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

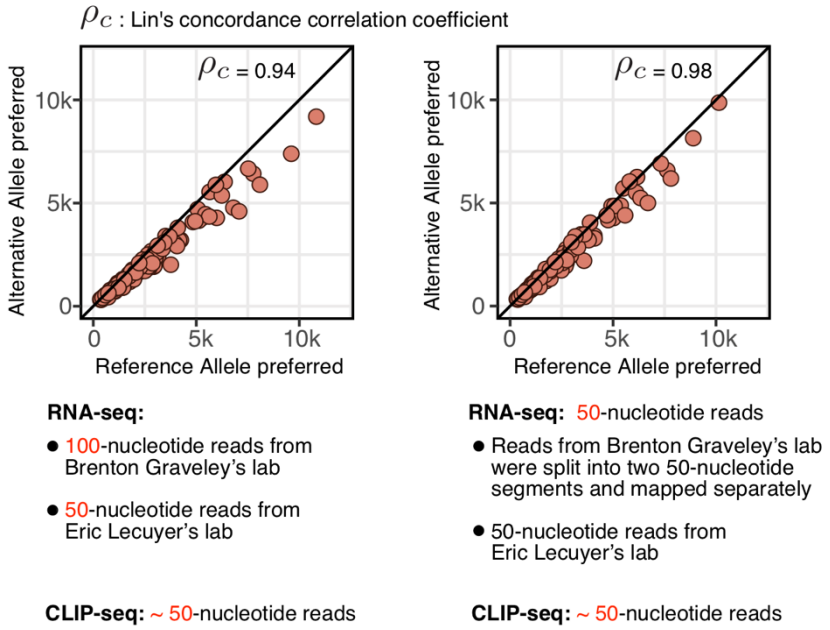

Figure S1: 100 bp RNA-seq reads in the ENCODE data were split into two 50 bp segments and mapped separately to alleviate systematic mapping bias for the reference over the alternative alleles in CLIP-seq data compared to the RNA-seq data.

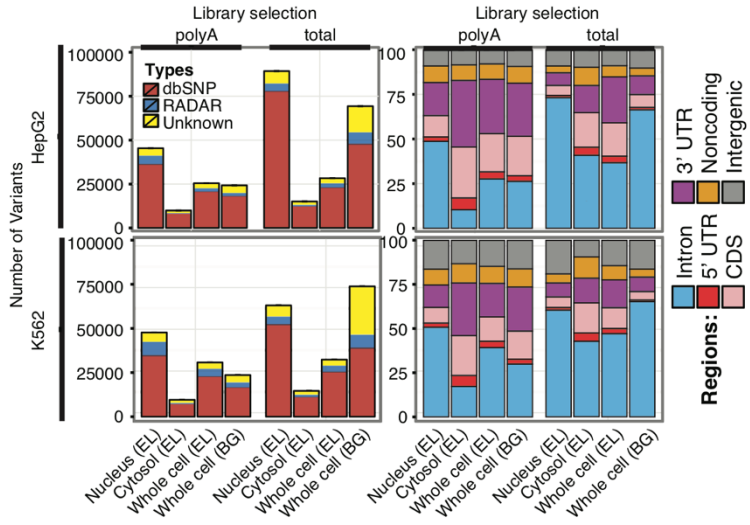

Figure S2: The highest number of SNPs were called from the nuclear RNA-seq data and the lowest number of SNPs from the cytosolic RNA-seq data in both HepG2 and K562 cell lines. SNPs from cytosolic polyA+ RNA-seq data were enriched for exonic regions within UTRs (Untranslated Regions) and CDS (Coding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

