Title

Expression and Splicing Mediate Distinct Biological Signals

Authors

Søren Helweg Dam[1], Lars Rønn Olsen[1], Kristoffer Vitting-Seerup[1,*]

*Corresponding Author

Affiliation

[1]Section for Bioinformatics, Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

ORCID

SHD: 0000-0003-0755-0016

LRO: 0000-0002-6725-7850

KVS: 0000-0002-6450-0608

Abstract

Through alternative splicing, most human genes produce multiple isoforms in a cell, tissue, and disease-specific manner. Numerous studies show that alternative splicing is essential for development, disease, and treatment. Recently, expression and splicing changes have been proposed as distinct mechanisms for regulating biological function, but no systematic analysis of this hypothesis exists. To facilitate paired analysis of expression and splicing changes, we developed pairedGSEA. We then used pairedGSEA to profile transcriptional changes in 100 representative RNA-seq datasets. Our systematic analysis demonstrates that splicing changes, on average, contribute to 48.1% of the biological signal in expression analyses. Gene-set enrichment analysis demonstrates that expression and splicing both convey shared and distinct biological signals. These findings establish alternative splicing as a major regulator of the human condition and suggest that most contemporary RNA-seq studies likely miss out on important biological findings. We anticipate our results will contribute to the transition from a gene-centric to an isoform-centric research paradigm.

Introduction

High-throughput sequencing of RNA (RNA-seq) has revolutionized our understanding of molecular biology. It has become the standard approach for unraveling the complexity of living organisms, as illustrated by tens of thousands of already published RNA-seq datasets[1,2]. The main use of RNA-seq is to compare samples in case/control settings by analyzing differential gene expression and differential gene splicing. Regardless of whether differential expression or differential splicing is analyzed, the result is that, between conditions, hundreds, if not thousands, of genes are found to change significantly. Most scientists use gene-set enrichment analysis (GSEA) to interpret these long gene lists. Hence GSEA has become a cornerstone within genomics because of its ability to extract biologically meaningful insights.

Differential splicing is caused by altering the RNA splicing of a gene. RNA splicing is an essential molecular mechanism that removes introns from precursor mRNA, joining exons to form the mature mRNA. Along with alternative transcription start and termination sites, alternative splicing enables genes to give rise to multiple transcripts (hereon jointly referred to as splicing). During the past decades, this has changed how we think of genes. A gene is no longer viewed as encoding just one transcribable product but instead as giving rise to several mRNA transcripts. It is estimated that 92–94% of human genes undergo alternative splicing[3] and that protein-coding genes produce, on average, seven transcripts[4]. However, analysis of long-read data suggests this is an understatement[5,6]. Adding further to the complicated picture, recent large-scale studies find that the splicing of the vast majority of human protein-coding genes is affected by naturally occurring genetic variation (SNPs)[7–9].

Alternative splicing often induces important functional alterations of the gene product. Alternative splicing is important for defining cell types[10], e.g., the difference between naïve and activated memory T cells, which are defined by the two CD45 isoforms, CD45RA and CD45RO, respectively[11]. The switch to the CD45RO isoform drastically increases the cell's T cell receptor-mediated activation[12]. Besides being important for defining cell types, alternative splicing is also central to cellular functions, such as apoptosis. Many apoptotic proteins, such as BLC-X[13] and caspase-2[14], exist in two opposing isoforms, one pro-apoptotic and one anti-apoptotic. As part of the apoptotic cascade, these genes switch from producing the anti-apoptotic to the pro-apoptotic isoform, thereby simultaneously removing the breaks and accelerating apoptosis[14].

Although most scientists analyze their RNA-seq data using differential gene expression followed by GSEA, alternative splicing is generally overlooked. Our literature analysis indicated that only 12% of articles that analyzed RNA-sequencing data in 2020 did any sub-gene level analysis (Figure S1).

However, an interesting pattern did emerge from the few studies analyzing both differential expression and splicing: The overlap between the differentially expressed and differentially spliced genes is rather low[5,9,15], suggesting that expression and splicing could mediate distinct biological signals. Such studies are, however, merely individual examples, and, to our knowledge, no systematic comparison of differential expression and differential splicing currently exists. In other words, we do not know how widespread or biologically relevant changes in splicing are.

Here, we set out to systematically analyze and compare differential expression and differential splicing. We chose to focus on bulk RNA-seq because it is currently the only widely available high-throughput approach where both the technology and analysis approaches are mature enough to analyze splicing. Across 100 representative RNA-seq datasets, we see that changes in splicing are pervasive and, compared to differential expression, mediate both shared, distinct and biologically meaningful signals. Our findings indicate that many scientists are underutilizing their RNA-seq data, thereby missing important biological insights.

Results

Enabling paired analysis of differential expression and splicing

To systematically compare differential splicing and differential expression, we assembled a robust analysis pipeline and collected 100 suitable datasets (Table S1). The analysis pipeline is implemented in the R package pairedGSEA (Figure 1A)(https://github.com/shdam/pairedGSEA). pairedGSEA uses state-of-the-art tools[16] to enable easy:

- Correction for both known and unknown unwanted experimental factors (e.g., batch effects) (via sva[17])
- Analysis of differential splicing and differential expression, ensuring that the results are comparable (via DEXSeq[18] and DESeq2[19], respectively)
- Analysis and comparison of the result of both expression and splicing for over-represented gene sets (via fgsea[20]).

