




# Post-transcriptional regulation supports the homeostatic expression of mature RNA

Zheng Su <sup>1</sup>, Mingyan Fang <sup>2,3</sup>, Andrei Smolnikov<sup>1</sup>, Fatemeh Vafaei <sup>1</sup>, Marcel E. Dinger<sup>1,4,\*</sup>, Emily C. Oates<sup>1,5,\*</sup>

<sup>1</sup>School of Biotechnology and Biomolecular Sciences, Faculty of Science, The University of New South Wales, Biological Sciences North Building (D26), Upper Kensington Campus, Sydney, New South Wales 2052, Australia

<sup>2</sup>BGI Research, Building 1, Future Science and Technology Innovation Mansion, No. 59, Science and Technology 3rd Road, East Lake High-tech Development Zone, Wuhan City, Hubei Province, 430074, China

<sup>3</sup>BGI Australia, L6, CBCRC, QIMR Medical Research Institute, 300 Herston Road, Herston, QLD 4006, Australia

<sup>4</sup>School of Life and Environmental Sciences, Faculty of Science, University of Sydney, F22 Life, Earth and Environmental Sciences (LEES) Building, Camperdown NSW 2050, Australia

<sup>5</sup>Department of Neurology, Sydney Children's Hospital, High St, Randwick NSW 2031, Australia

\*Corresponding authors. Emily C. Oates, Tel: +612 90654060, E-mail: [e.oates@unsw.edu.au](mailto:e.oates@unsw.edu.au); Marcel E. Dinger, Tel: +61293512222, E-mail: [marcel.dinger@sydney.edu.au](mailto:marcel.dinger@sydney.edu.au)

## Abstract

Gene expression regulation is a sophisticated, multi-stage process, and its robustness is critical to normal cell function and the survival of an organism. Previous studies indicate that differential gene expression at the RNA level is typically attenuated at the protein level through translational regulation. However, how post-transcriptional regulation (PTR) influences expression change during the RNA maturation process remains unclear. In this study, we investigated this by quantifying the magnitude of expression change in precursor RNA and mature RNA across a vast range of different biological conditions. We analyzed bulk tissue RNA sequencing data from 4689 samples, including healthy and diseased tissues from human, chimpanzee, rhesus macaque, and murine sources. We demonstrated that PTR tends to support homeostatic expression of mature RNA by amplifying normal tissue-specific expression of precursor RNA, while reducing expression change of precursor RNA in disease contexts. Our study provides insight into the general influence of PTR on gene expression homeostasis. Our analysis also suggests that intronic reads in RNA-seq studies may contain under-utilized information about disease associations. Additionally, our findings may assist in identifying new disease biomarkers and more effective ways of altering gene expression as a therapeutic strategy.

**Keywords:** post-transcriptional regulation; RNA expression; RNA-seq

## Introduction

Regulation of gene expression is a complicated, multi-stage process. DNA is initially transcribed into precursor RNA (pre-RNA, including subtypes such as precursor messenger RNA and precursor non-protein-coding RNA), subject to transcriptional regulation [1–4]. Pre-RNA is subsequently processed into mature RNA. This process is subject to post-transcriptional regulation (PTR), which determines the stability, abundance, localization and even sequence of mature RNA isoforms. PTR mechanisms include alternative splicing, 5' capping, alternative polyadenylation, binding of regulatory RNAs, RNA methylation and RNA editing, among others [5–9]. Subsequently, mature RNA is translated into protein, subject to translational and post-translational regulation.

The complexity of PTR activities varies across species and is not constant during evolution. Interestingly, highly complex organisms (comprised of multiple different cell types) tend to have more complex PTR processes than unicellular and simpler multicellular organisms [10–14]. While complicated, PTR should represent a significant energy burden to cells and may increase the chance of regulatory errors. Its evolutionary selection in complex organisms suggests that the benefits it provides may outweigh the associated energy costs and risks, or that it is not deleterious enough to be

removed by selective pressure. As one process of PTR, alternative splicing is known to expand the repertoire of proteins and increase the diversity of cellular function [15]. However, many other PTR processes are not directly involved in pre-RNA splicing. The evolutionary implications of these processes have not been fully elucidated.

Organisms have developed various mechanisms to maintain robustness against external disturbances and internal aberrations [16–18]. One such mechanism is the attenuation of protein expression change; expression change across biological conditions at the RNA level is reduced at the protein level [19–21]. A proposed explanation is that protein is under greater evolutionary pressure compared to RNA, and that this mechanism may have evolved to buffer fluctuations in gene expression at the RNA level, ensuring robustness of components closer to cellular function [19]. However, it remains unclear how gene expression differences change from pre-RNA to mature RNA, and what the role of PTR is in this process.

One possibility is that gene expression change across conditions is larger at the pre-RNA level compared to the mature RNA level. This may be because mature RNA is more closely related to cellular function and is therefore subjected to higher evolutionary

**Received:** August 30, 2024. **Revised:** October 31, 2024. **Accepted:** February 5, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

pressures, compared to pre-RNA. Another possibility is that gene expression change is smaller at the pre-RNA level and becomes amplified in mature RNAs. There is also a third possibility: that there is no consistent pattern in the relative gene expression change between pre-RNA and mature RNA levels, and it varies from condition to condition. Finding out which of these three possibilities is true is important, as it will provide us insight into the general roles of pre-RNA and mature RNA, as well as the role evolution has played in shaping transcription regulation and PTR. To determine which of these three possibilities is most likely occurring in nature, we comprehensively examined the influence of PTR on gene expression in various conditions, including normal and disease states, in human and other mammalian species.

## Materials and methods

### Sample selection and ethics statement

Datasets used in the study were obtained from the GTEx [22] and SRA [23] databases. For the GTEx dataset, bulk RNA-seq data in Analysis V8 was used; samples were filtered to retain only those with RNA Integrity Number (RIN) > 7, and tissues with greater than 100 samples were down-sampled to 100 samples for downstream analyses (Table S1, available online at <http://bib.oxfordjournals.org/>). BAM files were downloaded for GTEx samples. For the SRA datasets of the SRA human tissues (Table S2, available online at <http://bib.oxfordjournals.org/>), the non-human tissues (Table S3, available online at <http://bib.oxfordjournals.org/>) and the disease conditions (Table S4, available online at <http://bib.oxfordjournals.org/>) studies were manually selected from recount2 database [24] to include disease conditions with intermediate sample sizes. The meta data of SRA samples were extracted from annotation information in <https://sra-explorer.info/>, Expression Atlas [25] or from the corresponding literatures. FASTQ files were downloaded and used for the SRA dataset.

The studies involving human samples were conducted in compliance with ethical standards and procedures, with oversight and approval by the University of New South Wales (UNSW) Human Research Ethics Committee (Project No. HC230388).

### Data quality control

The downloaded FASTQ files were first quality controlled by FastQC [26]; samples that were marked as failed in 'Per sequence quality scores' by the software were discarded. BAM files of human samples were then quality controlled by RNA-SeQC [27], a program that enables filtering based on key measures of RNA-seq data quality, using criteria of Mapping Rate > 0.7, Base Mismatch < 0.02, High Quality Rate > 0.6, Exonic Rate > 0.3, Ambiguous Alignment Rate < 0.1, rRNA Rate < 0.4, Avg. Splits per Read < 0.4 and 5000 < Genes Detected < 30,000. The thresholds were determined according to the distribution of the values of each metric in the studied samples. Samples that failed to meet any of the criteria were excluded from downstream analyses.

### Preparation of annotation files

Gencode v38 gff annotation file [28] was used to extract human exon coordinates. To obtain the intron coordinates, the command 'genometools gt gff3 -retainids -addintrons -tidy' in toolkit genometools [29] was first used to add introns, then 'bedtools subtract' [30] was used to subtract exons from introns to ensure there was no overlapping exonic region in the final introns. For quantification of reads mapped to 5' end of pre-RNAs and mature RNAs, two introns and two exons at the 5' end of genes were selected, respectively, and were used to generate new gff files.

