RBP-Maps enables robust generation of splicing regulatory maps

Brian A Yee[1,2], Gabriel A Pratt[1,2,3], Brenton R Graveley[4], Eric L Van Nostrand[*1,2], Gene W Yeo[*,1,2,3]

1. Department of Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA
2. Institute for Genomic Medicine, University of California at San Diego, La Jolla, CA
3. Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA
4. Department of Genetics and Genome Sciences, Institute for Systems Genomics, UConn Health, Farmington, CT

[*]Correspondence should be addressed to Eric Van Nostrand (eric.vannostrand@gmail.com) and Gene Yeo (geneyeo@ucsd.edu).

Abstract

Alternative splicing of pre-messenger RNA transcripts enables the generation of multiple protein isoforms from the same gene locus, providing a major source of protein diversity in mammalian genomes. RNA binding proteins (RBPs) bind to RNA to control splice site choice and define which exons are included in the resulting mature RNA transcript. However, depending on where the RBPs bind relative to splice sites, they can activate or repress splice site usage. To explore this position-specific regulation, *in vivo* binding sites identified by methods such as cross-linking immunoprecipitation (CLIP) are integrated with alternative splicing events identified by RNA-seq or microarray. Merging these datasets enables the generation of a 'splicing map', where CLIP signal relative to a merged meta-exon provides a simple summary of the position-specific effect of binding on splicing regulation. Here, we provide **RBP-maps,** a software tool to simplify generation of these maps and enable researchers to rapidly query regulatory patterns of an RBP of interest. Further, we discuss various alternative approaches to generate such splicing maps, focusing on how decisions in construction (such as the use of peak versus read density, or whole-reads versus only single-nucleotide candidate crosslink positions) can affect the interpretation of these maps using example eCLIP data from the 150 RBPs profiled by the ENCODE consortium.

Introduction

After RNA is transcribed from DNA, intronic regions are removed and exons are joined together in the process of splicing. Most exons are constitutively spliced, meaning they are always included in the mature RNA transcript that is ultimately translated. However, recent estimates indicate that nine out of every ten human genes undergo alternative splicing in which alternative splice sites are utilized in a cell type- or condition-specific manner to create distinct RNA transcripts from the same pre-mRNA molecule (Wang et al. 2008). The key role of alternative splicing is further confirmed by the linkage of splicing regulation to numerous human diseases, including neurological disorders and many types of cancer (Scotti and Swanson 2016). Thus, understanding the regulatory patterns that control alternative splicing can give valuable insights into a variety of biological systems.

RNA binding proteins (RBPs) interact with RNA through recognition of sequence motifs, structures, and combinations thereof to regulate condition-specific alternative splicing. Thus, identifying the direct *in vivo* targets of RBPs can give insight into their mechanism of regulation. Most commonly, this is done through cross-linking and immunoprecipitation (CLIP), which pulls down an RBP of interest along with its bound RNA (Lee and Ule 2018). However, although *in vivo* targets in isolation can yield insights into potential roles for an RBP, integration of this data with RBP-responsive targets allows the identification of directly regulated targets, which can provide a deeper understanding of the mechanisms of regulation by an RBP. For regulation of alternative splicing, where RBP binding can cause either inclusion or exclusion of alternative exons, it is common to identify RBP-responsive events by knocking down or over-expressing the RBP and performing RNA-seq. Following sequencing, several algorithms have been

developed to discover changes in splicing among transcripts between conditions (Katz et al. 2010; Shen et al. 2014). These algorithms detect common splicing events, including skipped exons (SE), alternative 3' and 5' splice sites (A3SS, A5SS), retained introns (RI) and mutually exclusive exons (MXE), all of which contribute to increased diversity of the human proteome.

In addition to simple overlaps between the lists of RBP-responsive events and RBP-bound regions, it has become common to specifically query how the positional dependence of binding differentially effects alternative splicing of nearby events. Visualization of this location-dependent splicing regulatory information is often referred to as a "splicing map," which has become an important tool to visualize RNA binding activity over a collective set of genomic regions (Witten and Ule 2011). For a given splice type, a meta event is typically shown to visualize global binding across alternatively spliced exons containing a composite signal across a set of events. These meta-events are often comprised of vectorized windows aligned to the splice site and the flanking proximal exon/intron region (Yeo et al. 2009; Lim et al. 2011; Hauer et al. 2015). For example, skipped exon (SE) events are typically represented as four windows showing all upstream splice sites, all 5' and 3' splice sites of the cassette exon, and downstream splice sites, plus corresponding surrounding regions (usually 50nt into the exon, 200-300nt into the intron) (Xue et al. 2009; Witten and Ule 2011; Cereda et al. 2014; Park et al. 2016). Analogous approaches can be used to visualize the other types of splicing events as well, focusing on the splice site regions associated with the meta-event type.

In particular, splicing maps which overlap RBP binding with alternatively spliced events from RNA-seq provides insight into how RBPs can regulate these events differently depending on where they associate (Witten and Ule 2011). Early on, these maps were made by using motif enrichment as a proxy for RBP association, showing for example that the RBFOX family of RBPs appeared to encourage exon inclusion if associated downstream but increased exon exclusion if associated upstream (Yeo et al. 2009). However, these approaches are limited to RBPs with well-characterized binding motifs, which remains a small subset of RBPs overall. The use of CLIP to profile protein-RNA interactions directly have led to a rapid expansion of this area, with maps generated for a variety of RBPs that reveal a complex regulatory specificity for RBPs based on their location of binding (Witten and Ule 2011). However, due to the variety of CLIP-seq technologies and their current limitations, there remains a lack of consensus on numerous details regarding the calculations underlying the generation of such maps (Wheeler et al. 2017; Lee and Ule 2018).

Recently, the ENCODE project published eCLIP datasets for 150 RBPs across K562 and HepG2 cell types, as well as identification of alternative splicing for shRNA knockdown of 263 RBPs in HepG2 and K562 cell types (Van Nostrand et al. 2016; Van Nostrand et al. 2017a). As part of this effort to perform integrated analyses of RNA processing to map splicing regulatory patterns, we observed that many decisions could significantly alter the interpretation of the subsequent splicing map. Here, we describe RBP-Maps (https://github.com/yeolab/rbp-maps) a robust software tool to standardize the generation of these maps from ENCODE and other datasets in order to enable integration of CLIP and RNA-seq for non-expert users. Further,

we discuss numerous analysis options enabled by this tool, and how these decisions can shape the downstream generated splicing map.

Results

Generation of splicing regulatory maps with RBP-Maps

To enable simplified generation of splicing regulatory maps for the ENCODE eCLIP and RNA-seq datasets, we developed the RBP-maps software package. At its core, the program intersects a CLIP dataset provided by the user (either read densities in the form of bigwig coverage files or peaks in the form of bed files) and intersects them with any number of user-defined alternative splicing event files (Fig. 1a). The program then outputs a normalized summary figure displaying the average signal across all events (Fig. 1b), including data matrices containing the raw and normalized signal values for each event, as well as the mean across all events, for each position in the meta-exon map as a comma separated file in order to facilitate further downstream processing. The RBP-Maps package is publicly available (https://github.com/yeolab/rbp-maps), and contains details regarding installation setup requirements, usage, and examples for different alternative event types (including cassette or skipped exons, alternative 5' and 3' splice site usage, and retained introns).

As an essential component of this software, we have also provided a number of additional options, including read density normalization, window size, density outlier removal options, statistical significance calculation, and incorporation of multiple background event lists. In the following sections, we discuss how each of these options can affect the resulting splicing map and provide recommendations for usage.

