Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

Evidence for widespread subfunctionalization of splice forms in vertebrate genomes

Matthew J Lambert, Wayne O Cochran, Brandon M Wilde, et al.

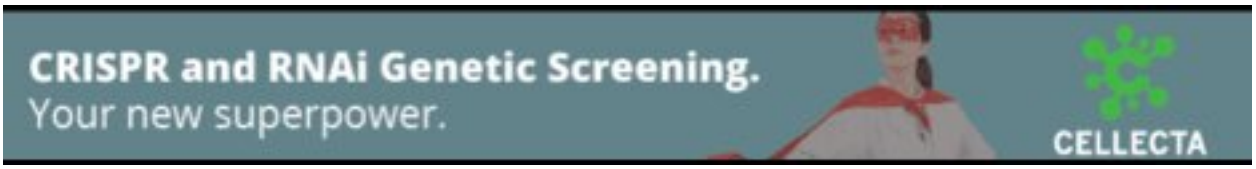
Genome Res. published online March 19, 2015
Access the most recent version at doi:10.1101/gr.184473.114

P<P Published online March 19, 2015 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

Evidence for widespread subfunctionalization of splice forms in vertebrate genomes.

Matthew J. Lambert[1], Wayne O. Cochran[2], Brandon M. Wilde[3], Kyle G. Olsen[3], Cynthia D. Cooper[1,4]

1. School of Biological Sciences, Washington State University, Pullman, WA 99164, USA

2. School of Engineering and Computer Science, Washington State University Vancouver, Vancouver, WA 98686, USA

3. Washington State University Vancouver, Vancouver, WA 98686, USA

4. School of Molecular Biosciences, Washington State University, Pullman, WA 99164, USA

Corresponding Author: Cynthia D. Cooper

Email: cdcooper@vancouver.wsu.edu

Phone: (360) 546-9342

Keywords: gene duplication, alternative splicing, subfunctionalization

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

Abstract:

Gene duplication and alternative splicing are important sources of proteomic diversity. Despite research indicating that gene duplication and alternative splicing are negatively correlated, the evolutionary relationship between the two remains unclear. One manner in which alternative splicing and gene duplication may be related is through the process of subfunctionalization, in which an alternatively spliced gene upon duplication divides distinct splice isoforms among the newly generated daughter genes, in this way reducing the number of alternatively spliced transcripts duplicate genes produce. Previously, it has been shown that splice form subfunctionalization will result in duplicate pairs with divergent exon structure when distinct isoforms become fixed in each paralog. However, the effects of exon structure divergence between paralogs have never before been studied on a genome-wide scale. Here, using genomic data from human, mouse and zebrafish we demonstrate that gene duplication followed by exon structure divergence between paralogs results in a significant reduction in levels of alternative splicing. In addition, by comparing the exon structure of zebrafish duplicates to the co-orthologous human gene, we have demonstrated that a considerable fraction of exon divergent duplicates maintain the structural signature of splice form subfunctionalization. Furthermore, we find that paralogs with divergent exon structure demonstrate reduced breadth of expression in a variety of tissues when compared to paralogs with identical exon structures and singletons. Taken together, our results are consistent with subfunctionalization partitioning alternatively spliced isoforms among duplicate genes and as such highlight the relationship between gene duplication and alternative splicing.

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

Introduction:

Gene duplication has long been known as an important source of genetic novelty leading to functionally distinct proteins (Ohno 1970). Alternative splicing, through the differential retention of exonic and in some cases intronic material, likewise generates novel protein sequences (Black 2003). There are several types of alternative splicing events including; cassette exons, alternative 5' and 3' splice sites, mutually exclusive exons and intron retention, with cassette exons being the most common in vertebrates. Interestingly, rates of alternative splicing correlate with organismal complexity and it is alternative splicing that bridges the gap between the approximately 22,000 genes within the human genome and the greater than 100,000 unique proteins that the human genome codes for (Keren et al. 2010). Given that both alternative splicing and gene duplication increase proteomic diversity, it is not surprising that several authors have hypothesized that the two processes may be evolutionarily related. One manner in which alternative splicing and gene duplication may be related is through subfunctionalization, wherein an alternatively spliced gene upon duplication partitions distinct splice isoforms among the newly generated daughter genes (paralogs) (Hughes 1994; Lynch and Force 1999). In the context of subfunctionalization, gene duplication substitutes for alternative splicing, thus freeing duplicate genes from the constraints of maintaining multiple functions (Hughes 1994; Hahn 2009)

Numerous individual gene case studies provide evidence that alternative splicing and gene duplication may be interchangeable. For example, Yu *et al*. (2003) demonstrate the subfunctionalization of the duplicated fugu (*Takifugu rubripes*) gene *syn2* by showing that the two copies correspond to distinct isoforms of the alternatively spliced orthologous human

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

gene. In a similar manner, Altschmied *et al*. (2002) and Hultman *et al*. (2007) demonstrate the subfunctionalization of the teleost genes *mitf* and *kitl*, respectively. Recently, Marshall *et al.* (2013) show that splice form subfunctionalization is not restricted to vertebrates. They report that in several fungal species two distinct proteins, SKI7 and HBS, are generated from a single alternatively spliced gene. In *Saccharomyces cerevisiae*, however these isoforms were partitioned into distinct genes after gene duplication and subsequent subfunctionalization.