Similarly, two introns and two exons at the 3' end were selected to generate new gff files for 3' end quantification.

For non-human species, the relevant gff files were downloaded from Ensembl Release 105 [31]. The same method used for human genes was used to extract intron coordinates of genes for non-human species.

### Quantification of pre-RNA and mature-RNA abundance

STAR v2.7.5a was used to align reads in FASTQ files to GRCh38 [32]. For human samples, the STAR index used by GTEx was downloaded from [gs://gtex-resources/STAR\\_genomes/](gs://gtex-resources/STAR_genomes/) via gsutil, and it was used in the analysis. For non-human species, STAR index was built using command STAR—runMode genomeGenerate with parameters of—genomeFastaFiles \$genomeFastaFiles—sjdbGTFfile \$sjdbGTFfile—sjdbOverhang 100—limitGenomeGeneRateRAM=45,000,000,000—genomeSAindexNbases, in which genomeSAindexNbases was set as 14. TOPMed RNAseq pipeline (<https://github.com/broadinstitute/gtex-pipeline>), which was used by GTEx, was used for read alignment and duplicates marking.

Quantification of exonic and intronic reads at gene level was performed using featureCounts [33], using parameters of -a GFF\_FILE -o OUTPUT -F GTF -t exon/intron—ignoreDup -p -J—minOverlap 10 INPUT.bam, where GFF\_FILE was the previously prepared exon or intron annotation file. As most of the intronic reads are from pre-RNAs, while exonic reads can be from both pre-RNAs and mature RNAs, we used intronic read counts as pre-RNA read counts. We then used the following formula to calculate the mature RNA read counts by assigning exonic reads to both pre-RNAs and mature RNAs:

$$C_{\text{mature-RNA}} = C_{\text{exonic reads}} - \left( \frac{C_{\text{intronic reads}}}{L_{\text{intron}}} \times L_{\text{exon}} \right)$$

Where  $C_{\text{mature-RNA}}$  is mature RNA read count for a gene,  $C_{\text{exonic reads}}$  and  $C_{\text{intronic reads}}$  were the counts of reads mapped to exonic and intronic regions of the genes respectively, and  $L_{\text{exon}}$  and  $L_{\text{intron}}$  were the total exon and intron length of the gene, respectively. The term  $C_{\text{intronic read}} / L_{\text{intron}} * L_{\text{exon}}$  is the exonic reads assigned to pre-RNAs, with pre-RNA abundance being estimated from intronic reads, and normalized by intron and exon lengths.

### Extraction and quantification of boundary reads

To extract the exon-exon boundary reads, a bed file was created, in which two lines were generated for each exon, with the first line recording the start position (same value for the second and third columns) while the second line recording the end position of the exon. Then 'bedtools intersect' command was used to extract reads in bam files that mapped to the start or end position of the exon, and mapping information was extracted from CIGAR string to ensure the reads had at least 10 bp overhang at both ends. The extracted boundary reads were then counted by featureCounts using parameters -a EXON.gff3 -o OUTPUT -F GTF -t exon—ignoreDup -p -J—splitOnly—minOverlap 10 EXTRACTED\_READS.bam.

Similarly, to extract exon-intron boundary reads, a bed file was created, in which two lines for each intron were generated, with the first line recording the start position minus 10 bp (the minimum overhang) while the second line recording the end position plus 10 bp of the intron. The 'bedtools intersect' command was used to extract reads in bam files that mapped to start or end position in the created bed file. Mapping information was extracted from CIGAR string to remove reads with any 'N'

(represents skipped bases on the reference). The extracted boundary reads were then counted by featureCounts using parameters -a INTRON.gff3 -o OUTPUT -F GTF -t intron—ignoreDup -p -J—nonSplitOnly—minOverlap 10 EXTRACTED\_READS.bam.

## Preprocessing for differential expression analysis and subsampling of mature RNA reads

The calculated pre-RNA and mature RNA reads were used for differential expression analysis. First, genes with total exon length  $\leq 200$  bp or total intron length  $\leq 200$  bp were excluded, only genes that had exon to intron length ratio between 0.001 and 10 were retained.

The number of mature RNA reads was significantly higher than pre-RNA reads (Fig. S1a available online at <http://bib.oxfordjournals.org/>). Although there was no apparent correlation between read counts and delta fold change (Pearson  $r = -0.08$ , Fig. S1b available online at <http://bib.oxfordjournals.org/>), to avoid any unexpected influence of read count on the analysis, we performed random subsampling of mature mRNA reads for each sample to balance the number of pre-RNA and mature RNA reads, and used the subsampled reads for differential expression analysis and differential expression magnitude quantification. To subsample mature RNA reads, for each sample, we first calculated the trimmed mean pre-RNA count and trimmed mean mature RNA count. To calculate the trimmed mean pre-RNA count, we reordered genes based on pre-RNA counts from the smallest to the largest, and removed 10% genes at both ends, then calculated the mean counts of the remaining genes, by formula:

$$\overline{C_{pre-RNA}} = \frac{1}{n - 2[k]} \sum_{[k]+1}^{n-[k]} C_{pre-RNA \bullet i}$$

Where  $\overline{C_{pre-RNA}}$  is the mean pre-RNA count for the sample,  $n$  is the number of genes,  $[k]$  is the integer closest to  $k = 0.1 * n$ , and  $C_{pre-RNA \bullet i}$  is the pre-RNA count for gene  $i$ , similarly, mean mature RNA count  $\overline{C_{mature RNA}}$  was calculated as:

$$\overline{C_{mature RNA}} = \frac{1}{n - 2[k]} \sum_{[k]+1}^{n-[k]} C_{mature RNA \bullet i}$$

Then the number of subsampled reads  $S_{mature RNA \bullet i}$  for the gene is one random sampling from binomial distribution if  $C_{pre-RNA} < \overline{C_{mature RNA}}$ .

$$S_{mature RNA \bullet i} \sim B(C_{mature RNA \bullet i}, p), \text{ where } p = \overline{C_{pre-RNA}} / \overline{C_{mature RNA}}$$

## Differential expression analysis

Differential expression analysis was performed using both DESeq2 [34] and edgeR [35], at pre-RNA and mature RNA levels, separately. In the analysis using DESeq2, low count genes with total read counts  $\leq 15$  were excluded, then differential expression analysis was performed using the 'DESeq' function, which performed Negative Binomial GLM fitting and Wald statistics. Shrunk  $\log_2$ FoldChanges were generated by the 'lfcShrink' function using estimator of 'apeglm' [36]. In the analysis using edgeR, only genes with 10 reads in  $0.7 * \text{min.group.size}$  samples were retained, where  $\text{min.group.size}$  was the size of the smaller group. Library size was normalized by method 'TMM', tagwise dispersions were estimated by function 'estimateDisp', then a quasi-likelihood negative binomial generalized log-linear

model (function glmQLFit()) was used to fit the count data and quasi-likelihood F-test (function glmQLFTest()) was used to test for differentially expressed genes.

The  $\log_2$ FoldChange values from differential expression analysis were used to calculate delta fold change using formula:

$$\text{Delta fold change} = |\log_2 FC_{mature RNA}| - |\log_2 FC_{pre-RNA}|$$

Where  $|\log_2 FC_{mature RNA}|$  was the absolute value of  $\log_2$  fold change observed at mature RNA level, and  $|\log_2 FC_{pre-RNA}|$  was the absolute value of  $\log_2$  fold change observed at pre mRNA level.

Only genes with  $|\log_2 FC| > 0.5$  and un-adjusted p-value  $< 0.05$  at either pre-RNA or mature RNA level were used to calculate the mean or median delta fold change for each comparison. To ensure that there were sufficient genes in the calculation of the mean or median delta fold change, only comparisons with more than 50 filtered genes were used in the analyses to generate Figs. 1, 2, and S2.

