Systematic evaluation of isoform function in literature reports of alternative splicing: Supplemental Data

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

Table of Contents

SECTION 1: MASTER SPREADSHEET LEGEND

File S2 contains all human and mouse studies we curated along with all information annotated for each study. In this section, we explain each column heading in File S2. Guidelines to adding publications to this spreadsheet can be found in Section 2.

- "Evidence_of_Functional_Distinctness" – This column indicates whether the study provided evidence of functional distinctness. The column is annotated as the following (for further information, see Methods and Figure 2):
    - If no depletion of a splice isoform occurred, then this column is left blank
    - "Positive" – If multiple isoforms of the same gene are individually depleted and more than two individual depletions cause a phenotype
    - "Potentially positive" – If the depletion of a single isoform causes a phenotype
    - "Negative" – If the depletion of the multiple isoforms of the same gene occurs but only one depletion causes a phenotype while the remaining depletions cause no isoform
- "Gene" – the gene symbol and gene name investigated in this study
- "PMID" – the PubMed identifier of the curated study
- "Isoform_Names" – a list of the isoforms studied in the curated study
    - "NA" – Study will be annotated as NA if the study does not investigate splice isoforms
- "Number_of_Reported_Isoforms" – the number of splice isoforms for the investigated gene according to the authors
    - "NA" – Study will be annotated as NA if the study does not investigate splice isoforms
- "Number_of_Functional_Isoforms_according_to_authors" – the number of splice isoforms for the investigated gene that the authors claim is functional
    - "NA" – Study will be annotated as NA if the study not investigate splice isoforms or the authors make no claim about how many are functional
- "Article_claims_functional_splice_isoforms" – whether the authors indicate that the splice isoforms of the investigated gene are functional
    - "NA" – Study will be annotated as NA if the study does not investigate splice isoforms, or if the study makes no claims on functionality
- "Evidence_of_Functionality" – curator's notes about how functionality was tested in this paper
- "Isoform_causes_disease" – the investigated splice isoforms are only expressed in disease conditions
- "Isoform_organism" – the organism that endogenously expresses the splice isoforms
- "Experimental_organism" – the organism the splice isoforms were tested in
- "Only_evidence_is_presence" – a "no" in this column indicates depletion experiments in the study
- "Type_of_experiments" – experiments used to characterize splice isoforms

- “Curator” – initials of individual who curated the study

SECTION 2: CURATION STANDARDS (INSTRUCTIONS TO FILL OUT MASTER SPREADSHEET)

1. Double check the Master Spreadsheet to ensure the paper you have curated has not already been curated.
2. If paper is a review article, fill in all column with NAs with the exception of study type and PubMed ID. Fill in Pubmed ID with PubMed ID and study type with "review article"
3. Note that even if the species is not of interest (i.e human or mouse), we still will annotate the experiment
4. Identify investigated gene(s). Fill out gene column with the NCBI gene name and PMID with the article's PubMed ID.
    a. If multiple genes are investigated, then one row per gene
    b. If a paper doesn't have a PMID, enter in the citation for the paper
    c. The gene name should be written to the standards of the organism in which the isoforms are endogenously expressed
5. Identify the number of splice isoforms for the investigated gene. Identify the names of the splice isoforms if possible.
    a. This is reported by the authors but may be different from the number of splice isoforms they actually test in the experiment
    b. Fill out '# of splice isoforms' column and 'isoform name" column
6. Identify if the paper is actually about identifying the function of the gene's **endogenously** expressed splice isoforms.
    a. Splicing papers are sometimes not about the gene's splice isoform's function but regulation of the alternative splicing of that gene.
        i. This can be further confusing if the investigated gene is a splicing regulator. Sometimes these splicing regulators have splice isoforms whose function is being studied.
    b. If the study is about the regulation of the splice isoforms and not function, then make a note of this in the Evidence of Functionality column. Fill out the remaining columns as NA. Stop curating the study.
    c. Endogenously expressed isoforms are ones that are expressed in a healthy wildtype population
        i. Splice isoforms expressed solely in a disease condition is **not** what we are looking for.
            1. Example: in cancer we observe novel splice isoform expressions and we don't report on this
            2. If the study is about disease associated splice isoforms, make a note of this in the 'Evidence of Functionality' column and mark the "isoform