In addition to those listed above, several genome-wide studies have examined the relationship between alternative splicing and gene duplication. Kopelman *et al.* (2005) and Su *et al.* (2006) provide support for the relationship between gene duplication and alternative splicing by showing that as gene family size increases, levels of alternative splicing decrease. This negative correlation was presumed to be the product of subfunctionalization reducing alternative splicing levels in duplicate genes (Su *et al.* 2006). Others, however, have questioned whether the inverse relationship between alternative splicing and gene duplication can be explained by subfunctionalization or if the negative correlation between the two is even valid (Talavera *et al*. 2007; Jin *et al.* 2008; Chen *et al.* 2011; Roux and Robinson-Rechavi 2011; Grishkevich and Yanai 2014). Of note, Roux and Robinson-Rechavi (2011), using a more complete dataset than either Kopelman *et al.* (2005) or Su *et al.* (2006), find that the largest gene families display the lowest levels of alternative splicing; however, singletons do not have the highest levels of alternative splicing. In fact, it would seem as though gene families of intermediate size undergo the most alternative splicing. This is difficult to reconcile with the hypothesis that gene duplication can substitute for alternative splicing, but suggests that less stringent purifying selection acting on newly formed duplicates may allow for the acquisition of novel splice forms (Roux and

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

Robinson-Rechavi 2011). They also argue that the apparent reduction in alternative splicing found in the largest gene families is not due to gene duplication reducing levels of alternative splicing, but is the product of a preduplication bias predisposing genes with initially low levels of alternative splicing to duplicate more frequently (Studer and Robinson-Rechavi 2009; Roux and Robinson-Rechavi 2011).

More recently, we have argued that the signature of splice form subfunctionalization for genes containing alternatively spliced cassette exons before duplication will be duplicate pairs with different numbers of exons. We have demonstrated that in zebrafish and fugu, duplicate pairs with different numbers of exons undergo less alternative splicing than singletons and duplicate pairs with identical numbers of exons. This reduction in alternative splicing is reflected in both the proportion of genes that undergo alternative splicing and mean number of transcripts per gene (Lambert *et al.* 2014).

The purpose of this study is to examine the relationship between alternative splicing and gene duplication in light of exon structure divergence after gene duplication on a genome-wide scale, in a variety of vertebrate species.

Results:

Exon divergent paralogs undergo less alternative splicing:

This study focuses on human (*Homo sapiens*), mouse (*Mus musculus*) and zebrafish (*Danio rerio*). These species were selected because they represent a wide range of vertebrates and for each there exists a considerable amount of resources, including well annotated genomes and an abundance of expression data. For each species we compiled a dataset that consists of

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

singletons, and sibling paralogs (see methods). Within the three species, between 4 and 26% of all genes are sibling paralogs that have diverged such that they now have different numbers of exons (fig 1a). Such change represents an extreme form of intron/exon structure divergence (Xu *et al.* 2012). We refer to these duplicate pairs with non-identical numbers of exons as exon divergent paralogs. Those sibling paralogs that have retained the same number of exons are referred to as nondivergent paralogs. Previously, we have shown for a subset of genes in zebrafish that exon divergent duplicate pairs undergo significantly less alternative splicing compared to single-copy genes and nondivergent paralogs. Furthermore, by comparing the intron/exon structure of zebrafish paralogs to their alternatively spliced orthologs in human, we demonstrated splice form subfunctionalization (fig. 1b) for a large portion of the exon divergent duplicate pairs (Lambert *et al*. 2014). Here, to test if the effects of splice form subfunctionalization are detectable at the level of entire genomes we calculated the mean number of transcripts per gene for the singletons, exon divergent and nondivergent paralogs in each of the three species. In every case, the exon divergent paralogs have significantly fewer alternatively spliced transcripts per gene compared to both singletons and nondivergent paralogs (fig 2a). In addition, we calculated the proportion of genes that are alternatively spliced – i.e. the fraction of genes with more than one transcript – for each group in each species. Again, we find that the exon divergent group consistently demonstrates a significant reduction in alternative splicing (fig 2b). Although the mean number of transcripts and proportion of genes with more than one transcript are both widely used to measure rates of alternative splicing, Su and Gu (2012) argue that a more accurate measure is the mean number of transcripts per gene, excluding genes with only a single transcript. Calculating levels of

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

alternative splicing in this manner is more sensitive to rare alternative splicing events that likely remain undetected by current sequencing methods. Furthermore, by excluding genes with only a single transcript we eliminate the effects of duplication via retrotransposition, which creates non-alternatively spliced, intronless daughter genes regardless of the alternative splicing status of the parental gene (Kopelman *et al*. 2005). We performed this calculation and found that the exon divergent paralogs still demonstrate the lowest levels of alternative splicing in each species (supplemental fig. S1). It should be noted that we have included all transcripts, protein coding and otherwise, in these calculations. However, filtering the dataset so that only protein coding transcripts are counted did not change the overall pattern. The exon divergent paralogs have the lowest levels of alternative splicing even when only protein coding transcripts are considered (supplemental fig. S2). Importantly, we find that the nondivergent paralogs consistently display the highest levels of alternative splicing, regardless of which measure of alternative splicing is used.

The above results suggest that gene duplication followed by exon structure divergence, but not gene duplication alone, may lead to a reduction in alternative splicing. However, there are several potentially confounding factors. First, it is known that gene families with many paralogs undergo very low rates of alternative splicing (Kopelman *et al*. 2005; Su *et al*. 2006; Jin *et al.* 2008; Chen *et al* 2011; Roux and Robinson-Rechavi 2011). Therefore, if the exon divergent paralogs are enriched in genes that belong to these large gene families, the importance of exon structure divergence would necessarily be called into question as the association between exon structure divergence and lower rates of alternative splicing may be spurious, due to the influence of gene family size. To test for the effects of gene family size we compared the

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

number of paralogs per gene for the exon divergent and nondivergent groups. Singletons, by definition, have only one member in each family. Figure 3a shows the number of paralogs per gene is significantly different between the two groups in each species, however it is the nondivergent paralogs, rather than the exon divergent paralogs, that have larger mean family sizes. These results indicate that family size is not responsible for the reduction in alternative splicing observed in the exon divergent group.

