The Evolutionary Fate of Alternatively Spliced Homologous Exons after Gene Duplication

Federico Abascal*, Michael L. Tress*, and Alfonso Valencia

Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

*Corresponding author: E-mail: fabascal@cnio.es; mtress@cnio.es.

Accepted: April 27, 2015

Abstract

Alternative splicing and gene duplication are the two main processes responsible for expanding protein functional diversity. Although gene duplication can generate new genes and alternative splicing can introduce variation through alternative gene products, the interplay between the two processes is complex and poorly understood. Here, we have carried out a study of the evolution of alternatively spliced exons after gene duplication to better understand the interaction between the two processes. We created a manually curated set of 97 human genes with mutually exclusively spliced homologous exons and analyzed the evolution of these exons across five distantly related vertebrates (lamprey, spotted gar, zebrafish, fugu, and coelacanth). Most of these exons had an ancient origin (more than 400 Ma). We found examples supporting two extreme evolutionary models for the behaviour of homologous exons after gene duplication. We observed 11 events in which gene duplication was accompanied by splice isoform separation, that is, each paralog specifically conserved just one distinct ancestral homologous exon. At other extreme, we identified genes in which the homologous exons were always conserved within paralogs, suggesting that the alternative splicing event cannot easily be separated from the function in these genes. That many homologous exons fall in between these two extremes highlights the diversity of biological systems and suggests that the subtle balance between alternative splicing and gene duplication is adjusted to the specific cellular context of each gene.

Key words: alternative splicing, gene duplication, protein diversity, homologous exons, subfunctionalization.

Introduction

Alternative splicing (AS) and gene duplication (GD) are two of the main mechanisms behind the diversification of protein function. Both can increase the numbers of proteins coded within genomes; GD creates initially redundant copies of genes that with time, and following different possible evolutionary paths, can diversify in sequence and function (Conant and Wolfe 2008; Innan and Kondrashov 2010), whereas AS allows genes to code for more than one distinct protein from the same locus (Smith and Valcárcel 2000). The relationship between GD and AS is not well understood, so analyzing the interconnection between the two processes may provide insights into their relative importance in the generation of new protein products.

As GD and AS are both repositories of protein diversity, interplay between the two can be expected. According to the interchangeable model (I-model), or function-sharing model, alternative isoforms that were originally coded within a single gene may separate into different genes after a GD event by means of differential retention of AS patterns in each duplicate gene. The extreme case would be the subfunctionalization of gene duplicates. Here, AS and GD might be regarded as interchangeable repositories of protein diversity. This model has received support from 1) genome-wide analyses reporting a negative correlation between AS and the size of protein families (Kopelman et al. 2005; Su et al. 2006; Talavera et al. 2007), and 2) reports of acceleration of AS divergence after GD (Zhang et al. 2010; Xu et al. 2012), although the validity of some of these results is debated (Talavera et al. 2007; Roux and Robinson-Rechavi 2011; Su and Gu 2012). More recently, Lambert et al. (2015) analyzed exon divergence of zebrafish gene duplicates that are co-orthologs of human genes. Although their analysis does support a general trend of splice isoform separation after GD, their results must be treated with caution as they were based on the comparison of heterogeneous transcriptome annotations that, in the case of zebrafish at least, are far from complete.
In the noninterchangeable (NI-model), the AS-encoded protein diversity is not distributed among gene duplicates. The underlying implication of this model is that it may not be favorable to separate alternative isoforms into different genes, which may indicate that AS in these genes is not just a means to encode protein diversity but also of controlling their expression. The importance of AS in these genes may be related to the balanced production of isoforms or other kinds of regulation linked to the splicing process that may not be attainable with independent genes. In contrast to the I-model or “Function-sharing model,” the NI-model has not been thoroughly investigated. The natural prediction of the NI-model is that alternative exons will be preserved by purifying selection after GD events.

To study the relationship between GD and AS we concentrated on characterizing the evolutionary conservation of mutually exclusive homologous exons (MEHEs), defined here as duplicated exons that are incorporated into alternatively spliced transcripts in a mutually exclusive manner. We chose to focus on MEHEs because they are potentially the most biologically relevant type of AS (Ezkurdia et al. 2012), and because they are particularly adequate for the systematic comparison of isoforms after GD.

Like gene duplicates, MEHEs can evolve new functions or experience a subfunctionalization process within the context of a single gene. As long as these alternative MEHEs have evolved different functions, full subfunctionalization may occur if each of the ancestral MEHEs is retained in a different gene after GD. Indeed, a handful of examples of subfunctionalization driven by splice isoform separation have been reported in the literature (Altschmied et al. 2002; Yu et al. 2003; Pacheco et al. 2004; Cusack and Wolfe 2007; Hultman et al. 2007; Marshall et al. 2013). Unfortunately, a study of true subfunctionalization is not possible in silico, because only experimental evidence can confirm that two different sequences have two different cellular functions. So here we have used the separation of MEHEs among gene duplicates as a proxy for subfunctionalization. As we cannot be sure that the separation of MEHEs is genuine subfunctionalization, the process of separating homologous exons after GD is referred to here as splice isoform separation.

