

Position-Associated GC Asymmetry of Gene Duplicates

Sergei N. Rodin, Dmitri V. Parkhomchuk

Department of Theoretical Biology, Beckman Research Institute of the City of Hope, 1500 East Duarte Road, Duarte, CA 91010, USA

Received: 13 November 2003 / Accepted: 24 March 2004 [Reviewing Editor: Dr. Manyuan Long]

Abstract. It is well known that repositioning of a gene often exerts a strong impact on its own expression and whole development. Here we report the results of genome-wide analyses suggesting that repositioning may also radically change the evolutionary fate of gene duplicates. As an indicator of these changes, we used the GC content of gene pairs which originated by duplication. This indicator turned out to be duplicate-asymmetric, which means that genes in a pair differ significantly in GC content despite their apparent origin from a common ancestor. Such an asymmetry necessarily implies that after duplication two originally identical genes mutated in opposite directions—toward GC-rich and GC-poor content, respectively. In mammalian genomes, this trend is definitely associated with presumably methylated hypermutable CpG sites, and in a typical GC-asymmetric gene pair, its two member genes are embedded in GC-contrasting isochores. However, we unexpectedly found similar significant GC asymmetry in fish, fly, worm, and yeast. This means that neither methylation alone nor methylation in combination with isochores can be counted as a primary cause of the GC asymmetry; rather they represent specific realizations of some universal principle of genome evolution. Remarkably, genes from pairs with the greatest GC asymmetry tend to be on different chromosomes, suggesting that the mutational difference between gene duplicates is associated with translocation of a new gene to a different place in the genome, whereas GC symmetric pairs demonstrate the opposite tendency. A recently emerged extra gene

copy is usually on the same chromosome as is its parent but quickly, by 0.05 substitution per synonymous site, either has perished or occupies a different chromosome. During this earliest posttranslocation period, the ratio of nonsynonymous/synonymous base substitutions is unusually high, suggesting a rapid adaptive evolution of novel functions. In a general context of evolution by gene duplication, our interpretation of this position-dependent GC asymmetry between duplicated genes is that evolution of redundant genes toward a new function has often been associated with their very early, postduplication repositioning in the genome, with a concomitant abrupt change in epigenetic control of tissue/stage-specific expression and an increase in the mutation rate. Of eight eukaryotic genomes studied, the most distinguished in this respect is the human genome.

Key words: Comparative genomics — Epigenetic regulation — Position effects — Isochores — GC content — Gene duplication

Introduction

The evolutionary significance of DNA duplication as a major source of new genes was recognized long before modern genome-wide studies (Haldane 1932; Muller 1935; Serebrovsky 1938; Ohno 1970). The problem of duplicate gene retention was also first addressed long ago (Haldane 1933). Indeed, since deleterious mutations occur far more frequently than advantageous ones, a new extra gene copy has a much higher chance of degrading into a functionless

pseudogene than gaining a new function or gradual divergence toward a tissue/stage-specific variant of an old function (Ohno 1972). In contrast to a single-copy gene, in which the natural selection easily senses deleterious mutations and eliminates them from a population, selection cannot act on recently duplicated genes if they have the same expression pattern as the original gene. Under the shelter of gene redundancy, any deleterious mutation is actually neutral, and therefore, instead of being eliminated by selection, it may become fixed by random drift. The longer natural selection remains relaxed for extra gene copies, the more likely are they to be pseudogenized. In order to escape pseudogenization, one of the new structurally indistinguishable duplicates must break its identity in the expression pattern and come back under the surveillance of selection, as soon as possible. Multigene families show us how this might come to pass.

Multigene families are the product of the functional divergence of gene duplicates. However, in a typical present-day family, the member genes do not face the “loss-or-gain” dilemma because usually each of them has a particular developmental period and/or tissue of expression during which its evolutionarily old and young relatives are inactivated. In fact, these gene-specific patterns of inactivation complement each other, comprising *en toto* the integral expression pattern of the family. This makes all gene members unique and hence visible for negative selection, despite their origin as formerly identical duplicates. Obviously, for each pair of homologous genes in a family, such a stage- and tissue-specific mutually complementary inactivation event had to have occurred once in the evolutionary past.

In principle, two major mechanisms are conceivable for such complementation: mutational (Force et al. 1999; Lynch and Force 2000; Lynch et al. 2001) and epigenetic (Rodin and Riggs 2003). Mutational complementation can be provided by degenerative mutations in different, relatively independent regulatory elements responsible for stage/tissue-specific expression. The originality of this duplication–degeneration–complementation (DDC) model is that it involves deleterious mutations and yet protects duplicates from degradation. Somatic epigenetic complementation (EC) could play the same antidegradation role under the assumption that newly produced structurally identical genes might not be identical with respect to the epigenetic regulation of their expression (Rodin and Riggs 2003). Complementary functional inactivation of duplicates can be provided by methylation (Rodin and Riggs 2003), homologous RNAi-mediated silencing (Carmichael 2003), and other processes involving heritable chromatin structure (Jenuwein and Allis 2001). A simplified EC model for two genes is shown in Fig. 1. In

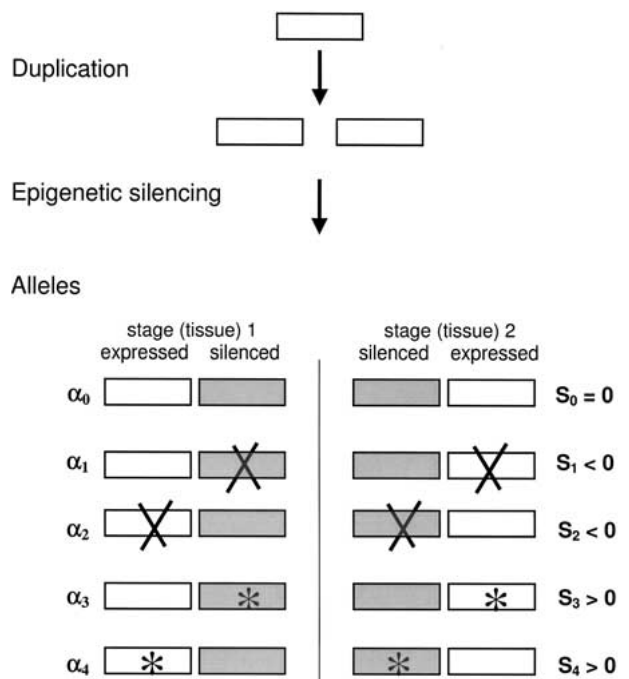


Fig. 1. Model of epigenetic protection of new gene duplicates from mutational degradation to pseudogenes. For simplicity, the case of only two tissue/stage-specific gene duplicates is shown. These two are shown functionally silenced (shaded boxes) in the tissue (stage)-complementary manner. X denotes a degenerative mutation; *, advantageous mutation. Selective values of the corresponding mutant alleles are indicated relative to the wild-type allele α_0 ($s = 0$): degenerative ($s < 0$), and advantageous ($s > 0$), respectively. In spite of redundancy, the tissue/stage-complementary inactivation of duplicates causes a degenerative mutation ($s < 0$) in alleles α_1 and α_2 exposable to negative selection. The latter recognizes and eliminates both from the population, thus making way for advantageous mutations ($s > 0$) in alleles α_3 and α_4 (Fig. 1). The positive selection spreads such adaptive alleles in the population, thus driving neofunctionalization of an extra gene copy. Otherwise, without epigenetic inactivation, the degenerative mutant alleles such as α_1 and α_2 are actually neutral (each has one functional gene copy) and can be fixed by random genetic drift. The detailed analysis of the model is described in Rodin and Riggs (2003).

fact, epigenetic complementation combines the evolutionary benefits of two states of a gene: in the single state, selection recognizes and eliminates frequent deleterious mutations, while the duplicated state allows selection to pick up and spread rare advantageous mutation, thus driving evolution of a new gene without losing an old one (Rodin and Riggs 2003).

