Exon-mediated activation of transcription starts

1 March 2019

Ana Fiszbein[1], Keegan S. Krick[1], Christopher B. Burge[1*]

[1] Department of Biology, Massachusetts Institute of Technology, Cambridge MA 02138

*Correspondence to: cburge@mit.edu

Summary

The processing of transcripts from mammalian genes often occurs near in time and space to their transcription. Here we describe a phenomenon we call exon-mediated activation of transcription starts (EMATS) that affects thousands of mammalian genes in which the splicing of internal exons impacts the spectrum of promoters used and expression level of the host gene. We observed that evolutionary gain of new internal exons is associated with gain of new TSSs nearby and increased gene expression. Inhibiting exon splicing reduced transcription from nearby promoters. Conversely, creation of new splice sites and enable splicing of new exons activated transcription from cryptic promoters. The strongest effects were associated with weak promoters located proximal and upstream of efficiently spliced exons. Together, our findings support a model in which splicing factors recruit transcription machinery locally to influence TSS choice, and identify exon gain, loss and regulatory change as major contributors to the evolution of alternative promoters and altered gene expression in mammals.

Introduction

The processing of RNA transcripts from mammalian genes often occurs nearby in time and space to their synthesis, creating opportunities for functional connections between transcription and splicing (Barbosa-Morais et al., 2012; Merkin et al., 2012; Oesterreich et al., 2016; Osheim et al., 1985). Several links between splicing and transcription are known and a key player on its coordination is the RNA polymerase itself (RNAPII). Transcription dynamics can influence splicing outcomes (Bentley, 2014; Kornblihtt et al., 2013) and chromatin structure can regulate splice site selection (Schor et al., 2013). However, more recent evidence suggests a "reverse-coupling" mechanism in which splicing feeds back to transcription. Adding an intron to an otherwise intron-less gene often boosts gene expression in plants, animals, and fungi via effects on transcription, nuclear export, mRNA stability, and/or translation (Furger et al., 2002; Shaul, 2017). Splicing can impact transcription elongation rates (Fong and Zhou, 2001) and in yeast, the presence of an intron can generate a transcriptional checkpoint that is associated with pre-spliceosome formation (Chathoth et al., 2014). Furthermore, recruitment of the spliceosome complex can stimulate transcription initiation by enhancing preinitiation complex assembly (Damgaard et al., 2008) and, inhibition of splicing can reduce levels of histone 3 lysine 4 trimethyl (H3K4me3), a chromatin mark associated with active transcription (Bieberstein et al., 2012).

Several components of the splicing machinery associate with RNA polymerase II (RNAPII) and other transcription machinery (Das et al., 2007; Emili et al., 2002; Kameoka et al., 2004; Morris and Greenleaf, 2000; Mortillaro et al., 1996; Neugebauer and Roth, 1997; Vincent et al., 1996). U1 and U2 snRNA associate with general transcription factors GTF2H (Kwek et al., 2002), GTF2F (Kameoka et al., 2004), and RNAPII CTD (Emili et al., 2002; Morris and

Greenleaf, 2000). In addition to its role in splicing, U1 snRNP acts as a general repressor of proximal downstream premature cleavage and polyadenylation (PCPA) sites (Gunderson et al., 1998; Kaida et al., 2010). The general absence of U1 snRNP binding sites in the antisense orientation from promoters determine that antisense transcripts generally terminate at PCPA sites, resulting in short unstable transcripts (Almada et al., 2013).

Although RNA splicing is a highly regulated process that influences almost every aspect of eukaryotic cell biology, transcript isoform differences across human tissues are predominantly driven by alternative start and termination sites (Reyes and Huber, 2018). Recent evidence suggests that transcription starts and splicing are coordinate (Anvar et al., 2018). However, whether exon splicing commonly impacts transcription start site (TSS) location and activity remains unknown. Here we describe a phenomenon we call exon-mediated activation of transcription starts (EMATS) in which the splicing of internal exons, especially those near gene 5' ends, influence gene expression by controlling which sites of transcription initiation are used. Our results demonstrate that exon splicing activates transcription initiation locally in thousands of mammalian genes enriched in neuronal development. Our findings also indicate that activation or repression of gene expression for research or therapeutic purposes may commonly be achievable by manipulation of splicing.

Results

Splicing is associated with increased gene expression and usage of TSS

We used a comparative approach to explore potential connections between splicing and TSS usage, examining transcript patterns in orthologous genes of mouse and rat that differed by the presence/absence of an internal exon (i.e. a non-terminal exon flanked by introns). Previously, we identified over one thousand such exons that were unique to the mouse transcriptome and not detected in RNA-seq data from diverse tissues of other mammals including rat, macaque, cow, etc., and are therefore likely to have arisen recently in the mouse lineage. A similar number of exons were unique to the rat. Most such evolutionarily new exons are located in 5' untranslated regions (UTRs) and are spliced in an alternative and tissue-specific fashion (Merkin et al., 2015). Between closely related species, we have observed that genes with evolutionarily new internal exons are associated with increased gene expression, but only in those tissues where the new exons are included (Figure 1A, S1A and Table S1) (Merkin et al., 2015). This trend was stronger for exons that were efficiently spliced – assessed by "percent spliced in" (PSI or ψ) values > 0.95, indicating that more than 95% of mRNAs from the gene include the exon (Figure 1B) – suggesting an association between extent of exon splicing and level of gene expression.

Grouping genes by their promoter structure, we found a positive association between inclusion of new exons and gene expression for genes with multiple TSSs while this association was not observed for genes with only one TSS (Figure 1C). Furthermore, our RNA-seq data (from (Merkin et al., 2012)) showed that genes with mouse-specific new exons were far more likely to have multiple TSSs compared to all expressed genes in mouse (Figure S1B and S1C).

We confirmed that genes with new mouse-specific exons are enriched for multiple TSSs using different methods to define TSSs, including H3K4me3 ChIP-seq peaks and data from high-resolution sequencing of polymerase-associated RNA (Start-seq) (Scruggs et al., 2015) (Figure 1D, S1D, Table S2). Genes with new rat-specific exons also had more TSSs per gene than rat genes overall (Figure S1E). Furthermore, genes that gained new species-specific exons were more likely to have gained TSSs in the same species, suggesting that the evolutionary gain of an internal exon is connected to evolutionary gain of TSSs nearby (Figure 1E and S1F).

To investigate this connection further, we examined the usage of new exons in relation to the number of TSSs used by a gene across tissues. We observed that genes containing mouse-specific exons used more distinct TSSs than their rat orthologs (Figure S1G), and that this association was specific to mouse tissues where the new exon was included with PSI > 0.05 (Figure 1F), showing a connection between splicing and TSS use. We also observed higher PSI values for new exons in genes with multiple alternative TSSs relative to genes with a single TSS (Figure 1H). Furthermore, we observed significantly increased gene expression levels in mouse relative to rat only in genes that gained TSSs in mouse (Figure 1G and S1I). Together, these observations indicate that use of new TSSs and splicing of new internal exons tend to occur in the same tissues, genes, and species, suggesting an intimate connection between splicing, increased gene expression and activation of new TSSs.

TSSs arise proximal and upstream of new exons

We observed a positional effect in which the increase in TSS count per gene was associated predominantly with new exons located in 5' UTRs rather than in 3' UTRs or coding

regions (Figure S2A). We examined the distribution of the locations of all mouse TSSs relative to the locations of mouse-specific new exons (Figure S2B), and compared it to the distribution of rat TSSs relative to sites homologous to the mouse-specific exons. This comparison showed an enrichment of TSSs in mouse within one or two kilobases (kb) upstream of new exons (Figure 2A, inset and S2C). Thus, evolutionary gain of new internal exons was specifically associated with gain of proximal, upstream TSSs.

We then asked about the relationship between splicing levels and usage of alternative TSS within the same gene. Considering relative TSS use ("TSS PSI", representing the fraction of transcripts from a gene that derive from a given TSS) we found that use of the most proximal upstream TSS (designated TSS –1) was positively correlated with new exon inclusion, especially for TSSs located within approximately 1 kb upstream of the new exon (Figure 2B and S2D). Furthermore, absolute expression of transcripts from nearby TSSs increased specifically in tissues where new exons were included at moderate or high levels (Figure 2C). These observations suggest a positive influence of splicing on nearby transcription (or possibly vice versa).

Manipulation of exon splicing impacts upstream transcription initiation

To directly test whether splicing impacts nearby transcription, we chose two candidate genes, *Gper1* (G protein-coupled estrogen receptor 1), and *Tsku* (Tsukushi, small leucine rich proteoglycan). These genes both have widespread, moderate expression and contain a mouse-specific 5' UTR internal exon whose splicing is positively correlated with the expression of the gene across mouse tissues (Spearman $\rho = 0.64$ and 0.57, respectively; Figure 3A and 3B left panels). When cultured mouse fibroblasts were treated with morpholino oligonucleotides (MO)

targeting splice sites of the new exons, exon inclusion decreased by about 4-fold in *Gper1* (Figure 3A) and *Tsku* (Figure 3B). Moreover, gene expression levels of these two genes were depressed to a similar extent (Figure S3A), consistent with a positive effect of exon inclusion on gene expression. We observed similar levels of repression of metabolically labeled nascent RNA (Figure S3A) as of steady state mRNA, indicating that the effect is primarily at the level of transcription rather than mRNA stability (Figure 3A and 3B). These observations support the idea that exon splicing positively impacts nearby transcription.

To confirm the directionality of this effect and to ask how splicing of new exons impacts the use of different TSSs, we used CRISPR/Cas9 mutagenesis to generate cell lines with mutations abolishing the inclusion of the new exon in the mouse *Stoml1* (Stomatin Like 1) gene, which has three alternative TSSs (Figure S3B). Notably, the three alternative TSSs of the *Stoml1* gene responded differently to inhibition of splicing of the new exon. The TSS in position –1 was down-regulated by 4-fold while downstream TSSs +1 and +2 were up-regulated to a similar extent in the mutant cell lines as measured by qPCR of nascent RNA (Figure 3C). Effects on antisense transcription in these mutant cell lines mirrored those observed for sense transcription (Figure 3C), suggesting that inclusion of the new exon enhances transcription from the upstream promoter in both directions. This pattern is distinct from some observations of intron-mediated enhancement in which sense-oriented introns specifically inhibited antisense transcription (Agarwal and Ansari, 2016), but is consistent with reported impacts on transcription initiation resulting from changes in the position of an intron in a reporter gene (Gallegos and Rose, 2017). The increase in downstream promoter activity was not observed in other genes studied and may reflect some sort of locus-specific (e.g., homeostatic) regulation of *Stoml1* expression. Levels of H3K4me3 and RNAPII decreased in the upstream TSS and increased in the downstream TSSs in

the mutant cell lines, consistent with the observed effects on nascent transcript production (Figure S3C).

Premature cleavage and polyadenylation (PCPA) can produce truncated, unstable transcripts, but can be inhibited by binding of spliceosome component U1 small nuclear ribonucleoprotein particle (snRNP) near (within one or two kb) of a PCPA site. If the observations above reflected effects of U1 snRNP or other splicing machinery on PCPA rather than on transcription, this would require the presence of PCPA sites near new exons ("nePCPA" sites) in affected genes. Using available polyA-seq data from five mouse tissues (Methods), we observed that only 8.6% of genes with new exons had evidence of a nePCPA site, a lower fraction than in a control set (Figure 3D). And for the subset of genes that contain nePCPA site(s), we did not observe differences in usage of the site between tissues where the new exon was spliced in from those where it was not (Figure 3E and S3D). Furthermore, we saw no relationship between the number of nePCPA sites and gene expression changes between mouse and rat (Figure S3E). Therefore, we found no evidence that effects on PCPA contribute significantly to EMATS. Since we observed similar effects on nascent RNA (in both sense and antisense orientations) as in steady state mRNA levels, altogether these results demonstrate that EMATS acts on transcription initiation.

Creation of a new splice site activates the use of a cryptic promoter nearby

We next sought to explore how splicing might affect use of different upstream TSSs. In the *Tsku* gene, the mouse-specific TSS in position –1 is located within 1 kb upstream of the mouse-specific exon, while the conserved TSS –2 is located further upstream. Analysis by 5' RACE showed that both TSSs are normally used at similar levels in mouse fibroblasts. However,

inhibiting inclusion of the new exon by MO produced a shift away from TSS –1 (Figure 4A and S4A). The strong down-regulation of transcription from TSS –1 observed by 5' RACE was confirmed by qRT-PCR of nascent RNA in both sense and antisense orientations (Figure S4B). This shift was accompanied by a 3-fold decrease in H3K4me3 (Figure 4B) in MO-treated cells. However, levels of H3K4me3 were unchanged near TSS –2, confirming that transcription from TSS –2 is not affected (Figure 4B). In cells treated with MOs, recruitment of general transcription factor 2F1 (GTF2F1) and RNAPII decreased by almost 3-fold near TSS –1 but were unchanged near TSS –2 (Figure 4C and SC). These observations suggest that splicing of the new exon contributes to recruitment of core transcription machinery to the proximal TSS –1. Moreover, the loss of signal of GTF2F1 and RNAPII near the new exon suggests that the inclusion of the new exon is associated with recruitment of transcription factors and higher levels of RNAPII as previously observed (Damgaard et al., 2008; Das et al., 2007). These observations demonstrate that splicing of new exons can regulate the usage of alternative TSSs, with predominant effects on the most proximal upstream promoter, consistent with the correlations observed in Figure 2.

