Global impact of unproductive splicing on human gene expression

Benjamin Fair [1,7], Carlos F. Buen Abad Najar [1,7], Junxing Zhao[2,5], Stephanie Lozano [1,6], Austin Reilly[1], Gabriela Mossian[1], Jonathan P. Staley [3], Jingxin Wang[2,5] & Yang I. Li [1,4] ✉

Alternative splicing (AS) in human genes is widely viewed as a mechanism for enhancing proteomic diversity. AS can also impact gene expression levels without increasing protein diversity by producing 'unproductive' transcripts that are targeted for rapid degradation by nonsense-mediated decay (NMD). However, the relative importance of this regulatory mechanism remains underexplored. To better understand the impact of AS–NMD relative to other regulatory mechanisms, we analyzed population-scale genomic data across eight molecular assays, covering various stages from transcription to cytoplasmic decay. We report threefold more unproductive splicing compared with prior estimates using steady-state RNA. This unproductive splicing compounds across multi-intronic genes, resulting in 15% of transcript molecules from protein-coding genes being unproductive. Leveraging genetic variation across cell lines, we find that GWAS trait-associated loci explained by AS are as often associated with NMD-induced expression level differences as with differences in protein isoform usage. Our findings suggest that much of the impact of AS is mediated by NMD-induced changes in gene expression rather than diversification of the proteome.

Alternative splicing (AS) has the potential to expand the number of functional peptides encoded in messenger RNA. Large-scale transcriptomics studies have confirmed that nearly all protein-coding genes generate multiple—sometimes dozens—of distinct mRNA isoforms. This finding is often interpreted as supporting the role of AS in diversifying the proteome; yet, most alternatively spliced isoforms are lowly expressed and lack cross-species conservation[1–6]. To explain these observations, multiple studies have suggested that the vast majority of isoforms are nonfunctional transcripts resulting from mis-splicing rather than regulated AS[4–9].

Mis-splicing from aberrant activation of unconserved 'cryptic' splice sites often introduces frameshifts in the mRNA coding sequence,

resulting in premature termination codons (PTCs). Consequently, downstream exon junction complexes, which would normally be displaced by translating ribosomes, recruit nonsense-mediated decay (NMD) machinery to the mRNA for degradation. Thus, most transcripts with one or more aberrant splicing events are considered to be 'unproductive', as they are expected to undergo rapid NMD.

Unproductive transcripts can also result from regulated AS[10–21]. For example, some splicing factors control AS of their own pre-mRNA, relying on the coupling between AS and NMD (AS–NMD) to autoregulate their expression levels[12–14,17,21]. However, regulated AS–NMD has only been documented in a handful of genes. While most genes have annotated unproductive isoform structures[22], the extent to which these

[1]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA. [2]Department of Medicinal Chemistry, University of Kansas, Lawrence, KS, USA. [3]Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago, IL, USA. [4]Department of Human Genetics, University of Chicago, Chicago, IL, USA. [5]Present address: Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA. [6]Present address: Center for Neuroscience, University of California Davis, Davis, CA, USA. [7]These authors contributed equally: Benjamin Fair, Carlos F. Buen Abad Najar. ✉e-mail: yangili1@uchicago.edu

isoforms influence gene expression levels, impact phenotypes and/or are important for organismal fitness is unknown. Assessing these questions is complicated partly because the rapid decay of unproductive isoforms obscures quantitative measurements of their splicing[7,23,24].

An early study of unproductive splicing estimated that up to a third of transcript isoforms inferred from expressed sequence tags supported unproductive rather than productive splicing and hypothesized that AS–NMD may be a widespread regulatory mechanism[7]. Since then, multiple studies have attempted to test this hypothesis using improved methods, for example, by measuring gene expression levels before and after knocking down core NMD factors such as UPF1 or UPF2. These studies revealed a modest impact on gene expression levels, noting that only a small fraction (<10%) of genes show appreciable change in gene expression levels upon knockdown[25,26], providing evidence against a widespread role of AS–NMD[27]. By contrast, recent studies[23,24,28–32] support partial redundancy between core NMD factors, which can obscure knockdowns of single NMD factors and underestimate the impact of AS–NMD on mRNA expression levels. Thus, the impact of AS on gene expression levels remains unclear[10,15,33–36].

Results

High-throughput measurements of AS before mRNA decay

To assess the impact of AS on steady-state gene expression levels, we must jointly consider multiple stages of gene regulation that reflect mRNA before and after the influence of cytoplasmic decay processes. To do this, we leveraged a large collection of molecular assays in lymphoblastoid cell lines (LCLs) derived primarily from 40–86 Yoruba individuals (Fig. 1a and Supplementary Fig. 1). These datasets[37–40] have been used to study the impact of genetic variants on molecular phenotypes and consists of measurements tracking major steps of mRNA biogenesis including chromatin activity at enhancers (H3K4me1 and H3K27ac combining chromatin immunoprecipitation sequencing (ChIP-seq)) and promoters (H3K27ac and H3K4me3 ChIP-seq), newly transcribed polyA RNAs (4sU pulse-labeled for 30 or 60 min) and steady-state mRNA levels (RNA sequencing, RNA-seq). However, these data fail to capture spliced mRNA before potential cytoplasmic degradation, preventing us from capturing rapidly degraded mRNA transcripts.

To measure AS of pre-mRNA splicing before mRNA decay, we used nascent RNA-seq (naRNA-seq). We obtained a total of 22.4 billion naRNA-seq reads across 86 LCLs. To further increase the temporal resolution of our dataset, we also collected CUT&Tag data from 95 LCLs to profile H3K36me3, a mark associated with active transcription elongation (Supplementary Fig. 2).

The correlation between gene expression measurements in steady-state RNA and gene expression measurement at previous stages of RNA processing reveals a clear temporal pattern, matching our expectation that our naRNA-seq data capture mRNA at an earlier stage

of maturation than 4sU sequencing labeled for 30 min (Fig. 1a). We further verified that naRNA-seq captures RNA associated with chromatin and nascent pre-mRNA in several ways (Extended Data Fig. 1), showing that our naRNA dataset allows high resolution analysis of splicing outcomes before cytoplasmic degradation.

Unproductive mRNA splicing is pervasive

We used naRNA-seq data to estimate the prevalence of 'unproductive splicing', that is, splicing outcomes that are expected to induce NMD of the host transcript. In contrast, 'productive splicing' is expected to preserve the proper reading frame of the host transcript. Thus, exon–exon junction reads from unproductive transcripts are expected to be depleted in steady-state RNA-seq, while the junction reads identified in naRNA would reflect unbiased rates of unproductive splicing. We find 0.44% of all junction reads overlapping protein-coding genes from naRNA-seq map to splice junctions uniquely annotated in NMD-targeted transcripts (Gencode annotations[22]) compared with 0.15% in steady-state RNA (Extended Data Fig. 2a). Moreover, we find that 2.4% of junction reads in naRNA are not attributable to annotated junctions of stable protein-coding transcripts, compared with 1.1% in 30 min labeled 4sU RNA and 0.60% in steady-state RNA (Fig. 1b). As only a small fraction of these 2.4% of junction reads can be attributed to annotated NMD transcripts, we sought to understand whether the remaining represent unproductive isoforms. We find that about half of these remaining junctions are annotated in transcripts that are expected to be NMD substrates, albeit, not explicitly defined as so by Gencode (for example, transcripts associated with a retained intron) (Fig. 1b, Extended Data Fig. 2b,c and Methods). Most of these splice junctions are greatly depleted in steady-state RNA (Supplementary Fig. 3a), consistent with their rapid decay. To categorize the unannotated splice junctions, we developed a method to predict their effect on transcript coding potential (Supplementary Methods). We find that these unannotated junctions are overwhelmingly expected to either frameshift or introduce a PTC in the coding sequence, thus, resulting in a transcript targeted by NMD. To confirm the quality of our productive/unproductive categorizations of splice junctions, we assessed the change in abundance of splice junctions in short hairpin RNA (shRNA)-induced knockdowns of core NMD components, by reprocessing RNA-seq data from single knockdowns of UPF1, SMG6 and SMG7 and a double knockdown (dKD) of SMG6 and SMG7 (ref. 24). As expected, annotated and unannotated splice junctions that we classify as productive are relatively unchanged upon NMD knockdown, while unproductive junctions increase in abundance (Supplementary Fig. 3b). Single knockdown of core NMD factors displayed much smaller abundances of unproductive splice junctions compared with double SMG6 and SMG7 knockdowns (Extended Data Fig. 3), suggesting functional redundancy between these NMD factors. Overall, we find a similar
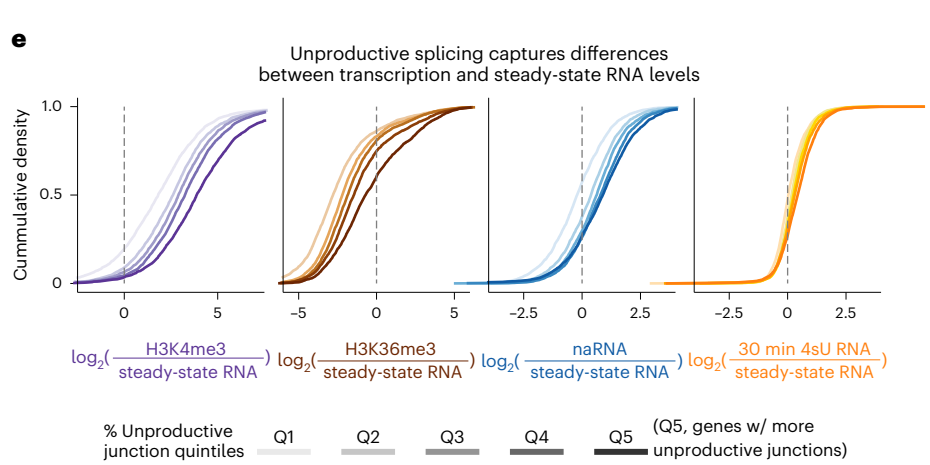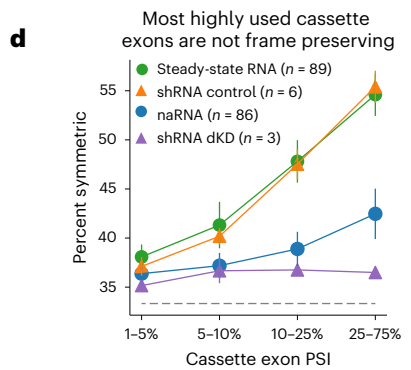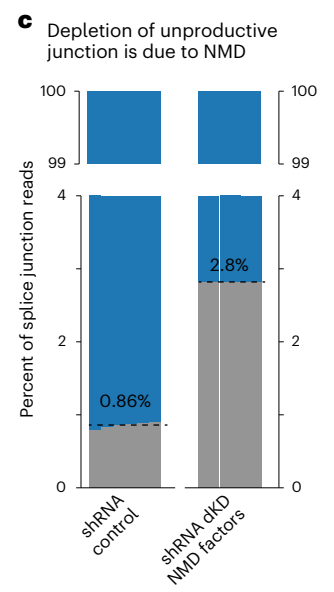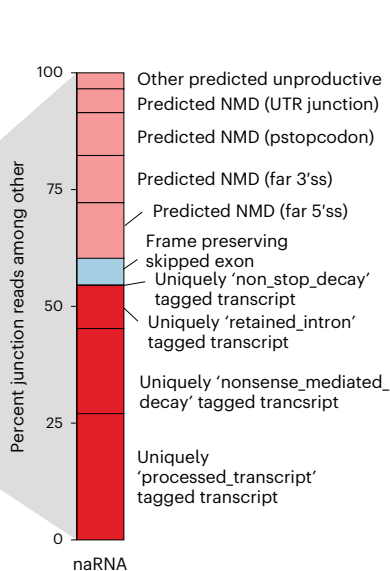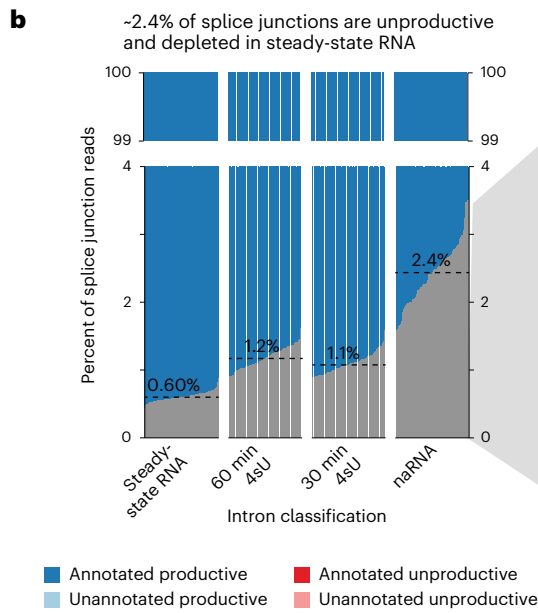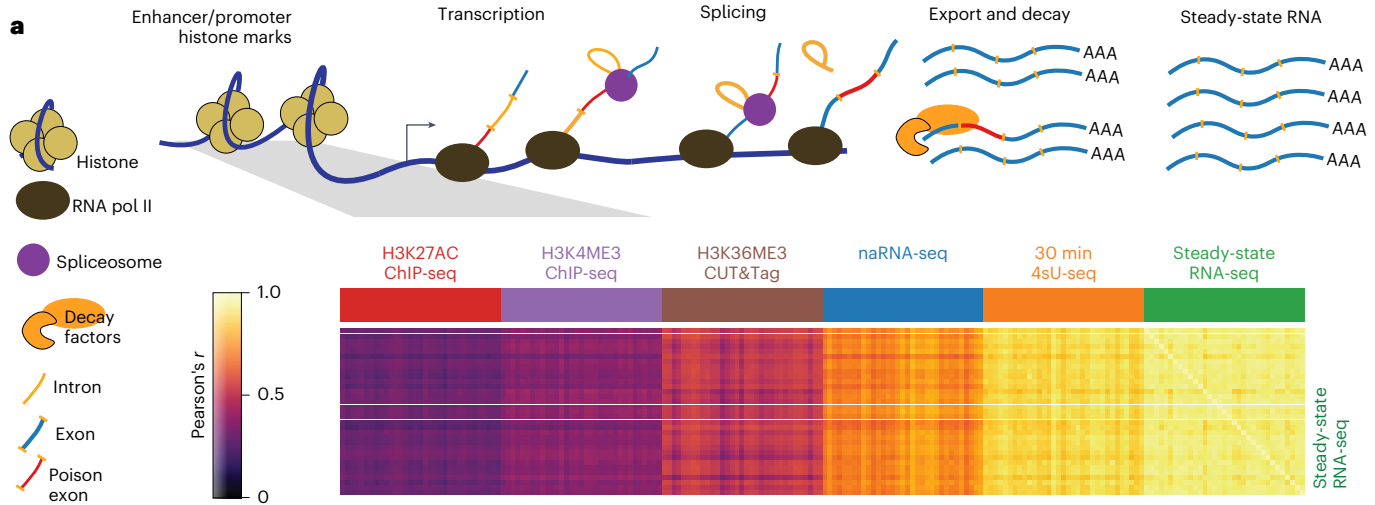

Fig. 1 | Genomic data captured before, during, and after transcription reveal an abundance of NMD isoforms. a, Subset of the population-scale datasets we analyzed, covering stages of mRNA biogenesis, from activation of enhancers/promoters (that is, H3K27ac and H3K4me3 ChIP-seq) to steady-state RNA (polyA RNA-seq). The gene expression correlation matrix (using promoter peak coverage for H3K27ac and H3K4me3, gene body coverage for H3K36me3 and exonic read coverage for RNA-seq) relative to steady-state RNA samples is shown as a heat map. **b,** Left: fraction of splice junction reads in each RNA-seq sample (columns, grouped by dataset) that are in Gencode-annotated productive transcript structures (blue) versus unannotated or annotated unproductive transcript structures (gray). The dashed lines indicate the median for each dataset. Right: for the 2.4% of splice junctions in naRNA-seq data that are not in annotated productive transcript structures, we checked for their unique presence in annotated unproductive transcript structures (for example, transcripts tagged by Gencode as 'retained_intron'), or if unannotated, we attempted to translate sequences surrounding the splice sites (Supplementary Methods). The stacked bars indicate fraction

of naRNA-seq splice junctions in each subcategory. **c,** Similar to **b**, comparing steady-state RNA from shRNA scramble control (n = 6) and shRNA dKD (n = 3) of SMG6 and SMG7 in HeLa cells[24]. **d,** Fraction of cassette exons that are symmetric (that is, length divisible by three) as function of their usage, estimated as percent spliced in PSI. The error bars represent standard error of values for LCL lines treated as replicates (circular markers, same dataset as **b**) and standard error across replicate shRNA knockdown experiments (triangular markers, same dataset as **c**). **e,** Cumulative distribution of log fold differences in steady-state gene expression versus gene transcription (measured by H3K4me3 promoter activity, naRNA, H3K36me3 or 30 m 4sU-labeled RNA), a proxy for degradation rate. Genes are grouped by quintiles based on percent of unproductive junction reads. The quintile of genes with the most unproductive splicing (darkest) show the strongest signature of mRNA degradation. The correlation between unproductive splicing and 30 min labeled 4sU RNA/ steady-state RNA is weaker than in comparisons using the other degradation rate proxies, consistent with rapid decay of unproductive transcripts.

enrichment of unproductive splice junctions in the dKD experiments (Fig. 1c) as in our naRNA (Fig. 1b), suggesting NMD is the primary mechanism explaining the abundance of unannotated junctions in naRNA.

