A Mammalian Genomic Signature Shaped by Single Nucleotide Variants Controlling Transcriptome Integrity and Diversity

Jian Yang[1,*], Samuel Ogunsola[2,*], Jason Wong[3,*], Aydan Wang[2], Roby Joehanes[4,5]
Daniel Levy[4,5], Shalini Sharma[3,†], Chunyu Liu[1,4 †], Jiuyong Xie[2, †]

[1]Department of Biostatistics, Boston University School of Public Health, Boston, MA, 02118, USA

[2]Department of Physiology & Pathophysiology, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, R3E 0J9, Canada

[3]Department of Basic Medical Sciences, College of Medicine-Phoenix, University of Arizona, Phoenix, AZ, USA.

[4] Framingham Heart Study, Framingham, MA 01702, USA

[5]Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892, USA

Keywords: Non-coding regions, REPAG, GWAS, cryptic splicing, 2nd transesterification step

The authors declare no potential conflicts of interest.

†: Correspondence to: shalinijs@arizona.edu, liuc@bu.edu, and xiej@umanitoba.ca, phone: +1 204 975-7774

Abstract

Many functional features of mammalian genomic sequences remain poorly understood. Here, we identify a widely evolved genomic signature, G-tract-AG motifs consisting of guanine tracts closely upstream of AG dinucleotides, which is significantly associated with single-nucleotide variants (SNVs) identified in genome-wide association studies (GWAS), particularly within non-coding regions (NCRs). Approximately 9,000 such G-tracts within human genes are disrupted by variants of the *cis*-splicing quantitative trait loci (sQTLs) in the Genotype-Tissue Expression (GTEx) project. Functionally, G-tracts repress splicing at the adjacent 3′AG, primarily by stalling the second transesterification step. Disruption of the G-tracts by SNVs relieves this repression, enabling splicing and generating novel transcript isoforms. These G-tract-disrupting SNVs are in *cis* across the majority of protein-coding genes and are among thousands of rare variants causing genetic diseases. Our findings provide mechanistic insights into the maintenance of transcriptome integrity while still allowing diversity by a mammalian-evolved genomic signature, particularly for the NCR SNVs in association with diverse traits and a new framework for their functional annotation.

Introduction

Many sequence features of mammalian genomes, predominantly composed of non-coding regions (NCRs), remain largely unknown (1). While cryptically spliced products from the NCR pose challenges to transcript integrity (2), novel mRNA isoforms driven by genetic variations could contribute to transcript diversity in evolution (3,4). It remains challenging, however, to identify the functional sequence motifs and single nucleotide variants (SNVs) in the NCR, particularly those reported in genome-wide association studies (GWAS) (5).

Several sequence motifs have been identified to prevent cryptic splicing of endogenous transcripts, such as pyrimidine-, UG- or CA-rich motifs (2,6-10). Lengthening uridine tracts also allow the emergence of new exons in primate evolution (4). Distinct from these motifs, G-tracts are made of $(G)_{\geq 3}$ purines. They are abundant in mammalian introns often as splicing enhancers (11-13), but not typically anticipated to be within the consensus sequence of the 3′ splice site (SS), $(Y)_{20}$NYAG in humans (Y: C or T; N: A, C, G or T) (14). When G-tracts do occur inside this region, they repress splicing as regulatory elements between the polypyrimidine tract (Py) and 3′ AG dinucleotides (REPAG) (15-19). REPAGs have been found within annotated 3′SS in eukaryotes from yeast to human and are enriched the most in mammals (18), likely contributing to the extensive use of alternative splicing (17,20,21).

Despite the abundance of G-tracts in mammalian introns (11), only about 2,000 REPAG-harbouring 3′SS have been identified in the human genome (17,18). Interestingly, in addition to annotated 3′SS, cryptic 3′SS in cancer tissues with splicing factor mutations also show REPAG enrichment (18). This observation suggests that additional G-tract-AG signatures may lie deeper within introns or other non-coding regions, where they could serve additional roles.

Here, we searched the human genome and identified millions of these signatures, which are largely in the NCRs conserved in mammals. We show their potent ability to repress or de-repress splicing depending on the SNVs, and highlight a distinct mechanism by which genetic

variants in these regions control splicing in their association with mammalian traits or diseases. These key sequences may also assist in the functional annotation of SNVs in NCRs (22-25).

Results

1. Human G-tract-AG signatures: genome-wide presence, enrichment in GWAS-mapped regions and conservation among mammalian species

To identify genome-wide G-tract-AG signatures, we searched both strands of the human genome based on the criteria applied to the annotated 3´SS (18) (Fig. 1A). Initially, we identified and evaluated 111 millions of AG dinucleotides within splice site-like sequences (3´SSL) that had splicing signals greater than 0.1, the lower boundary of most constitutive exons (S_Fig. 1), according to their MaxENT scores (26). Although numerous such 3´SSLs possess potential splicing signals, currently annotated 3´SS is only around 0.3 million (14). This discrepancy suggests that the majority of these 3´SSLs remain either unused or actively repressed in cells. Consistent with this repression hypothesis, roughly 6 million of these 3′SSLs (about 5.4%) harbor the G-tract-AG signature - an enrichment of 2,630 times over that observed in annotated 3′SS (18). This substantial enrichment underscores the potential functional importance of G-tract-AG signatures within these 3′SSL genomic regions.

To identify G-tract-AG signatures that are close to or disrupted by single nucleotide variants (SNVs), we first filtered their 3′SSL sequences (Fig. 1A, 23nt: 18nt + AG + 3nt, by the MaxENT algorithm(26)) for the presence of any common SNVs from the TOPMed (Trans-Omics for Precision Medicine, minor allele frequency MAF ≥ 0.05) database. This analysis identified approximately 1.8 million unique SNVs, including 41,353 that overlap with the GWAS hits. In particular, these SNVs were significantly enriched in the GWAS loci (32.1%)

compared to random SNVs in the human genome (30%, Fig. 1B, enrichment index, see also Methods). In contrast, there was no significant enrichment of SNVs in the CACA- or TGCATG-3′SSL in GWAS in parallel searches. Moreover, more than 95% of such GWAS SNVs were found in the NCR, mainly introns and intergenic regions (Fig. 1C). Further, within the genomic start-end boundaries of 3,202 GWAS genes, we identified 4,447 SNVs that directly disrupt the G-tracts [$(G)_{3\text{-}11}$] in the sense strand (*in cis*, Fig. 1A, S_Table Ia and S_Fig. 2). Together, these data support a significant association of the G-tract-AG signatures with the GWAS SNVs.

Manual examination of a group of G-tract-AG signatures indicated that most of them were in mammals only (97%, n = 37), based on the MultiZ-alignment sequences of 100 vertebrates from the UCSC Genome Browser (27) (S_Fig. 3). However, they exhibited higher divergence than those in the annotated 3′SS (15,17), for instance, about 25% of them in primates only. In the two representative examples of different levels of conservation (Fig. 1D), the G-tract-AG signature in *MAX* is conserved in 80.6% of the species (n = 62), and the signature in *QSOX1* is conserved in 7.3% only (n = 41). The G-tract disruption occurred in other species as well, by A, C or T in guinea pig and primates, among other changes within the 3′SSLs (see also S_Figs. 4 & 5). By MaxENT scores, the G-tract disruption increased the splicing signal to or close to that of the human SNV-T variants in a few species (*MAX* in guinea pig and *QSOX1* in chimpanzee, gorilla and orangutan). Therefore, the G-tract-AG signatures or the G-tract-disruptive SNVs are conserved in mammals to different extents. In some species, their 3′SSL splicing signal has also evolved to the level in humans.

These data, taken together with the REPAG repression of splicing (15,18), imply a much wider role for the G-tract-AG signature in the NCR of mammalian genomes, beyond its control of alternative splicing of annotated 3′SS (15-17).

2. G-tract repression and SNV-caused de-repression of splicing from the 3′AG

To determine whether the G-tract-disruptive SNVs are globally associated with splicing, we identified 156,380 such SNVs *in cis* within gene start-end boundaries in the human genome, and cross-referenced them with the *cis*-sQTLs (splicing quantitative trait loci, false discovery rate, FDR<0.05, n = 486,725 unique IDs) from the Genotype-Tissue Expression Project (GTEx) v10 (28). Of these, approximately 9,000 SNVs were significant sQTLs (5.62%, S_Table_Ib), representing a 1.58-fold enrichment compared to randomly selected SNVs (n=1,000 sets of 156,380 SNVs, matched by their MAF and transcription start sites (TSS), p = 2.2E-16, one sample t-test). Thus, these G-tract-disruptive SNVs are significantly enriched for sQTLs, supporting their potential role in splicing and impact on the transcriptome.

To validate the potential effects of the G-tracts and SNVs on splicing, we examined the two representative examples in mini-gene splicing reporter assays (Fig. 2 and S_Fig. 5), which we have employed to consistently demonstrate the G-tract repression of splicing in ten REPAG reporters (15,17,19) (S_Table II). The *MAX* SNV rs762810-T is located in the upstream 3′SS of exon 4 (E4), in a mammal-specific but much less conserved region than the E3 region (Fig. 2A). The E4 inclusion is at about 20% of the E3 by their RNA-Seq reads in blood cells, but invisible in the other 53 GTEx tissues including kidney and skin. Moreover, the intron 3 excision ratio is significantly higher in samples with the rs762810-T/T than those with the G/G genotype (S_Fig. 6) (28). The *MAX* G-tract repression and SNV de-repression effects were indeed validated in the splicing reporters (Fig. 2B): in human kidney HEK293T cells, the $(G)_4$-carrying reporter skipped the E4. In contrast, the $(G)_3$T-carrying reporter with a transversion of the last G to the rs762810-T resulted in 66% E4 usage (in molar ratio), and the $(A)_4$ reporter with G to A transitions resulted in 98% E4 usage. Thus, it is the $(G)_4$ that represses but the rs762810-T that de-represses splicing from the 3′ AG. The *QSOX1* rs3767199-T is located in

the first intron, 398nt downstream of exon 1 (S_Fig. 5C). Similar to the rs762810-T, this variant also de-repressed the G-tract repression of splicing from its 3′ AG splice site, albeit to a smaller extent in the reporter assay.

To visualize the SNV genotype effect on splicing in humans, we examined the 3′SS intron/exon ratios of RNA-Seq reads, as described previously (29), in genotype- and RNA-Seq data-matched whole blood cell samples from participants in the Framingham Heart Study (FHS) (50). We found intron/exon junction reads ratios of 1.0, 0.7, and 0.2 for individuals with G/G, G/T, or T/T genotypes of rs762810, respectively (Fig. 2C), indicating no splicing, partial splicing, or near-complete splicing of the 3′SS of the blood cell-specific exon. Together with the splicing reporter assay and GTEx aggregated data (Fig. 2B and S_Fig. 6B), this supports that the G-tract represses splicing from the 3′AG site and the rs762810-T de-represses splicing for E4 usage in human blood cells.

Taken together, these data support that the G-tract-AG signatures have dual functions in the NCR: (1) to repress splicing from the 3′AG with the G-tract as a functional REPAG to protect transcript integrity and (2) to permit splicing for the emergence of novel exon (e.g., E4 in *MAX*) or transcript isoforms in mammals upon disruption of the G-tract by a SNV for transcript diversity.

3. Splicing repression at the two transesterification steps by the REPAG

For the mechanism of splicing repression by the G-tracts, we have shown previously that a REPAG element with its bound hnRNP (heterogeneous nuclear ribonucleoprotein) H1 represses splicing and reduces Py-binding by an early spliceosome factor U2AF65 (15,17). The hnRNP H1 binding is disrupted by mutating the G-tract (15), and by the *MAX* SNV-T variant in structure modelling (S_Fig. 7). However, it is still unclear which of the two

transesterification steps of splicing is repressed by such G-tracts. Our *in vitro* assay of splicing intermediates indicated that both steps were repressed but primarily at the 2nd step.

In the *in vitro* splicing assay, we analyzed the kinetics of lariat formation in a pre-mRNA containing either the REPAG or its mutant as in our previous report (7) (Fig. 3A). Compared to the AdML vector, the presence of REPAG caused a significant delay by about 6 minutes in generating the lariat intermediate (lanes 1-3 vs. 4-6, and Fig. 3B). The delay was reversed upon mutating the REPAG sequence (lanes 7-9). Thus, REPAG delays the generation of the first step product, the lariat intermediate, which is consistent with the reduction of U2AF65 binding to the Py (15).