Another potentially confounding effect is based on the age of the duplication event that created the paralogs. Su *et al*. (2006) demonstrate that in humans there is a reduction in alternative splicing shortly after gene duplication. This reduction appears to be transient, as increasing time since duplication correlates with increasing levels of alternative splicing (Su *et al* 2006; Roux and Robinson-Rechavi 2011). This means that if exon divergent paralogs are enriched in young duplicates, the influence of duplication age may make it appear as though exon structure divergence between paralogs results in lower levels of alternative splicing. Figure 3b plots the proportions of exon divergent and nondivergent paralogs created during a variety of evolutionary epochs as indicated by Ensembl Compara (Vilella *et al*. 2009). Within each plot the proportions of divergent and nondivergent paralogs produced at each epoch are strikingly similar. Furthermore, in each species, except humans, there exists an enrichment of nondivergent paralogs in the most recent age categories. In humans, the proportion of nondivergent paralogs is slightly less than the proportion of exon divergent paralogs when considering the most recent duplication ages (*Homo sapiens* – Hominoidea), however this excess is only marginally significant ($\chi^2 = 6.06$, $p = .014$). These results demonstrate that exon

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

divergent paralogs are not enriched in young duplicates when compared to nondivergent paralogs and that time since duplication is not a confounding factor.

Lastly, Grishkevich and Yanai (2014) use several genetic parameters, specifically gene length and expression level, to explain the inverse relationship between gene duplication and alternative splicing. They argue that shorter genes are more likely to duplicate while being less likely to undergo alternative splicing. Likewise, they argue that genes that are expressed at low levels demonstrate low levels of alternative splicing and high rates of duplication. These arguments provide a mechanism for the preduplication bias first proposed by Roux and Robinson-Rechavi (2011). That is to say, the apparent reduction in alternative splicing found in large gene families is not due to gene duplication, rather it is a consequence of short genes and lowly expressed genes being more likely to duplicate, while being less likely to undergo alternative splicing. It is very unlikely that a preduplication bias is present in our dataset because in each species only the exon divergent duplicates show any reduction in alternative splicing, while the nondivergent duplicates consistently demonstrate the highest levels of alternative splicing (fig. 2). However, to demonstrate that gene length does not produce a confounding effect we calculated the mean length of each category - exon divergent, nondivergent and singletons - in each species. Were gene length a confounding factor we would expect the exon divergent paralogs to have the shortest lengths. Contrary to this expectation, we find that in each of the three species, exon divergent paralogs never have the shortest mean length (supplemental fig. S3). Using UniGene expression data we also calculated mean expression levels for each category in each species. We find that in accordance with Grishkevich and Yanai (2014) the exon divergent paralogs demonstrate low levels of expression,

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

however we believe this to be a consequence of subfunctionalization reducing the number of tissues in which the exon divergent paralogs are expressed (see below).

To further demonstrate that our results are not subject to these or any other potential preduplication biases, we queried the level of alternative splicing for all single-copy human genes that are orthologous to the genes that comprise the zebrafish dataset. In this manner, the human orthologs serve as a preduplication outgroup allowing us to approximate the ancestral state (Roux and Robinson-Rechavi 2011). These human orthologs were categorized based on the duplication state of the zebrafish genes, producing three groups; human orthologs of either zebrafish singletons, zebrafish nondivergent paralogs or zebrafish exon divergent paralogs. If any manner of bias were responsible for the reduction in alternative splicing witnessed by exon divergent paralogs, we would expect the human orthologs of the zebrafish exon divergent paralogs to have significantly lower levels of alternative splicing, when compared to the human orthologs of the zebrafish singletons or nondivergent paralogs. We find that the opposite is true, the human orthologs of the zebrafish exon divergent paralogs have the highest rates of alternative splicing (fig. 4). We performed the same analysis with the mouse dataset, again using orthologous human genes as a proxy for the ancestral state, and with human using mouse orthologs as the preduplication outgroup. In neither instance did we find a significant reduction in the levels of alternative splicing for the orthologs of the exon divergent paralogs (fig. 4). These results indicate that the reduction in alternative splicing witnessed in the exon divergent paralogs is not the product of a preduplication bias.

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

Exon divergent paralogs are expressed in fewer tissues:

The reduction in alternative splicing seen in the exon divergent paralogs suggests that splice form subfunctionalization may be driving down levels of alternative splicing. If this were true, we would expect the consequences of subfunctionalization to be manifest in other areas as well. For example, along with the partitioning of splice forms, expression profiles may also be divided among subfunctionalized duplicates. In this manner, subfunctionalization will result in duplicate genes with fewer transcripts and more restricted expression profiles. Indeed, shifts in expression after duplication have often been used as evidence for subfunctionalization (Blanc and Wolfe 2004; Casneuf *et al* 2006; Duarte *et al.* 2006; Hellsten *et al.* 2007; Wolfe and Semon 2008). To test if gene duplication followed by exon structure divergence results in the partitioning of expression domains, we analyzed the available UniGene (Pontius *et al.* 2003) EST expression profiles, which approximate expression patterns inferred from EST libraries. In line with our expectations, we find that in human and zebrafish the exon divergent paralogs are expressed in fewer tissue types than either singletons or nondivergent paralogs (fig 5). In mouse, the exon divergent paralogs are expressed in fewer tissues than singletons, however there is no significant difference in the number of tissues between exon divergent and nondivergent paralogs. It may be that the reduced expression profiles evidenced in the exon divergent group is simply an effect of exon divergent paralogs possessing fewer transcripts per gene. To test this we calculated the Pearson correlation coefficient between the number of tissues in which a gene is expressed and the number of transcripts a gene produces. We find that these parameters are correlated, however only very weakly (zebrafish r = .10, mouse r = .16, human r = .30), suggesting that a reduction in alternative splicing does not, in and of itself,

