However, GTEx metadata includes information on the number of megabases per sample, so we restricted selection to the sample with the largest size. GTEx metadata also included further information that we used; specifically, we further restricted selection to samples that (1) were hosted by National Center for Biotechnology Information (NCBI), (2) had matched genome sequencing data, (3) had an average spot length of 152 bp, (4) were not flagged by GTEx as a sample to remove (SMTORMVE), and (5) had an RNA integrity number (RIN) score greater than 6.

Given these restrictions, we selected up to 15 samples per tissue type for further analysis. We chose 15 samples as the maximum number of samples per tissue group because preliminary analysis using MAJIQ indicated that reproducibility for tissue-specific differential splicing analysis saturates with around 15 samples in GTEx (data not shown). Consequently, when there were more than 15 samples meeting the above criteria for a particular tissue type, we randomly selected 15 samples among them. For the other tissue types, we kept all samples meeting criteria for further analysis.

Sample read alignment

We aligned RNA-seq reads from the selected samples to the human genome for splicing analysis with MAJIQ using the following procedure. We downloaded selected samples as FASTQ files using SRA Tools (v2.9.6)[24]. We performed quality and adapter trimming on each sample using TrimGalore (v0.4.5)[25]. We used STAR (v2.5.3a)[26] to perform a two-step gapped alignment of the trimmed reads to the GRCh38 primary assembly with annotations from Ensembl release 94[27].

Gene expression quantification

We quantified gene expression in each of the samples. We quantified transcript abundances in transcripts per million (TPM) using Salmon (v0.13.1)[28] quasi-mapping on the trimmed reads for each sample with a transcriptome built from Ensembl release 94 annotations. We aggregated the quantifications to gene expression by taking the sum of abundances for the transcripts associated with each gene.

Splicing identification and quantification using MAJIQ

First, we used MAJIQ (v2.1)[19] with Ensembl release 94 annotations to identify/model the set of all possible annotated and de novo splicing events across our samples. We then quantified these splicing events for each sample, considering an event to be quantifiable for a given sample if it had at least one junction with at least ten supporting reads starting from at least three unique positions. We estimated the percent spliced in (PSI or Ψ) for the junctions and retained introns in each quantifiable splicing event. For the quantifiable LSVs, we used MAJIQ to estimate PSI. Meanwhile, we assigned $\Psi = 100\%$ to the quantifiable constitutive junctions, as they were the only choice for inclusion in their respective events.

Finally, we identified and filtered out ambiguous splicing events per sample. We defined ambiguous splicing events as events containing junctions or retained introns that were in quantified splicing events assigned to more than one gene. These ambiguous assignments occur because of the presence of overlapping genes, especially combined with the unstranded nature of the RNA-seq experiments we used. The resulting nonambiguous quantifications were used to identify relevant consistent and tissue-specific differences between CATs and non-CATs.

Identifying relevant splicing events

To determine the extent to which splicing in non-CATs is inadequately represented by splicing in CATs, we first defined which splicing events for each non-CAT we would consider changes in usage for. For each non-CAT, we consider the set of consistent splicing events, which are splicing events with a junction or retained intron that is highly included in nearly all samples for their tissue type. Specifically, we considered splicing events with a junction or retained intron quantified as $\Psi > 50\%$ in more than 85% of the samples for each non-CAT.

We then evaluated how well splicing quantified in CATs reflected splicing in these consistent splicing events. To do so, we identified the subset of these events for which usage in CATs was consistently low or unquantified. Specifically, we identified which events were either unquantifiable or had $\Psi < 10\%$ for the same junction or retained intron in more than 85% of the samples for each CAT. We call these splicing events inadequately represented in their respective CAT.

Analysis of genes with consistently used splicing events

We then aggregated information about these consistent and inadequately represented splicing events to their respective genes for each tissue. That is, we determined which genes had consistent splicing events in each non-CAT and the subset of these genes for which these events were inadequately represented for each CAT. We evaluated gene expression for the inadequately represented genes to assess how inadequately represented splicing related to low gene expression versus tissue-specific alternative splicing. We also evaluated which of the inadequately represented genes were annotated as disease-causing. We obtained our list of disease-causing genes by combining annotations from ClinVar[29] (gene annotations from the table named "gene_condition_source_id") and the Human Gene Mutation Database (HGMD) 2018.3[30] (inferred from variants classified as DM, the highest level of pathogenicity).