To dissect the impacts of individual splice sites and splicing levels, we created an exon corresponding to the mouse-specific new exon in the rat *Tsku* gene and assessed effects on transcription. In the rat *Tsku* locus, transcripts are predominantly transcribed from the distal TSS –2. However, the regions homologous to TSS –1 and the mouse-specific new exon have high sequence identity with the mouse genome: both 5' splice sites are present in rat, but no YAG is present in rat near the location of the mouse 3' splice site, likely preventing splicing (Figure S4D). To introduce the desired mutations, we cloned the 5' end of the rat *Tsku* gene upstream of the coding sequence of *Renilla* luciferase and recreated the 3' splice site that is present in the

mouse genome (rn + mm 3'ss), as well as a stronger 3' splice site (rn + strong 3'ss), while either maintaining or mutating the native rat 5' splice site sequence (mm 3'ss + mut 5'ss). Remarkably, the creation of a 3' splice site promoted the inclusion of an exon analogous to that observed in mouse in constructs with an intact 5' splice site (Figure S4E), indicating that this mutation is sufficient to create a new exon in the rat locus. In the presence of both 3' and 5' splice sites, but not when either splice site was absent, total gene expression levels increased, as measured by luciferase activity (Figure 4D). By 5' RACE analysis, TSS –1 is used at basal levels in the minigene. However, the mouse-specific exon activates the usage of TSS –1 by 3-fold in the presence of a 5' splice site, demonstrating that the effect on TSS usage depends on the inclusion of the mouse-specific exon rather than merely the presence of a 3' splice site sequence (Figure 4E and S4F). In some examples studied previously, species-specific alternative splicing alters protein function (Gracheva et al., 2011; Gueroussov et al., 2015). Our observations support a distinct evolutionary pathway in which after a mutation that generates a new internal exon, transcription from a distal upstream promoter triggers a positive feedback loop in which inclusion of the new exon activates a cryptic proximal upstream promoter, which produces additional transcripts that include the exon, further activating the new promoter. The new TSS produces novel transcript isoforms and higher gene expression in tissues where the upstream promoter is active and the exon is included (Figure 4F).

Efficiently spliced exons activate usage of weak proximal TSSs

To investigate the genomic scope of the relationship between splicing and alternative TSS usage, we asked whether the inclusion of alternative skipped exons (SE) in general – not just those that evolved recently – can influence start site selection. We identified 49,488 SEs

detected in mouse RNA-seq data, distributed across 13,491 genes (Table S3). Analyzing unique SEs with TSS-exon distances matching those of new exons, we observed no significant association between SE inclusion and use of proximal upstream TSSs overall (Figure 5A). In addition, we observed a symmetrical distribution of TSSs around the location of SEs, distinct from the upstream-biased distribution seen relative to new exons (Figure 5B). These differences suggest that genes with new exons have distinct properties that promote linkage of splicing with transcription.

Examining other features of loci with new exons, we observed that, although new exons tend to have lower PSI values than SEs overall (Figure S5A), those new exons with proximal upstream TSSs tended to have higher PSI values and stronger 5' splice sites (Figure S5B). Furthermore, while the distribution of TSS PSI values was similar in genes with new exons and genes with SEs generally (Figure S5C), those TSSs located proximal and upstream of new exons had lower average expression levels across tissues than TSSs in other locations (Figure 5C). These observations suggested that the link between splicing and TSS usage is most pronounced when the promoter is intrinsically weak and splicing activity is high. Similarly, previous studies have observed stronger intron-mediated enhancement in the presence of weaker promoters (Callis et al., 1987). To test this idea, we grouped SEs and their most proximal and upstream TSSs into four bins from weak to strong (quartiles of TSS PSI values) and analyzed the correlation between TSS PSI and SE PSI separately for each bin. Notably, we observed that TSS usage was most highly correlated with exon inclusion for TSSs with PSI values in the bottom quartile (Figure 5D, S5D and S5E) and for SEs with PSI values in the top quartile (Figure 5E and S5F). Thus, the EMATS observed for new exons occurs more generally for SEs, with robust effects observed when a weak promoter is located upstream of a highly included SE: 21,326

instances of this pattern in 3,833 genes occur in the mouse genome, and the strongest effects when the weak promoter is also proximal: 1777 genes (Figure 5F; Table S3).

To further investigate the distance dependence of splicing effects on TSS use, we analyzed changes in TSS usage when inhibiting the inclusion of a SE in the mouse *Tsku* locus located more than 6 kb downstream of the TSSs. Perturbations of the splicing of this exon caused no detectible changes in TSS usage (Figure S5G), consistent with a requirement for proximity for exons to influence TSS use. Considering another mouse gene, chosen because it contained multiple TSSs and SEs, we observed that inhibition of the stronger upstream SE in the *Zfp672* (Zinc Finger Protein 672) locus affected the usage of TSSs more dramatically than inhibition of the weaker downstream SE in the same gene (Figure 5G). A weaker distal TSS (TSS –2) was impacted to similar degrees as a stronger proximal TSS (TSS –1) by splicing perturbations of these SEs (Figure 5G). Together, these observations confirm that splicing of SEs can impact TSS use, particularly when the TSS is intrinsically weak, the SE is highly included, and the TSS is located proximal and upstream of the SE. The generalization of EMATS implies that targeted activation or repression of the expression of a gene for research or therapeutic purposes may be achievable by use of compounds such as antisense oligonucleotides or small molecules (Havens and Hastings, 2016) that enhance or inhibit the splicing of an appropriately located promoter-proximal exon.

To investigate the biological relevance of EMATS we analyzed the role of genes with EMATS structure in diverse biological processes. Interestingly, those 1777 mouse genes with the strongest EMATS potential are enriched for functions in brain development, neuron projection morphogenesis, and synapse assembly (Figure 6A). This observation suggests that EMATS may

contribute to neuronal differentiation, perhaps making use of neuron-specific splicing factors to reinforce neuronal gene expression programs.

Splicing factors impact TSS use and interact with transcription machinery

The link between splicing and TSS usage could be mediated by splicing machinery, splicing factors, or exon junction complex components, particularly those factors that interact with transcription machinery, transcription factors or chromatin. To explore these potential links between splicing factors and TSS use, we analyzed transcriptome-wide changes in alternative TSS usage following knockdown of RNA-binding proteins (RBPs) using data from a recent ENCODE project (Van Nostrand et al., 2018). This analysis detected large numbers of TSS changes (Figure S6A), consistent with previous observations in *Drosophila* cells (Brooks et al., 2015). Depletion of factors involved in RNA splicing generally impacted larger numbers of TSSs than did depletion of other RBPs (Figure 6B and S6B). Among the ten splicing factors associated with the largest numbers of changes in TSS usage (Figure S6C), we found neuron-specific splicing factors as PTBP1 and transcription-machinery interactors as HNRNPU. Using protein-protein interaction (PPI) data from the STRING database (Szklarczyk et al., 2015), we observed that those ten splicing factors interact with 65 other proteins, including subunits of RNAPII and general transcription factors (Figure 6C). Compared with the PPI partners of the ten splicing factors whose depletion affected the fewest TSSs, these 65 proteins were enriched for functions in enhancer binding, transcription factor activity and promoter proximal binding (Figure 6C, S6D and S6E). Together, these observations indicate that some splicing factors, including neuron-specific splicing factors, have wider impacts on promoter choice than previously recognized, and establish that these factors interact extensively with core transcription machinery.

Discussion

Here, we have shown that creation and regulation of an internal exon in a gene – during evolution, by directed mutation or regulation – can activate transcription from an upstream TSS and thereby boost expression levels, a phenomenon which we refer to as EMATS. Our study highlights several features of this relationship: (i) it requires exon splicing, not merely presence of a 5' or 3' splice site; (ii) it is more potent when the promoter is intrinsically weak and exon inclusion is high; (iii) it is sensitive to distance, occurring most robustly when exon and promoter are within a kilobase or two; and (iv) the above features occur in thousands of mammalian genes (Table S3).

A model that captures the observed features would involve recruitment of specific splicing factors, as neuron-specific splicing factors, to a transcript which is being transcribed and is therefore tethered to the gene locus (Fig. 6D). The splicing of exons can directly recruit core transcription machinery to the local vicinity, increasing local concentration and occupancy of RNAPII at nearby promoters to directly boost transcription (Damgaard et al., 2008; Fong and Zhou, 2001; Kwek et al., 2002). The analyses above document the existence of various splicing-related proteins that interact with core transcription machinery, and impact alternative first exon expression when depleted. Once activated in this way, a new TSS produces new transcript isoforms and higher overall expression of the gene in specific tissues, providing a substrate for the regulatory evolution of the gene.

Acknowledgments

We thank Phillip A. Sharp, Alberto R. Kornblihtt and members of the Burge lab for helpful discussions and comments, as well as Maria Alexis, Jason Merkin, Peter Freese and Brenton R. Graveley for data access. This work was supported by grants from the NIH to C.B.B. (grant numbers HG002439, GM085319), and by the Pew Latin American Fellows Program in the Biomedical Sciences (A.F.).

Author Contributions

A.F. and C.B.B designed the study and wrote the manuscript. A.F conducted computational analyses, designed and performed experiments. K.S.K. contributed technically to some experiments. C.B.B supervised the work.

Declaration of Interests

The authors declare no competing interests. Correspondence and requests for materials should be addressed to cburge@mit.edu

References

Agarwal, N., and Ansari, A. (2016). Enhancement of Transcription by a Splicing-Competent Intron Is Dependent on Promoter Directionality. PLoS Genet. *12*, e1006047.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360–363.

Anvar, S.Y., Allard, G., Tseng, E., Sheynkman, G.M., de Klerk, E., Vermaat, M., Yin, R.H., Johansson, H.E., Ariyurek, Y., Dunnen, den, J.T., et al. (2018). Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. Genome Biol. *19*, 46.

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.

Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. Nat. Rev. Genet. *15*, 163–175.

Bieberstein, N.I., Carrillo Oesterreich, F., Straube, K., and Neugebauer, K.M. (2012). First exon length controls active chromatin signatures and transcription. Cell Rep *2*, 62–68.

Brooks, A.N., Duff, M.O., May, G., Yang, L., Bolisetty, M., Landolin, J., Wan, K., Sandler, J.,

Booth, B.W., Celniker, S.E., et al. (2015). Regulation of alternative splicing in Drosophila by 56 RNA binding proteins. Genome Res. *25*, 1771–1780.

Callis, J., Fromm, M., and Walbot, V. (1987). Introns increase gene expression in cultured maize cells. Genes Dev. *1*, 1183–1200.

Chathoth, K.T., Barrass, J.D., Webb, S., and Beggs, J.D. (2014). A splicing-dependent transcriptional checkpoint associated with prespliceosome formation. Mol. Cell *53*, 779–790.

Damgaard, C.K., Kahns, S., Lykke-Andersen, S., Nielsen, A.L., Jensen, T.H., and Kjems, J. (2008). A 5' splice site enhances the recruitment of basal transcription initiation factors in vivo. Mol. Cell *29*, 271–278.

Das, R., Yu, J., Zhang, Z., Gygi, M.P., Krainer, A.R., Gygi, S.P., and Reed, R. (2007). SR proteins function in coupling RNAP II transcription to pre-mRNA splicing. Mol. Cell *26*, 867–881.

Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res. *22*, 1173–1183.

Emili, A., Shales, M., McCracken, S., Xie, W., Tucker, P.W., Kobayashi, R., Blencowe, B.J., and Ingles, C.J. (2002). Splicing and transcription-associated proteins PSF and p54nrb/nonO bind to the RNA polymerase II CTD. *8*, 1102–1111.

Fong, Y.W., and Zhou, Q. (2001). Stimulatory effect of splicing factors on transcriptional elongation. Nature *414*, 929–933.

Furger, A., O'Sullivan, J.M., Binnie, A., Lee, B.A., and Proudfoot, N.J. (2002). Promoter proximal splice sites enhance transcription. Genes Dev. *16*, 2792–2799.

Gallegos, J.E., and Rose, A.B. (2017). Intron DNA Sequences Can Be More Important Than the Proximal Promoter in Determining the Site of Transcript Initiation. Plant Cell *29*, 843–853.

Gracheva, E.O., Cordero-Morales, J.F., González-Carcacía, J.A., Ingolia, N.T., Manno, C., Aranguren, C.I., Weissman, J.S., and Julius, D. (2011). Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats. Nature *476*, 88–91.

Gueroussov, S., Gonatopoulos-Pournatzis, T., Irimia, M., Raj, B., Lin, Z.Y., Gingras, A.C., and Blencowe, B.J. (2015). An alternative splicing event amplifies evolutionary differences between vertebrates. Science *349*, 868–873.

Gunderson, S.I., Polycarpou-Schwarz, M., and Mattaj, I.W. (1998). U1 snRNP Inhibits Pre-mRNA Polyadenylation through a Direct Interaction between U1 70K and Poly(A) Polymerase. Mol. Cell *1*, 255–264.

Havens, M.A., and Hastings, M.L. (2016). Splice-switching antisense oligonucleotides as therapeutic drugs. Nucleic Acids Res. *44*, 6549–6563.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664–668.

Kameoka, S., Duque, P., and Konarska, M.M. (2004). p54(nrb) associates with the 5' splice site within large transcription/splicing complexes. Embo J. *23*, 1782–1791.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods *7*, 1009–1015.

Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M.J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat. Rev. Mol. Cell Biol. *14*, 153–165.

Kwek, K.Y., Murphy, S., Furger, A., Thomas, B., O'Gorman, W., Kimura, H., Proudfoot, N.J., and Akoulitchev, A. (2002). U1 snRNA associates with TFIIH and regulates transcriptional initiation. Nat. Struct. Biol. *9*, 800–805.

Merkin, J.J., Chen, P., Alexis, M.S., Hautaniemi, S.K., and Burge, C.B. (2015). Origins and impacts of new mammalian exons. Cell Rep *10*, 1992–2005.

Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science *338*, 1593–1599.

Morris, D.P., and Greenleaf, A.L. (2000). The splicing factor, Prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II. J. Biol. Chem. *275*, 39935–39943.

Mortillaro, M.J., Blencowe, B.J., Wei, X., Nakayasu, H., Du, L., Warren, S.L., Sharp, P.A., and Berezney, R. (1996). A hyperphosphorylated form of the large subunit of RNA polymerase II is associated with splicing complexes and the nuclear matrix. Proc. Natl. Acad. Sci. U.S.a. *93*, 8253–8257.

Neugebauer, K.M., and Roth, M.B. (1997). Transcription units as RNA processing units. Genes Dev. *11*, 3279–3285.

Oesterreich, F.C., Herzel, L., Straube, K., Hujer, K., Howard, J., and Neugebauer, K.M. (2016). Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. Cell *165*, 372–381.

Osheim, Y.N., Miller, O.L., and Beyer, A.L. (1985). RNP particles at splice junction sequences on Drosophila chorion transcripts. Cell *43*, 143–151.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat Protoc *8*, 2281–2308.

Reyes, A., and Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. Nucleic Acids Res. *46*, 582–592.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics *27*, 2325–2329.

Schor, I.E., Fiszbein, A., Petrillo, E., and Kornblihtt, A.R. (2013). Intragenic epigenetic changes modulate NCAM alternative splicing in neuronal differentiation. Embo J. *32*, 2264–2274.

Scruggs, B.S., Gilchrist, D.A., Nechaev, S., Muse, G.W., Burkholder, A., Fargo, D.C., and Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Mol. Cell *58*, 1101–1112.

Shaul, O. (2017). How introns enhance gene expression. Int. J. Biochem. Cell Biol. *91*, 145–155.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562–578.

Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A., Dominguez, D., et al. (2018). Biorxiv 1–111.

Vincent, M., Lauriault, P., Dubois, M.F., Lavoie, S., Bensaude, O., and Chabot, B. (1996). The nuclear matrix protein p255 is a highly phosphorylated form of RNA polymerase II largest subunit which associates with spliceosomes. Nucleic Acids Res. *24*, 4649–4652.

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.

Yu, H.-B., Yurieva, M., Balachander, A., Foo, I., Leong, X., Zelante, T., Zolezzi, F., Poidinger, M., and Ricciardi-Castagnoli, P. (2015). NFATc2 mediates epigenetic modification of dendritic cell cytokine and chemokine responses to dectin-1 stimulation. Nucleic Acids Res. *43*, 836–847.

Methods

RNA-seq analysis and genome builds

We used the RNA-seq data from 9 tissues from mouse and rat (3 individuals each) associated with Merkin et al. (Merkin et al., 2012), available at NCBI Gene Expression Omnibus (GEO) (accession no. GSE41637). Reads were mapped to the mm9 and rn4 genome builds, respectively,

and processed using TopHat (Trapnell et al., 2009) and Cufflinks (Trapnell et al., 2012). Alternative splicing patterns and PSI values were analyzed using MISO (Katz et al., 2010). Exons were defined as in Merkin et al. (Merkin et al., 2012), requiring FPKM ≥ 2 and meeting splice site junction read requirements implicit in the TopHat mapping. Exons with 0.05 < PSI < 0.97 in at least one tissue and two individuals were categorized as skipped exons (SE). Exons with PSI > 0.97 in all expressed tissues were defined as constitutive exons (CE), if the gene was expressed in at least three tissues and two individuals. Genomic and splicing ages were defined as in Merkin et al. (Merkin et al., 2015) by the pattern of species with genomic regions aligned to the exon or with an expressed region in the orthologous gene overlapping the aligned region, respectively. Open reading frames (ORFs) were annotated as in Merkin et al. (Merkin et al., 2012) and used to classify exons as located in the 5' UTR, 3' UTR or coding region.

Cell lines, cell culture and treatments

NIH3T3 and HeLa cells were grown in DMEM, with high glucose and pyruvate (Gibco), supplemented with 10% fetal bovine serum (FBS). Mouse CAD (Cath.-a-differentiated) cells were grown in DMEM/F12 (Gibco) supplemented with 10% FBS. For morpholino oligonucleotide (MO) treatment (Gene Tools), 20 μM of morpholino targeting 5' or 3' splice site or MO control was added with endoporter (Gene Tools) to cells plated at low confluence and left for 24 h.

CRISPR sgRNA design, genetic deletions and genotyping

CRISPR-Cas cell lines with the 5' splice site of *Stoml1* deleted were generated using the protocol described in Ran et al. (Ran et al., 2013). The single-guide RNA was designed in silico to target the 5' splice site using the CRISPR Design Tool (http://tools.genome-engineering.org) and cloned into a Cas9 expression plasmid (pSpCas9). After transfecting CAD cells with the plasmid expressing Cas9 and the appropriate sgRNA, clonal cell lines were isolated and insertion/deletion mutations were detected by the SURVEYOR nuclease assay. Positive clones detected were amplified by PCR, subcloned into TOPO-TA plasmids, and individual colonies were sequenced to reveal the clonal genotype.

RNA Extraction, RT-PCR and qPCR

Total RNA was extracted using the RNA-easy kit (Qiagen) according to the manufacturer's protocol. Reverse transcription using M-MLV reverse transcriptase (Invitrogen) and random primers was performed according to the manufacturer's instructions. For nascent RNA extraction, RNA was metabolically labeled with 5-Ethynil Uridine for 10 minutes using Click-iT (Invitrogen) and labeled RNA was extracted and amplified according to the manufacturer's instructions. Quantitative PCR analyses were performed with SYBR green labeling using a LightCycler 480 II (Roche).

ChIP and antibodies

Chromatin immunoprecipitation was performed using the MAGnify™ Chromatin Immunoprecipitation System (Invitrogen) according to the manufacturer's recommendations. For each immunoprecipitation, we used 10 µg of H3K4me3 antibody (PA5-17420 from Invitrogen), 10 µg of RNA polymerase II antibody (Ab817 from Abcam), 10 µg of Transcription Factor IIF1 (GTFIIF1) antibody (PA5-30050 from Invitrogen) and 10 µg of Rabbit IgG antibody (Invitrogen) as a negative control. DNA was purified and quantitative PCR analysis was performed with SYBR green labeling using a LightCycler 480 II (Roche). Immunoprecipitated chromatin was normalized to input chromatin and control IgG antibody.

5' RACE

5' RACE experiments were performed with 5' RACE System for Rapid Amplification of cDNA Ends (Invitrogen) using three gene-specific primers (GSP) that anneal to the known region and an adapter primer that targets the 5' end. Products generated by 5' RACE were subcloned into TOPO-TA vectors and individual colonies were sequenced.

Plasmids and luciferase activity assay

Rat *Tsku* genomic region and mutants were cloned into the psiCHECK backbone. For transfection assays, 1 μg plasmid was transfected into each well of a 6-well culture plate using Lipofectamine 2000 (Life Technologies) according to the manufacturer's recommendations and cells were harvest after 24 h. To measure luciferase activity, we used the Dual-Luciferase® Reporter Assay System (Promega).

Software for data analysis, graphical plots and statistical analyses

For data analysis we used R Bioconductor, BEDTools, SamTools, GenomicRanges and the Integrative Genomics Viewer. All statistical analyses were performed in R (v.3.4.2) and graphical plots were made using the R package ggplot2. Lower and upper hinges of box plots correspond to the 25th and 75th percentiles, respectively. The upper and lower whiskers extend from the hinge to the largest and lowest value no further than 1.5 × IQR (interquartile range), respectively. Notches give approximate 95% confidence interval for comparing the medians. Statistical significance is indicated by asterisks (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$, *****$p < 0.00001$).

Definition of species-specific exons

Evolutionarily new exons were identified as is Merkin et al. (Merkin et al., 2015). Genomic mappings of mouse and rat RNA-seq data were combined with whole-genome alignments to classify the species distribution of exons. Only internal exons were considered in this analysis, excluding first and last exons, and only unique exons were considered, excluding exons that arose from intra-genic duplications to avoid issues related to possibly inaccurate genome assemblies, annotations or read mappings. In all, 1,089 mouse exons were classified as mouse-specific exons and 1,571 rat exons were classified as rat-specific exons, as they were detected in RNA-seq data from mouse or rat, respectively, but not from any other species analyzed (Supplementary Tables 1, 2). Most genes that contained a new exon had only one, with 159 mouse genes and 276 rat genes containing more than one new exon.

Transcription start site annotation

TSSs in mouse were identified using Start-seq data from Scruggs et al. (Scruggs et al., 2015) downloaded from GEO (accession no. GSE62151); Start-seq uses high-throughput sequencing of nascent capped RNA species from the 5'-end, allowing for definition of TSSs at nucleotide resolution. TSSs were defined within 2,000 bp search windows centered on RefSeq-annotated TSSs, using the location to which the largest number of Start-RNA reads aligned. Very closely spaced TSSs with a distance of less than 50 nucleotides were considered as a single TSS in Fig. 1D. To identify TSSs in the same RNA-seq data used to classify new exons, we used data from Merkin et al. (Merkin et al., 2012) (GEO accession no. GSE41637) mapped with TopHat combined with Ensembl annotations. As in Merkin et al. (Merkin et al., 2012), Cufflinks version 1.0.2 was used to identify novel transcripts. The set of TSSs from each library identified from transcripts as the start site of the first exon were combined with the existing Ensembl annotations and merged into a single set of annotations using Cuffcompare (Roberts et al., 2011). Cufflinks was then applied to each library to quantitate the same set of transcripts (Supplementary Table 3). The number of TSS were also estimated by the number of H3K4me3 peaks assigned to each gene with ChIP data from Yu et al. (Yu et al., 2015) (GEO accession no. GSE59896 and GSE59998).

New exon inclusion, TSS usage, and species-specific expression

We considered genes with new exons as all genes with a new exon with PSI > 0.05 in any of the 9 tissues sequenced. We grouped genes as control genes with no new exons and genes with new exons divided by whether the exon was included or excluded in a given tissue. We calculated the number of TSSs used in each gene in each tissue and considered genes that gained TSSs in mouse, genes that gained TSSs in rat, and genes with same number of TSSs in both species based on the numbers of TSSs for each species in each gene in each tissue, or when considering all tissues together. Gene expression in mouse was compared to that in rat by taking the ratio of expression in mouse to expression of the orthologous gene in the analogous tissue in rat.

Definition of new exon-proximal cleavage and polyadenylation sites

Polyadenylation sites were identified using available polyA-seq data from five mouse tissues (brain, liver, kidney, muscle, testis) (Derti et al., 2012). Only reads aligning to unique loci were retained and ends of reads within 25 nt of each other on the same strand were clustered. Polyadenylation sites were considered to be new exon-proximal cleavage and polyadenylation (nePCPA) sites if they were located within 2 kb upstream or downstream of a new exon, and as skipped exon-proximal cleavage and polyadenylation (sePCPA) sites if they were located within 2 kb upstream or downstream of skipped exons.

Effects on nascent and steady state RNA levels

Effects on transcription initiation should be reflected in nascent RNA, while effects on RNA stability would only be visible in steady state mRNA. In the *Tsku* gene, nascent RNA levels were reduced to a similar extent as steady state mRNA (Fig. 2d, Extended data Fig. 3b, Extended data Fig. 5a-d), in both sense and antisense orientations. For other genes studied here, *Stoml1* and *Gper1*, we also observed similar effects on nascent RNA in sense and antisense directions (Fig. 2c, Extended data Fig. 3b, Extended data Fig. 4a-c). Furthermore, the model invoking inhibition of PCPA involves U1 snRNP binding at a 5' splice site, but we observed increased gene expression from creation of a 3' splice site. Thus, our observations are consistent with splicing-dependent regulation of transcription initiation but not with models involving PCPA.

Data availability

The RNA-seq data from 9 tissues from mouse and rat associated with Merkin et al. (Merkin et al., 2012) is available at GEO (accession no. GSE41637). The Start-seq data from Scruggs et al. (Scruggs et al., 2015) is available at GEO (accession no. GSE62151), as well as the H3K4me3 data from Yu et al. (Yu et al., 2015) (accession no. GSE59896 and GSE59998). PolyA-seq data from five mouse tissues is available in Derti et al (Derti et al., 2012). Data of evolutionarily new exons is available in Merkin et al. (Merkin et al., 2015) as well as here in Supplementary Tables 1 and 2.

