Human Molecular Genetics, 2002, Vol. 11, No. 13 1561–1567

Common exon duplication in animals and its role in alternative splicing

Ivica Letunic[†], Richard R. Copley[†] and Peer Bork*

EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany

Received December 2001; Revised and Accepted March 16, 2002

When searching the genomes of human, fly and worm for cases of exon duplication, we found that about 10% of all genes contain tandemly duplicated exons. In the course of the analyses, 2438 unannotated exons were identified that are not currently included in genome databases and that are likely to be functional. The vast majority of them are likely to be involved in mutually exclusive alternative splicing events. The common nature of recent exon duplication indicates that it might have a significant role in the fast evolution of eukaryotic genes. It also provides a general mechanism for the regulation of protein function.

INTRODUCTION

The evolutionary significance of introns in eukaryotic genes has been the subject of much debate. Arguments include, for example, that introns facilitate recombination (1,2) or enable exon shuffling between different genes (3). Although duplicated exons have been observed within genes, we here report that this is an unexpectedly common phenomenon and that tandemly duplicated exons within genes have frequently retained or acquired functionality. This mechanism complements gene duplication as a source of evolutionary novelty (4) (Table 1). In particular, we show that a substantial amount of exon duplication events lead to alternative splicing; in the majority of instances we can predict mutually exclusive alternative splicing.

Table 1. Exon statistics for *Homo sapiens*, *Drosophila melanogaster* and *Caenorhabditis elegans*

Feature	*H. sapiens*	*D. melanogaster*	*C. elegans*
Annotated genes	29 181	13 744	19 315
Annotated exons			
Total	191 656	52 043	115 385
Duplicated	5 544	3 078	3 669
Novel exons	2 064	371	647
Additional pseudoexons[a]	1 242	19	317
Genes with duplicated exons[b]	3 140 (10.7%)	976 (7.1%)	1 418 (7.5%)

Annotation was taken from standard distribution sites for each organism (Ensembl v3.26 for human, Flybase release 2 for fly and Wormbase WS54 for worm).
[a]Exons containing stop codons that are unlikely to be transcribed.
[b]This includes the novel exons, which increases the number of duplicated exons by 37.2%, 12% and 17.6% in human, fly and worm respectively.

RESULTS

Detection of duplicated exons

To detect cases of tandemly duplicated exons, we successively compared each exon of each gene to its neighboring exons in the same gene, for the *Homo sapiens* (5), *Drosophila melanogaster* (6) and *Caenorhabditis elegans* (7) genomes (see Materials and Methods for details of data sources). As eukaryotic gene predictions are not perfect, and previous studies have shown that many annotated introns may contain unannotated exons that are involved in alternative splicing (8), we also searched each exon against adjacent introns.

The comparison identified 12 291 cases of tandem duplication in the annotated exons of human, fly and worm. In addition, we detected 4660 exon homologous regions within introns in all three species. We refer to these latter regions as 'unannotated' exons. We also detected 66 cases of duplications (data not shown), where the DNA sequence was 100% identical. We attributed this to artifacts of the assembly procedure and did not include such regions in further analyses.

Of the 4660 unannotated exons, 1578 contained stop codons. These likely pseudoexons were excluded from further analysis (Table 1). To confirm the functionality of the remainder of the unannotated exons, we searched for their presence in expressed sequence tags (ESTs) and cDNAs. As many as 35.1% of the unannotated exons were found in ESTs (compared with 41% for the annotated ones) and 21% were contained in well-annotated human RefSeq (9) sequences (compared with 63.3% of the annotated ones). The set of unannotated exons thus contains both known exons that are not included in current genome-specific databases, and inferences, based on homology, of novel exons not included in any database. As there is no difference in the methodology used to identify these subsets,

*To whom correspondence should be addressed. Email: bork@embl-heidelberg.de
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

we continue to refer to both categories as 'unannotated'. The lower coverage of unannotated exons by cDNA might reflect the practice of current gene prediction schemes that use cDNAs as evidence of gene structure without further analysis of the genomic region involved i.e. the presence of an exon in cDNA will lead to its inclusion as an annotated exon within genome databases.