Thus, we estimate that ~2.3% of splicing events target transcripts for NMD, as measured in naRNA, compared with ~0.55% in steady-state RNA (Extended Data Fig. 2c).

a Population-scale measurements of transcription and post-transcriptional RNA processing

b ~2.4% of splice junctions are unproductive and depleted in steady-state RNA

c Depletion of unproductive junction is due to NMD

d Most highly used cassette exons are not frame preserving

e Unproductive splicing captures differences between transcription and steady-state RNA levels

$$\log_2\left(\frac{H3K4me3}{\text{steady-state RNA}}\right) \quad \log_2\left(\frac{H3K36me3}{\text{steady-state RNA}}\right) \quad \log_2\left(\frac{naRNA}{\text{steady-state RNA}}\right) \quad \log_2\left(\frac{30 \text{ min } 4sU \text{ RNA}}{\text{steady-state RNA}}\right)$$

We further describe the abundance of unproductive splicing revealed by naRNA in a qualitative manner by analyzing exons whose length is divisible by three ('symmetric exons') and, thus, are frame preserving whether the exon is skipped or included. Previous studies using steady-state RNA-seq data found that highly included and conserved alternatively spliced exons ('cassette exons') are biased toward being symmetric[1–4,41], suggesting that there is selective pressure to maintain coding frame for highly included cassette exons (Extended Data Fig. 4a). Indeed, ~55% of highly included cassette exons are symmetric compared with only ~35% of rarely used cassette exons (percent spliced in (PSI) <1%). However, we find that in naRNA the fraction of symmetric exons is low, under 40%, even for the most highly included cassette exons (Fig. 1d and Extended Data Fig. 4b). We find similar results in steady-state RNA after knockdown of NMD factors (Fig. 1d and Extended Data Fig. 4b). These observations show that in the absence of NMD, there is no bias for alternatively spliced exons to be symmetric. These findings question the widely held belief that the main function of AS is protein diversification.

Having observed such abundant unproductive splicing, we wondered to what extent AS–NMD influences gene expression levels genome wide. We find that the genewise fraction of unproductive splicing correlates with multiple measures of degradation rate (Fig. 1e), and we conservatively estimate that AS–NMD explains at least 9% of post-transcriptional gene expression variance across genes (Supplementary Note 1).

Despite the abundant effects of AS–NMD on gene expression levels, it is not clear to what extent AS–NMD is functionally regulated and optimized through selection versus a result of splicing mistakes, which nonetheless impact gene expression. To better understand this, we explored which classes of genes and introns are most prone to AS–NMD. We found that highly expressed (Extended Data Fig. 5a,b) and evolutionarily constrained genes (Extended Data Fig. 5c) have among the lowest rates of unproductive splicing. Because evolutionarily constrained genes are expected to be under strong selective pressure to be tightly and accurately regulated[42], the fact that we see the least unproductive splicing in them suggests that unproductive splicing generally represents molecular noise, rather than a form of regulation. However, we find exceptions—for example, the highly conserved SR-family proteins have among the highest rates of unproductive splicing even when controlling for expression (Extended Data Fig. 5b and Supplementary Note 2), consistent with previous reports of conserved AS–NMD-based autoregulatory loops in these genes[12–14,17,19,21]. Still, splicing regulators represent only a small fraction of genes with the highest rates of unproductive splicing (Supplementary Note 2). Further, we observe that introns with weak splice sites (Supplementary Fig. 4) and long introns (Fig. 2a) have higher unproductive splicing rates, consistent with splicing errors caused by appreciable competition between bona fide functional splice sites and cryptic splice sites. We also reasoned that if unproductive splicing is regulated and optimized, most unproductive splicing would be derived from a single alternative event that could be efficiently regulated by cis-regulatory elements, as is the case for SR protein autoregulatory loops. Yet, for most genes, we were unable to attribute a majority of unproductive splice events to a single splice junction (Fig. 2b and Supplementary Fig. 5).

Based on these observations, we propose a model in which every protein-coding intron of a gene can generate unproductive junctions at a specific 'error' rate, which depends on intron length, splice site strength and other genomic features. This model predicts that unproductive junctions will compound along the entire length of pre-mRNA, making long genes—or genes with many introns—more likely to produce unproductive transcripts than those with fewer introns (Fig. 2c). Under the simplest binomial model, assuming that 2.3% of splicing events are unproductive, we predicted that ~17% of mRNA transcripts would be unproductive for a typical human gene with eight introns. To test this prediction, we analyzed long-read RNA-seq (LRS) data from

datasets collected using Oxford Nanopore technology[23,43]. We investigated the number of transcripts containing one or more unproductive junctions, binned LRS reads by their number of junctions and calculated the fraction of reads with an unproductive junction for each bin to estimate the abundance of unproductive transcripts. We first analyzed LRS naRNA from chromatin cell fractions from K562 cells[43]. We found that ~15% of naRNA reads spanning a typical human gene contained one or more unproductive junctions (Fig. 2d). Importantly, we also investigated LRS of steady-state mRNA following shRNA knockdown of NMD machinery[23] and found that ~15% of reads spanning a typical human gene had one or more unproductive junctions (Fig. 2d). We further corroborated these findings by analyzing the small subset of LRS reads that probably represent full-length mRNA transcripts, which we directly evaluated for adherence to known molecular rules governing NMD[44] (Extended Data Fig. 6). Taken together, we estimate that for a typical multi-intronic gene, ~15% of all transcript molecules generated are NMD targets, though rates of unproductive splicing vary greatly across genes. For example, at lowly expressed genes, we estimate the fraction of NMD targets for transcripts with ten exons is 25% of mRNA molecules, and for transcripts with 15 or more exons, this number increases to over 50% (Supplementary Fig. 6).

Artificially induced cryptic splicing widely triggers NMD

Our findings collectively suggest that a large fraction of AS events result in unproductive transcripts that are targeted by NMD for rapid degradation. To validate this finding experimentally, we measured the effects of a splice-switching drug on gene expression levels in LCLs (Fig. 3a). Risdiplam was originally identified as a small molecule splicing modulator that upregulates *SMN2* by activating a productive GA|GU 5′ splice site (5′ss) at the expense of an unproductive isoform[45,46]. Recent studies[47,48] find that at higher doses, risdiplam's effects are less selective, extending to GA|GU 5′ss in many other genes. We predicted increases in these cryptic GA|GU splice sites would generally lead to an abundance of unproductive, NMD-targeted transcripts and, consequently, decreased expression of the affected genes. To test this, we treated LCLs with risdiplam and sequenced steady-state RNA at eight increasing doses of risdiplam and naRNA at two doses (Fig. 3a). We observed widespread, genome-wide induction of GA|GU 5′ ss in a dose-dependent manner but no effect on canonical AG|GU 5′ss or other noncanonical splice sites (Fig. 3b and Supplementary Fig. 7). For example, *MYB*, a transcription factor protooncogene targeted by a number of antitumor therapies, contains a risdiplam-induced exon (Fig. 3c). This exon is uniquely annotated in an unproductive transcript isoform, and *MYB* is downregulated in a dose-dependent manner.

In total, we identified 316 risdiplam-induced cassette exons, 281 (92%) of which are unannotated. We found that 257/281 (91%) of the unannotated risdiplam-induced exons are within the coding region of the transcript (not in UTRs). Of these, 214/257 (61%) result in a frameshift, matching the expectation that two-thirds of cryptic splicing changes should produce a frameshift. An additional 56/281 (20%) of the unannotated exons contain an in-frame PTC and are also expected to induce NMD (Fig. 3d). The effects of these risdiplam-induced exons on expression matched our expectation that inclusion of predicted frame-shifting- and in-frame-PTC-containing exons generally results in downregulation of steady-state steady RNA but not naRNA (Fig. 3e and Supplementary Fig. 8). This set of 230 unproductive risdiplam-induced exons in 219 genes included some with robust expression effects in potentially disease-modifying genes (Supplementary Fig. 9). Notably, these 219 risdiplam-regulated genes were not limited to gene families traditionally thought of as 'druggable' by protein-targeting small molecule drugs (for example, kinases) (Fig. 3f and Supplementary Fig. 10a). Consistent with our observation that unproductive junctions accumulate across the length of the transcript and that long introns and long genes are most susceptible to AS–NMD, we found that introns and genes hosting these risdiplam-induced exons tend to be longer
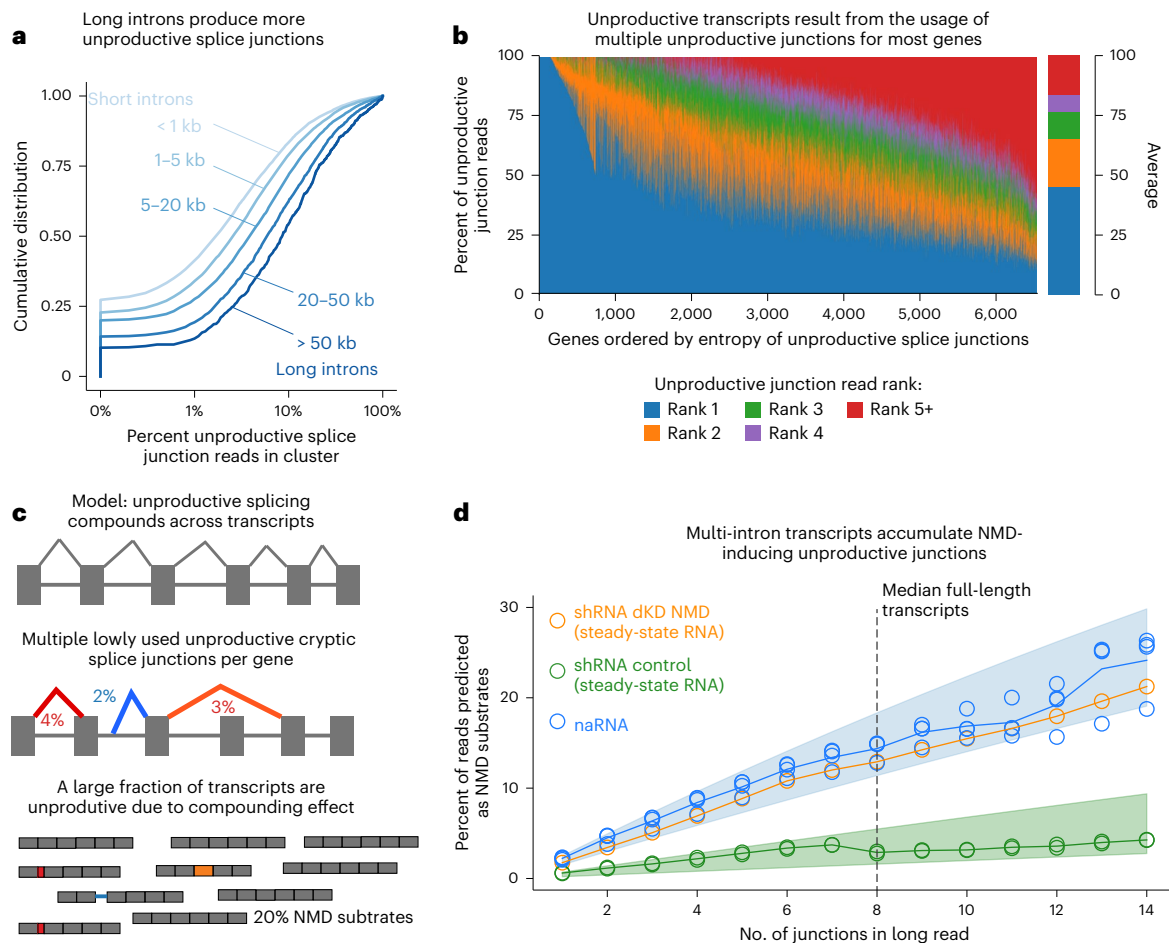
Fig. 2 | Unproductive mis-splicing accumulates across transcripts.
a, Correlation between percent of unproductive junction reads in
LeafCutter clusters versus length of the most used productive intron in
the cluster (Spearman's rho coefficient of 0.19; two-sided correlation test
P value = 1.3×10^{-243}). Correlation presented as cumulative distribution of
percent unproductive splice junctions, for increasing groups of intron length.
b, Percent unproductive splice junctions for each gene that are attributed
to the most common (rank 1) unproductive junction, the second most
common unproductive junction and so on. **c**, Model of unproductive splicing
compounding across multi-intronic transcripts. A low unproductive rate at many
independently spliced introns produces a high rate of unproductive molecules

at the transcript level. **d**, Nanopore long-read sequencing quantifies the percent
of full-length reads that are targeted by NMD, defined as containing at least one
unproductive junction, as a function of the number of splice junctions in the
read. The vertical dashed line marks eight splice junctions, corresponding to a
typical full-length human transcript. The total RNA were isolated from shRNA-
mediated dKD of NMD factors *SMG6* and *SMG7* or shRNA scramble control[23]
in HeLa cells. The naRNA data were from K562 cells[43]. Multiple points of the
same color indicate replicate experiments. The blue and orange shaded area
represents the binomial expectation when assuming 1.5–2.5% and 0.2–0.7% of
unproductive junction reads at each independent junction.

than matched controls (Fig. 3g and Supplementary Fig. 10). Finally,
we identified an asymmetry in risdiplam-induced post-transcriptional
effects (Extended Data Fig. 7) favoring downregulation. Thus, random
splicing changes, such as those introduced by similar nonspecific
splice-switching drugs or mutations, usually result in NMD-targeted
transcripts and subsequent downregulation, as opposed to frame
preserving or frame correcting (for example, *SMN2*) transcript changes.

Global effects of sQTLs on gene expression levels
The prevalence of unproductive splicing, along with the observation
that unproductive splicing anti-correlates with gene expression levels
genomewide, predicts that genetic effects on RNA splicing would often
impact RNA expression levels. To test this prediction, we used quantita-
tive trait loci (QTL) mapping (Fig. 4a) to identify genetic variants associ-
ated with expression (eQTLs) and splicing (splice junction abundance,
sQTLs) in naRNA-seq, 4sU sequencing and steady-state RNA-seq data. To
better distinguish splicing-mediated expression effects from transcrip-
tional effects, we mapped histone QTLs (hQTLs), reflective of variants
impacting promoter and enhancer activity (H3K27ac, H3K4me1 and

H3K4me3) and transcription across gene bodies (H3K36me3). In total,
we identified 57,981 QTLs for 620,020 tested molecular traits (Supple-
mentary Fig. 11). Consistent with previous work, we find a large fraction
of eQTLs are explained by transcriptional regulation, as indicated by the
high degree of sharing between eQTL and hQTL signals. For example,
we estimate that 67% of eQTL-containing genes (eGenes, steady-state
RNA) contain hQTL effects at the gene's promoter (Storey's $\pi_1 = 0.67$;
Supplementary Fig. 12a), consistent with transcriptional regulation. As
expected, we observe a strong concordance in the direction of hQTL
and eQTL effects, wherein hQTL alleles that increase H3K27ac signal
at the promoter overwhelmingly have corresponding upregulating
signals at the level of mRNA (Fig. 4b). The remaining steady-state RNA
eQTLs that do not have hQTL signal probably function during or shortly
after transcription, as the effects of an additional 24% of steady-state
eQTLs are also detected in 30 min 4sU-labeled RNA ($\pi_1 = 0.91$; Sup-
plementary Fig. 12a).