Increasing data indicate that gene expression strongly depends on the local chromatin environment formed by the same or even different chromosomes within the nucleus (Cockell and Gasser 1999; Brown et al. 2001). The EC model actually suggests that the duplication event itself may change this environment for duplicates by bringing one of them farther from or closer to tissue- and stage-specific regulatory sites. Such position effects are expected to be much more common for translocated than for tandem duplicates. Indeed, in eukaryotic genomes, the control of gene transcription includes not only elements adjacent to

the transcribed part of a gene, but also very distant regulatory elements (Levine and Tjian 2003). Therefore, the greater the distance of repositioning of an extra gene duplicate, the more likely it will cease to be controlled by the former “parental” regulatory elements and come under the command of different regulatory elements. In short, repositioning of a gene duplicate should more often than not change its stage and tissue-specific expression and, consequently, aid selection by keeping this duplicate intact while “waiting” for rare new function-prospective mutations to arise (Rodin and Riggs 2003).

The DDC-based preservation of gene duplicates definitely does not require position changes, whereas such changes might be a signature of the epigenetic complementation model (Rodin and Riggs 2003). In this paper we directly address this issue, using as a criterion the GC content of gene duplicates. The rationale behind this approach is that genomes of vertebrates consist of isochores, usually rather long (>300-kb) segments of DNA differing in GC content (Bernardi et al. 1985; Bernardi 2000). A gene duplicate introduced into a different isochore evolves toward the GC content of its new residence. The EC model predicts that if early repositioning of an extra gene copy does epigenetically alter its expression, renders preservation, and allows subsequent functional divergence, then pairs of homologous genes should differ, quite often and significantly, in GC content. The genome-wide analysis of gene pairs described below confirms this prediction.

Materials and Methods

The sequence data for protein-coding genes were retrieved from proteomes of human (*Homo sapiens*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), fish (*Fugu rubripes*), worm (*Caenorhabditis elegans*), fly (*Drosophila melanogaster*), plant (*Arabidopsis thaliana*), and yeast (*Saccharomyces cerevisiae*). In this study we totally excluded intronless retrosequences since the majority of them are processed nonfunctional pseudogenes (Li 1997). All the data are available at the NCBI Web site. GenBank annotation files contain chromosome positions of the genes used in the study.

The gene duplicates were identified at the amino acid level by matching each gene with all others in a proteome and by clustering similar genes using the Blastclust program available at NCBI. A gene of inquiry was considered to belong in a given cluster of duplicates if its alignment with at least one other gene already included in the cluster spans no less than 80% of its length and has more than 60% amino acid identity. These homology criterion numbers represent not too divergent gene duplicates and a sufficiently long similarity region in order to make the analysis of nucleotide alignments sensible. In some analyses (Fig. 3), we partitioned duplicates further, into “young” and “old” duplicate groups. The young group included duplicates having more than 95% identical amino acids, whereas the similarity of more diverged duplicates in the old group was in the range of between 60 and 95 identical amino acids. The genes having no detectable similarity to each other formed the “single” gene group.

After clustering duplicates, we formed the sets of gene pairs for analyses. In the case of clusters consisting of only two genes, it was

straightforward. However, for clusters of larger sizes there is a problem of overrepresentation since different pairs from one cluster may represent the same duplication event, thus being not independent (Fig. 2). Of course, one can limit consideration to clusters of size two, thus ignoring the bias, but this certainly results in loss of information and might worsen the overall statistics. Instead, to minimize the bias, we met halfway and used the clusters that did not contain more than five genes. The resulting sample represented 80% of paralogous duplicates, i.e., had a very good coverage of duplication events (see, inset Fig. 2). Besides, as illustrated in the auxiliary scheme (Fig. 2), we formed two sets of gene pairs, cumulative and serial. The cumulative set is simply a set of all possible gene pairs; it is suitable for analyses of the general GC distribution (Fig. 3), isochore affiliation (Fig. 4), and mutational asymmetry of gene duplicates (Fig. 6), i.e., where it makes no difference whether gene pairs or duplication events are taken into account. In analyses where the combinatorial excess of gene pairs within large clusters might distort the real picture (Figs. 7 and 8) we preferred to keep track of the actual number of duplications and therefore used serial pairing (Fig. 2). The program for constructing a serial set, first, randomly chooses a gene as an outgroup in a cluster and, second, finds all its pairs, then excludes this gene with its pairs from further iterations, from the rest of the genes randomly chooses the next outgroup gene, finds its pairs, then excludes them, and so on—the cycle is iterated until the very last gene pair is found. In the resulting serial set, the number of gene pairs will be equal to that of duplication events minus one.

The human genome has a pronounced isochore structure so that it is very heterogeneous in GC content compared with other genomes. Alternatively, the nematode genome is so GC homogeneous that no distinct isochores *per se* can be detected. It seemed rather tempting to compare these two genomes with respect to the GC content within the upstream 5' and downstream 3' flanking regions of the corresponding gene duplicates. Accordingly, in order to retrieve data on quite lengthy flanking sequences, we used the whole-genome assemblies (from NCBI) of the human and nematode genomes.

As a preliminary rough approach to the estimation of selection pressure, we used the ratio R/S, where R is the number of mutations resulting in amino acid replacement and S is the number of synonymous mutations, both calculated per base substitution site. $R/S > 1$ indicates positive selection, which favors nonsilent mutations during evolution of new functions; $R/S < 1$ reflects the pressure of negative selection, which guards old functions; and $R/S = 1$ corresponds to neutral evolution with relaxed selection. In practice, synonymous vs. nonsynonymous substitution analysis was performed as follows. Since it is critical to preserve codon-to-codon alignment for two nucleotide sequences, first we aligned two corresponding amino acid sequences using the BestFit utility from the Wisconsin Package. It uses the local homology algorithm of Smith and Waterman (1981). Based on that amino acid alignment, we created the nucleotide sequences alignment by our Perl script. This nucleotide alignment was then fed to the Diverge utility from the Wisconsin Package, which estimated the pairwise number of synonymous and nonsynonymous substitutions per site based on the method described by Li et al. (1985).

It should be noted that, naturally, we experimented with different criteria of clustering duplicates as well as sampling and splitting them into young and old pairs. The major phenomenon described below—repositioning-associated GC asymmetry of duplicates—along with some other trends turned out to be surprisingly robust to the criteria.

Results

Our main finding described below is that in very different organisms (from yeast to human) a pair of