The size distribution of both annotated and unannotated duplicated exons shows no significant difference to the size distribution of all exons in any of the analyzed genomes. The majority of the duplicated exons fall into the 30- to 50-amino-acids range. We do not see any obvious bias towards certain functional classes of gene products (data not shown).

Evidence of functional constraint

As probably more than 50% of the genes in animals contain more than one splice form (5,10), it is likely that, in many cDNAs, not all exons are present. The similarity of the ratios above for inclusion of annotated and 'unannotated' exons in expressed sequence are indicative that the bulk of the latter set can be regarded as functional. In order to provide more support for this inference, we performed an independent functionality test based on the ratio of synonymous versus non-synonymous mutations per codon. We calculated the K_a/K_s ratio for both the annotated and the candidate exons using the method of Yang and Nielsen (11). The ratios of both annotated and candidate exons are similar (in contrast to presumed pseudoexons containing a stop codon) (Fig. 1). We conclude that the majority are under selective pressure, and thus likely to be functional. In this way, we identified 2064, 371 and 647 unannotated candidate exons in human, fly and worm respectively, and have thus increased the number of duplicated exons in the three organisms by 37.2%, 12% and 17.6%. These figures are likely to be lower limits due to the restricted sensitivity of homology search methods for short sequences.

Duplicated exons and alternative splicing

The common nature of duplicated, likely functional exons suggests that they provide a useful and evolutionarily accessible way of meeting functional demands. One possible way in which they might do this is via complex splicing patterns. Mutually exclusive alternative splicing of these exons would allow any given part of a protein product corresponding to a single exon to be replaced with a modified version, providing a framework for the regulation of differing

