Supplemental Material

Alternative Splicing Modulated by Genetic Variants Demonstrates Accelerated Evolution Regulated by Highly Conserved Proteins

By Hsiao et al.

Supplemental Methods

RNA-Seq data and mapping

Paired-end RNA-Seq data (2x76nt) from seven human cell lines (GM12878, K562, HeLa, HepG2, HUVEC, NHEK, and H1-hESC) were downloaded from the ENCODE data repository (www.encodeproject.org) under the ENCODE Data Coordination Center accession number ENCSR037HRJ, or NCBI GEO accession number GSE30567, where two biological replicates are available for all cell lines except H1-hESC. The reads were mapped using a stringent mapping method described in our previous work (Bahn et al. 2012). Briefly, each end of the paired-end reads was mapped to the hg19 genome and transcriptome using BLAT (Kent 2002) and Bowtie (Langmead et al. 2009), allowing 12 mismatches in each read to collect as many mapping positions as possible. The following mapping parameters were used: BLAT (version 3.4): -minIdentity=75 -tileSize=11; Bowtie (version 0.12.3): -k 80 -e 140 -n 3 -l 20. After initial mapping, we checked for proper pairing with the correct orientations and distance of at most 500,000 base pairs between reads in a pair. For those reads passing the pairing filter, we collected only the read pairs that uniquely mapped as a pair with fewer than five mismatches on each read but did not map anywhere else as a pair with fewer than 12 mismatches on each read to eliminate ambiguous mappings to repetitive genomic regions and false mapping results due to sequencing errors, genetic variations, or RNA editing sites.

Prediction of eGMAS events

The eGMAS analysis was applied to the CA+ and NA+ data of each cell line. Biological replicates were combined, and duplicated reads were removed while retaining the one with best mapping quality. The analysis was conducted using our previous method (Li et al. 2012). Since most cell lines (except GM12878) have no genome-sequencing data, we used our in-house pipeline to identify cell line-specific SNVs and combined them with the common SNVs from dbSNP (Sherry et al. 2001). As the ENCODE RNA-Seq data are strand-specific, we used an updated version of our previous method that can handle strand-specific data to identify the eGMAS SNVs. A package for this method is available at: https://github.com/cyruschan/ASARP.

PhastCons analysis

To evaluate the conservation level of GMAS exons and flanking introns, we separated the exons into coding and non-coding types. For each GMAS exon, we picked a random human control exon that satisfies 3 requirements: (1) being an AS exon, (2) harboring at least one SNV in the flanking intronic regions (within 200nt from exon-intron boundaries), (3) being the same type (coding or non-coding) as the GMAS exon. For both GMAS and control exons, we calculated the PhastCons scores extracted from the UCSC database (46-way PhastCons) (Siepel et al. 2005).

Percent sequence identity of GMAS exons

We calculated the percent sequence identity of GMAS exons by parsing the 46-way mutilz

alignments downloaded from the UCSC Genome Browser (Kent et al. 2002). We compared the human GMAS exons to eight other commonly used organisms, namely chimpanzee, rhesus, mouse, rat, opossum, platypus, chicken, and frog, spanning 350 million years in evolutionary history. As controls, we randomly selected control exons in a similar way as described in the PhastCons analysis. For each exon, we extracted pair-wise sequence alignment of human vs. another species from the 46-way alignment and calculated the percentage of identical nucleotides relative to the entire length of the exon.

Calculation of d_N/d_S of coding exons

For coding GMAS exons, we obtained the pair-wise alignments of human vs. one of the other species (chimpanzee, rhesus macaque or mouse) based on multiz alignments and calculated the d_N, d_S, and d_N/d_S ratios using the yn00 program in the PAML package (version 4.8) (Yang 2007). Coding control exons were chosen randomly as alternatively spliced exons with similar percent spliced-in (PSI) values as GMAS exons. PSI was calculated using human RNA-Seq data of brain, heart or liver tissues as described in Barbosa-Morais et al. (Barbosa-Morais et al. 2012) The PSI of a control exon is required to be within 10% of that of the GMAS exon. A total of 1000 sets of control exons were randomly generated in this way with the d_N/d_S calculation repeated for all sets. This analysis was carried out for the three tissues separately.

Tajima's D, F_{ST}, and iHS of GMAS SNVs

Tajima's D of GMAS SNVs for CEU (Utah residents (CEPH) with northern and western European ancestry), FIN (Finnish in Finland), GBR (British in England and Scotland), and TSI (Toscani in Italia) populations were obtained from the dbPSHP database (Li et al. 2014). As controls, we randomly sampled the same number of control SNVs from the SNVs included in the GMAS analysis (i.e., those with adequate read coverage, located within or next to alternatively spliced exons, and not residing in genes with overall allele-specific expression). We compared the Tajima's D distributions of GMAS SNVs and controls using the Kolmogorov–Smirnov test. For the GMAS and controls SNVs, we also calculated the Weir & Cockerham's F_{ST} (fixation index) values (Weir and Cockerham 1984) between the CEU and YRI (Yoruba in Ibadan, Nigera) populations and the integrated haplotype score (iHS) (Voight et al. 2006) for the CEU population based on the HapMap phase 2 data.

GMAS SNVs relative to GWAS SNPs

The GWAS catalog and the LD information of the CEPH (Utah residents with ancestry from northern and western Europe) population were obtained from the NHGRI GWAS page at http://www.genome.gov/gwastudies/ on Dec. 17, 2014, and the International HapMap Project (The International HapMap Consortium 2003), respectively. To determine whether a GMAS SNV is in LD with a GWAS SNP, we required both variants to be covered by a single LD block which passed the thresholds $D' > 0.9$ and $r^2 > 0.8$, and additionally, the distance between the GMAS SNV and GWAS SNP being less than 200kb.

To identify GWAS traits whose SNPs are more often in LD with GMAS SNVs than expected, we randomly sampled 1000 sets of control SNVs from all testable SNVs for the GMAS analysis

(i.e., those with adequate read coverage, located within or next to alternatively spliced exons, and not residing in ASE genes). The number of control SNVs (denoted as X) in LD with GWAS SNPs for each trait was determined, which was compared to that of the GMAS SNVs (denoted as x). An empirical P-value was calculated as $P = \Pr(X \geq x)$.