The 100 analyzed datasets (see methods) represent what scientists typically publish, covering a wide range of study types. Examples include using different growth media, inhibition or overexpression experiments, and the effects of treating various human diseases with different drugs (Supplementary Table S1). We applied pairedGSEA to all 100 datasets resulting in 199 comparable analyses of differential expression and differential splicing (since some datasets had more than two

conditions)(Supplementary data). Interestingly, we found that every dataset had genes significantly affected by confounding factors (e.g., batch effects) (median 598 genes, Figure 1B). Indeed, if these confounders were not considered, the false discovery rate (FDR) would, on average, be ~16.7% instead of the expected 5% (Figure 1C). In other words, batch effects are pervasive, and, if not considered, the results of differential analysis will largely be unusable.

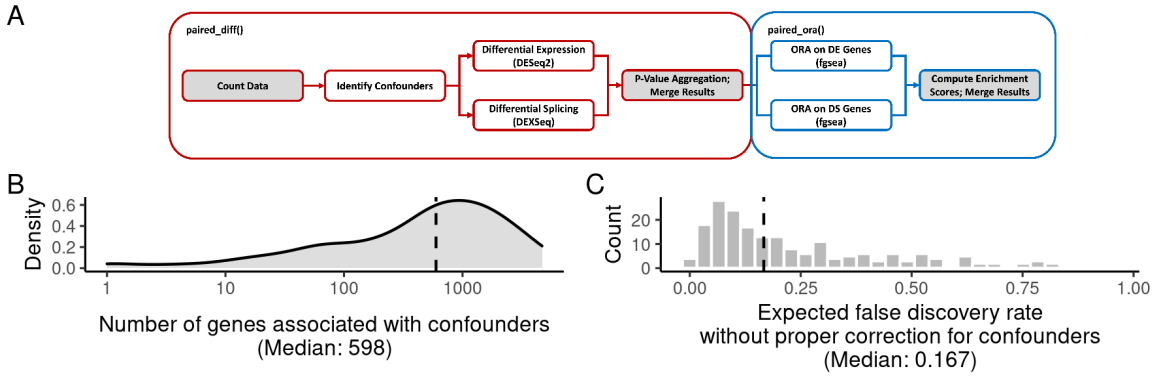

Figure 1. Enabling paired analysis of differential expression and splicing. **A**) Flowchart of the pairedGSEA R package and its functions (red and blue rounded squares). Data and functionality are indicated by gray and white boxes, respectively. **B**) The distribution of the number of significantly differentially expressed genes only found when not corrected for confounders (aka. False positives) across the 199 comparisons. C) Histogram of the false discovery rate when not correcting for confounders. Significance is defined as having an FDR-adjusted p-value of <0.05.

Differential Splicing is Omnipresent

To systematically compare differential splicing and differential expression, we extracted the significant results (FDR Corrected P-values < 0.05) from all 199 analyses. We find that, on average, 4327 (25.1% of tested) genes are significantly differentially expressed, 2247 (13.0% of tested) genes are significantly differentially spliced genes, and 1252 (7.2% of tested) genes are significantly affected by both expression and splicing (Figure 2A-B). Interestingly, this large overlap shows that, on average, 33.6% of differentially expressed genes are also significantly affected by alternative splicing (Figure 2C). Importantly, the splicing changes are not just lowly expressed isoforms; the changing isoforms contribute, on average, 45.6% of the parent gene expression (Figure 2D). Thus, more than 1 in 3 significantly differentially expressed genes contain splicing differences that could dramatically change the gene function through, e.g., dominant-negative splice variants, highlighting the need to consider splicing changes.

Across the 199 comparisons, our systematic analysis finds that 48.2% of the combined biological signal detected is at least partially mediated by changes in splicing (Figure 2E). It is, therefore, not surprising

that 93.0% of all multi-isoform genes are significantly differentially spliced in at least one dataset, indicating that almost all genes use changes in splicing during the adaptation to new circumstances.

Taken together, this indicates that changes in splicing are an integral part of the response pattern for all genes that is just as widespread as expression changes.