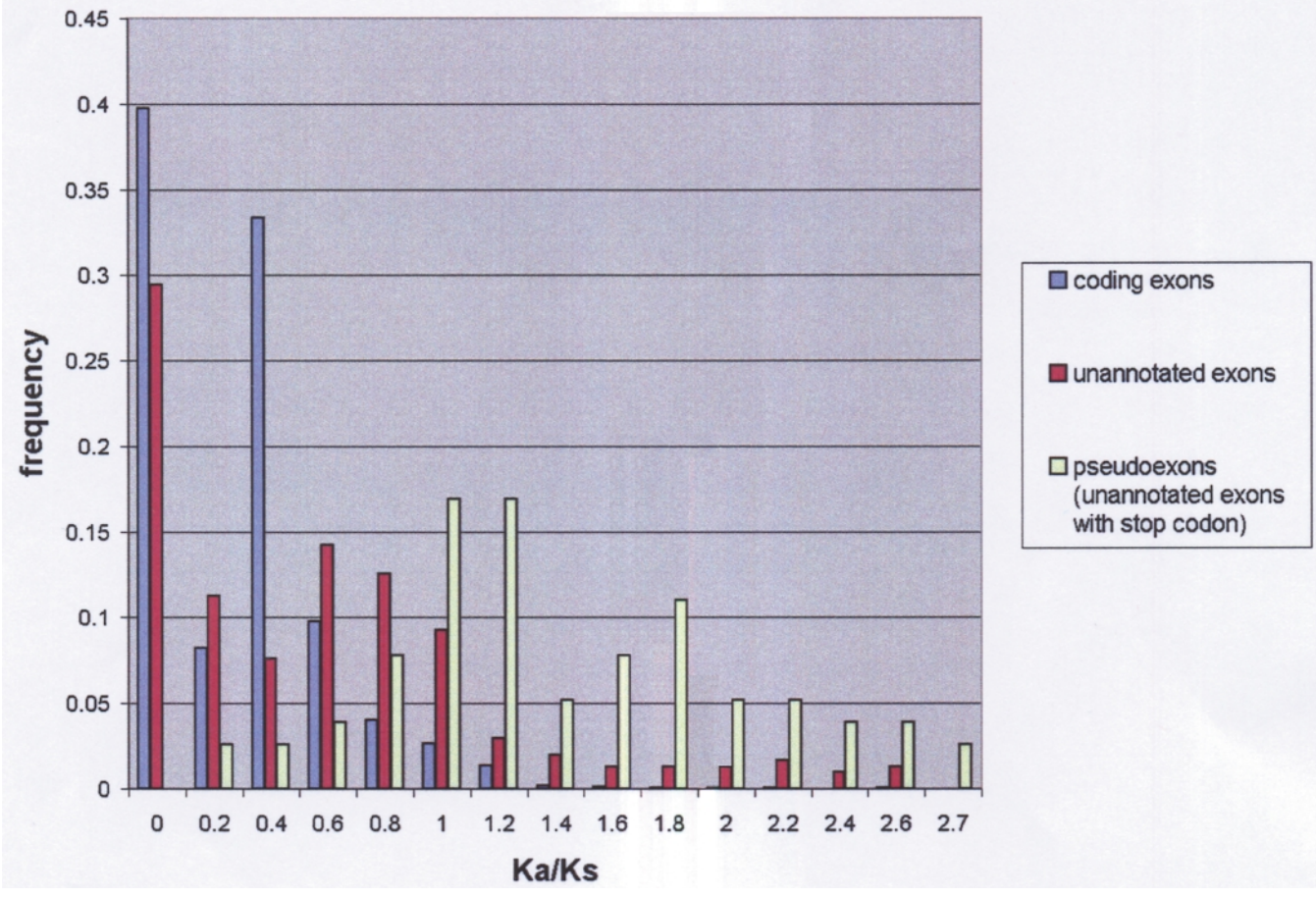

Figure 1. K_a/K_s ratios in *H. sapiens* tandem exon pairs. Only duplicates with over 70% amino acid sequence identity have been compared, since otherwise K_a/K_s ratios are unreliable. Maximum-likelihood analysis of the data shows that about 70% of the unannotated exons are functional.

Table 2. Comparison of annotated and unannotated exon duplicates in human

	Annotated	Unannotated
Exon numbers		
Total	5544 tandem pairs	2064 tandem pairs
Subset analyzed[a]	1823 tandem pairs (3260 exons)	963 tandem pairs (963 novel exons)
Evidence for expression of single exons of subset		
cDNA (RefSeq) matches	2065 (63.3%)	202 (21%)
EST matches	1335 (41%)	338 (35.1%)
Evidence for simultaneous expression of both exons of subset		
cDNA (RefSeq) matches	618 (33.9%)	4 (0.4%)
EST matches	342 (18.8%)	6 (0.6%)
Evidence for alternative splicing in subset		
Annotated[b]	849	NA
Different intron phase	498 annotated (58.6% of 849)	600 observed (62.3%)
Same intron phase	351 annotated (41.4% of 849)	NA[c]

[a]Derived as follows (see also Materials and Methods). For unannotated exons, initial hits within the introns were expanded to match the original query exon's length; 10 nucleotide regions at the beginning and the end of hits were checked for conserved splice site sequences (GT and AG); only the hits with both splice sites present were analyzed. For annotated exons, pairs with $E < 10^{-5}$ where blast hit covered at least 90% of the exon sequence were analyzed.
[b]Implying mutually exclusive alternative splicing based on Ensembl annotation; note that this number is a lower limit given current estimates that at least 50% of all human genes are subjected to alternative splicing.
[c]If exon pairs are in the same intron phase, they could be either coexpressed or alternatively spliced. We found evidence of coexpression in only 5 out of 363 cases (1.3%); therefore we assume that the majority of all unannotated exons are mutually exclusively spliced.

biochemical and physical properties, within the context of pre-existing genes.

To test this hypothesis, we extracted a subset of the duplicated annotated and unannotated exons with high sequence similarity, so that exon borders can be reliably predicted. Given that both tandem exons are functional and provided that the splice sites remain conserved, if the lengths of tandemly repeated exons are not exact multiples of 3, they are unlikely to be incorporated into the same transcript in the same reading frame (doing so would be likely to lead to a frameshifted product), and so we infer that they are likely to be mutually exclusively spliced. In human, 3260 annotated and 963 unannotated exons were sufficiently similar to each other (see Materials and Methods) to be studied in detail (Table 2). 62.3% of the unannotated exons and 58.6% of the annotated duplicated exons, which are known to be subjected to mutually exclusive alternative splicing, have lengths that are not exact multiples of 3. As a further test, we searched the EST and cDNA databases for evidence of simultaneous expression of unannotated duplicated exons. Only six such cases were detected (0.6%), suggesting that for the majority of tandemly duplicated exons, mutually exclusive alternative splicing can be predicted (Table 2).

