ARTICLE

Discovery of Allele-Specific Protein-RNA Interactions in Human Transcriptomes

Emad Bahrami-Samani[1,2] and Yi Xing[1,2,3,*]

Gene expression is tightly regulated at the post-transcriptional level through splicing, transport, translation, and decay. RNA-binding proteins (RBPs) play key roles in post-transcriptional gene regulation, and genetic variants that alter RBP-RNA interactions can affect gene products and functions. We developed a computational method ASPRIN (Allele-Specific Protein-RNA Interaction) that uses a joint analysis of CLIP-seq (cross-linking and immunoprecipitation followed by high-throughput sequencing) and RNA-seq data to identify genetic variants that alter RBP-RNA interactions by directly observing the allelic preference of RBP from CLIP-seq experiments as compared to RNA-seq. We used ASPRIN to systematically analyze CLIP-seq and RNA-seq data for 166 RBPs in two ENCODE (Encyclopedia of DNA Elements) cell lines. ASPRIN identified genetic variants that alter RBP-RNA interactions by modifying RBP binding motifs within RNA. Moreover, through an integrative ASPRIN analysis with population-scale RNA-seq data, we showed that ASPRIN can help reveal potential causal variants that affect alternative splicing via allele-specific protein-RNA interactions.

Introduction

Natural genetic polymorphisms can diversify the transcriptome and proteome among individuals by altering the post-transcriptional processing and modification of RNA.[1] Such regulatory variation can cause disease, modify disease risk, or affect therapeutic response.[2,3] Thus, the discovery of genetic variants that affect post-transcriptional RNA regulation may reveal causal mechanisms underlying phenotypic variability and disease pathogenesis in human populations.[4,5]

RNA-binding proteins (RBPs) are key regulators of post-transcriptional RNA processing and modification.[6] RBPs participate in various steps of RNA regulation, including splicing, transport, translation, and decay, thus determining the fate of RNAs after transcription.[7] RBPs bind to their RNA targets via defined sequence and/or structural motifs.[8]

The predominant technology for transcriptome-wide mapping of RBP-RNA interactions is CLIP-seq.[9–12] Multiple variants of CLIP-seq (HITS-CLIP,[9] PAR-CLIP,[10] iCLIP,[11] and eCLIP[12]) aimed at improving library efficiency and reducing artifacts have been used to define the RBP-RNA binding landscape of hundreds of RBPs across different cell types and species. These variants of CLIP experiment are all fairly similar in essence, which is cross-linking RBP and its targets for a more stringent washing of unbound RNA followed by high-throughput sequencing, but due to their technical differences and biases, deliver slightly different datasets, as detailed in Chakrabarti et al.[13]

Previous studies have investigated the effects of genetic variants on post-transcriptional regulation, primarily using a sequence motif-based approach. Jian et al.[14] reviewed eight bioinformatics tools that predict splice-altering single nucleotide variants in the human genome. These methods use information about highly conserved splicing regulatory elements (5′ and 3′ splice sites and branch point signals) as well as auxiliary *cis*-acting elements recognized by *trans*-acting RBPs[14] to predict the effects of genetic variants on alternative splicing. Some other recent studies used defined binding motifs of RBPs to predict variants that alter RBP-RNA interactions.[15,16] However, as RBP binding motifs are typically short (4–6 nucleotides) and degenerate, methods based on RBP motifs are expected to have a low accuracy and high noise.[17]

We developed ASPRIN (Allele-Specific Protein-RNA Interaction), a computational method to identify genetic variants that alter RBP-RNA interactions via a joint analysis of CLIP-seq and RNA-seq data. The premise of ASPRIN is that the allelic ratio in CLIP-seq data compared to that in RNA-seq data of the same cell type can reflect the effects of genetic variants on RBP-RNA interactions. We performed a systematic ASPRIN analysis of ENCODE CLIP-seq (eCLIP) and RNA-seq data for 166 RBPs in two cell lines. One advantage of eCLIP is that it is designed to enrich for fragments that are truncated at the cross-link location,[12] although the degree of enrichment is RBP dependent, while in some other types of CLIP experiments such as PAR-CLIP, the characteristic T-to-C mutation at the cross-link sites[10] introduces additional complications to allele-specific analysis. ASPRIN identified genetic variants that alter RBP-RNA interactions by modifying conserved RBP binding sites. Moreover, through an integrative ASPRIN analysis with population-scale RNA-seq data, we showed that ASPRIN can help reveal causal variants that affect alternative splicing via allele-specific protein-RNA interactions.

[1]Department of Microbiology, Immunology & Molecular Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA; [2]Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [3]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
*Correspondence: xingyi@email.chop.edu
https://doi.org/10.1016/j.ajhg.2019.01.018.

Material and Methods

Calling Variants from RNA-Seq Data

The total RNA-seq data for HepG2 whole-cell preparations from two different labs (ENCODE: ENCSR468ION and ENCSR181ZGR), a HepG2 cytosolic fraction (ENCODE: ENCSR862HPO), a HepG2 nuclear fraction (ENCODE: ENCSR061SFU), K562 whole-cell preparations from two different labs (ENCODE: ENCSR000AEN and ENCSR885DVH), a K562 cytosolic fraction (ENCODE: ENCSR860DWK), and a K562 nuclear fraction (ENCODE: ENCSR040YBR) were downloaded from the ENCODE website.

The GATK Best Practices workflow for calling single-nucleotide polymorphisms (SNPs) and indels on RNA-seq data was used with minor modifications.[18] Briefly, the datasets were mapped using STAR v.2.5.2a,[19] and total RNA-seq data from all fractions and all labs were merged to make one large RNA-seq dataset for each cell line. The rest of the pipeline included adding read groups, sorting, marking duplicates, and creating the index using Picard tools v1.134 (see Web Resources), followed by splitting and trimming, subsequent reassignment of mapping qualities, indel realignment, base recalibration, and finally, calling variants using GATK pipeline.[18] Mapping and variant calling statistics are given in Table S1.

Filtering SNPs

In our analyses, we removed false positive SNPs due to sequencing errors, alignment artifacts, and RNA editing events. Only heterozygous variants in the RNA-seq data that matched known SNPs in the NCBI SNP Database (dbSNP)[20] were kept. Potential RNA editing events were labeled and removed by intersecting called heterozygous variants with the RADAR (Rigorously Annotated Database of A-to-I RNA editing) RNA editing database.[21] However, ASPRIN can run in different modes to consider variants that are SNPs, RNA editing events, or both.

eCLIP Data Analysis

For pre-processing ENCODE eCLIP data, the standard operating procedure (SOP) published on the ENCODE website was followed. In brief, (1) adaptors were trimmed using cutadapt v1.10,[22] (2) a second round of adaptor cutting was performed to control for double ligation events, (3) the resulting reads were mapped to the human-specific version of Repbase[23] using STAR 2.5.2a[19] to remove repetitive elements and other repetitive reads, as well as to control for spurious artifacts from rRNA, (4) reads mapped to repetitive regions were filtered out of the resulting output from STAR, and (5) PCR duplicates were further removed using random-mers that were provided in the names of the reads. The raw read files available at the ENCODE data portal are already pre-processed, and random-mers that can reveal PCR duplicates are removed from the reads and put in the read names. This information can be used for removing PCR duplicates that are mapped to the same genomic location.

Mapped reads for each replicate were sorted, merged, and indexed, and the resulting mapped reads file was used as input for ASPRIN. In our analysis of ASPRIN results, when we needed the peaks, second (paired-end) reads were used to perform peak-calling using Piranha,[24] with a bin size of 1 nt so we can achieve single-nucleotide resolution in peak-calling for the eCLIP data. We considered significant peaks to be those that had a corrected p value of less than 0.01. Mapping and peak calling statistics for RBPs in HepG2 and K562 cell lines are given in Tables S2 and S3, respectively.

ASPRIN Allelic Ratio Test

For each RBP, ASPRIN counts the number of reads that cover each allele in the CLIP-seq and RNA-seq datasets and forms a contingency table with (1) the number of reads covering the reference allele in CLIP-seq, (2) the number of reads covering the alternative allele in CLIP-seq, (3) the number of reads covering the reference allele in RNA-seq, and (4) the number of reads covering the alternative allele in RNA-seq. The result of Fisher's exact test for each SNP shows whether a particular SNP is significantly differentially bound by an RBP. For each RBP, ASPRIN p values are corrected for multiple hypothesis testing using the Benjamini-Hochberg method and SNPs with q value < 0.1 are reported as significantly differentially bound, or "ASPRIN SNPs."

Assessing the Robustness of ASPRIN

To measure the error associated with the used variant filtering method, RNA-seq datasets for the GM12878 cell line were downloaded from SRA (SRA: SRR307897 and SRR307898) and the complete genotype for this cell line was downloaded from the 1000 Genomes (1000G) project website.[25] We performed variant calling as described above and intersected the set of called variants with 1000G SNPs, dbSNP, and RADAR.

To investigate the choice of RNA-seq protocol and how it may affect the power of ASPRIN, in addition to the total RNA-seq data, we also downloaded polyA+ mRNA-seq data for the same cell lines, fractions, and laboratories: HepG2 whole-cell preparations from two different labs (ENCODE: ENCSR985KAT and ENCSR561FEE), a HepG2 cytosolic fraction (ENCODE: ENCSR931WGT), a HepG2 nuclear fraction (ENCODE: ENCSR058OSL), K562 whole-cell preparations from two different labs (ENCODE: ENCSR000AEO and ENCSR545DKY), a K562 cytosolic fraction (ENCODE: ENCSR384ZXD), and a K562 nuclear fraction (ENCODE: ENCSR530NHO). To normalize for read number and length, we sampled n number of reads from all of these datasets, ten times, where n was the minimum number of reads among these datasets. The RNA-seq libraries that had 100-nucleotide reads (from Brenton Graveley's lab) were also truncated to 50 nucleotides, to have the same read length as the RNA-seq libraries with 50-nucleotide reads (from Eric Lecuyer's lab). We then called variants from all these datasets and compared the number of called variants and the regions in which these variants were located. We also ran the ASPRIN pipeline on all eCLIP datasets with these ten subsampled RNA-seq datasets using only cytosolic polyA+ mRNA-seq and nuclear total RNA-seq, to compare the number of ASPRIN SNPs that can be called using these two distinct RNA-seq sets representing different RNA species and subcellular fractions.

To investigate the cross-linking bias and its potential effects on our analysis, for any ASPRIN SNP that was associated with at least one of the 75 RBPs in the HepG2 cell line, we counted for how many RBPs this SNP was (1) called significant with preference for the reference allele, (2) called significant with preference for the alternative allele, (3) not called significant, and (4) not present in enough reads to pass the filters for the ASPRIN analysis.

ASPRIN SNP Enrichment or Depletion in Genomic Regions

We measured the enrichment of ASPRIN SNPs in different genomic regions using Fisher's exact test. For RBP x and region r, we counted (1) the number of ASPRIN SNPs for x in r and (2) the rest of ASPRIN SNPs for x. In addition, for the background, we

counted (3) the number of ASPRIN SNPs in *r* for the rest of the RBPs and (4) the number of ASPRIN SNPs in any region except *r* for the rest of the RBPs. Then, we used Fisher's exact test to measure the significance of enrichment or depletion of ASPRIN SNPs in region *r* for RBP *x* compared to the average expectation.

Measuring RBP Sequence Specificity

We determined the sequence specificity of RBPs as the information content of the motif obtained by *de novo* motif discovery in the high-quality binding sites as defined by the Piranha peak caller.[24] For each RBP, peaks output was obtained using Piranha. Then the genomic region (intron, 5′ UTR, coding segment, 3′ UTR, noncoding RNA, and intergenic sequence) containing each peak was assigned to them. All peaks in noncoding or intergenic regions were filtered out and the highest peak in each gene was selected as the representative peak of that RBP binding to the gene. Finally, top 1,000 peaks based on the corrected p value reported by Piranha were selected as the set of high-quality peaks. Zagros[8] was then used for *de novo* motif discovery, using sequence and secondary structure information. The parameters were window size 6 and top 10 motifs (-w 6 –n 10), and we selected the top motif reported by Zagros as the discovered motif for that RBP. Information content for each RBP consensus motif is obtained by taking the average information content over all positions within the consensus sequence and for each position defined by Shannon's entropy. RBPs with consensus sequences that had more information content were considered to have higher sequence specificity.