Avoidance of duplication within RBP-responsive events

The first decision in the generation of a splicing map is the selection of set(s) of alternative events, which is specified by the '--annotations' option. By default, alternative splicing input files are in the RMATS JunctionCountsOnly.txt file format (for SE, RI, A3SS, A5SS, and MXE event types), although support is also available for MISO format (Katz et al. 2010; Shen et al. 2014). Any number of event lists can be provided, each of which will be separately processed and plotted together in the subsequent splicing map. This enables the user to include not only the experimental event set (for example, events included or excluded upon RBP knockdown), but also various control sets of constitutive or alternative exons not responsive to the RBP for comparison purposes (see further discussion below).

We have found that outputs from many standard RNA-seq splicing analysis tools require pruning of events in order to avoid duplication of CLIP signals in the resulting splicing map, as some software reports multiple 'events' for the same gene that in fact overlap (often due to these events sharing one (or multiple) exon-exon junctions). As a consequence, distinct splicing events can share genomic coordinates, which would result in integrating the same eCLIP signal multiple times in the subsequent regulatory map. We observed that these overlapping events accounted for an average of 22% of the total number of events reported by RMATS among all submitted ENCODE datasets (Fig. 2a), suggesting that there could be significant double- (or

more) counting of single eCLIP peaks if overlapping events were not removed. Therefore, we group overlapping events and select the event with the highest inclusion junction count as the exemplar to be incorporated into the splicing map using the included subset_rmats_junctioncountsonly.py script (Fig. 2b). We observed that the most common source of these events were exons which shared exclusion junctions and had variable 5' or 3' splice sites, many of which had extremely low inclusion levels.

To show the effect on an example splicing map, we considered ZC3H8 in K562 cells. Plotting Z3CH8 CLIP against an unfiltered set of 239 events identified as significantly included (change in Percent Spliced In ($|\Delta\Psi|$) ≥ 0.05, FDR ≤ 0.1 and $p \leq 0.05$) upon knockdown of ZC3H8 causes a peak in the global signal in the proximal upstream intron of the meta-downstream exon (Fig. 2c). However, this appears to be the result of intersecting CLIP signal multiple times across 50 overlapping events, as this peak was no longer observed when we removed overlapping events. Thus, we have found that such double-counting of eCLIP signal can cause artifacts in splicing maps if not properly accounted for.

Reads versus peaks

There are two major alternative approaches to how CLIP signal is utilized in a splicing map: as either read density (in various processed or normalized forms) or as the density of significantly enriched peaks. To enable researchers to implement both of these approaches, RBP-Maps can run in two modes: --peak mode (which takes a bigbed file that describes significantly enriched regions of CLIP signal identified from any standard CLIP analysis toolkit), and --density mode (which accepts read densities formatted as two standard bigwig files, one for each strand).

Conversion of read density into computationally identified peaks or clusters, using one of a variety of peak-calling algorithms, is a standard step of CLIP analysis (De and Gorospe 2017). The use of peaks provides two appealing benefits for simplified creation of splicing maps. First, because peaks identify regions where IP signal is significantly enriched over a background model, they mitigate noise in read density signal by focusing on regions of significant enrichment (Park et al. 2016). Second, by compressing CLIP signal to a single binary value indicating the presence of a peak at each position, each event is weighted equally in the resulting average signal trace, removing the need for further normalization of the CLIP signal to control for relative abundance or differential enrichment.

For peak-based maps, a count of peaks that overlap alternatively spliced regions is plotted as a histogram at every position, with the final value as the fraction of events that contain a peak at each position (i.e. the total count divided by the number of regions). Considering the ENCODE dataset, we observe that a subset of RBPs show clear splicing maps based on peak density alone: for example, RBFOX2 shows enrichment for peaks downstream of the 5' splice site of knockdown-excluded exons (Fig. 2d). Thus, for some sufficiently deeply sequenced CLIP datasets from proteins with distinct binding patterns, peak-based maps offer the most succinct way of integrating CLIP and alternative splicing data (Fig. 2d). These RBPs typically bind

directly to, or near splice sites and provide high position-specific overlap between splice events and CLIP peak regions, yielding a consistent signal across the meta-event.

However, we noted that other RBPs had results that varied between peak- and read-based maps. For example, a map based on read density for SRSF9 in HepG2 cells showed enrichment at knockdown-excluded exons, consistent with the general role of SR proteins as enhancing exon inclusion (Ibrahim et al. 2005). However, a peak-based map provides limited insight, as only 4 knockdown-excluded exons are overlapped by a significantly enriched reproducible eCLIP peak (Fig. 2e). This is a common occurrence among ENCODE datasets, as the mean percentage of peaks intersecting RBP knockdown-altered skipped exon regions (3' end of the upstream exon to the 5' end of the downstream exon) is less than 1 percent (with a slightly higher average of 1.3% and 1.9% for known splicing regulators and spliceosomes, respectively), limiting the power of the peak-based approach (Fig. 2f). Even decreasing the stringency threshold for peak identification from 8-fold to only 2-fold enriched in IP versus input background yields only a slight increase in the median number of peaks overlapping events from 4 to 12 (Fig. 2g). Thus, even though peak-based maps are often simpler (both conceptually as well as computationally), read density-based maps remain highly useful due to increased signal (Fig. 2d). In both cases, 50 to 100 or more alternatively spliced events are generally required to yield robust maps.

Read density-based approaches: normalization

Splicing regulatory maps based off of read density are generated by RBP-Maps --density mode, in which the user data is provided as bigwig format read density files (one for each strand) for both IP and (if available) paired input (or other control) experiments. However, read density does not inherently include normalization against background or provide regions of enrichment as compared to peaks. Therefore, we have made three CLIP density normalization options available as part of --density mode in RBP-Maps: "raw" values (option [0]), which illustrates a splicing map using just CLIP read densities (and is the same method used for peak-based maps), "subtraction" normalization ((option [1], default for density), which subtracts normalized size-matched input read densities from its corresponding CLIP IP read density, and "entropy" normalization (option [2]), which calculates information content-based fold enrichment of CLIP read density over corresponding input.

We observe that in some instances, density maps look similar regardless of normalization method. This is particularly true in cases where size-matched input background is low, as is the case for intronic-binding RBPs that are often being profiled in studies of splicing regulatory networks. For example, subtracting input from CLIP signal from an RBFOX2 skipped/cassette exon map yields little differences in peak position, aside from changes in scaling (Fig. 3a). However, experiments that include a size-matched input can leverage this information to correct for common background artifacts, including the typical observation of non-enriched read density at abundant exonic regions (Van Nostrand et al. 2016). Indeed, applying different normalization methods to an HNRNPK splicing map does change the shape of binding upstream and downstream of the cassette exon (Fig. 3b, c, e, f). Thus, although using IP read

density only can yield reasonable splicing regulatory maps, incorporation of a paired input is recommended.

To consider the effect of normalization methods, we compared two strategies: background subtraction and entropy-based enrichment (Fig. 3d). The subtraction method first calculates the difference between density values of the IP and its corresponding input, then scales these values for each event to sum to one, equalizing each region's contribution to the overall splicing map, similar to existing normalization methods (Licatalosi et al. 2008) . This prioritizes the global relative shape of binding enrichment (Fig. 3e). In contrast, we tried a second method in which we did not normalize per event, but instead calculated the entropy (or relative information content) in IP versus input at each position for each event as $p_i \times log_2(\frac{p_i}{q_i})$, where p_i and q_i are the fraction of total reads in IP and input respectively that map overlapping position *i*. The final averaged map was then calculated as the position-wise mean over these information scores across all events (Fig. 3f). This information content maintains the strength of binding, meaning that events with greater read density will dominate the final average.

As expected, we found that the summarized map using the entropy-based method would often be highly dominated by highly abundant CLIP signals at only a small number of events (Fig. 3g). In contrast, the subtraction method proved to be an effective approach, yielding more robust signals than peak-based maps while being more resistant to over-weighting single events (Fig. 3h). Thus, the subtraction method provides a mechanism to incorporate paired size-matched input (or other control datasets) into the standard read density-based splicing regulatory map framework.