Data download, quality control, alignment, quantification and differential expression analysis was performed using UNSW Australia's Katana supercomputing facility [37].

## Correlation with RNA integrity and gene length

RNA integrity of the samples was measured by Transcript Integrity Number [38], which was calculated from the sequencing reads using RSeQC [39]. Median Transcript Integrity Numbers across all the transcripts were used to measure the RNA integrity at sample level. For each comparison of gene expression between tissues or conditions, the mean Transcript Integrity Number value across all samples was used to measure the RNA integrity of the comparison. To evaluate the contribution of RNA integrity on the mean delta fold change values, a linear model of

$$\text{Mean delta fold change} = \beta_0 + \beta_1 \text{database} + \beta_2 \text{condition} + \beta_3 \text{TIN} + \varepsilon$$

was fitted, where database was the source of data (SRA or GTEx), condition was either diseases or tissues and TIN was the mean Transcript Integrity Number of a comparison. Magnitude of contribution was measured by the absolute value of the coefficient  $\beta_3$ .

Correlation between mean delta fold change and total/mean length of exons/intron and total length of genes in GTEx tissues were explored using scatter plot and linear regression, and the strength of the correlation was measured by R-squared values.

## Gene set analysis, statistical analysis, and visualization

The human housekeeping gene list was downloaded from HRT Atlas v1.0 database [40], genes in file 'Housekeeping genes', which were genes stably expressed across 52 tissues and cell types, were used. Orthologs of human housekeeping gene in non-human species were used as their housekeeping genes. Ortholog information was retrieved from Orthologous Matrix (OMA) database [41], specifically, orthologs between each non-human species and human were queried at <https://omabrowser.org/oma/genomePW/>, only orthologs with 1:1 mapping relations were used.

Pathway and gene ontology enrichment analysis was performed using Enrichr [42] via their API, gene set library of 'KEGG\_2021\_Human' [43] was used for pathway analysis, while libraries of 'GO\_Cellular\_Component\_2023', 'GO\_Biological\_Process\_2023' and 'GO\_Molecular\_Function\_2023' [44] were used

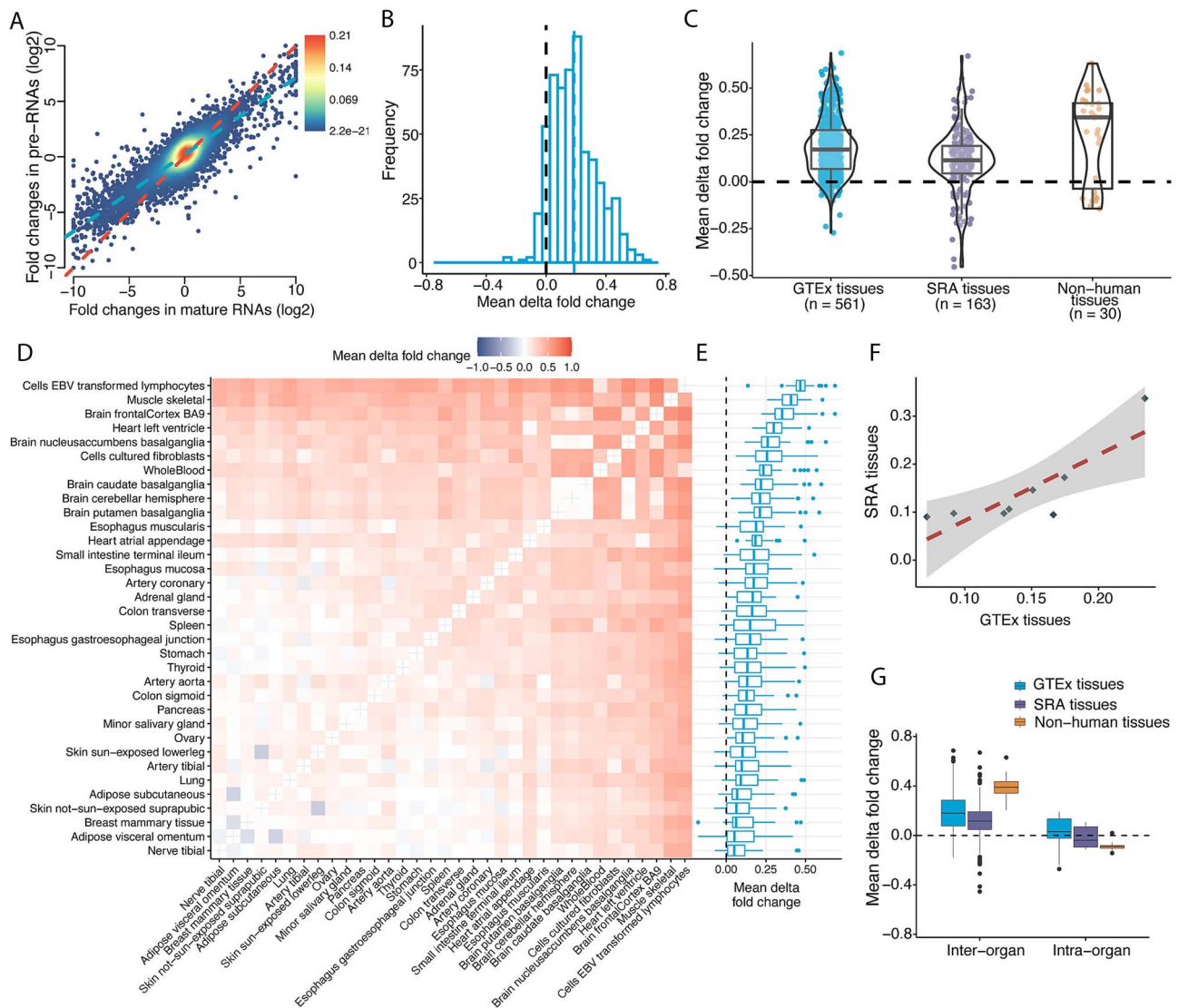


Figure 1. Differential expression magnitude between tissues in pre-RNAs and mature RNAs. (a) An illustrative example showing amplified differential expression magnitude in mature RNAs between brain caudate basal ganglia and heart atrial appendage tissues in the GTEx dataset. Each dot represents a gene; the zero delta fold change reference line and the regression line are shown with dashed lines. (b) Distribution of mean delta fold changes across GTEx tissue comparisons, with the zero delta fold change and mean value lines represented by dashed lines. (c) Comparison of the distribution of mean delta fold changes across GTEx tissues, SRA tissues, and non-human mammal tissues. Each data point is a comparison between two tissues. (d) Heatmap showing mean delta fold changes for all GTEx tissue comparisons. (e) Boxplot depicting the distribution of mean delta fold changes in inter-organ comparisons, ordered by their median values. (f) Correlation of the median values of mean delta fold changes between GTEx and SRA datasets, calculated using the tissues present in both datasets. (g) Distribution of mean delta fold changes in inter-organ and intra-organ comparisons in three datasets.

for gene ontology analysis. In the analysis to determine whether the genes with amplified (or reduced) expression change are consistent across different tissues and disease states, for each tissue, we selected all 33 differential expression comparisons involved. Then we identified the top 1000 genes most frequently observed with amplified (or reduced) expression change in those 33 comparisons. For each identified gene, we calculated the percentage of tissues in which the gene was observed with amplified (or reduced) expression change. For the diseases dataset, only the diseases with >30 genes observed amplified (or reduced) expression change were considered, to reduce the bias in the result caused by studies with small number of genes with amplified (or reduced) expression change. For disease enrichment analysis, geneset databases of Clinvar [45], OMIM [46] and GWAS Catalog [47] from Enrichr [42] were used for analysis.

Statistical analysis was performed in R version 4.1.2, visualization was done using packages ggplot2 [48], ggpubr [49], ggfortify [50], ggsci, ggstatsplot [51], grid, ggthemes, and RColorBrewer.

