The neoselectionist theory of genome evolution

Giorgio Bernardi*

Molecular Evolution Laboratory, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy

Edited by Tomoko Ohta, National Institute of Genetics, Mishima, Japan, and approved April 4, 2007 (received for review February 22, 2007)

The vertebrate genome is a mosaic of GC-poor and GC-rich isochores, megabase-sized DNA regions of fairly homogeneous base composition that differ in relative amount, gene density, gene expression, replication timing, and recombination frequency. At the emergence of warm-blooded vertebrates, the gene-rich, moderately GC-rich isochores of the cold-blooded ancestors underwent a GC increase. This increase was similar in mammals and birds and was maintained during the evolution of mammalian and avian orders. Neither the GC increase nor its conservation can be accounted for by the random fixation of neutral or nearly neutral single-nucleotide changes (i.e., the vast majority of nucleotide substitutions) or by a biased gene conversion process occurring at random genome locations. Both phenomena can be explained, however, by the neoselectionist theory of genome evolution that is presented here. This theory fully accepts Ohta's nearly neutral view of point mutations but proposes in addition (*i*) that the AT-biased mutational input present in vertebrates pushes some DNA regions below a certain GC threshold; (*ii*) that these lower GC levels cause regional changes in chromatin structure that lead to deleterious effects on replication and transcription; and (*iii*) that the carriers of these changes undergo negative (purifying) selection, the final result being a compositional conservation of the original isochore pattern in the surviving population. Negative selection may also largely explain the GC increase accompanying the emergence of warm-blooded vertebrates. In conclusion, the neoselectionist theory not only provides a solution to the neutralist/selectionist debate but also introduces an epigenomic component in genome evolution.

base composition | isochores | nearly neutral theory | neutral theory

The most famous sentence from *The Origin of Species* (1), "This preservation of favourable variations and the rejection of injurious variations I call Natural Selection," suggests a dichotomy in the fate of "variations" and is generally interpreted accordingly. This sentence was immediately followed, however, by another one that still is only exceptionally quoted: "Variations neither useful nor injurious would not be affected by natural selection." I.e., Darwin distinguished not two, but three kinds of variations: advantageous, deleterious, and neutral. Whereas advantageous variations expand in the progeny (by positive, or Darwinian, selection), the deleterious ones tend to disappear (by negative, or purifying, selection), and the neutral ones may come out of their limbo to be fixed (like the advantageous variations), or to disappear (like the deleterious ones). Incidentally, the concept of neutral variations is absent in Wallace (2).

Neutral variations were simply obliterated by the selectionists (the neo-Darwinians) (3, 4), although not by all of them (see ref. 5). They were, however, resurrected by the neutral theory, which broke the long predominance of the selectionist theories (see ref. 6). Indeed, Kimura (7, 8) claimed that "the main cause of evolutionary change at the molecular level—changes in the genetic material itself—is random fixation of selectively neutral or nearly neutral mutants." Darwin's "survival of the fittest" was replaced by Kimura's "survival of the luckiest," and Darwinian and neo-Darwinian evolution were substituted by "non-Darwinian evolution" (9). Kimura's revolutionary proposal started a neutralist/selectionist debate, which is still going on (see ref. 10) and which concerns the important issue of the role of chance in evolution. A very significant modification of the neutral theory was the nearly neutral theory of Ohta (11, 12), who proposed that a substantial fraction of changes are caused by random fixation of nearly neutral changes, a class that "includes intermediates between neutral and advantageous, as well as between neutral and deleterious classes" (13). The first four schemes of Fig. 1 display a qualitative picture of the classical theories just mentioned.

All classical theories proposed that natural selection acted on the "phenotype" [called the "classical phenotype" by Bernardi and Bernardi (14), to distinguish it from the "genome phenotype" presented below]. Whether visualized through hereditary morphological traits (by Darwin), genetic characters (by the neo-Darwinians), or protein and DNA sequences (by Kimura), the classical phenotype is the result of the expression of single genes or of a small number of genes, and its evolution was seen as essentially due to single-nucleotide mutations.

In contrast, the "neoselectionist theory" arose by comparing the genome phenotypes (14), namely the "compositional patterns" or "compositional landscapes" of whole genomes. The "compositional approach" was initially made possible by a high-resolution analysis (see refs. 15–18) of eukaryotic genomes, in which DNA was fractionated by ultracentrifugation in Cs_2SO_4 density gradients in the presence of sequence-specific ligands (Ag^+ or, later, BAMD [bis(acetato-mercuri-methyl)dioxane]). Although experimentally sophisticated, this approach was conceptually simple, because it concerned the frequency of ligand-binding oligonucleotides, a property correlated with the base composition of DNA. This compositional approach was extended by computer work to sequences of genes and whole genomes as soon as they became available.

The compositional approach led to the following: (*i*) the discovery of compositional compartmentalization, compositional correlations, and compositional evolution; (*ii*) evidence that genome evolution does not only proceed by point mutations, but also by regional changes; and (*iii*) the development of the neoselectionist theory of genome evolution. The three points under (*i*) will be summarized in the sections below by using our results on the vertebrate genome.

