Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop

Timothy Sterne-Weiler†*[1], Robert J. Weatheritt†[1,2,4], Andrew Best[1], Kevin C. H. Ha[1,3], and Benjamin J. Blencowe*[#1,3]

[1] Donnelly Centre, University of Toronto, Toronto, M5S 3E1 Canada
[2] MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge, CB2 0QH UK
[3] Department of Molecular Genetics, University of Toronto, Toronto, M5S 3E1 Canada
[4] EMBL Australia, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, New South Wales 2010, Australia

† These authors contributed equally
* Co-corresponding Authors
[#] Lead Contact and Senior Author

Correspondence:
Donnelly Centre, University of Toronto
Toronto, ON, Canada
Office 416-978-3016
Lab 416-978-7150
Fax 416-946-5545
Email: b.blencowe@utoronto.ca; tim.sterne.weiler@utoronto.ca

SUMMARY

Alternative splicing (AS) is a widespread process underlying the generation of transcriptomic and proteomic diversity and is frequently misregulated in human disease. Accordingly, an important goal of biomedical research is the development of tools capable of comprehensively, accurately and efficiently profiling AS. Here, we describe Whippet, an easy-to-use RNA-Seq analysis method that rapidly – with hardware requirements compatible with a laptop – models and quantifies AS events of any complexity without loss of accuracy. Using an entropic measure of splicing complexity, Whippet reveals that one-third of human protein coding genes produce transcripts with complex AS events involving co-expression of two or more principal splice isoforms. We observe that high-entropy AS events are more prevalent in tumor relative to matched normal tissues, and correlate with increased expression of proto-oncogenic splicing factors. Whippet thus affords the rapid and accurate analysis of AS events of any complexity, and as such will facilitate future biomedical research.

INTRODUCTION

High-throughput RNA sequencing (RNA-seq) technologies are producing vast repositories of transcriptome profiling data at an ever expanding pace (Silvester et al., 2018). This explosion in data has enabled genome-wide investigations of the role of alternative splicing (AS) in gene regulation and its dysregulation in human diseases and disorders. Initial investigations using RNA-seq data revealed that ~95% of human multi-exon gene transcripts undergo AS (Pan et al., 2008; Wang et al., 2008). These and more recent studies analyzing ribosome-engaged transcripts and quantitative mass spectrometry data suggest that AS is a major process underlying the generation of transcriptomic and proteomic complexity (Floor and Doudna, 2016; Liu et al., 2017; Sterne-Weiler et al., 2013; Weatheritt et al., 2016; reviewed in Blencowe, 2017). Furthermore, numerous AS events belonging to co-regulated and evolutionarily conserved exon networks have been shown to provide critical functions in diverse processes (Baralle and Giudice, 2017; Tapial et al., 2017).

A major challenge confronting genome-wide investigations of AS is that existing methods for analyzing RNA-seq data require extensive computational resources and expertise. For example, widely employed tools involve alignment of reads to a transcriptome or reference genome, followed by quantification by downstream methods that estimate percent spliced in (PSI, Ψ) values for each AS event, such as cassette exons, alternative 5′ and 3′ splice sites, and retained introns. These steps can be time-consuming and typically present a bottleneck when analyzing large datasets.

Recent developments in transcript expression quantification have circumvented traditional alignment steps by extracting k-mers (i.e. all possible sequences of length k) from reads to identify possible transcripts of origin. Such methods can decrease

processing times by 10-100 fold (Bray et al., 2016; Patro et al., 2017). However, their accuracy relies on whole 'transcript-level' annotation models (i.e. models that record the precise location of intron and exon boundaries, and spliced junctions, for all transcripts), which are incomplete for the majority of species, and inconsistent among even the best annotated species. The lack of complete annotation models can thus confound the accurate detection and quantification of AS events when using transcript-level methods. More widely used methods for the RNA-seq analysis, focusing on the local detection and quantification of AS events, are referred to below as 'event-level' approaches (**Figure S1A;** Katz et al., 2010; Tapial et al., 2017; Wang et al., 2017). These methods can achieve considerable accuracy for simple AS events (Vaquero-Garcia et al., 2016), yet existing tools are computationally inefficient in comparison with transcript-level methods, and most utilize predetermined simple binary models (i.e. a single alternative exon surrounded by two constitutive exons), making them poorly suited for the analysis of complex AS patterns.

In light of these challenges, an important goal for understanding how transcriptomes shape biological processes is to develop methods capable of the accurate analysing simple and complex AS patterns with high efficiency. To address these challenges, we have developed Whippet, an easy-to-use event-level software tool for the accurate and efficient detection, and quantification of AS events of any complexity. Whippet has computational requirements compatible with a laptop computer and is capable of analysing reads streamed from web-accessible data files by entering a file accession number. Another feature of Whippet is that it uses an entropic measure of AS to facilitate the accurate profiling of AS. We demonstrate the utility of Whippet in the discovery of previously uncharacterized AS complexity in vertebrate transcriptomes associated with the regulation of tandem domains and other

protein sequence features, as well as a remarkable increase in AS complexity in cancer transcriptomes.

DESIGN

Efficient quantification of alternative splicing using Whippet

Whippet models transcriptome structure by building `Contiguous Splice Graphs` (CSGs). These are directed graphs whose nodes are non-overlapping exonic sequences, and edges (i.e. connections between nodes) represent splice junctions or adjacent exonic regions (**Figures 1A and 1B**). Splice graphs allow single isoforms to be represented as paths through nodes in the graph (Heber et al., 2002; Trapnell et al., 2010; Vaquero-Garcia et al., 2016). Whippet's CSGs extend the concept of splice graphs to a lightweight data structure that indexes the transcriptome for fast and modular alignment of raw RNA-seq reads across splice junctions (**Figures 1B and 1C**). To facilitate indexing, Whippet defines incoming and outgoing boundary types (e.g. 5′ or 3′ splice sites, or transcription start or end sites; refer to **Figure 1B** legend for details) that specify the theoretical connectivity through the CSG for each node (**Figure 1B**; **Figure S1B**). For each 5′ or 3′ splice site boundary, Whippet's CSG index records an upstream or downstream k-mer respectively, so as to enable efficient spliced read alignment across all possible splice junctions, including junctions that do not occur within annotated transcripts but which combine annotated donor or acceptor splice sites (**Figures 1B-1D; Figure S1C-D**; **Methods** for details). For example, Whippet's CSG index for the human genome hg19 build can represent AS events from >1.3 million exon-exon junctions in >2.3 billion theoretically possible isoform

paths, whereas only ~100K of these paths are found in GENCODE v25 TSL1 annotated transcripts.

After alignment, a Whippet AS event is defined as the collective set of a node's skipping or connecting edges (e.g. edge 1-3 skips node 2, and edges 1-2 and 2-3 connect to node 2 in **Figure 1E**; see **Methods**). When enumerating paths through a node's AS event, it is possible that multiple paths share common (i.e. ambiguous) edges (e.g. edges 1-2 and 3-4 are shared among multiple paths in **Figure 1E**). Therefore, to accurately quantify all AS events, the proportional abundance of each path is determined using maximum likelihood estimation by the expectation-maximization (EM) algorithm (see **Methods**). The percent spliced in (PSI, Ψ; range 0.0 to 1.0) value of a node is then calculated as the sum of the proportional abundance of the paths containing the node (**Figure 1E**).

RESULTS

Whippet facilitates accurate analysis of alternative splicing

To assess Whippet's accuracy, we compared its Ψ values with those measured from RT-PCR data, and commonly used RNA-seq event-level analysis tools (Irimia et al., 2014; Katz et al., 2010; Wang et al., 2017; Vaquero-Garcia et al., 2016) – which quantify Ψ using reads that directly map to an AS event – as well as transcript-level tools (Trincado et al., 2018), which estimate Ψ based on reads mapping across entire transcripts (see **Methods S1** and **Figure S2A-G** for details of mapping benchmarking). RT-PCR- and RNA-seq-derived Ψ values were both from adult mouse liver and cerebellum, as well as from stimulated and unstimulated human Jurkat T-cell line samples (Vaquero-Garcia et al., 2016). Notably, Whippet and the

other event-level tools display ~2.5 fold lower median error profiles compared to transcript-level methods, including SUPPA2 (Trincado et al., 2018) and Whippet_TPM, an approach developed in the present study to afford direct comparisons of transcript-level Ψ estimates that maintain Whippet's node definitions (**Figure 2A; Table S1; Figure S2H and S3A-B; Methods**).

Benchmarking against RT-PCR Ψ values, while informative, is limited by the relatively small sample set (n=162), the types of the events assessed, and possible intrinsic technical biases introduced by PCR. To address this, we assessed the accuracy of Whippet relative to other tools when comparing their Ψ values against synthetic (i.e. 'ground truth') Ψ values simulated from RNA-seq data obtained from a reference transcriptome annotation (GENCODE v25 TSL1 for hg19; **Methods**).

In contrast to results from benchmarking against RT-PCR data, we find that transcript-level methods perform with similar accuracy to event-level approaches, including Whippet, when using simulated RNA-seq data (compare **Figures 2A** and **2B**). This discrepancy is likely due to the artificial nature of the simulation, where the exact transcript-annotations used to generate the reads are provided to the quantification software. In the analysis of RNA-seq data from biological samples, the quantification software will likely be challenged by discrepancies between the annotation model and the set of true transcripts present in the sample (e.g. **Figure 2C** shows that a large percentage of alternative splice junctions in vertebrate species are not annotated in Ensembl). To investigate such effects, we simulated RNA-seq reads with ground truth Ψ values using one annotation set (RefSeq Release 84 for hg19), and created an index database for each quantification program using another annotation set (GENCODE v25 TSL1 for hg19). Notably, in this comparison (and the inverse comparison in **Figure S3C**) there is a 2-2.5 fold increase in error rate for

estimating Ψ values using transcript-level methods, but minimal change in error rate for any of the event-level tools, including Whippet (**Figure 2B; Figure S3D**). We conclude that differences in transcript reference annotations can confound estimates for Ψ values when using transcript-level methods, whereas event-based methods are largely insensitive to this issue.

The analyses so far used widely employed transcript annotations from human and mouse, which are among the most complete for any species. To assess Whippet's performance when analyzing species with less extensively annotated transcripts, we applied it to RNA-seq data (Brawand et al., 2011) from five of the same tissues from gorilla, chimp, opossom, chicken, as well as from mouse and human. While ~12% of alternative exon-exon junctions aligned by Whippet in human and mouse are unannotated, the percentage of unannotated AS junctions is in the range of 40-80% in the other species (**Figure 2C**). These observations further indicate that transcript-level tools, and event-level tools reliant on annotated AS events, fail to detect a considerable amount of unannotated transcript diversity in vertebrates. In contrast, Whippet can detect and accurately quantify AS events involving numerous unannotated splice junctions represented by pairings of combinations of splice sites from its CSG indices (see also below).

The benchmarks described so far focus on "simple" AS events, such as single cassette alternative exons flanked by pre-defined constitutive exons that have binary splicing outcomes. However, many AS events involve splice sites that are variably paired with two or more other sites. Whippet provides output metrics designed to quantify such AS complexity in two related ways. First, it classifies AS events into discrete bins of complexity based on the number of enumerated paths from the event (i.e. $n = \lceil log_2(paths) \rceil$, such that K($n$) can produce at most 2^n spliced outcomes for

K1, …, K6; **Figure 2D**). Second, it calculates a Ψ-dependent measure of AS complexity using Shannon's entropy (i.e., entropy = $-\sum_i \Psi_i \, log_2 \Psi_i$ such that the maximum entropy for an event in K(*n*) is *n*; **Figure 2E; Figures S4A and S4B**). This entropic measure conveniently formalizes the total number of possible outcomes for an event, and the degree of their proportional contribution to the transcriptome in a read-depth and read-length independent manner (**Figures S4C and S4D**)

To assess whether Whippet accurately quantifies AS events with increasing degrees of complexity and entropy, we simulated RNA-seq datasets and corresponding Ψ values for events in the formalized categories (K1, …, K6) of increasing complexity and distributed entropy (**Figures 2D-E; Figure S4E**). In contrast to other methods tested, the accuracy of Whippet-derived estimates for Ψ does not decrease as the complexity and entropy of the simulated AS events increases. This difference in performance is because Whippet has the unique feature among the event-level approaches tested of employing the EM algorithm to assign reads that are ambiguously shared between multiple paths through high entropy AS events. This capability translates as a ~2-3 fold greater accuracy for Whippet in the quantification of K2-K6 events compared to other tested methods (**Figures 2E, 2F and S4F**).

To further assess Whippet's performance relative to other methods, we next investigated whether transcript-level methods potentially achieve comparable accuracy when provided with a predefined annotation set that comprehensively represents complex events. To test this, we built a transcript annotation set from combinatorial Whippet graph paths (N4 annotation file, **Methods**). While this annotation set allows SUPPA2 to detect unannotated AS events, its error rate in estimating Ψ values is still four-fold higher than Whippet's (**Figures 2F and Figure S4E-F**).

To experimentally validate Whippet-derived predictions of high AS event entropy, RNA-seq data (Raj et al., 2014) from mouse neuroblastoma (N2a) cells were analyzed and 10 events with different predicted degrees of entropy and complexity involving tandem arrays of alternative exons were tested by RT-PCR (**Methods**). Notably, 56/61 (91.8%) of the amplified spliced products were predicted by Whippet, whereas five (8.2%) of the expected isoforms were not detected. Of the detected products, 32 (52.5%) are consistent with annotated isoforms and 24 (39.3%) correspond to novel isoforms (**Figure 2G and Figure S5A**). Collectively, these data demonstrate that Whippet is an accurate method for the analysis of both simple and complex AS events.

Efficiency of Whippet

To assess Whippet's efficiency, we benchmarked speed and memory usage relative to published AS quantification methods. When analyzing several paired-end RNA-seq datasets from HeLa cells with increasing read depth (~15M, ~25M and ~50M), Whippet quantifies AS from a raw paired-end 25M RNA-seq read dataset in 43 minutes while using less than 1.5GB of memory on a typical computer with a single core (Dual-Core AMD Opteron(tm) Processor 8218, 2.5 GHz, 60GB RAM, 1,024KB cache). This represents a considerable increase in performance over other tested event-level tools, and is of comparable performance to transcript-level methods (**Figures 2H** and **S5B-C; Table S2**). For example, MISO, the most highly-cited event-level tool, in combination with the read aligner STAR, took days and used 30 GB of memory to analyze the same data (**Figure 2H; Figure S5C**), whereas the fastest transcript-level methods took approximately 20 minutes. It is important to note that when provided with annotation sets for complex AS events (e.g. N4 annotation

file) the runtime and memory usage of transcript-level methods were greater than that of Whippet (**Figure 2H; Figure S5C**). Moreover, on a personal laptop with a solid-state hard drive (Macbook Pro 3.1 GHz Intel i7), Whippet quantified the ~25M dataset in 15 minutes using downloaded data files, and in 31 minutes when streaming data from the internet after inputting the SRA identifier. The considerably longer time taken to analyze the same data by MISO and some of the other event level tools may be influenced by the hardware used to run these programs. The unique features of Whippet thus obviate the use of high-performance computational clusters for the quantitative profiling of AS using RNA-seq data.

Taken together with the assessment of accuracy, the results indicate that Whippet offers advantages over other methods in terms of its capacity to reliably and efficiently detect and quantify AS events.

Detection of high-entropy, tissue-regulated AS events

Because previously described tools were not designed for the formalized quantitative profiling of AS complexity, we used Whippet to investigate the prevalence and possible biological relevance of high-complexity AS events in mammalian transcriptomes. To this end, we applied Whippet to an analysis of 60 diverse human and mouse tissue RNA-seq datasets (**Table S3; Figure 3A and S6A**). Remarkably, of more than 13,000 analyzed human protein coding genes, 42.68% harbor an AS event predicted to have an entropy > 1.0 (i.e. two or more expressed isoforms) in at least one tissue (**Figure S6B**; see **Methods**). Moreover, 4,101 (30.1%) of these genes co-express at least two major isoforms at similar levels in one or more of the *same* tissue (**Figure 3B, Figure S6C**; **Methods**). The majority (~20%) of the events are predicted to undergo substantial tissue-dependent changes in splicing

entropy (**Figure 3C**), without concurrent changes in expression of the corresponding genes (**Figure 3D**; R^2 = 0.074, Pearson correlation). These results contrast with previous proposals that the vast majority of mammalian genes express a single, major splice variant (Gonzalez-Porta et al., 2013; Tress et al., 2017), and instead are consistent with data indicating that a substantial fraction of genes express multiple major isoforms either within or between different cell and tissue types (Tapial et al., 2017; Vaquero-Garcia et al., 2016; Wang et al., 2008). However, new isoforms generated by high entropy AS events detected by Whippet further increase the estimated fraction of genes predicted to express multiple major isoforms compared to previous estimates (e.g. up to ~40% vs. ~18% in Tapial et al. 2017). Supporting the possible biological relevance of these AS events, the corresponding genes are enriched in functions associated with the cytoskeleton, extracellular matrix organization, cell communication, signaling and muscle biology (**Figure 3E**, p-values < 0.05; corrected FDR).

To further investigate the possible significance of high-entropy AS events detected by Whippet, we analyzed their evolutionary conservation using RNA-seq data from six of the same tissues from seven vertebrate species (Brawand et al., 2011), comparing entropy values for the orthologous exons (1,304 'low-entropy' [< 1.0] and 369 'high-entropy' [>1.5] exons, **Figure 4A, Figure S6D-E**) in each species. This revealed a significantly greater concordance in both Ψ and entropy values for orthologous AS events between the analyzed species than expected by chance, when compared to randomly-permuted sets of exons from the same data (**Figures 4B-C**, low-entropy AS events: $p < 2.2 \times 10^{-16}$; high-entropy AS events: $p < 4.3 \times 10^{-4}$, Kolmogorov–Smirnov test; **Figure S6F-G**; see **Methods**). Thus, overall, the degree of entropy of low- and high-complexity AS events detected and quantified by

Whippet is conserved across vertebrate species, implying that these patterns may often be functionally important.

We next asked whether these events are potentially translated. Due to the extremely limited coverage of currently available mass spectrometry data (Blencowe, 2017), Whippet was applied to RNA-seq data from HeLa mono- and polysomes, as well as from whole cell, nuclear, and cytosolic fractions (Floor and Doudna, 2016). This analysis reveals comparable distributions of AS event entropy across all samples (**Figure 4D**; d < 0.25, Cohen's D statistic, Nuclear vs. High Polyribosome), suggesting that high-entropy AS events contribute substantially to the translated transcriptome. Furthermore, the enrichment of high entropy AS events within the 5′UTRs of transcripts (**Figure S6H**, $p < 4.37 \times 10^{-38}$, Fisher's exact test) suggests possible roles in the regulation of translation.

High-entropy alternative splicing regulates genes with extensive domain repeats and disordered regions

Given previous evidence for important roles of AS in rewiring protein-protein interaction networks among other functions (Buljan et al., 2012; Ellis et al., 2012; Yang et al., 2016), we next investigated whether increasing levels of AS event entropy are associated with specific protein structural features. We observe a significant monotonic increase in the frequency of overlap with intrinsically disordered regions as a function of increasing entropy of AS events (**Figure 5A**; $p < 1.02 \times 10^{-43}$, Mann-Whitney U test, low-entropy [<1.0] vs highest-entropy [> 2.0] events; **Figure S7A**). As expected, an inverse trend is observed for overlap with structured domains (**Figure 5A**, $p < 1.78 \times 10^{-41}$, Mann-Whitney U test). However, an interesting exception is that highest-entropy AS events (entropy > 2.0) display

significant overlap with tandem repeat domains (**Figure 5A** $p < 2.14 \times 10^{-05}$, Mann-Whitney U test; **Figure S7A**), particularly nebulin-like and epidermal growth factor (EGF)-like domains (p-values < 0.05, Fisher-exact test). Further analysis of the highest-entropy (>2.0) AS events overlapping tandem protein domain repeats reveals that they are significantly more likely to arise from exon duplication than lower-entropy (<2.0) events (**Figure 5B**, $p < 4.57 \times 10^{-42}$, Fisher's exact test; **Figures S7B-C**). As an example, high-entropy AS events overlap two classes of tandem repeat domains, LDL-receptor class A and EGF-like domains, within the low-density lipoprotein receptor-related protein 8 (Lrp8). These events were confirmed by RT-PCR analysis (**Figure 5C**). Moreover, supporting their likely functional importance, one of them is differentially regulated by the neural and muscle-enriched splicing factor Rbfox2 (**Figure 5D**). These data thus provide evidence for important roles for Whippet-detected, high-entropy AS events in the expansion of proteomic diversity, principally through changes to intrinsically disordered protein regions and combinatorial changes to the composition of tandem arrays of specific-classes of protein domains.

High-entropy AS events display prototypical alternative splicing signals

We hypothesized that high-entropy AS events may be associated with specific sequence features that facilitate their complex patterns of regulation. To investigate this, we binned AS events by entropy and compared the strengths of their 3′- and 5′-splice sites, flanking intron lengths, and exonic splicing enhancer (ESE) and silencer (ESS) motif densities. Interestingly, the highest-entropy AS events show significant decreases in 3′- and 5′-splice site strength compared to low-entropy AS events (**Figure 6A**; $p < 3.73 \times 10^{-4}$ and 1.83×10^{-3}, Mann-Whitney U test). Additionally, we

observe monotonic decreases in flanking intron length (**Figure 6B**, $p < 1.78 \times 10^{-18}$, Mann-Whitney U test, highest vs lowest entropy events) and ESS motif density (**Figure 6C**; ESS: $p < 6.06 \times 10^{-05}$; Mann-Whitney U test, highest vs lowest entropy events) as a function of increasing entropy. In contrast, the density of ESE elements displayed a monotonic increase between low- and high- entropy AS events (**Figure 6C**; ESE: $p < 4.20 \times 10^{-06}$, Mann-Whitney U test, lowest vs highest entropy events). These results suggest that weak splice sites, reduced intronic length, and altered frequencies of exonic splicing elements, are important features underlying the regulation and function of high-entropy AS events (**Figure 6D**).