Interestingly, the second step of splicing was also significantly delayed by REPAG but far more than would be expected from a simple delay of the first-step (Fig. 3B). The half-life of lariat intermediates in conversion to the lariat was approximately 10, 21, and 45min for the AdML, REPAG mutant, and REPAG reporters, respectively (Mean, ±1.2, ±4.4, ±6.5, SEM, n = 6, 3 and 4, Fig. 3C), indicating a significant delay by the REPAG. Additionally, a less pronounced second-step repression was also observed for the REPAG mutant, likely due to the longer spacer between the Py and 3′ AG as reported (30).

Together, the results from the splicing intermediate assays demonstrate that the REPAG unusually represses both transesterification steps with a stronger effect on the 2nd step of splicing.

4. Widespread *in cis* disruption of the G-tracts of the genomic signatures by SNVs

The impact of the genomic G-tract-AG signature is likely widespread on mammalian genes. In the GWAS, the 3,202 genes that contain G-tract-disrupting SNVs *in cis* are strongly enriched for a variety of health-related traits including aging, heart disease, and cancer by

DAVID functional clustering analysis (Fig. 4A, S_Table Ia). About 77% of them are protein-coding genes. In the human genome, the 156,380 G-tract-disrupting SNVs are *in cis* within 23,479 unique genes, including about 71% of the 20,078 protein-coding genes in the hg38 (Fig. 4B, S_Table Ib). These figures together with the sQTL enrichment and functional data on splicing (Figs. 2-3 and S_Fig. 5) imply widespread impact by the G-tract-AG signatures and their SNVs on the transcriptomes and proteomes. For example, the *MAX* E4 usage contributes to two transcripts, resulting in truncated protein isoforms lacking the nuclear localization signal (NLS) or dimerization domain (S_Fig. 8) (27,31-33). Several similarly truncated MAX isoforms showed cell growth-promoting instead of the tumor suppressor effect by the full length protein (32,34-36). Moreover, the E4-promoting T-variant is associated with higher blood cell counts in GWAS (37), enriched among individuals with melanomas and associated with E4 ectopic inclusion in melanoma tumor cells (S_Fig. 9), implying its association with cell growth as well. The usage of the *QSOX1* cryptic splice site caused by the rs3767199-T variant would lead to intron retention and nonsense-mediated mRNA decay reducing full-length protein levels (S_Fig. 5A). Related to this anticipation, reduced protein level in *Qsox1* knockout mice caused lower pulse pressure, an effect also associated with the rs3767199-T variant in GWAS (38,39).

Beyond the genome-wide common SNVs, we also identified 2,145 rare genetic variants (MAF<0.01) that disrupt the G-tracts of the signatures in the human gene mutation database (HGMD) and ClinVar (40,41) (Fig. 4B, S_Fig. 10 and S_Tables III). For instance, the c.392G>T variant of the glucose-6-phosphate dehydrogenase (*G6PD*) gene causes a neutral change of amino acids (G131V) but deficient enzyme activity in hemizygotes (42-44). The discrepancy could be well explained by RNA level changes instead, by aberrant splicing due to T variant-disruption of the G-tract resulting in 44aa deletion at the N-terminal region (Fig.

4C). The variant enzyme will lose the Y112 for $NADP^+$ binding and catalytic activity of the dehydrogenase (45).

Conclusion

Taken together, our data define an evolved mammalian genomic signature (G-tract-AG) enriched for GWAS SNVs and GTEx sQTLs, and demonstrate that G-tract-disruptive SNVs act through a distinct molecular mechanism to control splicing. The G-tract, as a two-step tight splicing repressor, not only regulates alternative splicing of annotated 3′SS as previously reported(15), but also represses the latent AG splice sites from splicing in the introns of the genome. The 3′ AG becomes a splice site upon disruption of the G-tract by SNVs, leading to the generation of novel transcript isoforms. Therefore, it is likely that the widespread G-tract-AG bipartite motifs regulate mammalian transcriptome integrity and diversity depending on the SNV.

Our finding provides not only a functional genomic signature to explain how transcriptome integrity and diversity are controlled in mammalian species by the signatures and their SNVs, but also a valuable tool for annotating similar sequence signatures and SNVs in the genomes. Identifying more of such signatures will help cover a much larger proportion of GWAS SNVs such as those in the CACA-3′SSL (Fig. 1B) or even coding regions (Fig. 1C, and 4D), or in other species (Fig. 1D). It will be particularly helpful for the identification of functional SNVs currently underestimated by prediction algorithms, like the ones in the *MAX* and *QSOX1* that were scored below the confidence level as were 99% of the GWAS SNVs (Fig. 2A, S_Fig. 11 and S_Table IV), or not even scored by SpliceAI (23). Further genome-wide

functional annotation of the signatures will provide a global view of the balanced control between the maintenance of transcriptome integrity and the evolution of transcriptome diversity, as well as genetic differences among individuals or mammalian species. Moreover, detailed mechanistic studies will also provide further insights into the signature's unusual effect on the 2nd step of splicing, as found in other cases (46,47). Together, these studies on the signatures could help bridge the mammalian genome with transcriptome integrity and diversity but also advance personalized medicine through understanding and correcting cryptic splicing caused by SNVs (48).

Materials and Methods

***In vitro* splicing assay**: To the 3′ splice site of the adenovirus splicing reporter (49), we inserted the REPAG gggtgggggagtg upstream of the PRMT5 exon 3 and its mutant aaataaaaaagtg (15). The *in vitro* splicing was carried out in HeLa nuclear extracts as reported (49). The optimal time range for the time course measurements of lariat products was determined by pre-tests to avoid RNA degradation after prolonged incubation *in vitro*.

Human genome search and analysis for G-tract-AG harbouring 3´SSL: GRCh38/hg38 was searched of both strands by identifying $(G)_{\geq 3}$ within 1 to 13nt upstream of AG dinucleotides(18). The 3´SS 23nt (18nt upstream plus 3nt downstream of AG) region was filtered for MaxENT 3´SS strength measurement(26), and for the presence of TOPMed common SNVs (MAF > 0.05)(50) . Briefly, to identify potential 3′SS AG sites, we developed a pipeline to first search for the genomic coordinates of AG dinucleotides on both strands across the human genome. We then output 23nt sequences covering the AG (5′- 18nt + AG + 3nt -3′, e.g. 5′- ttccaaacgaacttttgtaggga-3′), whose likelihood to function as a valid 3′SS was quantified by using MaxEntScan (26).

To map SNVs within the 23nt 3′SSL sequences, we next used the comprehensive TOPMed SNV annotation to retrieve allele and MAF information and to combine the SNVs reported in the GWAS Catalog(5). The TOPMed SNV annotation has been developed from deep sequencing on over 200,000 human genomes, with an average coverage of approximately 30-fold (50). In this study, we selected SNVs with MAF ≥ 0.05 from the TOPMed SNV annotation to identify any SNVs in the 23nt region. To obtain MaxENT scores for the 23nt sequences with SNV alleles, we replaced the reference alleles with the alternative alleles.

To identify G-tract-AG-harbouring 3′SSL, we extracted the 23nt sequences (Fig. 1A) containing $(G)_{\geq 3}$ (G-tract) between the 6th and 18th nucleotides using our in-house pipeline. The consensus sequence motifs were obtained using the online MEME motif analysis (51).

Enrichment analysis of G-tract-AG-harbouring 3′SSL among GWAS loci, GTEx sQTLs or health traits/diseases. Of the hundreds of thousands of SNVs reported in the GWAS Catalog (5,52), some entries are missing allele information or allele frequencies. Therefore, we annotated the GWAS SNVs with the TopMed SNV annotation. We selected SNVs with MAF ≥ 0.05 and *p*-value ≤ 5E-08 from the GWAS Catalog, resulting in 128,991 distinct GWAS SNVs. We categorized the MAFs of GWAS SNVs into five groups and transcription start site (TSS) into four groups, creating 20 matching groups based on these classifications (53).

We next generated reference SNVs by collecting all TopMed SNVs with MAF ≥ 0.05 and assigned them into the 20 matching groups to the identified GWAS SNVs. To minimize bias, we randomly selected 1000 sets of 128,991 SNVs, having approximately the same MAFs (bin boundaries: 0.05, 0.1, 0.2, 0.3, 0.5) and distance to TSS (bin boundaries: 0kb, 1kb, 10kb, 100kb, 1Mb). We calculated the proportion of GWAS SNVs on G-tract-AG-harbouring 3′SSL (n = 29,284) and then the average proportion of the 128,991 SNVs selected by matching on G-tract-AG-harbouring 3′SSL across 1000 sets. The fold-enrichment was calculated as the ratio of the proportion of GWAS SNVs on the 3′SSL to the average proportion of the matched SNVs across the 1000 sets.

GTEx v10 sQTLs of 54 tissues were filtered to obtain 486,725 unique sQTLs, which was used to identify sQTLs among the genome-wide 156,380 SNVs that disrupt the G-tract of G-tract-AG signatures. One thousand random sets of SNVs matched by number, MAF and distance to TSS were selected similarly as above in parallel searches. Enrichment ratio was calculated similarly as well.

Disease enrichment/cluster analysis was carried out using the DAVID functional clustering analysis with the Genetic Association Database (54,55)

Study participants: This study used the 1,544 Third Generation cohort participants (mean age 47±9 years, 52% women, 100% American White) (56), at the second examination cycle (2008–2011). Protocols for genetic material collection during regular health examinations were approved by the Institutional Review Board at Boston Medical Center. All participants provided written informed consent for genetic studies, and the research was conducted in accordance with relevant guidelines and regulations.

RNA-Seq in the Framingham Heart Study: Peripheral whole blood samples were collected at the second examination cycle. RNA was isolated from peripheral whole blood using PAXgene™ tubes (PreAnalytiX, Hombrechtikon, Switzerland) to collect 2.5 mL samples from FHS participants (56). Samples were incubated at room temperature for 4 hours for RNA stabilization and then stored at −80 °C. Prior to extraction, tubes were thawed at room temperature for 16 hours. Total RNA was extracted with a PAXgene Blood RNA Kit at the FHS Genetics Laboratory. Following centrifugation and washing, white blood cell pellets were lysed in guanidinium buffer. RNA quality was assessed by absorbance at 260 and 280 nm using a NanoDrop ND-1000, and integrity was evaluated with the Agilent Bioanalyzer 2100 microfluidic.

RNA sequencing was conducted at the University of Washington Northwest Genomics Center, an NHLBI TOPMed reference laboratory, following standard protocols. Data processing took place at the same facility. Base calls were generated on the NovaSeq 6000, and unaligned BAM files were converted to FASTQ format using SamTools bam2fq. Sequence quality was assessed with the FASTX-toolkit, and alignment to GRCh38 with GENCODE release 30 was performed using STAR. FastQC provided initial quality control metrics, and

STAR created a genome index to generate BAM files at the gene and transcript levels(57). The TOPMed RNA-seq pipeline utilized RNA-SeQC for standard quality control metrics from aligned reads(58).

Whole genome sequencing: DNA from peripheral whole blood from 7,215 FHS participants underwent whole genome sequencing (WGS) at the Broad Institute, a TOPMed contracting laboratory. Details on DNA extraction, processing, and variant identification were previously described (50). Consistent sequencing and data processing criteria were applied, aligning DNA sequences to human genome build GRCh38. The resulting BAM files were sent to TOPMed's Informatics Research Center (IRC) for re-alignment and standardized BAM file generation. This study utilized WGS data from Freeze 10a, which included approximately 208 million SNPs across TOPMed cohorts.

Visualization of RNA-Seq reads corresponding to rs762810 (G>T) genotypes: We identified the same FHS participants with both RNA-Seq and WGS data. We then randomly selected three participants with GG, GT, and TT genotypes of rs762810 (G > T, MAF = 0.25 for T). We applied Integrative Genomics Viewer (IGV) (59) to visualize the RNA-Seq reads count in response to the genotypes of rs762810, GG, GT, and TT.

Structure modelling of protein-RNA complexes: We used AlphaFold 3.0(60) to model the hnRNP H1 protein-RNA structure and changes by the SNV. We first confirmed the prediction accuracy by aligning the top-ranked AlphaFold models with the corresponding crystal or NMR structures of a group of RNA-binding proteins in complex with RNA oligos (≤ 30nt) from the protein data bank PDB(61). The root mean square deviation (RMSD) was 0.559 angstrom (± 0.049, mean ± SEM, n = 21 pairs of predicted and PDB structures) after alignment using PyMOL (Molecular Graphics System, Version 3.0 Schrödinger, LLC). This deviation is significantly less than the threshold of 3.0 ($p < 0.0001$, one sample two-tailed t-test) for

structural differences (62). We then modelled the protein-RNA complex structures between hnRNP H1 qRRM1 and qRRM2 and the REPAG-containing RNA oligos 5´-UGGGGCU-3´ and variant 5´-UGGGUCU-3´of the 3´SS of *MAX* intron 3.

