
Global impact of unproductive splicing on human gene expression

In the format provided by the
authors and unedited

Supplementary Information

Supplementary Notes

[Supplementary Note 1: Contribution of alternative splicing coupled with nonsense-mediated decay to post-transcriptional regulation of gene expression levels](#)

[Supplementary Note 2: Splicing of unproductive isoforms is upregulated in splicing factors and downregulated in translation factors](#)

Supplemental Figures

Supplementary Methods

[Illumina short read RNA-sequencing data](#)

[Histone modification ChIP-seq and CUT&Tag data](#)

[Molecular trait quantification](#)

[Classification of unannotated splice junctions](#)

[eQTL calling on GTEx gene expression data](#)

[Oxford Nanopore Technologies long read RNA-sequencing data](#)

[Scoring splice site strengths with MaxEntScan](#)

[Ranking of NMD junctions by contribution and entropy calculation](#)

[Analysis of alternative splicing and symmetry of cassette exons](#)

[Measurements of unproductive junctions in Illumina short-read data](#)

[\$\pi\$ 1 sharing of eQTLs between RNA-seq datasets](#)

[\$\pi\$ 1 sharing of eQTLs and hQTLs at TSS](#)

[Colocalization of molQTLs](#)

[Enrichment of genomic annotations amongst QTLs](#)

[Gene set enrichment and characteristics of genes with risdiplam-induced exons](#)

References

Supplementary Notes

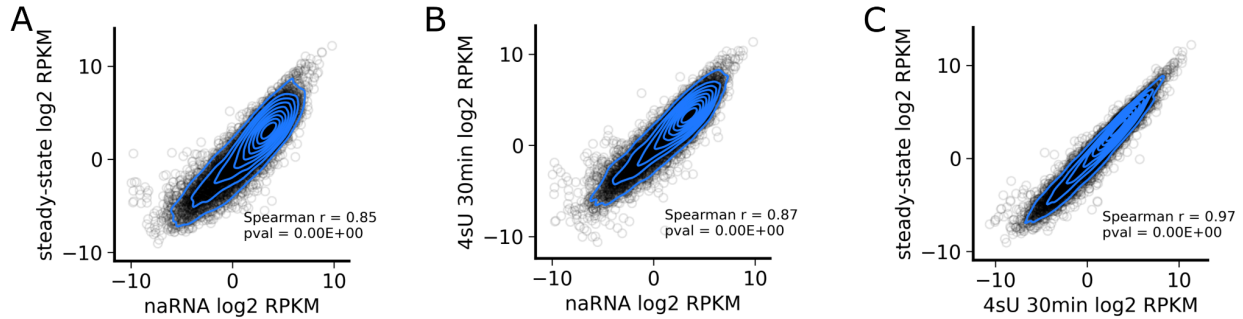
Supplementary Note 1: Contribution of alternative splicing coupled with nonsense-mediated decay to post-transcriptional regulation of gene expression levels

In this supplementary note, we use linear regression to explore the impact of NMD in gene expression. We find that at least ~9% of the variance in post-transcriptional gene regulation in RNA levels across genes is explained by AS-NMD.

Standard RNA sequencing measures polyadenylated steady-state mRNA molecules. These mRNA have undergone splicing, and most transcripts targeted by nonsense-mediated decay (NMD) have already been degraded. In contrast, chromatin-associated RNA sequencing captures nascent RNA before degradation, and during co-transcriptional splicing. We hypothesize that transcripts targeted by NMD explains part of the expression differences between standard RNA-seq, and nascent RNA sequencing (naRNA-seq). 4sU labeled RNA sequencing, at 30 minute and 60 minute labeling timepoints¹, sequencing captures recently transcribed RNA. We expect recently transcribed RNA to have a higher portion of NMD junctions than standard RNA-seq, but less than naRNA-seq. Indeed, we observe that genes with high % NMD junction reads in naRNA have a higher log2FC between naRNA and RNA-seq, than genes with low % NMD junction reads in naRNA (Main Text, Figure 1G). This effect still exists but it is greatly diminished when comparing 4sU labeled RNA-seq with standard RNA-seq. A lingering question is whether this higher percent of NMD junction reads has a global effect in gene expression as RNA matures. Here we implement a simple regression model to show that at least ~9% of the variance in degradation rates observed across genes can be attributed to NMD activity levels as measured by % of unproductive junctions. This is likely an underestimate due to confounding factors and regression dilution. Additionally, technical differences are also expected between naRNA and steady-state RNA measurements further biasing our estimate downwards.

Correlation in gene expression across different RNA-seq assays

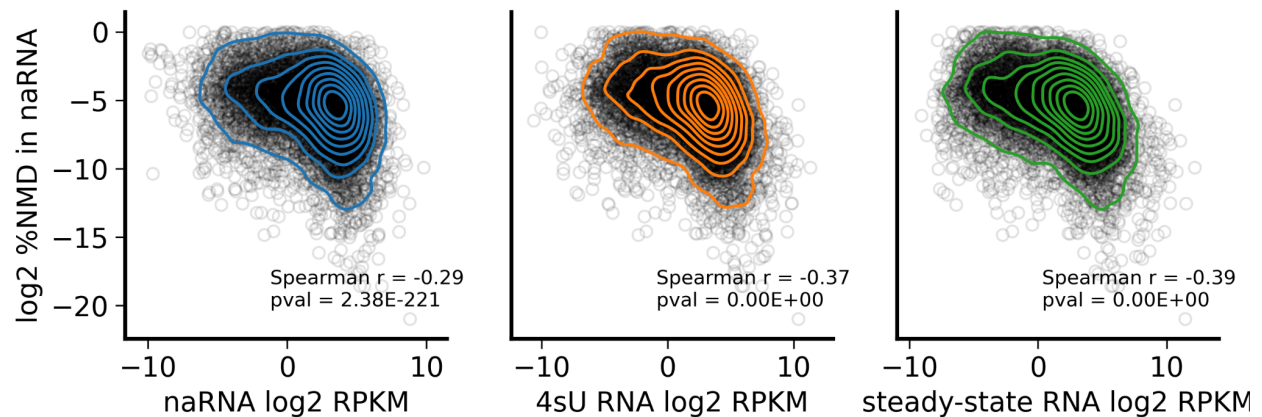
We normalized gene expression in the naRNA-seq, 4sU labeled RNA-seq, and standard RNA-seq data to RPKM as described in the methods. For each protein coding gene, we consider the median log2 RPKM expression across all the samples in each assay. We observed that gene expression is highly correlated among all RNA-seq assays, with a Pearson r of 0.85 between naRNA-seq and standard RNA-seq (Supplementary Note 1 Figure 1A). Gene expression in 4sU labeled RNA-seq has a Pearson r of 0.87 and 0.97 with naRNA-seq and standard RNA-seq, respectively (Supplementary Note 1 Figure 1B,C). This corroborated the notion that 4sU labeled RNA-seq captures RNA at a stage between nascent RNA and steady-state mature mRNA.



Supplementary Note 1 Figure 1: Correlation in gene expression levels measured using different RNA sequencing assays. (A) naRNA-seq vs standard RNA-seq. (B) naRNA-seq vs 4sU labeled (30 minutes) RNA-seq. (C) 4sU labeled (30 minutes) RNA-seq vs standard RNA-seq.

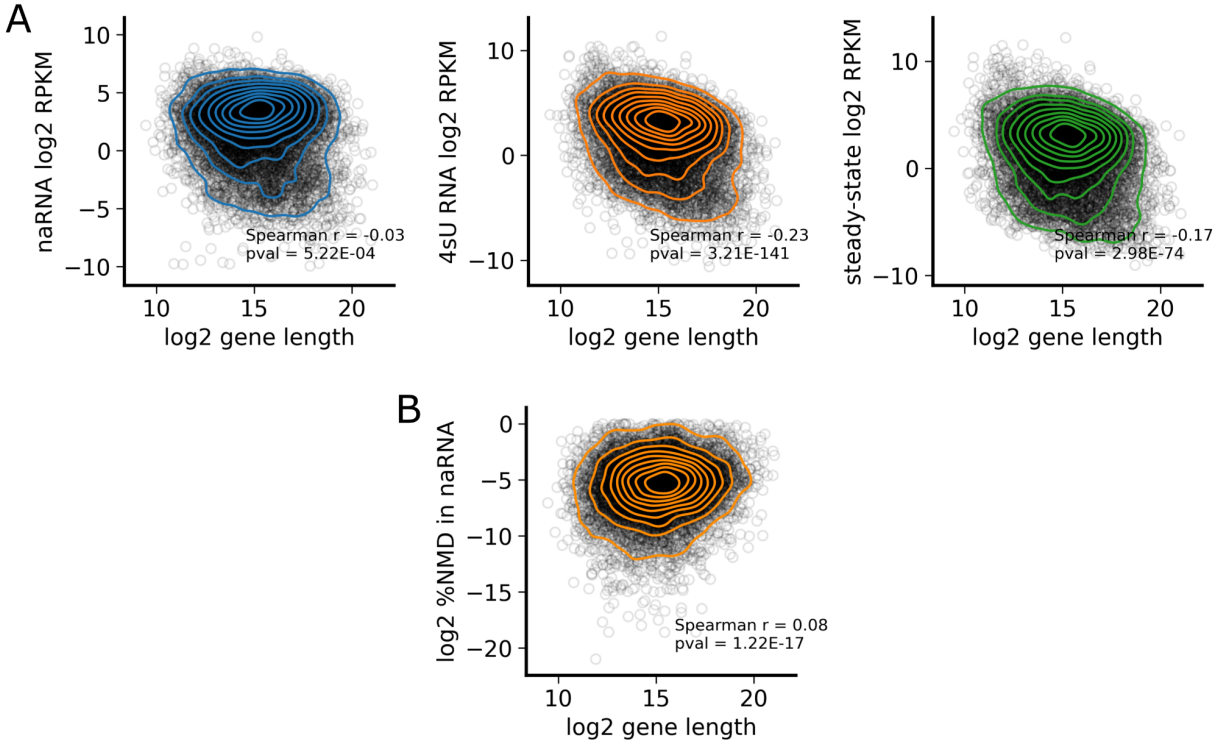
Gene expression and gene length are correlated with NMD transcript levels

The percent of NMD splice junction reads in a nascent RNA is negatively correlated with a gene's expression (Main Text, Figure 1D). This correlation strengthens across the RNA lifetime (Supplementary Note 1 Figure 2), as the effects of degradation become increasingly more prominent in increasingly mature RNA (Main Text, Figure 1G).



Supplementary Note 1 Figure 2: Correlation between the percent of NMD junction reads in naRNA-seq, and gene expression across the multiple RNA-seq assays.

Standard RNA sequencing and 4sU labeled RNA sequencing were both performed using polyA selection, which is known to introduce a 3' sequencing bias. Accordingly, gene length is negatively correlated with gene expression in both 4sU labeled and standard RNA sequencing, but this correlation is not observed in naRNA-seq (Supplementary Note 1 Figure 3A). Conversely, the percent of NMD junction reads in naRNA is slightly positively correlated with gene length (Supplementary Note 1 Figure 3B). For this reason, gene length is a potential confounder for the effect of NMD in the differences in gene expression between naRNA-seq and standard RNA-seq. To account for this, we used gene length as an extra covariate in our regression.



Supplementary Note 1 Figure 3: (A) Correlation between gene length and gene expression. (B) Correlation between gene length and percent of NMD junction reads in naRNA-seq.

Linear regression reveals global effect of NMD in gene expression

To determine the contribution of NMD to the post-transcriptional regulation of gene expression, we obtained the residuals of linear regressions of the measured expression levels of all genes at each stage of RNA maturity, versus the measured expression at earlier stages, adding the log2 of gene length as a covariate:

$$\text{steady-state RNA} \sim 4\text{sU RNA} + \text{gene length},$$

$$\text{steady-state RNA} \sim \text{naRNA} + \text{gene length},$$

$$4\text{sU RNA} \sim \text{naRNA} + \text{gene length},$$

Then we performed a second regression of the residual on the log2 percent of NMD junction reads in naRNA-seq data:

$$\text{resid}_{\text{steady-state RNA v 4sU}} \sim \log2 \% \text{NMD naRNA},$$

$$\text{resid}_{\text{steady-state RNA v naRNA}} \sim \log2 \% \text{NMD naRNA},$$

$$\text{resid}_{\text{naRNA v 4sU}} \sim \log_2 \% \text{NMD naRNA},$$

To ensure that the percent of variance attributed to NMD occurs post-transcriptionally, we also performed a regression on the coverage of H3K27ac and H3K4me3 at the transcription start site of each gene, and the H3K36me3 coverage across the gene body:

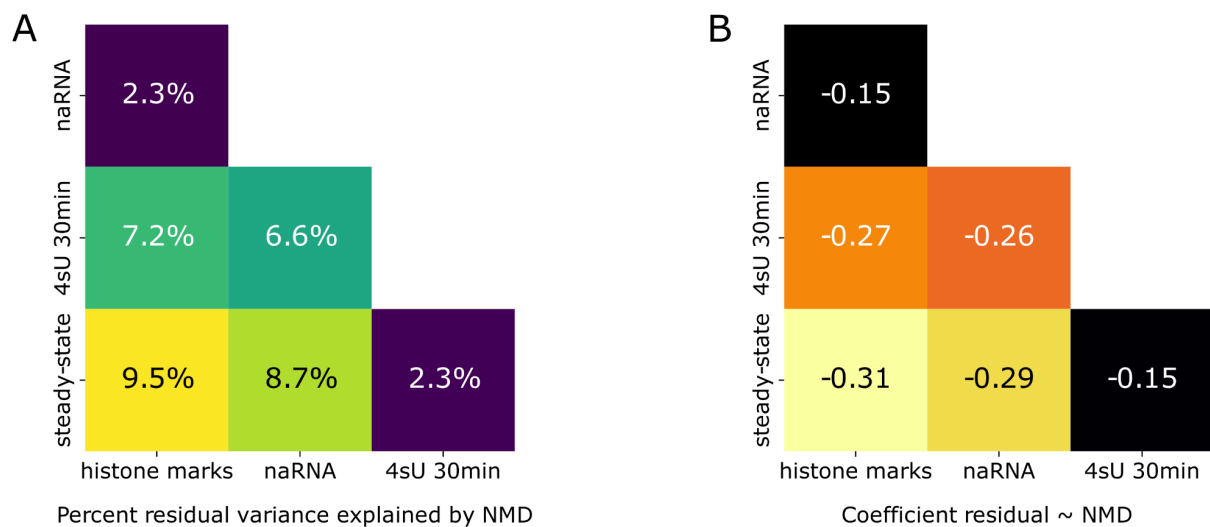
$$\text{RNA-seq} \sim \text{H3K27ac} + \text{H3K4me3} + \text{H3K36me3} + \text{gene length},$$

In all cases, nonsense-mediated decay levels, as measured by % of unproductive junctions, were negatively correlated with the residuals, implying that excess NMD reduces expression levels as the RNA matures (Supplementary Note 1 Figure 4A). The slope is stronger between assays that are further apart in the regulatory cascade. Using R^2 scores, we found that 8.7% of the residual variance of regressing the standard RNA-seq expression versus naRNA-seq expression and gene length is explained by the percent of unproductive junction reads in naRNA-seq. This percentage is smaller (6.6%) for the residual variance between 4sU labeled RNA-seq and naRNA-seq, and for that between standard RNA-seq and 4sU labeled RNA-seq (2.3%, Supplementary Note 1 Figure 4B).

Moreover, the percent of residual variance attributed to NMD after regressing the RNA-seq assays on histone modifications is small (2.3%) in naRNA-seq, as expected. The variance explained is bigger for 4sU labeled RNA-seq, and for standard RNA-seq (7.2% and 9.5%, respectively) (Supplementary Note 1 Figure 4B), which is consistent with the residual between chromatin marks and 4sU-seq and standard RNA-seq capturing RNA degradation levels.

Discussion

The results of this analysis confirm that NMD is an important post-transcriptional regulatory mechanism of gene expression level. NMD has global effects across the transcriptome, suggesting that splicing plays a greater role in gene regulation than previously thought. Due to technical biases, regression dilution and other confounding factors, our analysis underestimates the percent of variance attributable to NMD.



Supplementary Note 1 Figure 4:(A) Regression coefficient of the residual variance explained by the percent of NMD junction reads in naRNA-seq. (B) Percent of residual variance from $Y \sim X + \text{gene length}$ explained by the percent of NMD junction reads in naRNA-seq.