Global increases in high-entropy AS in cancer

Aberrant splicing is a hallmark of cancer and contributes to numerous aspects of tumor biology (Ladomery, 2013; Oltean and Bates, 2014). Cancer associated changes in AS have been linked to altered expression of RNA binding proteins, some of which are oncogenic or act as tumor suppressors, as well as to splicing-sensitive disease mutations that impact the levels or activities of other cancer-associated genes (Sebestyen et al., 2016; Sterne-Weiler and Sanford, 2014). Despite extensive evidence for altered AS in cancer (Climente-Gonzalez et al., 2017; Dvinge et al., 2016), the extent to which these changes relate to altered levels of splicing complexity has not been previously determined. Accordingly, we applied Whippet to compare AS entropy using RNA-seq data (**Table S3**) from 15 matched tumor and control liver samples of patients with hepatocellular carcinoma (HCC), the third leading cause of cancer deaths worldwide. Remarkably, this analysis revealed a significant and reproducible (i.e. between replicate samples) increase in AS event entropy and number of unannotated alternative exon-exon junctions detected in tumor compared

to control samples (**Figures 7A-7C; Figure S7D**; 4.30 x 10^{-18}, Mann-Whitney U test) with only a relatively small degree of correlating change in the expression levels of the corresponding genes (**Figure S7E;** R^2 = 0.412, Pearson Correlation Coefficient). Genes with the largest AS entropy changes display significant enrichment for functions known to be dysregulated in liver cancer, including DNA repair and cell-cycle regulation (**Figure 7D**; p-values < 0.05; corrected FDR).

Further investigation revealed AS events previously identified as aberrant in cancer samples (**Figure 7E**), including those associated with over-expression of the splicing regulator SRSF1 (Anczukow et al., 2015; Das and Krainer, 2014). Consistent with this observation, differential gene expression analysis revealed a number of RNA binding proteins, including SRSF1, that are significantly over-expressed in tumor compared to control samples (**Figure 7F-7G; Figure S7F**; DESeq2, False Discovery Rate-adjusted p-values < 0.01). To further investigate the possible role of SRSF1 over-expression in the expansion of AS entropy observed in the cancer samples, we used Whippet to analyze RNA-seq data (Anczukow et al., 2015) from an MCR-10A cell line over-expressing SRSF1. This revealed a significant increase in high-entropy AS events associated with SRSF1 over-expression (**Figure 7H**; p < 9.41 x 10^{-9}, Mann-Whitney U test, compared to control) and a significant overlap with events differentially regulated between tumor versus normal tissues (**Figure 7I**; p < 2.09 x 10^{-5}, Fisher's exact test). These data thus indicate that overall splicing entropy increases in specific tumor types in response to changes in the expression of oncogenic splicing regulators, such as SRSF1. These results further illustrate how Whippet's unique capacity for the efficient and quantitatively accurate profiling of high entropy AS patterns can provide insight into how transcriptomes are altered in different biological contexts.

DISCUSSION

Advancements in RNA-seq analysis have involved the generation of tools that estimate Ψ values from transcript-level expression information (Trincado et al., 2018). While such methods are efficient, we observe that they are subject to increased error rates as a result of inaccuracies in standard transcript annotation models. In contrast, event-level tools are insensitive to most annotation inaccuracies since they only consider reads that directly map to splice junctions, exons, or introns forming an AS event. In the present study, we describe Whippet, a graph- and indexing-based event-level approach for the rapid and accurate quantitative profiling of AS. Whippet applies the concept of lightweight algorithms (Bray et al., 2016; Patro et al., 2014) to splicing quantification using RNA-seq data. As such, it eliminates the requirement for extensive computational resources typically required for read alignment steps. It further affords an unprecedented degree of accuracy in the profiling of complex AS events, in part through the use of entropy as metric for the formalized analysis of AS complexity. Collectively, these attributes of Whippet facilitated the discovery and characterization of transcriptomic complexity and associated features in the present study.

Our results indicate that high-entropy AS events occur more frequently in vertebrate transcriptomes than previously appreciated (Nellore et al., 2016; Vaquero-Garcia et al., 2016), affecting up to 40% of human genes. In contrast to previous proposals that the vast majority of mammalian genes express a single major splice isoform (Gonzalez-Porta et al., 2013; Tress et al., 2017), our results from employing Whippet reveal that at least one-third of human and mouse genes simultaneously express multiple major isoforms. The results further suggest that many of these events

are biologically significant since their AS entropy levels are frequently tissue-regulated, conserved, and the corresponding variant transcripts are highly expressed.

Previously documented examples of high entropy AS events include those that control of the biophysical properties of giant proteins that form muscle fibers (Buck et al., 2010; Li et al., 2012). Many of the high entropy events detected by Whippet are also reminiscent of well-studied examples in other systems, such as the splice variants generated by tandem arrays of alternative exons in the Drosophila *DSCAM* gene (Bolisetty et al., 2015). In this example, high-entropy AS events overlap tandemly repeated immunoglobulin-like domains that function as interaction surfaces in neural circuit assembly (Hattori et al., 2008). Our results suggest that targeting of tandemly repeated domains by high-entropy AS may represent a widely used mechanism to modulate the functions of multi-domain proteins. We further provide evidence that large repertoires of transcripts from high-entropy AS events is particularly prominent in post-mitotic tissues, and likely contributes to intricate networks of regulation and cell-cell interactions in these tissues.

Alterations in splicing by spliceosomal gene mutations and over-expression of RBPs contribute to the transcriptomic dysfunction characteristics of myelodysplastic syndromes and related cancers (Inoue et al., 2016). We demonstrate a significant increase in AS event entropy in hepatocellular carcinoma, affecting genes that function in DNA damage and spindle formation, and relate these changes to the mis-regulation of the splicing factor SRSF1. These data may reflect an overall loss of splicing fidelity in cancers and exemplify how the formalization of AS entropy is important when evaluating changes in global splicing patterns (Ritchie et al., 2008). For example, such measures of entropic splicing change may be valuable in future diagnostic techniques for precision medicine.

In summary, Whippet enables the efficient and accurate profiling of simple to complex AS events. As such, it is expected to significantly facilitate future biomedical research. Whippet's ability to rapidly quantify raw read data as a stand-alone software package on a personal computer further renders genome-wide analyses of AS more accessible to the scientific community. In this regard, we believe that Whippet will represent a valuable tool until long-read sequencing protocols (Byrne et al., 2017; Tilgner et al., 2017) offer comparable sequencing depth and efficiency as short read analysis methods.

Limitations

A limitation of Whippet is that it only detects and analyses AS events represented by splice sites in a CSG index. However, it can detect and quantify previously unknown AS events representing novel combinations of splice junctions derived from the indexed splice sites. Moreover, CSG indices can be supplemented beyond standard annotation sets with new splice sites (and therefore novel exons) mined using *de novo* spliced read aligners (Dobin et al., 2013; Kim et al., 2015); see Methods S1 and **Figure S1E**). This approach is expected to be useful in the analysis of AS from poorly annotated species, as well as disease-altered transcriptomes harboring aberrant splicing patterns.

ACKNOWLEDGEMENTS

We gratefully acknowledge M. Irimia and P. Melsted for valuable suggestions and testing of the Whippet software. We also thank G. Bader, N. Barbosa-Morais, U. Braunschweig, S. Gueroussov, T. Gonatopoulos-Pourtnatzis and B. Harpur for helpful discussions and comments of the manuscript. This work was supported by grants from

the CIHR and Canada First Excellence Fund to B.J.B.. Additional support was provided by CIHR postdoctoral fellowships (T.S.W., R.J.W., A.B.), a C.H. Best Postdoctoral Fellowship (T.S.W.), Marie Curie IOF Fellowship (R.J.W.), and EMBO Long-Term Fellowship (A.B.), Ontario Graduate Scholarship, and CIHR Frederick Banting and C.H. Best Canada Graduate Scholarship (K.C.H.H). B.J.B holds the University of Toronto Banbury Chair in Medical Research.

AUTHOR CONTRIBUTIONS

T.S.W conceived, designed, and implemented the Whippet software, with contributions from R.J.W. T.S.W., R.J.W.; K.C.H.H. simulated data and benchmarked accuracy and performance. R.J.W. and T.S.W. designed and performed computational analyses, with input from B.J.B. A.B. performed experimental validations. T.S.W, R.J.W and B.J.B. wrote the manuscript with input from the other authors.

DECLARATION OF INTERESTS

The authors declare no competing interests

References:

Anczukow, O., Akerman, M., Clery, A., Wu, J., Shen, C., Shirole, N.H., Raimer, A., Sun, S., Jensen, M.A., Hua, Y., *et al*. (2015). SRSF1-Regulated Alternative Splicing in Breast Cancer. Molecular cell *60*, 105-117.

Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol.

Blencowe, B.J. (2017). The Relationship between Alternative Splicing and Proteomic Complexity. Trends Biochem Sci *42*, 407-408.

Bodenhofer, U., Kothmeier, A., and Hochreiter, S. (2011). APCluster: an R package for affinity propagation clustering. Bioinformatics *27*, 2463-2464.

Bolisetty, M.T., Rajadinakaran, G., and Graveley, B.R. (2015). Determining exon connectivity in complex mRNAs by nanopore sequencing. Genome biology *16*, 204.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., *et al*. (2011). The evolution of gene expression levels in mammalian organs. Nature *478*, 343-348.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature biotechnology *34*, 525-527.
Buck, D., Hudson, B.D., Ottenheijm, C.A., Labeit, S., and Granzier, H. (2010). Differential splicing of the large sarcomeric protein nebulin during skeletal muscle development. J Struct Biol *170*, 325-333.
Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., and Babu, M.M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Molecular cell *46*, 871-883.
Byrne, A., Beaudin, A.E., Olsen, H.E., Jain, M., Cole, C., Palmer, T., DuBois, R.M., Forsberg, E.C., Akeson, M., and Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nature communications *8*, 16027.
Climente-Gonzalez, H., Porta-Pardo, E., Godzik, A., and Eyras, E. (2017). The Functional Impact of Alternative Splicing in Cancer. Cell reports *20*, 2215-2226.
Das, S., and Krainer, A.R. (2014). Emerging functions of SRSF1, splicing factor and oncoprotein, in RNA metabolism and cancer. Molecular cancer research : MCR *12*, 1195-1204.
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.
Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433-3434.
Dvinge, H., Kim, E., Abdel-Wahab, O., and Bradley, R.K. (2016). RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer *16*, 413-430.
Ellis, J.D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., *et al*. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. Molecular cell *46*, 884-892.
Ferragina, P., Manzini, G., Mäkinen, V., and Navarro, G. (2004). An Alphabet-Friendly FM-Index. In String Processing and Information Retrieval: 11th International Conference, SPIRE 2004, Padova, Italy, October 5-8, 2004 Proceedings, A. Apostolico, and M. Melucci, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 150-160.
Floor, S.N., and Doudna, J.A. (2016). Tunable protein synthesis by transcript isoforms in human cells. eLife *5*.
Frazee, A.C., Jaffe, A.E., Langmead, B., and Leek, J.T. (2015). Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics *31*, 2778-2784.
Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome biology *14*, R70.
Grant, G.R., Farkas, M.H., Pizarro, A.D., Lahens, N.F., Schug, J., Brunk, B.P., Stoeckert, C.J., Hogenesch, J.B., and Pierce, E.A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics *27*, 2518-2528.
Hattori, D., Millard, S.S., Wojtowicz, W.M., and Zipursky, S.L. (2008). Dscam-mediated cell recognition regulates neural circuit formation. Annu Rev Cell Dev Biol *24*, 597-620.

Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic acids research *38*, e131.
Heber, S., Alekseyev, M., Sze, S.H., Tang, H., and Pevzner, P.A. (2002). Splicing graphs and EST assembly problem. Bioinformatics *18 Suppl 1*, S181-188.
Inoue, D., Bradley, R.K., and Abdel-Wahab, O. (2016). Spliceosomal gene mutations in myelodysplasia: molecular links to clonal abnormalities of hematopoiesis. Genes & development *30*, 989-1001.
Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallieres, M., Tapial, J., Raj, B., O'Hanlon, D., *et al*. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell *159*, 1511-1523.
Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods *7*, 1009-1015.
Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome research *21*, 1360-1374.
Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature methods *12*, 357-360.
Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology *14*, R36.
Ladomery, M. (2013). Aberrant alternative splicing is another hallmark of cancer. Int J Cell Biol *2013*, 463786.
Letunic, I., Copley, R.R., and Bork, P. (2002). Common exon duplication in animals and its role in alternative splicing. Human molecular genetics *11*, 1561-1567.
Letunic, I., Doerks, T., and Bork, P. (2015). SMART: recent updates, new developments and status in 2015. Nucleic acids research *43*, D257-260.
Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics *12*, 323.
Li, S., Guo, W., Schmitt, B.M., and Greaser, M.L. (2012). Comprehensive analysis of titin protein isoform and alternative splicing in normal and mutant rats. J Cell Biochem *113*, 1265-1273.
Liu, Y., Gonzalez-Porta, M., Santos, S., Brazma, A., Marioni, J.C., Aebersold, R., Venkitaraman, A.R., and Wickramasinghe, V.O. (2017). Impact of Alternative Splicing on the Human Proteome. Cell reports *20*, 1229-1241.
Love, M.I., Hogenesch, J.B., and Irizarry, R.A. (2016). Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. Nature biotechnology *34*, 1287-1291.
Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology *15*, 550.
Nellore, A., Jaffe, A.E., Fortin, J.P., Alquicira-Hernandez, J., Collado-Torres, L., Wang, S., Phillips Iii, R.A., Karbhari, N., Hansen, K.D., Langmead, B., *et al*. (2016). Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome biology *17*, 266.
Oltean, S., and Bates, D.O. (2014). Hallmarks of alternative splicing in cancer. Oncogene *33*, 5311-5318.
Pachter, L. (2011). Models for transcript quantification from RNA-Seq. In ArXiv e-prints.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature genetics *40*, 1413-1415.
Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature methods *14*, 417-419.
Patro, R., Mount, S.M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature biotechnology *32*, 462-464.
Pellegrini, M., Renda, M.E., and Vecchio, A. (2012). Ab initio detection of fuzzy amino acid tandem repeats in protein sequences. BMC Bioinformatics *13 Suppl 3*, S8.
Raj, B., Irimia, M., Braunschweig, U., Sterne-Weiler, T., O'Hanlon, D., Lin, Z.Y., Chen, G.I., Easton, L.E., Ule, J., Gingras, A.C., *et al*. (2014). A global regulatory mechanism for activating an exon network required for neurogenesis. Molecular cell *56*, 90-103.
Ritchie, W., Granjeaud, S., Puthier, D., and Gautheret, D. (2008). Entropy measures quantify global splicing disorders in cancer. PLoS computational biology *4*, e1000011.
Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome biology *12*, R22.
Sebestyen, E., Singh, B., Minana, B., Pages, A., Mateo, F., Pujana, M.A., Valcarcel, J., and Eyras, E. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome research *26*, 732-744.
Silvester, N., Alako, B., Amid, C., Cerdeno-Tarraga, A., Clarke, L., Cleland, I., Harrison, P.W., Jayathilaka, S., Kay, S., Keane, T., *et al*. (2018). The European Nucleotide Archive in 2017. Nucleic acids research *46*, D36-D40.
Sterne-Weiler, T., Martinez-Nunez, R.T., Howard, J.M., Cvitovik, I., Katzman, S., Tariq, M.A., Pourmand, N., and Sanford, J.R. (2013). Frac-seq reveals isoform-specific recruitment to polyribosomes. Genome research *23*, 1615-1623.
Sterne-Weiler, T., and Sanford, J.R. (2014). Exon identity crisis: disease-causing mutations that disrupt the splicing code. Genome biology *15*, 201.
Tapial, J., Ha, K.C.H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Vallieres, M., Permanyer, J., Sodaei, R., Marquez, Y., *et al*. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome research *27*, 1759-1768.
Tilgner, H., Jahanbani, F., Gupta, I., Collier, P., Wei, E., Rasmussen, M., and Snyder, M.P. (2017). Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. Genome research.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology *28*, 511-515.
Tress, M.L., Abascal, F., and Valencia, A. (2017). Most Alternative Isoforms Are Not Functionally Important. Trends Biochem Sci *42*, 408-410.
Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J., and Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome biology *19*, 40.

Vaquero-Garcia, J., Barrera, A., Gazzara, M.R., Gonzalez-Vallinas, J., Lahens, N.F., Hogenesch, J.B., Lynch, K.W., and Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. eLife *5*, e11752.
Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.
Wang, J., Pan, Y., Shen, S., Lin, L., and Xing, Y. (2017). rMATS-DVR: rMATS discovery of differential variants in RNA. Bioinformatics *33*, 2216-2217.
Weatheritt, R.J., Sterne-Weiler, T., and Blencowe, B.J. (2016). The ribosome-engaged landscape of alternative splicing. Nat Struct Mol Biol.
Wootton, J.C. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem *18*, 269-285.
Xiong, H.Y., Lee, L.J., Bretschneider, H., Gao, J., Jojic, N., and Frey, B.J. (2016). Probabilistic estimation of short sequence expression using RNA-Seq data and the positional bootstrap. bioRxiv.
Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G.M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y.A., Murray, R.R., *et al*. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell *164*, 805-817.
Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Journal of computational biology : a journal of computational molecular cell biology *11*, 377-394.

MAIN FIGURES LEGENDS

Figure 1 – Overview of Whippet algorithm

(A) Example gene model with two alternative isoforms and Whippet's node assignments, as indicated by number and separated by dashed lines. Gene models can be supplemented beyond standard annotation sets with new splice sites and novel exons mined using *de novo* spliced read aligners (see also **Figure S1E**).

(B) The Contiguous Splice Graph (CSG) for the same gene model in (A). Each CSG node has two boundaries: an incoming (left side of node, label pointing upwards) and outgoing (right side of node, pointing downwards), and these have 'Soft' or 'Hard' alignment extension properties (see D). Boundary types are designated as Hard or Soft depending on whether or not genomic sequence separates two neighboring nodes, respectively. All 5′ SpliceSite and 3′ SpliceSite boundaries have k-mer indices

(colored lines) that are used for spliced read alignment (*middle top*). Lines with arrows indicate potential connectivity (edges) between nodes (*middle bottom*).

(C) A single transcriptome Full-text index in Minute space (FM-Index) is built from concatenated CSG sequences, with solid lines indicating separation between each CSG (*bottom*).

(D) Diagram of CSG alignment, which is seeded from a raw RNA-seq read, and then extended in both directions. Alignments can extend through Soft, but not Hard boundaries. The two read k-mers flanking a spliced node boundary are used to return the set of compatible nodes for spliced junction extension (see **Methods** for CSG alignment rules).

(E) Example Whippet AS event (*top*) for a node *N*, defined as the set full set of spliced edges aligned (in an RNA-seq dataset) between the nodes farthest upstream or downstream for connecting (bolded labels) or skipping edges (regular labels). To determine Percent Spliced In (Ψ) of some node *N*, all paths through the AS event are enumerated (*bottom left*), and quantified through convergence of the expectation-maximization (EM) algorithm (*bottom right*) (see **Methods**). Paths including the node *N* are bolded. mle, maximum likelihood estimate.

Figure 2 – Whippet benchmarking against event-level and transcript-level approaches

(A) Cumulative distribution plot comparing percent spliced in (Ψ) values from RT-PCR data with Ψ values quantified from RNA-seq data. RT-PCR and RNA-seq data were generated from the same samples of mouse liver and cerebellum, as well as from stimulated and unstimulated human Jurkat T-cell line samples (Vaquero-Garcia et al., 2016). By default, all benchmarked programs were supplied with the full Ensembl

GRCh37.73 annotation file, unless indicated otherwise (see **Table S4**). Cumulative distribution plots describe the proportion of data (*y-axis*) less than or equal to a specified value (*x-axis*). Dotted y-axis lines mark the lower quartile, median, and upper quartile (25%, 50%, 75%) values respectively. Cumulative Freq F(x), cumulative distribution function.

(B) Bar plots showing the absolute error rate of quantification algorithm Ψ values compared to simulated 'ground truth' (i.e. known) Ψ values. Errors bars represent the standard error of the mean. RSEM, RNA-seq by Expectation Maximization (Li and Dewey, 2011).

(C) Bar graph displaying the fraction of unannotated junctions (with two or more supporting reads) as a total of all junctions identified by Whippet across six vertebrate species (Brawand et al., 2011). "Error bars" represent the y-axis value range for a cumulative number of tissues, one (lower bound of the error bar) to five (height of the bar). Source of annotation (left to right): panTro4 Ensembl; monDom5 Ensembl; galGal4 Ensembl; gorGor3 Ensembl; hg19 GENCODE v27 tsl1; mm10 GENCODE VM11 Basic

(D) Formalization of AS complexity into discrete categories K(n). n, theoretical number of alternative nodes and $K(n) = 2^n$ spliced-outcomes for a given AS event. Schematics of K(n) show constitutive exons (dark grey) and alternative exons (light gray) with curved lines representing all potential exon-exon junctions.

(E) Cumulative distribution of entropy scores (i.e. entropy = $-\sum_i \Psi_i \, log_2 \Psi_i$) detected by Whippet for simulated AS events of different categories of complexity according to (D). See **Figure 2a** legend for a description of cumulative distribution plots.

(F) Comparison of the ability of different RNA-seq analysis methods to detect AS events from simulated reads (**Methods**) of complexity as defined in (D). Bar plots

show the absolute mean error rate as a function of increasing complexity of AS. Error bars indicate standard error. Ψ, Percent Spliced In.

(G) RT-PCR analysis confirms the numerous splice isoforms in N2a cells for AS events of increasing levels of complexity and matching Whippet predictions for the maximal complexity and entropy (*far right*). Boxes to right of gels display UCSC (*left*, *blue*) and Whippet (*right*, *yellow*) *in silico* predictions based on expected primer amplification products (**Methods**). Colored boxes (*blue and yellow*), correct predictions; black boxes, possible missed predictions. Diagrams below show exon structures of analyzed AS events with approximate positions of RT-PCR primers. Predicted constitutive and alternative exons are in dark and light gray, respectively (see legend in panel D).

(H) Comparison of the log-scaled "core" time requirements (i.e. taking into account time spent using multiple cores) for running Whippet relative to published methods for RNA-seq splicing quantification when analyzing 15M, 25M or 50M paired-end RNA-seq read datasets (see **Methods and Table S3**).

Figure 3 – Tissue-regulation of high entropy events detected using Whippet.