Gene ontology (GO) analysis

We used our previous approach (Lee et al. 2011) to identify enriched GO terms in our GMAS gene list. Briefly, for each target gene in our list, we resampled a control from all the genes with at least one AS event whose transcript length and GC content differences are within 5% of those of the target. We then calculated the frequency of occurrence of each GO term in the control set. This process was repeated 10,000 times to generate an expected frequency distribution for each GO term, and an empirical P-value was calculated as described above for the GWAS trait analysis. The significance cutoff for choosing enriched GO terms was 1/Total GO terms considered.

Nucleotide sequence translation and protein domain search

We used EMBOSS Transeq (Rice et al. 2000; Goujon et al. 2010) to translate the identified GMAS exons in all six possible reading frames, and the translated protein sequences were analyzed to detect significant protein domains using Pfam (Finn et al. 2006). For each GMAS exon, we randomly sampled an AS exon as control requiring its exon length and GC content to be within 10% of those of the GMAS exon. We compared the frequencies of each protein domain hits among the GMAS exons to the frequencies among the controls. To calculate an empirical P-value, we repeated the random control selection 10,000 times and calculated the P-value as $\Pr(X \geq x)$ where X are x represent the protein domain frequencies in the controls and GMAS exons respectively. The P-values were corrected using the Bonferroni approach. The significant protein domains were defined as those encoded by at least two non-overlapping GMAS exons and with a corrected P-value < 0.05.

Analysis of intrinsically disordered regions (IDRs)

We used the IUPred tool (Dosztányi et al. 2005) to determine the overlap of GMAS exons with IDRs. Disordered score for each amino acid was calculated with a range from 0 to 1, with 1 representing the most disordered. An IDR is normally defined as a region with a score > 0.5. We calculated an average disordered score for each GMAS exon and defined it as being in an IDR if the average score was > 0.5. The control exons were chosen by the same criteria as described in the protein domain analysis, and an empirical P-value was calculated similarly.

Identification of GMAS-associated RBPs

Position weight matrices (PWMs) of known or predicted RBP binding motifs were obtained from RBPDB (Cook et al. 2011) and cisBP-RNA (Ray et al. 2013) databases. The PWMs were used to score the 7-mers overlapping GMAS SNVs. For each 7-mer, two versions of its sequences were created, each harboring the reference or the variant allele of the SNV. The motif score for each version was calculated using the PWM of each RBP. The difference between

motif scores of the alternative alleles was then determined for each RBP. As controls, we sampled 1,000 7-mers each overlapping a random intronic SNV that was at least 5,000 base pairs away from known exons. To decide thresholds for significant binding scores and score differences, we generated the distributions of binding scores and score difference for each RBP using the control SNVs. The significance thresholds were set to be the 95-percentile of each distribution. For each pair of sequences overlapping a GMAS SNV, we determined its associated RBP(s) by requiring: (1) the RBP binding score of either reference or variant allele exceeded the significance threshold of binding scores of the RBP (or 1, if this threshold is less than 1), and (2) the RBP score difference of the alternative alleles exceeded the significant score difference threshold (or 1, if this threshold is less than 1).

Conservation levels of GMAS-associated splicing factors (SFs)

We collected a list of SFs from previous publication (Han et al. 2013). We used the Protein Residue Conservation Prediction (PRCP) method (Capra and Singh 2007) to score the conservation level of the SFs of interest. Specifically, we generated the multiple sequence alignments (in amino acid sequences) of SFs via the Ensembl-Compara Perl API (Flicek et al. 2014). The multiple sequence alignments were analyzed by PRCP to calculate position-specific amino acid sequence conservation scores of the entire protein. The conservation scores of the RNA binding domains of each SF were parsed from the PRCP output based on the RNA binding domain sequences obtained from UniProt (The UniProt Consortium 2014).

Calculation of information content (IC) of SF binding motifs

The IC of SF binding motifs were calculated from the PWMs provided by RBPDB and cisBP-RNA using the seqLogo package (version 1.34.0) (Bembom 2015). The overall IC of a sequence motif was calculated as the averaged IC value of all nucleotides of the motif. Overall IC was calculated for each of the top 15 GMAS-associated SFs (affecting > 5 GMAS events) and for non-GMAS-associated SFs (as controls). Next, the IC values of nucleotide positions overlapping GMAS SNVs were extracted respectively. These values were compared to those of random nucleotide positions within the associated sequence motifs. To account for differences in overall motif consensus strength of different SFs, we normalized the individual-nucleotide IC value by the average IC of all positions of the corresponding sequence motif.

***SRSF1* RNA-seq analysis**

FASTQ files of human *SRSF1* knockdown (KD) and control RNA-seq (two replicates for each sample) were downloaded from the ENCODE database (www.encodeproject.org). We mapped the reads by RASER using the default parameters of the "obviously best" mapping option (Ahn and Xiao 2015). We first removed duplicated reads in each replicate of the samples, and the replicates of the same sample were combined for further analysis. We then calculated PSI values of *SRSF1*-regulated GMAS exons in both the KD and control samples (Katz et al. 2010).

***SRSF1* eCLIP-seq analysis**

Human *SRSF1* eCLIP-seq alignment files (bam format) and eCLIP-seq peak files (BED format) were downloaded from the ENCODE database (www.encodeproject.org). First, we searched for dbSNPs in the eCLIP reads (after removal of duplicate reads and low-quality reads) and retained those SNPs that had two alleles present in the reads, with each allele associated with at least 1 or 2 reads. For these SNPs, we further required a total read counts of at least 20 to have reasonable statistical power in detecting allele-specific bias (Li et al. 2012). Allelic bias of these SNPs in the eCLIP reads was tested similarly as in our previous study (Li et al. 2012). In addition, we calculated the distance of eCLIP peaks to the GMAS exons by BEDTools (Quinlan and Hall 2010) "closest" function using the peak BED files as input.