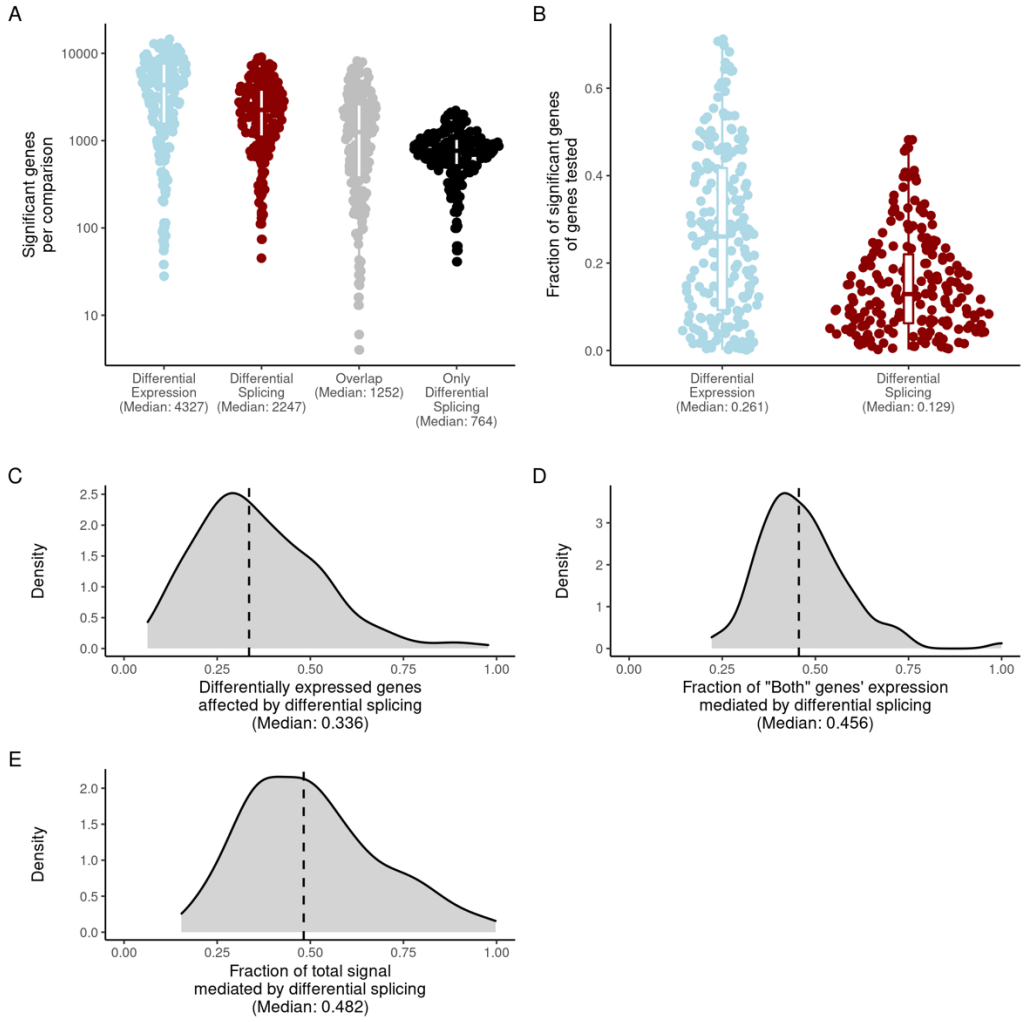

Figure 2. Differential Splicing is Omnipresent. **A**) The number of significant genes for each comparison across analyses. **B**) The fraction of genes tested that were deemed significant for each analysis. **C**) The fraction of differentially expressed genes that are also differentially spliced. **D)** For this analysis, we only considered genes that were both differentially expressed and spliced ("both" genes). For each gene in each condition, we calculated the fraction of the parent gene expression that originate from differentially spliced transcripts. The fraction was averaged across conditions, and for each condition we extracted the median fraction. **E**) For each comparison, we extracted the genes that were differentially expressed and/or differentially spliced (total signal). We then calculated the fraction of these genes that were differentially spliced. Across all panes significance is defined as having an FDR-adjusted p-value of <0.05. For all plots medians are indicated.

Splicing and Expression Regulate Distinct Biological Processes

Since genes are rarely independent functional entities, we next used GSEA to analyze gene sets (Gene ontology terms, KEGG pathways, etc.) affected by differential splicing and differential expression. This reduced the result of the differential expression from, on average, 1829 significant gene sets and the differential splicing result to, on average, 305 gene sets (Figure 3A). On average, ~230 gene sets overlap between analyses, indicating that many cellular circuits are regulated through both expression and splicing. The remaining ~70 identified gene sets were solely modulated through differential splicing, again suggesting that splicing mediates distinct information (Figure 3A).

To illustrate this point, we analyzed the study by Tian *et al*.[21], where human small airway epithelial is treated with TGFβ: a known regulator of telomere length[22,23]. Surprisingly, the gene-set enrichment analysis showed no significant associations between differentially expressed genes and telomere-related gene sets. However, the telomere-related gene sets were significantly enriched among the differentially spliced genes (Figure 3B), indicating that the regulation of telomeres by TGFβ is mediated through changes in splicing[24].

