INTRA-GENIC EXON DUPLICATIONS IN THE HUMAN TRANSCRIPTOME

ANKE BUSCH

Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

Department of Microbiology and Molecular Genetics, University of California, Irvine, CA 92697, USA

PETER F. STADLER*

Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany
Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark
Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
studla@bioinf.uni-leipzig.de

Tandem duplications are a major source of variance in genomic organization accounting for a large fraction of gene duplications. They also act on the smaller scale of individual exons, playing a significant role in the rapid evolution of eukaryotic genes. Here we show that duplicated exons are not necessarily placed in adjacent position, i.e., exon duplication is not restricted to tandem duplications of individual exons. The fraction of duplicated exons within genes increases with gene size, indicating that genes with very large numbers of exons arose from a series of intra-genic duplication events of one or several unrelated exons. Exons at both the 3' and the 5' end of coding regions are duplicated less frequently than internal exons. Duplicated exons show elevated levels of alternative splicing, indicating an intimate relation with it. The largest proteins, comprising nearly repetitive exons with large copy numbers, are typically structural proteins.

Keywords: Exon duplication, human, alternative splicing, duplication pattern.

Introduction

In the first systematic study on duplicated exons [1], Letunic, Copley, and Bork found that about 10% of all human, fly, and worm genes contain tandemly duplicated exons. The duplication rate of genomic regions entirely internal to annotated genes, furthermore, is consistent with the rate of gene duplication (about 10^{-3} to 10^{-2}

*corresponding author

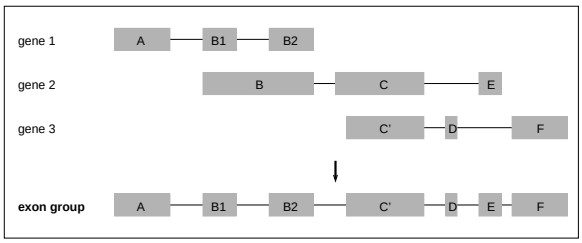

Fig. 1. Combination of entries with overlapping exons to one single entry: an exon group. Gray boxes indicate exons, while lines between them symbolize intronic regions. The two separated exons B1 and B2 in gene 1 are retained in the final exon group instead of exon B, which includes an intron retention. Furthermore, the smaller version of exon C is kept.

events per gene and Myr) [2]. Exon duplications thus are a major contribution to the evolution of protein-coding genes. An important evolutionary consequence of such segmental duplication is the subsequent remodeling of the gene structure e.g. through the activation of latent splice sites [2]. Tandem exon duplication thus is often associated with alternative splicing to reduce the possible deleterious impacts on transcript/protein structure [1,3,4].

Case studies on individual genes, e.g. drosophilid Gr39a [5] or metazoan tropomyosin [6] show that several rounds of exon duplications can lead to complex gene structures and evolutionary histories. Here we extend the earlier quantitative analyses beyond tandem duplications of individual exons and investigate in detail the patterns of duplicated exons within human genes.

Materials and Methods

Construction of exon groups

The RefSeq gene annotation track for human (GRCh37/hg19), which includes information on the exon positions of each exon, was downloaded from the UCSC genome browser [7,8]. There are several complications that make it undesirable to use this annotation directly: In many cases, identical gene names refer to several distinct but overlapping exons as a result of alternative splicing. Identical gene names are also encountered with completely different exon coordinates, e.g. as a consequence of very recent gene duplications. Examples are shown in Supplemental File 1. In the case of intron retention, furthermore, two distinct exons may be contained in a single large one from a different isoform of the same gene. Since we are interested here in local duplication events of exons, we adopted the following convention: If 95% of exon A is contained in another exon B, we conclude A is nearly contained in B. In this case we replace the larger exon by the smaller one.

Since overlapping RefSeq gene annotations would lead to an over-counting of exons that belong to more than one annotated gene, we define an exon group as a collection of exons that are connected by common genes, see Figure 1. In practise,

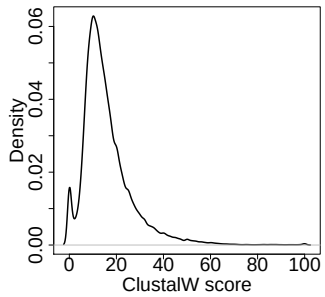

Fig. 2. Density of the `ClustalW` similarity score distribution of all intra-genic exon pairs.

we construct a graph whose vertices are the exons retained after filtering for near inclusion. We connect two such exons whenever they, or exons which they have replaced, are contained in the same annotated gene. An exon group is now defined as connected component in this graph. In order to compute the exon groups efficiently we first construct a graph of all exons, with arcs connecting two vertices whenever their corresponding exons appear in a common gene. Then we test for overlapping exons. Whenever A replaces B, A inherits all connections of B.

