Available online at www.sciencedirect.com
SCIENCE DIRECT®

Gene 317 (2003) 203–208

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

The major shifts of human duplicated genes

Kamel Jabbari[a], Edda Rayko[a], Giorgio Bernardi[a,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 1 March 2003; received in revised form 2 April 2003; accepted 12 May 2003

Abstract

Since many gene duplications in the human genome are ancient duplications going back to the origin of vertebrates, the question may be asked about the fate of such duplicated genes at the compositional genome transitions that occurred between cold- and warm-blooded vertebrates. Indeed, at that transition, about half of the (GC-poor) genes of cold-blooded vertebrates (the genes of the gene-dense "ancestral genome core") underwent a GC enrichment to become the genes of the "genome core" of warm-blooded vertebrates. Since the compositional distribution of the human duplicated genes investigated (1111 pairs) mimics the general distribution of human genes (about 50% GC_3-poor and 50% GC_3-rich genes, the border being at 60% GC_3), we considered two possibilities, namely that the compositional transition affected either (i) about half of the copies on a random basis, or (ii) preferentially only one copy of the duplicated genes. The two possibilities could be distinguished if each copy is put into one of two subsets according to its GC_3 level. Indeed, in the first case, the two distributions would be similar, whereas in the second case, the two distributions would be different, one copy having maintained the ancestral GC-poor composition, and one copy having undergone the compositional change. Using this approach, we could show that, by far and large, one copy of the duplicated genes preferentially underwent the GC enrichment. This result implies that this copy, which had possibly acquired a different function and/or regulation, was preferentially translocated into the gene-dense compartment of the genome, the "ancestral genome core", namely the "gene space" which underwent the compositional transition at the emergence of warm-blooded vertebrates.

Keywords: Ancestral genome core; Human duplicated gene; Genome transition

1. Introduction

Investigations from our laboratory (see Bernardi, 2003, for a general review) have shown that: (1) the genomes of vertebrates comprise two "gene spaces", a small gene-rich space and a large gene-poor space; (2) the gene-rich space of the genome of cold-blooded vertebrates (the "ancestral genome core") is only slightly GC-richer compared to the gene-poor space, whereas this difference is very strong in warm-blooded vertebrates; indeed, (3) two independent compositional transitions affected the "ancestral genome core" of the cold-blooded vertebrate ancestors of mammals and birds, which comprises about half of the genes (and the associated non-coding regions, both intra- and inter-genic) leading to a considerable increase in GC (the molar ratio of guanine + cytosine in DNA) to form the "genome core" of the genomes of warm-blooded vertebrates (see Fig. 1a).

Recent results indicate that at least one round of genome duplication occurred in early chordates (650–350 Mya), 13% of human genes still being recognizable as duplicates (McLysaght et al., 2002). Along the same line, the phylogenetic analysis of 749 vertebrate gene families (Gu et al., 2002) led to the identification of a pattern characterized by two waves (I, II) and an ancient component (900–750 Mya). Wave I (430–80 Mya) represents a recent gene family expansion by tandem or segmental duplications, whereas wave II (750–430 Mya), a rapid paralogous gene increase in the early stage of vertebrate evolution, supports the idea of genome duplication(s). Further analyses indicated that large- and small-scale gene duplications both made a significant contribution during the early stage of vertebrate evolution.

It has been argued (Friedman and Hughes, 2001; see also Smith et al., 1999) that a 'slow shuffle' (individual gene duplications followed by transpositions to form "paralogons") is a more parsimonious explanation of the current

Abbreviations: CDS, coding sequences; GC, molar fraction of guanine and cytosine; GC_3, GC in third codon positions of protein-coding genes; Mya, million(s) years ago.

* Corresponding author. Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121, Naples, Italy. Tel.: +39-81-583-3300; fax: +39-81-245-5807.
E-mail address: bernardi@szn.it (G. Bernardi).

