Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Gene 365 (2006) 41 – 47

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

The evolution of introns in human duplicated genes

Edda Rayko [a,1], Kamel Jabbari [a,1], Giorgio Bernardi [b,*]

[a] *Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, F-75005 Paris, France*
[b] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Naples, Italy*

Received 22 April 2005; received in revised form 7 July 2005; accepted 7 September 2005
Available online 13 December 2005
Received by T. Gojobori

Abstract

In previous work [Jabbari, K., Rayko, E., Bernardi, G., 2003. The major shifts of human duplicated genes. Gene 317, 203–208], we investigated the fate of ancient duplicated genes after the compositional transitions that occurred between the genomes of cold- and warm-blooded vertebrates. We found that the majority of duplicated copies were transposed to the "ancestral genome core", the gene-dense genome compartment that underwent a GC enrichment at the compositional transitions.

Here, we studied the consequences of the events just outlined on the introns of duplicated genes. We found that, while intron number was highly conserved, total intron size (the sum of intron sizes within any given gene) was smaller in the GC-rich copies compared to the GC-poor copies, especially in dispersed copies (i.e., copies located on different chromosomes or chromosome arms). GC-rich copies also showed higher densities of CpG islands and Alus, whereas GC-poor copies were characterized by higher densities of LINEs. The features of the copies that underwent the compositional transition and became GC-richer are suggestive of, or related to, functional changes.

Keywords: Genome; Isochores; CpG islands; Alus; LINEs

1. Introduction

Since many gene duplications present in the human genome are ancient duplications going back to the origin of vertebrates (see Postlethwait et al., 2004, for a review), a question may be asked about the fate of such duplicated genes after the compositional transitions that occurred between the genomes of cold- and warm-blooded vertebrates (see Bernardi, 2004, for a review). Indeed, at those transitions, the gene-dense "ancestral genome core" of cold-blooded vertebrates underwent a GC enrichment to become the "genome core" of warm-blooded vertebrates (see Fig. 1).

We could show that, by far and large, one copy of the duplicated gene underwent a GC enrichment, the other copy keeping the original GC level (Jabbari et al., 2003). We hypothesized that the former copy was preferentially translocated into the gene-dense compartment of the genome, the "ancestral genome core", namely the gene space which underwent the compositional transition (GC enrichment) at the emergence of warm-blooded vertebrates (see Fig. 2). This hypothesis was based on three assumptions: (i) that duplication occurred more frequently in the gene-poor compartment of the genome of the cold-blooded ancestors, the "ancestral genome desert", as suggested by the abundance of gene duplications in GC-poor pericentromeric regions (see Jabbari et al., 2003, for references); (ii) that one copy acquired a new function; and (iii) that transposition of the duplicated copy into the "ancestral genome core" rather than into the "genome desert" (the gene-poor part of the genome; see Fig. 1) was generally preferred. The latter assumption was based on the observation that the chromatin of the "ancestral genome core" is in an open configuration (as shown by the results of Federico et al., in press, on the genomes of *Rana esculenta* and *Podarcis sicula*), and that integration of retroviral sequences preferentially occurs into open chromatin (see Rynditch et al., 1998; Tsyba et al., 2004). This assumption

[*] Corresponding author. Tel.: +39 081 5833402; fax: +39 081 2455807.
E-mail addresses: erayko@biologie.ens.fr (E. Rayko), kjabbari@biologie.ens.fr (K. Jabbari), bernardi@szn.it (G. Bernardi).

[1] Present address: Département de Biologie, Ecole Normale Supérieure, 46 Rue d'Ulm, F-75230 Paris, France.

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.09.038

was also justified by the fact that it could account for the higher gene density of the “ancestral genome core” compared to the “genome desert” (see Fig. 1).

As far as structural and functional consequences of the major shift of duplicated genes are concerned, the fact that the majority of duplications are ancient duplications, suggested that the copy that experienced a GC shift might exhibit the same properties as any GC-rich gene of the human genome (see Bernardi, 2004 for a review), namely an enrichment in Alus (Jabbari and Bernardi, 1998) and flanking CpG-islands (Aïssani and Bernardi, 1991a,b), a decrease in LINEs (Pavlicek et al., 2001) and a shortening of introns (Duret et al., 1995).