Minigene constructs

Genomic regions encompassing the candidate GMAS exon and ~450nt upstream and downstream flanking introns were amplified using genomic DNA from the corresponding cell line where the iGMAS exon was predicted. Primers used in this study are listed in Supplemental Table 3. After double digestion by HindIII and SacII or EcoRI and SacII, the DNA fragment was sub-cloned into a splicing reporter (Fig. 2C) (Wang et al. 2006; Xiao et al. 2009). Final constructs were sequenced to ensure that a pair of plasmids containing the two alternative alleles of the SNV was obtained.

Transfection, RNA extraction, Reverse transcription, and PCR

Twelve-well HeLa cells were transfected with 2 μg Minigene plasmid when cells are 90% confluence using Lipofectamine 2000 (Life Technologies). Cells were harvested 24 h post transfection and total RNA was subsequently extracted using the TRIzol method (Life Technologies). cDNA was made from 5 μg of total RNA by reverse transcription and one twentieth of cDNA was used as template to amplify both inclusion and skipping forms, if exist, of the candidate iGMAS exon by PCR within 25 cycles (Supplemental Table 3).

Gel electrophoresis and quantification

Five microliter of PCR product was loaded onto 5% polyacrylamide gel and electrophoresis at 70 volt for one and a half hours. The gel was then stained with SYBR® Safe DNA Gel Stain (Life Technologies) for half an hour before imaging via Syngene SYBRsafe program (Syngene). Expression levels of spliced isoforms were estimated using the ImageJ software (http://imagej.nih.gov/ij/). Inclusion rate (% inclusion) of the target exon was calculated as the intensity ratio of upper/(upper+lower) bands.

Purification of recombinant human SRSF1 protein and FF-DD mutant

The wild-type human SRSF1 sequence was amplified from human embryonic stem cell H1 cDNA, and subcloned into the pET28b bacterial expression vector. The SRSF1 FF-DD RNA binding mutant (Cho et al. 2011) was generated using the overlapping extension PCR method. Sequences of the expression constructs were confirmed by Sanger sequencing. PCR primers are listed in Supplemental Table 5.

Wild-type and FF-DD mutant SRSF1 bacterial overexpression constructs were transformed into BL21 (DE3) Gold competent cells (Agilent). Protein induction was carried out via 1mM IPTG treatment in 100mL cultured cells (O.D = 0.8) for 4hr at 215 rpm at 28°C. Next, cultured cells were centrifuged at 7000 × g for 5 min at 4°C and the pellets were resuspended with ice-cold 10mL lysis buffer (2 × PBS, 5mM DTT, 10% glycerol, 0.1 × protease inhibitor cocktail, 100µg/mL lysozyme, 100U DNAse I, 0.1% IGEPAL CA-630). After 30 min incubation on ice, the lysate was disrupted using three times sonication at 30% amplitude for 20sec with 1sec pulse. Subsequently the lysates was centrifuged at 15,000 × g for 30 min at 4°C. The supernatant was collected and filtered using 0.45µm syringe filter. The sample was loaded into cOmplete His-Tag purification column (Roche) and washed with 20 mL buffer A (2 × PBS, 5mM DTT, 10% glycerol, 0.1 × protease inhibitor cocktail, 0.1% IGEPAL CA-630). The sample was eluted with 500mM imidazole in buffer A. Purify of recombinant wild type SRSF1 and FF-DD mutant was checked by Coomassie staining and western blot using anti-HIS antibody (Santa Cruz Biotech, cat # sc-8036, 1:500 dilution). Clean fractions (E3 and E4, Supplemental Fig. 7A) were combined (~3mL). Salt and small size of non-specific proteins were removed by 20K Slide-A-Lyzer dialysis cassette with 1L Buffer A in the cold room overnight. Protein concentration was measured by Qubit protein assay kit and Qubit 2.0 fluorometer (ThermoFisher Scientific).

In vitro transcription of *SRSF1* target RNA

We selected 4 target RNAs (in genes *TACC2*, *HSPD1*, *RMND5B*, and *PIK3C3*) harboring GMAS SNVs that overlap *SRSF1* binding motif. 100uM of sense and antisense oligos including T7 promoter (Supplemental Table 5) were annealed with oligo annealing buffer (10mM TRIS-HCl pH 8.0, 1mM EDTA pH 8.0, 100mM NaCl) at 95°C for 5 min in a heat block then cooled slowly to 28°C for 2hr. Annealed oligos were purified using 6% PAGE gel and concentrated using Zymoclean Gel DNA recovery kit (Zymo Research). In vitro transcription was performed using 1µg of annealed oligos and HiScribe T7 high yield RNA synthesis kit. In vitro synthesized RNAs was treated with 10U RNAse-free DNAse I (ThermoFisher Scientific) at room temperature for 15 min, then purified by RNA clean & concentrator-5 Kit (Zymo Research). Next, RNA samples were treated with 10U shrimp alkaline phosphatase (NEB) at 37°C for 30 min and then labeled with 0.3µl of gamma 32P-ATP (7000Ci/mmol, MP Biomedicals) using 10U T4 polynucleotide kinase (NEB). Subsequently RNA probes were purified by 6% Urea PAGE extraction and RNA clean & concentrator-5 Kit. RNA concentration was measured by Qubit 2.0 fluorometer (ThermoFisher Scientific).

Electrophoretic Mobility Shift Assay (EMSA)

The purified RNA probes (20pmol) and recombinant SRSF1 protein (0, 0.37, 0.75, 1.5, and 3.0µM) or the FF-DD mutant (0, 0.37, 0.75, 1.5, and 3.0µM) were incubated in 15µl of buffer A (2 × PBS, 5mM DTT, 10% glycerol, 0.1 × protease inhibitor cocktail, 10U RNAse inhibitor, 0.1% IGEPAL CA-630) at 28°C for 30 min, then loaded onto 6% TBE-PAGE including 10% glycerol run at 75V for 1.5hr. The gel was processed without drying, covered with clear folder and exposed to X-ray film at -80°C.

Supplemental References

Ahn J, Xiao X. 2015. RASER: Reads Aligner for SNPs and Editing sites of RNA. *Bioinformatics* btv505–.