(A) Symmetrical heatmap of pairwise correlations of normalized splicing entropy scores across multiple human tissues. Heatmap showing affinity propagation clustering of pairwise similarities between entropy scores. Colored bars surrounding heatmap indicate clusters defined by the dendrogram. Darker blue, stronger correlation in splicing entropy; lighter blue, weak or no correlation. r1, replicate 1; r2, replicate 2

(B) Plot of ranked genes (*x*-axis) ordered by their maximum minor / major isoform relative expression ratio across all tissues (*y*-axis) at different minimum cut-offs

(*color*-scale), for the number of reads mapping to exon-exon junctions corresponding to the AS event. Dashed line, 45:55% ratio cutoff (equivalent to a minor / major ratio of 0.818; see **Methods**).

(C) Bar plot displaying maximum variance of splicing entropy per gene (n = 11,421), revealing that >20% of genes exhibit extensive variance in AS entropy of AS across human tissues. Genes lacking major changes in entropy are not shown.

(D) Scatter plots of change in AS entropy across tissues versus change in expression level of the corresponding genes. Red line, best-fit linear regression. R-squared value calculated using Pearson Correlation Coefficient.

(E) Functional analysis for GO, REACTOME and KEGG functional categories of genes with large changes in splicing entropy (>2.0) across human tissues. P-value, corrected FDR hypergeometric test.

Figure 4 – Alternative splicing entropy is evolutionarily conserved and high entropy events are potentially translated

(A) Distribution of the number of unique conserved exons with genomic coordinate 'liftover' across at least three vertebrate species (human, chimp, gorilla, mouse, opossum, platypus, and chicken). Conserved exons are counted in discrete bins by their maximum entropy in any of the species.

(B) Cumulative distribution plots comparing the cross-species variance of entropy values among the same tissue in seven vertebrates (at least three species present per-event) as compared to a permuted null control. See **Figure 2A** legend for a description of cumulative distribution plots.

(C) Distributions for the cross-species variance of entropy values (*y*-axis) for conserved exons, binned by maximal entropy values (*x*-axis), and compared to a

control set of the same data but with permuted AS event labels for each species (*color-scale*). All two-sided KS-test p-values are less than epsilon (2.2×10^{-16}), except for the bin (1.5,3] whose p-value was 4.6×10^{-4}. (*bottom*) Same as (*top*) except the distributions plotted contain the cross-species variance of Ψ-values (*y*-axis) for the same conserved exons. All two-sided KS-test p-values are less than epsilon (2.2×10^{-16}), except for the bin [1.5,3] whose p-value was 4.3×10^{-2}. Boxplots display the interquartile range as a solid box, 1.5 times the interquartile range as vertical thin lines, the median as a horizontal line, and the confidence interval around the median as a notch.

(D) Violin plots of the distribution of splicing entropy in different cellular compartments and ribosome (mono- and polysome) fractions. Kernel density is displayed as a symmetric curve, with white dots indicating the median, black box the interquartile range, and black lines the 95% confidence interval.

Figure 5 – High entropy splicing events encode unique protein features

(**A**) Cumulative distribution plots showing frequency of overlap of AS events (with different degrees of entropy) within intrinsically disordered regions (IDRs) of proteins (*left*) structured single protein domains (*center*), and structured tandemly repeated protein domains (*right*). See **Figure 2A** legend for a description of the cumulative distribution plots ($n > 368$).

(B) Bar plot showing frequency at which exons undergoing AS with different degrees of entropy (based on Whippet analysis of tissue RNA-seq data in **Figure 3**) show evidence of duplication. Numbers of AS events analyzed indicated above plots. See **Figure 5A** for color legend.

(C) Domain diagram for Lrp8 (Low-density lipoprotein receptor-related protein 8)

based on SMART annotation. Dotted boxes describe area of proteins undergoing high entropy splicing in different tissues types. Domain diagram below illustrates exons undergoing splicing within N2a cells and position of primers for RT-PCR validation below. CNS, Central Nervous System; E, embryonic day; LDL, Low-Density Lipoprotein; EGF, Epidermal Growth Factor

(D) RT-PCR analysis confirms the presence of putative Lrp8 spliced isoforms in N2a cells. Diagrams below show exon structures of analyzed AS events with approximate positions of RT-PCR primers indicated. See **Figure S5** for full gel.

Figure 6 – Exons within high entropy splicing events have unique splice site features

(A) Plots showing the cumulative distribution of 3′-splice site (3′ss) strength (*left*) and 5′-splice site (5′ss) strength (*right*) estimated using MaxEntScan (Yeo and Burge, 2004) and binned by maximum splicing entropy scores (*bottom panels*). The median 3′ss strength for AS events with different degrees of splicing entropy are plotted as colored lines (*top panels*). See **Figure 2A** legend for a description of cumulative distribution plots (n >1064).

(B) Boxplot displaying the distribution of exon length (*top*) and intron length (*bottom*) surrounding exons binned by maximum entropy of AS. See **Figure 6A** for color legend. nt, nucleotide. n as in **Figure 6A.** See **Figure 4C** for descriptions of boxplots.

(C) Cumulative distribution plots of exonic splicing regulatory elements in AS events with different degrees of AS event entropy. Scores calculated based on the density of exonic splicing enhancers (*top*) and exonic splicing silencers (*bottom*) per nucleotide (**see Methods)**. Motifs extracted from Ke et al. 2011. See **Figure 6A** for color legend, and **Figure 2A** legend for a description of cumulative distribution plots. n as in

Figure 6A.

(D) Mechanistic model for the regulation of low-entropy (simple binary) AS events versus high-entropy (complex) AS events by *cis*-acting elements and other sequence features. Exons are represented by boxes and introns by lines, with *cis*-regulatory elements and relative splice site strength indicated by color.

Figure 7 – Increases in high entropy splicing in cancer is associated with over-expression of the essential splicing factor, SRSF1

(A) Boxplot showing percentage of high entropy AS events (> 1.5) within each replicate identified from Whippet analysis of RNA-seq data comprising 15 matched tumor and control samples. Black dots represent individual data. See **Figure 4C** for descriptions of boxplots.

(B) Cumulative proportion of unannotated alternative splice junctions (with two or more supporting reads) identified across matched tumor and control RNA-seq samples. See **Figure 2A** legend for description of cumulative distribution plots.

(C) Heatmap of splicing entropy values for events with significant changes ($p<0.05$, Mann-Whitney U test) between tumor and control samples (n=657).

(D) Bar plots of enriched functional categories for genes harboring AS events with significant entropy changes (p-values < 0.05, Mann-Whitney U test) from panel (**C**) identified from RNA-seq analysis of 15 matched tumor and control samples. P-values were corrected using false discovery rate (FDR) multiple hypothesis testing correction (n = 657).

(E) Schematic diagrams of two genes showing significant changes in AS event entropy between tumor and matched control samples. Domain structure extracted from SMART database. Light blue arrows and boxes indicate increased occurrence of

splicing regulation in tumor samples. For BIN1, dashed boxes indicate protein regions predicted to be regulated by splicing in control (grey box) and cancer samples (cyan box). EZH2, Histone-lysine N-methyltransferase EZH2; BIN1, Myc box-dependent-interacting protein 1

(F) Differential gene expression analysis for selected RNA-binding proteins (GO:0000380) identified from RNA-seq analysis of 15 matched tumor and control samples. Genes with blue bars display reduced expression in cancer samples, red bars show increased expression in cancer samples, and gray bars show no significant difference between control and tumor samples.

(G) Boxplot showing normalized read counts for SRSF1. See **Figure 4C** for descriptions of boxplots.

(H) Boxplot showing relative complexity of transcriptomes, as measured by distribution of entropy scores for high quality AS events, between SRSF1 over-expression (OE) sample and matched control (n=1,998). Statistical test = Mann-Whitney U test. See **Figure 4C** for descriptions of boxplots

(I) Bar plot showing percentage of events from plot (H) with differential splicing changes between SRSF1 OE (over-expression) and matched control samples that overlap with splicing changes in tumor samples from (C), as compared to the number of overlapping events expected by chance (n = 1,998). Statistical test = Fisher's exact test.

STAR METHODS:

CONTACT FOR REAGENT AND RESOURCE SHARING

Requests should be directed to and will be fulfilled by Lead Contact Benjamin

Blencowe (b.blencowe@utoronto.ca).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines and Cell Culture

Neuro-2A (N2A) cells are a male, mouse neuroblastoma cell line, and were grown in DMEM supplemented with 10% FBS, sodium pyruvate, non-essential amino acids and penicillin/streptomycin. Cells were maintained at 37°C with 5% CO_2. An authenticated N2A cell line was purchased from ATCC (catalog number: ATCC CCL-131).

Short interfering RNA knockdown and RT-PCR

Mouse Neuro2A (N2A) cells were transfected with SMARTpool siRNAs (Dharmacon) (50nM final concentration) using Lipofectamine RNAiMAX (Invitrogen), as recommended by the manufacturer. A non-targeting siRNA pool (siNT) was used as a control. Cells were harvested at 48 hours post transfection and total RNA was extracted using RNeasy columns (QIAGEN). Semi-quantitative RT-PCR was performed using the QIAGEN One-Step RT-PCR kit as per the manufacturer's instructions, using 50ng total RNA in a 20uL reaction. Products were resolved on 2-4% agarose gels and bands were quantified using Image Lab (BioRad) or ImageJ. Predictions of band sizes were based on *in silico* PCR using data from from the UCSC Genome Browser (http://genome.ucsc.edu) server after combining exons from Whippet predictions. Only predictions supported by multiple sources of evidence (i.e. RT-PCR, Whippet and UCSC) were included in figures (see **Key Resources Table** for details of primers used).

METHOD DETAILS

RNA-seq simulation

To simulate RNA-seq reads transcriptome wide, we used RSEM (Li and Dewey, 2011) to quantify the benchmark dataset SRR2300536 (a ~25M read depth RNA-seq dataset from HeLa cell line). With the RSEM parameters and gene expression distributions obtained from this quantification (RSEM *estimated_model_file*, *estimated_isoform_results*, and *theta*), we used RSEM's *rsem-simulate-reads* to simulate 50M paired-end reads for each of two hg19 annotation builds: Gencode v25 TSL1, and RefSeq Release 84. In order to calculate 'ground truth' (i.e. known) Ψ values for Whippet nodes, we used the Whippet_TPM method on the ground truth isoform TPM values provided by the RSEM simulator.

To investigate the accuracy and capability of AS quantification tools, we simulated transcripts with AS-events of increasing complexity. To formalize AS events into discrete classes of complexity K(n) = 2^n splicing-outcomes for K1 through K6, we randomly chose 500 CSGs of each complexity class with at least n total internal nodes (not including nodes with TxStart or TxEnd node boundaries). From those CSGs, we randomly chose a set of n consecutive internal nodes and created partial transcript sequences from the first internal node to the last internal node, with all combinations of n internal nodes. In the case of nodes with Soft boundary types, less than 2^n total combinations were created, since nodes whose incoming edge is a Soft 5' Splice Site cannot be included in the transcript unless the adjacent upstream node is also included. Similarly, a node whose outgoing edge is a Soft 3´SpliceSite requires the adjacent downstream node to be included. Given the six sets of simulated events of complexity K(n) (where n = 1, ... ,6), we used polyester (Frazee et al., 2015)

(read length = 100, error rate = 0) to simulate RNA-seq reads from the simulated transcripts for each gene (see Methods S1 for extended details).

Combinatorial gene model

To investigate engineering *de novo* AS analysis capability for transcript-level methods, we utilized Whippet's CSGs (in the Whippet/bin/simulation/*whippet-combinatorial.jl* script) to enumerate combinatorial graph paths for each pair of TxStart and TxEnd boundaries. While we successfully simulated combinatorial paths for a sliding window of four, five, six, eight, and ten nodes, we used four nodes throughout the manuscript (referred to as the 'N4 annotation Gene Transfer Format [GTF]'). This was the largest number of nodes in a sliding window for which, due to memory usage issues, we were able to successfully build indices using transcript-level methods.

Benchmarking

All genomic and transcriptomic sequences, as well as GTF files, were downloaded from the Ensembl database. The following genome builds were used: Hg19 GRCh37.p12 (v73) and Mm10 GRCm38.p4 (v84) using the full Ensembl GRCh37.73 annotations for all programs unless otherwise stated in the analysis or in the online instruction manual for that program (e.g. **Figure 2A** uses the full Ensembl annotation sets by default, while **Figure 2B** restricts each program to GENCODE v25 TSL1 or RefSeq Release 84 as specifically stated; see **Table S4**). Exon annotations (including genomic annotations) were downloaded from Ensembl using BioMart.

All benchmarking was performed on a Sun Microsystem X4600M2 server with 8 AMD Dual-Core 8218 CPU @2.6GHz, total 16 cores and 64GB RAM. The

local hard disk was SATA 73GB, 10K RPM. Identical paired-end HeLa data of increasing read-depths were employed for all resource usage benchmarking (see **Table S3**). All programs were run with default settings with additional settings described in **Table S4**. The default linux package "time" (/usr/bin/time – e.g. http://man7.org/linux/man-pages/man1/time.1.html) was used to measure the resource usage of each program. See Methods S1 for extended details, and **Tables S2** and **S5** for results.

Benchmarking of mapping success was performed using the program Benchmarker for Evaluating the Effectiveness of RNA-Seq Software (BEERS) (http://www.cbil.upenn.edu/BEERS/) and simulated reads based on hg19 GRCh37.73 Ensembl transcriptome data. Simulated reads were generated using "reads_simulator.pl" with substitution frequency (parameter "-subfreq") error rates of 0.001, 0.005 and 0.01, respectively and a read depth of 1,000,000. For resource and mapping benchmarks the program "time" was used (see above and **Methods S1** for details, and **Table S6** for results).

RT-PCR and RNA-seq data used in comparisons of Ψ values were generated from samples prepared from mouse cerebellum and liver tissue, as well as from stimulated and unstimulated human Jurkat T-cell line cells (Vaquero-Garcia et al., 2016). ΔΨ values were calculated by comparing Ψ values between the mouse cerebellum and liver tissues samples or between the stimulated and unstimulated human Jurkat T-cells. Only simple events (as defined by MAJIQ as involving a total of three exon-exon junctions) were included in the analysis.

Tissue-wide analysis of splicing

Low-entropy AS events are defined by an entropy value less than 1.0. High entropy events (for description of entropy of AS events see **Figure 2D**, **Figure S4C-D** and **Methods S1**) are defined as events with an entropy score of greater than 1.5, and differential entropy requires a change of entropy of greater than 1.0 (unless stated). Highest entropy events are defined as those greater than 2.0. Only events with a Whippet confidence interval width of less than 0.2, and Ψ values of over 0.05 and under 0.95 were included in the analyses. Analyses were limited to core exons (CE), as defined by Whippet. An exception to this rule is when assessing the fraction of genes co-expressing two or more major isoforms. For this analysis, due to observation in **Figure S6B**, we used a minimum read cut-off of 20 in main text (see **Figure 3B** and **S6C** for additional cut-offs).

Tissue RNA-seq data analyzed in **Figure 3 and Figure S6** were from the Illumina Bodymap2 dataset and supplemented with human tissue RNA-Seq data from Kunming Institute of Zoology (**Table S3**). The maximum change in splicing entropy between tissues is the comparison of the lowest entropy of an exon/node compared to the highest entropy for same exon/node between tissues. This is therefore not a measure of tissue-specificity but rather a measure of maximum variability for the number of well-expressed exon-exon junctions an exon may have across tissues.

The analysis of how many genes co-express at least two isoforms at similar levels was calculated using the above tissue specific data. For an event to be considered as co-expressed the two principal isoforms must be expressed at similar levels (within a 10% range). Expression was assessed based on assigned reads. All types of splicing events were considered.

Tissue-wide heatmaps were generated by affinity propagation clustering using the R package (apcluster) with pairwise similarities as correlations (corSimMat and r=2) and negative correlations taken into account.

Feature analysis of high entropy AS events

For all amino acid residues in a protein, a score for predicted intrinsic disorder is computed using IUPred (Dosztanyi et al., 2005). Amino acid residues with a score larger than 0.4 were considered as disordered. For each coding exon the proportion of total residues that are predicted to be disordered was estimated. Domain data extracted from SMART database (Letunic et al., 2015).

MaxEntScan (Yeo and Burge, 2004) was used to estimate the strength of 3′ and 5′ splice sites. 5′ splice site strength was assessed using a sequence including 3nt of the exon and 6nt of the adjacent intron. 3′ splice site strength was assessed using a sequence including -20nt of the flanking intron and 3nt of the exon. Exonic splicing silencer or exonic splicing enhancer densities were extracted from motifs quantified in (Ke et al., 2011). To calculate exonic splicing enhancer and silencer densities, all motifs defined by Ke et al. were summed together and normalized by the number of exonic nucleotides.

Analysis of cancer data

Hepatocellular carcinoma (HCC) and control data were from a transcriptome profiling study undertaken by the University of Hong Kong (see **Table S3**). For **Figure 7A**, all events with sufficient reads (n>10) across multiple samples (more than 2) that showed evidence of AS ($0.05< \Psi<0.95$) were included in the analysis. These criteria were used throughout **Figure 7**, with the exception of **Figure 7B**, when all

exons required at least 3 reads to support identification. For **Figure 7B**, unannotated alternative exon-exon junctions were extracted from the Whippet '.jnc' file.

Differential complexity between control and tumour samples across 15 replicates described in **Figure 7C** was assessed. Only samples with a significant difference (Mann-Whitney U test $p < 0.01$) and a median entropy difference between control and tumour samples of at least 0.5 were considered differential. To identify differentially expressed genes, read counts for transcripts (calculated by Whippet) were combined and DESeq2 (adjusted p-value < 0.05) was used. SRSF1 over-expression data (Anczukow et al., 2015) was analyzed by Whippet. Only events with high entropy (> 1.5) in either the control or over-expression study were included in the analysis. Events with detected aberrant splicing in **Figure 7I** are displayed in **Figure 7C**.

QUANTIFICATION AND STATISTICAL ANALYSIS

Contiguous Splice Graph Index

The central data structure underlying the alignment and quantification capabilities of *Whippet* is the Contiguous Splice Graph (CSG). This directed acyclic (i.e. except when circular splicing detection is enabled) graph structure is composed of all non-overlapping exon intervals, which are each defined as separate 'nodes'. Nodes in the CSG are connected by edges, defined as either splice junctions or adjacent exonic regions. All nodes are arranged consecutively in a single sequence based on genomic coordinates (see **Algorithm S1**). As such, a CSG sequence built from a set of annotated transcripts may not necessarily resemble any of the individual transcript sequences. Each transcript sequence can however be defined by a sequential

series of nodes through the graph. Whippet defines node boundaries (one upstream and one downstream, flanking either side of the node sequence) to describe the incoming and outgoing connectivity to other nodes. Whether an edge can exist between two nodes is defined by their incoming and outgoing 'boundary-types'. Node boundary-types are formally made up of two properties: a classification and an alignment property. The classification property can be a transcription start (TxStart), transcription end (TxEnd), donor splice site (5′SpliceSite), or acceptor splice site (3′SpliceSite) (**Figure S1B and Table S7**). The alignment property is one of two categories: 'Soft' or 'Hard'. Soft boundaries are node boundaries adjacent to other nodes in the genomic sequence. For example, in **Figure 1B**, nodes 3 and 4 have Soft outgoing and incoming edges, respectively. This is because in an annotated transcript they are part of the same exon (i.e. zero nucleotides exist between the end position of node 3 and the start position of node 4 in the genomic sequence). In contrast, Hard boundaries exist when one or more genomic nucleotides separate the nodes. For example, there is a Hard boundary between nodes 2 and 3 in **Figure 1B** because genomic sequence separates the nodes. The compatibility of two boundary-types is determined by three simple rules: (1) All outgoing 5′SpliceSite boundaries are compatible with all incoming 3′SpliceSite boundaries, (2) Soft boundaries are compatible with adjacent neighboring Soft boundaries, and (3) no Hard boundary is compatible with any other boundary except in the case of Rule #1 (Methods S1 for extended details). This distinction between CSG Hard and Soft boundaries allows boundary type-specific rules to be utilized for alignment extension. After building all CSGs, the CSG Sequences are concatenated into a single Multi-CSG sequence that is used to create a transcriptome Full-text index in Minute space (CSG FM-Index) (Ferragina et al., 2004) for full-text substring searches.

Whippet aligns RNA-seq reads to the CSG index by performing ungapped extensions from alignment seed sequences mapped to the CSG FM-Index (Methods S1 and **Algorithm S2** for details). Using the CSG index, Whippet is able to efficiently align spliced reads to any combination of nodes in a CSG. To facilitate this, reads are aligned across spliced edges using nucleotide k-mers flanking annotated 5′ or 3′ splice-site node boundaries. Each 5′ or 3′ splice-site flanking k-mer indexes each of two global hash-tables (i.e. associative maps) that link to a list of (gene, node) tuples, respectively (**Figure 1D**). Spliced read alignment uses read k-mers at an alignment node boundary to match compatible nodes from the same gene (note all nodes with outgoing 5′ splice sites are compatible with all nodes with incoming 3′ splice sites) (**Figure 1D**, **Figure S1**; see Methods S1 for extended details). Read alignment in this manner affords considerable efficiency by storing minimal data while supporting *de novo* AS event identification.

AS event definition and PSI quantification

After all reads have been assigned full or partial (for multi-mapping reads) counts to the edges in a CSG (see Methods S1 for details of isoform-level quantification and multi-mapping read assignment), AS events are next built *de novo* to quantify AS. In order to define an AS event for a node, the set of edges connecting to – and skipping over the target node (*N*) – are collected, where the read count of a skipping edge must be ≥ 1% of the maximal connecting edge read count. The AS event built *de novo* for each node (referred to here as the 'target node' of the event) is initially defined by the span of the edges that directly connect or skip the target node. Whippet iteratively collects all edges that fall within the span of previously defined directly connecting or skipping edges (**Figure 1E**). Whippet then performs the same

procedure for each non-target node within the AS event, extending the AS event as necessary to encompass all auxiliary edges, including edges for non-target nodes that do not directly skip or connect to the target node (**Figure 1E**). The set of paths through the AS event are then enumerated using **Algorithm S3** (see Methods S1).