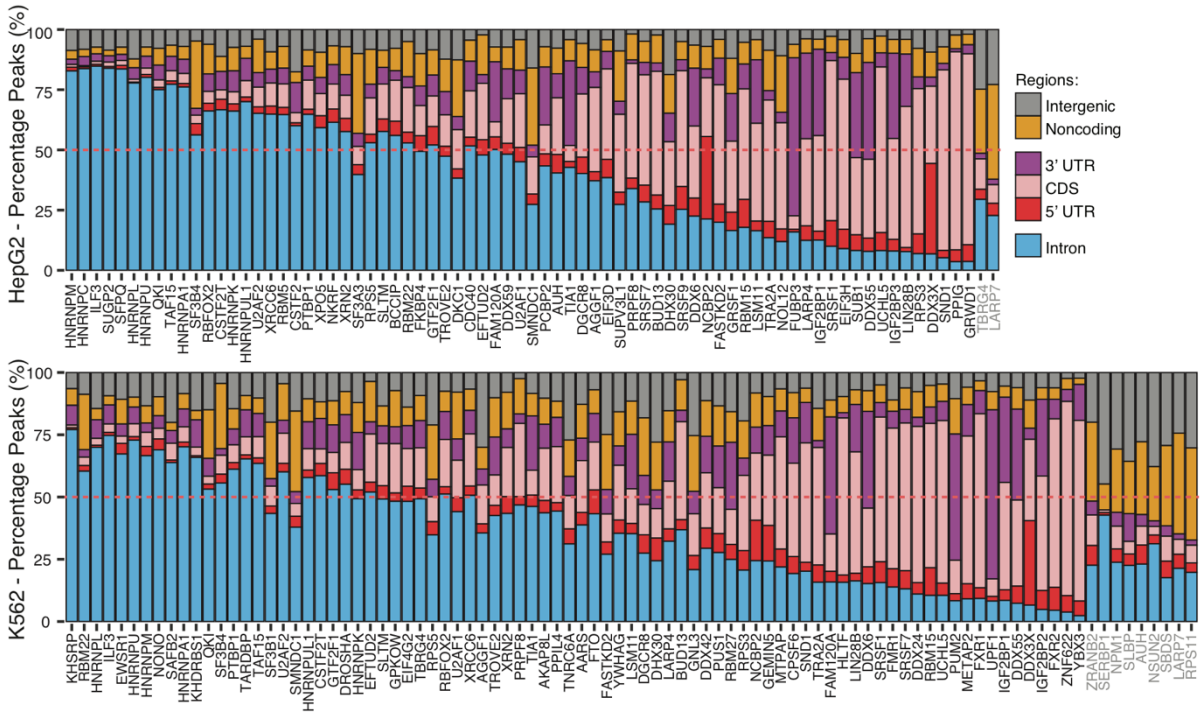

Figure S3: The complete distributions of peaks in different regions for all RBPs.

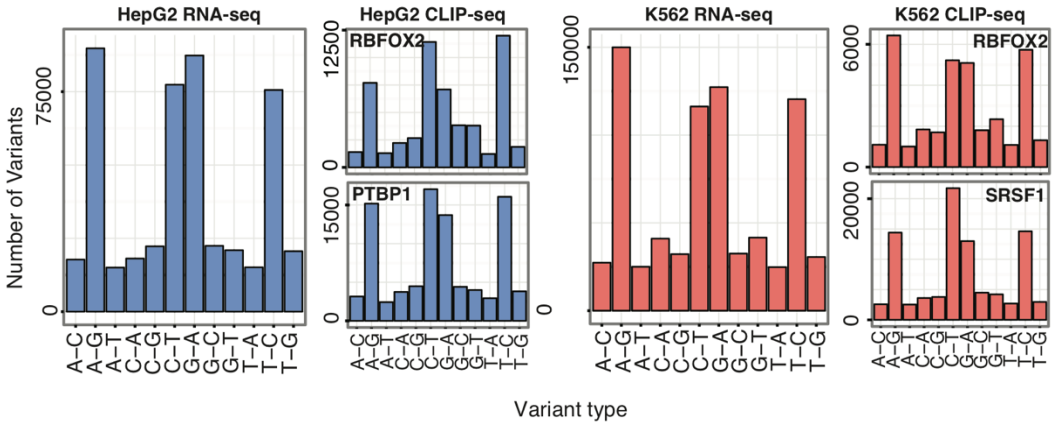

Figure S4: Distribution of variant types called from CLIP-seq and RNA-seq data in HepG2 and K562.