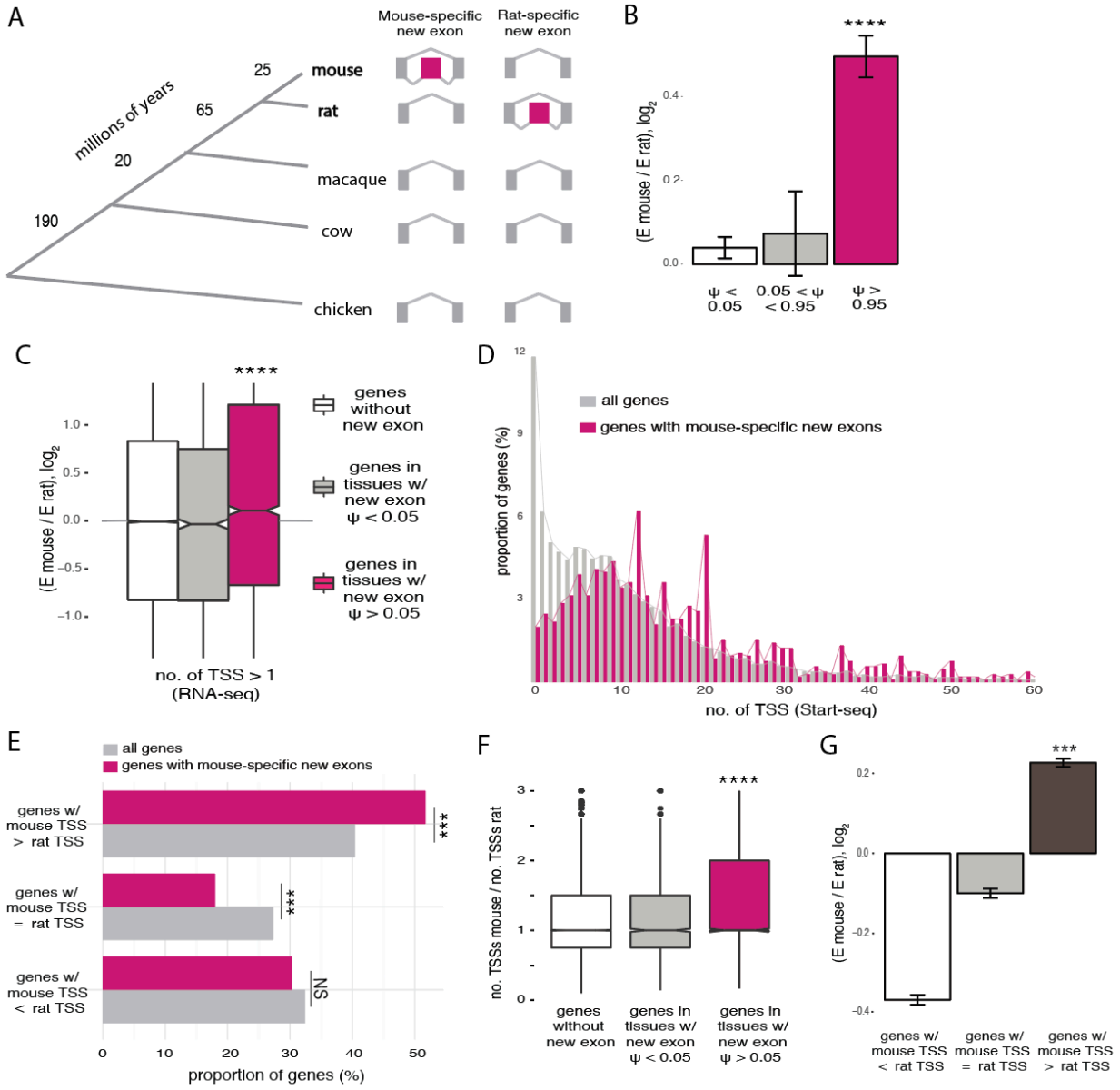

Figure 1. Splicing is associated with increased gene expression and usage of TSS.
a, A phylogenetic tree representing the main species used for dating evolutionarily new exons and approximate branch lengths in millions of years. The patterns of inclusion/exclusion used to infer mouse-specific new exons (n = 1089) and rat-specific new exons (n = 1517) are shown. **b**, Fold change in gene expression between mouse and rat in 9 tissues (brain, heart, colon, kidney, liver, lung, skeletal muscle, spleen and testes) for genes with mouse-specific new exons, binned by the PSI (ψ) value of the new exon in each tissue. **** $p < 0.0001$ by one-way ANOVA, Tukey post hoc test. **c**, Fold change in gene expression between corresponding tissues of mouse and rat in genes with multiple TSSs in mouse (no. of TSS > 1) for mouse control genes with no new exons (white), genes with mouse-specific new exons in tissues where inclusion of the new exon is not detected, PSI < 0.05 (grey), and genes with new mouse-specific exons in tissues were the exon is included, PSI > 0.05 (pink). *** $p < 0.001$ by one-way ANOVA, Tukey post hoc test. **d**, Distribution of the number of TSSs per gene using Start-seq data from murine macrophages for all genes expressed in mouse and genes with mouse-specific new exons. TSS

peaks located within 50 bp apart were merged. Genes with mouse-specific new exons have increased numbers of TSSs ($p < 2.2e^{-16}$ by Kolmogorov-Smirnov test). **e**, Proportion of genes that gained TSSs in mouse (mouse TSS > rat TSS), genes that lost TSSs in mouse (mouse TSS < rat TSS) and genes with same number of TSSs (mouse TSS = rat TSS) for all genes expressed in both species and genes with mouse-specific new exons. *** $p < 0.001$ by one-way ANOVA, Tukey post hoc test. **f**, Fold change in the number of TSSs used per gene between mouse and rat for 9 tissues, for mouse control genes with no new exons (white), genes with mouse-specific new exons in tissues where inclusion of the new exon is not detected, PSI < 0.05 (grey), and genes with mouse-specific new exons in tissues were the exon is included, PSI > 0.05 (pink). **g**, Fold change in gene expression between mouse and rat for genes that lost TSSs in mouse (white), genes with same number of TSSs in both species (grey) and genes that gained TSSs in mouse (brown).
See also Figure S1.

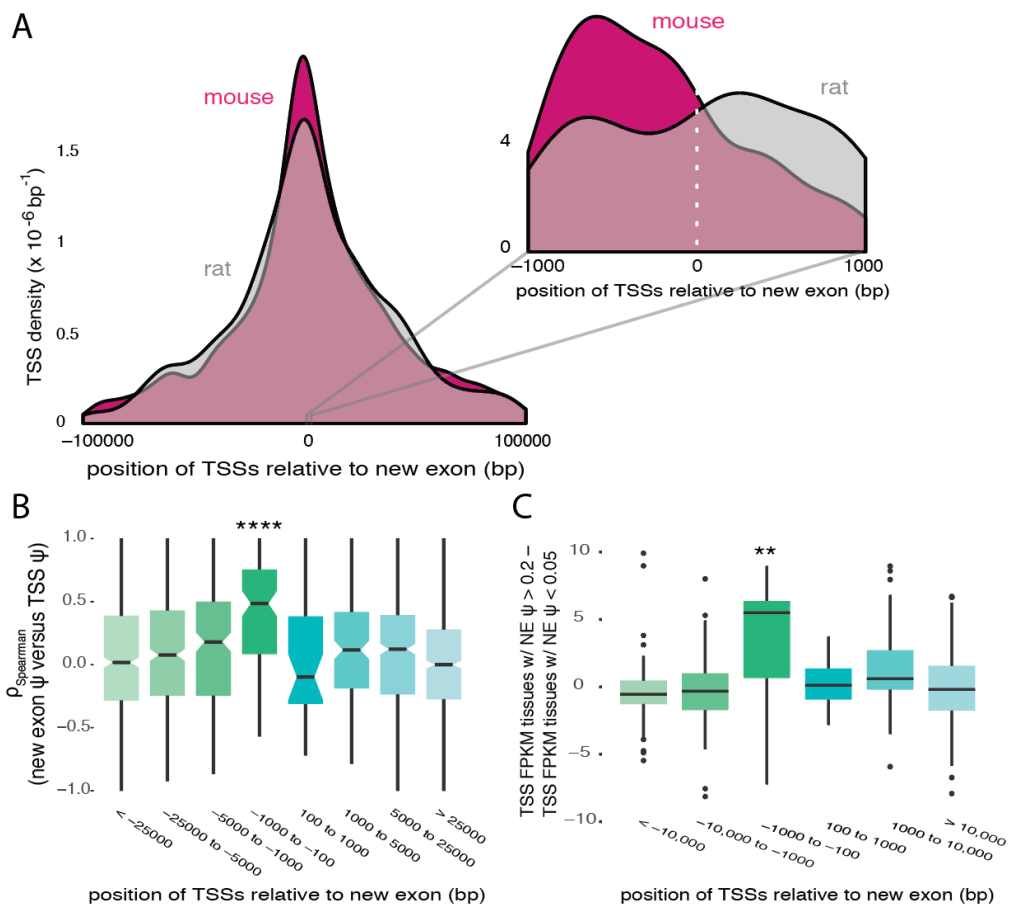

Figure 2. TSSs arise proximal and upstream of new exons.
a, Histogram of TSS locations in mouse (pink) and rat (grey) for genes with mouse-specific new exons in all 9 tissues, centered on start of mouse new exon or homologous genomic position in rat. Inset shows zoom-in of locations within 1 kb of new exon. Distributions were smoothed with kernel density estimation by ggplot2 with default parameters. **b**, Spearman correlations between TSS PSI and new exon PSI across mouse tissues, for TSSs binned by position relative to mouse-specific exon. **c**, Difference in expression (in units of fragments per kilobase of exon per million mapped reads, FPKM) in mouse tissues for transcripts including TSSs in tissues where new new exon is moderately or highly included (PSI > 0.2) versus tissues where new exon is excluded (PSI < 0.05), grouped by TSS location relative to new exon.
See also Figure S2.

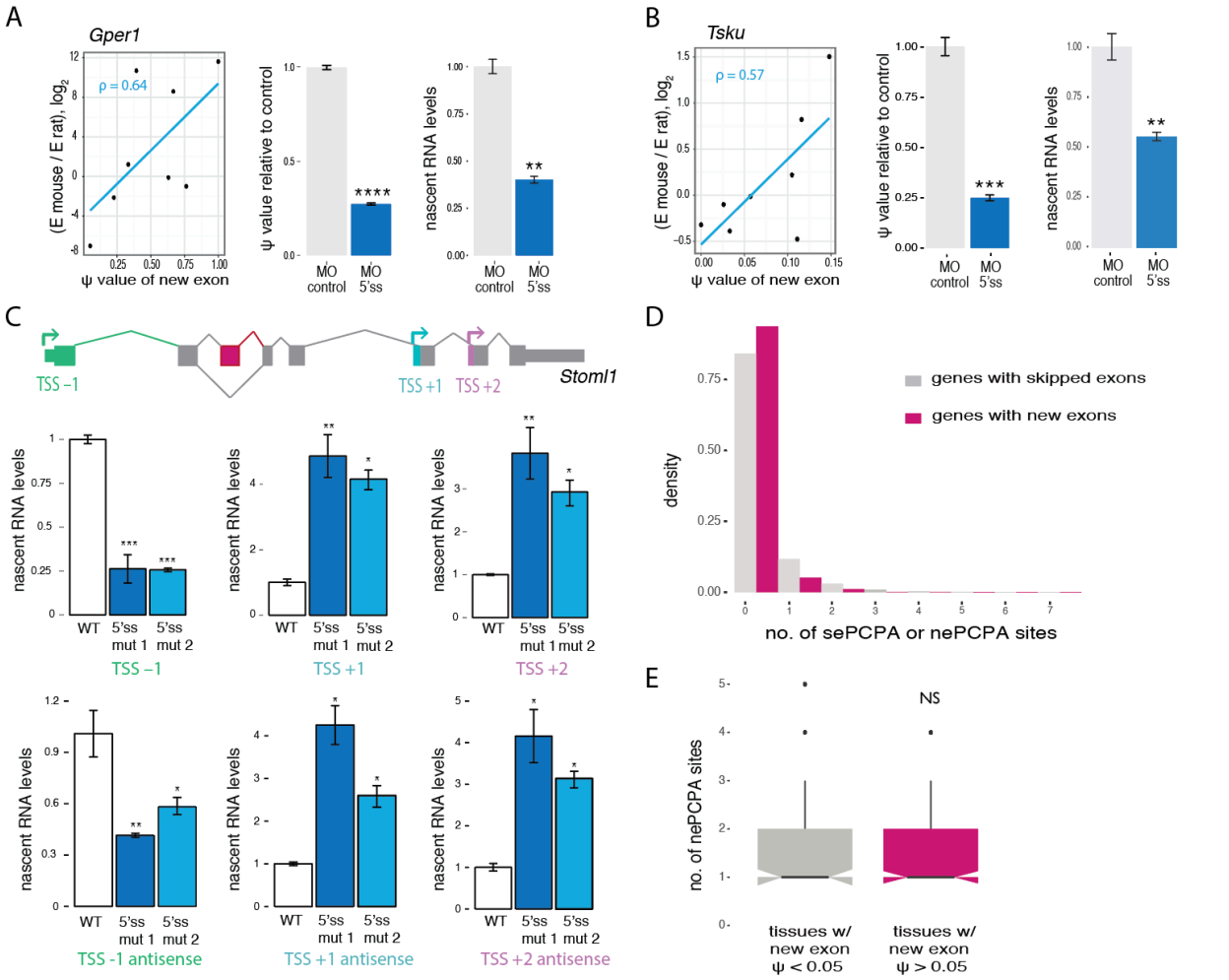

Figure 3. Manipulation of exon splicing impacts upstream transcription initiation.
a, (Left) Relationship between fold change in gene expression between mouse and rat and new exon PSI value across 9 tissues for *Gpr30* gene. (Right) qRT-PCR analysis of fold change in new exon PSI value (middle) and gene expression (right) in nascent RNA metabolically labeled for 10 minutes with 5-ethynyl uridine, following treatment of NIH3T3 cells with MO targeting new exon 5' splice site relative to control treatment. Mean ± SEM of displayed distributions, n=3 biological replicates. ** $p < 0.01$, *** $p < 0.001$, ****$p < 0.0001$ by one-way ANOVA, Tukey post hoc test. **b**, As in (**a**) for mouse *Tsku* gene. **c**, Fold change in nascent sense (top) and antisense (bottom) RNA levels of *Stoml1* in CAD cells measured by qPCR in RNA metabolically labeled for 10 minutes with 5-ethynyl uridine and normalized using housekeeping genes GAPDH, HPRT and HSPCB. Wild type cells in white and CRISPR-cas cells with mutations in the 5' splice site of the new exon in blue. Mean ± SEM of displayed distributions, n=3 independent experiments. A schematic diagram of *Stoml1* showing the exon-intron organization is shown at the top. **d**, Distribution of the number of polyadenylation sites used per gene located 2 kb upstream/downstream of a control set of mouse genes with skipped exons (grey, sePCPA) and genes with mouse-specific new exons (pink, nePCPA). Definitions of sePCPA and nePCPA are provided in Methods. **e**, Distribution of the number of polyadenylation sites used 2 kb upstream/downstream of new exons per gene in tissues new exon is excluded (PSI < 0.05, grey) and tissues with inclusion of new exons (PSI > 0.05, pink) for genes with new exons and at least one nePCPA. Distributions are not significantly different by Kolmogorov-Smirnov test.
See also Figure S3