Supplementary Note 2: Splicing of unproductive isoforms is upregulated in splicing factors and downregulated in translation factors

Alternative splicing (AS) of transcripts targeted for nonsense-mediated decay (NMD) is a mechanism that affects gene expression post-transcriptionally. NMD is largely considered a quality control mechanism to remove mRNA molecules that encode truncated proteins. Some AS-NMD events are highly conserved, which suggests that these AS-NMD events have a regulatory role. In this Supplementary Note we use polynomial regression and Gene Ontology analysis to identify what genes produce the most or least transcripts subject to NMD. We find evidence of excess of AS-NMD in genes encoding splicing factors, and depletion in genes encoding translation factors.

Prior to the present study, the regulated alternative splicing of NMD targeted isoforms was considered a relatively uncommon occurrence, affecting mostly splicing factors²⁻⁵. Our analysis of nascent RNA-sequencing data revealed that AS leading to transcripts targeted for NMD represent 2.4% of all splice junction reads. We estimated that ~20% of all mRNA molecules produced are subject to NMD. The majority of protein coding genes present NMD isoforms, with 11,328 out of 14,000 protein coding genes analyzed having at least one NMD junction read, and 6,549 genes with at least 1% and less than 20% of reads predicted to induce NMD (we consider genes with less than 20% NMD splice junction reads because higher numbers could correspond to genes with very low expression or to genes with little to no protein-coding isoforms). These results put together suggest that AS-NMD events are widespread across the transcriptome, in greater abundance than previously described. Still, one remaining question is to what extent AS-NMD events play a regulatory function in gene expression versus a mere consequence of erroneous RNA splicing.

Genes with highly conserved poison exons have high percentage of NMD transcripts

Arguably the best studied instance of regulated AS-NMD are the genes in the Serine/arginine-rich splicing factors (SRSF) family. The majority of genes in this family can produce isoforms with alternative splicing events - usually cassette exons, that introduce premature termination codons (PTC), known as poison exons - that lead to NMD^{2,3,5,6}. Cross-linking and immunoprecipitation (CLIP) studies have shown that proteins of the SRSF family bind to their own poison exons, promoting their inclusion⁷⁻¹⁰. As a result, AS of NMD isoforms has been proposed as a post-transcriptional mechanism of gene expression auto-regulation in these genes. In addition to the SRSF family, AS events leading to NMD transcripts are enriched in splicing factors and chromatin factors⁴. These include the poison exons of the key spliceosome component SNRNP70¹¹, and in SMNDC1, a splicing factor involved in the assembly of the mature spliceosome^{11,12}. Other splicing factors autoregulate their own activity through other means, such as MBNL1, which represses inclusion of its own exon carrying a nuclear localization signal protein domain¹³.

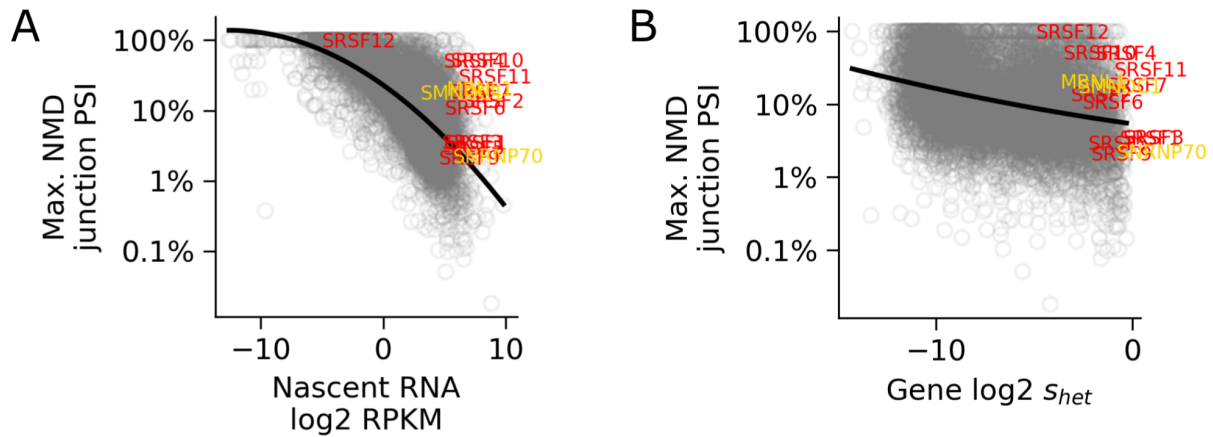
In this paper we show that two factors are highly correlated with the PSI of NMD junctions: (i) gene expression and (ii) the evolutionary constraint of the gene (ExtendedData 5). As a result, lowly expressed genes dominate the highest quantiles of NMD junction PSI. To account for these confounding factors, we performed a quadratic polynomial regression of each gene's maximum

junction PSI on the gene's expression in log2 RPKM, and the gene's constraint score in log2 s_{het}^{14} as follows:

$$\max(\text{junction PSI}) \sim a * \text{RPKM} + b * s_{het} + c * \text{RPKM} * s_{het} + d * \text{RPKM}^2 + e * s_{het}^2 + \text{const}$$

The residual of this regression indicates the deviation in the usage of NMD junctions (as measured by the maximum PSI of a NMD junction) from the expected usage based on the gene's expression and evolutionary constraint. Nine out of the eleven genes in the SRSF family with annotated splice junctions had a positive residual, indicating that they have higher NMD junction PSI than expected (Supplementary Note 2 Figure 1). Three of them: SRSF4, SRSF10 and SRSF11 had more than three standard deviations above the average residual. MBNL1, SNRNP70 and SMNDC1 also had positive residuals, with MBNL1 having more than two standard deviations above the average residual. SRSF5 and SRSF9 had a negative residual within one standard deviation from the mean. SRSF8 was excluded given that it does not have annotated protein coding splice junctions.

With the exception of SRSF12 and SMNDC1, all aforementioned splicing factors are in the highest quartile of gene expression level. And with the exception of SRSF12, all of them are also in the highest quartile of evolutionary constraint. The combination of high expression, high evolutionary constraint and high production of NMD transcripts likely played a role in why AS-NMD events in these genes have been extensively described. These results show that the highly conserved AS-NMD events in splicing factors are used significantly more frequently than in genes with similar levels of gene expression and evolutionary constraint. This is consistent with the notion that the ability of splice factors to regulate their own expression is an important example of regulated AS-NMD.



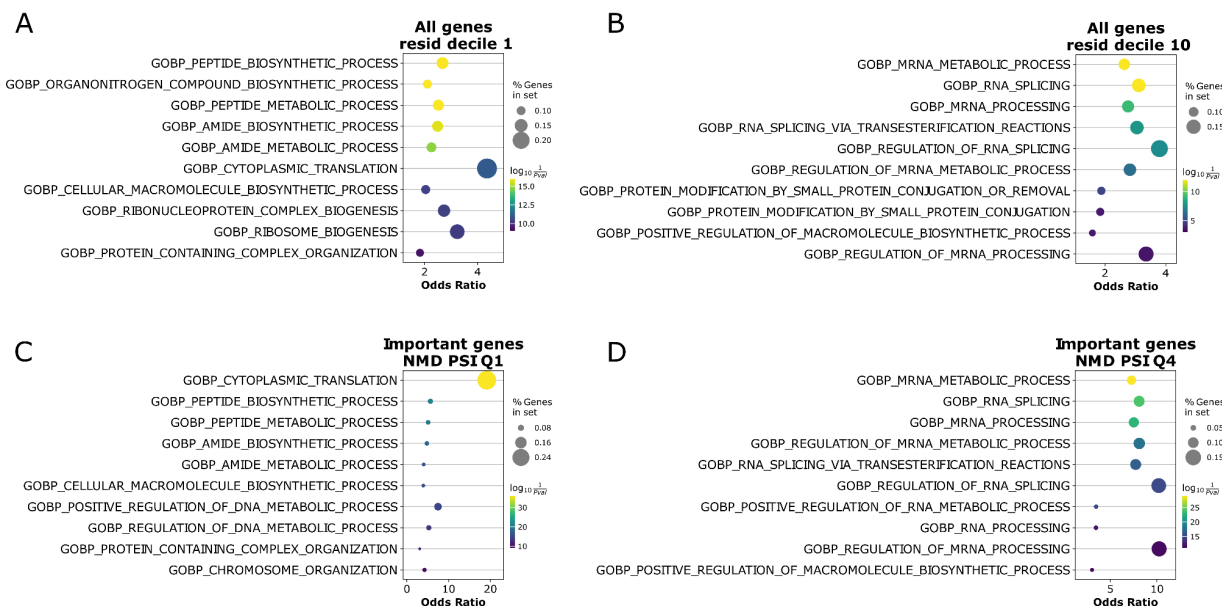
Supplementary Note 2 Figure 1: Splicing factors with highly conserved cassette exons tend to have more AS-NMD splicing than predicted by (A) gene expression and (B) evolutionary constraints alone. Gene names in red are from the SRSF genes, while names in orange correspond to MBNL1, SMNDC1 and SNRNP70. Black lines correspond to the polynomial regression of the maximum junction PSI on the corresponding X-axis feature. It is worth noticing that the polynomial regression used for the rest of the analyses was done simultaneously on the gene expression level and evolutionary constraint score.

Splicing factors have high AS-NMD, while cell cycle genes have low AS-NMD

To expand our previous results, we investigated what type of genes are the most affected by AS-NMD events, and which genes are the least affected. We reasoned that genes with the lowest residuals from the polynomial regression are the least affected by AS-NMD, possibly due to AS-NMD splicing suppression. Meanwhile, genes with the highest residual score could be genes whose AS-NMD is upregulated.

From the 14,000 protein coding genes that we selected for our main analysis, 11,328 had AS-NMD events in naRNA-seq data. We performed Gene Ontology enrichment analysis in the bottom and top decile of genes ranked by the residual from the polynomial regression (i.e., the 1,133 genes with the lowest regression residual, and the 1133 genes with the highest regression residual). For this, we used the Enrichr¹⁵ implementation in GSEApv¹⁶. We used the Biological Process subset of the C5: Gene Ontology signature collection from the MSigDB database^{17,18}. We found that the top enriched tags on the bottom decile are associated with peptide and amide biosynthesis, ribosome biogenesis, and translation (Supplementary Note 2 Figure 2A). In contrast, the most enriched tags on the top decile are associated with RNA metabolism, processing, and splicing, as well as protein modification (Supplementary Note 2 Figure 2B).

To corroborate these results, we also explored Gene Ontology enrichment in groups of highly expressed genes with high evolutionary constraint. From these, we selected 1,225 genes that are both in the top quartile of expression in our naRNA-seq data, and in the top quartile of the s_{het} evolutionary constraint metric from GeneBayes¹⁴. From these genes, we selected the genes at the bottom or top quartile of NMD junction PSI (306 genes on each quartile). Once again, we found that genes in the bottom quartile of NMD splicing junctions are enriched for translation and protein biosynthesis tags (Supplementary Note 2 Figure 2C), while genes on the top NMD quartile are enriched for tags in mRNA metabolism and splicing (Figure 2D).

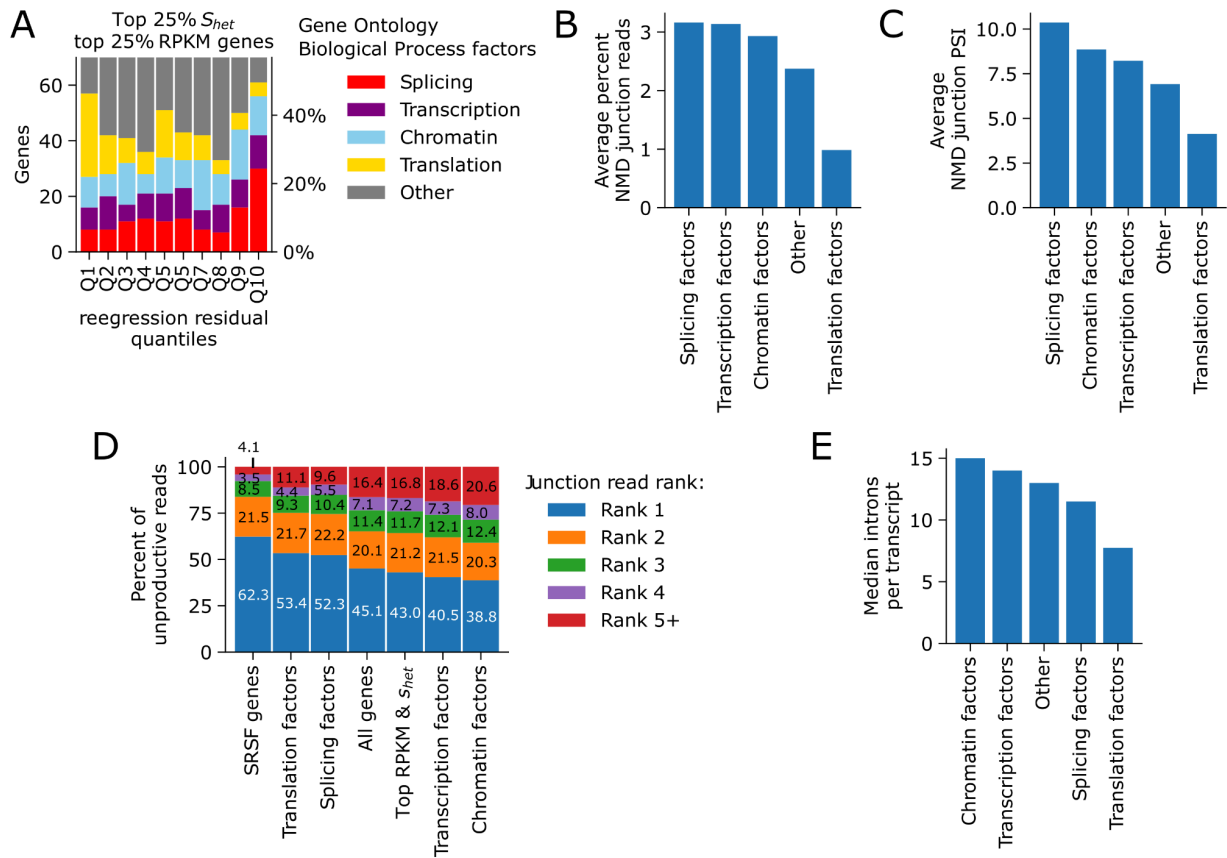


Supplementary Note 2 Figure 2: Gene Ontology analysis using MSigDB's C5 collection of tags, selected for Biological Process events. (A) All genes in the bottom decile of polynomial residual. (B) top decile of residual. (C) Tags in top expressed and evolutionary conserved genes, in the bottom NMD junction PSI. (D) Tags in top expressed and evolutionary conserved genes, in the top NMD junction PSI.

To further explore the contribution of specific types of genes, we selected genes with Gene Ontology Biological Process tags associated with transcription (transcription factors), RNA splicing (splicing factors), chromatin remodeling (chromatin factors) and RNA translation into proteins (translation factors), from the 1,225 genes in the top RPKM and s_{het} score quartiles. Genes in the top decile of the regression residual were enriched for splicing factors (enrichment = 2.43, hypergeometric test p-value = 1.9×10^{-7}), while genes in the bottom decile were enriched for translation factors (enrichment = 2.64, hypergeometric test p-value = 2.1×10^{-8} ; Supplementary Note 2 Figure 3A). Previous studies reported that regulated AS-NMD events are more common in splicing factors, RNA binding proteins, and chromatin factors. Our results further show that indeed splicing and chromatin factors have a larger percent of NMD junction reads (Supplementary Note 2 Figure 3B) and a larger NMD junction PSI (Supplementary Note 2 Figure 3C) than other types of genes.

On average, the most abundant NMD junction of a gene contributes 45.1% of all NMD junction reads, while 16.4% are contributed by reads in the fifth and lower rank (Main figure 2B). This suggests that splicing is error-prone and that genes tend to produce multiple NMD junction reads. Splicing factors have a higher contribution from the top ranked NMD junction, with 52.3%, and a smaller contribution from the fifth and lower ranked NMD junctions, with 9.6% (Supplementary Note 2 Figure 3D). The difference is even higher in SRSF genes, with 62.3% of the NMD junction reads on average coming from the top junction, and only 4.1% from the fifth and lower ranked NMD junctions. Translation factors also have a higher than average contribution from the top ranked NMD junction, while transcription and chromatin factors have a higher than average NMD contribution from junctions ranked fifth and lower. Interestingly, coding transcripts from splicing factors and translation factors tend to have fewer introns than chromatin and transcription factors (Supplementary Note 2 Figure 3E).

