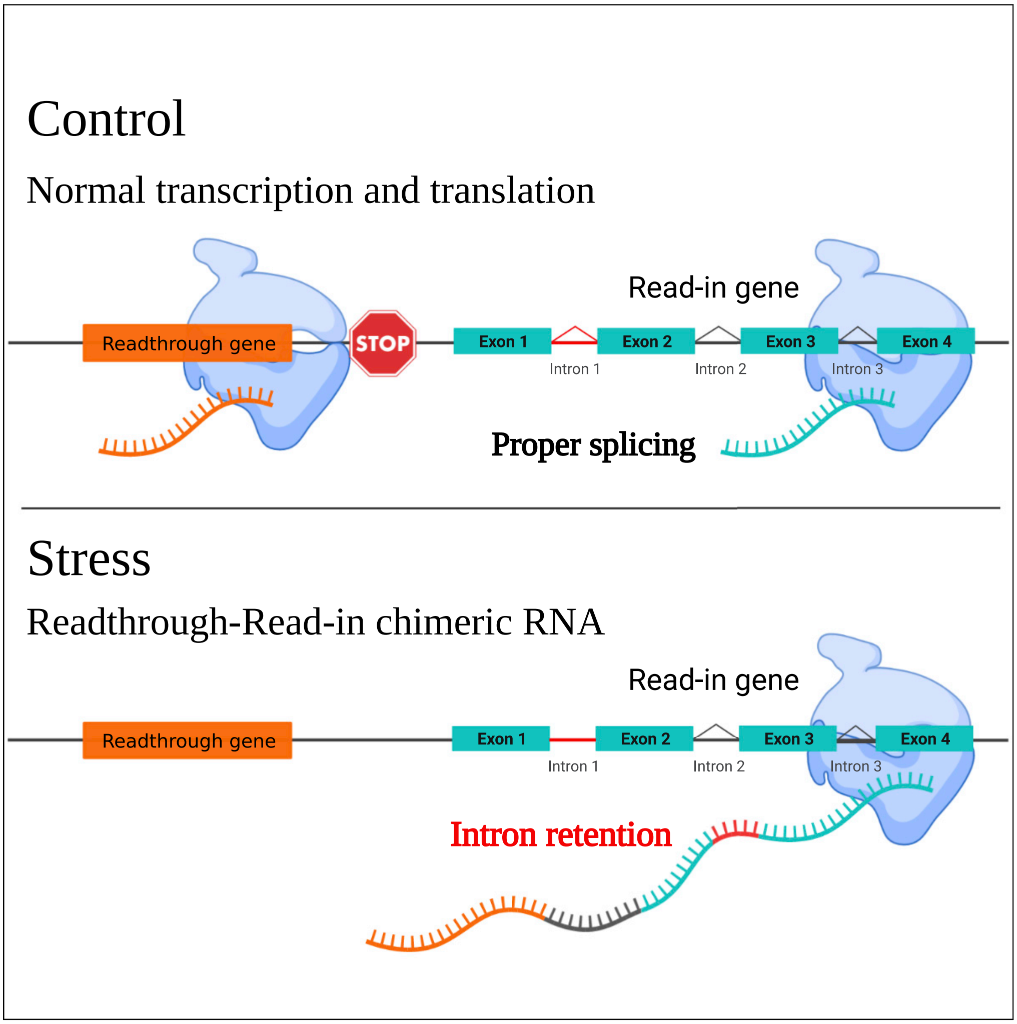
iScience

CellPress
OPEN ACCESS

Article

Stress-induced transcriptional readthrough into neighboring genes is linked to intron retention



Shani Hadar,
Anatoly Meller,
Naseeb Saida,
Reut Shalgi

reutshalgi@technion.ac.il

Highlights

Stress-induced readthrough transcription can extend into downstream genes ("read-in")

Read-in genes have short, GC-rich introns

Read-in genes show marked intron retention during stress

Demarcation of exon-intron junctions by H3K36me3 is absent in read-in genes first introns

Hadar et al., iScience *25*, 105543
December 22, 2022 © 2022 The Authors.
https://doi.org/10.1016/j.isci.2022.105543

iScience

CellPress
OPEN ACCESS

Article

Stress-induced transcriptional readthrough into neighboring genes is linked to intron retention

Shani Hadar,[1,2] Anatoly Meller,[1,2] Naseeb Saida,[1] and Reut Shalgi[1,3,*]

SUMMARY

Exposure to certain stresses leads to readthrough transcription. Using polyA-selected RNA-seq in mouse fibroblasts subjected to heat shock, oxidative, or osmotic stress, we found that readthrough transcription can proceed into proximal downstream genes, in a phenomenon previously termed "read-in." We found that read-in genes share distinctive genomic characteristics; they are GC-rich and extremely short , with genomic features conserved in human. Using ribosome profiling, we found that read-in genes show significantly reduced translation. Strikingly, read-in genes demonstrate marked intron retention, mostly in their first introns, which could not be explained solely by their short introns and GC-richness, features often associated with intron retention. Finally, we revealed H3K36me3 enrichment upstream to read-in genes. Moreover, demarcation of exon-intron junctions by H3K36me3 was absent in read-in first introns. Our data portray a relationship between read-in and intron retention, suggesting they may have co-evolved to facilitate reduced translation of read-in genes during stress.

INTRODUCTION

Gene expression is known to be highly regulated in response to stress.[1] In the past decade, it is becoming increasingly clear that post-transcriptional RNA processing is also tightly regulated in response to stress,[2] including splicing,[3–5] and polyadenylation.[6] In particular, intron retention is dynamically regulated in various environmental and physiological conditions.[5,7,8] In heat shock, for example, widespread intron retention in more than a thousand genes was shown to occur, leading to the accumulation of polyadenylated, stable, intron-containing mRNAs in the nucleus.[5] Indeed, one of the prevalent fates of intron-containing mRNAs is the prevention of nuclear export, leading to nuclear retention.[9]

More recently, it was found that several stress conditions, including heat shock, osmotic, oxidative stress,[10,11] and hypoxia,[12] lead to pervasive transcriptional readthrough, resulting in long continuous transcripts that can extend up to thousands of kilobases downstream to gene ends and affecting thousands of genes in human and mouse cells.[10,11] This phenomenon, which also happens during viral infection[13,14] and in renal carcinoma,[15] is thought to occur due to reduced efficiency of polyadenylation.[16] Nevertheless, although it was shown to be regulated, rather than the result of a random failure,[11] and even though several characteristics of readthrough genes have been identified,[11] the underlying selectivity of stress-induced transcriptional readthrough still remains elusive. Key genomic characteristics of readthrough-affected genes include depletion of polyadenylation motifs downstream to gene ends,[10,11,13] open chromatin marks past gene ends,[11,17] and close proximity to neighboring genes.[11]

Although several underlying pathways have been found to contribute to transcriptional readthrough during viral infection,[18] and in osmotic stress,[19] the consequences of this fairly newly identified phenomenon are still largely a mystery. Antisense regulation by readthrough transcripts has been proposed as one potential consequence[11] and was demonstrated to occur during senescence.[20] Interestingly, using nascent RNA-seq performed during HSV-1 viral infection, it was observed that readthrough transcription can extend into neighboring genes, which were therefore termed "read-in" genes.[13] Read-in was also shown to occur during influenza virus infection using rRNA-depleted RNA-seq data.[21] Nevertheless, whether

[1]Department of Biochemistry, Rappaport Faculty of Medicine, Technion–Israel Institute of Technology, Haifa 31096, Israel

[2]These authors contributed equally

[3]Lead contact

*Correspondence: reutshalgi@technion.ac.il

https://doi.org/10.1016/j.isci.2022.105543

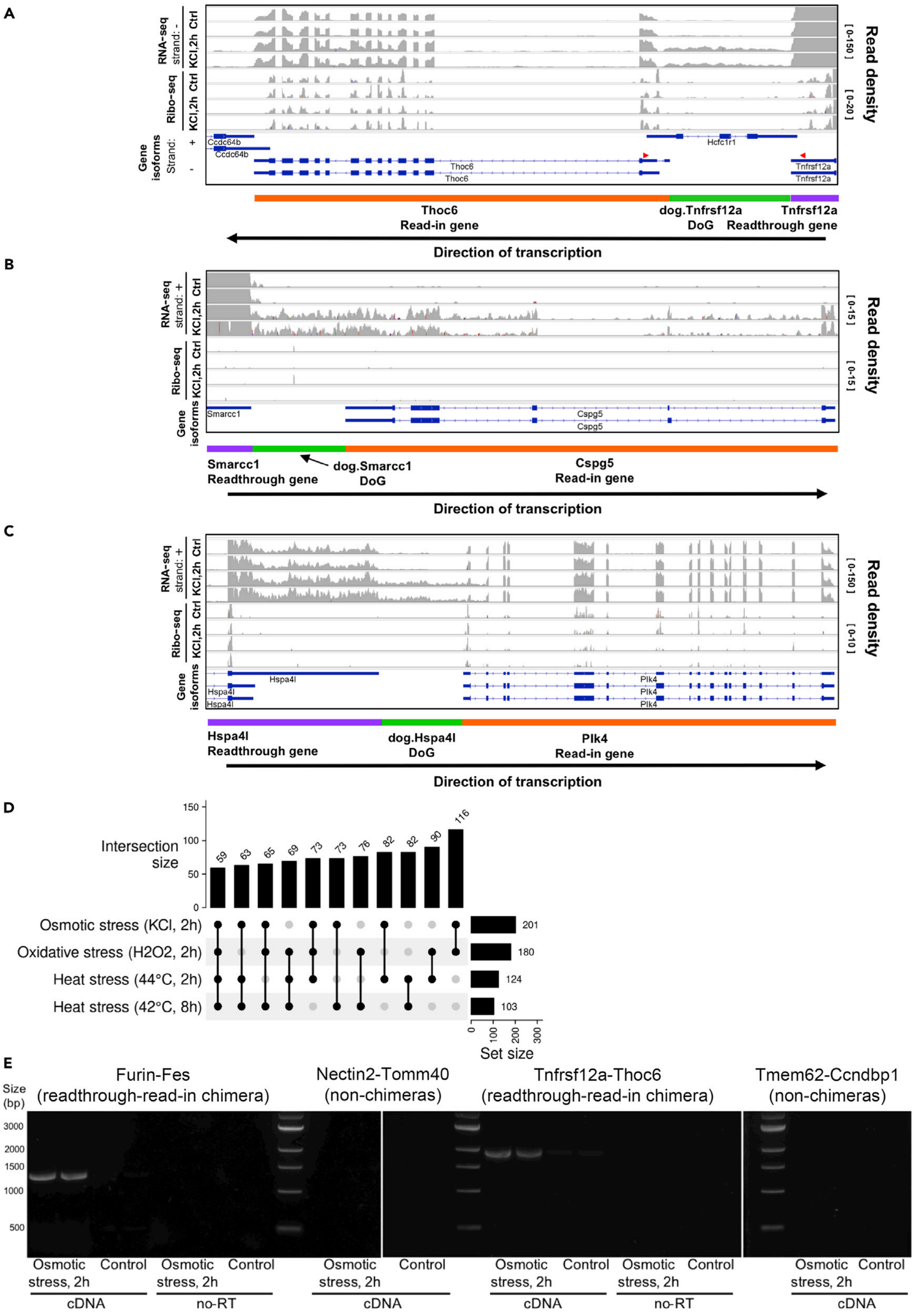

A
RNA-seq strand: -
Ctrl
KCl,2h
Ribo-seq
Ctrl
KCl,2h
Gene isoforms
Strand: +
-
Ccdc64b
Ccdc64b
Hcfc1r1
Thoc6
Thoc6
Tnfrsf12a
Tnfrsf12a
[0-150]
[0-20]
Read density
Thoc6
Read-in gene
dog.Tnfrsf12a
DoG
Tnfrsf12a
Readthrough gene
Direction of transcription
B
RNA-seq strand: +
Ctrl
KCl,2h
Ribo-seq
Ctrl
KCl,2h
Gene isoforms
Smarcc1
Cspg5
Cspg5
[0-15]
[0-15]
Read density
Smarcc1
Readthrough gene
dog.Smarcc1
DoG
Cspg5
Read-in gene
Direction of transcription
C
RNA-seq strand: +
Ctrl
KCl,2h
Ribo-seq
Ctrl
KCl,2h
Gene isoforms
Hspa4l
Hspa4l
Hspa4l
Plk4
Plk4
Plk4
[0-150]
[0-10]
Read density
Hspa4l
Readthrough gene
dog.Hspa4l
DoG
Plk4
Read-in gene
Direction of transcription
D
Intersection size
150
100
50
0
59
63
65
69
73
73
76
82
82
90
116
Osmotic stress (KCl, 2h)
Oxidative stress (H2O2, 2h)
Heat stress (44°C, 2h)
Heat stress (42°C, 8h)
201
180
124
103
0
100
200
300
Set size
E
Size (bp)
3000
2000
1500
1000
500
Furin-Fes
(readthrough-read-in chimera)
Nectin2-Tomm40
(non-chimeras)
Tnfrsf12a-Thoc6
(readthrough-read-in chimera)
Tmem62-Ccndbp1
(non-chimeras)
Osmotic stress, 2h
Control
Osmotic stress, 2h
Control
cDNA
no-RT
Osmotic stress, 2h
Control
cDNA
Osmotic stress, 2h
Control
Osmotic stress, 2h
Control
cDNA
no-RT
Osmotic stress, 2h
Control
cDNA

Figure 1. Stress leads to readthrough into downstream neighboring genes

(A–C) Read density plots (gray) for three examples of read-in genes, shown using IGV plots (Integrative Genomic Viewer 2.12.2[50]), in control and osmotic stress (KCl, 2 h) for expression (RNA-seq) and translation (Ribo-seq). Data are strand specific; strand is indicated. On the bottom, gene annotation tracks (in blue) are shown with the gene names. Regions of interest are highlighted with colors: the DoG region is highlighted in green, the readthrough gene in purple, and the read-in gene in orange. RT-PCR primer locations used in (E) are indicated in red for Tnfrsf12a-Thoc6 (A).