causes a disease" column as yes. Fill out the remaining columns as NA and then stop curating. Annotate "study type" column as "disease association"

d. This can be further complicated if the **upregulation of an endongenously expressed splice isoform** causes a disease. Certain studies can identify a splice isoform that is necessary for the overall function of the gene at the wildtype expression level but at an upregulated level, there's a disease phenotype. We are still interested in these splice isoforms.
    i. Another source of confusion is that often time a lowly expressed wildtype splice isoform can be slightly detrimental but has no effect on the cell as it is lowly expressed. But when the splice isoform is upregulated or an overexpression study, a disease is caused because now the splice isoform is abundant.

7. Determine which splice isoforms the authors use for their experiments and whether or not the authors are claiming the splice isoforms are functional
    a. Fill out the "# of functional isoforms" column
8. Determine the organism the splice isoform has been studied in and fill out the "Organism" column
    a. This is generally the organism where the splice isoform is endogenously expressed
        i. However, studies will occasionally test their splice isoforms in multiple organisms.
        ii. Add this to the organism column (e.g mouse, rat, ferret)
9. In the 'study types' column, fill out the experiment that was performed to investigate the splice isoforms. Common experiments we have seen are: knockdown, knockdown (1 isoform), knockdown (non-isoform-specific), knockout, knockout (1 isoform), knockout (non-isoform-specific), rescue, rescue (1-isoform), rescue (non-isoform-specific) overexpression, tissue distribution, subcellular localization, activity assay, protein interaction, disease association, regulation of isoform expression, detection, structural characterization, mutation isoform absence, immunodepletion, timecourse distribution
    a. If multiple experiments were used, then annotate column with all experiments. The main experiment type should be listed first.
        i. Usually if the experiments depleted a splice isoform, then we want that experiment listed first.
            1. Example: if an experiment investigated isoform tissue distribution and knockdown, then "knockdown, tissue distribution" is what the column should be annotated with
        ii. If any of the experiments used eliminated the expression of a single isoform and looked for an effect, then mark the 'Only evidence is presence' as no. These are the studies we are most interested in.

iii. If more than one splice isoform is shown to be necessary (i.e the absence of each splice isoform causes a phenotype) then mark this study in the “Evidence for Functional Distinctness?” column as “positive”.

iv. If only one splice isoform is depleted, then mark this study in the “Evidence for Functional Distinctness column as “potentially positive”

b. If the splice isoforms cause a different function from each other or have the same function as each other, then fill out the appropriate column to reflect that (“same function” column or “different function” column)

10. Fill out the Evidence of Functionality column with a short, concise description of the study in present tense. This description will often explain how the isoforms were molecularly characterized and what function was tested. If the paper does provide evidence of a “Gold Standard Gene”, please include in the description which figure(s) best shows this evidence.
11. Fill out “sequence accession” column with any sequence accession information provided to the investigated splice isoforms.

SECTION 3: PROBLEMATIC CASES OF LINKING FDSIs TO ENSEMBL

The following file describes the functionally distinct splice isoforms (FDSIs) that we failed to link to an appropriate Ensembl Transcript ID, and the investigators of the curated literature provided sequence accessions.

RNA binding protein, fox-1 homolog 1 (*Rbfox1*)

According to Hamada and colleagues, mouse *Rbfox1* has two FDSIs: Rbfox1-isoform5 (aliases: Rbfox1_C, RBFOX1-isoform2) and Rbfox1-isoform1 (alias: Rbfox1_N) (1). Hamada and colleagues performed knockdown experiments of both splice isoforms. In vivo knockdown of Rbfox1-isoform1 in mice resulted in cortical neuron radial migration and terminal translocation defects. In vitro knockdown of Rbfox1-isoform1 in hippocampal neurons reduced spine density, and the length of dendrites and primary axons. In-vitro Rbfox1-isoform5 knockdown increased the number of stubby-shaped immature spines only, while Rbfox1-isoform1 knockdown increased the number of filopodia-like immature spines as well.