In order to quantify the AS event paths $i \in I$, we utilize the set of edges E in the event and the read count c_e assigned to each edge $e \in E$. Counts for each unique edge e that exist in only one path i are assigned fully. However, non-unique edges found in multiple paths have counts initially divided among their compatible paths with uniform probability, and then the maximum likelihood for the relative expression of each AS event path is estimated using the expectation-maximization (EM) algorithm. We define a compatibility matrix $\mathbf{y}_{e,i} = 1$ for an edge e existing in a path i, and $\mathbf{y}_{e,i} = 0$ otherwise (Bray et al., 2016). We define the length of path i as proportional to the number of edges in the path such that: $j_i \propto \sum_{e \in E} \boldsymbol{y}_{e,i}$ (see Methods S1 for extended details). The probability α of observing reads from an AS event path i with relative expression level ψ_i is then defined by $\alpha(i) = \frac{\psi_i \, j_i}{\sum_{p \in I} \psi_p j_p}$. The following likelihood function is therefore iteratively optimized in the EM algorithm:

$$\mathcal{L}(\alpha) \propto \prod_{e \in E} \left(\sum_{i \in I} \boldsymbol{y}_{e,i} \, \frac{\alpha(i)}{j_i} \right)^{c_e}$$

In the M-step, the relative expression of each path (ψ_i) is given by:

$$\psi_i = \frac{\sum_{e \in E} \alpha(e,i) \, c_e}{j_i}$$

In the E-step, the probability α of observing reads from an edge e and path i are:

$$\alpha(e,i) = \frac{\boldsymbol{y}_{e,i} \, \psi_i}{\sum_{p \in I} \boldsymbol{y}_{e,p} \, \psi_p}$$

The percent-spliced-in Ψ of the node *n* is then calculated as the sum of the normalized relative expression of the paths containing the node $\{I_n \subset I\}$:

$\Psi_n = \sum_{i \in I_n} \hat{\psi}_i$, where $\hat{\psi}_i = \frac{\psi_i}{\sum_{p \in I} \psi_p}$.

It's important to note that the this represents a generative model for RNA-seq count data, assuming that counts from each edge are drawn independently from a multinomial distribution. While this assumption will not always be satisfied (e.g. for reads that span multiple edges), assuming independence among edges simplifies the problem space considerably and in turn does not adversely affect the accuracy of the quantifications.

Whippet_TPM

To calculate PSI values for Whippet nodes from the Transcript Per Million (TPM) values calculated by transcript-level analysis tools such as Kallisto/Salmon (Bray et al., 2016; Patro et al., 2017) (in the Whippet/bin/simulation/*whippet-quant-bytpm.jl* script, a.k.a 'Whippet_TPM'), we utilize the quantification concepts described for SUPPA (Trincado et al. 2018). Briefly, $\Psi_n = \frac{\sum_{i \in I_n} \tau_i}{\sum_{i \in I} \tau_i}$, where *n* is the node being quantified, *I* is the set of transcripts in the gene, I_n is the set of transcripts containing node *n* in the gene, and τ_i is the TPM of transcript *i*. To simplify this script, only nodes guaranteed to be quantified correctly are used, i.e. Whippet_TPM only quantifies nodes with 3´SpliceSite incoming and 5´SpliceSite outgoing boundary types.

Statistical Analysis

Gene function enrichment analysis (**Figures 3b and 4d**) was performed using g:Profiler (with the python package: gprofiler; http://biit.cs.ut.ee/gprofiler), which uses a hypergeometric test with multiple hypothesis testing correction, as originally described by Benjamini and Hochberg. Mann-Whitney U non-parametric statistical tests were used for comparing distributions (R query: Wilcox.test <default parameters>) in **Figures 4A, 6B, 6C, 7A, 7C, 7H** and, **Figure S7**. An exception was in **Figure 2A** and **Table S1** when analyzing repeated measurements (e.g. in RT-PCR comparisons), in which case the Wilcoxon signed rank test was used (R query: Wilcox.test – signed=T). Kolmogorov–Smirnov (KS) tests were used in Figure S9. Fisher's exact test (R query: fisher.test) was used for comparing two nominal variables in a small population in **Figure 7I** and **Figure S6**. DESeq2 (Love et al., 2014) tested for differential gene expression using negative binomial generalized linear models with a multiple hypothesis testing correction, as originally described by Benjamini and Hochberg. The adjusted p-value cut-off was 0.05. Heatmaps were generated using Affinity Propagation clustering with the R package "apcluster". Clustering was based on either pairwise similarities of correlations (Pearson), or mutual pairwise similarities of data vectors, measured as the negative Euclidean distance. Correlations were assessed using Pearson Correlation Coefficient.

Additional Resources

Further benchmarking and methods details are described in Methods S1. Protocol is available at http://github.com/timbitz/Whippet.jl

DATA AND SOFTWARE AVAILABILITY

Whippet is implemented in the high-level, high-performance dynamic programming language Julia (julialang.org) and is freely available as open-source software under the MIT license (Git repository: http://github.com/timbitz/Whippet.jl). The analysis scripts and simulated data used in this study are available at http://figshare.com/articles/Whippet_analysis_scripts/5711683

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Lipofectamine RNAiMAX	Invitrogen	Cat# 13778030
SMARTpool siRNAs	Dharmacon	N/A
Critical Commercial Assays		
One-Step RT-PCR	Qiagen	Cat# 210210
RNeasy Mini Kit	Qiagen	Cat# 74104
Experimental Models: Cell Lines		
Human: HeLa	N/A	N/A
Mouse: Neuro2A	ATCC	ATCC CCL-131
Deposited Data		
Public RNA-seq data used in paper	Table S3	Table S3
Oligonucleotides		
Slmap: Forward:GAGCGCACTCAGGAAGAGTT Reverse: TTCCTTTGCTTTTGCCTGAT	This paper	N/A
Slmap (Control): Forward:GAGCGCACTCAGGAAGAGTT Reverse:TTCCTGCTCAGTCATTTCAAAC	This paper	N/A
Eps15l1: Forward:TTGGAACCCTAGACCCCTTT Reverse:CTTTTTCACTCTCCCGCTTG	This paper	N/A
Asap1: Forward:GCCCGCGATGGAATAATG Reverse:TGAGGAAGAGGCACAGGTCT	This paper	N/A
Eml4: Forward:TCCTGTATAACCAATGGAAGTG Reverse:CATTGTAATTGGCCGACCTC	This paper	N/A
Atp8a1: Forward:CGGTCGTTACACAACACTGG Reverse:GGCCAAGTTCCTCATTCAGA	This paper	N/A
Sfl1: Forward:TCATGCCTCACAAAACTGGA Reverse:CCATAGCCAGCCTCTGTACC	This paper	N/A
Mapt: Forward:AATGGAAGACCATGCTGGAG Reverse:GCCACACTTGGAGGTCACTT	This paper	N/A

Lrp8: Forward:CGGAGAGAAGGACTGTGAGG Reverse:CAGTGCAGATGTGGGAACAG	This paper	N/A
Gtf2ird1: Forward:CCCCAACACCTATGACATCC Reverse:CGCTTGGGAATGTTGTCTTT	This paper	N/A
Rbms3: Forward:GAGACAGGGTCAGAGCAAGC Reverse:AAACCGGAGGCCAACTAACT	This paper	N/A
Cask: Forward:AGGGAAATGCGAGGGAGTAT Reverse:GTCATCCTTGGCTGGATCAT	This paper	N/A

Software and Algorithms		
Whippet	This paper	https://github.com/timbitz/Whippet.jl
Whippet_TPM	This paper	https://github.com/timbitz/Whippet.jl
Supplemental scripts and simulated data	This paper	http://figshare.com/articles/Whippet_analysis_scripts/5711683
Julia	N/A	http://www.julialang.org
BioJulia	N/A	https://github.com/BioJulia
MAJIQ	(Vaquero-Garcia et al., 2016)	https://majiq.biociphers.org/
rMATS	(Wang et al., 2017)	http://rnaseq-mats.sourceforge.net/
MISO	(Katz et al., 2010)	http://genes.mit.edu/burgelab/miso/
VAST-TOOLS	(Tapial et al., 2017)	https://github.com/vastgroup/vast-tools
BENTO	(Xiong et al., 2016)	https://github.com/PSI-Lab/BENTO-Seq
SUPPA	(Trincado et al., 2018)	https://github.com/comprna/SUPPA
Kallisto	(Bray et al., 2016)	https://pachterlab.github.io/kallisto/
STAR	(Dobin et al., 2013)	https://github.com/alexdobin/STAR
HISAT	(Kim et al., 2015)	https://ccb.jhu.edu/software/hisat
TOPHAT	(Kim et al., 2013)	http://ccb.jhu.edu/software/tophat
BEERS	(Grant et al., 2011)	http://cbil.upenn.edu/BEERS/
Polyester	(Frazee et al., 2015)	https://github.com/alyssafrazee/polyester
RSEM	(Li and Dewey, 2011)	https://github.com/deweylab/RSEM
DESeq2	(Love et al., 2014)	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
IUPred	(Dosztanyi et al., 2005)	http://iupred.enzim.hu/
MaxEntScan	(Yeo and Burge, 2004)	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html
apcluster	(Bodenhofer et al., 2011)	https://cran.r-project.org/web/packages/apcluster/index.html

PTRStalker	(Pellegrini et al., 2012)	http://bioalgo.iit.cnr.it/index.php?pg=ptrs
SEG	(Wootton, 1994)	http://www.biology.wustl.edu/gcg/seg.html
Image Lab	BioRad	Cat# 1709691

Other		
Parameters for software used	Table S7	Table S7
Supplemental Methods	Methods S1	Methods S1
Additional Benchmarks	Methods S1	Methods S1

