Systematic evaluation of isoform function in literature reports of alternative splicing

Authors: Shamsuddin A. Bhuiyan[1,2], Sophia Ly[1], Minh Phan[1], Brandon Huntington[1], Ellie Hogan[1], Chao Chun Liu[1], James Liu[1], Paul Pavlidis*[1,2]

[1] Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[2] Department of Psychiatry, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

*To whom correspondence should be addressed. Tel: (604) 827-4157 Email: paul@msl.ubc.ca

ABSTRACT

Though most genes in mammalian genomes have multiple isoforms, an ongoing debate is whether these isoforms are all functional as well as the extent to which they increase the functional repertoire of the genome. Here we present a literature-based analysis of experimental evidence for functionally distinct splice isoforms (FDSIs), that is, the gene has multiple splice isoforms necessary for the gene's overall function. We established a curation framework for evaluating experimental evidence of FDSIs, and analyzed over 700 human and mouse genes, biased towards genes that are prominent in the alternative splicing literature. We found experimental evidence for functionally distinct isoforms in fewer than 10% of the curated human and mouse genes. Furthermore, many of these isoforms were not referenceable to a specific isoform in the Ensembl, database that forms the basis for much computational research. Our results imply a lack of support from the experimental literature for claims that alternative splicing increases the functional diversity of the genome.

INTRODUCTION

There is currently a debate as to whether mammalian genes in general produce more than one functional isoform (1–3). The mere presence of multiple isoforms in public sequence databases is clearly insufficient to settle the question. Arguments against widespread functional alternative isoforms include the fact that the limited fidelity of the splicing machinery causes the stochastic generation of "junk" isoforms (4, 5). Analyses using proteomics and molecular evolution approaches have also failed to support the expression and the conservation of most splice isoforms (3, 6–9). Nevertheless, the question remains open because lack of evidence is not generally accepted as evidence of no function (1). Beyond the question of whether most genes have more than one functional isoform is a critical issue: whether these isoforms increase the functional repertoire of genes, or are merely functionally redundant. In this paper we take steps to address the gap between the commonplace assumption that most genes have more than one distinct functional product and the evidence-based reality.

Establishing whether a gene has functionally distinct isoforms requires experimental validation. While databases that contain information on transcript isoforms gather information on isoform features, none attempt to assess functionally distinct isoform reports from the experimental literature. For example, Ensembl, RefSeq, and Uniprot catalog and annotate splice isoforms based on evidence that they exist as a transcript or protein (10–12). However, the existence of a splice isoform alone does not provide direct support for its functionality, much less functional distinctness.

To establish the extent to which splice isoforms increase the functional repertoire of the genome, we need data on whether genes have *functionally distinct splice isoforms (FDSIs).* Identification of genes with FDSIs requires experimental support to demonstrate the necessity of each splice isoform. A classical method to determine the function of a given gene is to knock it out and observe the phenotypic

consequence (13, 14). This idea readily extends to isoforms; if a single isoform is made absent and that isoform is required for the normal function of the gene, then a consequence (phenotype) would be expected. A gene has FDSIs if two or more isoforms meet this criterion independently (Figure 1A). Note that a negative result from such an experiment is not evidence of a lack of functional distinctness, as it is possible the functional distinction between the isoforms may be eventually discovered. Still, it is of obvious importance to evaluate the state of the literature support for the commonplace claim that alternative splicing increases the functional repertoire of the genome.

Here we present a literature-based analysis of experimental evidence for functionally distinct splice isoforms (FDSIs) for over 700 human and mouse genes. Despite a gene selection strategy that was highly biased towards genes suggested to have multiple functional isoforms, we found experimental evidence for FDSIs for fewer than 10% of genes.

MATERIAL AND METHODS

Determining the type of functional distinctness

We developed a scheme to describe non-mutually exclusive types of functional distinctness found in genes with FDSIs. We recognize two general biological mechanisms by which functional distinctness could arise, schematized in Figure 1A, and elaborated on further below: “expression-pattern distinctness” or “biochemical distinctness”. Figure 1B outlines our full scheme for classifying functional distinctness. The subclasses we identified were designed to accommodate how functional distinctness is reported in the literature we curated, that is, we did not create this classification wholly *ab initio*. We determined the type of functional distinctness using the publication which provided the evidence for FDSI, but some cases required an inference based on other literature by the authors. We stress that a gene can have multiple types of functional distinctness. For example, biochemically distinct isoforms could also have expression pattern distinctness. We annotated as many types of functional distinctness as were provided by the literature reports.

Expression-pattern distinctness

Expression-pattern distinctness requires the condition-dependent expression of isoforms of a single gene. Generally, in this category, splice isoforms of the same gene have functional relevance in distinct conditions. We further specified expression-pattern distinctness as “subcellular-localization-specific”, “cell-type-specific”, “tissue-specific”, “developmental stage-specific”, and “other-condition-specific”. Thus genes with cell-type-specific FDSIs express their splice isoforms in distinct cell-types, and the elimination of expression of either splice isoform causes a phenotype in their respective cell-type (Figure 1A). Crucially, the splice isoforms’ final products could be identical (that is, they are not biochemically distinct). However,

they are still functionally distinct because they have at least partly non-overlapping expression patterns so one cannot fully compensate for the other.