Mini-gene splicing reporter assay in HEK293T cells: G-tract, SNV or mutant reporters of *MAX* intron 3 or *QSOX1* intron 1 were made using the Gibson mutagenesis system(63), and verified by Sanger sequencing. Briefly, plasmids were transfected into HEK293T cells for overnight before RNA extraction for RT-PCR assay of spliced variants, as previously reported(64).

Plasmid Construction: Minigene splicing reporters were constructed using the Gibson Assembly Master Mix (New England Biolabs(63)) following the manufacturer's protocol. Linear PCR fragments were generated using primers with overlapping sequences corresponding to the *ApaI/BglII* linearized pDUP175 vector. The fragments were designed with the NEBuilder Assembly Tool.

For the pDUP175-MAX-E4-WT plasmid, a genomic DNA fragment containing exon 4 (81bp) and partial flanking introns (27 bp upstream and 22 bp downstream) was PCR amplified from human genomic DNA. The primers used were MAXE4-WT-DUP175 (5′-CTTGGGTTTCTGATAGGGCCCTGACCTCTCCCCTTCTTGGGGCTCAGGTG-3′) and MAX_3′SS_DUP175_REV (5′-TCTCTGTCTCCACATGCCCAACCTGGAGTCTCTGGTAC-3′).

For the pDUP175-STREX-QSOX1-3′SS-WT plasmid, the 3′ splice site of the STREX exon from the pDUPST1 vector was replaced by a 27-nt sequence from intron 1 of QSOX1 using PCR with primers QSOX1_DupST1_3SS'_WT (5′-CTTGGGTTTCTGATAGGGCCCTGACACCTTACCGTTTCCTGGGCCAGCCAAGATG

TCCATCTAC-3′) and DUP175_DupST1_3′SS_REV (5′-TCTCTGTCTCCACATGCCCAAGACATGGAACTGTGTGTG-3′). Mutant plasmids were created by amplifying the PCR fragments from the wild-type plasmid using complementary mutant-specific primers and cloned using the Gibson Assembly Master Mix. All plasmid constructs were verified through DNA sequencing.

Cell Culture and Transfection: Human embryonic kidney (HEK293T) cells were cultured at 37°C with 5% CO_2. The cells were grown in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin-glutamine (PSG). For transfections, 1µg of plasmid DNA was used per well in a 6-well plate, and transfection was performed using Lipofectamine 3000 (Invitrogen), according to the manufacturer's protocol.

RT-PCR: Reverse transcription was carried out based on a previously described procedure (7). Briefly, for reverse transcription, 1µg of total RNA (for MAX) or cytoplasmic RNA (for QSOX1) was added to a 10µl RT reaction and incubated at 45°C for 50min, followed by a 5-minute incubation at 90°C. For PCR amplification, 1µl of the RT product was used in a 24µl reaction. The PCR cycling was set for 30 cycles using primers DUP8a (5′-GACACCATGCATGGTGCACC-3′) and DUP10 (5′-CAAAGGACTCAAAGAACCTCTG-3′). PCR products were resolved on agarose gels, visualized under UV light, and photographed using a digital camera. Band intensities were quantified using ImageJ software (65). Exon inclusion efficiency was calculated as the molar percentage of the included products by their intensity relative to the total intensity of both included and skipped products after normalization to the production lengths.

Acknowledgements: We thank the donors for the Framingham Heart Study for sharing their genotype and RNA-Seq information, and Dr. Ren-Jang Lin for helpful comments. The work is supported by a discovery grant (RGPIN-2022-05023) by the Natural Sciences and Engineering Council of Canada (NSERC) to J.X., by NIH (GM127464 and P30CA023074) and Arizona Biomedical Research Centre (ABRC: RFGA2022-010-30) to S.S., and by NIH R01AA028263 and R01HL15569 to C.L. The Framingham Heart Study (FHS) was supported by NIH contracts N01-HC-25195, HHSN268201500001I, and 75N92019D00031. RNA-seq was supported in part by the Division of Intramural Research (D. Levy, Principal Investigator) and an NIH Director's Challenge Award (D. Levy, Principal Investigator).

Author contributions: JX conceptualized the project with LC and SS. JY, SO and JW carried out the bioinformatics analyses, cell culture or *in vitro* splicing reporter assays, respectively. AW contributed to the REPAG MEME and AlphaFold3 structural analyses, RJ and DL contributed to the FHS RNA-Seq data. JX wrote the manuscript, with critical revisions by LC and SS. All authors have read and approved the manuscript.

References

1. Shabalina, S.A. and Spiridonov, N.A. (2004) The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol*, **5**, 105.
2. Ling, J.P., Pletnikova, O., Troncoso, J.C. and Wong, P.C. (2015) TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science*, **349**, 650-655.
3. Gilbert, W. (1978) Why genes in pieces? *Nature*, **271**, 501.
4. Attig, J., Ruiz de Los Mozos, I., Haberman, N., Wang, Z., Emmett, W., Zarnack, K., Konig, J. and Ule, J. (2016) Splicing repression allows the gradual emergence of new Alu-exons in primate evolution. *Elife*, **5**.
5. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Gunes, O., Hall, P., Hayhurst, J. *et al.* (2023) The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res*, **51**, D977-D985.
6. McClory, S.P., Lynch, K.W. and Ling, J.P. (2018) HnRNP L represses cryptic exons. *RNA*, **24**, 761-768.
7. Lei, L., Cao, W., Liu, L., Das, U., Wu, Y., Liu, G., Sohail, M., Chen, Y. and Xie, J. (2018) Multilevel Differential Control of Hormone Gene Expression Programs by hnRNP L and LL in Pituitary Cells. *Mol Cell Biol*, **38**.
8. Das, U., Nguyen, H. and Xie, J. (2019) Transcriptome protection by the expanded family of hnRNPs. *RNA Biol*, **16**, 155-159.
9. Zarnack, K., Konig, J., Tajnik, M., Martincorena, I., Eustermann, S., Stevant, I., Reyes, A., Anders, S., Luscombe, N.M. and Ule, J. (2013) Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, **152**, 453-466.
10. Ling, J.P., Chhabra, R., Merran, J.D., Schaughency, P.M., Wheelan, S.J., Corden, J.L. and Wong, P.C. (2016) PTBP1 and PTBP2 Repress Nonconserved Cryptic Exons. *Cell Rep*, **17**, 104-113.
11. Yeo, G., Hoon, S., Venkatesh, B. and Burge, C.B. (2004) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A*, **101**, 15700-15705.
12. McCullough, A.J. and Berget, S.M. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol Cell Biol*, **17**, 4562-4571.
13. Martinez-Contreras, R., Fisette, J.F., Nasim, F.U., Madden, R., Cordeau, M. and Chabot, B. (2006) Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS Biol*, **4**, e21.
14. Nguyen, H., Das, U., Wang, B. and Xie, J.Y. (2018) The matrices and constraints of GT/AG splice sites of more than 1000 species/lineages. *Gene*, **660**, 92-101.
15. Sohail, M. and Xie, J. (2015) Evolutionary emergence of a novel splice variant with an opposite effect on the cell cycle. *Mol Cell Biol*, **35**, 2203-2214.
16. Sohail, M. and Xie, J. (2015) Diverse regulation of 3' splice site usage. *Cell Mol Life Sci*, **72**, 4771-4793.
17. Sohail, M., Cao, W., Mahmood, N., Myschyshyn, M., Hong, S.P. and Xie, J. (2014) Evolutionarily emerged G tracts between the polypyrimidine tract and 3' AG are splicing silencers enriched in genes involved in cancer. *BMC Genomics*, **15**, 1143.
18. Nguyen, H. and Xie, J. (2018) Widespread Separation of the Polypyrimidine Tract From 3' AG by G Tracts in Association With Alternative Exons in Metazoa and Plants. *Front Genet*, **9**, 741.
19. Romano, M. (2010) G runs in cystathionine beta-synthase c.833C/c.844_845ins68 mRNA are splicing silencers of pathogenic 3' splice sites. *Biochimica et biophysica acta*, **1799**, 568-574.
20. Xie, J. (2014) Differential evolution of signal-responsive RNA elements and upstream factors that control alternative splicing. *Cell Mol Life Sci*, **71**, 4347-4360.

21. Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R. *et al.* (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**, 1587-1593.
22. Sullivan, P.J., Gayevskiy, V., Davis, R.L., Wong, M., Mayoh, C., Mallawaarachchi, A., Hort, Y., McCabe, M.J., Beecroft, S., Jackson, M.R. *et al.* (2023) Introme accurately predicts the impact of coding and noncoding variants on gene splicing, with clinical applications. *Genome Biol*, **24**, 118.
23. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B. *et al.* (2019) Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, **176**, 535-548 e524.
24. Dawes, R., Bournazos, A.M., Bryen, S.J., Bommireddipalli, S., Marchant, R.G., Joshi, H. and Cooper, S.T. (2023) SpliceVault predicts the precise nature of variant-associated mis-splicing. *Nat Genet*, **55**, 324-332.
25. Zeng, T. and Li, Y.I. (2022) Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biol*, **23**, 103.
26. Yeo, G. and Burge, C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*, **11**, 377-394.
27. Nassar, L.R., Barber, G.P., Benet-Pages, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T. *et al.* (2023) The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res*, **51**, D1188-D1195.
28. Consortium, G.T. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, 580-585.
29. Ma, J., Huang, A., Yan, K., Li, Y., Sun, X., Joehanes, R., Huan, T., Levy, D. and Liu, C. (2023) Blood transcriptomic biomarkers of alcohol consumption and cardiovascular disease risk factors: the Framingham Heart Study. *Hum Mol Genet*, **32**, 649-658.
30. Merendino, L., Guth, S., Bilbao, D., Martinez, C. and Valcarcel, J. (1999) Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG. *Nature*, **402**, 838-841.
31. Kato, G.J., Lee, W.M., Chen, L.L. and Dang, C.V. (1992) Max: functional domains and interaction with c-Myc. *Genes Dev*, **6**, 81-92.
32. Makela, T.P., Koskinen, P.J., Vastrik, I. and Alitalo, K. (1992) Alternative forms of Max as enhancers or suppressors of Myc-ras cotransformation. *Science*, **256**, 373-377.
33. Mathsyaraja, H., Freie, B., Cheng, P.F., Babaeva, E., Catchpole, J.T., Janssens, D., Henikoff, S. and Eisenman, R.N. (2019) Max deletion destabilizes MYC protein and abrogates Emicro-Myc lymphomagenesis. *Genes Dev*, **33**, 1252-1264.
34. Arsura, M., Deshpande, A., Hann, S.R. and Sonenshein, G.E. (1995) Variant Max protein, derived by alternative splicing, associates with c-Myc in vivo and inhibits transactivation. *Mol Cell Biol*, **15**, 6702-6709.
35. Babic, I., Anderson, E.S., Tanaka, K., Guo, D., Masui, K., Li, B., Zhu, S., Gu, Y., Villa, G.R., Akhavan, D. *et al.* (2013) EGFR mutation-induced alternative splicing of Max contributes to growth of glycolytic tumors in brain cancer. *Cell Metab*, **17**, 1000-1008.
36. Zhang, H., Fan, S. and Prochownik, E.V. (1997) Distinct roles for MAX protein isoforms in proliferation and apoptosis. *J Biol Chem*, **272**, 17416-17424.
37. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.H., Raffield, L.M., Tardaguila, M., Huffman, J.E. *et al.* (2020) The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell*, **182**, 1214-1231 e1211.
38. Caillard, A., Sadoune, M., Cescau, A., Meddour, M., Gandon, M., Polidano, E., Delcayre, C., Da Silva, K., Manivet, P., Gomez, A.M. *et al.* (2018) QSOX1, a novel actor of cardiac protection upon acute stress in mice. *J Mol Cell Cardiol*, **119**, 75-86.