Outlier removal

Particularly with the relative information content method, we observed that individual highly abundant positions at single events could dominate the composite signal. Manual inspection suggested that these typically arose from snRNAs, miRNAs, and other multi-copy or highly abundant transcripts or pseudogenes present within these intronic regions. For example, we observed a single site of significant enrichment approximately 250bp downstream of knockdown-excluded skipped exons in HNRNPC splice maps (Fig. 4a). Upon further inspection, we noticed that this signal came exclusively from a single event near a snoRNA (Fig. 4b). To address this, we performed outlier removal on the top and bottom 2.5% signal at each position across each splicing map, which removed extreme outliers and revealed signal consistent with the splicing-repressive role of HNRNP proteins (Fig. 4c). While keeping the middle 95% of values appears to work in removing these artifacts in most ENCODE datasets, the (--conf) parameter can be adjusted to define an alternative outlier threshold. Although this was more critical in generating reliable maps using the relative information metric, we found that it also tended to decrease noise when using the background-subtraction method as well.

Choice of background events for comparison

Interpretation of a splicing map requires the use of some sort of background control in order to contrast binding of the RBP around RBP-responsive exons to a set of non-responsive

ones. Many studies have indicated that the typical alternative exon is fundamentally different from a typical constitutively spliced exon, with altered exon and intron size, weaker 5' and 3' splice sites, higher sequence conservation, and higher RBP binding (Yeo et al. 2005). Thus, although the process of selection of these background events is often treated as so basic as to not warrant further discussion in publications, we have found that the selection of a proper background for comparison is essential for proper interpretation. To explore the effect of comparison against different event backgrounds, we generated five sets of control exons: constitutive exons (which that had no exclusion observed in any of 29 scrambled shRNA control RNA-seq datasets in HepG2 or 29 in K562), 'native' cassette exons that were alternatively spliced under normal conditions (0.05 < inclusion < 0.95 in at least half of control RNA-seq datasets), and three subgroups of native events: 'included native' (inclusion > 0.67 in at least half of control datasets), 'central native' (0.33 < inclusion < 0.67 in at least half of control datasets), and 'excluded native' (inclusion < 0.33 in at least half of control datasets).

As an example of the effect of background choice, we considered the splicing regulatory maps of Serine and Arginine Rich Splicing Factor 1 (SRSF1). We observe that SRSF1 eCLIP signal is higher at exons excluded upon SRSF1 knockdown than those included upon SRSF1 knockdown, consistent with its known role in exon inclusion (Ibrahim et al. 2005; Zhou and Fu 2013) (Fig. 5a). Next, considering SRSF1 eCLIP signal at these different background event lists, we observed a clear pattern where exons with higher average inclusion (constitutive or native included groups) had higher SRSF1 eCLIP signal, whereas those with lower inclusion (native excluded) had far less SRSF1 eCLIP signal. Further, we observed that whereas exons excluded upon SRSF1 knockdown had higher SRSF1 eCLIP signal relative to any of the four cassette exon classes, they had lower SRSF1 signal than constitutive exons (Fig. 5b). This clearly demonstrates the impact of background choice, as if the background selected was largely composed of constitutive exons one might believe that SRSF1 is depleted at knockdown-excluded events, whereas the use of a native cassette exon background indicates enriched SRSF1 binding at knockdown-excluded events. As the latter conclusion is better supported by the differences between alternative and constitutive exons more broadly as well as previous knowledge about the exon inclusion promoting role of SRSF1, we believe (based on this and other examples) that using a background of native alternative exons is preferred in nearly all situations.

Statistical significance models

Once the proper background has been selected, RBP-Maps can test up to two conditions (i.e., significantly included and significantly excluded cassette events) and show position-wise significance against an indicated background using the --sigtest and --bgnum options (which select the 0-indexed number order corresponding to the events to test and the set of events to use as a background respectively). Different models are used for the peak-based and density-based approaches. For peak-based maps, a Fisher's exact test is used at each position along the meta-event to test whether the fraction of events with a peak at that event is significantly altered relative to the selected background using the '--sigtest fisher' option (Fig. 6a). For density-based maps, users can perform a Kolmogorov-Smirnov test (--sigtest ks),

which provides users a way to visualize significance as a heatmap of p-values (Fig. 6b). However, we found that this test was not ideal as it tended to yield false positive significance for datasets with many altered events, and conversely was poor at identifying positions (such as the +67 position for RBFOX2) where many events showed no change but a subset showed dramatic change (Fig. 6b) . Therefore, we implemented an additional non-parametric test (--sigtest permutation) by performing a random sampling (*n*=1000) of a chosen background (typically the native cassette exon set). This allows users to generate confidence bounds and significance based on a null distribution of samples of alternative events, which better captures the true variability in signal (Fig. 6c).

Whole reads vs 5' read ends

During CLIP, reverse transcriptase enzymes often terminate at the site of protein-RNA crosslinking, which causes the 5’ end of reads to correspond to the site of RBP-RNA interaction (with some variability due to the positioning of available crosslinkable amino acids and bases within the binding site) (Konig et al. 2010; Van Nostrand et al. 2017b). Thus, an additional advantage to the use of read density is the ability to utilize these crosslink-diagnostic events to improve the resolution of the resulting splicing map (Konig et al. 2010; Wang et al. 2010). To test this, we re-generated splicing maps using just the 5' ends of each read, and observed variable results depending on the RBP. For example, we observed a significant increase in resolution in the splicing map for U2AF2, which resolved specifically to the intronic 3’ splice site region as opposed to overlapping the alternative exon (Fig. 7a). However, for other RBPs (even those such as RBFOX2, which has previously been shown to crosslink directly to its *in vitro* binding motif (Weyn-Vanhentenryck et al. 2014)) we observed that using 5’ read ends yielded a similar structure with dramatically increased noise relative to using whole reads (Fig. 7b). Thus, these results suggest that this method can improve resolution for some RBPs (particularly those with highly specific splice site-proximal binding), but that factors with broader crosslinking and binding patterns may suffer unacceptable loss of signal.

Discussion

The ability to profile both RNA processing and RNA binding protein association transcriptome-wide *in vivo* has revolutionized our ability to study the mechanisms of RNA processing. Integration of *in vivo* RBP targets identified by methods such as CLIP and RBP-responsive targets by knockdown or over-expression followed by RNA-seq or microarray, coupled with bioinformatics analysis techniques, has enabled the mapping of position-dependent regulatory principles for RBPs. For alternative splicing, this is typically referred to as a ‘RNA splicing map,’ which visualizes the average binding signal across RBP-responsive alternatively spliced events to simply summarize the role of that RBP on splicing regulation. Although many tools have been described to implement this approach, we found that incorporation of paired input datasets generated as part of eCLIP profiling required additional optimization. Therefore, we developed the RBP-Maps software package to enable users to implement a variety of normalization techniques and optimizations we observed to improve analysis of the ENCODE data resource.

Although this work focuses on describing the use of RBP-maps (and the associated options) with respect mapping the position-specific effect of RBP association on splicing regulation, the same approaches can be directly applied to position-specific regulation of 3' or 5' end processing, RNA stability and translation, or any other aspect of RNA processing regulated by RBPs. For example, polyadenylation analysis implicated splicing regulator NOVA in regulation of alternative polyadenylation (Licatalosi et al. 2008) and recently yielded insight into how binding of TARDBP/TDP43 shows differential regulation of alternative polyadenylation based on whether binding is close to, or further downstream of, a potential polyadenylation site (Rot et al. 2017). As it becomes easier to directly assay translation rates, RNA half-lives, and other aspects of RNA processing transcriptome-wide under RBP-modifying conditions, such RNA processing maps are likely to yield further insights into the complex regulatory code of RBP association.