Biochemical distinctness

Biochemical distinctness is defined as differences in biochemical properties or activities, and which cannot compensate for each other even if co-expressed in the same condition. We further specified biochemical distinctness as “protein domain change”, “dominant negative”, “subcellular localization”, and “protein terminus change”. Genes categorized as FDSIs with distinct protein-domains indicate that each splice isoform has a unique structural or functional unit in their final protein product. We manually extracted information about the specific protein domain from the authors providing the evidence of functional distinctness. In some cases, this could involve the presence or absence of one or more protein domains. Genes categorized as “protein terminus change” indicates that the FDSIs’ final protein product differ from each other either in their C-terminus or their N-terminus. These changes to the C- or N-termini usually do not affect the presence or absence of protein domains (or the paper did not make any note of changes to protein domains). Genes with dominant-negative FDSIs have splice isoforms with antagonistic phenotypes. Typically, these splice isoforms regulate each other’s function. The loss of one splice isoform generally affects the function of the other splice isoform.

Literature selection

On July 17th 2017, we generated a “starting set” of publications associated with human and mouse genes to curate using PubMed e-utilities and the search term “alternative splicing”. From here curation was both “gene-centric” and “paper-centric.”

Gene-centric curation

The gene-centric approach attempted to curate all relevant studies associated with a specific gene. PubMed linked each study from our starting set to a specific gene which provided a list of genes with literature. We selected the following 14 human genes and their mouse orthologues from to curate because the literature discussed them profusely or due to suggestions from the community: *AR, ANK3, ARL6, BDNF, BRCA1, CALCA, BCL2L11, MAPT, NRXN1, TP53, PTPRF, PTPRS, SRSF1* and *XBP1*.

Paper-centric curation

The paper-centric approach attempted to curate literature likely enriched for evidence of genes with FDSIs. Using this approach, we make no attempt to curate all relevant reports for any specific gene. As a targeted source of literature likely to be enriched for functional evidence, we used review articles on the function of alternative splicing that provided citations for 439 genes (3, 15–18). We further extended

paper-centric curation with specific search phrases in PubMed. Search terms were: “functionally distinct splice isoforms”, “CRISPR alternative splicing”, “alternative splicing knockdown” and “alternative splicing knockout.” These queries identified an additional 260 papers for our starting set of papers.

Curation process

For each paper, a trained curator first identified general features of that study by manually extracting the following information: the investigated gene (the gene with splice isoforms), the reported number of the splice isoforms for the gene, the names used by the authors for the splice isoforms, the number of splice isoforms specifically investigated in the paper (“the investigated isoforms”), the experiments performed, the organism where the gene was identified, the organism or cell line used for the experiments, and any claims of functional distinctness.

Next, using a decision tree (Figure 2), we annotated each paper as to whether the data provided positive evidence of functional distinctness for the investigated splice isoforms. We sought evidence where the loss of one isoform (via knockdown, knockout or other means of isoform-specific depletion) produced a phenotype in the test system. Many studies failed to provide evidence of functional distinctness. For example, overexpression of a single isoform may cause a phenotype, but these experiments have questionable relevance to native conditions (as an example, see study by Scotton and colleagues (19)). Similarly, rare mutations can result in aberrant isoforms that cause disease (e.g. cancer); these also lack relevance to the normal function of the gene as we have defined it (as an example, see study by Cogan and colleagues (20)). If any studies provided evidence where investigators depleted multiple splice isoforms of a single gene but only one splice isoform caused a phenotype, we classified the gene as having *negative* support for FDSIs.

For our definition of genes with FDSIs, we required evidence for the independent depletion of at least two splice isoforms of the same gene. If the curated study investigated the outcome of the absence of a single isoform for a given gene, then that study alone insufficiently provides evidence of FDSIs. While such studies demonstrate an existence of a single functional isoform, the support for FDSIs requires data on at least two isoforms from the same gene. However, we subsequently attempted to identify a second paper that provided functional evidence for a different splice isoform of the same gene. In situations where a second paper identified evidence of a different functional splice isoform, we recorded the gene as having FDSIs.

Linking FDSIs to Ensembl

If a paper provided positive evidence for FDSIs, we linked the splice isoforms with the appropriate Ensembl transcript ID. Generally, studies provided GenBank or RefSeq accession IDs and these accession IDs linked to Ensembl. In the absence of an accession ID, we referred to the literature for

sequence information about the splice isoforms and aligned splice isoform sequences to Ensembl using ClustalOmega (https://www.ebi.ac.uk/Tools/msa/clustalo/).

Curator validation

To ensure consistent curation, we evaluated the curators. These tests consisted of all curators curating the exact same randomly selected 50 papers. After the test, we addressed any discrepancy between curators and we updated the curation standards with any necessary clarifications (curation standards provided in File S1). This evaluation process was conducted three times. We also further scrutinized papers annotated as providing positive evidence of a gene with FDSI to eliminate any false positives.

RESULTS

Landscape of the alternative splicing literature

To generate a starting set of papers to curate, we queried PubMed on August 2017 using the term "alternative splicing". We found 19,049 human studies and 8,197 mouse studies representing 12,891 human genes and 7,585 mouse genes. While the median number of papers per gene was one, there was a large variance (Figure S1 and Figure S2). Most human genes (7,738) had only one such paper associated with them, while some have up to 100 (for example, SRSF1). We also observed that genes with many "alternative splicing"-mentioning papers tend to have many papers in PubMed overall (Spearman's rank correlation = 0.55, Figure S3). For example, we identified 86 studies linked to human *TP53* with the term "alternative splicing" (rank 2), but this is not particularly remarkable because overall, PubMed contains 8,261 studies linked to TP53 – the most studies for any single gene. This suggests, unsurprisingly, that heavily studied genes tend to have more research done on their splicing.