Compositional Compartmentalization: Isochores and Chromosomal Bands, and Gene Distribution

The genomes of mammals are compositionally compartmentalized (16), in that they are "mosaics of isochores" (18), fairly homogeneous regions, originally estimated as longer than 300 kb (the upper limit that could be measured by the approach used at that time; see previous section). These regions could be assigned to five compositionally narrow families that cover a very wide GC range (34–59% GC in the human genome). This discontinuous compartmentalization was in sharp contrast with the

Author contributions: G.B. designed research, performed research, analyzed data, and wrote the paper.

The author declares no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

*To whom correspondence should be sent. E-mail: bernardi@szn.it.

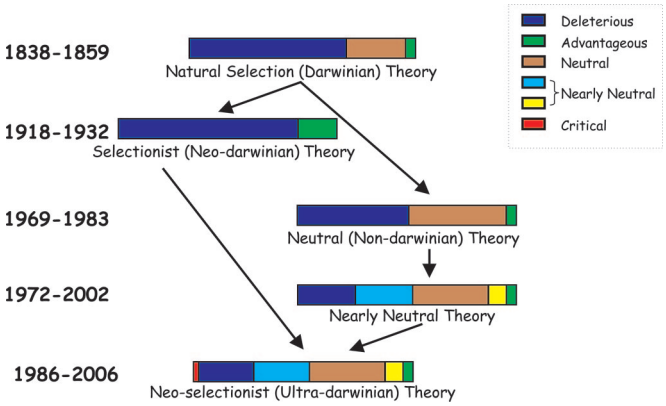

Fig. 1. Darwin postulated the existence of deleterious, advantageous, and neutral changes. The neo-Darwinians (or selectionists) neglected neutral changes. These were reintroduced and amplified by Kimura (7, 8), who developed the neutral theory of evolution (a non-Darwinian evolution, according to ref. 9). The nearly neutral theory was proposed by Ohta (12, 13, 102) to include intermediates between neutral and advantageous, as well as between neutral and deleterious changes. In the neoselectionist theory, the critical changes are responsible for the transition from point mutations to regional changes (modified from refs. 12 and 103).

then-predominant view of a continuous compositional change of DNA along chromosomes.

Recent results (19) have confirmed these conclusions at the sequence level by mapping isochores on human chromosomes and by providing information on the number (≈3,200), the average size (≈1 Mb), and the GC levels of isochores. The standard deviations of GC levels are ≈1% GC within isochores covering 85% of the genome, and ≈2% GC in the remaining, mostly GC-rich, isochores. If isochores are pooled in 1% GC bins, their distribution confirms (19) that they belong in the five families previously described (20) (see Fig. 2). Isochores are, in fact, the ultimate chromosomal bands because below the 100 kb level the GC profiles of chromosomes (especially of the GC-richest regions) are characterized by large fluctuations due to coding sequences, introns, interspersed repeats, etc. Interestingly, the number of isochores is close to that, 3,000, of the highest-resolution bands observed by Yunis *et al.* (21) in early prophase, and the average size of an isochore is equal to the average size of a "site of replication" as defined by Ma *et al.* (22).

It is generally accepted that the molecular basis of cytogenetic bands is not well understood. Isochores do allow one, however, to define the standard Giemsa and reverse bands at a 850- or 400-band resolution on purely compositional grounds (23), thus ending the debate on whether the chromosomal bands are due to DNA, as thought by Caspersson (see ref. 24), or to the associated proteins (25). In addition, the isochore map allowed us to detect a nested structure that concerns not only contiguous isochores but also contiguous high-resolution bands (23).

The assessment of gene density in compositional DNA fractions (see Fig. 2) led to the discovery (20, 26, 27) that genes are not uniformly distributed in the mammalian genome, contrary to what was previously thought. Indeed, in the human genome almost two-thirds of the protein-coding genes are concentrated in the GC-richest isochore families H2 and H3 (only representing 15% of the genome), which was called the "genome core" (28, 29). The rest is spread over the vast (≈85%) GC-poor "empty space," or "empty quarter" (from the name of the Arabian desert), eventually called the "genome desert" (30–32), namely the GC-poor isochore families L1, L2, and H1. These two "gene spaces" (30) are different not only in gene density but also in a number of other basic properties, which are summarized in Fig. 2. Recent data have revealed that the isochore map (19) matches the replicon cluster map of Watanabe *et al.* (33), and that both isochore size and average GC of isochore families are conserved in vertebrates (M. Costantini, F. Auletta, and G.B., unpublished data), reinforcing the concept that isochores represent "a fundamental level of genome organization" (34).