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

account for a restricted expression profile. In addition, to demonstrate that the restricted expression profiles found in the exon divergent paralogs is not simply a consequence of the selected dataset, we performed the same analysis using the egenetics/SANBI (Kelso *et al*. 2003) human gene expression dataset. Again, we find that exon divergent paralogs are expressed in fewer tissue types (supplemental figure S4).

Structural evidence for splice form subfunctionalization:

The correspondence of exon structure between unique copies in a given sibling pair to distinct alternatively spliced transcripts from a preduplication outgroup gene serves as a signature of splice form subfunctionalization. To determine if the exon divergent duplicates maintain this signature of subfunctionalization we compared the exon structure of a subset of zebrafish exon divergent duplicates to their co-orthologous gene in human, looking specifically for instances in which each duplicate in a given sibling pair shares common exon structure with distinct alternatively spliced transcripts form the single human ortholog. Despite the fact that the independent gain and loss of introns and exons along each lineage is expected to mask the signature of subfunctionalization, we find that 25% of those tested demonstrate a pattern that is indicative of splice form subfunctionalization (Supplementary Table 1). Figure 6 illustrates the splice form subfunctionalization of the gene cyclin M2 (*cnnm2*) in zebrafish. The human ortholog of cyclin M2, *CNNM2*, indicates that prior to duplication the single copy gene coded for both a long and a short transcript. The long transcript contains eight exons and codes for a peptide that is 875 amino acids long. The short transcript contains seven exons and codes for an 853 amino acid peptide, however this transcript does not express the sixth exon found in the long transcript. In zebrafish, *cnnm2a* codes for a single transcript that contains eight exons, all

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

of which correspond to the eight exons found in the long transcript of the human ortholog. The zebrafish paralog of *cnnm2a, cnnm2b*, shares nearly identical exon structure with the short human transcript. The differential retention of the human cassette exon in the zebrafish duplicates suggests that the *cnnm2* paralogs in zebrafish have been retained due to splice form subfunctionalization. Interestingly, tissue-specific expression patterns of zebrafish *cnnm2a* and *cnnm2b* (Arjona *et al.* 2013) largely mirror the expression patterns of the alternatively spliced transcripts of human *CNNM2* (de Baaji *et al*. 2012).

Discussion

In 2005, Kopelman *et al*. demonstrated that a negative correlation between alternative splicing and gene duplication exists in the human genome. Later, Su *et al*. (2006) expanded upon this work and used the asymmetric evolution of alternative splicing between paralogs as evidence for subfunctionalization. Since then, a number of reports have been published that challenge whether the inverse correlation between alternative splicing and gene duplication is a consequence of subfunctionalization. For example, Roux and Robinson-Rechavi (2011) invoke a preduplication bias and progressive acquisition of splice forms to explain the inverse relationship between alternative splicing and gene duplication. In a similar manner, Grishkevich and Yanai (2014) suggest that the reduction in alternative splicing seen in large gene families is explained by the propensity of short genes and lowly expressed genes, both of which are less likely to be alternatively spliced, to undergo duplication at higher rates.

Here, by considering exon divergent paralogs apart from other genes we have focused on those duplicate pairs that are more likely to have been retained via subfunctionalization and as such

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

have more power to detect the consequences of subfunctionalization. It should be noted that we do not argue that all duplicates with different numbers of exons are the product of subfunctionalization. Indeed, recent research suggests that exon structure divergence after duplication is much too common for subfunctionalization to be the single driver of divergence (Xu *et al.* 2012; Wang *et al.* 2013). For example, processes such as intron/exon gain and loss and tandem exon duplication have the potential to cause exon structure divergence between paralogous genes, but would not necessarily be involved in splice form subfunctionalization. Nevertheless, we argue that changes in intron/exon structure after duplication should be accounted for and in doing so we show that paralogs with different numbers of exons undergo the lowest rates of alternative splicing, as measured by mean number of transcripts per gene as well as the proportion of genes with more than one transcript. We also show that exon divergent paralogs have restricted expression profiles, which is also consistent with subfunctionalization. Additionally, we demonstrate that the reduction in alternative splicing that accompanies exon structure divergence after gene duplication is free of all confounding factors that have previously been used to dismiss the relationship between alternative splicing and gene duplication.

Furthermore, we used the alternative splicing levels of single-copy orthologs that diverged before duplication to approximate the ancestral preduplication state. In this way, we compared the number of isoforms between human genes, either orthologous to zebrafish singletons, zebrafish nondivergent paralogs or zebrafish exon divergent paralogs. We find that the human orthologs of the zebrafish exon divergent paralogs display the highest rates of alternative splicing, indicating that the preduplication state may have been enriched in

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

alternatively spliced isoforms. This is compelling evidence in support of our hypothesis that exon divergence subsequent to duplication substitutes for alternative splicing.