Curation Summary

We manually curated primary studies which provide evidence for the function of splice isoforms. As described in Methods, we selected genes and publications for curation in a manner that we expected should enrich for documentation of functional distinctness – for example, using review articles on splicing function. The curation process primarily focused on determining whether the elimination of expression of each splice isoform from a single gene caused an observable phenotype. Table 1 provides a summary of the knowledgebase as of January 20th, 2018 and File S2 contains a list of all curated studies. In total, we curated 1,031 human and mouse studies. This encompasses 829 human studies (528 genes) and 246 mouse studies (206 genes). On average, we have curated a median of 1 study per curated human gene and 1 study per curated mouse gene (mean = 1.5 studies and 1.2 studies, respectively). Our curation evaluations (see Methods) revealed that the curators agree on the interpretation of a paper 98% of the time. Errors were generally false positives for functional distinctness, which we addressed in the final review (see Methods).

Identification of twenty human genes with direct evidence of functionally distinct splice isoforms

By definition, a gene with functionally distinct splice isoforms (FDSIs) has at least two splice isoforms necessary for the gene's normal function. We find that genes with such evidence are rare: about 3% of curated human genes (9% of mouse genes) have FDSIs, based on reports in a total of 64 studies out of 1,031 studies (note that 128 studies depleted only one splice isoform of a gene). We provided the full list of the 20 human genes and the 19 mouse genes with FDSIs in Table 2 with additional information in File S3. RNAi knockdown experiments provided support for over 75% of these FDSIs, while the remaining FDSIs were characterized using gene knockouts combined with isoform-specific rescue.

As mentioned, we biased our gene and paper selection in such a way that our estimate of 3% (9% for mouse) might be too high. To help clarify this issue, we randomly selected 100 human genes (from those that had at least one alternative splicing related paper) for gene-centric curation (listed in Table S1). Of these 100 genes, only two genes (PML and DPF3, 2%, of the curated genes) had experimental evidence of FDSIs.

Genes tend to express functionally distinct splice isoforms in the same condition

To further explore "functional distinctness" in splicing, we identified non-mutually exclusive types of functional distinctness between FDSIs of the same gene, summarized in Table 5. We classified two main types of distinctness, expression-pattern distinctness and biochemical distinctness. Genes with expression-pattern distinct FDSIs have splice isoforms necessary for specific conditions while genes with biochemical distinctness have FDSIs with distinct biochemical properties that cannot compensate for each other even when co-expressed (for further description see Methods and Figure 1).

The majority of genes (25/39) have biochemically distinct FDSIs, rather than expression-pattern distinct FDSIs. We identified "dominant-negative" as the most common subtype of biochemically distinct FDSIs (11/39 genes). For example, the mouse gene *Enc1* has two FDSIs, named "57 kDa" and "67 kDa" by the authors, interacting in the Wnt-signalling pathway (21). Knockdown of 57 kDa promoted osteoblast mineralization while the knockdown of 67 kDa inhibited osteoblast mineralization.

In contrast to biochemical distinctness, we identified fewer cases of genes with expression-pattern distinct FDSIs. Only a total of 16 human and mouse genes had FDSIs in which the distinctness arises from distinct expression patterns. For example, the mouse gene *Myh10* has two FDSIs, named B1 and B2 by the authors (22). Cells in the brainstem express B1 to promote normal migration of facial neurons, while cells in the cerebellum expressed B2 to promote normal cerebellar Purkinje cell development.

Challenges linking FDSIs to sequence databases

We attempted to link all identified FDSIs back to Ensembl transcript identifiers, and were successful in 25/39 cases. Our process was as follows. First, in the studies for six genes, investigators provided a GenBank or RefSeq ID. We were able to map three of these to Ensembl (which includes GenBank and RefSeq data), but not for the other three (for more details see File S3), accounting for three of the 14 failures. Next, for 33 genes where no accession was provided, we used sequence alignment or other information to identify likely matches (See Methods). This was successful in 22 cases. In a further 9 cases, we were able to determine a sequence by referring to other papers by the same authors. Despite extensive efforts, we were unable to find matching Ensembl transcripts for the isoforms of 14 genes. This situation was not specific to Ensembl as we failed to link isoforms of 8 genes to UniProt; see File S3.

DISCUSSION

This paper represents progress towards documenting and evaluating the breadth of evidence for functionally distinct splice isoforms (FDSIs) for human and mouse genes. The inspiration for our study was strong arguments against the likelihood of most genes having multiple functional isoforms, contrasted with the ubiquitous claim that splicing vastly increases the functional repertoire of the genome (3, 23). This led us to ask where this latter claim comes from: while surely there are interesting cases of multi-isoform genes, has this been optimistically extrapolated to the entire genome? Our analysis suggests this is the case. While we were not surprised that there is no evidence of FDSIs for most genes, we were surprised by the low fraction for which there is supporting data, a mere 3% in human genes and 9% in mouse genes. Regardless of whether this number hold true with more curated studies, by contributing a list of genes with documented functionally distinct isoforms, we start to identify the scope of the gaps, the parameters for future experimental work, and assist computational methods that require training examples.