To better interpret molecular mechanisms of eQTLs while account-
ing for linkage disequilibrium, we performed a multitrait colocalization
analysis to identify molecular QTLs that probably share causal variants

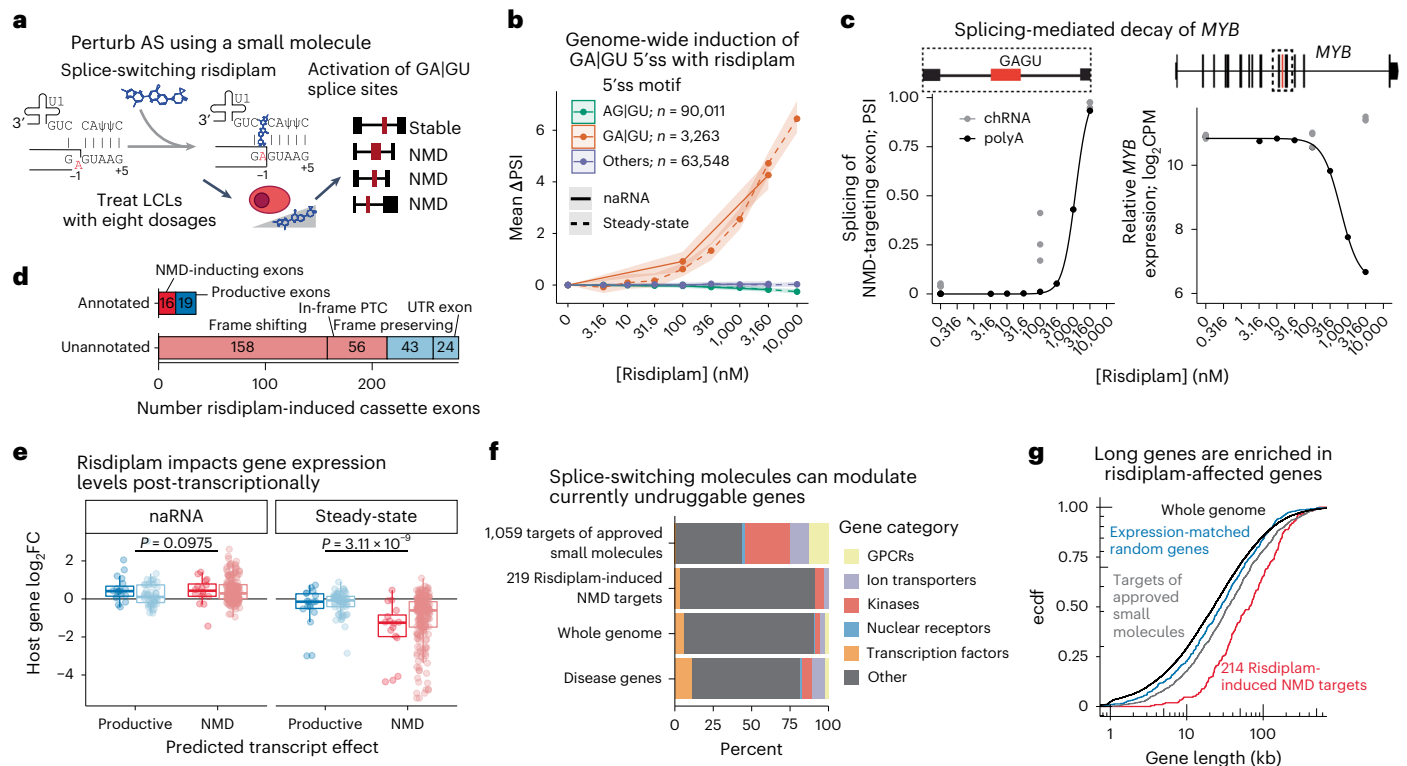**Fig. 3 | Risdiplam-induced splicing alterations mediate expression changes at hundreds of genes. a**, Overview of risdiplam-based approach to assess pervasiveness of NMD after splicing perturbations. LCLs were treated with eight doses of risdiplam. Splicing changes at cryptic exons assessed for NMD potential, and gene expression changes were estimated. **b**, Genome-wide mean splicing dose-dependent changes at various classes of 5′ss in naRNA and steady-state RNA. Bootstrapped 95% confidence intervals shaded around the mean activation level across n introns in each group. **c**, Left: dose-dependent splicing response at a risdiplam-induced exon in *MYB*. Right: dose-dependent expression response of *MYB*. **d**, Predicted translation result of 316 risdiplam-induced exons. Unproductive exons (expected to induce NMD) versus those that maintain transcript stability are red and blue, respectively. Annotated and unannotated exons are dark and light colors, respectively. **e**, Empirically measured effect of host gene expression ($\log_2$ fold-change) as measured in steady-state RNA and naRNA in the presence of risdiplam at 3.16 μM. Each point is a gene hosting a

risdiplam-induced exon colored by the same exon classifications used in **d**. P values for two-sided Mann–Whitney U test, comparing productive ($n = 85$ genes hosting annotated or unannotated risdiplam-induced exons) and NMD sets ($n = 219$ genes) are shown. The box represents median and innerquartiles. The whiskers extend from hinge to most extreme value no greater than 1.5× IQR from hinge. **f**, Conventional small molecule drugs usually operate at the level of protein binding and are disproportionately skewed for particular classes of 'druggable' genes (for example, G-protein-coupled receptors, GPCRs) and against other classes (for example, transcription factors). Risdiplam-induced NMD targets are more representative of all genes. Disease genes with therapeutic potential by downregulation (Online Mendelian Inheritance in Man dominant negative genes) are similarly distributed across categories of previously 'druggable' genes. **g**, Cumulative distribution of gene length for genes with predicted NMD-induced exons, a similarly sized set of expression-matched genes, all genes or a set of gene targets for Food and Drug Administration (FDA)-approved small molecules.

(Fig. 4a, Supplementary Fig. 13a and Methods). For example, at the *CCHRC1* locus, we identify an eQTL, an hQTL and an sQTL (Fig. 4a). While both the sQTL and hQTL are nominally significant for an eQTL association, colocalization analysis reveals that only the hQTL and eQTL are probably caused by the same genetic variant (posterior probability of full colocalization, PPFC >0.5). This suggests that transcriptional regulation explains the eQTL effect, while the sQTL constitutes a distinct genetic effect that does not explain the primary eQTL signal. Across all 3,970 steady-state RNA eQTLs, we find 831 (~20%) colocalized with an hQTL (PPFC >0.5) (Fig. 4c and Supplementary Fig. 13a). We do not interpret this to suggest that only 20% of eQTLs are caused by transcriptional regulation but rather that 20% represents a lower limit of the degree of transcriptional regulation. Nevertheless, we reasoned that the abundance of eQTL colocalizations with hQTLs relative to their colocalizations with other molecular QTLs may be informative. When considering eQTLs that colocalized with any molecular QTL (molQTL) we tested, 62% colocalized to an hQTL, consistent with our previous finding that approximately two-thirds of eQTLs are driven by chromatin effects. Among the remaining 38%, which we hereafter refer to as 'post-transcriptional eQTLs', nearly half colocalized with an sQTL (Fig. 4c and Supplementary Fig. 13b), suggesting that AS may be a major

contributor to interindividual variation in gene expression levels. By comparison, alternative polyadenylation QTLs (apaQTLs) colocalized with only ~5% of the non-hQTL colocalizations, suggesting that alternative polyadenylation plays a comparatively minor role.

The degree of sQTL–eQTL colocalization suggests that a large number of genetic variants may impact gene expression levels through AS. However, we and others have noted that alternative transcription initiation or polyadenylation can alter splicing quantifications and vice versa[40,49–51]. Furthermore, preferential decay of specific mRNA isoforms may manifest as sQTL–eQTL colocalizations[52], without being mediated by AS. To better assess whether splicing changes causally drive eQTL–sQTL colocalizations, we asked which genomic annotations are most enriched among post-transcriptional eQTLs, compared with transcriptional eQTLs as controls. As expected, post-transcriptional eQTLs are strongly depleted in enhancers and promoters. In terms of enriched genomic regions, we find significant enrichment of post-transcriptional eQTLs near polyadenylation sites and in splice donor, branch site and splice acceptor regions (Supplementary Fig. 13c). While polyadenylation sites and splice sites are similarly enriched among post-transcriptional eQTLs over transcriptional eQTLs, splice sites account for a far greater fraction of
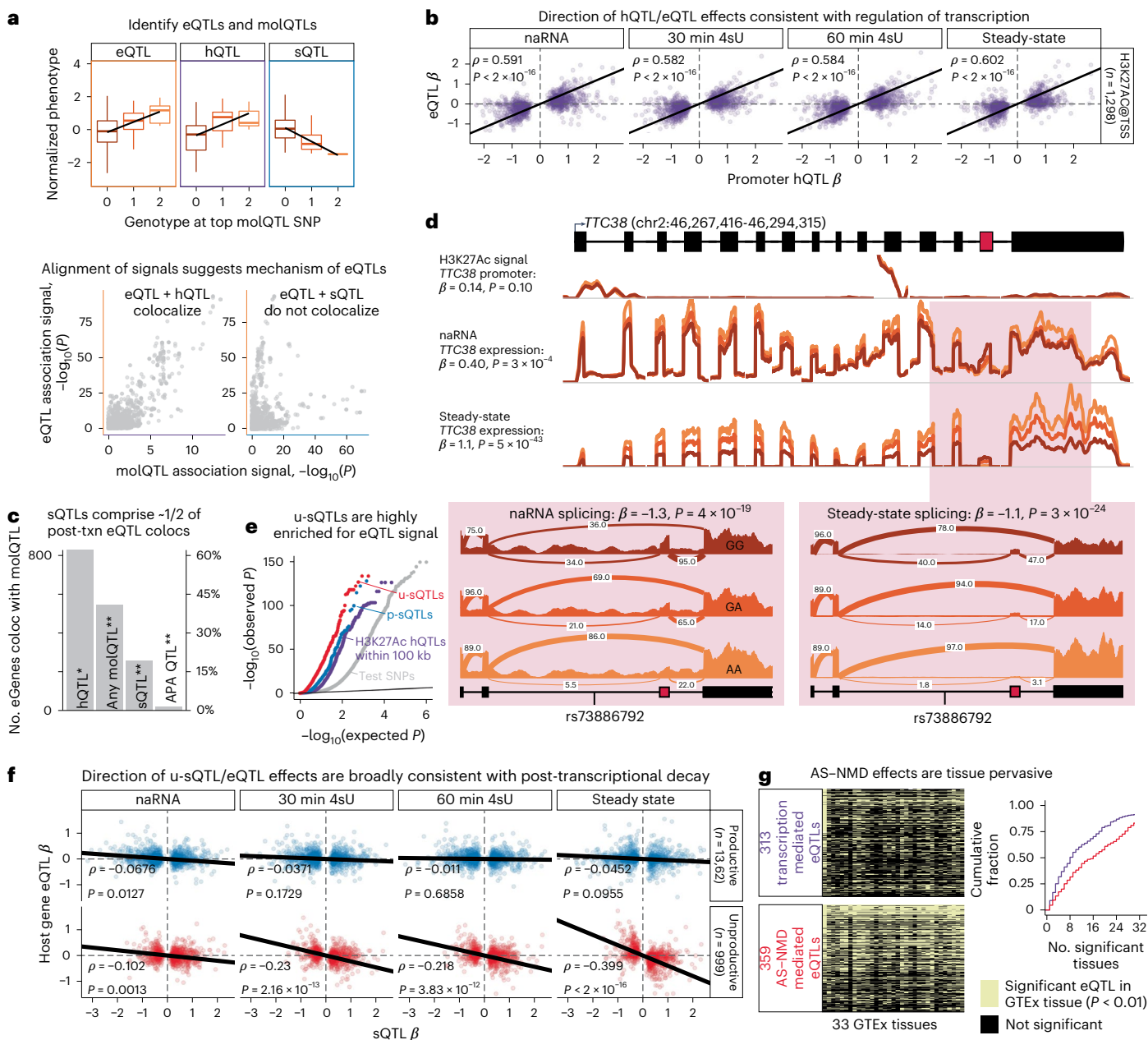
Fig. 4 | Genetic variants that alter expression post-transcriptionally are enriched in splice-altering variants that are predicted to induce NMD.
a, Approach for identifying mechanisms of gene regulation, using data from *CCHCR1* locus as illustrative example. Genetic variants that associate with chromatin peak height, RNA expression, splicing and so on (molQTL) are identified. The boxplots grouped by genotype at lead SNP depict innerquartiles of phenotype values with whiskers extending to most extreme value no greater than 1.5× IQR from hinge. Multitrait colocalization (Methods) compares alignment of molQTL signals (scatter plots[54]) to identify molQTLs that probably share causal SNP. **b**, Effect size (β) of H3K27ac hQTLs at promoter versus eQTL effect sizes estimated in each RNA-seq dataset. Correlation summarized with Spearman's rho coefficient and two-sided significance test. **c**, Tally of steady-state eGenes by their colocalizing molQTLs. A total of 831 eGenes colocalize (coloc) with an hQTL (*, H3K27ac, H3K4me1, H3K4me3 or H3K36me3), and 518 'post-transcriptional' (post-txn) eGenes (**) do not colocalize with any hQTL but do colocalize with other molQTL, some of which may be sQTL or apaQTL.

d, Example eQTL functioning through AS–NMD from poison exon (red). H3K27ac ChIP-seq and RNA-seq coverage grouped by genotype of lead eQTL SNP. The pink inset region (chr22:46,289,343-46,294,094) shows effect on splicing of poison exon, depicted as sashimi plots with relative usage (intronic PSI) of splice junction arcs. **e**, QQ plot shows inflation of eQTL signal among groups of SNPs (lead SNPs for p-sQTLs and u-sQTLs within the host gene, H3K27ac QTLs within 100 kb of test gene or random test SNPs). **f**, Similar to **b**. Top: effect size of p-sQTLs (only significantly affecting productive splice junctions) versus host gene eQTL β. Bottom: effect size of u-sQTLs (significantly affecting at least one unproductive splice junction) versus effect on host gene expression. **g**, eQTLs consistent with transcriptional regulation (eQTL/hQTLs, purple) in discovery dataset (LCLs) compared with eQTLs consistent with AS–NMD (eQTL/u-sQTLs, red). Left: each eQTL (SNP:gene pair, rows) assessed for eQTL effects in GTEx tissues (columns). Right: cumulative distribution of number tissues with significant effects. The *P* value for a two-sided Mann–Whitney test is shown.

post-transcriptional eQTLs (Supplementary Fig. 13d). These observations indicate that changes in RNA splicing causally drive many of these post-transcriptional eQTLs.

As an illustrative example of a splicing-mediated eQTL, we highlight the *TTC38* gene (Fig. 4d) for which the lead eQTL variant has no detectable effect on promoter activity. Rather, we observed a clear

effect of the lead eQTL variant on splicing. The allele associated with decreased expression level is also associated with an increase in splicing of unproductive introns. These unproductive splice junctions are substantially more abundant in naRNA than steady-state RNA, consistent with their rapid degradation, though the sQTL effect size is similar in naRNA and steady-state RNA (Fig. 4d). More generally, when we stratify sQTLs by whether they affect an unproductive splice junction (henceforth 'u-sQTLs'), versus merely switching between alternate productive protein-coding isoforms (henceforth 'p-sQTLs'), we find that u-sQTLs are particularly enriched in eQTL signal for the host gene (Fig. 4e and Supplementary Fig. 14a), with even stronger eQTL signal than H3K27ac QTLs within 100 kb of the gene (Fig. 4e). Furthermore, though we identify similar numbers of p-sQTLs versus u-sQTLs, u-sQTLs explain 77% of sQTL colocalizations with post-transcriptional eQTLs, compared with 23% for p-sQTLs ($P = 5.9 \times 10^{-12}$, hypergeometric test; Supplementary Fig. 14b). These results again suggest that AS–NMD is a major contributor to the genetic regulation of expression levels.

To further validate that u-sQTLs affect gene expression levels through post-transcriptional decay, we considered the concordance of normalized effect sizes between u-sQTLs and eQTLs (Methods). As expected, p-sQTL variants are largely inert with respect to expression levels (Fig. 4f). In contrast, u-sQTL effects are strongly anticorrelated with expression, such that alleles increasing the unproductive junction are associated with decreased expression of the host gene (Fig. 4f, steady-state facet). These effects on expression are most apparent in steady-state RNA, weaker in our 4sU RNA-seq data and largely absent in naRNA (Fig. 4f and Supplementary Fig. 15), which is expected, as NMD occurs post-transcriptionally and in the cytoplasm.

While we observed an overwhelming enrichment of eQTLs among u-sQTL SNPs (Fig. 4e), previous studies[37,50,53] have found that the leading eQTL and sQTL signals for a given gene tend to be independent. To reconcile these observations, we provide two mutually compatible explanations: (1) an eGene's top sQTL signal is often a p-QTL, not an expression-modifying u-sQTL (Supplementary Fig. 16a,b) and (2) many genes have multiple eQTLs[53,54], and u-sQTLs may contribute to expression even when they are not the lead eQTL.

Given reports that lead eQTLs can sometimes change across tissues[53,55] and our previous observation that sQTL effects are generally more stable across tissues than eQTLs[56], we hypothesized that the effects of splicing-mediated eQTLs would be more consistent across tissues than the effects of transcription-mediated eQTLs. Indeed, we find that eQTLs mediated by splicing mechanisms are eQTLs in a larger number of GTEx tissues than eQTLs mediated by transcription regulation (Fig. 4g and Extended Data Fig. 8), indicating that the regulatory impact of variants that function through AS–NMD are more tissue-pervasive than those that function through transcription. Such variants could be relevant for interpreting genome-wide association study (GWAS) signals that have yet to be explained by eQTLs in known cell-types.

AS–NMD impacts GWAS traits

Given the pervasive effects of AS–NMD on gene expression, we wondered to what extent AS that affects complex phenotypes are mediated by NMD versus protein-diversifying stable isoforms. To answer this question, we compiled summary statistics from 45 GWAS for blood and immune-related traits and evaluated the enrichment of GWAS signal among various classes of molecular QTLs: eQTLs, H3K27ac QTLs, p-sQTLs and u-sQTLs. For example, using multiple sclerosis GWAS summary statistics, we find enrichment of u-sQTLs on par with that of p-sQTLs, hQTLs and eQTLs (Fig. 5a). Similar results are found in most complex traits we examined (Supplementary Fig. 17).