Compositional Correlations: The Genomic Code

Positive compositional correlations were found to hold between coding and contiguous noncoding sequences, both intergenic and intragenic (20). Moreover, GC_1, GC_2, and GC_3 (the GC levels of the three codon positions) were shown to be correlated with each other (14, 35). Such general rules were called (36) "genomic code" (not to be confused with the "genetic code"), a definition later extended to the correlations between the GC levels of the three codon positions with amino acid composition

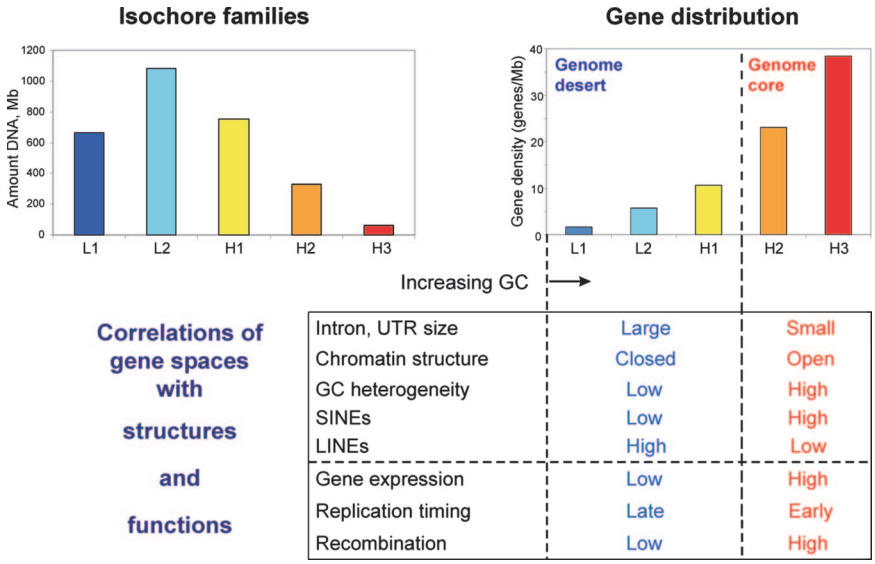

Fig. 2. DNA and gene distribution in the isochore families of the human genome. The major structural and functional properties associated with each gene space are listed (in blue for the genome desert and in red for the genome core). The top frames are modified from M. Costantini, F. Auletta, and G.B. (unpublished data). SINEs, short interspersed sequences; LINEs, long interspersed sequences.

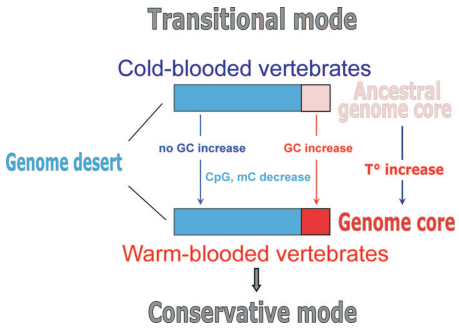

Fig. 3. Scheme of the compositional evolution of vertebrate genomes. At the transition from cold- to warm-blooded vertebrates, the gene-dense, moderately GC-rich ''ancestral genome core'' (pink box) became the gene-dense, GC-rich genome core (red box), but the GC-poor and gene-poor (blue box) genome desert did not undergo any major compositional change. This transitional (or shifting) mode, which was accompanied by an overall decrease of CpG doublets and methylcytosine, was followed by a conservative mode of genome evolution in which compositional patterns were maintained (modified from ref. 32).

(37), hydrophobicity (38), and secondary structures of the encoded proteins (39, 40). Interestingly, the genomic code allows predicting the composition of flanking sequences and introns from the composition of coding sequences (and vice versa), as well as predicting protein properties from the composition of the corresponding coding sequences (and vice versa). Along the same line, plots of GC_2 vs. GC_3 revealed clusters of protein-coding genes corresponding to each isochore family and forming "gene landscapes" (41) that have led to simple proteomic checks for detecting noncoding RNA (42). Needless to say, all these results provided strong evidence for the genome as an integrated ensemble with little or no room left for what Ohno called "junk DNA" (43).

Compositional Evolution: Genome Phenotypes, and the Transitional and Conservative Modes of Evolution

In vertebrates, the compositional patterns of isochores and of coding sequences, as well as the GC_3 levels of orthologous coding sequences, were found to be strikingly different between fish/amphibians on the one hand and mammals/birds on the other, reptiles showing intermediate patterns (17, 44–47). More precisely, at the transition between cold- and warm-blooded vertebrates, the moderately GC-rich ancestral genome core of cold-blooded vertebrates underwent a remarkable GC increase to become the genome core of mammals and birds (48) (see Fig. 3). This implied the existence of a "transitional" (or "shifting") "mode" of evolution that took place independently in the ancestral genomes of mammals and birds, and yet led to similar isochore patterns that were maintained in mammalian and avian genomes (49, 50). In contrast, the genome desert did not undergo any major compositional change (48), except for the decrease of methylcytosine and CpG, which affects the whole genome (see below).

Compositional differences among genomes from vertebrates belonging to the same classes (fishes, amphibians, etc.) could, however, be observed as the result of "whole genome (or horizontal) shifts" (51–53). These shifts apparently are the consequence of changes in genome size, in environmental conditions, in mutation rates, and in the intensity and duration of the mutational AT bias, i.e., the predominance of GC→AT vs. AT→GC changes (54–63). This bias is likely to be due, as in the case of "mutator mutations" of prokaryotes (64), to mutations in the genes coding for subunits of the replication machinery (65).