Bahn JH, Lee J-H, Li G, Greer C, Peng G, Xiao X. 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22**: 142–50.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**: 1587–93.

Bembom O. 2015. seqLogo: Sequence logos for DNA sequence alignments. *Bioconductor*. http://www.bioconductor.org/packages/release/bioc/html/seqLogo.html (Accessed June 17, 2015).

Capra JA, Singh M. 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**: 1875–82.

Cho S, Hoang A, Sinha R, Zhong X-Y, Fu X-D, Krainer AR, Ghosh G. 2011. Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proc Natl Acad Sci U S A* **108**: 8233–8.

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. 2011. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* **39**: D301–8.

Dosztányi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**: 3433–4.

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al. 2006. Pfam: clans, web tools and services. *Nucleic Acids Res* **34**: D247–51.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res* **42**: D749–55.

Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R. 2010. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res* **38**: W695–9.

Han H, Irimia M, Ross PJ, Sung H-K, Alipanahi B, David L, Golipour A, Gabut M, Michael IP, Nachman EN, et al. 2013. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**: 241–5.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**: 1009–15.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–64.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler a. D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Lee J-H, Gao C, Peng G, Greer C, Ren S, Wang Y, Xiao X. 2011. Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. *Circ Res* **109**:

1332–41.

Li G, Bahn JH, Lee J-H, Peng G, Chen Z, Nelson SF, Xiao X. 2012. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res* **40**: e104.

Li MJ, Wang LY, Xia Z, Wong MP, Sham PC, Wang J. 2014. dbPSHP: a database of recent positive selection across human populations. *Nucleic Acids Res* **42**: D910–6.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–2.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**: 172–7.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308–11.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–50.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426**: 789–796.

The UniProt Consortium. 2014. UniProt: a hub for protein information. *Nucleic Acids Res* **43**: D204–212.

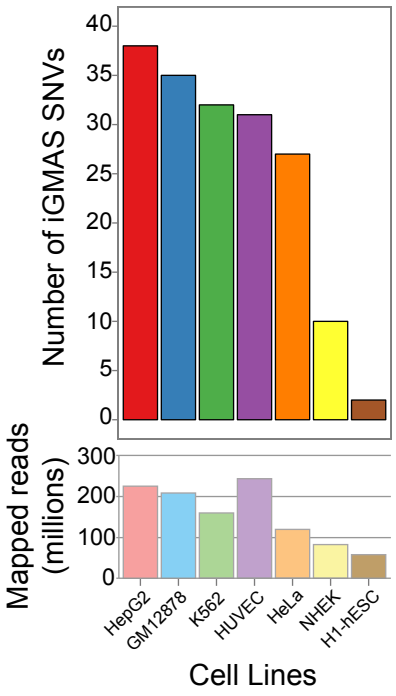
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. ed. L. Hurst. *PLoS Biol* **4**: e72.

Wang Z, Xiao X, Van Nostrand E, Burge CB. 2006. General and specific functions of exonic splicing silencers in splicing control. *Mol Cell* **23**: 61–70.

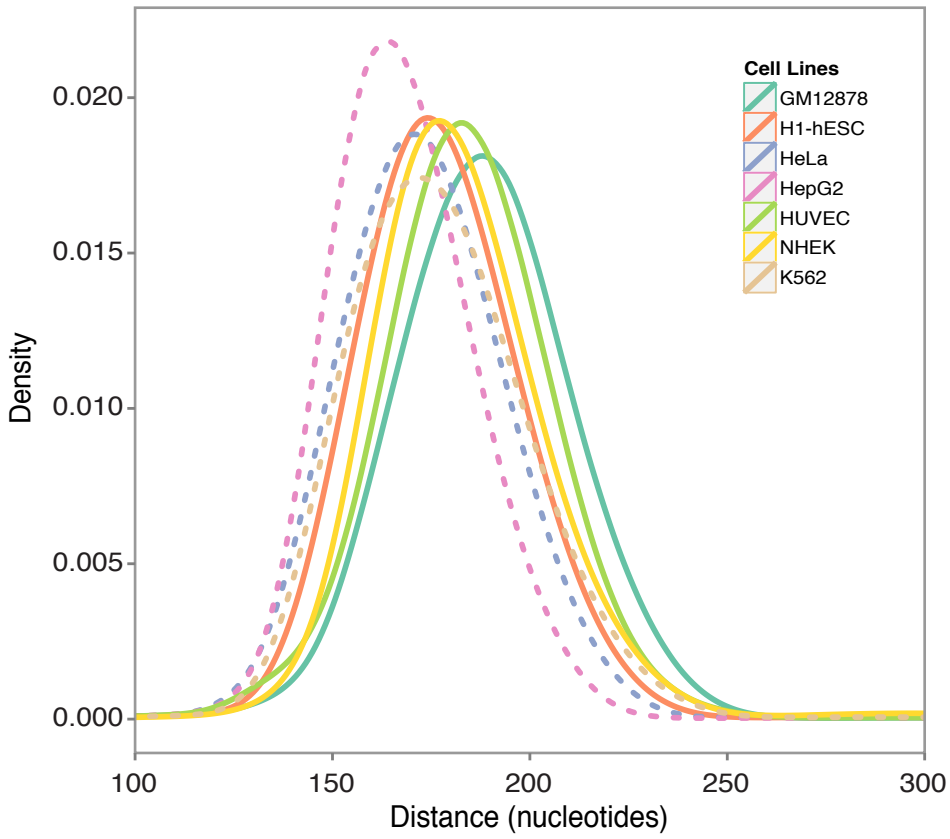
Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution (N Y)* **38**: 1358–1370.

Xiao X, Wang Z, Jang M, Nutiu R, Wang ET, Burge CB. 2009. Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat Struct Mol Biol* **16**: 1094–100.