It should be noted that by studying paralogs with divergent exon numbers we have focused on the subfunctionalization of genes that originally contained cassette exons. In truth, paralogs can have the same number of exons and still represent structural variants. However, this pattern indicates that the ancestral state (before duplication and subfunctionalization) included mutually exclusive exons or alternative 3' or 5' splice sites. We chose not to study these other types of alternative splicing because cassette exons represent the most common form of alternative splicing (Sorek *et al.* 2004; Pohl *et al.* 2013) and because detecting subfunctionalization of alternative 3' and 5' splice sites as well as mutually exclusive exons is computationally very difficult. By omitting other types of alternative splicing events, our estimates of the effects of subfunctionalization are likely conservative, meaning that gene duplication may reduce alternative splicing levels even more than reported here.

One point worth making is that the magnitude of the effect of subfunctionalization varies considerably between species. These differences are especially evident when comparing human and zebrafish. For example, in the human dataset, alternative splicing levels are reduced nearly 50% when comparing exon divergent and nondivergent paralogs. In zebrafish there is only a 6% reduction. This difference is likely the consequence of the relative contribution of gene duplication and alternative splicing to genome expansion in each lineage. Zebrafish have very high rates of gene duplication (Lu *et al.* 2012) and display very low levels of alternative splicing (Lu *et al*. 2010), while the inverse is true for humans (Shoja and Zhang 2006; Keren *et al*. 2010) . Therefore, it is not surprising that the effects of subfunctionalization will be

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

amplified in a genomic background replete with alternative splicing, yet relatively free of duplicate genes, as is the case with the human genome. Another possibility is that differences in gene annotation quality between the three species may be responsible for the differences apparent in our dataset. Ensembl's gene annotation pipeline relies heavily on the quality of the genome assembly, which can be assessed in several different ways, including scaffold and contig N50 as well as the number of ESTs and cDNAs aligned to the genome. The human genome has the highest quality assembly with substantially longer scaffold and contig N50 values, as well as many more aligned cDNAs and ESTs than either the mouse or zebrafish genomes. Considering that the human genome has the highest quality annotation, as well as the lowest proportion of exon divergent duplicates, it may be that some of the mouse and zebrafish exon divergent sibling pairs are the product of dubious annotation. However, the human genome demonstrates the largest effect of exon structure divergence between duplicates, meaning that we witness the greatest reduction in rates of alternative splicing in the human exon divergent duplicates. This suggests that any annotation artifacts present in the mouse and zebrafish dataset may obscure rather than enhance the effects of splice form subfunctionalization and demonstrates that the reported results of exon structure divergence between paralogs are not the product of poor annotation.

In conclusion, we find that exon divergent duplicate genes undergo the lowest levels of alternative splicing and have restricted expression profiles, which is consistent with subfunctionalization partitioning distinct isoforms among duplicates. Considering that duplicate gene retention via subfunctionalization is thought to be relatively rare (Wolfe and Semon 2008), it is not surprising that the relationship between gene duplication and alternative

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

splicing was not readily detected in previous genome-wide studies that treated all duplicates as equals.

Methods:

Sequence analysis:

All sequences were obtained from the Ensembl database release 75. Ensembl Compara was used to group all genes in each species as either paralogs or singletons. Paralogs in each species were then subject to a species specific reciprocal BLASTP search in an attempt to locate the most closely related sibling paralogs. BLAST e-values indicate that in each species most sibling paralogs are closely related. For example, 65% of the human, 66% of the mouse and 74% of zebrafish BLAST e-values for sibling paralogs were below E-100, while 86%, 85% and 90% were below E-50 in human, mouse and zebrafish, respectively. Ensembl BioMart (Smedley *et al.* 2009) was used to count the total number of transcripts, the number of protein coding transcripts and the number of exons for each gene. For independent estimates of the number of exons per gene we used the UCSC Genome Browser (Kent *et al.* 2002) to query RefSeq. In addition, we used ExoLocator (Khoo *et al.* 2014) to verify exon counts for a subset of genes in each species. Paralogs were classified as exon divergent if none of the transcripts contain equal numbers of exons. Information regarding duplication age, gene family size and length was obtained directly from Ensembl.

We used Ensembl Compara to identify orthologous preduplication outgroup genes in mouse, with which we tested for a preduplication bias in humans. For the exon divergent and nondivergent paralog groups we considered only genes that had duplicated since the

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

divergence of human and mouse. We used a BLASTP mutual best hit approach to verify co-ortholog assignments. In this way, we only considered human duplicate pairs that identified a common paralog in mouse by both tree reconciliation and mutual best hit. This same analysis was performed for zebrafish using human single-copy orthologs to estimate the ancestral state and for mouse, again using human orthologs to estimate the ancestral state.

We employed a two-step process to test for structural evidence of splice form subfunctionalization. First, exon sequences of protein coding transcripts for exon divergent zebrafish duplicates as well as the single human co-ortholog were aligned using tBLASTx. We used an expect value of E -8 as a cutoff for determining homologous exons between species, although most expect values were much lower. We then used blat and the UCSC Genome Browser to verify these alignments against both the human genome and the zebrafish genome. This analysis was performed on the 100 zebrafish exon divergent sibling pairs with the most closely related human co-ortholog.