We carefully curated a list of human MEHEs, most of which are predicted to be relevant on the basis of evolutionary conservation, and analyzed the conservation of MEHEs using sequence similarity searches in five distantly related vertebrate species, including lamprey, fugu, zebrafish, spotted gar, and coelacanth. Within this data set we focused on GD events to assess the prevalence of the NI- and I-models. We identified cases of splice isoform separation by looking for differential conservation of MEHEs among gene duplicates. We also identified genes in which MEHEs were preserved after duplication. We discuss the biological implications of each model.

Materials and Methods

We explored the human genome using Ensembl version 75 from February 2014 (Flicek et al. 2013) and compared CCDS annotations to identify genes with MEHEs. CCDS annotations represent high-quality transcript annotations for which the EBI, the NCBI (National Center for Biotechnology Information), the WTSI, and the UCSC (the University of California–Santa Cruz) reached a consensus (Pruitt et al. 2009). Although CCDS annotations are not complete, restricting the set of transcripts to these cases avoids including rare and low frequency human transcript variants. We discarded those CCDS that were part of another CCDS. We found that 5,322 genes contained more than one nonredundant CCDS. We sorted the resulting CCDS by length and defined the longest one as reference. Other transcripts were compared against the reference transcript to identify coding exons that code for at least ten amino acids and are present in one transcript but not in the other, that is, that are MEES. We then identified whether the resulting pairs of MEES were homologous (MEHEs) using BLAST v2.2.25 (Altschul et al. 1997) comparisons, with an e value threshold of 0.005. To simplify the analysis and to avoid the inclusion of false positives related to annotation problems, we restricted the set to those genes for which we identified just one pair (or set) of MEHEs in the reference transcript. The final set contained 97 genes.

Five vertebrate species, all distantly related to human, were selected to explore the evolutionary conservation of MEHEs. Selected taxa included lamprey (Smith et al. 2013), spotted gar (Amores et al. 2011), zebrafish (Howe et al. 2013), fugu (Aparicio et al. 2002), and coelacanth (Amemiya et al. 2013), and were retrieved from Ensembl v75. The genomes of these target species were scanned using TBLASTN without low complexity filtering (−F F) and with an e-value threshold of 0.1 to find similarity matches to query human MEHEs. We merged overlapping similarity hits using bedtools v2.2.17 (Quinlan and Hall 2010) and determined whether they overlapped (or were close enough to) annotated genes. To set a distance threshold for assigning nonoverlapping hits to neighbor genes, we calculated the 95 percentile of gene lengths for each species and required a hit to be closer than that threshold (this threshold is highly variable between target species: 27,992, 41,583, 96,255, 109,349, and 151,632 bp for fugu, lamprey, spotted gar, zebrafish, and coelacanth, respectively). We carefully reviewed those cases in which similar hits were ambiguously assigned to multiple neighbor genes. Finally, we identified those cases in which multiple nonoverlapping hits belonged to the same gene, as these cases are candidates for having conserved MEHEs.

For genes with multiple similar hits (usually 2) to the query MEHEs, we determined whether each hit was most similar to each of the query MEHEs. In addition, we determined whether the genes to which hits were assigned were orthologs of the query human gene according to EnsemblCompara (Vilella et al. 2013). We also compared against the reference transcript to identify coding exons that code for at least ten amino acids and are present in one transcript but not in the other, that is, that are MEES. We then identified whether the resulting pairs of MEES were homologous (MEHEs) using BLAST v2.2.25 (Altschul et al. 1997) comparisons, with an e value threshold of 0.005. To simplify the analysis and to avoid the inclusion of false positives related to annotation problems, we restricted the set to those genes for which we identified just one pair (or set) of MEHEs in the reference transcript. The final set contained 97 genes.

Five vertebrate species, all distantly related to human, were selected to explore the evolutionary conservation of MEHEs. Selected taxa included lamprey (Smith et al. 2013), spotted gar (Amores et al. 2011), zebrafish (Howe et al. 2013), fugu (Aparicio et al. 2002), and coelacanth (Amemiya et al. 2013), and were retrieved from Ensembl v75. The genomes of these target species were scanned using TBLASTN without low complexity filtering (−F F) and with an e-value threshold of 0.1 to find similarity matches to query human MEHEs. We merged overlapping similarity hits using bedtools v2.2.17 (Quinlan and Hall 2010) and determined whether they overlapped (or were close enough to) annotated genes. To set a distance threshold for assigning nonoverlapping hits to neighbor genes, we calculated the 95 percentile of gene lengths for each species and required a hit to be closer than that threshold (this threshold is highly variable between target species: 27,992, 41,583, 96,255, 109,349, and 151,632 bp for fugu, lamprey, spotted gar, zebrafish, and coelacanth, respectively). We carefully reviewed those cases in which similar hits were ambiguously assigned to multiple neighbor genes. Finally, we identified those cases in which multiple nonoverlapping hits belonged to the same gene, as these cases are candidates for having conserved MEHEs.