0378-1119/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0378-1119(03)00704-2

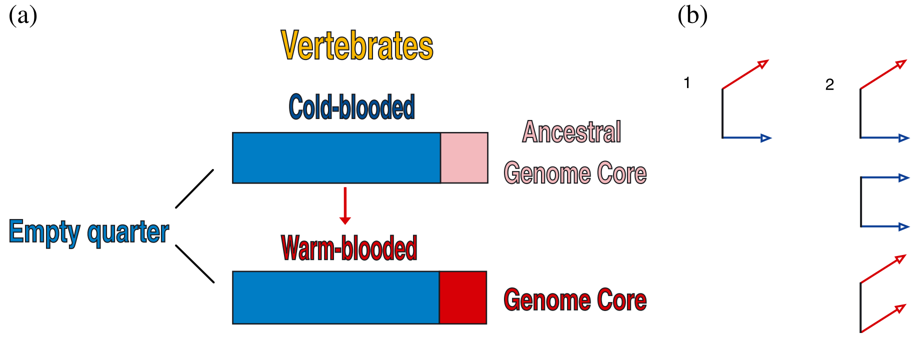

Fig. 1. (a) Scheme of the compositional transition from cold- to warm-blooded vertebrates. The “empty quarter” (genome desert) of their genomes is GC-poor and gene-poor (blue box) and essentially did not undergo any compositional change. The gene-dense, moderately GC-rich “ancestral genome core” (pink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

structure of the human genome than is a ‘big bang’ (duplication of the whole genome or of substantial sections of it). McLysaght et al. (2002) noticed, however, that the parsimony test (Gu and Huang, 2002; Meyer and Schartl, 1999; Wolfe, 2001) will, regardless of which model is correct, always favor the slow shuffle whenever the density of duplicated genes in a genome (or paralogon) is below 50%, as is the case in the well-documented paleopolyploids.

Accordingly, assuming that many gene duplications in the human genome are ancient duplications going back to the origin of vertebrates (Ohno, 1970; see also Wolfe, 2001 and Prince and Pickett, 2002, for recent reviews), a hypothesis which we could demonstrate to be correct, the question may be asked about the fate of duplicated genes at the compositional transitions discussed above.

Since the compositional distribution (GC% at third codons positions) of all human duplicated genes investigated here is similar to the compositional distribution of human genes, we considered two possibilities, namely, that the compositional transition either (i) preferentially affected one copy of the duplicated genes, or (ii) about half of both copies. More precisely, the possible events in the second case are presented in the scheme of Fig. 1b, which indicates that the genes in which one copy preferentially underwent the compositional transition were accompanied by the pairs of genes which either maintained or changed their composition.

The first case would indicate that one copy, which had possibly acquired a different function and/or regulation through subfunctionalization (Force et al., 1999; Avaron et al., 2003) or function partitioning (Wagner, 2000) or other mechanisms, was selected for the compositional change, the second copy keeping the ancestral GC-poor composition. In contrast, the second case would indicate no such preference, but rather a randomness in the genes that underwent the compositional transition.

The present work, concerning duplicated human genes, indicates that, indeed, one copy preferentially underwent the compositional transition, implying a selection between the two copies and possibly a different structure/function/regulation of one copy relative to the other one. Some additional implications will be discussed in the conclusions.

2. Materials and methods

2.1. Data sets of duplicated genes

Human gene families were retrieved, using the ACNUC retrieval system (Gouy et al., 1985), from HOVERGEN (Duret et al., 1994) release April 19, 2000 (an old release having the advantage of not being contaminated by putative genes). In this database, genes are clustered in families according to their homology level. Pairs of coding sequences (CDS) with similar sizes (the difference in sizes being arbitrarily set at $\leq 7\%$) were collected from each family. When explicitly defined, partial CDS, isoforms and alternatively spliced genes, as well as pseudogenes, were excluded from the data set. Also, the chromosomal location and the related literature were checked whenever possible.