Gene expression:

UniGene EST expression profiles were downloaded from UniGene (ftp://ftp.ncbi.nih.gov/repository/UniGene/, 4/18/2014). To estimate expression levels we queried the profile for each species. In humans there are 45 potential tissues that serve as sources for cDNA libraries. In mouse and zebrafish there are 43 and 8 potential tissue types, respectively. Following the methods of Chain *et al.* (2011) we calculated three summary statistics regarding gene expression; total expression levels (*T*), expression intensity (*I*) and evenness (*E*). The value *T* measures transcripts per million summed across all tissue restricted

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

libraries in which a gene is expressed ($T = \Sigma L_i$). It should be noted that genes that are highly expressed in one or a few tissues will have comparable *T* values to those gene that are expressed in many tissues at lower levels. To distinguish between the two we use the value *I*, which is calculated as a weighted average of expression levels: $I = \Sigma L_i^2 / \Sigma L_i$. The value $T/I = E$ indicates the breadth of a gene's expression profile and is used here as a measure of the effective number of tissues a gene is expressed in.

Acknowledgments:

The authors thank Dr. Ruth Phillips and Kersten Peterson for their thoughtful comments on the manuscript. MJL is supported by the Washington State University graduate student professional development award and the Washington State University College of Arts and Sciences Research and Creative Activity Enhancement Fund awarded to CDC. MJL conceived of the study and wrote the manuscript. MJL, WOC, KGO and BGW performed the experiments. CDC edited the manuscript.

Disclosure declaration:

The authors declare that there is no conflict of interest.

Figure Legends:

Figure 1. Gene duplication and exon structure divergence. (A) Stacked bar chart indicating the proportions of exon divergent paralogs, nondivergent paralogs and singletons in each species. A total of 14671, 14445 and 10115 genes comprise the datasets for human, mouse and zebrafish respectively. (B) A model of exon structure divergence. An alternatively spliced gene prior to duplication codes for both a long and a short transcript. After duplication the long transcript becomes fixed in duplicate 1, while the short transcript becomes fixed in duplicate 2. In this way, each duplicate codes for only half of the ancestral alternative splicing repertoire. The dashed lines represent alternative splicing and the solid lines represent constitutive splicing events that have become fixed after duplication.

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 2. The relationship between exon structure divergence among paralogs and levels of alternative splicing. The exon divergent paralogs demonstrate significantly lower levels of alternative splicing compared to nondivergent paralogs and singletons as measured by mean number of transcripts (A), and the proportion of genes that are alternatively spliced (AS) (B). Student's *t*-test was used to calculate the significance of the differences in mean number of transcripts. χ^2 was used to calculate the significance of the differences in proportion of genes that are alternatively spliced. The asterisks indicate the significance of the difference as compared to exon divergent paralogs, ** $p < .01$ * $p < .05$, the error bars represent SEM.

Figure 3. Family size and duplication age do not explain the reduction in alternative splicing. (A) Histograms comparing gene family size between exon divergent and nondivergent paralogs. In each species the nondivergent paralogs have larger mean family sizes. Student's *t*-test was used to calculate the significance of the differences in mean number of paralogs. The asterisks indicate the significance of the difference as compared to exon divergent paralogs, **$p < .01$ *$p < .05$, the error bars represent SEM. (B) Histograms showing the proportion of all exon divergent paralogs (black) created at each evolutionary epoch in each species. These values are compared against the proportion of nondivergent paralogs (white) produced at each epoch within each species (Note, for each species the black bars sum to 1, as do the white bars). Duplicate age increases along the x-axis (not to scale). These indicate that the exon divergent paralogs are not consistently enriched in young duplicates and that time since duplication is not a confounding effect.

Figure 4. The potential for a preduplication bias. Histograms comparing the levels of alternative splicing in single-copy orthologs that serve as a proxy for the ancestral state. (A) Human orthologs of zebrafish exon divergent paralogs (black) display higher levels of alternative splicing compared to the human orthologs of zebrafish nondivergent paralogs (white) and zebrafish singletons (grey). (B) Human orthologs of mouse exon divergent paralogs (black), nondivergent paralogs (white) and singletons (grey). (C) Mouse orthologs of human exon divergent paralogs (black), nondivergent paralogs (white) and singletons (grey). The asterisks indicate the significance of the difference as compared to the singleton orthologs of the exon divergent paralogs, **$p < .01$ *$p < .05$, the error bars represent SEM.

Figure 5. Exon divergent paralogs display restricted expression profiles. Exon divergent paralogs are expressed in fewer tissue types than nondivergent paralogs and singletons in each species. The p-values indicate the significance of the difference as compared to exon divergent paralogs. Student's *t*-test was used to calculate levels of significance. The asterisks indicate the significance of the difference as compared to exon divergent paralogs, **$p < .01$ *$p < .05$, the error bars represent SEM.

Figure 6. Structural evidence of splice form subfunctionalization. The zebrafish paralogs, *cnnm2a* and *cnnm2b*, correspond to distinct isoforms of the alternatively spliced orthologous human gene, *CNNM2*. The dashed box highlights the alternatively spliced cassette exon from the human gene that has been subfunctionalized in the zebrafish paralogs. The dashed lines connect homologous exons. Exons lengths are indicated either below or above exons. Introns are not to scale.

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

References:

Altschmied, J., Delfgaauw, J., Wilde, B., Duschl, J., Bouneau, L., Volff, J.N., and Schartl, M. 2002. Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. Genetics 161: 259–267.

Arjona FJ, Chen YX, Flik G, Bindels RJ, Hoenderop JG. 2013. Tissue-specific expression and in vivo regulation of zebrafish orthologues of mammalian genes related to symptomatic hypomagnesemia. Pflugers Arch. 465(10):1409-21.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. Annu. Rev. Biochem. 72, 291–336.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell. 16:1679–1691.

Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. Genome Biol. 7:R13.

Chain FJ, Dushoff J, Evans BJ. 2011. The odds of duplicate gene persistence after polyploidization. BMC Genomics 12;12:599.

