Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 31 (2004) 788–793
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Short Communication

Inaccurate reconstruction of ancestral GC levels creates a "vanishing isochores" effect☆

Fernando Alvarez-Valin,[a] Oliver Clay,[b] Stéphane Cruveiller,[b,c] and Giorgio Bernardi[b,*]

[a] *Secciòn Biomatemàtica, Facultad de Ciencias, Montevideo 11400, Uruguay*
[b] *Laboratory of Molecular Evolution, Stazione Zoologica "Anton Dohrn," Villa Comunale, 80121 Napoli, Italy*
[c] *Atelier de Génomique Comparative, Génoscope, Centre National de Séquençage, 2 rue Gaston Cremieux, CP 5706, 91057 Evry Cedex, France*

Received 9 August 2003; revised 19 January 2004

Abstract

It has recently been proposed, based on an analysis of orthologous gene sequences from closely related species, that GC-rich regions of primate and cetartiodactyl genomes are becoming GC-poorer, i.e., that GC-rich isochores are now vanishing in these lineages. We review an artefact of parsimony-based ancestor reconstruction in GC-rich DNA, and show that its magnitude explains the apparent vanishing of the GC-richest regions in cetartiodactyls, even if they are in fact at compositional equilibrium. The presently available data do not allow the disequilibrium hypothesis to be entirely ruled out in primates, yet, as we argue here, second-order artefacts can accumulate. They are therefore likely to explain many if not all of the observations, rendering unnecessary the general hypothesis of vanishing GC-rich isochores in mammals.

1. Introduction

Three recent papers address, and report evidence for, a new hypothesis of "vanishing isochores" in mammals. According to these papers, long regions of DNA (isochores) in primates, cetartiodactyls, and rodents are now simultaneously losing their GC base pairs.

The reasoning adopted in the three papers (Duret et al., 2002; Smith et al., 2002; Webster et al., 2003) exemplifies an important effect, or methodological artefact, that is often downplayed in routine reconstructions of ancestral states in GC-rich DNA. The artefact elegantly shows the limitations of sequence-based reconstructions, even where the sequences come from recently diverged species.

In all three papers, the authors reach or accept a conclusion that the GC-richest regions of primate genomes are disappearing, i.e., that the GC-rich regions are undergoing a decrease in GC level. In one of the papers, the same effect is reported also for artiodactyls. The papers imply that if current substitution patterns persist, they will eventually lead to lower GC levels, or erode GC, in these regions and in the third positions of the genes that the regions contain: we will refer to this as the "vanishing GC-rich isochore" hypothesis. It corresponds to one, but only one, possible interpretation of results presented earlier by Lander et al. (2001), in which GC levels of repetitive elements were found to be lower in GC-rich human DNA than in their respective consensi. Lander et al. (2001), Bernardi (2001), and Alvarez-Valin et al. (2002) discuss an alternative interpretation of the same results, namely a genome-wide mutational (not substitutional) bias towards AT, which appears more congruent with other observations on mammalian genomes.

We have reanalyzed the DNA sequence data used in the study of Duret et al. (2002), but we do not reach their conclusions. In particular, our results do not suggest that GC-rich regions are vanishing in cetartiodactyl genomes: the inferred substitution bias towards AT is no more than expected from parsimony artefacts under the null hypothesis of compositional equilibrium. This can be shown already by using a conservative, first-order approximation of the expected bias.

☆ Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2004.01.016.

* Corresponding author. Fax: +39-081245-5807.
E-mail address: bernardi@szn.it (G. Bernardi).

1055-7903/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2004.01.016

2. Results and discussion

2.1. First-order calculation of expected equilibrium $(GC \rightarrow AT)/(AT \rightarrow GC)$ ratios inferred by parsimony

All three papers under consideration (Duret et al., 2002; Smith et al., 2002; Webster et al., 2003) use parsimony to infer GC levels in ancestors from aligned orthologous DNA sequences. It has been cautioned before, backed by analytical calculations (Eyre-Walker, 1998) and simulations (Galtier and Gouy, 1998), that the use of GC-rich sequences to infer ancestral sequences via parsimony, or via naive maximum likelihood scenarios, will lead to overestimates of ancestral GC levels, as an artefact. This artefact is relevant, for example, in genes located in GC-rich regions of mammalian genomes, where GC levels of third codon positions (GC_3) can easily attain values between 75 and 100%. In this case, ancestral GC information will be quickly lost. In fact, a simple caricature of substitutions under perfect (i.e., unrealistically strict) compositional conservation can be stimulated by swapping red (GC) and black (AT) cards within aligned sequences of playing cards: apparent erosion of GC can then be inferred by parsimony yet it is obviously an artefact since card swaps within a sequence cannot change that sequence's GC level (number of red cards). In general, information loss during card shuffling can proceed quickly, as in a phase transition (Diaconis, 1996).