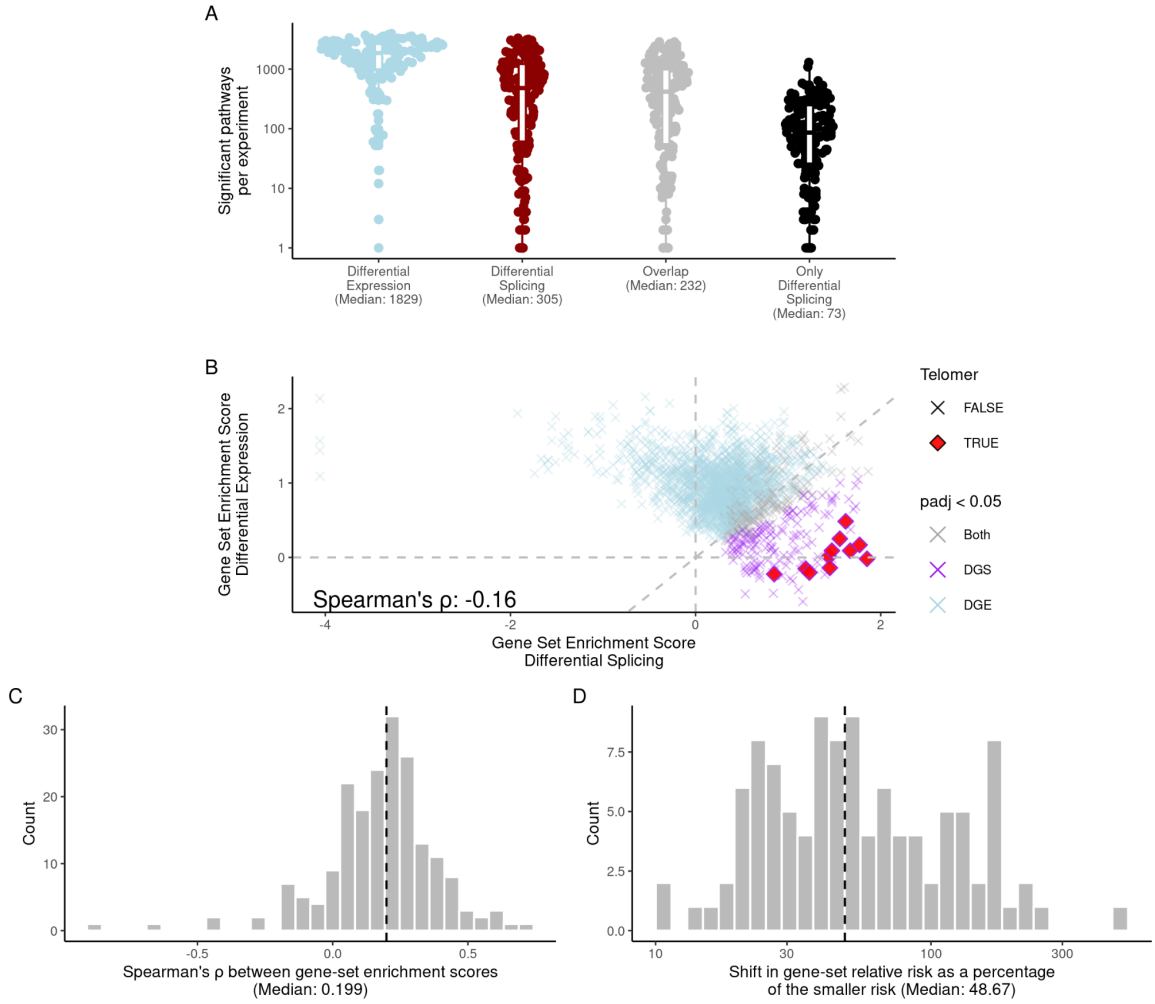

Figure 3. Splicing and Expression Regulate Distinct Biological Processes. **A**) The number of significant pathways found for each analysis and comparisons thereof. **B**) For each gene set (dot), the gene-set enrichment scores from the differential splicing (x-axis) and differential expression (y-axis) GSEA. Only gene sets that were significantly enriched among differentially spliced or differentially expressed genes (indicated by color) are shown. The shape highlights gene sets containing the word "telomer". The Spearman's correlation is indicated in the lower left corner. DGS = "Differential Gene Splicing", DGE = "Differential Gene Expression". **C**) Histogram of the Spearman's correlations between gene-set enrichment scores for gene sets significantly enriched among differentially expressed or differentially spliced genes. **D**) For each comparison, we calculated the median differences between the differential expression and differential splicing gene set enrichment scores as the percent change of the smallest enrichment score.

If a gene set is equally regulated by expression and splicing, we expect the gene-set enrichment scores to be very similar. Intriguingly, we find that the gene-set enrichment scores for the splicing and expression GSEA have a slightly negative correlation in the TGFβ study (Spearman's ρ of -0.16, Figure 3B). Inspired by this, we analyzed the correlation of gene-set enrichment scores for all significant gene sets within each of the 199 comparisons. We found that gene-set enrichment scores of splicing and expression changes were mostly uncorrelated (Median correlation: 0.2, Figure 3C). In addition, the gene-set enrichment scores were also substantially different (Figure 3D), suggesting little or no shared

regulation. These observations also hold true when considering gene sets only found via differential expression or differential splicing (Figure S2-3). Interestingly, the gene sets found significantly enriched among both differentially expressed and spliced genes seem to act differently. They had much smaller enrichment score differences and a Spearman's correlation of 0.72 (Figure S2-3), suggesting co-regulation of these pathways. Thus, our systematic analysis across 100 representative RNA-seq datasets indicates that differential expression and differential splicing act both independently and jointly to regulate biological function.