For genes with multiple similar hits (usually 2) to the query MEHEs, we determined whether each hit was most similar to each of the query MEHEs. In addition, we determined whether the genes to which hits were assigned were orthologs of the query human gene according to EnsemblCompara (Vilella et al. 2013).
et al. 2009). We also annotated whether query and target genes were part of the same phylogenetic tree in EnsemblCompara and whether, if not orthologous, the alternative paralogous relationship was set as confident according to the EnsemblCompara pipeline (we determined whether the gene is considered an ortholog of a different human gene that the query gene with confidence). Uncertain cases were carefully reviewed.

To date the origin of the MEHEs, we assumed that when two species share a pair of MEHEs these have not been acquired independently. This is equivalent to inferring ancestral character states with Dollo parsimony (Farris 1977). Phylogenetic analyses (particular cases) and the degree of similarity (in general) support this assumption. In certain cases, the presence of human paralogs with the same MEHEs allowed dating the evolutionary origin of MEHEs at the corresponding GD event, which may be an older age than that inferred looking at the presence of MEHEs in the five target species.

We conducted detailed evolutionary analyses to validate and characterize potential splice isoform separation cases, that is, cases in which gene duplicates retained or lost different MEHEs. Multiple sequence alignments were built with MAFFT v7.123b (Katoh and Standley 2013), handled and visualized with Jalview (Waterhouse et al. 2009). Phylogenetic trees were reconstructed for exons and/or genes under maximum likelihood with Phyml v3.1 (Gouy et al. 2010; Guindon et al. 2010), using 1,000 replicates of nonparametric bootstrapping and choosing the best-fit model of evolution with ProtTest v2.4 (Abascal et al. 2005). The selection of taxa varied depending on each particular case (alignments and trees are available from the author upon request). Tree figures were prepared with FigTree (http://tree.bio.ed.ac.uk/software/figtree/, March 2015).

Results

AS of MEHEs Is Highly Conserved in Vertebrates

We identified 97 genes with a single set of MEHEs from among the set of human CCDS (consensus coding sequences) transcripts (see Materials and Methods). To assess the evolutionary conservation of the corresponding AS events, we relied on direct sequence searches against target genomes with TBLASTN rather than comparing annotations of the isoforms in the corresponding species because the gene annotations of all species apart from human are still not close to being complete. For each BLAST hit we determined whether it corresponded to annotated or new genes and/or exons, and whether the corresponding genes were considered orthologs or paralogs of the query human gene in the EnsemblCompara database. We carefully analyzed each of the cases.

We assessed the validity of relying on sequence similarities rather than on comparison of transcript annotations to trace the evolution of MEHEs across species. Careful curation revealed that in a few cases (4) the MEHEs were conserved even though TBLASTN was not able to detect them, mainly because these exons were too short or highly divergent. Despite this, our assessment of transcript annotation qualities in target species showed that sequence-based approaches are still much better. We found that transcript annotations are usually incomplete and, importantly, of very heterogeneous quality across species. We estimated the number of nonannotated genes and exons (see supplementary material, Supplementary Material online) and found that although 93.7% of the MEHEs identified with TBLASTN were annotated in fugu, only 53.6% of the TBLASTN-identified MEHEs were annotated in lamprey (supplementary table S1, Supplementary Material online).

For 84 of 97 genes we found that both MEHEs are present in at least one of the five target species (fig. 1 and supplementary table S1, Supplementary Material online). We determined that for almost all of the cases orthologous relationships could be established between each of the target MEHEs and each of the query MEHEs, implying that the large majority of MEHEs have not duplicated independently in different lineages (see supplementary material, Supplementary Material online). Hence, we can infer that these 84 MEHEs (or the majority of them) originated at least 400 Ma.

In 41 of the 84 cases, the MEHEs were conserved in all four species of jawed vertebrates. MEHE conservation reached lamprey in 28 genes, despite the distant relationship between lamprey and human (~500 Ma; Kumar and Hedges, 2011). Up to 80 cases have been conserved in at least one bony fish, more frequently in spotted gar (77 cases) than in fugu and zebrafish (56 and 54, respectively; fig. 1B). The larger number of losses in teleosts is probably the result of the whole-genome duplication experienced in their ancestor.