(D) Read-in gene sets are common between stresses. UpSet plot,[23] visualizing the sizes of set intersections, of the overlap of the four stresses with the largest number of identified read-in genes, shows significant overlaps between them, with 59 read-in genes shared by all four ($p = 3.33e-268$, exact test of multiset intersection[24]). See Figure S2E for overlap with additional conditions.

(E) RT-PCR gels performed for the predicted osmotic stress readthrough-read-in chimeras Furin-Fes and Tnfrsf12a-Thoc6 and the non-chimeras Nectin2-Tomm40 and Tmem62-Ccndbp1. The four pairs were PCR amplified from cDNA of osmotic stress (2h) or control cells, in replicates, using primers spanning the intergenic region. Primers locations for Tnfrsf12a-Thoc6 are indicated in red in panel (A) and for the other pairs in Figures S3A–S3C (see STAR Methods and Table S5). Positive bands are shown for both predicted readthrough-read-in chimeras in the stressed cells, whereas both negative non-chimeras did not show any amplification, and no bands were observed in the no-RT controls (RNA samples without reverse transcriptase) validating the lack of genomic DNA contamination. Amplicons expected sizes: 1864 and 1344 for Tnfrsf12a-Thoc6 and Furin-Fes, respectively, and 1598 and 986 for the Nectin2-Tomm40 and Tmem62-Ccndbp1, respectively; ladder sizes are indicated on the left.

these chimeric read-in transcripts can be detected as mature mRNAs, and what their potential fate is, has remained unknown.

Here we show that stress induces readthrough into neighboring read-in genes, which are evident by the presence of mature, polyadenylated, mRNA transcripts in RNA-seq data, and can potentially generate chimeric readthrough-read-in transcripts. Read-in genes tend to reside close to their upstream readthrough genes, as expected. However, read-in is not simply a function of distance from the readthrough gene, as there are genes proximally downstream to readthrough genes that do not show read-in. We found that read-in genes have distinct genomic characteristics; they are significantly short, they have fewer introns and their introns are shorter, and they tend to be GC-rich. In addition, genes with a high degree of read-in transcription are largely translationally inhibited. Importantly, we found that read-in genes display marked intron retention, especially in their first introns. Our analyses showed that although introns of read-in genes share characteristics associated with intron retention, namely they are short and GC-rich, the high degree of intron retention observed could not be simply explained by these properties. We further identified that regions upstream of read-in genes are highly enriched with H3K36me3 chromatin mark, which is typical of actively transcribed gene bodies. Moreover, demarcation of exon-intron junctions by a sharp decrease in H3K36me3 was largely absent from first introns of read-in genes. Finally, we showed that read-in gene properties are conserved in human. As intron retention is known to be associated with nuclear retention, we speculate that these properties were evolutionarily retained, and therefore facilitate nuclear retention of read-in transcripts, thereby preventing potential aberrant translation resulting from readthrough of chimeric read-in genes during stress conditions.

RESULTS

Stress induces transcriptional readthrough into neighboring genes

To compare the transcriptional and translational responses to stress, we analyzed polyA-selected mRNA-seq, as well as ribosome footprint profiling (Ribo-seq), performed in mouse NIH3T3 fibroblasts subjected to heat shock (42–44°), oxidative stress (H2O2), or osmotic stress (KCl), for acute (2 h) or sustained (7–8 h) treatments. Comparison of the fold changes at the level of the mRNA versus the level of translation revealed a population of mRNAs with marked induction of expression level, with no change at the level of translation, which were apparent mainly at the acute responses to all three stress conditions, as well as at the sustained response to heat shock (Figure S1). As we previously characterized widespread transcriptional readthrough in these conditions,[11] we hypothesized that these may represent readthrough genes (also termed DoGs, downstream of genes containing transcripts). However, mapping of DoGs using the DoGFinder tool[12] showed a very small overlap between DoG-generating genes and the aforementioned population. Instead, manual examination of individual genes from these populations showed that some of these genes were located downstream of readthrough regions, where the DoGs seem to continue into the downstream gene (Figures 1A–1C, S2A–S2D, and S3A). Such a phenomenon has been documented before using analysis of nascent RNA following HSV-1 infection,[13] a condition known to induce transcriptional readthrough, and was termed "read-in."[13] However, our data included sequencing of mature, polyA-selected mRNAs. We therefore sought to characterize the read-in phenomenon more globally in this dataset. To that end, we set to systematically identify read-in genes, which we defined as genes with substantial RNA-seq read coverage at the "read-in region," i.e., the region 1 kb upstream to their most upstream isoform transcription start site (TSS, see STAR Methods), and which were additionally

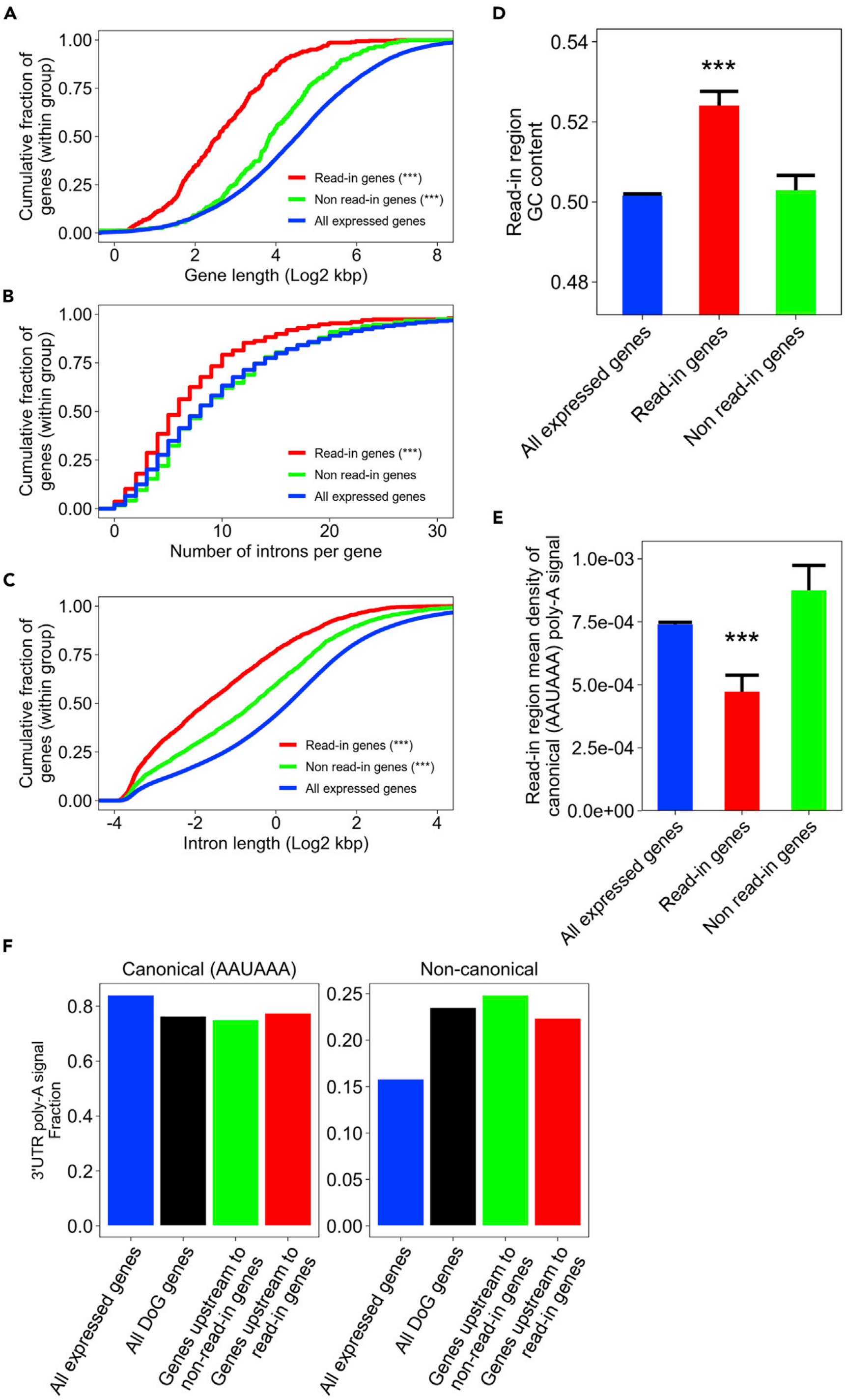


Figure 2. Read-in genes are significantly short, with fewer shorter introns, and GC-rich

(A–C) Distribution of feature lengths (log2 kbp) of read-in (red), non-read-in (green), and all expressed genes (blue) presented as cumulative distribution function (CDF) plots. p values were calculated using Wilcoxon rank-sum test, between either read-in or non-read-in distributions versus all expressed genes; ***p < 0.001, p indicated when significant (p < 0.05). Shown are the length of the entire gene (A): p (read-in versus expressed) = 3.23e-73, p (non-read-in versus expressed) = 2.33e-8, p (read-in versus non-read-in) = 1.47e-23, number of introns per gene (B), p (read-in versus expressed) = 2.16e-9, p (non-read-in versus expressed) = 4.2e-1, p (read-in versus non-read-in) = 1.37e-6, and introns lengths distributions (C) p (read-in versus expressed) = 1.87e-294, p (non-read-in versus expressed) = 3.37e-73, p (read-in versus non-read-in) = 1.32e-49.

Figure 2. *Continued*

(D) GC content was significantly higher in the 1kb upstream regions of read-in genes; p values were calculated using Wilcoxon rank-sum test; ***p < 0.001, p (read-in versus expressed) = 2.49e-10, p (non-read-in versus expressed) = 7e-1, p (read-in versus non read-in) = 1.87e-5.

(E) PolyA signals were significantly depleted in read-in regions (1 kb upstream regions), compared to the corresponding regions upstream to all expressed genes. p (read-in versus expressed) = 1.05e-9, p (non-read-in versus expressed) = 4.51e-1, p (read-in versus non-read-in) = 8.77e-7. Additional variants of the polyA signal showed similar trends, Figure S5J. p values were calculated using Wilcoxon rank-sum test; ***p < 0.001.

(F) The frequencies of the canonical polyA signal AAUAAA (left panel) versus non-canonical polyA signals (right panel), within the 3′ UTRs of genes in each group, showed higher frequencies of non-canonical polyA signals in 3′ UTRs of DoG-producing genes, as well as non-read-in and read-in genes (p < 0.05, chi-square test) compared with all expressed genes (see Table S2 for all p values). (D–F) Data are presented as mean +/− SEM of regions of interest of all genes within respective groups.

overlapping a readthrough event (a DoG) on the same strand (see STAR Methods). This process identified overall 307 read-in genes, which were most abundant in the acute responses to all three stresses, and in the sustained heat shock (Table S1 and Figure 1D). Notably, this number is much higher than the number of read-in genes expected by chance given our selection criteria (307 versus 11, see STAR Methods for additional details). The sets of read-in genes showed significant overlap between the four conditions (Figure 1D, 59 genes, p = 3.33e-268, see also Figure S2E), with a higher overlap between the acute stress conditions (Figure 1D, 73 genes, p = 6.65e-202). Overall, 63 read-in genes were identified in control conditions, of which only one was exclusive to non-stressed conditions (Figure S2E).