The compositional transition between cold- and warm-blooded vertebrates could not be accounted for by the random nucleotide substitutions of the neutral theory (see also below) and required, therefore, another explanation. The "thermodynamic stability hypothesis" (14) proposed that the GC increase of the genome core accompanied the emergence of homeothermy and simultaneously provided the selective advantages of increased stabilities of DNA, RNA, and proteins. Indeed, high GC levels in coding sequences not only increase the stability of DNA and of the stem structures of transcribed RNAs, as demonstrated by the GC increases of corresponding stems of ribosomal RNAs from cold- to warm-blooded vertebrates (66, 67), but also favor amino acids that thermodynamically stabilize proteins, as indicated by observations of Argos *et al.* (68) and Zuber (69) that were recently confirmed by Nishio *et al.* (70, 71).

This selectionist explanation was supported (*i*) by the similar genome changes occurring in the independent lines of mammals and birds; (*ii*) by the decreases of CpG doublets and methylcytosine, which are correlated with increasing body temperatures in vertebrates ranging from Antarctic fishes to mammals (72–76); (*iii*) by the variable compositional heterogeneity and methylation levels (intermediate between those of fishes/amphibians and mammals/birds) of the genomes from reptiles (51, 47, 77), which are known to have different body temperatures and thermal regulations; (*iv*) by the mammalian-like isochore organization of the genomes of subtropical and tropical insects (*Drosophila*, *Anopheles*) (78, 79); and (*v*) by the increase of GC levels that accompanies the increase in optimal growth temperatures in many families of prokaryotes (80).

The thermodynamic stability hypothesis stressed the molding effect of the environment on the genome (14). In fact, our use of the old-fashioned distinction between cold- and warm-blooded vertebrates was meant to suggest that the cause of genome changes was "body temperature." The hypothesis could also explain why, at the emergence of a stable body temperature ≈40°C (homeothermy), the "open chromatin" of the genome core, which has an expanded configuration at the center of the interphase nucleus, required a GC increase to be stabilized, whereas the "closed chromatin" of the GC-poor genome desert did not, being stabilized by its own compact structure packed against the nuclear membrane (81). The different chromatin structure could also explain why retroviral sequences initially integrate in GC-rich regions (see ref. 32 for a review).

The transitional mode of evolution stopped in mammals after they emerged from their common ancestor, suggesting that an equilibrium was reached (the decline of the isochore structure in some mammalian orders, proposed by some authors, still is under debate) (see refs. 33 and 82–85). The evidence for this compositionally "conservative mode" of evolution was initially provided by the similarity of the compositional distribution of coding sequences and of GC_3 levels of orthologous coding sequences of mammals from different orders (44). Recent results showed that, in addition, the patterns of orthologous (or syntenic) isochores are conserved in different mammalian orders (e.g., Fig. 4), including the human/mouse case in which one of the two genomes, the mouse genome, is characterized by a great acceleration in substitution rate (86, 87). This is an astonishing result in view of the mutational AT bias of vertebrates and of the fact that the species under consideration have been diverging for 90 million years (88). Because the neutral substitution rate has been estimated as 3.9×10^{-9} substitutions per year (10, 89), ≈70% nucleotide substitutions took place. This would approximately correspond to 54% sequence identity if one accounts for multiple and parallel substitutions. Some differences in size and in GC levels between syntenic isochores do exist, but they are largely accounted for by inserted mobile elements, and by the whole-genome shifts mentioned above. Needless to say, the remarkable conservation of both isochore size and average base composition of isochore families in all vertebrates (M. Costan-

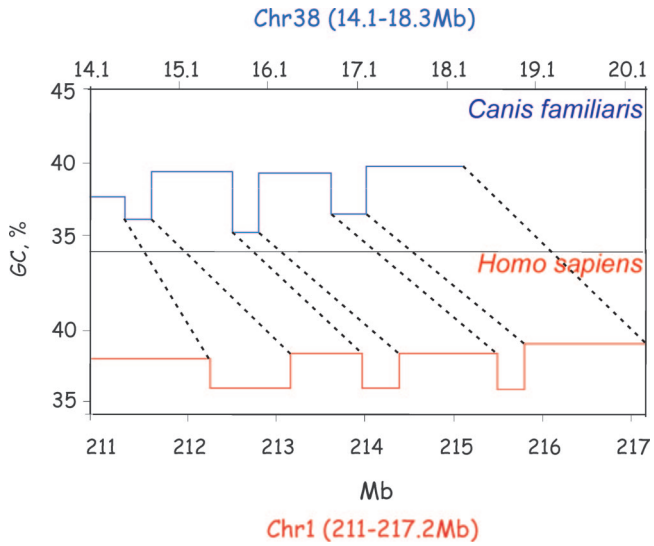

Fig. 4. An example of isochore conservation between syntenic chromosome regions of dog and human (from M. Costantini, F. Auletta, and G.B., unpublished data).