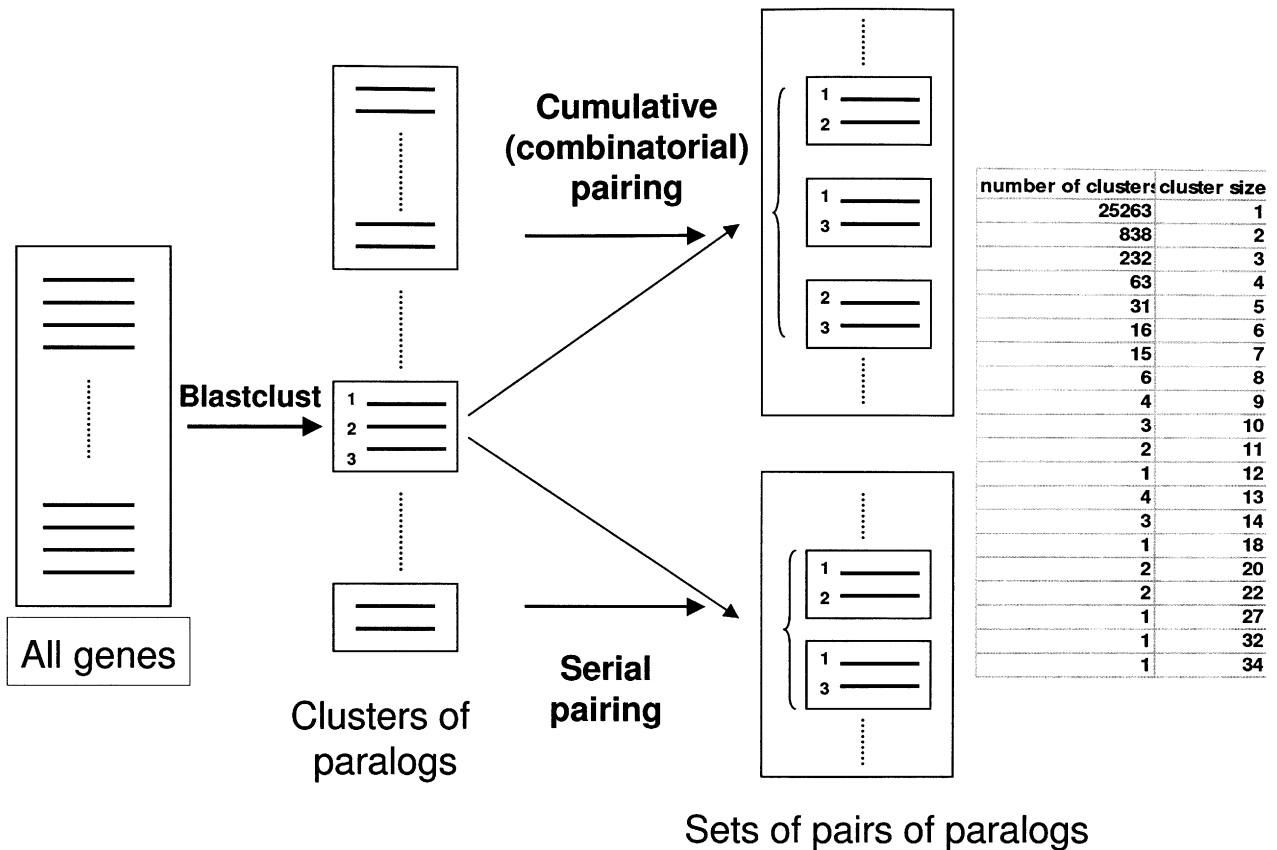


Fig. 2. The scheme illustrating the difference between cumulative and serial sets of paralogous gene pairs. Shown on the right is the distribution of size of gene clusters in the human genome.

homologous genes once produced by duplication unexpectedly often shows a great difference in GC content (thus suggesting a significant gene-biased mutational pressure), and most strikingly, this difference between former twin genes is apparently associated with their repositioning in the genome. Accordingly, we call this phenomenon the repositioning produced asymmetry of gene duplicates in GC content or, for brevity, the GC asymmetry.

The General GC-Asymmetric Pattern

A typical pair of aligned homologous genes (α -1 and γ -2 human actin genes) with a strong bias in the GC content is shown in Fig. 3A. All seven nucleotide differences detected within the short fragment appeared to occur at the mostly degenerated third codon position (Fig. 3A), where selection is supposed to have been far more relaxed than at the first and second positions. The whole-genome GC3 frequency distributions of single (nonduplicated) and duplicated human genes in the GC content at the third codon position are shown in Fig. 3B. As one would expect, single genes and very young gene duplicates (>95% amino acid identity) exhibit quite similar distribu-

tions. More diverged duplicates and older duplicates gain a new conspicuous feature: their GC3 distribution is bimodal (Fig. 3B), and furthermore, the corresponding two peaks become more distinct with duplicates aging (Fig. 3B). This pattern does not depend on the criterion of dividing genes in young and old pairs. Shown in Fig. 3 are simply consecutive illustrative points of the GC content dynamic during the evolution from single genes through young duplicates to older ones. Three other GC distributions, the GC1 and GC2 (first and second codon positions) as well as the GCi (intronic) have the same, although less pronounced, trend toward the progressing-with-time bimodality (not shown).

Since this trend is also typical for rat and mouse duplicates, we supposed that it might be associated with the large-scale isochore structure of the mammalian genome. To test this hypothesis, we analyzed the GC level of isochores in which GC-asymmetric genes are placed. Figure 4 illustrates a typical situation for the human genome: (1) the GC-rich member of a gene pair is embedded in the GC-rich isochore, whereas (2) its GC-poor counterpart is embedded in the GC-poor isochore, and (3) as in single genes (Aota and Ikemura 1986), the coding region of duplicates has a GC content exceeding that in the

A

	Tyr	Glu	Leu	Pro	Asp	Gly	Gln	Val	Ile	Thr	Ile	Gly	asn	Asp	Arg	Phe	Arg	Cys	Pro	Glu	Thr
alpha1	...TAC	GAG	CTG	CCA	GAC	GGG	CAG	GTG	ATC	ACC	ATC	GGC	AAC	GAG	CGC	TTC	CGC	TGC	CCG	GAG	ACG ...
gamma2	...TAT	GAG	CTG	CCA	GA	GGG	CAG	GT	ATC	ACC	AT	GGC	AA	GAG	CGC	TTC	CGC	TGC	CC	GAG	AC ...

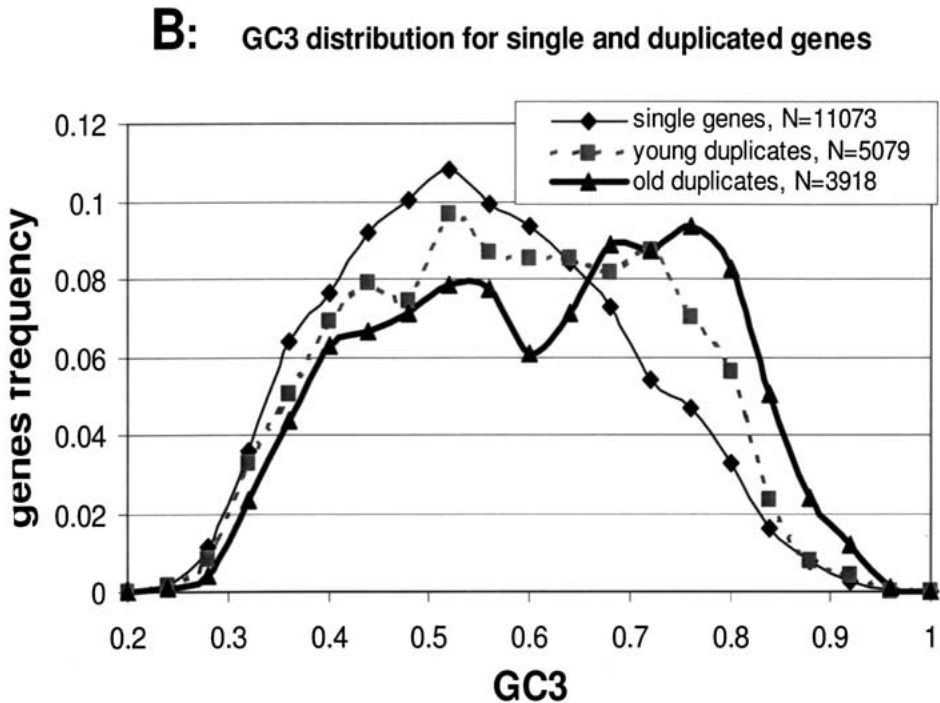


Fig. 3. GC asymmetry of human gene duplicates. **A** An example of the GC3 asymmetry: a fragment of aligned α -1- and γ -2-actin genes is shown with all seven nucleotide differences (in boldface) occurring in silent positions and all but one at CpG sites. **B** GC3

frequency distributions of human single and duplicated genes. Young duplicates have more than 95% identical amino acids, whereas old duplicates are all those in the range between 60 and 95% (see Materials and Methods for detail).

flanking regions. Quite curiously, at the larger scale of genome-wide scanning we found very few cases where the GC content was lower in the gene than in the surrounding isochores.

Thus, the GC content of gene duplicates in mammalian genomes is undoubtedly associated with their isochores organization. Therefore, it was quite a surprise when for gene duplicates from nonisochoreic genomes of fish, fly, worm, plant, and yeast, we observed virtually the same bimodal type of GC3 distribution as shown in Fig. 3, suggesting that there should be some universal causes of the phenomenon (see also Position Effects and Discussion). Consistent with this, the GC spectra of flanking regions for the most GC-asymmetric gene pairs in the nonisochoreic genome of *C. elegans* turned out to be qualitatively similar to those in the isochoreic genome of *H. sapiens* (Fig. 4). The difference is a smaller distance between GC-rich and GC-poor flanks of nematode duplicates, most likely reflecting a higher GC homogeneity of *C. elegans*. Preliminarily, we therefore speculate that *C. elegans* also has the isochores-like

domains but its small genome does not contain enough “junk” (constraint-free) intergenic DNA for isochores to become easily observable.