Motif Enrichment Analysis

Motif enrichment analysis was done using the STORM software.[26] As described above the top 6-nucleotide motif discovered by Zagros in top 1,000 peaks for each RBP was used as the consensus motif for that RBP. STORM can use the motif position weight matrix output from Zagros directly and calculate the enrichment of that motif in the set of input sequences.

For each SNP, a sequence of 11 nucleotides centered at the SNP (windows containing all 6-mer positions in the genome that include the SNP) was extracted. Then for each sequence we flipped the center nucleotide, the SNP, to the alternative allele. Therefore, for each RBP, two sets of sequences were formed, that are pairwise identical, except for the center position that contains two alleles of the SNP. One set contains the alleles with low-affinity binding and the other contains the alleles with high-affinity binding. Then, STORM was run using the corresponding consensus motif for each RBP in two sets of sequences for the said RBP to assess the difference in motif score. Parameters for STORM can be set in a way to find the top occurrence of a motif per sequence (-n 1 -q) in single stranded mode (-S) for RNA. For each RBP, we only considered SNPs that have positive scores in both high and low binding affinity sequences to filter out SNPs occurring outside the binding site. For each SNP the maximum motif score among all six possible windows in the high binding affinity sequence and its corresponding motif score in the low binding affinity sequence were selected to produce the boxplots of motif scores for each RBP in each position of the motif. We also defined a motif impact score for each RBP and its associated ASPRIN SNP set as the maximum difference in average motif score between the two alleles with high versus low binding affinity in the window of six nucleotides overlapping the ASPRIN SNP.

Splicing Quantitative Trait Loci (sQTLs) Analysis

To demonstrate the utility of ASPRIN in finding relevant SNPs that may cause changes in splicing, we analyzed ASPRIN SNPs in HepG2 cell line and sQTLs calculated from population-scale RNA-seq data in liver as part of the GTEx consortium.[27] RNA-seq and genotype data of liver tissues from 71 individuals (GTEx v6) were downloaded, mapped to the hg19 genome and Percent Spliced In (Psi) values were calculated for each splicing event in each individual. We selected events requiring the condition $Max(Psi) - Min(Psi) > 0.1$ over all individuals. Then, for each splicing event, GLiMMPS[28] was run on SNPs within a 400-kb window centered on the splicing event. The false discovery rate (FDR) was estimated using a permutation procedure to obtain the null hypothesis. In each of the ten permutations, we shuffled the individuals' genotypes so that each individual would have a randomly assigned genotype. We then ran GLiMMPS to obtain the sQTLs on the permutated data and recorded the minimum p value for each exon over all *cis* SNPs in each permutation and used this set of p values as the empirical null distribution for estimating the FDR. Using an FDR threshold of 10%, we calculated the p value cutoff t such that $P(p_0 < t)/P(p_1 < t) = 0.1$, where $P(p_0 < t)$ is the fraction of expected p values from the null distribution less than t and $P(p_1 < t)$ is the fraction of observed p values less than t from the real data. For each splicing event, the sQTLs were defined as the SNPs that have p values less than the cutoff. The linkage disequilibrium (LD) with all the ASPRIN SNPs was calculated and used for selecting only the exons that had sQTLs in high LD with ASPRIN SNPs ($r^2 > 0.8$). The LD map was created using a CEU population.[29] Exons for events in which the ASPRIN SNP is near the exon were further filtered with the criteria that the ASPRIN SNP is within a window of 500 nucleotides around the alternative splicing event. The windows were defined for each alternative splicing event as follows: (1) skipped exon: 500 nucleotides into the introns on each side of the skipped exon; (2) mutually exclusive exons: 500 nucleotides into the introns on each side of two mutually exclusive exons; (3 and 4) alternative 5′ or 3′ splice sites: 500 nucleotides into the introns on each side of the longer exon; and (5) intron retention: 500 nucleotides into the exons on each side of the retained intron. The numbers of each type of alternative splicing event that pass the filters are given in Table S4.

Genome-wide Association Study (GWAS) Signals

23,444 GWAS SNPs with p values $< 10^{-5}$ were downloaded from the NHGRI GWAS catalog[29] and PLINK v1.08p[30] was used to calculate the LD between ASPRIN SNPs and GWAS SNPs on the LD map that was created using a CEU population.[29] SNPs in high LD ($r^2 > 0.8$) with GWAS SNPs were reported as GWAS-correlated ASPRIN SNPs.

Results

ASPRIN Pipeline for Detecting Allele-Specific Protein-RNA Interactions

The discovery of allele-specific protein-RNA interactions in ASPRIN is based on the rationale that if a particular SNP creates or disrupts an RBP binding site, we would expect to observe a difference in the allelic ratio of the SNP in the CLIP-seq reads compared to the corresponding RNA-seq reads from the same cell type. A schematic

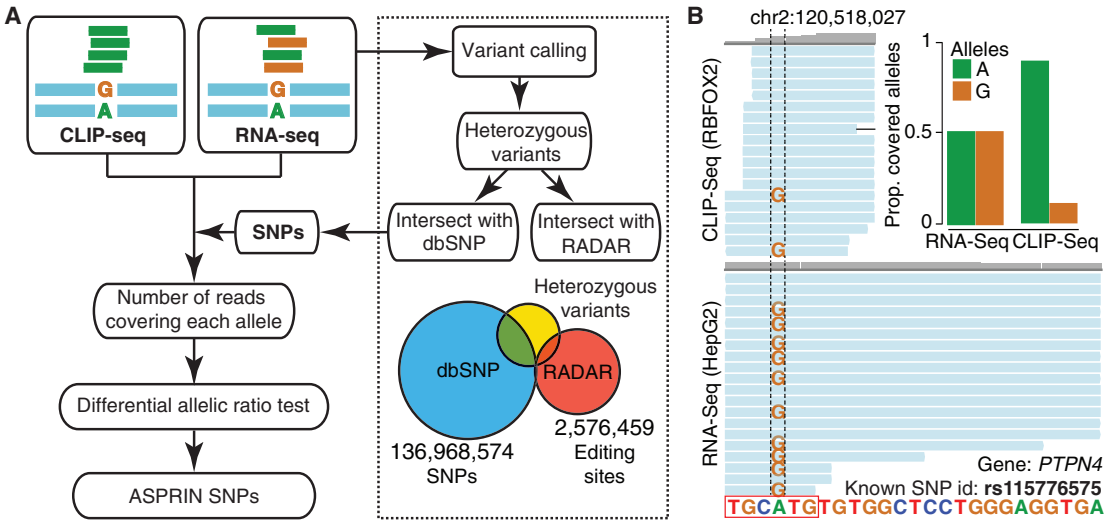

Figure 1. The ASPRIN Pipeline for Identifying Allele-Specific Protein-RNA Interactions from CLIP-Seq and RNA-Seq Data
(A) Flowchart of the ASPRIN pipeline: variants are called from RNA-seq data, and heterozygous variants are intersected with dbSNP to obtain a list of high-confidence SNPs and intersected with RADAR to filter out potential A-to-I RNA editing events. For each SNP, ASPRIN counts the number of reads in the CLIP-seq and RNA-seq data that support each allele. An allelic ratio test then assesses whether one allele is significantly more preferred for RBP binding.
(B) An A-to-G SNP (rs115776575) disrupts a consensus RBFOX2 binding site in *PTPN4*. This disruption of binding is illustrated in the difference in the numbers of reads containing each allele in CLIP-seq reads, while equal numbers of reads contain each allele in the RNA-seq data.

diagram of ASPRIN is provided in Figure 1A. Briefly, to call SNPs, RNA-seq reads were mapped to the human genome and transcriptome, and single nucleotide variants (SNVs) were called using the GATK pipeline[18] (see details in Material and Methods). We then applied stringent filters to remove false positive SNPs contributed by potential sequencing errors, alignment artifacts, and RNA editing events. Specifically, heterozygous variants in RNA-seq data that matched known SNPs in dbSNP were kept,[20] while potential RNA editing events were removed by intersection with the RADAR RNA editing database.[21] After this set of high-confidence SNPs was generated, CLIP-seq reads were mapped and reads supporting the reference or alternative allele in the CLIP-seq data were counted. Additionally, because RNA-seq reads are typically longer than CLIP-seq reads, we split the 100 bp RNA-seq reads in the ENCODE data into two 50 bp segments and mapped them separately to count reference and alternative alleles in the RNA-seq data, to alleviate systematic mapping bias for the reference over the alternative alleles in CLIP-seq data compared to the RNA-seq data. Indeed, by splitting 100 bp RNA-seq reads, the mapping bias was largely removed (Figure S1). Finally, we tested each SNP site with at least ten reads (sum of two alleles) in both the RNA-seq and CLIP-seq data for significant difference in allelic ratio via Fisher's exact test of allelic read counts in RNA-seq versus CLIP-seq data. After correcting for multiple hypothesis testing, we reported SNPs with corrected p values of less than 0.1 as ASPRIN SNPs (Figure 1A). An example result for the HepG2 cell line is an A-to-G SNP (rs115776575) in *PTPN4* (MIM: 176878) that disrupts a highly conserved "A" nucleotide in the "TGCATG" consensus motif of RBFOX2. While the allelic ratio between "A" and "G" was 1:1 in the RNA-seq reads, the "G" allele represented only 10.5% of the CLIP-seq reads (Figure 1B), consistent with RBFOX2 binding to the TGCATG motif, and that the A-to-G SNP at the fourth nucleotide position of the motif disrupts RBFOX2 binding.

ASPRIN Is Robust in Discovering SNPs Involved in Allele-Specific Protein-RNA Interactions

We evaluated various issues that may affect the performance of ASPRIN, such as errors arising from calling variants from RNA-seq data, choice of RNA-seq protocols, and potential artifacts due to the cross-linking step in CLIP-seq experiments. First, since whole-genome genotype data are not available for most of the cell types with CLIP-seq data, we assessed our SNP calling procedure using RNA-seq data alone. To obtain a ground truth for this assessment, we called SNVs using RNA-seq data for the GM12878 cell line (SRA accessions SRR307897 and SRR307898), for which high-quality whole-genome genotype data are available from the 1000G project.[25] After calling SNVs in GM12878 using our pipeline, we intersected the set of heterozygous variants with known SNPs in GM12878 from the 1000G project[25] and known A-to-I RNA editing sites in the RADAR database[21] to investigate the distribution of different variant types. As shown in Figure 2A, 63.2% of the called SNVs were known SNPs and 23.8% were known RNA editing events. The remaining 13.0% were unknown variants that did not match any 1000G SNPs or RADAR sites and the distribution of

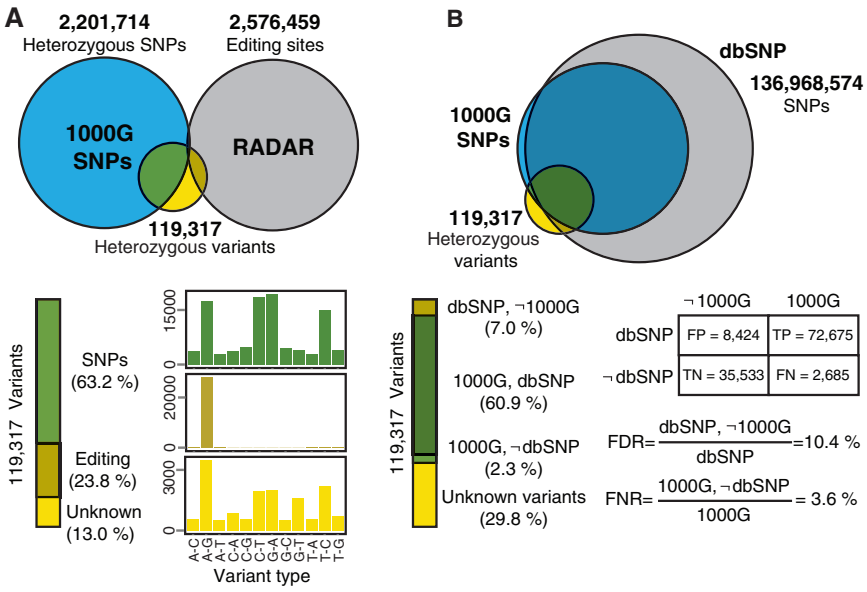

Figure 2. RNA-Seq Variants Called in the GM12878 Cell Line

dbSNP and RADAR were used as external references to obtain a set of high-confidence SNPs from RNA-seq variant calling in the absence of matching genotype data. (A) Intersection of variants with the 1000G SNPs and RADAR RNA editing events as well as the distribution of variant types over all 12 possible single-nucleotide changes.