Discussion

In this study, we have used 100 representative RNA-seq datasets to show that differential splicing can significantly affect all multi-isoform genes. Almost half of the observed biological signal goes at least partly through splicing, indicating that changes in splicing are an integral part of biological responses. Since we estimate that 88% of recent articles about RNA-seq are not doing any sub-gene-level analysis, our results indicate a huge opportunity gap. Reanalyzing all these studies might host a bountiful range of missed discoveries important for both the fundamental and the translational sciences.

Our results are obtained through a systematic analysis of RNA-seq data. Therefore, some observed changes could exist without being translated to the protein level. Using bottom-up proteomics, many research groups have tried to determine how much splicing is reflected at the protein level, resulting in a lively debate on the topic. From the onset, this discussion is difficult because of the high sequence similarity of protein isoforms and low coverage of proteomics[25]. Furthermore, peptides generated by standard bottom-up proteomics are heavily depleted for peptides that cross exon-exon junctions, making it harder to detect splicing differences[26]. The result of the discussion is that there are multiple papers on either side of the argument respectively concluding that there is a lack of proteomics evidence for splicing[27,28] or ample proteomics evidence for splicing[29–31]. This disagreement is highlighted by commentary papers challenging articles on both sides of the argument[32,33]. The central point in this discussion seems to be how to produce trustworthy but not overly conservative results; unfortunately, no consensus seems to be on the horizon.

Until such consensus is reached, we believe it is worth looking at other data types to estimate the abundance of splicing in proteins. Firstly, it is worth noting that top-down proteomics, which does not suffer from the above-described shortcomings, easily identifies and distinguishes between protein isoforms[34,35]. Prominent examples are Tran *et al.*[36] and Yu *et al.*[37] who, respectively, identified isoforms for 22.44% and 46.5% of the proteins they identified (1043 and 628, respectively; our calculations, see

methods). Secondly, we have previously found that transcriptome-derived isoform switches are 16 times more frequent among lab-validated isoform switches than expected by chance (P-value < 3.4e-35, Fisher exact test)[38], indicating that transcriptional approaches do indeed capture validated isoform switches. Thirdly, there are a growing number of studies specifically designed to detect and analyze the impact of changes in protein isoforms. Large-scale studies of protein isoforms using technologies such as yeast two-hybrid have shown that the interaction partners for isoforms are very different from each other[39,40]. Furthermore, several research groups find good correspondence between transcriptional and protein changes when analyzing isoform-targeted proteomics and Ribo-seq[35,41–43]. Given this evidence, we find it highly likely that the majority of the transcriptional variation we described here will also be reflected in the proteome.

The analysis presented in this paper uses state-of-the-art bioinformatics tools. Even so, it is well established that differential expression encompasses more statistical power than differential splicing[16]. This means that, compared to differential splicing, we expect to find a larger fraction of genes that change expression simply because it is easier to test statistically. In other words, our estimates of the frequency and importance of splicing are probably underestimated.

Most systems biology relies on databases of annotation to make biological inferences. Almost all such databases, especially the most used ones, only contain gene-level annotation. Since we find splicing is both frequent and significant, our tool provides an essential way to extend the functionality of these databases. And, while some people have started to work on transcript-level databases[44–46], there is still much to do.

Our results show that splicing changes are omnipresent and biologically relevant. It is, therefore, clear that splicing should be considered when possible. It also highlights a massive caveat with technologies relying on only capturing the 5' or 3' end of transcripts. This is especially true for single-cell RNA-seq, where most datasets cannot be used to assess changes in splicing. In further support, the few single-cell datasets that can be used to analyze splicing show that splicing analysis leads to novel findings, including novel cell types[10]. Thus, it is clear that much more research into splicing is needed.

Our results indicate that the biological role of alternative splicing is *on par* with the importance of changes in gene expression, similar to what has previously been estimated from analysis of genetic data[9]. This has significant implications for most aspects of life sciences and equally affects wet labs doing mechanistic single-gene work and consortia investigating population-level genomics. Ultimately, utilizing the mostly untapped information hidden in differential splicing could pave the way for new clinical strategies within disease diagnosis, therapy, and precision medicine. For instance, detecting a switch in isoform usage in TP53 could indicate cancerous activity, cancer-specific isoforms could pose

as targets for novel immunotherapies[47], and expression of specific isoforms could increase the risk of relapse or even render some therapies ineffective[48]. All sciences must update how they work, moving from the current gene-centric research paradigm toward a more modern isoform-centric one.