These results show that, despite both splicing and chromatin factors having more NMD junction reads overall, splicing factors have a larger contribution from a single AS-NMD event, while chromatin factors have a more uniform contribution of splice junctions. This suggests that there are AS-NMD events in splicing factors that are consistently spliced, which implies that the process is regulated. In contrast, in chromatin factors there are multiple AS-NMD events contributing to the overall production of NMD isoforms. These could be the result of errors in the splicing of a larger number of introns per transcript. Finally, our results also provide evidence that AS-NMD events are suppressed in translation factors.

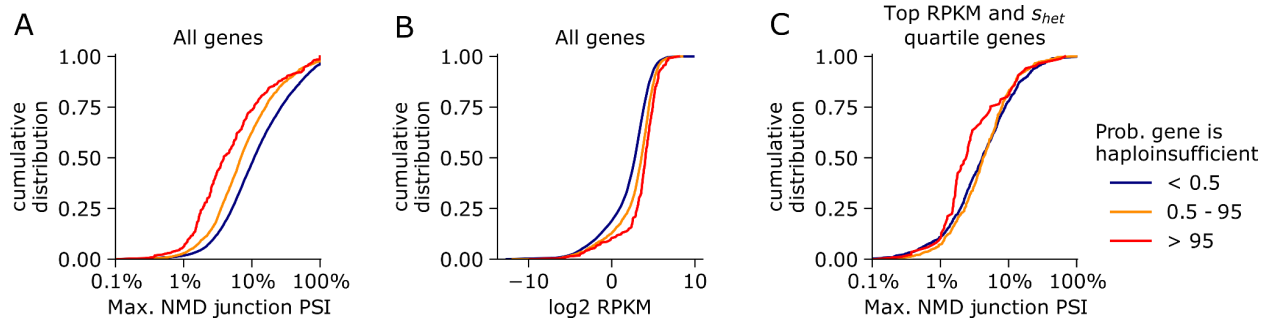


Supplementary Note 2 Figure 3: AS-NMD in splicing and translation factors. (A) percent of genes from each tag present in polynomial regression residual. For clarity, the barplot shows up to 70 genes per decile, out of a total of 123. (B) Average percent of NMD junction reads on multiple gene categories. (C) Average percent of NMD junction PSI on multiple gene categories. (D) distribution of NMD junction ranks across multiple gene categories. (E) Median number of introns per coding transcript by each gene category.

Haploinsufficient genes have lower levels of AS-NMD

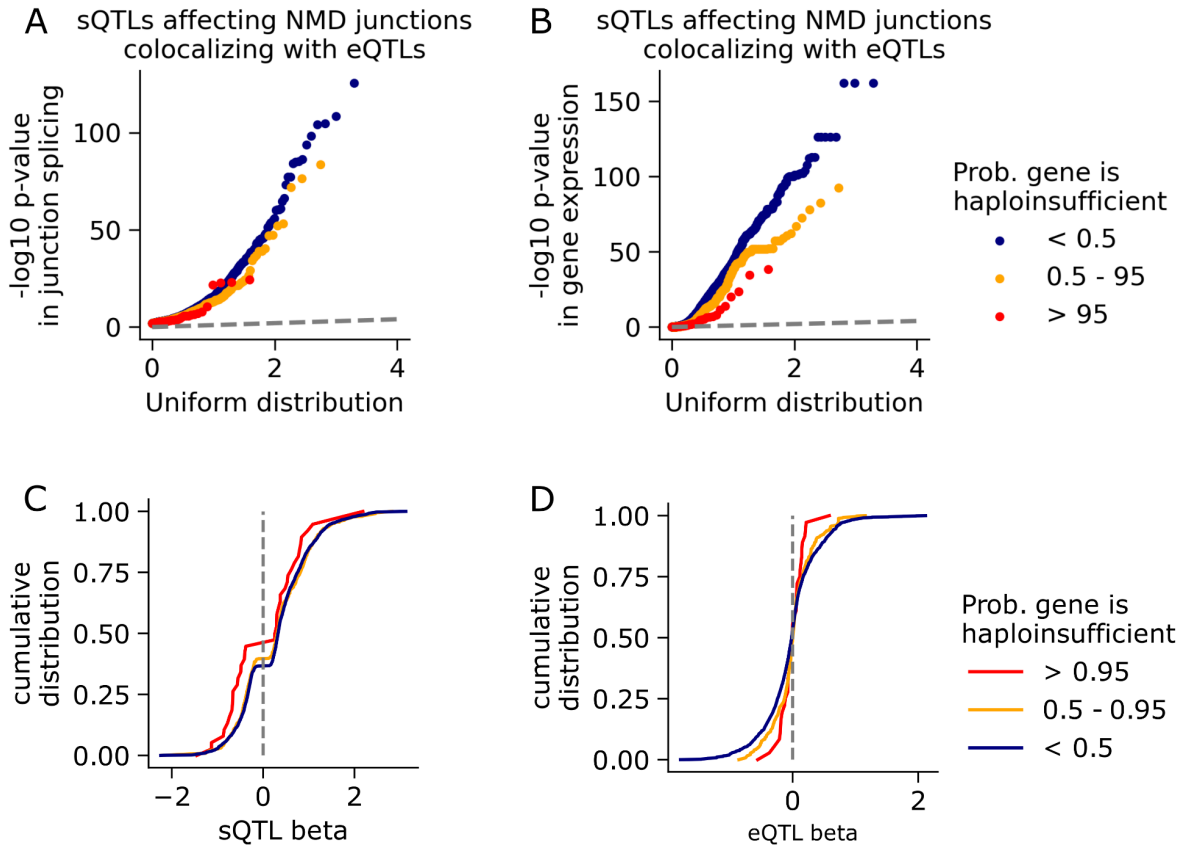
Genes are considered haploinsufficient if deletion or loss-of-function mutations on one single copy lead to reduction of fitness. This means that for haploinsufficient genes, the dosage or functional RNA and/or protein produced by one single copy is not enough to maintain the gene's proper function. Since haploinsufficient genes are likely more sensitive to changes in gene dosage than genes that can maintain proper function with one single functional copy. We reasoned that haploinsufficient genes may exhibit fewer AS-NMD events than haplosufficient genes. To test this hypothesis, we used the haploinsufficiency prediction scores from Huang et al. 2010¹⁹ to sort genes according to their probability of being haploinsufficient. We found that genes with high probability of being haploinsufficient in general have a lower NMD splice junction PSI (Supplementary Note 2 Figure 4A). However, this effect could be influenced by higher expression levels of haploinsufficient genes compared to haplosufficient genes (Supplementary Note 2 Figure 4B). Indeed, when we analyze genes in the top expression and evolutionary constraint quartile, the difference in splice junction PSI for the most used NMD

junction is reduced, although haploinsufficient genes still have a lower PSI (Supplementary Note 2 Figure 4C).



Supplementary Note 2 Figure 4. AS-NMD in haploinsufficient genes. (A) Cumulative distribution of the percent spliced-in of the highest used NMD junction in genes, stratified by their probability of being haploinsufficient. Red lines correspond to genes with high probability of being haploinsufficient. Orange lines are genes that are more likely than not to be haploinsufficient. Blue lines correspond to genes that are likely not haploinsufficient. (B) Cumulative distribution of gene expression in log2 RPKM in genes. (C) Cumulative distribution of the percent spliced-in of the highest used NMD junction in genes in the top quartiles of expression and evolutionary constraint.

In our main analysis, we demonstrate that splicing QTLs (sQTLs) affecting NMD associated splice junctions are more likely to have an effect on the gene's expression level, than sQTLs affecting protein coding junctions. Given that AS-NMD events affect haploinsufficient genes, we asked what is the impact that sQTLs affecting NMD splice junctions in haploinsufficient genes have on gene expression. We found that, although sQTLs affecting NMD splice junctions in haploinsufficient genes have a similar strength of effect than their counterparts in genes that are not haploinsufficient (Supplementary Note 2 Figure 5A), their effect on gene expression is weaker (Supplementary Note 2 Figure 5B). Interestingly, sQTLs affecting NMD junctions in haploinsufficient genes tend to decrease intron splicing rather than increase it, when comparing them with sQTLs affecting NMD junctions in genes that are not haploinsufficient (Supplementary Note 2 Figure 5C). NMD sQTLs in haploinsufficient genes also tend to result in smaller changes in gene expression levels, when compared to genes that are not haploinsufficient (Supplementary Note 2 Figure 5D).



Supplementary Note 2 Figure 5: sQTLs affecting NMD junctions in haploinsufficient genes. (A) QQ plot of the p-values for sQTLs affecting NMD splice junctions, stratified by the gene's probability of being haploinsufficient. (B) QQ plot of the p-values of the effect of these sQTLs on the expression of the gene containing the NMD splice junction. (C) Cumulative distribution of the sQTL effect size on the affected NMD splice junctions. (D) Cumulative distribution of the sQTL effect size on the expression of the genes containing the NMD splice junctions.

Discussion

AS-NMD events are pervasive, affecting the majority of protein coding genes. Functional and evolutionary evidence suggest that some AS-NMD events are regulated and they have a role in maintaining homeostasis in the levels of gene expression, as it is the case of the highly conserved poison exons in the SRSF genes and other splicing factors. Here we show that splicing factors have a higher percentage of NMD splice junction reads, and a higher NMD splice junction PSI, than other types of genes with similar levels of gene expression and evolutionary constraint. This increase in NMD splicing comes primarily from one or two NMD splice junctions, which supports the hypothesis that AS-NMD events actively regulate the expression levels of splicing factors. In contrast, translation factors have smaller percentages of NMD splicing and lower NMD splice junction PSI than other genes with similar levels of expression and evolutionary constraint. This suggests that AS-NMD events are suppressed in translation factors.

Supplementary Figures

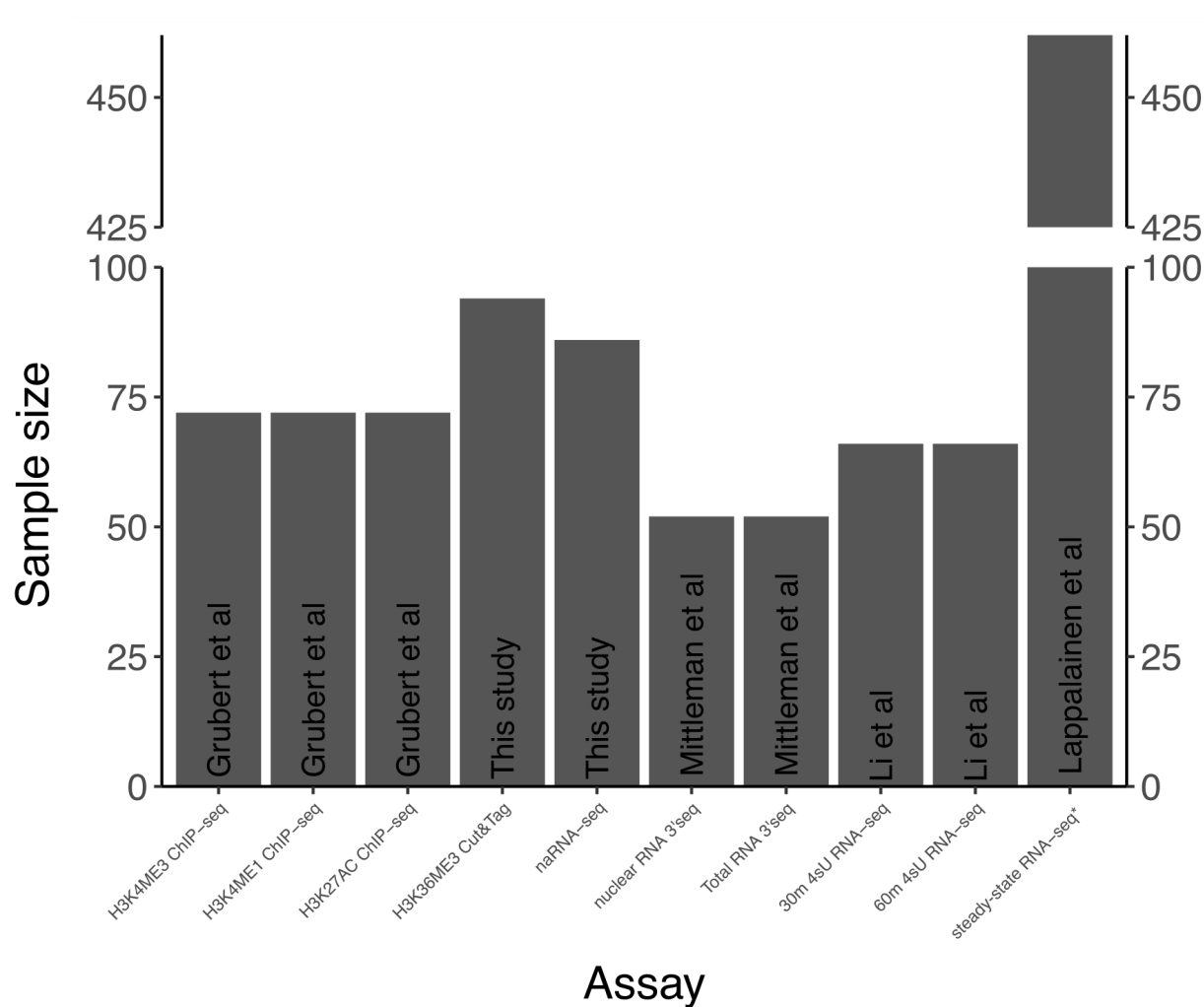


Figure S1. Data sources and sample size. Source publication, assay, and sample size used for QTL analyses^{1,20-22}. All samples are lymphoblastoid cell lines of Yoruba ancestry, except *poly RNA-seq dataset contains 89 Yoruba ancestry cell lines, and 362 non Yoruba ancestry cell lines. Some analyses (Methods) use only Yoruba ancestry cell lines of this dataset.

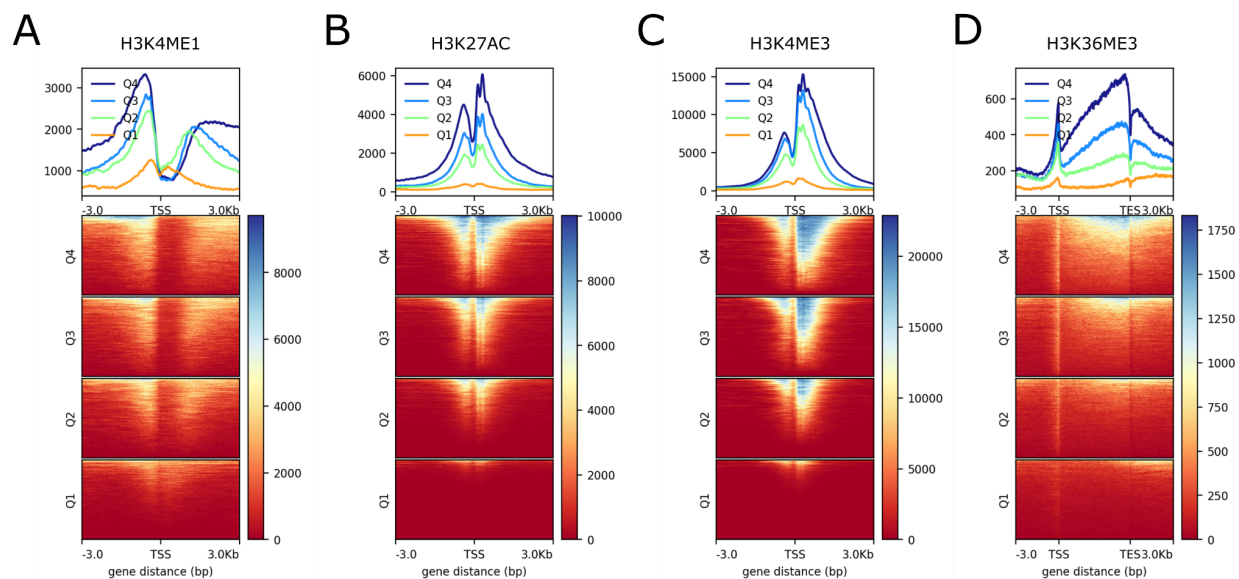


Figure S2. Chromatin profiling metagene plots. (A) H3K4me1 (enhancer mark) ChIP-seq signal. Genes grouped by expression quartile, as determined by polyA-RNA-seq RPKM values of the top 14000 expressed genes. (B) H3K27ac (promoter/enhancer mark) ChIP-seq signal. (C) H3K4me3 (promoter mark) ChIP-seq signal. (D) H3K36me3 CUT&Tag signal. Enrichment at the promoter region may represent some cross-reactivity with other marks.