We carefully revised the 13 cases for which no conservation was detected in any of the target species to check whether the MEHEs appeared later in the human lineage or whether the lack of significant sequence similarity was due to low sequence conservation and/or exons that were too short. We found that 4 of these 13 cases were indeed present in at least one of the target species. Consequently, we ended with a total of 88 of 97 cases of human MEHEs of ancient origin (90.7%). We did not include these four cases as part of the comparative analysis because we have no objective way to establish their conservation across the five target species.

Splice Isoform Separation by Retention of Different Homologous Exons after GD

We identified cases in which alternative isoforms ancestrally coded by a single gene became separated into different genes by means of GD coupled to differential loss and conservation of MEHEs in each paralog. In such cases, protein diversity initially encoded through AS becomes distributed in different
genes, supporting the I-model. We found a total of ten cases of this kind (Table 1), nine of which are new. We also identified a case (CUX1) of differential conservation of nonhomologous mutually exclusive exons (MEEs). Among the 11 cases, the following 7 experienced complete splice isoform separation: CALU, CUX1, MARVELD3, PGM1, PDLIM3, RNF128, and U2AF1. In the remaining four cases (CACNA1C/D, CDC42, FYN, and SLC8A3) splice isoform separation was detected between two paralogs but at the same time other paralogs conserved the ancestral pattern of AS of MEHEs. These 11 cases

**Table 1**
The 11 Cases in Which Each Duplicated Gene Lost or Retained One of the Ancestral MEHEs in a Concerted Manner (Splice Isoform Separation) Are Shown, Indicating Which Genes and Lineages Are Affected

| Human Gene | Origin of MEHEs | Human Exons (GRCh38) | Differential Conservation of Ancestral MEHEs in Lineage (Genes) |
|------------|-----------------|----------------------|-------------------------------------------------------------|
| CACNA1C, CACNA1D | Vertebrates (12:2504435–2504539, 12:2504841–2504945; 12:2633628–2633712, 12:2634296–2634374) | Vertebrates (CACNA1S and CACNA1F) | 
| CALU | Jawed vertebrates (7:128754455–128754528) | Vertebrates (CALUA and CALUB) | 
| CDC42 | Jawed vertebrates (1:22091427–22091517) | Zebrafish (CDC42L and CDC42L2) | 
| CUX1 | Bilaterians (7:101816011–102258233) | Vertebrates (many genes, e.g., FRK vs. SRC, YES1 . . . ) | Lamprey, spotted gar, zebrafish, fugu (also other vertebrates) |
| FYN | Chordates (6:111700103–111700268) | Teleosts (PGM1 and PGM5) | 
| MARVELD3 | Vertebrates (16:71640389–71641027, 16:71643192–71643480) | Zebrafish and cave fish (Otophysa) (RNF128A and ENSDARG00000029890) | 
| PDLIM3 | Chordates (?) (4:185508298–185508562, 4:185514702–185514890) | Platypus (ENSOANG00000006867, ENSOANG00000013438) | 
| PGM1 | Vertebrates (?) (6:13623460–63623760) | Spotted gar, coelacanth . . . (SLC8A4b, SLC8A2a) | 
| RNF128 | Vertebrates (14:70060835–7006939, 14:70063822–7006929) | Zebrfish and cave fish (Otophysa) (RNF128A and ENSDARG00000029890) | 
| SLC8A3 | Vertebrates (21:6493043–6493110, 21:6492130–6492197) | Fugu, tilapia and stickleback (Percomorphaceae?) (U2AF1 and ENSTRUG00000013815) | 
| U2AF1 | Vertebrates (21:6493043–6493110, 21:6492130–6492197) | Fugu, tilapia and stickleback (Percomorphaceae?) (U2AF1 and ENSTRUG00000013815) | 

**NOTE.**—Genes in bold indicate cases undergoing complete splice isoform separation.

*CUX1* is a case of a homologous but not of nonhomologous MEEs.
represent a very significant increase with respect to those reported in the literature. The cases of PGM1 and CUX1 are described in the supplementary material, Supplementary Material online (supplementary figs. S1 and S2, Supplementary Material online), whereas CALU, MARVELD3, and CACNA1C/CACNA1D are described below. Splice isoform separation of PDLIM3 occurred in platypus and splice isoform separation (and subfunctionalization) of U2AF1 has been already described in the literature (Pacheco et al. 2004).