Supplementary Figures

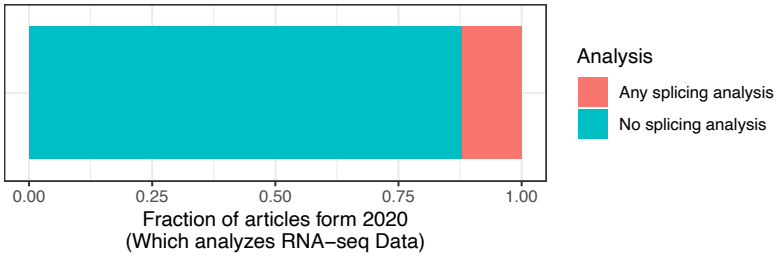

Figure S1: Summary of the manual literature review of articles from 2020 which analyzes RNA-seq data.

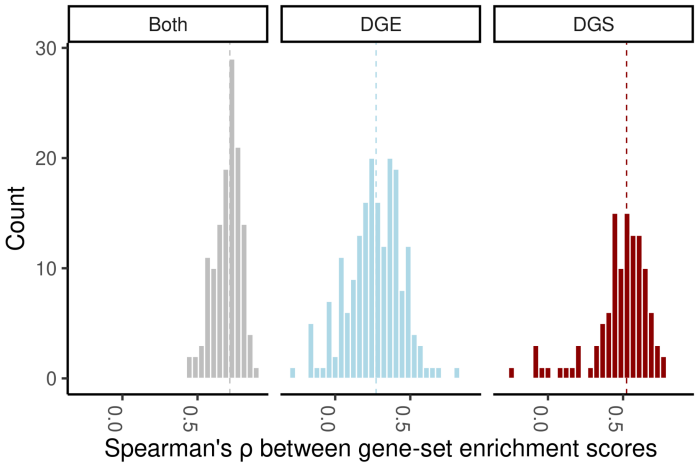

Figure S2: Histogram of the Spearman's correlations between gene-set enrichment scores for gene sets significantly enriched among differentially expressed or differentially spliced genes. Correlations are computed separately for gene sets found with either differentially spliced or expressed genes or both (sub-plots). Significance is defined as having an FDR-adjusted p-value of <0.05.

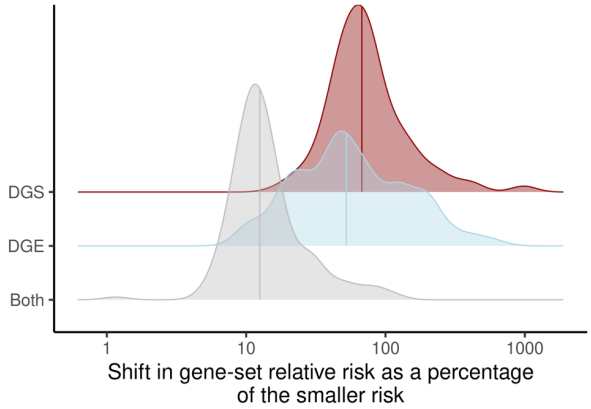

Figure S3: For each comparison and data subset (y-axis), we calculated the median differences between the differential expression and differential splicing gene set enrichment scores as the percent change of the smallest enrichment score (x-axis). Data subsets are defined as gene sets that were significantly enriched among differentially spliced genes, expressed genes, or both.

Methods

Systematic literature survey

For the systematic literature survey, we used the same approach as in Vitting-Seerup *et al.*[38], except this time we analyzed articles from the first 6 months of 2020. Briefly, we used PubMed to search for articles about RNA-seq and isoforms. We randomly selected 50 articles only found in the RNA-seq search and 50 articles also found in the isoform search. The, we manually profiled how the data was analyzed, including whether transcript quantification was done and whether any sub-gene level analysis was done. The results were extrapolated to the expected fractions of all articles about RNA-seq data as described in Vitting-Seerup *et al.*[38].

Curating representative datasets

To construct a collection of high-quality datasets, we considered >14,500 human RNA-seq studies[1]. From these, we computationally extracted high-quality datasets containing at least two conditions. High-quality datasets were defined as follows: 1) All samples within a dataset were predicted as bulk RNA-seq data with >90% certainty. 2) All samples having at least 40% of reads aligned to the human transcriptome. 3) All samples having at least 10 million aligned reads. 4) The study has a maximum of 50 samples (as DEXSeq do not scale well with increasing sample size). Next, we randomly selected 100 RNA-seq datasets and manually determined which groups to compare, extracting maximum 4 comparisons per dataset. The result was 199 comparisons from 100 datasets (Table S1).