The Ensembl database contains two splice transcripts for mouse *Rbfox1*: ENSMUST00000056416.7 and ENSMUST00000115841.9. The transcript ENSMUST00000056416.7 has a CDS length of 4048 bp and a protein length of 417 amino acids. The other transcript, ENSMUST00000115841.9, has a CDS length of 1,192 bp and a protein length of 396 amino acids.

We linked Rbfox1-isoform1 to Ensembl transcript ID ENSMUST00000115841.9 for Rbfox1. Hamada and colleagues constructed the plasmid for Rbfox1-isoform1 from GenBank submission AY659954. According to GenBank, Rbfox1-isoform1 has a CDS length of 1188 bp and a protein length of 396 amino acids. The amino acid sequence found in GenBank for Rbfox1-isoform1 match exactly to the amino acid sequence for an Ensembl transcript ID ENSMUST00000115841.9 – though the CDS sequences do not match.

We failed to link Rbfox1-isoform5 to any Ensembl transcripts. Hamada and colleagues constructed the plasmid for this splice isoform from the sequence reported in the GenBank entry AY659955. According to this GenBank record, Rfox1-isoform5 has a CDS length of 1,241 bp and protein length of 373 amino acids which does not match either transcript reported in the Ensembl entry for Rbfox1.

Two other reports further provided positive evidence of FDSIs for *Rbfox1*. Hamada and colleagues (2015) again performed in vitro knockdown experiments for Rfox1-isoform5 (2). Lee and colleagues performed isoform specific rescues using Rbfox1-isoform1 and Rbfox-isoform5 (3). Neither reports provided GenBank accessions however they provided sequence information which matched the isoforms reported in Hamada and colleagues (2016).

Ryanodine receptor 3 (*Ryr3*)

According to Dabertrand and colleagues, mouse *Ryr3* has two FDSIs: RYR3L and RYR3S (4). Dabertrand and colleagues performed knockdown experiments for both FDSIs of Ryr3 in mouse duodenum myocytes. The elimination of RYR3L decreased the upstroke and amplitude velocity of caffeine-induced calcium response while the elimination of RYR3S increased the upstroke and amplitude velocity of caffeine-induced calcium response.

The Ensembl database contains 12 splice variants for mouse *Ryr3*. Of these splice variants, the six splice are protein coding. Five of the protein coding splice variants have a protein length greater than 4,000 amino acids while the remaining splice variant has a protein length of 46 amino acids.

We failed to link RYR3L to any Ensembl transcript. Dabertrand and colleagues constructed the plasmid for RYR3L from sequence data provided in the GenBank entry AF111166. The protein sequence provided in this GenBank entry has a length of 454 amino acids. Note, that the European Nucleotide Archive (ENA) has linked the sequence for this GenBank entry to two Ryr3 transcripts in Ensembl: ENSMUST00000080673 and ENSMUST00000208151. The amino acid length for the proteins encoded by these Ensembl transcripts are 4,863 amino acids and 4,834 amino acids, respectively. We failed to understand how ENA linked RYR3L to these Ensembl transcripts.

We also failed to link RYR3S to any Ensembl transcript. For their construction of the RYR3S plasmid, Dabertrand and colleagues referred to previous work by Coussin and colleagues (5). Coussin and colleagues constructed their sequence from the GenBank entry X83934. This GenBank entry connected us to a RYR3S protein sequence with a length of 288 amino acids. Despite failing to link this protein for RYR3S to any Ensembl entry, we linked the RYR3S sequence to the RefSeq entry XM_619795. However, this RefSeq removed this entry from their database due to insufficient evidence and replaced it with a new sequence.