## Results

### Influence of PTR on tissue-specific gene expression

To investigate how PTR influences tissue-specific expression in humans, we examined gene level expression of protein-coding and non-protein-coding genes across various human tissues. We randomly selected and downloaded bulk RNA-seq data of 3221 samples representing 34 tissues from the GTEx database, with ~100 samples per tissue (Table S1, available online at <http://bib.oxfordjournals.org/>, Methods). Sequencing reads were aligned to

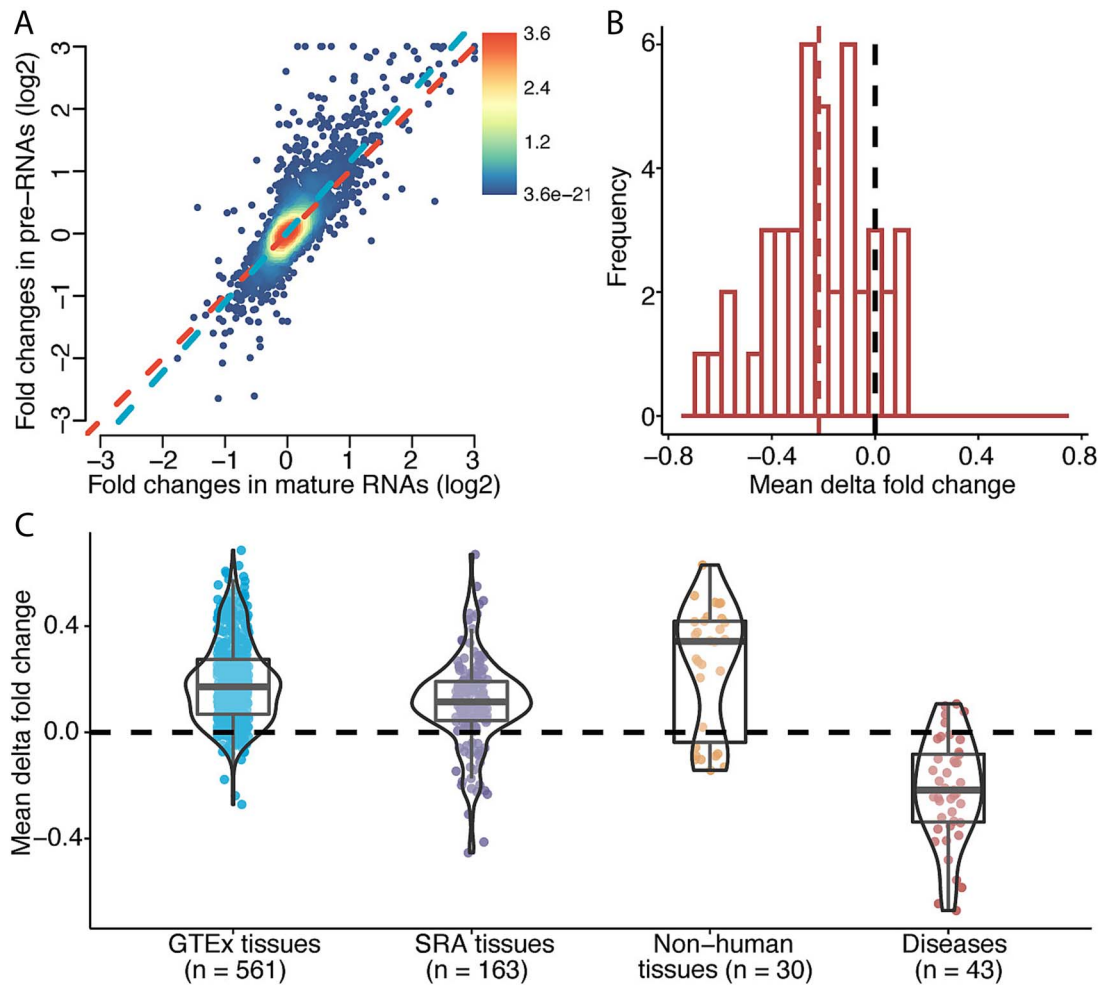


Figure 2. Reduction of expression dysregulation magnitude in mature RNAs in disease conditions. (a) Illustrative example of reduced differential expression magnitude, in *Mycobacterium smegmatis* infection samples. Each dot represents one gene; the zero delta fold change reference line and the regression line are shown with dashed lines. (b) Distribution of mean delta fold changes across various disease conditions, with the zero delta fold change and mean value lines represented by dashed lines. (c) Comparative analysis of the distribution of mean delta fold changes across GTEx tissues, SRA tissues, non-human tissues, and disease conditions. Each data point is a comparison between two groups (e.g. two tissues or healthy group versus disease group).

the human reference genome, and reads mapped to genic regions were quantified. We separated the mapped reads into intronic reads and exonic reads, and used them to infer pre-RNA and mature RNA abundance, which has been previously shown to be a robust approach [52]. This has taken the advantage of the finding that intronic reads in RNA-seq are not technical artifacts but reflect transcriptional activity and pre-RNA expression [52, 53]. As most of the intronic reads are from pre-RNAs, while exonic reads can be from both pre-RNAs and mature RNAs, we assigned exonic reads to both pre-RNAs and mature RNAs according to the calculated pre-RNA abundance (Methods).

We then identified genes that were differentially expressed between tissues, analyzing pre-RNA and mature RNA levels separately. We tested for differential expression between all possible tissue combinations; 561 pairwise comparisons in total. We measured the magnitudes of expression change for pre-RNA and mature RNA levels separately. To quantify the effect of PTR on these magnitudes, we derived a metric named delta fold change. It is the difference in expression fold changes (using absolute values to account for negative fold changes) between pre-RNAs and mature RNAs (Methods). For a specific gene, a delta-fold change value greater than zero indicates that the magnitude of

expression change in mature RNAs is greater than in pre-RNAs, and vice versa.

Interestingly, we found that the mean delta fold changes were greater than zero in most tissue comparisons, indicating that the magnitude of expression change in mature RNAs is generally larger than in pre-RNAs (Fig. 1a, b,  $P < 1e-16$ , one-sided one-sample Wilcoxon signed-rank test). This observation suggests that PTR tends to amplify the expression differences between tissues.

To ensure that this amplification was not an artifact arising from a peculiar characteristic of the GTEx dataset analyzed, we expanded our analysis to human tissue data from other studies. We analyzed another human tissue dataset (here referred to as the SRA tissue dataset) of 172 samples from two different studies, one with 18 tissues and another with five tissues (Table S2 available online at <http://bib.oxfordjournals.org>). In this dataset, we observed a significant post-transcriptional differential expression amplification pattern similar to that observed in the GTEx dataset (Fig. 1c,  $P = 1.05e-14$ , one-sided one-sample Wilcoxon signed-rank test). Moreover, to determine whether the pattern is unique to human species, or is a general property existing in other mammal species, we analyzed RNA-seq data from five different tissues of three other mammal species, including *Macaca mulatta*, *Mus*

*musculus* and *Pan troglodytes* (Table S3 available online at <http://bib.oxfordjournals.org/>), and observed a similar significant differential expression amplification pattern (Fig. 1c,  $P = 1.34e-5$ , one-sided one-sample Wilcoxon signed-rank test). We also investigated whether our observations could arise from analytical biases, and the results are detailed in the following section.

To further explore the possible correlation between differential expression amplification and the tissue of origin, we compared the mean delta fold changes between different tissues and organs in GTEx dataset. We observed a non-uniform distribution, with some specific tissues, such as the brain, muscle and heart tissues, typically showing larger delta fold changes (Fig. 1d, e). This suggests that PTR amplifies the expression change in these tissues to a greater extent. To determine if this observation can be replicated in independent datasets, we compared the results from GTEx tissues with those from the human SRA tissue dataset. We found that mean delta fold changes between tissues in both datasets were highly correlated (Fig. 1f,  $R = 0.82$ ,  $P = 0.01$ , Pearson correlation).