To better resolve the mechanisms at these GWAS loci, we used multitrait colocalization to identify GWAS signals that colocalize with hQTLs, eQTLs, sQTLs or various combinations of QTLs. Across all complex traits, approximately 70% of GWAS loci could not be colocalized

with any molecular QTL (Fig. 5b), in line with previous studies with similar sample sizes[57,58]. Approximately 18% of GWAS loci colocalize with either an hQTL, eQTL or some combination of molecular QTLs (Fig. 5b). The remaining 12% of loci colocalize with an sQTL but not an hQTL, consistent with splicing-mediated impacts on traits. We next sought to assess whether these GWAS/sQTL loci possess characteristics consistent with an AS–NMD mechanism versus protein diversification.

Notably, we find that these sQTLs largely affect low-usage splice junctions, such that 57% of these splice junctions are spliced in at PSI <5% (Fig. 5c). This observation is consistent with that of a recent study[57] and naturally poses questions as to how such low-usage isoforms might impact traits. We found that most of these low-usage sQTLs are u-sQTLs that alter the balance of unproductive and productive isoforms, consistent with AS–NMD mediating these loci. We also found that the allelic effects of u-sQTLs that colocalize with GWAS loci, similar to that of u-sQTLs in general (Fig. 4f), were anticorrelated with their effects on gene expression levels, again consistent with NMD (Fig. 5d). These u-sQTLs are also enriched among the GWAS/sQTL loci that also colocalize with eQTL signal (Extended Data Fig. 9a). For example, a reticulocyte-count-associated GWAS signal colocalizes with a u-sQTL in *NUDT14* gene, as well as *NUDT14* eQTL signal (Extended Data Fig. 10a,b). As expected, the allele that increases usage of the unproductive splice junction is associated with *NUDT14* downregulation (Fig. 5c). While the effect on splicing is similar in both steady-state RNA-seq and naRNA-seq (Extended Data Fig. 10c), the effect on *NUDT14* expression is only apparent in steady-state RNA (Extended Data Fig. 10d), again consistent with AS–NMD.

More generally, u-sQTLs that colocalize with GWAS and eQTL signals tend to also display lower usage in steady-state RNA than naRNA (Extended Data Fig. 9b), with sQTL and eQTL effects that are also consistent with AS–NMD (Extended Data Fig. 9c). In contrast, the sQTLs that colocalize with GWAS but not eQTL do not share these characteristics and probably function by tuning the expression levels of alternative protein-coding isoforms. Given that there are a similar number of these GWAS loci that colocalize with both sQTL and eQTL (and not hQTL), as compared with just sQTLs (Fig. 5b), we conclude that AS–NMD carries similar importance as splicing-mediated protein diversification for complex organism-level traits.

Discussion

The molecular impact of AS has been challenging to verify experimentally, as it has been difficult to study the function of individual isoforms at the protein level, and the rapid decay of unproductive isoforms obscure their quantification at the mRNA level. Through detailed analysis of molecular measurements that capture the major steps of RNA maturation, we found that aberrant splicing produces remarkably high levels of unproductive transcripts bearing a PTC. Unproductive mRNAs account for around 15% of all mRNA transcripts from the average human gene, even exceeding 50% for many long genes expressed at low levels. These estimates may be surprising, given earlier studies that utilized single knockdown of *UPF1* or *UPF2* to identify a relatively small subset of AS–NMD-regulated genes[25–27,59], suggesting that unproductive isoforms are produced at such low rates that expression levels of most genes are unaffected by AS–NMD[27]. However, we find that early estimates are consistent with incomplete inhibition of NMD due to partial redundancy between core NMD factors. Indeed, the levels of unproductive splicing in steady-state mRNA stabilized from double *SMG6* and *SMG7* knockdowns—but not that of single *UPF1* knockdowns—were nearly identical with that estimated in our nascent mRNA dataset. Thus, previous studies using knockdowns of single NMD factors appear to have substantially underestimated the impact of AS–NMD.

Notably, we also show that an important previous observation—that highly used AS exons are enriched for being frame-preserving—is largely due to NMD surveillance rather than selection at the level
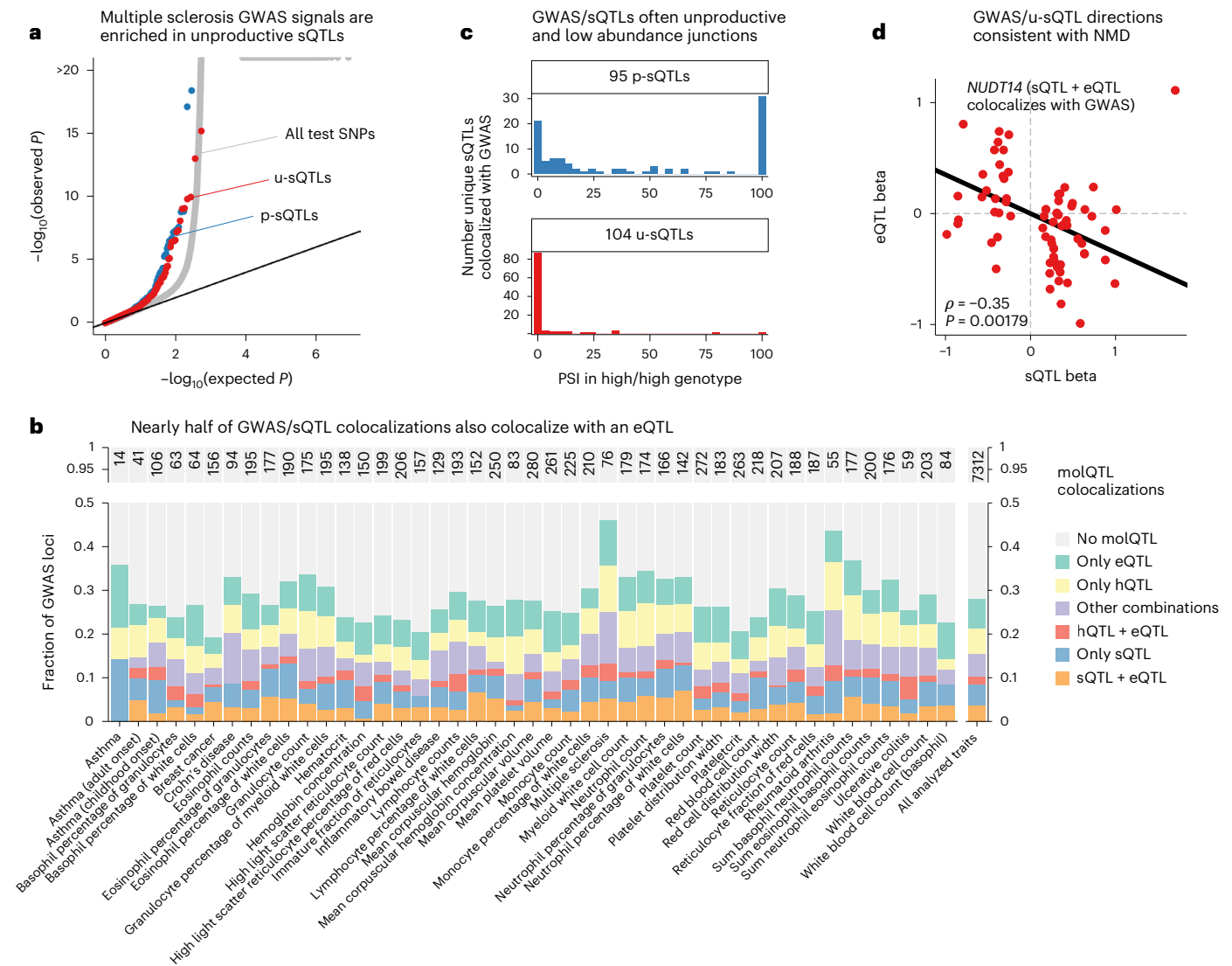
a Multiple sclerosis GWAS signals are enriched in unproductive sQTLs

c GWAS/sQTLs often unproductive and low abundance junctions

d GWAS/u-sQTL directions consistent with NMD

b Nearly half of GWAS/sQTL colocalizations also colocalize with an eQTL

Fig. 5 | Splicing-mediated NMD contributes to complex trait biology. a, QQ plot of multiple sclerosis GWAS signal, grouped by categories of SNPs. p-sQTLs that impact the balance of protein-coding isoforms and u-sQTLs that impact usage of unproductive splice junctions are similarly inflated for GWAS signals. **b**, Fraction of GWAS loci that colocalize with various sets of molQTLs in each of 45 blood or immune-related traits. Number of loci for which colocalization was attempted is indicated at the top of each column. 'Other combinations' includes loci that colocalize with alternative polyadenylation QTLs, hQTLs and sQTLs or other combinations that may include sQTLs and other molQTLs and are difficult

to interpret mechanistically. **c**, Histogram of usage of unique sQTL junctions that colocalize with a GWAS signal, grouped by sQTL type. Intronic PSI (junction read count divided by most abundant junction in LeafCutter cluster) for each junction was summarized as the median from steady-state RNA samples which are homozygous for the PSI-increasing allele. Many GWAS sQTLs, especially u-sQTLs, have low PSI, even in samples with genotypes that favor higher usage. **d**, Effect size (β) of sQTLs and eQTLs for distinct u-sQTLs that colocalize with a GWAS signal. Correlation was summarized with Spearman's rho coefficient and two-sided significance test.

---

of splicing regulation. Thus, our study suggests that the molecular impact of AS is largely shouldered by NMD, which regulates protein output by targeting unproductive transcripts for degradation. Supporting this view, we identified nearly as many genetic variants that impact production of these unproductive transcripts as compared with those that tune the balance of stable mRNA isoforms. Importantly, this observation also holds for sQTLs that colocalize with GWAS signals to influence organism-level traits. These unproductive sQTLs can have variable effects on expression, wherein small changes in splicing are sometimes associated with large changes in expression or vice versa. While the biological basis for this variability is unknown, we speculate that multiple mechanisms, including transcriptional adaptation[60] and differential efficiency of NMD[44], may be at play. Nonetheless, the general tissue-pervasive nature of these AS−NMD-mediated eQTLs compared with transcription-mediated eQTLs may have particular relevance both in enhancing the phenotypic impact of AS−NMD-based

regulation and in mapping the regulatory mechanisms underlying genetic associations for complex traits.

What fraction of mRNA isoforms encode functionally diverse peptides has long been under debate[7,27,35,36,61]. Novel isoforms continue to be discovered as RNA-seq experiments increase in depth, making this question particularly timely. Our observation that introns are often mis-spliced into substrates of NMD indicates that splicing is inherently noisy and that novel isoforms uncovered by RNA-seq generally do not encode functional proteins. This view is consistent with previous findings that most AS events are lowly used and do not show cross-species conservation.

Still, we posit that future research will reveal a preponderance of cases where regulated AS functions by tuning protein expression levels rather than by creating protein diversity, as the sheer abundance of AS−NMD events presents opportunities for evolution to co-opt AS−NMD as a functional regulatory mechanism. Regulated AS−NMD has been

identified in the past, but it has largely been found in genes encoding splicing regulators. While we confirmed that splicing regulators are enriched among genes with high levels of unproductive transcripts, splicing regulators represent only a small fraction of all genes with very high levels of unproductive transcripts. We predict that future work using long-read sequencing of RNA across multiple species, stages of maturation and biological systems will provide us with a much more complete understanding of the mechanisms by which AS functionally impacts cellular function, organismal phenotypes and evolution.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01872-x.

References

1. Barbosa-Morais, N. L. et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).
2. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**, 1593–1599 (2012).
3. Mudge, J. M. et al. The origins, evolution, and functional potential of alternative splicing in vertebrates. *Mol. Biol. Evol.* **28**, 2949–2959 (2011).
4. Bénitère, F., Necsulea, A. & Duret, L. Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans. *Evol. Biol.* https://doi.org/10.1101/2022.12.09.519597 (2022).
5. Saudemont, B. et al. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biol.* **18**, 208 (2017).
6. Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* **6**, e1001236 (2010).
7. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA* **100**, 189–192 (2003).
8. Mendell, J. T., Sharifi, N. A., Meyers, J. L., Martinez-Murillo, F. & Dietz, H. C. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat. Genet.* **36**, 1073–1078 (2004).
9. Melamud, E. & Moult, J. Stochastic noise in splicing machinery. *Nucleic Acids Res.* **37**, 4873–4886 (2009).
10. García-Moreno, J. F. & Romão, L. Perspective in alternative splicing coupled to nonsense-mediated mRNA decay. *Int. J. Mol. Sci.* **21**, 9424 (2020).
11. Kalyna, M. et al. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res.* **40**, 2454–2469 (2012).
12. Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C. & Brenner, S. E. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446**, 926–929 (2007).
13. Lareau, L. F. & Brenner, S. E. Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Mol. Biol. Evol.* **32**, 1072–1079 (2015).
14. Leclair, N. K. et al. Poison exon splicing regulates a coordinated network of SR protein expression during differentiation and tumorigenesis. *Mol. Cell* **80**, 648–665.e9 (2020).
15. McGlincy, N. J. & Smith, C. W. J. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem. Sci.* **33**, 385–393 (2008).
16. Nasif, S., Contu, L. & Mühlemann, O. Beyond quality control: the role of nonsense-mediated mRNA decay (NMD) in regulating gene expression. *Semin. Cell Dev. Biol.* **75**, 78–87 (2018).
17. Sureau, A. SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. *EMBO J.* **20**, 1785–1796 (2001).
18. Tabrez, S. S., Sharma, R. D., Jain, V., Siddiqui, A. A. & Mukhopadhyay, A. Differential alternative splicing coupled to nonsense-mediated decay of mRNA ensures dietary restriction-induced longevity. *Nat. Commun.* **8**, 306 (2017).
19. Yan, Q. et al. Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. *Proc. Natl Acad. Sci. USA* **112**, 3445–3450 (2015).
20. Yang, R. et al. Upregulation of SYNGAP1 expression in mice and human neurons by redirecting alternative splicing. *Neuron* **111**, 1637–1650.e5 (2023).
21. Ni, J. Z. et al. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.* **21**, 708–718 (2007).
22. Frankish, A. et al. GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
23. Karousis, E. D., Gypas, F., Zavolan, M. & Mühlemann, O. Nanopore sequencing reveals endogenous NMD-targeted isoforms in human cells. *Genome Biol.* **22**, 223 (2021).
24. Colombo, M., Karousis, E. D., Bourquin, J., Bruggmann, R. & Mühlemann, O. Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways. *RNA* **23**, 189–201 (2017).
25. Tani, H. et al. Identification of hundreds of novel UPF1 target transcripts by direct determination of whole transcriptome stability. *RNA Biol.* **9**, 1370–1379 (2012).
26. Rehwinkel, J., Letunic, I., Raes, J., Bork, P. & Izaurralde, E. Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. *RNA* **11**, 1530–1544 (2005).
27. Pan, Q. et al. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.* **20**, 153–158 (2006).
28. Drechsel, G. et al. Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the *Arabidopsis* steady state transcriptome. *Plant Cell* **25**, 3726–3742 (2013).
29. Chen, C. et al. UPF3A is dispensable for nonsense-mediated mRNA decay in mouse pluripotent and somatic cells. *Life Sci. Alliance* **6**, e202201589 (2023).
30. Metze, S., Herzog, V. A., Ruepp, M.-D. & Mühlemann, O. Comparison of EJC-enhanced and EJC-independent NMD in human cells reveals two partially redundant degradation pathways. *RNA* **19**, 1432–1448 (2013).
31. Yi, Z. et al. Mammalian UPF3A and UPF3B can activate nonsense-mediated mRNA decay independently of their exon junction complex binding. *EMBO J.* **41**, e109202 (2022).
32. Boehm, V. et al. SMG5-SMG7 authorize nonsense-mediated mRNA decay by enabling SMG6 endonucleolytic activity. *Nat. Commun.* **12**, 3965 (2021).
33. Blencowe, B. J. Alternative splicing: new insights from global analyses. *Cell* **126**, 37–47 (2006).
34. Ezkurdia, I. et al. Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.* **14**, 1880–1887 (2015).
35. Blencowe, B. J. The relationship between alternative splicing and proteomic complexity. *Trends Biochem. Sci.* **42**, 407–408 (2017).

36. Tress, M. L., Abascal, F. & Valencia, A. Most alternative isoforms are not functionally important. *Trends Biochem. Sci.* **42**, 408–410 (2017).

37. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).

38. Grubert, F. et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* **162**, 1051–1065 (2015).

39. The Geuvadis Consortiumet al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

40. Mittleman, B. et al. Alternative polyadenylation mediates genetic regulation of gene expression. *Genomics* https://doi.org/10.1101/845966 (2019).

41. Long, M., Rosenberg, C. & Gilbert, W. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl Acad. Sci. USA* **92**, 12495–12499 (1995).

42. Glassberg, E. C., Gao, Z., Harpak, A., Lan, X. & Pritchard, J. K. Evidence for weak selective constraint on human gene expression. *Genetics* **211**, 757–772 (2019).

43. Drexler, H. L., Choquet, K. & Churchman, L. S. Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores. *Mol. Cell.* https://doi.org/10.1016/j.molcel.2019.11.017 (2019).

44. Lindeboom, R. G. H., Vermeulen, M., Lehner, B. & Supek, F. The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. *Nat. Genet.* **51**, 1645–1651 (2019).