Pairwise alignments of exons

For each exon group, we extracted all exonic sequences and determined pairwise alignments scores using `ClustalW` [9,10]. Recall that `ClustalW` implements a global pairwise alignment with free end gaps. This reduces spurious hits caused e.g. by short repetitive sequences that are frequently observed in local alignments. The final pairwise scores are thus calculated as the number of identical residues in the optimal alignment divided by the number of alignment edges, i.e., the number of matches and mismatches in the alignment. Note that we obtain top scores whenever (a) a copy of one exon is contained in the other one, or (b) when a suffix of one copy is a prefix of the other one. On the other hand low scores are obtained when only a relatively short infix matches.

In order to find an appropriate similarity cutoff that separates similar/duplicated exon pairs from unrelated exons, we analyzed the distribution of all similarity scores, Fig. 2. We accept only those exon pairs as highly similar, and hence almost certainly homologous, that are more similar than the 99% quantile of the score distribution (cutoff score = 56). Since this `ClustalW` score is an estimate of the pairwise sequence identity, we consider this cutoff to be quite conservative.

A potentially confounding factor in the analysis of local duplication are repetitive elements embedded in exonic sequences. We therefore removed highly similar exon pairs consisting of repetitive elements. To this end, we downloaded the repeat annotation track from the UCSC Genome Browser [7,8] that was produced with

`RepeatMasker` [11] and aligned it to the exons. Additionally, very short exons were also removed. Here, the length cutoff was chosen according to an analysis of the length distribution of all exons. We chose a length cutoff of 54 nucleotides, which represents the 5% quantile of the length distribution. Finally, all exons were removed whose sequence contained more than 10% undetermined nucleotides (letters other than A, C, T, G). The remaining highly similar exon pairs are interpreted as duplicates of each other.

All analyses reported in this contribution are based on these filtered exon groups.

Paralogs

In order to address the question whether the number of paralogs correlates with the number of duplicated exons in an exon group, we downloaded all human homology data from the `HomoloGene Database` [12,13] and extracted all human paralogs. Since an exon group may contain more than one gene, we extend the definition of paralogy to exon groups whenever two exon groups are linked by a pair of annotated paralogous genes. For each exon group we furthermore computed the maximal and average number of human paralogs that are annotated for its member genes.

Patterns of duplicated exons

Since exons are linearly arranged along the genome, exon groups can also be represented as strings so that each exon is labeled by a single letter that uniformly labels duplicated exons. Exons without duplicates, are of little interest for us and hence are represented by a special character "`-`". When focusing on the duplicated exons only, we ignore the unduplicated exons altogether and hence delete the "`-`" characters. In the following, a pattern is a string of length at least 2 since, in this section, we are only interested in the occurences of more complex duplication events than just copying single exons. For each group, we search the resulting contracted string for the longest duplicated pattern, i.e., a substring of length ≥ 2 that occurs at least twice in the string representing the exon group and no longer duplicated substring can be found. If it is equal to `AB`, the longest duplication event in this group corresponds to a duplication of the two exons `A` and `B`, which occur twice or multiple times in the same order in this exon group, e.g. `AB-CAB`.

A duplicated pattern `AA` refers to two copies of the same exon that can be found twice or multiple times in the exon group, e.g. `-AA-BAA`. Please note, we are not interested in duplicated substrings of length 1 (e.g. `A`) since, here, we concentrate on duplication events more complicated than a single exon duplication. Furthermore, we say that a pattern is irreducible if it does not contain multiple copies of shorter patterns. For instance, the irreducible pattern `AA` cannot be reduced to `A` since the latter is not a pattern because pattern have length ≥ 2. For each exon group, we determine also the longest irreducible pattern that appears at least twice, i.e., that is duplicated. In addition, we searched, in each exon group, for the most frequent duplicated pattern. By construction, shorter patterns will occur more or at least

equally often compared to larger ones. If two patterns of different size appeared equally often, we preferred the longer one. In order to compare the patterns between different exon groups we standardize them by translating the distinct letters in the pattern to `A`, `B`, `C`, etc., in their order of appearance. For instance `BBACBXA` is recoded as `AABCADB`.

GO term analysis

We sorted all 1607 exon groups including duplicates according to their maximal copy number of one exon and according to their overall fraction of duplicated exons. In each subset, we extracted all genes the groups are composed of. In order to search for common gene ontology terms (GO terms), we neglected members of the same gene family within one exon group. All remaining genes were analyzed for common GO terms using a web-based version of `GOTermFinder` [14].

Results

Duplicated exons are frequent

We processed the RefSeq annotation to combine overlapping annotations in the same reading direction and to remove redundancies at exon level as detailed in the Methods section, obtaining 210887 exons that are subdivided into 21288 exon groups (see Materials & Methods), each representing a genomic locus that hosts one or a few related RefSeq genes. Of these, 18698 exon groups include more than one exon. Duplicate exons were identified as pairs with a `ClustalW` similarity score in the top percentile of the similarity score distribution, leading to 3823 exon groups (18%) that contain at least one duplicated exon, consistent with previous estimates [1,2]. After filtering for repetitive elements and other possible sources of artifacts (see Methods for details), 2216 exon groups were returned to the background set of exon groups not including unambiguous duplicates. The remaining 1607 groups including duplicated exons comprise 10611 homologous exon pairs, i.e., many of them contain multiple copies of the same exon. 274 pairs are identical and in 17 additional pairs, the sequence of one copy is completely contained in another one. These 291 pairs are either very young exons, indicate copy number variants, or might in some cases reflect problems with the genome annotation.