In our earlier tissue comparisons, we noted that some RNA-seq data was generated from different regions from within the same organ, including brain, skin, heart, esophagus, colon, artery and adipose tissue. Tissues from within the same organ are generally expected to have higher biological similarity compared to tissues from different organs. To explore a possible association between expression change and biological similarity, we divided these comparisons into inter-organ and intra-organ groups, and analyzed whether they had similar delta fold change distributions. Our analysis revealed that inter-organ delta fold changes were significantly greater than intra-organ delta fold changes in GTEx tissues, SRA tissues and non-human tissue datasets (Fig. 1g,  $P = 5.12e-11$ ,  $P = 0.013$  and  $P = 6.99e-08$ , respectively, one-sided Wilcoxon rank sum test). This finding suggests a possible association between biological differences and the magnitude of expression change by PTR.

## Influence of PTR on disease gene expression dysregulation

After observing amplification of expression change by PTR in normal tissues, we sought to ascertain if PTR had the same effect in disease states. To answer this question, we investigated gene expression across a wide range of disease (or abnormal) conditions. To systematically investigate the behavior of PTR in these conditions and broaden the biological relevance of the study, we collected the bulk tissue RNA-seq data from 1118 tissue and cell culture samples from 28 previous disease studies (Table S4, available online at <http://bib.oxfordjournals.org/>). The samples were from 48 different non-homeostatic disease conditions, including various types of cancers, genetic diseases, infections, immune diseases, or from gene manipulation studies. We found that, in contrast to amplification of expression change observed in normal tissues, most disease conditions showed a reduction in the magnitude of expression change (Fig. 2,  $P = 3.05e-9$ , one-sample t-test). This suggests that PTR tends to reduce disease associated expression dysregulation in pre-RNAs.

## Investigating analytical robustness

To substantiate the observation that PTR amplifies tissue-specific expression but reduces disease expression change, and to rule out the likelihood of artifacts arising from defects or bias in the analysis methods, we performed comprehensive experiments to determine the impact of different factors on the analysis results.

First, to demonstrate that our findings were robust to the selection of differential expression analysis tools or algorithms, we used the tool edgeR, in addition to DESeq2, which was used to generate our previous results, for analysis. EdgeR uses a different algorithm and estimates fold change without shrinkage. This analysis yielded a result consistent with our original findings (Fig. Sa).

Second, considering that the fold change calculation for lowly expressed genes may be inaccurate and may lead to a disproportionately large magnitude change that bias the analyses, we applied different gene filtering criteria ( $\log_2\text{CPM} > 0$  and  $\log_2\text{CPM} > 5$ ) to include or exclude lowly expressed genes. Both datasets produced similar results (Fig. Sb), suggesting that lowly expressed genes did not significantly impact the overall observation.

Third, since our previous results were based on analysis using subsampled exonic reads for mature RNA abundance calculation, to investigate potential bias in the subsampling process, we also re-performed the analysis using all exonic reads, and found that it led to similar results (Fig. S3 available online at <http://bib.oxfordjournals.org/>, Fig. Sc).

Fourth, considering we used all reads mapped to any region of introns or exons for RNA abundance calculations, reads mapped to internal regions of exons could arise from pre-RNAs or mature RNAs and this may cause errors in read assignment. To mitigate the possible bias arising from these errors, we re-performed the analysis using only boundary reads. Those reads can be more unambiguously assigned to either pre-RNAs or mature RNAs. In this analysis, we used only intron-exon boundary reads to represent pre-RNAs, and only exon-exon boundary reads to represent mature RNAs. The analysis yielded results similar to our original results (Fig. S3, available online at <http://bib.oxfordjournals.org/>, Fig. Sd). We also compared the mean delta fold change values derived from subsampled data with those calculated from exon-intron and exon-exon boundary reads, and found they have strong correlation (Pearson  $r = 0.94$ , Fig. Sg).

Fifth, we investigated the possible impact of three prime bias in the sequencing data on the analysis results. This form of bias is characterized by greater sequencing coverage at the 3' end of genes compared to 5' end, caused by the poly-A selection method used for mRNA fragment enrichment in some studies. We performed the investigation by using only reads that mapped to either the 5' end or 3' end of the genes to estimate RNA abundance. This analysis produced similar results (Fig. S3, available online at <http://bib.oxfordjournals.org/>, Fig. Sd), indicating minimum impact of three prime bias on the results.

Sixth, considering the possible confounding factor of relatively larger sample sizes in each group for tissue comparisons, compared to disease condition comparisons, we repeated the GTEx tissue analysis by randomly selecting only five samples for each tissue. This analysis yielded a result consistent with our original observations, suggesting that sample size did not bias the general observation (Fig. Se).

Seventh, to address the potential bias introduced by mean values, which can be more easily influenced by outliers, we calculated the median delta fold changes for the differential expression magnitudes in each comparison. We found that median fold changes recapitulated the same phenomenon as mean fold changes (Fig. Sf).

Furthermore, to assess the potential impact of RNA integrity on the observed result, we checked the correlation between the mean delta fold changes and mean Transcript Integrity Number, a score measuring RNA integrity calculated from sequencing reads.

We found that the contribution of Transcript Integrity Number to mean delta fold changes was minor compared to the differences between tissues or diseases (coefficient of  $-0.0070$  versus  $0.38$  in linear regression model) (Fig. S4 available online at <http://bib.oxfordjournals.org/>).

Lastly, considering that the length of exons and introns may influence the physical accessibility of the PTR regulatory machinery, it may influence the effectiveness of PTR. To explore this possible association, we examined the correlation between mean delta fold change and mean/total length of exons/introns of genes, as well as total length of the gene, and found that there was only a minor correlation (Fig. S5, available online at <http://bib.oxfordjournals.org/>).

In summary, our findings remained consistent across various checks against potential analytical biases, underscoring that our conclusions were robust with respect to the applied analytical approach.

### Geneset and pathway analysis

Our observation that PTR amplifies expression differences between normal tissues but reduces expression change in disease states suggests an interesting dynamic. In homeostatic states, PTR may boost tissue-specific functions, while in non-homeostatic states (such as disease), it may help restore a homeostatic state. To further investigate which genes contribute to the observed differences in PTR consequences, we split the gene set into 2233 housekeeping genes, with the remaining 31,625 genes designated non-housekeeping genes (Methods). As housekeeping genes maintain the basal functions of the cells, their expression might be less influenced by PTR.

We first examined data from GTEx tissues and found that housekeeping genes had smaller delta fold changes compared to non-housekeeping genes (Fig. 3a). However, we also noticed that housekeeping and non-housekeeping genes had different abundance and fold change distributions (Fig. 3b, c), which may have interfered with the delta fold change results. Therefore, we subsampled non-housekeeping gene sequencing reads, such that they were of equivalent abundance and fold change distribution to non-housekeeping genes (Fig. 3e, f). After subsampling, housekeeping genes only had slightly smaller delta fold changes compared to non-housekeeping genes (Fig. 3d). The same analysis was performed on the SRA tissues, non-human tissues, and disease condition datasets. It was found that housekeeping genes and non-housekeeping genes have similar delta fold change distributions (Fig. 3g–i), suggesting PTR has similar effect on housekeeping and non-housekeeping genes.

Finally, we examined the enriched pathways and gene ontology for gene sets categorized by differential expression and delta fold change directions. They included categories of up-regulation and down-regulation, as well as categories of amplified or reduced expression change in normal tissues or disease conditions. We did not observe a clear indication that the enrichment was limited to specific pathways or gene ontology (Supplementary spreadsheet Table S5, available online at <http://bib.oxfordjournals.org/>). We also performed enrichment analysis on the affected genes categorized by tissue or disease and observed that some of them showed enrichment in tissue-specific or disease-specific functions (Tables S6–S10, available online at <http://bib.oxfordjournals.org/>). For instance, in brain tissues, the affected genes were enriched in tissue-specific pathways like Glutamatergic synapse, GABAergic synapse, Dopaminergic synapse, and Synaptic vesicle cycle. However, this observation was not consistent across tissues and diseases, suggesting that

PTR tends to maintain the homeostatic states of a wide range of cellular processes, and that more evidence may help to illustrate their implications in biological functions. To determine whether the genes with amplified or reduced expression change are consistent across different tissues and disease states, we analyzed the percentage of tissues or diseases in which each gene exhibited the expression change. We found that most of these genes appeared in less than 20% of tissues or diseases, indicating they are generally not consistent across different conditions. However, some genes, such as SLC7A2, JAG1, LGALS3, TPM2, VIM, and XBP1, consistently showed the expression change across multiple tissues (Fig. S6, available online at <http://bib.oxfordjournals.org/>, Tables S11–S14).