45. Palacino, J. et al. SMN2 splice modulators enhance U1–pre-mRNA association and rescue SMA mice. *Nat. Chem. Biol.* **11**, 511–517 (2015).

46. Campagne, S. et al. Structural basis of a small molecule targeting RNA for a specific splicing correction. *Nat. Chem. Biol.* **15**, 1191–1198 (2019).

47. Ishigami, Y. et al. Specificity, synergy, and mechanisms of splice-modifying drugs. *Biophysics* https://doi.org/10.1101/2022.12.30.522303 (2022).

48. Ottesen, E. W. et al. Diverse targets of *SMN2* -directed splicing-modulating small molecule therapeutics for spinal muscular atrophy. *Nucleic Acids Res.* **51**, 5948–5980 (2023).

49. Alasoo, K. et al. Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *eLife* **8**, e41673 (2019).

50. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.* **12**, 727 (2021).

51. Kubota, N. & Suyama, M. Mapping of promoter usage QTL using RNA-seq data reveals their contributions to complex traits. *PLoS Comput. Biol.* **18**, e1010436 (2022).

52. Sun, B. & Chen, L. Mapping genetic variants for nonsense-mediated mRNA decay regulation across human tissues. *Genome Biol.* **24**, 164 (2023).

53. The GTEx Consortiumet al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

54. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).

55. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).

56. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).

57. El Garwany, O. et al. Low-usage splice junctions underpin immune-mediated disease risk. *Genomics* https://doi.org/10.1101/2023.05.29.542728 (2023).

58. Mu, Z. et al. The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome Biol.* **22**, 122 (2021).

59. Weischenfeldt, J. et al. Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. *Genome Biol.* **13**, R35 (2012).

60. El-Brolosy, M. A. et al. Genetic compensation triggered by mutant mRNA degradation. *Nature* **568**, 193–197 (2019).

61. Tress, M. L., Abascal, F. & Valencia, A. Alternative splicing may not be the key to proteome complexity. *Trends Biochem. Sci.* **42**, 98–110 (2017).

Methods

We confirm that this research complies with all relevant ethical regulations. Ethical approval was not required because the Yoruba in Ibadan, Nigeria (population code YRI) panel of cell lines used here (derived from consenting individuals with publicly available genotypes[62]), from which molecular sequencing data was generated, are not considered human subjects research by the institutional review board at the University of Chicago.

Statistics and reproducibility

No statistical method was used to predetermine sample size for novel data (naRNA-seq, H3K36ME3 CUT&Tag, RNA-seq and naRNA-seq at various risdiplam doses). We note that our sample sizes are similar to previous QTL studies using this panel of cell lines[37,38,40] and risdiplam transcriptomics studies[45,48]. For eQTL and sQTL mapping using naRNA, we excluded line NA18855. This was not a predetermined decision. Rather, principal component analysis revealed this sample as an outlier, and differential expression analysis against other lines identified differentially expressed genes clustered along large sections of chromosomes, suggestive of chromosomal abnormalities. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment, as cell lines (for QTL studies) and treatment were required knowledge for carrying out the experiments. When data could not be assumed to follow assumptions for parametric tests, nonparametric tests were used. Tests for significance of spearman correlation were obtained using 'cor.test(…, method = 's')' in R.

Molecular trait quantification

We aligned the Illumina short-read RNA-seq datasets from LCL and HeLa cells to the human genome version GRCh38 with Gencode v34 annotation using STAR[63] version 2.7.7a. For the LCL lines, we used STAR's WASP mode[64] to account for genetic variants. We mapped ChIP-seq and CUT&Tag reads to the human genome using HISAT[65] version 2.2.1, and we used Hornet (https://github.com/TheFraserLab/Hornet, a reimplementation of WASP pipeline[64]) to find reads that overlap with SNPs for remapping. For the ChIP-seq datasets, we used MACS2 (ref. 66) version 2.2.7.1 to call peaks.

We quantified gene expression and histone modification coverage of H3K27ac, H3K4me1 and H3K4me3 using featureCounts[67] version 2.0.3. We used 'bedtools multicov' to count the number of H3K36me3 CUT&Tag reads overlapping each protein-coding gene. For each phenotype, we normalized the raw counts into counts per million (CPM) and selected the top 14,000 autosomal genes with the highest median CPM in standard RNA-seq data for downstream analysis.

To quantify splicing efficiency of introns, we used SPLICE-q[68] version 1.0.0 to calculate the reverse intron expression ratio in protein-coding genes. To quantify splicing in the RNA-seq datasets, we extracted junction reads by running regtools version 0.5.2 (ref. 69) on the filtered Binary Alignment Map (BAM) files. We obtained intron clusters using Leafcutter[56]

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 of the annotated protein-coding transcript. See the 'Classification of unannotated splice junctions' section in Supplementary Methods for a detailed description.

See Supplementary Methods for a detailed description of our mapping and quantification pipeline and for a description of our analysis of Oxford Nanopore Technologies long-read RNA-seq data.

Molecular *cis*-QTL mapping

To prepare the LCL RNA-seq data for eQTL mapping, we *Z*-score normalized the top 14,000 expressed protein-coding genes samples.

Then, we enforced normality of the data by applying the rank-based inverse normal transform to each sample across all genes. We utilized the same procedure to normalize the gene coverage quantification for H3K36me3. For hQTLs from ChIP-seq data, we normalized the read counts quantified by featureCounts at the peaks called by MACS2.

For sQTL mapping, we used the prepare_phenotype_table.py script from Leafcutter. This script calculates the PSI of each junction, it applies *Z*-score normalization to each junction across all samples, and enforces normality by rank-normalizing each sample across all junctions.

To build a covariate matrix for QTL mapping for each dataset, we applied principal component analysis on the standard and rank-normalized feature-by-sample matrix. We selected the number of principal components that explain more variance than in a version of the data matrix obtained by randomly permuting each feature across the samples.

For molecular *cis*-QTL calling, we used as input for each dataset the standard and rank-normalized data matrices, the principal components (PCs) obtained as previously described and the corresponding VCF files from the 1000 Genomes Project.

We ran QTLTools[70] version 1.3.1 for molecular *cis*-QTL mapping, using both the permutation and nominal pass versions. For the permutation pass, we used 1,000 permutations using the permute 1000 flag. For the nominal pass, we used the nominal 1 flag to obtain the QTL statistics of all SNPs irrespective of their *P* value. For eQTLs and hQTLs, we used a *cis*-window of 100,000 bp. For sQTLs and splicing efficiency QTLs, we used a *cis*-window of 10,000 bp. For the permutation pass of sQTLs, we used the grp-best flag to get only the statistics of the best hit per intron cluster. For the permutation pass, we applied the Benjamini–Hochberg correction to the adjusted β distribution *P* values to account for the false discovery rate.

For analyses that require unstandardized effect size estimates, we considered the top QTL SNP from the QTLtools pass using standardized and rank-normalized data matrices and reran QTLtools using the same PC covariates but providing unstandardized expression matrix (that is, $\log_2$RPKM for expression or gene level PSI for splicing), such that the β estimate is in more directly interpretable units.

See 'eQTL calling on GTEx gene expression data' section in Supplementary Methods for details on the QTLTools runs on GTEx data.

naRNA-seq experimental methods

Cell growth. LCLs were grown in Roswell Park Memorial Institute media (RPMI) + glutamine + penicillin–streptomycin + 20% fetal bovine serum. Cells were grown in four batches, with approximately 20–35 cell lines per batch. The day before collection, cell cultures were counted and normalized across cell lines to approximately 35 million live cells, supplemented with media to 50 ml, such that cells were in log-phase growth with approximately 50 million cells per culture at the time of collection.

Isolation of naRNA. We collected naRNA from LCLs by first isolating nuclei through a sucrose cushion followed by high-salt washes to dissociate nucleoplasm and weakly bound RNAs and proteins from chromatin with slight modifications from previous protocols for cellular fractionation[71], as detailed below.

Cells were collected by centrifugation (300*g* for 3 min) in 50 ml conical tubes. Cells were washed twice in cold phosphate-buffered saline with 1 mM EDTA. We reasoned that the inclusion of EDTA (and exclusion of magnesium) in buffers used during cellular fractionation would inhibit splicing during sample processing. However, we have noted that chelation of magnesium yields fragile nuclei that are more prone to premature bursting, and care should be taken to pipet gently in subsequent steps. Washed cell pellets were resuspended in 400 µl BufferA (10 mM pH 7.5 HEPES, 10 mM KCl, 10% (v/v) glycerol, 11.6% (w/v) sucrose, 1 mM dithiothreitol (DTT), 1× ROCHE complete protease inhibitor) and transferred to 2 ml 96-well plates for convenience. An equal

volume of BufferA supplemented with 0.2% TritonX was gently mixed to the resuspended cells, bringing the final TritonX concentration to 0.1% for a 12 min incubation on ice with periodic inversion. Nuclei were isolated by centrifugation (1200g, 5 min). Supernatant was discarded, and the nuclei pellet was washed with 500 µl BufferA, followed by centrifugation (1,200g, 5 min), removal of supernatant, and resuspension in 250 µl nuclear resuspension buffer (20 mM pH 7.5 HEPES, 50% (v/v) glycerol, 75 mM NaCl, 1 mM DTT, 0.5 mM EDTA, 1× protease inhibitor). An equal volume of NUN (high salt) buffer (50 mM pH 7.5 HEPES, 1 M urea, 300 mM NaCl, 1 mM DTT, 1× protease inhibitor) was added to nuclei pellets and gently mixed, following by 5 min incubation on ice and periodic inversion. Chromatin pellets were isolated by centrifugation (1,200g, 5 min). Nonchromatin-bound supernatant was removed. Chromatin pellets were washed with 500 µl BufferA supplemented with 0.2% NP-40, followed by centrifugation and removal of supernatant. The usually insoluble chromatin pellet was resuspended in 100 µl nuclear resuspension buffer, added to 1 ml Trizol, and stored in 1.5 ml centrifuge tubes at −20 °C for further processing.

Trizol samples were vigorously mixed with periodic heating at 50 °C until the pellet dissolved. After adding 200 µl chloroform, vigorously mixing and centrifugation (16,000g, 15 min), the aqueous phase was transferred to clean tubes or 96-well plates. An equal volume of ethanol was mixed, and samples were bound to Zymo spin I-96-XL plate (Catalog no. C2010) or individual Zymo Spin ii (Catalog no. C1008-50) columns by centrifugation (2,000g, 2 min). Columns were washed twice with 500 µl wash buffer (80% ethanol, 10 mM pH 8.0 Tris buffer). Samples were treated with DNAseI (2.5 µl RQ1 Promega DNAseI, 1× DNAse buffer in 25 µl total volume) while bound to columns and incubated at room temperature for 15 min. Columns were washed with RNA binding buffer (2 M guanidinium, 75% isopropanol), followed by two washes with 80 µl wash buffer and a dry spin. RNA was eluted by incubation with 25 µl water for 5 min followed by centrifugation. RNA yield was quantified by NanoDrop, typically yielding 5–20 µg naRNA per sample.

RNA-seq library preparation. rRNA was depleted using Lexogen ribocop v2 kit according to the manufacturer's protocol. RNA was eluted in 8 µL water, all of which was used as input for NEB Ultra Directional II RNA-seq kits according to manufacturer's protocol. Fragmentation time was adjusted from the recommended 15 min to 5 min to obtain larger insert sizes. Samples were pooled and sequenced on a single NovaSeq flow-cell (2 × 150 bp paired end reads) by UChicago sequencing core.

H3K36me3 CUT&Tag
Cells were grown as described above (naRNA-seq experimental methods) and 100,000 cells were frozen in 10% dimethylsulfoxide (DMSO). CUT&Tag was performed as previously described[72] (detailed protocol at https://doi.org/10.17504/protocols.io.z6hf9b6) using polyclonal antibody ('Reporting summary' section), with the following modifications. Rather than 13 polymerase chain reaction (PCR) cycles, we used 14 cycles. We determined this on the basis of a test qPCR with ~10% of the pre-PCR library, estimating that 14 PCR cycles with the remaining 90% would yield fluorescent signal at about halfway to the plateau, ensuring we have enough DNA material to quantify and sequence.

Molecular QTL sharing
Sharing of molQTLs between datasets and colocalization. We used Storey's π_1 statistic[73], an estimate of the fraction of non-null hypothesis from a distribution of P values, to get the fraction of molecular QTLs that are shared across multiple phenotypes (for example, 'what fraction of eQTLs discovered in steady-state RNA are eQTLs in naRNA?'). See 'π_1 sharing of eQTLs between RNA-seq datasets' and 'π_1 sharing of eQTLs and hQTLs at TSS' sections in Supplementary Methods for details.

We simultaneously assessed colocalization of molQTLs around each gene using hyprcoloc[74]. Summary statistics for a 100 kb

cis-window surrounding the gene were obtained for each molQTL from QTLtools nominal pass. Only molQTLs with a permutation pass $P < 0.01$ were considered for colocalization.

Effect size concordance of sQTLs and eQTLs. For each cluster with a significant sQTL intron, we classified the sQTL as a u-sQTL if it contains at least one sQTL intron (false discovery rate (FDR) <10%) in an unproductive intron. sQTLs with a nominally significant hQTL ($P < 0.01$, for any H3K4me3, H3K27ac or H3K36me3 trait) were filtered out. To avoid plotting nonindependent sQTLs, we selected only a single sQTL intron per cluster, retaining the intron with the largest absolute value of sQTL β. The top SNP of remaining sQTLs was used to look up the corresponding eQTL effect size in the host gene. Test for significance of correlation was performed with 'cor.test(method='spearman')' in R.

Effects of AS–NMD-mediated eQTLs and transcription-mediated eQTLs across tissues. We first identified a set of 359 eQTLs, which we were most confident are mediated by AS–NMD. Specifically, the steady-state eQTL must colocalize with an u-sQTL within the gene ('Colocalization of molQTLs' section in Supplementary Methods) and not with any hQTL. Furthermore, the lead SNP for the colocalizing u-sQTL junction must not be nominally significant ($P > 0.01$) for any hQTL test feature (considering all H3K34me3, H3K27ac peaks within 100 kb of the gene and H3K36me3 for the gene body). The top colocalized eQTL/u-sQTL SNP was then assessed for effect size across GTEx tissues ('eQTL calling on GTEx expression data' section in Supplementary Methods). We similarly identified a set 313 eQTLs which we were most confident are mediated by expression effects: the steady-state eQTL must colocalize with an hQTL at the eQTL gene's promoter (see 'π_1 sharing of eQTLs and hQTLs at TSS' section in Supplementary Methods for linking hQTLs to promoters; H3K4me3, H3K27ac and H3K4me1 peaks were considered). The top colocalized eQTL/hQTL SNP was then assessed for effect size across GTEx tissues ('eQTL calling on GTEx expression data' section in Supplementary Methods). To assess effect sizes across tissues, we used the (not quantile-normalized) log$_2$CPM matrix for eQTL mapping with QTLtools, enabling interpretable comparisons of effect sizes (β) across tissues in units of log$_2$ fold-change per alt allele.

Colocalization of molQTLs with GWAS loci. Summary statistics for GWAS[75–81] were downloaded from the GWAS Catalog or other source datasets. One mega-base windows centered at lead GWAS SNPs were determined as previously described[58], using a lead SNP threshold of 5×10^{-8} to consider locus for colocalization. Summary statistics for all molQTL features within each GWAS locus window were obtained using QTLtools nominal pass, after selecting only those with a QTLtools permutation pass $P < 0.01$. Hyprcoloc was used with default settings to colocalize molQTLs and GWAS signals. Loci were categorized as 'Only hQTL', 'Only eQTL' or 'Only sQTL', if the only molQTLs to colocalize with the GWAS signal were either hQTLs (H3K36me3, H3K27ac, H3K4me3 or H3K4me1), eQTLs (in naRNA, 4sU or steady-state RNA-seq) or sQTLs (in naRNA, 4sU or steady-state RNA-seq), respectively. If all of the molQTLs could be classified as eQTLs or hQTLs, or eQTLs and sQTLs, the loci was classified as 'hQTL+eQTL' or 'sQTL+eQTL', respectively. All other loci with a molQTL colocalization were classified as 'other combinations'. We chose to classify these loci as 'other combinations' because sQTLs that also colocalize with both eQTL and hQTL or with apaQTL or other combinations are hard to interpret, and we wanted to refrain from suggesting these loci may be mediated by AS–NMD or alternative protein isoforms caused by AS.

Risdiplam dosage series experiment
Cell growth, library preparation and RNA-seq. LCL growth, naRNA isolation and conversion to sequencing libraries were performed as described above (naRNA-seq experimental methods). Risdiplam was

added 24 h before cell collection. Total ('steady-state') RNA was converted to sequencing libraries using NEB polyA capture kit (product no.) followed by NEB Ultra directional II RNA-seq library kits. Steady-state RNA-seq libraries were sequenced by (2 × 150 bp paired end). naRNA-seq libraries were sequenced by UChicago sequencing core (2 × 150 bp paired end).