Examination of RNA-seq gene coverage plots (Figures 1A–1C, S2A–S2D, and S3A) suggested that readthrough-read-in gene pairs generate continuous transcripts, forming chimeric RNAs. To validate this possibility, we chose two such readthrough-read-in gene pairs, Tnfrsf12a-Thoc6 (Figure 1A) and Furin-Fes (Figure S3A), and performed RT-PCR in untreated (control) or osmotic stress (2 h)-treated cells, using primers spanning from the end of the readthrough gene to the beginning of the read-in gene (red arrows, Figures 1A and S3A, Table S5, see STAR Methods). As control non-chimeras, we selected two gene pairs, Tmem63-Ccndbp1 and Nectin2-Tomm40 (Figures S3B and S3C respectively), with a similar, or shorter, intergenic distance between them, and that did not show any readthrough in our data albeit robust expression in osmotic stress. In addition, no-RT controls (parallel RNA samples without Reverse Transcriptase) were used to eliminate the possibility of DNA contamination. Positive bands were observed for both predicted readthrough-read-in gene pairs in the stressed cells, substantiating the presence of chimeric RNAs, whereas both negative non-chimeras did not show any amplification (Figure 1E). The absence of genomic DNA contamination was validated using PCR of RNA samples without reverse transcriptase (no-RT, Figure 1E). Thus, RT-PCR indeed established the presence of specific chimeric RNA transcripts in osmotic stress (Figure 1E).

To get a sense of whether read-in transcripts tended to terminate early, or whether they continued through the entire gene, we examined the distribution of read density ratios between the first and the last exons in read-in genes and compared it with that of all expressed genes. We found the distributions were highly similar or slightly shifted compared with the distributions of all expressed genes (Figure S3D), indicating that for the majority of read-in transcripts, transcription mostly continued to the end of the read-in gene. Overall, these results suggest that most read-in genes were detectable as chimeric transcripts with their upstream DoG and readthrough gene.

Read-in genes are exceptionally short and GC-rich

Next, we sought to examine whether read-in genes demonstrated specific genomic characteristics. Examination of the distribution of distances to their upstream gene ends (on the same strand) showed that, as expected, read-in genes tended to be much closer to their upstream neighboring genes. Although shifted, the distributions of distances to the same-strand proximal gene for read-in and for DoGs without read-in overlapped (Figure S4). This suggested that read-in is not merely a consequence of proximity to a readthrough gene. We therefore wanted to identify additional properties that characterized read-in genes. In order to control for potential confounding factors, we decided to compare the group of read-in genes with a control group of genes that were downstream of DoGs, and had a similar constraint on the distances to their upstream readthrough gene ends, but were not classified as read-in genes in any of the conditions we tested. We termed this control group ‘‘non read-in genes’’ (see STAR Methods). Our first analysis showed that read-in genes tended to be significantly short (Figure 2A), with a median length of about

6 kb, which is less than one-third of the median length of expressed genes (22.9 kb). We note that this result was not due to the specific parameters used in the read-in identification process (see STAR Methods). Non-read-in genes showed an intermediate length, with a median of 14.4 kb. Next, we asked in which gene regions this difference manifested, and, although exon lengths of read-in genes did not differ from the background of expressed genes (Figure S5B), all other gene regions examined were significantly shorter (Figures 2C, S5A, S5C, and S5D). Most significantly, read-in genes had fewer introns (Figure 2B), which were also much shorter (Figure 2C), whereas non-read-in genes had a similar number of introns as the entire population of expressed genes (Figure 2B), with a median length of 0.7 kb, which was much longer than that of read-in genes (0.3 kbs) but shorter than the median of the entire population (1.2 kb, Figure 2C).

Next, we focused on the read-in region, i.e., the 1 kb region upstream to read-in genes, and found that this region was significantly GC-rich compared with both the background of all expressed genes as well as non-read-in genes (Figure 2D). Read-in genes also had GC-rich introns (Figure S5E) and were overall significantly GC-richer than expressed genes (Figure S5F). DoGs upstream to read-in genes also tended to be GC-rich compared with other DoGs, as well as DoGs upstream of non-read-in genes (Figure S5G), and their corresponding upstream genes were GC-richer (Figures S5H and S5I). Thus, it seems that the genomic environment of read-in genes tends to be significantly GC-rich.

Previous studies have shown that DoG regions, i.e., readthrough regions downstream of the canonical polyA site of the readthrough gene, tend to be depleted with polyA signals compared to intergenic regions right downstream of non-readthrough genes polyA sites.[10,11,13] One interpretation of this result was that multiple polyA signals downstream to gene ends could assist with promoting efficient termination in times when the termination process is partly impaired. We therefore analyzed the polyA signal density in the read-in regions and found that, here too, polyA signals were significantly depleted compared with corresponding background regions of all expressed genes, as well as those of non-read-in genes (Figures 2E and S5J). We then turned to examine the 3′UTRs of DoG-producing genes upstream of read-in genes. Interestingly, a higher proportion of them contained non-canonical polyA signals, as opposed to the canonical one, AAUAAA, in comparison to the proportion in all expressed genes 3′ UTRs (Figure 2F, chi-square p = 3.39e-5, see STAR Methods). However, this tendency was not specific to 3′ UTRs upstream of read-in genes but was rather apparent for all DoG-producing genes (Figure 2F, chi-square p = 2.99e-17, compared with all expressed genes). Therefore, higher prevalence of non-canonical polyA signals, which are considered weaker,[25,26] in 3′ UTRs may promote readthrough in stress conditions, and the scarcity of extra polyA signals downstream to gene ends may further contribute to read-in into the downstream gene.

Genes with a high degree of read-in are translationally inhibited in stress conditions

Having observed several examples of read-in genes that showed low levels of translation in stress (Figure 1C), or lack of translation altogether albeit having robust expression in stress conditions (Figures 1B, S2A–S2C, and S3A), we asked if this trend could be generalized for the entire group of read-in genes. Our analyses showed that, following acute osmotic stress, read-in genes showed a significant induction at the mRNA level; however, their translation was unchanged (Figure S6A). In acute heat shock and oxidative stress, read-in genes were induced at the level of expression, whereas at the level of translation, a significant but milder effect, on average, was observed (Figures S6B and S6C). Next, we asked whether read-in is associated with a translational shutoff during stress. A comparison between mRNA expression and translation levels showed that only a subset of the read-in genes was translationally inhibited (Figures 3A and S7). Manual examination of several examples indicated that, in some cases, the expression of the upstream DoG seemed to be similar to that of the read-in gene (as in Figures 1B, S2A, and S2B), whereas in other cases, the expression of the read-in gene was clearly higher than that of its upstream DoG (as in Figures 1A and S2D). We therefore sought to quantify this effect for all read-in genes, reasoning that read-in genes mRNAs with similar levels of expression as their upstream DoGs are probably mostly expressed due to readthrough transcription, whereas those with a higher expression than their upstream DoG represent a mix of readthrough transcription and independent expression. We therefore calculated a metric termed ''read-in estimation,'' which is simply the read density in the read-in region (1 kb upstream to the TSS of the read-in gene) divided by the read density over the entire read-in gene mRNA (see STAR Methods). Stratifying read-in genes into three equal-sized groups according to their read-in estimation values showed that the subset of genes with high read-in estimation values, meaning that most of their expression could be attributed to readthrough transcription, had the lowest levels of translation in all conditions (Figures 3A, 3B, S7, and S8). Furthermore, the higher the read-in

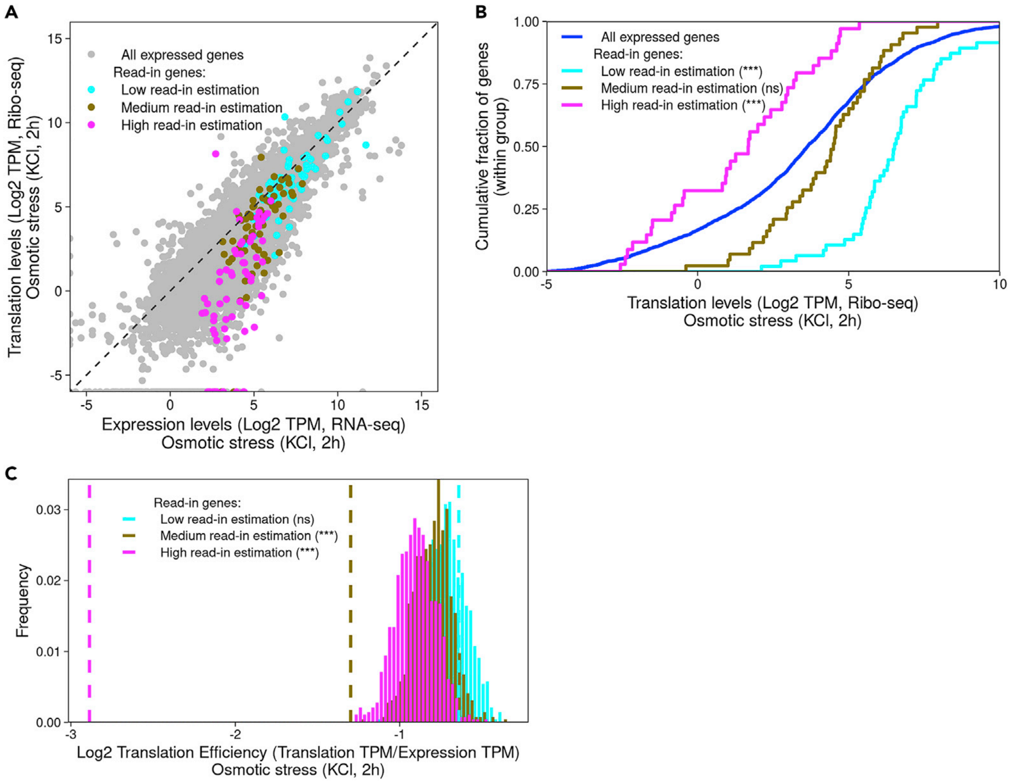


Figure 3. Read-in genes tend to be lowly translated

(A) Expression versus translation levels (Log2 TPM values of RNA-seq on the x axis and Ribo-seq on the y axis) of genes during osmotic stress (KCl, 2h). Read-in genes were stratified by their read-in estimation values (cyan, brown, and magenta) corresponding to 33% quantiles of read-in estimation values (see STAR Methods, Table S3). All expressed genes are shown in gray. High read-in estimation read-in genes tend to have lower levels of translation given their expression levels. Similar trends were also found in other conditions (Figure S7).

(B) CDF plot of the translation levels (Ribo-seq TPM, in log2) of genes with different read-in estimation groups in osmotic stress (2 h) showed an inverse correlation between the degree of read-in and the level of translation. p values were calculated using Wilcoxon rank-sum test, ***p < 0.001, **p < 0.01, (see Table S2 for exact p values). Similar trends were also found in other conditions (Figure S8).

(C) Random sampling analysis of translation efficiency (translation normalized to expression), where randomly sampled groups were matched for levels of expression with their corresponding read-in estimation group (high, medium, and low, color-coded, see STAR Methods) in osmotic stress (2 h). Dashed lines represent the mean translation efficiency value for each read-in group. The analysis demonstrated a significantly lower translation efficiency for high and medium read-in estimation genes (***p < 0.001 for both, see Table S2 for exact p values) compared to the expected translation levels given their expression levels (solid colored distributions). Similar trends were also found in other conditions (Figure S9).

estimation was, the lower the translation levels were in all conditions (Figures 3A, 3B, S7, and S8). To understand whether these translation levels could merely reflect the expected translation given the levels of the mRNA, we calculated the ratios between translation and expression levels (often termed ''Translation Efficiency'' or ''TE'') and compared the mean TE values of high, medium, or low read-in estimation genes with the mean TE values of randomly sampled groups of mRNAs with expression levels that were matched to those within each read-in gene group (see STAR Methods). We found that, although for read-in genes with low read-in estimation values translation levels were no different than expected given their corresponding mRNA levels, read-in genes with high read-in estimation values were significantly less translated than expected given their expression levels in all conditions (Figures 3C and S9). Genes with intermediate levels of read-in estimation showed significantly lower translation-to-expression ratios than expected in the three acute stress conditions and in sustained heat shock (Figures 3C and S9). Therefore, read-in genes with high read-in estimation, i.e., for which expression is mainly due to readthrough, are indeed translationally repressed.