The pairedGSEA R package

A comprehensive guide on the functionalities of pairedGSEA is available in the package vignette (https://github.com/shdam/pairedGSEA). The purpose of pairedGSEA is to make a baseline vs. case paired differential expression and splicing analysis simple. It assumes you have already preprocessed and aligned your sequencing data to obtain transcript level counts. Running pairedGSEA with default settings will detect potential batch effects (or other confounders) using sva[17]. If any such are detected, they will be added to the design matrix used for the differential analyses. pairedGSEA will then compute the differential expression and differential splicing on transcript level using DEXSeq[18] and DESeq2[19], respectively. DESeq2 is run using a likelihood-ratio test using a reduced model where the condition information is removed. The results are extracted with a baseline vs. case contrast. The model used in DEXSeq adds the interaction between the transcript counts and the condition as well as the confounding variables. DEXSeq does not allow a definition of a baseline as that is more abstract in differential splicing; however, pairedGSEA ensures the log fold changes correlate between the two analyses.
In the final step of the paired differential expression and splicing analysis, pairedGSEA aggregates the transcript p-values to gene level using Lancaster aggregation[49] with base means as weights. The differential expression log fold changes are aggregated using a weighted mean with base means as weights. Again, the log fold change of differential splicing is a bit more abstract; therefore, it was chosen to keep the log fold change of the transcript with the lowest p-value as the log fold change of the corresponding gene. Then, the p-values are adjusted, separately for the two analyses, by false discovery rate using Benjamini-Hochberg procedure, and the two results are merged into a single object. The results can then be directly used in the GSEA part of pairedGSEA, where the significant genes are extracted by a user-defined adjusted p-value cutoff (defaulted to 0.05), and two over-representation analyses are run using the fora function from the fgsea package[20], one for differentially expressed genes and one for differentially spliced genes. Genes found in both will be included in both over-representation analyses. But before doing so, a list of gene sets is needed. pairedGSEA provides a function to extract gene sets from MSigDB[50], but users may use any database they prefer. After running the over-representation analyses, pairedGSEA computes an enrichment score for each gene set as the log2 relative risk. Specifically, it is calculated as: Log2((overlap / gene_set_size) / (significant_genes / total_genes_analyzed) + 0.06).

Analysis of curated dataset

All data sets were obtained from a local copy of the ARCHS4 v11 database of transcript counts[1]. The Genome Reference Consortium Human Build 38 was used to retrieve transcript-to-gene associations, minimizing the number of lacking associations otherwise found in the ARCHS4 database. The baseline and case condition and experiment titles were retrieved from the manually curated metadata described in the Data selection section. Paired differential analysis was run with pairedGSEA using default settings. The inbuilt wrapper for MSigDB extraction was used to create a list of gene sets from the 'C5'-collection of Homo Sapiens gene sets, which were subsequently used in pairedGSEA's over-

representation analysis implementation. When considering enrichment score shifts/differences, calculations were done without the log2 transformation.

Confounders influence analysis

To evaluate the impact of accounting for confounders in the data, differential gene expression was recomputed for all 199 comparisons as described above without the initial step of searching for confounding variables.

Counting isoforms in top-down proteomics datasets

We obtained the isoform level supplementary data of Tran *et al.*[36] and Yu *et al.*[37]. For the Yu *et al*. data, we determined the number of isoforms for each protein by counting how many distinct proteoforms were annotated without considering post-translational modifications (PTMs). For the Tran *et al.* data, it is not possible to infer the exact number of isoforms with the information they provide. Instead, we counted the number of genes where isoforms were needed to explain the number of proteoforms annotated (called "species" in this data) given the number of PTMs annotated. Specifically, if the number of annotated proteoforms was larger than the number of possible combinations of PTMs (and no PTMs), we defined the protein as having more than one isoform. Note that the Tran *et al.* result is very conservative and less trustworthy than the Yu *et al*. estimate.

Data availability

All data is available as supplementary table or at Zendo. Specifically, for each of the 199 comparisons, a Rds data objects with the result of running pairedGSEA has been uploaded to Zenodo. pairedGSEA is available at GitHub (https://github.com/shdam/pairedGSEA).

Competing interests

The authors declare no competing interests.

Author contributions

KVS conceived the study. KVS did the literature survey on the previous use of RNA-seq. SHD and KVS curated the metadata of studies. SHD did all analyses. SHD, LRO, and KVS interpreted the results and wrote the manuscript.

References

1. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun* 9, 1366 (2018).

2. Wilks, C. *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing. doi:10.1101/2021.05.21.445138.

3. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476 (2008).

4. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res* 49, gkaa1087- (2020).

5. Glinos, D. A. *et al.* Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* 1–8 (2022) doi:10.1038/s41586-022-05035-y.

6. Hardwick, S. A. *et al.* Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. *Nat Biotechnol* 40, 1082–1092 (2022).

7. Consortium, T. Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330 (2020).

8. Qi, T. *et al.* Genetic control of RNA splicing and its distinct role in complex trait variation. *Nat Genet* 1–9 (2022) doi:10.1038/s41588-022-01154-4.

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

10. Booeshaghi, A. S. *et al.* Isoform cell-type specificity in the mouse primary motor cortex. *Nature* 598, 195–199 (2021).

11. Schaub, A. & Glasmacher, E. Splicing in immune cells—mechanistic insights and emerging topics. *Int Immunol* 29, 173–181 (2017).

12. Rheinländer, A., Schraven, B. & Bommhardt, U. CD45 in human physiology and clinical medicine. *Immunol Lett* 196, 22–32 (2018).

13. Stevens, M. & Oltean, S. Modulation of the Apoptosis Gene Bcl-x Function Through Alternative Splicing. *Frontiers Genetics* 10, 804 (2019).

14. Schwerk, C. & Schulze-Osthoff, K. Regulation of Apoptosis by Alternative Pre-mRNA Splicing. *Mol Cell* 19, 1–13 (2005).