Here we considered, therefore, the consequences of the events just outlined on the size of the introns and on the frequencies of CpG islands, Alus and LINEs located in the introns of duplicated genes. The changes that we found in the GC-rich copies compared to the GC-poor copies further support the idea that translocations preferentially took place from the gene-poor “ancestral genome desert” to the gene-rich “ancestral genome core”. Moreover, they are suggestive of functional changes in the copies that underwent the compositional change.

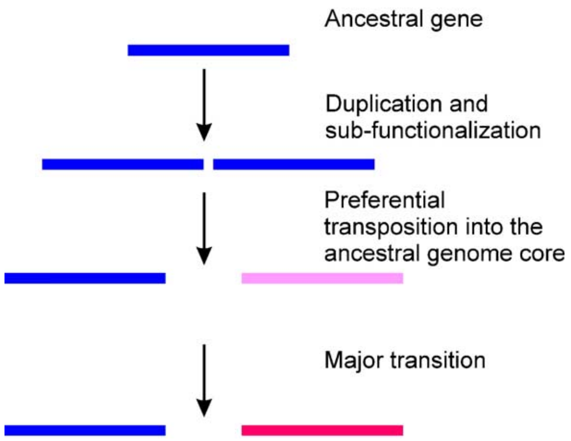
Ancestral gene
Duplication and sub-functionalization
Preferential transposition into the ancestral genome core
Major transition

Fig. 2. A scheme of the most frequent pathway following ancient gene duplication (blue bars). One copy is supposed to be preferentially transposed into the ancestral genome core (pink bar), which then undergoes the major compositional transition (red bar) (modified from Jabbari et al., 2003).

2. Materials and methods

We retrieved 3947 complete human coding sequences from HOVERGEN (Duret et al., 1994), with gene family annotation. We collected pairs of coding sequences (CDS) with similar size (the difference in size being lower than, or equal to, 7%) from each family. Redundant and alternatively spliced CDS were disregarded. This led to a data set of 412 pairs of duplicated human genes representing 214 gene families and including 596 individual CDS. Exon and intron sizes were calculated using their coordinates along the gene sequences, as indicated in EMBL annotations. We excluded a few pseudogenes that were present in the group of intronless genes. We should notice that only the exons containing the initiator methionine were considered as first exons, last exons being those ending with a stop codon (Supplementary Table 1).

CpG islands were searched using the program cpgplot http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/cpgplot.html. This program defines, by default, a CpG island as a region where the GC level is over 50%, the calculated Observed/Expected (O/E) CpG ratio is over 0.6, these conditions holding for a minimum of 200 bases. Alu and LINE sequences located within introns were identified using RepeatMasker (A. Smit and P. Green, unpublished, http://ftp.genome.washington.edu/RM/RepeatMasker.html).

χ^2-test for means comparisons with a two-sided hypothesis was used to test the statistical significance; 1% standard errors are indicated (see figures). We used correlation analysis to test the statistical significance of the association between the