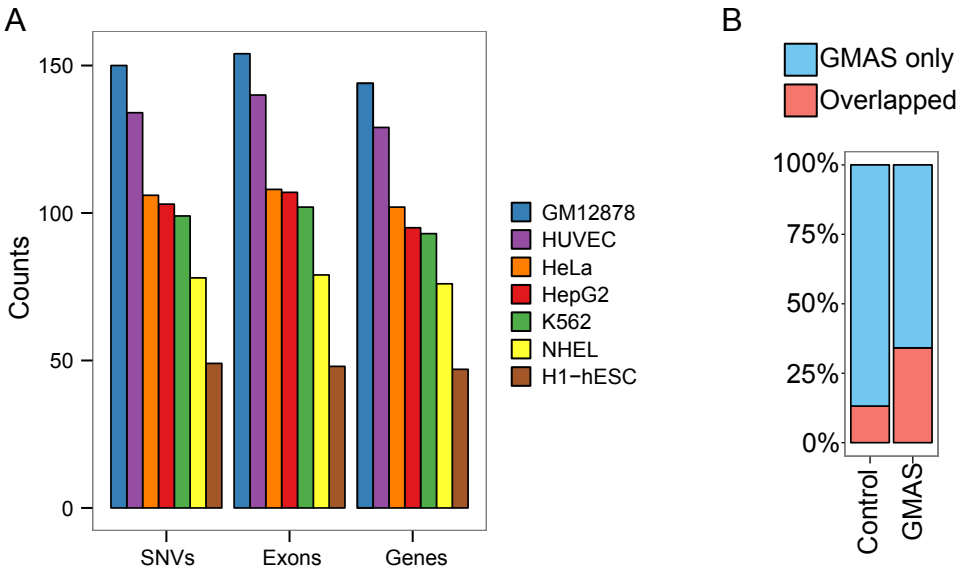
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–91.



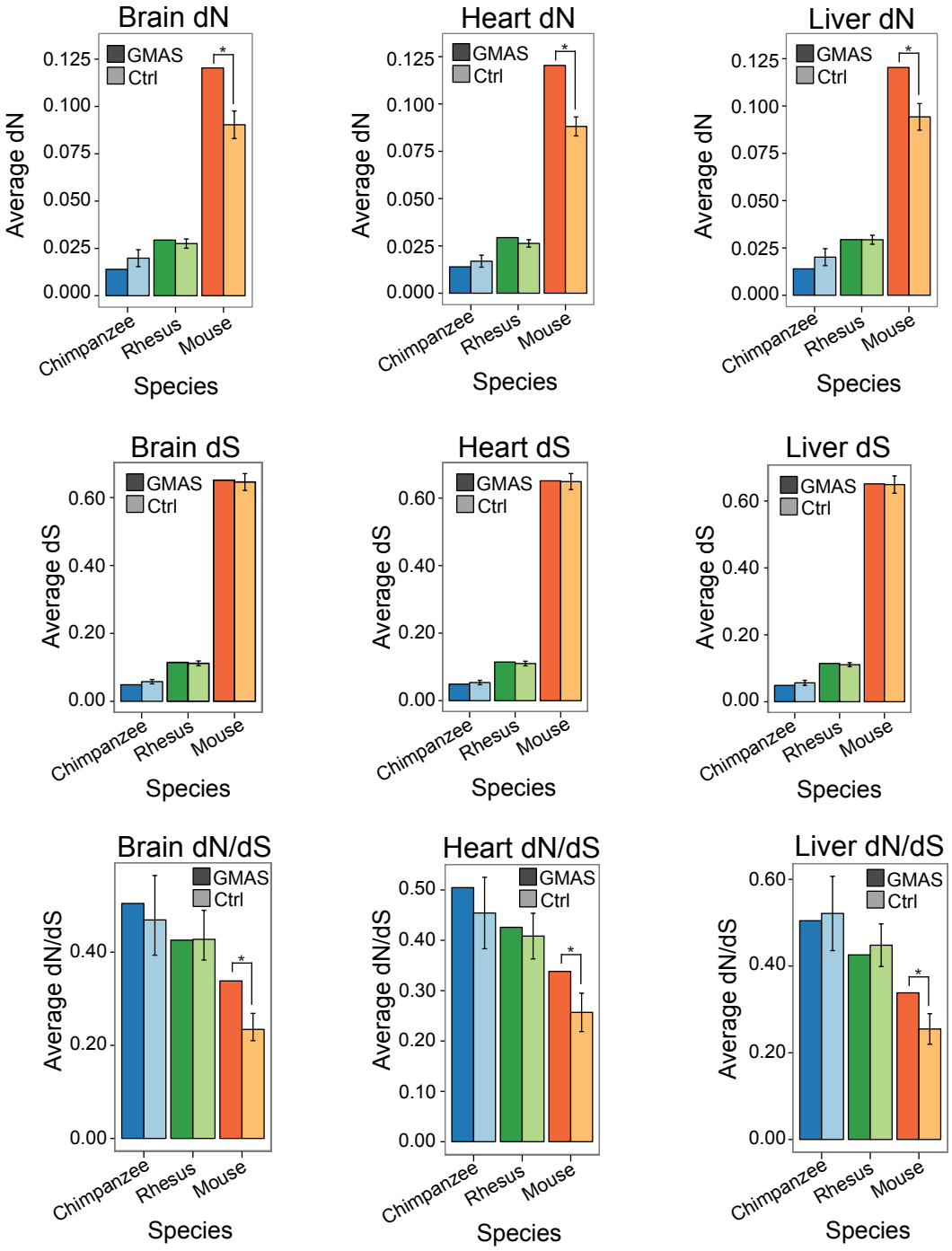
Supplemental Figure 1. iGMAS identified in seven ENCODE cell lines. The top panel shows the number of iGMAS SNVs identified in each of the seven ENCODE cell lines. The bottom panel shows the number of uniquely mapped reads in the NA- RNA-Seq data for the corresponding cell lines. The number of iGMAS events detected is generally dependent on the uniquely mapped read depth of a cell line, as expected.