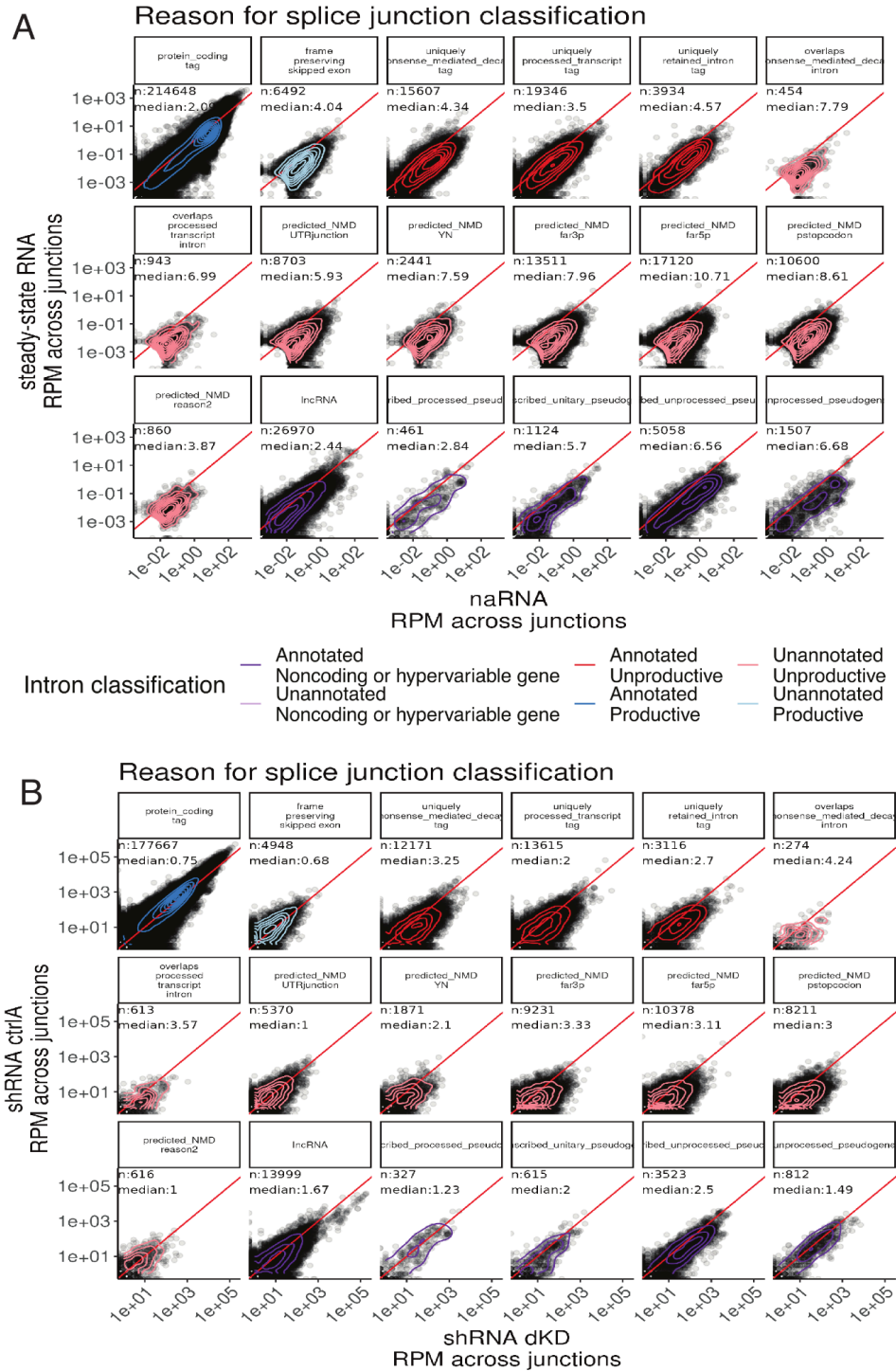


Figure S3. Splice junction abundances across datasets are consistent with classifications of splice junctions as productive or unproductive. (A) Splice junctions (introns) are classified into 6 types based on whether they are annotated, predicted to create a productive transcript, and by whether they are in a protein-coding gene (colors, Supplemental Methods). For each subcategory of these six classes, a scatter plot depicts relative junction prevalence in naRNA vs steady-state RNA. Each point is a unique splice junction. The median fold changes across all n junctions in each category labeled in each plot. Splice junctions annotated or predicted as productive are generally similarly present in naRNA versus polyA RNA. Splice junctions annotated as unproductive are generally depleted from polyA RNA. (B) Similar to (A), but showing relative change in scramble shRNA control compared to shRNA double knockdown of NMD factors *SMG6* and *SMG7*²³.

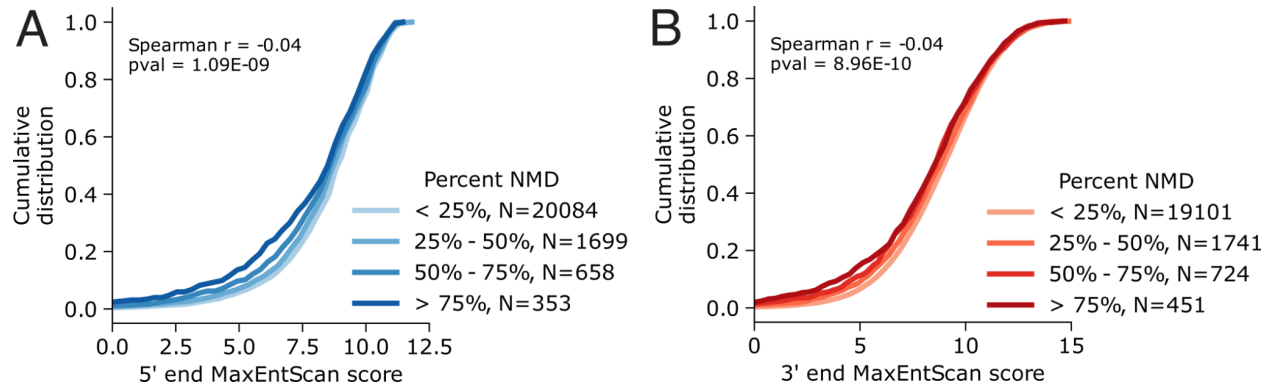


Figure S4. Splice site strength correlates with percent of unproductive splice junctions. (A) Rate of unproductive splicing overlapping annotated productive introns is negatively correlated with strength of the productive intron's 5' splice site (MaxEntScan). Correlation summarized with spearman correlation coefficient and two-sided correlation test P value, and visually presented as cumulative distribution of percent unproductive (NMD-inducing) splice junctions, grouped by MaxEntScan score quintiles. (C) Same as (B), but for 3' splice sites.

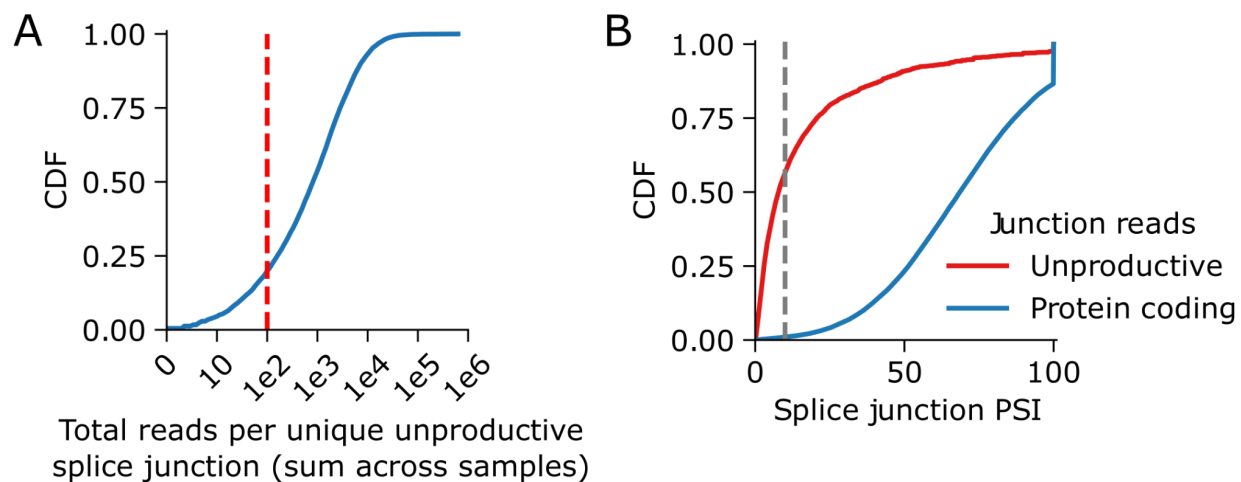


Figure S5. Distribution of unproductive splice junction usage. (A) Cumulative distribution of the total number of reads mapping to each unique unproductive splice junction across all naRNA-seq samples. We filtered out the ~20% of splice junctions with fewer than 100 reads (red) for the unproductive junction contribution analysis in (B): Distribution of the splice junction PSI of unproductive and productive splice junctions, averaged across all naRNA-seq samples. The PSI of a junction is the number of reads mapping to the junction, divided by the maximum number of reads mapped to any junction in the same gene. The junction with the highest number of reads in a given gene has a junction PSI of 100%. More than 50% of unproductive junctions have a junction PSI of 10% or less (gray line).

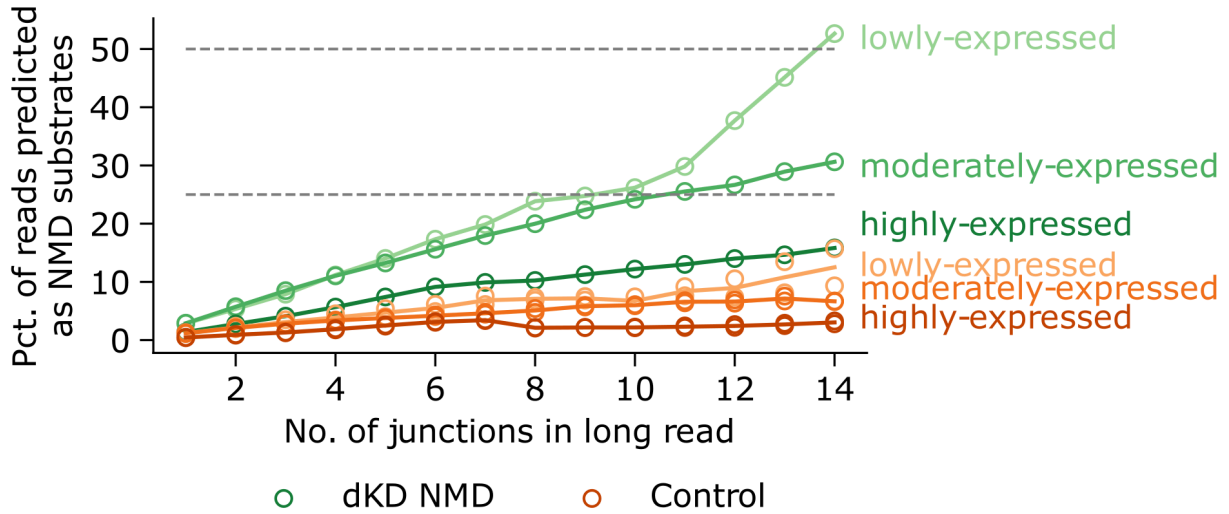


Figure S6. Over 50% of transcripts from lowly-expressed long genes are unproductive. Percent of full-length Nanopore reads that are targeted by NMD, as a function of the number of splice junctions in the read, and stratified by gene expression. dKD NMD corresponds to the shRNA SMG6/SGM7 double knockdown data in HeLa cells²⁴, and control corresponds to the shRNA scrambled controls from the same dataset. Expression rankings are based on the RPKM expression of shRNA *SMG6/SGM7* double knockdown in HeLa cells using short-read data²⁴. Lowly-expressed corresponds to genes in the bottom quartile (0.01-1.3 RPKM), moderately-expressed are genes in the second and third quartiles (1.3-15.6 RPKM), and highly-expressed are genes in the top quartile of expression (>15.6 RPKM). The gray lines mark the 25% and 50% thresholds of mRNA molecules targeted for NMD.

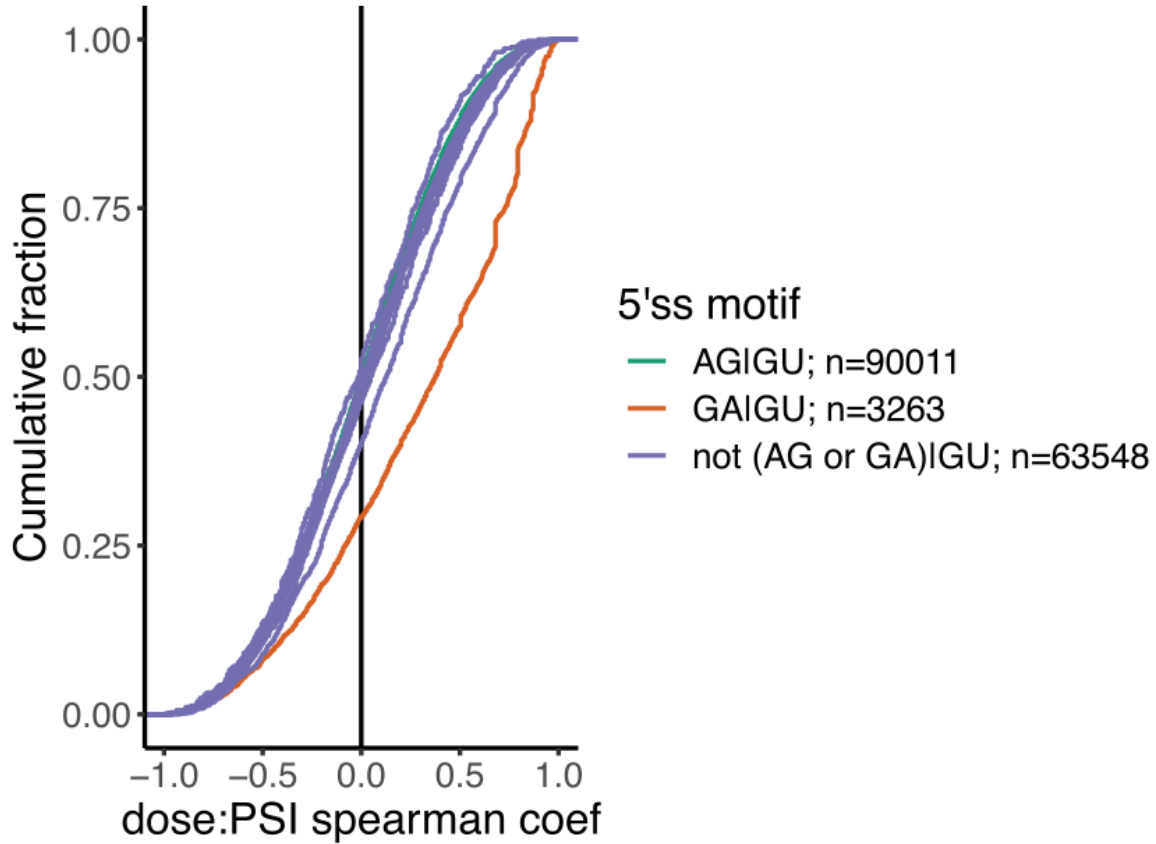


Figure S7. Genomewide activation of GA|GU 5' splice sites. Cumulative distribution of dose vs splicing PSI spearman correlation coefficients amongst splice junctions, grouped by their 5' splice site motif. All 5' splice site motifs contain GU at position +1:+2. 16 possible dinucleotides at position -2:-1 are plotted. The most common sequence (AG|GU, which forms Watson-Crick base-pairs with U1 snRNA) has a similar number of positive and negative correlations. Only GA|GU splice junctions as a group have a noticeable shift in distribution towards positive correlations between dose and splicing rate. The 641 splice junctions with significant ($FDR < 10\%$) dose-response correlation coefficients, corresponding with a spearman coefficient > 0.775) were chosen for further analysis (a subset of which are associated with cassette exons, see Methods).