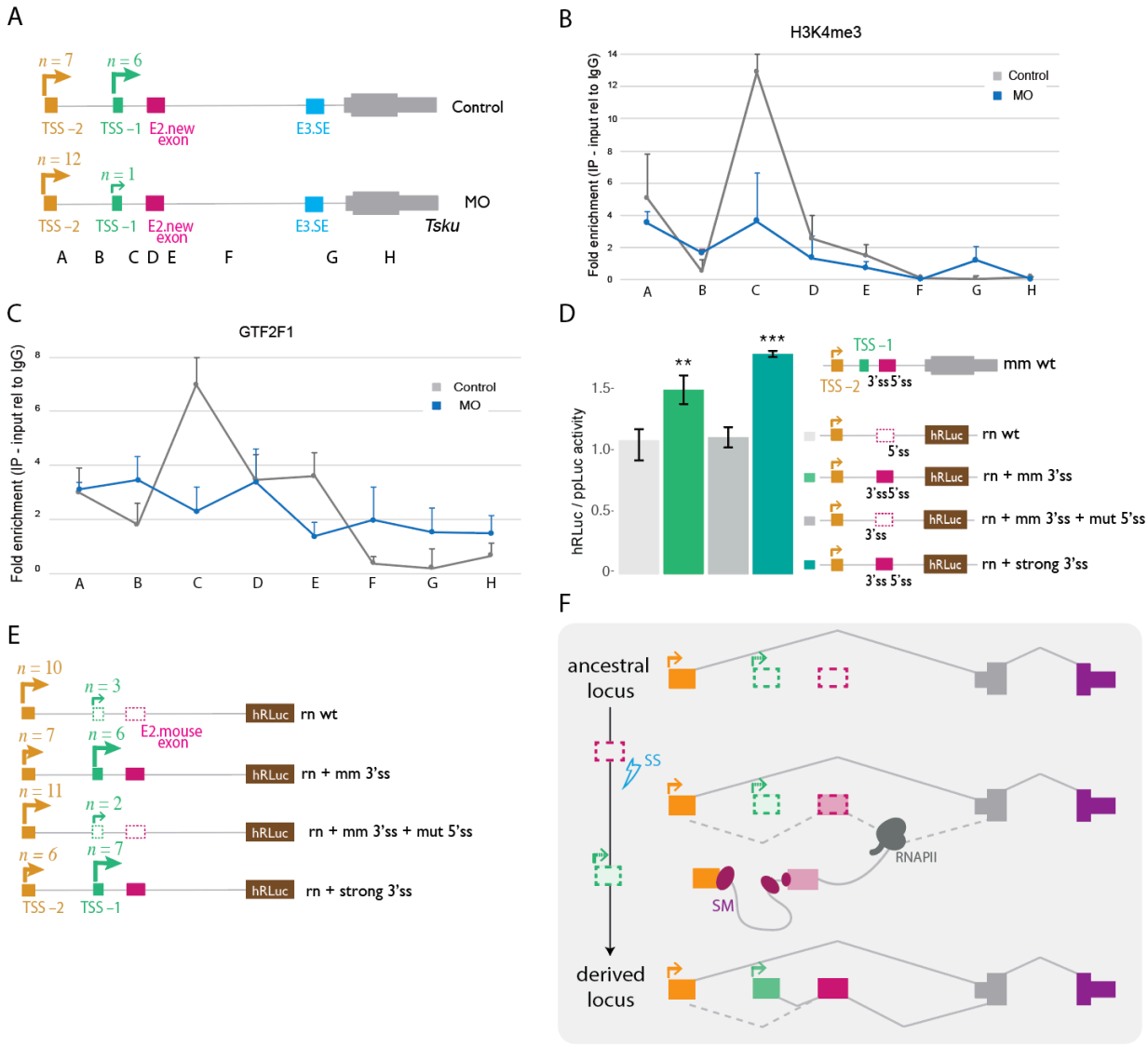

Figure 4. Creation of a new splice site activates the use of a cryptic promoter nearby

a, Schematic of 5' RACE products and quantity of clones obtained for each TSS in control NIH3T3 cells and cells transfected with MO targeting the 3' and the 5' splice sites of the new exon in *Tsku*, *n*=2 biological replicates. **b,c**, ChIP-PCR analysis of H3K4me3 (**b**) and GTF2F1 (**c**) in *Tsku* gene in NIH3T3 cells for regions indicated in (**a**). Mean ± SD of two independent immunoprecipitations normalized to input and mean value for control IgG antibody are shown. Data shown for control cells (grey) and cells treated with MOs targeting the 3' and 5' splice sites of the new exon (blue). **d**, Luciferase activity of in HeLa cells transfected with the hybrid constructs of the *Tsku* gene (right). Promoter activities of the corresponding constructs (corrected for transfection efficiency) are presented as fold increase of *Renilla* luciferase activity relative to firefly luciferase activity (encoded on the same plasmid). Mean ± SD for *n*=3 independent experiments. **e**, Schematic of 5′ RACE analysis showing TSS usage and quantity of clones obtained in NIH3T3 mouse cells transfected with plasmids expressing the corresponding rat *Tsku* mutants. **f**, Model showing how creation of a splice site (ss) during evolution promotes the gain of an new internal exon and an upstream TSS. In the model, exon recognition by the splicing machinery (SM) in

transcripts from the distal promoter activates TSS(s) located proximal and upstream of the exon. At steady state in the derived locus, inclusion of the new exon in transcripts initiating from both TSSs reinforces initiation at the proximal upstream TSS.
See also Figure S4.

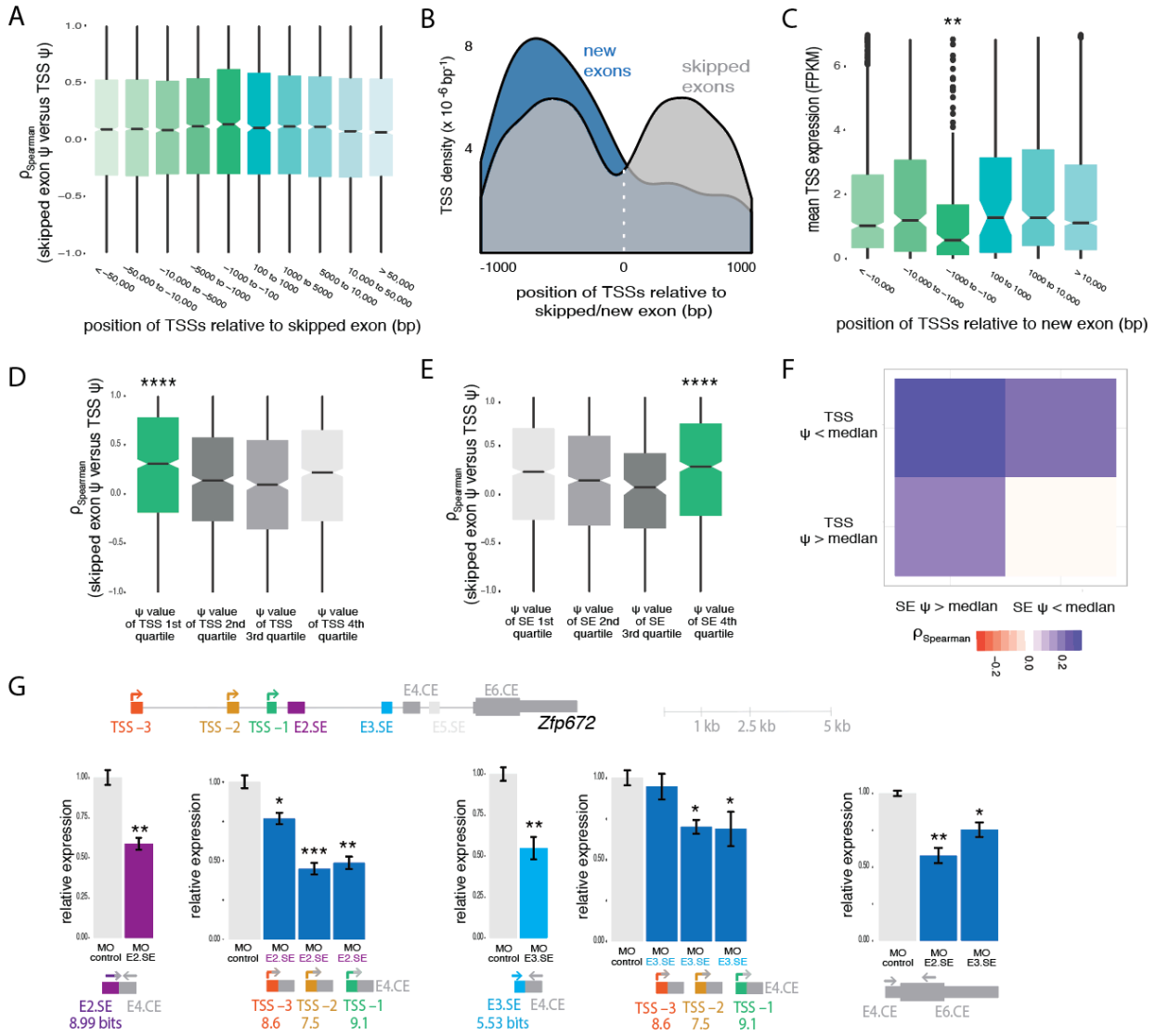

Figure 5. Efficiently spliced exons activate usage of weak proximal TSSs

a, Spearman correlations between TSS PSI (n = 49911) and skipped exons (SE, n = 13491) PSI in the same gene across mouse tissues for all TSSs used in genes with SEs, binned by genomic position relative to the SE. **b**, Comparison of distributions of TSS positioning in 9 tissues between genes with mouse-specific new exons (pink) and genes with SEs in mouse (grey). The 0 is set for the start coordinate of the new exon/skipped exon. Distributions were smoothed with Kernel density estimation. **c**, Expression of alternative first exons (AFE) for all TSSs in genes with mouse-specific new exons in tissues where the new exon is included (PSI > 0.05), binned by position relative to the new exon. **d**, Spearman correlation between TSS PSI and SE PSI in the same gene across mouse tissues for TSSs within 1kb upstream of the SE, binned by quartiles of mean TSS PSI. **e**, Same as (**d**) but binned by quartiles of mean SE PSI. **f**, Heat map showing the median Spearman correlation between TSS PSI and SE PSI in the same gene across mouse tissues for SEs with at least one TSS located upstream, in four groups, according to whether the mean TSS PSI (across tissues) and the mean SE PSI were greater than or less than the corresponding median values. **g**, Exon-intron organization of mouse *Zfp672* gene. qRT-PCR analysis of expression of

Zfp672 in NIH3T3 cells normalized to expression of housekeeping genes HPRT and HSPCB. Data for control cells and cells treated with MO targeting the indicated splice sites (E4.CE and E6.CE, E5.SE is not included in NIH3T3 cells). Inclusion levels of the skipped exons, as well as levels of exon-excluding transcripts from the alternative TSSs (TSS –3, TSS –2, TSS –1) and total gene expression are shown. Scores of 5' splice sites of skipped exons and first exons are listed in bits. Mean ± SEM of displayed distributions for $n=3$ independent experiments.
See also Figure S5.