Data access and software

Sequencing data used for this analysis are available in dbGaP under accession phs000424 and ArrayExpress

under accessions E-MTAB-4840 and E-MTAB-7031. Software versions, resources, and specific parameters used are listed in Table S1. The analysis was implemented for reproducible execution as a Snakemake pipeline[31]. Links to source code for the analysis and online resource, MAJIQ-CAT, are listed in Table S2 and have been deposited to Zenodo[32].

RESULTS

Our sample selection procedure yielded a data set with $n = 801$ RNA-seq samples for 53 non-CATs and 3 CATs (Table S3). Seven hundred sixty-two samples came from GTEx, 30 fetal brain samples came from HDBR, and 9 fetal heart samples came from E-MTAB-7031. We selected and processed 15 samples for each tissue except for bladder, cervix (ectocervix and endocervix), fetal heart, and fallopian tube, where we selected all available samples that met our criteria.

Across all samples, we identified a total of 239,406 quantifiable splicing events (124,909 LSVs and 114,497 constitutive junctions) in 25,494 genes. Per sample, we quantified a median of 116,153 splicing events (65,481 LSVs and 50,719 constitutive junctions) in 12,872 genes. We then identified and removed ambiguous splicing events with junctions or retained introns associated with multiple genes, leaving a total of 223,590 splicing events (117,728 LSVs and 105,862 constitutive junctions) in 26,643 genes with a per-sample median of 107,174 splicing events (60,591 LSVs and 46,825 constitutive junctions) in 12,049 genes.

Among quantified LSVs and constitutive junctions, we identified in each non-CAT a median of 73,669 junctions or retained introns in 9966 genes that were consistently used (Fig. **2a**, S1; Table S4). Looking at these same events in CATs, we found that 27.7% were inadequately represented in at least one CAT (3925 or 40.2% of genes) (Table S5); 4.4% were inadequately represented by all CATs (609 or 6.3% of genes).

We compared the quantities of inadequately represented splicing per CAT and non-CAT. The median percentage of genes with consistently spliced events that were inadequately represented across non-CATs was 10.8% for fibroblasts in comparison to 17.5% for EBV-transformed lymphocytes and 34.4% for whole blood (Fig. **2b**). The percentage was lowest in fibroblasts and highest in whole blood for each non-CAT except for spleen, for which the percentage was lowest in EBV-transformed lymphocytes (Fig. S2). Considering all consistently spliced junctions/retained introns across non-CATs, we found that the percentage of inadequately represented splicing was also lowest in fibroblasts and highest in whole blood (Fig. **2c**).

We examined potential strategies for decreasing inadequately represented splicing compared with current practice. To evaluate the benefit of potentially acquiring and sequencing two CATs instead of one, we quantified for each CAT the percentages of inadequately represented events and genes that were not inadequately represented in the other two CATs (Figs. S3–S5). We also considered how primary skin types, which are typically not accepted by clinical laboratories for nondermatologic conditions, would perform as alternative CATs by calculating their percentages of inadequately represented junctions and genes (Fig. S6).

We further investigated the expression and pathogenicity of inadequately represented genes. The maximum median gene expression across inadequately representing CATs was less than 1 TPM in 52.1% of inadequately represented genes (Fig. **3a**). However, an average of 5.8% of inadequately represented genes per non-CAT (217 genes) are expressed with greater than 10 TPM. Meanwhile, a median of 29.2% of inadequately represented genes per non-CAT were annotated as disease-causing in either ClinVar or HGMD (Fig. **3b**).

To facilitate the interrogation of specific genes and splicing variations of interest by clinicians, we developed MAJIQ-CAT (https://tools.biociphers.org/majiq-cat). MAJIQ-CAT is an online resource that provides panels with which users can select genes and non-CATs to look at how well CATs represent splicing across their genes of interest, both globally (Fig. **4a**) and looking at individual splicing events in a specific gene (Fig. **4b**). Genes can be selected from predefined lists of genes (e.g., from ClinVar, ClinGen, etc.) or custom lists provided by the user either interactively or by uploading a text file. Non-CATs can be selected similarly. Changes to these inputs automatically regenerate plots and tables describing the consistent and inadequately represented genes. Individual genes can further be explored by clicking their names to load an additional page that displays their tissue-specific gene expression and splicing events. For example, if a laboratory was interested in studying intellectual disability as a phenotype, they could focus on brain non-CATs and genes associated with the corresponding Human Phenotype Ontology (HPO) term (HP:0001249), finding 1232 genes with consistent splicing in at least one of the brain tissues (Fig. **4a**). They could further look for genes that are expressed but inadequately represented by filtering the table by expression; in this example, setting a minimum of TPM >10 yields a list of 139 genes. Clicking into one of the resulting genes (e.g., *MEF2C*) leads to another page with comparisons of splicing in the CATs and the brain tissues, demonstrating where the inadequately represented splicing events are and the distributions of PSI in each tissue for each event (Fig. **4b**). We developed additional, more detailed example scenarios to demonstrate how to use MAJIQ-CAT in the supplementary information.