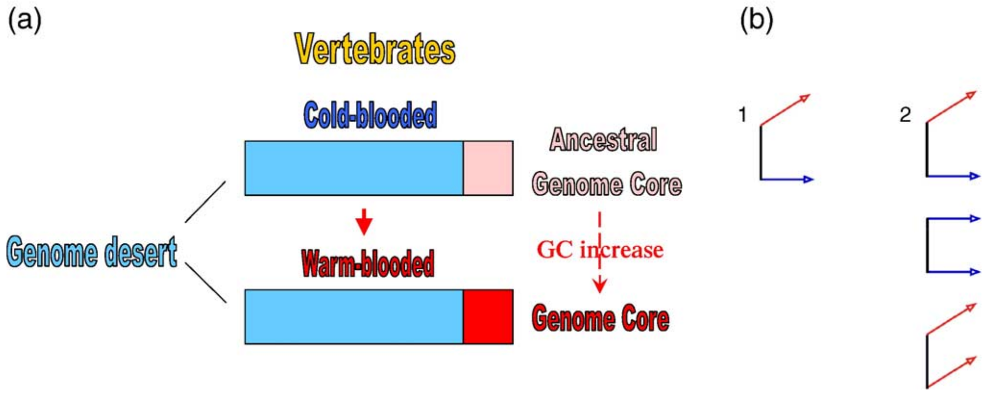

Fig. 1. (a) Scheme of the compositional genome transitions that took place between cold- and warm-blooded vertebrates. The “genome desert” of cold-blooded vertebrates is GC-poor and gene-poor (blue box) and essentially did not undergo any compositional change. In contrast, the gene-dense, moderately GC-rich “ancestral genome core” (pink box) underwent a compositional change into a gene-dense, GC-rich “genome core” (red box). (b) A model of the two situations hypothesized for duplicated genes at the transitions from cold- to warm-blooded vertebrates. (1) One copy of each pair preferentially underwent the transition (red arrow), the other copy maintaining its original low GC level (blue arrow). (2) In addition to situation 1, both copies underwent the transition or maintained their original low GC level (modified from Jabbari et al., 2003).

parameters analysed here. Regression equations and p values are also given.

3. Results and discussion

3.1. Intron number and size

The 412 human duplicated gene pairs retrieved were split, as in previous work, into two sets according to GC_3 (the GC level of third codon positions in coding sequences). In each pair, the set 1 copy was GC_3-rich compared with the set 2 copy. For 6 gene pairs, both copies had the same GC_3 content and were randomly assigned to one of the two sets. The number of exons/introns was the same in 253 (61.4%) pairs and different in 159 (38.6%) pairs; in 35 pairs (22.0%) from the latter set, one counterpart was intronless.

As shown in Fig. 3, intron number is remarkably conserved, as previously shown in Arabidopsis/Maize orthologous genes (Carels and Bernardi, 2000). The correlation between intron number in the GC-rich and the GC-poor sets has a coefficient, r, of 0.94 and a slope of 0.91. Most interestingly, in the case of different pairs intron number tended to be systematically higher in the GC_3-poor copies. In particular, intronless genes were GC_3-rich, their GC_3-poor counterparts comprising several introns.

A comparison of GC_3 levels showed that the differences between GC_3-rich and GC_3-poor copies (both for average coding sequences and for average exons; Fig. 4a,b) were larger for dispersed than for clustered genes, namely for the genes located on the same or on different chromosomes or chromosome arms, respectively (Fig. 4a). Likewise, a comparison of total intron size (the sum of intron sizes within any coding sequence) showed that the difference between the GC-rich and the GC-poor sets for dispersed gene pairs (9.5 kb vs. 15.5 kb), was larger than for clustered gene pairs (5.7 kb vs. 7.1 kb) (Fig. 4c). This could be expected since the former are the result of more recent duplications. Similar results were observed for individual duplicated intron pairs (Fig. 4d). Interestingly, dispersed copies were more than twofold as frequent as clustered copies (see Table 1). This ratio is much lower than our previous estimate (Jabbari et al., 2003). This difference is,

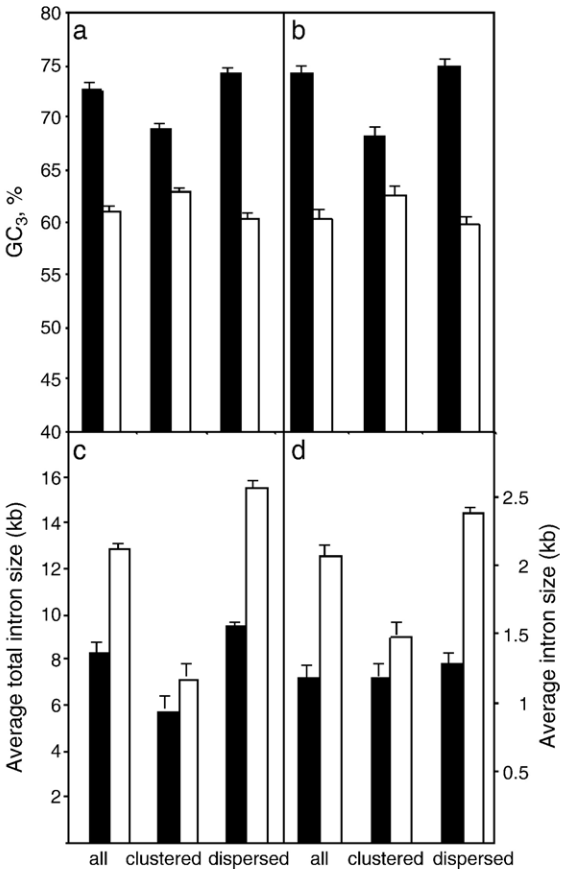

Fig. 4. Average GC_3 levels for average coding sequences (a) and for average exons (b). (c) and (d) refer to the average size of total and individual introns, respectively. Black and white bars concern GC-rich and GC-poor copies, respectively, of duplicated genes. The difference in average intron size between GC_3-poor and GC_3-rich clustered sets (c) is not statistically significant (p=0.07).

