GENETICS

Widespread occurrence of hybrid internal-terminal exons in human transcriptomes

Ana Fiszbein[1,2]*, Michael McGurk[1], Ezequiel Calvo-Roitberg[3], GyeungYun Kim[2], Christopher B. Burge[1]*, Athma A. Pai[3]*

Messenger RNA isoform differences are predominantly driven by alternative first, internal, and last exons. Despite the importance of classifying exons to understand isoform structure, few tools examine isoform-specific exon usage. We recently observed that alternative transcription start sites often arise near internal exons, often creating "hybrid" first/internal exons. To systematically detect hybrid exons, we built the hybrid-internal-terminal (HIT) pipeline to classify exons depending on their isoform-specific usage. On the basis of splice junction reads in RNA sequencing data and probabilistic modeling, the HIT index identified thousands of previously misclassified hybrid first-internal and internal-last exons. Hybrid exons are enriched in long genes and genes involved in RNA splicing and have longer flanking introns and strong splice sites. Their usage varies considerably across human tissues. By developing the first method to classify exons according to isoform contexts, our findings document the occurrence of hybrid exons, a common quirk of the human transcriptome.

INTRODUCTION

The composition of mRNAs in a eukaryotic cell is highly dynamic and diverse. More than 95% of multiexonic mammalian genes express multiple mRNA isoforms and isoform usage is often regulated in a tissue- and context-specific manner. Despite the increased focus on alternative splicing mechanisms, it was recently observed that isoform differences across human tissues are predominantly driven by alternative mRNA start and end sites (*1*–*3*). Alternative splicing of internal exons was estimated to explain tissue-dependent transcript differences for only ~35% of the genes, while alternative transcription start and end sites accounted for the majority of tissue-dependent isoform usage (*1*). The modulation of transcripts through usage of alternative mRNA start and end sites is highly conserved during evolution, with alternative terminal exon usage occurring for most genes across fungal, plant, insect, and mammalian species (*4*–*6*). Furthermore, accumulating evidence suggests that a substantial fraction of internal exons are used as first exons (FEs) or last exons (LEs) in different cell contexts (*7*, *8*). The usage of polyadenylation sites in introns can lead to the conversion of an internal exon into a 3′ terminal exon (*7*), and cryptic promoters that arise during evolution may lead to conversion of internal exons to FEs (*8*). However, much less is known about the prevalence, usage, and evolution of this class of hybrid exons.

The usage of alternative transcription start and end sites can contribute to the regulation of the transcriptome and proteome in many ways. Few alternative terminal exons have large coding capacity. Instead, alternative first exons (AFEs) or alternative last exons (ALEs) are more likely to influence the open reading frame through introduction of alternative start codons or induce protein truncation with premature stop codons. However, the vast majority of alternative terminal exons contribute to alternative 5′ or 3′ untranslated region (UTR) composition and consequent regulation of posttranscriptional mRNA processes (*9*–*11*). Although UTRs do not directly contribute to protein sequence, they harbor crucial regulatory sequences [including those for RNA binding proteins and microRNAs (*12*)] and are thus modulators of mRNA subcellular localization (*13*), stability (*14*), and translation (*15*) in vivo. As expected, given the many ways that alternative terminal exons can influence mRNA and protein diversity, the dysregulation of transcription start and end sites has been associated with several human diseases, including cancers, and neurological disorders (*16*, *17*). Furthermore, the usage of alternative terminal ends is dynamically regulated after exposure to environmental or immune stimuli (*18*, *19*), suggesting that these mechanisms may also play an important role in cellular responses and remodeling. Given these roles of terminal exon composition in the regulation of both the transcriptome and the proteome, it is important to be able to properly classify which exons are used as terminal or internal exons across cell types and cellular contexts.

Many experimental strategies have been developed to specifically identify transcription starts and ends by anchoring on the 5′ 6-methyl-guanosine cap [e.g., Cap Analysis of Gene Expression (CAGE)] or the 3′ polyA-tail (e.g., 3p-seq). However, these experiments do not simultaneously quantify expression of mRNA molecules or provide information about the isoforms associated with the terminal site usage. Furthermore, these specialized protocols are laborious, making them less practical than RNA sequencing (RNA-seq) for general use across cell types and cellular conditions. Given the widespread use and availability of RNA-seq datasets, there is significant interest in using RNA-seq data to identify and quantify the usage of alternative terminal events (*20*–*23*). However, identifying terminal exons from short-read transcriptomic alignments and quantifying their usage remain challenging. Of specific interest is how to uniquely classify exons as first, internal, and last, which we refer to as the "exon-type problem." These issues are especially problematic for hybrid exons, which would be classified as internal in RNA-seq data and terminal in data from the specialized terminal end sequencing techniques highlighted above, despite being used as both terminal and internal exons in different transcripts within a cell type. Even the newest classification and quantification algorithms (*23*–*25*) are not able to differentiate between purely terminal exons and exons with hybrid

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Biology, Boston University, Boston, MA, USA. [3]RNA Therapeutics Institute, University of Massachusetts Medical School, Worcester, MA, USA.
*Corresponding author. Email: anafisz@bu.edu (A.F.); cburge@mit.edu (C.B.B.); athma.pai@umassmed.edu (A.A.P.)

features or quantify the proportional usage as a terminal or internal exon.

Here, we introduce the hybrid-internal-terminal (HIT) index, a straightforward approach to identify and classify exons as hybrid, internal, or terminal on the basis of short-read RNA-seq data. By modeling the ratio of exon-exon splice junction reads (SJRs), the HIT index pipeline reliably classifies exons and identifies thousands of hybrid first-internal or hybrid internal-last exons. By comparing HIT index exon classifications with classifications from specialized techniques to identify transcript ends, we show that our approach accurately classifies exons on the basis of their isoform-specific usage. After applying the HIT index to RNA-seq data across human tissues, we are able to characterize features that differentiate hybrid terminal exons. Overall, the widespread identification of hybrid exons expands the repertoire of tissue-specific isoforms and highlights the diversity of the human transcriptome.

RESULTS

The regulation of mature RNAs can occur at multiple levels within cells, including quantitatively by varying gene expression levels and qualitatively by varying isoforms through alternative transcription start site (TSS), splice site, and/or cleavage site usage. Advances in high-throughput RNA-seq have provided unprecedented insights into mRNA levels across cell types, the internal structure of isoforms, and previously unknown alternatively spliced exons. However, it is still challenging to use short-read RNA-seq data to also identify previously unidentified terminal ends of mRNAs—specifically, 5′ TSSs and 3′ transcription end or polyadenylation sites—or quantify their alternative usage. This is particularly difficult for hybrid terminal exons, which can be used as either terminal or internal exons in the same cell type. The ability to measure terminal site usage simultaneously with mRNA levels and isoform usage would enable a complete picture of the dynamics of both alternative UTRs and internal coding exons. To overcome these challenges, we set out to develop an approach that robustly identifies and quantifies hybrid exons, internal exons, and exons used only as terminal exons.

To do so, we decided to leverage informative SJRs in short-read polyA-selected RNA-seq data to classify exon positioning and usage within an isoform. SJRs in RNA-seq data inform about the connectivity between two exons that have been spliced together. Internal exons should approximately have the same number of SJRs connected to both upstream and downstream flanking exons. At the other extreme, terminal exons should have SJRs connected only to gene-proximal flanking exons (downstream flanking exons for FEs or upstream flanking exons for LEs). Thus, we reasoned that hybrid exons should have an intermediate distribution between FE and internal or between internal and LE, with both upstream and downstream SJRs present, but with a skew toward more gene-proximal SJRs depending on the terminal positioning of the hybrid exon. Building on these assumptions, we calculated the bias in upstream and downstream SJRs for each annotated exon (Fig. 1A). This involves three primary steps: (i) annotating meta-exons from a transcript annotation, (ii) extracting overlapping upstream and downstream SJRs for each exon, and (iii) calculating the SJR ratio to classify exons and estimate percent spliced in (PSI) values for alternative terminal exons. Because exact exon boundaries are confounded by imprecision in (or alternative) TSSs, 3′ and 5′ splice sites, and/or transcription end sites (TES), we use nonredundant meta-exons (generated by collapsing all overlapping annotated exons) with an added buffer region. Statistical significance of the SJR ratio is estimated using parametric bootstrapping to calculate confidence intervals and two *P* values (Materials and Methods). With this formalization, exons used solely as FEs or LEs have an SJR ratio of −1 or 1, because they only have downstream or upstream SJRs, respectively. Internal exons, which should have roughly equal numbers of upstream and downstream SJRs, have an SJR ratio near 0. Notably, this approach allows us to classify hybrid exons, which exhibit |SJR ratio| values between 0 and 1 commensurate with the fraction of transcripts that include a hybrid exon as a terminal or internal exon.

The HIT index accurately classifies simulated exon data

To assess the performance of our approach, we simulated mRNAs with a range of gene structures (varying numbers, usage, and lengths of terminal, hybrid, alternatively spliced, and constitutive exons) and gene expression levels and simulated short reads using a range of library preparation parameters, including different fragment sizes and read lengths (fig. S1 and Materials and Methods). We then calculated the SJR ratio for each simulated exon using exon-exon junction reads as described above and classified exons as terminal exons that serve solely as FEs or LEs in transcripts, internal exons, and hybrid exons that can be used as either terminal or internal exons in different transcripts. Classifications were performed as described in Materials and Methods, using different SJR ratio thresholds when specified. Overall, classifications based on the SJR ratios recapitulate expected exon classifications at loose to moderate SJR ratio thresholds that represent less than a 50% imbalance (|SJR ratio| < 0.5, Fig. 1B). Conversely, the classifications capture a smaller portion of hybrid exons at stringent thresholds of greater than 50% imbalance (|SJR ratio| > 0.5), but with higher confidence that each exon is truly hybrid. Because terminal-only exons lack gene-distal junction reads and thus have extreme SJR ratios of 1 or − 1 (for FEs or LEs, respectively), all terminal exons with reasonable read coverage can be reliably classified regardless of the chosen threshold. However, there is a trade-off between the identification of hybrid exons and internal exons, where more hybrid exons are correctly identified at less stringent thresholds at the expense of some misidentification of internal exons.

We next sought to use these simulation data to identify systematic biases and characteristics of the data in situations where the SJR ratio is unable to correctly classify hybrid or internal exons. The sensitivity for identifying terminal exons increased with more stringent |SJR ratio| thresholds, coupled to slightly reduced specificity in terminal exon classification (Fig. 1C). Given that >95% of terminal-only exons (Fig. 1B) are correctly classified, the variance in predictive values is entirely driven by hybrid exon classification. Both the false-positive and false-negative rates of terminal exon classification are less than 10% when using |SJR ratio| ≥ 0.3. However, the increased ability to correctly classify terminal exons across |SJR ratio| thresholds occur at the expense of the sensitivity, but not specificity, of internal exon classification. In both cases, overall specificity and sensitivity scale with the read depth of libraries (Fig. 1C) and the minimum number of SJRs required (fig. S2A). On the basis of these trends, we further probed the parameters that influenced internal exon and hybrid exon classification across the chosen thresholds.