Chen TW, Wu TH, Ng, WV, Lin WC. 2011. Interrogation of alternative splicing events in duplicated genes during evolution. BMC Genomics. 12 Suppl 3:S16.

De Baaji JH, Stuiver M, Meij IC, Lainez S, Kopplin K, Venselaar H, Muller D, Bindels RJ, Hoenderop JG. 2012. Membrane topology and intracellular processing of cyclin M2 (CNNM2). J Biol Chem. 287(17):13644-55.

Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW. 2006. Expression pattern shifts following duplication indicative ofsubfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. Mol Biol Evol. 23:469–478.

Force A, Lynch M, Pickett FB, Armores A, Yan Y, Postlewhait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151:1531-1545.

Grishkevich V, Yania I. 2014. Gene length and expression level shape genomic novelties. Genome Res. doi:10.1101/gr.169722.113

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered. 100(5):605-617.

Hellsten U, Khokha MK, Grammer TC, Harland RM, Richardson P, Rokhsar DS. 2007. Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog Xenopus laevis. BMC Biol. 5:31.

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication.Proc. Biol. Sci. 256, 119–124.

Hultman KA, Bahary N, Zon LI, Johnson SL. 2007. Gene duplication of the zebrafish kit ligand and the partitioning of melanocyte development functions to kit ligand a. PloS Genet. 3(1):e17.

Jin L, Kryukov K, Clemente JC, Komiya, Suzuki Y, Imanishi T, Ikeo K, Gojobori T. 2008. The evolutionary relationship between gene duplication and alternative splicing. Gene 427, 19–31.

Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel CV, McCarthy MI, Hide T, Hide W. 2003. eVOC: a controlled vocabulary for unifying gene expression data. Genome Res. 13(6A):1222-30.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002 The human genome browser at UCSC. Genome Res. 12(6):996-1006.

KerenH, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. Nat. Rev. Genet. 11(5), 345-55.

Khoo AA, Ogrizek-Thomas M, Bulovic A, Korpar M, Gurler E, Slijepcevic I, Sikic M, Mihalek I. 2014. ExoLocator—an online view into genetic makeup of vertebrate proteins. Nucleic Acids Res. 42(Database issue):D879-81.

Kopelman, NM, Lancet, D, Yanai, I 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat. Genet. 37, 588–589.

Lambert MJ, Olsen KG, Cooper CD. 2014. Gene duplication followed by exon structure divergence substitutes for alternative splicing in zebrafish. Gene 546, 271-276

Marshall AN, Montealegre MC, Jiménez-López C, Lorenz MC, van Hoof A. 2013. Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes. PLoS Genet. 9(3):e1003376.

Lu J, Peatman E, Tang H, Lewis J, Liu Z. 2012. Profiling of gene duplication patterns of sequenced teleost genomes: evidence for rapid lineage-specific genome expansion mediated by recent tandem duplications. BMC Genomics. 13:246-257.

Lu J, Peatman E, Wang WQ, Yang Q, Abernathy J, Wang SL, Kucuktas H, Liu ZJ. 2010. Alternative splicing in teleost fish genomes: same-species and cross-species analysis and comparisons. Mol Genet Genomics. 283:531-539.

Ohno S. 1970. Evolution by gene duplication. Berlin (Germany): Springer-Verlag.

Pohl M, Bortfeldt RH, Grutzmann K, Schuster S. 2013. Alternative splicing of mutually exclusive exons – a review. Biosystems. 114(1):31-8.

Pontius JU, Wagner L, Schuler GD. 2003. UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information.

Roux, J, Robinson-Rechavi M. 2011. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. Genome Res. 21, 357–363.

Downloaded from genome.cshlp.org on June 4, 2026 . Published by Cold Spring Harbor Laboratory Press

Sémon M1, Wolfe KH. 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. Proc Natl Acad Sci U S A. 105(24):8333-8.

Shoja V. Zhang L. 2006. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. Mol. Biol. Evol. 11, 2134-41.

Sorek R, Shemesh R, Cohen Y, Basechess O, Ast G, Shamir R. 2004. A non-EST based method for exon-skipping prediction. Genome Res. 14:1617-23.

Studer RA, Robinson-Rechavi M. 2009. How confident can we be that orthologs are similar, but paralogs differ? Trends Genet. 25:210-16.

Su Z, Wang, J, Yu J, Huang X, Gu X. 2006. Evolution of alternative splicing after gene duplication. Genome Res. 16:182–9.

Su Z, Gu X. 2012. Revisit on the evolutionary relationship between alternative splicing and gene duplication. Gene. 504(1):102-6.

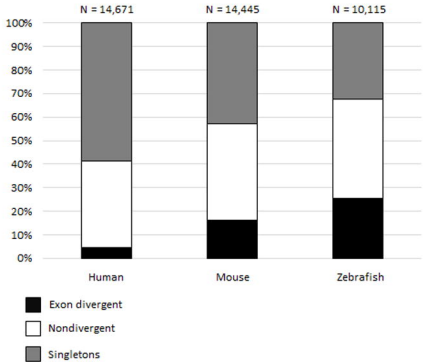
Talavera D, Vogel C, Orozco M, Teichmann SA, de la Cruz X. 2007. The (in)dependence of alternative splicing and gene duplication. PLoS Comput Biol. Mar 2;3(3):e33.

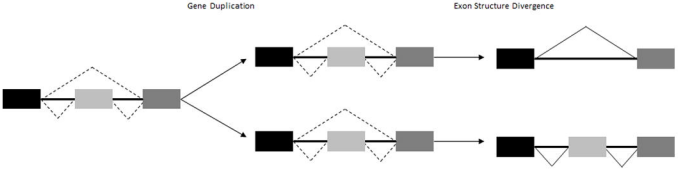
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19: 327–335.