Identification and quantification of risdiplam-induced exons. All splice junctions containing GA|GU in reference genome sequence at the 5'ss were assessed for a significant positive dose:response correlation using leafcutter's intron excision ratio as the response. Significance was assessed using R's 'cor.test(…, method='spearman')' function, and P values were adjusted for multiple testing with Storey's q-value method. All significant ($q < 0.1$) GA|GU introns with a leafcutter-clustered splice junction that has a 3' splice site (3'ss) 500 bp upstream of the GA|GU splice donor were considered as risdiplam-induced cassette exons. Splicing at these cassette exons was requantified using the cassette exon PSI metric (see 'Analysis of alternative splicing and symmetry of cassette exons' in Supplementary Methods).

Prediction of transcripts effects. We utilized Gencode transcript structures and their predictions for coding potential ('basic'-tagged transcripts being deemed as productive) to annotate cassette exons that use an annotated GA|GU downstream splice junction. Cassette exons with an unannotated downstream GA|GU splice junction were translated in-frame using the most-expressed 'basic'-tagged transcript (transcript quantifications derived from Salmon[82]) as a reference, using custom scripts. For simplicity, we classified unannotated exons as unproductive if the exon-included translation is shorter than the exon–excluded translation. While splicing in UTRs may trigger NMD by 50 nt rule, all risdiplam-induced exons in UTRs are, by definition, already in an intron-containing UTR and, thus, we did not classify those as unproductive.

Quantification of expression and splicing. Gene expression was quantified using featureCounts as described above. naRNA-seq samples were analyzed using edgeR[83], with two contrasts: (1) 100 nM risdiplam versus DMSO and (2) 3,160 nM risdiplam versus DMSO contrast. The unconventional experimental design of the titration series experiment (a single replicate at eight doses) precluded use of many standard differential splicing or differential gene expression analysis approaches. For those samples, we fit splicing quantifications (in units of cassette exon PSI) and gene expression quantifications (in units of log2CPM) to a four-parameter log-logistic curve using the drc package[84] in R: 'drc::drm(formula = cpm ~ dose, fact=LL.4(), …)'. Effect size estimates and standard errors at 100 nM and 3,160 nM were extracted from the model fits using 'predict(…, se.fit=T)'. FDR was estimated with Storey's q-value. Genes were classified as transcriptionally regulated if they had an absolute $\log_2 FC$ between naRNA and steady-state RNA less than 1.5, with FDR <0.1 in both naRNA and steady-state RNA. If genes had FDR <0.1 in steady-state RNA but with an effect size difference between naRNA and steady-state greater than $\log_2(1.5)$, they were classified as post-transcriptionally downregulated. Similarly, genes had FDR <0.1 in steady-state RNA but with an effect size difference between naRNA and steady state less than $\log_2(1.5)$, they were classified as post-transcriptionally upregulated.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Publicly available data sequence data generated as part of this study (naRNA-seq and H3K36ME3 Cut&Tag) is publicly available and has been deposited in Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) under accession GSE252006. Other publicly data utilized in this study included genotypes were downloaded from the 1000 Genomes project[62] (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/). FASTQ files of short-read RNA-seq data of shRNA dKD of *SMG6* and *SMG7* in HeLa cells and shRNA controls from a previous study[24] (SRA accession SRP083135) are available at https://www.ncbi.nlm.nih.gov/sra. Other publicly available short-read sequencing data are described in Supplementary Fig. 1. For these datasets, we obtained FASTQ files of standard short-read RNA-seq data from the following accession numbers: ENA project accession PRJEB3365 (steady-state RNA-seq produced by the GEUVADIS consortium), SRA project accession PRJNA268086 (H3K4me1, H3K4me3 and H2K27ac ChIP-seq) and SRA project accession PRJNA302818 (4sU RNA-seq). The 3' sequencing APA data were obtained as a sample by peak expression matrix from authors of a previous study[40] (SRA accession number SRP223759). The data were aligned to GRCh38 and transcript release v34 annotations from Gencode (https://www.gencodegenes.org/human/). Some analyses (Supplementary Methods) also utilized v37 annotations.

Code availability

Pipelines and all original code are available at Zenodo via https://doi.org/10.5281/zenodo.12571961 (ref. 85) and also at Github via https://github.com/bfairkun/ChromatinSplicingQTLs/.

References

62. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
63. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
64. Van De Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
65. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
66. Zhang, Y. et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
67. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
68. De Melo Costa, V. R., Pfeuffer, J., Louloupi, A., Ørom, U. A. V. & Piro, R. M. SPLICE-q: a Python tool for genome-wide quantification of splicing efficiency. *BMC Bioinf.* **22**, 368 (2021).
69. Cotto, K. C. et al. Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nat. Commun.* **14**, 1589 (2023).
70. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
71. Pandya-Jones, A. & Black, D. L. Co-transcriptional splicing of constitutive and alternative exons. *RNA N. Y. N.* **15**, 1896–1908 (2009).
72. Kaya-Okur, H. S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1–10 (2019).
73. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
74. Foley, C. N. et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat. Commun.* **12**, 764 (2021).
75. Ishigaki, K. et al. Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat. Genet.* **54**, 1640–1651 (2022).

76. Ferreira, M. A. R. et al. Genetic architectures of childhood- and adult-onset asthma are partly distinct. *Am. J. Hum. Genet.* **104**, 665–684 (2019).

77. Demenais, F. et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* **50**, 42–53 (2018).

78. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).

79. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).

80. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).

81. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429.e19 (2016).

82. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

83. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

84. Ritz, C., Baty, F., Streibig, J. C. & Gerhard, D. Dose–response analysis using R. *PLoS ONE* **10**, e0146021 (2015).

85. Fair, B. & Buen Abad Najar, C. F. bfairkun/ChromatinSplicingQTLs: Fair_et_al_NatGen2024. *Zenodo* 10.5281/ZENODO.12571961 (2024).

86. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat. Genet.* 1–12 (2024).

Acknowledgements

Author contributions

B.F. and C.F.B.A.N. contributed equally. Y.I.L., J.W. and J.P.S. jointly supervised research. B.F., C.F.B.A.N., Y.I.L., J.W. and J.P.S. conceived and designed the experiments. B.F., S.L., A.R., G.M. and J.Z. performed the experiments. B.F., C.F.B.A.N. and Y.I.L. performed statistical analyses, analyzed the data and wrote the paper.

Competing interests

The authors declare no competing interests. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01872-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01872-x.

Correspondence and requests for materials should be addressed to Yang I. Li.

Peer review information *Nature Genetics* thanks Boxiang Liu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at www.nature.com/reprints.
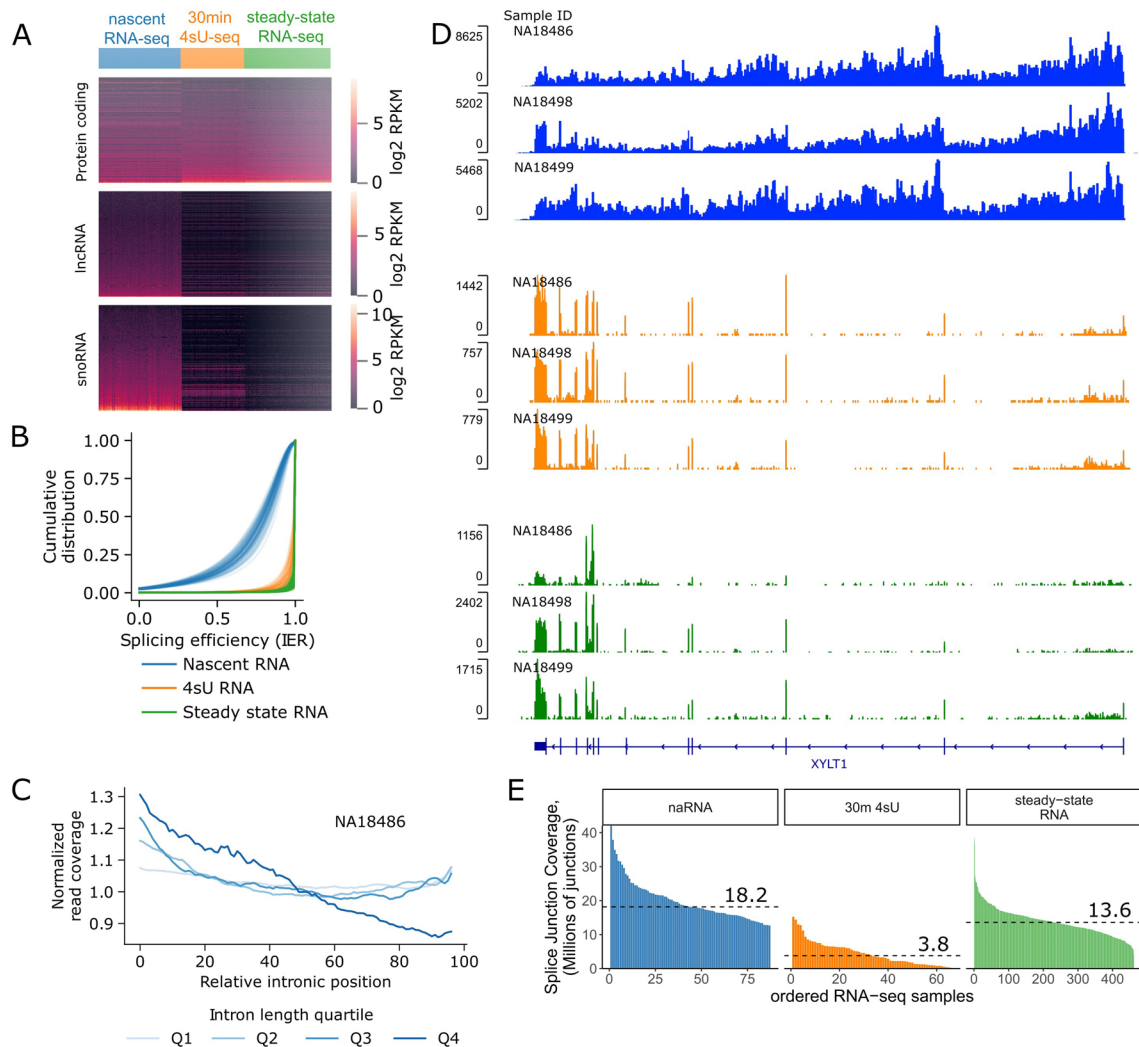
Extended Data Fig. 1 | Overview of nascent RNA-seq. (**a**) Nascent RNA-seq (naRNA-seq) captures nuclear-retained, non-polyadenylated, and rapidly decayed RNAs (snoRNAs, lncRNAs), that are absent from labeled and steady-state RNA-seq datasets. Each column represents an RNA-seq sample, grouped by the dataset type, each row a different gene. (**b**) naRNA-seq transcripts are only partially spliced. The splicing efficiency metric is based on the ratio of spliced and unspliced (intron:exon junction) reads, and varies between 0 and 1, with 1 indicating all reads are spliced. The cumulative distribution of splicing efficiency across all introns in expressed genes, for each RNA-seq sample from naRNA, recently transcribed RNA (30 min 4sU), and steady-state RNA. (**c**) Meta-intron coverage plot in LCL naRNA-seq sample NA18486 confirms the expected 5′ bias in intronic coverage genome-wide, consistent with the nascent nature of naRNA transcripts. Longer introns are naturally expected to have steeper slopes than short introns when intron lengths are rescaled for metaplot. (**d**) naRNA-seq transcripts are only partially spliced. Example sawtooth pattern in nascent RNA in the gene *XYLT1*. The nascent nature of transcripts in naRNA creates a 5′ bias in coverage, and in combination with co-transcriptional splicing, creates a sawtooth pattern of coverage. (**e**) Number of exon-exon splice junction reads in RNA-seq samples. The median in each dataset is marked with a labeled dashed line. Though naRNA-seq is only partially spliced, our deeper sequencing of naRNA-seq results in high coverage of splice junctions, allowing measurements of splicing before cytoplasmic decay.
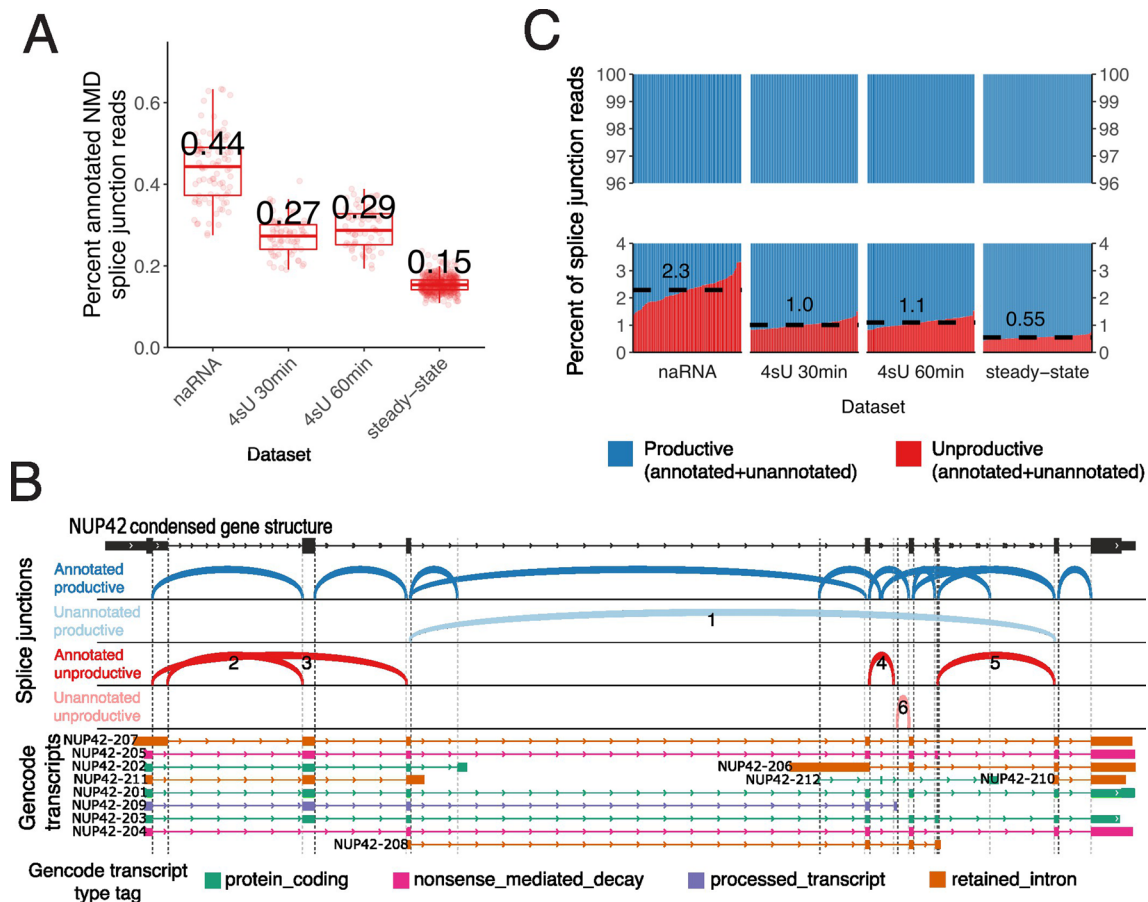
Extended Data Fig. 2 | Classification and quantification of splice junction classes across datasets. (**a**) The percent of splice junctions in each sample that are uniquely attributable to transcripts tagged as 'nonsense_mediated_decay' (Gencode v37). Box and whiskers show quartiles for LCL samples (individual jittered points) in each RNA-seq data-type (n=86, 66, 66, and 462 for naRNA, 4sU 30min, 4sU 60min, and steady-state RNA-seq, respectively). Median for each data-type is labeled. (**b**) Splice junctions (arcs) overlapping the *NUP42* gene illustrate approach (Supplemental Methods) for classifying splice junctions. Annotated splice donors and splice acceptors are marked with vertical dashed lines in dark and light gray, respectively. Annotated productive junctions are defined by their presence in at least one transcript with the value of 'protein_coding' in the Gencode transcript type tag. Unannotated productive junctions are not in any Gencode transcripts, and skip exons in the principal isoform such that the reading frame is maintained (that is, splice junction marked with 1).