Sad1 and UNC84 domain containing 1 (*SUN1*)

According to Nishioka and colleagues, human *SUN1* has three FDSIs: SUN1_916, SUN1_888 and SUN1_785 (6). They performed knockdown experiments for all three splice isoforms in HeLa cells as well as further knockdown experiments for SUN1_916 and SUN1_888 in MDA-MB-231 cells. The experiments revealed that all three splice isoforms were necessary for proper cell migration.

The Ensembl database contains 35 transcripts for human *SUN1* and 17 transcripts are protein-coding. The largest protein-coding transcript encodes for a splice isoform of 822 amino acids.

We linked SUN1_785 to the Ensembl transcript ID ENST00000401592.5. Nishioka and colleagues constructed the transcript from the RefSeq ID NM_001130965 and the transcript encoded for a 785

amino acid protein. The Ensembl database had already linked this RefSeq ID to ENST00000401592.5 and we ensured that the RefSeq protein matched the Ensembl protein using ClustalOmega.

We failed to link SUN1_916 and SUN1_888 to any Ensembl record. Nishioka and colleagues provided the GenBank accessions EAW87177.1 and AB648918, respectively, as the source for plasmid construction. According to their GenBank entries, SUN1_916 encoded a 916 amino acid protein while SUN1_888 encoded for a protein with 888 amino acids. No Ensembl record matched either of these transcripts or proteins. Furthermore, we found no notes on the GenBank entries to explain the absence of these splice isoforms on the Ensembl database.

SECTION 4: SUPPLEMENTAL TABLES AND FIGURES

ACTR2	CSF2RA	GPX6	MBD3L5	POLE	STC2
AGPAT2	CTSD	GRK4	MDM4	PPP1R32	SUMF1
AGTR1	CUL3	HBA1	MEG9	PRKAR1A	SUMO3
ALG3	DGKI	HMGB1P6	MGAT5B	PTGS2	SYNCRIP
ARHGEF26	DNAJB11	HMGB2	MKLN1	RABL6	TBC1D2
C1QTNF7	DPEP3	HMGB3P18	MLLT10P1	REG1A	THRB
CCDC71L	DSC3	HRH3	MORN4	RTBDN	TRIM25
CD82	ELMOD3	IKBKB	MTA2	SCARB1	TSPY1
CDH10	EME1	INTS7	MYC	SCIN	URB2
CHRD	EXD2	IPP	NAPA	SLC12A3	VEGFC
CHRNE	FAM19A1	KIF21A	NAPG	SLC26A2	VTA1
CHST12	FAM96B	KRR1	NOTCH1	SMARCB1	XAGE-4
CLEC10A	FANCD2	KRT72	NR2E3	SOST	ZCCHC8
CLHC1	FGD3	LRRN4	NYAP2	SPERT	ZNF441
CMBL	FILIP1	MAL	PARP15	SRRD	ZNF506
COL6A4P1	FSCB	MAPKAPK2	PITX2	SRSF11	
CRNN	GLE1	MBD3L4	PKD2	STAP1	

Table S1: 100 random human genes for gene-centric curation. We randomly selected these genes from the set of genes that linked to the publications retrieved from the PubMed query “alternative splicing”. We also gene-centrically curated the mouse orthologues of these genes.