(B) The variant filtering steps in the ASPRIN pipeline yield low false discovery and low false negative rates.

all 12 possible single-nucleotide changes suggested that these unknown variants represented a mixture of SNPs and RNA editing events (Figure 2A). As shown in Figure 2B, 89.6% of the called SNVs that were in the dbSNP were also present in the 1000G data for GM12878, suggesting an upper bound of 10.4% for the false discovery rate of our RNA-seq-based SNP calling procedure. Moreover, 3.6% of the called SNVs for GM12878 present in the 1000G data were not in the dbSNP, suggesting that the use of the dbSNP had a minimal impact on the false negative rate of SNP identification. Collectively, our data suggest that, by using dbSNP and RADAR as filters, we can obtain a set of high-confidence SNPs from our RNA-seq variant calling in the absence of matching genotype data.

Next, we investigated issues that may affect the power of ASPRIN for calling SNPs and identifying allele-specific protein-RNA interactions. Specifically, the choice of RNA-seq protocol may affect the power of ASPRIN depending on the binding location of a given RBP within the RNA. For instance, a cytosolic polyA+ RNA-seq library would be appropriate for an RBP that predominantly binds to exons within mRNAs in the cytosol, but not for an RBP that predominantly binds to introns within precursor mRNAs in the nucleus. To investigate the most appropriate RNA-seq protocols and libraries, we randomly sampled equal numbers of reads from polyA+ and total RNA-seq libraries of distinct subcellular fractions (nucleus, cytosol, and whole-cell) from the HepG2 cell line and performed SNP calling and ASPRIN analysis on the sampled RNA-seq data. For both polyA+ and total RNA-seq libraries, we called the highest number of SNPs from the nuclear RNA-seq data and the lowest number of SNPs from the cytosolic RNA-seq data (Figure 3A). The lowest number of SNPs was called from cytosolic polyA+ RNA-seq data (Figure 3A); these SNPs were enriched for exonic regions within UTRs (untranslated regions) and CDS (coding segments) and depleted for intronic regions within pre-mRNAs (Figure 3B). A similar trend was observed for the K562 leukemia cell line (Figure S2). On the other hand, as reads of cytosolic polyA+ RNA-seq libraries were concentrated within CDS and UTR regions, such data may have better power for detecting allele-specific protein-RNA interactions of RBPs that bind predominantly to exons. As expected, the nuclear fraction of the total RNA-seq library provided a much greater power for ASPRIN analysis of an RBP that binds predominantly to introns (HNRNPM), while ASPRIN analyses of an RBP that binds predominantly to exonic regions (YBX3) identified similar numbers of ASPRIN SNPs from the cytosolic polyA+ RNA-seq library and the nuclear total RNA-seq library (Figure 3C). Furthermore, after calling peaks, we sorted all RBPs in both cell lines based on the ratio of exonic (CDS and UTR regions) to intronic peaks. The complete distributions of peaks in different regions for all RBPs are shown in Figure S3 and we excluded RBPs for which more than 50% of peaks fell in intergenic regions and noncoding RNAs. We observed a positive correlation (Pearson correlation coefficient = 0.34, p value < 0.0001) between binding of an RBP to exonic regions and the relative power of identifying significant ASPRIN SNPs using cytosolic polyA+ RNA-seq libraries, despite large variation among individual RBPs (Figure 3D).

Finally, we evaluated potential false positives that may arise from the cross-linking step in CLIP-seq experiments. Specifically, the sequences in the CLIP-seq libraries may be altered by mutation or deletion at the cross-linking site.[9–11] We noted that in the eCLIP protocol used for generating the ENCODE CLIP-seq data, the majority of fragments were truncated at the cross-linking site rather than containing mutations or deletions.[12] Nonetheless, we investigated this issue further by calling SNVs from the ENCODE eCLIP data and comparing the distribution of variant types to that of the RNA-seq data and observed a similar distribution (Figure S4). Another possible source of artifacts is cross-linking bias that may shift the read count toward specific nucleotides in the CLIP-seq data. However, 70% of ASPRIN SNPs were called significant for only one RBP. Only 6% of ASPRIN SNPs were called

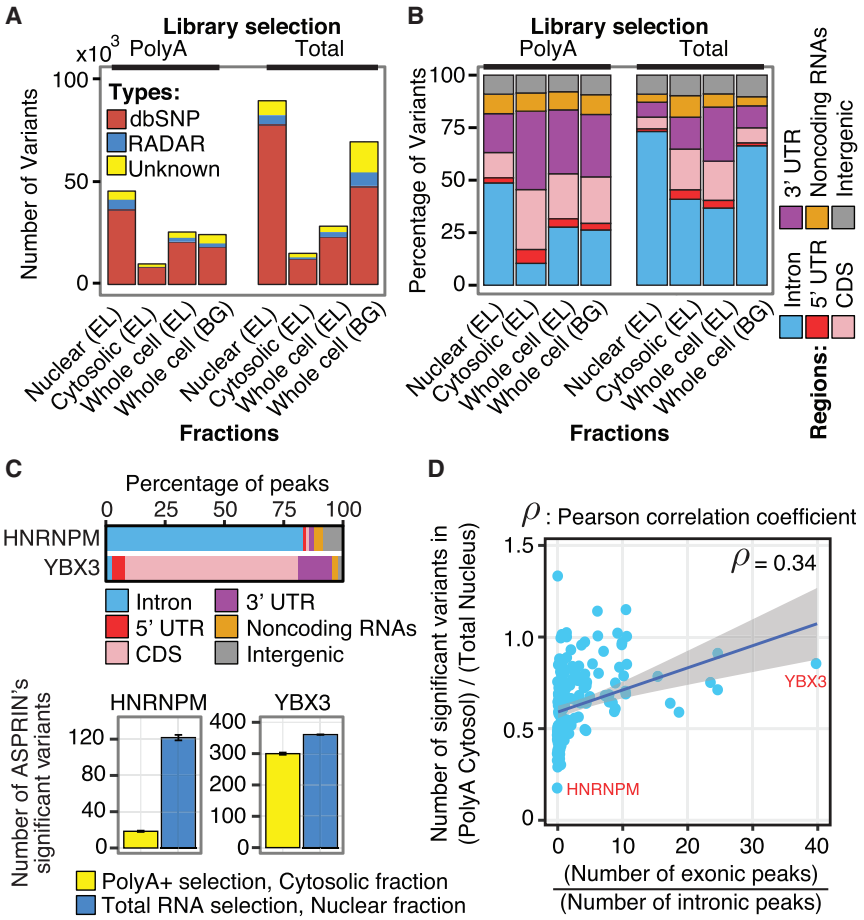

Figure 3. RNA-Seq Variants Called from Different RNA-Seq Libraries of the HepG2 Cell Line

Two methods of library selection (polyA+ and total RNA) in different subcellular fractions (nucleus, cytosol, and whole-cell fractions from two different labs: EL = Eric Lecuyer's lab at Institut de Recherches Cliniques de Montréal, and BG = Brenton Graveley's lab at University of Connecticut).

(A) Numbers of variants called from different RNA-seq libraries and their intersections with dbSNP and RADAR.

(B) Distribution of called variants in different genomic regions.

(C) Numbers of significant ASPRIN variants from polyA+ cytosolic or total RNA nuclear RNA-seq libraries for an RBP that binds predominantly to intronic regions (HNRNPM) and an RBP that binds predominantly to exonic regions (YBX3). Standard error of the mean is indicated as the error bar for each library selection method and subcellular fraction.

(D) The ratio of ASPRIN SNPs found using polyA+ cytosolic RNA-seq libraries to ASPRIN SNPs found using total RNA nuclear RNA-seq libraries increases as the ratio of exonic to intronic peaks increases.

significant for more than five RBPs. Among these SNPs, the same allele was preferred by all RBPs in 87% of the SNPs, whereas in the remaining 13%, different alleles were preferred by different RBPs (Figure S5). Overall, these data suggest that the fraction of ASPRIN SNPs that may be attributable to CLIP-seq cross-linking bias is small.

To assess the reproducibility of ASPRIN using different eCLIP replicates, we ran ASPRIN on all the ENCODE data and each eCLIP replicate separately. For each pair of datasets, we calculated the normalized intersection over union of the number of ASPRIN SNPs to show for each eCLIP replicate which dataset shows the highest degree of agreement. As shown in Figures S6 and S7, ASPRIN is reproducible between replicates in both cell lines.

ASPRIN Identifies Functionally Relevant SNPs for Different Classes of RBPs

To assess the potential functional relevance of the ASPRIN results, we investigated the positional distribution of ASPRIN SNPs for different classes of RBPs. To this end, we classified RBPs based on their known functions,[31] and we defined genomic regions as follows: (1) 5′ UTRs, (2) upstream proximal intronic regions (500 nucleotides upstream of an internal exon), (3) coding regions, (4) downstream proximal intronic regions (500 nucleotides downstream of an internal exon), (5) 3′ UTRs, (6) distal intronic regions (more than 500 nucleotides away from exons on both sides), (7) noncoding RNAs, and (8) intergenic regions. Then, for each RBP, we calculated the enrichment of ASPRIN SNPs in different genomic regions (see details in Material and Methods). As expected, ASPRIN SNPs were more enriched in regions to which RBPs bind to perform their known functions (Figure 4). For instance, in the HepG2 cell line, we observed an enrichment (p value < 0.001) of ASPRIN SNPs in the 5′ UTR for translation regulators such as DDX3X and NCBP2, with 27.1% and 16.0% of their ASPRIN SNPs found within the 5′ UTR, respectively. Multiple classes of splicing factors showed distinct patterns of positional distributions for their ASPRIN SNPs. We observed an enrichment of ASPRIN SNPs in upstream proximal intronic regions for branch point recognition factors such as SF3B4 (30.5%), U2AF2 (12.2%), U2AF1 (8.8%), and SF3A3 (11.8%). Similarly, ASPRIN SNPs were enriched in the downstream proximal intronic regions for RBPs that are part of the 5′ splice site machinery such as PRPF8 (22.0%), EFTUD2 (14.8%), and RBM22 (11.7%). There was an enrichment of ASPRIN SNPs in coding regions for several splicing regulators that primarily bind to coding exons, such as SRSF1 (40.7%) and TRA2A (29.0%). For RBFOX2, we observed an enrichment of ASPRIN SNPs in both upstream and downstream proximal intronic regions (6.0% and 15.1%, respectively), as we expect RBFOX2 to bind to either region to promote exon skipping or