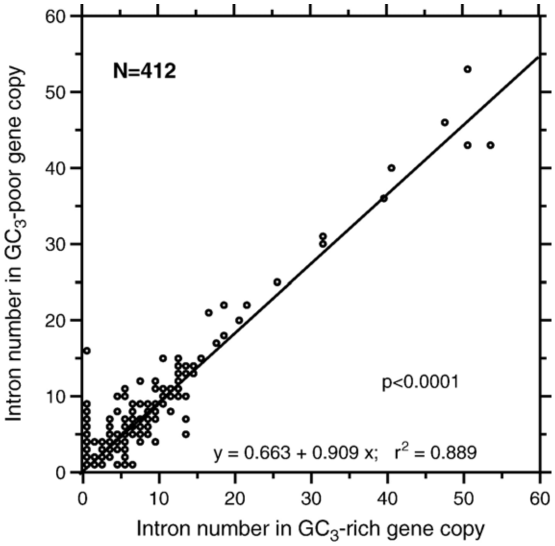

Fig. 3. Plot of intron number (nb) of GC-rich copies against intron number of GC-poor copies of human duplicated genes.

Table 1
CpG islands, Alus and LINEs in duplicated gene pairs

Element tested	Paralog localization	Gene pairs number	Number and % of pairs with one CpG islands- or Alu/LINE-containing copy	Pairs number where CpG islands- or Alu/LINE-containing copy is	
				GC_3-richer	GC_3-poorer
CpG islands	Any	365	99 (27.1%)	65 (65.7%)	34 (34.3%)
	Clustered	115	18 (15.6%)		
	Dispersed	250	81 (32.4%)		
Alus	Any	310	107 (34.5%)	43 (40.2%)	64 (59.8%)
	Clustered	106	31 (29.2%)		
	Dispersed	204	76 (37.3%)		
LINEs	Any	310	112 (36.1%)	48 (42.0%)	64 (58.0%)
	Clustered	106	27 (25.5%)		
	Dispersed	204	85 (41.7%)		

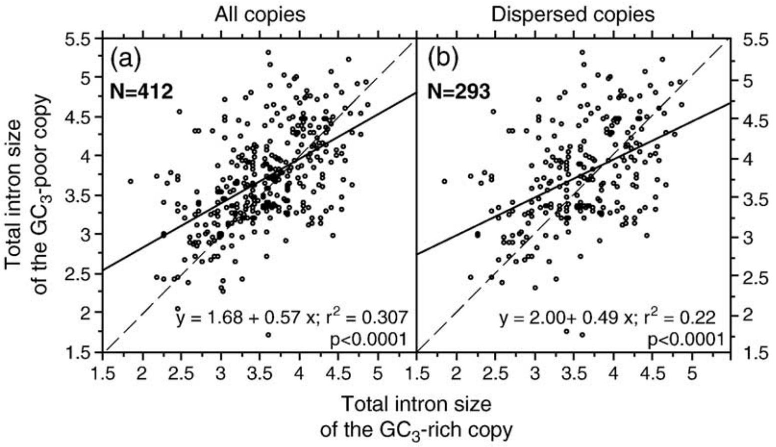

Fig. 5. Plot of total intron size of GC-rich and GC-poor copies for all 412 pairs (a); and for only 293 pairs of dispersed genes (b).

however, due to the fact that our previous estimate was based on the actual chromosomal localization of all copies, whereas in the present work genes referred to as "molecule = DNA" in EMBL (i.e. genomic sequences) were taken into consideration, leading to an overestimate of clustered copies.

We also analysed the correlation between total intron size in the GC-rich and the GC-poor sets (412 pairs). Since total intron size varies very much among the copies (from 0 bp to 212.5 kb), we used for this analysis logarithmic scales, as in Yu et al. (2002). As shown in Fig. 5a, the regression coefficient r^2, 0.307, was very significant ($p < 0.0001$). This value was lower, 0.22, when clustered copies were discarded, but the p value still was <0.0001 (Fig. 5b). Expectedly, the slope of the regression line was also lower.