Exons that are solely used as internal exons should have a fairly even distribution of upstream and downstream junction reads;

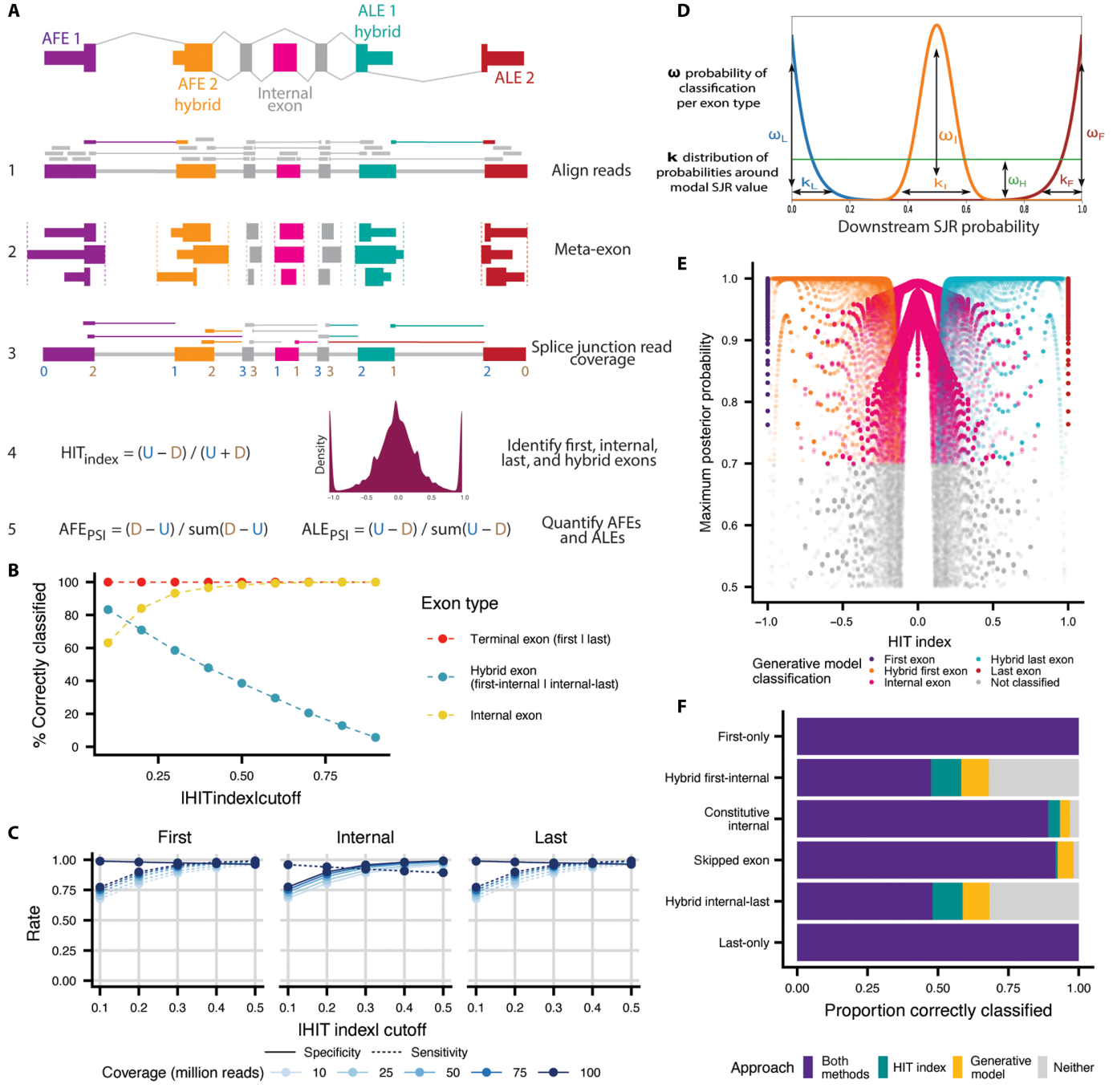

Fig. 1. HIT index performs well with simulated data. (**A**) Schematic of the steps involved in the HIT index pipeline, including (1) alignment of reads to (2) meta-exons created by collapsing overlapping annotated exon boundaries, (3) extracting SJRs, and calculating (4) SJR ratio and (5) PSI values for each exon. (**B**) Percent of simulated exons correctly classified (*y* axis) within each exon type across variable SJR ratio thresholds (*x* axis). (**C**) True-positive (sensitivity, dotted lines) and true-negative (specificity, solid lines) rates (*y* axis) for simulated exon classification across moderate SJR ratio thresholds (*x* axis), for different read depths. (**D**) Schematic of the mixture of beta-binominal distributions of downstream SJR counts (*x* axis) that represent the generative model probabilities (*y* axis) that an exon belongs to a particular exon type, where the modes of the distributions are 0, 0.5, and 1 for last-only, internal-only, and first-only exons, respectively, and hybrid exons are those that fall outside these distributions. (**E**) The relationship between the SJR ratio (*x* axis) and the maximum posterior probability from the generative model (*y* axis) for various simulated exon types. (**F**) The proportion of different exon types that are correctly classified by only the SJR ratio, the generative model, both models, or neither.

however, we see that as many as 40% of internal exons are misclassified at less restrictive |SJR ratio| thresholds, with as much as a 30% imbalance of junction reads. We reasoned that misclassification of internal exons might be associated with their position in the gene and therefore evaluated the misclassification of exons stratified by their order in the gene (first, second, or third internal exon), the length of the exon, and the lengths of the upstream or downstream flanking exons (fig. S2, B and C). Internal exons are more likely to be

misclassified as either FEs or LEs when the length of upstream or downstream exon, respectively, is shorter. Furthermore, this misclassification is more extreme when the library has larger insert sizes (fig. S2B) or is sequenced with shorter read lengths (fig. S2C). These trends suggest that our ability to classify internal exons that are near terminal exons is affected by RNA-seq edge effects, whereby there is a depletion of reads at the edges of molecules caused by size selection to achieve a narrow distribution of fragment lengths before sequencing. We do not see these same trends for skipped exons (SEs), which were always flanked by constitutive exons in the simulations. Instead, the ability to classify an SE as an internal exon is dependent on the PSI of the SE, where low SE inclusion results in a moderately higher frequency of misclassification (<10%) due to decreased junction read coverage (fig. S2D). To account for constitutive internal exon misclassification, we incorporated an edge-effect flag, which uses piecewise regression to identify exons that are likely to be affected by the edge effect (fig. S1C, Materials and Methods) and allows the end user to decide whether to include or exclude those exons from downstream analyses.

Because the SJR ratio represents the imbalance between upstream and downstream junction reads, the choice of SJR ratio threshold has a large impact on the ability to identify hybrid exons. To assess this impact, we stratified the percent of hybrid exons correctly classified by SJR ratio thresholds and the percent usage of the hybrid exon as an FE, measured by the PSI value (fig. S3A). Regardless of SJR ratio threshold, lower usage of the hybrid exon as an FE rather than an internal exon reduces the ability to classify the exon as a hybrid exon, because junction reads are less skewed. At more stringent thresholds (|SJR ratio| > 0.5), the detection of hybrid exons with an FE PSI < 0.5 drops below 25%. When we evaluate the classification of hybrid exons at a moderate SJR ratio threshold of 0.3, this relationship becomes even more evident, where the assignment to hybrid decreases as PSI decreases, and these exons are classified as solely internal exons.

To alleviate these issues, we developed a probabilistic framework to classify exons by directly modeling the probability of any exon being a hybrid, internal, or terminal exon considering read coverage and the ratio of downstream junction reads to total junction reads (Materials and Methods and Fig. 1D). Specifically, this generative approach relies on jointly modeling the ratio of downstream junction reads to all junction reads for first, internal, and last exons, assuming that these exons are derived from probability distributions with a mode of 1, 0.5, and 0, respectively. Thus, the downstream junction read probabilities can be modeled as arising from a mixture of these three distributions and exons can be assigned a posterior probability of belonging to each distribution. Hybrid exons are detected as outliers relative to each of the nonhybrid distributions, enabling the estimation of a posterior probability of being either a first-internal or internal-last hybrid exon. This method accounts for the intrinsic noisiness in sequencing data and the overdispersion of read counts relative to simple binomial or Poisson count distributions. By moving away from a strict SJR ratio threshold and allowing for flexibility in the classification probability of an exon, this framework enables exons to be correctly classified across the distribution of SJR ratio (Fig. 1E). Most exons, including all terminal exons, are correctly classified by both the SJR ratio and generative model (Fig. 1F). However, combining the two approaches improves the classification of hybrid and internal exons, where the SJR ratio has more power for exons with lower read coverage, while the generative model has more power to classify hybrid exons with moderate PSI values (fig. S3B). Together, the combined approach correctly classifies 68% of simulated hybrid exons and 97% of internal exons. Moving forward, we use a joint classification scheme, called the HIT index pipeline, that integrates the strengths of the SJR ratio and generative model probabilities (Materials and Methods) and henceforth refer to the SJR ratio as the HIT index.

Identification of hybrid exons in human transcriptomes

We applied the HIT index pipeline to a high-coverage human RNA-seq data from LCL (lymphoblastoid cell line) cells to evaluate the performance of our new method in classifying bona fide terminal, internal, and hybrid exons. From a total of 353,677 annotated meta-exons, we discarded 50% that are not expressed and 0.3% that are located in single-exon genes (Fig. 2A). Using an absolute HIT index cutoff of 0.3, we classified ~60,000 internal, ~16,000 terminal, and ~30,000 hybrid exons. (Fig. 2B). We divided hybrid exons into hybrid FE-internal and hybrid internal-LE, each corresponding to ~13% of total expressed exons (Fig. 2C). Consistent with the simulated data, meta-exons classified as hybrid using the SJR ratio alone have a significantly higher probability of being classified as hybrid by the generative model (Fig. 2, D and E). We note that when a retained intron occurs adjacent to an internal exon, SJR ratios might be skewed and lead to a false hybrid classification. However, in our analyses, we did not observe enrichment of retained introns around hybrid exons (fig. S4D). Overall, most exon classifications are statistically robust, with a low probability that the observed HIT index arose at random from an internal exon with the same SJR count (Materials and Methods and Fig. 2F). However, as expected, the HIT index alone has lower power to identify hybrid exons relative to exons solely used as internal or terminal (fig. S4, A and B). To study the properties of hybrid exons, in the following, we consider only high-confidence hybrid exons, which we define as being classified as a hybrid using the HIT index and a first-internal or internal-last posterior probability greater than 0.5 when using the generative model (Fig. 2G). With these conservative criteria, we can reliably classify 3.8% of exons as hybrid in this dataset (fig. S4C).

Reference techniques validate HIT index classifications

To evaluate the accuracy of the HIT index classifications from RNA-seq data with orthogonal datasets, we used CAGE (*26*) and 3′ end sequencing (*27*) datasets for direct assessments of TSS and TES usage, respectively. All HIT index–classified FE and hybrid FE-internal exons overlapped with CAGE-identified TSSs, while all LE and hybrid internal-LE exons overlapped with 3′ sequencing (3′-seq)–identified TESs (Fig. 3, A and B). Reassuringly, terminal and hybrid exons classified as first or last did not significantly overlap the TESs or TSSs, respectively, supporting high specificity of HIT index classifications. Moreover, neither gene expression thresholds nor |HIT index| cutoffs change TSS and TES specificity significantly (fig. S5). Overall, these orthogonal reference techniques validated HIT index classifications, supporting near-perfect specificity in determining true positive TSS and TESs with this approach. The HIT index also reliably identified hybrid exons, significantly expanding the landscape of exons that can act as FE and LE.