The low fraction of genes surveyed for which we found evidence of FDSIs (~3-9%) agrees with the general sense that we still have limited concrete evidence of more than one functional splice isoform per gene (3, 8, 15). Furthermore, we only considered genes for which some literature exists for their isoforms, so the range 3% to 9% is relative to genes that have at least one publication about them associated with splicing, but not all genes. Based on our query of PubMed using e-utilities, we estimate that one-third of human protein-coding genes have never had any type of experimental study of differences among their isoforms (including specific studies of expression patterns).

One might question whether the fraction 3% will rise substantially as we continue our curation efforts, but we hypothesize a lower true fraction. First, we aimed the gene-centric aspect of our curation at genes mentioned in review articles or otherwise prominent genes, and thus should be biased towards genes

with experimentally-backed function, yielding an over-estimate. Second, the gene-centric survey of 100 randomly-selected human genes yielded only two genes with evidence of FDSIs. Third, we retrieved a median of only one study per gene from PubMed. Since the genes with FDSIs tended to be genes with relatively more associated studies (Table 2), genes with few associated studies seem less likely to yield existing positive evidence for functional distinctness. Fourth, investigators face technical and/or resource challenges when testing the functional distinctness of isoforms, requiring either the ability to conduct isoform-specific depletion experiments, or isoform-specific rescue following a complete gene knockout. Reasonably, one might suppose that in many cases the experiments have not been done. The essential problem remains that most genes simply have not had their isoforms tested in such a way as to establish distinct functions.

We also sought *negative* evidence of genes having FDSIs from experiments where the depletion of only one splice isoform causes a phenotype while the depletion of the remaining splice isoforms of the same gene causes no phenotype. However, we only identified eight human genes and eight mouse genes from 16 studies with this type of evidence in our current curation of 1,075 studies (Table 4). Since most studies consider only one type of functional assay, it remains possible that tests of different functions would yield positive results for these genes. Nevertheless, the “file-drawer effect” – a type of publication bias towards only positive results – potentially plays a role in the dearth of negative evidence (24).

Evaluating the evidence for FDSIs at the gene level

After the curation of over 1,000 alternative splicing studies, we identified 20 human genes and 19 mouse genes with evidence for functionally distinct splice isoforms, mostly determined by RNAi knockdown experiments. RNAi knockdowns naturally align with our definition of a functional splice isoform and how researchers traditionally determine function in molecular biology. One question that arises in discussing RNAi is target specificity and efficacy. In most, but not all, of the papers we curated as having FDSIs, the authors demonstrate the target specificity of their siRNA to effectively deplete a single isoform. We raise this as a reminder that reports of evidence for functional distinctness may vary in quality.

Isoform-specific rescues demonstrating functional distinctness provide an alternative option to knockdown studies but the method has limitations when determining whether the splice isoforms rescue distinct phenotypes. In some studies, splice isoforms of the same gene clearly rescued distinct functions. For example, Candi and colleagues performed rescue experiments on Tp63-null mice (25). The knockout of Tp63 impeded the development of skin. In the rescue experiments, the splice isoform ΔNp63 restored the skin’s basal layer while the TAp63 restored the skin’s upper layers. In contrast, other studies rescued the same phenotype with each splice isoform, which makes evidence of functional distinctness unclear. For example, in the investigation by Coldwell and colleagues, each splice isoform of *EIF4G1* (eIF4G1e and eIF4G1f) rescued the phenotype of translation by restoring the translation rate (26). However, it remains

unclear whether this constitutes as evidence of functional distinctness. Since both splice isoforms rescued the same phenotype, they appear functionally redundant. Nevertheless, in cases such as these, we accepted the claim of the authors that the has FDSIs.

Some studies we curated only used overexpression of specific isoforms (293 studies) or examined expression patterns of isoforms (209 studies). Such experiments are important and can provide evidence consistent with functional distinctness, but they do not meet our criteria. We draw a parallel to the standards of evidence for characterizing gene function, in which evaluation of a loss of function is the gold standard (27). We argue that the same criteria used to establish gene function must be applied to isoforms.

Types of functional distinctness in FDSIs

It has been speculated that many poorly-characterized isoforms may have function because genes express splice isoforms in specific conditions, perhaps yet to be studied (1, 28, 29). It is therefore relevant that the minority (16) genes had functional distinctness due to condition-specificity. This may simply be due to a lack of study of condition-specific studies, as it might be generally easier to study isoforms expressed in the same conditions. Our results thus point to a potential gap in the literature.

Disconnect between the literature and gene databases

In one-third of the genes with FDSIs, the isoforms studied in a paper could not be matched to transcripts in Ensembl (as mentioned, not an Ensembl specific problem; 20% of genes with isoforms could not be matched to Uniprot). Conversely, Ensembl contains many transcripts that the literature ignores. This observation has fairly serious implications for basing splice isoform research on the contents of Ensembl (or related databases). If one developed experiments to functionally test the splice isoforms of the genes we identified to have FDSIs based on Ensembl transcripts for that gene, their experiments would not contain the correct FDSIs in at least one-third of the genes. In bioinformatics research, computational methods that make predictions based on Ensembl transcripts might be valueless to experimental biologists as Ensembl does not reflect the literature. Large-scale databases specific for alternative splicing, such as the Alternative Splicing encyclopedia (ASpedia) and the APPRIS database, tend to anchor to Ensembl (30, 31). Of note, previous discussions used APPRIS to understand the functional impact of alternative splicing (3). The disconnect between Ensembl the literature also impacts datasets not specific to splicing but where splice isoform information is important. For example, the GTEx consortium provides transcript-level quantification based on the Ensembl transcriptome (32). The FDSIs that are not in Ensembl are therefore not included in GTEx. Given the few known cases of genes with FDSIs, it remains crucial that computational resources contain FDSIs and experimentalists ensure that they submit their sequence data to these resources.