DISCUSSION

In this study, we present a comprehensive analysis of RNA splicing events that consistently occur in clinically inaccessible tissues, focusing on how corresponding events take place in clinically accessible tissues. While clinicians and scientists are often interested in what takes place in the inaccessible tissues as part of disease pathology, laboratories can only measure the accessible tissues as a proxy. Thus, these results inform clinicians and scientists as to where RNA-seq is limited, especially with respect to previously underappreciated tissue-specific splicing, and suggest when specific clinically

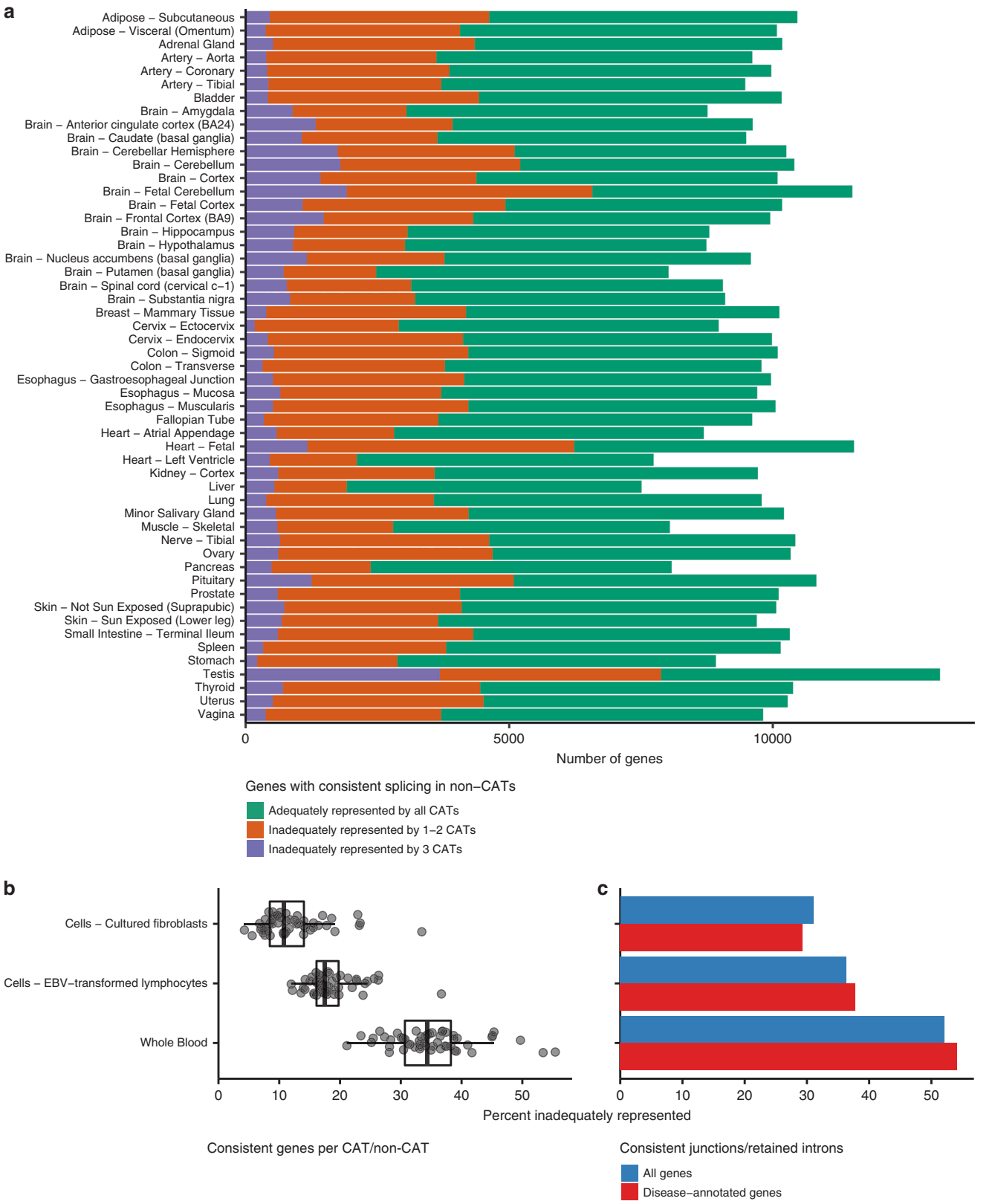

Fig. 2 Mapping transcriptome variations identified in clinically accessible tissues (CATs) vs. non-CATs. (**a**) Of an average of 9966 genes with consistently spliced events per non-CAT, 3925 (40.2%) were inadequately represented in at least one CAT, with 609 (6.3%) being inadequately represented by all CATs. (**b**) The percentages of genes with consistently spliced events that were inadequately represented over the 53 non-CATs were lowest in fibroblasts and highest in whole blood. (**c**) The percentage of junctions/retained introns that were consistently used in at least one non-CAT that were inadequately represented by each CAT was lowest in fibroblasts and highest in whole blood. *EBV* Epstein–Barr virus.