We then compared HIT index classifications to annotated exon classifications (see Materials and Methods). While most HIT index terminal and internal exons were annotated as the same, the majority of hybrid exons were annotated solely as internal exons in genome

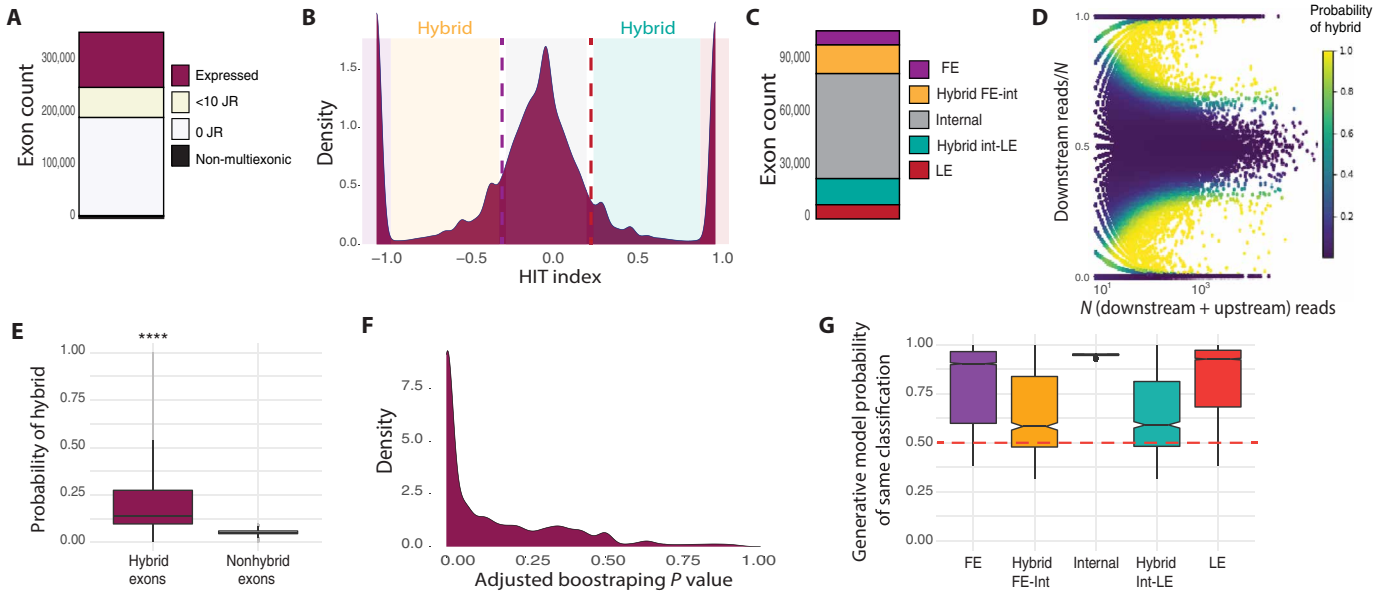

Fig. 2. Classifying hybrid exons in RNA-seq data. (**A**) The number of exons in multiexonic genes expressed in human lymphoblastoid cell lines (LCLs). (**B**) Distribution of HIT indices (*x* axis) for expressed LCL exons (>10 junction reads). (**C**) The classification of expressed exons into each of the five exon types including both hybrid categories. (**D**) Distribution of the ratio of downstream SJRs to total (*N*) SJRs (*y* axis) relative to *N* (*y* axis) across expressed exons. Heatmap represents the probability that an exon is classified as a hybrid exon by the generative model. (**E**) Probability that an exon is classified as hybrid by generative model (*y* axis) for exons classified as hybrid or nonhybrid by the HIT index metric alone (*x* axis). (**F**) Distribution of *P* values estimating the probability that the observed HIT index arises from a null distribution of internal exons (*x* axis). (**G**) Probability of being classified in the same category by the HIT index metric and the generative model (*y* axis) for all five exon types (*x* axis). Statistical significance of one-way analysis of variance (ANOVA), Tukey post hoc test, is indicated by asterisks (****$P < 0.0001$).

annotations (Fig. 3C). This suggests that current genome annotations substantially underestimate the existence of hybrid exons. A similar underestimation of hybrid exons was observed when conditioning on annotated exons and comparing to HIT index classifications (Fig. 3D). Consistent with previous work (*28*), we find that only a minority of annotated exons are expressed in a single cell type and, of these, many hybrid exons are misclassified as internal exons. Together, our new classifications show that a reliance on genome annotations in RNA-seq analyses underestimates the occurrence of alternative terminal exons by misclassifying hybrid exons as internal.

We also used published Oxford Nanopore direct RNA long-read sequencing data from LCLs to evaluate our classifications. Despite the high frequency of truncated reads in long-read sequencing (fig. S6), we were able to validate the HIT index classifications in selected genes that we highlight here (Fig. 4). First, the *SFPQ* gene encodes a member of the proline/glutamine-rich family of splicing factors and is associated with amyotrophic lateral sclerosis (*29*), Alzheimer's disease, and frontotemporal dementia (*30*). The 10th exon of *SFPQ* is used as a terminal exon in the longest *SFPQ* isoform and is essential, as isoforms that skip exon 10 are noncoding or degraded by nonsense-mediated decay (NMD). When exon 10 is excluded, splicing occurs to a group of exons in the 3′UTR, with exon 14 used as an ALE. We found that exon 12 is a hybrid FE-internal exon, used as an internal exon in transcripts that skip only exon 10 and as an FE in transcripts that skip exon 10 and use exon 14 as an LE (Fig. 4A). Because SFPQ binds the intron between exon 9 and downstream putative NMD-triggering exons (Fig. 4B), it has been hypothesized that SFPQ might regulate its own alternative splicing and polyadenylation (*31*). This proposed model of self-regulation suggests that the hybrid usage of *SFPQ* exons could contribute to the modulation of levels of active SFPQ protein.

In another example, *PTPN6* encodes a member of the protein tyrosine phosphatase (PTP) family that functions as an important regulator of multiple signaling pathways in hematopoietic cells and has a hybrid FE-internal exon and several splicing isoforms. We identified three AFEs, including a hybrid FE-internal second exon that is associated with expression of the predominant isoform (fig. S7A). Next, we found an unannotated hybrid internal-LE that acts as the predominant LE in LCLs for *RPS3A* (Fig. 4C), whose full-length isoform encodes a component of the 40*S* ribosome subunit. When the hybrid exon is used as an LE, the mRNA lacks a stop codon and is subjected to nonstop decay (Fig. 4D). Thus, regulation of the hybrid exon of *RPS3A* controls the levels of full-length protein available for incorporation into ribosomal small subunits. Last, *CCT8* encodes the theta subunit of the CCT chaperonin, which is involved in the transport and assembly of newly synthesized proteins. In LCLs, we observed expression of two AFEs and a hybrid internal-LE that produces a truncated protein (fig. S7B). These examples illustrate how hybrid exons can play an important role in modulating gene expression and isoform prevalence, affecting mRNA and protein levels.

Hybrid exons occur downstream of CpG islands and upstream of polyA sites

To understand the distribution of hybrid exons across human genes, we analyzed the molecular features of the genes in which they occur. We first focused on the gene architecture of these genes, specifically the number of overall exons and internal exons. We observed that hybrid exons tend to occur in genes with 15 or more internal exons (Fig. 5A and fig. S8A). However, genes with more internal exons often host only one hybrid exon (fig. S8B), suggesting that additional features govern the regulation of hybrid exons. As expected, hybrid

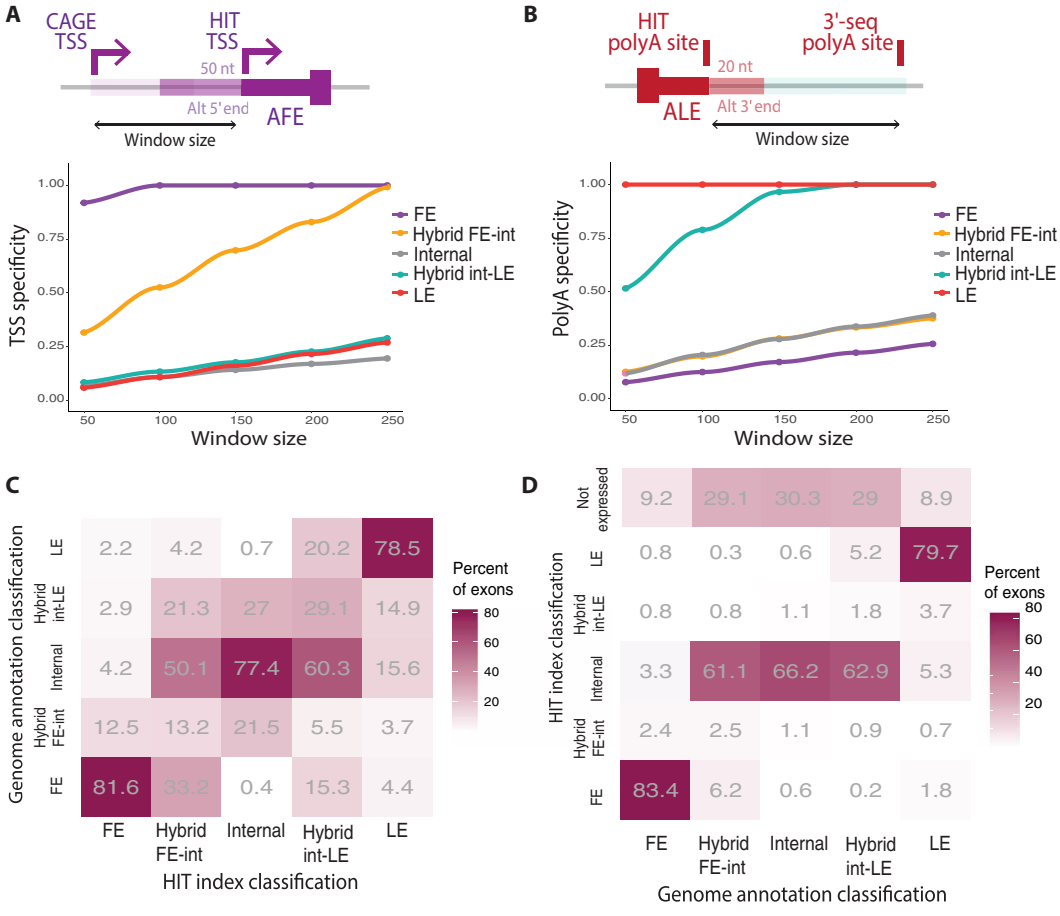

Fig. 3. Evaluation of HIT index performance. TSS specificity (**A**) and polyA specificity (**B**) for all five exon types across window sizes, where specificity is calculated as the fraction of HIT index TSSs (start coordinate of identified FE) or HIT TESs (end coordinate of identified LE) that fall within the given distance away from any CAGE-seq identified TSS or 3′seq-identified polyA site. (**C**) Heatmap of the percent of exons classified by the HIT index that are annotated as belonging to the same or different categories in the latest genome annotation. Column's percentage can add up to more than 100% when the same exon is annotated in more than one category. (**D**) Heatmap of the percent of exons annotated in the latest genome annotation that are classified into the same or different category by the HIT index or not expressed in LCL cells.