The same analysis based on intron by intron comparisons in the data set of 253 duplicated pairs where the duplicated genes had the same number of exons/introns (1213 comparisons in total), showed (Fig. 6a) a very weak r^2 value (0.059; $p < 0.0001$). This value dropped to 0.0009 when clustered copies were excluded (not shown).

The possible common histories of indels between the two copies of a gene prompted us to express intron size not only in absolute values as in Fig. 6a, but also in relative values (i.e. in the relative contribution of individual intron size to the total intron size of a gene). In such analysis (Fig. 6b), the r^2 value increased considerably ($r^2 = 0.294$) and points tended to cluster around the diagonal. This value was lower, 0.234, yet still significant ($p < 0.0001$), when clustered copies were disregarded (Fig. 6c).

3.2. Intron positions in duplicated genes

Duplicated genes having the same number of exons/introns (253 pairs) contained a total of 1478 paralogous exons, of which 830 pairs (56.2%) were found to be of the same size. For 568 paralogous exons, size difference was a multiple of 3, implying that intron phase (intron position relative to the position within codons) was the same. In other words, for 1398 out of 1478 (94.6%) paralogous exons compared, intron position and/or intron phase were the same. This demonstrates that the gene organization of the 253 pairs of duplicated genes under consideration is very highly conserved. The conservation of the intron positions in paralogous genes has a biological significance, since clear instances of exon/intron structure

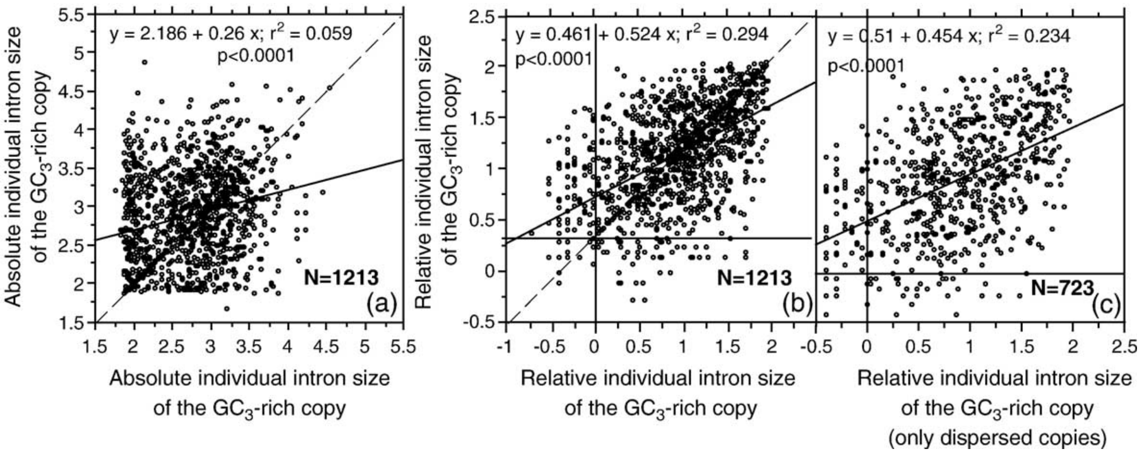

Fig. 6. Plot of individual intron size of GC-rich copy against the intron size of GC-poor copy. (a) and (b) plots of absolute and relative individual intron size, respectively; all 1213 copies were compared; (c) plot of relative individual intron size only for 723 dispersed copies.

conservation in vertebrates were demonstrated even in the case of low or non-significant paralogous protein sequence similarity (Betts et al., 2001). It should be noted that most cases of discordant introns in homologous genes (5.4% in this data set) were considered to be artefactual (Stoltzfus et al., 1997); nevertheless, intron sliding (the relocation of intron–exon boundaries) might well be a real phenomenon (Rogozin et al., 2000; Boudet et al., 2001).

3.3. Exon and intron size variation according to their position in duplicated genes

Among 253 pairs of duplicated genes having the same number of exons/introns, 215 (85%) pairs contained genes with at least 3 exons. For these pairs, the sizes of first, last and internal paralogous exons were compared. The first, last and internal exons were found to have the same size in 35.8%, 40.5% and 66.5% of exon pairs, respectively. This shows that exon size is more conserved for internal exons than for external ones.