### Discussion

In this study, we examined the expression of pre-RNA and mature RNA across tissues in human and other mammalian species. We discovered that PTR tends to amplify normal (non-disease state) tissue-specific expression. We also explored the impact of PTR in a variety of diseases and found that it tends to reduce expression change in diseases. The attenuation effect of PTR on expression change in diseases aligns with the stabilization effect observed in protein translation regulation in previous studies [16–18]. Recent work by Sánchez-Escabias et al. [54] demonstrated that co-transcriptional splicing efficiency is a gene-specific feature influencing mature mRNA levels, and that it can be regulated by the TGF $\beta$  signaling pathway. While their study focuses on splicing kinetics within individual genes, our research examines the broader impact of PTR across different tissues and disease states. Together, these findings highlight the significant role of RNA processing in gene expression regulation at both the gene-specific and systemic levels.

Our thorough examination of analytical factors confirmed the robustness of these results against different analysis parameters and algorithms. Moreover, this phenomenon was consistently observed across different datasets and species, suggesting that the findings are unlikely to have arisen from analysis methodology bias, differences in RNA integrity or possible batch effect in the datasets. Opposing effects were observed in normal tissues compared to diseases. This suggests the observations were unlikely to have resulted from differences in sequencing read mappability between introns and exons, as those differences should affect both diseased and normal tissue analyses equally. Taken together, our study suggests that PTR tends to support the homeostatic expression of mature RNAs in normal tissues and in disease states. In our gene set enrichment analysis, we observed that genes affected by PTR showed some enrichment in tissue-specific or disease-specific pathways or gene ontologies; however, this pattern was not consistent across different tissues or diseases. Additional data from future studies may help to elucidate their role in biological functions.

The magnitude of expression differences of pre-RNA in disease states is larger compared to mature RNA. It might therefore be possible to identify new disease-associated dysregulated genes by examining intronic sequences in RNA-seq datasets. This finding also suggests that post-transcriptional regulators, such as non-coding RNAs and polyadenylation, may play roles different to those previously assumed in disease studies. It is possible that the increased expression of post-transcriptional regulators in disease states reduces abnormal gene expression, offering a protective rather than a pathogenic effect, contrary to what traditional association analysis suggests. This might account for

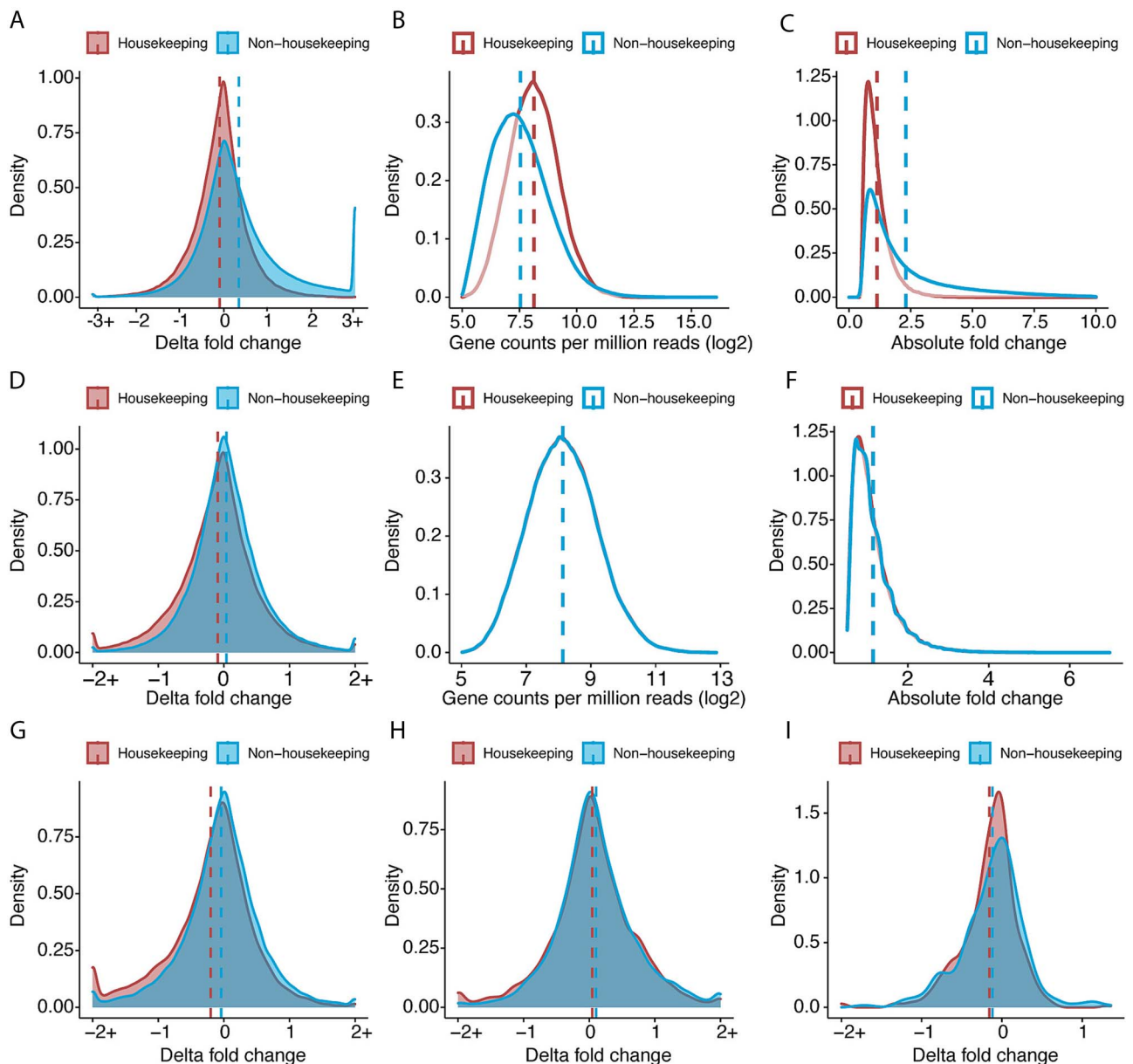


Figure 3. Distribution of delta fold changes in housekeeping and non-housekeeping genes. (a–c) Distribution of delta fold changes,  $\log_2$ CPM and absolute fold changes in housekeeping genes and all non-housekeeping genes in GTEx tissues. (d–f) Distribution of delta fold changes,  $\log_2$ CPM and absolute fold changes in housekeeping genes and subsampled non-housekeeping genes in GTEx tissues. (g–i) Distribution of delta fold changes in housekeeping genes and subsampled non-housekeeping genes in SRA tissues, non-human tissues and disease conditions. Dashed vertical lines represent mean values for each distribution.

the challenges in targeting these genes for therapeutic purposes and their limited effectiveness as disease diagnostic or prognostic biomarkers [55–58]. Our study also suggests the value of analyzing both pre-RNA and mature RNA levels. The fact that PTR seems to mitigate expression dysregulation in disease states implies that disease-associated gene expression dysregulation might be primarily caused by transcriptional regulation. Thus, targeting transcriptional regulation processes might be an effective strategy for restoring gene expression balance.