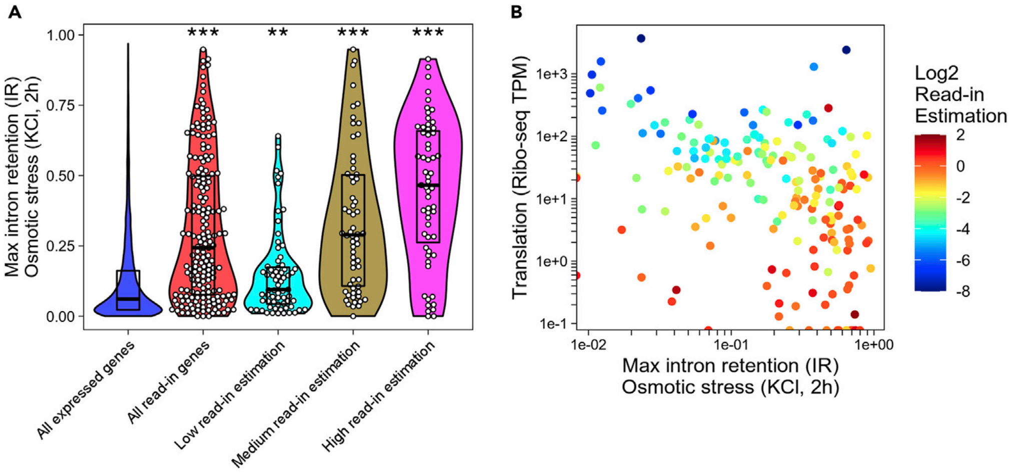


Figure 4. Read-in genes show marked intron retention

(A) Violin plots demonstrate significantly higher degrees of intron retention for all groups compared with all expressed genes, which increase with the extent of read-in estimation. Each gene is represented by the maximal value of intron retention among all its introns in osmotic stress (2 h). Boxes indicate median, 25th and 75th percentiles for each of the groups. Wilcoxon rank-sum test p value calculated for each group versus all expressed genes, ***$p < 0.001$, *$p < 0.05$, see Table S2 for exact p values. Similar trends were also observed in other conditions (Figure S11).

(B) Scatterplot of translation levels (log2 Ribo-seq TPM, y axis) versus maximal intron retention (log2 IR value, x axis) for read-in genes in osmotic stress (2 h) exhibit a significant negative correlation ($R = -0.407$, $p(R) = 2.4e\text{-}08$). The color axis shows the read-in estimation values (in log2), further demonstrating that the more the intron is retained the higher the read-in estimation tends to be ($R = 0.475$, $p(R) = 3.55e\text{-}11$). Similar trends were also observed in other conditions (Figure S14). The relationship between translation levels and intron retention is shown in Figure S13.

Read-in genes show marked intron retention

Looking at several individual examples of read-in genes, we noticed that they showed high levels of intron retention in their first intron (Figures 1B, 1C, and S2A–S2D) and occasionally for other introns (Figures 1A, 1B, and S2B). We therefore asked how prevalent intron retention is in read-in genes, in a systematic manner. We first calculated the IR value (intron retention) for each intron in the genome, as a metric to evaluate the extent of intron retention (see STAR Methods). Indeed, we found that, overall, read-in genes tended to show much higher levels of intron retention compared with other genes (Figures 4A and S11), in all conditions. Furthermore, intron retention was not restricted to first introns, although the highest levels of intron retention were found in first introns (Figure S12). Interestingly, the higher the read-in estimation levels were, the higher the degree of intron retention was observed (Figures 4A and S11). Because intron retention is known to lead, in many cases, to nuclear retention,[5,9] this could explain why genes with high read-in estimation levels are lowly translated. Indeed, translation levels showed a negative correlation with intron retention, also for the background population of all expressed genes, although intron retention was always higher within read-in genes (Figure S13). Examination of all three factors together—read-in estimation, intron retention, and translation levels—showed that, indeed, the higher the degree of intron retention was, the lower the translation levels were, and the higher the read-in estimation was for read-in genes (Figures 4B and S14).

Intron retention in read-in genes is much higher than expected given their intronic features

We found that read-in genes have particularly short introns, which were significantly GC-rich (Figures 2C and S5E). It was previously shown that these two properties, i.e., short intronic length and high GC content, are associated with an increased tendency for intron retention.[9] We analyzed the entire set of introns and indeed observed that these tendencies are largely recapitulated in our data (Figure S15). We therefore asked whether these properties alone could explain the marked degree of intron retention that we observed in read-in genes. To answer this, we performed a random sampling test, where sets of introns with the same length and GC content distributions as those of read-in genes were randomly sampled from the set of all expressed genes, and the mean degree of intron retention was calculated for each random set (see STAR Methods). This would allow us to specifically control for the two intronic properties known to be associated with intron retention. Surprisingly, this analysis showed that these two features

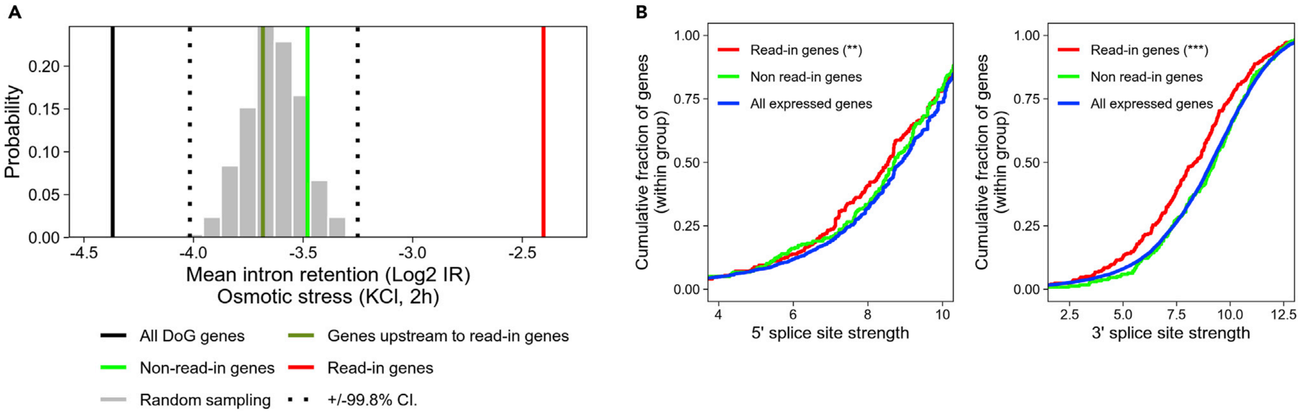


Figure 5. Read-in genes intron retention levels are much higher than expected given their genomic characteristics

(A) A thousand randomly sampled comparison groups, matched for GC content, and intron lengths distributions as in osmotic stress (2 h) read-in genes, were generated from the set of all expressed genes (see STAR Methods). For each randomly sampled comparison group, the mean value of the maximum intron retention (log2 IR) per gene was calculated and plotted as a histogram (gray). Confidence intervals (+/−99.8%, corresponding to 3*STD) are presented as dashed black lines. The mean value of the maximum intron retention (log2 IR) of read-in genes (red) is significantly higher than the distribution of the mean intron retention values, even when controlling for GC content and intron lengths (***p < 0.001); however, those of other groups were either no different than the background (non-read-in genes in green and genes upstream to read-in genes in olive) or even lower than the read-in genes-matched controls (all DoG-producing genes in black). Similar trends were also observed in other conditions (Figure S16).

(B) CDF plots of 5′ (left) and 3′ (right) splice site strengths, calculated as MaxEnt scores[27] (see STAR Methods), showed significantly weaker splice site strengths distribution for first introns in read-in genes (red) compared with all expressed genes (blue), whereas non-read-in genes splice site strengths (green) were similar to those of all expressed genes. p values were calculated using Wilcoxon rank-sum test; ***p < 0.001, **p < 0.01. For 5′ splice sites: p (read-in versus all expressed) = 4.78e-3, p (non-read-in versus all expressed) = 0.12, for 3′ splice sites: p (read-in versus all expressed) = 7.8e-7, p (non-read-in versus all expressed) = 0.77.

could not explain the high degree of intron retention observed in read-in genes, which was significantly higher than expected in all conditions (Figures 5A and S16), and in particular in the acute stress conditions (with z-scores of 4.04, 7.4, and 9.65, number of STDs above the mean of the expected population, for acute heat shock, oxidative, and osmotic stresses, respectively, Figure 5A). Moreover, these differences were specific to read-in genes, as non-read-in genes, and the DoG-generating genes upstream to read-in genes were no different from the background (Figures 5A and S16).

We then turned to examine another property that may partly explain the extreme intron retention that we observed in read-in genes, namely splice site strength[27] (see STAR Methods). We quantified splice site strength for all first introns using MaxEnt.[27] Our analysis showed that first introns in read-in genes have significantly weaker splice sites, in both 5′ as well as 3′ splice sites, compared with the background distribution of first introns, whereas first introns in non-read-in genes were no different than the entire population (Figure 5B). However, although retained introns in both read-in genes and expressed genes demonstrated overall weaker 5′ splice site strengths than non-retained introns, this trend was much noisier for the 3′ splice site (Figure S19). Thus, although a higher proportion of read-in genes have weaker 5′ and 3′ splice sites, the impact of this on read-in-associated intron retention still requires further investigation

Regions upstream to read-in genes show higher H3K36me3 chromatin marks

Next, we sought to identify chromatin features that may be specific to read-in genes and read-in regions. To that end, we examined histone modification ChIP-seq data in mouse cell lines (see STAR Methods) available through the ENCODE project database.[28] It was recently shown that proximal genes on the same strand tend to have distinct chromatin features, as the region downstream of one gene end intermixes with the promoter of its downstream neighbor.[29] Because read-in genes tend to be proximal to their same-strand upstream neighboring genes (Figure S4), we wanted to control for any potential confounder. Thus, we generated an additional control group, comprised of expressed genes that were matched for their intergenic distances (to their upstream genes on the same strand), to the distribution of read-in genes (see STAR Methods). We then tested, for each chromatin mark, whether it tended to be more present within read-in regions (the 1 kb upstream to the TSS of the gene's most upstream isoform) of read-in genes compared with respective regions of non-read-in genes, as well as compared with the respective regions

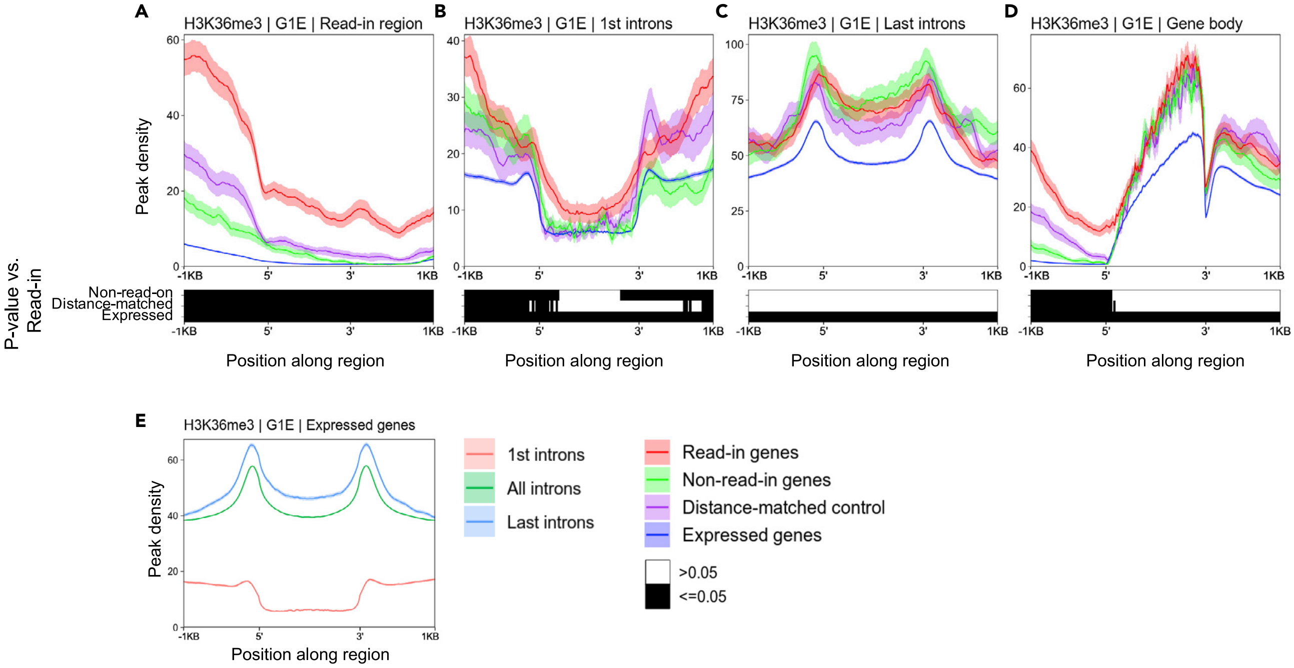


Figure 6. H3K36me3 profiles show enrichment upstream to read-in genes and diminished demarcation of exon-intron junctions in read-in genes first introns

(A–D) Peak density profiles (mean and STE of H3K36me3 peaks across all genes within a group) of H3K36me3 histone modifications in G1E cell line (using ENCODE ChIP-seq data) are shown within (A) read-in regions, (B) first introns, (C) last introns, and (D) entire gene body, demonstrating significant enrichment of H3K36me3 in read-in regions of read-in genes compared with the respective regiong of all other control groups (see STAR Methods). Each region was normalized to the same length and plotted in the center, with the addition of flanking regions of 1 kb on each side. See Figure S20 for additional cell lines showing the same trends. Bottom panels show FDR-corrected Wilcoxon ranksum p values for the differences along the positions between read-in genes profile and each of the other groups.