39. Giri, A., Hellwege, J.N., Keaton, J.M., Park, J., Qiu, C., Warren, H.R., Torstenson, E.S., Kovesdy, C.P., Sun, Y.V., Wilson, O.D. *et al.* (2019) Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat Genet*, **51**, 51-62.
40. Bacolla, A., Zhu, X., Chen, H., Howells, K., Cooper, D.N. and Vasquez, K.M. (2015) Local DNA dynamics shape mutational patterns of mononucleotide repeats in human genomes. *Nucleic Acids Res*, **43**, 5065-5080.
41. Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C. *et al.* (2020) ClinVar: improvements to accessing data. *Nucleic Acids Res*, **48**, D835-D844.
42. Li, Z., Huang, Z., Liu, Y., Cao, Y., Li, Y., Fang, Y., Huang, M., Liu, Z., Lin, L. and Jiang, L. (2023) Genotypic and phenotypic characterization of glucose-6-phosphate dehydrogenase (G6PD) deficiency in Guangzhou, China. *Hum Genomics*, **17**, 26.
43. Bancone, G., Chu, C.S., Somsakchaicharoen, R., Chowwiwat, N., Parker, D.M., Charunwatthana, P., White, N.J. and Nosten, F.H. (2014) Characterization of G6PD genotypes and phenotypes on the northwestern Thailand-Myanmar border. *PLoS One*, **9**, e116063.
44. Chiu, D.T., Zuo, L., Chao, L., Chen, E., Louie, E., Lubin, B., Liu, T.Z. and Du, C.S. (1993) Molecular characterization of glucose-6-phosphate dehydrogenase (G6PD) deficiency in patients of Chinese descent and identification of new base substitutions in the human G6PD gene. *Blood*, **81**, 2150-2154.
45. Wei, X., Kixmoeller, K., Baltrusaitis, E., Yang, X. and Marmorstein, R. (2022) Allosteric role of a structural NADP(+) molecule in glucose-6-phosphate dehydrogenase activity. *Proc Natl Acad Sci U S A*, **119**, e2119695119.
46. Lallena, M.J., Chalmers, K.J., Llamazares, S., Lamond, A.I. and Valcarcel, J. (2002) Splicing regulation at the second catalytic step by Sex-lethal involves 3' splice site recognition by SPF45. *Cell*, **109**, 285-296.
47. Dybkov, O., Preussner, M., El Ayoubi, L., Feng, V.Y., Harnisch, C., Merz, K., Leupold, P., Yudichev, P., Agafonov, D.E., Will, C.L. *et al.* (2023) Regulation of 3' splice site selection after step 1 of splicing by spliceosomal C* proteins. *Sci Adv*, **9**, eadf1785.
48. Burnight, E.R., Gupta, M., Wiley, L.A., Anfinson, K.R., Tran, A., Triboulet, R., Hoffmann, J.M., Klaahsen, D.L., Andorf, J.L., Jiao, C. *et al.* (2017) Using CRISPR-Cas9 to Generate Gene-Corrected Autologous iPSCs for the Treatment of Inherited Retinal Degeneration. *Mol Ther*, **25**, 1999-2013.
49. Martelly, W., Fellows, B., Kang, P., Vashisht, A., Wohlschlegel, J.A. and Sharma, S. (2021) Synergistic roles for human U1 snRNA stem-loops in pre-mRNA splicing. *RNA Biol*, **18**, 2576-2593.
50. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M. *et al.* (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, **590**, 290-299.
51. Bailey, T.L., Johnson, J., Grant, C.E. and Noble, W.S. (2015) The MEME Suite. *Nucleic Acids Res*, **43**, W39-49.
52. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*, **47**, D1005-D1012.
53. Paul, D.S., Albers, C.A., Rendon, A., Voss, K., Stephens, J., HaemGen, C., van der Harst, P., Chambers, J.C., Soranzo, N., Ouwehand, W.H. *et al.* (2013) Maps of open chromatin highlight cell type-restricted patterns of regulatory sequence variation at hematological trait loci. *Genome Res*, **23**, 1130-1141.
54. Becker, K.G., Barnes, K.C., Bright, T.J. and Wang, S.A. (2004) The genetic association database. *Nat Genet*, **36**, 431-432.

55. Sherman, B.T., Hao, M., Qiu, J., Jiao, X., Baseler, M.W., Lane, H.C., Imamichi, T. and Chang, W. (2022) DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*, **50**, W216-W221.
56. Liu, C., Joehanes, R., Ma, J., Wang, Y., Sun, X., Keshawarz, A., Sooda, M., Huan, T., Hwang, S.J., Bui, H. *et al.* (2022) Whole genome DNA and RNA sequencing of whole blood elucidates the genetic architecture of gene expression underlying a wide range of diseases. *Sci Rep*, **12**, 20167.
57. 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.
58. Lun, A.T., Chen, Y. and Smyth, G.K. (2016) It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. *Methods Mol Biol*, **1418**, 391-416.
59. Robinson, J.T., Thorvaldsdottir, H., Wenger, A.M., Zehir, A. and Mesirov, J.P. (2017) Variant Review with the Integrative Genomics Viewer. *Cancer Res*, **77**, e31-e34.
60. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J. *et al.* (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, **630**, 493-500.
61. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.
62. Magala, P., Klevit, R.E., Thomas, W.E., Sokurenko, E.V. and Stenkamp, R.E. (2020) RMSD analysis of structures of the bacterial protein FimH identifies five conformations of its lectin domain. *Proteins*, **88**, 593-603.
63. Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd and Smith, H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*, **6**, 343-345.
64. Liu, L., Nguyen, H., Das, U., Ogunsola, S., Yu, J., Lei, L., Kung, M., Pejhan, S., Rastegar, M. and Xie, J. (2024) Epigenetic control of adaptive or homeostatic splicing during interval-training activities. *Nucleic Acids Res*.
65. Schneider, C.A., Rasband, W.S. and Eliceiri, K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat Methods*, **9**, 671-675.

Figure Legends

Figure 1. Distribution of G-tract-AG signatures in the human genome and GWAS-mapped regions, and in mammals. **A**. Diagram of the genome-wide search for G-tract-AG signatures. (Upper) A successful search within the Ensembl-annotated 3´SS with the criteria: the consensus motifs G-tract and AG of identified REPAG-3´SS (n = 1000). (Lower) Search criteria for the G-tract, AG positions/ranges within the 23nt 3´SS-like sequences in the genome, with the result summary and G-tract consensus motifs at the bottom. #: n = 700, E-value =1.9e-478; ##: n = 419, E-value =3.6e-122. **B**. Enrichment index of SNVs within G-tract-3´SSL among GWAS SNVs (input n = 128, 991, MAF ≥ 0.05, p < 5E-08). Random SNVs from the TOPMed (n = 1000 sets, each with 128, 991 SNVs (RND), matched by MAF bins and TSS distance, excluding GWAS SNVs), were used as background for calculating the enrichment ratios of the abundance (GWAS/RND). p-value: by one-sample t-test. NS: non-significant. **C**. Genomic location profile of the GWAS SNVs within G-tract-3´SSL (n = 41,353), showing their mainly intron and intergenic locations. **D**. The G-tract-3´SSL of the *MAX* (Left) or *QSOX1* (Right) aligned among mammalian species, with human SNV variant sequences as the top two. 3´SSL MaxENT scores are to the right of each sequence. Black dot: location of potential branch point. Dotted box: exon, and heavy black line: intron. Shaded box: exon already emerged in humans as a tissue-restricted alternative exon.

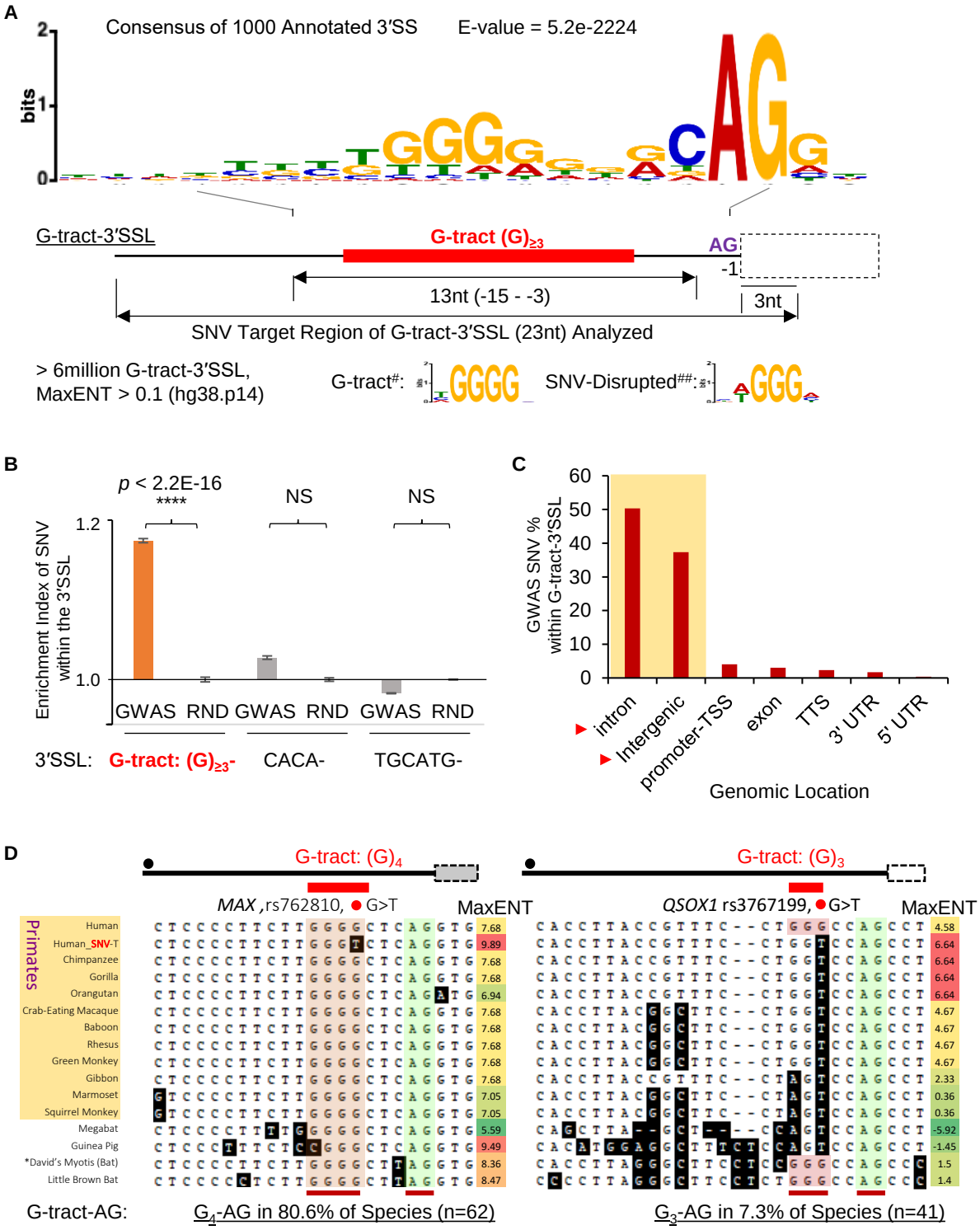
Figure 2. Functional validation of the NCR G-tract-AG signatures and SNVs: G-tract repression of splicing from the AG and disruption of the repression by SNV. **A**. Diagram of the SNV rs762810 in the genomic structure of the *MAX* gene transcripts, with the mammal-specific SNV region in the enlarged snapshots of GTEx RNA-Seq peaks and multiz alignment of 100 vertebrates from the UCSC Genome Browser. *: transcripts containing E4 (ENST00000557746.5 and NM_001407112.1). None of the other 53 GTEx tissues reaches the same level of E4 RNA-Seq peaks as in the blood. **B**. Splicing reporter assay

of the G-tract-3´SS of the SNV allele in human embryonic kidney HEK293T cells. Intron/exon structure and sequence around the rs762810 loci with either the allele of G or T are above the agarose gels, which show the RT-PCR of spliced products. **C**. The SNV effect in genotype-RNA-Seq matched human samples of white blood cells from the FHS cohort, demonstrating the RNA-Seq reads at the G-tract-3´SS with different genotypes of the SNV alleles.

Figure 3. REPAG and SNV effects on the two transesterification steps of splicing. **A**. Left, diagram of the *in vitro* splicing reporters, AdML vector, REPAG-WT and REPAG-Mut and their 3´SS sequences. Right, examples of the *in vitro* splicing lariat products and pre-mRNA. N = 3-6 experiments of the three reporters. **B**. Time course of lariat product generation in the 1st and 2nd steps, after normalizing each product to its peak band intensities during the time course of the experiment. $^{***}p < 0.001$, $^{*}p < 0.05$, n = 8 pairs in total at two (10 and 15min) time points. **C**. Time course of the conversion of the 1st step lariat intermediate to the 2nd step lariat, after normalizing the lariat intermediate percentages of total lariats (1st step lariat intermediate plus 2nd step lariat) in each reaction to the highest percentage during the time course of the experiment. $^{a}p < 9E-09$, n = 12 pairs in total; $^{b}p < 4E-06$, $^{c}p < 0.001$, n = 9 pairs in total, at three (15, 30 and 45min) time points of 3-6 independent experiments. Blue dots and SEM (standard error of the mean) error bars: half conversion time (min). The best fitted curve (with minimal Chi-squared values, $X^2 < 0.007$) was used to calculate the mean half conversion time of the 1st step lariat intermediates in each experiment by their equations in Igor Pro (Wavemetrics).