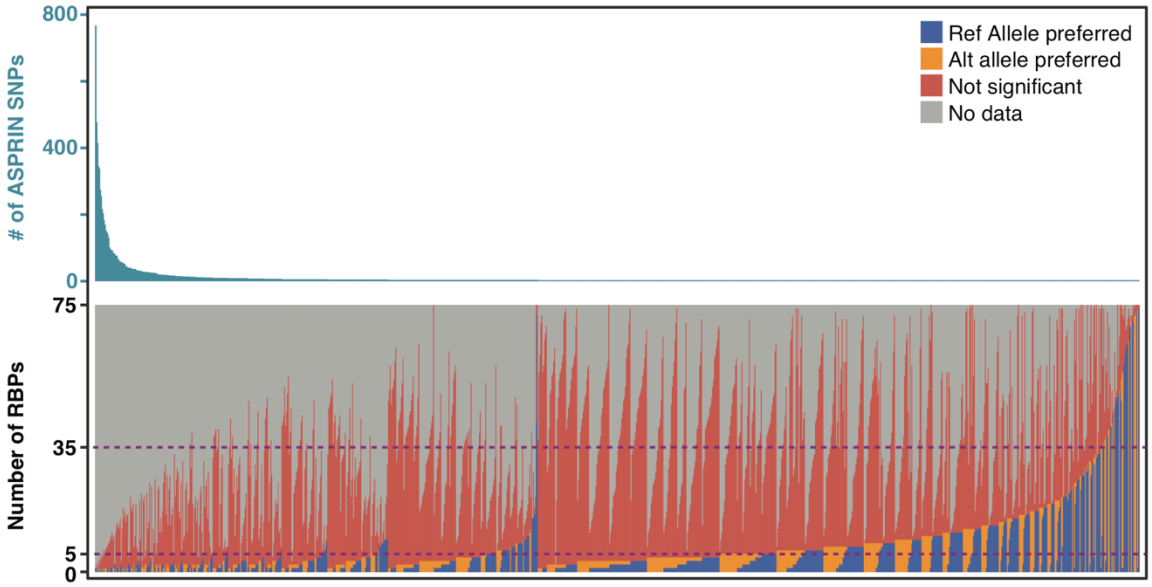

Figure S5: Distribution of ASPRIN outcome on all ASPRIN SNPs for all RBPs in HepG2 cell line.

HepG2:

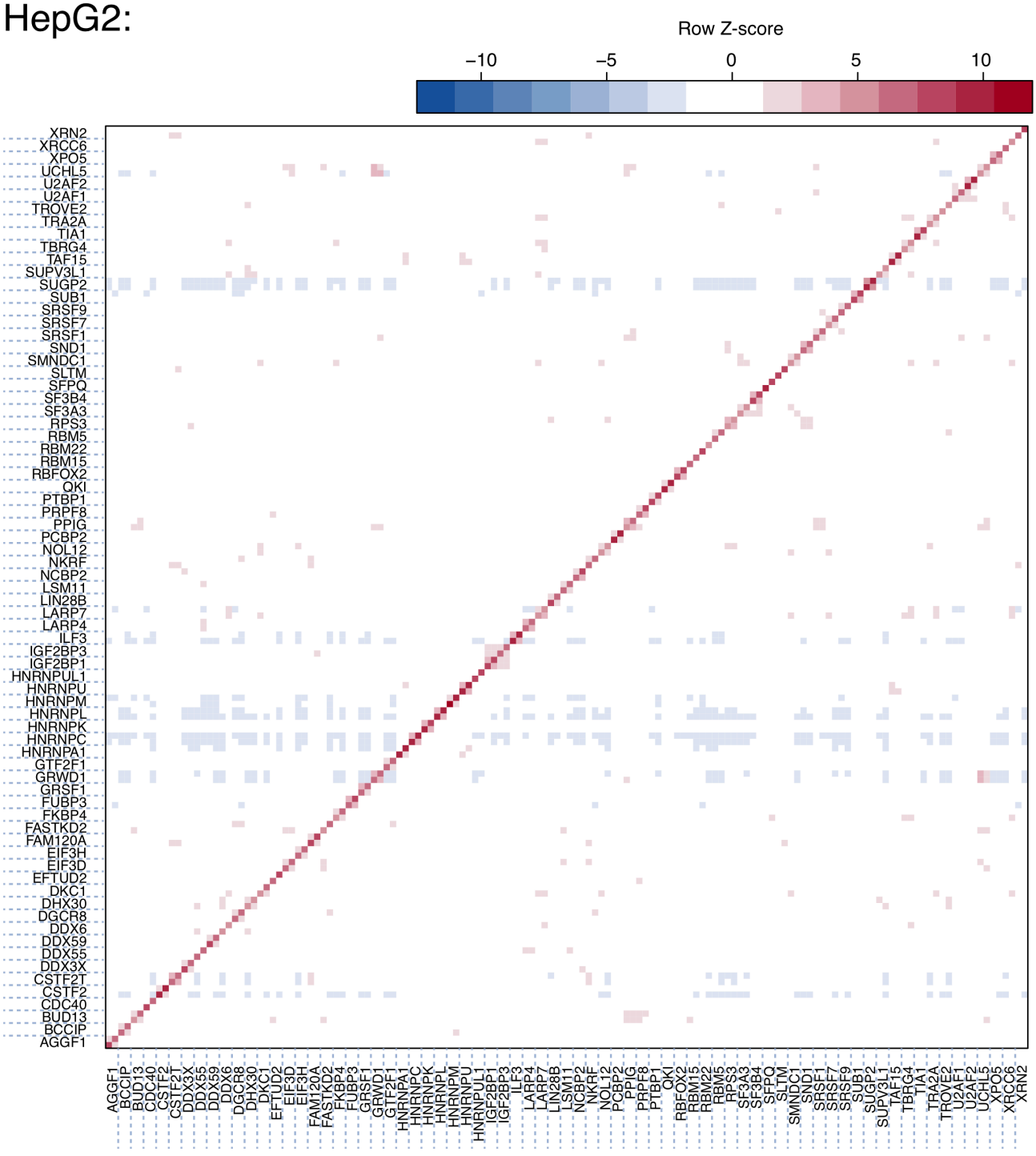

Figure S6: Intersection over union ASPRIN SNPs for all pairs of two replicates of all eCLIP data sets in HepG2.

K562:

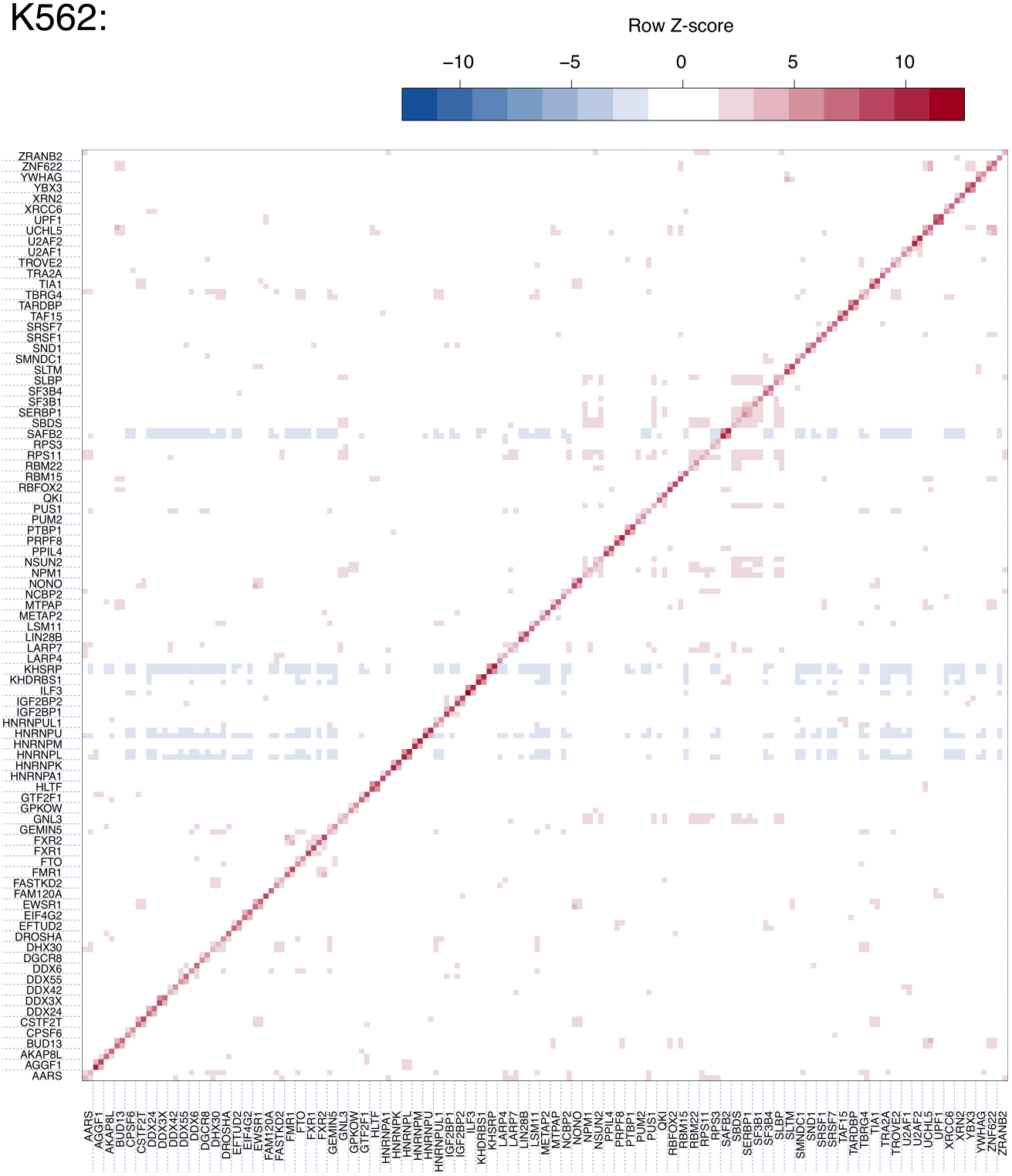

Figure S7: Intersection over union ASPRIN SNPs for all pairs of two replicates of all eCLIP data sets in K562.