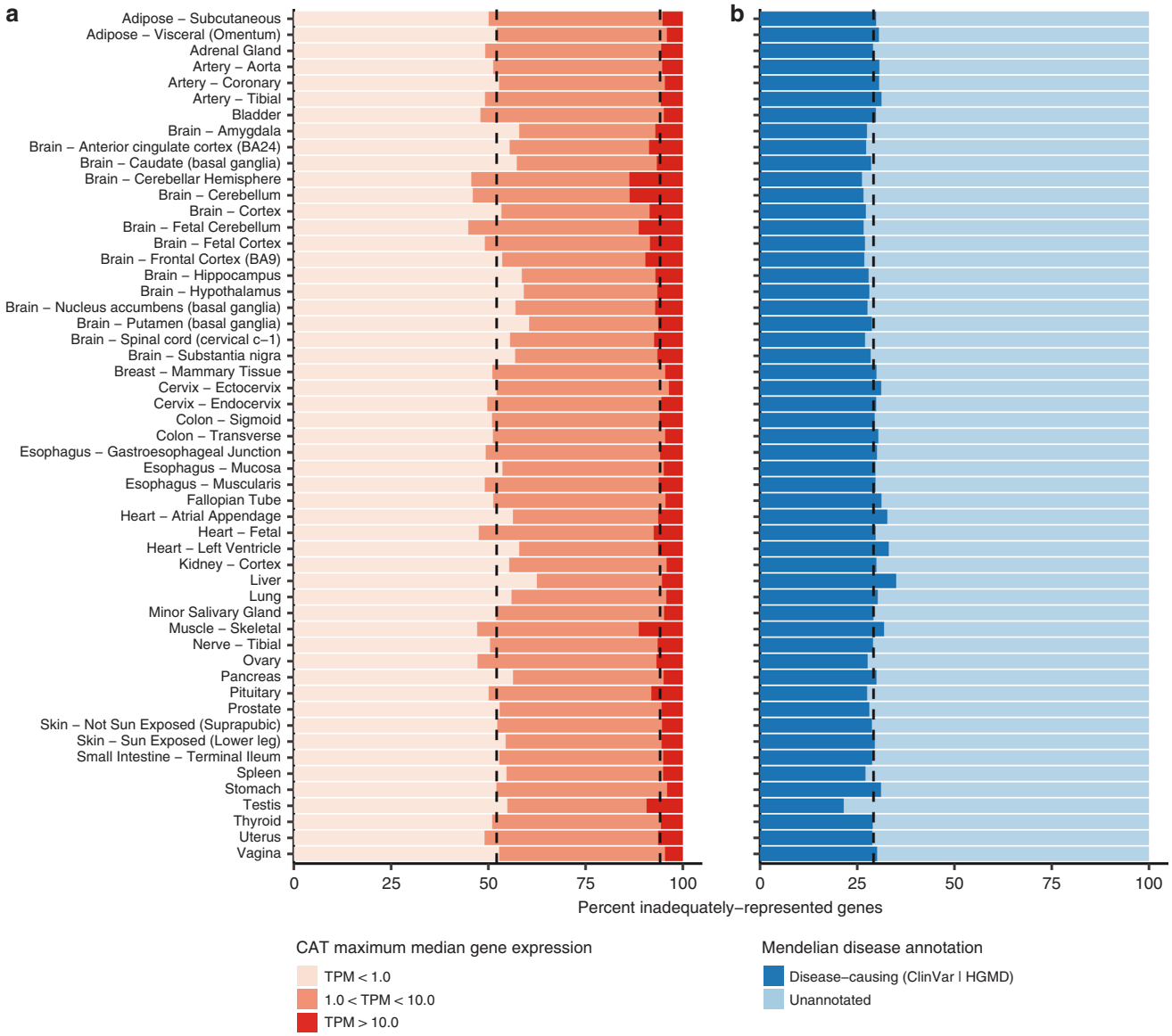

Fig. 3 Expression and pathogenicity of inadequately represented genes. (**a**) The majority of inadequately represented genes are lowly expressed (TPM < 1) in clinically accessible tissues (CATs), but an average of 217 genes (5.8%) are well expressed (TPM > 10) in at least one inadequately representing CAT. (**b**) An average of 29.2% of inadequately represented genes are annotated as disease-causing. *HGMD* Human Gene Mutation Database.

accessible tissues should be preferred over others or when alternative approaches to clinical RNA-seq are needed. By making the results interactively accessible through MAJIQ-CAT, we enable clinicians and scientists to more directly explore how these limitations impact specific genes and tissues of interest to them.

Previous studies analyzing patient RNA-seq from CATs demonstrated that these data could be used to identify rare disease genes and variants from a variety of disease categories[14,16]. These studies demonstrate that although clinicians are typically limited to using CATs for patient RNA-seq, RNA-seq in those tissues can still improve the molecular diagnostic rate for suspected Mendelian disorders by identifying changes that were not identified using exome or genome sequencing alone. Our study provides an orthogonal, but related, result. Because we are typically limited to using CATs for patient RNA-seq, there are splicing events found in disease-relevant tissues and genes that will consistently be a blind spot in such studies.

Our study found that 40.2% of genes with consistent splicing events per non-CAT are inadequately represented by at least one CAT. This implies that clinicians and scientists interested in how one of the inadequately represented genes are spliced in the non-CAT in patients need to be careful about which clinically accessible tissues they measure as a proxy because at least one of the accessible tissues will not represent the splicing events well. We show that many of these genes are considered disease-causing (29.2% of the