Figure 1

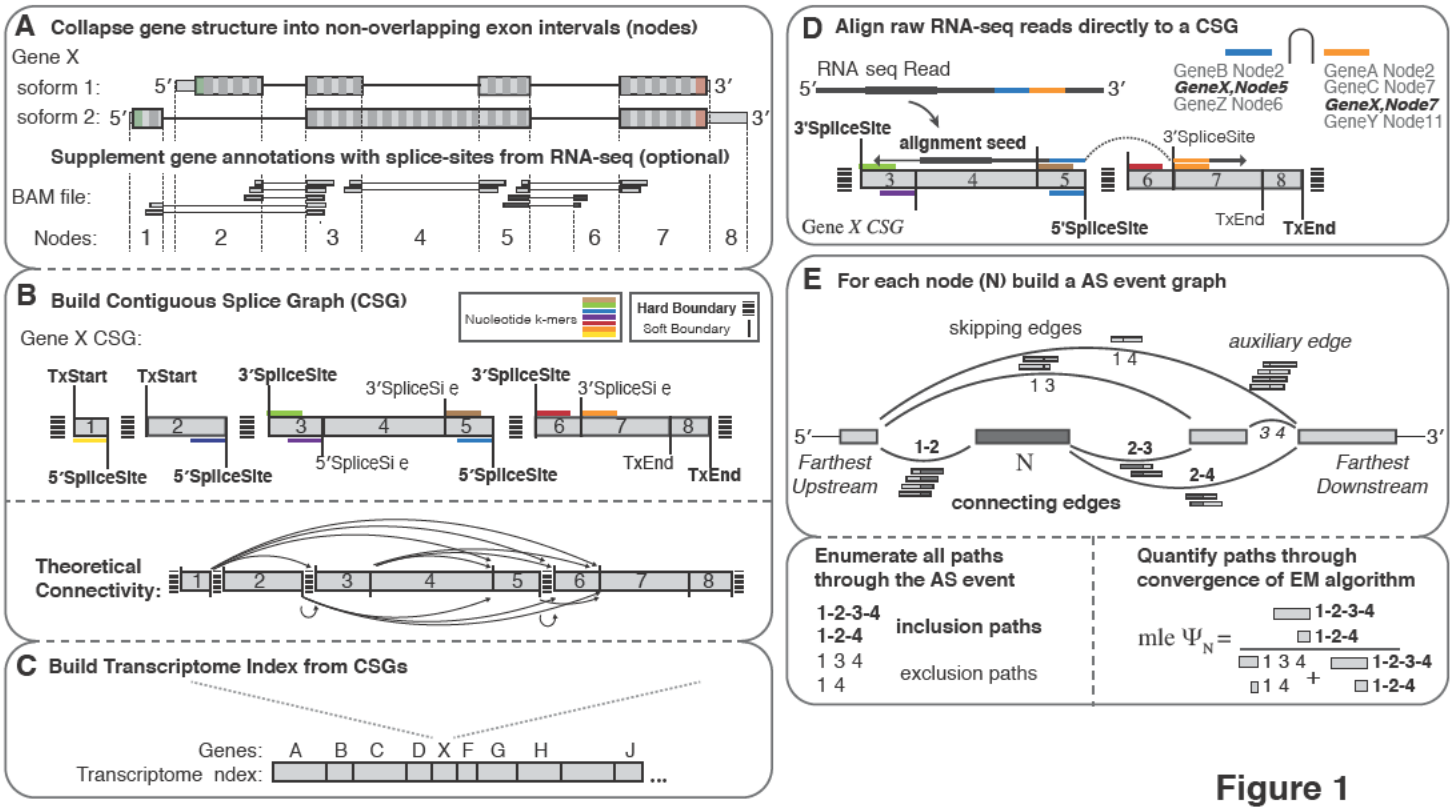

Figure 1

Figure 2

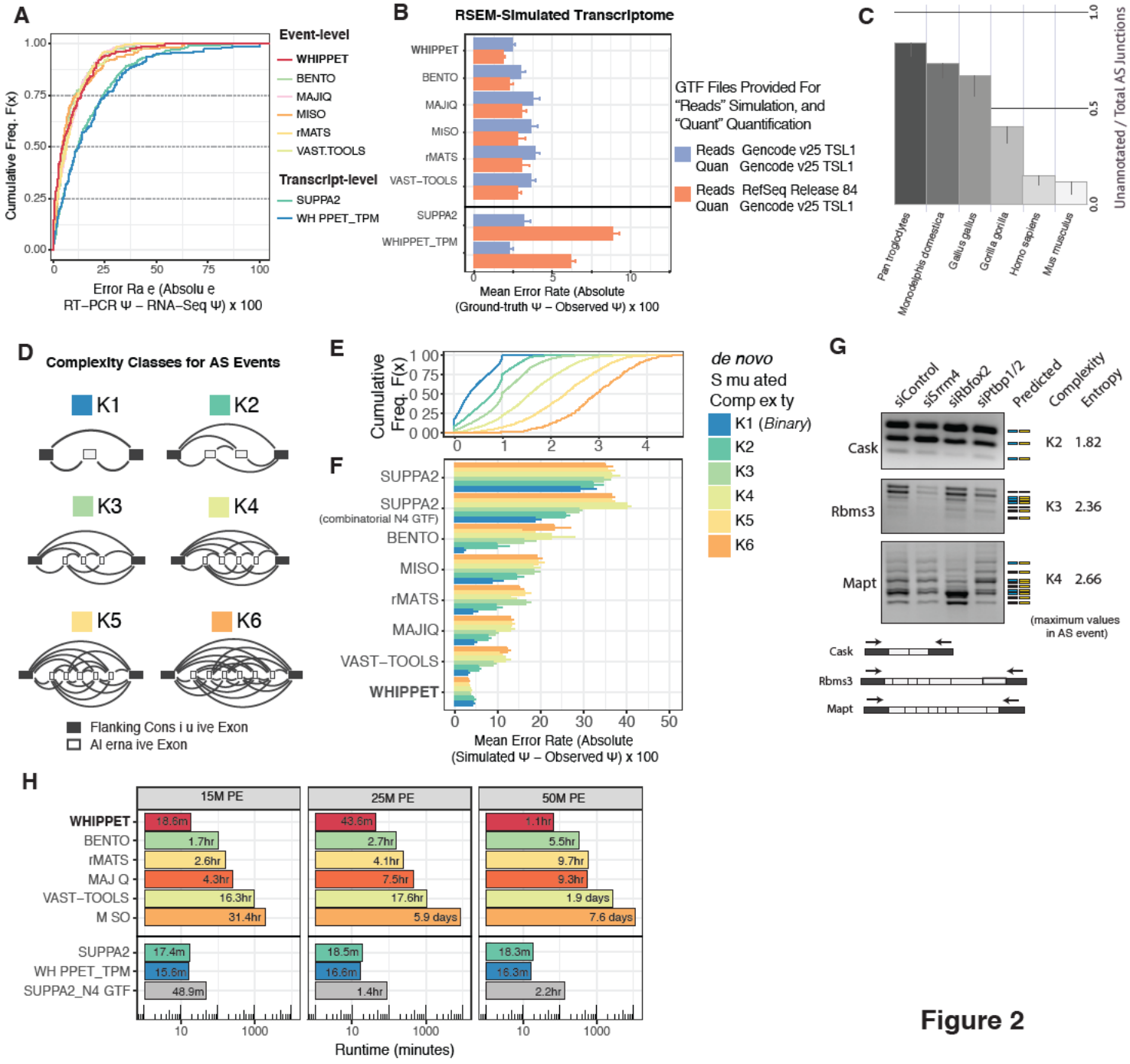

Figure 2

Figure 3

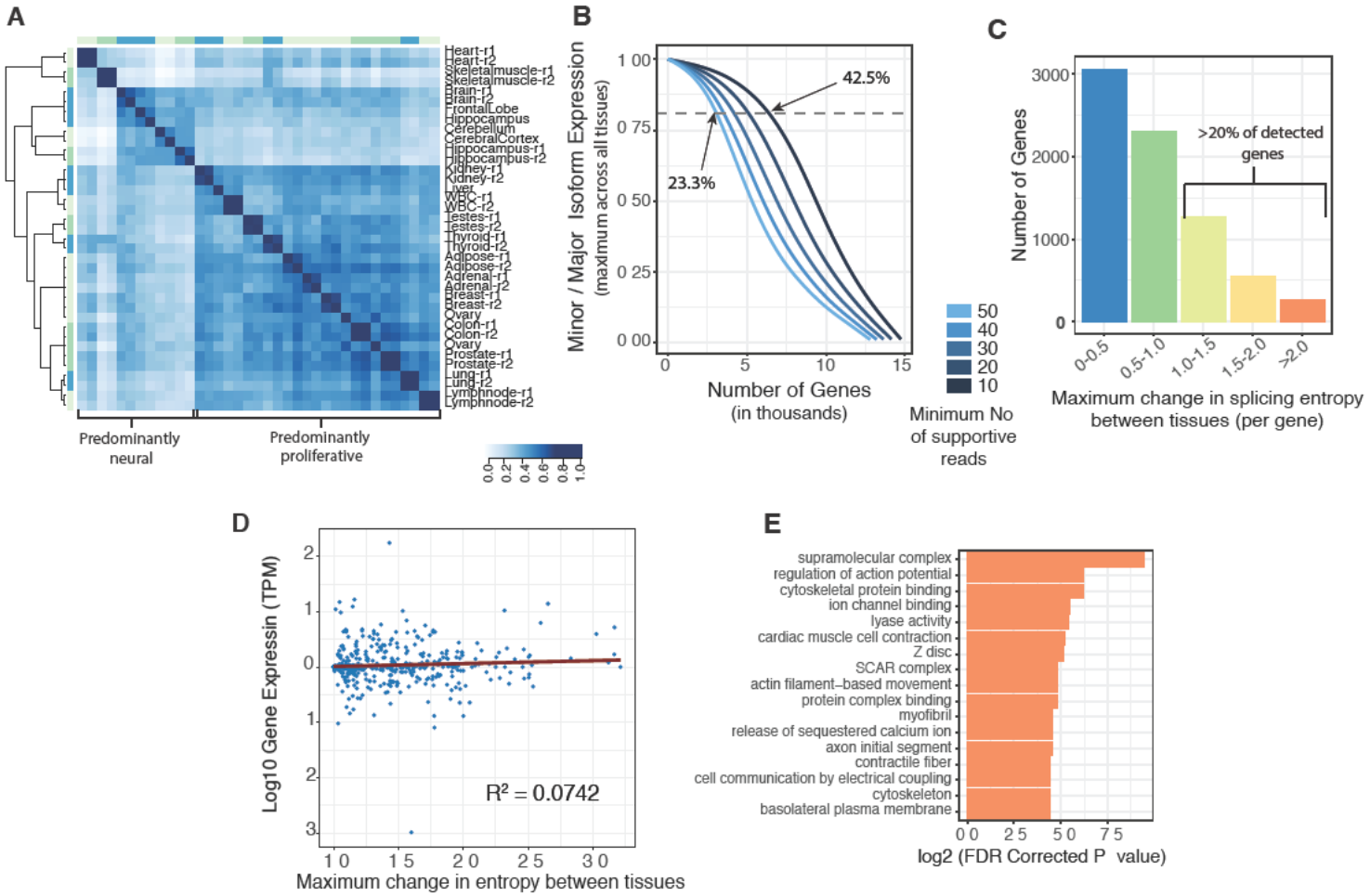

Figure 3

Figure 4

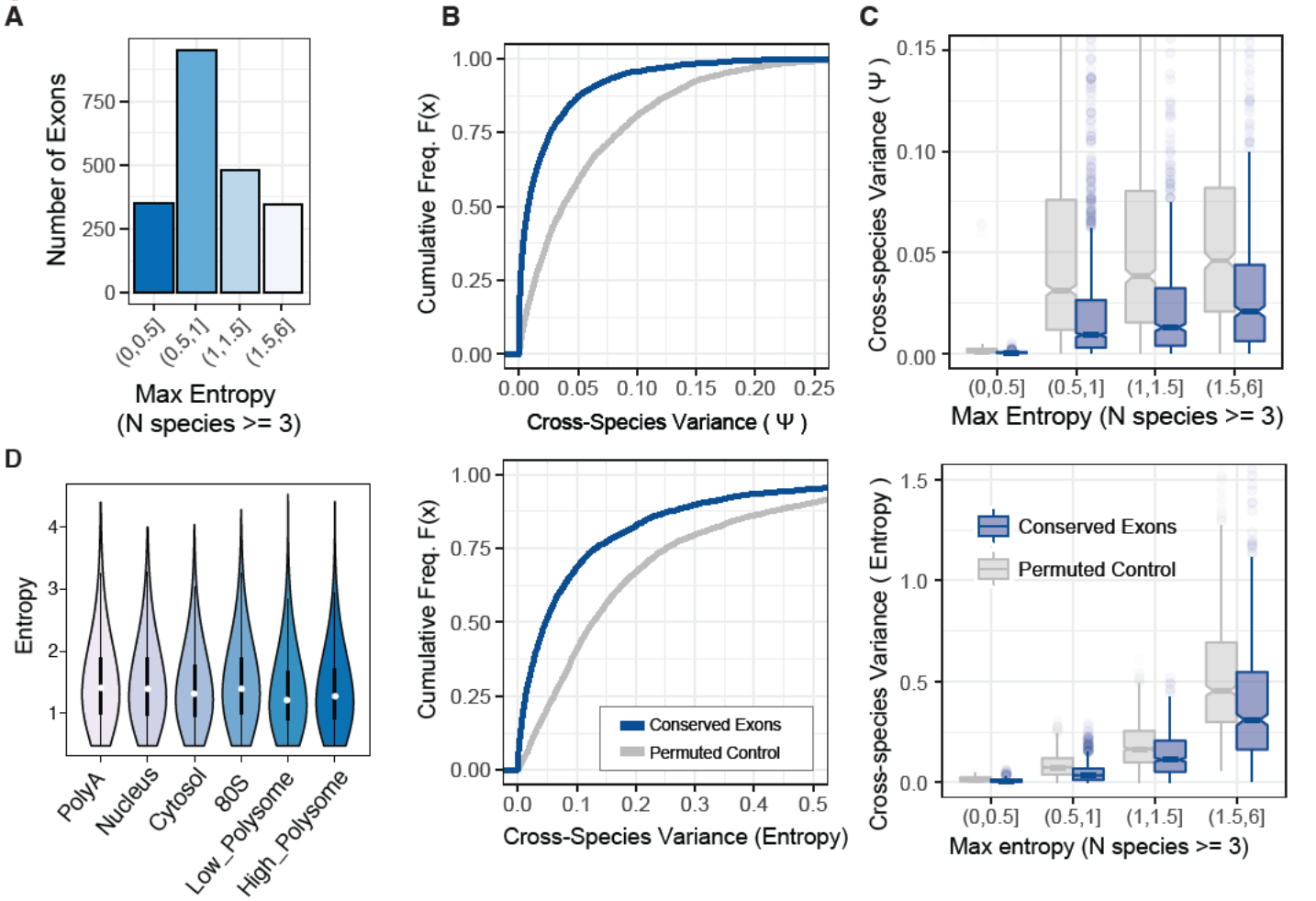

Figure 4

Figure 5

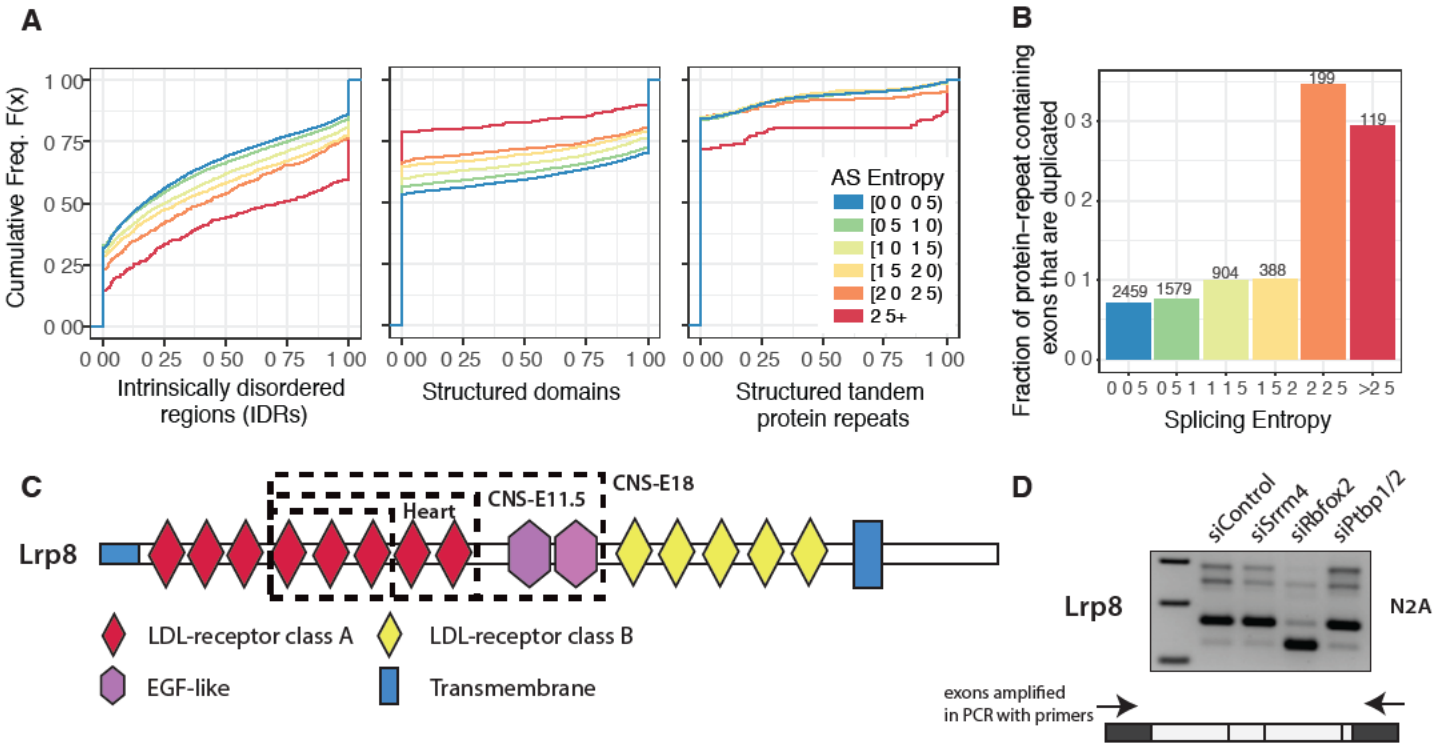

Figure 5

Figure 6

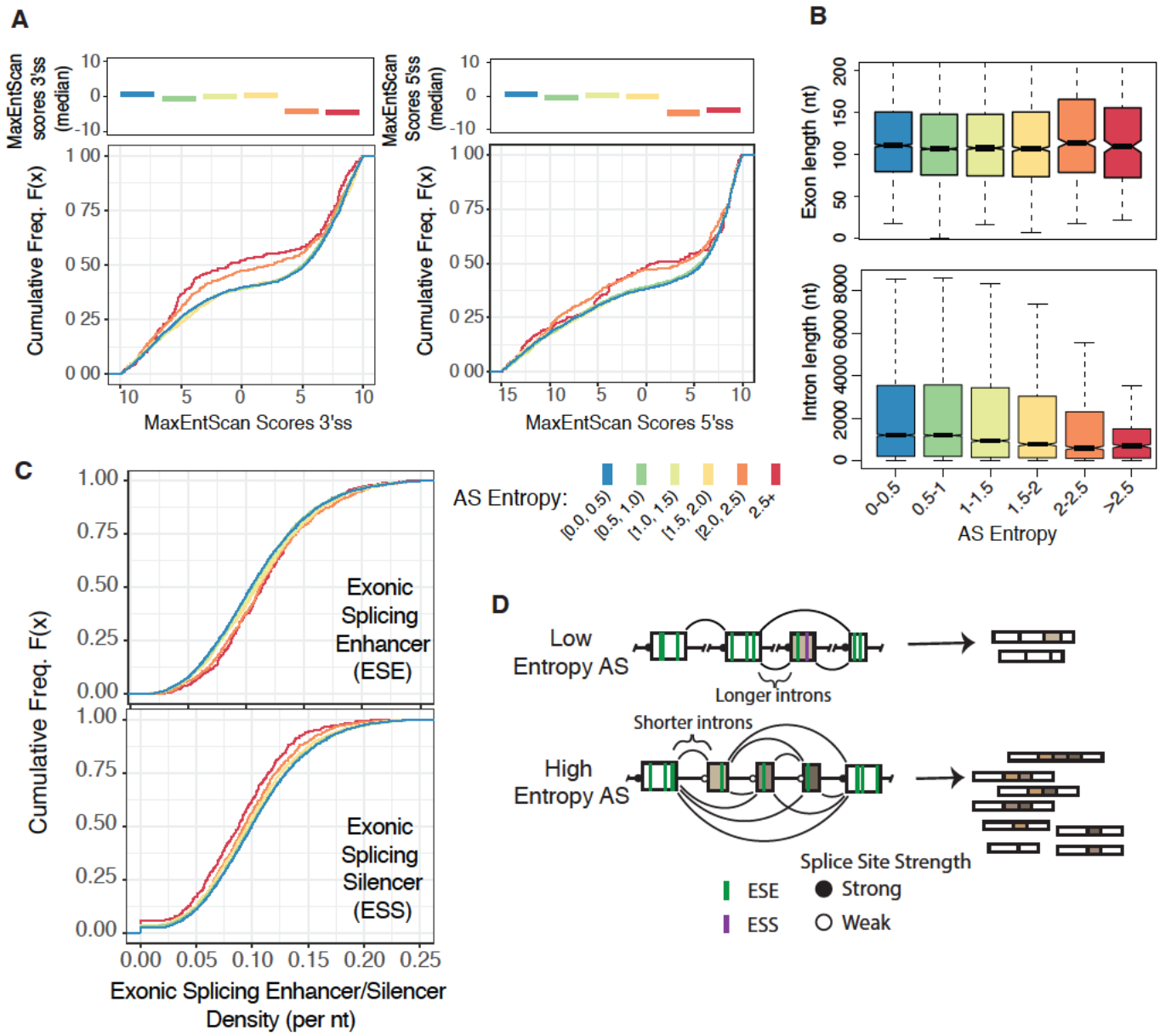

Figure 6

Figure 7

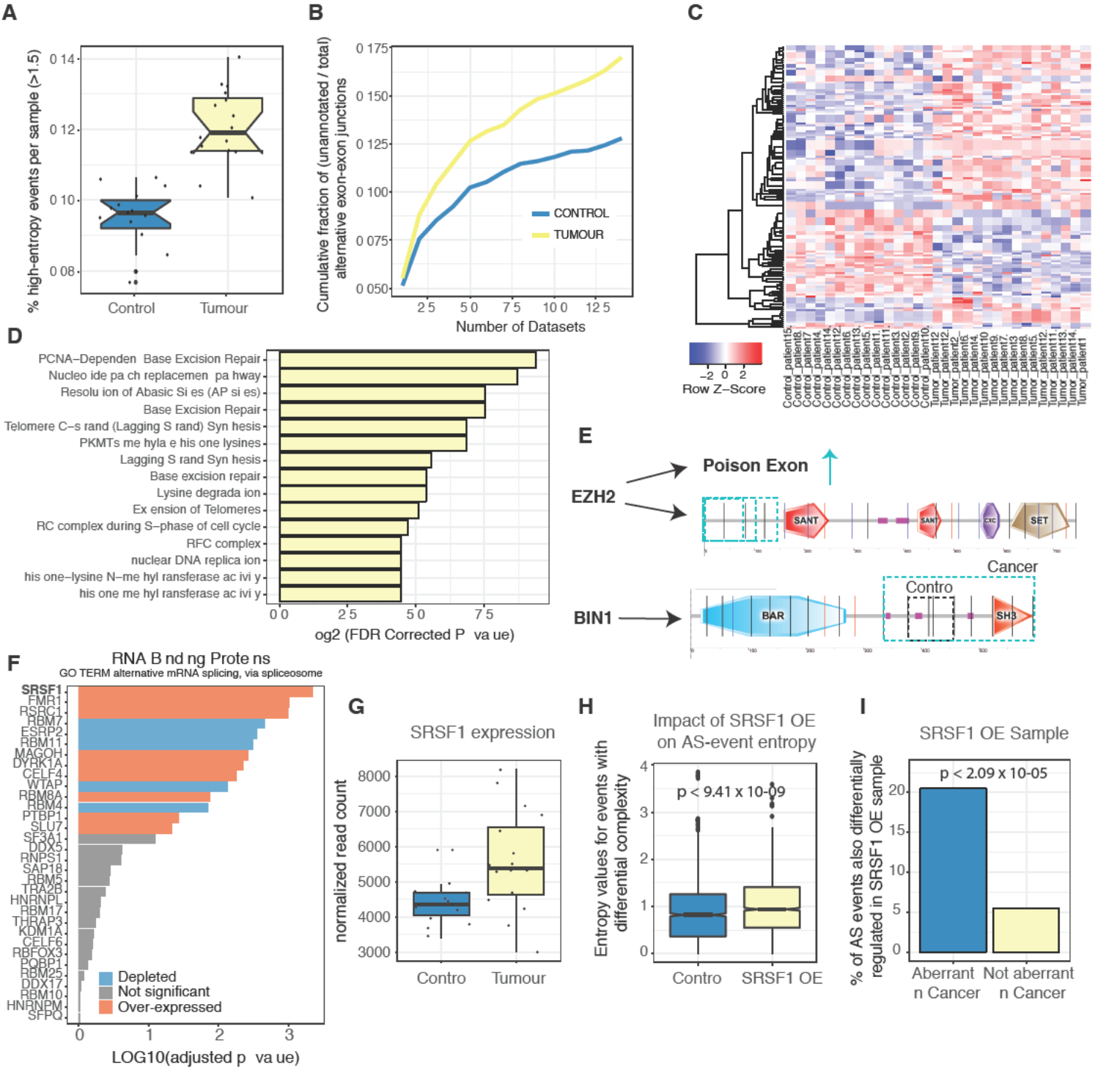

Figure 7

METHODS S1, Related to STAR METHODS and Figures 1, 2, 3 and 5

Contiguous Splice Graph Index (Extended):

To build a Contiguous Splice Graph **(**CSG) from the transcript annotations for a gene, let X represent the set of exon (start, end) intervals (a, b) from those transcripts (where $a \leq b$ for all intervals). The start and end positions for all exons in the gene are also stored in four sets of sorted (in increasing order) positions: F, the first a position for each transcript, L the last b position for each transcript, A the rest of the a positions in the transcript, and B the rest of the b positions in each transcript. Next, we define a single ordered vector of all exon start and end positions $V = \{F \cup L \cup A \cup B\}$ where adjacent positions in V have monotonically increasing order (e.g. $v_i \leq v_{i+1}$ for all i in 1 to $|V|$, $v \in V$). A CSG is built by iterating through the positions in V and building nodes and assigning boundary types based the set membership of adjacent positions v_i and v_{i+1}. The pseudocode for a positive strand gene is provided as an example below (**Algorithm S1**):

``` **Algorithm S1.**

*csg* = Vector(∅)

**For** *i* **in** length(*V*) - 1:
    **If** Is_exonic_interval($v_i$, $v_{i+1}$):
        *incoming.position* = $v_i$
        *incoming.class* = Boundary_class($v_i$)
        *outgoing.position* = $v_{i+1}$
        *outgoing.class* = Boundary_class($v_{i+1}$)

        **If** $i > 1$ **AND** Is_exonic_interval($v_{i-1}$, $v_i$):
            *incoming.property* = Soft
        **Else**:
            *incoming.property* = Hard

        **If** $i <$ length($V$) – 1 **AND** Is_exonic_interval($v_{i+1}$, $v_{i+2}$):
            *outgoing.property* = Soft
        **Else**:
            *outgoing.property* = Hard
```

```
            csg.Append( Node( incoming, outgoing ) )
return csg


Function Boundary_class( element ):
    If element in A:
        return 3′SpliceSite
    Elseif element in B
        return 5′SpliceSite
    Elseif element in F
        return TxStart
    Elseif element in L
        return TxEnd

Function Is_exonic_interval( start, end ):
    If there exists an exon in X with subinterval(start , end):
        Return True
    Else:
        Return False
```

Unannotated Splice Sites from BAM:

Since many non-model organisms have poorly annotated genomes, CSG indexes built from standard annotation files may miss a considerable number of splice sites and exons. In order to partially overcome this limitation, Whippet allows users to provide a spliced read alignment file in BAM (sorted and indexed) format from an independent *de novo* capable spliced read aligner (e.g. STAR, HISAT) using the `--bam` parameter in `whippet-index.jl`. Whippet iterates through the introns (CIGAR string intron operation `N`) of spliced reads within the boundaries of each gene's annotated upstream and downstream transcription start and end sites (Vaquero-Garcia et al., 2016). In order to decrease artifacts due to ambiguities among overlapping gene annotations (e.g. head-to-head or tail-to-tail regions) with unstranded RNA-seq data, Whippet's default is to require that each spliced read in the BAM file has a splice-site that is known (e.g. found in the annotation set, or novel but already added by an upstream read). This stringency can be overridden using the `--bam-both-novel`

parameter, allowing all splice-sites in a spliced read alignment to be novel, and also the `--bam-min-reads` parameter can be used to specify a minimum number of supporting reads for a splice site. These can be used in order to make Whippet's indexing behavior more closely matched to the defaults for other methods (Vaquero-Garcia et al., 2016).

Formally, building a CSG with supplemented splice-sites from BAM follows the same methodology as from GTF annotation, with a few exceptions. Given a novel splice site *i* added to *A* or *B* (a 3' splice site would be added to *A* and a 5' splice site to *B*), and by definition therefore, added to *V*, the functions in **Algorithm S1**: `Is_exonic_interval($v_i$, $v_{i+1}$)` and `Is_exonic_interval($v_{i-1}$, $v_i$)` return true when the interval falls on the CIGAR string match operation `M` side of the spliced read alignment's intron (e.g. not on the gap side the novel splice-site's RNA-seq alignment).

To test the performance of utilizing BAM supplemented indexes, we used the RSEM read simulator `rsem-simulate-reads` to generate a set of 5M paired RNA-seq reads (using GENCODE v25 TSL1, and the same parameters as used in **Figure 2B**) and aligned them (using the `--sam` output parameter to produce a BAM file) with Whippet to an index built with the same annotation file. Filtering the annotation set for the genes with 500 or more simulated reads, we obtained a new annotation file for 2,432 expressed genes with high simulated read coverage. Supplying this filtered annotation file to `whippet-index.jl` alongside the BAM file (using the `--bam` parameter), Whippet identifies 66,888 annotated splice sites and erroneously finds 3 "new" splice sites (false positives), at an FDR of 0.0048% (3 / 66,891), and 33 false positive splice sites with the `--bam-both-novel` flag, at an FDR of 0.049% (33 / 66,921). Similarly applying Whippet using annotation files where the internal exons

have been randomly removed at a variable rate (from 10% to 90%), we observe >95% Recall (also called True Positive Rate, TPR; TP / [TP+FN]) of splice-site identification with 30% exons down-sampled for the default settings, and nearly total Recall (>99%) at all levels of down-sampling using the `--bam-both-novel` flag. This analysis illustrates the effectiveness of GTF+BAM supplemented indices compared to GTF alone (**Figure S1E**).

Contiguous Splice Graph alignment & seeding:

Since a major computational bottleneck of read alignment is the full-text search for alignment seed location, we focused on making the seeding process as efficient as possible. Whippet therefore will not attempt to seed any window containing a FASTQ nucleotide entry with a QUAL score below the minimum threshold (Phred Q=4, P=0.398 by default), and will iteratively scan the read until a seed of acceptable QUAL distribution is obtained. After choosing a seed of (constant seed-length n) at read offset i ($S_{i,n}$), Whippet is very restrictive in the number of matching loci (Count($S_{i,n}$)) that a valid seed is allowed to have (by default $1 \leq \text{Count}(S_{i,n}) \leq \textit{MaxLocations}$) in order to avoid exhaustive searching in highly repetitive sequence space (where the default for *MaxLocations* is 4; parameter `--seed-tol`). For failed seeds, the reverse complement of the seed is tried $S_{i,n} = \text{RevComp}(S_{i,n})$, and then if that fails, the read offset is incremented by a constant number of nucleotides specified by *SeedIncrement* (parameter `--seed-inc`, default is 18; **Algorithm S2**). If the *SeedIncrement* is set to 1, all positions in the read can be seeded. However, this is highly inefficient, given that the full-text search of a seed is the main computational bottleneck of read alignment, and the reason that a given seed will not have valid matches would be shared by most overlapping seeds. Therefore, the default heuristic is to only search non-overlapping seeds.

Seeds themselves can match across boundaries between adjacent nodes in a CSG. This may produce a slight potential for bias, as a seed can map across a constitutive junction, but not an alternative junction. This is due to the presence of constitutive junctions, but not alternative junctions, in the FM-Index. To decrease this potential form of bias, reads are not considered 'spliced reads' (i.e. to be used as evidence of splicing) until they are formally extended across a boundary using the CSG alignment k-mer logic. In theory, there would be a greater likelihood of mapping the first, and possibly second seed in a read for inclusion reads rather than skipping reads. However, this bias is likely to be negligible, since an exclusion read that does not map with its first or second seed, will simply attempt another upstream/downstream seed, one of which should eventually map to the flanking exon and extend properly in the forward and reverse directions. In this way, we have instituted the same rules for a constitutive inclusion alignment as an alternatively spliced read alignment.