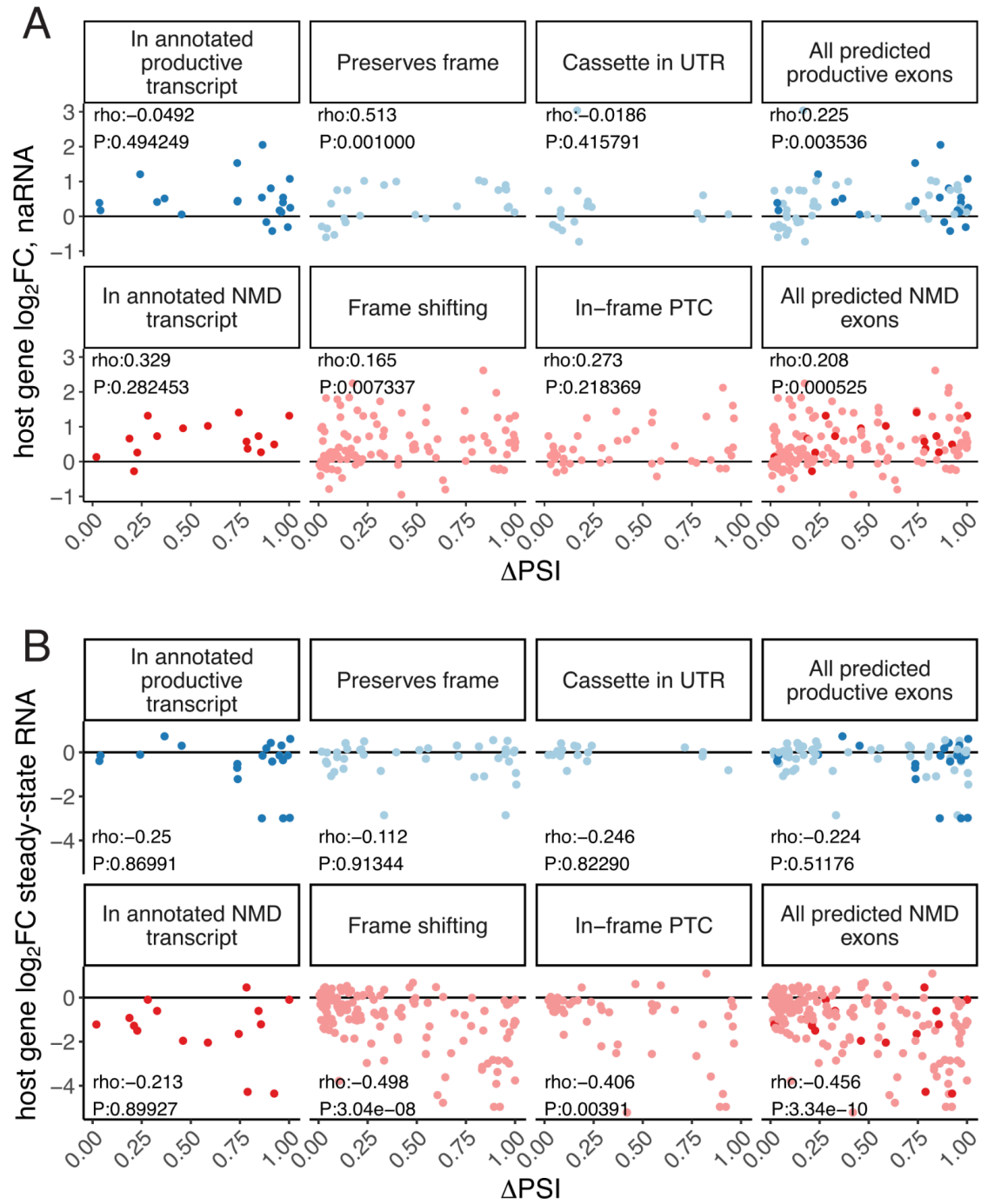


Figure S8. Expression and splicing effects of risdiplam-induced exons. (A) The change in percent cassette exon spliced-in (Δ PSI) vs the host-gene expression as measured in naRNA at 3160nM risdiplam dose for 305 induced cassette exons. Exons are grouped into facets for each possible reason the exon is predicted as either NMD-inducing (red) or stable (blue). Annotated exons are darker colors. In general, no correlation is observed in between splicing of cassette exons and host-gene expression in either NMD-inducing or stable exons. (B) Same as A, but effects measured in steady-state RNA. In general, a negative correlation is observed for NMD-inducing-, but not stable cassette exons. Correlations summarized with Spearman's rho correlation coefficient and P-value of two-sided correlation test.

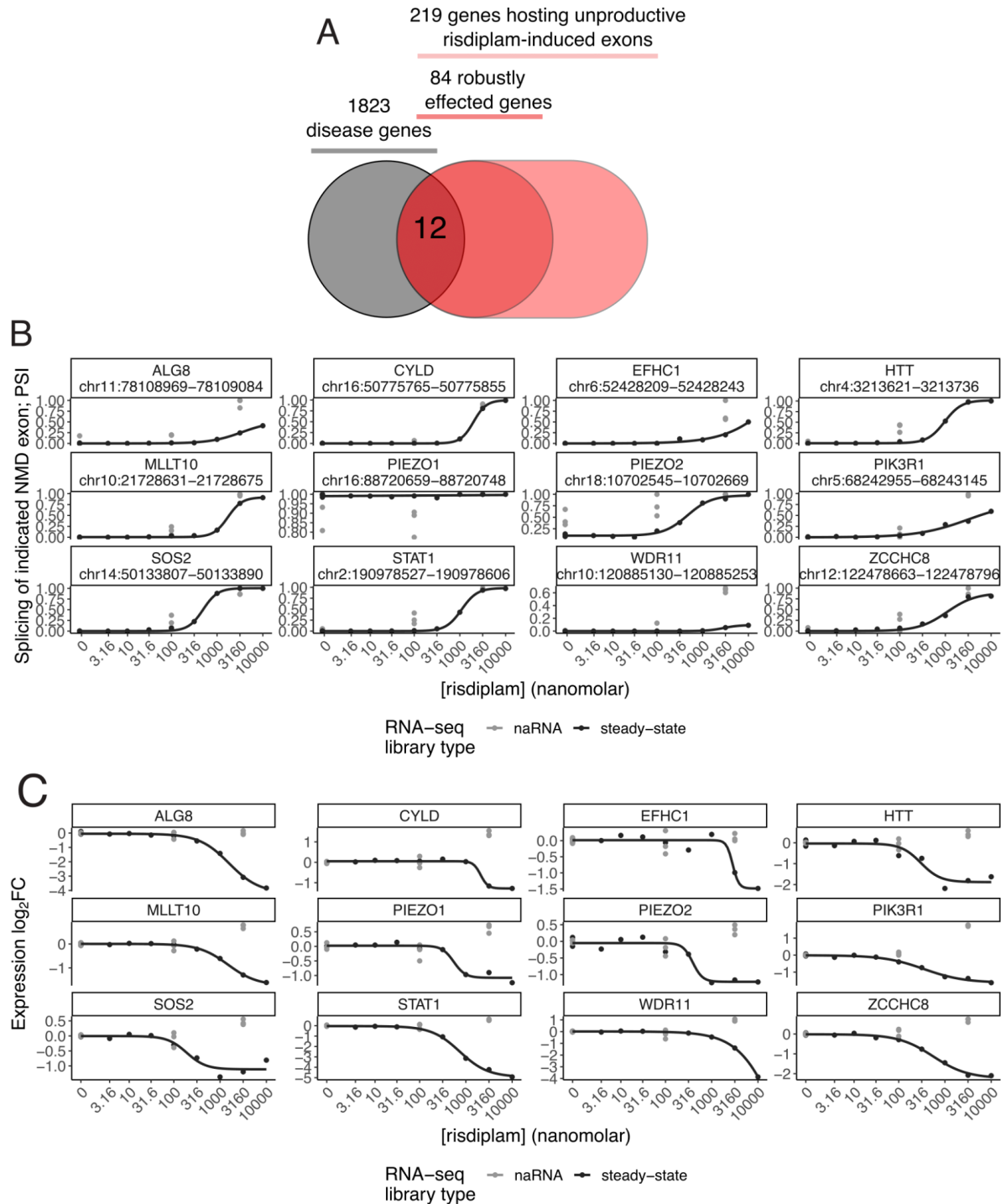


Figure S9. Dose-response curves of disease genes. (A) The set of 219 significantly down-regulated genes which have an identifiable splicing change predicted to induce NMD were further subsetted into 84 genes with robust effects, defined as $FDR < 10\%$ with greater than two-fold down-regulation in steady-state RNA after 3160nM risdiplam treatment. These 84 genes were intersected with a set of 1823 disease genes which may produce therapeutic benefit if down-regulated (Supplemental Methods), resulting in 12 candidate genes, including *HTT* a known target of risdiplam and analogs under current investigation as a therapy for Huntington's disease. Disease genes are defined by the presence of dominant negative alleles in the OMIM database²⁵. (B) Dose-response curve shown for the splicing of the predicted NMD-inducing cassette exon among the candidate genes. Dose-response log-logistic model fit based on PSI estimates in polyA RNA-seq, while PSI estimates from naRNA-seq samples are shown as lone points. (C) similar to (A), but measuring host-gene expression for the induced cassette exons. As expected, the down-regulating effect is polyA-specific, consistent with post-transcriptional regulation by NMD.

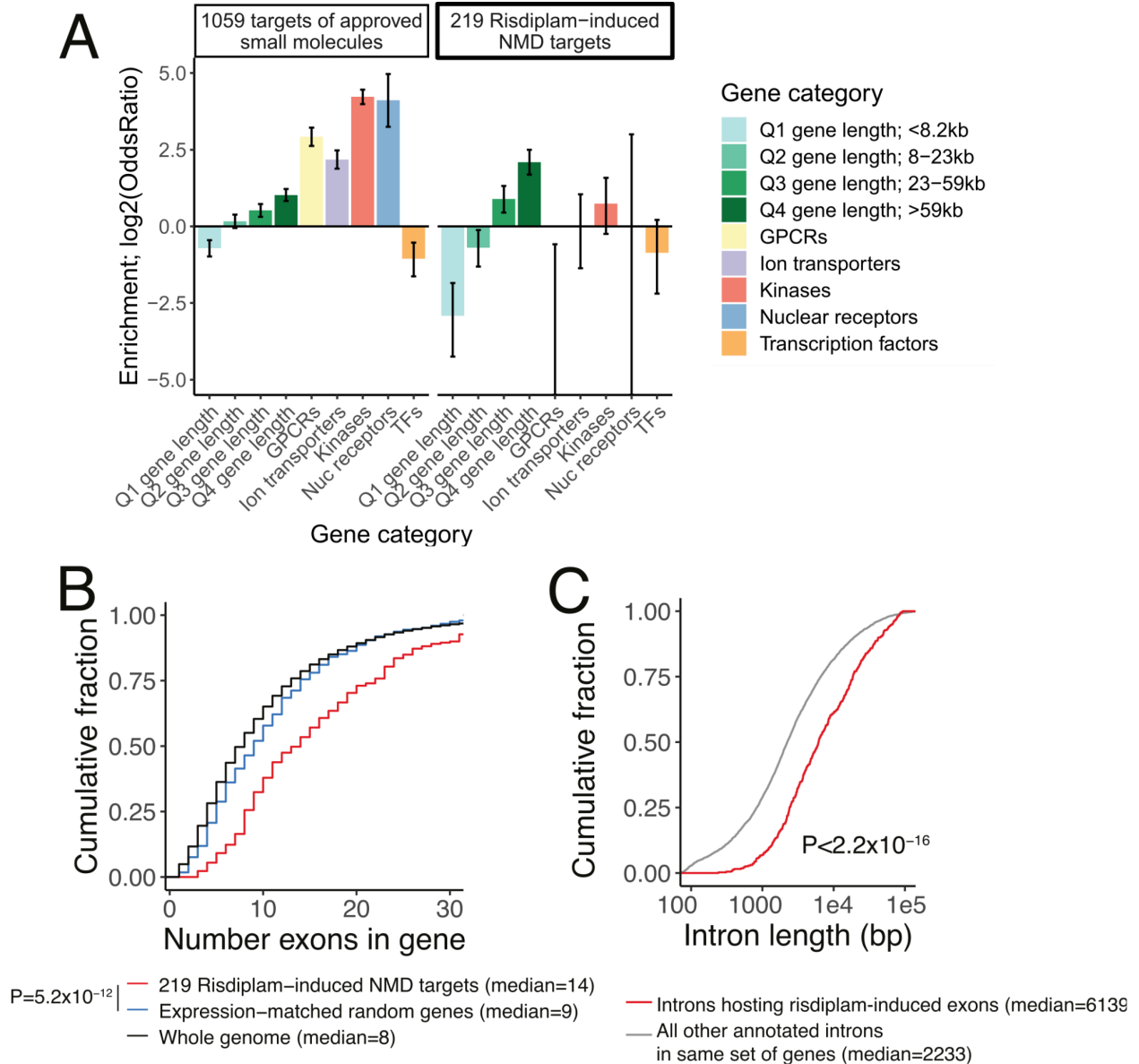


Figure S10. Risperidone induced NMD exons are more likely to occur in long genes and introns. (A) The enrichment of sets of genes among approved small molecule targets (small molecule drugs that primarily function at the level of protein binding), risperidone-induced predicted NMD targets (defined by identification of and annotation of splicing changes), and a larger set of risperidone-induced post-transcriptionally down-regulated genes (defined by changes in gene expression in naRNA and steady-state RNA). Gene sets defined by quartiles of gene length, and gene ontology categories (Methods). Error bars represent bootstrapped (resampling genes within each group) 95% confidence intervals. (B) Cumulative distribution of the number of exons in the gene (using the highest expressed protein_coding transcript isoform as a reference) for genes with a risperidone-induced exon, a similarly sized set of expression-matched set of genes, or all protein_coding genes in the genome. P value from two-sided Mann Whitney U-test. (C) Cumulative distribution of the length of introns that host risperidone-induced exons, or as a control, all other annotated introns in the same set of genes. P-value from two-sided Mann Whitney U-test.

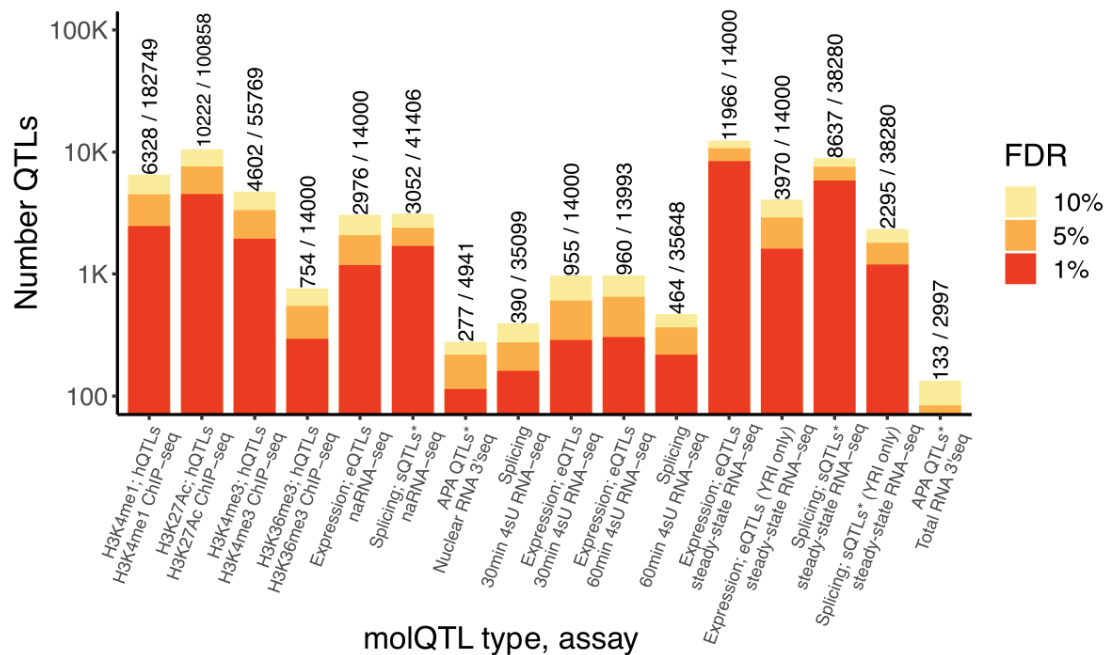


Figure S11. Number of QTLs. The number of QTLs at various false discovery rate thresholds. Numbers indicate the number of QTLs at 10% FDR, and the total number of test features. Numbers of *sQTLs and apaQTLs are counted once per local QTL. That is, an sQTL that affects multiple introns in the same LeafCutter cluster is only counted once.