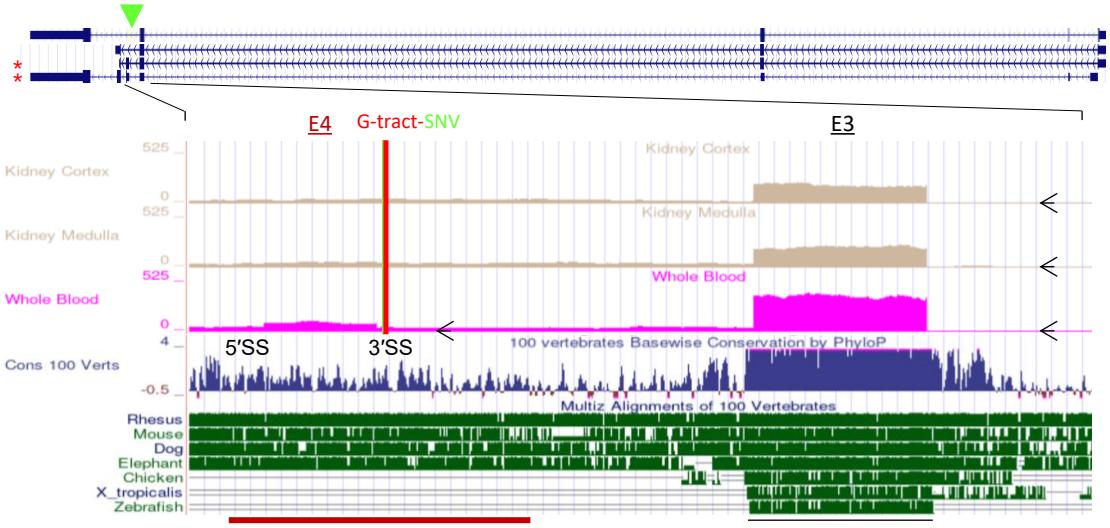
Figure 4. Disruption of the G-tracts of the signatures by human genetic variants *in cis* among GWAS-mapped genes, protein-coding genes and genetic disease genes. **A**. Significantly enriched GAD (Genetic Association Database) disease classes among the GWAS-mapped genes that harbor G-tract-AG signatures *in cis* with their G-tracts directly disrupted by GWAS SNVs, by DAVID functional

clustering analysis. **B**. Percentages of the GWAS-mapped or genome-wide genes that harbor such signatures and SNVs. N = 3,202 and 23,479 genes for GWAS and the genome, respectively. **C**. Total number of such SNVs *in cis* within the host genes among the databases screened. **D**. A reasonable explanation provided by the G-tract-AG signature and its disruption to account for the unexplainable protein deficiency and genetic disease caused by a rare genetic variant, the SNV c.392G->T of the *G6PD* gene. Disruption of the G-tract motif by the variant causes partial exon skipping and protein truncation at the N-terminal. The AlphaFold3.0-predicted structure of the wild type or missense variant protein aligned well (RMSD = 0.925 or 0.885, respectively), but the truncated protein failed to do so (green, RMSD = 14.500) with the corresponding N-terminal region (HQSD……FREDQI) of the electron microscopy structure (magenta, PDB #7UAG).


A
Consensus of 1000 Annotated 3′SS
E-value = 5.2e-2224
bits
G-tract-3′SSL
G-tract (G)≥3
AG
-1
13nt (-15 - -3)
3nt
SNV Target Region of G-tract-3′SSL (23nt) Analyzed
> 6million G-tract-3′SSL,
MaxENT > 0.1 (hg38.p14)
G-tract#:
SNV-Disrupted##:
B
p < 2.2E-16
****
NS
NS
Enrichment Index of SNV within the 3′SSL
1.2
1.0
GWAS RND GWAS RND GWAS RND
3′SSL: G-tract: (G)≥3- CACA- TGCATG-
C
GWAS SNV % within G-tract-3′SSL
60
50
40
30
20
10
0
intron
Intergenic
promoter-TSS
exon
TTS
3′UTR
5′UTR
Genomic Location
D
G-tract: (G)4
G-tract: (G)3
MAX ,rs762810, G>T
MaxENT
QSOX1 rs3767199, G>T
MaxENT
Primates
Human
Human_SNV-T
Chimpanzee
Gorilla
Orangutan
Crab-Eating Macaque
Baboon
Rhesus
Green Monkey
Gibbon
Marmoset
Squirrel Monkey
Megabat
Guinea Pig
*David's Myotis (Bat)
Little Brown Bat
G-tract-AG:
G4-AG in 80.6% of Species (n=62)
G3-AG in 7.3% of Species (n=41)

A

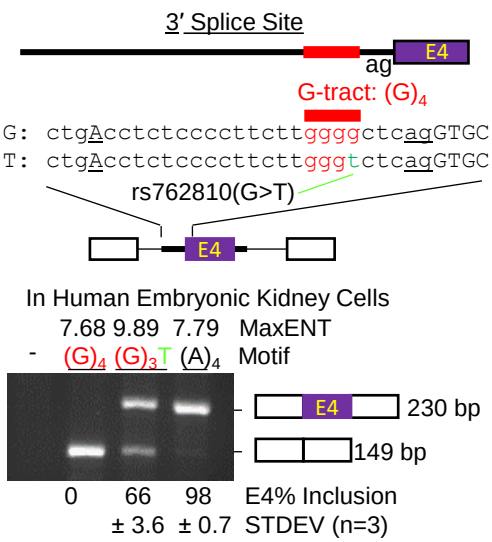
GWAS rs762810, Eosinophil percentage of white blood cells, *p* = 6E-10, SpliceAI = 0.33, in the *MAX* gene



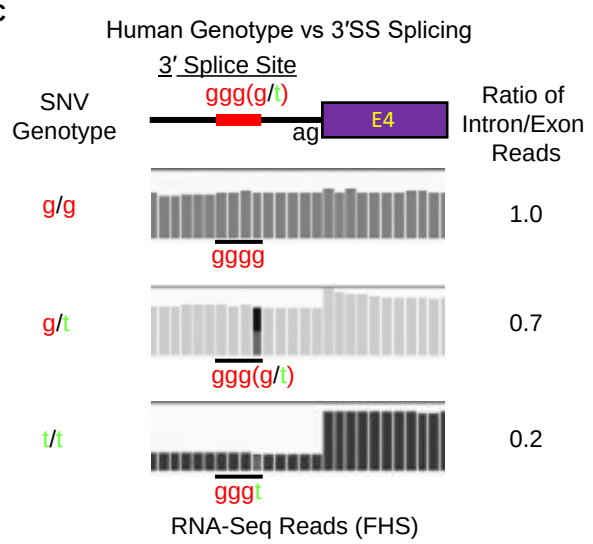
E4 Region in Mammals only

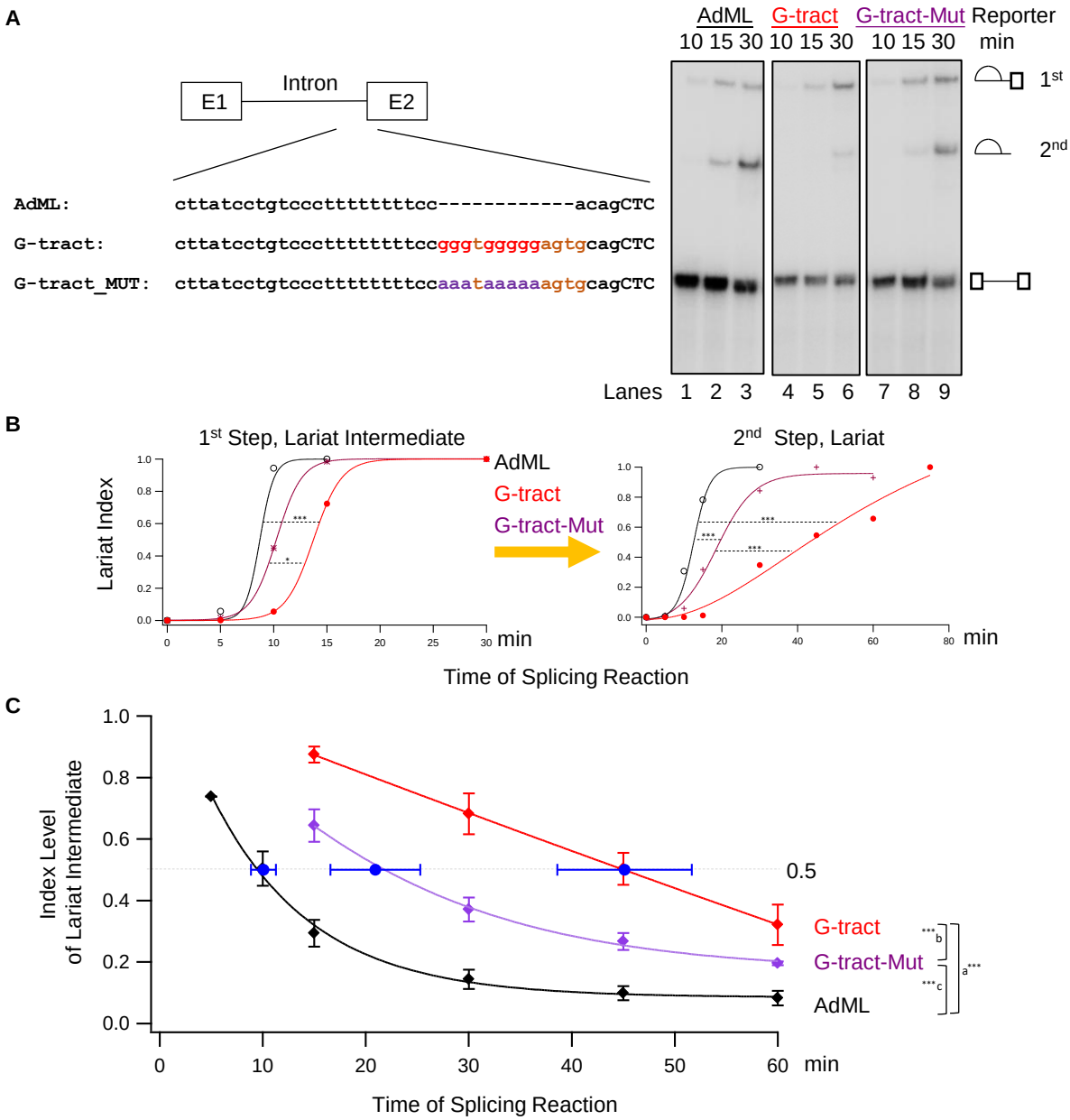
E3 in Vertebrates

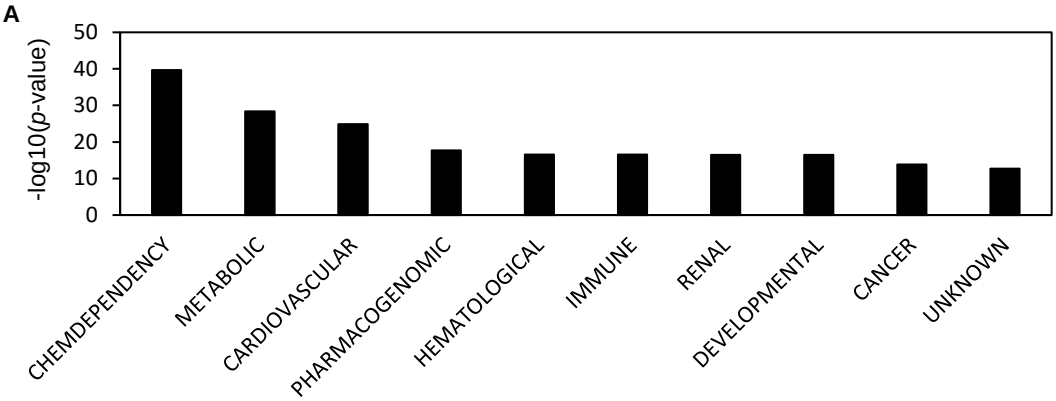
B

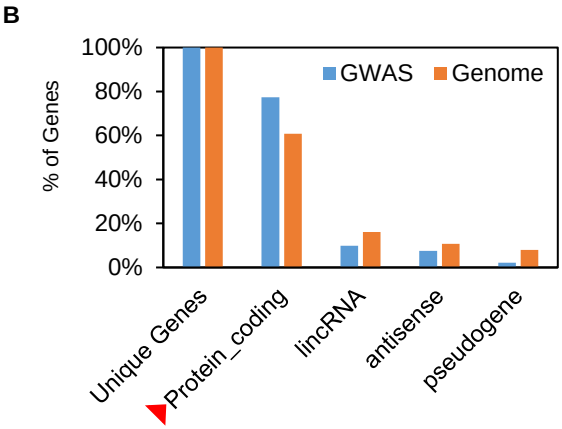


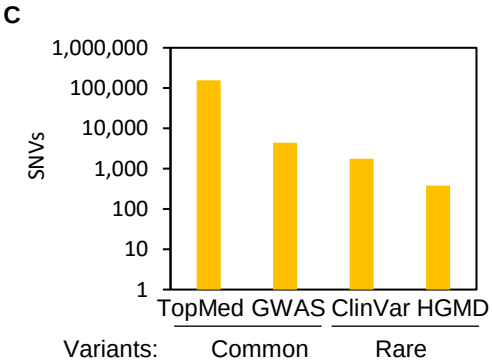
C




A
AdML
G-tract
G-tract-Mut
Reporter
10 15 30 10 15 30 10 15 30 min
Intron
E1
E2
AdML: cttatcctgtccctttttttcc----------acagCTC
G-tract: cttatcctgtccctttttttccgggtgggggagtgcagCTC
G-tract_MUT: cttatcctgtccctttttttccaaataaaaaagtgcagCTC
1st
2nd
Lanes 1 2 3 4 5 6 7 8 9
B
1st Step, Lariat Intermediate
2nd Step, Lariat
Lariat Index
AdML
G-tract
G-tract-Mut
min
Time of Splicing Reaction
C
Index Level of Lariat Intermediate
0.5
G-tract
G-tract-Mut
AdML
min
Time of Splicing Reaction