FE-internal exons occur toward the 5′ end of genes compared to hybrid internal-LE exons, although hybrid exons are generally located at internal positions relative to terminal-only exons (Fig. 5B and fig. S8, C and D). Furthermore, hybrid FE-internal exons often occur downstream of FEs and upstream of internal exons, while hybrid internal-LE exons often occur downstream of internal exons and upstream of LEs (fig. S8E). Introns that flank hybrid exons tend to have intermediate length, such that hybrid exons have longer flanking introns compared to internal exons but shorter flanking introns relative to terminal exons. This trend may allow for more genomic space for proximal sequence elements that regulate the usage of hybrid relative to internal exons (Fig. 5, C and D). Hybrid exons also tend to have equally strong or stronger splice sites than terminal exons, reinforcing the idea that the genomic characteristics of hybrid exons fall in between internal and terminal exons (Fig. 5, E and F).

We next sought to identify candidate genomic features involved in the regulation of hybrid exons, by looking for motifs enriched in regions flanking hybrid exons relative to regions flanking internal exons. For transcription to initiate from a hybrid FE-internal TSS, the immediately upstream sequence may act as a promoter. A recent study showed that ~60% of random sequences are 1 nucleotide (nt) away from being able to serve as active promoters in bacteria (*32*). In mammals, core promoters are still not fully understood but include structurally and functionally diverse sequences composed of a variety of DNA elements, including CpG islands (*33*–*35*). We observed that the regions upstream of hybrid FE-internal exons are enriched for motifs with long stretches of Cs and Gs (Fig. 5G), with an observed-to-expected CpG ratio greater than 60%, significantly more CpG islands (at least 200 nt of CGs enrichment), and an average of ~50% GC content (Fig. 5H). These CpG islands tend to be located immediately upstream of hybrid FE-internal start coordinates (fig. S8F). To test whether the enrichment of Cs and Gs upstream of hybrid FE-internal exons simply reflects high GC content of the host gene, we repeated the analysis using hybrid FE-internal and internal exons with matching locations within transcripts. In both cases, regions upstream of hybrid FE-internal and internal exons increased in GC content when we moved toward the 3′ end of genes, but regions upstream of hybrid FE-internal exons are still enriched for motifs high in Gs and have higher GC content compared to regions upstream of internal exons (fig. S8G). Our results do not address causality, but they suggest that CpG-rich regions upstream of internal

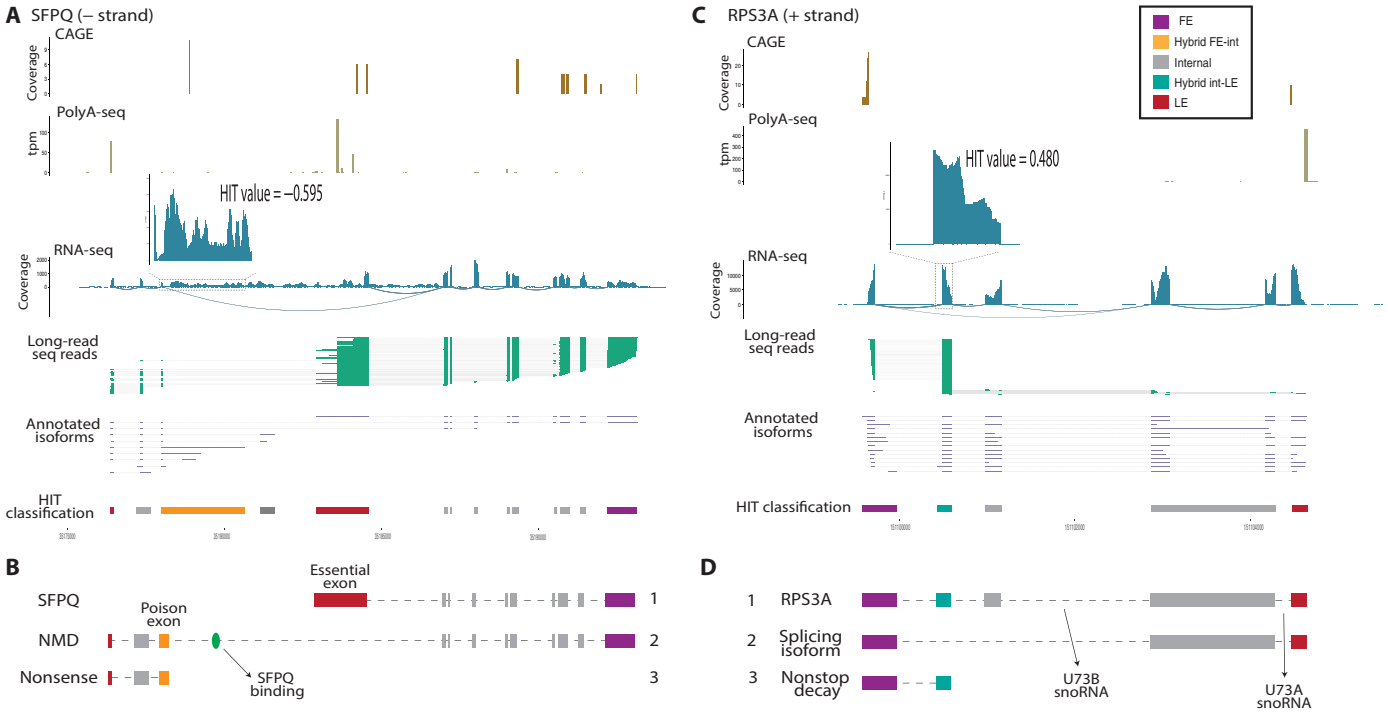

Fig. 4. Representative examples of hybrid exons. CAGE-seq peaks, polyA-seq peaks, RNA-seq coverage and junction reads, long-read sequencing reads, annotated isoforms, and HIT index classification for SFPQ (**A**) and RPS3A (**C**). There is no clear polyA-seq peak evidencing the use of the hybrid internal-LE as a terminal exon because the 3′-seq and the HIT index classifications correspond to different cell types. tpm, tags per million. (**B**) SFPQ, encoded by isoform 1 of *SFPQ*, binds its pre-mRNA cotranscriptionally and switches splicing toward the poison hybrid FE-internal exon. (**D**) Two alternative variants of *RPS3A*, where isoform 3 uses a hybrid LE-internal as a terminal exon and leads to nonstop decay.

exons are associated with a gain of hybrid usage and novel transcription initiation sites. To independently examine whether a sequence upstream of a hybrid FE-internal exon can act as a promoter, we looked at the localization of histone modification chromatin immunoprecipitation sequencing (ChIP-seq) peaks for modifications that are known to mark different regulatory regions. We find that, across all marks, the proportion of 500-nt regions upstream of hybrid FE-internal exons that have ChIP-seq peaks are more similar to FEs to internal exons (fig. S9). Notably, regions upstream of FE and hybrid FE-internal exons show substantial overlap with H2AFZ and H3K4me3 peaks, which occur at TSSs and promoters, respectively.

We also found that both FE and hybrid FE-internal exons less often contain ATG codons that start the open reading frame (ORF) compared to internal exons (fig. S8H), suggesting that they mostly affect UTRs. However, hybrid FE-internal exons have higher frequency of upstream ORFs (uORFs), which are key regulators of gene expression and often activate or inhibit translation of the main ORF downstream. Conversely, sequences downstream of hybrid internal-LE exons are enriched in motifs rich in As and Ts (Fig. 5I), have slightly lower CpG dinucleotide content, and have significantly more predicted polyA sites than internal exons (Fig. 5J). Our findings show that the gain of hybrid internal-LE usage is associated with the presence of downstream polyA sites. In addition, we observed that the primary sequences of human hybrid exons are evolutionarily more conserved than the sequence of terminal exons but less conserved than internal exons (Fig. 5K). Our findings demonstrate that hybrid exons have intermediate properties and suggest a functional relevance of hybrid exons in gene expression programs during evolution.

Widespread usage of intratissue and intertissue hybrid exons

Alternative mRNA isoforms are often considered as signatures of different tissues or cell types. Thus, we wanted to evaluate the differential usage of hybrid exons across human tissues. To do so, we used our HIT index pipeline to analyze RNA-seq data from the Genotype-Tissue Expression (GTEx) Project (*36*). We observed that the distribution of the HIT index does not vary significantly across human tissues (Fig. 6A), but the numbers of high-confidence hybrid exons show a tissue-specific pattern. Both types of hybrid exons are more predominant in testes, colon, and brain tissues, all tissues that are known to have extensive alternative isoform usage (Fig. 6B). The brain enrichment is consistent with neuronal genes being larger, with longer introns, and undergoing more alternative RNA processing. While both types of hybrid exons showed similar tissue specificity, we observed significantly more hybrid FE-internal exons than hybrid internal-LE exons per tissue (Fig. 6B), with twice as many observed in brain tissues, suggesting that hybrid FE-internal exons are more prevalent in the human genome. The greater variability of hybrid FE-internal exons across tissues suggests that these exons may be more tissue-specifically regulated than hybrid internal-LE exons.

Hybrid exons might be more likely to arise in genes with particular functions or tissue-specific expression patterns. To understand whether the same genes consistently use hybrid exons across tissues, we calculated the fraction of exons expressed in brain tissues that