DISCUSSION

Taking all exons together, the above numbers show that as many as 10.7%, 7.1% and 7.5% of the annotated genes in *H. sapiens*, *D. melanogaster* and *C. elegans* respectively contain conserved duplicated exons. Up to 83 copies per gene can be found, the most dramatic case being the well-known DSCAM in *D. melanogaster*, where multiple mutually exclusive exons lead to enormous numbers of splice variants (12). In such cases, alternative exons have not been annotated within the genomic databases.

The results indicate that exon duplication, and mutually exclusive alternative splicing of the duplicated exons, is a relatively common phenomenon. This suggests that mutually exclusive alternative splicing has arisen as a means of producing regulated modular changes of protein functionality, independent of domain boundaries or other structural constraints. How might the control of such splicing have evolved?

Evolutionary scenarios

Cases where tandemly duplicated exons have lengths that are not exact multiples of 3 will lead to nonsense transcripts if both exons are incorporated into the product, and so are likely to be strongly selected against, providing pressure for the evolution of regulated mutually exclusive splicing. Cases where known mutually exclusive exons have lengths that are exact multiples of 3 pose an interesting evolutionary question: When will duplicated exons be incorporated into the same transcript, and when will they be subjected to mutually exclusively splicing? In theory, both duplicated exons could be incorporated into the same transcript with no loss of reading frame; however, in cases where the exon does not correspond to a structural domain, it seems likely that in some instances, the translated polypeptide will not be able to fold to form a functioning product. This will again lead to selection for regulation of the splicing of the duplicated exons, leading to controlled mutually exclusive splicing. However, in other observed instances of mutually exclusive splicing, especially when the duplicated exon is short, or falls in a loop region, it is plausible that both exons could be incorporated into a folded protein product. For mutually exclusive splicing to occur in such instances, it is necessary to invoke some form of regulation, from the moment of their duplication, to prevent both being incorporated in the same transcript. One possible explanation is that the duplicated exons are flanked by introns of different types (i.e. U2- and U12-type introns spliced by the major and minor

spliceosomes). When duplicated, such exons will not be able to be spliced together, since the donor site of the 5′ exon and acceptor of the 3′ exon will form a hybrid 'intron'. It is thought that neither spliceosome is capable of removing such regions (13). This appears to be the scenario responsible for mutually exclusive splicing in human stress-activated protein kinase (JNK 1) (data not shown). As U12-type introns are believed to be capable of evolving into U2 types (14), this theory may also explain some instances of mutually exclusive splicing of other exons where the flanking introns are the same phase and both U2 type.

Structural and functional implications

Protein structure can provide a direct means by which to assess the functional significance of the residue changes effected by alternative splicing. By analyzing the duplicated exon within the context of the 3D structure of the entire protein domain in which it resides, we are able to propose functional consequences and adaptive significance for some of the alternative splicing events. Here we present two such examples.

Our automated screening identified a duplicated exon in human glycine receptor α-2 (NCBI RefSeq NP_002054). The duplicated exon in this gene has been noted previously as a candidate for alternative splicing (15). It encodes a duplicated region of 22 amino acids, of which only two are variable, substituting a VT pair for IA (Fig. 2A). cDNA evidence (GenBank accession nos AA463244 and AV729257; UCSC genome browser http://genome.ucsc.edu) confirms that the duplicated exons are mutually exclusively spliced into the gene products. By using the recently solved structure of molluscan acetylcholine-binding protein (AChBP) as a template for the homologous glycine receptors (16), we were able to establish the structural location of the duplicated exon within the protein fold, and find that the two-amino-acid substitution occurs on the inner face of the internal channel formed by pentamers of this domain (Fig. 2B, C) (16). Intriguingly, comparison with the alignment of all family members suggests that the exon encoding the 'IA' variant is the only one for which a non-polar residue is found in the second location (data not shown). It thus seems likely that the alternative splicing event is used to regulate the precise properties of the channel, perhaps to create a constitutively inactive form.