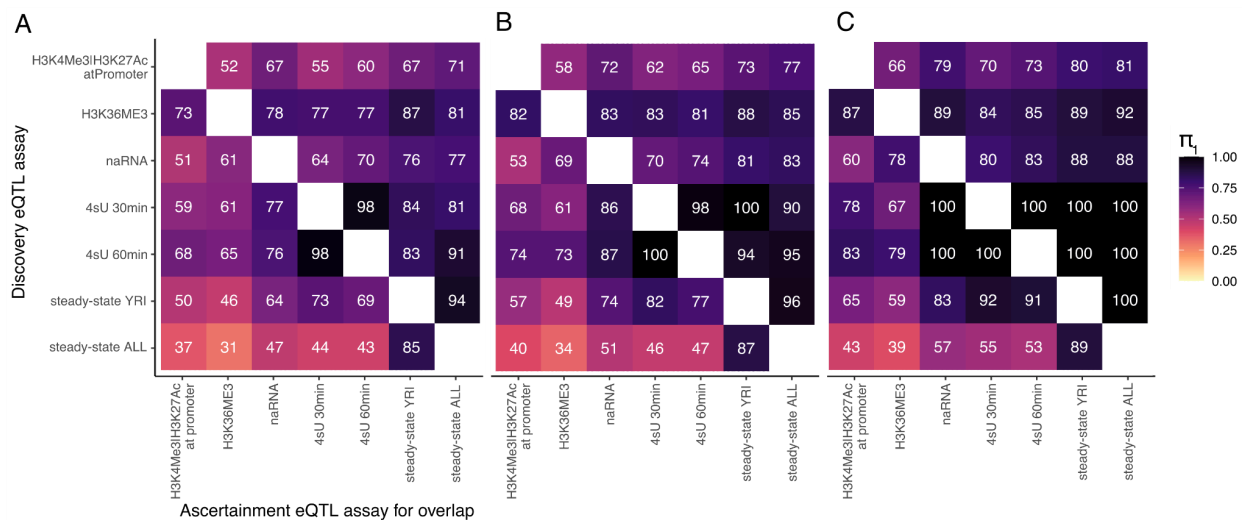
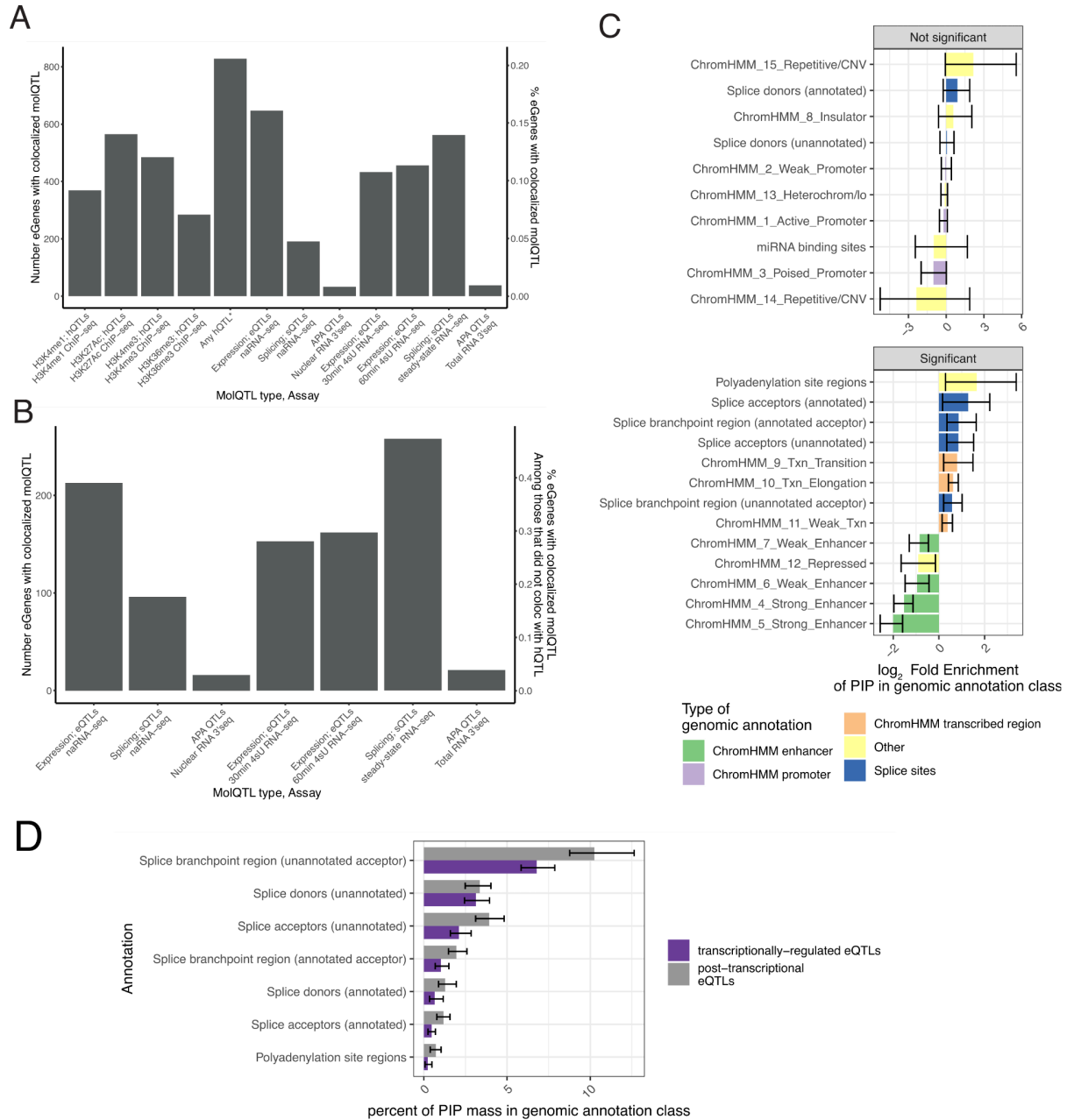


Figure S12. π_1 sharing across phenotypes that measure gene expression. (A) QTLs discovered at 10% FDR (rows) were assessed for overlap by measuring the π_1 statistic for the corresponding SNP:gene pair in different assays (columns). Promoter marks include H3K27ac, and H3K4me3, and consider any annotated promoter for the corresponding gene (Methods for details). (B, C) same as A, using 5% and 1% FDR threshold for discovery QTLs, respectively. Overall, the effect of promoter hQTLs are apparent throughout subsequent datasets (upper right of square). By contrast, effects discovered in steady-state RNA are less likely to be apparent as promoter hQTL signals (lower left of square), suggesting the existence of post-transcriptional regulation in determining steady-state eQTL signals. Steady-state eQTLs were mapped using either n=89 (similar sample size as other assays) ancestry-matched samples (YRI), or using n=453 mixed ancestry samples.



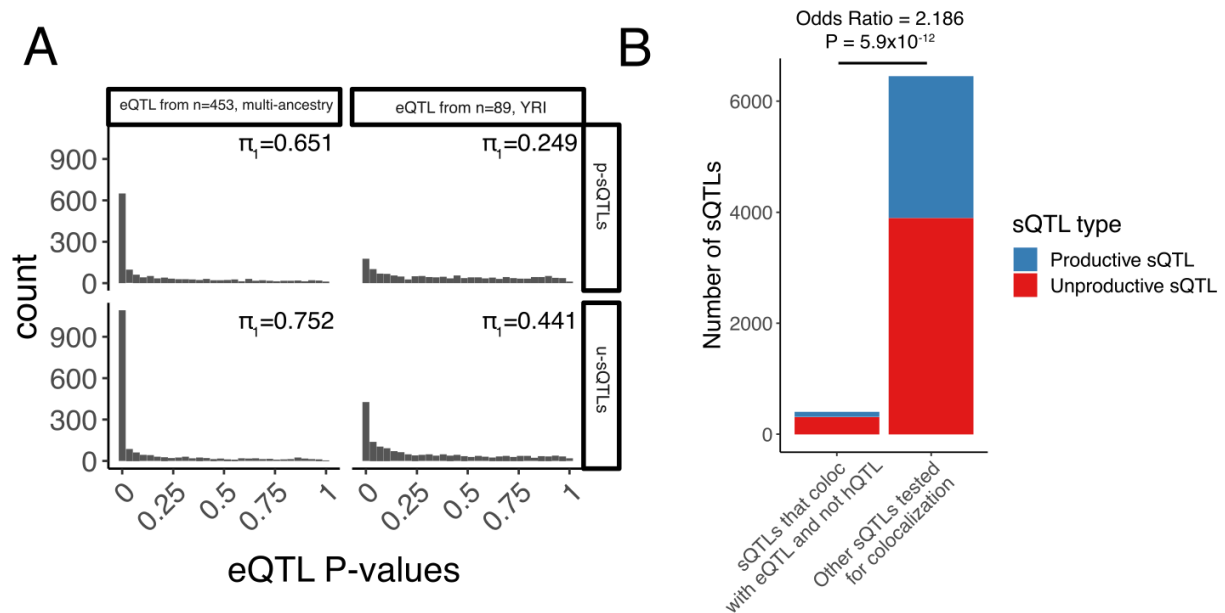


Figure S14. Enrichment of eQTL signals among sQTLs that affect unproductive splice junctions. (A) u-sQTLs (sQTLs that alter the balance of unproductive and productive isoforms, defined as sQTLs in LeafCutter clusters that contain an sQTL in an unproductive splice junction and are not nominally any hQTL) are enriched for eQTL signal, as evidenced by inflation for small P-values, compared to p-sQTLs (which contain significant sQTLs only in productive splice junctions). π_1 statistic labeled in top right of P-value histograms. Importantly, unlike colocalization analysis in (B), the general inflation of small eQTL P-values implicitly accounts for sQTLs that may explain non-primary eQTLs. The effect is observed when estimating eQTL signal from all n=452 samples, as well as in the more power-limiting case when n=89 ancestry-matched samples. (B) sQTLs that colocalize with an eQTL are enriched in u-sQTLs compared to p-sQTLs. If a sQTL was tested for eQTL colocalization in multiple genes (which both overlap the splice junction), it is tallied multiple times. P-value from hypergeometric test for over-representation.

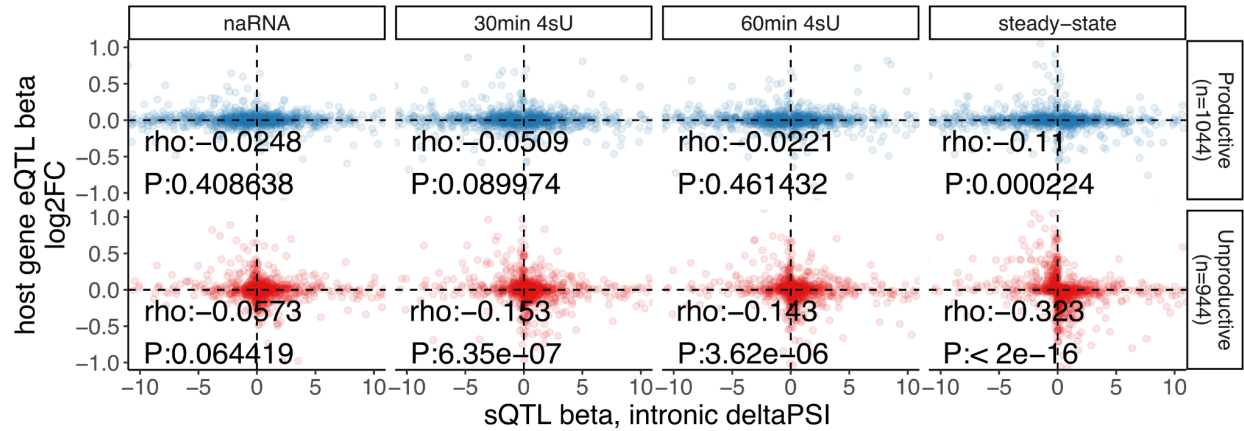


Figure S15. Correlation of splicing and expression- effect sizes amongst u-sQTL and p-sQTLs.

(Top row) Effect size of p-sQTLs (all sQTL introns in LeafCutter cluster are productive splice junctions) versus the effect on host gene expression. eQTL beta is in un-normalized units of log₂ expression fold-change. sQTL beta is in un-normalized units of gene-wise splice-junction PSI (PSI of a junction at the gene level, Methods). (Bottom row) Effect size of u-sQTL (sQTLs which significantly influence at least one unproductive splice junction) versus effect on host gene expression. Correlation between PSI and log₂FC is strongest in steady-state u-sQTLs, consistent with expectation. Effects assessed in each RNA-seq dataset relative to the top sQTL SNP, using the unproductive splice junction for u-sQTLs. Correlations summarized with Spearman rho correlation coefficient and two-sided correlation test P-value.

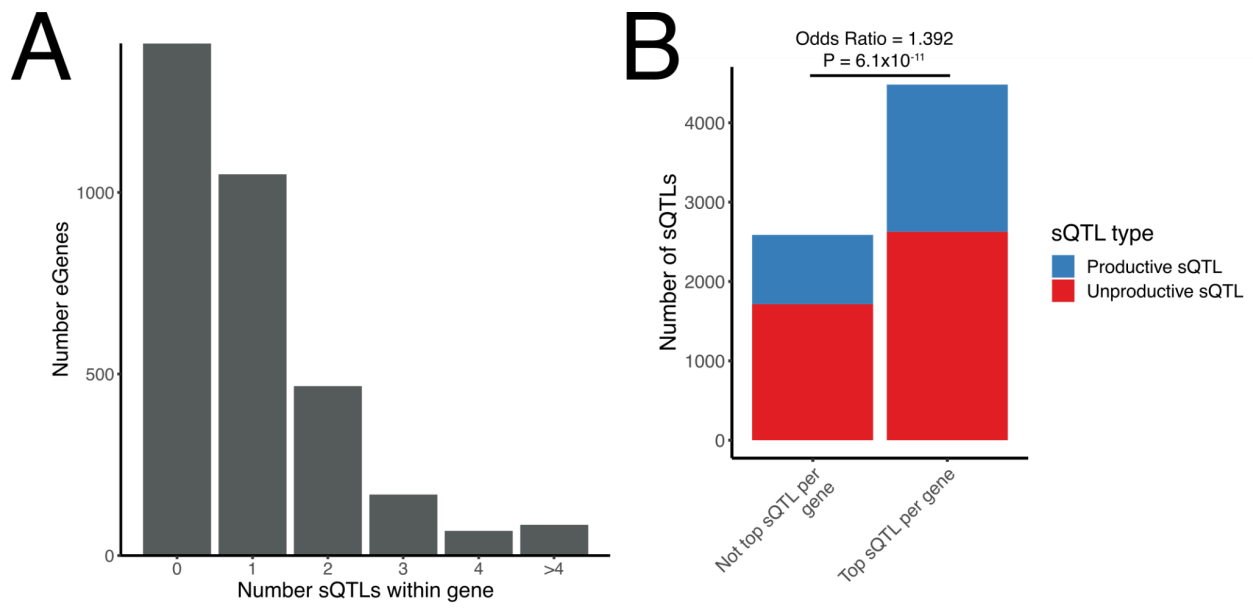


Figure S16. eGenes with multiple sQTLs. (A) Number of sQTLs (distinct sQTL clusters) discovered amongst the 4025 eGenes (steady-state RNA-seq). While there are often more than one independent sQTLs per gene, (C) primary sQTLs (most significant sQTL cluster per gene) are enriched for p-sQTLs over u-sQTLs, compared to non-primary sQTLs. One possible explanation for this enrichment is that unproductive isoforms are rapidly degraded, resulting in attenuated unproductive splice isoform abundances steady-state RNA-seq such that u-sQTL signals are relatively difficult to detect. P value from hypergeometric test for enrichment.