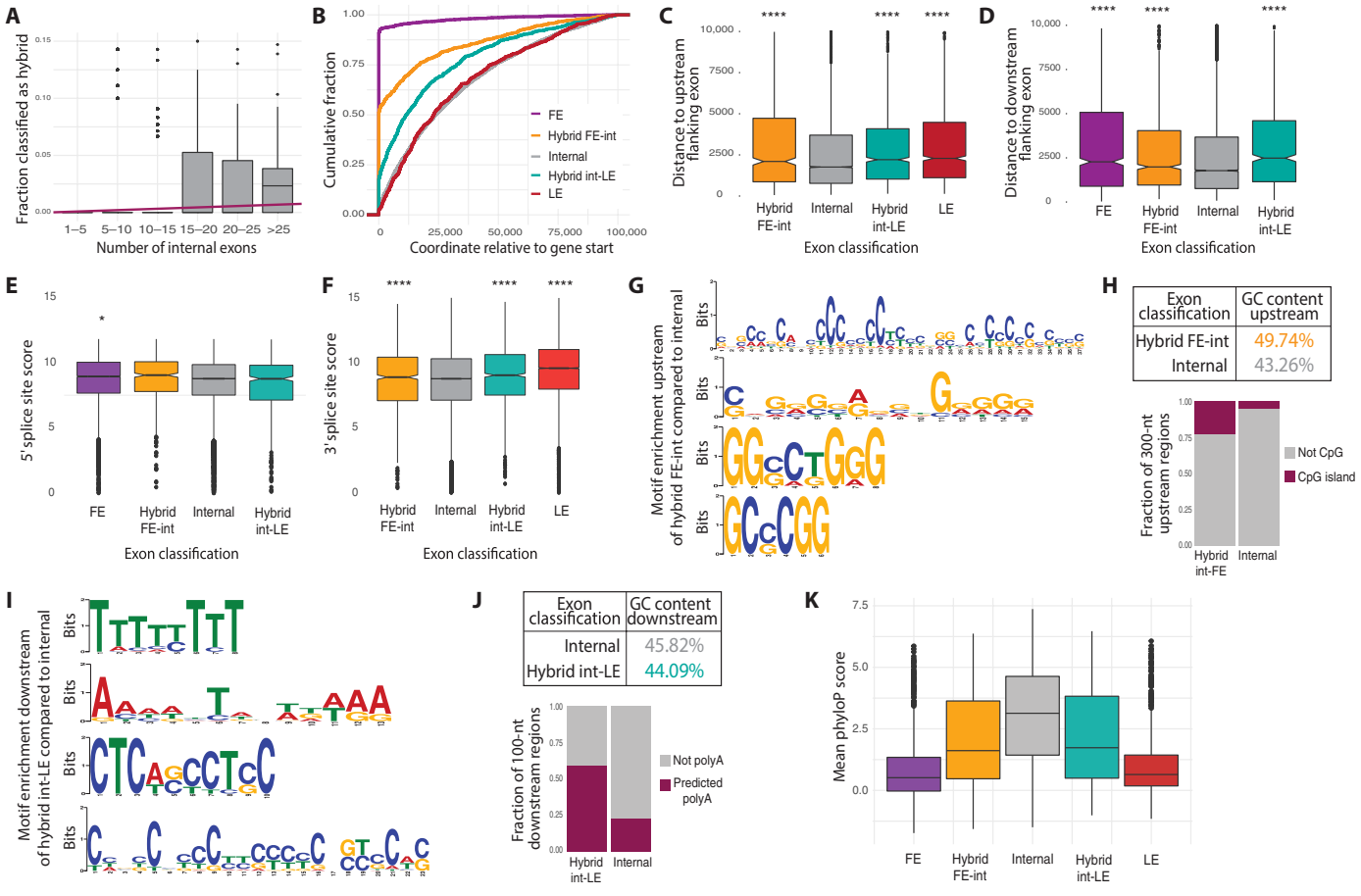

Fig. 5. Genomic properties of hybrid exons. (**A**) Distribution of the fraction of exons classified as hybrid per gene (*y* axis) binned by the number of internal exons in the same gene (*x* axis) and linear regression trend. (**B**) Cumulative distribution function of distances to the start of the gene (*x* axis) for each of the five exon types across genes expressed in LCLs. Gene start is defined as the start coordinate of the most upstream first exon. (**C**) Distance from the end of the upstream flanking exon to the start of each meta-exon (*y* axis), binned by exon type (*x* axis). (**D**) Same as (C) but distance from the end of each meta-exon to the start of the downstream flanking exon. 5′ splice site (**E**) and 3′ splice site (**F**) maxEnt scores (*y* axis) for meta-exons across exon types in LCL (*x* axis). (**G**) Enriched motifs in the 100–nucleotide (nt) region upstream of hybrid FE-internal relative to all HIT index classified internal exons. (**H**) GC content in the 100-nt region upstream of hybrid FE-internal or HIT index internal exons and fraction of 300-nt regions upstream of each exon that include a CpG island (table). (**I**) Enriched motifs in the 100-nt region downstream of hybrid internal-LE relative to HIT index internal exons. (**J**) GC content (*y* axis) in the 100-nt region downstream of hybrid internal-LE or HIT index internal exons and fraction of 100-nt regions downstream of each exon that include a predicted polyA site (table). (**K**) Distribution of the mean phyloP scores based on a 30-way primate genome comparison of all primary sequences of exons in five different categories. Statistical significance of one-way ANOVA, Tukey post hoc test, is indicated by asterisks (*$P < 0.01$ and ****$P < 0.0001$).

are classified as the same or different exon type in heart. We observed that, although the majority of exons expressed in brain tissues are not expressed in heart, most expressed terminal or internal exons are classified as the same exon type in heart. In contrast, exons identified as hybrid FE-internal in brain are equally likely to be classified as FE, hybrid FE-internal, or internal in heart. Similarly, hybrid internal-LE exons in brain tissues are mostly classified as LE, followed by hybrid internal-LE and internal in heart (Fig. 6C). Expanding this analysis to all pairwise tissue comparisons, both types of hybrid exons are more likely to be classified as terminal or internal exons than hybrid exons in other tissues, suggesting that hybrid exons are more likely to be alternatively used across tissues with hybrid usage tending to be somewhat tissue-restricted (Fig. 6D). We note that because hybrid exons are more difficult to detect, the occurrence of hybrid exons may be underestimated. Exons classified as FEs and LEs in brain tissues are both mostly classified as internal in other tissues, while brain internal exons are mostly classified as FE and LEs. Similarly, brain-specific hybrid exons are mostly classified as internal and FE or LE (depending on their hybrid type; Fig. 6E). Last, the HIT index classifications of exons in the same tissue across different individuals in the GTEx dataset are significantly more similar than the exon classifications of a given individual across tissues (fig. S10, A to C), suggesting that our classifications are robust to technical variance across libraries and that ~20 to 25% of exon classifications are tissue-specific. Together, our findings demonstrate that hybrid exons are more likely to be used as both terminal and internal exons in a tissue-specific manner, while being used exclusively as terminal or internal exons in other tissues. This finding is consistent with the observation that hybrid exons tend to have properties intermediate between internal and terminal exons, allowing for more flexibility in isoform usage.

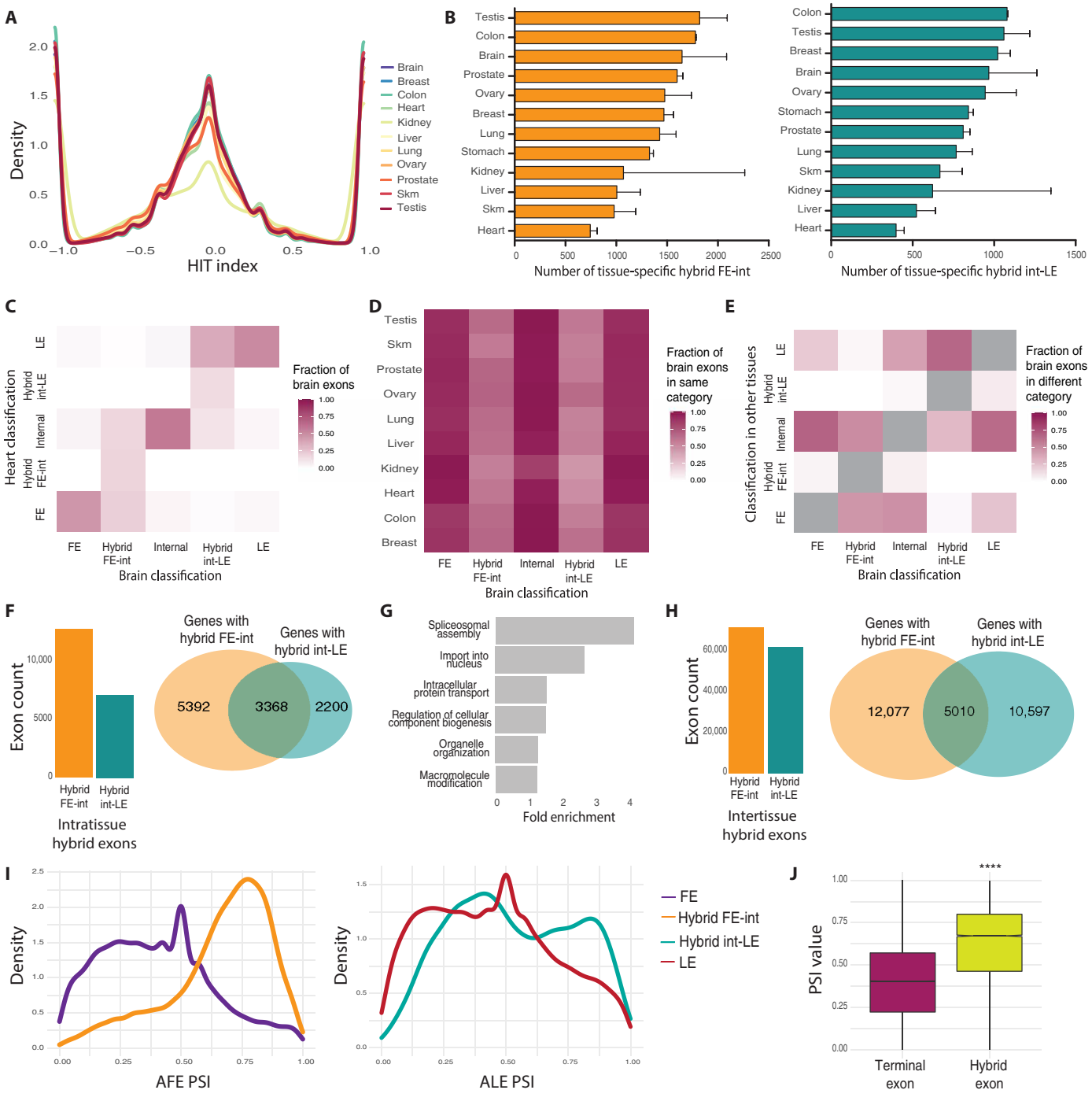

Fig. 6. Landscape of hybrid exons across human tissues. (**A**) HIT index distributions of exons expressed across 11 human tissues. (**B**) The number of tissue-specific hybrid FE-internal (left) and hybrid internal-LE exons (right) classified by the HIT index across 11 human tissues. (**C**) Heatmap of the fraction of exon-type classifications for exons in brain tissues (*x* axis) classified as the same or different category in heart tissues (*y* axis). (**D**) Heatmap of the fraction of classifications of exons in brain that are classified in the same category in the other 10 human tissues analyzed. (**E**) Heatmap of the fraction of classifications of exons in brain that are classified in any of the four other categories across the other 10 human tissues analyzed together. (**F**) The number of hybrid FE-internal and hybrid internal-LE exons used in both categories (internal and terminal exon) within the same tissue (left) and the overlap between genes with hybrid FE-internal and genes with hybrid internal-LE exons within the same tissue (right). The overlap is 1.7 above background expectation ($P < 1.6 \times 10^{-20}$). (**G**) Fold enrichments for GO categories that are significantly enriched (adjusted *P* values <0.05) among human genes with both types of intratissue hybrid exons. (**H**) The number of internal exons that are classified as FE or LE in any of the other 10 tissues analyzed (left); the overlap between genes with intertissue hybrid FE-internal and genes with intertissue hybrid internal-LE (right). The overlap is 1.3 above background expectation ($P < 1.6 \times 10^{-20}$). (**I**) Distribution of PSI values for AFE (left) and ALE exons (right), including hybrid exons across human tissues with multiple AFEs and ALEs. (**J**) Distribution of PSI values (*y* axis) of hybrid exons relative to terminal exons in genes with multiple FEs and/or LEs across all human tissues analyzed. Statistical significance of one-way ANOVA, Tukey post hoc test, is indicated by asterisks. Skm, skeletal muscle.