Fig. 4 MAJIQ-CAT enables clinicians and scientists to explore inadequate representation of splicing by clinically accessible tissues (CATs) in specific genes and tissues of interest. (**a**) MAJIQ-CAT allows users to choose from predefined or custom gene sets and tissues (left) to quantify and understand the user-specific relevant limitations of RNA-seq in different accessible tissues (right). (**b**) Users can further explore individual genes for tissue-specific differences in gene expression and splicing. Shown here is a closer look at the gene *MEF2C*, with a violin plot of its expression in CATs and selected non-CATs (left) and violin plots of percent splicing inclusion (PSI) for one of its inadequately represented splicing events (right). See main text for more details.

inadequately represented genes); thus, understanding these limitations is increasingly clinically relevant as RNA-seq enters clinical practice.

Considering these 40.2% of genes with inadequately represented splicing, the majority (52.1%) were associated with low gene expression (TPM < 1), as expected. However, we still find that 217 genes per non-CAT are highly expressed (TPM > 10) but spliced differently in CATs. The limitations of these genes for clinical RNA-seq would be missed by previous expression-first analyses, highlighting the novelty and impact of our splicing-first analysis.

For the other 59.8% of genes, we note that splicing in CATs may still not always adequately represent splicing in non-CATs. While they may not pass the stringent thresholds we set to define inadequately represented splicing present in most samples for a CAT ($\Psi < 10\%$ or unquantifiable in more than 85% of samples), splicing inclusion may take intermediate values or be highly variable between samples. Furthermore, even for splicing variations that are similar between tissues, they may still involve different tissue-specific regulation by different tissue-specific factors. Thus, while we might expect variants in tissue-independent splicing sequence elements

(e.g., canonical splice sites) to impact the different tissues similarly, variants in tissue-specific splicing enhancers or silencers could lead to tissue-specific defects that would not be represented by CATs.