A total of 7255 exons with at least one partner within its exon group was identified. Groups without duplicated exons consist of 9.92 exons on average, while groups including duplicates are composed of 24.08 exons on average. We note that the size of the exon groups strongly correlates with both the length of the longest annotated CDS (Pearson correlation coefficients $\rho = 0.810$ and $\rho = 0.730$ for groups including and not including duplicates, respectively) and with the mean lengths of annotated coding sequences (Pearson $\rho = 0.806$ and $\rho = 0.729$, respectively, for the two groups), see Supplemental File 2. Thus, exon duplication significantly contributes to the evolution of large proteins.

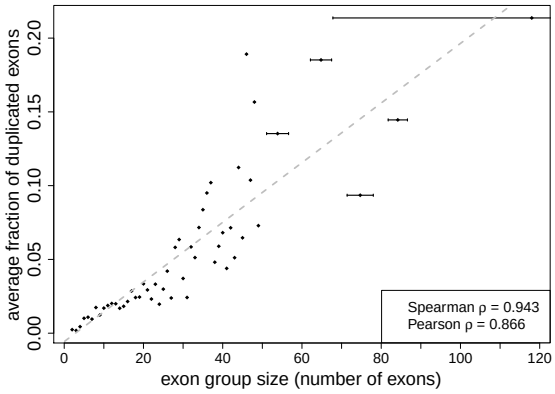

Fig. 3. Relation between the size of an exon group and the fraction of included duplicated exons. All groups including between 50 and 59, 60 and 69, 70 and 79, 80 and 89, and 90 or more exons where grouped together. Their variances around their mean sizes are shown as vertical lines. The gray dashed line indicates the linear regression.

For all exon groups with at least one duplicated exon we determined the total number of exons that have a duplicate, i.e., a homologous exon, within the same exon group. Dividing by the total number of exons in these exons groups yields a surprisingly high fraction of 23.07%, indicating that genes that can incorporate duplicated exons are prone to accept multiple additions of duplicate exons. Indeed, Figure 3 shows that the average fraction of duplicated exons is strongly correlated with the size of the exon group (Spearman $\rho = 0.943$, Pearson $\rho = 0.866$), so that very large exon groups are dominated by duplicated exons. Among these, we find several genes coding for collagens, e.g. COL4A2, COL4A4, and COL27A1, all of which contain large numbers of copies of a single exon. Other examples of this type are the fibrillin FBN3 and the nebulin NEB, a giant protein component of the cytoskeletal matrix. Multiple copies of a pair of unrelated exons are found in the apolipoprotein(a) precursor LPA, which contains 5-50 copies of kringle-type domains depending on the individual [15]. Our analysis shows that each of the 15 kringle-type domains of the isoform included in the RefSeq gene list consists of two consecutive exons (see also Figure 7A).

Fig. 4A shows that most duplicate exons have nearly the same length: in 62.8% of the exon pairs the shorter copy reaches as least 80% of the length of the longer exon, i.e., duplicated exons are predominantly incorporated as a whole. Possibly this results from the fact that exons often represent structural as well as functional domains [16]. Interestingly, the distribution shows a second peak centered at a length ratio of 0.5, for which we have no explanation. We furthermore observe the expected positive correlation between sequence similarity and length similarity (Spearman $\rho = 0.509$, Kendall $\tau = 0.376$, $p < 10^{-100}$, see Supplemental File 3). Many of the

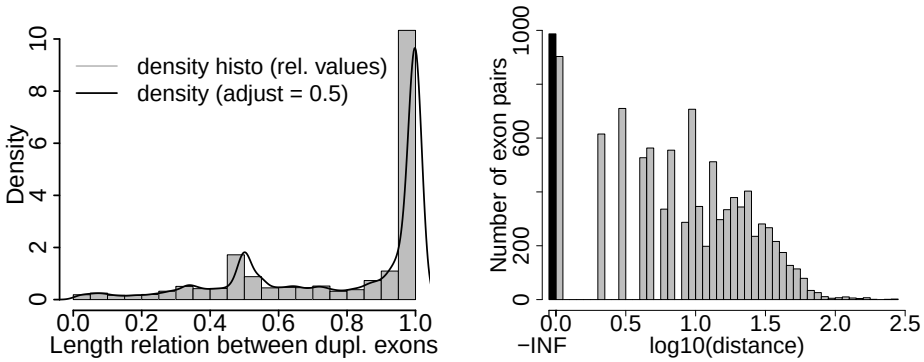

Fig. 4. (A) Distribution of relative sizes of duplicated exons. The relative length is computed by dividing the length of the shorter exon by the length of the larger one. (B) Log-scale histogram of the distance between duplicated exons measured as the number of exons located between the duplicates. The red bar shows the number of duplicates that are neighbored to each other and thus have a distance of 0, which leads to a log-distance of $-\infty$.

duplicated exons are adjacent or separated by only a small number of intervening exons, Fig. 4B.