Methods

Identification of significantly altered splicing events

Datasets used included 203 RBPs with both eCLIP and knockdown/RNA-seq performed in the same cell type and released by the ENCODE project at https://www.encodeproject.org (Supplementary Table 1) (Van Nostrand et al. 2017a). Alternatively spliced (AS) events were identified from rMATS JunctionCountsOnly files obtained from the ENCODE DCC (see accession identifier ENCSR413YAF for listings of all rMATS output files). Significant AS events were defined as having a p-value > 0.05, FDR > 0.1 and change in exon inclusion level (also referred to as Percent Spliced In, or $|\Delta\Psi|$) > 0.05. Elimination of overlapping splicing events was performed by identifying groups of overlapping AS events and selecting the event with the highest inclusion junction count (IJC) among the overlapped events using the bedtools (v2.26) command merge (-o collapse -c 4) and pybedtools (v0.7.9). Positive IncLevelDifference ($\Delta\Psi$) indicates that the skipped exon is more included upon RBP knockdown, while negative $\Delta\Psi$ indicates that the exon is more excluded.

Generation of control events

A number of background references for cassette exon comparisons were generated, including: 'constitutive' cassette exons defined as exons in GENCODE v19 which had no exclusion observed in any of 29 scrambled shRNA control RNA-seq datasets in HepG2 or 29 in K562 (7,351 events in HepG2 and 7,888 in K562); 'native' cassette exons defined as exons in GENCODE v19 with $0.05 < \Psi < 0.95$ in at least half of control shRNA RNA-seq datasets for that cell type (1,805 events in HepG2 and 2,222 in K562); 'included native' with inclusion > 0.67 in at least half of control datasets (1,137 events in HepG2 and 1,451 in K562); 'central native' with 0.33 < inclusion < 0.67 in at least half of control datasets (256 events in HepG2 and 292 in K562); and 'excluded native' with inclusion < 0.33 in at least half of control datasets (357 events in HepG2 and 439 in K562). All numbers reflect events remaining after removing overlapping events as described above.

Splice map generation

Multiple approaches to generating RBP splicing maps were tested. For all methods, eCLIP signal (either read density or peak presence) was first identified for 350nt windows flanking the relevant exon/intron boundaries, extending a maximum of 50nt into each exon and 300nt into each intron. For shorter exons (<100nt) and introns (<600nt), signal was only counted until the boundary of the neighboring feature. For cassette (skipped) exons, the relevant regions included the upstream, cassette, and downstream exon, creating four windows: the 3' end of the upstream exon, the 5' end of the cassette exon, the 3' end of the cassette exon, and the 5' end of the downstream exon, resulting a total vectorized region of 1400nt (350*4).

For peak based splicing maps, each position within each vectorized region was marked as 1 if it was within a peak (requiring p-value ≤ 0.001 and fold-enrichment ≥ 8 in IP versus input), and 0 otherwise. These values are then summed and divided by the total number of events at each position to obtain the final splicing map. For Figure 2F, peaks with relaxed thresholds of fold-enrichment ≥ 2 were also used.

For read density-based methods, IP and input read density (normalized as reads per million uniquely mapped, non-PCR duplicate reads) was identified at each position within the cassette exon region described above. For the background subtraction approach, input sample read density was subtracted from IP sample read density to result in difference values at every position through the event region. These values were then normalized in order to equally weigh each event by dividing the value at each position by the sum of absolute values across all 1400 positions (plus a pseudocount of 1 read, normalized to reads per million, at each position) to obtain the normalized enrichment profiles for each event. For the relative information approach, per-position information was calculated using the equation $p_i \times log_2(\frac{p_i}{q_i})$, where p_i and q_i are the per-position read probabilities at a given coordinate for IP and size matched input, respectively. To conservatively address positions with zero reads in either IP or input, a pseudocount of one read (normalized to total input read number) was added to each position before calculating IP and input read probabilities. Then, for both background subtraction and relative information approaches, values at each base across all events were sorted, removing the highest (2.5%) and lowest (2.5%) outlier values before calculating the mean across all events that is shown as the final splicing map.

To generate 5' end splicing maps, density of 5' read ends were identified using genomeCoverageBed (bedtools v2.26). Read end coverage was then used as input to the above pipeline, including background subtraction, outlier removal, and averaging across all events.

Modeling significance between RBP-responsive and native events

Significance tests for peak-based maps were computed using the fisher_exact() function based on a two by two contingency table at each position *i* based on four conditions: RBP-responsive events with peak at position *i*, RBP-responsive events without peak at position *i*, native events with peak at position *i*, native events without peak at position *i*.

Significance and confidence intervals for read density-based approaches were performed in two ways. For overall significance, a Kolmogorov Smirnov test as performed

comparing the outlier-removed normalized values for RBP-responsive events versus native events at each position using stats.ks_2samp() (Python scipy.stats module v1.1.0). To calculate significance and confidence intervals based on a bootstrapping approach, a random sample (with replacement) of *n* background events was selected, where *n* is the number of significant AS events in the test condition. These events then underwent outlier removal by filtering the top and bottom 2.5% of values, followed by calculating the mean at each position across all random events. This was repeated 1000 times to create a distribution of randomly sampled native event means at each position. By default, the 0.5^{th} and 99.5^{th} percentile values at each position were used to identify positions where RBP-responsive event maps were significantly different than native events. When multiple test conditions are present (e.g. included events and excluded events), this approach was performed separately for each, yielding a 'max' and 'min' value for each condition. For visualization of both knockdown-included and knockdown-excluded splicing maps on the same plot, the highest 'max' and lowest 'min' value was conservatively used to visualize error boundaries.

Acknowledgements

We would like to thank members of the Yeo lab for insightful comments and suggestions. This work was funded by the National Human Genome Research Institute ENCODE Project as a grant U54HG007005 to GWY and BRG. GAP is supported by the NSF graduate research fellowship. ELVN is a Merck Fellow of the Damon Runyon Cancer Research Foundation (DRG-2172-13) and is supported by a K99 grant from the NHGRI (HG009530). GWY was partially supported by grants from the NIH (HG007005, NS075449).

ELVN and GWY are co-founders and consultants for Eclipse BioInnovations Inc. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. The authors declare no other competing financial interests.

References:

Cereda M, Pozzoli U, Rot G, Juvan P, Schweitzer A, Clark T, Ule J. 2014. RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing. *Genome biology* **15**: R20.

De S, Gorospe M. 2017. Bioinformatic tools for analysis of CLIP ribonucleoprotein data. *Wiley interdisciplinary reviews RNA* **8**.

Hauer C, Curk T, Anders S, Schwarzl T, Alleaume AM, Sieber J, Hollerer I, Bhuvanagiri M, Huber W, Hentze MW et al. 2015. Improved binding site assignment by high-resolution mapping of RNA-protein interactions using iCLIP. *Nature communications* **6**: 7921.

Ibrahim EC, Schaal TD, Hertel KJ, Reed R, Maniatis T. 2005. Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 5002-5007.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods* **7**: 1009-1015.

Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology* **17**: 909-915.

Lee FCY, Ule J. 2018. Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Molecular cell* **69**: 354-369.

Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X et al. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**: 464-469.

Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 11093-11098.

Park JW, Jung S, Rouchka EC, Tseng YT, Xing Y. 2016. rMAPS: RNA map analysis and plotting server for alternative exon regulation. *Nucleic acids research* **44**: W333-338.

Rot G, Wang Z, Huppertz I, Modic M, Lence T, Hallegger M, Haberman N, Curk T, von Mering C, Ule J. 2017. High-Resolution RNA Maps Suggest Common Principles of Splicing and Polyadenylation Regulation by TDP-43. *Cell reports* **19**: 1056-1067.