Since all genes in the CSG index are stored in positive-strand orientation, only the seed needs to be reversed to subsequently search the entire index for the opposite strand. For paired-end sequencing reads, the same principles apply, except that mate-pair seeds ($M_{j,n}$) are locked into a relative orientation to one another (*fwd_read & rev_mate* by default, but can be changed by the user-defined parameter `--pair-same-strand`), and only loci within the mate-pair mapping index distance (not genomic distance) where || Location($S_{i,n}$) – Location($M_{j,n}$) || <= *MaxMateDistance* (user-defined parameter `--pair-range` with default of 5000nt), are returned. The pseudocode for the seeding procedure is provided in **Algorithm S2** below:

```**Algorithm S2.**

**user-defined constant** *MaxSeeds = 4*
**user-defined constant** *MaxLocations* = 4
```

```
user-defined constant SeedIncrement = 18
user-defined constant MinQual = 4

i = 1
seed_tries = 0
While i+n-1 ≤ length(Read) AND seed_tries ≤ MaxSeeds:
        If minimum( Qual_{i,...,i+n-1} ) > MinQual:
                seed_tries = seed_tries + 1
                If 1 ≤ Count(S_{i,n}) ≤ MaxLocations:
                        return Locations(S_{i,n}), PositiveStrand
                Elseif 1 ≤ Count( RevComp(S_{i,n}) ) ≤ MaxLocations:
                        return Locations(RevComp(S_{i,n})), NegativeStrand
        i = i + SeedIncrement
```

Given a set of mapped transcriptome loci from seeding a sequencing read, Whippet performs ungapped extension of each alignment seed, storing only the offset of the alignment in the read, the offset of the alignment in the transcriptome, and the alignment's path through CSG nodes. An alignment path is defined as the vector of nodes, such that each node records the {Gene, Node, Score}, where the Score refers to a set of {Matches, Mismatches, and the Mismatch_probability_sum}. The *Mismatch_probability_sum* is a running sum of the probability of a correct base call for each mismatched nucleotide base quality q in the alignment (i.e. $\sum_q 1 - 10^{-phred(q)/10}$). This is a heuristic count used to fully penalize mismatches at FASTQ positions that have high quality, whereas low quality read positions (such as the IUPAC nucleotide N) are only partially counted towards the mismatch threshold. *Alignments* are extended first in the *forward* and then in the *reverse* direction from a single position at the seed offset. Alignments are ungapped and local, and extensions continue until either the edge of the read or a Hard boundary that fails spliced extension is reached, or the *Mismatch_probability_sum* exceeds the *MismatchThreshold* (default 3.0, parameter `-X`). An alignment is considered valid if its score exceeds the *MinimumScoreThreshold* (default is Matches / read length = 0.6, parameter `--score-min`) or contains at least one spliced edge in its path. Upon

alignment extension past a Soft node boundary where splicing is possible (e.g. 5′SpliceSite in the *forward* extension direction, and 3′SpliceSite in the *reverse* direction), the extension conditionally traverses past the boundary. If the alignment ends in the neighbor node within a specified distance (k-mer size) from the node's boundary, then the neighbor node is removed from the path, and k-mer spliced extension from the previously skipped Soft boundary is attempted (**Figure S1D**). If spliced alignment extension fails, the node is removed from the alignment and the alignment is truncated at the previous node boundary. This heuristic is intended to reduce potential for bias by requiring the same number of matches past an unspliced Soft boundary, as is required for a spliced extension utilizing k-mers. If multiple Soft boundaries are traversed before the alignment fails, k-mer spliced extension is attempted on each previous neighbor node. For spliced extension, the sorted list of (gene, node) tuples for the k-mer flanking the node boundary is intersected with the sorted list of (gene, node) tuples for the next adjacent k-mer (if sufficient read length exists). Compatible (gene, node) tuples for intersection must share the same gene, where the node with an incoming 3′ SpliceSite boundary lies downstream of the current node in the CSG. Spliced alignment extension continues to all compatible downstream nodes recursively, returning only the best scoring alignment path. If the circular splicing `--circ` parameter is enabled, Whippet is uniquely able (compared to other splice graphs approaches which are acyclic) to align reads from circular RNA products where the read is spliced from a downstream 5′ SpliceSite to an upstream 3′ SpliceSite. In order to calculate Ψ for a circular splicing event, Whippet simply outputs the marginal distribution of the circular edge counts out of the total edges for a given outgoing 5′ SpliceSite.

TPM quantification and multi-mapping reads:

For each alignment seed, an alignment is returned. Multi-mapping reads with multiple *valid* alignments whose scores are within 5% from the maximal scoring alignment are subsequently treated as *repetitive alignments*. Since multi-mapping reads suggest that a sequencing read could have derived from one of multiple paralogous gene loci, we utilize the Expectation Maximization (EM) algorithm to iteratively maximize the likelihood of the relative abundance of all CSGs. Since each CSG can produce paths of various lengths, we utilize the set of all transcript annotations *T* to quantify gene expression at the transcript-level. The EM algorithm alternates between estimating the fraction of multi-mapping read counts belonging to each transcript (E-step) and calculating the relative abundance of all transcripts given the total and partial counts allocated to each (M-step) (Pachter, 2011). In the first step, the probability of observing a read from a given transcript *i* with relative expression level μ_i is given by $\alpha(i) = \frac{\hat{\mu}_i \tilde{l}_i}{\sum_{t \in T} \hat{\mu}_t \tilde{l}_t}$ where $\tilde{l}_i$ refers to the approximate effective length of the transcript given an average read length $\bar{m}$ and a transcript length l_i, where $\tilde{l}_i = l_i - \bar{m} + 1$. We then define a multi-mapping compatibility matrix $\mathbf{y}_{r,i} = 1$ for a read *r* that maps to an annotated transcript *i* and $\mathbf{y}_{r,i} = 0$ otherwise (Bray et al., 2016). The probability of observing a specific multi-mapped read *r* from a transcript *i* in its compatible set of transcripts is then estimated in the (E-step) of the algorithm by:

$$\alpha(r, i) = \frac{\boldsymbol{y}_{r,i}\, \hat{\mu}_i}{\sum_{t \in T}\, \boldsymbol{y}_{r,t}\, \hat{\mu}_t}$$

In the (M-step) of the algorithm, the relative abundance of each annotated transcript is calculated by summing all of the full and partial read counts over all reads *R* for each transcript *i*:

$$\mu_i = \frac{\sum_{r \in R} \alpha(r, i)}{\tilde{l}_i}$$

The likelihood function is given by:

$$\mathcal{L}(\alpha) \propto \prod_{r \in R} \sum_{t \in T} \boldsymbol{y}_{r,t} \frac{\alpha(t)}{\tilde{l}_i}$$

The transcripts-per-million (Li and Dewey, 2011) (TPM) of each transcript is then calculated as:

$$\tau_i = \hat{\mu}_i 10^6 \text{ where } \hat{\mu}_i = \frac{\mu_i}{\sum_{t \in T} \mu_t}$$

The TPM at the gene-level is thus the sum of the transcript TPMs in the gene. To seed the EM algorithm, a uniform probability across compatible transcripts is assigned for each read, followed by the standard M-step, and subsequent EM-steps until the end condition, $(\tau_{i,iter} - \tau_{i,iter-1}) < 0.1$ for all i, or a user-defined max-iterations is reached (default = 10,000; which was never reached in our testing). Since a major bottleneck of each EM step is iterating through the set of reads, and since some sets of reads share transcript compatibility, we can decrease the number of EM iterations by grouping reads into equivalence classes (Bray et al., 2016; Pachter, 2011): sets of reads in R with the same transcript compatibilities $\mathbf{y}_{r,i}$ for all i. Defining the set of equivalence classes Q, the read count for each equivalence class $q \in Q$ as c_q, and a new compatibility matrix where $\mathbf{y}_{q,i} = \mathbf{y}_{r,i}$ for all reads in the equivalence class q, we can re-write the likelihood function as:

$$\mathcal{L}(\alpha) \propto \prod_{q \in Q} \left(\sum_{t \in T} \boldsymbol{y}_{q,t} \frac{\alpha(t)}{\tilde{l}_i} \right)^{c_q}$$

In order to further decrease this computational bottleneck, we implement a separate convergence condition for each equivalence class based on the logic that while convergence of the entire set of transcripts may not *yet* be reached, it is possible that the relative assignments of a single equivalence class *has* already converged. For any equivalence class that meets the end condition: absolute $\alpha(q,i)_{\text{iter-1}} - \alpha(q,i)_{\text{iter}} < 0.0001$

for all transcripts i compatible with q, we fully assign the read count(s) and remove the specified read or equivalence class from further EM iterations.

AS event path enumeration:

After collecting all edges for an event E as a vector of connected nodes, where E_i refers to one edge in the AS event, we build a vector V that contains the minimum set of non-redundant paths (each path V_i contains the set of nodes in the connected path) through the AS event using the edges in E. To build the set of paths V through the AS event, we implement a speed and memory efficient variant of a Breadth-First graph enumeration algorithm (see **Algorithm S3**).

```**Algorithm S3.**

**Function** has_terminal_overlap( a, b ):
    **return** first(a) == last(b) OR first(b) == last(a) ? **True** : **False**

**Function** enumerate_paths( $E$, $V$=copy!($E$) ):
    $R$ = Vector( ∅ )
    $i$ = 1
    **While** $R$ does not equal $V$:
        If $i$ > 1:
            $V$ = $R$
            $R$ = Vector( ∅ )

        **For** $j$ **in** 1 **to** length($V$):
            Added = False
            **For** $k$ **in** 1 **to** length($E$):
                **If** has_terminal_overlap( $V_j$, $E_k$ ):
                    Added = **True**
                    push!( $R$, $V_j \cup E_k$ )

            **Unless** Added **is** True:
                push!( $R$, $V_j$ )
        $i$ = $i$ + 1
    **return** $R$
```

AS event definition and PSI quantification (Extended):

Unlike quantification at the transcript-level (as described above), PSI quantification uses only edge counts. Therefore to account for the length of enumerated AS event paths in nucleotides (analogous to the effective length of a transcript in nucleotides for transcript-level quantification), we consider that the number of positions a read of average length $\overline{m}$ can map in order to cover an edge with k-mer length k is a constant $\overline{m} - 2k + 1$. We next assume that this is the same for all spliced edges. Since spliced reads covering multiple edges are counted fully at each edge, and since the range of mappable positions for each edge can overlap for proximal neighboring edges, the length of the path in nucleotides is proportional to the number of edges in a path. We therefore define j_i as the length of path i which is proportional to the number of edges in the path (using the compatibility matrix $\boldsymbol{y}$) such that: $j_i \propto \sum_{e \in E} \boldsymbol{y}_{e,i}$ (as used in the **Methods**).

For tandem UTR event types (e.g. TS: tandem transcription start site, TE: tandem transcription end site), Ψ values are calculated using all reads that map to the set of TS or TE nodes being quantified, and the effective length of the path in nucleotides is used directly. For tandem UTR events, the likelihood function is therefore identical to that provided for the quantification of AS events with two exceptions: (1) $j_i = l_i - \overline{m} + 1$, so the length is equal to the full length of the path in nucleotides, and (2) E refers to the set of equivalence classes (not edges) for reads with the same path compatibility $\boldsymbol{y}_{e,i}$.

Since the EM-algorithm provides only a point-estimate for Ψ without a depth-dependent measure of variance, we utilize the conjugate posterior distribution of the binomial likelihood as a means to compute a read-count derived confidence interval (CI) over Ψ. Given a total read depth for an AS event of N reads which can either support inclusion of node n, $inc \in I_n$, or support exclusion, $exc \in \{I - I_n\}$, the number of

inclusion reads N_{inc} are binomially distributed such that $N_{inc} \sim Binomial(n=N, p=\Psi)$. Given a uniform prior distribution of $P(\Psi) = Beta(\alpha=1, \beta=1)$, we obtain a posterior distribution, $P(\Psi|N_{inc}) \propto P(N_{inc}|\Psi)P(\Psi)$, where $P(\Psi|N_{inc}) = Beta(N_{inc} + \alpha, N_{exc} + \beta)$. A 90% confidence interval (between 5% and 95%) is then calculated through the quantile distribution of the posterior. This output allows a user to more easily filter for a subset of nodes that have a minimum read depth to estimate Ψ within some range of expected confidence.

Bias correction

RNA-seq protocols can result in a number of biases. When comparing quantifications of RNA-seq libraries prepared using different kits or protocols, or from different batches, these biases can become pronounced and affect downstream analyses (Love et al., 2016). While bias correction methods are not enabled by default, Whippet does correct for bias in RNA-seq data when the `--biascorrect` flag is enabled. One source of bias is due to the generation of cDNA using random hexamer priming, which has been shown to induce sequence-specific biases at the 5′ end of Illumina reads (Hansen et al., 2010; Roberts et al., 2011). To account for this 5′ sequence bias, Whippet implements the method of Hansen et al. (Hansen et al., 2010), in which all read counts are adjusted after read-alignment but before any quantification takes place. Since AS event quantification is often based on only a small number of exon-exon junctions, unusually high or low GC-content in the regions flanking exon-exon junctions could also have a profound effect on quantification (Love et al., 2016). Therefore, for AS event quantification only, Whippet will additionally correct for GC-content bias using a fast heuristic method. Briefly, Whippet corrects each read count by the ratios of the expected distribution of

GC-content (for 50nt windows) to the observed distribution of GC-content as follows: The Whippet CSG index stores GC-content in 20 bins of 5% GC-content intervals (0-5%, 6-10%, 11-15%, etc.) for all 50-mers in each annotated transcript. After transcript-level expression quantification, a single expected distribution of GC-content is calculated as the normalized expression-weighted sum of the distributions of all transcripts. Additionally, an observed distribution for GC-content is calculated as the normalized distribution of GC-content across sliding windows of 50-mers from all reads (or set of all read pairs). To correct GC-content bias, edge counts are subsequently adjusted by the mean of the (expected / observed) ratios for all of the GC-content windows stored for the reads in the edge.

AS event types:

In order to define the nature of an alternative node in an AS event, a number of discrete categories are utilized. These include AS specific types for alternative 5′ or 3′ splice-sites, Core-exon nodes (which may be a whole exon or part of an exon with flanking alternative splice sites that are used), or a retained intron. Additionally, alternative 5′ or 3′ ends are annotated as either alternative first or last exons, or tandem 5′ or 3′ untranslated regions (UTR). **Table S7** provides a list of the two-letter symbols for each node type and their formal definition as a set of flanking boundary types.

AS event complexity and RNA-seq simulation (Extended):

To simulate AS-events with known Ψ-values using polyester (error rate = 0), we randomly subsampled one of the *n* alternative nodes from each gene and assigned it a random Ψ-value sampled from a Beta distribution. Since we simulate both low and

high complexity events K1, … K6, we observe that at higher complexities, the assignment of a Ψ-value to one node has indirect effects on the Ψ-values of the neighboring *n* alternative nodes. In order to achieve a near uniform coverage of total Ψ-values for each complexity K(*n*), the α and β parameters of the Beta distribution were selected *ad hoc* accordingly. For K <= 2, we used Beta(α=0.9, β=0.9) which produces a near uniform distribution over Ψ. For K(*n*) in the range of interval $n \in [3, 5]$, it was necessary to use a distribution skewed more towards 0.0 and 1.0 (such as Beta(α=0.7, β=0.7)), since a uniform distribution of initial Ψ-values resulted in a bell-shaped curve centered on Ψ=0.5. For K ≥ 6, this effect was substantially increased, requiring a more skewed initial distribution, Beta(α=0.2, β=0.2). Transcripts containing the sampled exon were randomly assigned relative expression values such that their total expression would be proportional to the pre-assigned Ψ-value. Similarly, the remaining transcripts were randomly assigned expression values such that their total expression is proportional to (1 - Ψ). To simulate a controlled variable range of gene expression, each gene was randomly assigned a coverage multiplier value from a uniform distribution between 5x and 60x. Subsequently, RNA-seq reads of length 100nt were simulated for each transcript in both single and paired-end modes. For **Figure S4C-S4D**, read lengths of 50nt and 75nt were additionally simulated.

Benchmarking (Extended):

Resource usage benchmarks for STAR and TOPHAT included conversion to a sorted BAM file (as this step is required for majority of splicing quantification algorithms) whereas quantification time was removed for Whippet in comparison to alignment programs. The only exception to this rule was in **Figure S2D** when

quantification time and time to produce sam output was included. When necessary, initial read alignment was done by STAR and same BAM output file used by all splicing quantification algorithms. The default linux package "time" (/usr/bin/time – e.g. http://man7.org/linux/man-pages/man1/time.1.html) was used to measure the resource usage of each program. The running (CPU) time was calculated by combining the User time and the System time. The Maximum resident set size was used as measurement of "Maximum memory used".

Benchmarks of the mapping success used the program Benchmarker, and for each aligner, default parameters were used with exceptions in **Table S4**. Using bedtools (https://github.com/arq5x/bedtools2), the bam output from aligners were compared directly to BEERS true alignment files filtered to contain only reads overlapping junctions (derived from "junctions read set" supplied by BEERS). This is considered the "ALL SPLICED READS" version in **Figure S2D**. The second sample "W/ INTACT KMERS" was constructed by removing reads with annotated substitutions (substitutions derived from BEERS substitution dataset). This sub-setting is necessary because Whippet, like other k-mer methods, requires perfect matches in order for the CSG alignment to work. We find that for the set of reads with the intact k-mer set, Whippet's mapping accuracy is > 97% (see **Table S6**), which is comparable to other RNA alignment software (**Figure S2D**). The analysis shown in **Figure S2D** suggests that the decrease in mapping performance observed for "ALL SPLICED READS" at higher error rates is due to a lack of error tolerance in the k-mers flanking exon-exon junctions and is therefore unlikely to disproportionally impact individual gene family clusters. This conclusion is supported by our observation of a strong similarity across the genome between the full distribution of "ALL SPLICED READS" and the subset of reads not successfully aligned by

Whippet (see **Figure S2E**). Therefore, such error harboring reads would not adversely affect Whippet's quantification, other than to slightly decrease statistical power for high error rate datasets.

The original reads are located within the data folder here:

https://figshare.com/articles/Whippet_analysis_scripts/5711683

To manually recreate these reads please use the following command:

perl reads_simulator.pl 1000000 refseq_stimul_Beer-reads_0.001 -palt 0 -indelfreq 0 -subfreq 0.001 -error 0 -configstem refseq -outdir beers_simulator/refseq_config/

perl reads_simulator.pl 1000000 refseq_stimul_Beer-reads_0.005 -palt 0 -indelfreq 0 -subfreq 0.005 -error 0 -configstem refseq -outdir beers_simulator/refseq_config/

perl reads_simulator.pl 1000000 refseq_stimul_Beer-reads_0.01 -palt 0 -indelfreq 0 -subfreq 0.01 -error 0 -configstem refseq -outdir beers_simulator/refseq_config/

Only simple cassette exons events were considered in the RT-PCR analysis to ensure best possible mapping quality. Simple events are those containing just three exon-exon junctions, defined by MAJIQ. Across comparison we only used core cassette exons, as this type of event is described by all methods. In general, only PSI values for programs in which the event coordinates exactly matched the "ground-truth" event were considered. For MAJIQ, because PSI values can be duplicated in different local splicing variations, we used the most compatible PSI value after ensuring complete overlap of coordinates. In general, it is difficult to directly compare between splicing events measured by different methods. To alleviate concerns on these issues we firstly removed comparisons of number of detected events, as this is

very dependent on mapping fidelity. Secondly, we developed Whippet_TPM, which uses same coordinates scheme as Whippet, to ensure that differences observed between transcript- and event-level programs are not solely due to coordinate matching issues. Thirdly, we used the Wilcoxon signed rank as the statistical test because it is paired and therefore only assesses directly comparable events.

Tissue-wide analysis of splicing (Extended):

For the polysome analysis of entropy, monosome and polysome samples combined based on sub-groups identified within original paper (Floor and Doudna, 2016). This included 80S (monosomes), low polysomes (two-four ribosomes), high polysomes (five-eight+ ribosomes), and total cytoplasmic RNA. Additional nuclear and whole-cell HeLa fractions originating from a different paper were also analysed as a comparison (see **Table S3**).

For the analysis of correlation of expression and entropy values in tissue and cancer data, TPM values calculated by Whippet were used. Correlation was assessed using Pearson Correlation Coefficient.

Functional analysis:

Functional analysis was undertaken using the functional enrichment analysis tool g:Profiler (http://biit.cs.ut.ee/gprofiler). Genes identified as containing mammalian-classifying splicing events were compared to a background of multi-exon genes conserved within vertebrates. Structured controlled vocabularies from Gene Ontology organization, as well as information from the curated KEGG and Reactome databases were included in the analysis. Only functional categorizes with more than five members and fewer than 2,000 members were included in the analysis.

Significance was assessed using the hypergeometric test with the multiple testing correction method created by Benjamini and Hochberg.

Feature analysis of high entropy events (Extended):

For all positions in a protein low complexity regions were calculated using SEG (Wootton, 1994). Amino acid residues not within ordered annotated protein domains, putative transmembrane domains, signal peptides and coiled coil regions were considered as low complexity regions. For each exon, the ratio of amino acids annotated as within a low complexity region is estimated. Tandem protein repeat regions within structured regions were identified using the PTRStalker algorithm for de-novo detection of fuzzy tandem repeats (Pellegrini et al., 2012) and filtered using IUPred (score < 0.4).

Exon duplication events were identified using approach described previously (Letunic et al., 2002). In brief, exon were considered duplicates if (i) within the same gene body (ii) blastn comparison had an e-value of less than 0.0001 (iii) 80% similarity in length. Exons with peptide repeats were extracted and maximum entropy value (across tissues) identified. For each entropy bin, the fraction of duplicated exons was calculated.

SUPPLEMENTARY FIGURE AND TABLE LEGENDS

Figure S1 – Related to Figure 1

(A) Schematic of transcript-level (ie. Whippet_TPM; *left*) vs. event-level (ie. Whippet; *right*) quantification paradigms. Here, an incomplete transcript-level annotation set (*left*) can unintentionally produce interaction between distal transcript features (e.g. an internal exon and an alternative polyadenylation site). In contrast, the event-level method built from the same annotation set calculates Ψ independently, using only the subset of reads directly mapping to the exons and splice-junctions forming the AS for quantification. Red coverage plots above each gene schematic illustrate relative read depth, and curved lines indicate exon-exon junction mapping read depth. TPM, transcripts-per-million.

(B) Exemplar gene and isoform annotations. The node diagram below displays how such a complex set of splicing patterns can be collapsed into a single set of nodes for a CSG. Incoming and outgoing boundary types associated with the node set in the panel are annotated upstream and downstream of the node sequence respectively, with hard boundary types bolded (see Methods for definitions of boundary types).

(C) Top panel illustrates the full CSG sequence from panel (B). Bottom panel shows the node "event type" annotations for the CSG nodes in panels B-C according to the incoming and outgoing boundary types listed in **Table S7**. AF, alternative first exon; AA, alternative acceptor splice site; CE, Core exon; RI, retained intron; AD, alternative donor splice site; TE, tandem alternative polyadenylation site.

(D) Graphical overview of the CSG alignment algorithm. High FASTQ-QUAL region (*black box*) of sequencing reads are used to seed to the Whippet Index. Alignment extension occurs in the forward and reverse directions, and spliced extension is allowed as necessary to bridge

spliced boundaries. Unspliced alignment extension can occur past a soft boundary that also allows spliced extension (i.e. a 5′SpliceSite for forward extension, and a 3′SpliceSite for reverse extension). If the alignment fails during unspliced extension past a soft spliced boundary, then the new node is removed from the alignment and spliced extension proceeds from the spliced boundary of the previous node(s) (see **Methods** for details).

(E) Simulation of *de novo* splice site recovery by `whippet-index.jl` using the `--bam` parameter and a GTF file with a variable percentage (10-90%) of internal exons randomly removed. The *y*-axis shows Recall (i.e. True Positive Rate; TPR—TP / [TP+FN]) as a function of the percentage of internal exons sampled out (*x*-axis). "One Annotated" refers to the default settings in Whippet v0.11, while "Both Novel" uses the `--bam-both-novel` parameter flag, relaxing stringency (see **Methods S1**).