(E) H3K36me3 peak density profiles of first, last, and all introns of all expressed genes, demonstrating a basin-like shape, with peaks around 5′ and 3′ exon-intron junctions which sharply decrease toward the intron body.

of the intergenic distance-matched controls (see STAR Methods). This comparison resulted in a single significant chromatin mark, which was present in significantly higher proportions in read-in genes read-in regions compared with the corresponding regions of both control groups: H3K36me3 (Table S4), a typical mark of actively transcribed gene bodies.[30] This trend was significant in multiple cell line datasets from ENCODE (see Table S4 for p values). Furthermore, quantitative enrichment was also significant when examining the H3K36me3 profiles (Figures 6A and S20A). The levels of expression of the three groups were similar (Figure S21C), and examination of the H3K36me3 patterns throughout read-in gene bodies showed similar H3K36me3 profiles to that of both control groups (Figures 6D and S20D), ruling out differences in expression levels as confounders of this effect. These results indicated that, indeed, the regions upstream to read-in genes tend to be more prone to active transcription.

H3K36me3 demarcation of 5′ and 3′ exon-intron junctions is absent in read-in genes first introns

Interestingly, H3K36me3 was previously linked to regulation of splicing,[31–33] although the exact mechanism underlying the mutual effects between H3K36me3 and the spliceosome are still largely elusive. It was demonstrated that H3K36me3 demarcates exon bodies, as its profiles were shown to peak at the middle of the exon, in line with nucleosome profiles.[33] This previous literature, as well as our findings above, prompted us to examine H3K36me3 profiles within introns. We were particularly interested in introns of read-in genes, as they showed pronounced intron retention, especially in the first introns (Figures 4, 5, and S12). In general, H3K36me3 profiles along introns tended to show a basin-like shape, with peaks around 5′ and 3′ exon-intron junctions, which sharply decreased toward the intron body (Figures 6E and S21A). Because there is a sharp increase in H3K36me3 marks along gene bodies from 5′ to 3′ (as seen in Figures 6D and S20D), we separately examined the first and the last introns. We observed that, indeed, the last intron H3K36me3 profiles were much higher than those of the first intron, and their shapes were somewhat different (Figures 6E and S21A). Nonetheless, from an intron-centric perspective, both first and last intron profiles showed a drop in the 5′ and 3′ exon-intron borders (Figures 6E and S21A). Thus, a sharp decline in H3K36me3 seems to demarcate the 5′ and 3′ exon-intron junctions.

We next compared the H3K36me3 profiles within the first and last introns of read-in genes with those of both control groups. The H3K36me3 profiles in the last introns of read-in genes were overall similar in their shapes and heights to the control groups in all cell lines (Figures 6C and S20C). Interestingly though, in the first intron, we found that, although both control groups maintained the basin-shape profiles typical to first introns, the shape of the H3K36me3 profile around first introns of read-in genes did not preserve the sharp boundaries at the exon-intron junctions (Figures 6B and S20B). Instead, it showed a steady decline toward the middle of the intron (Figure 6B, see Figure S20B for other cell lines), which was also evident from the slope analysis (Figure S21B). Thus, the sharp decline that generally demarcates 5′ and 3′ exon-intron junctions, was absent in first introns of read-in genes.

Read-in genes features are conserved in human

Finally, we asked whether genomic features of read-in genes that we identified in mouse were conserved in the human genome. First, we looked for conservation of the pairing between the readthrough-read-in gene pairs and found that, although the overall conservation of gene organizations (tandem genes on the same strand) between mouse and human is high (91.87%), the fraction of conserved pairs of readthrough-read-in genes is significantly higher (97.3%, chi-square p = 0.0021). Because read-in genes tend to be close to their upstream genes (Figure S4), we checked the fraction of conserved pairs within proximal pairs in mouse (see STAR Methods) and found that it was 93.8%, still significantly lower than the fraction of conserved readthrough-read-in gene pairs (chi-square p = 0.0346). We then asked whether the distances between readthrough-read-in gene pairs tended to be more conserved. To that end, we calculated the distances of readthrough gene ends to the read-in gene starts (using their most downstream and most upstream isoforms, respectively), and looked at the differences between them, in mouse compared with human. To strictly control for the fact that read-in genes tended to be close to their upstream genes, we sampled from all expressed tandem gene pairs, multiple random groups with a similar distribution of intergenic distances to that of readthrough-read-in gene pairs defined in mouse, and asked if the distances are overall more or less conserved in human. We found that, overall, intergenic distances tended to be longer in human, with a median of 418 bases in the matched controls; however, for readthrough-read-in gene pairs, they were only 207 bases longer (median, Figure 7A, p = 0.0045, see STAR Methods), indicating a constraint on their divergence. We then continued to look at the conservation of intron lengths of read-in genes in human. Intron lengths of the human orthologs of read-in genes were overall significantly shorter than the background (Figure 7B), in agreement with our findings in mouse (Figure 2C). When examining conservation of their lengths, again using the same randomly sampled controls as above, we found that introns were generally longer in human, by 207 bases (on average per gene, median) for the matched controls, whereas for read-in genes, the median difference was only 70 bases (Figure 7C, p = 1e-4, see STAR Methods). This indicated that introns of read-in genes might be under evolutionary pressure to remain short.

Thus, genomic properties of read-in genes, including their organization, distance to their upstream readthrough genes, and intron lengths, are broadly conserved in human.

DISCUSSION

Under stress conditions, reduced polyadenylation and termination efficiencies lead to widespread readthrough transcription.[16] This phenomenon leads to read-in into downstream neighboring genes, as we identified here, which tend to have short introns and high GC content, and present marked intron retention. Importantly, previous studies focusing on transcriptional readthrough,[10,11] some of which indicating that readthrough transcription may continue into downstream genomic loci,[13] have mainly relied on the analysis of pre-mRNA, using either rRNA-depleted nuclear RNA[10,11,21] or nascent RNA-seq.[13] In the current study, we utilized mature, polyA-selected mRNA-seq, which enabled us to establish the existence of read-in transcripts in the cell as mature, stable, RNAs.

Previous studies found that housekeeping genes are generally short.[34] However, read-in genes did not significantly overlap with the set of housekeeping genes (hypergeometric p > 0.999, housekeeping genes taken from[34]), and pathway analysis showed no significant enrichment with any particular functional category. Interestingly, early studies highlighted a correlation between GC-content and gene-dense genomic

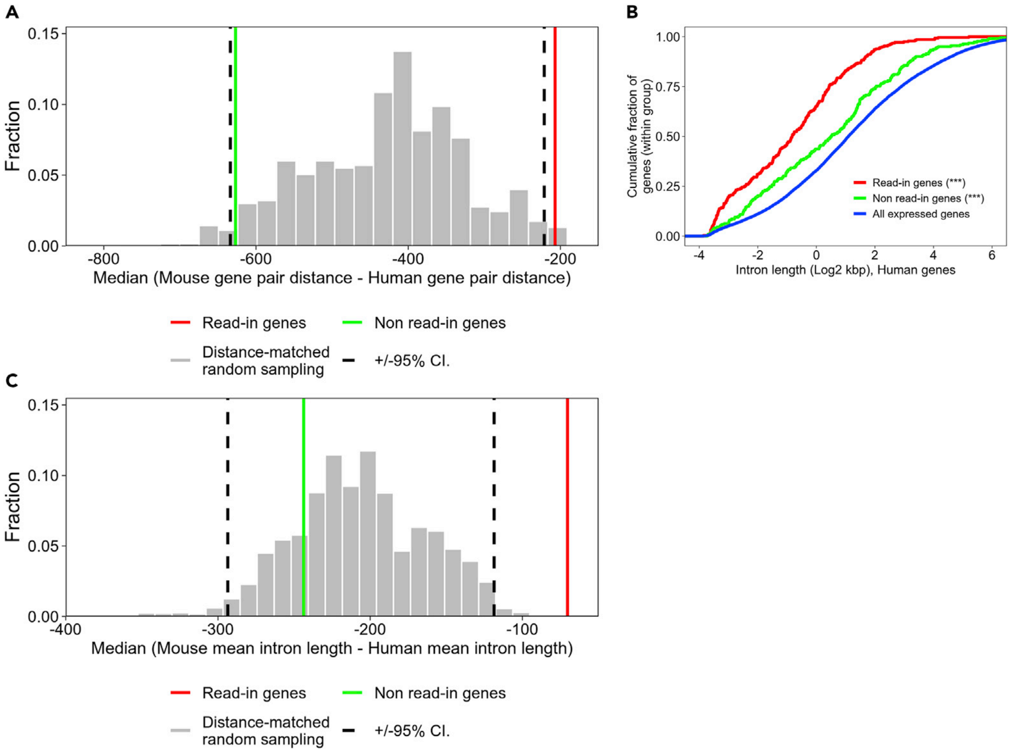


Figure 7. Conservation of read-in genes features in human

(A) Ten thousand randomly sampled comparison groups, matched for intergenic distances to the distribution of readthrough-read-in genes intergenic distances were selected, and the median of the differences between mouse and human intergenic distances was calculated to generate the background distribution (gray). +/−95% confidence intervals (CI) indicated by dashed lines. The median intergenic distance difference between mouse and human is significantly more conserved, i.e. closer to zero, for read-in genes (red line, p = 0.0045), whereas that of non-read-in (green line) is no different from the background.

(B) As in Figure 2C, intron lengths of the human orthologs of the set of read-in genes defined in mouse are significantly shorter than both all expressed genes orthologs (p = 2.52e-40) and non-read-in genes orthologs (p = 3.21e-9). ***p < 0.001, Wilxocon rank-sum test.

(C) Conservation analysis of intron lengths of read-in genes showed that read-in genes intron lengths (mean intron length per gene, red line) were significantly more conserved (the difference between mouse and human is closer to zero) compared with the background distribution (p < 0.0001, using random sampling test, gray shows the background districution as in A, +/−95% confidence intervals [CIs] are indicated by dashed lines), whereas that of non-read-in genes (green line) is no different from the background.

regions.[35] Nevertheless, when we examined non-read-in genes, which have a similar gene proximity distribution to that of read-in genes, we observed that upstream regions of non-read-in genes have the same GC content as that of the expressed genes background (Figure 2D). When looking at the gene itself, the GC content of non-read-in genes was higher than that of the expressed genes background but still significantly lower than the GC content of read-in genes (Figure S5F). Thus, read-in genes seem to be GC-richer than expected even given their tendency to be proximal to their upstream genes.

Our data showed that a higher fraction of 3′UTRs upstream to DoG-producing genes contained more weaker, non-canonical, polyA signals (Figure 2F). In addition, in the read-in regions, read-in genes had a significantly lower frequency of polyA signals, both canonical and non-canonical, compared with non-read-in genes (Figures 2E and S5J). This could contribute to the increased tendency for readthrough when polyadenylation and termination efficiency are compromised, which is further enhanced upstream to read-in genes.