The CALU gene is ubiquitously expressed and encodes a protein (calumenin) distributed throughout the secretory pathway (Vorum et al. 1999), known to inhibit vitamin-K-dependent protein carboxylation (Wajih et al. 2004) and involved in protein sorting and folding (Tsukumo et al. 2009; Wang et al. 2012). CALU contains six calcium-binding EF-hand domains, the first of which is coded by one of two MEHEs (fig. 2A and B). Little is known about the functional role of the splicing of MEHEs exons in CALU. Calumenin MEHEs may be differentially expressed in human primary tumors (Dutertre et al. 2010). This, together with the observation that CALU is a phosphorylation substrate of v-Src (Shah and Shokat 2002), suggests that it may participate in signal transduction pathways related to transformation (Honoré 2009).

Genomic BLAST revealed similarities specific to each of the human alternative exons within the corresponding CALU loci of spotted gar and coelacanth, allowing us to date the origin of the MEHEs of CALU to the ancestor of jawed vertebrates. We found that the pattern of AS was lost in fugu and zebrafish. There are two orthologs of human CALU in zebrafish (CALUA and CALUB), which originated from a duplication event in the ancestor of teleosts (one of these duplicates was later lost in fugu; fig. 2C). Interestingly, each zebrafish ortholog specifically retained one of the ancestral alternative exons while losing the other. By exploring other species that present multiple orthologs to human CALU we also found differential exon losses in all other teleosts but tetraodon.
and stickleback, which, as fugu, lost one of the duplicated genes (fig. 2C). Hence, the process of splice isoform separation took place in the ancestor of teleosts right after the duplication of CALU.

MARVELD3 belongs to the occludin family, whose members are components of tight junctions (Steed et al. 2009) and share a MARVEL domain that contains four transmembrane helices and is typically involved in membrane apposition events (Sánchez-Pulido et al. 2002). MARVELD3 acts by coupling tight junctions to the MEKK1–JNK (c-Jun-N-terminal kinase) pathway, so determining cell behavior and survival (Steed et al. 2009). Indeed, MARVELD3 is downregulated during epithelial–mesenchymal transition in human pancreatic cancer cells (Kojima et al. 2011) and loss of MARVELD3 expression increases cell migration and proliferation, whereas re-expression reverts the metastatic phenotype (Steed et al. 2009).

Our analysis revealed a complex evolutionary history for MARVELD3, and we had to consider other vertebrates to clarify it. The pattern of AS, previously reported as specific to mammals (Steed et al. 2009), is also observed in coelacanth and Xenopus. In all vertebrates but mammals, that is, in lamprey, ray-finned fishes, coelacanth, Xenopus and reptiles (including birds), there are two MARVELD3 genes instead of one. With the exception of coelacanth and Xenopus, species with duplicated MARVELD3 show no AS. Interestingly, the phylogenetic reconstruction of these exons reveals two clearly defined lineages (groups of orthology), each covering the whole set of analyzed vertebrates. In species with duplicated genes, each of the two separated exons maps to a different group of orthology. In species with AS, each alternative exon maps to each group of orthology. In coelacanth and Xenopus both things happen, as one of their duplicated genes conserved the AS pattern. Although other alternative hypotheses could be proposed, we believe that the most parsimonious interpretation for this complex scenario is that originally, in the ancestor of vertebrates, MARVELD3 acquired the pattern of AS. Then, this ancestral gene duplicated and one of the paralogs lost the pattern of AS. Later, after the split of the major vertebrate lineages, some lineages lost the paralog that had no AS (mammals) whereas other lineages lost one of the AS isoforms from the paralog that did have AS (fig. 3). According to the phylogenetic tree, splice isoform separation occurred at
least three times (at the ancestors of lamprey, ray-finned fishes, and reptiles) whereas a single gene loss event took place in the ancestor of mammals. Remarkably, despite several gene and exon losses, both original splice isoforms have been always kept, either within the same or different genes, which might be taken as an indication of their biological relevance and functional independence.

AS of MEHEs Is Conserved in Human Paralogs

The cases of differential conservation of MEHEs in duplicated genes reveal that the protein diversity encoded with AS can be distributed between independent genes. To explore the validity of the alternative NI-model, we looked for cases in which MEHEs were conserved between paralogs after GD.

We identified 21 clusters of human paralogs, comprising 54 genes, with the same pattern of MEHEs (table 2). The great majority of these paralogs duplicated a long time ago, in the ancestor of jawed vertebrates or earlier. These MEHEs are of special interest because they have ancient origins and have been conserved along different gene lineages. Hence, these are genes for which AS may be resilient to GD and support the NI-model. The following examples illustrate how relevant the AS of MEHEs of these genes might be.

The strongest support for the NI-model comes from the JNKs, AMPA glutamate receptors, and myocyte-specific enhancer proteins (MEF2s). In these cases, MEHEs were conserved in all the members of their families despite ancient GD events. In the case of JNKs (MAPK8, MAPK9, and MAPK10), the MEHEs code for part of the kinase domain (fig. 4). The biological significance of these MEHEs may relate to different ligand-binding specificities (Gupta et al. 1996), but remains unclear (Seki et al. 2012). In the case of AMPA glutamate receptors (GRIA1, GRIA2, GRIA3, and GRIA4), MEHEs already existed in the ancestor of vertebrates (already reported in Chen et al. 2006) and code for the flip and flop exons (supplementary fig. S5, Supplementary Material online). Although these exons have almost identical amino acid sequences, their use yields important functional variations (Partin et al. 1996).