We also calculated the absolute average difference size for first, last and internal paralogous exons. Again, external exons were found to be more subject to variations than internal ones: the average size differences of first and last exons were similar (12.5 and 13.9 bp, respectively), whereas that of internal exons was smaller (only 7.3 bp; see Fig. 7a). Incidentally, it should be noted that among 2602 exons of 215 gene pairs studied, the average sizes of first, internal and last exon were 138, 135 and 204 bp long, respectively. This agrees with the observation that the last exons tend to be the largest ones as previously observed among exons of human genes on chromosomes 21 and 22 (Chen et al., 2002).

For 155 (155/253 = 61%) pairs of genes with at least 3 introns, we compared also the sizes of paralogous introns having different positions in genes. The absolute average size differences for the first, internal and last introns were observed to be 3.3 kb, 1.7 kb and 1.4 kb, respectively (see Fig. 7b). Therefore, the first introns vary considerably more compared to the others. In this connection, it should be mentioned that among all 2036 introns of the 155 genes pairs studied, the average size of first introns (3.5 kb) was more than twofold larger than those of internal and last introns (1.5 kb and 1.3 kb, respectively). The largest size of first introns was demonstrated earlier for human genes on chromosomes 21 and 22 (Chen et al., 2002).

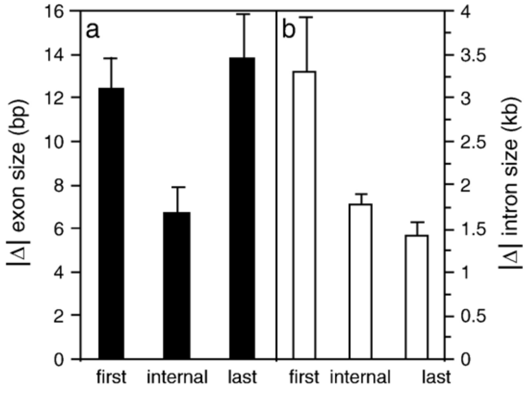

Fig. 7. First, last and internal exon/intron size conservation. (a) Absolute average exon size difference (bp); (b) absolute average intron size difference (kb). The internal exon/intron size is statistically different from first and last exon/intron size ($p < 0.001$).

3.4. CpG-islands

All the 596 genes studied in this work were analysed for the presence of clusters of unmethylated CpGs or "CpG islands" (Bird et al., 1985; Gardiner-Garden and Frommer, 1987) in introns. Six genes were found to be totally covered by CpG islands. In the case of the other 590 genes, 2580 introns (those larger than 200 bp, the minimal size of CpG islands detected by the program used, Alus being excluded) were considered. This led to the identification of 221 intronic CpG islands, 81 of them overlapping with adjacent exons, and 140 being purely intronic.

To estimate the conservation of intronic CpG islands in the compared duplicated genes, we selected 365 pairs in which both copies contained introns larger than 200 bp and could correspond to CpG islands.

As shown in Table 1, in 99 out of 365 (27.1%) pairs, only one of the two copies harbored intronic CpG islands. Interestingly, this was true for only 15.6% of pairs with clustered copies, but for 32.4% of pairs with dispersed copies (Table 1). It should be noted that in 65 out of 99 (65.7%) pairs considered, the GC_3 level of the paralog containing intronic CpG island was higher than that of the paralog not containing CpG island, suggesting the preferential formation of intronic CpG islands in the GC_3-rich copy (see also following section).

Although outside the scope of present work, we also estimated the conservation of CpG islands in 5′ region of duplicated genes. We selected 342 gene pairs where 500 nucleotides were known upstream of the start codon for both copies. In 70 out of 342 (20%) pairs analysed, both copies harbored 5′ CpG islands. Interestingly, in 43 out of 70 (61.4%) pairs, the GC_3-rich copy had a longer 5′ CpG island, and only in 38.6% of pairs, the GC_3-poor copy had a longer 5′ CpG island. In addition, the average 5′ CpG island size of the GC_3-rich copy was larger (923 bp) than that of the GC_3-poor copy (732 bp); *p*-value of the pairwise comparisons test is 0.0025.