Conclusions

To our knowledge, this report represents the first effort to curate the literature on splicing in order to determine the genes where splicing increases the genome's functional potential. Such individual reports have been generally ignored in the debate about the function of alternative splicing, which has instead focused on databases and high-throughput data sets. Our estimate that only 3% of human and 9% of mouse genes have evidence for functionally distinct isoforms serves both a sobering reminder of the limited evidence, and a motivation for increased experimental efforts to settle the debate. At the same time, we also recognize there are likely genes with FDSIs that we did not curate and should be included in our database. We invite contributions from the community of suggestions for papers or genes to curate, which can be sent to the authors.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors would like to thank and acknowledge Dr. Sanja Rogic, Dr. Lilah Toker, Dr. Shreejoy Tripathy and Benjamin Callaghan for their useful feedback throughout the project.

FUNDING

This work was supported by a Natural Sciences and Engineering Research Council (http://www.nserc-crsng.gc.ca/) CREATE scholarship for the University of British Columbia Bioinformatics Training Program (awarded to SB), and Natural Sciences and Engineering Research Council Discovery grant (RGPIN-2016-05991) and National Institutes of Health (www.nih.gov) grant MH111099 to PP. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Blencowe,B.J. (2017) The Relationship between Alternative Splicing and Proteomic Complexity. *Trends Biochem. Sci.*, **0**.

2. Tress,M.L., Abascal,F. and Valencia,A. (2017) Most Alternative Isoforms Are Not Functionally Important. *Trends Biochem. Sci.*, 10.1016/j.tibs.2017.04.002.

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

4. Hsu,S.-N. and Hertel,K.J. (2009) Spliceosomes walk the line: splicing errors and their impact on cellular function. *RNA Biol.*, **6**, 526.

5. Melamud,E. and Moult,J. (2009) Stochastic noise in splicing machinery. *Nucleic Acids Res.*, **37**, 4873–4886.

6. Abascal,F., Tress,M.L. and Valencia,A. (2015) Alternative splicing and co-option of transposable elements: the case of TMPO/LAP2α and ZNF451 in mammals. *Bioinformatics*, **31**, 2257–2261.

7. Pickrell,J.K., Pai,A.A., Gilad,Y. and Pritchard,J.K. (2010) Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLOS Genet*, **6**, e1001236.

8. Reyes,A., Anders,S., Weatheritt,R.J., Gibson,T.J., Steinmetz,L.M. and Huber,W. (2013) Drift and conservation of differential exon usage across tissues in primate species. *Proc. Natl. Acad. Sci.*, **110**, 15377–15382.

9. Saudemont,B., Popa,A., Parmley,J.L., Rocher,V., Blugeon,C., Necsulea,A., Meyer,E. and Duret,L. (2017) The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biol.*, **18**, 208.

10. Aken,B.L., Ayling,S., Barrell,D., Clarke,L., Curwen,V., Fairley,S., Banet,J.F., Billis,K., Girón,C.G., Hourlier,T., *et al.* (2016) The Ensembl gene annotation system. *Database*, **2016**, baw093.

11. Bely,B., Martin,M.J. and Apweiler,R. (2010) Source of annotations in the UniProt Knowledgebase. *F1000Posters*, **1**.

12. Zhao,S. and Zhang,B. (2015) A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, **16**, 97.

13. Alberts,B., Johnson,A., Lewis,J., Raff,M., Roberts,K. and Walter,P. (2002) Studying Gene Expression and Function.

14. Shehu,A., Barbará,D. and Molloy,K. (2016) A Survey of Computational Methods for Protein Function Prediction. In Wong,K.-C. (ed), *Big Data Analytics in Genomics*. Springer International Publishing, pp. 225–298.

15. Kelemen,O., Convertini,P., Zhang,Z., Wen,Y., Shen,M., Falaleeva,M. and Stamm,S. (2013) Function of alternative splicing. *Gene*, **514**, 1–30.

16. Kovacs,E., Tompa,P., Liliom,K. and Kalmar,L. (2010) Dual coding in alternative reading frames correlates with intrinsic protein disorder. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 5429–5434.

17. Ramanouskaya,T.V. and Grinev,V.V. (2017) The determinants of alternative RNA splicing in human cells. *Mol. Genet. Genomics*, 10.1007/s00438-017-1350-0.

18. Stamm,S., Ben-Ari,S., Rafalska,I., Tang,Y., Zhang,Z., Toiber,D., Thanaraj,T.A. and Soreq,H. (2005) Function of alternative splicing. *Gene*, **344**, 1–20.

19. Scotton,P., Bleckmann,D., Stebler,M., Sciandra,F., Brancaccio,A., Meier,T., Stetefeld,J. and Ruegg,M.A. (2006) Activation of muscle-specific receptor tyrosine kinase and binding to dystroglycan are regulated by alternative mRNA splicing of agrin. *J. Biol. Chem.*, **281**, 36835–36845.

20. Cogan,J., Austin,E., Hedges,L., Womack,B., West,J., Loyd,J. and Hamid,R. (2012) Role of BMPR2 alternative splicing in heritable pulmonary arterial hypertension penetrance. *Circulation*, **126**, 1907–1916.