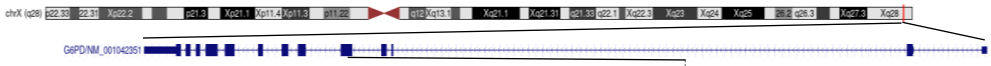




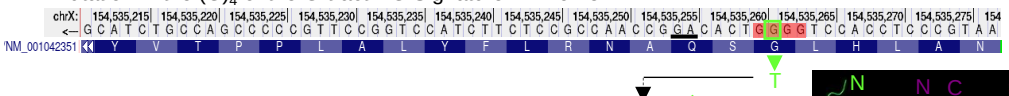


D

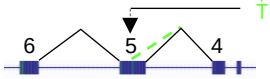
- *G6PD*, NM_001042351.1, c.392G>T, rs137852341, chrX:154535261, G131V, < 20% activity in hemizygotes



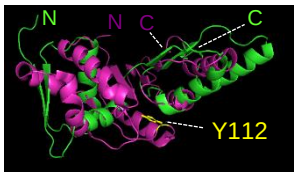
- Mutation in the $(G)_4$ of the G-tract-AG Signature in Exon 5



- More dramatic change explained by G-tract mutation: increasing 3′SSL MaxENT score from 5 to 6.9 and aberrant splicing, deletion of an N-terminal peptide region critical for $NADP^+$ binding and catalysis.

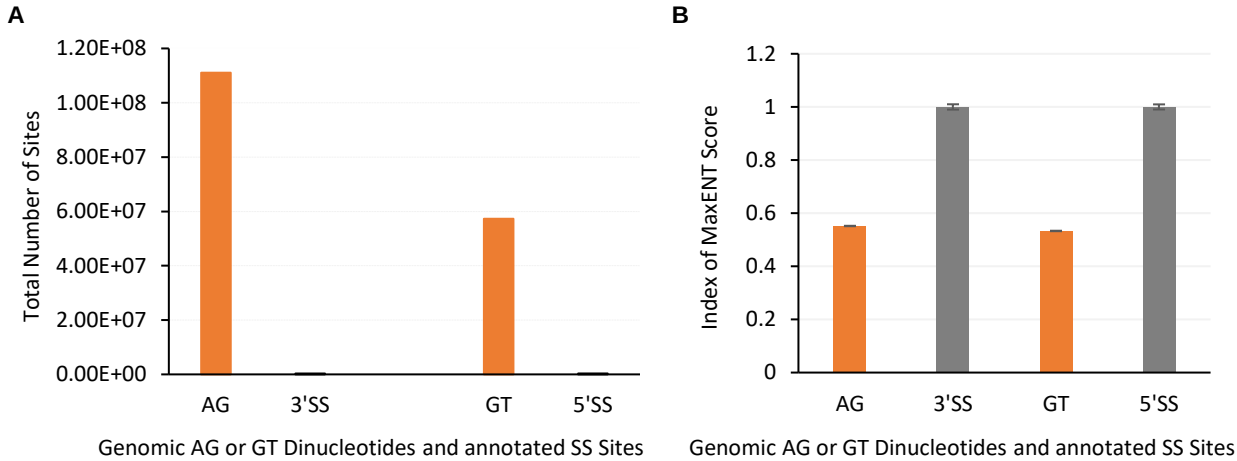


In-Frame Deletion
(…PFFK-(del44aa)-ANRL…)

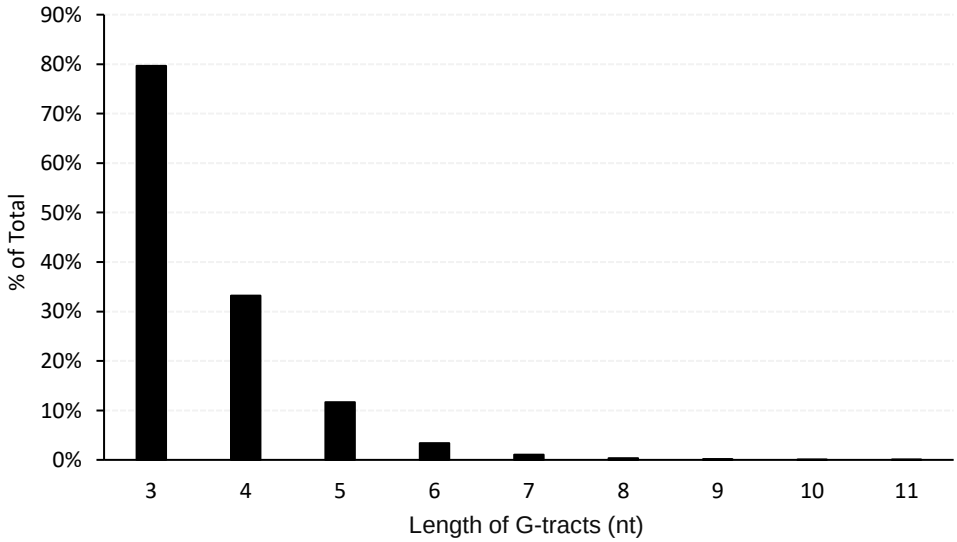


Aberrant splicing -> 132nt deletion -> 44aa deletion (90 – 133, Y112) and structural change

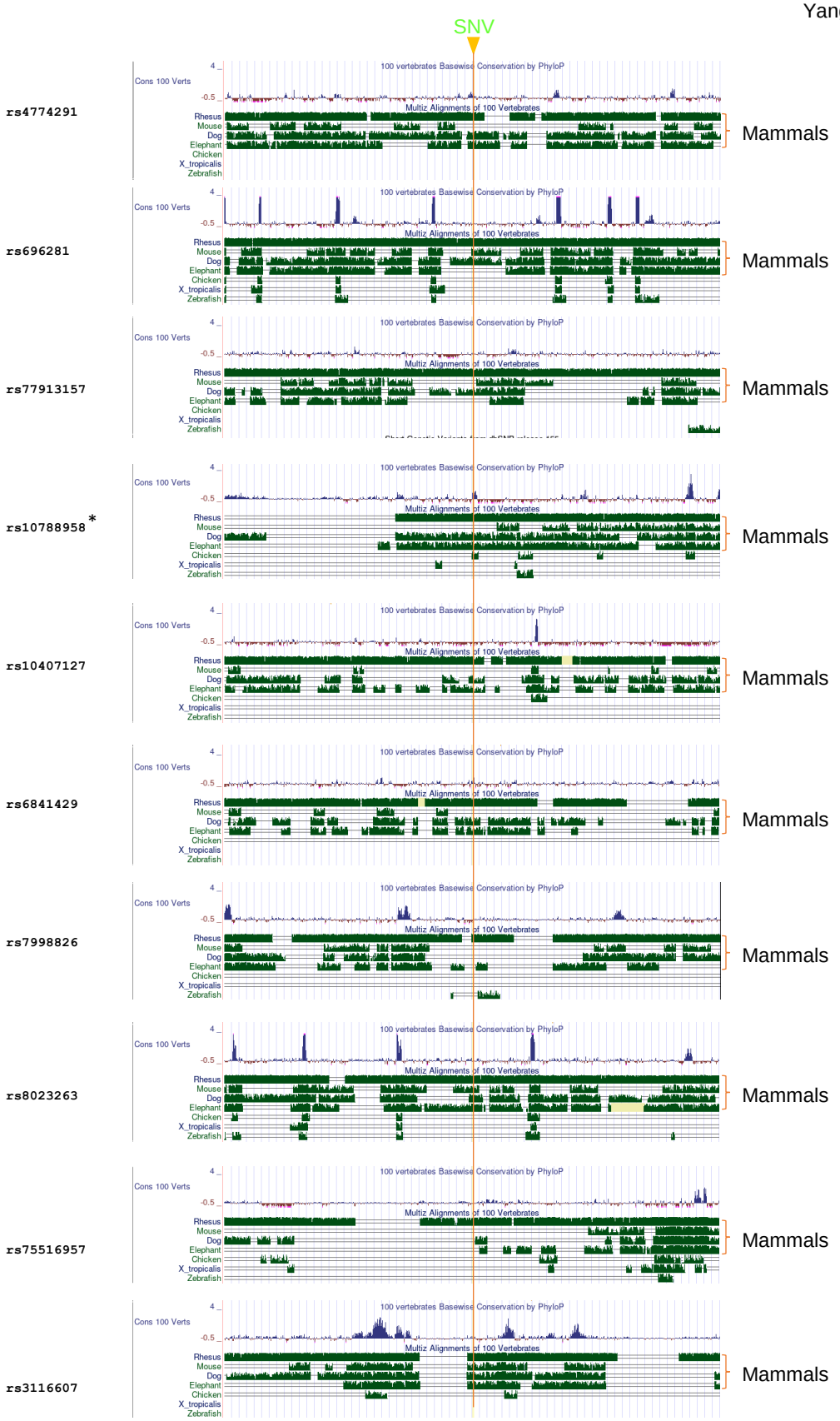
Supplementary Figures (S_Figs.)