Duplicated exons and gene diversity

On average, between 1/5 and 1/4 of an exon group with duplicates is composed of duplicated exons. We observe that exon groups with duplicated exons more often contain multiple annotated genes than exon groups without recognizable duplicates, Fig. 5. The difference is highly significant (Wilcoxon rank test $p < 2.2 \times 10^{-16}$). In groups not including duplicated exons, each exon is overlapped by 1.55 genes on average, while in groups including duplicates we found 1.72 overlapping genes per exon on average.

For each exon we determined whether there are known events of alternative splicing, such as skipping or alternative 3' or 5' splice sites. The fraction of potentially alternatively spliced exons is much higher among duplicated exons than on not duplicated exons (18.0% vs. 13.4%, Fisher's exact test $p < 2.2 \times 10^{-16}$). When analysing different types of alternatively spliced exons separately, we see the most dramatic differences between duplicated and un-duplicated exons among the skipping events. In contrast, alternative splice site events (3' or 5') show a smaller difference even though they are still significant. Taken together, these data demonstrate that exon duplication contributes to the complexity of alternative splicing predominantly by including additional, facultatively skipped exons.

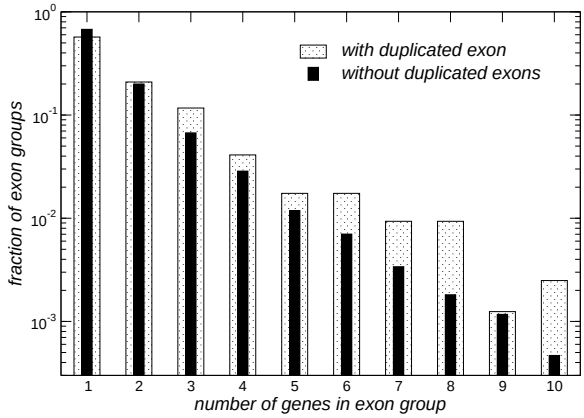

Fig. 5. Exon groups with duplicated exons are more often covered by many genes than exon groups without duplicates (Wilcoxon rank test $p < 2.2 \times 10^{-16}$).

Exon duplication and paralogous genes

Only 63 of the 1607 exon groups (3.9%) have paralogs. Pearson correlation coefficients around 0.1 show that there is no significant correlation between exon duplication at the same genomic locus and the distribution of paralogous genes elsewhere in the genome. This implies that a protein's evolutionary flexibility w.r.t. to the incorporation of an extra domain is not related with its retention rate after duplication, i.e., its propensity for sub- or neofunctionalization after a gene duplication [17]. In 51 of these 63 exon groups, duplicate exons are present also in the paralogous exon group. A comparison of the fraction of duplicated exons in paralogous groups shows a Pearson correlation of 0.63, see Supplemental File 4. This suggests that the tolerance to exon duplication is determined by structural or functional properties of the protein, which typically do not change dramatically as a result of gene duplication.

Distribution of duplicate exons within exon groups

Many exon groups show duplications of more than a single exon. In order to analyze this phenomenon in more detail we looked at the patterns of duplication more closely. To this end we encode each exon group as a string so that homologous duplicated exons are represented by the same letter, see Materials and Methods for details. Exons without duplicates (shown as '-' in the examples below) are ignored in the analysis since they either were not involved in duplications or have diverged beyond recognition.

We determined all duplicated patterns, i.e., all substrings that appear at least twice with a length of at least 2 (since in this analysis we are only interested in dupliation patterns more complicated than a single exon duplication). This ap-

Table 1. Frequent duplication patterns. The left part of the table lists the top 10 longest duplicated and irreducible patterns and the number N of exon groups in which they occur. The right hand part of the table lists the most frequently duplicated patterns within one exon group, the number N of exon groups in which they occur and the average number f of occurrences per exon group.

	longest irreducible		most frequent		
rank	pattern	N	pattern	N	f
1	AB	100	AA	156	7.1
2	AAA	71	AB	118	2.5
3	AA	58	ABC	27	2.7
4	ABC	26	ABCD	9	2.1
5	ABCD	15	ABCDE	5	2.2
6	ABA	9	ABCDEF	4	2.0
7	ABCDE	8	ABB	2	3.0
8	AAABA	6	ABCDEFGH	2	2.0
9	AABA	5	ABCDEFGD	1	5.0
10	ABCDEF	4	ABA	1	2.0

proach provides only a crude approximation of the true duplication history, which would require a detailed phylogenetic analysis. Nevertheless, the simple combinatorial analysis already provides useful insights.

A pattern is irreducible if it does not consist two or more copies of shorter patterns. The "longest irreducible patterns" are those irreducible patterns that cannot be written as substrings of longer irreducible patterns. We can interpret these as (an approximation of) the blocks that are duplicated as a unit. Table 1 lists the most frequent duplicated patterns and the most frequent longest irreducible patterns in exon groups. The full tables are available as Supplemental File 5. As expected, we observe that the frequency of irreducible blocks quickly decreases with pattern length.