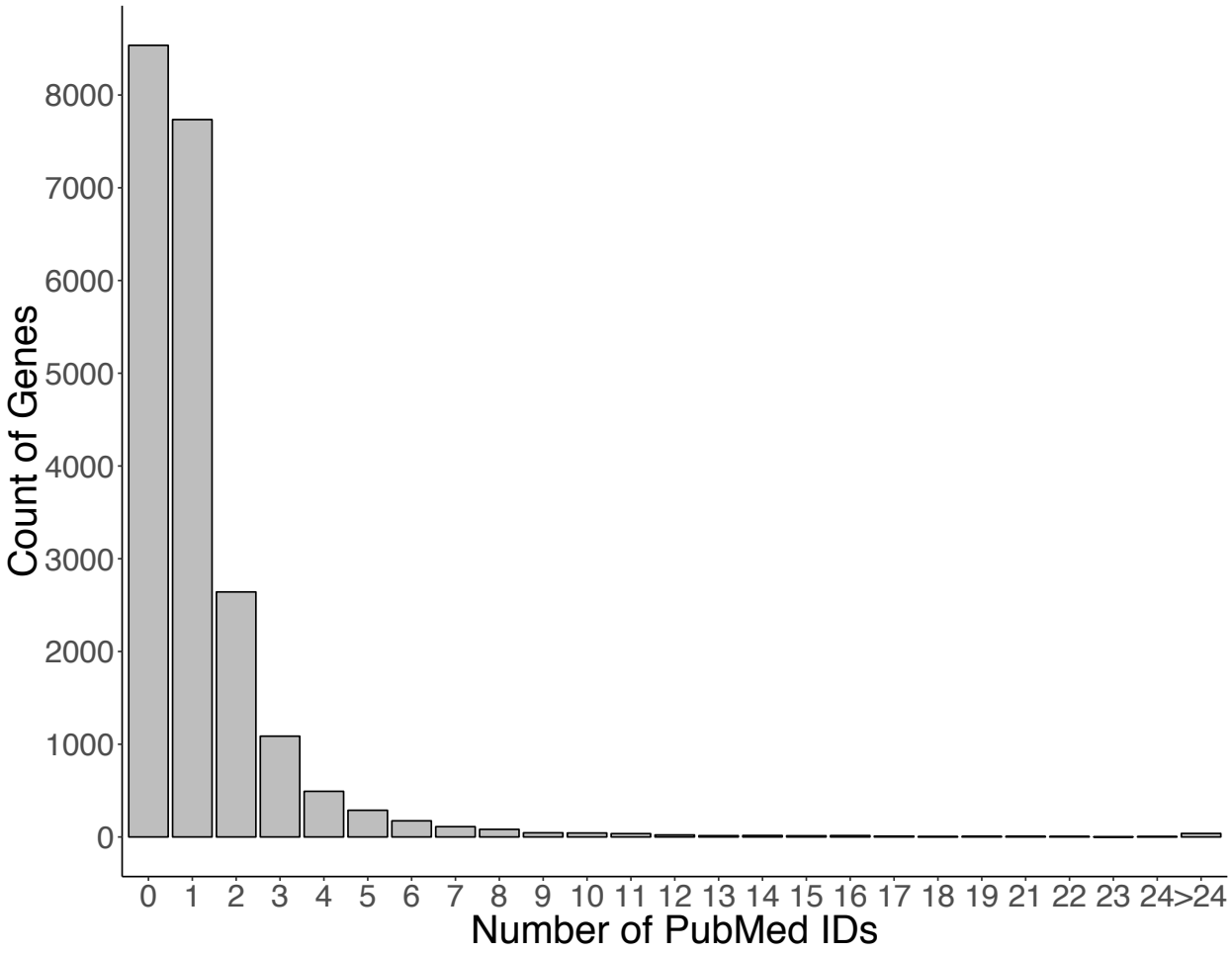

Figure S1: Most human genes have one study linked to alternative splicing in PubMed. The total number of human studies retrieved with the term “alternative splicing” was 19,049. These studies linked to 12,891 human genes. Genes (taken from Ensembl) that were not retrieved from this query were labelled as 0.