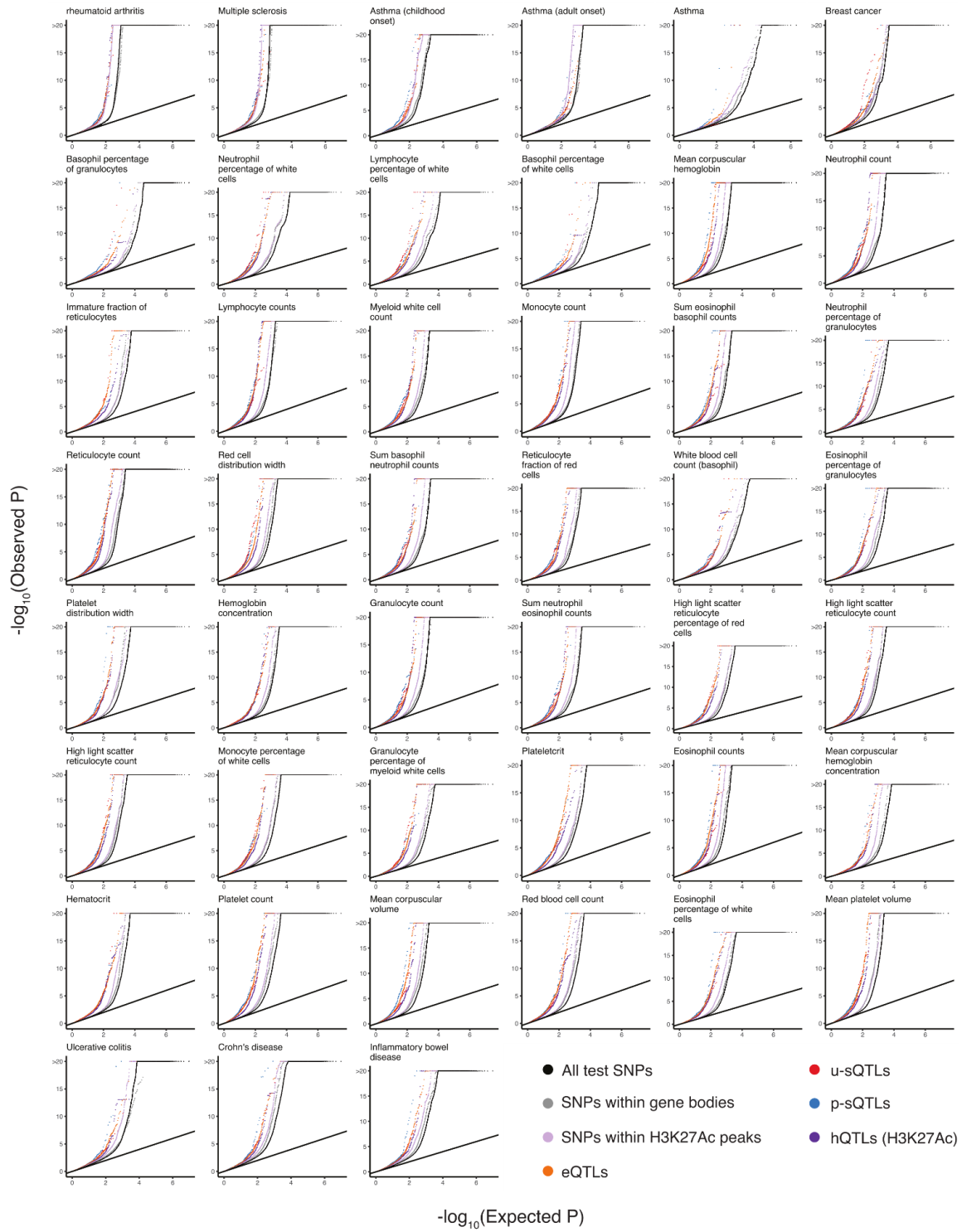


Figure S17. Enrichment of GWAS signal amongst molQTLs. QQ-plot of GWAS association signals for 45 blood and immune related traits, for different sets of SNPs. QTLs refer to the top SNP for each class of molQTLs. u-sQTLs and p-sQTLs represent the sQTLs that affect an unproductive splice junction or only productive (protein-coding) splice junctions, respectively.

Supplementary Methods

Illumina short read RNA-sequencing data

We aligned the Illumina short read RNA-sequencing datasets using STAR²⁶ version 2.7.7a with the same parameters as the ENCODE project for Illumina short read sequencing:

```
--outFilterType BySJout
--outFilterMultimapNmax 20
--alignSJoverhangMin 8
--alignSJDBoverhangMin 1
--outFilterMismatchNmax 999
--outFilterMismatchNoverReadLmax 0.04
--alignIntronMin 20
--alignIntronMax 1000000
--alignMatesGapMax 1000000
```

For the LCL lines, we used STAR's WASP mode²⁷ with the genotype data from the VCF files of the 1000 Genomes Project. In addition to the previously outlined parameters, we added the following:

```
--waspOutputMode SAMtag
--outSAMAttributes NH HI AS nM XS vW
```

After alignment, we filtered the BAM files to retain primary alignments. For the LCL cell lines mapped with STAR WASP mode, we additionally filtered for the alignments that passed the WASP filtering, by retaining those with the vW:i:1 tag.

Histone modification ChIP-seq and CUT&Tag data

To analyze ChIP-seq and CUT&Tag data, we mapped the reads to the human genome using HISAT²⁸ version 2.2.1. We used the `--no-spliced-alignment` flag to prevent the insertion of junction reads. For the paired end ChIP-seq datasets (H3K27ac, H3K4me1 and H3K4me3), we used the `--no-discordant` flag, and allowed a maximum insert size of 1000 bp. We used the Horner (a reimplement of WASP pipeline²⁷, <https://github.com/TheFraserLab/Horner>) `find_intersecting_snps.py` script to find reads that overlap with SNPs for remapping. These reads were realigned using HISAT with the parameters outlined above, and the reads that mapped to different locations were discarded. For the ChIP-seq datasets, we used MACS2²⁹ version 2.2.7.1 to call peaks, using default parameters for paired-end reads. We called narrow peaks for H3K27ac and H3K3me1 data, and broad peaks for H3K4me3.

Molecular trait quantification

We quantified gene expression and histone modification coverage of H3K27ac, H3K4me1 and H3K4me3 using featureCounts³⁰ version 2.0.3. For quantifying gene expression in the LCL RNA-seq datasets, we used the Gencode v34 primary assembly annotation. We used the `--ignoreDup` flag to ignore duplicate reads, and the `--primary` flag to limit the counts to primary alignments only. For steady-state RNA-seq and naRNA-seq we used the `-p` flag for paired end reads. For naRNA-seq data we additionally used the `-s 2` parameter to signal that the data is reversely stranded. For quantifying ChIP-seq coverage of histone modifications, we used the peaks called by MACS2 as the annotation for featureCounts. Unlike other histone modifications, H3K36me3 covers the entire gene body instead of being concentrated at

peaks. We used bedtools multicov to count the number of H3K36me3 CUT&Tag reads overlapping the entire gene body of the defined in the Gencode basic gene annotation for protein-coding genes.

To quantify splicing in the RNA-seq datasets, we extracted junction reads by running regtools version 0.5.2³¹ on the filtered BAM files, requiring a minimum intron length of 20 bp. We merged the resulting .junc files into a single database of observed splice junctions and splice junction read counts across all samples. For alternative splicing analysis, we used the .junc files from the LCL samples to obtain intron clusters using Leafcutter's³² leafcutter_cluster_regtools_py3.py script. We used the read counts for each intron to quantify the percent spliced-in (PSI) of each splice junction in a cluster.

To quantify splicing efficiency of introns, we used SPLICE-q³³ version 1.0.0 to calculate the reverse intron expression ratio (IER) in protein-coding genes using the Gencode v34 basic chromosomal annotation. By default, SPLICE-q uses the highest filtering settings for IER quantification, which selects only introns that do not overlap any exons of the same gene or of any other gene.

Classification of unannotated splice junctions

We developed a method to predict the effect of splice junctions on transcript coding potential. Our method attempts to reconcile junctions identified from short-read RNA-seq with introns of annotated transcripts and predicts whether the junction is compatible with the open reading frame (ORF) of the annotated protein-coding transcript (Gencode v37). Specifically, we classify every annotated intron into one category in the following order of priority: protein_coding > processed_transcript > lncRNA > unprocessed_pseudogene > retained_intron > nonsense_mediated_decay, such that only introns that are uniquely used within transcripts labeled as nonsense_mediated_decay are classified to belong in that category. If an intron belongs to more than one category, it is classified to be in the category with the highest priority.

For splice junctions with coordinates do not match exactly that of an annotated intron, we separated junctions into four different categories for classification:

- (1) Junctions that only overlap with introns within the 5' or 3' UTRs are classified as UTR junctions and are not classified as NMD inducing (though in theory, new junctions in the 3'UTRs may trigger NMD). Additionally, junctions that only overlap with introns from transcripts classified as processed_transcripts, retained_intron or nonsense_mediated_decay are classified as the category with the highest priority. All remaining unannotated junctions that overlap with an intron that flanks an annotated coding exon are classified according to the methods described in (2), (3), or (4).
- (2) Junctions for which both 5' and 3' splice sites (ss) are annotated but are not used by a single intron in an annotated transcript. For these junctions, we translate the resulting mRNA using the frame from the upstream annotated exon CDS until the annotated end of the downstream annotated exon and classify the junction as NMD-inducing if an in-frame stop codon is observed.
- (3) Junctions for which only the 5'ss or 3'ss is annotated. For junctions with annotated 5'ss, we find all overlapping introns of annotated protein-coding transcripts for which the 3'ss is within 60 nt of the junction 3'ss. If no such intron/transcript exists, then the junction is predicted to be NMD inducing. If one or more annotated intron exists, we translate the resulting mRNA up to the end of the annotated downstream exons (or up to an annotated stop codon). We classify the junction as NMD-inducing if all possible resulting mRNA harbor an in-frame stop codon. For junctions with annotated 3'ss, we similarly translate the resulting mRNA, setting the frame from the downstream

annotated exon, and up to the start of the upstream exon. Again, we classify the junction as NMD-inducing if all possible resulting mRNA harbor an in-frame stop codon.

- (4) Junctions for which neither the 5'ss or 3'ss are annotated. Similar to (3), we first attempt to find introns that overlap with these junctions, allowing 60 nt to differ from the 5' or 3' ends or junction ends to be within annotated exons. Junctions that do not fit this criteria are classified as NMD-inducing. For the remainder of junctions, we translate the resulting mRNA and classify the junction as NMD-inducing if the resulting mRNA harbors an in-frame stop codon.

eQTL calling on GTEx gene expression data

To map eQTLs in GTEx data³⁴, we downloaded the publicly available raw count matrices of gene expression in Illumina short read RNA-seq data from the GTEx consortium. We normalized the gene counts for each tissue gene independently, using the standard and rank-normalization of the log2 CPM data described above. For each tissue, we ran QTLtools in nominal pass with the parameters outlined above.

Oxford Nanopore Technologies long read RNA-sequencing data

Long-read Oxford Nanopore Technology sequencing data was obtained from published sources. The total RNA data from double knockdown of *SMG6* and *SMG7* in HeLa cells (SRA accession SAMEA8691113), as well as two control experiments (SRA accessions SAMEA8691110 and SAMEA8691111), was obtained from a previous publication²⁴. Long-read naRNA-seq data in K562 cells (SRA accession SRP171702) was obtained from Drexler et al³⁵. Reads were mapped to the human genome version GRCh38 using minimap 2.24³⁶ preset parameters for spliced long reads: -x splice. We used the flag -a to produce BAM files with CIGAR strings for downstream analysis.

We used a custom python script to extract the splice junctions from each read using the CIGAR string from the BAM files produced by minimap2. Long-read sequencing is prone to high rate of sequencing errors and misalignments. Accordingly, we only considered splice junctions that matched splice junctions observed in protein-coding genes in our short-read analysis. We removed reads that did not match any junctions. Two out of six of the 4sU naRNA samples from Drexler et al had fewer than 10,000 unique long reads left and were removed from downstream analysis. The remaining naRNA 4sU samples had ~20,000 unique long reads or more matching at least one short-read junction.

Reads with at least one NMD-associated splice junction were considered NMD substrates. Due to the variable length of long-reads, not all reads cover a full-length transcript. As a result, the lack of NMD-associated splice junctions in one read does not guarantee that the read comes from a protein-coding transcript. For this reason, we calculated the percent of transcripts confirmed to be NMD substrates (i.e., the percent of transcripts with at least one NMD junction) for reads with 1 to 14 junctions. For any k number of junctions per read, we bootstrapped the reads with k or more junctions to get an estimate of observed NMD transcripts.

We compared the observed percent of NMD transcripts with the theoretical probability of a read presenting an NMD-associated splice junction. The theoretical probability is:

$$\text{Probability that a read has NMD junction} = 1 - (1 - p)^k$$

Where k is the total number of junctions in the read, and p is the proportion of splice junctions that are NMD-associated. To generate shaded regions that show binomial expectation with probability p (related to Figure 2D), we used $p=1.5\%$ to $p=2.5\%$. We justify this with the observation that in short read data we find that (across naRNA samples, a median of) 2.3% of splice junctions are unproductive junctions, which under the simple binomial model would yield $\sim 17\%$ of unproductive full-length transcripts for transcripts with 8 junctions. In long read data, rather than 2.3% of unproductive junctions, we observe a range from 1.8% to 2.3% of unproductive junctions, which we (unnecessarily, but conservatively) round to 1.5-2.5%, to yield a binomial expectation of 11-18% of unproductive full length transcripts. We believe these simple binomial models with simple numbers are faithful to the data while also best communicating the general idea that we wish to convey: that a relatively low mis-splicing rate can compound across multi-intronic genes.

For analyses involving full-length Nanopore reads (related to Extended Data Figure 6) we converted alignments to bed files using `bedtools bamtobed` with `-bed12` flag. We then filtered for reads whose 5' end is within 25nt of an annotated transcription start site (Gencode comprehensive annotations) and whose 3' end is within 50nt of an annotated transcript end site. To prevent spurious splicing alignments, we further filtered out reads that contain a splice junction not observed in any of our short read RNA-seq datasets. We then obtained the reference sequence for each alignment using `bedtools getfasta` and used regular expressions in custom scripts to identify the first annotated open reading frame in each read, and apply the NMDFinderB decision tree³⁷ to classify reads according to the position of the stop codon relative to splice junctions. In addition to the five categories described in NMDFinderB, we also considered two additional classes: (1) No CDS, for cases when there is no annotated start codon within the read, and (2) No stop, for cases when there is a start codon but no in-frame stop codon within the read. To re-assign similar classifications to individual junctions, we used the most common classification of full-length length reads containing the junction of interest, requiring at least three full-length reads. To estimate the degradation efficiency (Extended Data Figure 6C) for junctions of these seven classes, we compared the junction RPM (reads per million for a junction, across all junctions for the denominator) in naRNA versus steady-state RNA-seq, or in NMD dKD versus control. The degradation efficiency estimate is defined as the median $\log_2\text{FC}$ of the junction RPM, relative to the median $\log_2\text{FC}$ of the junction RPM observed for the “Last exon” class (that is, the “Last exon” class is set to 0 by construction, and a value of 1 indicates that the class of junctions is two fold more abundant in naRNA compared to steady-state, relative to the median fold difference observed for “Last exon” junctions).

Scoring splice site strengths with MaxEntScan

We scored the 5' and 3' splice site strength ends of all splice junctions in our data using `maxentropy`, the python wrapper of MaxEntScan³⁸

For each 5' splice sites, we used a 9 nucleotide sequence consisting in 3 nucleotides in the exon and 6 nucleotides in the intron at the splice site. For 3' sites, we used 23 nucleotide sequences, with 3 nucleotides in the exon and 20 nucleotides in the intron.

Ranking of NMD junctions by contribution and entropy calculation

Out of the 14000 protein-coding genes that we selected for downstream analysis, 11563 had NMD associated junction reads in our naRNA-seq data. From these, we selected the 6549 protein-coding genes with an average percent of NMD junction reads between 1 and 20% across all naRNA-seq samples. From these, we selected the genes with a total of one hundred or more NMD junction reads aggregated from the 86 naRNA-seq samples (average of at least 1.16 reads per sample). This resulted in a subset of

6537 protein-coding genes. For a given gene, we obtained the percent of NMD junction reads contributed by each unique NMD splice junction as follows:

$$\text{junction NMD contribution} = 100 \frac{\text{junction NMD reads}}{\sum_{\text{all NMD junctions in gene}} \text{junction NMD reads}}$$

We ranked the junctions by their percent of NMD junction reads contribution. For each gene, we calculated the NMD junction read's entropy as the Shannon entropy of the fraction of the contribution of NMD reads from each junction:

$$H = - \sum_{\text{all junctions } i} \log(p(i))p(i)$$

where

$$p(i) = \frac{\text{junction } i \text{ NMD reads}}{\sum_{\text{all NMD junctions in gene}} \text{junction NMD reads}}$$

Analysis of alternative splicing and symmetry of cassette exons

We downloaded the annotation of alternative splicing events for the human hg38 genome from VastDB³⁹. We selected cassette exon events that fall within the coding region of protein-coding genes. We used VastDB's annotation of the cassette exons to determine whether each exon is symmetric (i.e., if the length of the exon in bp is a multiple of 3).

For the Illumina short read RNA-seq datasets, we calculated the cassette exon PSI as follows:

$$\text{Cassette exon PSI} = 100 \frac{I1 + I2}{I1 + I2 + 2SE}$$

Where I1 and I2 are the splice junction read counts overlapping the first and second splice junction supporting cassette exon inclusion respectively, and SE is the splice junction read counts overlapping the splice junction supporting cassette exon exclusion. For each dataset independently, we retained the cassette exons that have at least one read overlapping any of the three junctions in at least 50% of the samples. To ensure that both junction reads are similarly used in the cassette exons, we discarded the exons in which the average splice junction PSI from I1 and I2 (defined as the average of $100 \cdot I1 / (I1 + I2 + SE)$ and $100 \cdot I2 / (I1 + I2 + SE)$) across all samples respectively; not to be confused with the cassette exon PSI) differs by more than 33%.