Our data set contains 62 distinct longest duplicated patterns. These are the longest patterns that occur at least twice in the whole string representing an exon group independent of whether it is possible to split it in further subpatterns. Out of these 62 patterns, 23 consist only of tandem duplications, i.e., of runs of As. The largest of those longest duplicated pattern of the form $\mathtt{A}_k$ consists of $k = 52$ exons. The 62 patterns are constructed from 29 distinct longest irreducible patterns, the majority of which correspond to exon pairs AA and AB. The abundance of AB and ABC patterns shows that contiguous exon sub-groups are duplicated just like individual exons. When searching for the most frequently duplicated pattern in one exon group, we found short and simple patterns like AA and AB overrepresented, which can easily be explained by the nature of the analysis.

The frequency with which duplicated exons are incorporated depends on their position in the exon group. Both first and last exons are much less likely to have duplicates in the same exon group. A less pronounced decrease is also seen for the second and second-last exons, respectively. Otherwise the chance to encounter

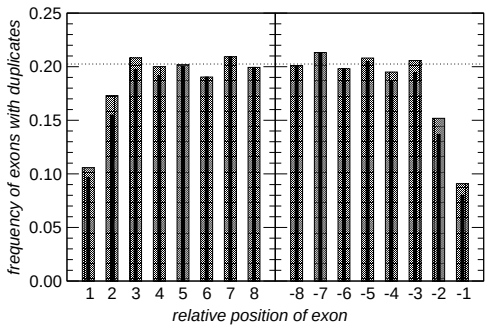

Fig. 6. Frequencies of exons with duplicates as a function of their position in an exon group. A reduction of duplicated exons is clearly visible at both ends. The thin bars refer to a dataset from which exon groups with less than 6 exons were removed.

duplicates is position independent (see Fig. 6). We repeated the computation after removing exon groups with fewer than 6 exons in order to avoid that the first three positions overlap with the last three positions. Since exon groups with duplicated exons and a small total number of exons are rare, we see the same result.

This distribution is expected when duplicate exons are randomly generated by segmental duplication. Immediately after duplication, copies of first and last exons lack a 5' splice acceptor and a 3' splice donor, respectively. Hence we expect that they are less frequently incorporated into transcripts. The second exons from either end often contain the start or stop codons, respectively, and hence are partially non-coding. Hence they often should fit poorly when copied to an internal position in a protein coding sequence. As a consequence, we expected that their incorporation is more strongly selected against.

Some genes exhibit elaborate duplication patterns recording a complex history, see Figure 7 for two rather different examples. A large exon group consisting of a single gene with 40 exons codes for a serine proteinase that constitutes a substantial portion of lipoprotein(a) (LPA) and contains 15 copies of kringle-type domains. All 15 copies of the domain are found by our analysis and identified as copies of each other. Each domain consists of two exons indicated by 'B' and 'C' in Fig. 7A. For this protein dramatic variations among individuals in the number of copies of the kringle-type domains (from 5 to 50 [15]) have been reported. A recent systematic review of 40 studies showed that the risk of coronary heart disease is 2-fold higher for people with shorter isoforms ($\leq$ 22 kringle-type domains) than for those with larger proteins ($>$ 22).

In the second example, a total of 10 exons form three overlapping genes of the cancer/testis antigen family 45. Two genes, CT45A2 and CT45A3, are disjoint, while CT45A4 uses the first three exons of CT45A2 and the last exon of CT45A3. The CT45 gene family, first described in [18], includes six members, three on the positive strand (CT45A1-3) and three on the opposite strand (CT45A4-6). They are