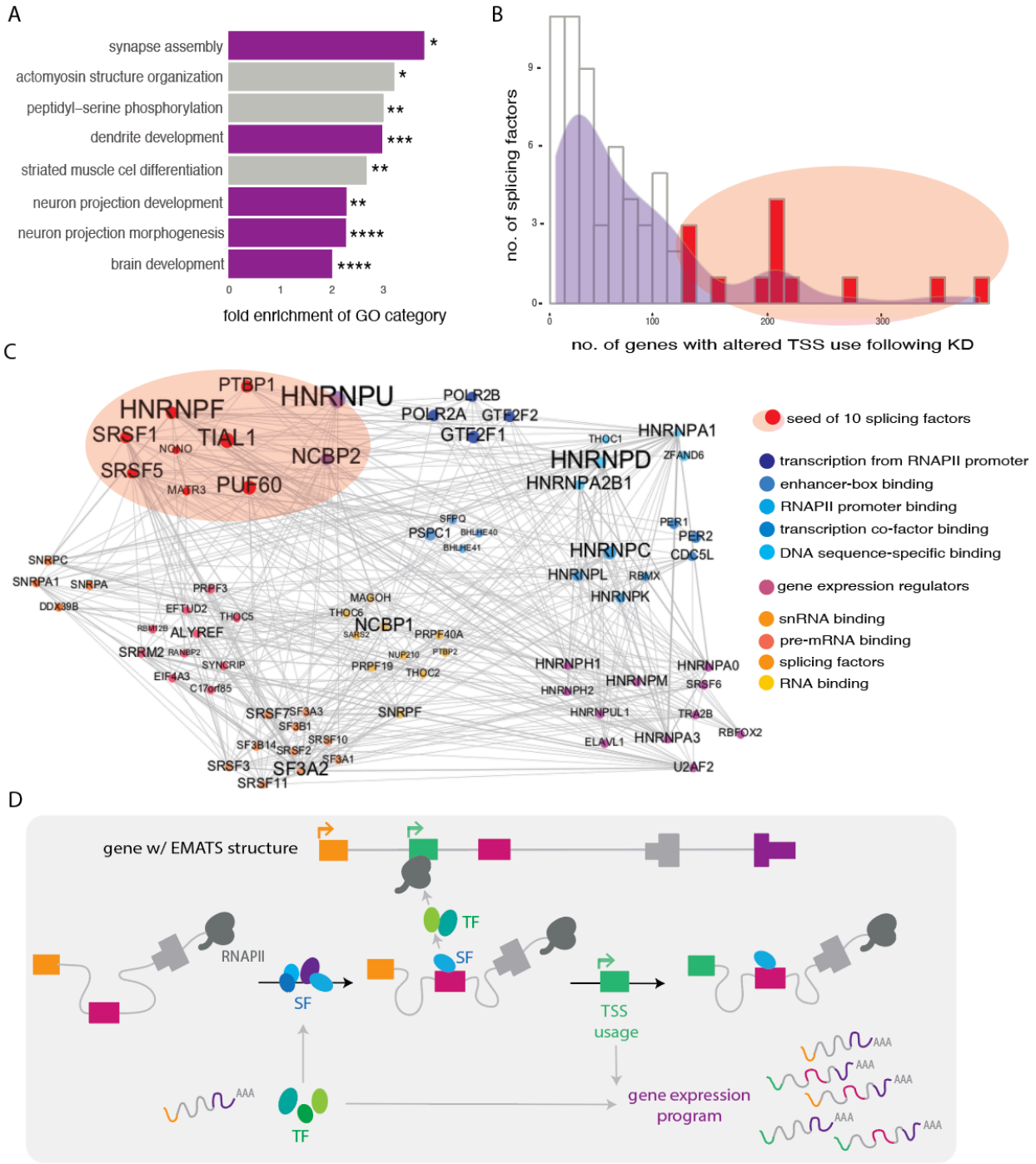

Figure 6. A subset of splicing factors impact TSS use and interact with transcription machinery.

a, Gene Ontology analysis of 1777 genes with the strongest EMATS effect. Fold enrichments shown for the most significant categories with asterisk indicating adjusted *p*-values and color indicating relation to neuron development. **b**, Histogram of number of genes with significant changes in alternative first exon usage following depletion of 67 splicing factors. Mean between two cell lines (HepG2 and K562) is plotted for each RBP (top ten splicing factors with greatest number of changes shown in red). **c**, Protein

interaction network for the top 10 splicing factors from (b), colored by Gene Ontology category. Nodes represent proteins and links represent the interactions among them. Node size and label size are proportional to protein connectivity. The selected 10 splicing factors in red primarily interact with other 65 proteins, generating a network with 75 nodes and 424 edges, a diameter of 5, an average weighted degree of 6.133, an average clustering coefficient of 0.386 and an average path length of 2.103. Protein interaction data are from STRING (Szklarczyk et al., 2015) (using experimentally determined, database annotated, homology, gene fusion and automated text mining interactions). Networks were built using Gephi (http://gephi.org). **d**, Model for the role of EMATS in dynamic gene expression programs. During development, transcription factors (TF) influence gene expression directly and indirectly by regulating splicing factors. In genes with EMATS structure, exon recognition by SF in transcripts from the distal promoter recruits TF and the transcription machinery, creating a high local concentration of RNAPII and core TF (consistent with constraints on promoter-exon distance) that activates weak TSS(s) located proximal and upstream of the exon. In differentiated cells, changes in alternative splicing with concomitant increased usage of TSSs up-regulate the expression levels of EMATS genes and their alternative isoforms.

See also Figure S6.

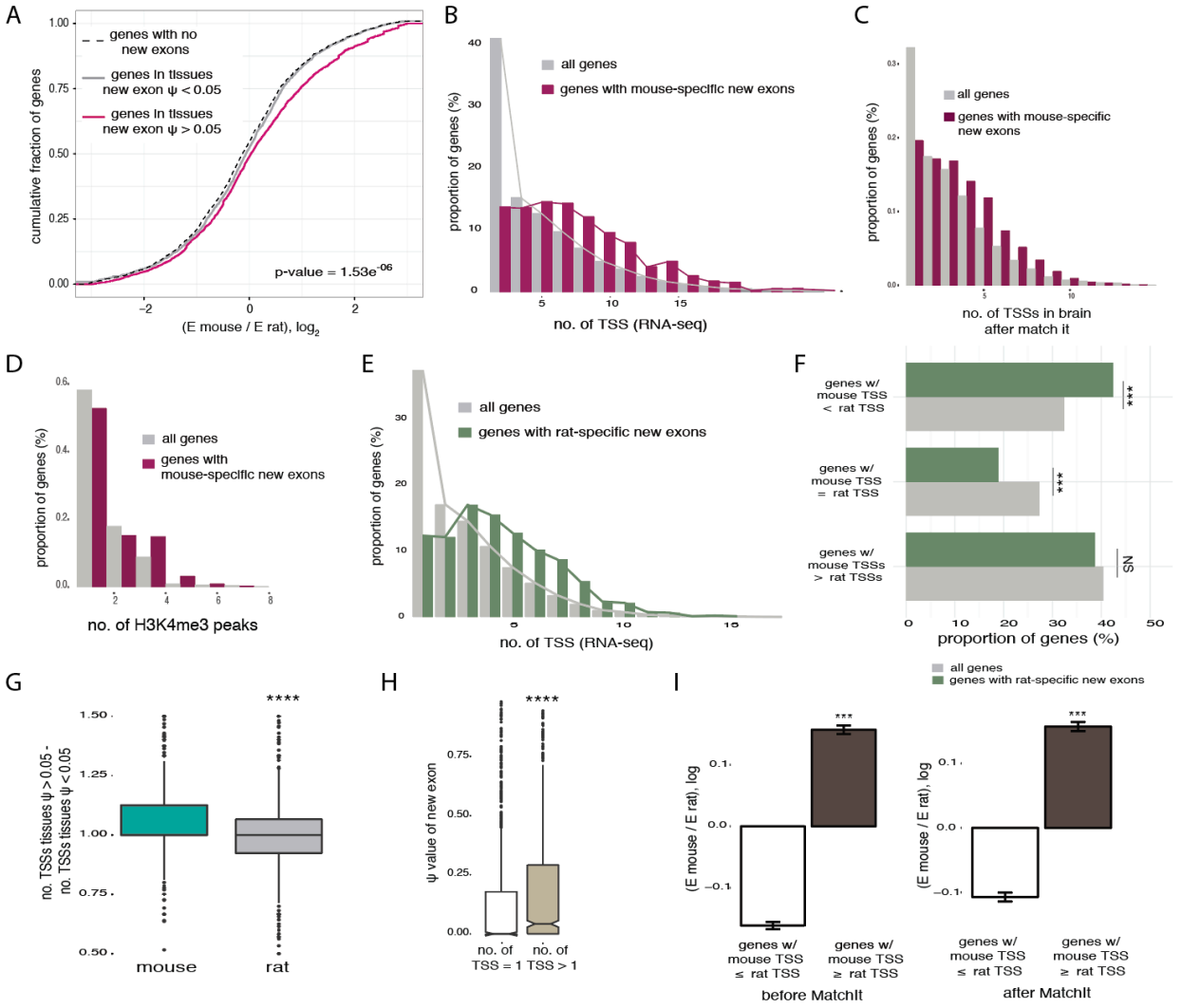

Figure S1. Related to Figure 1

a, Fold change in gene expression between mouse and rat for mouse control genes with no evolutionarily new exons (black, dotted line), genes with mouse-specific new exons in tissues where inclusion of the new exon is not detected, PSI < 0.05 (grey), and genes with new mouse-specific exons in tissues were the exon is included, PSI > 0.05 (pink). Statistical significance by Mann-Whitney U test is indicated between genes with mouse-specific new exons in tissues with PSI < 0.05 and tissues with PSI > 0.05. **b**, Distribution of the number of TSSs per gene using RNA-seq data across multiple species and multiple tissues, for all genes expressed in mouse and genes with mouse-specific new exons. Distributions are significantly different by Kolmogorov-Smirnov test. **c**, Distribution of the number of TSSs per gene in the mouse brain using RNA-seq data, for all genes expressed in mouse and genes with mouse-specific new exons, after matching the distribution of gene expression levels between the two groups using the MatchIt package in R. Distributions remain significantly different by Kolmogorov-Smirnov test after matching the gene expression levels between the groups, demonstrating that, independent of gene expression, genes with mouse-specific new exons are enriched in multiple TSSs. **d**, Distribution of the number of H3K4me3 peaks per gene using H3K4me3 ChIP-seq data for all genes expressed in mouse (grey) and genes with mouse-specific new exons (dark red). Distributions are significantly different by Kolmogorov-Smirnov

test. Genes with mouse-specific new exons are enriched in H3K4me3 peaks. **e**, Distribution of the number of TSSs per gene for all genes expressed in rat (grey) and genes with rat-specific new exons (green). Genes with rat-specific new exons are enriched in multiple TSSs (by Kolmogorov-Smirnov test). **f**, The proportion of genes with fewer TSSs in mouse (genes w/ mouse TSS < rat TSS), genes with the same number of TSSs in both species (genes we mouse TSSs = rat TSS), and genes that have more TSSs in mouse, for all genes expressed in both species (gray) and for genes with rat-specific new exons (green). Statistical significance indicated by asterisks corresponds to one-way ANOVA, Tukey post hoc test (NS = not significant). **g**, Fold change in the number of TSSs used per gene between tissues where mouse-specific exons are included (PSI > 0.05) and excluded (PSI < 0.05), for mouse genes and for the same tissues in rat. Evolutionary gain of internal exons and of TSSs are associated, but only in those tissues where new exons are included. **h**, Distribution of PSI values of new exons binned by the number of TSSs used in the same gene, for 9 tissues pooled together in mouse. **i**, Distribution of the fold change in gene expression levels between mouse and rat for genes with fewer or same number of TSSs used in mouse than rat (white) and genes with more TSSs used in mouse than rat (brown), before (left panel) and after (right panel) balancing the distribution of gene expression levels in mouse between the groups by MatchIt. Evolutionarily change in gene expression remain significantly different when balancing gene expression levels in mouse between groups, demonstrating that, independently of gene expression levels in one species, genes gaining TSSs in mouse have increased gene expression levels compared to rat.