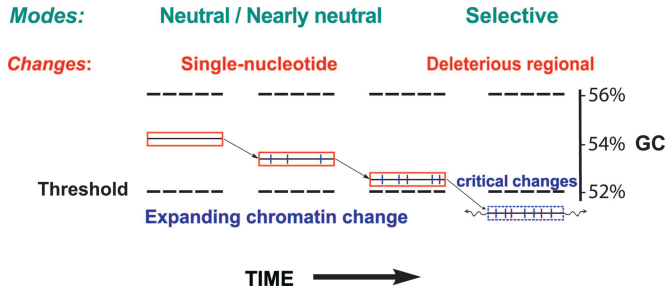

Fig. 5. Time course of typical compositional changes of a GC-rich region from a warm-blooded vertebrate in the conservative mode of evolution. In an early phase, the average GC level of the region, initially visualized at its compositional optimum (arbitrarily set here at 54% GC), is decreasing because of the mutational AT bias (the vertical blue bars crossing the black DNA line represent the "excess" GC→AT changes), but remains within a tolerated range (whose arbitrary thresholds are indicated by the thick horizontal broken lines). In a late phase, the average GC level trespasses the lower threshold (arbitrarily fixed here at 52% GC), because of the last changes, the critical changes. The chromatin (red boxes) then undergoes a structural change (broken blue box) that is deleterious for transcription and replication (see text). Until then, the changes may be neutral or, more probably, nearly neutral (modified from ref. 32).

tini, F. Auletta, and G.B., unpublished data) adds two very important properties to the general picture just presented.

The conservative mode of evolution was originally explained by "negative selection acting at a regional (isochore) level to eliminate any strong deviation from presumably functionally optimal composition of isochores" (44), a general conclusion supported by Gillespie (90). It should be stressed that the compositional conservation of syntenic isochores (see Fig. 4) not only reinforces the original idea of negative selection but also implies a functional role of noncoding sequences in the genome, which is also suggested by a number of other considerations (see, for example, ref. 91). Objections to selection and alternative hypotheses (in particular the biased gene conversion) (see ref. 84) for the formation and maintenance of the GC-rich isochores of mammals and birds have been discussed in detail elsewhere and shown not to hold (31, 32).

In conclusion, selection appears to account for both the formation and the maintenance of isochores, two phenomena that cannot be explained by the neutral theory, which, incidentally, was elaborated without any knowledge of isochore patterns. In fact, comparisons of orthologous mammalian genes or proteins (such as those used at that time) unfortunately missed not only the compositional heterogeneity of mammalian genes within the same genome, but also the compositional changes that took place between cold- and warm-blooded vertebrates. However, whereas a random fixation of neutral mutants (or a biased genome conversion) could not generate GC-rich isochores and would erode the isochore pattern over evolutionary time, the majority of mutations *per se* could only be neutral and nearly neutral. Indeed, in genomes in which noncoding sequences may represent 99% of DNA, no single-nucleotide change could have a strong enough selective advantage or disadvantage, at least in the vast majority of cases. There was, therefore, a need to reconcile the neutrality or near-neutrality of the point mutation process with selection at the regional level.

Neoselectionist Theory of Evolution

Let us consider now the conservative mode and follow the evolution of a gene-dense, GC-rich region from a warm-blooded vertebrate. Initially, the region is assumed to be at a compositionally acceptable GC level (Fig. 5). Because of the local accumulation of AT-biased mutations, some regions will drift toward lower GC values over evolutionary time. This ratchet process goes on until a certain GC threshold is trespassed (we called "critical changes" the changes responsible for trespassing the threshold). This leads to a regional change in chromatin structure that can spread (see refs. 92 and 93 for the spreading of silent chromatin) and interfere with DNA replication. Needless to say, the regional changes in chromatin structure would also have deleterious effects on the expression of genes in the region (see ref. 94 for chromatin structure control on gene expression and ref. 95 for insulators). The end result of this "epigenomic process" is negative selection on the carriers of such deleterious changes, the survivors showing a conserved isochore pattern. Interestingly, the proposed selection process does not create an impossible "genetic load" (96), as is the case for the selectionist theory of the neo-Darwinians, and does not substantially increase the "cost of natural selection" (97) because of its low frequency. Incidentally, the development of the model of Fig. 5 and some pioneering views along similar lines (91, 98) were described elsewhere (32).

To sum up, the essential points of the model are the deleterious chromatin changes, and the negative selection of the carriers of such changes. At present, the best explanations for initiating the regional chromatin changes are the compositional drift due to the AT bias and the trespassing of a lower GC threshold. It is possible, however, to think about other ways that allow one to reach the crucial changes in chromatin structure.

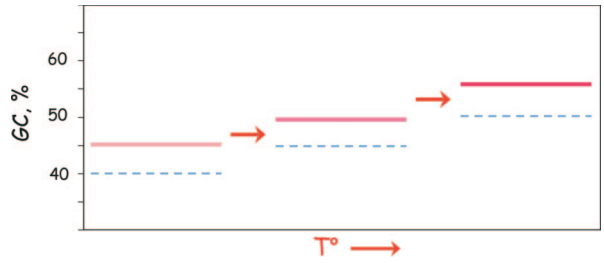

Fig. 6. A scheme of the transitional mode of evolution describing the GC increase of a gene-dense DNA region during the emergence of homeothermy. The basic feature is an increase in the GC level of the lower threshold (broken blue line) by a ratchet mechanism, which leads to an increased GC level of the region (pink-to-red lines).

For example, the insertion of a retroviral sequence or a mobile element might conceivably replace nucleotide substitutions as critical changes.