Our data showed that read-in is associated with intron retention (Figures 4 and 5). This may be combined with splicing inhibition under stress, as was shown to occur in heat shock,[5] to increase intron retention even further. Interestingly, Alpert et al. showed that when Nab2, a protein involved in polyA cleavage and polyA tail length, was depleted in yeast, readthrough transcripts invade downstream genes, and these read in transcripts were not able to be spliced, likely due to the wrong intron-exon structure.[36] Thus, it seems that introns are not recognized by the splicing machinery when they occur in fusion transcripts in yeast, whereas in mammalian cells, some, but not all, introns, and in particular first introns, fail to be recognized, as we present here during naturally occurring stress-inducing, readthrough promoting conditions. These further highlight the mechanistic links between read-in and splicing inhibition.

Our current study revealed specific sequence and chromatin features related to read-in. We found increased H3K36me3 active transcription chromatin marks upstream to read-in genes, indicative of these region being predisposed to active transcription, in multiple mouse cell lines (Figure 6A and S20A). Interestingly, mutations in SETD2, a major H3K36 methyltransferase, were previously associated with increased transcriptional readthrough in renal carcinomas.[15] Furthermore, H3K36me3 has also been linked to regulation of splicing in various ways,[33,37–39] and indeed, we uncovered that a sharp decrease in H3K36me3 demarcates the 5′ and 3′ exon-intron junctions in general; however, this sharp decrease is absent particularly in first introns of read-in genes (Figures 6B and S20B). Although the mechanistic relationship of H3K36me3 with the spliceosome is still unclear, tumors with a global defect in H3K36 trimethylation have shown marked increase in intron retention.[40] In light of these findings, our observations raise the intriguing possibility that spliceosome recognition might be compromised in cases where the H3K36me3 differences are absent at the exon-intron boundaries, a hypothesis that remains to be explored.

Finally, it has been shown that the Cap-binding complex has a critical role in splicing of the first intron.[41,42] It is therefore possible that, due to read-in transcripts being continuous with their upstream DoG-producing transcripts, the lack of a cap in proximity to the first read-in gene intron impedes its efficient splicing. This possibility too remains to be explored in future experiments.

The use of paired mRNA-seq and ribosome profiling allowed us to specifically determine that mRNAs with high degrees of read-in show reduced translation (Figure 3). This result raises the question of what would be the underlying mechanism that prevents translation of these read-in transcripts. Intron retention was shown to be conserved in evolution and correlated with proteome complexity.[43] In addition, widespread intron retention can be induced in response to some stress conditions, such as heat shock, leading to the accumulation of stable, polyadenylated, intron-containing mRNAs, which are retained in the nucleus.[5] Intron retention in mRNAs may lead to a variety of fates, most of them resulting in translation inhibition.[9] In some instances, nuclear-residing, retained-intron mRNA (previously termed detained-introns) were shown to be spliced in response to specific stimulations.[7,44] In rare cases, intron-containing mRNAs might be exported from the nucleus, where they are subjected to nonsense-mediated decay (NMD).[8,45] Nonetheless, the most common fate of intron-retained mRNAs is nuclear retention,[9] where they may subsequently be degraded in an NMD-independent manner.[46] Here, using polyA-selected mRNA-seq data representing a steady-state snapshot of polyadenylated, mature, whole-cell mRNAs, it seems that read-in transcripts were not degraded postmaturation, but rather were stable enough to allow their detection as mature, polyadenylated transcripts. Thus, although it cannot be excluded that these transcripts are subjected to NMD, it is likely that they are indeed nuclear-retained, which explains their reduced translation. To what extent stress-induced read-in transcripts are nuclear retained still remains to be determined.

We showed that genomic features of read-in genes are largely conserved between mouse and human. It is therefore possible that read-in genes have co-evolved with genomic properties that favor much higher degrees of intron retention, thereby aiding the nuclear retention of read-in gene transcripts during readthrough-promoting stress conditions. Our findings suggest that read-in genes, as previously shown for readthrough, are not merely a by-product of stress-induced failure in polyadenylation and transcription termination. The properties of read-in genes may facilitate the quality control for read-in transcripts, ultimately allowing stress-induced readthrough transcription to persist, while preventing the nuclear export of read-in transcripts, and precluding their unwarranted translation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105543.

ACKNOWLEDGMENTS

We thank members of the Shalgi lab for fruitful discussions throughout the project. We thank Flonia Levy-Adam for critical reading of the manuscript. Funding: this work was supported by the European Research Council, ERC StG [grant number 677776], and by the ISF grant 694/18.

AUTHOR CONTRIBUTIONS

R.S. conceived and supervised the study. N.S. performed RNA-seq and RT-PCR experiments. S.H. and A.M. performed all computational analyses. R.S. wrote the paper with the help of A.M.

DECLARATION OF INTERESTS

The authors declare no conflict of Interest.

Received: March 16, 2022
Revised: July 10, 2022
Accepted: November 7, 2022
Published: December 22, 2022

REFERENCES

1. López-Maury, L., Marguerat, S., and Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nat. Rev. Genet. *9*, 583–593.

2. Biamonti, G., and Caceres, J.F. (2009). Cellular stress and RNA splicing. Trends Biochem. Sci. *34*, 146–153. S0968-0004(09)00006-1 [pii]. https://doi.org/10.1016/j.tibs.2008.11.004.

3. Nevo, Y., Kamhi, E., Jacob-Hirsch, J., Amariglio, N., Rechavi, G., Sperling, J., and Sperling, R. (2012). Genome-wide activation of latent donor splice sites in stress and disease. Nucleic Acids Res. *40*, 10980–10994. https://doi.org/10.1093/nar/gks834.

4. Sabath, N., Levy-Adam, F., Younis, A., Rozales, K., Meller, A., Hadar, S., Soueid-Baumgarten, S., and Shalgi, R. (2020). Cellular proteostasis decline in human senescence. Proc. Natl. Acad. Sci. USA *117*, 31902–31913. https://doi.org/10.1073/pnas.2018138117.

5. Shalgi, R., Hurt, J.A., Lindquist, S., and Burge, C.B. (2014). Widespread inhibition of posttranscriptional splicing shapes the cellular transcriptome following heat shock. Cell Rep. *7*, 1362–1370. https://doi.org/10.1016/j.celrep.2014.04.044.

6. Di Giammartino, D.C., Shi, Y., and Manley, J.L. (2013). PARP1 represses PAP and inhibits

polyadenylation during heat shock. Mol. Cell *49*, 7–17. https://doi.org/10.1016/j.molcel.2012.11.005.

7. Boutz, P.L., Bhutkar, A., and Sharp, P.A. (2015). Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev. *29*, 63–80. https://doi.org/10.1101/gad.247361.114.

8. Wong, J.J.L., Ritchie, W., Ebner, O.A., Selbach, M., Wong, J.W.H., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., et al. (2013). Orchestrated intron retention regulates normal granulocyte differentiation. Cell *154*, 583–595. https://doi.org/10.1016/j.cell.2013.06.052.

9. Monteuuis, G., Wong, J.J.L., Bailey, C.G., Schmitz, U., and Rasko, J.E.J. (2019). The changing paradigm of intron retention: regulation, ramifications and recipes. Nucleic Acids Res. *47*, 11497–11513. https://doi.org/10.1093/nar/gkz1068.

10. Vilborg, A., Passarelli, M.C., Yario, T.A., Tycowski, K.T., and Steitz, J.A. (2015). Widespread inducible transcription downstream of human genes. Mol. Cell *59*, 449–461. https://doi.org/10.1016/j.molcel.2015.06.016.

11. Vilborg, A., Sabath, N., Wiesel, Y., Nathans, J., Levy-Adam, F., Yario, T.A., Steitz, J.A., and Shalgi, R. (2017). Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. Proc. Natl. Acad. Sci. USA *114*, E8362–E8371. https://doi.org/10.1073/pnas.1711120114.

12. Wiesel, Y., Sabath, N., and Shalgi, R. (2018). DoGFinder: a software for the discovery and quantification of readthrough transcripts from RNA-seq. BMC Genom. *19*, 597. https://doi.org/10.1186/s12864-018-4983-4.

13. Rutkowski, A.J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel, P., Efstathiou, S., Zimmer, R., Friedel, C.C., and Dölken, L. (2015). Widespread disruption of host transcription termination in HSV-1 infection. Nat. Commun. *6*, 7126. https://doi.org/10.1038/ncomms8126.

14. Zhao, N., Sebastiano, V., Moshkina, N., Mena, N., Hultquist, J., Jimenez-Morales, D., Ma, Y., Rialdi, A., Albrecht, R., Fenouil, R., et al. (2018). Influenza virus infection causes global RNAPII termination defects. Nat. Struct. Mol. Biol. *25*, 885–893. https://doi.org/10.1038/s41594-018-0124-7.

15. Grosso, A.R., Leite, A.P., Carvalho, S., Matos, M.R., Martins, F.B., Vítor, A.C., Desterro, J.M.P., Carmo-Fonseca, M., and de Almeida, S.F. (2015). Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. Elife *4*, e09214. https://doi.org/10.7554/eLife.09214.

16. Rosa-Mercado, N.A., and Steitz, J.A. (2022). Who let the DoGs out? - biogenesis of stress-induced readthrough transcripts. Trends Biochem. Sci. *47*, 206–217. https://doi.org/10.1016/j.tibs.2021.08.003.

17. Hennig, T., Michalski, M., Rutkowski, A.J., Djakovic, L., Whisnant, A.W., Friedl, M.S., Jha, B.A., Baptista, M.A.P., L'Hernault, A., Erhard, F., et al. (2018). HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes. PLoS Pathog. *14*, e1006954. https://doi.org/10.1371/journal.ppat.1006954.

18. Wang, X., Hennig, T., Whisnant, A.W., Erhard, F., Prusty, B.K., Friedel, C.C., Forouzmand, E., Hu, W., Erber, L., Chen, Y., et al. (2020). Herpes simplex virus blocks host transcription termination via the bimodal activities of ICP27. Nat. Commun. *11*, 293. https://doi.org/10.1038/s41467-019-14109-x.

19. Rosa-Mercado, N.A., Zimmer, J.T., Apostolidi, M., Rinehart, J., Simon, M.D., and Steitz, J.A. (2021). Hyperosmotic stress alters the RNA polymerase II interactome and induces readthrough transcription despite widespread transcriptional repression. Mol. Cell *81*, 502–513.e4. https://doi.org/10.1016/j.molcel.2020.12.002.

20. Muniz, L., Deb, M.K., Aguirrebengoa, M., Lazorthes, S., Trouche, D., and Nicolas, E. (2017). Control of gene expression in senescence through transcriptional read-through of convergent protein-coding genes. Cell Rep. *21*, 2433–2446. https://doi.org/10.1016/j.celrep.2017.11.006.

21. Roth, S.J., Heinz, S., and Benner, C. (2020). ARTDeco: automatic readthrough transcription detection. BMC Bioinf. *21*, 214. https://doi.org/10.1186/s12859-020-03551-0.

22. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

23. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. IEEE Trans. Vis. Comput. Graph. *20*, 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248.

24. Wang, M., Zhao, Y., and Zhang, B. (2015). Efficient test and visualization of multi-set intersections. Sci. Rep. *5*, 16923. https://doi.org/10.1038/srep16923.

25. Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res. *10*, 1001–1010. https://doi.org/10.1101/gr.10.7.1001.

26. Graber, J.H., Cantor, C.R., Mohr, S.C., and Smith, T.F. (1999). In silico detection of control signals: mRNA 3′-end-processing sequences in diverse species. Proc. Natl. Acad. Sci. USA *96*, 14055–14060. https://doi.org/10.1073/pnas.96.24.14055.

27. Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394. https://doi.org/10.1089/1066527041410418.

28. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

29. Nissani, N., and Ulitsky, I. (2022). Unique features of transcription termination and initiation at closely spaced tandem human genes. Mol. Syst. Biol. *18*, e10682. https://doi.org/10.15252/msb.202110682.

30. Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nat. Rev. Genet. *12*, 7–18. https://doi.org/10.1038/nrg2905.

31. Kim, S., Kim, H., Fong, N., Erickson, B., and Bentley, D.L. (2011). Pre-mRNA splicing is a determinant of histone H3K36 methylation. Proc. Natl. Acad. Sci. USA *108*, 13564–13569. https://doi.org/10.1073/pnas.1109475108.

32. Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science *327*, 996–1000. https://doi.org/10.1126/science.1184208.

33. Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased chromatin signatures around polyadenylation sites and exons. Mol. Cell *36*, 245–254. https://doi.org/10.1016/j.molcel.2009.10.008.