Annotated unproductive junctions are unique to Gencode transcripts not tagged with 'protein coding'. Splice junction 2 is unique to *NUP42-207*, a 'retained_intron' tagged transcript. This splice junction uses a deep intronic 5′ss, creating a premature termination codon. Junctions 3 and 5 are unique to transcripts tagged as 'nonsense_mediated_decay', and junction 4 is unique to a transcript tagged with 'processed_transcript'. All other junctions are classified as Unannotated unproductive. We attempted to translate the resulting transcripts that use these junctions, finding that they overwhelmingly introduce frameshift or in-frame stop codons (Supplemental Methods), such as the splice junction 6 which we predict to introduce a frameshift. (**c**) Similar to (**b**), where sample is represented as a column, grouped by dataset type, and the fraction of splice junction reads that are either productive (annotated or unannotated, classified as in (**b**), blue) or unproductive (annotated or unannotated, classified as in (**b**), red). The median in each dataset is marked with a dashed line and labeled.
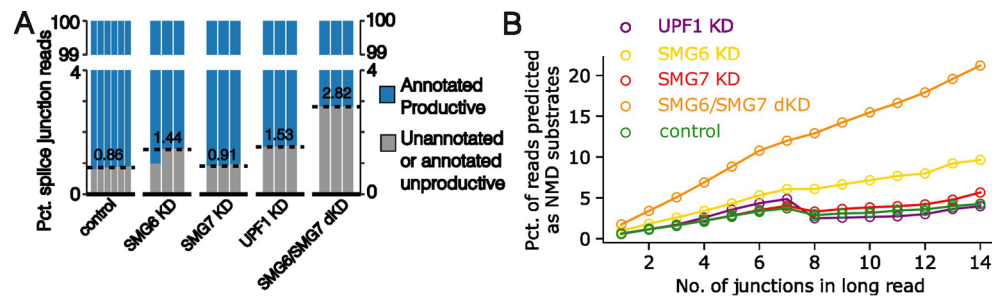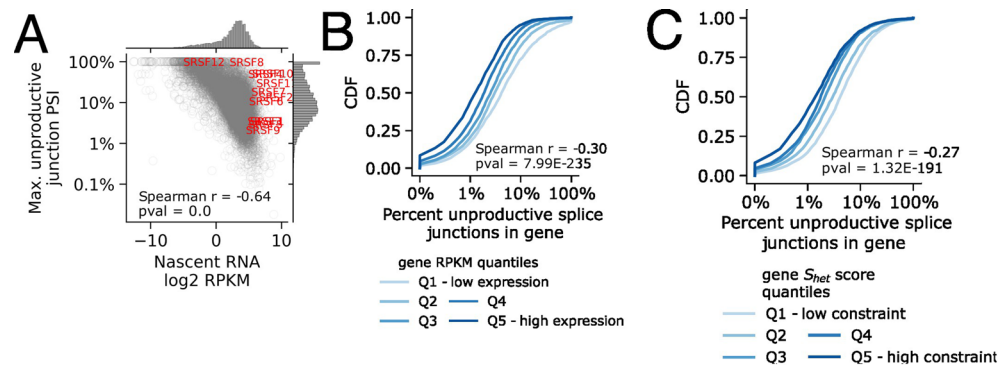
Extended Data Fig. 3 | Percentage of unproductively spliced reads upon knockdown of NMD factors. (a) Fraction of splice junction reads in each short read steady-state RNA-seq sample that are in Gencode-annotated productive transcript structures (blue), versus unannotated or annotated unproductive transcript structures (gray). Biological replicates represented by each column, with dashed lines to indicate median for each group. NMD factors were knocked-down (KD) singly or as double knockdown (dKD) with shRNA in HeLa cells with an shRNA scramble control[24]. (b) Nanopore long-read sequencing quantifies the percent of full-length reads that are targeted by NMD, defined as containing at least one unproductive junction, as a function of the number of splice junctions in the read. Knockdown experiments of similar design as in (a)[23].
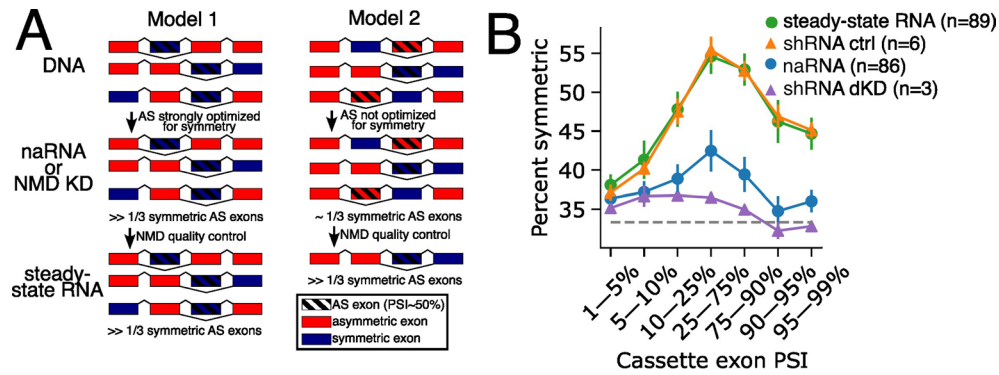
Extended Data Fig. 4 | Enrichment of symmetric exons among alternatively-spliced exons (AS exons) is largely the result of NMD, rather than optimized splicing. (a) Two opposing models to explain the observation that AS exons (defined as PSI ~ 50%) are enriched for symmetric exons (for example, length divisible by three) in steady-state RNA: (**Model 1**) AS exons are strongly enriched for symmetric exons, or (**Model 2**) AS exons are not more likely to be symmetric than random expectation or constitutive exons (~1/3 symmetric), but NMD efficiently eliminates frame-shifting AS exons such that they appear enriched for symmetric exons in steady-state RNA but not RNA that directly measures splicing outcomes without the influence of degradation, such as naRNA or RNA after knockdown of NMD factors. In both models, we expect that constitutive exons (PSI~100%) are not strongly enriched for symmetric exons since the reading frame can cross exon-boundaries without consequence if exons are truly constitutive. Although constitutive and AS exons cannot be experimentally distinguished by observations of DNA, we included the gene structures at the DNA level because the two models imply differing selection pressures on DNA sequence to maintain (Model 1), or not maintain (Model 2) accurate frame-preserving AS patterns. (**b**) Fraction of exons that are symmetric as a function of their usage, estimated as percent spliced in (PSI). Error bars represent standard error across LCL lines treated as replicates, and the standard error for replicate shRNA knockdown experiments (triangular markers, data from[24]). Unlike in steady-state RNA, the enrichment for symmetric exons among AS exons is not apparent in naRNA or NMD KD.

Extended Data Fig. 5 | Unproductive splicing is less abundant in highly transcribed and highly constrained genes. (a) Correlation between gene expression and the maximum junction PSI of any unproductive junction in the gene, a proxy for percent of unproductive transcripts. 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%. SRSF genes (red) are well-known examples[12,21] of genes with high gene expression and high unproductive junction PSI. **(b)** Highly expressed genes have a lower unproductive splicing rate, as measured by the genewise percent of splice junction reads that are unproductive. Correlation summarized with spearman correlation coefficient and P value. Correlation visually presented as cumulative distribution of percent unproductive splice junctions, grouped by expression quintiles. **(c)** Similar to B, showing that genes with a higher S_{het} score[86] (suggesting more selective constraint) have a lower unproductive splicing rate. Correlations in each panel summarized with spearman correlation coefficient and two-sided correlation test P value.
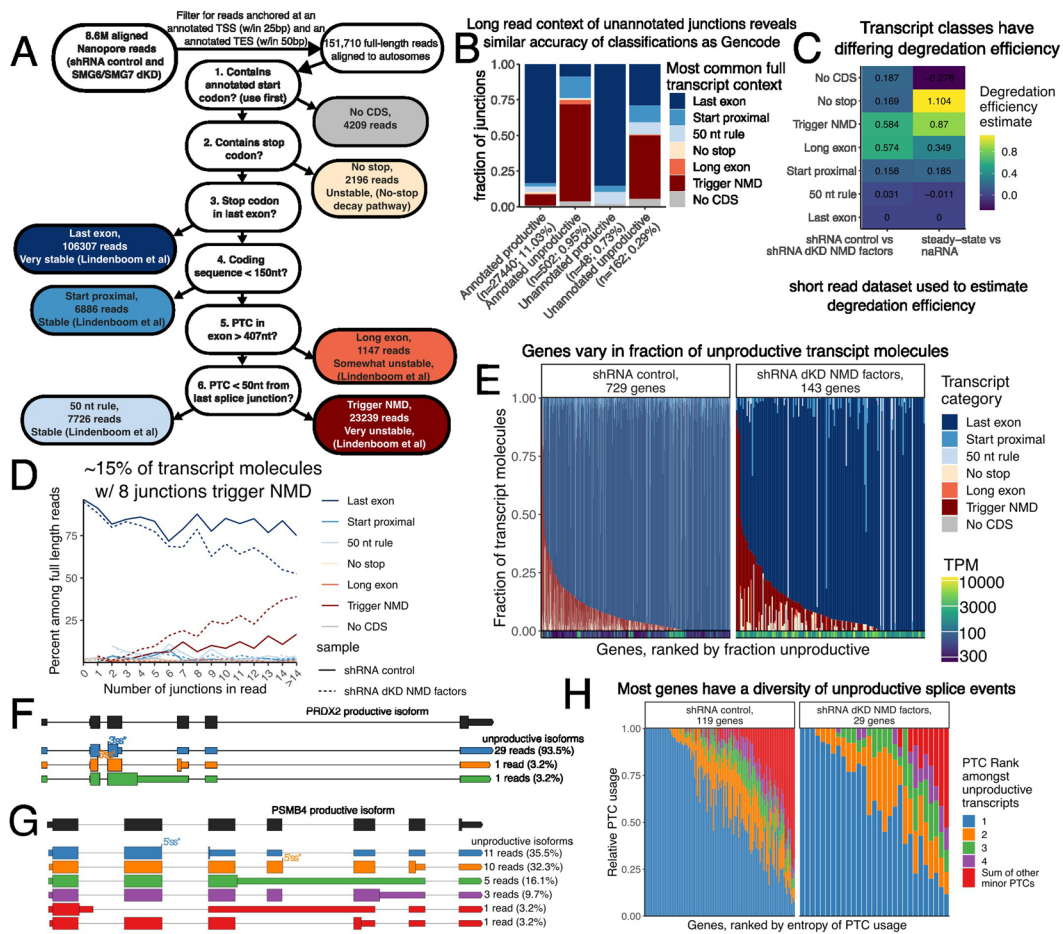
Extended Data Fig. 6 | Assessing AS-NMD prevalence and diversity using full-length (FL) long reads. (a) Pipeline to assess NMD status of transcripts (See Supplemental Methods). Aligned Oxford Nanopore cDNA reads of SMG6/SMG7 double knockdown (dKD) and shRNA control[23] filtered for FL reads (anchored at annotated transcript termini). FL reads were translated from the first annotated start codon, classified with decision tree (steps 2-6 are previously established rules regarding NMD-targeting efficiency of transcripts[44]) into seven transcript categories referenced in (**b-e**). Categories on right are qualitatively considered 'unproductive' in (**e-h**); categories on the left are 'productive'. Number reads in categories shown. (**b**) Each splice junction observed in short read data was classified as productive or unproductive (and annotated or unannotated) (Supplemental Methods, Extended Data Fig. 2). FL reads were used to assess accuracy of junction-level classifications by considering most common transcript categories of FL reads containing that junction (requiring >2 FL reads). Fraction unique junctions in each short-read category matching most common context (transcript categories) plotted as bars. Limited number

FL reads means only n junctions in each category (% of total in category) were assessed, noted on x-axis. (**c**) Relative degradation efficiency of each category estimated by comparing the median relative splice junction abundance in control vs dKD[24], and steady-state vs naRNA, short read data. Consistent with previous reports[44], categories differ in degradation efficiency. (**d**) Percent FL reads in each category, as function of number splice junctions in read. (**e**) Fraction reads belonging in each category across genes (columns). Only genes with >20 reads considered. The dKD sample has fewer reads, and therefore stronger ascertainment bias, with 143 highly expressed genes (TPM, colored rug) passing this filter. (**f**) Isoform structure and relative abundance (read count, and percent of each isoform among unproductive reads in control samples) of unproductive isoforms derived from the *PRDX2*, where 93.5% of unproductive transcripts derive from most-common PTC-inducing splice junction (blue). (**g**) Same as (**f**), for *PSMB4*, which has greater diversity of unproductive splice junctions. (**h**) Diversity of unproductive isoforms amongst all genes (columns) with at least 20 unproductive reads in control and dKD samples, respectively.
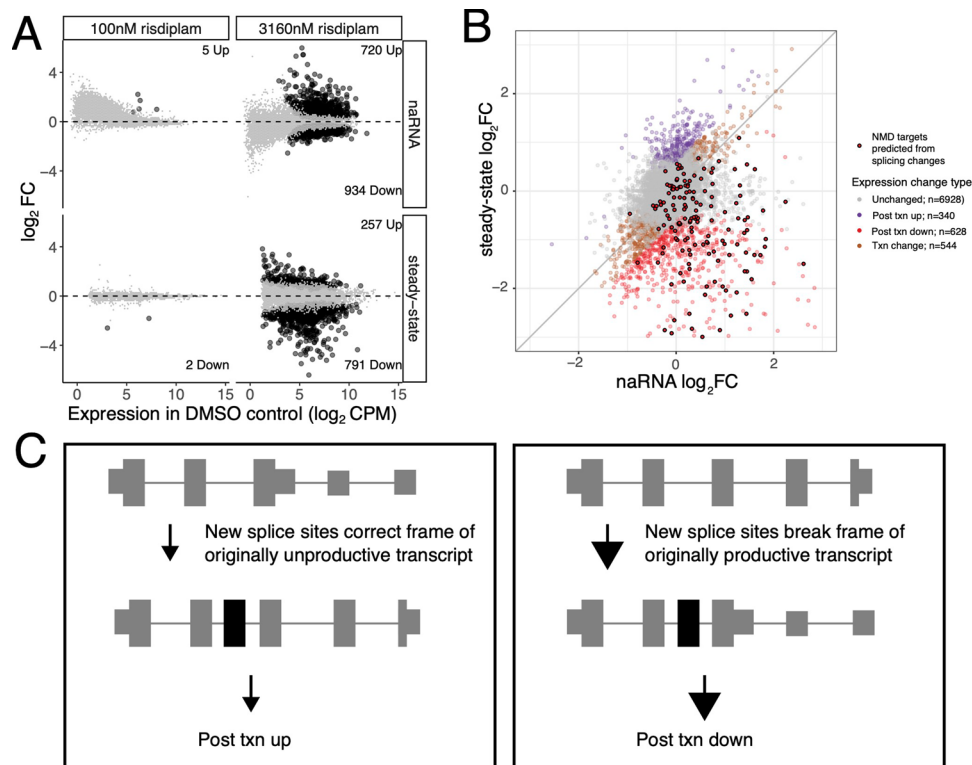
Extended Data Fig. 7 | Risdiplam primarily induces down-regulating-post-transcriptional changes. (**a**) MA-plots of differential expression upon treatment with low (100 nM) or high (3160 nM) dose of risdiplam in steady-state RNA or naRNA. Overlapping non-significant tests are reduced to gray hexbin, while significant tests (FDR<1%) represented as black dots. Number of significant up- or down-regulated genes is labeled to emphasize that while there are similar numbers of up- and down-regulated genes in naRNA, in steady-state RNA there is a relative over abundance of down-regulated genes. (**b**) Genes are classified by their effect size and significance (See Methods) of expression changes as measured in naRNA or steady-state RNA-seq after being treated with 3160nM risdiplam. Transcription (Txn)-based gene expression changes are defined as having similar and significant effects as measured in naRNA and polyA RNA. Genes regulated post-transcriptionally have stronger effects in steady-state RNA.