This procedure produced a data set of 1111 pairs of duplicated genes. These genes represented 465 families and included 1589 individual human CDS. Two hundred

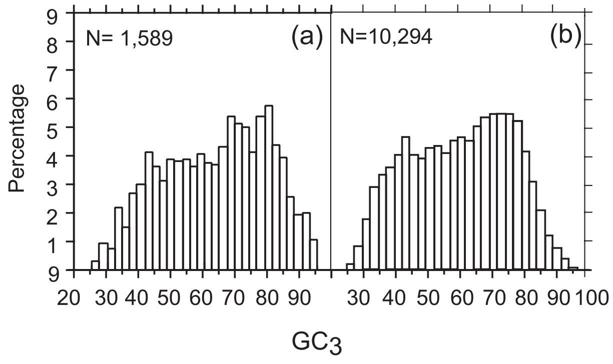

Fig. 2. Distribution of GC_3 values (a) of the complete data set of duplicated genes used in this work (b) of 10,294 human genes from the RefSeq database (Pruitt and Maglott, 2001).

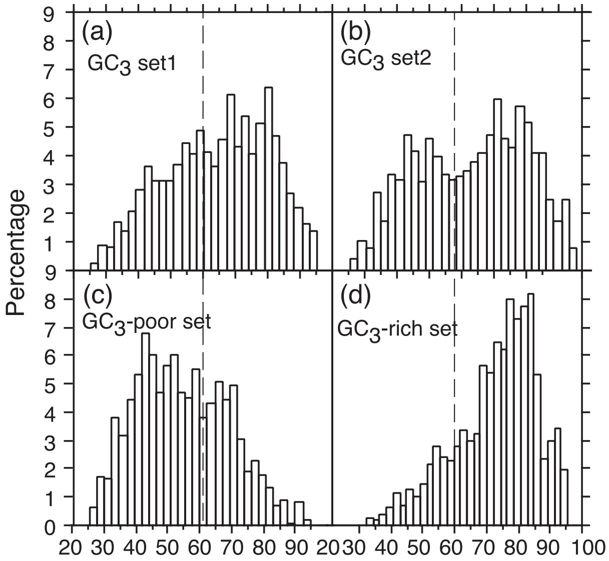

Fig. 3. Distribution of third codon positions (GC_3%) of (a) set 1 containing the longer counterparts, (b) set 2 containing the shorter counterparts of 1111 CDS pairs, (c) the GC_3-poorest set, and (d) the GC_3-richest set of 1111 CDS pairs.

and forty of these families contained two members, the remaining two hundred and twenty-five families being made up of more than two members; in the latter case, all possible comparisons, with CDS having similar size, were considered.

CDS size was used as an arbitrary criterion for arranging paralogous CDS, the longer counterpart being put in set 1 and the shorter in set 2. For the 147 pairs (13%) that had identical sizes, they were assigned randomly to one of the two sets. Average differences in size were small, 530 vs. 520 codons, with however, large standard deviations, 475 and 466, respectively. It is important to remark that, while the size constraint leads to losing members of the same family with a size difference >7%, such constraint favors extraction of bona fide gene family members.

2.2. Sequence analysis

Amino acid sequences were aligned using ClustalW (version 1.81, Thompson et al., 1994) and back-translated into their corresponding nucleic acid alignments. To calculate the number of synonymous vs. non-synonymous substitutions, the method of Nei and Gojobori (1986) was used. If the synonymous distances had a value ≥ 0.75, alignments were excluded from the analysis, because saturation had been reached. Distance estimation by maximum likelihood (Yang, 1997) was also used to confirm the results on the correlation of synonymous vs. non-synonymous distances.

For each gene pair compared, base frequency differences at third codon positions (ΔGC_3) and frequency differences for each amino acid (Δaa) were calculated. Hydropathy values of proteins were calculated according to Kyte and Doolittle (1982), as in Lobry and Gautier (1994).

3. Results

3.1. Compositional distribution of the data set

Fig. 2a shows the distribution of GC_3 values of the complete data set of duplicated genes used in this work. The distribution is very similar to that of all genes extracted from the RefSeq database (Pruitt and Maglott, 2001), which include all "curated" genes that had been confirmed, as well as "provisional" genes checked only once (Fig. 2b).

When the data set was split into two sets corresponding to each of the two gene copies present according to an arbitrary criterion, such as CDS size, the GC_3 histograms of the two sets are rather similar and show only small differences with that of the GC_3 distribution of all human genes (compare Fig. 3a and b with Fig. 2a).