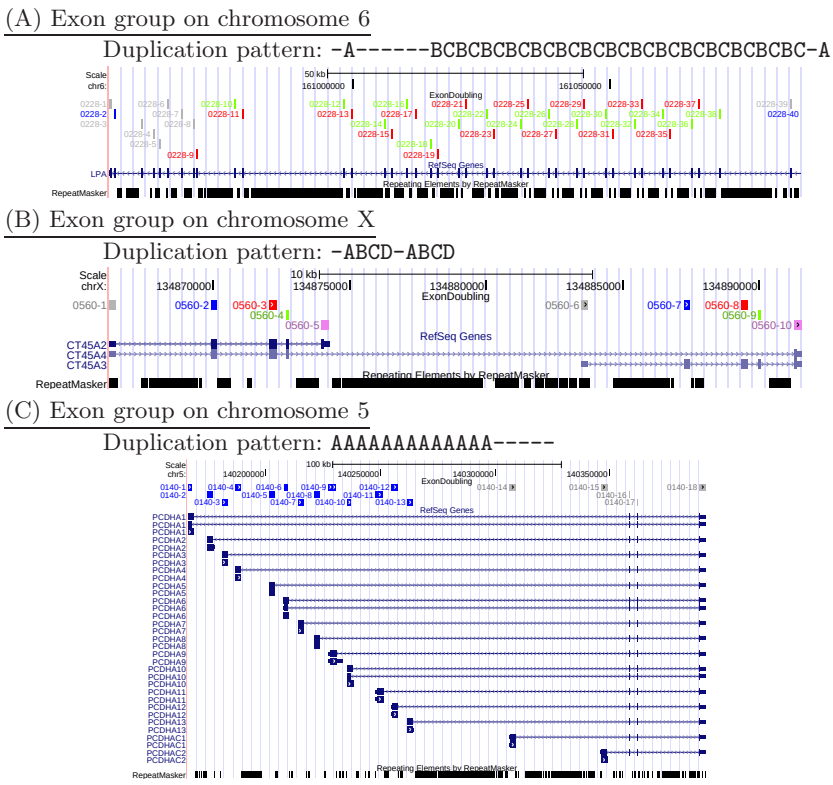

Fig. 7. Examples of complex exon groups. (A) A serine proteinase that inhibits the activity of tissue-type plasminogen. The encoded protein constitutes a substantial portion of lipoprotein(a) (LPA). (B) Cancer/testis antigen family 45 (CT45A2,CT45A3,CT45A4). (C) Protocadherin alpha gene cluster represented as a single exon group composed of tandem copies of the same exon.

products of recent gene duplication events and differ only by 2 to 12 nucleotides [18]. Surprisingly, a second identical copy of CT45A4 can be found on the positive strand consisting of the first three exons of CT45A2 and the last exon of CT45A3. It is shown in Fig. 7B.

The three largest exon groups comprise 316, 183, and 149 exons (a list of all exon groups including more than 80 exons can be found in Supplemental File 6). The largest number of exons belongs to the many isoforms of titin (TTN), a large abundant protein of striated muscle. Our analysis emphasizes that titin is a highly repetitive gene even if it does not detect the super-repeat pattern discussed in [19].

The 183 exons of different gene variants of nebulin, a giant protein component of the cytoskeletal matrix, include repeated modules. 97% of its polypeptides are arranged into simple repeats or super repeats [20]. The duplication pattern highlights the highly repetitive structure of the gene. It does not, however, indicate the super repeats.

While the first two cases of large exon groups are examples of highly repetitive proteins with high fractions of duplicated exons, our analysis does not detect such a repetitive structure in the third-largest exon group. Among the 149 exons belonging to SYNE1 (spectrin repeat containing, nuclear envelope 1) only four exons have obvious duplicates. SYNE1 is expressed in skeletal and smooth muscle and characterized by the presents of spectrin repeats [21]. Although ancestrally related, the sequences of the individual exons have diverged beyond our detection threshold, an effect that is confounded by the fact that many of the SYNE1 exons are shorter than the cut-off value of 54 nucleotides. SYNE1 thus shows that our analysis is conservative and focuses in particular on recent exon duplications.

Figure 7C shows a exon group built of the 15 cadherin superfamily genes of the protocadherin alpha gene cluster. This cluster is known to consist of 13 highly similar and two more distantly related coding sequences. The tandem array of 15 variable exons is followed by three downstream exons shared by all genes in the cluster [22,23]. These features are perfectly represented by our exon group analysis. The first 13 exons of the group are indicated as duplicates, while exons 14 and 15 are not. When taking a closer look on similarities, we find relatively high similarities also between these exons and the first 13 exons, which are only slightly below the cutoff of 56, while the similarity score of each of the first 15 exons to each of the last three (unrelated) exons is far below the threshold.

Exon Duplication and Gene Function

Among the set of exon groups we found 77 that include at least 10 duplicates of the same exon. 71 of them only include one RefSeq gene, while of the remaining six exon groups two consist of unrelated genes and four are composed of members of the same family. In order to search for common gene ontology terms (GO terms), we neglected members of the same gene family within one exon group and ended up with 79 gene names that were analyzed for common GO terms using a web-based version of `GOTermFinder` [14]. All GO terms with a p-value[a] $p < 10^{-15}$ are presented in Table 2, a complete list of all GO terms with $p < 10^{-5}$ is given in Supplemental File 7. Not surprisingly, collagens and components of the extracellular matrix are the most common types of proteins with highly repetitive exons.

The same analysis was performed for the following sets of exon groups: (i) groups with at most 2 duplicates of an exon; (ii) groups with at most 5 copies of a single exon containing at least one exon that has three or more copies; and (iii) exon groups with at most 9 copies of a single exon with at least one exon that has 6 or more copies. As above, we removed multiple members of the same gene family within one exon group. Set (i), containing genes with few duplicates of the same exon, covers 1217 different genes. Table 3 lists the highly significant GO terms with a p-value of $< 10^{-15}$, a more complete lists is is given in Supplemental File 7.

[a] A Bonferroni correction for multiple testing is applied already by `GOTermFinder`.