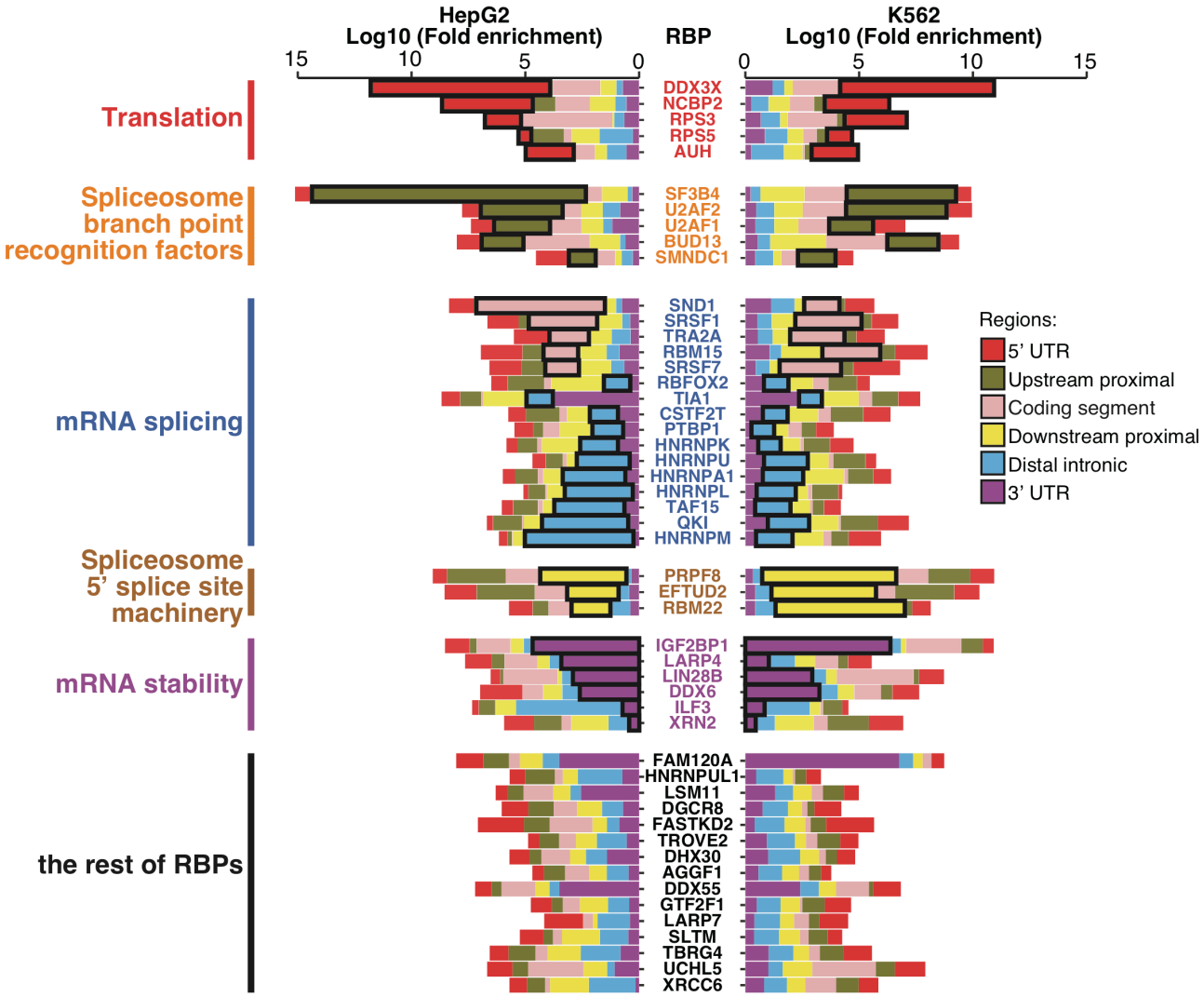

Figure S8: Enrichment of ASPRIN SNPs in genomic region in RBPs in both HepG2 and K562 cell lines.