Figure S2 – Related to Figure 2

(A) Comparison of maximum memory usage (*y-axis*) and log-scaled time (*x-axis*) for several published methods for RNA-seq read alignments when searching only for known (Gencode GRCh37.13) exon-exon junctions (*left*) and when novel junction finding features are enabled (*right*) (see data in **Table S5** and program parameters in **Table S4**). GB, gigabyte.

(**B**) A bar plot showing the percentage increase in speed when featured alignment software only searched for known exon-exon junctions

(**C**) A bar plot showing the number exon-exon junctions identified by MAJIQ when the RNA-seq alignment tool HISAT used only known junctions compared to when HISAT also searched for novel junctions. K(*n*), complexity category.

(**D**) Comparison of the mapping success over exon-exon junctions (*y-axis*) and log-scaled time (*x-axis*) requirements of Whippet relative to several published methods for RNA-seq read alignment. Reads were simulated using BEERS (Grant et al., 2011) (http://www.cbil.upenn.edu/BEERS/) with substitution frequency rate of 0.001 (*left*), 0.005 (*center*) and 0.01 (*right*). Only simulated reads overlapping exon-exon junctions by at least Whippet's k-mer size are considered in mapping success. Reads were divided into two categories: all spliced reads and those reads with perfect intact kmers (i.e. no substitutions). Whippet speed test for mapping success rate includes the output of alignment in SAM format (Whippet parameter `--sam`). See **Table S6** for data.

(**E**) Bar plot showing the distribution of all simulated reads reads used in the benchmark in (D), and the distribution of only those reads unaligned by Whippet across the genome.

(F) Plot of the reproducibility of PSI values when comparing RNA-seq from two conditions. A differentially included event is considered replicated if it maintains a rank at least as high as *N* in biological replicates, where *N* is the set size. (see Vaquero-Garcia et al. 2016 for details).

(G) A bar chart shows the number of AS events identified by each method and used in (F).

(H) A cumulative distribution plot comparing RT-PCR Ψ (percent spliced in) values to RNA-seq quantified Ψ values. Data was extracted from the same samples from human liver and cerebellum (Vaquero-Garcia et al., 2016). Same as in **Figure 2A** but without applying any stringency criteria to filter events. See **Figure 2A** legend for a description of cumulative distribution plots.

Figure S3 – Related to Figure 2

(A) Comparison of Whippet with other state-of-the art published splicing algorithms and

Whippet_TPM (see main text **Methods**) for Ψ (percent spliced in) predictions from RNA-seq data and corresponding Ψ estimates from RT-PCR. ΔΨ (change in percent spliced in) denotes difference in PSI measurements from the two data types, when comparing between liver and cerebellum samples, as well as between stimulated and unstimulated human Jurkat T-cell line samples. Regression line is shown as dotted line whereas diagonal is solid line (see **Table S1** for R-squared values)

(B) Same data as in panel (A) but showing Ψ rather than ΔΨ.

(**C**) Bar plots showing the absolute change (*left*) and absolute relative change (*right*) in error rate of quantification algorithms Ψ compared to ground truth Ψ. Errors bars represent the standard error of the mean. Reads were simulated using RSEM (RNA-seq by Expectation Maximization) based on annotation from either Gencode or RefSeq. Indices used for quantification based on RefSeq annotation

(D) Same as (C) except indices used for quantification based on Gencode annotation (an alternative visualization of **Figure 2B**)

Figure S4 – Related to Figure 2

(A) Plot of entropy (*y-axis*) vs. percent-spliced-in (Ψ) (*x-axis*) for a simple binary (K1) AS event.

(B) Plot of maximum entropy (*color-scale*) vs. percent-spliced-in (Ψ) (*x-axis,y-axis*) for a K2 event with two alternative exons *a* and *b* and two independent values for the percent-spliced-in of each exon, Ψ_a and Ψ_b.

(C) Proportion of AS events with entropy values in discrete ranges (*color-scale*) for transcriptome wide simulated RNA-seq data set from sub-sampled read-depth in millions (*left*) and truncated read-lengths (*right*).

(D) Total count of the number of AS events detected with the RNA-seq datasets from panel (C).

(**E**) Comparison of the ability of different RNA-seq analysis methods to detect AS events from artificial reads (**Methods**) of simulated complexity as defined in **Figure 2D**. Bar plots show the total number of AS events detected.

(F) Extension of **Figure 2F**. Scatter plots showing correlations between simulated ground truth Ψ values and RNA-seq predicted Ψ values by multiple published splicing algorithms at various levels of complexity K(*n*). (see **Methods** for details on data simulation).

Figure S5 – Related to Figure 2

(A) Extension of **Figure 2G**, showing uncropped gels for events in the main figure as well as additional examples. RT-PCR analysis confirms the presence of complex splicing events in N2a cells at increasing levels of complexity matching Whippet predictions. Event type, gene name, complexity type and entropy score are shown above each events. Control SImap demonstrates that complexity is not just due to number of exons monitored. Boxes to right of gels display UCSC (*left*) and Whippet (*right*) predictions based on primer sequences (see **Methods**). Colored boxes represent correct predictions whereas black boxes suggest missed predictions. Diagrams below show exon structures of analyzed genes with approximate positions of RT-PCR primers indicated. Predicted constitutive and alternative exons are indicated in dark and light gray, respectively.

(B) Comparison of the maximum memory-usage (*y-axis*) and log-scaled time (*x-axis*) requirements of Whippet relative to several published methods for RNA-seq read alignment using datasets comprising 15 million (M) paired-end (PE) RNA-seq reads. GB, gigabyte.

(**C**) Extension of **Figure 2H**. Comparison of the maximum memory-usage (*y-axis*) and log-

scaled time (*x-axis*) requirements of Whippet relative to published methods for RNA-seq read alignment and splicing quantification when aligning 15 M, 25M and 50M PE RNA-seq reads. GB, gigabyte.

Figure S6 – Related to Figures 3 and 4

(A) Symmetrical heatmap of pairwise correlations of normalized AS event entropy as in panel A except across multiple mouse tissues. See **Figure 4A** legend for details.

(B) Graph of the percentage of genes (*y*-axis) harboring an AS event with entropy > 1 (*top*) or a minor / major isoform co-expression ratio > 0.818 (*bottom*) as a function of various AS event filter thresholds (*x*-axis—*left:* Confidence Interval < *x*, *right:* Read Count > *x*). ALL_AS also includes Core Exons, Alternative Donor (AD) nodes, Alternative Acceptor (AA) nodes, Retained Introns (RI), and Alternative Last (AL) exon nodes.

(C) Ranked plot of genes (*x*-axis) by their maximal AS event minor / major isoform relative expression ratio (*y*-axis) at various minimum read cut-offs, defined as mapped reads overlapping exon-exon junctions. The plot only includes events categorized as Core Exon events. Dashed line indicates the 45:55% ratio cutoff (equivalent to a minor / major ratio of 0.818).

(D) Distribution of the number of unique conserved exons with genomic coordinate 'liftover' across at least three vertebrate species (human, chimp, gorilla, mouse, opossum, platypus, and chicken). Number of unique conserved exons is plotted by tissue (*x*-axis), and binned by the number of species with direct coordinate 'liftover' (*color-scale*).

(**E**) Distribution of the number of unique conserved exons with genomic coordinate 'liftover' across at least three vertebrate species (human, chimp, gorilla, mouse, opossum, platypus, and chicken). Conserved exons are counted in discrete bins by their average entropy in any of the

species. (*bottom*)

(**F**) Distributions for the cross-species variance of entropy values (*y*-axis) for conserved exons, binned by mean entropy values (*x*-axis), and compared to a control set of the same data but with permuted AS event labels for each species (*color-scale*). All two-sided KS-test p-values are less than epsilon (2.2×10^{-16}), except for the bin (1.5,3] whose p-value was 4.6×10^{-4}. (*bottom*) Same as (*top*) except the distributions plotted contain the cross-species variance of Ψ-values (*y*-axis) for the same conserved exons. All two-sided KS-test p-values are less than epsilon (2.2×10^{-16}), except for the bin (1.5,3] whose p-value was 4.3×10^{-2}. See **Figure 4C** for descriptions of boxplots.

(G) Cumulative distribution plots of (*left*) Ψ-values for all conserved exons in all species and tissues, (*right*) mean value of Ψ for a conserved exon across all species in a given tissue. The *left* panel shows an identical distribution to permuted control, while the *right* panels shows less concordance in Ψ-values among permuted data. See **Figure 2A** legend for a description of cumulative distribution plots.

(H) Bar plots showing distribution of events with entropy >1.0 and low entropy (<1.0) events within the 5′-UTR, CDS and 3′-UTR of transcripts across human and mouse tissues. NS, not significant (Fisher's exact test); *, $p < 0.05$; ***, $p < 1 \times 10^{-5}$; CDS, coding sequence; UTR, untranslated region

Figure S7 – Related to Figures 5, 6 and 7

(A) Cumulative distribution plots showing frequency of overlap of AS events (with different degrees of entropy) within the low complexity (LC) regions of proteins (*top*) and unstructured tandemly repeated protein domains (*bottom*). See **Figure 2A** legend for a description of the

cumulative distribution plots (n > 368). See **Figure S7B** for color legend.

(B) Violin plot showing the number of exons encoded by a gene body at different degrees of splicing entropy (maximum splicing entropy observed within gene body used to bin genes. See **Figure 4D** for description of violin plots.

(C) Violin plot showing that genes with higher complexity splicing events tend to be younger (or more recent gene duplication events). See **Figure S7B** for color legend. See **Figure 4D** for description of violin plots.

(D) Stacked bar plot showing the proportion of AS events within bins of increasing splicing entropy across 15 matched tumor and control RNA-seq samples.

(E) Scatter plots of change in entropy of AS events between control and cancer samples versus the change in expression level of the harboring gene. Dotted lines represent 0.5-fold change. Red points are genes displaying a 0.5 fold change in both entropy and expression. Grey points do not. R-squared value calculated using Pearson Correlation Coefficient. TPM = transcripts per million.

(F) Full list of DESeq2 differential gene expression analysis (Love et al., 2014) between tumor samples and matched controls for selected RNA-binding proteins (GO:0000380). Genes with blue bars show reduced expression in cancer samples, red bars show increased expression in cancer samples, and grey bars show no significant difference between control and tumor samples.

Table S1, Related to Figure 2.
P-values, quantiles and r-squared for benchmarking against RT-PCR data

Table S2, Related to Figure 2.
Speed and memory benchmarks for Percent Spliced In (Ψ) quantification algorithms

Table S3, Related to Figures 2,3, 4 and 7
Description of publicly available datasets used in paper

Table S4, Related to Figure 2
Versions of programs used in benchmarking with parameters used for alignment and splicing quantification. Default settings, paths, Fastq files excluded. All programs run on 8 cores unless stated.

Table S5, Related to Figure 2
Speed and memory benchmarks for RNA-seq aligners

Table S6, Related to Figure 2
Error Rate for reads produced by BEERS RNA-seq simulator

Table S7, Related to STAR methods
Node and event types defined by Whippet flanking edges

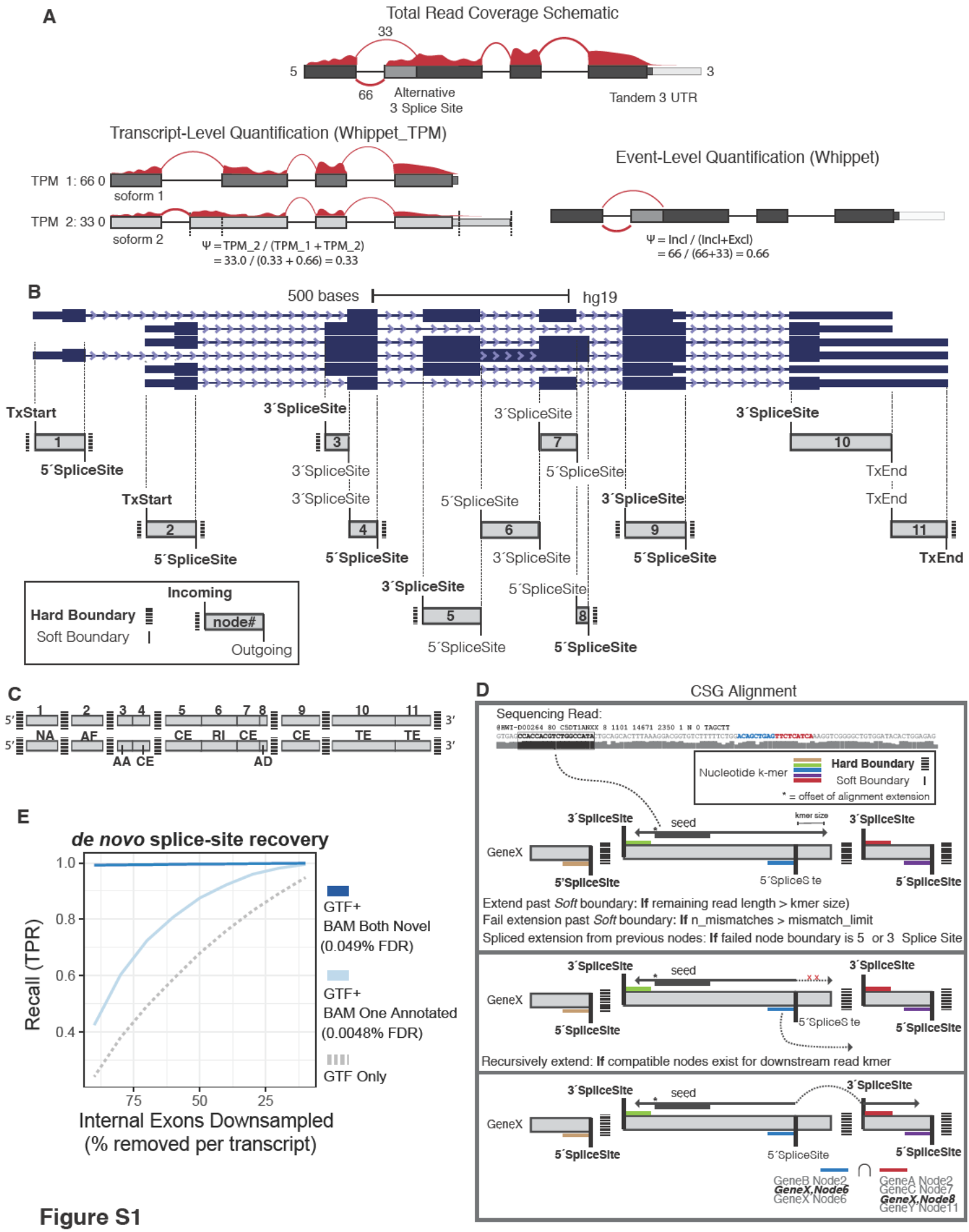
A
Total Read Coverage Schematic
33
5
3
66
Alternative
3 Splice Site
Tandem 3 UTR
Transcript-Level Quantification (Whippet_TPM)
TPM 1: 66 0
soform 1
TPM 2: 33 0
soform 2
Ψ = TPM_2 / (TPM_1 + TPM_2)
= 33.0 / (0.33 + 0.66) = 0.33
Event-Level Quantification (Whippet)
Ψ = Incl / (Incl+Excl)
= 66 / (66+33) = 0.66
B
500 bases
hg19
TxStart
5´SpliceSite
3´SpliceSite
TxEnd
Incoming
Hard Boundary
Soft Boundary
node#
Outgoing
C
NA
AF
AA CE
CE
RI
CE
AD
CE
TE
TE
D
CSG Alignment
Sequencing Read:
Nucleotide k-mer
Hard Boundary
Soft Boundary
* = offset of alignment extension
seed
kmer size
GeneX
Extend past Soft boundary: If remaining read length > kmer size)
Fail extension past Soft boundary: If n_mismatches > mismatch_limit
Spliced extension from previous nodes: If failed node boundary is 5 or 3 Splice Site
Recursively extend: If compatible nodes exist for downstream read kmer
GeneB Node2
GeneX.Node6
GeneX Node6
GeneA Node2
GeneC Node7
GeneX.Node8
GeneY Node11
E
de novo splice-site recovery
Recall (TPR)
1.0
0.8
0.6
0.4
75
50
25
Internal Exons Downsampled
(% removed per transcript)
GTF+
BAM Both Novel
(0.049% FDR)
GTF+
BAM One Annotated
(0.0048% FDR)
GTF Only

Figure S1

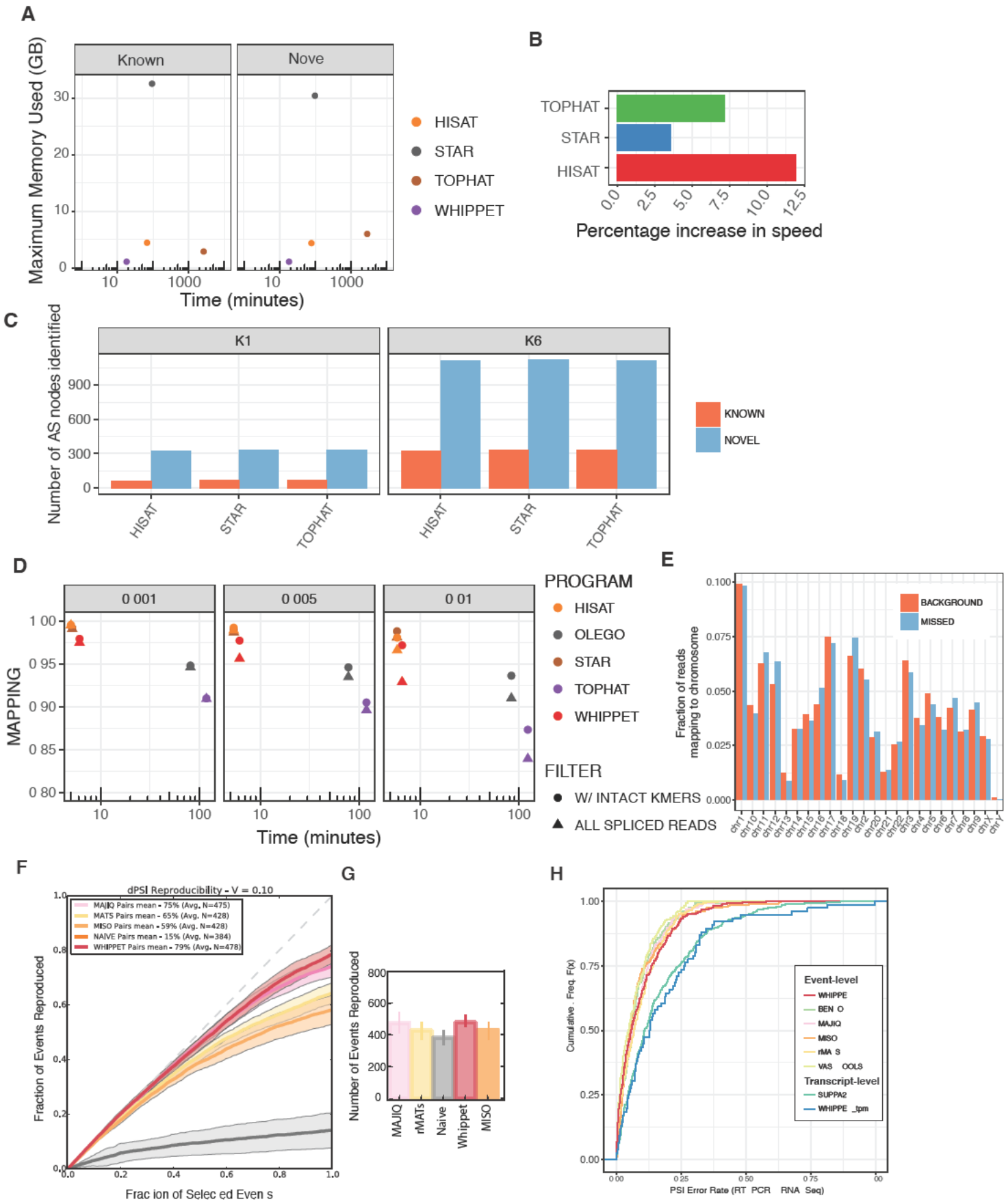
A
Known
Nove
Maximum Memory Used (GB)
Time (minutes)
HISAT
STAR
TOPHAT
WHIPPET
B
TOPHAT
STAR
HISAT
Percentage increase in speed
C
K1
K6
Number of AS nodes identified
KNOWN
NOVEL
D
0 001
0 005
0 01
MAPPING
Time (minutes)
PROGRAM
HISAT
OLEGO
STAR
TOPHAT
WHIPPET
FILTER
W/ INTACT KMERS
ALL SPLICED READS
E
BACKGROUND
MISSED
Fraction of reads mapping to chromosome
F
dPSI Reproducibility - V = 0.10
Fraction of Events Reproduced
Frac ion of Selec ed Even s
G
Number of Events Reproduced
MAJIQ
rMATs
Naive
Whippet
MISO
H
Cumulative Freq. F(x)
Event-level
Transcript-level
PSI Error Rate (RT PCR RNA Seq)