Because hybrid exons allow for increased isoform diversity, we reasoned that hybrid exons might arise within genes that would benefit from flexibility in RNA processing. We first asked whether genes with one type of hybrid exon are more likely to have another type of hybrid in the same tissue. Genes with hybrid FE-internal exons are 1.3-fold more likely to have hybrid internal-LE exons than expected by chance (overlap P value $<9.8 \times 10^{-85}$) (Fig. 6F). Genes with both types of intratissue hybrid exons are enriched in spliceosome components and splicing factors (Fig. 6G). The enrichment of hybrid exons within genes encoding for factors involved in splicing may contribute to the autoregulation of these factors, as proposed for *SFPQ* above, or may contribute to tissue-specific splicing programs. More broadly, an exon could be used as a terminal exon in one tissue, but as an internal exon in another tissue, which we term intertissue hybrid exons. We identified more than 96,000 intertissue hybrid exons across the 12 tissues we analyzed. Consistent with our observations for intratissue hybrids, genes with intertissue hybrid FE-internal exons are significantly more likely to have an intertissue hybrid internal-LEs (Fig. 6H). These findings suggest that gene structural (e.g., gene length and number of exons) or functional properties that favor evolution and maintenance of both types of hybrid exons are related. Last, we asked what proportion of isoforms are likely to use hybrid exons as terminal exons. Unexpectedly, we observed that hybrid exons are used in a relatively high proportion of isoforms compared to terminal exons, as measured by PSI values. When used as terminal exons, hybrid exons tend to produce the predominant gene isoforms across all tissues analyzed, suggesting that the switch between terminal and internal exon usage between tissues or conditions can have profound impacts on gene regulation (Fig. 6, I and J, and fig. S10, D and E). Our findings indicate that tissue-specific hybrid exons drive large changes in isoform usage and are likely to have major impacts on mRNA and protein synthesis.

DISCUSSION

Here, we describe the first computational pipeline to identify and classify hybrid exons, a largely undescribed type of exons, from RNA-seq data. The HIT index uses SJRs to classify exons as terminal exons that serve solely as FEs or LEs in a transcript, internal exons, or hybrid exons. We find that tens of thousands of human exons are intertissue hybrid exons that can serve as either terminal or internal exons in a transcript. However, technical biases due to library preparation and low proportional usage of exons (figs. S1 to S3) may lead to an underestimation of the true number of hybrid exons in human tissues. Moreover, increased tissue and cell-type sampling with higher coverage sequencing data could increase our power to identify intra- and intertissue hybrid exons and the contexts in which they are used.

Independent genome-wide high-throughput sequencing datasets targeting terminal site usage and long-read RNA-seq were used to validate our classifications and hybrid exons in individual genes. We obtained 100% TSS and polyA specificity for all terminal exons including hybrid exons (Fig. 3). Consistent with previous studies (*37*–*40*), the high frequency of truncated reads in the long-read sequencing prevented us from using these data to globally validate terminal or internal site usage. Combined with high error rates and potential coverage biases, long-read direct RNA-seq anchored on the polyA tail leads to substantial 5′ truncation of transcripts, reflected in 5′ ends of reads being located farther, on average, from annotated terminal sites than 3′ ends of reads, and read lengths that are smaller than average transcript lengths of the corresponding annotated isoforms (fig. S6, A and B). These biases significantly affect terminal exon identification and quantification. Thus, only 50% of our HIT index classifications are matched using long-read sequencing methods (fig. S6C), which could be due to either the aforementioned lack of power in long reads for terminal exon identification, limitations of the HIT index classifications, or a combination of both. Despite the great promise of long-read sequencing technologies, short-read sequencing datasets are still necessary to classify exons. Furthermore, the HIT index may provide a useful tool to calibrate terminal end truncation in long-read RNA-seq datasets.

We found that hybrid exons are an inherently flexible exon type due to their ability to serve as either terminal or internal exons within a transcript. Given their potential to increase both the coding capacity and regulatory potential within a genome, hybrid exons are an underappreciated way in which cells can expand their transcriptome and proteome to fine-tune alternative isoform usage from a single gene. Because hybrid exons can influence either UTRs or coding sequences depending on their usage, they likely experience different selective pressures than either internal-only or terminal-only exons. Their dual regulation by both splicing machinery and transcription or polyadenylation machinery (for hybrid FE or LE exons, respectively) suggests that hybrid exons lie in regions with flexible or multipurpose regulatory sequences.

Our analysis detected CG-rich sequences upstream of hybrid FE-internal and polyA-rich motifs downstream of hybrid internal-LE exons, indicating that hybrid exons are enriched for specific sequence features that may enable their use as either terminal or internal exons. These features are generally associated with transcription initiation and polyadenylation, respectively, but we did not specifically identify any auxiliary sequences canonically associated with mRNA splicing (apart from their canonical splice sites). These findings suggest that there might be sites involved in the competition between splicing and transcription/polyadenylation, which may be influenced by as yet unidentified sequence features, local epigenetic marks, DNA confirmation, RNA secondary structure, or coordination with other RNA processing events on the same transcript. Evolutionary gain or loss of a single splice site may, in some cases, be sufficient to convert a terminal exon to a hybrid exon or vice versa, so that exon "hybridization" and "dehybridization" may occur at least as often as evolutionary gain of new exons (*8*, *41*). Further studies of the evolutionary histories of these flexible hybrid exons may shed light on how new exons or isoforms arise in tissues or species and the mechanisms by which they are recognized by transcription factors, splicing components, and termination machinery.

Together, our findings demonstrate the unique role that hybrid exons play in human transcriptomes and highlight that our understanding of transcriptome complexity is far from complete.

MATERIALS AND METHODS

HIT index pipeline for exon classification

The HIT index approach uses the ratio between SJRs that overlap the beginning or end of exons to classify exons as being first, hybrid first, internal, hybrid last, or last exons. The HIT index pipeline involves three primary steps: (i) annotating meta-exons from a transcript annotation, (ii) extracting overlapping junction reads for each exon, and (iii) calculating the SJR ratios and generative model probabilities to classify exons and estimating PSI values for alternative

terminal exons. The full pipeline is available as a set of python scripts, with detailed usage instructions, at github.com/thepailab/HIT index and https://doi.org/10.5281/zenodo.5658060.

Exon definition and junction read counting

First, for each gene in a transcriptome reference file (i.e., gtf), annotated exons with overlapping boundaries are merged with pybedtools (*42*) to create a meta-exon. The boundaries of these meta-exons are used to extract junction reads. To account for ambiguity in the exact position of TSSs, splice sites, or mRNA cleavage sites, an optional buffer region is added to the boundaries of meta-exons, with a recommended distance of 50 nt upstream of the 5′ end of the exon and 30 nt downstream of the 3-end of the exon. Second, SJRs (split between two regions) are extracted using samtools (*43*). For each exon, SJRs that span the 5′ boundary (with at least 10 nt overlap) are counted as upstream SJRs and SJRs that span the 3′ boundary (with at least 10 nt overlap) are counted as downstream SJRs, regardless of the junction-specific boundaries (i.e., specific splice sites used). For example, an exon with two alternatively used 5′ splice sites would have an upstream SJR count that is the sum of junction reads derived from splicing to both 5′ splice sites. SJRs that begin and end in the same meta-exon are considered spurious and not used for further analysis.

SJR ratio estimation

The SJR ratio and associated metrics are calculated for each exon with at least five total SJRs. Specifically, the SJR ratio is calculated as the ratio of the imbalance in upstream and downstream SJRs over the total SJRs for the exon, as defined in Eq. 1

$$\text{SJR ratio} = \frac{\text{SJR}_{\text{up}} - \text{SJR}_{\text{down}}}{\text{SJR}_{\text{up}} + \text{SJR}_{\text{down}}} \quad (1)$$

To assess the robustness of the SJR ratio estimation, we report two parametric bootstrapping approaches with a null assumption that most exons are likely internal. First, bootstrapped SJR ratios for each exon are calculated from reads randomly subsampled with replacement, with a pseudocount added to both upstream and downstream junction read counts. Seventy-five percent, 90%, and 95% confidence intervals are calculated from the distribution of bootstrap statistics. The bootstrapped confidence interval *P* value (pval_CI) is calculated as the number of times that bootstrapped statistics are less than the |SJR ratio threshold|, representing the probability that the confidence interval overlaps the SJR ratio distribution classifying internal exons. Second, bootstrapped SJR ratios are calculated from reads sampled from an internal exon with the same SJR count as each exon. The bootstrapped *P* value from a null internal exon distribution (pval_internal) is calculated as the number of times that the bootstrapped statistics are as extreme or more extreme than the observed SJR ratio, representing the probability that the observed SJR ratio arises from a null distribution (internal exons). Both bootstrapping approaches are conducted with a user-defined number of bootstraps (1000 suggested samples).

Edge-effect flag

Simulated data uncovered biases in the SJR ratio estimation that existed specifically for exons near the edges of transcripts, where an RNA-seq edge effect led to a depletion of upstream or downstream SJRs for exons near the 5′ or 3′ ends of transcripts, respectively (fig. S1C). To account for this, the HIT index pipeline flags exons that have a higher probability of being affected by the edge effect and allows the user to decide whether to include them in downstream analyses. Exons are flagged on the basis of their distance to the transcript end and an empirical determination of the distance at which SJR ratios are biased due to the edge effect, which is expected to be approximately one fragment length away from the end of the transcript. Specifically, piecewise linear regression is used to find the inflection point between the two subsets of exons: (i) exons unaffected by edge effect–induced SJR biases, with equal upstream and downstream SJRs on average, and (ii) exons affected by edge effects, with a linear relationship between the exon's distance from the transcript end and the SJR bias. Exons with a 5′ or 3′ distance lower than the inflection point and not classified as terminal-only exons (|HIT index| = 1) are flagged as edge-effect exons.

Probabilistic framework

Given technical biases in the data, we developed a generative approach that estimates the probability that the ratio of downstream SJRs to the total SJRs arises from a particular exon type. Rather than assume that the ratio of downstream SJRs is generated from exons with some fixed probabilities, we assume that first-only, internal-only, and last-only exons generally have downstream SJR probabilities that are near 1, 0.5, and 0, respectively, but vary among the different exons in each class. Assuming the modes of these distributions are likely 1, 0.5, and 0, the model estimates two features from the data: (i) how concentrated downstream SJR probabilities are around these modes and (ii) the probabilities with which exons belong to each of these classes. Using these two assumptions, we model downstream SJR probabilities as arising from a mixture of these three distributions. Hybrid exons are then defined as exons poorly explained by or outliers relative to any of the three nonhybrid distributions.