The opposing effects of PTR observed in diseases compared to normal tissues suggest that mature RNA abundance may be more flexible and can be more easily shaped by regulation. This effect could be due to the greater separation of PTR from cellular functions and thus being under less evolutionary pressure, compared to protein translation regulation. The differences in magnitudes

of expression change across tissues may be associated with the intensity and uniqueness of PTR amongst tissues. This is consistent with the observation that the brain shows seemingly outlying post-transcriptional characteristics compared to other tissues, as indicated by its highly complex lncRNA, miRNA expression profiles and splicing patterns [59–61].

Although in the study a variety of methods have been employed to minimize the impact of possible confounding factors, some may still have influenced the results. For instance, transcription and processing of pre-RNAs is a continuous process, so some RNAs are in intermediate states between pre-RNAs and mature RNAs. Moreover, intron retention events may also confound the analysis. Nevertheless, as the fraction of intermediate and intron retention-impacted RNAs is likely to be relatively small [62, 63], and both groups in each comparison are evenly affected by such



confounding factors, the general observations and conclusions of the study are unlikely to have been significantly affected by these issues. Moreover, intron retention could theoretically reduce the magnitude of the delta fold change by decreasing the observed expression difference between pre-RNA and mature RNA levels, potentially biasing the delta fold change towards zero. However, intron retention should not alter the direction of the delta fold change. Therefore, our key conclusions—that PTR amplifies expression changes (delta fold change >0) in normal tissues and reduces expression changes (delta fold change <0) in disease contexts, should remain valid despite the potential influence of intron retention. Additionally, although we characterized dozens of disease conditions, in the future, examination of more disease samples might provide deeper insight into PTR in different disease states, perhaps revealing that the magnitude of expression change attenuation varies in different types of disease states and in different disease-affected tissues.

Furthermore, another limitation of our study is that it focuses on humans and three other mammalian species. Future studies can explore whether the observed patterns of PTR influence are applicable to non-mammalian species or more diverse mammalian lineages. The impact of PTR may correlate with the evolutionary divergence time of species. Expanding the analysis to include species across a broader evolutionary spectrum could provide more insights into the relationship between PTR and evolution. Additionally, our study is limited by analyzing data at the gene level instead of the transcript level. Quantifying pre-RNA abundance at the transcript level is challenging with standard short-read RNA-seq data because few reads map to intronic regions, and assigning these reads to specific transcripts is difficult due to shared introns among transcripts. Future studies using long-read sequencing or higher coverage datasets could provide insights into transcript-level PTR and tissue-specific mRNA isoforms.

#### Key Points

- Demonstrates how post-transcriptional regulation (PTR) maintains mature RNA homeostasis.
- PTR amplifies normal tissue-specific gene expression but attenuates expression change in diseases.
- Intronic reads in RNA-seq may reveal novel disease-associated genes missed by mature RNA analysis.

## Supplementary Data

Supplementary data are available at *Briefings in Bioinformatics* online.

## Acknowledgements

We acknowledge Professor Daniel G. MacArthur and Professor Marc Wilkins for providing advice on the research plan. We thank Dr. Richard Edwards for helping with the computation resources and Keiran Rowell for assistance with data storage. We thank Zhixin Liu from Stats Central, Mark Wainwright Analytical Centre of UNSW for providing statistical consulting service. We also thank Dr. Scott Berry from Single Molecule Science at UNSW for providing advice on the research strategies. We acknowledge the UNSW Australia's Katana supercomputing facility for providing

the computational infrastructure. Z.S. acknowledges funding support from The UNSW (University Postgraduate Award scholarship).

## Author contributions

Z.S., E.O., M.D., and M.F. designed the research plan, E.O., M.D., and F.V. supervised the project and provided advice on the research strategies and analysis result. Z.S. and M.F. performed data analysis and wrote the manuscript. All authors provided advice and helped shape the analysis and manuscript.

Conflict of interest: F.V. declares commercial association with OmniOmics.AI Pty Ltd.

## Funding

This work was supported by The UNSW through the University Postgraduate Award scholarship.

## Data availability

Gene expression datasets quantifying change at the pre-RNA and mature RNA levels across various conditions, including normal tissues and diseases, are available for access at <https://zenodo.org/records/13119710>.

## Code availability

The source code used to perform analysis and generate the results in the study is available on GitHub: [https://github.com/suzheng/PTR\\_RNA\\_seq](https://github.com/suzheng/PTR_RNA_seq).

## References

1. MacGillivray AJ, Paul J, Threlfall G. Transcriptional regulation in eukaryotic cells. *Adv Cancer Res* 1972;**15**:93–162. [https://doi.org/10.1016/S0065-230X\(08\)60373-5](https://doi.org/10.1016/S0065-230X(08)60373-5).
2. Kouzarides T. Chromatin modifications and their function. *Cell* 2007;**128**:693–705. <https://doi.org/10.1016/j.cell.2007.02.005>.
3. Emilsson V, Thorleifsson G, Zhang B. et al. Genetics of gene expression and its effect on disease. *Nature* 2008;**452**:423–8. <https://doi.org/10.1038/nature06758>.
4. Gonzalez-Sandoval A, Gasser SM. On TADs and LADs: spatial control over gene expression. *Trends Genet* 2016;**32**:485–95. <https://doi.org/10.1016/j.tig.2016.05.004>.
5. Graveley BR. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* 2001;**17**:100–7. [https://doi.org/10.1016/S0168-9525\(00\)02176-4](https://doi.org/10.1016/S0168-9525(00)02176-4).
6. Bird JG, Zhang Y, Tian Y. et al. The mechanism of RNA 5' capping with NAD<sup>+</sup>, NADH and desphospho-CoA. *Nature* 2016;**535**:444–7. <https://doi.org/10.1038/nature18622>.
7. Colgan DF, Manley JL. Mechanism and regulation of mRNA polyadenylation. *Genes Dev* 1997;**11**:2755–66. <https://doi.org/10.1101/gad.11.21.2755>.
8. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;**116**:281–97. [https://doi.org/10.1016/S0092-8674\(04\)00045-5](https://doi.org/10.1016/S0092-8674(04)00045-5).
9. Tian B, Manley JL. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* 2017;**18**:18–30. <https://doi.org/10.1038/nrm.2016.116>.