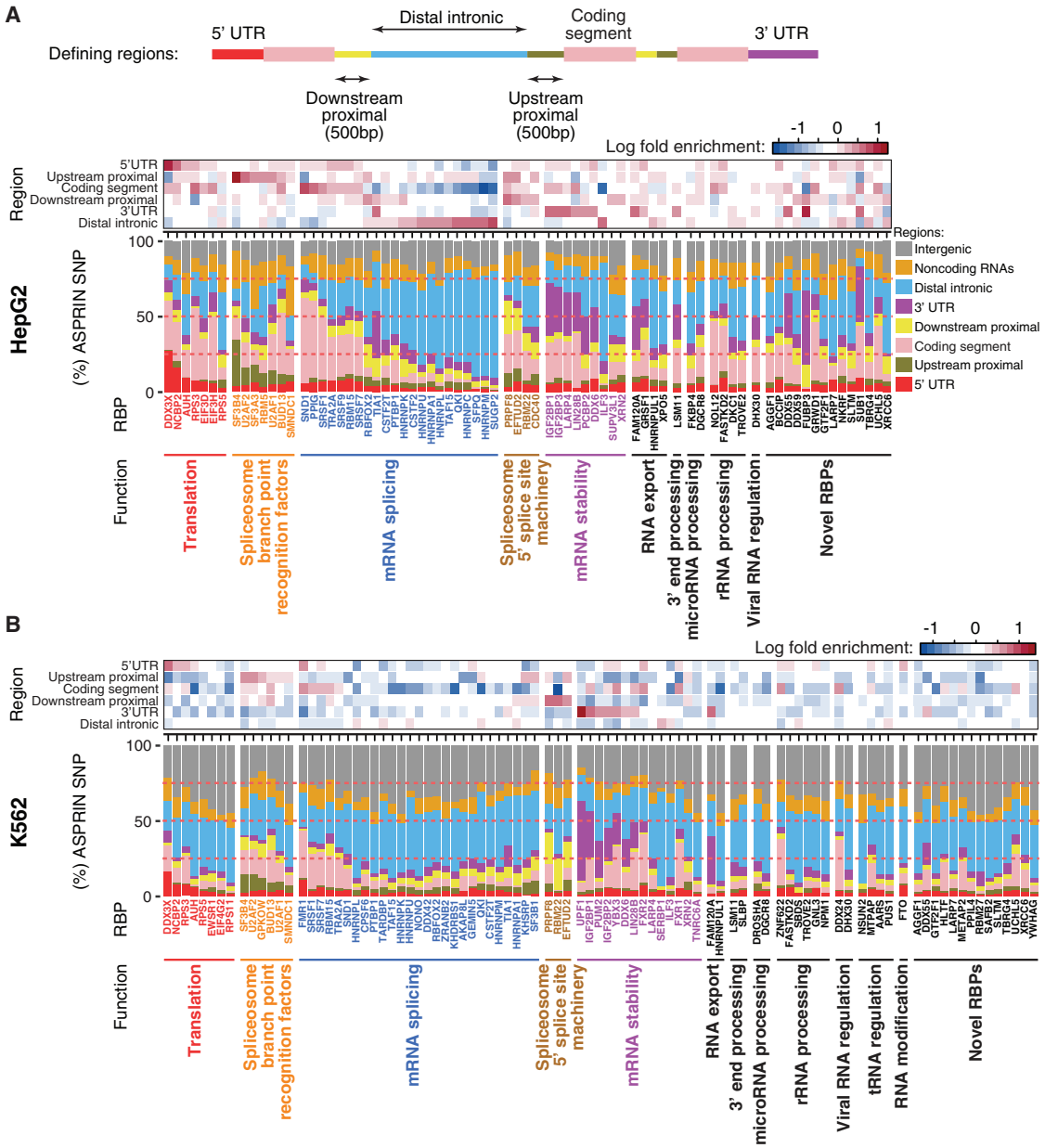

Figure 4. Enrichment of ASPRIN SNPs in Different Genomic Regions

Positional distributions of ASPRIN SNPs for different classes of RBPs in HepG2 (A) and K562 (B) cell lines. The top diagram of the figure depicts the different genomic regions used in the analysis. RBPs were classified based on their known functions.[31] In both panels the enrichment of ASPRIN SNPs for each RBP in different genomic regions is shown as heatmaps for color coded log fold enrichment (top) and barplots for percent of total ASPRIN SNPs (bottom).

inclusion, respectively. The ASPRIN SNPs of HNRNP proteins were enriched in distal intronic regions and depleted in coding regions, which fits that these RBPs predominantly bind to distal intronic regions. Finally, RBPs that regulate mRNA stability, such as IGF2BP proteins and LIN28, showed an enrichment of ASPRIN SNPs in the 3′ UTR (Figure 4A). We observed a similar pattern in the K562 cell line, where the same RBPs in both cell lines show the similar pattern of regional preference (Figures 4B and S8). The numbers of ASPRIN SNPs for each RBP in HepG2 and K562 are provided in Figure S9.

ASPRIN SNPs Affect RBP Consensus Motifs

To explore the potential molecular mechanisms by which ASPRIN SNPs affect protein-RNA interactions, we investigated the effects of ASPRIN SNPs on RBP consensus motifs. We predicted that if an RBP binds to RNAs in a highly sequence-specific manner, then variants within the conserved RBP consensus motif are likely to affect binding. First, we called peaks from ENCODE CLIP-seq data using Piranha[24] and performed *de novo* motif discovery on called peaks using Zagros[8] to obtain a 6-nucleotide consensus motif for each RBP. We then calculated the information

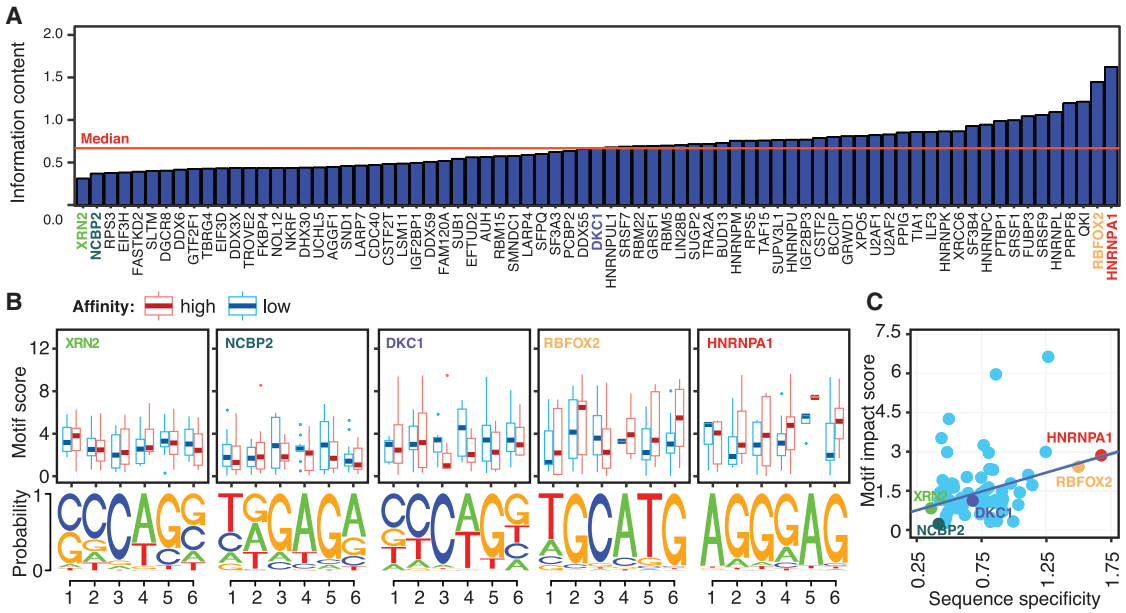

Figure 5. The Effect of ASPRIN SNPs on RBP Consensus Motifs

(A) RBPs in the HepG2 cell line, sorted based on the sequence specificity (i.e., information content) of their consensus motif. For each RBP, the information content was calculated by taking the average of the information content for each position within the motif, calculated using Shannon's entropy.

(B) Boxplots comparing the consensus motif scores for alleles with high and low binding affinity. Two RBPs with the lowest sequence specificity (XRN2 and NCBP2), one RBP with the median sequence specificity (DKC1), and two RBPs with the highest sequence specificity (RBFOX2 and HNRNPA1) are shown. The consensus motif obtained from the top 1,000 peaks for each RBP is represented at the bottom of each graph. The middle line of the boxplot represents median value. The low and high ends of the box represent the 25% and 75% quantiles, respectively. The two whiskers extend to 1.5 times the interquartile range.

(C) As sequence specificity increases, we observe a larger difference between the consensus motif scores of the high-affinity versus low-affinity ASPRIN alleles.

content of the consensus motif, defined as the average information content of each position within the 6-nucleotide motif, as a measure of sequence specificity (see details in Material and Methods). Figure 5A shows the RBPs in HepG2, sorted by the sequence specificity of their consensus motifs. Among all RBPs, HNRNPA1 and RBFOX2 had the highest sequence specificity of their consensus motifs, and they are known to bind to highly conserved AGGGAG[32] and TGCATG[33] motifs, respectively. Next, for all ASPRIN SNPs of a given RBP, we obtained two sets of sequences that corresponded to the two alleles, i.e., one with high binding affinity and the other with low binding affinity. Finally, we used the position weight matrix that was obtained for all RBP consensus motifs by Zagros and calculated the motif scores for the two sets of sequences using STORM[26] (Figure S10 and Material and Methods). Figure 5B shows the motif scores of five RBPs with high (HNRNPA1, RBFOX2), median (DKC1), and low (NCBP2, XRN2) consensus motif sequence specificity. Variants in different positions within the consensus motif did not seem to affect binding equally. For example, for HNRNPA1, variants in position 5 of the motif had a more significant effect on binding than did variants in other positions. This result shows that not all positions in the consensus motif contribute equally to RBP-RNA interactions.

To further explore the relationship between the ASPRIN SNPs and RBP consensus motifs, we defined a motif impact score for each RBP and its associated ASPRIN SNP set as the maximum difference of average motif score between the two alleles with high versus low binding affinity in the window of six nucleotides overlapping the ASPRIN SNP (see details in Figure S10). We observed a positive correlation (Pearson correlation coefficient = 0.29, p value < 0.05) between the motif impact score and the sequence specificity of a given RBP's consensus motif (Figure 5C), suggesting that for highly sequence-specific RBPs, ASPRIN SNPs tend to affect binding by altering the consensus binding motifs within the RNA. For instance, in the case of HNRNPA1 and RBFOX2, we observed a higher motif score for alleles with higher binding affinity, while for NCBP2 and XRN2, we did not observe noticeable differences in motif scores between the two alleles in any position of their consensus motif (Figure 5C).

ASPRIN Can Help Reveal Causal Variants Affecting Alternative Splicing

Finally, we investigated whether ASPRIN can help reveal causal genetic variants that affect post-transcriptional gene regulation. For this analysis, we focused on the genetic variation of alternative splicing. A series of population-scale transcriptome studies have revealed widespread

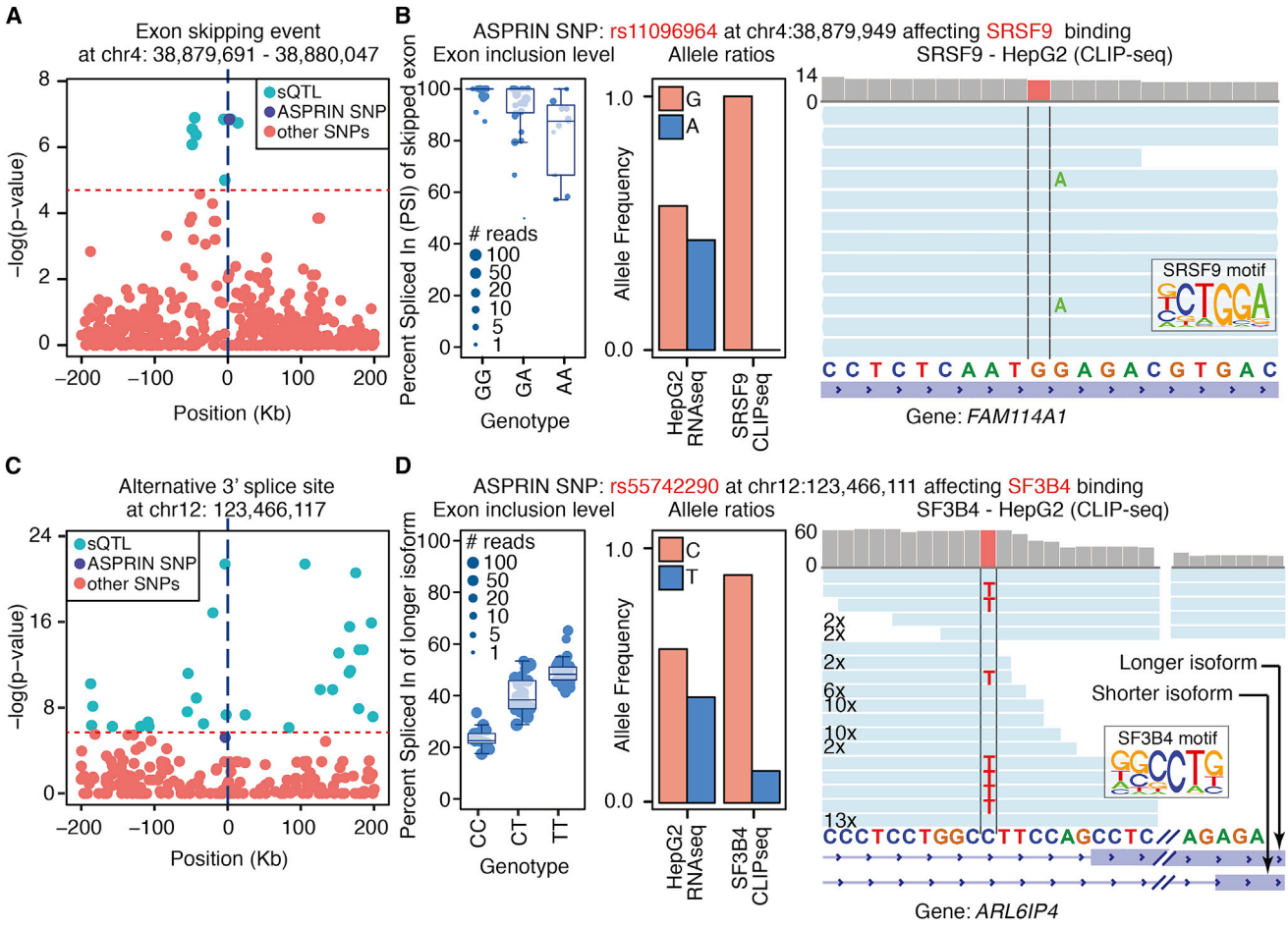

Figure 6. ASPRIN Helps Reveal Causal Variants Affecting Alternative Splicing

(A) Distribution of GLiMMPS p values around the exon skipping event in *FAM114A1*. For each SNP, the p value indicates the significance of correlation between genotype and exon inclusion level within a 400-kb window centered on the splicing event.