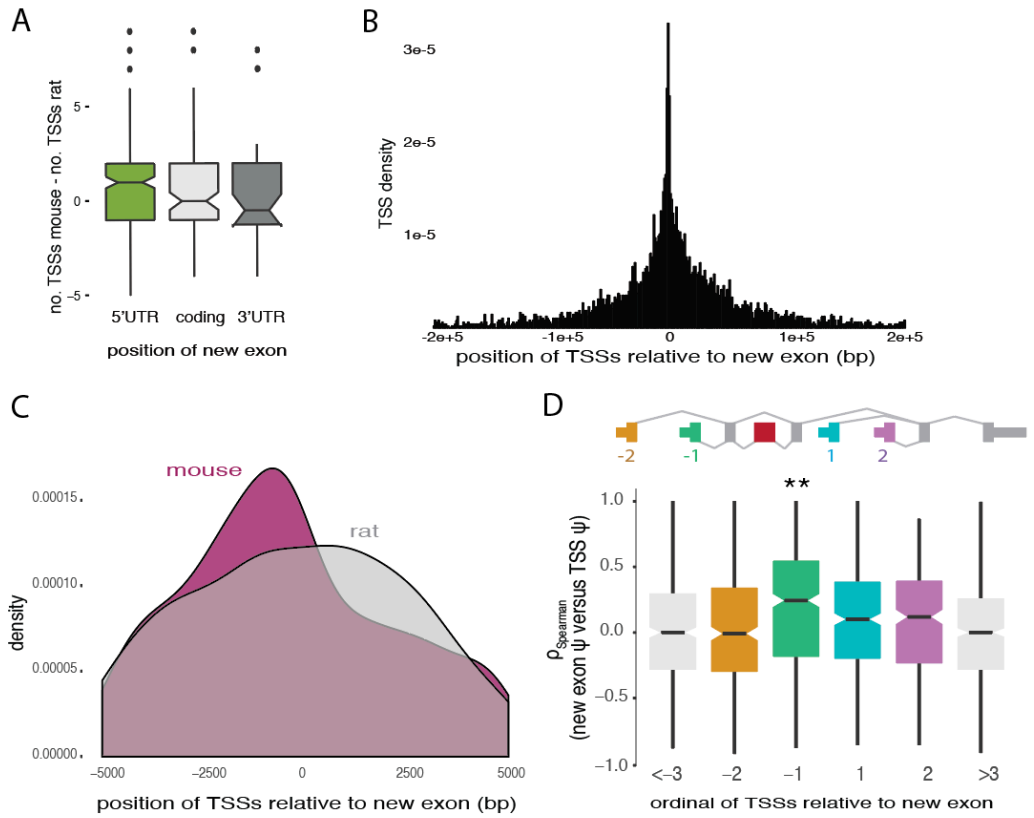

Figure S2. Related to Figure 2

a, Ratio between number of TSSs used in mouse and in rat for genes with mouse-specific evolutionarily new exons, binned by location of the exon within the gene. Increased number of TSS is associated with new exons located in the 5' UTR. **b**, TSS position relative to the start coordinate of the new exon in genes with mouse-specific new exons, for all TSSs used in 9 tissues in mouse. **c**, Comparison of distributions of TSS positions within 5 kb upstream and downstream of new exons between mouse (dark red) and rat (grey) for genes with mouse-specific new exons in all 9 tissues. The 0 position is at the start coordinate of the mouse-specific new exon (mouse) or at the location homologous to this position (rat). Distributions were smoothed with Kernel density estimation. **d**, Spearman correlations between the usage of a particular TSS and the PSI value of the new exon across multiple tissues for all TSSs used in genes with mouse-specific new exons, binned by their relative position to the new exon with negative numbers for TSSs located upstream of the new exon and positive numbers for TSSs located downstream of the new exon.

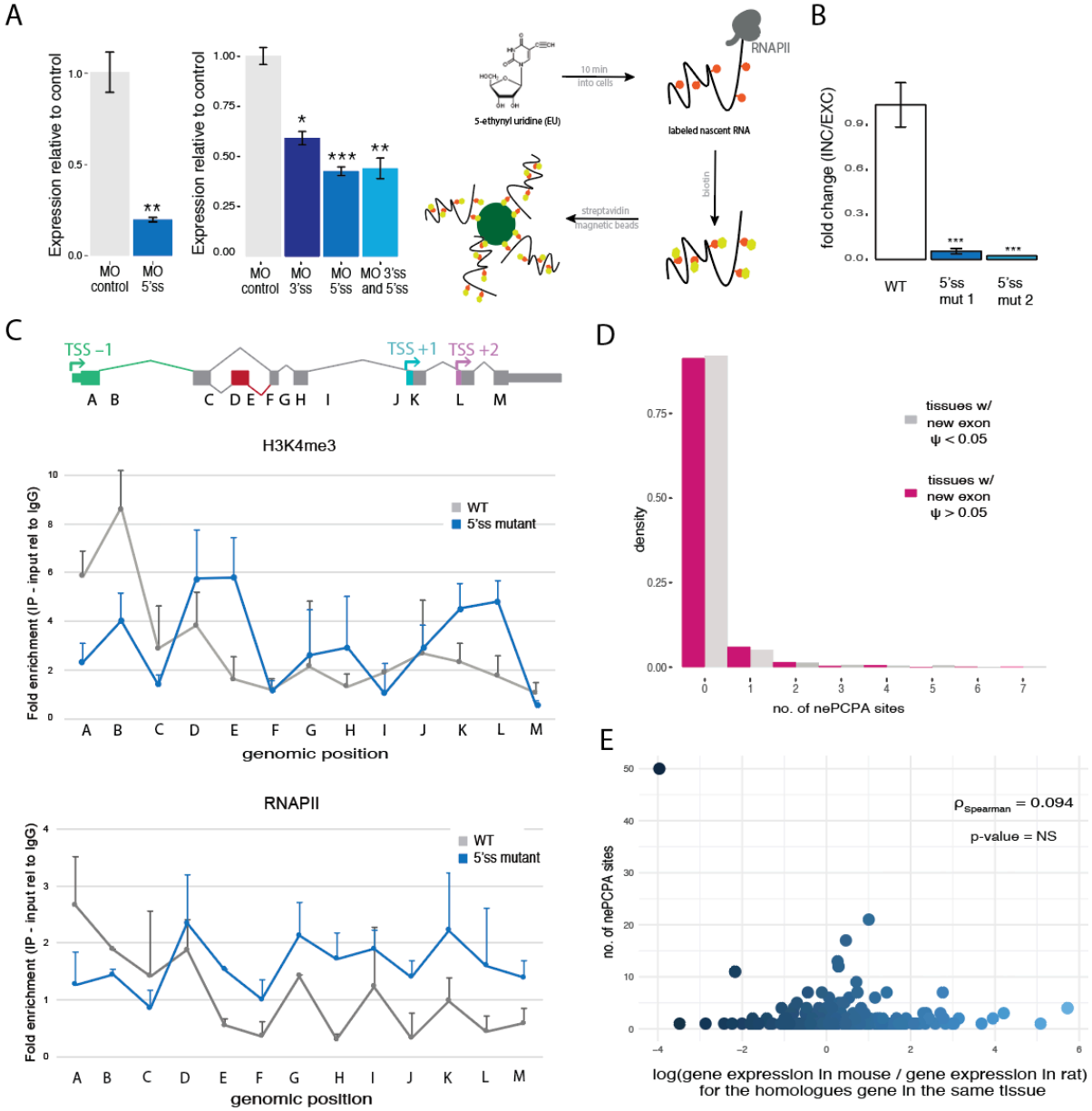

Figure S3. Related to Figure 3

a, A diagram representing the technique used to label nascent RNA with 5-ethynyl uridine and pull down the nascent RNA with the click-it method. Fold change in RNA levels of *Gpr30* (left) and *Tsku* (right) in NIH3T3 cells measured by qPCR. NIH3T3 cells were transfected with 20 µM morpholino (MO) targeting the 5' splice site of the new exon in Gpr30 or 20 µM MO targeting the 3' and/or the 5' splice sites of the new exon in *Tsku* for 24 h. Mean ± SEM of displayed distributions, n=3 biological replicates. Statistical significance indicated by asterisks corresponds to one-way ANOVA, Tukey post hoc test. **b**, Relative exon inclusion / exon exclusion of the mouse-specific new exon in *Stoml1* gene is shown, measured by qPCR of nascent RNA in wild type CAD cells and cells with CRISPR/cas-mediated mutations in the 5' splice site of the new exon. Mean ± SEM is shown for n=3 biological replicates. Statistical significance indicated by asterisks corresponds to one-way ANOVA, Tukey post hoc test. **c**, H3K4me3 and RNAPII profiles in Stoml1 gene in CAD cells determined by ChIP assay followed by qPCR with the regions

indicated in the top panel. Values of two independent immunoprecipitations normalized to input and the mean value for control IgG antibody are shown for each region. Wild type cells (grey) and cells with CRISPR/cas-mediated mutations in the 5' splice site of the new exon (blue) are shown. d, Distribution of the number of polyadenylation sites used 2 kb upstream/downstream of new exons per gene in tissues new exon is excluded (PSI < 0.05, grey) and tissues with inclusion of new exons (PSI > 0.05, pink) for all genes with new exons. Distributions are not significantly different by Kolmogorov-Smirnov test. e, Scatter plot showing the relationship between the number of nePCPA sites and the fold change in gene expression levels between mouse and rat. These variables are not significantly associated by Spearman correlation test. Polyadenylation sites for 5 tissues in mouse were analyzed using polyA-seq data (Derti et al., 2012).

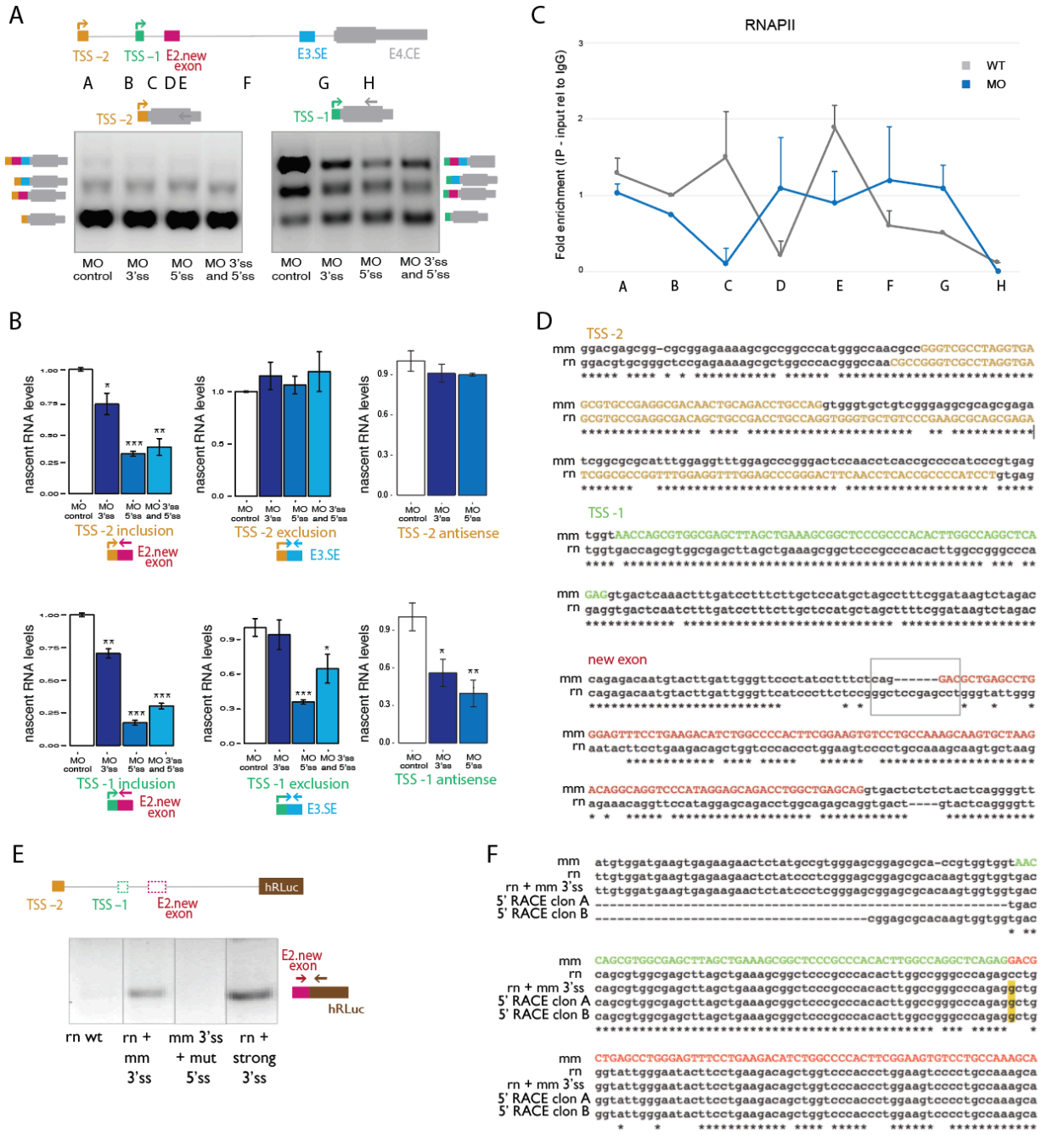

Figure S4. Related to Figure 4

a, Isoform expression for *Tsku* gene in NIH3T3 cells, measured by RT-PCR with primers targeting TSS –2 (top), TSS –1 (bottom) and exon E4.CE. Cells were transfected for 24 h with 20 µM MO targeting the 3' and/or 5' splice sites of the new exon as indicated. **b**, Fold change in inclusion and exclusion levels of the mouse-specific new exon in *Tsku* gene and antisense transcription levels from both TSSs measured by isoform-specific qPCR of nascent RNA with primers illustrated by arrows. Exon exclusion levels are measured from both alternative first exons to the following skipped exon downstream the mouse-specific new exon. NIH3T3 cells were transfected with control MO or MO targeting the 3' and/or 5' splice sites of the new exon. A decrease in the inclusion level of transcripts starting at TSS –2 is compensated by an increase in the exclusion levels, while total level of transcripts starting at TSS –1 is reduced by MO

treatment. Mean ± SEM of displayed distributions, n=3 biological replicates. Statistical significance indicated by asterisks corresponds to one-way ANOVA, Tukey post hoc test. **c**, RNAPII profile in *Tsku* gene in NIH3T3 cells determined by ChIP assay followed by qPCR with the regions indicated in panel **a**. Values of two independent immunoprecipitations normalized to input and the mean value for control IgG antibody are shown for each region. NIH3T3 cells transfected for 24 hours with 20 μM control MO or MO targeting both 3' and 5' splice sites of the new exon. **d**, Alignments and identity between mouse (mm) and rat (rn) of the DNA sequence of TSS –2, TSS –1 and mouse-specific new exon in the *Tsku* gene. **e**, Splicing patterns of the *Tsku* gene in HeLa cells transfected with the hybrid constructs shown in Figure 4d. The creation of the mouse 3' splice site (rn + mm 3'ss) or of a stronger 3' splice site (rn + strong 3'ss) of the mouse-specific new exon in the rat sequence promotes the inclusion of the mouse-specific new exon in the rat context only when the wild type 5' splice site is maintained (but not in the mm 3'ss + mut 5'ss construct). **f**, Sequence of the 5' end of *Tsku* transcripts generated by 5' RACE in HeLa cells transfected with rat *Tsku* constructs with the 3' splice site of the mouse-specific new exon (5' RACE clone A, clone B) aligned to the mouse sequence (mm) and the rat sequence (rn). For 80% of the sequenced transcripts, the 5' end mapped 1 bp upstream of the position of mouse TSS –1 (clone A), while in the remainder the 5' end mapped 19 bp upstream of (clone B).