10. Gaiti F, Calcino AD, Tanurdžić M. et al. Origin and evolution of the metazoan non-coding regulatory genome. *Dev Biol* 2017;**427**: 193–202. <https://doi.org/10.1016/j.ydbio.2016.11.013>.
11. Gaiti F, Fernandez-Valverde SL, Nakanishi N. et al. Dynamic and widespread lncRNA expression in a sponge and the origin of animal complexity. *Mol Biol Evol* 2015;**32**:2367–82. <https://doi.org/10.1093/molbev/msv117>.
12. Sebé-Pedrós A, Ballaré C, Parra-Acero H. et al. The dynamic regulatory genome of *Capsaspora* and the origin of animal multicellularity. *Cell* 2016;**165**:1224–37. <https://doi.org/10.1016/j.cell.2016.03.034>.
13. Gaiti F, Jindrich K, Fernandez-Valverde SL. et al. Landscape of histone modifications in a sponge reveals the origin of animal cis-regulatory complexity. *Elife* 2017;**6**:e22194. <https://doi.org/10.7554/eLife.22194>.
14. Chen L, Bush SJ, Tovar-Corona JM. et al. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. *Mol Biol Evol* 2014;**31**: 1402–13. <https://doi.org/10.1093/molbev/msu083>.
15. Kelemen O, Convertini P, Zhang Z. et al. Function of alternative splicing. *Gene* 2013;**514**:1–30. <https://doi.org/10.1016/j.gene.2012.07.083>.
16. El-Brolosy MA, Stainier DYR. Genetic compensation: a phenomenon in search of mechanisms. *PLoS Genet* 2017;**13**:e1006780–0. <https://doi.org/10.1371/journal.pgen.1006780>.
17. Ishikawa K, Makanae K, Iwasaki S. et al. Post-translational dosage compensation buffers genetic perturbations to stoichiometry of protein complexes. *PLoS Genet* 2017;**13**:e1006554. <https://doi.org/10.1371/journal.pgen.1006554>.
18. Sztal TE, Stainier DYR. Transcriptional adaptation: a mechanism underlying genetic robustness. *Development* 2020;**147**:147. <https://doi.org/10.1242/dev.186452>.
19. Khan Z, Ford MJ, Cusanovich DA. et al. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science (New York, NY)* 2013;**342**:1100–4. <https://doi.org/10.1126/science.1242379>.
20. Battle A, Khan Z, Wang SH. et al. Genomic variation. Impact of regulatory variation from RNA to protein. *Science (New York, NY)* 2015;**347**:664–7. <https://doi.org/10.1126/science.1260793>.
21. Zhang B, Wang J, Wang X. et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014;**513**:382–7. <https://doi.org/10.1038/nature13438>.
22. Consortium GT. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science (New York, N.Y.)* 2020;**369**:1318–30. <https://doi.org/10.1126/science.aaz1776>.
23. Leinonen R, Sugawara H, Shumway M. et al. The sequence read archive. *Nucleic Acids Res* 2010;**39**:D19–21. <https://doi.org/10.1093/nar/gkq1019>.
24. Collado-Torres L, Nellore A, Kammers K. et al. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol* 2017;**35**:319–21. <https://doi.org/10.1038/nbt.3838>.
25. Papatheodorou I, Fonseca NA, Keays M. et al. Expression atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res* 2018;**46**:D246–51. <https://doi.org/10.1093/nar/gkx1158>.
26. Andrews S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Cambridge, United Kingdom: Babraham Bioinformatics, Babraham Institute, 2010. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
27. DeLuca DS, Levin JZ, Sivachenko A. et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* 2012;**28**:1530–2. <https://doi.org/10.1093/bioinformatics/bts196>.
28. Frankish A, Diekhans M, Jungreis I. et al. GENCODE 2021. *Nucleic Acids Res* 2021;**49**:D916–23. <https://doi.org/10.1093/nar/gkaa1087>.
29. Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinform* 2013;**10**: 645–56. <https://doi.org/10.1109/TCBB.2013.68>.
30. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2. <https://doi.org/10.1093/bioinformatics/btq033>.
31. Howe KL, Achuthan P, Allen J. et al. Ensembl 2021. *Nucleic Acids Res* 2021;**49**:D884–91. <https://doi.org/10.1093/nar/gkaa942>.
32. Dobin A, Davis CA, Schlesinger F. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
33. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30. <https://doi.org/10.1093/bioinformatics/btt656>.
34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:1–21. <https://doi.org/10.1186/s13059-014-0550-8>.
35. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40. <https://doi.org/10.1093/bioinformatics/btp616>.
36. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* 2019;**35**:2084–92.
37. PVC (Research Infrastructure) US. Katana, UNSW, Sydney 2010.
38. Wang L, Nie J, Sicotte H. et al. Measure transcript integrity using RNA-seq data. *BMC Bioinformatics* 2016;**17**:58. <https://doi.org/10.1186/s12859-016-0922-z>.
39. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 2012;**28**:2184–5. <https://doi.org/10.1093/bioinformatics/bts356>.
40. Hounkpe BW, Chenou F, de Lima F. et al. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res* 2020;**49**:D947–55. <https://doi.org/10.1093/nar/gkaa609>.
41. Altenhoff AM, Glover NM, Train C-M. et al. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res* 2018;**46**:D477–85. <https://doi.org/10.1093/nar/gkx1019>.
42. Xie Z, Bailey A, Kuleshov MV. et al. Gene set knowledge discovery with enrichr. *Current protocols* 2021;**1**:e90. <https://doi.org/10.1002/cpz1.90>.
43. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30. <https://doi.org/10.1093/nar/28.1.27>.
44. Aleksander SA, Balhoff J, Carbon S. et al. The gene ontology knowledgebase in 2023. *Genetics* 2023;**224**:iyad031. <https://doi.org/10.1093/genetics/iyad031>.
45. Landrum MJ, Chitipiralla S, Brown GR. et al. ClinVar: improvements to accessing data. *Nucleic Acids Res* 2020;**48**:D835–d844. <https://doi.org/10.1093/nar/gkz972>.
46. Amberger JS, Bocchini CA, Schiettecatte F. et al. OMIM.org: online Mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* 2015;**43**:D789–98. <https://doi.org/10.1093/nar/gku1205>.

47. Sollis E, Mosaku A, Abid A. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* 2023;**51**:D977–d985. <https://doi.org/10.1093/nar/gkac1010>.
48. Wickham H. In: Gentleman R, Hornik K, Parmigiani G (eds.), *ggplot2: Elegant Graphics for Data Analysis*. Berlin, Germany: springer, 2016. <https://doi.org/10.1007/978-3-319-24277-4>.
49. Kassambara A, Kassambara MA. Package 'Ggpubr', R Package Version 0.12020. 6. Available online at <https://github.com/kassambara/ggpubr>.
50. Tang Y, Horikoshi M, Li W. Ggfortify: unified interface to visualize statistical results of popular R packages. *R J* 2016;**8**:474. <https://doi.org/10.32614/RJ-2016-060>.
51. Patil I. Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software* 2021;**6**:3167. <https://doi.org/10.21105/joss.03167>.
52. Gaidatzis D, Burger L, Florescu M. et al. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat Biotechnol* 2015;**33**:722–9. <https://doi.org/10.1038/nbt.3269>.
53. Ameur A, Zaghlool A, Halvardson J. et al. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol* 2011;**18**:1435–40. <https://doi.org/10.1038/nsmb.2143>.
54. Sánchez-Escabias E, Guerrero-Martínez JA, Reyes JC. Co-transcriptional splicing efficiency is a gene-specific feature that can be regulated by TGF $\beta$ . *Communications Biology* 2022;**5**:277. <https://doi.org/10.1038/s42003-022-03224-z>.
55. Bonneau E, Neveu B, Kostantin E. et al. How close are miRNAs from clinical practice? A perspective on the diagnostic and therapeutic market. *EJIFCC* 2019;**30**:114–27.
56. Backes C, Meese E, Keller A. Specific miRNA disease biomarkers in blood, serum and plasma: challenges and prospects. *Mol Diagn Ther* 2016;**20**:509–18. <https://doi.org/10.1007/s40291-016-0221-4>.
57. Lan H, Lu H, Wang X. et al. MicroRNAs as potential biomarkers in cancer: opportunities and challenges. *Biomed Res Int* 2015;**2015**:1–17. <https://doi.org/10.1155/2015/125094>.
58. Winkle M, El-Daly SM, Fabbri M. et al. Noncoding RNA therapeutics—challenges and potential solutions. *Nat Rev Drug Discov* 2021;**20**:629–51. <https://doi.org/10.1038/s41573-021-00219-z>.
59. Schaefer B, Sun W, Li YS. et al. The evolution of post-transcriptional regulation. *Wiley Interdisciplinary Reviews: RNA* 2018;**9**:e1485. <https://doi.org/10.1002/wrna.1485>.
60. Zimmer-Bensch G. Emerging roles of long non-coding RNAs as drivers of brain evolution. *Cells* 2019;**8**:1399. <https://doi.org/10.3390/cells8111399>.
61. Ludwig N, Leidinger P, Becker K. et al. Distribution of miRNA expression across human tissues. *Nucleic Acids Res* 2016;**44**:3865–77. <https://doi.org/10.1093/nar/gkw116>.
62. Middleton R, Gao D, Thomas A. et al. IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol* 2017;**18**:51. <https://doi.org/10.1186/s13059-017-1184-4>.
63. Braunschweig U, Barbosa-Morais NL, Pan Q. et al. Widespread intron retention in mammals functionally tunes transcripts. *Genome Res* 2014;**24**:1774–86. <https://doi.org/10.1101/gr.177790.114>.