Mutational Asymmetry

Each of the GC-asymmetric gene pairs has a single common gene ancestor. Obviously, the GC asymmetry appeared due to mutations that occurred after duplication in independently diverging paralogous genes. It is also obvious that when dealing not with a genealogical tree of a multigene family but with only pairs of aligned GC-asymmetric genes without knowing their nearest ancestor, one cannot distinguish direct and reverse base substitutions at every individual site: both could occur. Yet, at any rate, the GC asymmetry necessarily suggests the mutational asymmetry. Figure 5 makes this clear by the example of two homologous genes and their common ancestor. Assume (for simplicity) that all three of these consist of C and T bases only and consider the case of an extreme asymmetry of descendant genes, i.e., the

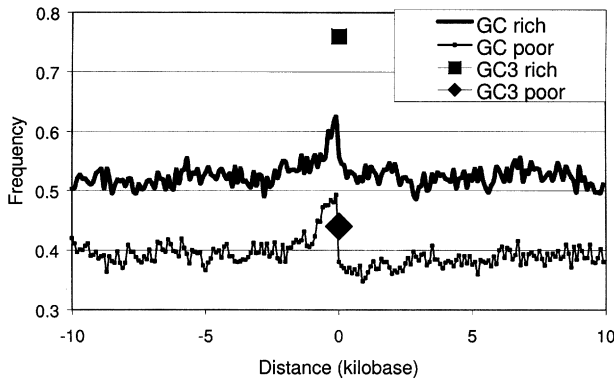


Fig. 4. The GC3 content of human gene duplicates and GC spectra of their flanking regions. All human gene pairs were sorted out by decreasing GC3 difference between members of pairs. Shown are two GC frequency spectra averaged for top 100 of the most asymmetrical pairs: the GC-rich (upper spectrum) and GC-poor (lower spectrum) members of the pairs, respectively. Genes are indicated by square and diamond. The GC levels of upstream and downstream flanking regions of the genes within isochores were calculated in a 100-bp-long window.

sites in which one gene differs from another are all Cs in the first and all Ts in the second (Fig. 5). The bases at these sites in the ancestral gene are unknown (and cannot be even roughly identified for two genes in principle), but apparently there are three qualitatively different possibilities: the ancestral bases are either all C, or all T, or part C and part T. What matters here is that whatever the ancestor might be, it is a strong mutational bias toward C→T in one pathway and/or T→C in another that eventually generates the base asymmetry of gene duplicates, i.e., as Fig. 5 clearly shows, the rates of C→T and T→C mutations are highly unequal in each of the two diverging genes.

The foregoing means that in order to test a gene pair for mutational asymmetry, one needs only to separately count the number of sites with this mutation, e.g., a C→T transition, when C is in the first gene of each pair and T in the second gene, and then the number of sites with a T→C transition when T is in the second gene and C in the first gene. If recently emerged identical genes mutate at equal rates, one would expect equal probabilities for these two results to arise, direct C→T and reverse C←T mutations. The only source of asymmetry would then be statistical fluctuations limited by binomial distribution. For two aligned genes with n specific base pair differences derived from the corresponding direct and reverse substitutions (e.g., C→T plus C←T), the probability of finding exactly x C↔T transitions is $p(x) = \binom{n}{x} \frac{1}{2^n}$, with a standard deviation of $\sigma = \frac{1}{2}\sqrt{n}$. In order to unify a large set of duplicates differing in n , i.e., to operate with the n -independent probability distribution, we used the distribution of normalized deviation from the mean value $U = (x - \frac{1}{2}n)/\sigma = (2x/\sqrt{n}) - \sqrt{n}$, known as deviation measured in

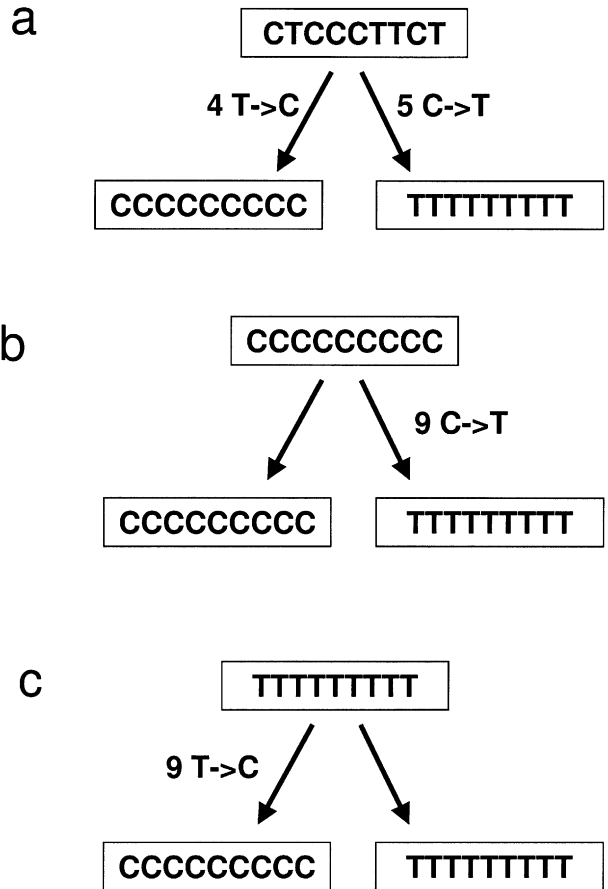


Fig. 5. The auxiliary scheme deducing the mutational bias from the base asymmetry of gene duplicates. Shown is an example of nucleotide differences in nine positions of two homologous genes. For simplicity, the extremely asymmetric case is presented when all these nine sites have C in one gene versus T in its duplicate. The corresponding hypothetical ancestors are indicated as being either maximally homogeneous (b and c) or heterogeneous (a: five C and four T). In each case, the asymmetry of the genes necessarily suggests unequal rates of direct and reverse substitutions (C→T and T→C) in diverging duplicates.

“sigmas.” For random mutations, the probability distribution $\phi(U)$ should be normal (for sufficiently large n) without bias, i.e., $\phi(U) = (1/\sqrt{2\pi})e^{-(1/2)U^2}$. The latter is the distribution in which about 95% of events should lie below two sigmas: $U < 2$.

The distribution of real deviations in the human genome appeared to be significantly wider than expected (Fig. 6A), especially at CpG dinucleotides. This might be due to methylation of cytosine at CpG sites that makes them highly unstable, having a very frequent spontaneous deamination of 5-mC to T (Rideout et al. 1990). CpG is a palindrome, so that if the deamination occurs on the nontranscribed strand of DNA, it directly results in a CpG→TpG transition; if the same, strand-mirror event occurs on the transcribed strand, it produces (in one round of replication) a complementary CpG→CpA transition (Yang et al. 1996). This pair of spontaneous mutations pre-

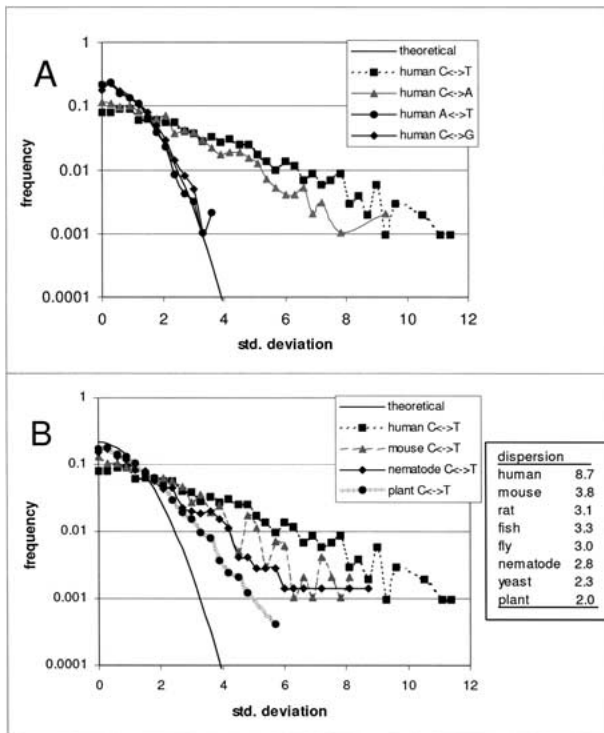


Fig. 6. The distribution $\phi(U)$ of normalized deviations from expected. Each point represents a frequency (Y -axis) of gene pairs with the corresponding “oddness” of pairwise asymmetry measured in sigmas (X -axis). Young duplicates (nucleotide identity $>80\%$) were excluded from the analysis, as they did not diverge enough to show a significant asymmetry and may contain future pseudogenes. A solid line shows the normal (Gaussian) distribution that is expected under the assumption of equal substitution rates in gene duplicates. If a point is higher than expected, this means that mutation rates are significantly duplicate-asymmetric in the corresponding pairs. **A** Distributions of $C \leftrightarrow T$, $C \leftrightarrow A$, $A \leftrightarrow T$, and $G \leftrightarrow C$ in gene pairs of the human genome ($G \leftrightarrow A$ and $G \leftrightarrow T$ are not shown as strand-complementary to $C \leftrightarrow T$ and $C \leftrightarrow A$). **B** $C \leftrightarrow T$ distributions of human, mouse, nematode, and mustard weed genomes. **Inset:** The dispersion of the $\phi(U)$ distribution for $C \leftrightarrow T$ transitions in different eukaryotic species.