We will now consider in more detail some features of the model. (*i*) First, until the threshold is reached, the mutational process essentially consists of neutral or nearly neutral changes. (*ii*) After the threshold is trespassed and chromatin changes have occurred, the effects will be more or less deleterious according to the affected genome regions and the extent of the epigenomic changes. (*iii*) The lowered GC level of a region may be revealed by a comparison with a syntenic region from another species or another population, so providing a test for the theory. (*iv*) The time required for the actual disappearance of the carriers of epigenomic changes from the population will depend on the severity of the deleterious changes, and should be comparable with the time needed for the elimination of classical deleterious changes in coding and regulatory sequences located in the same regions. (*v*) The deleterious effects of chromatin alteration may also concern physiological chromatin remodeling. For instance, in the course of spermatogenesis, chromatin is drastically remodeled when histones are replaced by protamines, the reverse process occurring after fertilization. It is conceivable that regional compositional changes in DNA/chromatin may interfere with this process and cause gamete or zygote death.

The model of Fig. 5 also helps understand the "transitional mode" of genome evolution (Fig. 6) by positing (*i*) a GC increase of the threshold, which accompanies the increasing body temperature of vertebrates on the way to the higher body temperature of mammals and birds; and (*ii*) a ratchet mechanism, which only requires negative selection because of the deleterious chromatin changes associated with sequences trespassing the threshold. In this case, negative selection, in combination with an environmental factor (body temperature), far from simply maintaining the "status quo" (and being "boring") (99) can help genome adaptation (and be "creative") (100). Two points should be stressed here. The first is that the GC increase process just outlined took a very long time and comprised at least two major phases, involving the compositional transitions of amphibians to reptiles and of reptiles to mammals/birds, respectively. The second point is that the process took place at the level of isochores, the common basic structure of all vertebrate genomes.

We know, however, that positive selection can also play a role in the transitional mode by increasing GC levels according to the thermodynamic hypothesis. For example, nonsynonymous changes characterized by GC increases are predominant over synonymous changes in many coding sequences and lead to amino acid replacements that favor protein stability in *Gillichthys seta*, a fish living at temperatures up to 40°C, compared with *Gillichthys mirabilis*, a congeneric fish living at a lower temperature (G. Bucciarelli, D. Costagliola, and G.B., unpublished data).

In conclusion, the major novelties of the neoselectionist theory are the demonstrations (*i*) that the compositional compartmentalization, the compositional correlations between coding and noncoding sequences, and the transitional and conservative modes of evolution cannot be accounted for by an AT-biased point mutation process occurring at random locations (nor by a random gene conversion process, for that matter); (*ii*) that the structural and evolutionary features of the vertebrate genome can only be explained by a regional process; incidentally, the main reason why the neutralist/selectionist debate went on for so long probably was its being focused on the neutrality/nonneutrality of point mutations; and (*iii*) that such a regional process can only be visualized as due to an epigenomic event such as a change in chromatin structure, which interferes with replication and transcription; the coincidence between isochores and replicon clusters explains why the functional problems may have an isochore dimension.

Needless to say, invoking the participation of chromatin in genome evolution is a paradigm shift that justifies calling the neoselectionist model (32) a theory (101). It should be noted that, strictly speaking, the definition of neoselectionist theory applies only to the "selective mode" of Fig. 5. In this case, the combination of the neoselectionist theory and the nearly neutral theory of Ohta should be called the "neosynthetic theory" (the synthetic theory combined Darwinism and Mendelism). For the sake of simplicity, the definition of neoselectionist theory might be extended, however, to incorporate the nearly neutral theory of Ohta, as depicted in the bottom scheme of Fig. 1, where the critical changes are indicated as the most deleterious ones because they affect whole genome regions instead of individual nucleotides.

It should be stressed that the neutral theory and the neoselectionist theory proposed here lead to very different general views of evolution. Indeed, although Kimura's view that "increases and decreases in the mutant frequencies are due mainly to chance" led him, logically, to the idea of the survival of the luckiest, obviously the control of natural selection on the genome phenotype leads us back to the survival of the fittest. In fact, the neoselectionist theory is an extension of Darwin's theory, and could be defined an "ultra-Darwinian theory," in that even neutral and nearly neutral changes are eventually controlled by natural selection. Some general implications of the neoselectionist theory concerning the deterministic vs. the stochastic nature of the evolutionary process, the central dogma, and the genetic code will be dealt with elsewhere.

Finally, the neoselectionist theory not only accounts for the organization and evolution of the vertebrate (and eukaryotic) genome but also makes some predictions, the main one being that an organism or a gamete may be endowed with a different degree of "genomic fitness" depending on the different numbers and locations of altered chromatin regions, some of the latter being possibly responsible for "genomic (not genetic) diseases" in which gene expression is not necessarily affected by nucleotide changes in the coding or in the promoter sequences, but by changes in chromatin structure.

I thank Tomoko Ohta for important suggestions, Emile Zuckerkandl for innumerable discussions and encouragement, and Fernando Alvarez-Valin, Giacomo Bernardi, Oliver Clay, Bernardino Fantini, and Vittorio Sgaramella for very useful comments.