Figure S2

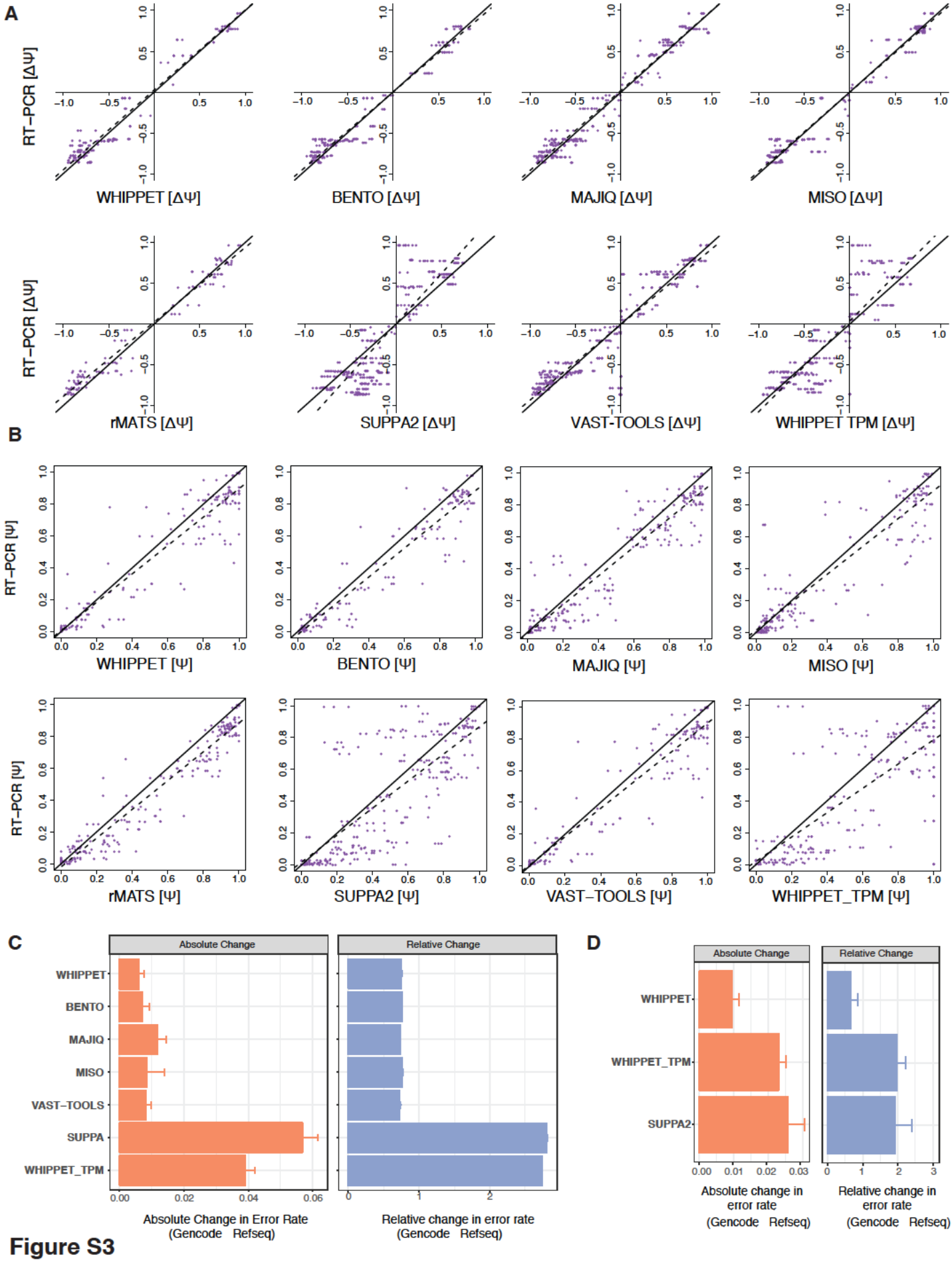
A
RT−PCR [ΔΨ]
WHIPPET [ΔΨ]
BENTO [ΔΨ]
MAJIQ [ΔΨ]
MISO [ΔΨ]
rMATS [ΔΨ]
SUPPA2 [ΔΨ]
VAST-TOOLS [ΔΨ]
WHIPPET TPM [ΔΨ]
B
RT−PCR [Ψ]
WHIPPET [Ψ]
BENTO [Ψ]
MAJIQ [Ψ]
MISO [Ψ]
rMATS [Ψ]
SUPPA2 [Ψ]
VAST−TOOLS [Ψ]
WHIPPET_TPM [Ψ]
C
Absolute Change
Relative Change
WHIPPET
BENTO
MAJIQ
MISO
VAST−TOOLS
SUPPA
WHIPPET_TPM
Absolute Change in Error Rate
(Gencode Refseq)
Relative change in error rate
(Gencode Refseq)
D
Absolute Change
Relative Change
WHIPPET
WHIPPET_TPM
SUPPA2
Absolute change in
error rate
(Gencode Refseq)
Relative change in
error rate
(Gencode Refseq)

Figure S3

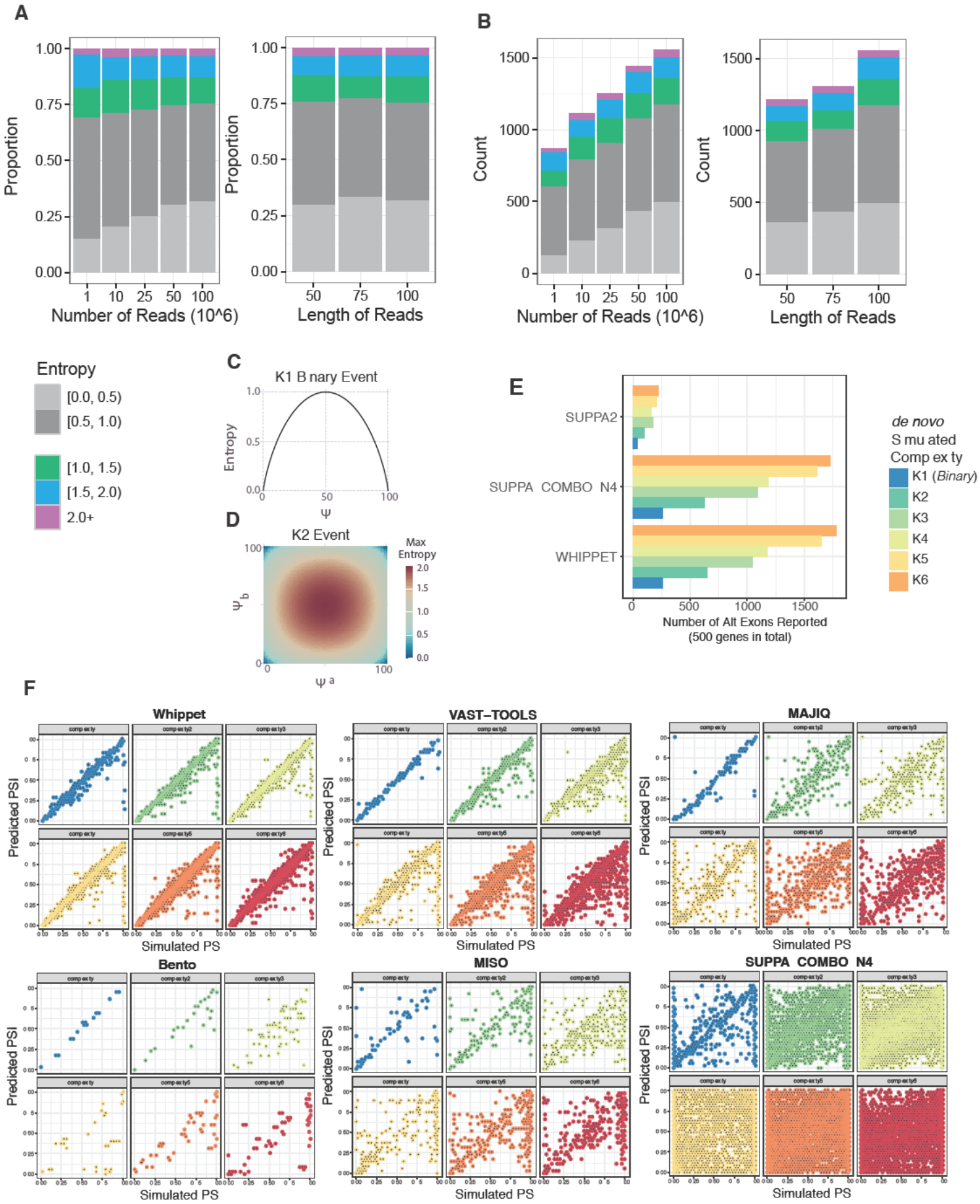
A
Proportion
Number of Reads (10^6)
Length of Reads
B
Count
Number of Reads (10^6)
Length of Reads
Entropy
[0.0, 0.5)
[0.5, 1.0)
[1.0, 1.5)
[1.5, 2.0)
2.0+
C
K1 B nary Event
Entropy
Ψ
D
K2 Event
Max Entropy
Ψ b
Ψ a
E
SUPPA2
SUPPA COMBO N4
WHIPPET
Number of Alt Exons Reported
(500 genes in total)
de novo
S mu ated
Comp ex ty
K1 (Binary)
K2
K3
K4
K5
K6
F
Whippet
VAST–TOOLS
MAJIQ
Bento
MISO
SUPPA COMBO N4
Predicted PSI
Simulated PS

Figure S4

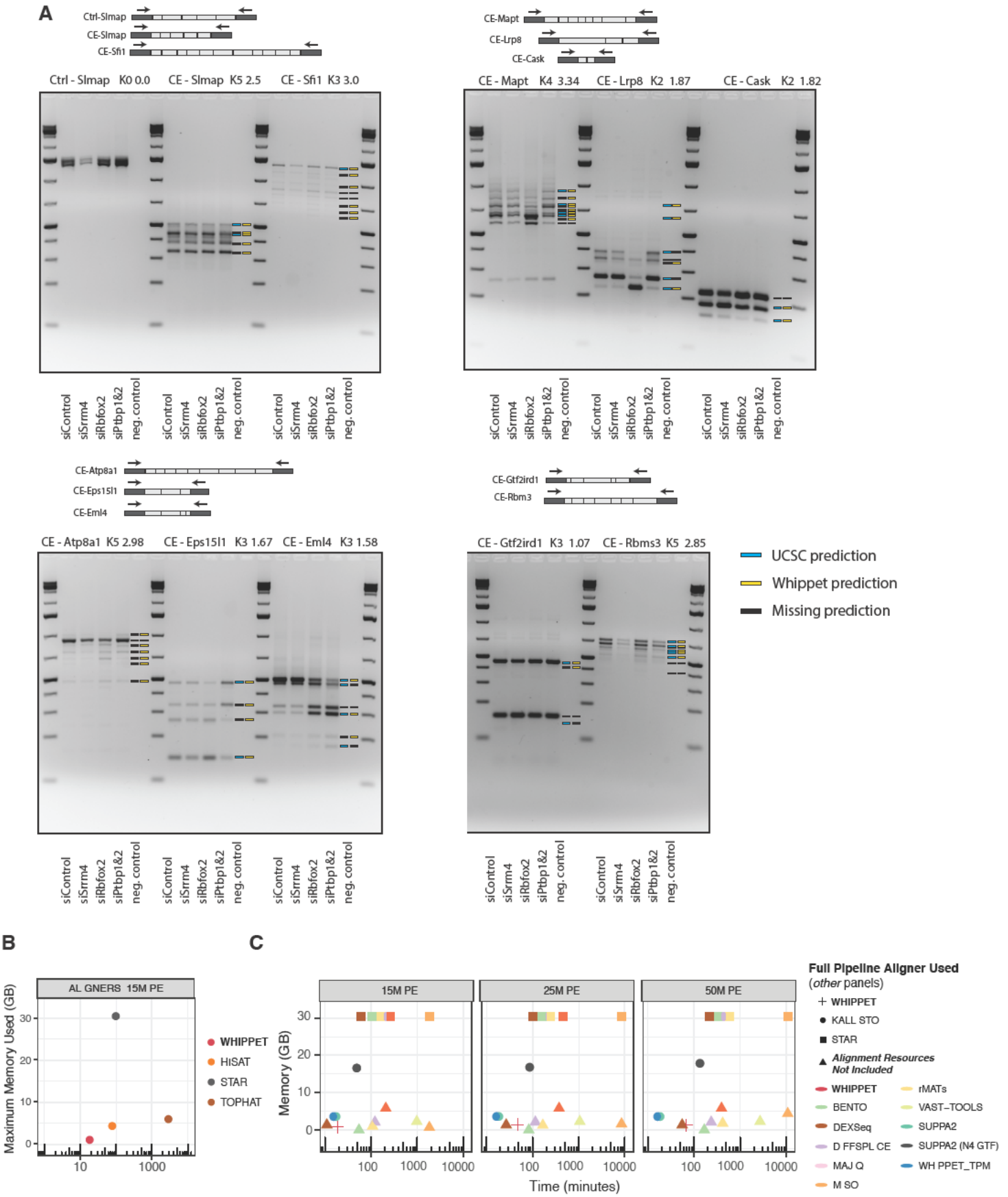
A
Ctrl-Slmap
CE-Slmap
CE-Sfi1
Ctrl - Slmap K0 0.0
CE - Slmap K5 2.5
CE - Sfi1 K3 3.0
CE-Mapt
CE-Lrp8
CE-Cask
CE - Mapt K4 3.34
CE - Lrp8 K2 1.87
CE - Cask K2 1.82
siControl
siSrrm4
siRbfox2
siPtbp1&2
neg.control
CE-Atp8a1
CE-Eps15l1
CE-Eml4
CE - Atp8a1 K5 2.98
CE - Eps15l1 K3 1.67
CE - Eml4 K3 1.58
CE-Gtf2ird1
CE-Rbm3
CE - Gtf2ird1 K3 1.07
CE - Rbms3 K5 2.85
UCSC prediction
Whippet prediction
Missing prediction
B
AL GNERS 15M PE
Maximum Memory Used (GB)
WHIPPET
HISAT
STAR
TOPHAT
C
15M PE
25M PE
50M PE
Memory (GB)
Time (minutes)
Full Pipeline Aligner Used
(other panels)
WHIPPET
KALL STO
STAR
Alignment Resources Not Included
WHIPPET
BENTO
DEXSeq
D FFSPL CE
MAJ Q
M SO
rMATs
VAST–TOOLS
SUPPA2
SUPPA2 (N4 GTF)
WH PPET_TPM

Figure S5

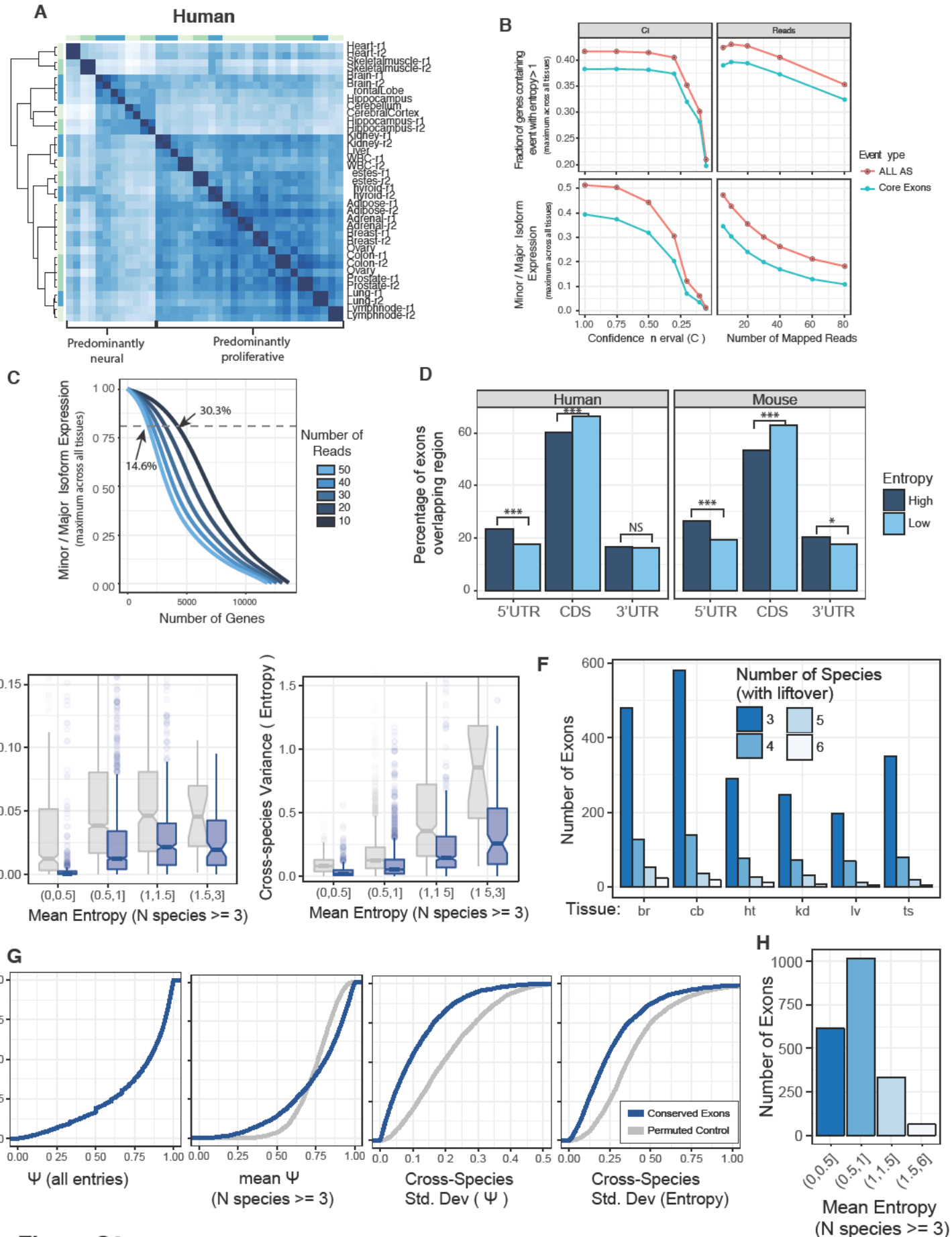
A
Human
Predominantly neural
Predominantly proliferative
B
Confidence n erval (C)
Number of Mapped Reads
Event ype
ALL AS
Core Exons
C
30.3%
14.6%
Number of Reads
Number of Genes
D
Human
Mouse
Entropy
High
Low
5'UTR
CDS
3'UTR
E
Mean Entropy (N species >= 3)
F
Number of Species (with liftover)
Number of Exons
Tissue:
br
cb
ht
kd
lv
ts
G
Ψ (all entries)
mean Ψ (N species >= 3)
Cross-Species Std. Dev (Ψ)
Cross-Species Std. Dev (Entropy)
Conserved Exons
Permuted Control
H
Mean Entropy (N species >= 3)

Figure S6

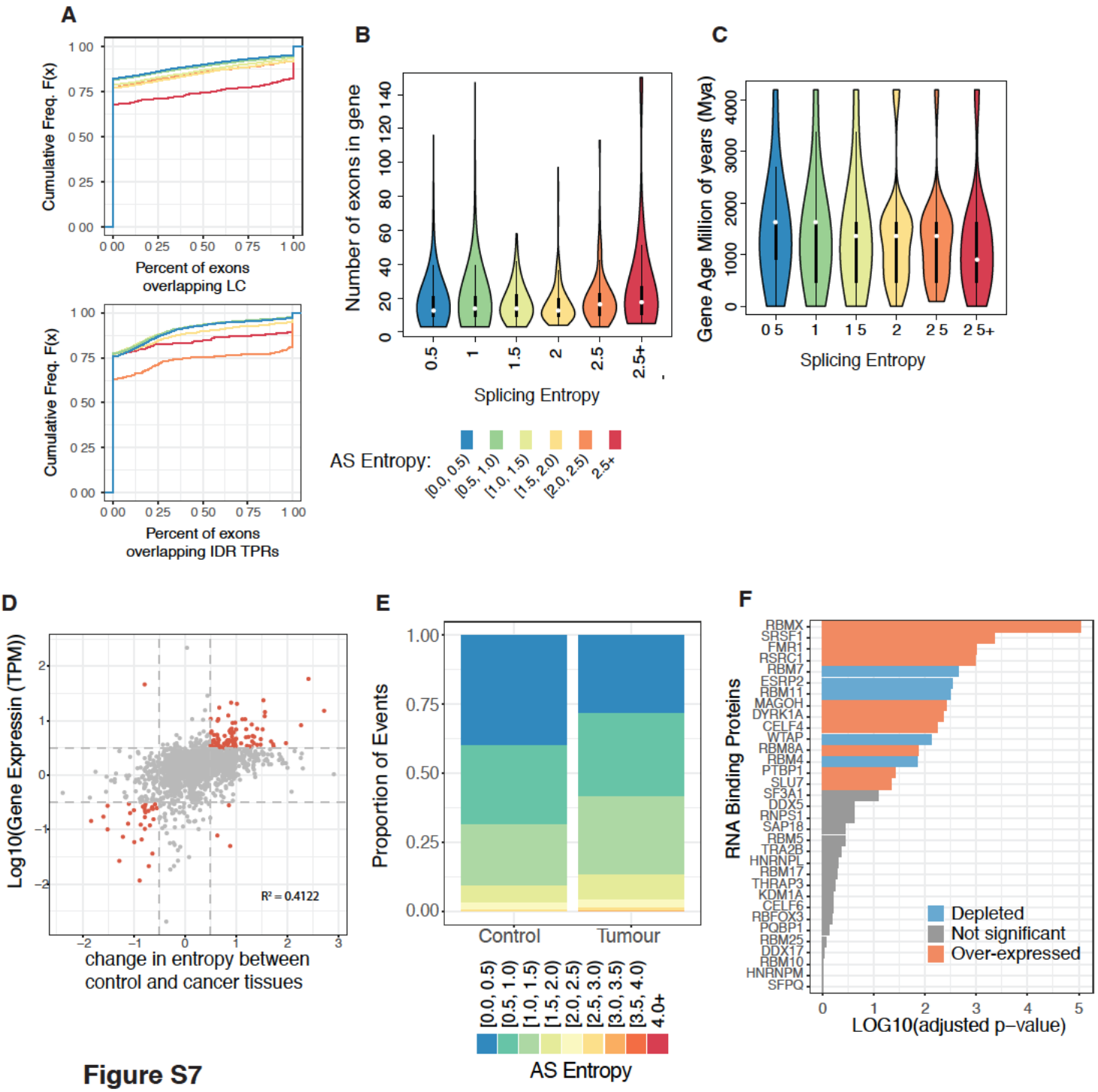
A
Cumulative Freq. F(x)
Percent of exons overlapping LC
Percent of exons overlapping IDR TPRs
B
Number of exons in gene
Splicing Entropy
AS Entropy:
[0.0, 0.5)
[0.5, 1.0)
[1.0, 1.5)
[1.5, 2.0)
[2.0, 2.5)
2.5+
C
Gene Age Million of years (Mya)
Splicing Entropy
D
Log10(Gene Expressin (TPM))
R² = 0.4122
change in entropy between control and cancer tissues
E
Proportion of Events
Control
Tumour
[0.0, 0.5)
[0.5, 1.0)
[1.0, 1.5)
[1.5, 2.0)
[2.0, 2.5)
[2.5, 3.0)
[3.0, 3.5)
[3.5, 4.0)
4.0+
AS Entropy
F
RNA Binding Proteins
RBMX
SRSF1
FMR1
RSRC1
RBM7
ESRP2
RBM11
MAGOH
DYRK1A
CELF4
WTAP
RBM8A
RBM4
PTBP1
SLU7
SF3A1
DDX5
RNPS1
SAP18
RBM5
TRA2B
HNRNPL
RBM17
THRAP3
KDM1A
CELF6
RBFOX3
PQBP1
RBM25
DDX17
RBM10
HNRNPM
SFPQ
Depleted
Not significant
Over-expressed
LOG10(adjusted p−value)

Figure S7