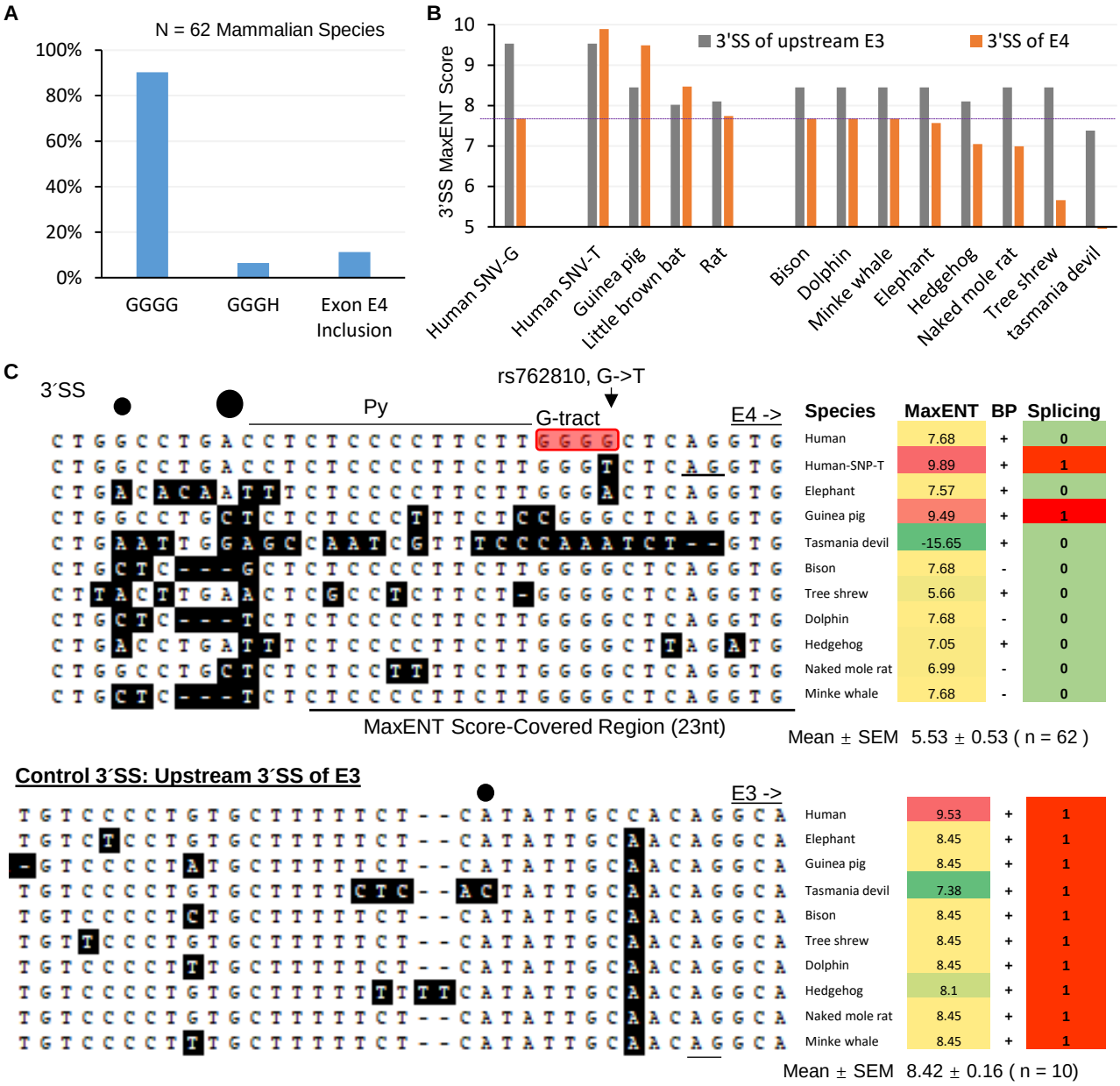
S_Figure 1. **168 millions of AG/GT dinucleotides as potential cryptic splice sites in the human genome. A**. Total number of AG/GT dinucleotides within SS-like sequences with MaxENT scores > 0.1 on both strands of the human genome hg38, with the relatively much smaller numbers of the annotated AG/GT 3′SS and 5′SS in the Ensembl for comparison. **B**. Index ratios of the MaxENT scores (Mean ± SEM) of sequences harboring the dinucleotides on either strand of chromosomes, assuming the AG/GT as the SS dinucleotides, relative to a subset of annotated constitutive exon's SSs (n = 5,010 exons) in the Chr. 1 of the human genome hg38. The means of the 3′SS and 5′SS scores are 6.7 (± 0.07, SEM, minimal -33.57 – maximal 15.33, >0.1: 92.6%) and 6.96 (± 0.07, SEM, minimal -26.61 – maximal 11.81, >0.1: 94.0%), respectively. The scores of these AG/GT dinucleotide sites in the genome are half of the constitutive SS's on average, posing significant challenges with the potential for cryptic splicing under appropriate conditions (e.g. *trans*-acting factors, signaling or cell/tissue types, etc. spatial or temporal changes).



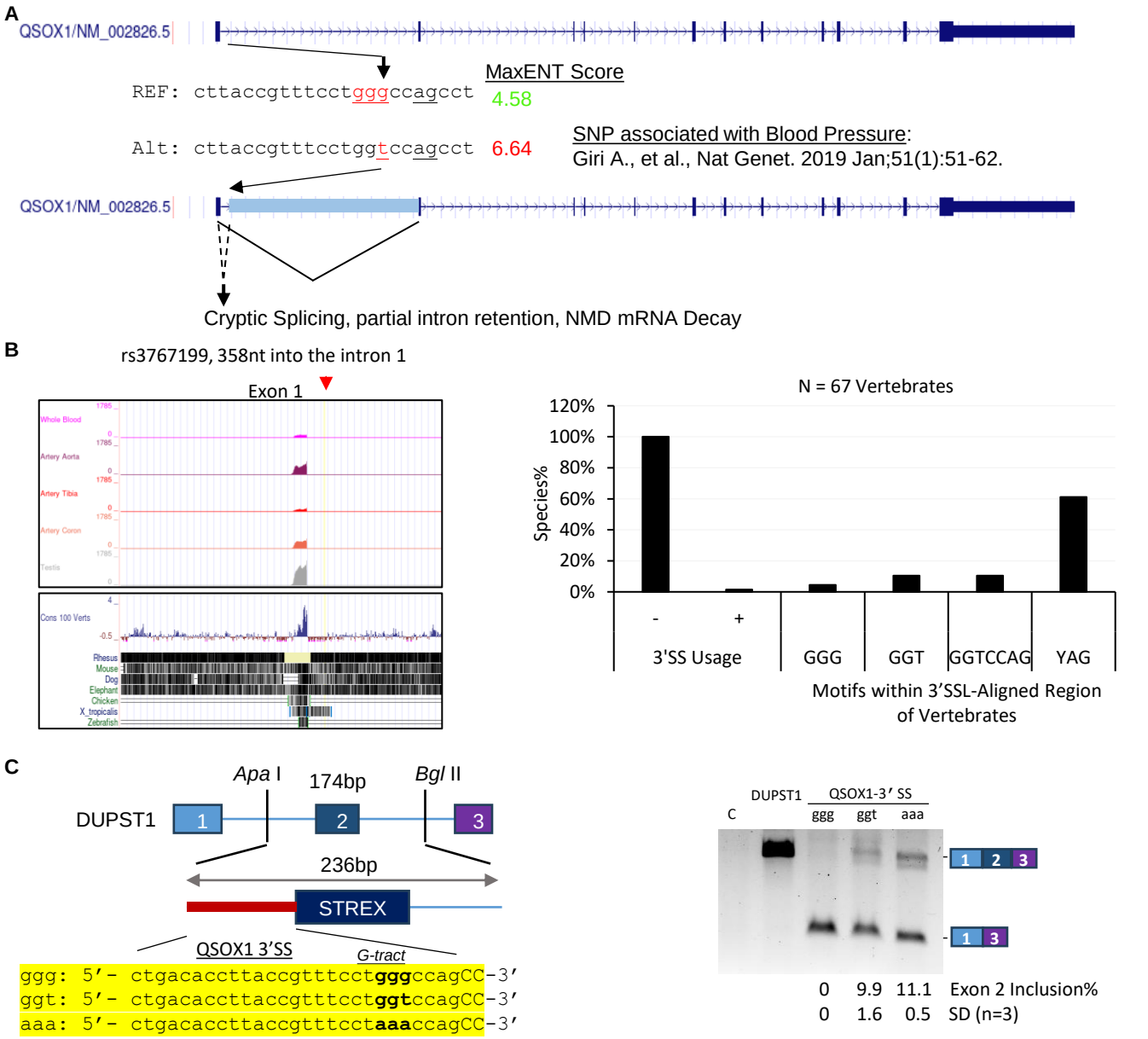
S_Figure 2. **Length (nt) distribution of the G-tracts, which are disrupted by GWAS SNVs, in the genomic signatures.** n = 6114 in total, from all GWAS SNVs with a *p*-value of less than 5E-08.



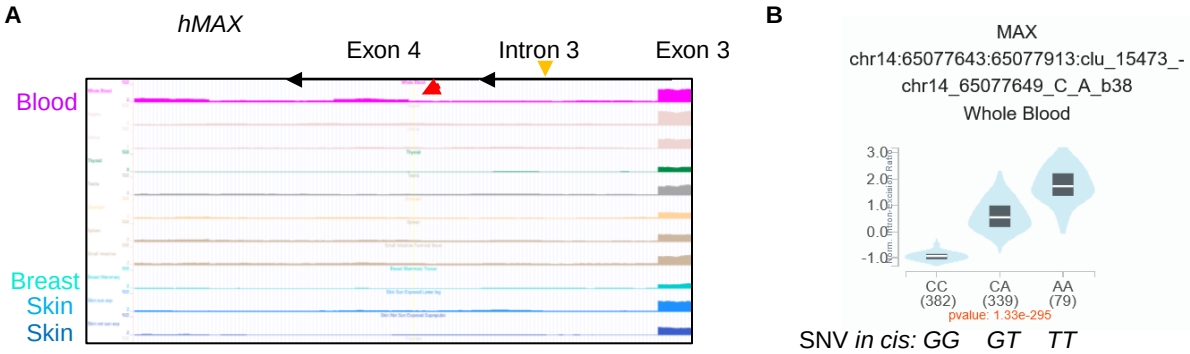
S_Figure 3. The human G-tract-AG signature is conserved in mammals only. Shown are 10 examples of the manually examined 37 SNVs that disrupt the G-tracts of the signatures in the MultiZ alignment of 100 vertebrates. Screenshots, taken from the UCSC Genome Browser, indicate their conservation in mammals only, in contrast to their flanking exon regions that are also in the other vertebrates like birds or fish. *: The rs10788958 alignment in chicken does not contain the G-tract nor AG.



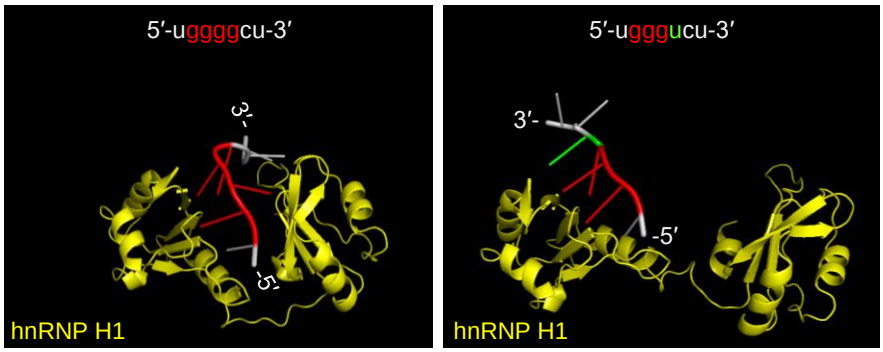
S_Figure 4. **Divergent sequence and 3´SS strength changes of the G-tract-3´SS of the *MAX* gene in some mammalian species**. **A**. Percentages of species containing the G-tract, its disruption and predicted exon 4 (E4) inclusion among 6 of 62 mammalian species, based on: higher 3´ and 5´ SS-combined MaxENT scores than the human reference sequence, presence of 3' AG and branch point motifs, shortened $(G)_4$- tract and intact ORF. H: A, C or T. **B**. In comparison with the human reference variant G (Left), 3´SS MaxENT scores of E4 of representative mammalian species that are higher (Middle) or lower (Right), in comparison with those of the upstream E3. **C**. Additional examples of 3´SS sequences showing the divergent changes of the *MAX* G-tract-3´SS motifs, besides the SNV-disruption of the G-tract among mammals, with human and guinea pig sequences for comparison. Upper panel: G-tract-3´SS alignment. Lower panel: the less divergent upstream 3´SS of exon 3 of the same species for comparison. Nucleotides different from the human reference genome sequences are shaded in black. Black dots: positions of potential branch points in some species. Py: polypyrimidine tract. Splicing prediction: by MaxENT scores, essential 3´SS motifs and/or transcript evidence. 1: positive or more likely; 0: negative or less likely. Note that here we use 3´SS since it is an authentic splice site in humans (see details on functional validation later).