In 117 out of 342 (52%) pairs, only one of the two copies harbored a 5′ CpG island. In 69 out of these 117 (59.0%) pairs, the 5′ CpG island-containing copy was the GC_3-rich copy. Note that these trends are in agreement with our previous results on CpG-island size and distributions in the human genome (Aïssani and Bernardi, 1991a,b; Jabbari and Bernardi, 1998).

3.5. Repetitive DNA sequences in introns

Using the program RepeatMasker, we performed a search and classification of repeated–Alu and LINE–elements in the introns studied. Such elements make up 10% and 15% of the human genome, respectively (Smit, 1996, 1999; Lander et al., 2001). Alu size being approximately 300 bp, only introns longer than 300 bp and thus susceptible to harbor Alus were selected. These 2298 introns (total size 3 Mb) corresponded to 501 genes.

In total, 1286 Alus and 1055 LINEs were detected. The average Alu and LINE density was 45.9 and 34.8 per 100 kb, respectively, 62.2% of Alus corresponding to AluS sequences (Batzer and Deininger, 2002) and 57.0% of LINEs corresponding to LINE1 sequences. Out of 2298 introns studied, 1211 (52.7%) contained neither Alus nor LINEs, whereas 433 introns (18.8%) harbored both Alus and LINEs.

It has been previously reported that Alu are more abundant in GC-rich isochores, whereas LINEs are preferentially located in GC-poor isochores (Soriano et al., 1983; Zerial et al., 1986), a point confirmed by later investigations (Jabbari and Bernardi, 1998; Smit, 1999; Gu et al., 2000; Lander et al., 2001; Pavlicek et al., 2001). In agreement with these results, GC_3-poorer genes contained more LINEs in their introns compared to GC_3-richer genes. In apparent contrast, also Alu sequences were more frequent in the introns of GC_3-poorer genes than in GC_3-richer genes. This finding is, however, explained by the shorter average size of the latter introns and by the lesser difference in GC of clustered compared to dispersed genes (see Fig. 4b). The contrast between the behaviours of repeated sequences and CpG islands is interesting, the latter arising by a local GC increase, the former by insertion into target sequences.

Interestingly, the average GC level of 345, 190 and 89 AluS-, AluJ- and AluY-containing introns, respectively (their sum is >454, as some introns can contain Alus from different families), is 47.7%, 47.8% and 47.4%, indicating no variation in different Alus subfamilies.

Finally, it should be mentioned that other interspersed repeats (like MIRs) are known to be more frequent in GC-rich isochores (Matassi et al., 1998).

4. Conclusion

The general problem of the fate of duplicated genes has been investigated in our laboratory by a novel approach, namely by studying the compositional changes undergone by duplicated copies and the changes that took place in the introns of the latter. The majority of the gene duplications under consideration here are the result of whole ancient genome duplications, so that most copies are found on different human chromosomes, dispersed copies being due to the separation of duplicates because of chromosomal rearrangements. In contrast with such duplicated genes, recent tandem duplications are clustered.

This approach differs from the classical one, which relies on the loss or change in function of duplicated genes, in that it is based on purely structural features. From this viewpoint, the most significant results concern CpG islands. Indeed, the frequency of CpG islands is about twofold higher in GC-rich copies and in dispersed copies compared to GC-poor and clustered copies, respectively. This finding by itself suggests differences in the regulation of genes which have, in addition, been enriched in GC, with consequent changes in the amino acid composition and hydrophobicity of the encoded proteins (Jabbari et al., 2003), as well as divergent gene expression (Makova and Li, 2003; He and Zhang, 2005).

One should finally mention that Rodin and Parkhomchuk (2004) reported a GC asymmetry for gene duplicates in human, a result similar to the GC_3 divergence of human gene duplicates (Jabbari et al., 2003); the other observation made by these authors is the relative GC-poorness of single copy genes (that we could confirm; data not shown) reinforcing our assumption of a GC-poor ancestral state.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version at doi:10.1016/j.gene.2005.09.038.

References

Aïssani, B., Bernardi, G., 1991a. CpG islands, genes, isochores in the genome of vertebrates. Gene 106, 185–195.

Aïssani, B., Bernardi, G., 1991b. CpG islands: features and distribution in the genome of vertebrates. Gene 106, 173–183.