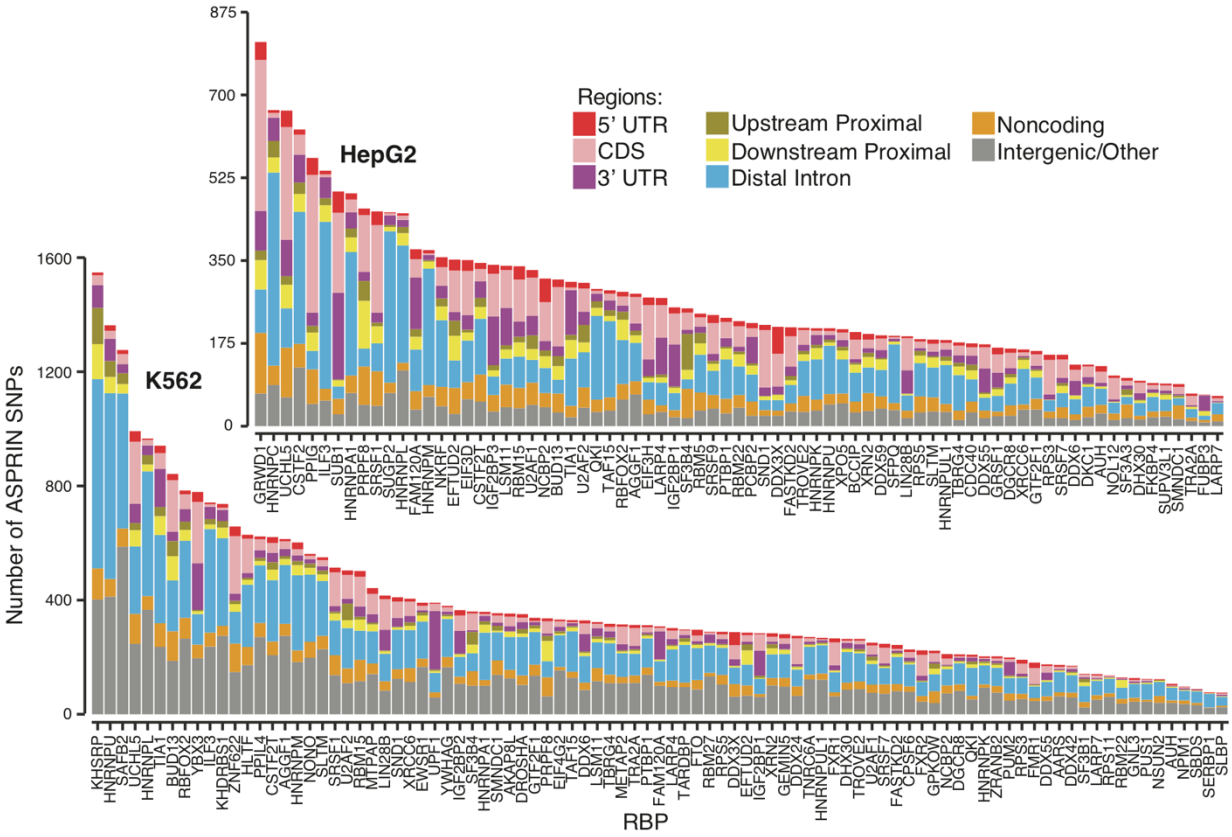

Figure S9: The numbers of ASPRIN SNPs for each RBP in HepG2 and K562.

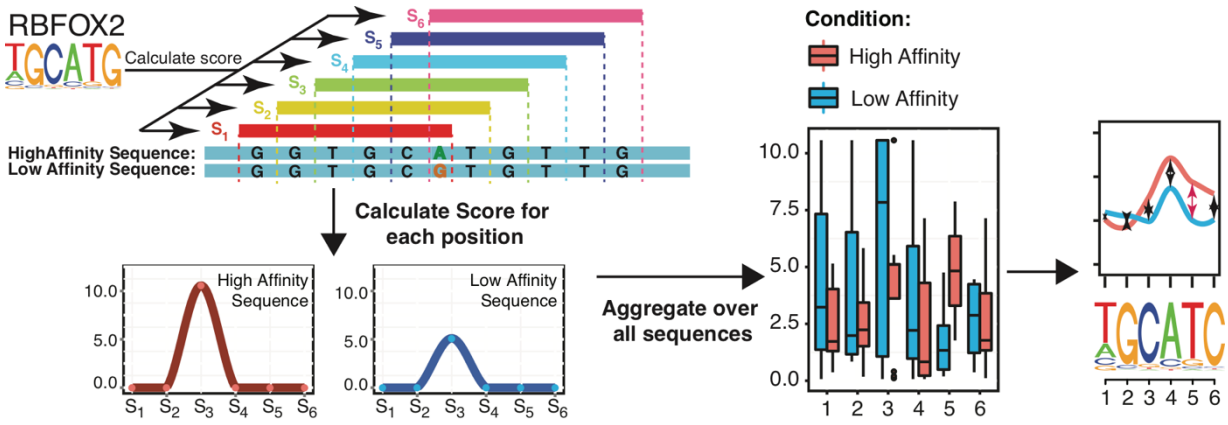

Figure S10: The procedure to investigate potential effects of ASPRIN SNPs on RBP consensus motifs.

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.