In contrast, when the data set is split into two sets according to GC_3 levels, the GC_3 distributions of the two sets are almost mirror images of each other (Fig. 3c and d).

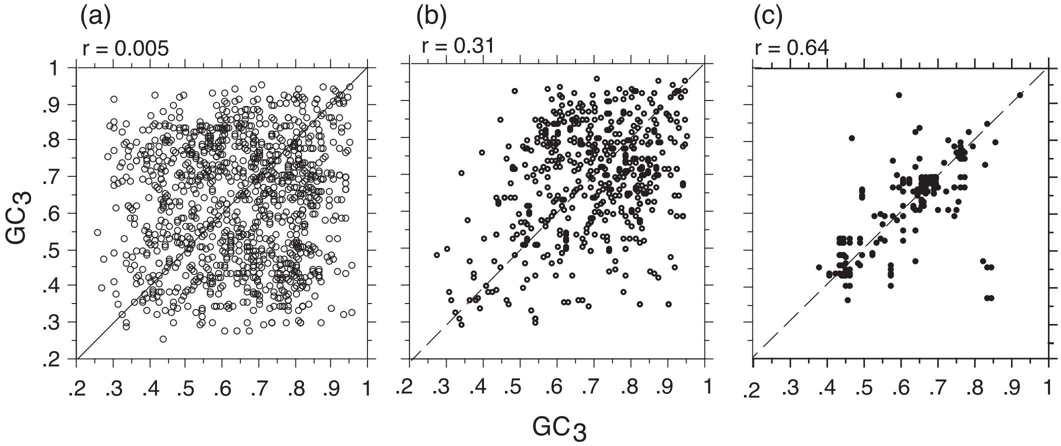

Fig. 4. Plot of GC_3 to GC_3 correlations (a) in the 1111 duplicated genes, (b) in duplicated CDS after exclusion of pairs with proportion of synonymous changes ≥ 0.75, (c) in tandemly duplicated genes.

3.2. Compositional distributions and correlations

As shown in Fig. 4a, no correlation was found between base frequencies at the third codon position of the two sets as obtained according to an arbitrary criterion such as size, whereas these correlations are highly significant between pairs of orthologous genes of species as distant as human and *Xenopus* (Bernardi et al., 1997; Cruveiller et al., 1999; Bernardi, 2000). After excluding sequences with a proportion of synonymous changes ≥ 0.75, a significant correlation ($r = 0.31$; $p < 0.001$) was, however, observed (Fig. 4b) and this correlation improved when analyzing the GC_3-set 1 vs. GC_3-set 2 plot of gene copies colocalized on the same chromosomes (Fig. 4c). Pairs showing a high proportion of synonymous changes (≥ 0.75), represented 54% of the total sample (1111 pairs), indicating that a substantial amount of gene duplicates had diverged extensively, as expected.

As far as the chromosomal distribution of family members is concerned, among the gene family members for which the location is known (715 gene pairs belonging to 344 families), only 5.9% (42 pairs belonging to 20 families) were found to be located on the same chromosome. Since 14 gene families had some members which colocalize and others which are dispersed on different chromosomes, the percentage of gene families with colocalized members is 9.5% (34 families). This indicates that the majority of gene families in the human genome are not only highly divergent but are also not located on the same chromosome.

Two additional observations made on duplicated genes concern: (i) the existence of a correlation ($r = 0.37$, $p < 0.0001$) between synonymous and non-synonymous distances (Fig. 5), a result which is confirmed when maximum likelihood distance is used ($r = 0.44$, $p < 0.0001$). This is in agreement with our previous proposal that synonymous and non-synonymous changes in mammals are under common constraints (Mouchiroud et al., 1995; Alvarez-Valin et al., 1998). The fact that the ratio dn/ds is less than 1 for almost all analyzed pairs, may indicate overdominance of negative (purifying) selection in the evolution of human gene families (see also Lynch and Conery, 2000; Kondrashov et al., 2002, for similar results). (ii) The increase in hydrophobicity of the proteins encoded by the duplicated genes whose GC_3 was increased at the compositional transition between cold- and warm-blooded vertebrates. When the set of 1111 gene pairs was sorted according to ΔGC_3 values and divided into five subsets of equal size ($n = 222$), and the ΔHb (hydrophobicity difference) was calculated for each subset, the results show that ΔHb values increase with increasing ΔGC_3 values (Fig. 6), the change being statistically very significant ($p < 0.0001$, Kruskal–Wallis non-parametric test). These results are in agreement with our previous observation that ΔGC_3 and ΔHb are correlated in a comparison *Xenopus* vs. human (Cruveiller et al., 1999) and with the fact that the GC_3 level of human coding sequences is correlated positively with protein hydropathy index (Jabbari et al., 2003).