15. Gilis, J., Vitting-Seerup, K., Berge, K. V. den & Clement, L. satuRn: Scalable analysis of differential transcript usage for bulk and single-cell RNA-sequencing applications. *F1000research* 10, 374 (2021).

16. Love, M. I., Soneson, C. & Patro, R. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000research* 7, 952 (2018).

17. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883 (2012).

18. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* 22, 2008–2017 (2012).

19. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014).

20. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *Biorxiv* 060012 (2021) doi:10.1101/060012.

21. Tian, B. *et al.* Analysis of the TGFβ-induced program in primary airway epithelial cells shows essential role of NF-κB/RelA signaling network in type II epithelial mesenchymal transition. *Bmc Genomics* 16, 529 (2015).

22. Cassar, L. *et al.* TGF-beta receptor mediated telomerase inhibition, telomere shortening and breast cancer cell senescence. *Protein Cell* 8, 39–54 (2017).

23. Ebeid, D. E. *et al.* Pim1 maintains telomere length in mouse cardiomyocytes by inhibiting TGFβ signalling. *Cardiovasc Res* 117, 201–211 (2020).

24. Cerezo, A., Kalthoff, H., Schuermann, M., Schäfer, B. & Boukamp, P. Dual regulation of telomerase activity through c-Myc-dependent inhibition and alternative splicing of hTERT. *J Cell Sci* 115, 1305–1312 (2002).

25. Yates, J. R., Ruse, C. I. & Nakorchevsky, A. Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. *Biomed Eng* 11, 49–79 (2009).

26. Wang, X. *et al.* Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity*. *Mol Cell Proteomics* 17, 422–430 (2018).

27. Tress, M. L., Abascal, F. & Valencia, A. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem Sci* 42, 98–110 (2017).

28. Fesenko, I. *et al.* Alternative splicing shapes transcriptome but not proteome diversity in Physcomitrella patens. *Sci Rep-uk* 7, 2698 (2017).

29. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* 509, 575–581 (2014).

30. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582–587 (2014).

31. Lau, E. *et al.* Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome. *Cell Reports* 29, 3751-3765.e5 (2019).

32. Ezkurdia, I., Vázquez, J., Valencia, A. & Tress, M. Analyzing the First Drafts of the Human Proteome. *J Proteome Res* 13, 3854–3855 (2014).

33. Blencowe, B. J. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends Biochem Sci* 42, 407–408 (2017).

34. Melani, R. D. *et al.* The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. *Science* 375, 411–418 (2022).

35. Chen, W. & Liu, X. Proteoform Identification by Combining RNA-Seq and Top-Down Mass Spectrometry. *J Proteome Res* 20, 261–269 (2021).

36. Tran, J. C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 480, 254–258 (2011).

37. Yu, D., Wang, Z., Cupp-Sutton, K. A., Liu, X. & Wu, S. Deep Intact Proteoform Characterization in Human Cell Lysate Using High-pH and Low-pH Reversed-Phase Liquid Chromatography. *J Am Soc Mass Spectr* 30, 2502–2513 (2019).

38. Vitting-Seerup, K. & Sandelin, A. The Landscape of Isoform Switches in Human Cancers. *Mol Cancer Res* 15, 1206–1220 (2017).

39. Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* 164, 805–817 (2016).

40. Ellis, J. D. *et al.* Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. *Mol Cell* 46, 884–892 (2012).

41. Han, Y. *et al.* Computation-assisted targeted proteomics of alternative splicing protein isoforms in the human heart. *J Mol Cell Cardiol* 154, 92–96 (2021).

42. Liu, Y. *et al.* Impact of Alternative Splicing on the Human Proteome. *Cell Reports* 20, 1229–1241 (2017).

43. Reixachs-Solé, M., Ruiz-Orera, J., Albà, M. M. & Eyras, E. Ribosome profiling at isoform level reveals evolutionary conserved impacts of differential splicing on the proteome. *Nat Commun* 11, 1768 (2020).

44. Karlebach, G. *et al.* An algorithmic framework for isoform-specific functional analysis. *Biorxiv* 2022.05.13.491897 (2022) doi:10.1101/2022.05.13.491897.

45. Louadi, Z. *et al.* DIGGER: exploring the functional role of alternative splicing in protein interactions. *Nucleic Acids Res* 49, gkaa768- (2020).

46. Ferrer-Bonsoms, J. A. *et al.* ISOGO: Functional annotation of protein-coding splice variants. *Sci Rep-uk* 10, 1069 (2020).

47. Barnkob, M. B., Vitting-Seerup, K. & Olsen, L. R. Target isoforms are an overlooked challenge and opportunity in chimeric antigen receptor cell therapy. *Immunotherapy Advances - In press* (2022).

48. Poulikakos, P. I. *et al.* RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). *Nature* 480, 387–390 (2011).

49. Yi, L., Pimentel, H., Bray, N. L. & Pachter, L. Gene-level differential analysis at transcript-level resolution. *Genome Biol* 19, 53 (2018).

50. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc National Acad Sci* 102, 15545–15550 (2005).