It is also important to note that the results described here are dependent on the limitations and technical biases of current practices and technologies for poly-A selected RNA-seq. For example, sequencing with greater depths or read lengths than is typically done in common practice could potentially increase detection of lower-expressed genes/splicing events. Likewise, alternative and/or future approaches for mRNA isolation or sequencing will differentially impact detection of splicing across the genome. In particular, protocols including globin-depletion of whole blood would likely improve its performance as a CAT because globin genes account for the majority of expressed transcripts in GTEx whole blood. Since these data are not available in GTEx, we plan to evaluate globin-depleted whole blood as a CAT for a future update to MAJIQ-CAT. It will be important to re-evaluate differences in what we can detect between tissues as emerging technologies, such as long-read sequencing, enter common practice and replace current protocols for measuring clinical transcriptomes.

One important conclusion from the analysis performed here is that for the 3316 genes per non-CAT that are inadequately represented by one or two CATs, at least one CAT offers a better representation of the gene's splicing than the others. Thus, our study implies that researchers interested in one of these genes and tissues should have a preference for which clinically accessible tissue to collect. Summarizing across all genes with consistent splicing, we found that fibroblasts almost always had the lowest percentage of inadequately represented genes. Thus, our results suggest that researchers interested in all genes and tissues equally should prefer collecting patient fibroblasts if possible. However, clinicians and scientists are often interested in specific genes or tissues relevant to a specific biological process. Our online resource, MAJIQ-CAT, will enable clinicians, scientists, and laboratories to interactively explore which CATs are most relevant for representing the biology they care about and which genes and splicing events are most affected.

Another important conclusion of this study is that there are 609 genes per non-CAT that are inadequately represented by all CATs. For these genes, using RNA-seq in any CAT as a proxy would have many limitations for studying splicing. In these cases, alternative approaches are likely necessary. One possible path forward is the use of in vitro differentiation/transdifferentiation of CRISPR-iPSCs/patient-derived cells toward tissue types of interest. Gonorazky et al. illustrated this possibility for transdifferentiated myotubes from patient fibroblasts as an alternative to skeletal muscle biopsy, although how these results would translate to other, more inaccessible, tissues remains to be explored[17]. Another possible path forward is the use of in silico models of splicing[33–39]. Previous works in several labs have developed models of tissue-specific splicing but do not directly train models on genetic variants[33–37]. Recent work by Cheng et al. directly trains models to predict splicing changes using genetic variants but does not account for tissue specificity[39]. Future developments combining aspects of these models to produce predictions of tissue-specific splicing as a consequence of genetic variants could help us understand potential splicing defects in those genes where we do not have a good proxy. These alternative strategies could be combined with other orthogonal approaches, including predicted variant pathogenicity, to further advance detection of splicing variants in these inadequately represented genes.

In summary, in this study, we demonstrated and quantified the limitations of CATs to serve as a proxy for non-CATs for RNA splicing measured by RNA-seq. We highlighted how alternative splicing contributes to these limitations in addition to tissue-specific gene expression. In addition, we developed and have made available an online resource, MAJIQ-CAT, that will allow clinicians and scientists to directly explore how these limitations affect specific genes and tissues of interest. MAJIQ-CAT will be of particular use for determining tissues to study for genes that are only inadequately represented in some but not all CATs. For the genes inadequately represented by all CATs, future work on alternative approaches to estimate splicing defects in patients will be necessary to improve clinical diagnoses.

SUPPLEMENTARY INFORMATION

The online version of this article (https://doi.org/10.1038/s41436-020-0780-y) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM128096. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. J.K.A. acknowledges salary support by NIH/Eunice Kennedy Shriver National Institute of Child Health (NICHD) fellowship F30HD098803.

DISCLOSURE

The authors declare no conflicts of interest.