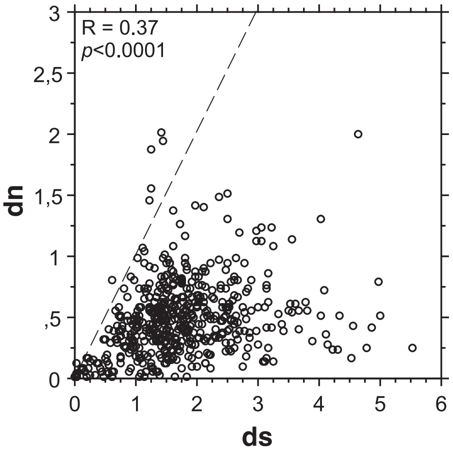

Fig. 5. Plot of synonymous vs. non-synonymous distances of human duplicated pairs. The correlation coefficient is given. The dashed line corresponds to the diagonal.

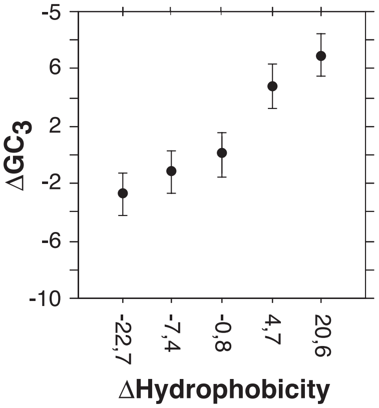

Fig. 6. Plots of average difference hydrophobicity (ΔHB) against average ΔGC_3, the data set were sorted according to ΔGC_3 and averaged in five classes of equal size.

4. Discussion and conclusions

The results obtained in these investigations allow us to draw some conclusions about the "major line" of events following the ancestral gene duplications (the "major line" disregards a number of phenomena, like recent duplications, because they do not blur the events concerning the majority of duplicated genes). This "major line" can be described as a succession of three steps which occurred in duplicated genes during the evolution of the vertebrate genome (see Fig. 7).

(i) Most of the original gene duplications of interest here seem to have occurred in the fish genomes or earlier in evolution. This is indicated by the compositional divergence which is so strong as not to allow to recognize any correlation between the copies of a given gene. Indeed, the concentration of points on or around the diagonal found when looking at GC_3 plots shown by the minority of genes

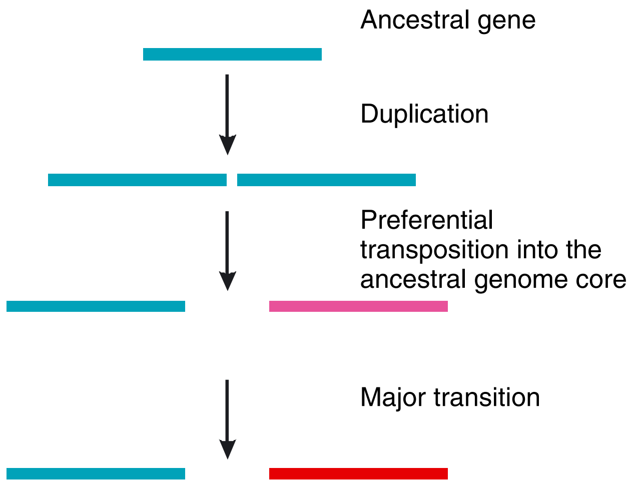

Fig. 7. A scheme of the most frequent pathway following ancient gene duplication (blue bars). One copy is supposed to be preferentially transposed into the ancestral genome core (pink bars), which then undergoes the major compositional transition (red bars).

which are localized on the same chromosomes (Fig. 4c) indicates that genes arising from recent duplications show very little compositional divergence. Moreover, this correlation can be observed between orthologous genes from human and *Xenopus* (Cruveiller et al., 1999) or human and zebrafish (Belle et al., 2002).