The *Drosophila* gene CG3487 encodes a putative trypsin-like serine protease. Analysis of the predicted transcript of this gene shows that it contains a duplicated final exon (Fig. 3A). Alignment with trypsin-like proteases of known structure shows that this region corresponds roughly with the C-terminal lobe. Although this could be considered a distinct structural entity, and so arguably both exons could be incorporated into the final protein product, we consider this unlikely. The active site of these proteins is a triad of residues (His, Asp, Ser). Of the three, the His and Asp occur in the N-terminal lobe of the protease, and the Ser in the C-terminal lobe. As these residues

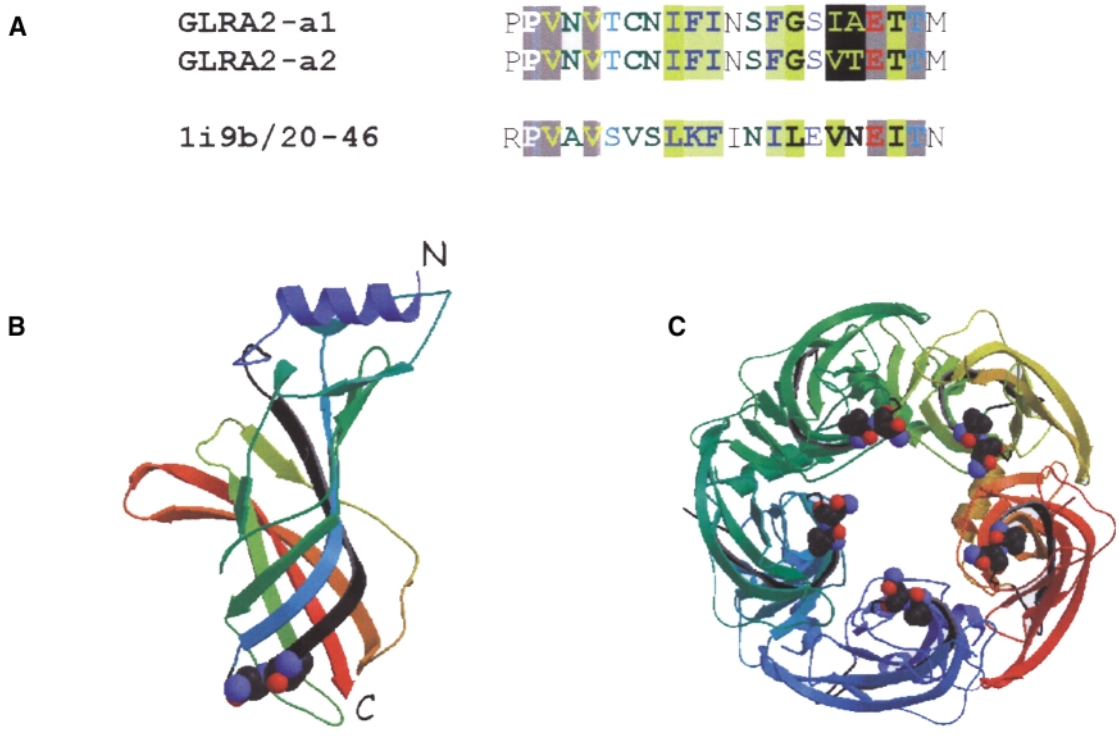

Figure 2. Alternative splicing of human glycine receptor α-2 (hGLRA2). (**A**) The duplicated exons aligned with a region of known structure from molluscan acetylcholine-binding protein (AChBP). The variable residues are highlighted in black. Other residues are colored using the CHROMA program with default parameters and an 80% consensus threshold (21). (**B**) The structure of the duplicated exon of GLRA2, marked in black, with the variant residues shown as space-filling, as inferred from the structure of AChBP (see text for details). (**C**) The structure of the biologically active pentameric form of AChBP, serving to represent how the alternatively spliced residues of GLRA2 would cluster in space.