(B) Plots indicating the correlation of exon inclusion level with genotype for the ASPRIN SNP, differential binding of SRSF9 to the ASPRIN SNP that is in high LD with the GLiMMPS sQTL, and CLIP-seq allelic coverage on the ASPRIN SNP illustrating the effect of the SNP on the RBP consensus motif. The middle line of the boxplot represents median value. The low and high ends of the box represent the 25% and 75% quantiles, respectively. The two whiskers extend to 1.5 times the interquartile range.

(C and D) Similar plots are shown for a GLiMMPS sQTL involving alternative 3′ splice site usage in *ARL6IP4*, along with an ASPRIN SNP with differential binding of SF3B4 that is in high LD with the sQTL.

alternative splicing variation among human individuals,[4] but it remains challenging to pinpoint the causal genetic variants underlying this splicing variation. To match our ASPRIN analysis of the HepG2 liver cell line, we analyzed liver RNA-seq data along with matching genotype data of 71 individuals from the GTEx consortium (v6). We performed a transcriptome-wide scan of splicing quantitative trait loci (sQTLs) using GLiMMPS[28] and obtained ASPRIN SNPs correlated with GLiMMPS sQTLs (see details in Material and Methods).

Our joint ASPRIN and GLiMMPS analyses revealed candidate causal SNPs that affected alternative splicing via allele-specific protein-RNA interactions. For example, GLiMMPS identified several SNPs that were significantly associated with an exon-skipping event in *FAM114A1*, one of which was an ASPRIN SNP (Figure 6A). The genotype at the ASPRIN SNP was significantly associated with the level of exon inclusion, with the GG and AA genotypes showing the highest and lowest levels of exon inclusion, respectively (Figure 6B). The ASPRIN analysis indicated that the G allele was associated with significantly greater binding by the splicing factor SRSF9 (Figure 6B), while the A allele disrupted binding at the highly conserved "G" nucleotide at the fourth position of the SRSF9 consensus motif (Figure 6B). Collectively, these data suggest that the G-to-A SNP disrupted the binding of the splicing activator SRSF9, leading to reduced inclusion of the *FAM114A1* exon. Similarly, we identified an ASPRIN SNP for the splicing factor SF3B4, which was significantly associated with an alternative 3′ splice site event in *ARL6IP4* (MIM: 607668) (Figures 6C and 6D). This C-to-T SNP was located seven nucleotides upstream of the intron-exon boundary and disrupted a highly conserved "C" nucleotide at the fourth position of the SF3B4 consensus motif. This was reflected by a much lower percentage of the T allele in the SF3B4 CLIP-seq data

than in the RNA-seq data and increased usage of an upstream cryptic 3′ splice site for the TT genotype (Figures 6C and 6D). Overall, our results show that ASPRIN can help pinpoint causal variants within a window of SNPs that are correlated with levels of alternative splicing and in high linkage disequilibrium with each other.

We further associated ASPRIN SNPs with GWAS SNPs.[29] Specifically, we used the LD map of a CEU population to calculate LD correlations between all ASPRIN SNPs and SNPs associated with diseases and traits in the NHGRI GWAS catalog.[29] Tables S5 and S6 show all ASPRIN SNPs in high LD ($r^2 > 0.8$) with GWAS SNPs in HepG2 and K562 cell lines, respectively. These tables can be used by researchers to narrow down their search for candidate causal SNPs from GWAS signals of human traits or diseases.

Discussion

We report ASPRIN, a computational tool for identifying genetic variants that may affect RBP-RNA interactions, by quantifying and contrasting the allelic ratios of heterozygous SNPs in CLIP-seq versus RNA-seq data. Unlike previous work that relied on short RBP consensus motifs,[15,16] ASPRIN adopts a data-driven approach to directly observe the allelic preference of RBPs in CLIP-seq data, using matching RNA-seq data from the same cell type as the control. Our comprehensive ASPRIN analysis of 166 RBPs in two ENCODE cell lines identified 55,646 candidate allele-specific protein-RNA interaction events. These events may provide valuable information for interpreting causal signals underlying human transcriptomic variation and phenotypic diversity. Of note, recent population transcriptomic studies (such as the GTEx project[27]) have revealed widespread genetic variation of gene expression and RNA processing in human populations, but identifying the causal SNPs underlying such regulatory variation remains difficult. The ASPRIN analysis provides an independent source of information that may assist the fine mapping of SNPs associated with gene expression levels or RNA processing patterns. In this work, we present two example cases in which the ASPRIN analysis reveals the likely causal variant responsible for splicing QTLs in the human liver. Future studies integrating other layers of RNA regulatory processes may reveal ASPRIN SNPs that causally impact other aspects of RNA processing and metabolism in human cells.

Supplemental Data

Supplemental Data can be found with this article online at https://doi.org/10.1016/j.ajhg.2019.01.018.

Acknowledgments

The authors thank Drs. Levon Demirdjian and Ying Nian Wu for insightful discussions and the ENCODE Consortium and the ENCODE production laboratories for generating the eCLIP and RNA-seq data. This work was supported by National Institutes of Health grants R01GM088342 and U01CA233074 to Y.X. E.B.S. was partly supported by National Institutes of Health T32 Tumor Cell Biology Training Grant (T32CA009056).

Declaration of Interests

Y.X. is a scientific co-founder of Panorama Medicine Inc.

Received: August 22, 2018
Accepted: January 29, 2019
Published: February 28, 2019

Web Resources

ASPRIN source code, https://github.com/Xinglab/ASPRIN
dbSNP, https://www.ncbi.nlm.nih.gov/projects/SNP/
ENCODE, https://www.encodeproject.org/
OMIM, http://www.omim.org/
Picard, http://broadinstitute.github.io/picard/
RADAR database version 2, http://lilab.stanford.edu/GokulR/database/Human_AG_all_hg19_v2.txt
Repbase, https://www.girinst.org/downloads/
SRA, https://www.ncbi.nlm.nih.gov/sra

References

1. Glisovic, T., Bachorik, J.L., Yong, J., and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett. *582*, 1977–1986.
2. Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell *136*, 777–793.
3. Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends Genet. *24*, 416–425.
4. Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The expanding landscape of alternative splicing variation in human populations. Am. J. Hum. Genet. *102*, 11–26.
5. Wang, G.-S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat. Rev. Genet. *8*, 749–761.
6. Hentze, M.W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. Nat. Rev. Mol. Cell Biol. *19*, 327–341.
7. Moore, M.J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. Science *309*, 1514–1518.
8. Bahrami-Samani, E., Penalva, L.O., Smith, A.D., and Uren, P.J. (2015). Leveraging cross-link modification events in CLIP-seq for motif discovery. Nucleic Acids Res. *43*, 95–103.
9. Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.
10. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.-C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141.
11. König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. *17*, 909–915.

12. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods *13*, 508–514.
13. Chakrabarti, A.M., Haberman, N., Praznik, A., Luscombe, N.M., and Ule, J. (2018). Data science issues in studying protein–RNA interactions with CLIP technologies. Ann. Rev. Biomed. Data Sci. *1*, 235–261.
14. Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. *42*, 13534–13544.
15. Mao, F., Xiao, L., Li, X., Liang, J., Teng, H., Cai, W., and Sun, Z.S. (2016). RBP-Var: a database of functional variants involved in regulation mediated by RNA-binding proteins. Nucleic Acids Res. *44* (D1), D154–D163.
16. Singh, B., Trincado, J.L., Tatlow, P.J., Piccolo, S.R., and Eyras, E. (2018). Genome sequencing and RNA-motif analysis reveal novel damaging noncoding mutations in human tumors. Mol. Cancer Res. *16*, 1112–1124.
17. Bahrami-Samani, E., Vo, D.T., de Araujo, P.R., Vogel, C., Smith, A.D., Penalva, L.O., and Uren, P.J. (2015). Computational challenges, tools, and resources for analyzing co- and post-transcriptional events in high throughput. Wiley Interdiscip. Rev. RNA *6*, 291–310.
18. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
19. 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.
20. Sherry, S.T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.
21. Ramaswami, G., and Li, J.B. (2014). RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. *42*, D109–D113.
22. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal *17*, 10–12.
23. Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 11.
24. Uren, P.J., Bahrami-Samani, E., Burns, S.C., Qiao, M., Karginov, F.V., Hodges, E., Hannon, G.J., Sanford, J.R., Penalva, L.O., and Smith, A.D. (2012). Site identification in high-throughput RNA-protein interaction data. Bioinformatics *28*, 3013–3020.
25. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
26. Schones, D.E., Smith, A.D., and Zhang, M.Q. (2007). Statistical significance of cis-regulatory modules. BMC Bioinformatics *8*, 19.
27. Ward, M.C., and Gilad, Y. (2017). Human genomics: Cracking the regulatory code. Nature *550*, 190–191.
28. Zhao, K., Lu, Z.X., Park, J.W., Zhou, Q., and Xing, Y. (2013). GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. Genome Biol. *14*, R74.
29. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.
30. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
31. Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Blue, S.M., Dominguez, D., Cody, N.A., Olson, S., Sundararaman, B., et al. (2017). A large-scale binding and functional map of human RNA binding proteins. bioRxiv. https://doi.org/10.1101/179648.
32. Burd, C.G., and Dreyfuss, G. (1994). RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. EMBO J. *13*, 1197–1204.
33. Damianov, A., Ying, Y., Lin, C.-H., Lee, J.-A., Tran, D., Vashisht, A.A., Bahrami-Samani, E., Xing, Y., Martin, K.C., Wohlschlegel, J.A., and Black, D.L. (2016). Rbfox proteins regulate splicing as part of a large multiprotein complex LASR. Cell *165*, 606–619.

The American Journal of Human Genetics, Volume *104*

Supplemental Data

Discovery of Allele-Specific Protein-RNA Interactions in Human Transcriptomes

Emad Bahrami-Samani and Yi Xing

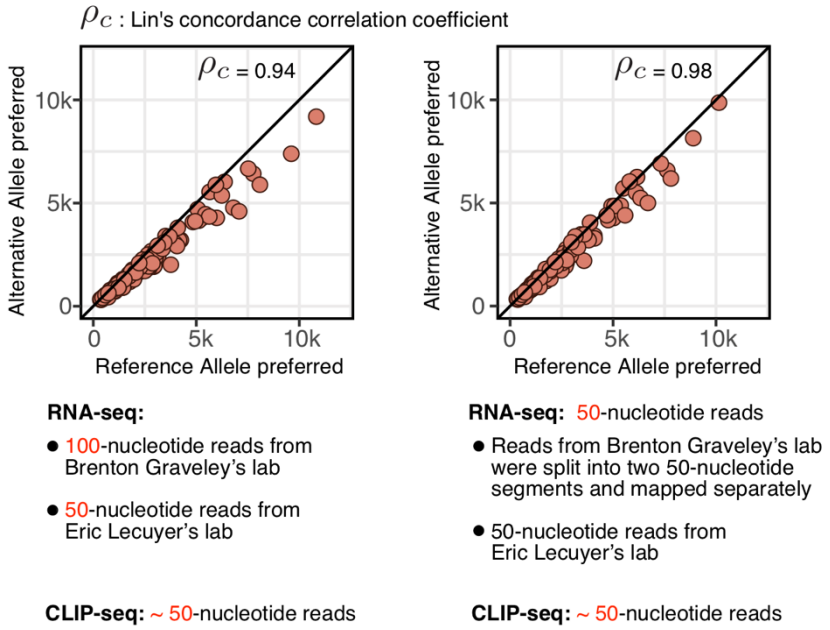

Figure S1: 100 bp RNA-seq reads in the ENCODE data were split into two 50 bp segments and mapped separately to alleviate systematic mapping bias for the reference over the alternative alleles in CLIP-seq data compared to the RNA-seq data.