Scotti MM, Swanson MS. 2016. RNA mis-splicing in disease. *Nature reviews Genetics* **17**: 19-32.

Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. 2014. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences of the United States of America* **111**: E5593-5601.

Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, Blue SM, Dominguez D, Cody NAL, Olson S et al. 2017a. A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv*.

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K et al. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods* **13**: 508-514.

Van Nostrand EL, Shishkin AA, Pratt GA, Nguyen TB, Yeo GW. 2017b. Variation in single-nucleotide sensitivity of eCLIP derived from reverse transcription conditions. *Methods* **126**: 29-37.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470-476.

Wang Z, Kayikci M, Briese M, Zarnack K, Luscombe NM, Rot G, Zupan B, Curk T, Ule J. 2010. iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS biology* **8**: e1000530.

Weyn-Vanhentenryck SM, Mele A, Yan Q, Sun S, Farny N, Zhang Z, Xue C, Herre M, Silver PA, Zhang MQ et al. 2014. HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell reports* **6**: 1139-1152.

Wheeler EC, Van Nostrand EL, Yeo GW. 2017. Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. *Wiley interdisciplinary reviews RNA*.

Witten JT, Ule J. 2011. Understanding splicing regulation through RNA splicing maps. *Trends in genetics : TIG* **27**: 89-97.

Xue Y, Zhou Y, Wu T, Zhu T, Ji X, Kwon YS, Zhang C, Yeo G, Black DL, Sun H et al. 2009. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Molecular cell* **36**: 996-1006.

Yeo GW, Coufal NG, Liang TY, Peng GE, Fu XD, Gage FH. 2009. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature structural & molecular biology* **16**: 130-137.

Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB. 2005. Identification and analysis of alternative splicing events conserved in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 2850-2855.

Zhou Z, Fu XD. 2013. Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma* **122**: 191-207.

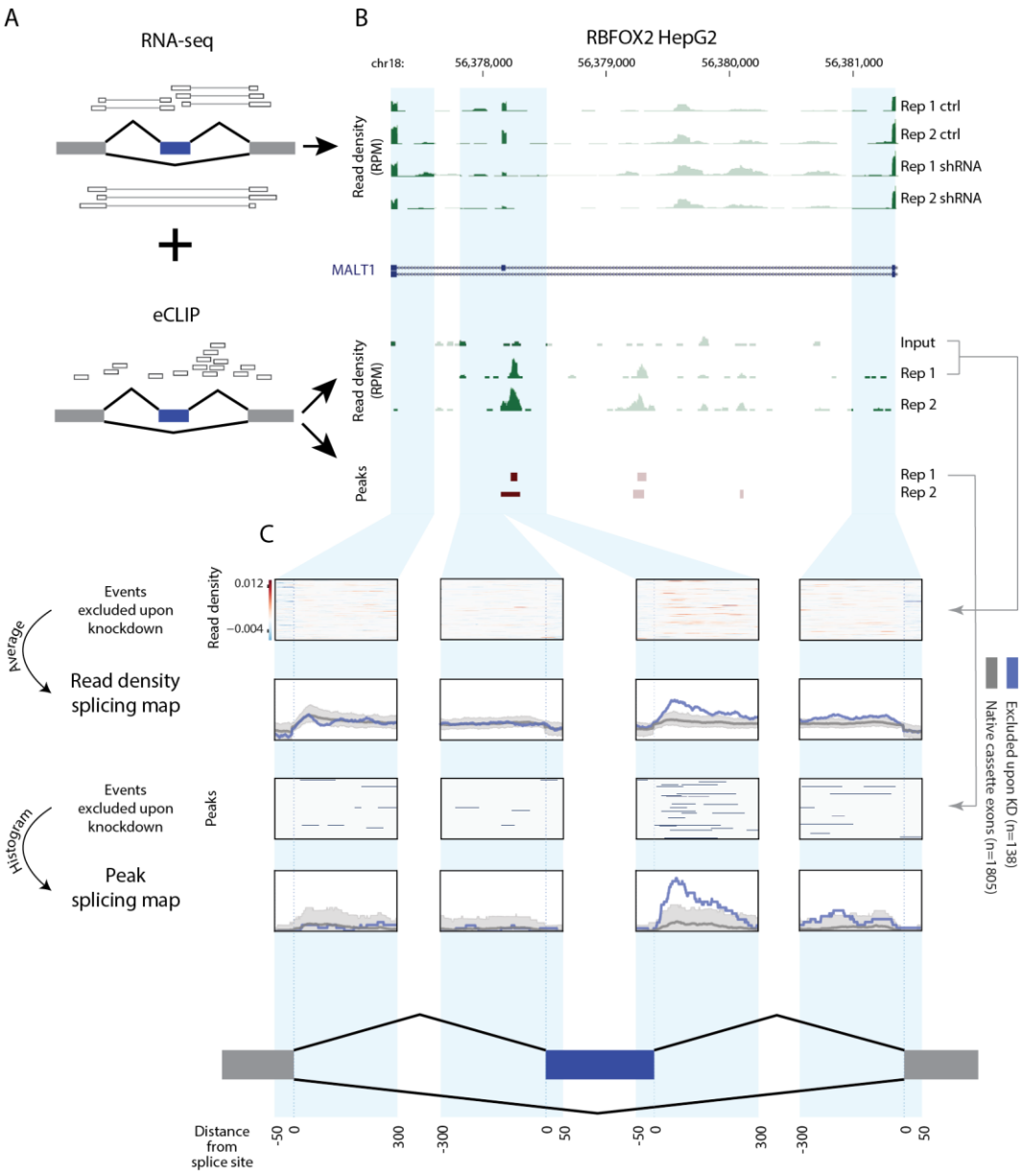

Figure 1. Splicing maps reveal position-dependent correlation between RBP binding and RBP-responsive targets. (A) Models showing (top) RNA-seq junction reads quantitating exon inclusion or exclusion and (bottom) eCLIP reads identifying 'peaks' as regions of enrichment. (B) Example derivation of a splicing map. (top) RNA-seq read density (in reads per million (RPM)) in RBFOX2 shRNA knockdown and (bottom) RBFOX2 eCLIP read density and peaks (enriched in immunoprecipitation versus paired input) for exon 7 in MALT1 (ENST00000348428.3) in HepG2 cells. (C) Integration across 138 RBP-responsive (excluded upon knockdown) events yields an averaged splicing map for (top) read density or (bottom) peak density.