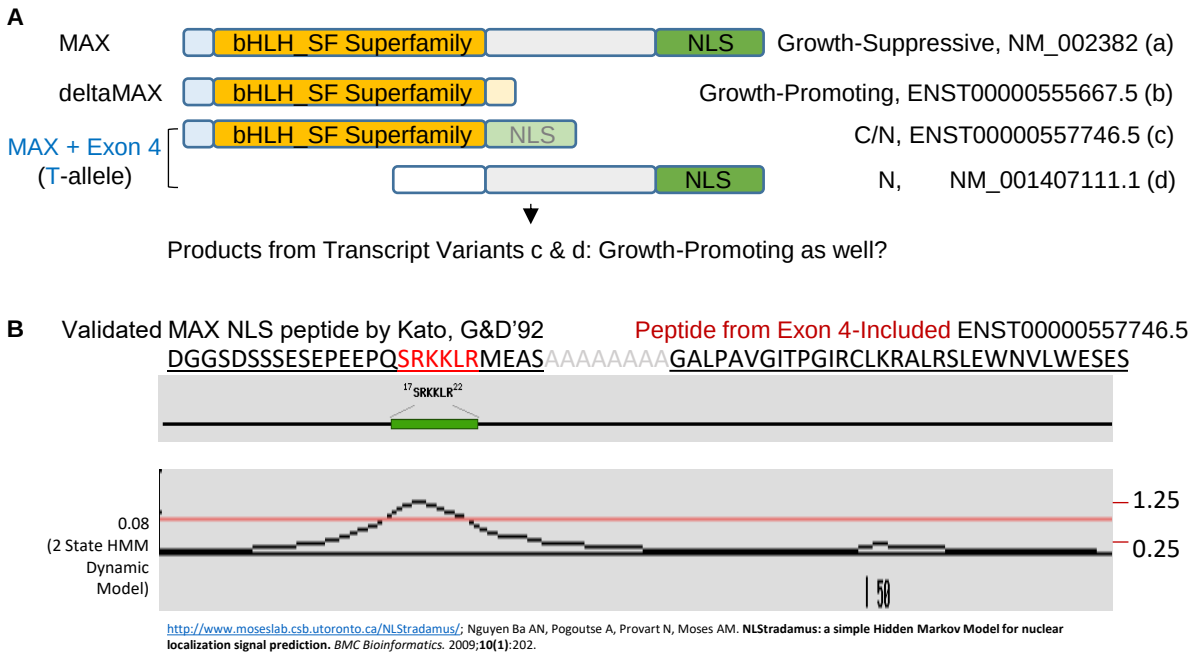
S_Figure 5. Validation of the effect of SNV rs3767199-T on splicing through the intronic G-tract-AG signature of the *QSOX1* gene by minigene splicing reporter assay. **A**. Diagram of the SNV region in the intron 1 of the *QSOX1* gene, adapted from the UCSC genome browser. **B**. Left, repression of the SNV-3′ SSL from cryptic splicing in human tissues: snapshot of the SNV region in GTEx showing RNA-Seq peaks of the upstream exon 1 in several tissues but none downstream the SNV, nor in the rest of the 54 tissues, with the aligned 100 vertebrate conservation profile below, both from the UCSC genome browser. Note: the GGG-AG signature is not in *X. tropicalis* though it is also aligned locally. Right, percentages of 67 vertebrate species with potential 3′ SS usage or the 3′ SSL motifs. The GGG and GGT are in only a small percentages of the species and all are in mammals only. 3′ SS usage (only with human SNV) based on: presence of 3′ AG, shortened G-tract and higher MaxENT 3′ SSL score than the human′s reference sequence. **C**. Left, diagram of the minigene splicing reporter, derived from the DUPST1 harboring the STREX exon (Xie & Black, *Nature*'01). Right, agarose gel of the RT-PCR products of cytoplasmic RNA of the splicing reporters transiently expressed in HEK293T cells. C: negative control with untransfected-cellular RNA. The $(G)_3$ repression was also observed when in front of a 51nt beta-globin exon.



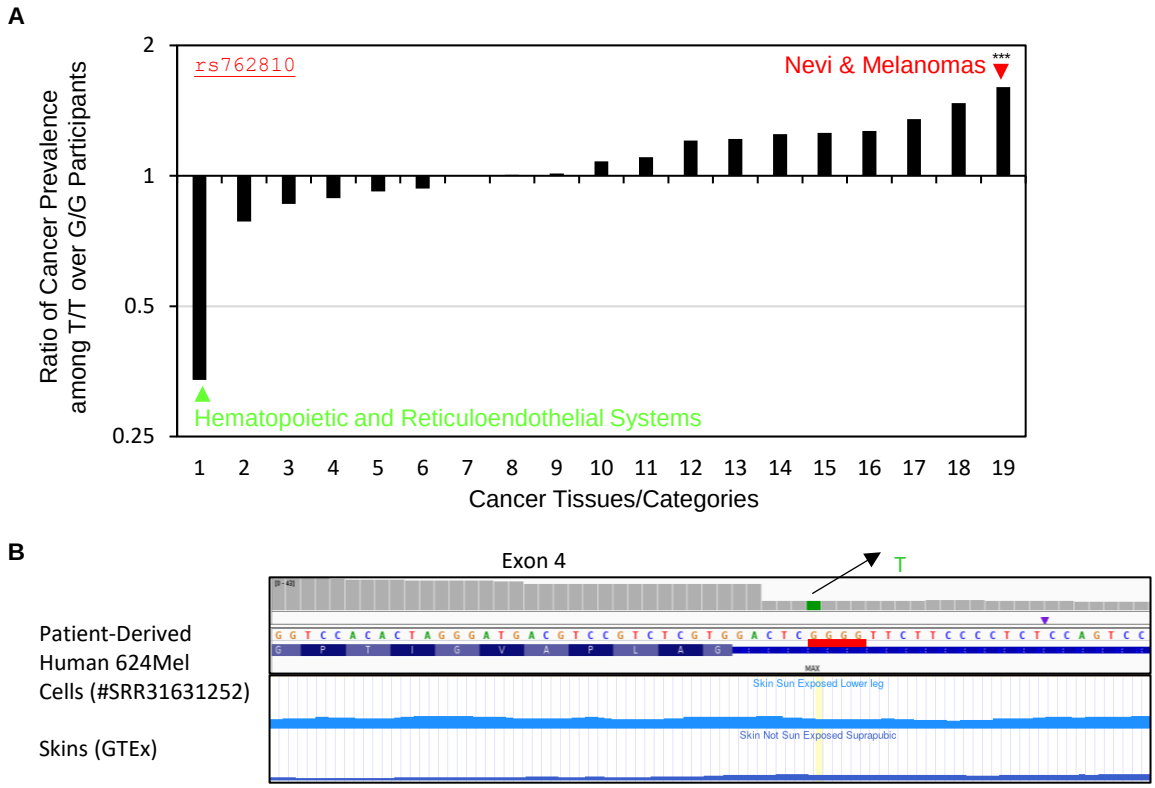
S_Figure 6. SNV-association with *MAX* exon 4 usage in blood cells in GTEx. **A**. Snapshot of the IGV view of the RNA-Seq reads peaks of the rs762810-T allele-associated exon 4 in the intron 3 of MAX ENST00000557746.5, among a number of GTEx tissues in the UCSC Genome Browser. The red arrowhead points to the location of the SNV and the upstream 3′SS of exon 4. Black arrows: direction of the *MAX* gene. **B**. GTEx result of the intron 3 excision ratios with different genotypes of the SNV rs762810.



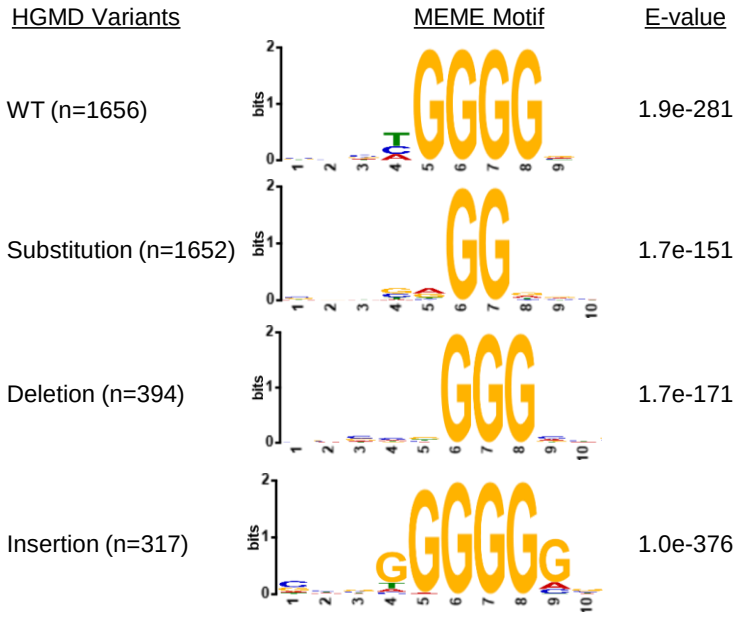
S_Figure 7. AlphaFold 3.0-predicted structural conformation changes in the interaction between the G-tract-G or -U RNA variants and the trans-acting splicing repressor hnRNP H1 qRRM1 and 2 domains. Shown are the MAX G-tract 5′-UGGGGCU-3′ (Gs in red) or its rs762810-T variant 5′-UGGGUCU-3′ (SNV U in green) in complex with the repressor. RMSD = 9.211 when aligned between the two complexes, indicating significant structural differences induced by the variants ($p < 0.0001$), in contrast to an average RMSD of 0.559 (±0.225, SEM) of 21 protein-RNA (≤ 30nt) complexes aligned with their experimental structures from the PDB in PyMol. The high-confidence RNA-protein complex show that the U-variant dramatically relaxed the G-tract RNA interaction with the trans-acting splicing repressor hnRNP H1. The complex switched from a compact key-lock pocket for the G-variant to a wide open configuration for the U-variant, in a manner similar to the crystal structure changes for different G-tracts (S. R. Penumutchu et al., *J Am Chem Soc* 140, 11661-11673 (2018)), suggesting destabilization of the G-tract-repressor interaction by the U-variant.



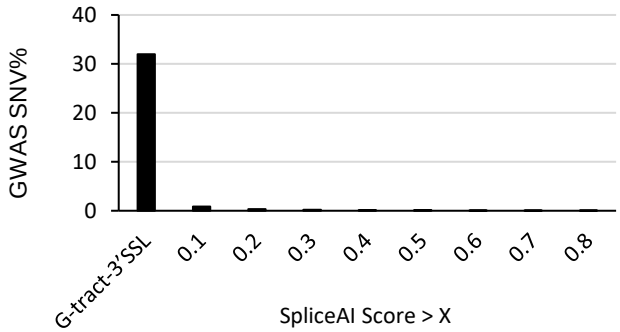
S_Figure 8. Domain loss of the MAX protein isoforms generated from the transcripts with the SNV-T allele-promoted exon 4 usage. **A**. Truncation of the COOH or NH_2 terminals in the isoforms from the exon 4-included transcripts, in comparison with the representative full length or short (delta) isoforms with known effects on cell growth. **B**. Significant loss of the NLS signal strength from the transcript ENST00000557746.5 containing exon 4. Prediction based on Nguyen et al.: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics.* 2009;10(1):202, with the known NLS peptide and ENST00000557746.5-encoded NLS-candidate peptide input together with a poly(Alanine) linker.



S_Figure 9. MAX variant genotype distribution among cancer patients of different tissues and exon 4 ectopic inclusion in melanoma cells with the T-variant. A. Tissue-dependent ratio of cancer prevalence among the T/T over G/G genotypes of SNV rs762810 of the *MAX* gene. Shown is the log2 ratios of the cancer prevalence among different tissues or histological categories over that among the cohort after normalization to (divided by) the allele frequencies in the cohort. From left to right: 1. hematopoietic and reticuloendothelial systems, 2. trachea bronchus, and lung, 3. only one cancer, 4. colon, 5. lymph nodes, 6. rectum, rectosigmoid junction, anal canal and anus, not otherwise specified, 7. control: all FHS participants with genotype data for SNP rs762810, 8. healthy controls from FHS with genotype data for SNP rs762810, 9. prostate gland, 10. urinary bladder, 11. adenocarcinoma, 12. female breast (excludes skin of breast), 13. basal cell carcinoma, 14. corpus uteri or endometrium, 15. skin, excludes skin of labia majora, skin of vulva, skin of penis and skin of scrotum, 16. with multiple cancers, 17. squamous cell carcinoma, 18. duct carcinoma, 19. nevi & melanomas, in an FHS cohort with sex- and age-matched genotype data for the SNV rs762810 (***: p < 4E-11, hypergeometric density test, n = 8,481 participants). **B**. Exon 4 ectopic inclusion (>50%) in patient melanoma-derived cells with the T variant only. Raw RNA-Seq sequence and BAM file derived from SRA #SRR31631252. Another sample with the T variant only and similar inclusion is #SRR31631251. The extra AG at the exon start is likely from the last two nts of exon 3 based on the gene sequence and RNA-Seq reads.



S_Figure 10. A group of G-tract mutations/variants identified in human genetic diseases in the HGMD database. Shown are the changes of the consensus motif sequences by G-tract mutations in comparison with the wild type sequences, as analyzed using the MEME. The mutations are from that published by Bacolla, et al., NAR'15 (43)5065.



S_Figure 11. Rarity of the G-tract-3′SSL SNVs among SpliceAI output results. Comparison of the percentages of G-tract-3′SSL associated SNVs (n=128,991 input, 32% with 3′SSL MaxENT >0.1, in the GWAS) and SpliceAI output of GWAS SNVs (n= 153,477 SNVs input). Only 166 of SpliceAI-predicted SNVs are above the confidence level of >0.5, including only 7 G-tract-disrupting SNVs. The MAX rs762810 was scored at 0.33, below the confidence level. Moreover, among the scored SNVs, only 0.1% of the SpliceAI scores were above 0.5 (S_Table II). The SNV-disrupted G-tract-3′SSLs are thus expected to add significantly to the list of splicing-regulating SNVs without the limitation of SNV distance to annotated splice sites by SpliceAI or other programs.