(ii) While in the original duplication events, genes were tandemly arranged, as suggested by many examples of recent duplications, a transposition of one copy is indicated by the fact that the majority of duplicated genes are found on different human chromosomes.

It is interesting to notice that segmental duplications are enriched within pericentromeric and subtelomeric regions in the human genome (Amann et al., 1996; Trask et al., 1998; Eichler et al., 1999; Horvath et al., 2000), this bias has been quantitatively tested recently in the working draft of the human reference sequence (Bailey et al., 2001), and enrichment levels were 4.7- to 11.8-fold; this bias appears to be more pronounced for interchromosomal than intrachromosomal duplications (for a review see Samonte and Eichler, 2002). Moreover, pericentromeric regions exhibited preferential duplication compared to subtelomeric regions.

(iii) In the majority of cases, only one copy underwent the compositional change. Indeed, if a majority of both genes from gene pairs had either undergone the compositional transition, or maintained the original composition, there would be little compositional difference between the two genes of each pair, and Fig. 3c and d would not be mirror images of each other. We know that compositional genome changes took place in the "ancestral genome core" of the ancestors of warm-blooded vertebrates. The compositional change of one copy occurred, therefore, in the copy located in the "ancestral genome core", whereas the other copy, maintaining the original low GC level remained in the "empty quarter", which did not undergo any compositional change.

To sum up, three steps could be identified in the major line of evolution of duplicated genes: (i) the original tandem duplications; (ii) the transposition of one copy of the duplicated genes into the "ancestral genome core"; and (iii) the compositional change of this copy.

These steps may suggest (i) that duplications occurred most frequently in the "ancestral empty quarter", possibly in agreement with the preferential duplications in GC-poor pericentromeric regions (see Saccone et al., 2002); (ii) that the duplicated copy acquired, in all likelihood, a new function; and (iii) that transposition of the duplicated copy into the open chromatin of the "ancestral genome core" was generally preferred, as in the case of retroviral integrations (see Tsyba et al., 2003); interestingly, these latter events led to a further gene enrichment in the "ancestral genome core". Expectedly, the increased hydrophobicity of the proteins encoded by genes located in the genome core were accompanied by a shortening of introns and by a preferential formation of CpG islands which will be described elsewhere (Rayko, Jabbari and Bernardi, in preparation).

A final point is that our results have no bearing on the controversy surrounding the classical two-round hypothesis of vertebrate genome duplications predating the origin of fishes (Ohno, 1970) nor on the debate between large-scale or small-scale duplication (see McLysaght et al., 2002; Gu et al., 2002).

References

Alvarez-Valin, F., Jabbari, K., Bernardi, G., 1998. Synonymous and non-synonymous substitutions in mammalian genes: intragenic correlations. J. Mol. Evol. 46, 37–44.

Amann, J., Valentine, M., Kidd, V.J., Lahti, J.M., 1996. Localization of chi1-related helicase genes to human chromosome regions 12p11 and 12p13: similarity between parts of these genes and conserved human telomeric-associated DNA. Genomics 32, 260–265.

Avaron, F., Thaeron-Antono, C., Beck, C.W., Borday-Birraux, V., Geraudie, J., Casane, D., Laurenti, P., 2003. Comparison of even-skipped related gene expression pattern in vertebrates shows an association between expression domain loss and modification of selective constraints on sequences. Evolut. Develop. 5, 145–156.

Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., Eichler, E.E., 2001. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 11, 1005–1017.

Belle, E.M., Smith, N., Eyre-Walker, A., 2002. Analysis of the phylogenetic distribution of isochores in vertebrates and a test of the thermal stability hypothesis. J. Mol. Evol. 55, 356–363.