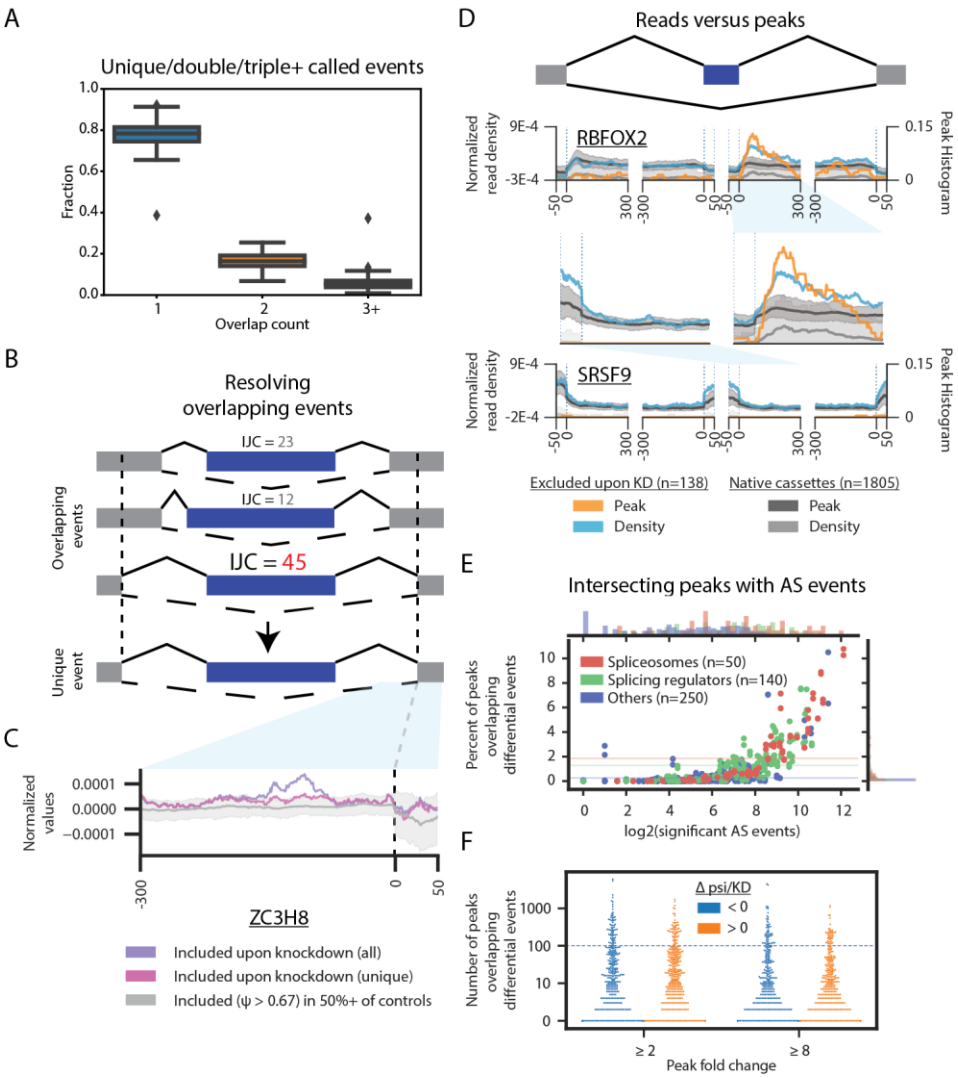

Figure 2. Event-driven options in creating splicing maps. (A) Boxplot indicates the distribution across 473 RBP knockdown RNA-seq datasets separated into included and excluded sets of events for the fraction of event regions that overlap one, two, or three or more differential event calls identified by rMATS. (B) Schematic indicates multiple overlapping events within one event region. The event with the highest inclusion junction count (IJC) is kept as the 'unique event'. (C) Example splicing map for ZC3H8 (in K562 cells) showing the difference in resulting map made by either (purple) including all rMATS-identified differential events or (pink) after discarding overlapping events. A set of natively included cassette exons which show exons which are included (≥67% percent spliced in / Ψ) in at least 50% of control RNA-seq experiments is shown in grey. (D) Splicing maps shown for (top) RBFOX2 in HepG2 and (bottom) SRSF9 in HepG2 cells. Maps made based on density of significantly enriched eCLIP peaks are shown in orange, with maps made using read density shown in blue. (E) Points indicate (x-axis) the number of significant RBP-responsive AS events versus (y-axis) the fraction with eCLIP peaks overlapping the event. Colors indicate RBP function annotations. (F) Violin plot indicates the number of RBP peaks overlapping RBP-responsive events for 203 eCLIP and

knockdown RNA-seq comparisons. Shown are distributions for peaks at least 2-fold or 8-fold enriched in IP versus paired input.

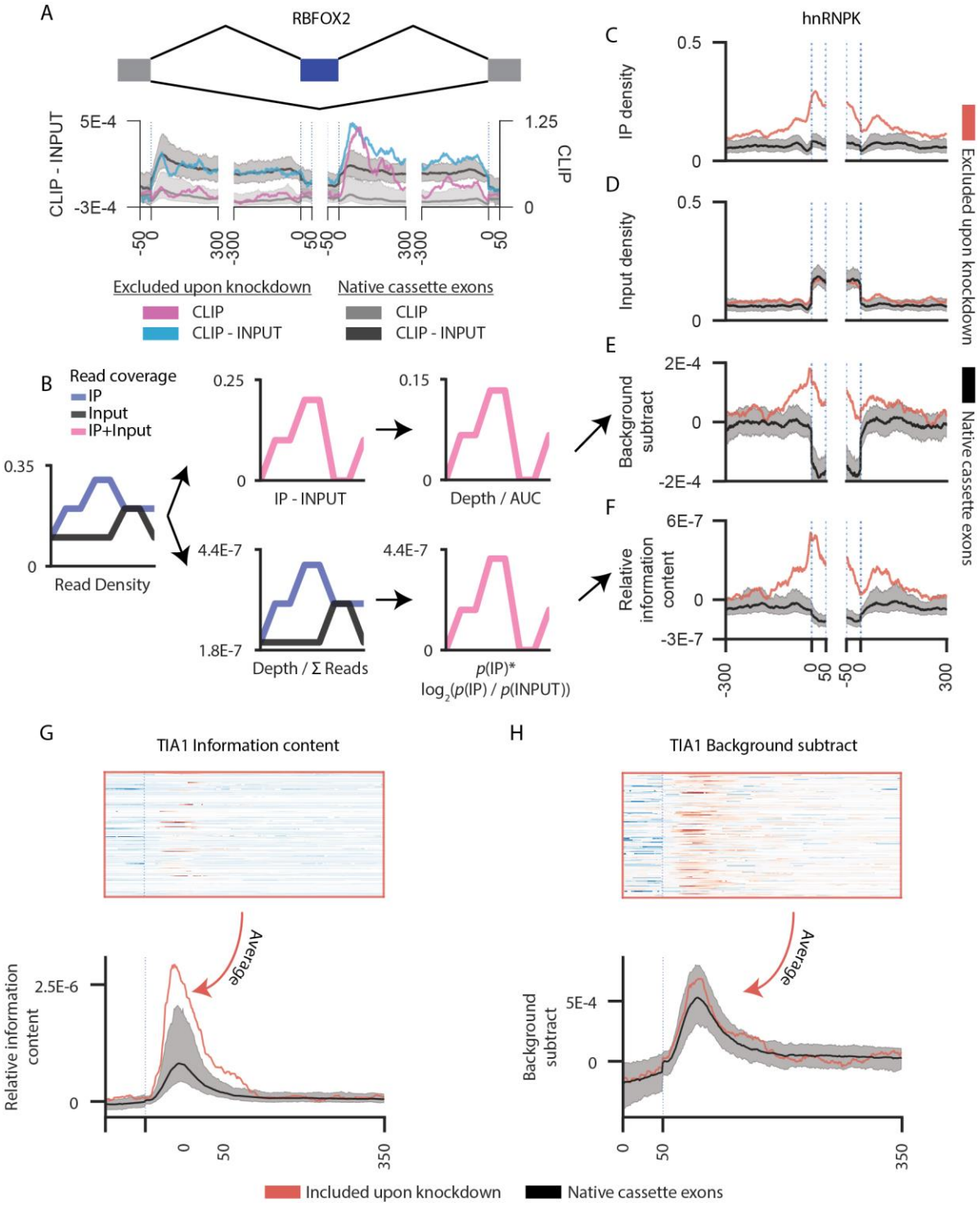

Figure 3. Strategies for normalizing read density maps against an input background. (A) Curves indicate splicing maps generated for cassette exons excluded upon RBFOX2 knockdown using either (pink) eCLIP read density alone or (blue) normalized read density after

comparing eCLIP read density versus size-matched input sample. (B) Schematic of the "background subtraction" versus the "information content" normalization for a single example event. (top) In the 'background subtraction' approach, input read density is subtracted from immunoprecipitation (IP) read density, then is normalized against area under the curve represented by read density. (bottom) in the 'information content' approach, read density is normalized to fraction of total reads in the dataset, followed by calculation of a relative information value at each position between IP and input. (C-F) Lines indicate differences observed upon generating splicing maps for excluded events upon HNRNPK knockdown using different inputs and normalization methods: (C) read density in eCLIP only, (D) read density of size-matched input only, (E) 'background subtraction' normalization, and (F) 'information content' normalization. (G) Heatmap shows (top) information content-normalized values and (bottom) corresponding average across the 5' splice site region of a meta cassette exon for TIA1 (HepG2). (H) As in G, but shown for background-subtraction normalized values.