21. Worton,L.E., Shi,Y.-C., Smith,E.J., Barry,S.C., Gonda,T.J., Whitehead,J.P. and Gardiner,E.M. (2017) Ectodermal-Neural Cortex 1 Isoforms Have Contrasting Effects on MC3T3-E1 Osteoblast Mineralization and Gene Expression. *J. Cell. Biochem.*, **118**, 2141–2150.

22. Ma,X., Kawamoto,S., Uribe,J. and Adelstein,R.S. (2006) Function of the neuron-specific alternatively spliced isoforms of nonmuscle myosin II-B during mouse brain development. *Mol. Biol. Cell*, **17**, 2138–2149.

23. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

24. Kennedy,D. (2004) The old file-drawer problem. *Science*, **305**, 451.

25. Candi,E., Rufini,A., Terrinoni,A., Dinsdale,D., Ranalli,M., Paradisi,A., De Laurenzi,V., Spagnoli,L.G., Catani,M.V., Ramadan,S., *et al.* (2006) Differential roles of p63 isoforms in epidermal development: selective genetic complementation in p63 null mice. *Cell Death Differ.*, **13**, 1037–1047.

26. Coldwell,M.J., Sack,U., Cowan,J.L., Barrett,R.M., Vlasak,M., Sivakumaran,K. and Morley,S.J. (2012) Multiple isoforms of the translation initiation factor eIF4GII are generated via use of alternative promoters, splice sites and a non-canonical initiation codon. *Biochem. J.*, **448**, 1–11.

27. Kopp,F. and Mendell,J.T. (2018) Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell*, **172**, 393–407.

28. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.

29. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

30. Hyung,D., Kim,J., Cho,S.Y. and Park,C. (2018) ASpedia: a comprehensive encyclopedia of human alternative splicing. *Nucleic Acids Res.*, **46**, D58–D63.

31. Rodriguez,J.M., Rodriguez-Rivas,J., Di Domenico,T., Vázquez,J., Valencia,A. and Tress,M.L. (2018) APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res.*, 10.1093/nar/gkx997.

32. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N., *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

.

TABLE AND FIGURES LEGEND:

Species	Curated genes	Genes with FDSIs	Studies curated	Study Type			
				Isoform Removal	Overexpression	Localization	Other study types
Human	528	20	829	132	271	83	343
Mouse	206	19	246	77	61	37	65
Total	734	39	1,031	209	332	120	408

Table 1: Curation of alternative splicing literature has reveals 20 human genes and 19 mouse genes with functionally distinct splice isoforms (FDSIs). The 20 human genes with FDSIs accounted for almost 3% of human genes annotated in this knowledgebase, while the 19 mouse genes accounted for 9% of the all mouse genes annotated. The majority of curated studies could be classified into three different types: “isoform removal”, “overexpression” and “localization”. Isoform removal studies have experiments where expression of at least one splice isoform is eliminated and a phenotypic change is evaluated. Overexpression studies have experiments where at least one splice isoform is overexpressed. This “abundance” of the splice isoform can cause a phenotype. Localization studies have experiments that characterize where in the cell or organism the splice isoform is expressed. A single study can report experiments with multiple study types. The total number of human and mouse studies curated do not sum to 1,075 studies because some publications investigated both human and mouse forms of a single gene.

	Gene	Number of FDSIs	Number of Ensembl Transcripts	Number of Studies	Mappable to Ensembl?	Gene	Number of FDSIs	Number of Ensembl Transcripts	Number of Studies	Mappable to Ensembl?
Human	*AR*	3	9	31	Yes	*BDNF*	3	19	12	No
	BIRC5	3	11	34	Yes	*BOK*	2	2	1	No
	CD44	2	39	58	Yes	*CFLAR*	2	25	6	Yes
	CSPP1	2	7	1	Yes	*DPF3*	2	16	2	Yes
	EIF4G1	2	38	6	No	*EIF4G2*	3	32	3	No
	KLF6	2	7	16	No	*MST1R*	2	15	11	Yes
	PML	2	22	12	Yes	*PGAM5*	2	4	1	Yes
	PRMT5	2	20	4	Yes	*STIM2*	2	12	2	No
	SUN1	2	35	2	No	*TICAM1*	2	2	1	No
	TICAM2	2	2	1	No	*TP63*	2	14	27	Yes
Mouse	*Calca*	2	7	10	Yes	*Cdc42*	2	2	8	Yes
	Enc1	2	1	3	No	*Homer1*	2	12	7	Yes
	Il1rap	2	7	4	Yes	*Lpin1*	2	8	5	No
	Lrp8	2	12	13	No	*Mecp2*	2	6	9	Yes
	Myh10	2	10	7	No	*Nf1*	2	9	5	Yes
	Opn4	2	3	2	Yes	*Oprm1*	3	31	20	Yes
	Rbfox1	2	2	22	No	*Robo3*	2	6	5	Yes
	Rock2	2	12	4	No	*Ryr3*	2	12	5	No
	Sirt3	2	10	3	No	*Snap25*	2	3	12	Yes
	Tp63	2	8	9	Yes					

Table 2: Genes with FDSIs identified. Studies have provided positive evidence of functional distinctness for these genes in experiments where individual splice isoforms were eliminated and a phenotypic change was observed. See File S3 for study demonstrating functional distinctness. “Number of FDSIs” indicates the number of splice isoforms where depletion of splice isoforms causes a phenotype. “Number of Ensembl Transcripts” indicates number of transcripts found in Ensembl entry for gene. “Number of studies” indicates the number of studies associated with the gene retrieved with the term “alternative splicing” on PubMed. The highest number of FDSIs found in a single gene is three. “Mappable to Ensembl” indicates genes where we successfully linked all FDSIs back to Ensembl.