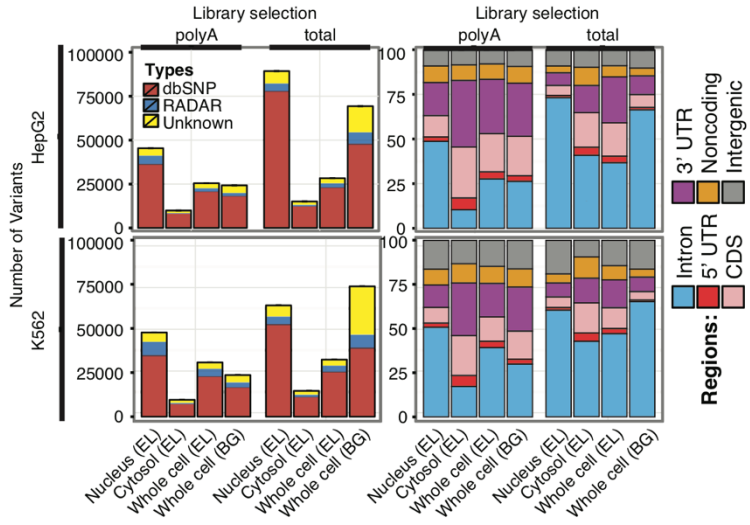

Figure S2: The highest number of SNPs were called from the nuclear RNA-seq data and the lowest number of SNPs from the cytosolic RNA-seq data in both HepG2 and K562 cell lines. SNPs from cytosolic polyA+ RNA-seq data were enriched for exonic regions within UTRs (Untranslated Regions) and CDS (Coding Segments) and depleted for intronic regions within pre-mRNAs.

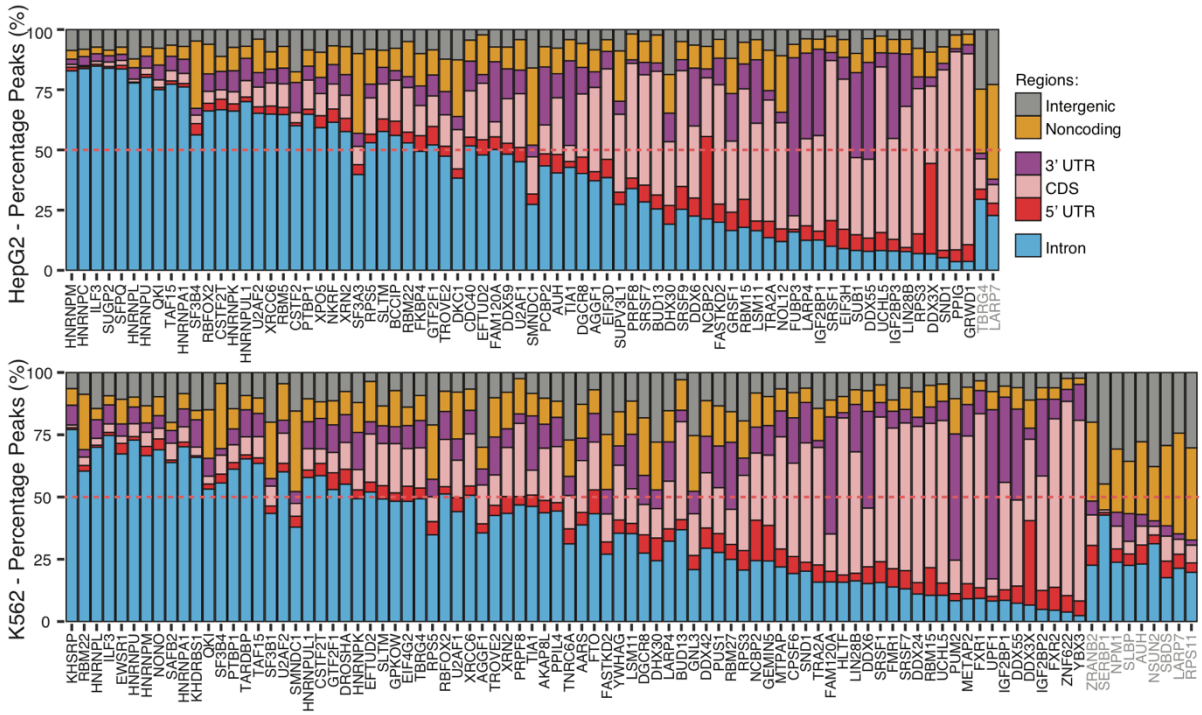

Figure S3: The complete distributions of peaks in different regions for all RBPs.

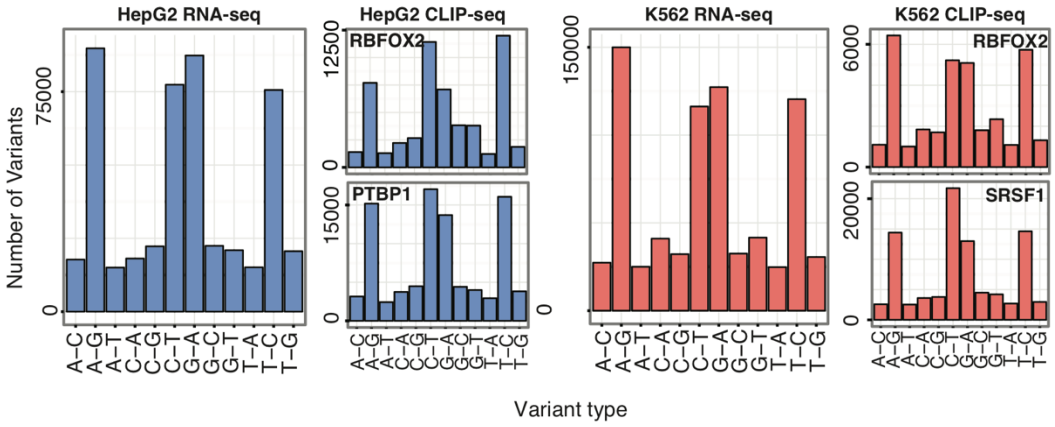

Figure S4: Distribution of variant types called from CLIP-seq and RNA-seq data in HepG2 and K562.

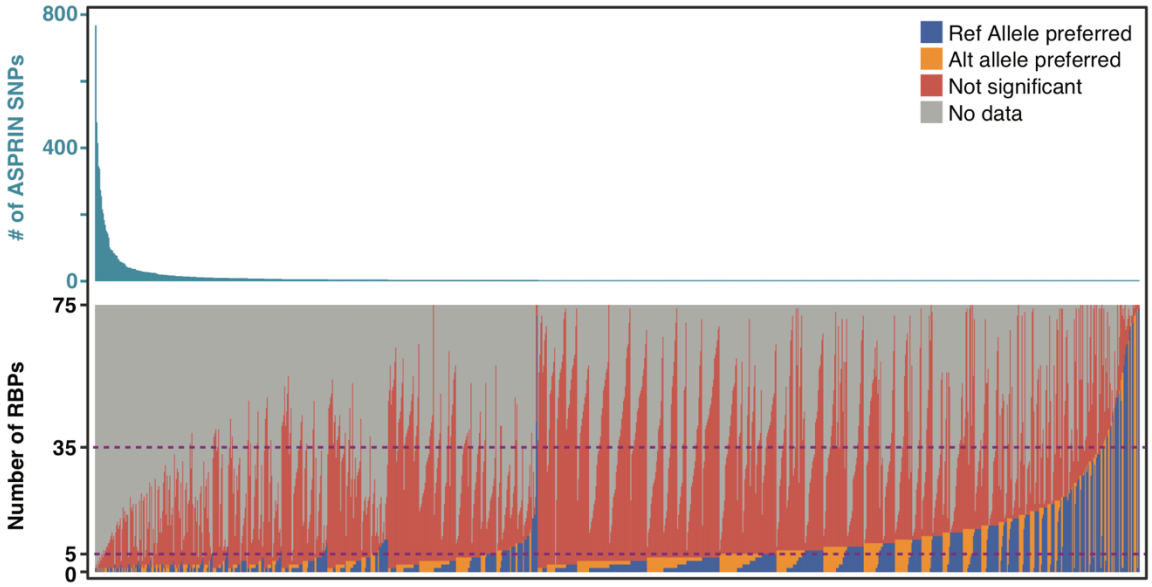

Figure S5: Distribution of ASPRIN outcome on all ASPRIN SNPs for all RBPs in HepG2 cell line.

HepG2:

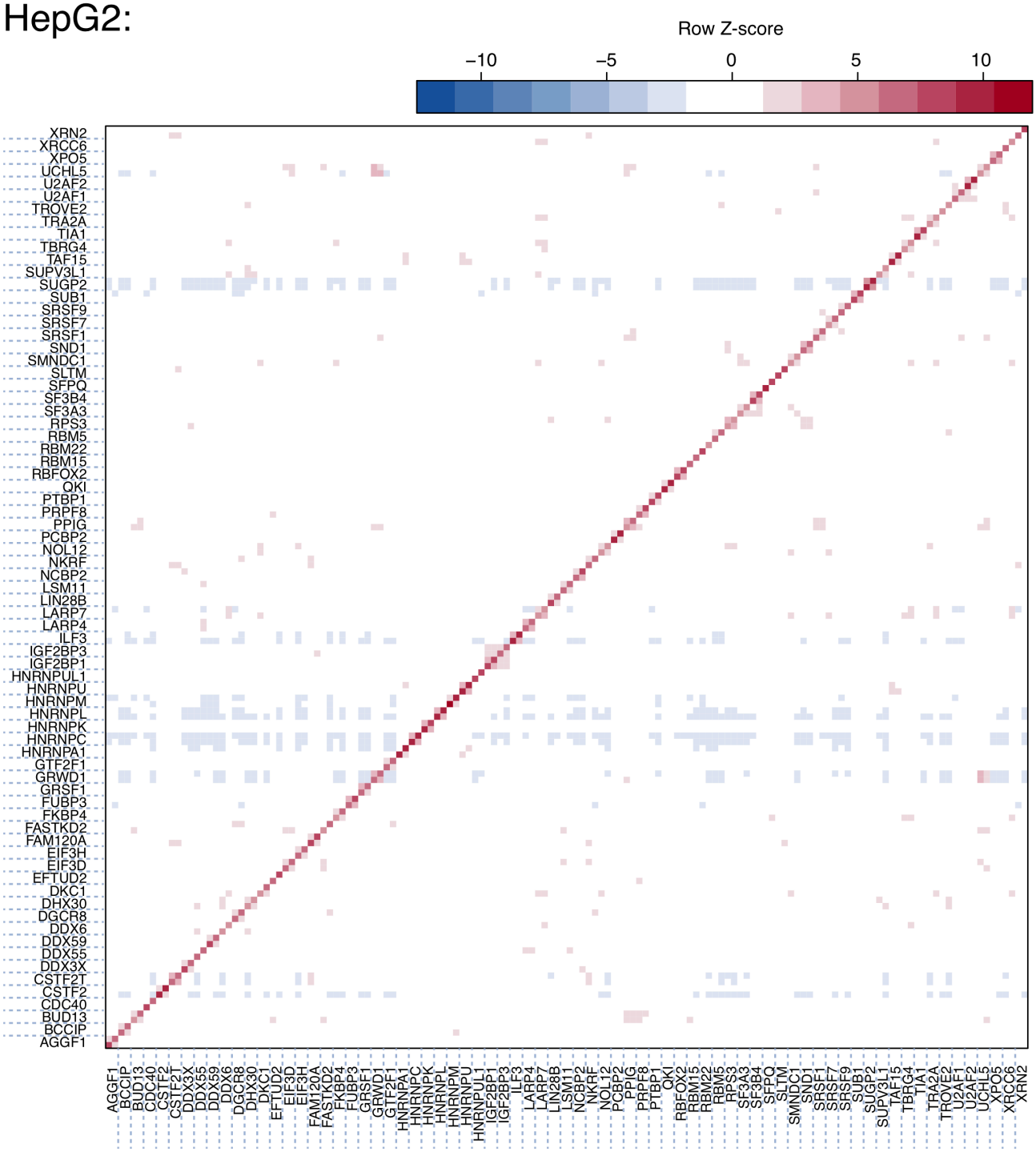

Figure S6: Intersection over union ASPRIN SNPs for all pairs of two replicates of all eCLIP data sets in HepG2.