Specifically, we define D_i as the number of downstream SJRs mapping to exon i of N_i total SJRs. We model D_i as arising from a mixture of beta-binomial distributions, representing the classes of first-only *F*, internal-only *I*, last-only *L*, and hybrid *H* exons, as defined in Eq. 2

$$D_I \sim \text{BetaBinomialMixture}(N_i, \nu, \kappa, \omega) \quad (2)$$

A Dirichlet prior is used for the probability, ω, that an exon belongs to each class, defined as ω~Dirichlet(α). The modes of the beta-distributions in the beta-binomial mixture are defined as (which are held constant as fixed assumptions) $\nu = \{\nu_F = 1, \nu_I = 0.5, \nu_L = 0, \nu_H = 0\}$. We want to use the data to learn the concentration of beta-distributions around these modes, κ, for first-only, internal-only, and last-only exons. We place relatively uninformative priors on these parameters, drawing them from half-normal distributions shifted to start at 3 because κ must be greater than 2 for the beta-distributions for κ_F and κ_L to not become uniform: κ_F, κ_I, $\kappa_L \sim \text{HalfNormal}(\text{loc} = 3, \sigma = 1000)$. In contrast, for the class of hybrid exons, we set $\kappa_H = 2$, which, when coupled with the model $\nu_H = 0$, corresponds to a uniform distribution under the assumption that, for hybrid exons, all downstream SJR probabilities are equally plausible.

We use automatic differentiation variational inference (ADVI) as implemented in the PyMC3 python package (*44*) to fit a full-rank approximation of the posterior distribution, allowing estimates of the posterior means and variation around this mean. Given a sample θ of *M* draws from the posterior distribution, defined as $\theta_i \sim P(\theta \mid D_i N_i)$, we can estimate the probability that an exon arose from a particular component distribution z_i based on observed D_i and N_i counts, as defined in Eq. 3

$$E[P(z_i \mid D_i, N_i, \hat{\theta})] = \frac{1}{M}\sum_j \frac{P(D_i \mid z_i, N_i, \hat{\theta}_j)}{\sum_{\zeta \epsilon Z} P(D_i \mid \zeta, N_i, \hat{\theta}_j)} \quad (3)$$

Last, hybrid exons are partitioned into two classes using the posterior over the underlying proportion of downstream SJRs, with P(HybridFI) and P(HybridIL) defined in Eqs. 4 and 5, respectively

$$P(\text{HybridFI}) = \int_{0.5}^{1} q \cdot P(q \mid \text{Hybrid}, D, N, \kappa_I) \cdot P(\text{Hybrid})\, dq \quad (4)$$

$$P(\text{HybridIL}) = \int_{0}^{0.5} q \cdot P(q \mid \text{Hybrid}, D, N, \kappa_I) \cdot P(\text{Hybrid})\, dq \quad (5)$$

The generative model leads to the following outputs: (i) the posterior mean estimate for the fraction of downstream SJRs, (ii) the posterior mean estimate for the fraction of downstream SJRs converted to the HIT index, and (iii) the posterior classification probabilities for first-only $P(F)$, first-internal $P(FI)$, internal $P(I)$, internal-last $P(IL)$, and last-only $P(L)$ exons.

Exon classification and PSI value calculations

Exons are classified into the following categories: "first," "FirstInternal_medium," "FirstInternal_high," "internal," "InternalLast_medium," "InternalLast_high," and "Last" exons, using a combination of thresholds across the SJR ratios, statistical confidence metrics, and posterior probabilities from the generative model, with all thresholds specified in the user-defined HIT_identity_parameters file. Specifically, terminal exons are defined as having an |SJR ratio| greater than or equal to the specified "HIT terminal" threshold and a bootstrapping variance value less than the specified "HIT pval." Hybrid exons are defined as having an |SJR ratio| greater than or equal to the specified "HIT hybrid" threshold, a |HIT index| less than the specified "HIT terminal" threshold, and an appropriate generative model posterior probability greater than the specified "prob_med" or "prob_high" thresholds, for medium- and high-confidence classifications, respectively. Last, internal exons are defined as all remaining exons, including those exons whose specified bootstrapped confidence interval overlaps zero. Exons classified as "first," "FirstInternal_medium," or "FirstInternal_high" are used to calculate PSI values for AFE usage, while exons classified as "last," "InternalLast_medium," or "InternalLast_high" are used to calculate PSI values for ALE usage, for each exon i of n total FEs or LEs as defined in Eqs. 6 and 7, respectively. PSI values specifically estimate the relative proportion of transcripts that use the exon as a terminal exon and are not intended to estimate the proportional terminal versus internal usage of a hybrid exon

$$\text{PSI}_{\text{AFE}} = \frac{\text{SJR}_{\text{down}_i} - \text{SJR}_{\text{up}_i}}{\sum_{i=0}^{n} \text{SJR}_{\text{down}_i} - \text{SJR}_{\text{up}_i}} \quad (6)$$

$$\text{PSI}_{\text{ALE}} = \frac{\text{SJR}_{\text{up}_i} - \text{SJR}_{\text{down}_i}}{\sum_{i=0}^{n} \text{SJR}_{\text{up}_i} - \text{SJR}_{\text{down}_i}} \quad (7)$$

If the user selects the --edge flag, exons flagged as potentially being affected by edge effects (see above) are not included in the PSI value calculations.

Simulations

We simulated 12 sets of artificial transcripts to assess the specificity and sensitivity of both the SJR ratio and generative model (fig. S1A). For each of these transcript architectures, we simulated short-read RNA-seq data from a range of different genomic and expression contexts: exon length (100 nt to 1 kb), proportional usage of AFEs or ALEs (0.05 to 0.95 PSI), proportional usage of SEs (0.05 to 0.95 PSI), library insert size (150 to 300 nt), read length (50 to 100 nt), and sequencing coverage (10 million to 100 million reads). Transcripts were also sampled from a range of gene expression levels that matched the distribution of TPMs in real RNA-seq data (fig. S1B). To generate short-read data from transcripts, we simulated several steps of Illumina library generation, including transcript fragmentation (using a modified Weibull distribution) and fragment size selection [as detailed in (*45*)]. The SJR ratio, generative model, and combined HIT index pipeline were applied to simulated read data as described above.

RNA-seq data processing and analyses

We used RNA-seq data from LCL human cells available at National Center for Biotechnology Information Gene Expression Omnibus [accession no. GSE30400 (*46*)] and several human tissues from different individuals from GTEx (*36*) downloaded through dbGaP under accession numbers SRR1068855, SRR1091865, SRR1366790, SRR1093075, SRR1375738, SRR1093527, SRR1437274, SRR1341721, SRR1434586, SRR1347481, SRR1347481, SRR1313449, SRR1080415, SRR1085975, SRR1313969, SRR1322373, SRR1362332, SRR1395999, SRR1441768, SRR1470273, SRR1082352 SRR1086140, SRR1317751, SRR1323087, SRR1366473, SRR1396700, SRR1460409, SRR1490246, SRR1071379, SRR1083632, SRR1090650, SRR1321351, SRR1336314, SRR1388190, SRR1416516, and SRR1464788. We excluded cases with different numbers of reads in read1 and read2 and selected the most recent releases in cases with multiple ones. Reads were mapped using Spliced Transcripts Alignment to a Reference (STAR) (*47*), guided by transcriptome coordinates in the genome annotation database (Ensembl GRCh37.72). The HIT index pipeline was run on resulting bam files, with default parameters, to classify exons into five categories. Expressed exons required SJR > 0 and classified as follows:

FE: HIT index == −1

Hybrid internal-FE: −0.3 > HIT index > −0.95 and probability of being hybrid > 0.5

Internal: −0.3 < HIT index < 0.3

Hybrid internal-FE: 0.3 < HIT index <0.95 and probability of being hybrid > 0.5

LE: HIT index == 1

All Gene Ontology (GO) analyses were performed using the GO knowledge base developed by the GO Consortium, using relevant expressed genes as a background set.

CAGE, 3p-seq, and polyA site data processing and analysis

To evaluate performance of the HIT index, start and end coordinates of all classified exons in LCL cells were interrogated for overlap on the same strand with TSS sites from CAGE and polyA sites from 3′-seq within 50, 100, 150, 200, and 250 base pairs. Specificity was calculated as the fraction of exons overlapping a bona fide TSS or polyA site within the indicated window size. TSS sites were identified as the nucleotide within a peak with the maximum score in CAGE data from the Functional Annotation of the Mammalian Genome (FANTOM) project (*26*), while polyA sites were also identified with nucleotide resolution in 3′-seq data downloaded from the polyAsite database (*27*). The polyA sites were identified using 3′-seq data generated in an immortalized B cell line from a donor.

Motif, splice site, context, and ORF analyses

DNA sequences flanking classified exon were extracted from the genome annotation database (Ensembl GRCh37.72) using GetFasta

function from BEDTools (*48*) with default parameters and strand specificity. Motif analyses were done with Multiple Em for Motif Elicitation (MEME) suite packages (*49*) and customized R scripts. The 5′ and 3′ splice site of each exon was mapped to hg38 reference genome using BEDTools (v.2.28.0), and splice site scores were calculated with MaxEntScan (*50*). Among all splice sites for overlapping exons including in each meta-exon, the exons with the highest splice site score were selected. ORF and uORF analyses were performed in R (v.4.0.5) using the "matchPattern" function from R package Biostrings and customized scripts. All occurrences of start and stop codons were annotated for each exon. All graphical plots were generated using the R package ggplot2.

ChIP-seq analysis

Histone modification profiles were examined using published ChIP-seq data for the GM12878 LCL generated by the Bernstein laboratory as part of the ENCODE consortium (*51*). Bed files for pseudoreplicated narrow peaks called by the Model-based Analysis of ChIP-seq (MACs) software were downloaded, and peak regions were intersected with the 500 nt upstream, exonic regions, and 500 nt downstream of all exons classified by the HIT index in the GM12878 LCL RNA-seq data. Specific files included H2AFZ (ENCFF377OJG), H3K4me3 (ENCFF998CEU), H3K4me1 (ENCFF321BVG), H3K36me3 (ENCFF475QVQ), and H3K27me3 (ENCFF571CZY).

Long-read RNA-seq analysis

A raw fastq file with base-called Oxford Nanopore direct RNA-seq data from the GM12878 LCL (*37*) was downloaded and mapped to the GRCh38 reference assembly using minimap2 (*52*). Ninety-eight percent of the 13.3 million total reads mapped with high-confidence primary alignments. Mapped reads were assigned to genes using the GRCh38 gtf annotations, and for downstream analyses, only reads mapping to protein coding transcripts were considered. To account for technical biases in the long-read sequencing data that lead to truncated or chimeric reads, we undertook a number of filtering steps to obtain a high-confidence set of reads. First, we calculated *z* scores for reads across the distribution of reads per gene to estimate how many SDs a mapped length is away from the mean mapped length of the gene. We discarded reads with mapped lengths 3 *z* scores away from the mean. Second, we compared the mapped length of a read against annotated coordinates using GRCh38 annotations. Reads with mapped lengths less than or greater than 1.5 kb from the smallest or largest annotated transcripts for the corresponding gene, respectively, were discarded. Last, we only considered genes with more than 10 reads. After applying these filters, we obtained 7.7 million reads across approximately 10,400 genes, with a mean read length of 1.1 kb. To identify first, internal, and last exons, each read was split into discrete exons using the CIGAR string. For each read, the most upstream and downstream exons were classified as FEs or LEs, respectively, and all other exons were considered as internal exons.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abk1752