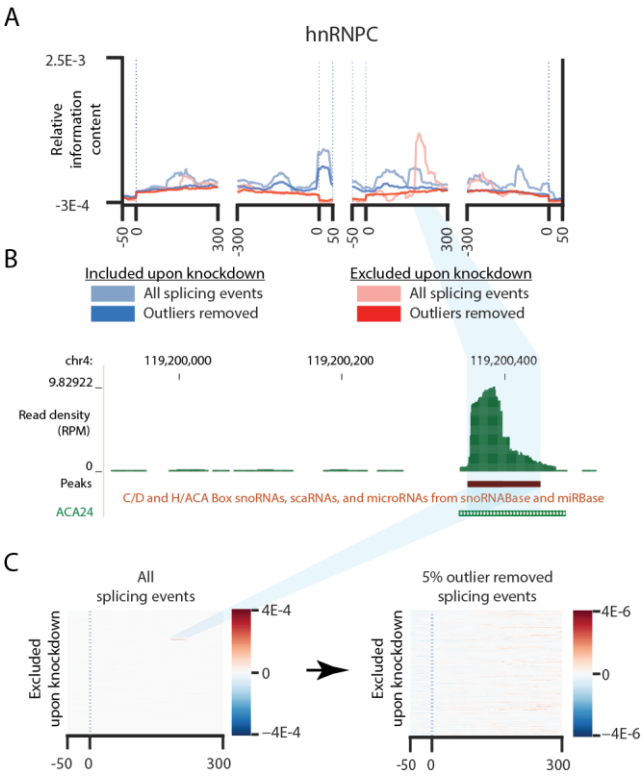

Figure 4. Removing outliers removes local artifacts that may confound global signal. (A) Figure shows splicing map of HNRNPC in HepG2 either including all events or excluding outliers (defined as the top and bottom 2.5% of values at each position). (B) Genome browser track shows an example outlier, HNRNPC HepG2 eCLIP read density at ACA24. (C) Heatmaps indicate normalized density tracks for all HNRNPC knockdown-excluded events (left) before and (right) after removal of outliers.

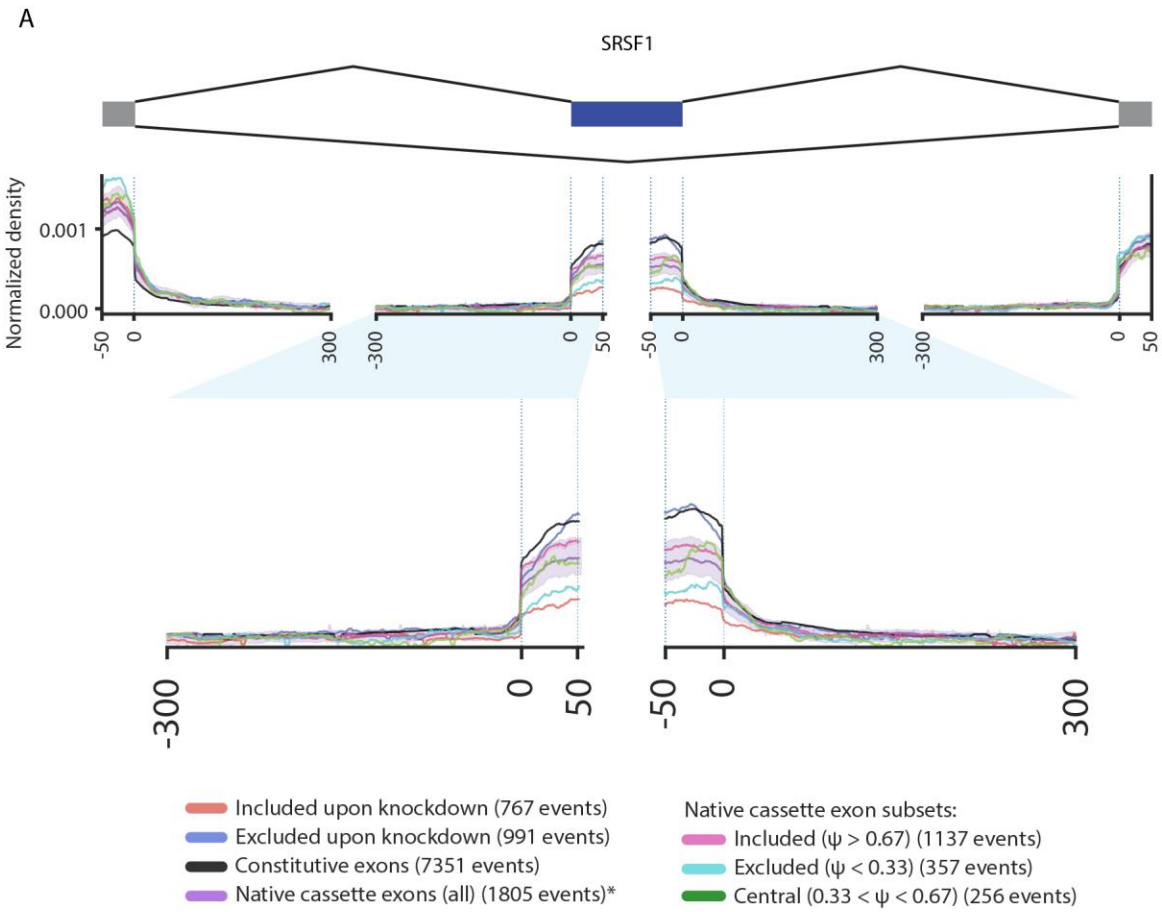

Figure 5. Choice of background affects interpretation of splicing maps. Lines indicate average normalized eCLIP signal at SRSF1 (red) knockdown-included and (blue) knockdown-excluded cassette exon events against four controls: constitutive exons (with no exclusion reads across multiple control RNA-seq datasets), native cassette exons with 0.05 < Percent Spliced In (Ψ) < 0.95 in at least half of ENCODE control RNA-seq datasets, and subsets of native cassette exons with average $\Psi < 0.33$ (excluded), $0.33 < \Psi < 0.67$ (central), and $\Psi > 0.67$ (included) in ENCODE control RNA-seq datasets.