REFERENCES

1. Clark MM, Stark Z, Farnaes L, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ Genom Med. 2018;3:16.
2. Yang Y, Muzny DM, Xia F, et al. Molecular findings among patients referred for clinical whole-exome sequencing. JAMA. 2014;312:1870–1879.
3. Farwell KD, Shahmirzadi L, El-Khechen D, et al. Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based

analysis: results from 500 unselected families with undiagnosed genetic conditions. Genet Med. 2015;17:578–586.
4. Retterer K, Juusola J, Cho MT, et al. Clinical application of whole-exome sequencing across clinical indications. Genet Med. 2016;18:696–704.
5. Alfares A, Aloraini T, Subaie LA, et al. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. Genet Med. 2018;20:1328.
6. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet. 2015;47:717–726.
7. Frésard L, Montgomery SB. Diagnosing rare diseases after the exome. Mol Case Stud. 2018;4:a003392.
8. Gloss BS, Dinger ME. Realizing the significance of noncoding functionality in clinical genomics. Exp Mol Med. 2018;50:97.
9. Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet. 2016;17:19–32.
10. Wang G-S, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet. 2007;8:749–761.
11. Fenwick AL, Goos JA, Rankin J, et al. Apparently synonymous substitutions in FGFR2 affect splicing and result in mild Crouzon syndrome. BMC Med Genet. 2014;15:95.
12. Raj T, Li YI, Wong G, et al. Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. Nat Genet. 2018;50:1584–1592.
13. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9:eaal5209.
14. Kremer LS, Bader DM, Mertes C, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat Commun. 2017;8:15824.
15. Hamanaka K, Miyatake S, Koshimizu E, et al. RNA sequencing solved the most common but unrecognized NEB pathogenic variant in Japanese nemaline myopathy. Genet Med. 2019;21:1629.
16. Frésard L, Smail C, Ferraro NM, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat Med. 2019;25:911–919.
17. Gonorazky HD, Naumenko S, Ramani AK, et al. Expanding the boundaries of RNA sequencing as a diagnostic tool for rare mendelian disease. Am J Hum Genet. 2019;104:466–483.
18. Gonorazky H, Liang M, Cummings B, et al. RNAseq analysis for the diagnosis of muscular dystrophy. Ann Clin Transl Neurol. 2016;3:55–60.
19. Vaquero-Garcia J, Barrera A, Gazzara MR, et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. eLife. 2016;5:e11752.
20. Norton SS, Vaquero-Garcia J, Lahens NF, Grant GR, Barash Y. Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates. Bioinformatics. 2018;34:1488–1497.
21. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–585.
22. Lindsay SJ, Xu Y, Lisgo SN, et al. HDBR expression: a unique resource for global and individual gene expression studies during early human brain development. Front Neuroanat. 2016;10:86
23. Pervolaraki E, Dachtler J, Anderson RA, Holden AV. The developmental transcriptome of the human heart. Sci Rep. 2018;8:15362.
24. NCBI. SRA-Tools. http://ncbi.github.io/sra-tools/. Accessed 12 Dec 2018.
25. Babraham Bioinformatics. Trim Galore! http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Accessed 12 Dec 2018.
26. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
27. Cunningham F, Achuthan P, Akanni W, et al. Ensembl 2019. Nucleic Acids Res. 2019;47(Database issue):D745–D751.
28. Love MI, Hogenesch JB, Irizarry RA. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. Nat Biotechnol. 2016;34:1287–1291.
29. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980–D985.
30. Stenson PD, Ball EV, Mort M, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003;21:577–581.
31. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2522.
32. Aicher JK, Jewell P, Vaquero-Garcia J, Barash Y, Bhoj EJ. Biociphers/Aicher2019-CAT-Splicing-Analysis: analysis for "Mapping RNA splicing variations in clinically-accessible and non-accessible tissues to facilitate Mendelian disease diagnosis using RNA-Seq." 2020. https://doi.org/10.5281/zenodo.3611492.
33. Barash Y, Calarco JA, Gao W, et al. Deciphering the splicing code. Nature. 2010;465:53–59. https://doi.org/10.1038/nature09000.
34. Barash Y, Blencowe BJ, Frey BJ. Model-based detection of alternative splicing signals. Bioinformatics. 2010;26:i325–i333.
35. Xiong HY, Alipanahi B, Lee LJ, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347:1254806.
36. Jha A, Gazzara MR, Barash Y. Integrative deep models for alternative splicing. Bioinformatics. 2017;33:i274–i282.
37. Zhang Z, Pan Z, Ying Y, et al. Deep-learning augmented RNA-seq analysis of transcript splicing. Nat Methods. 2019;16:307–310.
38. Jaganathan K, Panagiotopoulou SK, McRae JF, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176:535–548.
39. Cheng J, Nguyen TYD, Cygan KJ, et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome Biol. 2019;20:48.