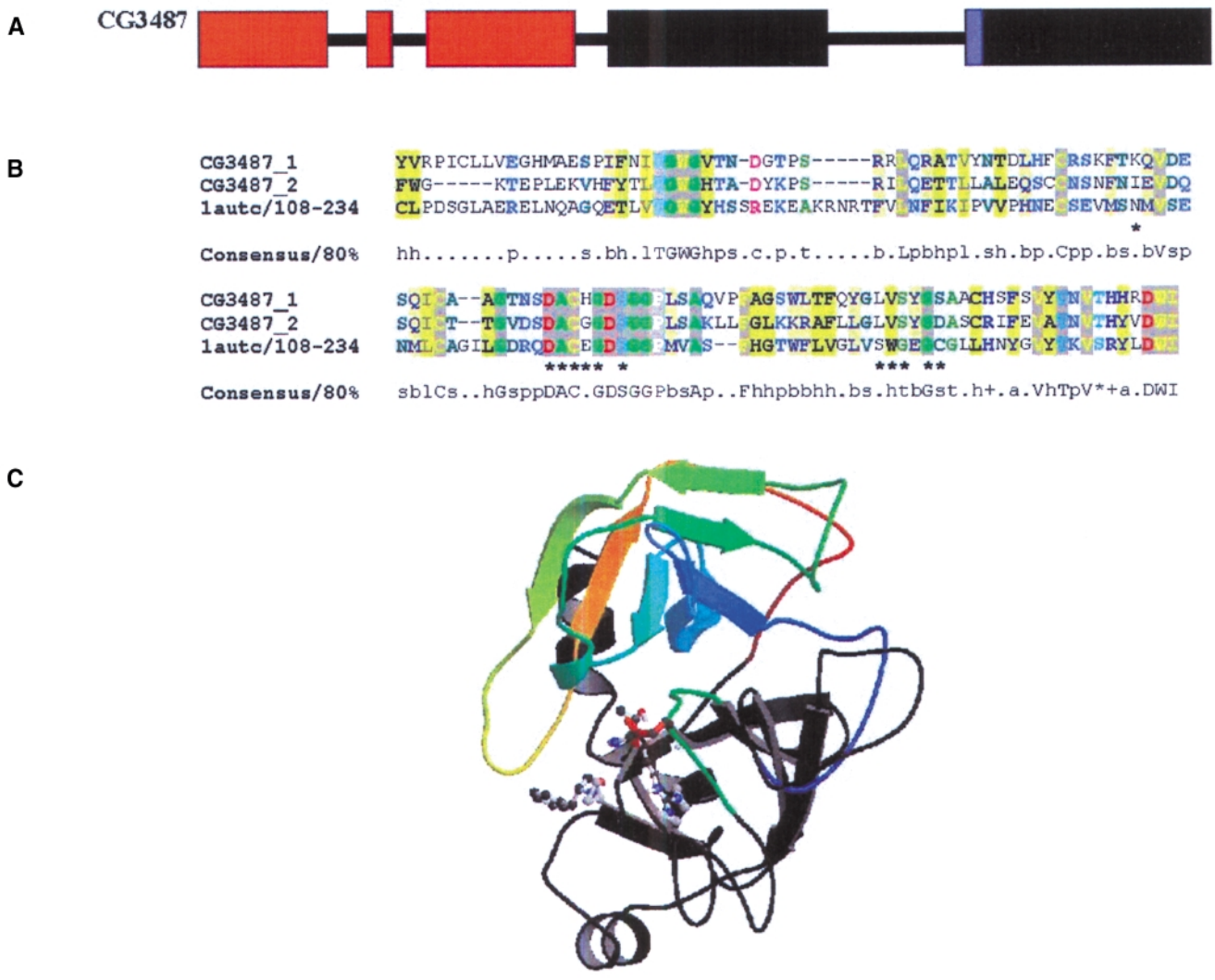

Figure 3. Duplicated exons in the *Drosophila* trypsin-like serine protease CG3487. (**A**) Gene structure, and proposed modification to splice site of final intron. The duplicated exons are shown in black. The blue bar shows the region of the terminal exon that was removed to make phases compatible with mutually exclusive splicing. (**B**) Alignment of duplicated exons and sequence of known structure, PDB code 1autC. Residues within 4.0 Å of the substrate are marked with an asterisk. (**C**) Structure showing location of duplicated exon (black), and location of functionally significant sequence variability between the two alternative exons (see text for details). The green insert in the black structure represents the most conserved region between the two exons, with the red residue showing the predicted site of the Gly→His substitution.