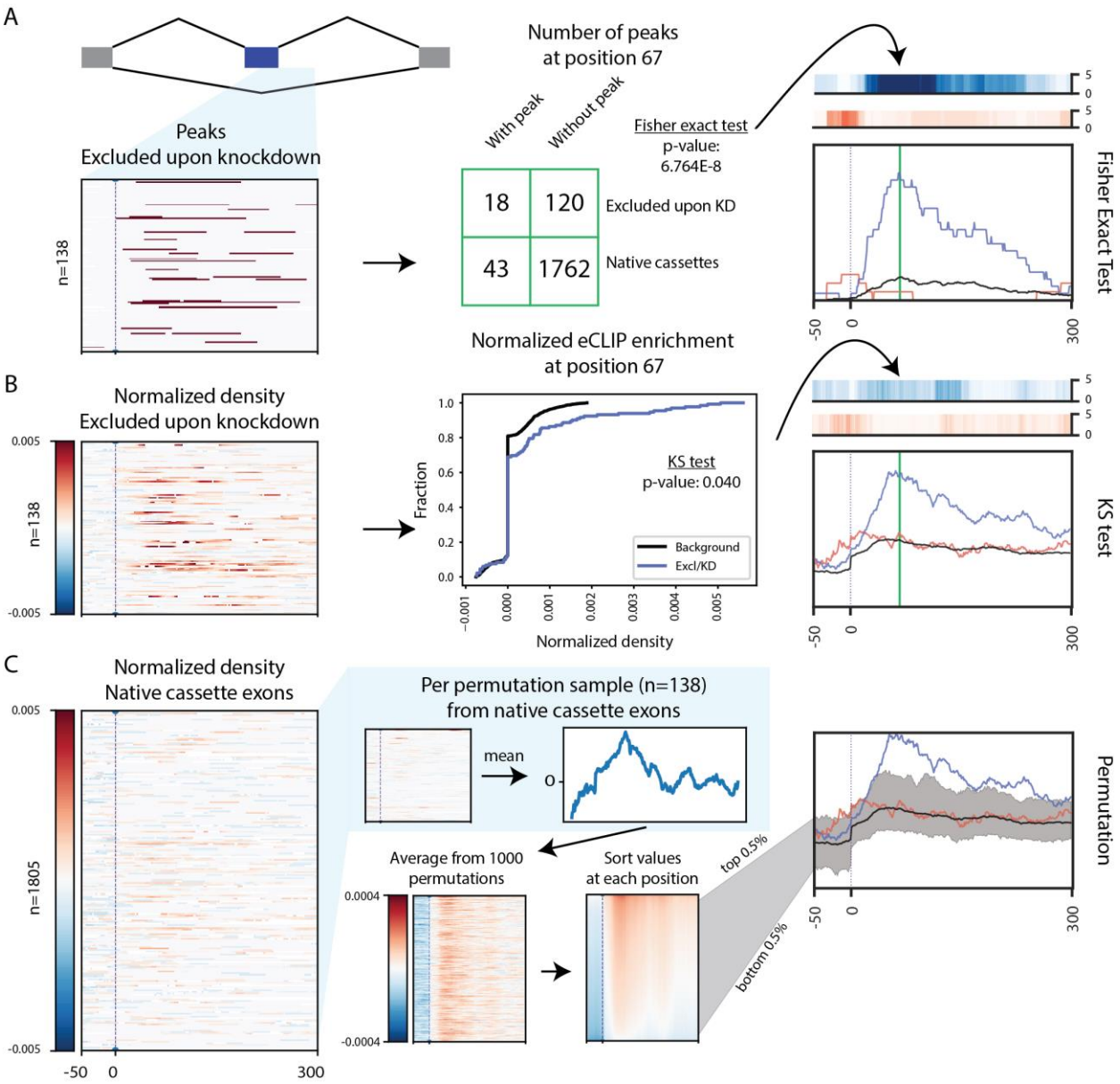

Figure 6. Significance models for splicing maps based on peak versus read density. (A) Schematic shows calculation of significance for peak-based splicing maps. (left) Peak positions are mapped across a set of significantly altered events (data shown is for exons excluded upon RBFOX2 knockdown in HepG2 cells). (center) At each position, a Fisher's Exact (or equivalent) test is performed between this set and some control set (e.g. native cassette exons; see further discussion in Figure 5). (right) Resulting significance can be plotted for all positions in the map for (blue) knockdown-excluded or (red) knockdown-included events. Significance is shown on a $-\log_{10}$ scale. (B) Significance calculation for read density maps using Kolmogorov-Smirnov test. (left) Normalized density is calculated for all knockdown-excluded events. (center) At each position, the distribution of normalized density is compared between knockdown-excluded and a control (native cassette exons). (right) Region-wide results are summarized similar to (A). (C) A bootstrapping strategy identifies confidence intervals for the control event list. (left) Normalized density is identified for the set of native cassette exons. (center-top) For each of 1000

permutations, a random sample of events is chosen (matching the number of knockdown-excluded events) and used to generate an average density map. (center-bottom) Average maps are collected for all 1000 permutations, and sorted at each position to identifiy 0.5% and 99.5% confidence bounds for the final map. (right) Native cassette exon density maps (along with confidence window) are then plotted along with maps identified from knockdown-excluded and included events.

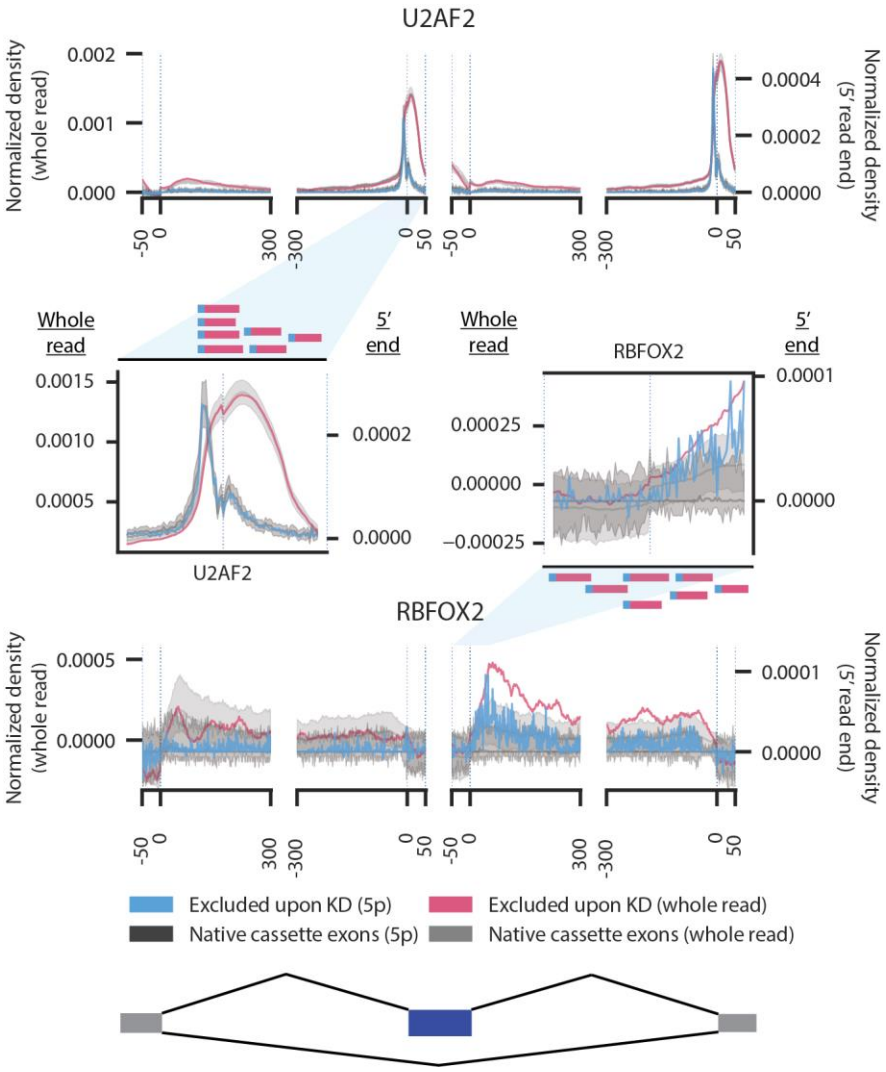

Figure 7. 5' read-based splice maps improve resolution of binding for some RBPs. Shown are splicing maps for (top) U2AF2 HepG2 eCLIP signal at exons excluded upon U2AF2 knockdown in HepG2 cells or (bottom) RBFOX2 HepG2 eCLIP signal at exons excluded upon RBFOX2 knockdown in HepG2 cells. Splicing maps were generated (red) using the entire read (as in previous figures), or (blue) using the 5' terminal position of reads only.

RBP-Maps enables robust generation of splicing regulatory maps

Brian Yee, Gabriel Pratt, Brenton Graveley, et al.

RNA published online November 9, 2018