vails over all others at m-CpG dinucleotides and it is this pair that shows the greatest bias in the $\phi(U)$ distributions (Fig. 6A). However, the strong bias remains even when CpG dinucleotides are excluded from the count: dispersion $\sigma^2 = 4.3$ for $C \rightarrow T$ transitions. This suggests that the phenomenon is not associated solely with these hypermutable sites.

Interestingly, not only are the same gene pairs biased in the most frequent $C \rightarrow T$ and $G \rightarrow A$ transitions, but also there is a colinear though smaller bias in $G \rightarrow T$ and complementary $C \rightarrow A$ transversions, (Fig. 6A), the $G \rightarrow T$ most likely originating on the nontranscribed strand and the $C \rightarrow A$ on the transcribed strand (Rodin et al. 2002). Only $G \leftrightarrow C$ and $A \leftrightarrow T$ transversions behave as expected for gene-symmetric mutagenesis (Fig. 6A). One possible explanation is that at least in isochoric genomes, neither $G \leftrightarrow C$ nor $A \leftrightarrow T$ mutations can change the GC content. Besides, $G \leftrightarrow C$ substitutions are self-strand-symmetric: regardless of whether a $G \rightarrow C$ or $C \rightarrow G$

transversion occurred, neither changes the G:C base pair itself. Therefore, one cannot distinguish the $G \rightarrow C$ transversion originating on the nontranscribed strand from the reverse event, $C \rightarrow G$, when the latter actually originated from $G \rightarrow C$ on the transcribed strand (Rodin et al. 2002). Accordingly, if in a given gene pair $G \leftrightarrow C$ transversions are indeed strongly biased toward one of the genes but their sequences contain nearly equal numbers of Gs and Cs (and most genes do), this bias is in fact unidentifiable. The same is true for the transversions pair $A \rightarrow T$ and $T \rightarrow A$.

Depletion of methylated CpGs contributes heavily to the mutational asymmetry of human gene duplicates. For example, the human γ -2 actin gene is obviously CpG-poor in comparison with α -1 (26 vs. 96 CpG sites, respectively). However, we unexpectedly found a similar mutational bias in all other species, independent of their methylation status and isochore organization (Fig. 6B). Although the most biased gene pairs occur in the methylated human genome (Fig. 6B and inset), which has a most spectacular isochore structure as well, the mouse genome (also methylated and isochoric) is actually indistinguishable from the unmethylated, isochoreless nematode genome. Moreover, with respect to the mutational asymmetry of gene duplicates, the methylated *A. thaliana* looks inferior even to *C. elegans* (Fig. 6B).

Position Effects

Thus, neither methylation alone, nor isochores, nor both can be a primary cause of the asymmetry described above; rather they represent specific realizations of some universal principle of genome evolution in eukaryotes. Very intriguing in this regard is a feature of duplicate genes, commonly shared by animal genomes (Table 1): genes from GC asymmetric pairs tend to be localized in different chromosomes, whereas genes from symmetric pairs demonstrate the opposite tendency. Since each of the pairs emerged once by duplication, this difference most likely reflects the earlier unknown role of position effects in evolution by gene duplication. Generally, we suggest that the chance of a duplicate to succeed in functional divergence, including a gradual evolution toward a new function, is greater for those which change their position in the genome. A distantly relocated gene copy is more likely to experience a different chromatin environment and a different epigenetic regulation than in the environment of the original functional gene. More specifically, a new position may positively influence the functional divergence of gene duplicates by means of their nonoverlapping stage/tissue-specific epigenetic inactivation, which makes the duplicates visible for natural selection and thus promotes the elimination of degenerative and

Table 1. Chromosomal localization of homologous gene pairs^a

Species	Top 50 asymmetric pairs		The rest of the pairs	
	Same chromosome	Different chromosomes	Same chromosome	Different chromosomes
<i>H. sapiens</i> (24) ^b	0	50	550	1108
<i>M. musculus</i> (21)	1	49	568	1142
<i>A. thaliana</i> (5)	21	29	1500	2555
<i>D. melanogaster</i> (5)	30	20	225	71
<i>C. elegans</i> (6)	22	28	746	275

^aTo minimize pseudogene influence, only the pairs with identity < 80% were considered (the trend remains the same when all the pairs are included). To avoid interference of conversion, only small multigene families (not containing more than five genes) were used.

^bNumber of chromosomes in the haploid set.

the fixation of advantageous mutations toward neofunctionalization (Rodin and Riggs 2003; Fig. 1). All these evolutionary effects are more likely when a duplicate gene transfers to a different chromosome and this may explain Table 1.

A closer “same vs. different chromosome” comparison of gene pairs revealed some striking details of their evolutionary dynamics. Assuming that silent mutations accumulate at a rate that is proportional to time, one can see in the human genome that (Figs. 7 and 8):

1. The majority of very recent duplicates are localized on the same chromosome, most likely as tandem repeated genes. Only about 10% of such very young duplicates occur on different chromosomes (Fig. 7).
2. With increasing antiquity (in the interval $0 < S < 0.02$) the number of syntenic duplicates drops steeply, whereas those on different chromosomes show a notable and rapid increase (Fig. 7).
3. Undoubtedly, the increase in the number of duplicates on different chromosomes occurs at the expense of those located on the same chromosome. However, the loss of syntenic duplicates considerably exceeds the establishment of nonsyntenic duplicate genes. This imbalance suggests that the majority of the newly born tandem duplicates perish if they stay in the same place, whereas translocation to more distant places, including other chromosomes, favors their survival.
4. Furthermore, the comparison of average R/S values in these two groups indicates that repositioning may not only “save” a gene duplicate from pseudogenization but also promotes its very fast adaptive evolution (Fig. 8: R/S = 2.25) driven by positive selection followed by a progressive decline to R/S < 1 that reflects the strong surveillance of negative selection. These dynamics are quite consistent with the EC model (Rodin and Riggs 2003). In contrast, a gene duplicate that stayed in the same place as its twin will most likely keep the previous epigenetic control so that its preservation

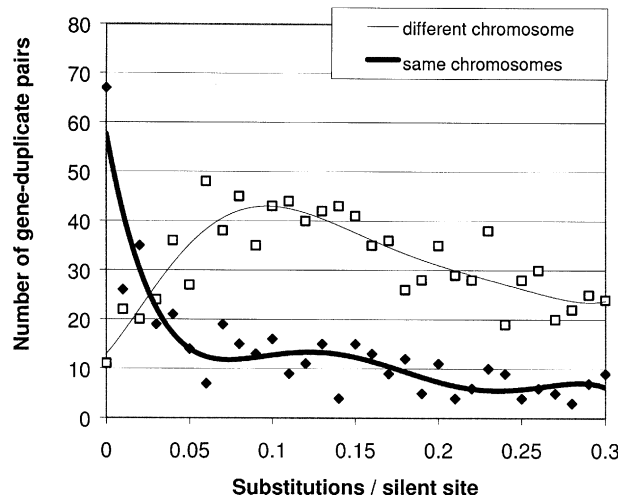


Fig. 7. The number of human duplicates as a function of the number of silent substitutions per silent site. The “same chromosome” (syntenic) group includes the pairs in which both duplicates are located on the same chromosome; accordingly, the pairs of duplicates located on different chromosomes constitute the “different chromosomes” (nonsyntenic) group.

could be provided mostly by DDC-like mechanisms (Force et al. 1999; Lynch et al. 2001). The average R/S ratio in the group of duplicates that did not change their chromosome gradually declines with time from the nearly “neutral” values (R/S = 1.3) (Fig. 8). As an approximate “quick and dirty” approach to estimation of synonymous divergence and in order to check the trends (Fig. 8) for consistency, we simply counted and compared substitutions in the third and second codon positions. The trends remained virtually the same.