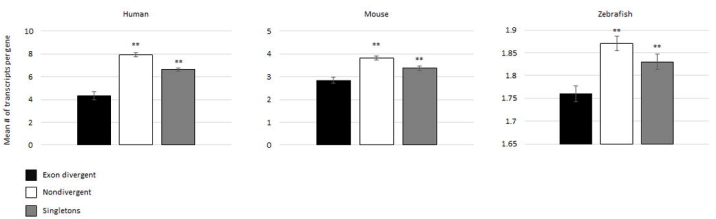
Wang Y, Tan X, Paterson AH. 2013. Different patterns of gene structure divergent following gene duplication in Arabidopsis. BMC Genomics. 14:652.

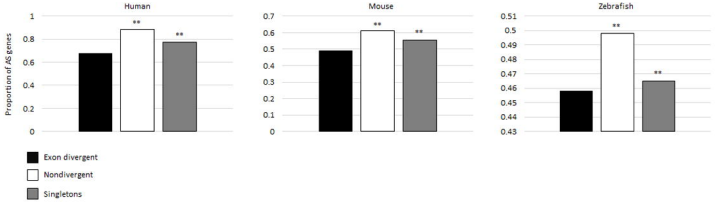
Xu G, Guo C, Shan H, Kong H. 2012. Divergent of duplicate genes in exon-intron structure. Proc Natl Acad Sci. 109(4):1187-92.

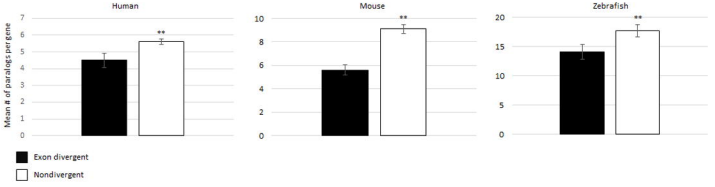
Yu WP, Brenner S, Venkatesh B. 2003. Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in Fugu. Trends Genet. 19, 180–183.


N = 14,671
N = 14,445
N = 10,115
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Human
Mouse
Zebrafish
Exon divergent
Nondivergent
Singletons

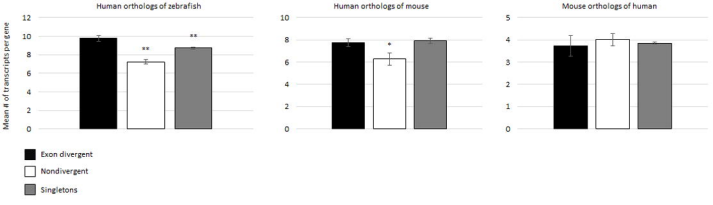

Gene Duplication
Exon Structure Divergence

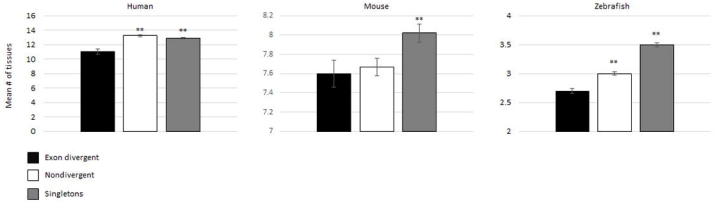

Human
Mouse
Zebrafish
Mean # of transcripts per gene
10
8
6
4
2
0
5
4
3
2
1
0
1.9
1.85
1.8
1.75
1.7
1.65
**
Exon divergent
Nondivergent
Singletons

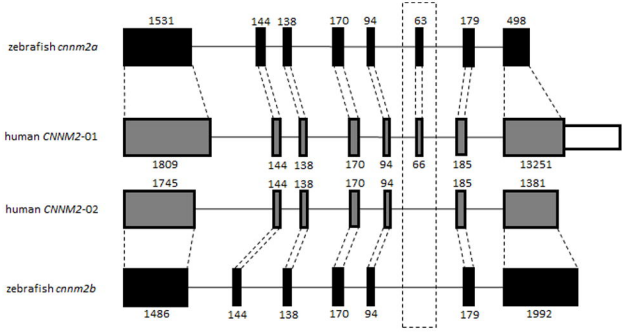

Human
Mouse
Zebrafish
Proportion of AS genes
1
0.8
0.6
0.4
0.2
0
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.51
0.5
0.49
0.48
0.47
0.46
0.45
0.44
0.43
**
Exon divergent
Nondivergent
Singletons


Human
Mouse
Zebrafish
Mean # of paralogs per gene
**
Exon divergent
Nondivergent

Human
Mouse
Zebrafish
Proportion of genes
Exon divergent
Nondivergent


Human orthologs of zebrafish
Human orthologs of mouse
Mouse orthologs of human
Mean # of transcripts per gene
12
10
8
6
4
2
0
**
**
10
8
6
4
2
0
*
5
4
3
2
1
0
Exon divergent
Nondivergent
Singletons


Human
Mouse
Zebrafish
Mean # of tissues
16
14
12
10
8
6
4
2
0
8.2
8
7.8
7.6
7.4
7.2
7
4
3.5
3
2.5
2
**
Exon divergent
Nondivergent
Singletons


zebrafish cnnm2a
1531
144
138
170
94
63
179
498
human CNNM2-01
1809
144
138
170
94
66
185
13251
human CNNM2-02
1745
144
138
170
94
185
1381
zebrafish cnnm2b
1486
144
138
170
94
179
1992