The importance of AS is particularly clear in the case of human ACTN2 and ACTN4 genes, which code for alpha-actinin 2 and 4, respectively. Alpha-actinins are important cytoskeletal proteins with multiple roles and many interacting partners. Interestingly, some of these partners also have MEHEs, for example PDLIM3, with MEHEs that affect a region involved in actinin-binding. The actinin family has two pairs of MEHEs that are distant in sequence, but close in the dimeric structure (fig. 5B). In human, ACTN4 has both pairs of MEHEs, ACTN1 and ACTN2 each share a different pair of MEHEs with ACTN4, and ACTN3 has no MEHEs.

Importantly, the MEHEs in ACTN2 and ACTN4 (and also in ACTN1 in fugu, spotted gar and coelacanth, but not in human or mouse ACTN1) have particularly ancient ancestry. Fruitfly and Caenorhabditis elegans have the same pattern of AS, which allows dating the origin of these MEHEs back to the ancestor of bilaterians (Barstead et al. 1991). This clearly points toward a key functional role of AS for alpha-actinins.

We identified other interesting examples, like the paralogs of the SCN2A gene or the fibroblast growth factor receptors, which are described in the supplementary material, Supplementary Material online (supplementary figs. S3–S5, Supplementary Material online).

The Complex Case of the CACNA1 Family MEHEs

The genes CACNA1C, CACNA1D, CACNA1S, and CACNA1F code for alpha subunits of voltage-gated calcium channels. These four paralogs form a monophyletic group that originated from GDs in the ancestor of jawed vertebrates. There are two pairs of MEHEs whose origin also dates back to the ancestor of vertebrates, but predating the GD events. The genes CACNA1C and CACNA1D both conserved the two pairs of ancestral MEHEs (fig. 6), whereas CACNA1F and CACNA1S experienced a process of loss/retention of different ancestral exons that, interestingly, affected the two pairs of MEHEs (fig. 6). CACNA1C presents an additional pair or MEHEs that may have evolved later in the ancestor of sarcopterygians, as it is conserved in coelacanth, Xenopus, and mammals.
| Human Paralogs | Description | Duplication Ancestor | Region Affected and AS Role |
|---------------|-------------|----------------------|-----------------------------|
| ACSL1, ACSL6  | Acyl-CoA synthetase long-chain | Jawed vertebrates     | Internal                     |
| ACTN1, ACTN2, | Alpha-actinin | Vertebrates           | Two pairs of internal MEHEs, Actin-binding domain (fig. 4), Tissue specificity (Waites et al. 1992) |
| ACTN4         |             |                      |                             |
| ASIC1, ASIC2  | Acid-sensing ion channel | Vertebrates           | 5 prime, N-terminus and first transmembrane helix of the channel |
| CACNA1A,     | Voltage-dependent L-type calcium channel subunit alpha-1 | Vertebrates           | Internal, Cytoplasmic C-terminal region, Fine tuning of channel properties (Lipscombe et al. 2013) |
| CACNA1B,     |             |                      |                             |
| CACNA1E      |             |                      |                             |
| CACNA1C, CACNA1D | Voltage-dependent L-type calcium channel subunit alpha-1 | Vertebrates           | Two pairs of internal MEHEs, End of first ion transport domain, beginning of last ion transport domain |
| CLDN10, CLDN18 | Claudin    | Vertebrates, MEHEs also found in C. savignii | 5 prime, PMP22_Claudin domain, Permeability for anions or cations (Günzel et al. 2009) |
| CYP4F2, CYP4F3 | Cytochrome P450, family 4, sub-family F | Catarrhini | Internal, Beginning of p450 domain |
| CYP4F2       |             |                      |                             |
| DEF8110, DEF8119 | Beta-defensin | Amniotes          | 3 prime, A signal peptide is shared between isoforms, with the extracellular domain, with many conserved Cys, is alternatively spliced |
| DNM1, DNM2   | Dynamin     | Vertebrates         | Internal, Dynamin_M domain |
| FGFR1, FGFR2, FGFR3 | Fibroblast growth factor receptor | Vertebrates, jawed vertebrates, MEHEs also found in tunicates | Internal, C-terminal half of the third Ig-like domain, Interaction with FGF and heparan sulfate proteoglycans (Olsen et al. 2004) |
| GNAL, GNAS   | Guanine nucleotide-binding protein | Jawed vertebrates | 5 prime, N-terminal region predicted disordered and beginning of G-alpha domain |
| GRIA1, GRIA2, GRIA3, GRIA4 | AMPA glutamate receptor | Vertebrates | Internal, Ligand-gated ion channel domain, Channel-gating kinetics (Partin et al. 1996) |
| ITGA3, ITGA6 | Integrin alpha | Vertebrates         | 3 prime, Cytoplasmic C-termini, Interaction with HPSS (Fukushi et al. 2004), Tissue specificity (De Melker et al. 1997) |
| MAPK8, MAPK9, MAPK10 | Mitogen-activated protein kinase/ JNK. | Vertebrates | Internal, Kinase domain, Different affinities for ATF-2, Elk-i and Jun transcription factors (Gupta et al. 1996) |
| MEF2A, MEF2C, MEF2D | Myocyte-specific enhancer factor. | Jawed vertebrates | Internal, Holliday junction regulator protein family C-terminal repeat |
| NRG1, NRG2   | Pro-neuregulin | Jawed vertebrates | Internal, Tissue specificity, cell localization, etc. (Liu et al. 2011) |
| PDLIM3, LDB3 | PDZ and UM domain protein 3 (ALP), UM domain-binding protein 3 (Enigma) | Chordates? (not in the same Ensembl tree) | Tissue specific AS affecting the small ZM domain responsible for alpha-actinin-2 binding (Faulkner et al. 1999) |
| SCN2A, SCN3A, SCN5A, SCN8A, SCN9A | Sodium channel protein subunit alpha | Amniotes, vertebrates | Internal, Beginning/middle of first ion transport domain, Developmental and tissue specificities (Gazina et al. 2010) |
| SLC44A2, SLC44A5 | Choline transporter-like protein | Vertebrates | 3 prime, Cytoplasmic C-terminal tail |
| SLC8A1, SLC8A3 | Sodium/calcium exchanger | Vertebrates | Internal in cax-beta motif. May modulate the dynamic properties of Ca2+ sensing (Khananshvili 2013) |
| TPM1, TPM2, TPM3, TPM4 | Tropomyosin alpha chain | Vertebrates | Several: 5 prime, internal, 3 prime, Developmental and tissue specificities (reviewed in Gunning et al. 2005) |