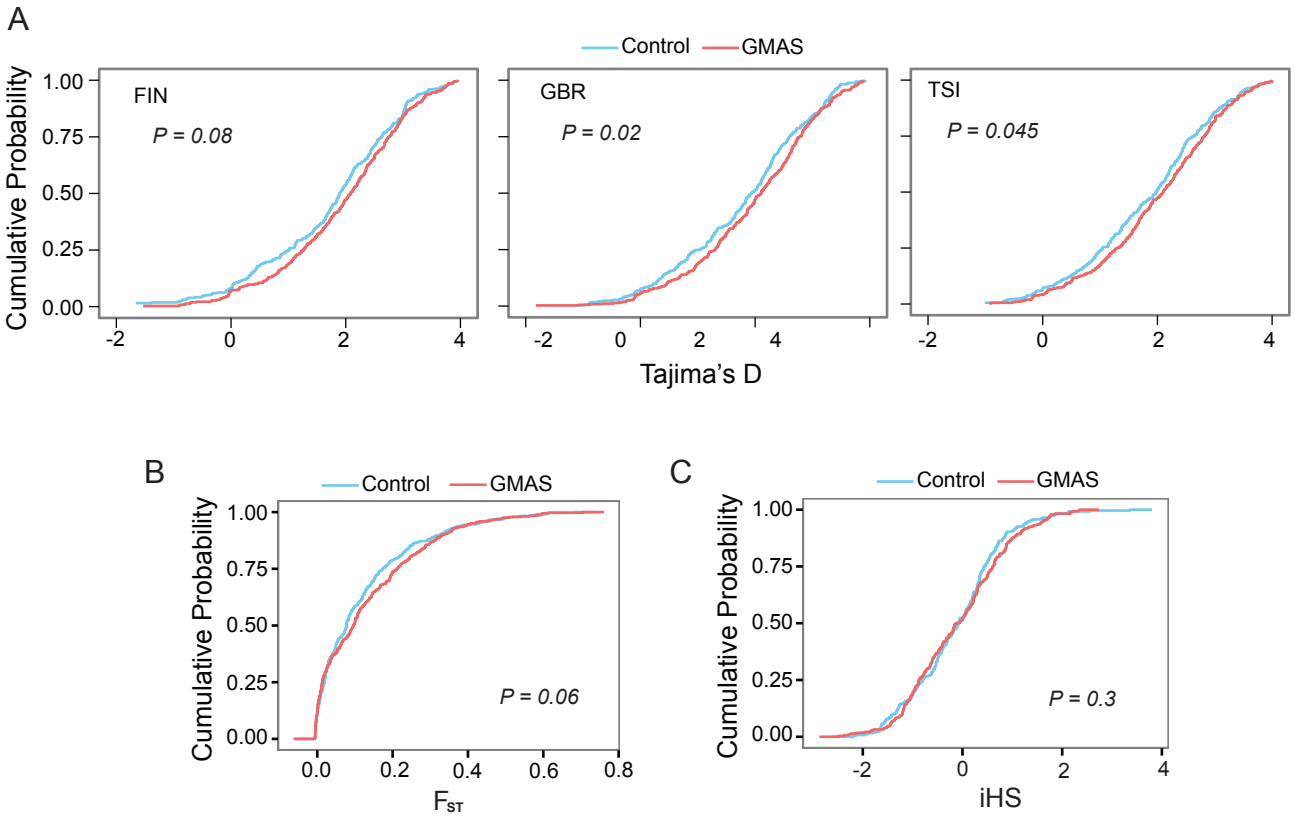
Supplemental Figure 2. Insert size of ENCODE NA- RNA-Seq (paired-end). Following read mapping, we calculated the insert size distributions of all paired-end reads from the ENCODE NA- RNA-Seq datasets. GM12878 has the largest insert size distribution, which is likely to give more power to detect the iGMAS SNVs than other cell lines.



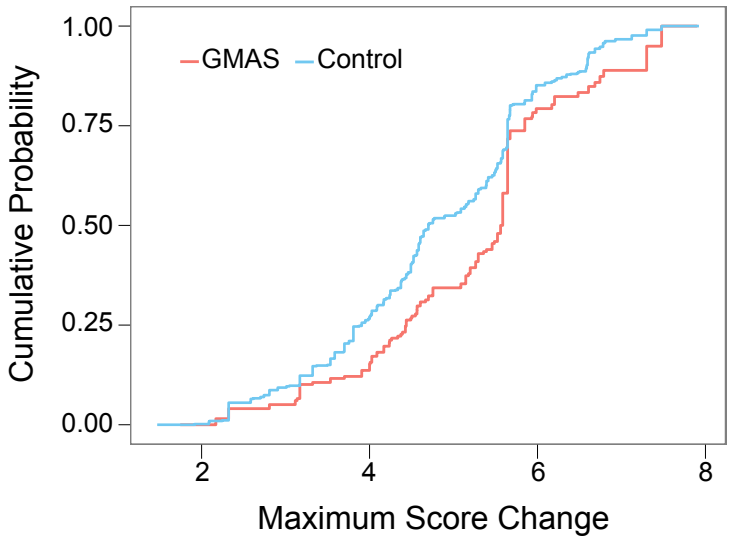
Supplemental Figure 3. GMAS events identified in seven ENCODE cell lines. (A) Number of GMAS SNVs, exons and genes identified in each cell line. Events shown include both iGMAS and eGMAS results. (B) Overlap between GMAS SNVs and sQTL results from prior studies (see Text). The same number of control SNVs were randomly sampled from all testable SNVs for GMAS analysis (i.e., those with adequate read coverage, being next to or within an alternatively spliced exon but not in an ASE gene). We compared the fraction of GMAS-sQTL overlap and the fraction of control-sQTL overlap and found that many more GMAS SNVs overlap with the previously known sQTLs than the controls (Chi-square test, $P < 0.0001$).