Bernardi, G., 2000. The compositional evolution of vertebrates. Gene 259, 31–43.

Bernardi, G., 2003. Structural and Evolutionary Genomics. Natural Selection in Genome Evolution. Elsevier, Amsterdam. In press.

Bernardi, G., Hughes, S., Mouchiroud, D., 1997. The major compositional transitions in the vertebrate genome. J. Mol. Evol. 44, S44–S51.

Cruveiller, S., Jabbari, K., D'Onofrio, G., Bernardi, G., 1999. Different hydrophobicities of orthologous proteins from *Xenopus* and human. Gene 238, 15–21.

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

Eichler, E.E., Archidiacono, N., Rocchi, M., 1999. CAGGG repeats and the pericentromeric duplication of the hominoid genome. Genome Res. 9, 1048–1058.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., Postlethwait, J., 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151 (4), 1531–1545.

Friedman, R., Hughes, A.L., 2001. Pattern and timing of gene duplication in animal genomes. Genome Res. 11, 1842–1847.

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., di Paola, G., 1985. ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput. Appl. Biosci. 1, 167–172.

Gu, X., Huang, W., 2002. Testing the parsimony test of genome duplications: a counterexample. Genome Res. 12, 1–2.

Gu, X., Wang, Y., Gu, J., 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat. Genet. 31 (2), 205–209.

Horvath, J., Viggiano, L., Loftus, B., Adams, M., Rocchi, M., Eichler, E., 2000. Molecular structure and evolution of an alpha/non-alpha satellite junction at 16p11. Hum. Mol. Genet. 9, 113–123.

Jabbari, K., Cruveiller, S., Clay, O., Bernardi, G., 2003. The correlation between GC_3 and hydropathy in human genes. Gene 317, 137–140 (this issue).

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., Koonin, E.V., 2002. Selection in the evolution of gene duplications. Genome Biol. 3, 0008.1–0008.9.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157 (1), 105–132.

Lobry, J.R., Gautier, C., 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. 22, 3174–3180.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155.

McLysaght, A., Hokamp, K., Wolfe, K.H., 2002. Extensive genomic duplication during early chordate evolution. Nat. Genet. 31 (2), 200–204.

Meyer, A., Schartl, M., 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Curr. Opin. Cell Biol. 11, 699–704.

Mouchiroud, D., Gautier, C., Bernardi, G., 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J. Mol. Evol. 40, 107–113.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 5, 418–426.

Ohno, S., 1970. Evolution by Gene Duplication. Springer-Verlag, Berlin.

Prince, V.E., Pickett, F.B., 2002. Splitting pairs: the diverging fates of duplicated genes. Nat. Rev., Genet. 11, 827–837.

Pruitt, K.D., Maglott, D.R., 2001. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 29, 137–140.

Saccone, S., Federico, C., Bernardi, G., 2002. Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. Gene 300, 169–178.

Samonte, R.V., Eichler, E.E., 2002. Segmental duplications and the evolution of the primate genome. Nat. Rev., Genet. 3, 65–72.

Smith, N.G.C., Knight, R., Hurst, L.D., 1999. Vertebrate genome evolution: a slow shuffle or a big bang? BioEssays 21, 697–703.

Thompson, J.D., Higgins, D.G., 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.

Trask, B.J., Massa, H., Brand-Arpon, V., Chan, K., Friedman, C., Nguyen O.T., Eichler, E.E., van den Engh, G., Rouquier, S., Shizuya, H., et al., 1998. Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. Hum. Mol. Genet. 7, 2007–2020.

Tsyba, L., Rynditch, A., Boeri, E., Jabbari, K., Bernardi, G., 2003. Genomic distribution of HIV-1 in genomes of infected individuals: correlation between localization in GC-poor isochores and high viremia are correlated. FEBS Lett. (submitted for publication).

Wagner, A., 2000. The role of population size, pleiotropy and fitness effects of mutations in the evolution of overlapping gene functions. Genetics 154 (3), 1389–1401.

Wolfe, K.H., 2001. Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. 2, 333–341.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.