34. Eisenberg, E., and Levanon, E.Y. (2003). Human housekeeping genes are compact. Trends Genet. *19*, 362–365. https://doi.org/10.1016/S0168-9525(03)00140-9.

35. Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. Gene *241*, 3–17. https://doi.org/10.1016/s0378-1119(99)00485-0.

36. Alpert, T., Straube, K., Carrillo Oesterreich, F., Herzel, L., and Neugebauer, K.M. (2020). Widespread transcriptional readthrough caused by Nab2 depletion leads to chimeric transcripts with retained introns. Cell Rep. *33*, 108324. https://doi.org/10.1016/j.celrep.2020.108324.

37. Iannone, C., and Valcárcel, J. (2013). Chromatin's thread to alternative splicing regulation. Chromosoma *122*, 465–474. https://doi.org/10.1007/s00412-013-0425-x.

38. Saldi, T., Cortazar, M.A., Sheridan, R.M., and Bentley, D.L. (2016). Coupling of RNA polymerase II transcription elongation with pre-mRNA splicing. J. Mol. Biol. *428*, 2623–2635. https://doi.org/10.1016/j.jmb.2016.04.017.

39. Zhou, H.L., Luo, G., Wise, J.A., and Lou, H. (2014). Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. Nucleic Acids Res. *42*, 701–713. https://doi.org/10.1093/nar/gkt875.

40. Simon, J.M., Hacker, K.E., Singh, D., Brannon, A.R., Parker, J.S., Weiser, M., Ho, T.H., Kuan, P.F., Jonasch, E., Furey, T.S., et al. (2014). Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects. Genome Res. *24*, 241–250. https://doi.org/10.1101/gr.158253.113.

41. Pabis, M., Neufeld, N., Steiner, M.C., Bojic, T., Shav-Tal, Y., and Neugebauer, K.M. (2013). The nuclear cap-binding complex interacts with the U4/U6.U5 tri-snRNP and promotes spliceosome assembly in mammalian cells. RNA *19*, 1054–1063. https://doi.org/10.1261/rna.037069.112.

42. Rambout, X., and Maquat, L.E. (2020). The nuclear cap-binding complex as choreographer of gene transcription and pre-mRNA processing. Genes Dev. *34*, 1113–1127. https://doi.org/10.1101/gad.339986.120.

43. Schmitz, U., Pinello, N., Jia, F., Alasmari, S., Ritchie, W., Keightley, M.C., Shini, S., Lieschke, G.J., Wong, J.J.L., and Rasko, J.E.J. (2017). Intron retention enhances gene regulatory complexity in vertebrates. Genome Biol. *18*, 216. https://doi.org/10.1186/s13059-017-1339-3.

44. Tan, Z.W., Fei, G., Paulo, J.A., Bellaousov, S., Martin, S.E.S., Duveau, D.Y., Thomas, C.J., Gygi, S.P., Boutz, P.L., and Walker, S. (2020). O-GlcNAc regulates gene expression by controlling detained intron splicing. Nucleic Acids Res. *48*, 5656–5669. https://doi.org/10.1093/nar/gkaa263.

45. Lykke-Andersen, S., and Jensen, T.H. (2015). Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. Nat. Rev. Mol. Cell Biol. *16*, 665–677. https://doi.org/10.1038/nrm4063.

46. Yap, K., Lim, Z.Q., Khandelia, P., Friedman, B., and Makeyev, E.V. (2012). Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. Genes Dev. *26*, 1209–1223. https://doi.org/10.1101/gad.188037.112.

47. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

48. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. *12*, 323. https://doi.org/10.1186/1471-2105-12-323.

49. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25. https://doi.org/10.1186/gb-2009-10-3-r25.

50. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. https://doi.org/10.1038/nbt.1754.

51. Gu, Z., Eils, R., Schlesner, M., and Ishaque, N. (2018). EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. BMC Genom. *19*, 234. https://doi.org/10.1186/s12864-018-4625-x.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
RNeasy Mini kit	Qiagen	#74104
DNAse	Qiagen	#79254
NEBnext Ultra II Directional RNA library prep kit	NEB	E7760
Deposited data		
RNA-seq and ribosome footprint profiling data	This paper	GEO: GSE197536
Ribosome footprint profiling data for heat shock	https://doi.org/10.1016/j.molcel.2012.11.028	GEO: GSE32060
Original DNA gel images	This paper	https://data.mendeley.com/datasets/znryjb62y3/1
Experimental models: Cell lines		
NIH3T3 mouse fibroblasts	ATCC	CRL-1658
Oligonucleotides		
Primers used for RT-PCR of Furin-Fes, forward: TGGACACGAGATAATGTTAGAGG	This paper	N/A
Primers used for RT-PCR of Furin-Fes, reverse: TGATTCCTGCTTCCTCCTCC	This paper	N/A
Primers used for RT-PCR of Tnfrsf12a-Thoc6, forward: CACGGAAACAACCATCTCCC	This paper	N/A
Primers used for RT-PCR of Tnfrsf12a-Thoc6, reverse: CTGTCGCCCAGCAACTAAGG	This paper	N/A
Primers used for RT-PCR of Nectin2-Tomm40, forward: CTGGGCATCTGGGTTGGGAATTT	This paper	N/A
Primers used for RT-PCR of Nectin2-Tomm40, reverse: CCGAGCGGCGGCAGAGTGAA	This paper	N/A
Primers used for RT-PCR of Tmem62-Ccndbp1, forward: CACCGATCATGGGCTTTCTG	This paper	N/A
Primers used for RT-PCR of Tmem62-Ccndbp1, reverse: CGGAGCAAGGAAGGGGAC	This paper	N/A
Software and algorithms		
R language	Version 4.2.1	https://www.r-project.org/
RStudio	Version 2022.02.3	https://www.rstudio.com/
DoGfinder	Wiesel et al., 2018[12]	https://github.com/shalgilab/DoGFinder
STAR	Dobin et al., 2013[47]	https://github.com/alexdobin/STAR
RSEM	Li and Dewey, 2011[48]	https://github.com/deweylab/RSEM
Bowtie2	Langmead et al., 2009[49]	https://bowtie-bio.sourceforge.net/bowtie2/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bedtools	Quinlan and Hall, 2010[22]	https://github.com/arq5x/bedtools2
Integrative Genomics Viewer (IGV)	Robinson et al., 2011[50]	https://software.broadinstitute.org/software/igv/
Other		
ENCODE CHIP-seq data for *Mus musculus* cell lines	https://www.encodeproject.org/	https://www.encodeproject.org/chip-seq-matrix/?type=Experiment&replicates.library.biosample.donor.organism.scientific_name=Mus+musculus&assay_title=Histone+ChIP-seq&assay_title=Mint-ChIP-seq&status=released

RESOURCE AVAILABILITY

Lead contact

Further information should be directed to the lead contact, Reut Shalgi (reutshalgi@technion.ac.il).

Materials availability

Requests for resources and reagents should be directed to the lead contact, Reut Shalgi (reutshalgi@technion.ac.il).

Data and code availability

- All RNA-seq and ribosome footprint profiling data were deposited in GEO, accession number GEO: GSE197536. Ribosome footprint profiling data for heat shock was taken from GEO, accession number GEO: GSE32060.
- Original DNA gel images have been deposited at Mendeley, dataset znryjb62y3 (Mendeley Data: https://data.mendeley.com/datasets/znryjb62y3/1), and are publicly available as of the date of publication.
- Custom scripts and any additional information required to reanalyze the data reported in this paper are available from the lead contact (reutshalgi@technion.ac.il) upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

NIH3T3 mouse fibroblast cells were seeded into 15cm plates. Cells were grown in DMEM medium (Sigma-Aldrich, D5796) supplemented with 10% Fetal Bovine Serum (Gibco, 16000044) and 1% Pen Strep (Sartorius, 03-031-1B) at 37°C and 5% CO2 conditions.

METHOD DETAILS

Datasets

Transcriptomics, using polyA selected RNA sequencing, and translatome mapping, using Ribosome footprint profiling, were conducted in duplicates, in two separate experiments, each having its own respective control. NIH3T3 mouse fibroblast cells were exposed to Heat shock (44°C for 2 h or 42°C for 8 h), or kept at 37°C as control. Additionally, NIH3T3 were exposed to Oxidative stress (0.2mM H2O2 for 2 or 7 h) and Osmotic stress (80mM KCl for 2 or 7 h). RNA was extracted using the RNeasy Mini kit (Qiagen, #74104) with on-column DNAse (Qiagen, #79254). Libraries were prepared using the NEBnext Ultra II Directional RNA library prep kit (NEB, E7760) with polyA selection, and sequenced using the BGI DNBseq platform, with read length of 100 bases. Illumina Nextseq 500 was used for QC and to validate the results from the BGI platform. Libraries were sequenced to a depth of 29–52M reads per library, and subsequent downsampling was performed using the DoGFinder package (see below). Ribosome footprint profiling was performed in house as in Sabath et al.,[4] and libraries were sequenced using an Illumina Nextseq 500 platform.

Mouse gene annotation

NCBI Refseq mm10 and UCSC RefSeq mm10 tables were downloaded from the UCSC genome browser website, using UCSC Table Browser data retrieval tool. To facilitate discovery of DoGs and read-in regions, NCBI Refseq annotation of mm10 mouse genome was combined with the UCSC Refseq annotation, by adding the isoforms unique to UCSC Refseq to the NCBI Refseq annotation. Annotations of non-translated short RNAs, such as snoRNAs, miRNAs and pseudogenes, were filtered out.

RNA-seq and Ribo-seq analysis – initial read mapping and transcript quantification

For RNA-seq analysis, preliminary filtering of sequences was performed. The raw reads were mapped to sequences of rRNAs using the STAR[47] aligner, with all mapped reads discarded. Following the filtering, the remaining raw reads were mapped to mm10 mouse genome, using the STAR aligner with the following parameters: –outFilterMatchNminOverLread 0.6. Transcriptome sequences were taken from the mm10 mouse genome annotation described above.

For Ribo-seq, preliminary filtering of non-mRNA footprint sequences was first performed. Raw reads post adapter trimming were mapped to sequences of rRNAs, tRNAs and snRNAs, using the STAR aligner, with all mapped reads discarded. Following filtering, the remaining reads were mapped to the coding sequences (CDS) of the mm10 mouse genome annotation described above, which were clipped by 30 nucleotides from the start and end of each CDS, since there is a positional bias of the ribosome in those areas. The mapping was done using the Bowtie2 aligner.[49] After alignment of RNA-seq and Ribo-seq samples, quantification was performed using the RSEM software package.[48] RSEM returns a raw read count and a normalized TPM (Transcripts Per Million) value for every gene and every isoform in the annotation. Expressed genes were defined as genes that had an RNA-seq TPM value larger than 4 in at least one sample (we note that they could potentially have lower TPM values in some of the samples, such as in Figure 3A). A total of 11,074 genes were defined as expressed using this criterion.

Identification of DoGs

Sorted bam files produced by RNA-seq STAR genomic alignment were used as input for the DoGFinder tool[12] in order to identify readthrough (DoG) regions. The annotation described above was used. We note that the DoGFinder tool downsampled all RNA-seq bam files (to a depth of 37M reads per library in the osmotic and oxidative stress experiments and corresponding controls, and to 29M reads per library in the heat shock experiments and corresponding controls) to have the same number of mapped reads, in order to avoid biases of DoG discovery between stress and control due to variable sequencing depths. The Get_DoGs function was run with the following parameters: -S -minDoGLen 100 -mode W -minDoGCov 0.6. DoGs RPKM were calculated using the DoGFinder tool. DoGs were filtered to have RPKM values of at least one and a total number of mapped reads of at least five. This resulted in an overall number of 1810 DoGs which were expressed in at least one condition, with, on average, 6.1% of all expressed genes having DoGs in each of the conditions. We note that, in cases where the DoG reached the TSS of a downstream gene on the same strand, the DoGFinder tool artificially determines the DoG end to be the TSS of the downstream gene.