For the ONT RNA-seq data, we calculated the cassette exon PSI of each exon as:

$$\text{Cassette exon PSI} = 100 \frac{I}{I + SE}$$

Where I is the total number of reads that contain both splice junctions supporting inclusion of the cassette exon on the same read, and SE is the total number of reads containing the splice junction supporting exclusion of the cassette exon.

We used these cassette exon PSI calculations to estimate the percent of symmetric cassette exons at different PSI ranges.

Measurements of unproductive junctions in Illumina short-read data

Through this paper we used multiple measurements of alternative splicing for different purposes:

- The **percent of unproductive splice junction reads per gene** correspond to the total number of unproductive splice junction reads in a gene divided by the total number of splice junction reads in the gene.
- The **PSI of a junction at the gene level** is the total number of reads mapping to that junction divided by the maximum number of reads mapping to any junction on the same gene.
- The **PSI of a junction at the Leafcutter intron cluster level** is the total number of reads mapping to that junction divided by the total number of reads mapping to any junction on the same intron cluster.
- The **PSI of cassette exons**, as described in the previous section.
- The **percent of unproductive transcripts** in long read sequencing data, calculated as the percent of transcripts that present one or more unproductive splice junctions.

π_1 sharing of eQTLs between RNA-seq datasets

Storey’s π_1 statistic⁴⁰, an estimate of the fraction of non-null hypothesis from a distribution of p-values, was used to estimate the fraction of features discovered one dataset (the “discovery dataset”) that have non-null effects in another dataset (the “ascertainment dataset”). For example, one may ask, “what fraction of eQTLs discovered in steady state RNA are eQTLs in naRNA?”. The P-value for each discovery eQTL (the nominal P-value for the top SNP:gene pair reported by QTLtools for each gene that passes a false discovery threshold for the genewise permutation test reported by QTLtools) was assessed in the ascertainment dataset to produce a distribution of P-values. π_1 (the complement of π_0), was estimated using the `pi0est` function from the `qvalue` package in R:

```
pi1 = 1 - pi0_est(p, pi0.method='bootstrap').
```

π_1 sharing of eQTLs and hQTLs at TSS

For assessing the sharing between hQTLs and eQTLs, we assessed hQTLs at peaks within 500bp of annotated TSS (the 5’ most end of Gencode ‘basic’-tagged transcript structures) as the promoter for each gene. There is often more than one such TSS for each gene. Furthermore, there are multiple hQTL assays (i.e., H3K27ac ChIP-seq, H3K4me3 ChIP-seq) which we collectively consider as an ascertainment feature for each discovery feature. Therefore, there is not a one-to-one mapping of discovery features to ascertainment features for purposes of estimating π_1 . To answer the question, “what fraction of eQTLs have at least one hQTL?”, we modified the approach used to estimate π_1 among eQTLs in RNA-seq datasets as follows to obtain a one-to-one mapping of discovery and ascertainment features: The minimum P-value among the n QTL test peaks that correspond to the promoter of a gene was considered a test statistic for obtaining a single ascertainment P-value for each gene. The distribution function of this test statistic under the null was estimated by repeatedly calculating $\min(x_1, x_2, \dots, x_n)$, where $x \sim \text{Uniform}(0,1)$. An empirical cumulative distribution function of this test statistic was calculated after 10,000 repetitions of this process to obtain a single ascertainment P-value for each gene, and π_1 was estimated as above.

Colocalization of molQTLs

We simultaneously assessed colocalization of molQTLs around each gene using `hyprcoloc`⁴¹. The following molQTL features were jointly considered for colocalization around each gene: Each H3K4me1, H3K4me3, & H3K27ac hQTL within 100kb of the gene, sQTLs introns fully contained in the gene body, H3K36me3 hQTLs (corresponding to the gene body), apaQTLs (features fully contained in the gene body), and eQTLs for the corresponding gene in each RNA-seq dataset (steady state RNA, 4sU 30m, 4sU 60m, naRNA). Summary statistics for a 100kb cis-window surrounding the gene were obtained for each molQTL were obtained using QTLtools nominal pass. Only molQTLs with a permutation pass P value < 0.1 were considered for colocalization. The `hyprcoloc::hyprcoloc` function was used in R with default settings to report clusters of colocalized molQTLs around each gene that satisfy a regional probability threshold ($P_r^*=0.5$ by default, corresponding to the probability threshold that all traits in the `hyprcoloc` iteration/cluster contain an association with a SNP) and regional alignment probability threshold ($P_a^*=0.5$ by default, corresponding to the probability threshold that all associations in the `hyprcoloc` iteration/cluster are aligned at a putative single causal SNP).

Enrichment of genomic annotations amongst QTLs

Genomic annotations include: ChromHMM annotations⁴² downloaded from UCSC genome browser based on ENCODE data from GM12878 LCL cell line, APA test peaks, splice donor regions (-3 to +7 from 5'ss), splice acceptor regions (-10 to 0 from 3'ss) and branchpoint regions (-40 to -10 3'ss) for annotated and unannotated splice sites, and miRNA binding sites (Downloaded from TargetScan v8.0⁴³). The fine-mapping posterior inclusion probabilities (PIP) output by `hyprcoloc` were obtained for clusters containing an eQTL and a hQTL (transcriptional eQTLs), and clusters that contain an eQTL but no hQTL (post transcriptional eQTLs). Fold enrichment was calculated as the total fraction of PIP in a genomic annotation for transcriptional eQTLs, compared to post-transcriptional eQTLs. Confidence intervals were bootstrapped by 1000 resamples of the sets of transcriptional and post-transcriptional eQTLs.

Gene set enrichment and characteristics of genes with risdiplam-induced exons

Gene sets were defined as follows: “Disease genes” are defined as genes from the OMIM genemap database²⁵ with a string match to ‘dominant’ in the phenotype column, reasoning that down-regulation of dominant negative genes is a reasonable shorthand to identify potential disease-modifying perturbations. Kinases (GO:0016301), transcription factors (GO:0003700), nuclear hormone receptors (GO:0016301), G-protein coupled receptors [GO:0004930 that are not in the set of olfactory receptors (GO:0004984)], and ion channels (GO:0015075) were downloaded from Msigdb¹⁷. Gene sets were filtered for tested expressed genes considered in differential splicing and differential expression testing. Significant enrichment of differentially expressed or spliced genes was assessed using a hypergeometric test. Gene targets of FDA approved small molecules (“Tier 1” in the source publication) were obtained from a previous publication⁴⁴. To determine the gene length and number of exons in genes hosting risdiplam-induced genes, we considered the number of exons and exonic gene length highest expressed Gencode isoform (basic annotations) using Salmon⁴⁵ version 1.10. For comparisons involving an expression-matched set of control genes, we used Salmon gene quantifications (sum of all transcripts) to identify the two non-test genes closest in expression to each gene in the test set.

References

1. Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. *Science* 352, 600–604. <https://doi.org/10.1126/science.aad9417>.
2. Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares, M. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.* 21, 708–718. <https://doi.org/10.1101/gad.1525507>.
3. Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* 446, 926–929. <https://doi.org/10.1038/nature05676>.
4. Yan, Q., Weyn-Vanhentenryck, S.M., Wu, J., Sloan, S.A., Zhang, Y., Chen, K., Wu, J.Q., Barres, B.A., and Zhang, C. (2015). Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. *Proc. Natl. Acad. Sci.* 112, 3445–3450. <https://doi.org/10.1073/pnas.1502849112>.
5. Sureau, A. (2001). SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. *EMBO J.* 20, 1785–1796. <https://doi.org/10.1093/emboj/20.7.1785>.
6. Lareau, L.F., and Brenner, S.E. (2015). Regulation of Splicing Factors by Alternative Splicing and NMD Is Conserved between Kingdoms Yet Evolutionarily Flexible. *Mol. Biol. Evol.* 32, 1072–1079. <https://doi.org/10.1093/molbev/msv002>.
7. Jumaa, H. (1997). The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. *EMBO J.* 16, 5077–5085. <https://doi.org/10.1093/emboj/16.16.5077>.
8. Änkö, M.-L., Müller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., and Neugebauer, K.M. (2012). The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol.* 13, R17. <https://doi.org/10.1186/gb-2012-13-3-r17>.
9. Brugiolo, M., Botti, V., Liu, N., Müller-McNicoll, M., and Neugebauer, K.M. (2017). Fractionation iCLIP detects persistent SR protein binding to conserved, retained introns in chromatin, nucleoplasm and cytoplasm. *Nucleic Acids Res.* 45, 10452–10465. <https://doi.org/10.1093/nar/gkx671>.
10. Leclair, N.K., Brugiolo, M., Urbanski, L., Lawson, S.C., Thakar, K., Yurieva, M., George, J., Hinson, J.T., Cheng, A., Graveley, B.R., et al. (2020). Poison Exon Splicing Regulates a Coordinated Network of SR Protein Expression during Differentiation and Tumorigenesis. *Mol. Cell* 80, 648–665.e9. <https://doi.org/10.1016/j.molcel.2020.10.019>.
11. Saltzman, A.L., Kim, Y.K., Pan, Q., Fagnani, M.M., Maquat, L.E., and Blencowe, B.J. (2008). Regulation of Multiple Core Spliceosomal Proteins by Alternative Splicing-Coupled Nonsense-Mediated mRNA Decay. *Mol. Cell. Biol.* 28, 4320–4330. <https://doi.org/10.1128/MCB.00361-08>.
12. Rappsilber, J., Ajuh, P., Lamond, A.I., and Mann, M. (2001). SPF30 Is an Essential Human Splicing Factor Required for Assembly of the U4/U5/U6 Tri-small Nuclear Ribonucleoprotein into the Spliceosome. *J. Biol. Chem.* 276, 31142–31150. <https://doi.org/10.1074/jbc.M103620200>.
13. Kino, Y., Washizu, C., Kurosawa, M., Oma, Y., Hattori, N., Ishiura, S., and Nukina, N. (2015). Nuclear localization of MBNL1: splicing-mediated autoregulation and repression of

- repeat-derived aberrant proteins. *Hum. Mol. Genet.* 24, 740–756.
<https://doi.org/10.1093/hmg/ddu492>.
14. Zeng, T., Spence, J.P., Mostafavi, H., and Pritchard, J.K. (2023). Bayesian estimation of gene constraint from an evolutionary model with gene features. Preprint,
<https://doi.org/10.1101/2023.05.19.541520> <https://doi.org/10.1101/2023.05.19.541520>.
 15. Xie, Z., Bailey, A., Kuleshov, M.V., Clarke, D.J.B., Evangelista, J.E., Jenkins, S.L., Lachmann, A., Wojciechowski, M.L., Kropiwnicki, E., Jagodnik, K.M., et al. (2021). Gene Set Knowledge Discovery with Enrichr. *Curr. Protoc. I*, e90. <https://doi.org/10.1002/cpz1.90>.
 16. Fang, Z., Liu, X., and Peltz, G. (2023). GSEApY: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* 39, btac757.
<https://doi.org/10.1093/bioinformatics/btac757>.
 17. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. <https://doi.org/10.1093/bioinformatics/btr260>.
 18. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. <https://doi.org/10.1073/pnas.0506580102>.
 19. Huang, N., Lee, I., Marcotte, E.M., and Hurles, M.E. (2010). Characterising and Predicting Haploinsufficiency in the Human Genome. *PLOS Genet.* 6, e1001154.
<https://doi.org/10.1371/journal.pgen.1001154>.
 20. Grubert, F., Zaugg, J.B., Kasowski, M., Ursu, O., Spacek, D.V., Martin, A.R., Greenside, P., Srivas, R., Phanstiel, D.H., Pekowska, A., et al. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* 162, 1051–1065.
<https://doi.org/10.1016/j.cell.2015.07.048>.
 21. Mittleman, B., Pott, S., Warland, S., Zeng, T., Kaur, M., Gilad, Y., and Li, Y.I. (2019). Alternative polyadenylation mediates genetic regulation of gene expression (Genomics)
<https://doi.org/10.1101/845966>.
 22. The Geuvadis Consortium, Lappalainen, T., Sammeth, M., Friedländer, M.R., ‘t Hoen, P.A.C., Monlong, J., Rivas, M.A., González-Porta, M., Kurbatova, N., Griebel, T., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511. <https://doi.org/10.1038/nature12531>.
 23. Colombo, M., Karousis, E.D., Bourquin, J., Bruggmann, R., and Mühlemann, O. (2017). Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways. *RNA* 23, 189–201.
<https://doi.org/10.1261/rna.059055.116>.
 24. Karousis, E.D., Gypas, F., Zavolan, M., and Mühlemann, O. (2021). Nanopore sequencing reveals endogenous NMD-targeted isoforms in human cells. *Genome Biol.* 22, 223. <https://doi.org/10.1186/s13059-021-02439-3>.
 25. Amberger, J.S., Bocchini, C.A., Scott, A.F., and Hamosh, A. (2019). OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* 47, D1038–D1043. <https://doi.org/10.1093/nar/gky1151>.
 26. 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. <https://doi.org/10.1093/bioinformatics/bts635>.
 27. Van De Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP:

- allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* 12, 1061–1063. <https://doi.org/10.1038/nmeth.3582>.
28. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. <https://doi.org/10.1038/s41587-019-0201-4>.
 29. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137. <https://doi.org/10.1186/gb-2008-9-9-r137>.
 30. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. <https://doi.org/10.1093/bioinformatics/btt656>.
 31. Cotto, K.C., Feng, Y.-Y., Ramu, A., Richters, M., Freshour, S.L., Skidmore, Z.L., Xia, H., McMichael, J.F., Kunisaki, J., Campbell, K.M., et al. (2023). Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nat. Commun.* 14, 1589. <https://doi.org/10.1038/s41467-023-37266-6>.
 32. Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* 50, 151–158. <https://doi.org/10.1038/s41588-017-0004-9>.
 33. De Melo Costa, V.R., Pfeuffer, J., Louloui, A., Ørom, U.A.V., and Piro, R.M. (2021). SPLICE-q: a Python tool for genome-wide quantification of splicing efficiency. *BMC Bioinformatics* 22, 368. <https://doi.org/10.1186/s12859-021-04282-6>.
 34. The GTEx Consortium, Aguet, F., Anand, S., Ardlie, K.G., Gabriel, S., Getz, G.A., Graubert, A., Hadley, K., Handsaker, R.E., Huang, K.H., et al. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330. <https://doi.org/10.1126/science.aaz1776>.
 35. Drexler, H.L., Choquet, K., and Churchman, L.S. (2019). Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Mol. Cell*, S1097276519308652. <https://doi.org/10.1016/j.molcel.2019.11.017>.
 36. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma. Oxf. Engl.* 34, 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
 37. Lindeboom, R.G.H., Vermeulen, M., Lehner, B., and Supek, F. (2019). The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. *Nat. Genet.* 51, 1645–1651. <https://doi.org/10.1038/s41588-019-0517-5>.
 38. Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 11, 377–394. <https://doi.org/10.1089/1066527041410418>.
 39. Tapial, J., Ha, K.C.H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Valli eres, M., Permanyer, J., Sodaei, R., Marquez, Y., et al. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res.* 27, 1759–1768. <https://doi.org/10.1101/gr.220962.117>.
 40. Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* 100, 9440–9445. <https://doi.org/10.1073/pnas.1530509100>.
 41. Foley, C.N., Staley, J.R., Breen, P.G., Sun, B.B., Kirk, P.D.W., Burgess, S., and Howson, J.M.M. (2021). A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat. Commun.* 12, 764.

<https://doi.org/10.1038/s41467-020-20885-8>.

42. Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* *12*, 2478–2492. <https://doi.org/10.1038/nprot.2017.124>.
43. McGeary, S.E., Lin, K.S., Shi, C.Y., Pham, T.M., Bisaria, N., Kelley, G.M., and Bartel, D.P. (2019). The biochemical basis of microRNA targeting efficacy. *Science* *366*, eaav1741. <https://doi.org/10.1126/science.aav1741>.
44. Finan, C., Gaulton, A., Kruger, F.A., Lumbers, R.T., Shah, T., Engmann, J., Galver, L., Kelley, R., Karlsson, A., Santos, R., et al. (2017). The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* *9*, eaag1166. <https://doi.org/10.1126/scitranslmed.aag1166>.
45. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* *14*, 417–419. <https://doi.org/10.1038/nmeth.4197>.