need to be in a precise geometric arrangement to achieve catalytic activity, it seems unlikely that both active-site serine residues (which are conserved in each of the exons) could be utilized simultaneously in the same triad. The phases of the penultimate and ultimate introns of the predicted gene structure are incompatible with alternative splicing; however, reanalysis of the gene prediction using CG14227, the closest homologue of CG3487, employing the genewise program (http://www.sanger.ac.uk/Software/Wise2/), shows that the boundaries of the final intron should be adjusted. This modification does make the gene structure compatible with alternative splicing. We aligned the repeated exons with the most similar sequence of known three-dimensional structure (Fig. 3B). Analysis of conservation and variation within the duplicated exons in light of the three-dimensional structure suggests that a Gly→His substitution between the two exons will have profound consequences for access to the active site substrate-binding pocket (Fig. 3C).

These examples show that mutually exclusive alternative splicing is likely to have direct functional consequences, suggesting that exon duplication and mutually exclusive splicing of the resulting duplicates provides a parallel route to gene duplication for subtly modifying protein function and increasing the range of biochemical control available to the cell.

Summary

These findings have important technical, functional and evolutionary implications. (i) Current gene prediction schemes

for higher eukaryotes should consider the phenomenon presented here, in particular if a prediction is based on homology to cDNA or proteins derived from it. (ii) As such a large number of duplicated exons show patterns of selection indicative of function, they are likely to be incorporated as tandem repeats in proteins [many examples are found in extracellular proteins when an exon corresponds to a domain (3)] and/or be used for alternative splicing. Indeed, in at least 60% of the cases, mutually exclusive splicing is the most likely scenario, enabling the modification of protein activity. (iii) The general phenomenon presented here suggests a mechanism for the fast evolution of genes and gene regulation via alternative splicing.

During the preparation of this paper, a complementary analysis of known cases of alternative splicing annotated in the SWISS-PROT database (17) concluded that mutually exclusive splicing caused by exon duplication contributes about 10% of the total of alternative splicing. We have shown that the vast majority of duplicated exons are subjected to alternative splicing. If one takes the current lower limit estimate that about 50% of all human genes are alternatively spliced (5) and considers that about 10% of all human genes contain duplicated exons, as much as 20% of alternative splicing might be due to exon duplications.

MATERIALS AND METHODS

Data generation and homology searches

The annotations of the genomes of *H. sapiens*, *D. melanogaster* and *C. elegans* were taken from standard distribution sites for each organism (Ensembl v3.26 for human, Flybase release 2 for fly and Wormbase WS54 for worm). Using various perl scripts, exons were extracted and similarity searches were carried out to identify homology to neighboring exons or within DNA adjacent to exons (Table 1). In both cases, significance was set at *E*-values of 0.001, using the bl2seq implementations of TBLASTN or BLASTP (18).

Analysis of exon functionality

A subset of the human set above was created for detailed analysis. For the annotated set, only those pairs that had a blast hit covering at least 90% of both exons with *E*-value lower than 10^{-5} were further considered. For the unannotated set (cases where annotated exons had a hit in the neighboring introns), blast hits within introns were first extended to match the length of the query exon. The regions of 10 nucleotides around the start and the end of the expanded hits were checked for presence of both standard splice-site sequences (GT–AG), thus allowing reliable identification of exon boundaries. In order to match the homology-based set with experimental data, ESTs and cDNA sequences were compared with our dataset using a restrictive threshold of 99% nucleotide identity over at least 30 nucleotides. Human ESTs were extracted from GenBank dbEST (19), and the human RefSeq database (9) was used a source of cDNAs. As matches to ESTs and cDNA could only partially support functionality of the predicted exons, an analysis of the nucleotide substitution rates was done.