K562:

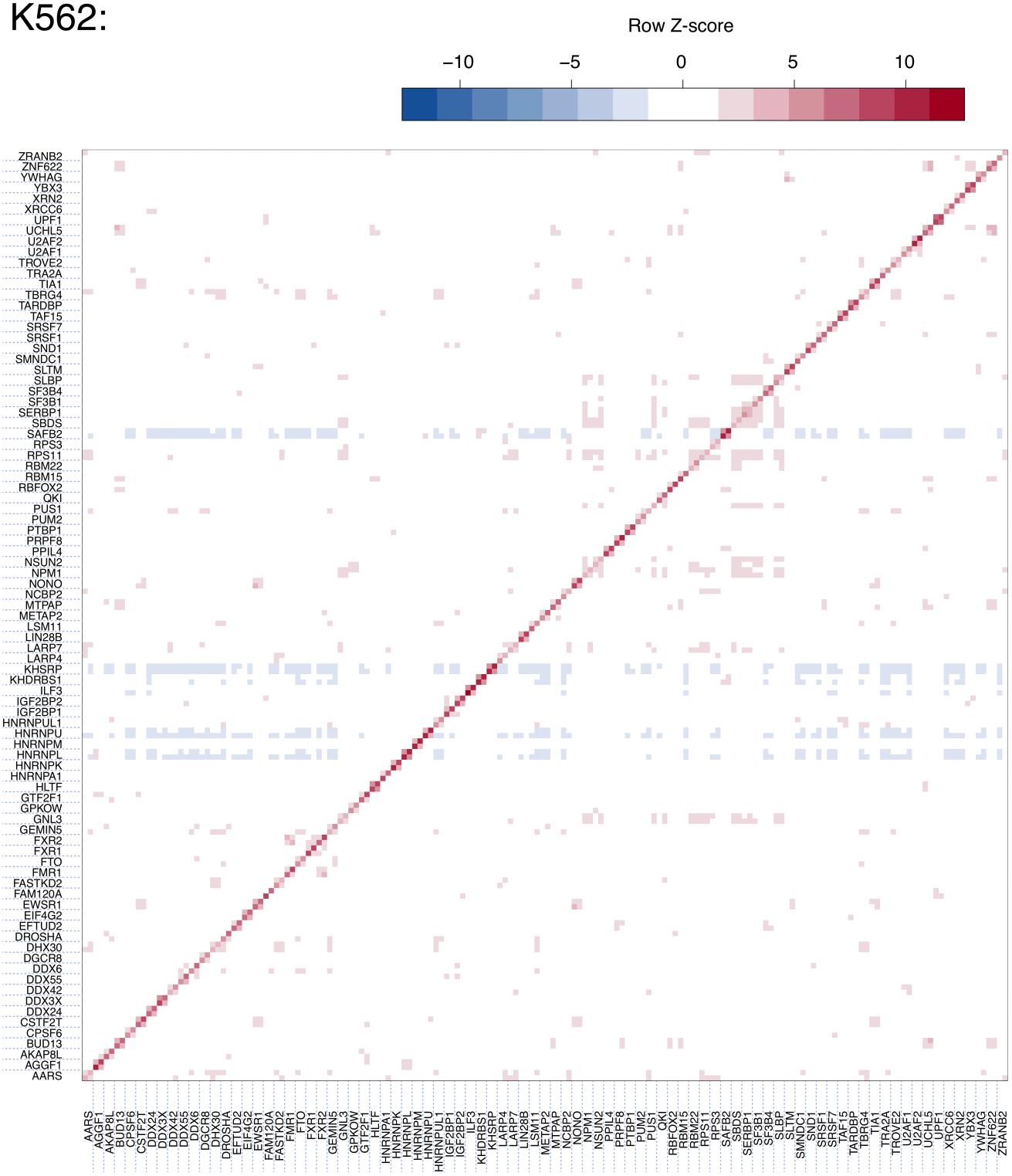

Figure S7: Intersection over union ASPRIN SNPs for all pairs of two replicates of all eCLIP data sets in K562.

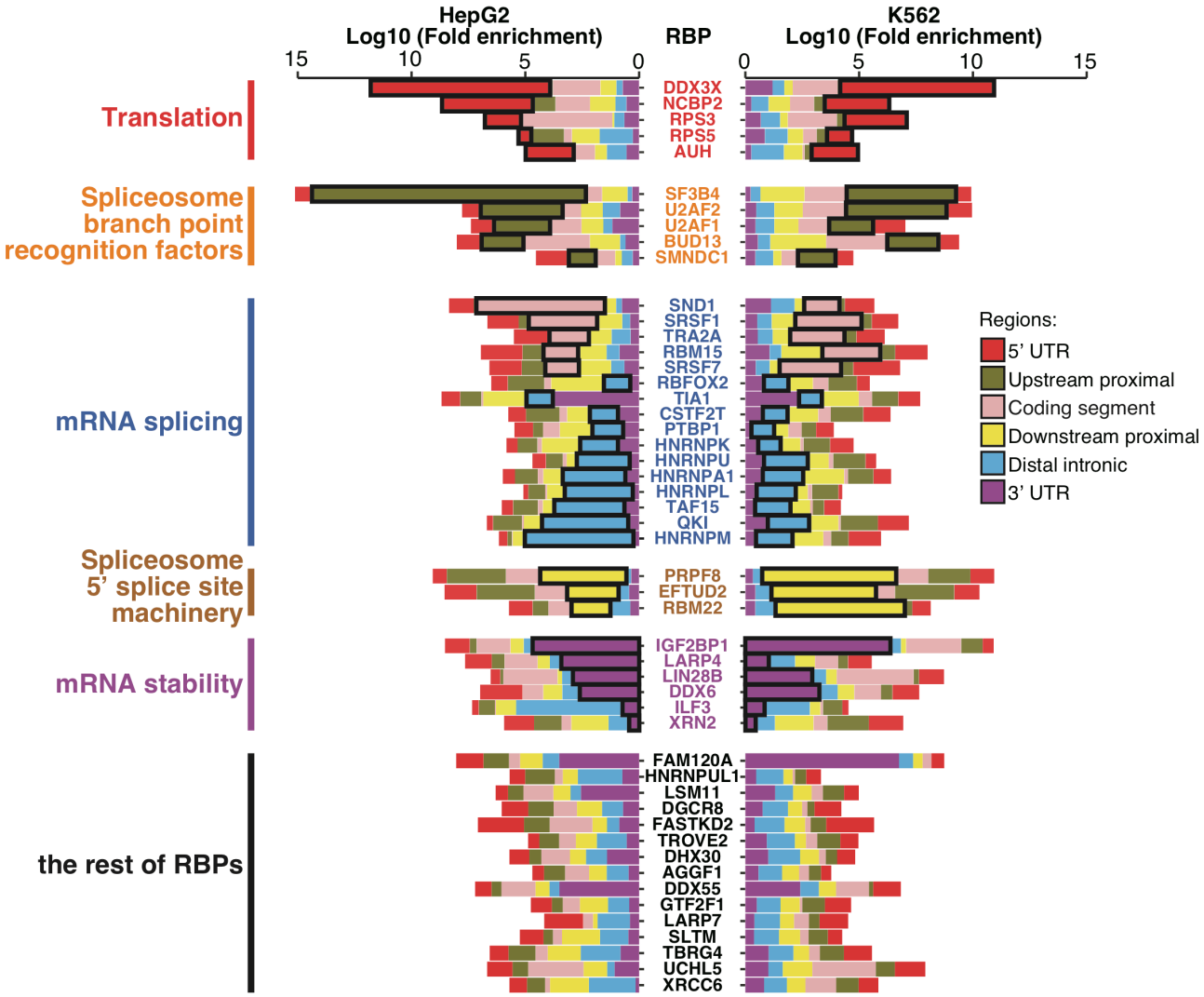

Figure S8: Enrichment of ASPRIN SNPs in genomic region in RBPs in both HepG2 and K562 cell lines.

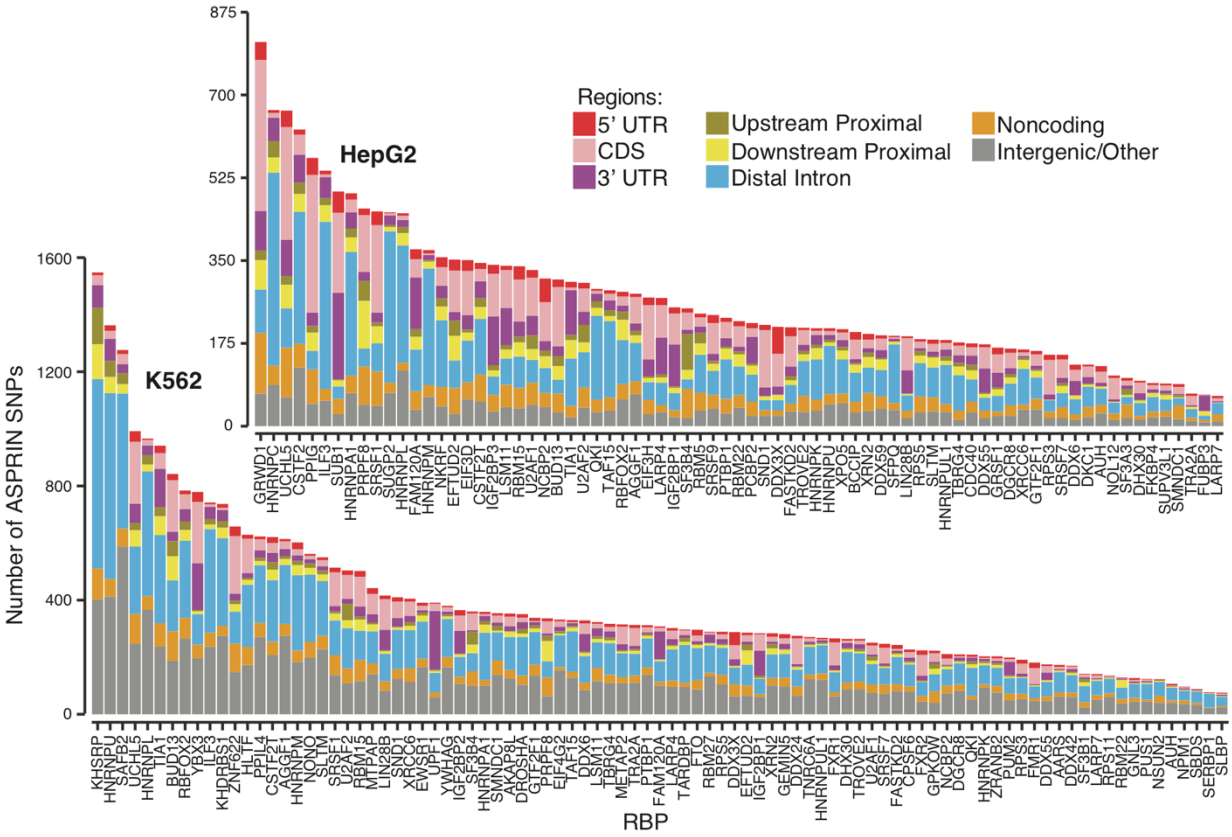

Figure S9: The numbers of ASPRIN SNPs for each RBP in HepG2 and K562.

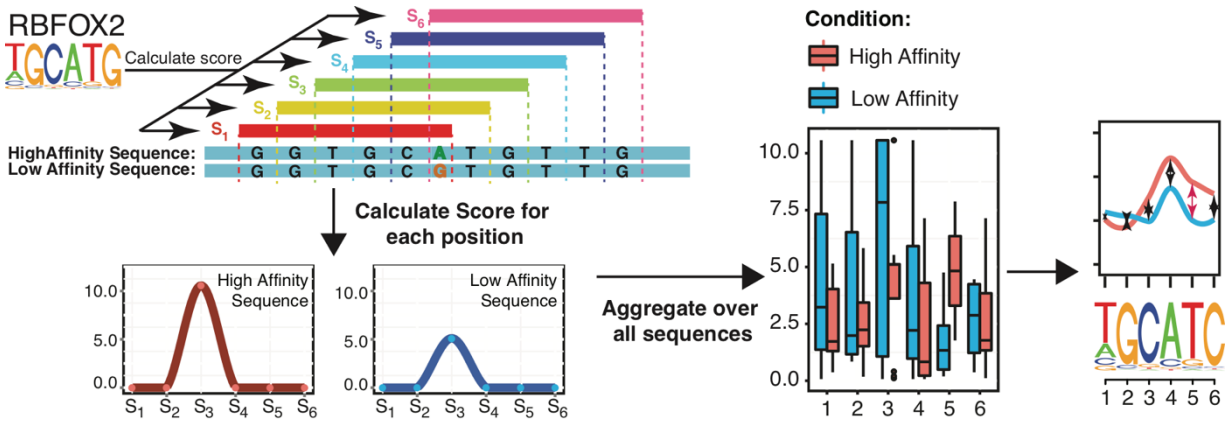

Figure S10: The procedure to investigate potential effects of ASPRIN SNPs on RBP consensus motifs.