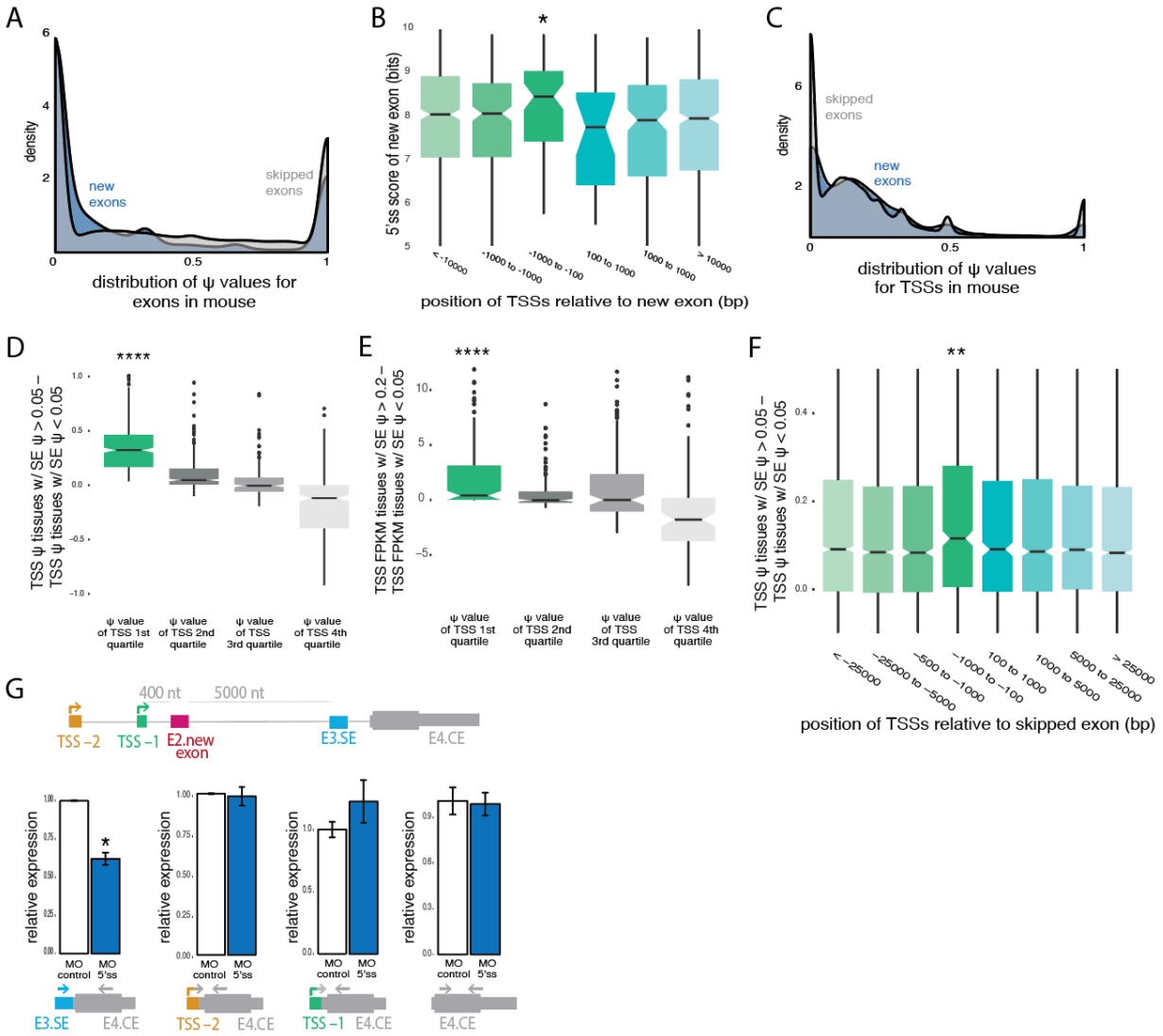

Figure S5. Related to Figure 5

a, Distribution of the PSI values of mouse-specific new exons (blue) and SE (grey) in mouse across 9 tissues. **b**, Distribution of 5' splice site scores of mouse-specific new exons, binned by the relative position to the next upstream TSS used in the same gene. 5' splice site scores were calculated using MaxEntScan (Yeo and Burge, 2004). **c**, Distribution of the PSI values of first exons associated with TSSs in genes with mouse-specific new exons (blue) and in genes with SEs in mouse (grey). **d**, **e**. Difference in TSS usage based on PSI value (**d**) and FPKM (**e**) in tissues with high versus low inclusion of skipped exons (SE), in the same gene across multiple tissues for proximal and upstream TSSs (within 1kb upstream the SE) used in genes with SEs in mouse, binned by quartiles of PSI values of the TSSs. **f**, Difference between TSS PSI values in tissues with high versus low inclusion of skipped exons (SE), for all weak TSSs (bottom quartile) used in genes with skipped exons in mouse, binned by their position relative to the SE. **g**, Fold change in inclusion (far left), exclusion levels (left center, right center) and total levels (far right) of the skipped exon in *Tsku* gene (E3.SE) from both TSSs measured by isoform-specific qPCRs with primers shown by arrows. Exclusion levels were measured from both alternative first exons relative to the next constitutive exon downstream of the skipped exon. NIH3T3 cells were transfected with control MO or MO targeting the 5' splice site of E3.SE. Mean ± SEM of displayed distributions, $n=3$

biological replicates. Statistical significance indicated by asterisks corresponds to one-way ANOVA, Tukey post hoc test.

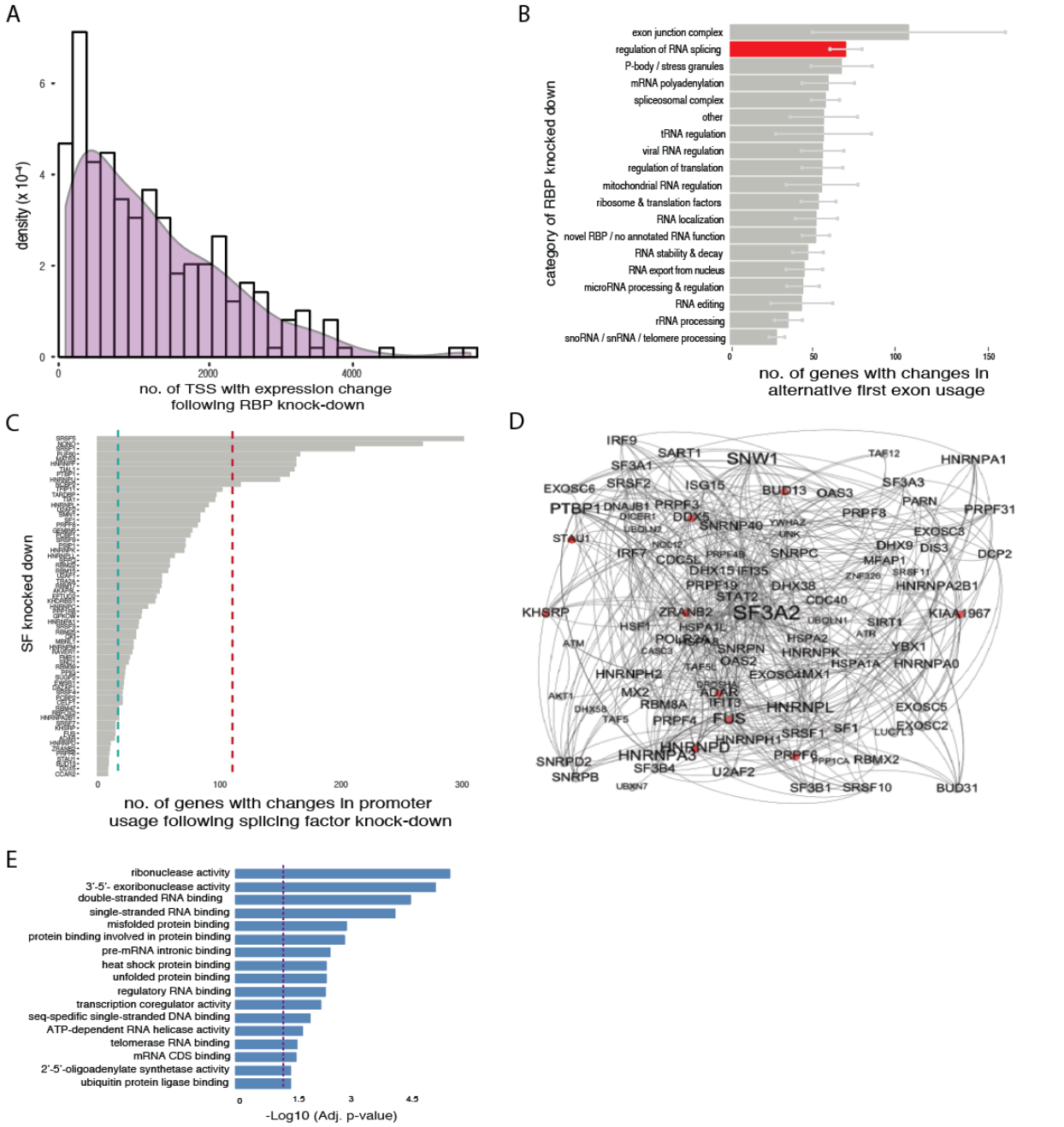

Figure S6. Related to Figure 6

a, Histogram and smoothed density of number of TSSs with significant expression change following depletion of each of 250 RNA binding protein genes. Mean two cell lines (HepG2 and K562) is plotted for each RBP. **b**, Distribution of the number of genes with significant changes in promoter usage associated with depletion of 250 RBPs, binned by Gene Ontology Biological Process categories of RBPs. Mean ± SEM between all RBPs in each GO category for two cell lines (HepG2 and K562) is plotted. **c**, Number of genes with significant difference in promoter usage associated with depletion of 67 splicing factors (SF). The red line indicates the cutoff for the top ten splicing factors driving the largest changes in promoter usage, while the green line indicates the cutoff for bottom ten control splicing factors driving

the fewest changes in promoter usage. **d**, Protein interaction network for 10 control splicing factors driving the fewest changes in promoter usage. The control 10 splicing factors in red primarily interact with 88 other proteins, generating a network with 98 nodes and 410 edges, a diameter of 3, an average weighted degree of 4.29, an average clustering coefficient of 0.43 and an average path length of 1.32. Nodes represent proteins and links represent the interactions among them. Node size and label size is proportional to the protein connectivity (number of interactions a protein establishes with others). Protein interaction data were collected from STRING (Szklarczyk et al., 2015) and networks were built using Gephi (http://gephi.org). **e**, Gene Ontology analysis of 88 proteins included in the interaction network shown in (**e**), excluding the 10 seed control splicing factors. Adjusted p-values shown for the most significant categories with dotted line indicating adjusted p-value < 0.05.

Table S1. Related to Figure 1

Mouse genes with new internal exons and their rat homologs. Columns A and H show the IDs for homologous mouse and rat genes with mouse-specific new exons, column B shows the locus of the new exon in mouse while column D shows the position of the new exon in the gene. Columns F and I show the average gene expression levels in brain for 3 individuals in mouse and rat, respectively. Column G shows the average PSI values in the mouse brain for 3 individuals.

Table S2. Related to Figure 1

Numbers of TSSs used in mouse and rat genes. Columns A and B show the IDs for homologous mouse and rat genes. Columns C and D show the numbers of TSSs used in mouse and rat, respectively, in a specific gene, pooling all nine sequenced tissues together.

Table S3. Related to Figure 5

	mouse SE, TSS	SE ψ > median	TSS ψ < median	TSS located upstream	TSS < 2kb upstream
no. of SE	49488	24744	13237	9621	3333
no. of TSS	58095	37266	18633	9510	2991
no. of SE-TSS pairs	223568	103801	42528	21326	4284
no. of genes	13491	9363	4973	3833	1777

Column A shows the number of SE expressed in the nine tissues sequenced in mouse and the number of genes in which they are distributed. Column B shows the number of these SE in which the average of PSI values across tissues is above the median of all SE and columns C the TSS paired with those SE with an average PSI across tissues below the median. Columns D and E reflect the subset of SE-TSS pairs and genes from previous columns in which the TSS is located upstream or proximal and upstream of the SE.