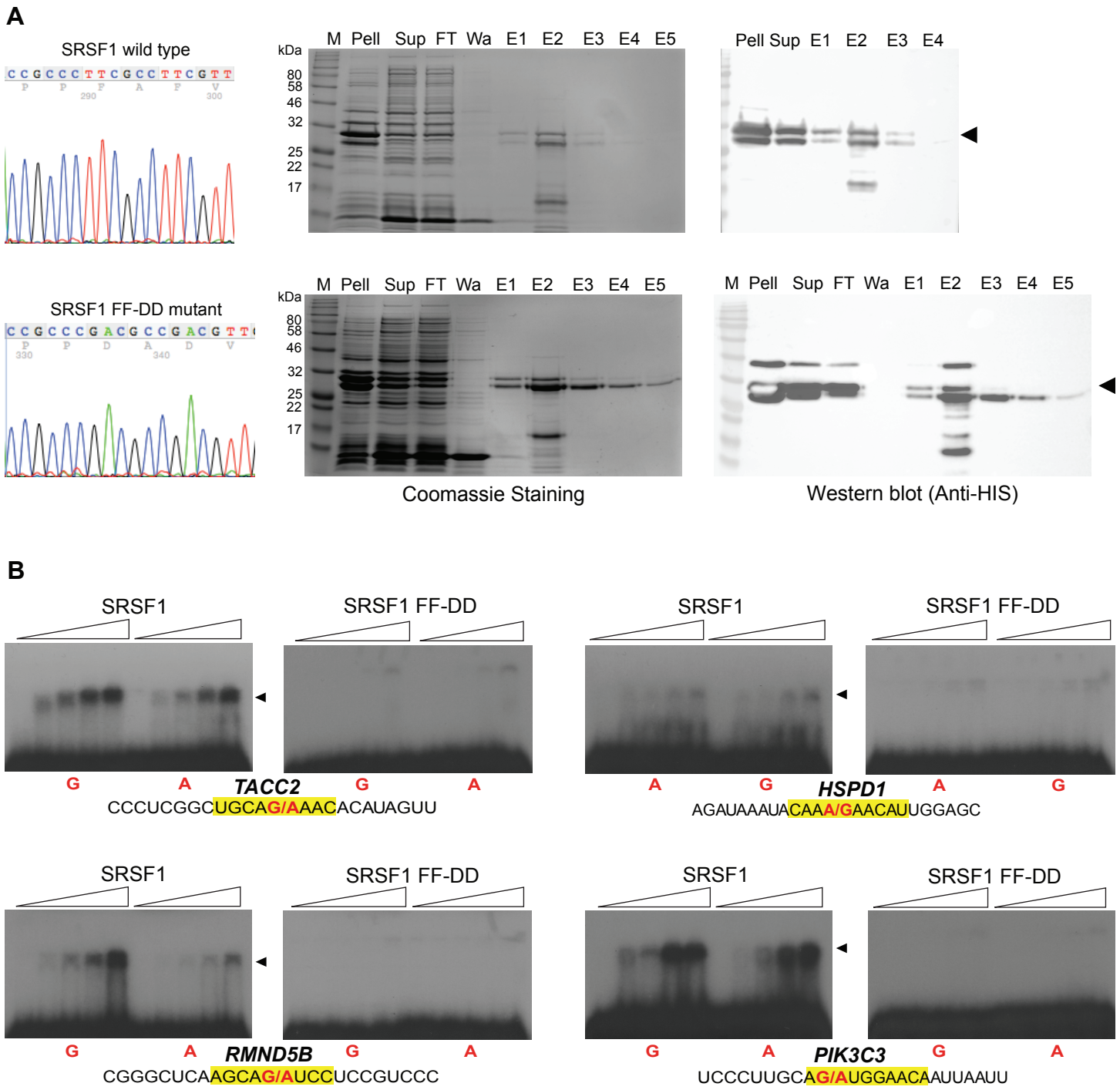
Supplemental Figure 4. dN, dS, dN/dS values of GMAS exons. The dN, dS, and dN/dS ratios were calculated using the yn00 module in the PAML package and the pair-wise alignments of human vs. chimpanzee (blue), rhesus macaque (green), or mouse (orange) (SI Materials and Methods). The dark blue, dark green, and dark orange bars correpond to GMAS coding exons, and the lighter colors correspond to the control sets. Control exons were randomly sampled from alternatively spliced coding exons with < 10% PSI difference from the corresponding GMAS exons. The PSI values were calculated from previously published RNA-Seq data of human tissues (SI Materials and Methods). We randomly sampled 1000 sets of control exons and calculated the average dN, dS, and dN/dS and SD (error bars) among the controls. Asterisks mark the statistically significant comparisons between GMAS coding exons and the 1000 random control sets (*P < 0.05, based on the empirical distribution of 1000 control sets). The dN/dS of GMAS coding exons between human and mouse are consistently higher than that of the controls with similar PSI in different tissues. The dN/dS difference is mainly due to the difference in dN instead of dS.