1. Darwin C (1859) *On the Origin of Species, a Facsimile of the First Edition with an Introduction by Ernst Mayr* (Harvard Univ Press, Cambridge, MA); reprinted (1964).
2. Wallace AR (1858) *J Proc Linn Soc (Zoology)* 3:53–62.
3. Fisher RA (1930) *The Genetical Theory of Natural Selection* (Clarendon, Oxford).
4. Haldane JBS (1932) *The Causes of Evolution* (Longmans, Green, and Co, London).
5. Wright S (1931) *Genetics* 16:97–159.
6. Mayr E (1963) *Animal Species and Evolution* (Belknap, Cambridge, MA).
7. Kimura M (1968) *Nature* 217:624–625.
8. Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge, UK).
9. King JL, Jukes TH (1969) *Science* 164:788–798.
10. Graur D, Li WH (2000) *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, MA), 2nd Ed.
11. Ohta T (1973) *Nature* 246:96–98.
12. Ohta T (1992) *Annu Rev Ecol Syst* 23:263–286.
13. Ohta T (2002) *Proc Nat Acad Sci USA* 99:16134–16137.
14. Bernardi G, Bernardi G (1986) *J Mol Evol* 24:1–11.
15. Corneo G, Ginelli E, Soave C, Bernardi G (1968) *Biochemistry* 7:4373–4379.
16. Filipski J, Thiery JP, Bernardi G (1973) *J Mol Biol* 80:177–197.
17. Thiery JP, Macaya G, Bernardi G (1976) *J Mol Biol* 108:219–235.
18. Macaya G, Thiery JP, Bernardi G (1976) *J Mol Biol* 108:237–254.
19. Costantini M, Clay O, Auletta F, Bernardi G (2006) *Genome Res* 16:536–541.

20. Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) *Science* 228:953–957.
21. Yunis JJ, Tsai MY, Willey AM (1977) in *Molecular Structure of Human Chromosomes*, ed Yunis JJ (Academic, New York).
22. Ma H, Samarabandu J, Devdhar RS, Acharya R, Cheng PC, Meng C, Berezney R (1998) *J Cell Biol* 143:1415–1425.
23. Costantini M, Clay O, Federico C, Saccone S, Auletta F, Bernardi G (2007) *Chromosoma* 116:29–40.
24. Caspersson TO (1989) *Am J Hum Genet* 44:441–451.
25. Bostock CJ, Sumner AT (1978) *The Eukaryotic Chromosome* (North–Holland, Amsterdam).
26. Mouchiroud D, D'Onofrio G, Aïssani B, Macaya G, Gautier C, Bernardi G (1991) *Gene* 100:181–187.
27. Zoubak S, Clay O, Bernardi G (1996) *Gene* 174:95–102.
28. Bernardi G (1993) in *Chromosomes Today*, eds Sumner AT, Chandley A (Chapman and Hall, London), Vol 11, pp 49–60.
29. Bernardi G (1995) *Annu Rev Genet* 29:445–476.
30. Bernardi G (2000) *Gene* 241:3–17.
31. Bernardi G (2001) *Gene* 276:3–13.
32. Bernardi G (2004) *Structural and Evolutionary Genomics, Natural Selection in Genome Evolution* (Elsevier, Amsterdam); reprinted (2005).
33. Watanabe Y, Fujiyama A, Ichiba Y, Hattori M, Yada T, Sakaki Y, Ikemura T (2002) *Hum Mol Genet* 11:13–21.
34. Eyre-Walker A, Hurst LD (2001) *Nat Rev Genet* 2:549–555.
35. D'Onofrio G, Bernardi G (1992) *Gene* 110:81–88.
36. Bernardi G (1990) *Biofutur* 94:43–46.
37. D'Onofrio G, Mouchiroud D, Aïssani B, Gautier C, Bernardi G (1991) *J Mol Evol* 32:504–510.
38. D'Onofrio G, Jabbari K, Musto H, Bernardi G (1999) *Gene* 238:3–14.
39. Chiusano ML, D'Onofrio G, Alvarez-Valin F, Jabbari K, Colonna G, Bernardi G (1999) *Gene* 238:23–31.
40. D'Onofrio G, Ghosh TC, Bernardi G (2002) *Gene* 300:179–187.
41. Cruveiller S, Jabbari K, Clay O, Bernardi G (2004) *Genome Res* 14:886–892.
42. Cruveiller S, Clay O, Jabbari K, Bernardi G (2007) *Proteomics* 7:361–363.
43. Ohno S (1972) *Brookhaven Symp Biol* 23:366–370.
44. Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988) *J Mol Evol* 28:7–18.
45. Aïssani B, Bernardi G (1991) *Gene* 106:173–183.
46. Aïssani B, Bernardi G (1991) *Gene* 106:185–195.
47. Hughes S, Clay O, Bernardi G (2002) *Gene* 295:323–329.
48. Bernardi G (1993) *Mol Biol Evol* 10:186–204.
49. Bernardi G (2000) *Gene* 259:31–43.
50. Cruveiller S, D'Onofrio G, Bernardi G (2000) *Gene* 261:71–83.
51. Bernardi G, Bernardi G (1990) *J Mol Evol* 31:265–281.
52. Bernardi G, Bernardi G (1990) *J Mol Evol* 31:282–293.
53. Bernardi G, Hughes S, Mouchiroud D (1997) *J Mol Evol* 44:S41–S51.
54. Gojobori T, Li WH, Graur D (1982) *J Mol Evol* 18:360–369.
55. Kunkel TA, Alexander PS (1986) *J Biol Chem* 261:160–166.
56. Eyre-Walker A (1999) *Genetics* 152:675–683.
57. Smith NG, Eyre-Walker A (2001) *Mol Biol Evol* 18:982–986.
58. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.* (2001) *Nature* 409:860–921.
59. Alvarez-Valin F, Lamolle G, Bernardi G (2002) *Gene* 300:161–168.
60. Arndt PF, Petrov DA, Hwa T (2003) *Mol Biol Evol* 20:1887–1896.
61. Arndt PF, Hwa T, Petrov DA (2005) *J Mol Evol* 60:748–763.
62. Santini S, Bernardi G (2005) *Gene* 346:51–61.
63. Lipatov M, Arndt PF, Hwa T, Petrov DA (2006) *J Mol Evol* 62:168–175.
64. Treffers HP, Spinelli V, Belser NO (1954) *Proc Natl Acad Sci USA* 40:1064–1071.
65. Cox EC, Yanofsky C (1967) *Proc Natl Acad Sci USA* 58:1895–1902.
66. Xia X, Xie Z, Kjer KM (2003) *Syst Biol* 52:283–295.
67. Wang HC, Xia X, Hickey D (2006) *J Mol Evol* 63:120–126.
68. Argos P, Rossmann MG, Grau UM, Zuber A, Franck G, Tratschin JD (1979) *Biochemistry* 18:5698–5703.
69. Zuber H (1981) in *Structural and Functional Aspects of Enzyme Catalysis*, eds Eggerer H, Huber R (Springer, Berlin), pp 114–127.
70. Nishio Y, Nakamura Y, Kawarabayasi Y, Usuda Y, Kimura E, Sugimoto S, Matsui K, Yamagishi A, Kikuchi H, Ikeo K, *et al.* (2003) *Genome Res* 13:1572–1579.
71. Nishio Y, Nakamura Y, Usuda Y, Sugimoto S, Matsui K, Kawarabayasi Y, Kikuchi H, Gojobori T, Ikeo K (2004) *Mol Biol Evol* 21:1683–1691.
72. Cacciò S, Jabbari K, Matassi G, Guermonprez F, Desgrès J, Bernardi G (1997) *Gene* 205:119–124.
73. Jabbari K, Cacciò S, Païs de Barros JP, Desgrès J, Bernardi G (1997) *Gene* 205:109–118.
74. Jabbari K, Clay O, Bernardi G (2003) *Gene* 317:161–163.
75. Jabbari K, Bernardi G (2004) *Gene* 333:179–181.
76. Varriale A, Bernardi G (2006) *Gene* 385:111–121.
77. Varriale A, Bernardi G (2006) *Gene* 385:122–127.
78. Jabbari K, Bernardi G (2000) *Gene* 247:287–292.
79. Jabbari K, Bernardi G (2004) *Gene* 333:183–186.
80. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G (2004) *FEBS Lett* 573:73–77.
81. Saccone S, Federico C, Bernardi G (2002) *Gene* 300:169–178.
82. Alvarez-Valin F, Clay O, Cruveiller S, Bernardi G (2004) *Mol Phylogenet Evol* 31:788–793.
83. Antezana MA (2005) *J Mol Evol* 61:834–836.
84. Duret L, Eyre-Walker A, Galtier N (2006) *Gene* 385:71–74.
85. Gu J, Li WH (2006) *Gene* 385:50–56.
86. Wu CI, Li W (1985) *Proc Natl Acad Sci USA* 82:1741–1745.
87. Gu X, Li WH (1992) *Mol Phylogenet Evol* 1:211–214.
88. van Rheede T, Bastiaans T, Boone DN, Hedges SB, de Jong WW, Madsen O (2006) *Mol Biol Evol* 23:587–597.
89. Fleishman SJ, Dagan T, Graur D (2003) *Mol Biol Evol* 20:1876–1880.
90. Gillespie JH (1991) *The Causes of Molecular Evolution* (Oxford Univ Press, New York).
91. Zuckerkandl E (1986) *J Mol Evol* 24:12–27.
92. Talbert PB, Henikoff S (2006) *Nat Rev Genet* 7:793–803.
93. Zuckerkandl E, Cavalli G (2007) *Gene* 390:232–242.
94. Felsenfeld G, Groudine M (2003) *Nature* 421:448–453.
95. Gaszner M, Felsenfeld G (2006) *Nat Rev Genet* 7:703–713.
96. Muller HJ (1950) *Am J Hum Genet* 2:111–176.
97. Haldane JBS (1957) *J Genet* 55:511–524.
98. Zuckerkandl E (1975) *J Mol Evol* 7:1–57.
99. Kondrashov AS (2000) *Trends Genet* 16:580.
100. Schmalhausen II (1949) *Factors of Evolution, the Theory of Stabilizing Selection* (Blakiston, Philadelphia).
101. Honderich T (1995) *The Oxford Companion to Philosophy* (Oxford Univ Press, Oxford).
102. Ohta T (1972) *J Mol Evol* 1:305–314.
103. Bromham L, Penny D (2003) *Nat Rev Genet* 4:216–224.