Supplementary methods

1. **Detailed steps of calling variants from RNA-seq data**

- We used the STAR 2-pass alignment method to map all the RNA-seq to the hg19 genome. In this procedure, reads are mapped to the standard genome index, then a new index is created using splice junction information contained in the file SJ.out.tab from the first pass. The resulting index is then used to produce the final alignments.

```
genomeDir=/path/to/hg19
mkdir $genomeDir
STAR --runMode genomeGenerate --genomeDir $genomeDir --genomeFastaFiles hg19.fa\
  --runThreadN <n>

runDir=/path/to/1pass
mkdir $runDir
cd $runDir
STAR --genomeDir $genomeDir --readFilesIn mate1.fq mate2.fq --runThreadN <n>

genomeDir=/path/to/hg19_2pass
mkdir $genomeDir
STAR --runMode genomeGenerate --genomeDir $genomeDir --genomeFastaFiles hg19.fa \
     --sjdbFileChrStartEnd /path/to/1pass/SJ.out.tab --sjdbOverhang t --runThreadN <n>
runDir=/path/to/2pass
mkdir $runDir
cd $runDir
STAR --genomeDir $genomeDir --readFilesIn mate1.fq mate2.fq --runThreadN <n>
```

- We merged **total RNA-seq** data from **all fractions** and **all labs** together to make one substantial RNA-seq data set for each cell line.

```
java -Xmx8g -jar picard.jar MergeSamFiles SORT_ORDER=coordinate I=fn1.bam I=fn2.bam O=fn.b
am
```

- **Add read groups, sort, mark duplicates, and create index**: The resulting SAM file is used to add read groups, perform the sorting, and creating the index.

```
java -jar picard.jar AddOrReplaceReadGroups I=star_output.sam O=rg_added_sorted.bam SO=coo
rdinate RGID=id RGLB=library RGPL=platform RGPU=machine RGSM=sample   java -jar picard.jar
MarkDuplicates I=rg_added_sorted.bam O=dedupped.bam CREATE_INDEX=true VALIDATION_STRINGENC
Y=SILENT M=output.metrics
```

- **Split'N'Trim and reassign mapping qualities**: To correct for the cases where the reads that should be split between two exons are mapped with an indel at the end.

```
java -jar GenomeAnalysisTK.jar -T SplitNCigarReads -R ref.fasta -I dedupped.bam -o split.b
am -rf ReassignOneMappingQuality -RMQF 255 -RMQT 60 -U ALLOW_N_CIGAR_READS
```

- **Base Recalibration and Indel Realignment**: This is for recalibrating bases established on a set of known SNPs and realign after the splitting step.

```
java -Xmx8g -jar GenomeAnalysisTK.jar -T BaseRecalibrator –I split.bam -R /path/to/hg19.fa
-knownSites dbsnp_150.vcf -o recalibration_report.grp

java -Xmx8g -jar GenomeAnalysisTK.jar -T PrintReads -R /path/to hg19.fa -I split.bam -BQSR
recalibration_report.grp -o recalibration.bam -U ALLOW_N_CIGAR_READS
```

- **Variant calling**: The final stage to perform the variant calling. Details can be found in supplementary methods.

```
java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R ref.fasta -I recalibration.bam -dontU
seSoftClippedBases -stand_call_conf 20.0 -o output.vcf
```

Command lines that we used are taken from GATK homepage:

https://software.broadinstitute.org/gatk/documentation/article.php?id=3891.

2. eCLIP data analysis

For mapping the ENCODE eCLIP data, we followed the Standard Operating Procedure (SOP) that was published and described in length on the ENCODE website. This procedure was also used and described in [1, 2]. In summary,

- Adaptors are trimmed using cutadapt v1.10,

```
cutadapt -f fastq --match-read-wildcards --times 1 -e 0.1 -O 1 --quality-cutoff 6 -m 18 -a
NNNNNAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -g CTTCCGATCTACAAGTT -g CTTCCGATCTTGGTCCT -A AACTT
GTAGATCGGA -A AGGACCAAGATCGGA -A ACTTGTAGATCGGAA -A GGACCAAGATCGGAA -A CTTGTAGATCGGAAG -A
GACCAAGATCGGAAG -A TTGTAGATCGGAAGA -A ACCAAGATCGGAAGA -A TGTAGATCGGAAGAG -A CCAAGATCGGAAGA
G -A GTAGATCGGAAGAGC -A CAAGATCGGAAGAGC -A TAGATCGGAAGAGCG -A AAGATCGGAAGAGCG -A AGATCGGAA
GAGCGT -A GATCGGAAGAGCGTC -A ATCGGAAGAGCGTCG -A TCGGAAGAGCGTCGT -A CGGAAGAGCGTCGTG -A GGAA
GAGCGTCGTGT -o filename_1_adaptorRemoved.fastq -p filename_2_adaptorRemoved.fastq filename
_1.fastq filename_2.fastq > filename.metrics
```

- Round 2 of cutting adaptors to control for double ligation events,

```
cutadapt -f fastq --match-read-wildcards --times 1 -e 0.1 -O 5 --quality-cutoff 6 -m 18 -A
AACTTGTAGATCGGA -A AGGACCAAGATCGGA -AACTTGTAGATCGGAA -A GGACCAAGATCGGAA -A CTTGTAGATCGGAAG
-A GACCAAGATCGGAAG -A TTGTAGATCGGAAGA -A ACCAAGATCGGAAGA -A TGTAGATCGGAAGAG -A CCAAGATCGGA
AGAG -A GTAGATCGGAAGAGC -A CAAGATCGGAAGAGC -A TAGATCGGAAGAGCG -A AAGATCGGAAGAGCG -A AGATCG
GAAGAGCGT -A GATCGGAAGAGCGTC -A ATCGGAAGAGCGTCG -A TCGGAAGAGCGTCGT -A CGGAAGAGCGTCGTG -A G
GAAGAGCGTCGTGT -o filename_1_adaptorRemoved_round2.fastq -p filename_2_adaptorRemoved_roun
d2.fastq filename_1_adaptorRemoved.fastq filename_2_adaptorRemoved.fastq > filename.round2
.metrics
```

- Resulting reads are mapped to human specific version of RepBase using STAR 2.5.2a [3] to remove repetitive elements, control for spurious artifacts from rRNA and other repetitive reads. Repbase is downloaded from: http://www.girinst.org/downloads/,

```
STAR --runMode alignReads --runThreadN 4 --genomeDir starindex_dbs/repbase/ --readFilesIn
filenamefilename_1_adaptorRemoved_round2.fastq filenamefilename_2_adaptorRemoved_round2.fa
stq --outSAMunmapped Within --outFilterMultimapNmax 30 --outFilterMultimapScoreRange 1 --o
utFileNamePrefix filename_adaptorRemoved_round2_rep.bam outSAMattributes All --outStd BAM_
Unsorted --outSAMtype BAM Unsorted --outFilterType BySJout --outReadsUnmapped Fastx --outF
ilterScoreMin 10 --outSAMattrRGline ID:foo --alignEndsType EndToEnd --alignEndsProtrude 10
ConcordantPair > filename_adaptorRemoved_round2_rep.bam
```

- Unmapped output from STAR rmRep are sorted to account for issues with STAR not outputting first and second mate pairs in order:

```
fastq-sort --id filename_adaptorRemoved_round2_rmRep.bamUnmapped.out.mate1 > filename_adap
torRemoved_round2_rep.bamUnmapped.out.mate1

fastq-sort --id filename_adaptorRemoved_round2_rmRep.bamUnmapped.out.mate2 > filename_adap
torRemoved_round2_rep.bamUnmapped.out.mate2
```

- The resulting files are then mapped using star

```
STAR --runMode alignReads --runThreadN 4 --genomeDir starindex_dbs/human19 --readFilesIn f
ilename_adaptorRemoved_round2_rep.bamUnmapped.out.mate1 filename_adaptorRemoved_round2_rep
.bamUnmapped.out.mate2 --outSAMunmapped Within --outFilterMultimapNmax 1 --outFilterMultim
apScoreRange 1 --outFileNamePrefix filename_adaptorRemoved_round2_rmRep.bam --outSAMattrib
utes All --outStd BAM_Unsorted --outSAMtype BAM Unsorted --outFilterType BySJout --outRead
sUnmapped Fastx --outFilterScoreMin 10 --outSAMattrRGline ID:foo --alignEndsType EndToEnd
--alignEndsProtrude 10 ConcordantPair > filename_adaptorRemoved_round2_rmRep.bam
```

- PCR-duplicates were further removed using the randommers that are in the names of the reads.

```
python barcode_collapse_pe.py --bam filename_adaptorRemoved_round2_rmRep.bam --out_file fi
lename_adaptorRemoved_round2_rmRep.rmDup.bam --metrics_file filename_adaptorRemoved_round2
_rmRep.rmDup.metrics
```

- Two replicates are then sorted, merged and indexed:

```
Samtools sort -o filename_rep1_aligned_sorted.bam filename_rep1_adaptorRemoved_round2_rmRe
p.rmDup.bam

Samtools sort -o filename_rep2_aligned_sorted.bam filename_rep2_adaptorRemoved_round2_rmRe
p.rmDup.bam

Samtools merge filename_aligned_sorted.bam filename_rep1_ aligned_sorted.bam filename_rep2
_ aligned_sorted.bam

Samtools index filename_mapped_sorted.bam
```

- When calling peaks are necessary, second (paired-end) read was used to perform peak-calling using Piranha [4], using a bin size of 1nt. We consider significant peaks to be those that have a corrected p-value less than 0.01. Mapping and peak calling statistics are given in supplementary table S2.

```
Piranha -s -v -b 1 filename_aligned_sorted_mate2.bed filename_peaks.bed
```

- We performed region-level analysis by intersecting peaks with annotated regions in Gencode (v19).

```
Python ASPRIN/src/scripts/find_SNP_region.py -g /path/to/gencode.v19.chr_patch_hapl_scaff.
annotation.gtf -i asprin_output.asp -o asprin_output_genes_and_regions.asp
```

3. Splicing analysis

- RNA-seq and genotype data of liver tissues from 71 individuals (GTEx v6) were downloaded and mapped to the hg19 genome:

```
hisat2 -x grch37_tran/genome_tran -1 individual_i_1.fastq -2 individual_i_2.fastq --no-mix
ed --no-discordant -t --no-unal --dta-cufflinks -p 8 -S individual_i.sam

samtools view -bS individual_i.sam | samtools sort -o individual_i_sorted.bam
```

- Percent Spliced In (Psi) values were calculated for each splicing event in each individual using rMATS:

```
python /path/to/rmats_pipeline/rmats/rmats/asevent.py
        --b1 Liver-b1.txt
        --gtf /path/to/Homo_sapiens.GRCh37.75.gtf
        --od output_dir -t paired --nthread 4 --readLength 76 --anchorLength 1
        --tmp temp_dir --task both
```

where Liver-b1.txt contains the comma separated list of all the bam files for individuals for which we had a genotype.

4. Motif enrichment analysis

To run Zagros for motif discovery using sequence and structure information, the secondary structure data must first be obtained and saved using the “thermo” program, which is provided within the Zagros package:

```
thermo -o input.str input.fa
```

or

```
thermo -c path/to/chrom_directory -o input.str input.bed
```

After this step, by providing both the target and secondary structure file to Zagros the motif discovery is performed based on both.

```
zagros -t input.str –o zagros_output.mat input.fa
```

or

```
zagros -c path/to/chrom_directory -t input.str –o zagros_output.mat input.bed
```

To calculate the enrichment of a consensus position weight matrix obtained from Zagros in a set of sequences, we use STORM [5], as follows:

```
storm -v -n 1 -q -h -S –s highAffinity_sequences.fa zagros_output.mat > storm_output_scores.mat
```

References:

1. Conway, A.E., et al., *Enhanced CLIP Uncovers IMP Protein-RNA Targets in Human Pluripotent Stem Cells Important for Cell Adhesion and Survival.* Cell reports, 2016. **15**(3): p. 666-679.
2. Van Nostrand, E.L., et al., *Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP).* Nature methods, 2016. **13**(6): p. 508-514.
3. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
4. Uren, P.J., et al., *Site identification in high-throughput RNA-protein interaction data.* Bioinformatics, 2012. **28**(23): p. 3013-3020.
5. Schones, D.E., A.D. Smith, and M.Q. Zhang, *Statistical significance of cis-regulatory modules.* BMC bioinformatics, 2007. **8**(1): p. 1.