Batzer, M.A., Deininger, P.L., 2002. Alu repeats and human genomic diversity. Nat. Rev., Genet. 3, 370–379.

Bernardi, G., 2004. Structural and Evolutionary Genomics. Natural Selection in Genome Evolution. Elsevier, Amsterdam.

Betts, M.J., Guigo, R., Agarwal, P., Russell, R.B., 2001. Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? EMBO J. 20, 5354–5360.

Bird, A.P., Taggart, M., Frommer, M., Miller, O.J., Macleod, D., 1985. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell 40, 91–99.

Boudet, N., Aubourg, S., Toffano-Nioche, C., Kreis, M., Lecharny, A., 2001. Evolution of intron/exon structure of DEAD helicase family genes in *Arabidopsis*, *Caenorhabditis*, and *Drosophila*. Genome Res. 11, 2101–2114.

Carels, N., Bernardi, G., 2000. Two classes of genes in plants. Genetics 154, 1819–1825.

Chen, C., Gentles, A.J., Jurka, J., Karlin, S., 2002. Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. Proc. Natl. Acad. Sci. U. S. A. 99, 2930–2935.

Duret, L., Mouchiroud, D., Gouy, M., 1994. HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res. 22, 2360–2365.

Duret, L., Mouchiroud, D., Gautier, C., 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. 40, 308–317.

Federico, C., Scavo, C., Cantarella, C.D., Motta, S., Saccone, S., Bernardi, G., in press. Gene-rich and gene-poor chromosomal regions have different locations in the interphase nuclei of cold-blooded vertebrates. Chromosoma.

Gardiner-Garden, M., Frommer, M., 1987. CpG islands in vertebrate genomes. J. Mol. Biol. 196, 261–282.

Gu, Z., Wang, H., Nekrutenko, A., Li, W.H., 2000. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. Gene 259, 81–88.

He, X., Zhang, J., 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169, 1157–1164.

Jabbari, K., Bernardi, G., 1998. CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene 224, 123–127.

Jabbari, K., Rayko, E., Bernardi, G., 2003. The major shifts of human duplicated genes. Gene 317, 203–208.

Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Makova, K.D., Li, W.H., 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. 13, 1638–1645.

Matassi, G., Lauda, D., Bernardi, G., 1998. Distribution of the mammalian-wide interspersed repeats (MIRs) in the isochores of the human genome. FEBS Lett. 439, 63–65.

Pavlicek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J.V., Bernardi, G., 2001. Similar integration but different stability of Alus and LINEs in the human genome. Gene 276, 39–45.

Postlethwait, J., Amores, A., Cresko, W., Singer, A., Yan, Y.L., 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends Genet. 20, 481–490.

Rodin, S.N., Parkhomchuk, D.V., 2004. Position-associated GC asymmetry of gene duplicates. J. Mol. Evol. 59, 372–384.

Rogozin, I.B., Lyons-Weiler, J., Koonin, E.V., 2000. Intron sliding in conserved gene families. Trends Genet. 16, 430–432.

Rynditch, A.V., Zoubak, S., Tsyba, L., Tryapitsina-Guley, N., Bernardi, G., 1998. The regional integration of retroviral sequences into the mosaic genomes of mammals. Gene 222, 1–16.

Smit, A.F., 1996. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. 6, 743–748.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.

Soriano, P., Meunier-Rotival, M., Bernardi, G., 1983. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. Proc. Natl. Acad. Sci. U. S. A. 80, 1816–1820.

Stoltzfus, A., Logsdon Jr., J.M., Palmer, J.D., Doolittle, W.F., 1997. Intron "sliding" and the diversity of intron positions. Proc. Natl. Acad. Sci. U. S. A. 94, 10739–10744.

Tsyba, L., Rynditch, A.V., Boeri, E., Jabbari, K., Bernardi, G., 2004. Distribution of HIV-1 in the genomes of AIDS patients. Cell. Mol. Life Sci. 61, 721–726.

Yu, J., Yang, Z., Kibukawa, M., Paddock, M., Passey, D.A., Wong, G.K., 2002. Minimal introns are not "junk". Genome Res. 12, 1185–1189.

Zerial, M., Salinas, J., Filipski, J., Bernardi, G., 1986. Gene distribution and nucleotide sequence organization in the human genome. Eur. J. Biochem. 160, 479–485.