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. A. Reyes, W. Huber, Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* **46**, 582–592 (2018).
2. S. A. Shabalina, A. Y. Ogurtsov, N. A. Spiridonov, E. V. Koonin, Evolution at protein ends: Major contribution of alternative transcription initiation and termination to the transcriptome and proteome diversity in mammals. *Nucleic Acids Res.* **42**, 7132–7144 (2014).
3. S. Pal, R. Gupta, H. Kim, P. Wickramasinghe, V. Baubet, L. C. Showe, N. Dahmane, R. V. Davuluri, Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res.* **21**, 1260–1272 (2011).
4. P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. T. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. M. Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schönbach, K. Sekiguchi, C. A. M. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, Y. Hayashizaki; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group), The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
5. A. Derti, P. Garrett-Engele, K. D. Macisaac, R. C. Stevens, S. Sriram, R. Chen, C. A. Rohl, J. M. Johnson, T. Babak, A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22**, 1173–1183 (2012).
6. D. Baek, C. Davis, B. Ewing, D. Gordon, P. Green, Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res.* **17**, 145–155 (2007).
7. B. Tian, Z. Pan, J. Y. Lee, Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.* **17**, 156–165 (2007).
8. A. Fiszbein, K. S. Krick, B. E. Begg, C. B. Burge, Exon-mediated activation of transcription starts. *Cell* **179**, 1551–1565.e17 (2019).
9. E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, C. B. Burge, Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
10. B. Tian, J. L. Manley, Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18**, 18–30 (2017).
11. A. G. Hinnebusch, I. P. Ivanov, N. Sonenberg, Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416 (2016).
12. K. Theil, K. Imami, N. Rajewsky, Identification of proteins and miRNAs that specifically bind an mRNA in vivo. *Nat. Commun.* **10**, 4205 (2019).
13. R. P. Jansen, mRNA localization: Message on the move. *Nat. Rev. Mol. Cell Biol.* **2**, 247–256 (2001).
14. A. Bashirullah, R. L. Cooperstock, H. D. Lipshitz, Spatial and temporal control of RNA stability. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 7025–7028 (2001).
15. K. Leppek, R. Das, M. Barna, Author Correction: Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.* **19**, 673 (2018).
16. D. Demircioğlu, E. Cukuroglu, M. Kindermans, T. Nandi, C. Calabrese, N. A. Fonseca, A. Kahles, K.-V. Lehmann, O. Stegle, A. Brazma, A. N. Brooks, G. Rätsch, P. Tan, J. Göke, A pan-cancer transcriptome analysis reveals pervasive regulation through alternative promoters. *Cell* **178**, 1465–1477.e17 (2019).
17. Z. Xue, R. L. Warren, E. A. Gibb, D. M. Millan, J. Wong, R. Chiu, S. A. Hammond, C. A. Ennis, A. Hahn, S. Reynolds, I. Birol, Pan-cancer analysis reveals complex tumor-specific alternative polyadenylation. *BioRxiv*, 160960 (2017).
18. A. A. Pai, F. Luca, Environmental influences on RNA processing: Biochemical, molecular and genetic regulators of cellular response. *Wiley Interdiscip. Rev. RNA* **10**, e1503 (2019).

19. E. K. Robinson, P. Jagannatha, S. Covarrubias, M. Cattle, R. Safavi, R. Song, K. Viswanathan, B. Shapleigh, R. Abu-Shumays, M. Jain, S. M. Cloonan, E. Wakeland, M. Akeson, A. N. Brooks, S. Carpenter, Inflammation drives alternative first exon usage to regulate immune genes including a novel iron regulated isoform of *Aim2*. *bioRxiv*, 2020.07.06.190330 (2020).
20. J. Vaquero-Garcia, A. Barrera, M. R. Gazzara, J. González-Vallinas, N. F. Lahens, J. B. Hogenesch, K. W. Lynch, Y. Barash, A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* **5**, e11752 (2016).
21. A. D. Tang, C. M. Soulette, M. J. van Baren, K. Hart, E. Hrabeta-Robinson, C. J. Wu, A. N. Brooks, Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.* **11**, 1438 (2020).
22. K. C. H. Ha, B. J. Blencowe, Q. Morris, QAPA: A new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol.* **19**, 45 (2018).
23. R. Goering, K. L. Engel, A. E. Gillen, N. Fong, D. L. Bentley, J. M. Taliaferro, LABRAT reveals association of alternative polyadenylation with transcript localization, RNA binding protein expression, transcription speed, and cancer survival. *bioRxiv*, 2020.10.05.326702 (2020).
24. Z. Qin, P. Stoilov, X. Zhang, Y. Xing, SEASTAR: Systematic evaluation of alternative transcription start sites in RNA. *Nucleic Acids Res.* **46**, e45–e45 (2018).
25. A. A. Cass, X. Xiao, mountainClimber identifies alternative transcription start and polyadenylation sites in RNA-Seq. *Cell Syst.* **9**, 393–400.e6 (2019).
26. M. Lizio, I. Abugessaisa, S. Noguchi, A. Kondo, A. Hasegawa, C. C. Hon, M. de Hoon, J. Severin, S. Oki, Y. Hayashizaki, P. Carninci, T. Kasukawa, H. Kawaji, Update of the FANTOM web resource: Expansion to provide additional transcriptome atlases. *Nucleic Acids Res.* **47**, D752–D758 (2019).
27. C. J. Herrmann, R. Schmidt, A. Kanitz, P. Artimo, A. J. Gruber, M. Zavolan, PolyASite 2.0: A consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Res.* **48**, D174–D179 (2019).
28. A. Joglekar, A. Prjibelski, A. Mahfouz, P. Collier, S. Lin, A. K. Schlusche, J. Marrocco, S. R. Williams, B. Haase, A. Hayes, J. G. Chew, N. I. Weisenfeld, M. Y. Wong, A. N. Stein, S. A. Hardwick, T. Hunt, Q. Wang, C. Dieterich, Z. Bent, O. Fedrigo, S. A. Sloan, D. Risso, E. D. Jarvis, P. Flicek, W. Luo, G. S. Pitt, A. Frankish, A. B. Smit, M. E. Ross, H. U. Tilgner, A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat. Commun.* **12**, 463 (2021).
29. R. Luisier, G. E. Tyzack, C. E. Hall, J. S. Mitchell, H. Devine, D. M. Taha, B. Malik, I. Meyer, L. Greensmith, J. Newcombe, J. Ule, N. M. Luscombe, R. Patani, Intron retention and nuclear loss of SFPQ are molecular hallmarks of ALS. *Nat. Commun.* **9**, 2010 (2018).
30. J. Lu, R. Shu, Y. Zhu, Dysregulation and dislocation of SFPQ disturbed DNA organization in Alzheimer's disease and frontotemporal dementia. *J. Alzheimers Dis.* **61**, 1311–1321 (2018).
31. D. Pervouchine, Y. Popov, A. Berry, B. Borsari, A. Frankish, R. Guigó, Integrative transcriptomic analysis suggests new autoregulatory splicing events coupled with nonsense-mediated mRNA decay. *Nucleic Acids Res.* **47**, 5293–5306 (2019).
32. A. H. Yona, E. J. Alm, J. Gore, Random sequences rapidly evolve into de novo promoters. *Nat. Commun.* **9**, 1530 (2018).
33. L. Vo Ngoc, C. J. Cassidy, C. Y. Huang, S. H. C. Duttke, J. T. Kadonaga, The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes Dev.* **31**, 6–11 (2017).
34. T. W. Burke, J. T. Kadonaga, Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.* **10**, 711–724 (1996).
35. A. Sandelin, P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki, D. A. Hume, Mammalian RNA polymerase II core promoters: Insights from genome-wide studies. *Nat. Rev. Genet.* **8**, 424–436 (2007).
36. The GTEx Consortium, The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
37. R. E. Workman, A. D. Tang, P. S. Tang, M. Jain, J. R. Tyson, R. Razaghi, P. C. Zuzarte, T. Gilpatrick, A. Payne, J. Quick, N. Sadowski, N. Holmes, J. G. de Jesus, K. L. Jones, C. M. Soulette, T. P. Snutch, N. Loman, B. Paten, M. Loose, J. T. Simpson, H. E. Olsen, A. N. Brooks, M. Akeson, W. Timp, Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
38. C. Soneson, Y. Yao, A. Bratus-Neuenschwander, A. Patrignani, M. D. Robinson, S. Hussain, A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* **10**, 3359 (2019).
39. C. Sessegolo, C. Cruaud, C. Da Silva, A. Cologne, M. Dubarry, T. Derrien, V. Lacroix, J.-M. Aury, Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Sci. Rep.* **9**, 14908 (2019).
40. K. A. Reimer, C. A. Mimoso, K. Adelman, K. M. Neugebauer, Co-transcriptional splicing regulates 3′ end cleavage during mammalian erythropoiesis. *Mol. Cell* **81**, 998–1012.e7 (2021).
41. J. J. Merkin, P. Chen, M. S. Alexis, S. K. Hautaniemi, C. B. Burge, Origins and impacts of new mammalian exons. *Cell Rep.* **10**, 1992–2005 (2015).
42. R. K. Dale, B. S. Pedersen, A. R. Quinlan, Pybedtools: A flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
43. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
44. J. Salvatier, T. V. Wiecki, C. Fonnesbeck, Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2**, e55 (2016).
45. A. A. Pai, T. Henriques, K. McCue, A. Burkholder, K. Adelman, C. B. Burge, The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture. *eLife* **6**, e32537 (2017).
46. J. Rozowsky, A. Abyzov, J. Wang, P. Alves, D. Raha, A. Harmanci, J. Leng, R. Bjornson, Y. Kong, N. Kitabayashi, N. Bhardwaj, M. Rubin, M. Snyder, M. Gerstein, AlleleSeq: Analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7**, 522 (2011).
47. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
49. T. L. Bailey, J. Johnson, C. E. Grant, W. S. Noble, The MEME suite. *Nucleic Acids Res.* **43**, W39–W49 (2015).
50. G. Yeo, C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
51. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
52. M. H. N. Yousefi, M. Goudarzi, S. A. Motahari, IMOS: Improved Meta-aligner and Minimap2 On Spark. *BMC Bioinformatics* **20**, 51 (2019).

Acknowledgments: We thank all members of the Burge laboratory, Fiszbein laboratory, and Pai laboratory for useful discussions and comments. **Funding:** This work was supported by the National Institutes of Health grant R35GM133762 (A.A.P.), Pew Latin American postdoctoral fellowship (A.F.), and Faculty funds from Boston University (A.F.). **Author contributions:** Conceptualization: A.F. and A.A.P. Methodology: A.F., M.M., C.B.B., and A.A.P. Investigation: A.F., M.M., E.C.-R., G.K., and A.A.P. Visualization: A.F., M.M., E.C.-R., G.K., C.B.B., and A.A.P. Supervision: A.F., C.B.B., and A.A.P. Writing—original draft: A.F. and A.A.P. Writing—review and editing: M.M., E.C.R., G.K., and C.B.B. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The HIT index python pipeline is also available at github.com/thepailab/HITindex and https://doi.org/10.5281/zenodo.5658060.

Submitted 10 July 2021
Accepted 23 November 2021
Published 19 January 2022
10.1126/sciadv.abk1752

ScienceAdvances

Widespread occurrence of hybrid internal-terminal exons in human transcriptomes

Ana FiszbeinMichael McGurkEzequiel Calvo-RoitbergGyeungYun KimChristopher B. BurgeAthma A. Pai

Sci. Adv., 8 (3), eabk1752. • DOI: 10.1126/sciadv.abk1752

Science Advances (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.