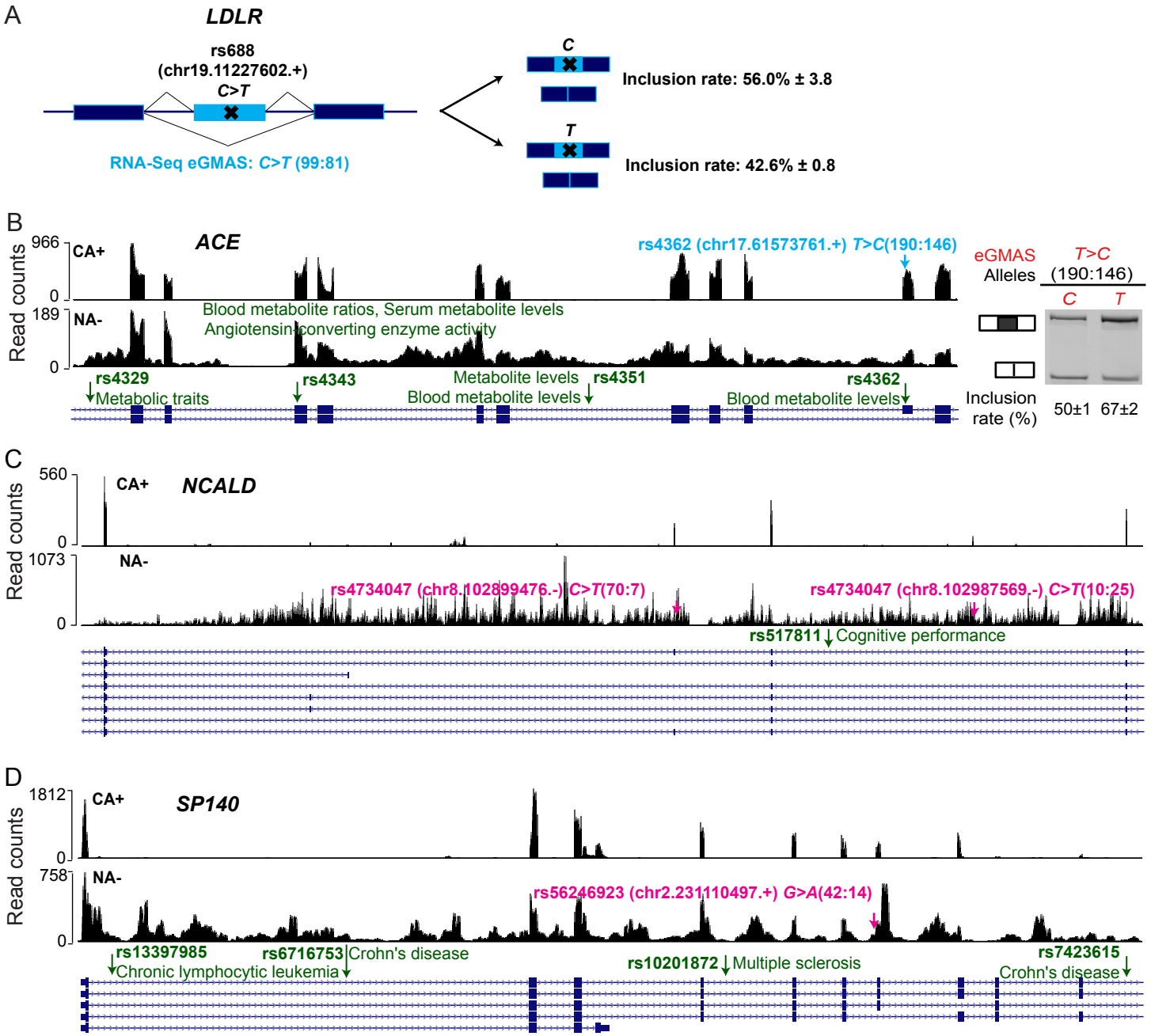
Supplemental Figure 5. Population genetic analysis of GMAS SNVs. (A) Tajima's D values of GMAS and control SNVs. Control SNVs were randomly sampled from all testable SNVs in GMAS analysis (i.e., those with adequate read coverage, reside next to or within an alternatively spliced exon, but not in an ASE gene). Results for 3 populations are shown (FIN, GBR and TSI), similarly as in Figure 4E (CEU population). *P* values were obtained via the Kolmogorov–Smirnov test. (B) Similar as (A), F_{ST} values of GMAS and control SNVs. (C) Similar as (A), iHS values of GMAS and control SNVs.



Supplemental Figure 6. GMAS-associated RBP binding score change. The maximum RBP binding score change (absolute value) for all possible 7-mer sequences due to the single nucleotide change at a GMAS SNV site is shown (red, SI Materials and Methods). Control SNVs were chosen from random intronic SNVs that are at least 5000nt away from an exon. Since these intronic SNVs are very far away from any exons, they are less likely involved in splicing regulation. The maximum binding score change was calculated for control SNVs in the same way as for GMAS SNVs. GMAS SNV-induced binding score changes are signficantly larger than those of the controls (Kolmogorov–Smirnov test, P = 1.3e-05).



Supplemental Figure 7. Recombinant SRSF1 bacterial overexpression and electrophoretic mobilty shift assay (EMSA). (A) Baterial overexpression of human SRSF1 and its RNA binding mutant FF-DD. Wild type SRSF1 and FF-DD mutant cDNAs were cloned into pET28b, confirmed by Sanger sequencing (left panel). Both proteins were purified from BL21 (DE3) GOLD using cOmpete His-Tag purification column. Detailed purification conditions are described in Supplemental Methods. Fractions from each purfication step were loaded onto 12% SDS-PAGE. (Pell: pellet; Sup: supernatant; FT: flow through; Wa: washing; E1-5: elution1-5). Purified SRSF1 and FF-DD mutant proteins were confirmed by Coomassie staining (middle pannel) and western blot (right pannel). Extra bands of smaller protein sizes are likely due to degradation. E3 and E4 fractions were used, followed by 20K dialysis. (B) EMSA using recombinant SRSF1 and FF-DD mutant and synthetic RNA fragments harboring alternative alleles of GMAS-SNVs. The specific alleles in each target and the sequences of the synthetic RNA fragments are shown below each gel image, where the SRSF1 sequence motif is highlighted in yellow and the two alleles of GMAS SNVs are written in red. RNA targets were *in vitro* transcibed and ^{32}P-labelled; 200ng of input RNA was incubated for 30 min at RT with 0, 0.37, 0.75, 1.5, and 3.0 uM of recombinant SRSF1 and FF-DD mutant proteins. RNA-protein complex was indicated by arrow heads.



Supplemental Figure 8. Examples of GMAS SNPs in LD with GWAS SNPs. (A) SNP (rs688) in *LDLR* was previously reported to affect splicing. Based on Zhu *et al*. (Hum Mol Genet 16(14):1765–72), the *C* allele is associated with higher inclusion level, which is consistent with eGMAS results. (B-D) Examples of validated GMAS events in our study that are in LD with GWAS SNPs (required LD block with D' > 0.9 and r^2 > 0.8, and the distance between the GMAS and GWAS SNPs is < 200kb). The locations of the GWAS SNPs are marked by the green arrows above annoted intron-exon strucutre (bottom), shown together with the SNP IDs and GWAS traits. The CA+ and NA- read distributions are shown as illustrated. Light blue arrows: eGMAS SNPs; Pink arrows: iGMAS SNPs. iGMAS events were identified in NA- data, whereas eGMAS events were identified in CA+ data. The GMAS SNP IDs, genomic coordinates, alleles and associated reads in CA+ or NA- data sets are also shown. Experimental validation results for *ACE* are shown, similarly as in Figure 2C. Validation results for *NCALD* and *SP140* are included in Figure 2C.