NOTE.—Groups in bold indicate cases in which all the paralogs descending from the last GD event conserved the ancestral MEHEs.
This case is particularly interesting because although conservation of AS within CACNA1C and CACNA1D supports the NI-model, the pattern of exon losses in CACNA1F and CACNA1S, which may be seen as a case of splice isoform separation, supports the I-model. A similar pattern occurs with the MEHEs of genes SLC8A3 and SLC8A1. These examples illustrate the complexity of the interaction between AS and GD. The evolutionary fate of MEHEs after GD may
depend on subtle characteristics of each gene and extreme models may not be realistic.

Discussion
We have determined that the AS of mutually exclusively spliced homologous exons is highly conserved in vertebrates. We found evidence of ancient ancestry (>400 Ma) for about 91% of the human MEHEs. Other studies that have compared AS across mammals have provided substantially lower estimates of evolutionary conservation of splicing events (Modrek and Lee 2002; Boue et al. 2003; Thanaraj et al. 2003; Pan et al. 2005; Yeo et al. 2005; Mudge et al. 2011). Modrek and Lee estimated that only 25% of all “minor” alternative exons (regardless of splice type) were conserved between mouse and human. This is in sharp contrast to the conservation of MEHEs, in particular since our taxa selection comprised species that are much more distantly related than human and mouse. The difference in conservation suggests that MEHEs are likely to be much more functionally relevant than other types of alternative exons. The relevance of MEHEs is also supported by strong evidence indicating that the corresponding alternative isoforms reach the protein level much more frequently than would be expected based on the background frequencies of annotated AS events in the transcriptome (Ezkurdia et al. 2012).

For the reasons stated above, MEHEs are particularly amenable to study the relationship between GD and AS and explore the validity of two extreme models. Although we found support for both the I- and NI-models, many cases fell between these two extremes (e.g., one of the duplicated genes conserved the two MEHEs, whereas the other lost one MEHE), reflecting subtle differences in the relative importance of AS, as might be expected from the diversity of biological systems and the large divergence times. Indeed, there were genes and splicing events that provided evidence for both extreme models. For example, within the family of CACNA1C, CACNA1D, CACNA1F, and CACNA1S, two of the paralogs (CACNA1C and CACNA1D) support the NI-model based on conservation of ancestral MEHEs, whereas the other two paralogs (CACNA1F and CACNA1S) support the I-model based on the process of complementary loss/retention of alternative exons.