Amino acid sequences of the tandemly duplicated exons were first aligned using ClustalW (20). Amino acid sequence alignments were used to reconstruct the DNA alignment by substituting amino acids with the corresponding codons from the genomic DNA. The alignments were analyzed using Yang and Nielsen's method (11) and K_a/K_s ratios were determined for three separate subsets: duplicates with both annotated exons, duplicates with unannotated exons and duplicates where unannotated exons contained a stop codon ('pseudoexons'). As K_a/K_s ratio reliability rapidly falls with sequence divergence, only pairs with over 70% amino acid sequence identity have been compared. On the other hand, for exons with lower similarity, one would expect an accumulation of stop codons if there were no functional constraints.

If the pool of unannotated exons is regarded as a mixture of true functional exons and 'pseudoexons', then the K_a/K_s distribution for unannotated exons can be viewed as a mixed distribution. Under this assumption and given histogram estimates for K_a/K_s distributions for functional exons and pseudoexons separately (Fig. 1), the fraction of coding exons in the set of unannotated exons was estimated using the maximum-likelihood method.

Analysis of mutually exclusive alternative splicing

All exons whose length was not an exact multiple of 3 (i.e. that had flanking introns of different phases) were considered as mutually exclusive alternative splicings because their fusion in one transcript would result in a frameshift. The phases of the introns flanking unannotated exons were determined by comparing the reading frame of the BLAST hit with the splice sites. All pairs that were considered to be alternatively spliced were checked against EST and cDNA databases for evidence of coexpression of both exons. Pairs were considered as coexpressed if they had an EST or cDNA match with 99% identity over at least 30 nucleotides on both sides of the splice site (Table 2).

ACKNOWLEDGEMENT

We thank Juan Valcarcel for helpful discussions.

REFERENCES

1. Comeron, J.M. and Kreitman, M. (2000) The correlation between intron length and recombination in drosophila. Dynamic equilibrium between mutational and selective forces. *Genetics*, **156**, 1175–1190.
2. Gilbert, W. (1978) Why genes in pieces? *Nature*, **271**, 501.
3. Patthy, L. (1999) Genome evolution and the evolution of exon-shuffling – a review. *Gene*, **238**, 103–114.
4. Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.
5. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
6. Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster. Science*, **287**, 2185–2195.
7. Consortium, T.C.e.S. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* Sequencing Consortium. *Science*, **282**, 2012–2018.

8. Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P. and Mattick, J.S. (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat. Genet.*, **24**, 340–341.
9. Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
10. Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J. and Bork, P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, **474**, 83–86.
11. Yang, Z. and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, **17**, 32–43.
12. Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E. and Zipursky, S.L. (2000) *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, **101**, 671–684.
13. Sharp, P.A. and Burge, C.B. (1997) Classification of introns: U2-type or U12-type. *Cell*, **91**, 875–879.
14. Burge, C.B., Padgett, R.A. and Sharp, P.A. (1998) Evolutionary fates and origins of U12-type introns. *Mol. Cell*, **2**, 773–785.
15. Cummings, C.J., Dahle, E.J. and Zoghbi, H.Y. (1998) Analysis of the genomic structure of the human glycine receptor alpha2 subunit gene and exclusion of this gene as a candidate for Rett syndrome. *Am. J. Med. Genet.*, **78**, 176–178.
16. Brejc, K., van Dijk, W.J., Klaassen, R.V., Schuurmans, M., van Der Oost, J., Smit, A.B. and Sixma, T.K. (2001) Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors. *Nature*, **411**, 269–276.
17. Kondrashov, F.A. and Koonin, E.V. (2001) Origin of alternative splicing by tandem exon duplication. *Hum. Mol. Genet.*, **10**, 2661–2669.
18. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
19. Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) dbEST – database for 'expressed sequence tags'. *Nat. Genet.*, **4**, 332–333.
20. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
21. Goodstadt, L. and Ponting, C.P. (2001) CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics*, **17**, 845–846.