Read-in genes

For each DoG-producing gene, a downstream genomic read-in region was defined as a 1kbp 5′ flanking region of the most upstream isoform of the proximal downstream gene on the same strand. In cases where the distance between genes was less than 1kpb, the entire region between the two genes was used. Coverage and read counts of the read-in regions were calculated using Bedtools.[22] Read-in genes were defined as genes located downstream of a DoG on the same strand, and for which the read-in region, namely the 1kb region upstream to the TSS of their most upstream isoform, overlapped with their upstream DoG, had an RPKM value of at least one (representing on average the top 2.69% of the distribution of RPKMs among all potential ‘‘read-in regions’’ within the population of all expressed genes) and read coverage of at least 0.6 (representing on average the top 3.7% of the distribution of read coverage among all potential ‘‘read-in regions’’ of the population of all expressed genes), in addition to requiring a TPM value of at least four for the read-in gene itself in the same condition. A set of all read-in genes was defined as all genes that were defined as read-in in at least one sample, resulting in a total of 307 read-in genes. To define read-in genes under a particular condition, a union of the read-in genes of the two replicate samples of that condition was taken.

To calculate the number of expected read-in genes given the selection parameters, we multiplied the average frequency for all of them within expressed genes, by the number of expressed genes, and the number of total samples:

$$\begin{aligned}&\%(\text{read} - \text{in} - \text{region RPKM} \geq 1) * \%(\text{read} - \text{in} - \text{region coverage} \geq 0.6) * \\ &\%(DoGs) * N(\text{all expressed genes}) * N(\text{total samples}) = \\ &0.026 * 0.037 * 0.061 * 11074 * 16 = 11.69\end{aligned}$$

This resulted in 11 genes, even without any further restrictions (which were inflicted on the read in gene group, i.e. to have their read-in region overlapping with their upstream DoG). Repeating this calculation with specific parameters of each of the conditions resulted in a similar number of 11 expected read-in genes.

RT-PCR of readthrough-read-in RNA chimeras

To validate the presence of readthrough-read-in RNA chimeras, NIH3T3 mouse fibroblast cells were exposed to osmotic stress (80mM KCl) for 2 h, or harvested in normal growth conditions (control), in replicates. RNA was extracted using the RNeasy Mini kit (Qiagen, #74104) with on-column DNAse (Qiagen, #79254), and cDNA was generated using MMLV reverse transcriptase (Promega, #M170A). A 100ng of cDNA was added to a 25ul PCR reaction using Phusion polymerase (NEB, #M0535L) according to manufacturer's instructions, for 25 cycles. no-RT control samples were generated similarly, but without the addition of the reverse transcriptase enzyme, to assess potential genomic DNA contamination. As shown in Figure 1E, no genomic DNA contamination was detected in the samples. The tested chimeras were chosen based on RNA-seq data, with both readthrough and read-in genes expressed in osmotic stress condition. Negative non-chimera tandem gene pairs were chosen to have similar intergenic distances and robust expression in osmotic stress condition, with no readthrough or read-in detected by RNA-seq. Primers were designed to amplify from the end of the readthrough gene, through the intergenic region to the start of the read-in gene (see primers in Table S5). PCR products were loaded on a 1% agarose gel and separated by electrophoresis. All gels were run simultaneously in the same running apparatus and exposed together (Figure 1E).

Non-read-in genes

A set of non-read-in genes was defined as a control group out of all expressed genes, by taking a set of all genes that were located downstream to DoGs on the same strand, but were not defined as read-in genes in any of the conditions. In addition, to further control for confounding factors related to gene proximity, this group of genes located downstream to DoGs was filtered by their distance from the end of their upstream DoG-producing gene, such that the distance was lower than the maximal distance between readthrough-read-in gene pairs (shown in Figure S4, a distance less than 12,190 bp). This resulted in a total of 241 non-read-in genes.

Feature analysis of gene sequences

For the analysis of gene length, coding sequence (CDS) length, exon length, intron length, 5′ and 3′ UTR lengths, number of introns and GC content of introns, the isoform most expressed in the RNA-seq control samples, as calculated by RSEM, was selected for each gene.

We examined the possibility that the result of read-in genes being short was confounded by our read-in selection parameters, and that longer read-in genes were polyadenylated before the end of the gene and therefore our selection criteria did not pick them up. We reasoned that if there were such longer read-in genes, with overall lower read coverage toward the end of the gene, they would not pass the required expression criteria of a TPM of 4 or above. Only 20 genes answered to all other criteria for read-in gene selection, while having a TPM <4, and they were shorter than read-in genes (9kb on average compared to 9.9kb of read-in genes). Thus, read-in gene length was not confounded by our selection criteria.

For the analysis of read-in regions GC content and polyA signals, the read-in region, defined as the 1kbp 5′ flanking region of the most upstream isoform of the gene, was used. In cases where the distance between genes was less than 1kpb, the entire region between the two genes was used.

Non-canonical polyA signals analysis

A list of non-canonical polyA signals was taken from[25] and further supplemented by all remaining sequences that were one base-pair different from the canonical AAUAAA signal, resulting in a set of 19 6-mers (AUUAAA, UAUAAA, AGUAAA, CAUAAA, GAUAAA, AAUAUA, AAUACA, AAUAGA, AAAAAG, ACUAAA, AAAAAA, AACAAA, AAGAAA, AAUUAA, AAUCAA, AAUGAA, AAUAAU, AAUAAC and AAUAAG). Frequency of occurrence of the canonical polyA signal (AAUAAA) and the set of all non-canonical polyA signals was calculated for the 3′UTR ends (50 bases upstream to annotated gene ends) of DoG-producing genes (or all expressed genes). For Figure 2F, each group of 3′UTR ends was divided into those that had a canonical polyA signal (left) and these that did not have it, but instead had at least one non-canonical polyA signal. We aggregated multiple 3′ UTR ends (50 bases upstream to annotated gene ends) when a gene had more than one 3′ end isoform, such that when at least one of them contained the canonical polyA signal, the gene was considered to have that signal. Chi-squared test was used to calculate the significance of the change in proportions of genes that had a canonical polyA, or any of the non-canonical polyA signals between different gene sets.

Read-in estimation values

For each read-in gene, and for each sample, read-in estimation value was defined as follows: the read density of the read-in region was divided by the read density of the most highly expressed isoform of the putative read-in gene in the specific sample. Read-in estimation values for a given stress were defined as the mean of the values of the two replicate samples.

$$\text{Read} - \text{in estimation} = \frac{\text{read} - \text{in region read density}}{\text{read} - \text{in gene (isoform) read density}}$$

When read-in genes were stratified according to their read-in estimation (low, medium and high), they were split into three equal-sized groups, and the values differed slightly according to the specific condition. For the osmotic stress, 2h (Figure 3B, 3C, and 4A), low read-in estimation values were less than 15.73%, high were above 67.48%, and medium was anything in between. For the cutoffs for each condition see Table S3.

Intron retention

For each intron and for each sample, the number of exon junction reads and intronic reads was extracted using the Bedtools software,[22] and read densities were calculated. IR (Intron Retention) was calculated using the following formula:

$$IR = \frac{\text{intron read density}}{\text{intron read density} + \text{flanking exon junction read density}}$$

Introns that had a flanking exon junction read density of zero were discarded. For each gene in each sample, the most highly expressed isoform was selected, and the intron with the highest IR value was selected to represent the maximum intron retention for this gene. For analysis of first introns retention, the first intron of the most highly expressed isoform was selected. Intron retention for a given condition was defined as the mean of the IR values in the two replicate samples of that condition.

Splice site strengths analysis

For splice site strength analysis, the most highly expressed isoform for each gene was chosen, and a custom R script was used to generate fasta files of splice site sequences of the first intron of the isoform. The MaxEnt software was used to calculate the 5′ and 3′ splice site strength scores.[27]

Analysis of read-in genes translation

To assess the difference in translation given the expression levels of read-in genes compared to all expressed genes, translation efficiency (TE, defined as translation TPM values divided by expression TPM values) was calculated for each gene in each sample. TE of a gene for a given stress was defined as the mean of the TE values in the two replicate samples. For each of the three read-in estimation groups (low, medium and high), mean log2 TE was calculated, and groups of 1000 random samples of TE values were taken from the set of expressed genes, while matching the range of TPM values to that of the read-in genes within the specific group. p-values were calculated by comparing the distributions of the mean log2 TE values of the random samples to the mean values of each of the read-in estimation groups. See Figure S10 for the distributions of TE values for the different groups of interest, together with those of all randomly-selected matched control groups (aggregated).

Controlled analysis of intron retention within read-in genes

Read-in genes were divided into 16 groups according to a combination of 4 quantiles of intron lengths and 4 quantiles of intronic GC content. Then, 1000 randomized comparison groups with the same GC content and intron lengths distributions as read-in genes were randomly picked from the set of all expressed genes without read-in genes (See Figure S17 for the distribution of GC content and intron length of the random controls vs. that of read-in genes). For each comparison group, the mean value of the maximum IR for each gene was calculated. See Figure S18 for the distributions of IR values for the different groups of interest, together with that of all randomly-selected matched control groups (aggregated).

Distance-matched control genes

A set of intergenic distance-matched control genes were defined by taking a set of genes (set A) which belonged to the all expressed genes set but were neither read-in nor non-read-in genes, and for which their upstream gene on the same strand was also defined as an expressed gene, but not a DoG-producing gene. From this set A, a subset of genes was randomly sampled, which had a similar distribution of intergenic distances as that of read-in genes. Sampling was performed by splitting the read-in genes upstream intergenic distances into 5 quantiles, and sampling from set A equal numbers of genes that belonged to each of the distance quantiles. This resulted in a total of 220 distance-matched control genes.

Histone modification analysis

In order to explore the histone modification landscape in genomic regions of interest, we used the ENCODE database of CHIP-seq experiments.[28] We downloaded the ChIP-seq histone experiments in all available mouse cell lines. The processed data was downloaded in the narrowPeak bed and bigWig formats. Bedtools[22] was used to find overlaps between ChIP-seq bed files and genomic regions of interest, such as read-in regions, intronic regions, etc. For each histone modification, first, an equality of proportions hypothesis test was performed to see if the histone modification tended to be more present within read-in regions, i.e. the 1kb upstream to the TSS of the most upstream isoform of the gene, in read-in genes compared to non read-in genes and compared to the intergenic distance-matched control group (see above). We considered significance only for modifications which showed a significantly higher proportion of presence in read-in genes compared to both background control groups in each specific cell line examined. To generate H3K36me3 profiles, the EnrichedHeatmap R package[51] was used. bigWig format ChIP-seq data files were mapped to regions of interest, such as gene bodies, read-in regions and introns, using the normalizeToMatrix function of the EnrichedHeatmap package, generating a length-normalized peak density values vector for each region of interest and each gene, with additional flanking regions of 1kb. Mean and standard error vectors for each of the three gene groups of interest (read-in genes, non-read-in genes and intergenic distance-matched control genes), as well as of all expressed genes, were then calculated and plotted (Figures 6, S20, and S21).

Profile difference p values between different gene groups were calculated, for each coordinate along the normalized region of the peak density plot, by performing a Wilcoxon ranksum of the peak density values in that specific coordinate (Figures 6 and S20, bottom p value panels, FDR corrected). Normalized peak density slope values of introns of different gene groups (Figure S21B) were generated by smoothing (R function *supsmu*, span = 0.175) and *Z* score normalizing the peak density vector in Figures 6 and S20B, and then calculating the slope at each point by denoting the difference between each vector element and the element before it.

Analysis of human orthologs

For an analysis of human gene orthologs, NCBI Refseq hg38 tables were downloaded from the UCSC genome browser website, using the UCSC Table Browser data retrieval tool, and orthologs were determined by name. Annotations of non-translated short RNAs, such as snoRNAs, miRNAs and pseudogenes, were filtered out. A table of genes and their upstream neighboring genes on the same strand for the human and mouse genomes was constructed. Conserved gene pairs were defined as gene orthologs that had the same upstream neighboring gene orthologs in the mouse and human genomes. As an additional control for proximal pairs, we restricted the mouse all expressed gene pairs to those that had an intergenic distance in mouse smaller than the cutoff used to generate the non read-in gene group (see above). Chi-squared test was then used to calculate the significance of the change in proportions of conserved pairs between different gene groups. For Figures 7A and 7C, random sampling of gene groups, with distribution of distances to their upstream neighboring genes matched to that of read-in genes (as described above) was performed 10,000 times. Median values of the differences in intergenic distances between mouse and human genes (7A), as well as median values of the differences in mean intron length (per gene) between mouse and human genes (7C) were calculated and plotted as the background, and compared to the corresponding median values of the read-in and non-read-in gene groups, in order to assess significance of conservation of distances.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical details of the experiments (significance tests and p values) can be found in the figure legends. Wherever asterisks are used to indicate significance, (*) indicates $p < 0.05$, (**) indicates $p < 0.01$ and (***) indicates $p < 0.001$.