There are more post-transcriptionally down-regulated genes than up-regulated genes, suggesting risdiplam-induced splice sites more often destabilize than stabilize host transcripts. Expectedly, there is significant overlap of the 219 NMD targets predicted from annotation of induced cassette exons among the post-txn down regulated genes (Odds Ratio=14.0, $P<2x10^{-16}$, hypergeometric test for over-representation). (**c**) Left: Post-txn up-regulated genes may arise by splicing changes (such as risdiplam-induced exons, depicted as black cassette exon) that relieve NMD with a frame-correcting exon at a gene that is originally spliced into an unproductive isoform. The open reading frame is depicted as thick regions of exons. Right: More commonly, risdiplam-induced splicing changes result in post-txn down-regulated genes, consistent with splicing changes that break the reading frame of genes that are originally spliced into productive isoforms.
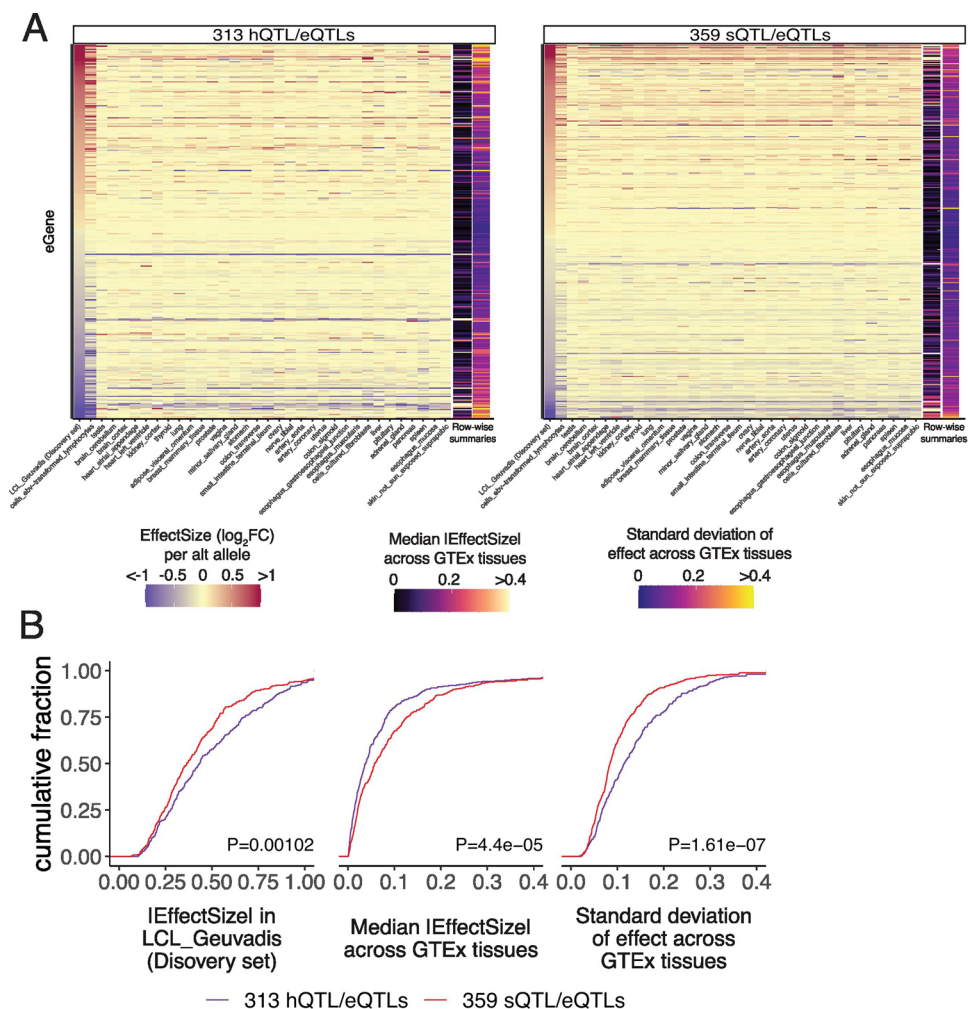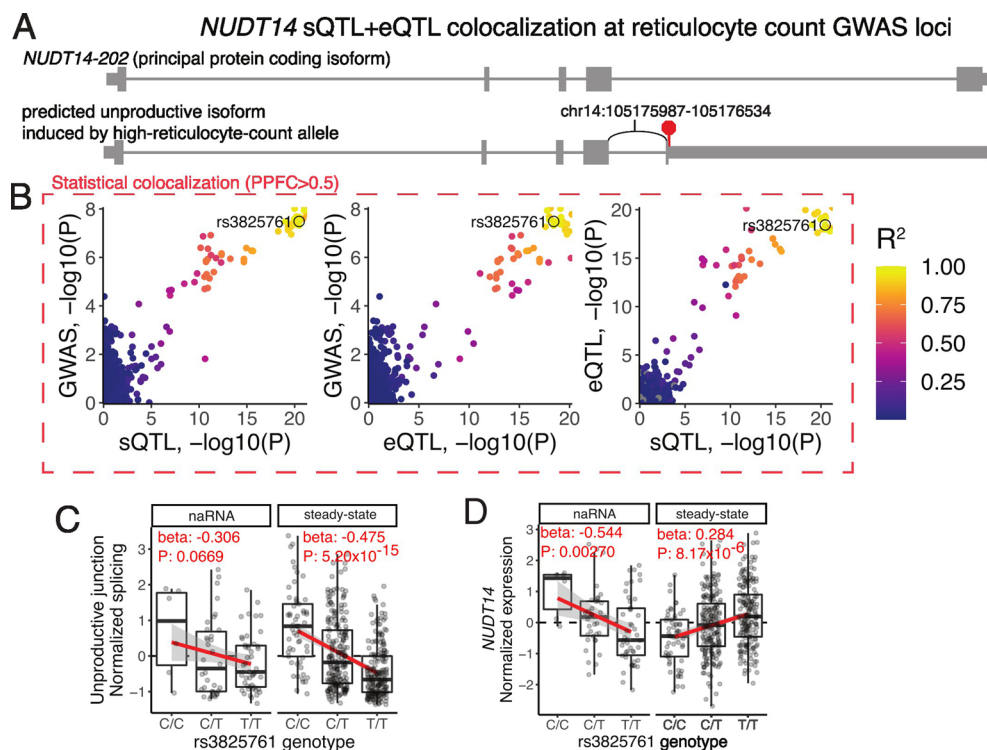
Extended Data Fig. 8 | Transcription-mediated eQTLs (hQTL/eQTLs) are more tissue-specific than splicing-mediated eQTLs (sQTL/eQTLs). (a) A set of transcription-mediated eQTLs (hQTL/eQTL colocalization) identified in our source dataset was compared to a set of splicing-mediated eQTLs (sQTL/eQTL colocalization without nominal hQTL signal), by estimating the SNP:gene effect across 38 GTEx tissues (columns) for each eGene (rows). Row-wise summary statistics were calculated and plotted as extra columns. **(b)** Row-wise summary statistics are plotted as cumulative distributions for visual contrast. From left to right, we see that (1) the absolute effect size in the Geuvadis LCL discovery dataset is slightly greater for hQTL/eQTLs than sQTL/eQTLs. Despite this, the sQTL/eQTLs have a (2) larger median effect size across GTEx tissues, and (3) have a smaller standard deviation of effect size across tissues. P values from two-sided Mann-Whitney test.
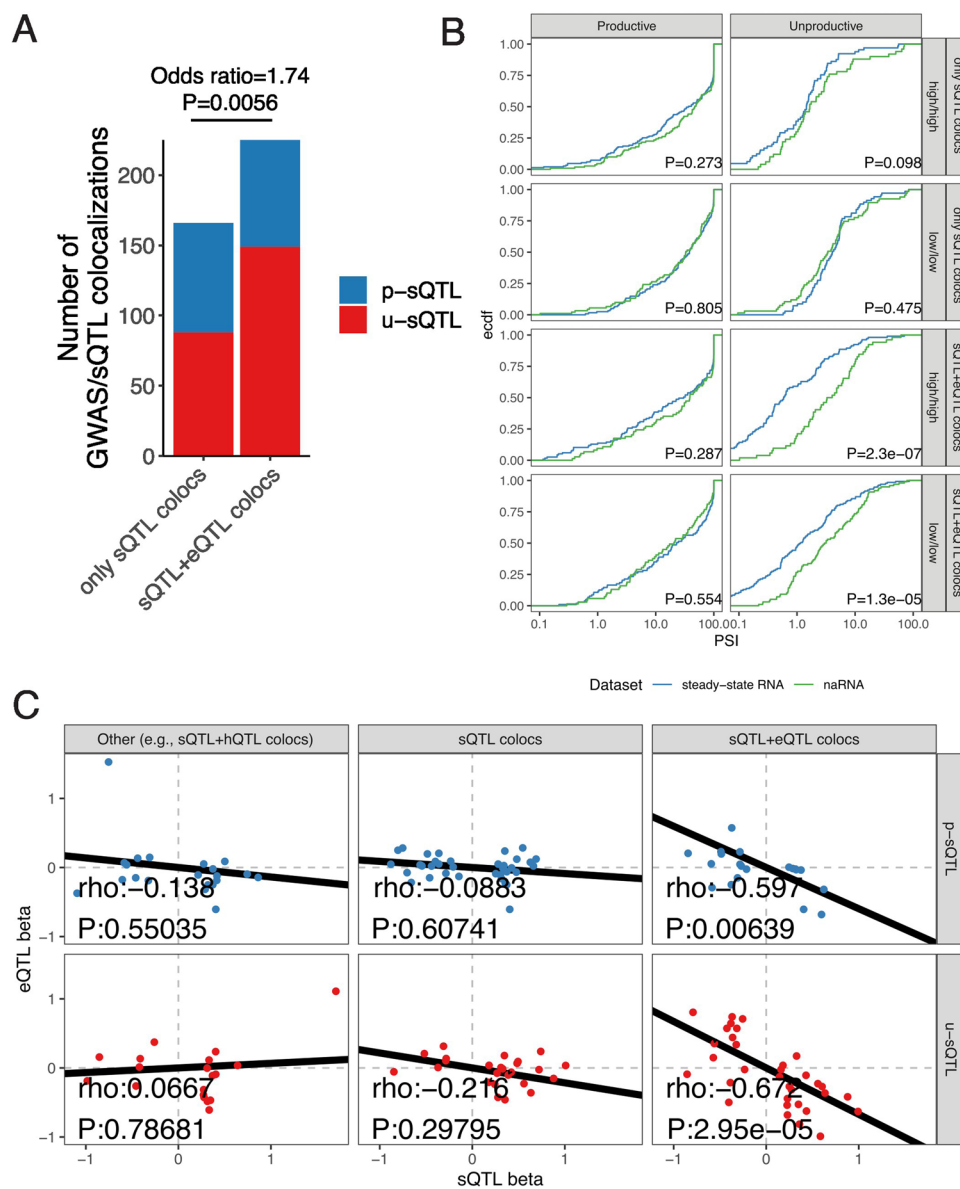
Extended Data Fig. 9 | GWAS/sQTL colocalizations that also colocalize with eQTL have characteristics consistent with AS-NMD. (a) sQTLs/GWAS colocalizations are more likely to come from u-sQTLs than p-sQTLs if the host gene eQTL also colocalizes in multi-trait colocalization analysis. P value from hypergeometric test for over-representation. **(b)** PSI distribution of introns as cumulative distribution plot for sQTLs that colocalize with eQTL and GWAS (sQTL+eQTL colocs) versus those that only colocalize with GWAS (sQTL colocs). We estimate PSI by averaging across samples with shared genotypes, either high/high genotypes or low/low genotypes (thus, avoiding confounding PSI estimates with different allele frequencies between datasets). sQTLs in unproductive introns that are sQTL+eQTLs have smaller PSI in steady-state RNA than naRNA, consistent with splicing-mediated decay at these GWAS loci transcripts. P-value for two-sided Mann-Whitney test. **(c)** Effect sizes of sQTLs that colocalize with GWAS, grouped by whether the GWAS signal also colocalizes with an eQTL (sQTL+eQTL colocs), whether it solely colocalizes with sQTL, or whether it also colocalizes with some other combination of traits (that is, sQTL + hQTL) in multi-trait colocalization analysis. Each junction is plotted once, even if it colocalizes with multiple GWAS loci across multiple traits. Correlation of effect sizes summarized as spearman rho correlation coefficient and two-sided correlation test P-value.

A *NUDT14* sQTL+eQTL colocalization at reticulocyte count GWAS loci

NUDT14-202 (principal protein coding isoform)

predicted unproductive isoform
induced by high-reticulocyte-count allele   chr14:105175987-105176534

B Statistical colocalization (PPFC>0.5)

C   naRNA   steady-state

beta: -0.306   beta: -0.475
P: 0.0669   P: 5.20x10⁻¹⁵

D   naRNA   steady-state

beta: -0.544   beta: 0.284
P: 0.00270   P: 8.17x10⁻⁶

Extended Data Fig. 10 | u-sQTL regulates *NUDT14* expression, likely contributing to reticulocyte count. *NUDT14* eQTL and sQTL. (**a**) Gene structure of *NUDT14-202*, the protein-coding isoform marked as the principal isoform by Gencode. Thick exonic regions mark the open reading frame. Using that isoform as a reference, we predicted the u-sQTL splice junction (labeled arc) that colocalizes with reticulocyte-count GWAS signal to introduce a premature stop codon (red octagon), created a transcript with a long 3′ UTR, inducing NMD. (**b**) Pairwise scatter plots depict the association between the GWAS signal, *NUDT14* eQTL signal, and chr14:105175987-105176534 junction sQTL signal. Each point is a SNP. All three traits colocalize in a single trait cluster in multi-trait colocalization (posterior probability of full colocalization, PPFC > 0.5,

see Supplemental Methods). SNPs colored according to linkage disequilibrium relative to the top fine-mapped SNP (rs3825761). (**c**) *NUDT14* sQTL boxplots showing unproductive splicing quartiles grouped by genotype show that the up-regulating effect of the C allele on the unproductive splice junction is present in steady-state RNA and naRNA, while the (**d**) down-regulating effect of the C allele on *NUDT14* expression is present in steady-state RNA but not naRNA, consistent with co-transcriptional splicing and post-transcriptional regulation by NMD. Box represents median and inner-quartiles. Whiskers extend from hinge to most extreme value no greater than 1.5 IQR from hinge. Beta and P-value from linear model to test association between genotype and normalized phenotype.

nature portfolio

Corresponding author(s): Yang Li

Last updated by author(s): Jun 3, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Code is available at https://github.com/bfairkun/ChromatinSplicingQTLs. This includes a reproducible pipeline with specific software versions and all new code. More description, including the exact version for all softwares used, is available in the Methods
Data analysis	Code is available at https://github.com/bfairkun/ChromatinSplicingQTLs. This includes a reproducible pipeline with specific software versions and all new code. More description, including the exact version for all softwares used, is available in the Methods

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Sequence data (naRNA-seq and H3K36ME3 Cut&Tag) is publicly available and has been deposited in Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/)

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender (identity/presentation), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	We aimed for a sex-balanced study design. naRNA-seq dataset contained 41 female, 46 male. H3K36me3 dataset contained 45 female, 48 male
Reporting on race, ethnicity, or other socially relevant groupings	We utilized lymphoblastoid cell lines (LCLs) in the NIGMS cell repository from Yoruban ancestry donors (YRI collection), as this collection of cell lines has been extensively molecularly characterized for molecular QTL studies. Genetic ancestry was not explicitly accounted for in QTL mapping analyses, though all YRI lines utilized are unrelated and from similar ancestry.
Population characteristics	See above
Recruitment	Cell lines were obtained from Coriell, under the MTA agreement: MTA000215_Li_Coriell_MTA_3_3_2021
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   ☐ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](#)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size is typical for similar molecular QTL studies: at least n>50, with greater sample sizes typically being better powered for more discoveries. We were able to achieve a sample size of n=93 liens for H3K36me3 Cut&Tag, and n=86 cell lines for naRNA-seq. No statistical method was used to predetermine sample size for novel data (naRNA-seq, H3K36ME3 CUT&Tag, RNA-seq and naRNA-seq at various risdiplam doses). We note that our sample sizes are similar to previous QTL studies using this panel of cell lines (ref 37,38,40) and risdiplam transcriptomics studies (ref 45,48)
Data exclusions	As described in Statistics and Reproducibility Methods section: For eQTL and sQTL mapping using naRNA, we excluded line NA18855. This was not a predetermined decision. Rather, principal component analysis revealed this sample as an outlier, and differential expression analysis against other lines identified differentially expressed genes clustered along large sections of chromosomes, suggestive of chromosomal abnormalities.
Replication	For various analyses, we replicated our initial analyses by reanalysis of published datasets (e.g., the conclusion that ~2.3% of splice junctions are unproductive in our initial naRNA-seq analysis was replicated in published RNA-seq of cell lines subject to knockdown of NMD-factors). Each of our experiments was performed once, with no additional replication. Risdiplam experiment was performed once, with biological replicates (seperate dishes) for each treatment dose for naRNA-seq, and a single replicate for each dose in steady-state RNA. Analyses utilized replicates for maximum power, and discoveries were not further replicated in independent datasets
Randomization	The experiments were not randomized, as molQTL studies typically do not require randomization for confounders, which are assumed to already be randomized with respect to genotype. Experiments with risdiplam used a single isogenic cell line, grown in a single well-controlled batch, such that randomization is not needed.
Blinding	The Investigators were not blinded to allocation during experiments and outcome assessment, as cell lines (for QTL studies) and treatment were required knowledge for carrying out the experiments

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	H3K36Me3 CUT&Tag used 1uL of the following antibody per sample: Abcam catalog no. ab9050, lot GR3386101-2.
Validation	We validated by analyzing meta gene profile of CUT&Tag sequencing data (Figure S2). Further, manufacturer states "Our Abpromise guarantee covers the use of ab9050 in the following tested applications: ICC/ICF, WB, ChIP" (https://www.abcam.com/products/primary-antibodies/histone-h3-tri-methyl-k36-antibody-chip-grade-ab9050.html?productWallTab=Abreviews)

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	LCL line GM12878. Coriell cell repository, NIGMS collection, YRI sub-collection LCL lines: NA18486,NA18497,NA18498,NA18499,NA18502,NA18504,NA18505,NA18507,NA18508,NA18510,NA18511,NA18516,NA18519,NA18520,NA18522,NA18523,NA18852,NA18853,NA18855,NA18858,NA18861,NA18862,NA18864,NA18867,NA18868,NA18870,NA18876,NA18877,NA18879,NA18881,NA18907,NA18909,NA18910,NA18912,NA18913,NA18915,NA18917,NA18923,NA18924,NA18934,NA19092,NA19093,NA19095,NA19096,NA19098,NA19099,NA19101,NA19102,NA19107,NA19108,NA19114,NA19117,NA19118,NA19119,NA19121,NA19122,NA19127,NA19128,NA19130,NA19131,NA19137,NA19138,NA19140,NA19141,NA19143,NA19146,NA19147,NA19150,NA19152,NA19153,NA19159,NA19160,NA19171,NA19184,NA19190,NA19193,NA19195,NA19196,NA19198,NA19200,NA19201,NA19203,NA19206,NA19207,NA19209,NA19210,NA19213,NA19214,NA19223,NA19225,NA19236,NA19238,NA19239,NA19247,NA19257
Authentication	Identity of all cell lines was verified from molecular profiling sequencing data (eg assessment of SNP genotypes in H3K36me# CUT&Tag or naRNA-seq data)
Mycoplasma contamination	cell lines were not tested for mycoplasma
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a