Support for the NI-model, in which protein diversity encoded by AS is not distributed among gene duplicates, came from 21 groups of human paralogs in which some or all duplicated genes conserved ancestral patterns of AS for long evolutionary periods. The best examples may be those of the JNKs and AMPA glutamate receptors. In both cases, all the multiple paralogs are related by ancient GD events but have conserved the ancestral MEHEs. The NI-model suggests that the control of expression by AS has a role that is tightly linked with the biological function of the gene.

Support for the I-model came from 11 examples in which each gene duplicate specifically retained one of the two ancestral MEHEs, that is, of splice isoform separation. As a result of the process of concerted loss and retention of ancestral MEHEs, the net protein diversity is conserved but distributed among different genes. For some genes, there may even be advantages to separating the alternative isoforms. At the very extreme of this model we found MARVELD3, for which the splice isoform separation process may have taken place independently in at least three different lineages.

Although we conducted no experimental confirmation, long-standing conservation of MEHEs very likely reflects the existence of functional differences between the homologous exons. If true, the identified cases of splice isoform separation would support a process of subfunctionalization in which ancestral functions have been partitioned between paralogs. The reported cases would add to the handful of cases of this kind reported in the literature (Altschmied et al. 2002; Yu et al. 2003; Pacheco et al. 2004; Cusack and Wolfe 2007; Hultman et al. 2007; Marshall et al. 2013). Evolutionary analysis may explore how these gene duplicates evolved once ancestral isoforms were uncoupled, and whether this uncoupling affected the evolution of accompanying constitutive exons and/or eventually had an adaptive value. In practical terms, having each human isoform represented by a distinct gene in a target species may facilitate the experimental characterization of each isoform function by, for instance, specific gene knockout experiments or gene expression analysis.

The curated analysis presented here throws light on the general aspects of the complex interplay between GD and AS as repositories of protein diversity, and also represents a guide for bettering our understanding of the role of AS for each specific gene.

Supplementary Material
Supplementary material, appendix, tables S1 and S2, and figures S1–S5 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgments
The authors acknowledge two anonymous referees for their valuable suggestions and corrections. This work was supported by the National Institute of Health (grant number U41 HG007234) and the Spanish Ministry of Economics and Competitiveness (grant number BIO2012-40205).

Literature Cited
Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.
Altschmied J, et al. 2002. Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. Genetics 161:259–267.

Support for the I-model came from 11 examples in which each gene duplicate specifically retained one of the two ancestral MEHEs, that is, of splice isoform separation. As a result of the process of concerted loss and retention of ancestral MEHEs, the net protein diversity is conserved but distributed among different genes. For some genes, there may even be advantages to separating the alternative isoforms. At the very extreme of this model we found MARVELD3, for which the splice isoform separation process may have taken place independently in at least three different lineages.

Although we conducted no experimental confirmation, long-standing conservation of MEHEs very likely reflects the existence of functional differences between the homologous exons. If true, the identified cases of splice isoform separation would support a process of subfunctionalization in which ancestral functions have been partitioned between paralogs. The reported cases would add to the handful of cases of this kind reported in the literature (Altschmied et al. 2002; Yu et al. 2003; Pacheco et al. 2004; Cusack and Wolfe 2007; Hultman et al. 2007; Marshall et al. 2013). Evolutionary analysis may explore how these gene duplicates evolved once ancestral isoforms were uncoupled, and whether this uncoupling affected the evolution of accompanying constitutive exons and/or eventually had an adaptive value. In practical terms, having each human isoform represented by a distinct gene in a target species may facilitate the experimental characterization of each isoform function by, for instance, specific gene knockout experiments or gene expression analysis.

The curated analysis presented here throws light on the general aspects of the complex interplay between GD and AS as repositories of protein diversity, and also represents a guide for bettering our understanding of the role of AS for each specific gene.

Supplementary Material
Supplementary material, appendix, tables S1 and S2, and figures S1–S5 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgments
The authors acknowledge two anonymous referees for their valuable suggestions and corrections. This work was supported by the National Institute of Health (grant number U41 HG007234) and the Spanish Ministry of Economics and Competitiveness (grant number BIO2012-40205).

Literature Cited
Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.
Altschmied J, et al. 2002. Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. Genetics 161:259–267.
Abascal et al.
Shah K, Shokat KM. 2002. A chemical genetic screen for direct v-Src substrates reveals ordered assembly of a retrograde signaling pathway. Chem Biol. 9:35–47.

Smith CW, Valcárcel J. 2000. Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem Sci. 25:381–388.

Smith JJ, et al. 2013. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nat Genet. 45:415–421, 421e1–e2.

Steed E, Rodrigues NTL, Balda MS, Matter K. 2009. Identification of MarvelD3 as a tight junction-associated transmembrane protein of the occludin family. J Biol Chem. 284:27500–27510.

Associate editor: Andreas Wagner