Remarkably, all other species exhibit the same contrasting difference between the two groups of gene pairs; their detailed comparison will be published elsewhere.

Position-Determined Mutation Rate

The mutational asymmetry shown in Fig. 5 and documented in Fig. 6 is evidence that two GC-

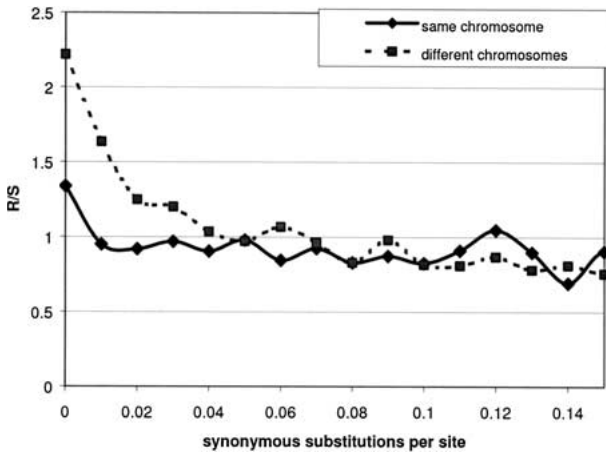


Fig. 8. Change in the R/S ratio with time (measured by a number of silent substitutions per silent site) for gene duplicates. The gene pairs are the same as in Fig. 6. Immediately after gene duplication, those copies transferred to different chromosomes tend to evolve faster.

asymmetric duplicates could both evolve at comparable mutation rates but in opposite directions. In reality, however, the balance seems to be strongly shifted to one of the genes due to a higher mutability or (and) selective advantage. We suppose there are at least three arguments for this imbalance. First, in evolution by gene duplication the typical situation is that one duplicate has to retain an old function, whereas its redundant copy may acquire new function-prone mutations. Accordingly, double translocation when both duplicates move (each to a new place) should rather often disrupt the regulatory system of an old gene, and therefore these cases are very rare (if exist at all). If so, only one of two duplicates would accumulate mutations to fit a new chromatin environment (e.g., the GC-contrasting isochore) (Fig. 9).

Second, G and C are generally more mutable than A and T (Li 1997). Methylated CpG dinucleotides, in particular, decay more easily than they form (Li 1997). Accordingly, at CpG sites the C→T and G→A transitions as well as the G→T and C→A transversions happen much more frequently than the reverse events: T→C, A→G, and T→G, A→C, respectively (Li 1997). Although with less contrast, the same inequality is true for these base substitutions at non-CpG sites (Li 1997).

Third, at least in mammalian genomes, even among young duplicates, we actually do not find AT-rich genes in GC-rich isochores. Does this mean that translocation of AT-rich genes to GC-rich isochores rarely happens or that these translocants do not survive for some unknown reasons? We favor the second possibility. Consistent with it is that ubiquitously expressed housekeeping genes tend to be located in GC-rich isochores, whereas strictly tissue-

specific genes preferentially occur in GC-poor isochores (Vinogradov 2003). We suppose, therefore, that the rarity of AT-rich duplicates in GC-rich isochores reflects a more general “evolutionary rule”: to come from general to specific is much easier than *vice versa*. At any rate, this tendency might be one more, external, cause of the general mutational directedness—from GC- to AT-rich duplicates.

Discussion

That the translocation of a gene often changes its expression and may strongly affect development has been known for 70 years (Muller 1930; Lewis 1950; Wilson et al. 1990). However, it remains unclear if such repositioning exerts any influence on the evolutionary fate of duplicated genes. That some duplicated genes have opposite GC contents is also a long-known fact (Ikemura and Aota 1988; Ellsworth et al. 1994; Li 1997). However, the universality of this GC asymmetry and the mechanism(s) producing and maintaining it are unknown. The present genome-scale study of gene duplicates fills both gaps; furthermore, it directly points to a plausible link between (1) the GC asymmetry of diverged duplicates and the (2) relocation of the extra gene copy soon after duplication and (3) its chance to survive and (4) eventually evolve a new function. *Arabidopsis thaliana* seems to be the only exception. Indeed, unlike other multicellular organisms studied, *A. thaliana* shows no evidence of significant differences in chromosomal localization between genes from symmetric and asymmetric gene pairs (Table 1). Yet one would expect this exception in as much as many gene duplicates in *A. thaliana* might originate by an ancient polyploidization event(s) (Grant et al. 2000; Lynch and Conery 2000; Wolfe 2001). It seems to us that global genomic doublings simply reproduce, at least originally, all the previous position relations between genes with mutually balanced expression and therefore actually do not change a local chromatin environment for every new copy of a gene.

In general, mutational and epigenetic complementary inactivations of duplicates are cooperative rather than antagonistic processes (Rodin and Riggs 2003). It is clear, however, that a repositioned duplicate gene has a greater chance to encounter a different chromatin environment and hence different epigenetic tissue- and stage-specific inactivation. We suppose, therefore, that the epigenetic (EC) model applies more to translocated duplicates, whereas tandem duplicates might be preserved mostly by the position-unspecific mutational (DDC) mechanism (Force et al. 1999; Lynch and Force 2000). The latter is amenable to direct experimental tests.

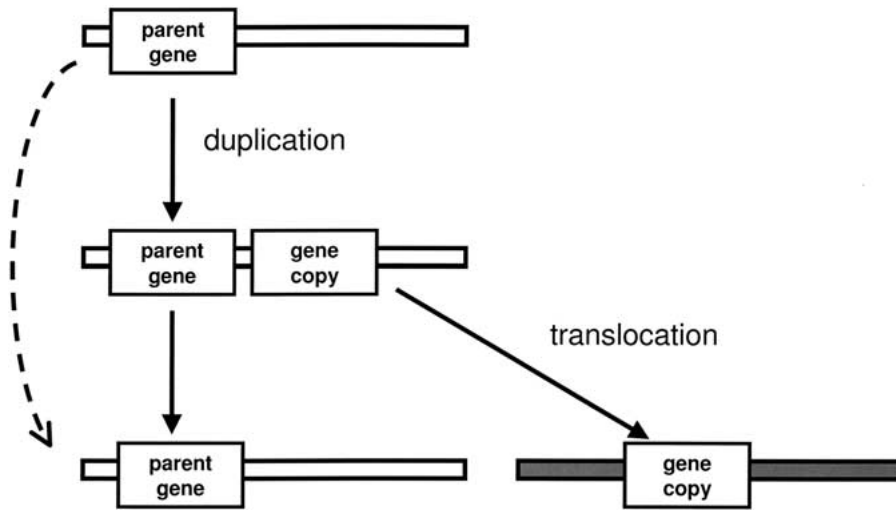


Fig. 9. The scheme in support of the hypothesis of an increased evolutionary rate in repositioned duplicates. Since the probability of simultaneous translocation of both new duplicates is negligible, only one is shown as moving to a new place, which could be on either the same or a different chromosome (the dashed arrow indicates the possibility of a duplicate's translocation at the origin,

i.e., without an intermediate phase of tandem duplicates). In any case, a duplicate is likely to be placed in a new chromatin environment and under different epigenetic control. In isochoric genomes, such repositioning may bring a duplicate into a different isochore (shaded). As a consequence, this may increase the evolutionary rate due to either selection, mutagenesis, or both.