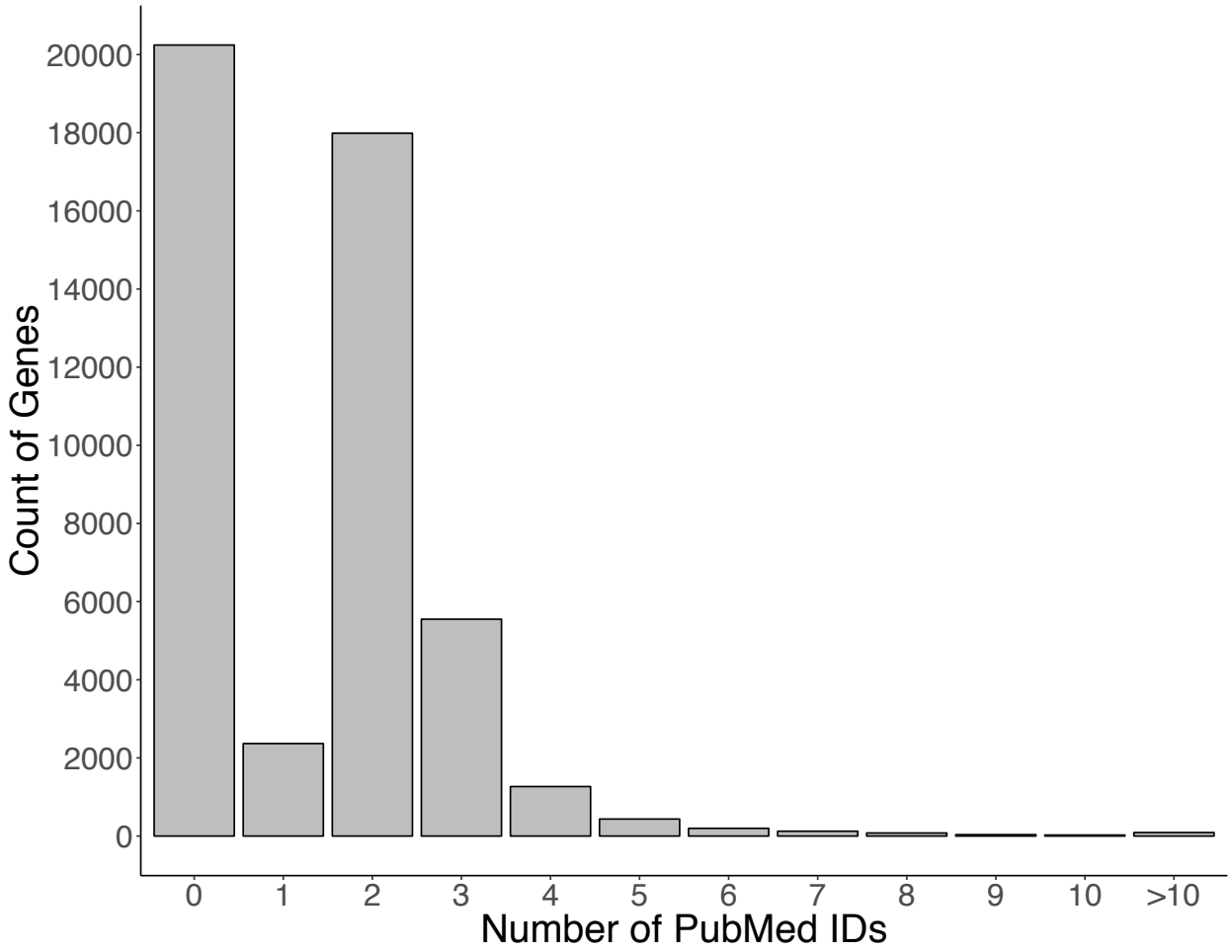

Figure S2: Most mouse genes have two study linked to alternative splicing in PubMed. The total number of mouse studies retrieved with the term “alternative splicing” was 8,203. These studies linked to 28,167 mouse genes. Note that this gene count, unlike the human gene count, include non-protein coding genes. This was likely due to six transcriptome-wide mouse studies which include the term “alternative splicing” and all mouse genes. Removal of these six studies from the query results led to only 7,585 mouse genes associated with a total of 8,197 “alternative splicing”-mentioning studies as described in the Results. Furthermore, after filtering these six studies, most mouse genes had one study which mentioned alternative splicing. We did not see a similar issue in our human query shown in Figure S1. Genes (taken from Ensembl) that were not retrieved from this query were labelled as 0.