Table 2. Shared GO terms among the 79 genes in exon groups that include at least 10 duplicates of the same exon. A) component GO terms; B) function GO terms; C) process GO terms.

	p-value	GO
A)	3.00×10^{-51}	collagen
	1.21×10^{-48}	extracellular matrix part
	7.62×10^{-45}	proteinaceous extracellular matrix
	3.58×10^{-42}	extracellular matrix
	2.93×10^{-29}	extracellular region part
	1.07×10^{-25}	extracellular region
	9.49×10^{-23}	fibrillar collagen
	2.03×10^{-21}	basement membrane
B)	3.66×10^{-39}	extracellular matrix structural constituent
	5.24×10^{-29}	structural molecule activity
C)	5.15×10^{-20}	extracellular matrix organization
	8.31×10^{-20}	cell adhesion
	8.60×10^{-20}	biological adhesion
	3.31×10^{-16}	extracellular structure organization

Most notably, the genes are frequently involved in a binding function, which is probably assisted by duplication of exons containing binding domains. In contrast, no significant overrepresentation of GO terms was found for sets (ii) and (iii).

We not only sorted exon groups according to their maximal copy number of one exon, but also analyzed them based on their overall fraction of duplicated exons. The 1607 exon groups including duplicates were divided into four subsets by their fraction f of duplicated exons: (i) $0 < f \leq 0.25$; (ii) $0.25 < f \leq 0.5$; (iii) $0.5 < f \leq 0.75$; and (iv) $0.75 < f \leq 1$. Subset (i), with a small fraction of duplicates, in particular includes genes involved in protein binding and components of the cytoskeleton. In contrast, genes of the highly duplicated subset (iii) are frequently parts of collagens or found in the extracellular matrix. No significant overrepresentations are found in groups (ii) and (iv). For (iv), this is a consequence of the small number of only 19 genes that are included in this subset. A complete list of all GO terms with $p < 10^{-5}$ is given in Supplemental File 7.

Discussion

In line with previous work [1,2] we observe that locally duplicated exons are a frequent phenomenon in the human genome. Our data show that duplication events are not restricted to tandem duplications of individual exons. Duplications frequently affect small groups of exons that appear to be copied as a unit. In concert with subsequent exon loss and exonization of intervening sequence elements large complex gene structures may arise.

The over-representation of annotated alternatively spliced exons links exon duplication to the evolution of alternative splicing, a connection that has been made

Table 3. Shared GO terms among the 1217 genes in exon groups that include at most 2 duplicates of the same exon. A) component GO terms; B) function GO terms; C) process GO terms.

	p-value	GO
A)	none	
B)	2.23×10^{-29}	binding
	3.91×10^{-19}	protein binding
C)	7.48×10^{-16}	cellular process

repeatedly based on different lines of evidence, see [3,1]. Our data also indicate an increased level of alternative splicing, in particular exon skipping, in response to abundant duplicate exons.

In the Duplication-Degeneration-Complementation (DDC) model [17], retention of paralogous copies of entire genes is explained by complementary loss of functionalities of the two copies. Alternative, mutually exclusive splicing of duplicate exons thus appears as a plausible first step leading towards the accommodation of duplicate exons in small proteins: Immediately after the duplication event the two exons are identical so that the alternative inclusion of either one of them leads to the same protein product, minimizing the fitness effect of duplication event. Assuming differential production of the splicing alternatives in different tissues, developmental stages, or environmental conditions, the duplication would subsequently provide an opportunity for divergent adaptive evolution of both copies, leading to the fixation of both copies. Indeed, duplicated exons are often associated with mutually exclusive alternative splicing [3,1]. A recent study comparing human and mouse genes shows, however, that tandem duplication usually does not by itself introduce alternative splicing. Rather, exons that are already subject to alternative splicing propagate this capability upon duplication [4].

The patterns of exon duplication observed in our data are consistent with a mode of evolution in which gene segments are randomly duplicated and duplicate exons are randomly incorporated into transcripts. Selection then purges copies that interfere with protein function. As expected from such a model, first and last exons are much less frequently incorporated, probably because they are less likely to be included into transcripts. Second and second-last exons have a reduced chance giving rise to functional duplicates because they often contain start and stop codons and hence non-protein-coding regions.

The observation that the abundance of paralogs does not correlate with the propensity to incorporate duplicate exons suggests that the selection pressures leading to the expansion of proteins by additional domains are unrelated to the reasons for retaining paralogous genes. As proteins with small copy numbers of duplicated exons are mostly involved in some type of binding, duplicated exons may contribute additional binding domains. The mechanism for retention of exon duplications may be more straight-forward in large proteins consisting of repetitive modules such as

nebulin. Here exon duplication changes the copy number of repetitive units which presumably are easily accommodated in such proteins as they are likely to cause only small changes of the protein's physical properties [24].

The correlation of duplicated exons with protein functions observed here at a very crude statistical level suggest to investigate this topic in more detail, focusing e.g. on the relationships of exon duplication and the protein domains entirely or partially encoded within the exons.

Authors' contributions

Both authors contributed to the design of the study, the interpretation of the data, and wrote the manuscript. AB performed the computational analysis.

Acknowledgments

This work was funded in part by the 6th Framework Programme of the European Union (projects 043312 "SYNLET") and by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD).