Another difference between DDC and EC models is that DDC tacitly suggests the preexistence of multiple regulatory elements before duplication. Moreover, consecutive extrapolation of the DDC backward to the evolutionary root of related genes inevitably leads us to the awkward conclusion that the very founder of any multigene family had the most complex tissue/stage-specific control of expression compared to all its descendants. This means that although DDC may preserve many gene duplicates in already well-evolved complexly regulated multigene families, it cannot explain the progressive evolution of the complexity itself. In paralogous genes, however, regulatory DNA evolves much faster than coding DNA (Hardison 1998), and this difference is consistent with the hypothesis that many new regulatory sequences are shaped by the positive selection at about the time of or just after gene duplication (Rodin and Riggs 2003). Indirect signs of not only degenerative but also generative evolution have been reported for regulatory sites of genes in the same multigene families (Skaer et al. 2002; Chiu et al. 2002). Furthermore, a recent genome-wide examination of duplicated genes in yeast revealed strong evidence for positive selection acting on cis-regulatory elements after duplication (Papp et al. 2003).

Table 1 and Figs. 7 and 8 unambiguously demonstrate a strong positive effect of repositioning on the fate of gene duplicates with respect to their survival and prospective functional divergence. The result is all the more remarkable if one takes into account that our test—comparison of GC-asymmetric duplicates located within the same chromosome with those located in different chromosomes—is

imperfect because the “same-chromosome” group certainly contains some admix of the repositioning cases when one of the genes in the pair did move to a new position and far enough from the previous place but within the same chromosome. This means that the real magnitude and effect of repositioning surpass even those shown in Table 1 and Figs. 7 and 8.

Yet, even underestimated, the repositioning effect is startlingly rapid (Figs. 7 and 8), in perfect agreement with the predictions of the EC model (Rodin and Riggs 2003). Generally consistent with this are increasing data showing that accelerated evolution of new repositioned gene duplicates is a general phenomenon (Long et al. 2003). Even among processed duplicates most of which are “dead on arrival” to a new genomic position, there are functional protein-coding retrogenes such as the chimeric *jingwei* in *Drosophila* and the *PGAM3* in primates that demonstrate at least an order of magnitude faster evolution and a very rapid emergence of a new expression pattern (Long and Langley 1993; Betran et al. 2002). Also consistent are two recent genome-scale findings for *Sacharomyces cerevisiae*. Using microarray data, Gu et al. (2002) did reveal a very rapid divergence in expression between yeast duplicated genes and Papp et al. (2003) found some evidence for positive selection acting on cis-regulatory elements after duplication and directing evolution toward the gain of functionally novel regulatory motifs. Interestingly, recent direct comparisons of over 100 human and chimpanzee genes showed that, on average, proteins from chromosomes that have undergone large structural rearrangements have been evolving more than

twice as fast as those from colinear chromosomes that preserved the same gene order (Navarro and Barton 2003). This difference supports the model of speciation in which chromosomal rearrangements trigger the separation of species by limiting gene flow within the rearranged region and, as a consequence, accumulating mutations by positive selection (Navarro and Barton 2003; Rieseberg and Livingstone 2003). Other explanations of accelerated evolution in rearranged chromosomes are also possible (Rieseberg and Livingstone 2003). Redundant genes seem to be of particular significance here since the primary negative effects of rearrangements on fitness, such as hybrid sterility, are reduced (Rieseberg and Livingstone 2003).

The GC asymmetry of gene duplicates sheds some light on the long-standing controversy about the major forces that generate GC-contrasting isochores. Two conflicting hypotheses have been put forward: the mutationalist (Filipksi 1987; Wolfe et al. 1989, Holmquist and Filipksi 1994) and the selectionist (Bernardi et al. 1985; Bernardi 2000) ones. Immediately after duplication, both genes are identical and have the same GC content. The asymmetry arises due to relocation of one of the duplicates to a new place that may differ from the original one in GC level, and as a consequence, the duplicate will change its mutational vector. Quite consistent with this position-specific mutation pressure is the hypothesis (Wolfe et al. 1989) that relates the GC content of genes to regions of early and late replication in which the nucleotide precursor pools have high and low GC contents, respectively. This hypothesis also explains the case of nonisochore genomes, again consistent with the universal, species-independent GC asymmetry of duplicates (Fig. 6B). At the same time, the striking deficiency of GC-poor young duplicates in GC-rich isochores may point to strong large-scale selection acting against duplicates of tissue-specific genes when they move from AT-rich to GC-rich isochores to which ubiquitously expressed housekeeping genes are predominantly mapped (Vinogradov 2003). The opposite direction of repositioning, from housekeeping to strictly tissue-specific regions, seems to be much less constrained or even favored by the positive selection in evolution. Our parallel phylogenetic analysis of two main hemoglobin gene loci, the housekeeping-like α - and the strictly erythrocyte-specific β -globin clusters supports this hypothesis (Rodin et al., submitted). Since the majority of new genes descend from duplicates of old ones, this kind of selection operating at the level of gene repositioning could actually use the difference in the G,C pool between regions of early and late replication and thus have greatly contributed to the global trend of reducing the genes' total GC content (Bird 1993).

Selection is at work as long as mutagenesis goes on. One of the new-place effects (reached by moving, for example, from G(C)-rich to G(C)-poor isochores) is a relative increase in the mutation rate in the duplicate compared with its rate in the previous place. One can perceive here a parallel with an unmethylated "master" *Alu* sequence and the high rate of independent mutations observed in its progeny: retransposed extensively methylated *Alu* repeats, most mutations occurring at methylated CpG dinucleotides (Deininger et al. 1992).

The increased mutagenesis rate does not imply a lessening of the role of selection in shaping the position-specific GC content of genes. On the contrary, judging *a posteriori*, the mutation rate increase is conducive to a selection-driven gradual gain of new functions by speeding up the process. Consistent with this is that nonsilent mutations also demonstrate gene asymmetry (data not shown), which is almost ideally colinear with the synonymous mutational bias. The current concepts of the molecular evolutionary clock (Zuckerkandl and Pauling 1966; Kimura 1983) and evolutionary distance between genes (Ratner et al. 1996; Li 1997) do not take these position effects into account.

As already mentioned, our results suggest that although newly duplicated genes can evolve asymmetrically in both directions, there is a general rather distinct trend of mutations from (GC) to (AT) rather than the reverse. Figure 3A represents a typical case of this bias within a short fragment of aligned human α -1- and γ -2-actin genes. All seven implied substitutions are synonymous, thus excluding any strong interference of selection. All but one are at CpG sites, and most might have occurred in the pathway leading to the γ -2 actin gene. The asymmetry for the entire alignment is even more impressive: silent C:G \rightarrow T:A transitions and C:G \rightarrow A:T transversions at CpG sites might be greatly biased to the γ -2 gene (38 vs. 0 transitions and 27 vs. 4 transversions in γ -2 and α -1 genes, respectively).

Our preliminary comparative analysis of some highly asymmetric gene pairs in different species confirms this trend. For example, the *H. sapiens* vs. *R. norvegicus* comparison indicates that, after separation of these two species, the mutation rate of the putatively methylated heat shock 8 gene is still at least twice as high as the rate of its unmethylated partner, the heat shock 1A gene. However, some other gene pairs (e.g., α 1- and γ 2-actins) evolve at nearly equal rates in these two species, thus suggesting that after duplication and repositioning of one of the duplicates, their mutational asymmetry reached the saturation level very quickly and apparently long before primate and rodent divergence. A closer phylogenetic analysis of embryonic and adult globins in α - and β -like gene clusters shows that rapid satura-

tion is the rule, rather than an exception (Rodin et al. submitted).

Metaphors such as “the right gene in the right place for the right development” need no comment. The genome-wide GC asymmetry of duplicated genes described in this paper points to the importance of position effects for the “right” evolution as well. In this regard, our own species having the most spectacularly isochoric and methylated genome notably outruns the others; even the orthologous gene pairs of other mammalian species (mouse and rat, for example) are less asymmetrical (Fig. 6). We suggest that, among other possibilities, this highest GC asymmetry combined with epigenetic differentiation of young gene duplicates (Rodin and Riggs 2003) might have been established in the human lineage of evolution as a kind of “internal compensation” (Rodin 1991; Ratner et al. 1996) for small effective population size and long development.