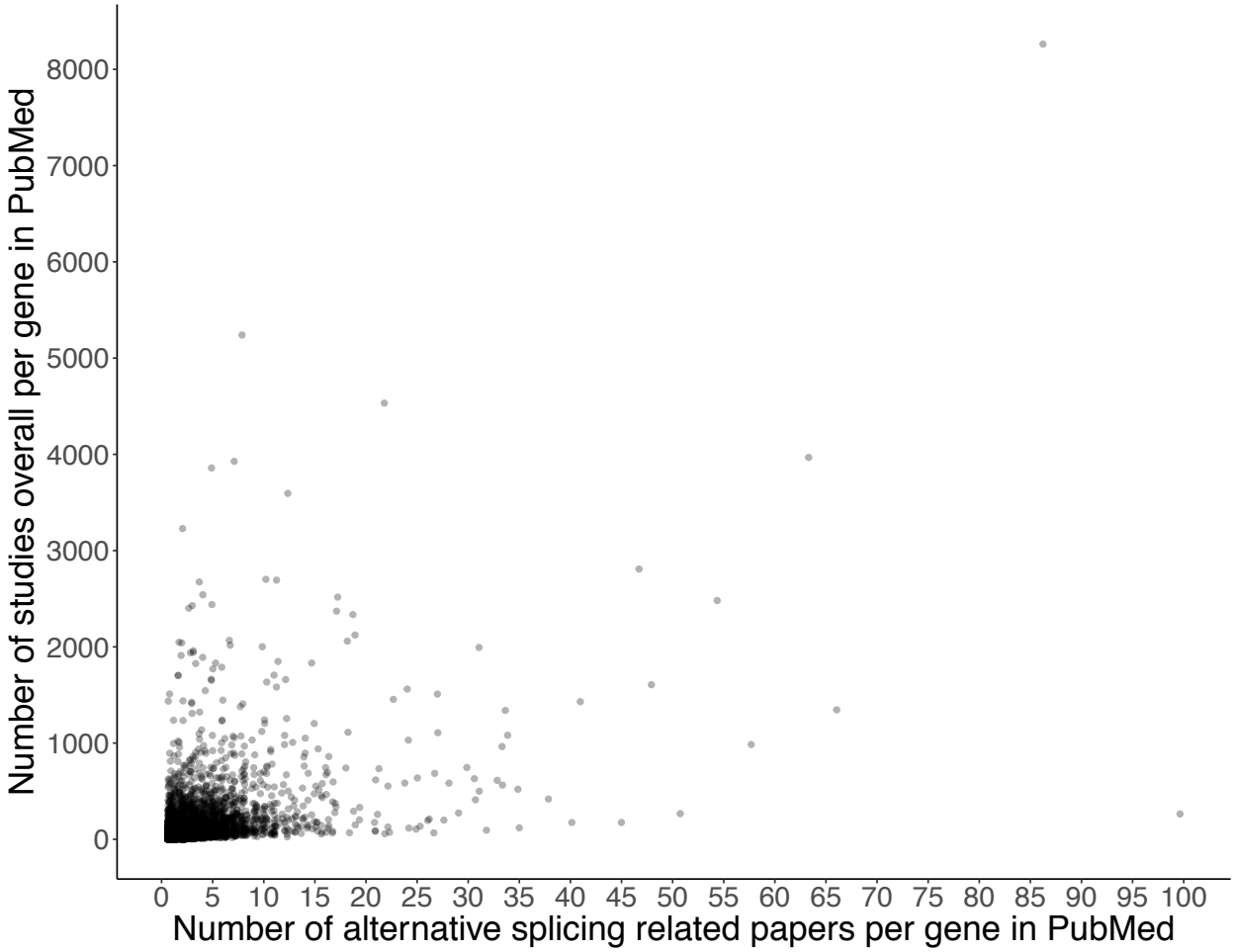

Figure S3: Human genes commonly studied in the context of alternative splicing are commonly well-studied in general. Based on our query of PubMed, we compared the number of papers which mentioned “alternative splicing” to the overall number of studies for a given gene (Spearman’s rank correlation = 0.55).