	Types of distinctness	**Human genes**	**Mouse genes**
Distinct expression patterns	Cell-type-specific	*AR*	
	Developmental-stage-specific	*CD44*	*Myh10, Robo3*
	Cellular localization	*BIRC5, CSPP1, PRMT5, PML*	*Myh10, Rbfox1, Robo3, Sirt3*
	Tissue-specific	*MST1R*	*Calca, Rock2*
	Other-condition-specific	*BOK*	
Biochemically distinct	Protein domain	*CFLAR, DPF3, EIF4G1, TICAM1, TP63*	*Lrp8*
	Dominant negative	*BIRC5, KLF6, Nf1, PRMT5, STIM2, SUN1, TICAM*	*Enc1, Nf1, Robo3, Ryr3, Tp63*
	Protein terminus change	*BDNF, EIF4G2, IL1RAP, PGAM5*	*Mecp2, Oprm1, Pn4*

Table 3: Most genes with FDSIs have biochemically distinct splice isoforms. Genes with FDSIs were categorized on functional type based on the literature that reported on the FDSIs using the scheme outlined in Figure 1. Genes categorized as “distinct expression patterns” express FDSIs in specific conditions. Genes categorized as “biochemically distinct” have FDSIs whose functional distinctness is a consequence of biochemical differences in their final protein product. Genes can be categorized as both “distinct expression patterns” and “biochemically distinct” such as Myh10 and Robo3.

Gene	Experimental method	Tissue/Cell Type	Reference (Pubmed ID)
Ank3	Isoform-specific rescue	Neuron	25552556
Ar	Knockdown	Prostate cancer cell line	20823238
Ccnd1	Isoform-specific rescue	Embryonic fibrobast	21200149
Dab1	Isoform-specific rescue	Neuron	28968791
Dntt	Isoform-specific rescue	Bone marrow	11136823
FANCE	Isoform-specific rescue	Breast cancer cell line	26277624
MADD	Isoform-specific rescue	Kidney	16682944
Pcdha1	Isoform-specific rescue	Brain	18973563
PDE4D	Knockdown	Kidney	16030021, 17673687
PEX19	Isoform-specific rescue	Fibroblast	11883941
Pparg	Knockdown and isoform-specific rescue	Adipose	11782442
RAP1GSDS1	Knockdown	Breast cancer cell line	24197117
RREB1	Knockdown	Bladder	21703425
SIRT1	Knockdown	Colon cancer cell line	22124156
Smad2	Isoform-specific rescue	Embryonic stem cells	15630024
STAT1	Knockdown	Embryonic cells	21914475

Table 4: Genes with evidence failing to support FDSIs (negative results). These genes had multiple isoforms tested however only one splice isoform caused a change in phenotype.

Figures:

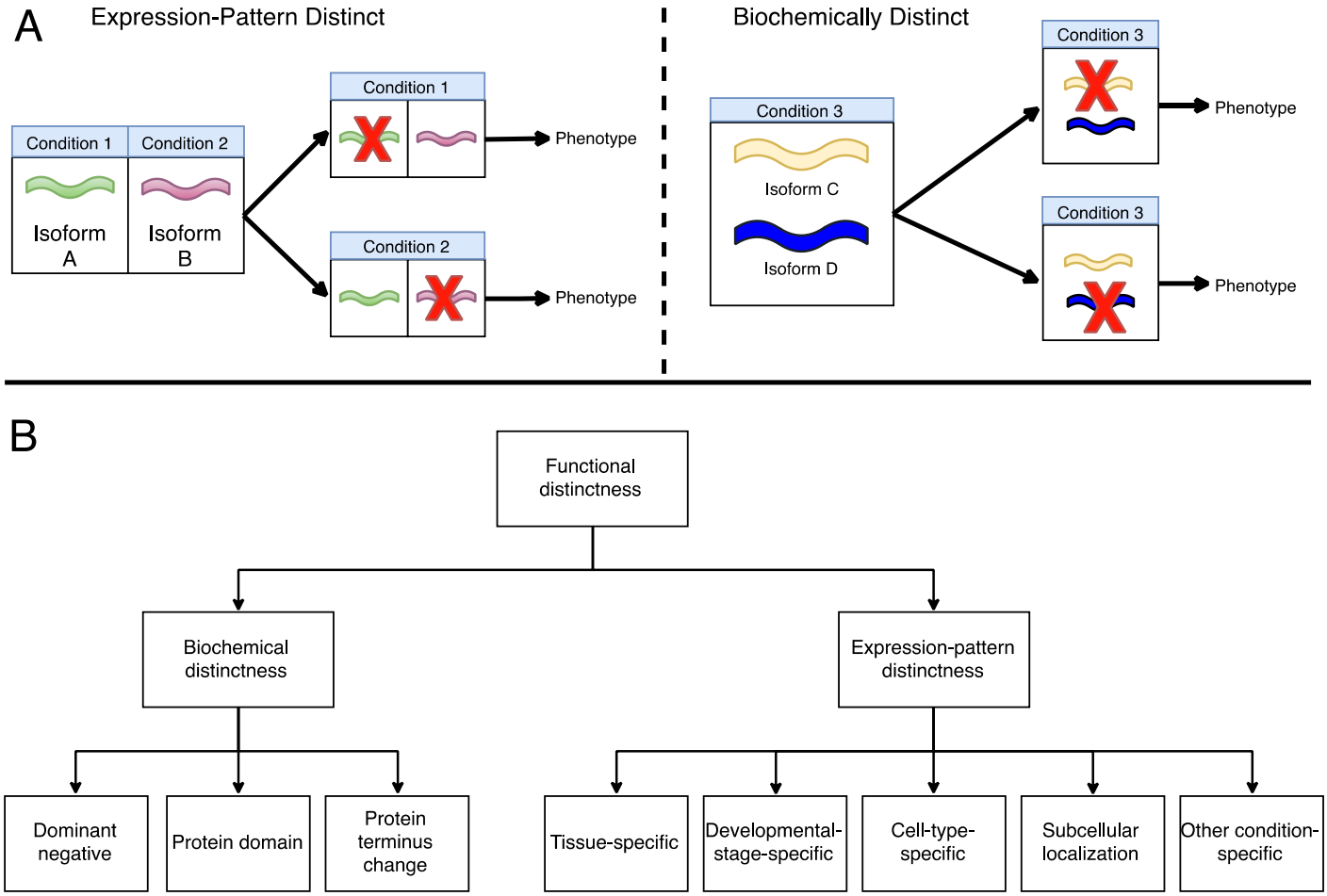

Figure 1: Non-mutually exclusive types of functional distinctness for literature reported genes with FDSIs. **A)** Generally, the distinctness of FDSIs of the same gene can be attributed to expression-pattern distinctness or biochemical distinctness. Expression-pattern distinctness is defined as a gene having specific splice isoforms necessary in distinct conditions. The depletion of the splice isoform in its distinct condition causes a phenotype. Biochemical distinctness is defined as a protein structure difference between splice isoforms of the same genes. While the FDSIs of the gene can be expressed in the same condition, the depletion of either splice isoform causes a phenotype. **B)** For genes with FDSIs, we categorized the specific subtypes of functional distinctness which contributed to the distinctness between the splice isoforms of the gene (summarized in Table 3). Expression-pattern distinctness can be further categorized as “cell-type-specific”, “tissue-specific”, “developmental-stage-specific”, “subcellular localization-specific” and “other condition-specific”. Biochemical distinctness can be further categorized as “dominant-negative”, “protein domain” and “protein terminus change”.

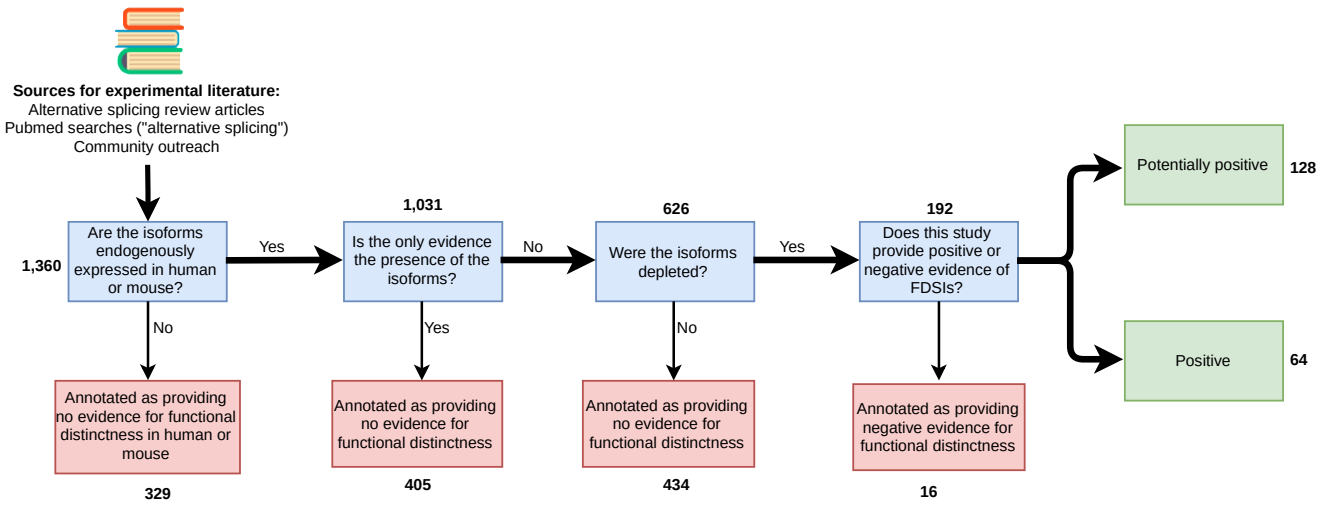

Figure 2: **Literature curation scheme.** We sought papers which study the functional distinctness of a single human or mouse gene's splice isoforms. Positive studies are those that provide evidence where multiple splice isoforms of a single gene are depleted and at least two isoforms show a phenotype. Potentially positive studies are studies where the depletion of only one isoform causes a phenotype. We annotated studies as providing negative evidence for functional distinctness when investigators deplete multiple splice isoforms of the same gene but only one produces an observable phenotype. The numbers in bold represent the number of studies in each category.