In conclusion, despite almost common expectations, comparative genome studies reveal that the number of genes is not commensurate with the phenotypic complexity of organisms. Of course, this G-value paradox (Hahn and Wray 2002; Betran and Long 2002) does not disprove the classic idea of progressive evolution by gene duplication but, rather, readdresses its direct application. Emerging evidence indicates that indeed it is not the total number of genes, but rather the diversity of their expression patterns, which correlates with organismal complexity. The expression diversity in turn depends on (i) the number of those genes that encode transcriptional factors and, accordingly, (ii) the number of their cis-control elements—promoters, enhancers, silencers, etc (Levine and Tjian 2003). Importantly, this source of functional diversity as well as exon reshuffling and alternative splicing (Baltimore 2001; Graveley 2001) are each, in one way or another, based on DNA duplications. We will next undertake a detailed analysis of the mutational and positional asymmetry in multigene families that are directly involved in the regulation of gene expression.

Note in proof. After acceptance of this paper, we found that shortly before its submission, K. Jabbari, E. Rayko, and G. Bernardi hypothesized that GC asymmetry of many gene duplicates in the human genome might reflect their ancient translocations in the GC-rich ancestral genome core in contrast to our results suggesting that these would usually be AT-rich isochores. These two hypotheses actually complement rather than exclude each other because most of the gene duplications analyzed by Jabbari et al. (2003) have occurred long before the transition from cold- to warm-blood vertebrates, while our interpretation refers to relatively young duplicates.

Acknowledgments. We thank A. Riggs, G. Holmquist, and A. Rodin for stimulating discussions and comments on the manuscript. Our special thanks go to S. Bates for reading the manuscript.

References

- Aota S, Ikemura T (1986) Diversity in G+C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res* 14:6345–6355
- Baltimore D (2001) Our genome unveiled. *Nature* 409:814–816
- Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958
- Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* 241:3–17
- Betran E, Long M (2002) Expansion of genome coding regions by acquisition of new genes. *Genetica* 115:65–80
- Betran E, Wang W, Jin L, Long M (2002) Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. *Mol Biol Evol* 19:654–663
- Bird AP (1993) Functions for DNA methylation in vertebrates. *Cold Spring Harbor Symp Quant Biol* 58:281–285
- Brown KE, Amoils S, Horn JM, Buckle VJ, Higgs DR, Merckenschlager M, Fisher AG (2001) Expression of α and β -globin genes occurs within different nuclear domains in haemopoietic cells. *Nature Cell Biol* 3:602–606
- Carmichael GG (2003) Antisense starts making more sense. *Nat Biotechnol* 21:371–372
- Chiu C-H, Amemiya C, Dewar K, Kim C-B, Ruddle FH, Wagner GP (2002) Molecular evolution of the HoxA cluster in the three major gnathostome lineages. *Proc Natl Acad Sci USA* 99:5492–5497
- Cockell M, Gasser SM (1999) Nuclear compartments and gene regulation. *Curr Opin Genet Dev* 9:199–205
- Deininger PL, Batzer MA, Hutchinson CA 3rd, Edgell MH (1992) Master genes in mammalian repetitive DNA amplification. *Trends Genet* 8:307–311
- Ellsworth DL, Hewett D, Li W-H (1994) Evolution of base composition in the insulin and insulin-like growth factor genes. *Mol Biol Evol* 11:875–885
- Filipski J (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosomal binding and chromatin compactness in germline cells. *FEBS Lett* 217:184–186
- Force A, Lynch M, Pickett B, Amores A, Yan Y-I, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545
- Grant D, Cregan P, Shoemaker RC (2000) Genome organization in dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. *Proc Natl Acad Sci USA* 97:4168–4173
- Graveley BR (2001) Alternative splicing: Increasing diversity in the proteomic world. *Trends Genet* 17:100–107
- Gu Z, Nicolae D, Lu H, Li W-H (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* 18:609–613
- Hahn MV, Wray GA (2002) The g-value paradox. *Evol Dev* 4:73–75
- Haldane JBS (1932) *The causes of evolution*. Longman and Green, London
- Haldane JBS (1933) The part played by recurrent mutation in evolution. *Am Nat* 67:5–19
- Hardison R (1998) Hemoglobin from bacteria to man: Evolution of different patterns of gene expression. *J Exp Biol* 201:1099–1117

- Holmquist GP, Filipksi J (1994) Organization of mutations along the genome: A prime determinant of genome evolution. *Trends Ecol Evol* 9:65–69
- Jabbari K, Rayko E, Bernardi G (2003) The major shifts of human duplicated genes. *Gene* 317:203–208
- Jenuwein T, Allis CD (2001) Translating the histone code. *Science* 293:1074–1080
- Ikemura T, Aota S-I (1988) Global variation in G+C content along vertebrate genome DNA: Possible correlation with chromosome band structures. *J Mol Biol* 203:1–13
- Kimura M (1983) *The neutral theory of molecular evolution*, Cambridge University Press, New York
- Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424:147–151
- Lewis EB (1950) The phenomenon of position effect. *Adv Genet* 3:73–115
- Li W-H (1997) *Molecular evolution*, MA Sinauer Associates, Sunderland, MA
- Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* 2:150–174
- Long M, Langley CH (1993) Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* 260:91–95
- Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: Glimpses from the young and old. *Nat Rev Genet* 4:865–875
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155
- Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473
- Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* 159:1789–1804
- Muller HJ (1930) Types of visible variations induced by X-rays in *Drosophila*. *J Genet* 22:299–335
- Muller HJ (1935) The organization of chromatin deficiencies as minute deletions subject to insertion elsewhere. *Genetics* 17:237–252
- Navarro A, Barton NK (2003) Chromosomal speciation and molecular divergence—Accelerated evolution in rearranged chromosomes. *Science* 300:321–324
- Ohno S (1970) *Evolution by gene duplication* Springer, Berlin
- Ohno S (1972) So much “junk” DNA in our genome. *Brookhaven Symp Quant Biol* 23:366–370
- Papp B, Pal C, Hurst LD (2003) Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet* 19:417–422
- Ratner V, Zharkikh A, Kolchanov N, Rodin S, Solovyov V, Antonov A (1996) *Molecular evolution*. Springer, Berlin
- Rideout WM III, Coetzee GA, Olumi AF, Jones PA (1990) 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science* 249:1288–1290
- Rieseberg LH, Livingstone K (2003) Chromosomal speciation in primates. *Science* 300:267
- Rodin SN (1991) *Idea of coevolution*. Novosibirsk, Nauka (in Russian)
- Rodin SN, Riggs AD (2003) Epigenetic silencing may aid evolution by gene duplication. *J Mol Evol* 56:718–729
- Rodin SN, Rodin AS, Juhasz A, Holmquist GP (2002) Cancerous hyper-mutagenesis in p53 genes is possibly associated with transcriptional bypass of DNA lesions. *Mutat Res* 510:153–168
- Rodin SN, Parkhomchuk DV, Rodin AS, Riggs AD (2004) Molecular recapitulation: Position – dependent evolution of genes in α - and β -like globin clusters. Submitted for publication
- Serebrovsky AS (1938) Genes scute and achaete in *Drosophila melanogaster* and the hypothesis of their divergence. *Proc Russ Acad Sci* 19:77–81 (in Russian)
- Skaer N, Pistillo D, Gibert J-M, Lio P, Wulbeck C, Simpson P (2002) Gene duplication at the achaete-scute complex and morphological complexity of the peripheral nervous system in Diptera. *Trends Genet* 18:399–405
- Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147:195–197
- Vinogradov AE (2003) Isochores and tissue-specificity. *Nucleic Acids Res* 31:5212–5220
- Wagner A (1998) The fate of duplicated genes: loss or new function. *Bio Essays* 20:785–788
- Wilson C, Bellen HJ, Gehring WJ (1990) Position effects on eukaryotic gene expression. *Annu Rev Cell Biol* 6:679–714
- Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* 2:333–341
- Wolfe KH, Sharp PM, Li W-h (1989) rates of synonymous substitution in plant nuclear genes. *J Mol Evol* 29:208–211
- Yang AS, Jones PA, Shibata A (1996) The mutational burden of 5-methylcytosine. In: Russo VEA, Martienssen RA, Riggs AD (eds) *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, Plainview, NY, pp 77–94
- Zuckerkindl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) *Evolving genes and proteins*. Academic Press, New York, pp 97–166