References

1. Ivica Letunic, Richard R. Copley, and Peer Bork. Common exon duplication i n animals and its role in alternative splicing. Human Mol. Genetics, 11:1561–1567, 2002.
2. Xiang Gao and Michael Lynch. Ubiquitous internal gene duplication and intron creation in eukaryotes. Proc Natl Acad Sci U S A., 106:20818–20823, 2009.
3. Fyodor A. Kondrashov and Eugene V. Koonin. Origin of alternative splicing by tandem exon duplication. Human Mol. Genetics, 10:2661–2669, 2001.
4. T Peng and Y Li. Tandem exon duplication tends to propagate rather than to create de novo alternative splicing. Biochem. Biophys. Res. Commun., 383:163–166, 2009.
5. A Gardiner, D Barker, R K Butlin, W C Jordan, and M G Ritchie. Evolution of a complex locus: exon gain, loss and divergence at the Gr39a locus in Drosophila. PLoS One, 3:e1513, 2008.
6. M Irimia, I Maeso, P W Gunning, J Garcia-Fernàndez, and S W Roy. Internal and external paralogy in the evolution of tropomyosin genes in metazoans. Mol Biol Evol, 27:1504–1517, 2010.
7. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. Genome Res, 12(6):996–1006, 2002.
8. A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2006. Nucleic Acids Research, 34(Database issue):D590–598, 2006.
9. J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTALW: 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, 1994.
10. M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J.

Gibson, and D. G. Higgins. Clustal W and Clustal X version 2.0. Bioinformatics, 23:2947–2948, 2007.
11. A.F.A. Smit, R. Hubley, and P. Green. RepeatMasker Open-3.0. http://www.repeatmasker.org, 1996-2004.
12. E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova.and L. Wagner, E. Yaschenko, and J. Ye. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 37(Database issue):D5–15, 2009.
13. E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, L. Y. Geer, W..Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A..Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 38:D5–16, 2010.
14. E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics, 20:3710–3715, 2004.
15. S. Erqou, A. Thompson, E. Di Angelantonio, D. Saleheen, S. Kaptoge, S. Marcovina, and J. Danesh. Apolipoprotein(a) isoforms and the risk of vascular disease: systematic review of 40 studies involving 58,000 participants. J Am Coll Cardiol, 55(19):2160–7, 2010.
16. Mingyi Liu, Heiko Walch, Shaoping Wu, and Andrei Grigoriev. Significant expansion of exon-bordering protein domains during animal proteome evolution. Nucl. Acids Res., 33:95–105, 2005.
17. Allan Force, Michael Lynch, F. Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. Genetics, 151:1531–1545, 1999.
18. Y.-T. Chen, M. J. Scanlan, C. A. Venditti, R. Chua, G. Theiler, B. J. Stevenson, C. Iseli, A. O. Gure, T. Vasicek, R. L. Strausberg, C. V. Jongeneel, L. J. Old, and A. J. G. Simpson. Identification of cancer/testis-antigen genes by massively parallel signature sequencing. Proc Natl Acad Sci U S A, 102:7940–7945, 2005.
19. S. Labeit, M. Gautel, A. Lakey, and J. Trinick. Towards a molecular understanding of titin. EMBO J, 11:1711–1716, 1992.
20. K. Donner, M. Sandbacka, V.-L. Lehtokari, C. Wallgren-Pettersson, and K. Pelin. Complete genomic structure of the human nebulin gene and identification of alternatively spliced transcripts. Eur J Hum Genet, 12:744–751, 2004.
21. Q. Zhang, J. N. Skepper, F. Yang, J. D. Davies, L. Hegyi, R. G. Roberts, P. L. Weissberg, J. A. Ellis, and C. M. Shanahan. Nesprins: a novel family of spectrin-repeat-containing proteins that localize to the nuclear membrane in multiple tissues. J Cell Sci, 114:4485–4498, 2001.
22. K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res, 35(Database issue):D61–65, 2007.
23. K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott. NCBI Reference Sequences:

current status, policy and new initiatives. Nucleic Acids Res, 37(Database issue):D32–36, 2009.

24. K. Donner, M. Sandbacka, V.-L. Lehtokari, C. Wallgren-Pettersson, and K. Pelin. Complete genomic structure of the human nebulin gene and identification of alternatively spliced transcripts. Eur J Hum Genet, 12:744–751, 2004.

Supplemental Files

Supplemental File 1.

Examples of RefSeq gene annotations that need filtering. a) different gene locations of the same gene (e.g. in case of gene duplications); b) identical gene names referring to several distinct but overlapping exons (alternative splicing).

Supplemental File 2.

Relation between exon group size and length of its coding sequence(s).

Supplemental File 3.

Correlation between relative size and similarity of the duplicate exons.

Supplemental File 4.

Correlation of the fraction of duplicated exons in paralogous exon groups. The x-axis always shows the smaller value among the two groups. The gray line represents the linear regression.

Supplemental File 5.

Complete tables of exon duplication patterns

Supplemental File 6.

List of all large exon groups with more than 80 exons.

Supplemental File 7.

List of significant GO terms characterizing exon groups depending on their maximal copy number of the same exon as well as on their general fraction of duplicated exons.