The three papers mention the parsimony artefact. The authors are well aware that, in order to demonstrate non-equilibrium, the substitution ratios (inferred $GC \rightarrow AT$)/(inferred $AT \rightarrow GC$) will have to significantly exceed the expected, and quite dramatic, artefact going in the same direction. The authors imply that they have taken care of this problem: Duret et al. (2002) and Webster et al. (2003) present different calculations suggesting that there is no need for concern. It is the purpose of this technical note to question this conclusion and, as a consequence, the general vanishing of GC-rich isochores in mammals.

It is clear that the whole discussion hinges on a technical point: are the GC-rich (apparently orthologous) sequences—coding sequences in Duret et al. (2002), also non-coding sequences in Smith et al. (2002) and Webster et al. (2003)—close to each other at all sites analyzed? More precisely, has enough information been conserved, in each region of each sequence, that one can confidently conclude the presence of higher GC levels in the past? In other words, are we sure that the relevant substitution rates have not been underestimated? Local erosion of information must be carefully quantified. If unchecked, such loss of information will masquerade as apparent erosion of GC.

The expected proportion of substitutions that parsimony infers to be $GC \rightarrow AT$ rather than $AT \rightarrow GC$, when there is true compositional equilibrium among three species (one outgroup and two ingroup species), rises immediately and steeply from its initial, correct value 0.5 at time 0 (i.e., distance 0) to its final value f, the present-day GC level of the region. At high GC levels, parsimony therefore reports a gross overestimate of the true proportion, which should always remain at 0.5 if there is compositional equilibrium. Correspondingly, the apparent substitution ratio $(GC \rightarrow AT)/(AT \rightarrow GC)$ rises steeply from 1 to $f(1-f)$. This can be seen by deriving exact calculations for different scenarios (Eyre-Walker, 1998) or by simulations (F. Alvarez-Valin, unpublished results). Both approaches show that the parsimony artefact does not necessarily diminish, for example, when four instead of three sequences are aligned, or when the ingroup distance is much smaller than the outgroup distance: a distance ratio far from 1 can actually exacerbate the effect. When the sequences are in disequilibrium, parsimony again yields quantitatively wrong results. An accurate estimate of the distance between the species, at the sites under examination (e.g., third or synonymous positions showing a GC/AT difference), is therefore crucial when calculating the magnitude of the parsimony artefact.

We reanalyzed the GC-rich cetartiodactyl and primate data of Duret et al., keeping track of the expected artefact due to parsimony. We found no evidence for a genome-wide decrease in GC-rich DNA in cetartiodactyls, even when we assumed that substitution rates do not vary within the genes (an approximation that tends to systematically underestimate the expected artefact) and that all alignments are perfect. In primates, this first-order approximation alone could not resolve the issue. We did not reanalyze the published rodent data, since murids and other myomorphs have indeed undergone a well-studied compositional shift, during which GC-rich DNA has been eroded (see, e.g., Douady et al., 2000; and references therein) We question the new hypothesis that a similar erosion affected, and still affects, the GC-rich DNA of the primate and cetartiodactyl lineages.

In carefully aligned, available sequences of the GC-richest class (average $GC_3 > 75\%$; supplementary table of Duret et al., 2002), relatively few sites differ in GC between the two ingroup species, i.e., are G or C in one ingroup species and A or T in the other ingroup species. Duret et al. found and analyzed 212 sites of this type for cetartiodactyls, and 47 for primates. We followed their protocol, as we understood it, and summarize the results in Table 1: parsimony gives the impression that many more sites underwent a substitution from GC to AT than from AT to GC. The average ratios reported by Duret et al. (3.2 in cetartiodactyls, 3.7 in primates) seem impressive, yet Table 1 also shows that the apparent ratio for cetartiodactyls is just what one would expect, to a first-order (conservative) approximation, from the parsimony artefact.

Table 1
First-order analysis of 25 very GC_3-rich genes in cetartiodactyls[a]

Outgroup	Sequence name		Distance 1	Distance 2	Species average GC_3	A	B	C	Expected ratio GC → AT/AT → GC
	Ingroup 1	Ingroup 2							
AF064555.PE1	BTPPI.PE1	OANIGFII1.PE1	0.115	0.336	92.7	0	2	2	7.608
AF181964	BT39469	Y13958	0.114	0.487	90.5	0	5	11	7.187
SSD158	BTCYB561.PE1	OAD157	0.032	0.148	90.2	0	1	3	2.577
AB038652.PMCT7	BTTRYPTMR.PE1	OAR18224.PE1	0.173	0.584	88.6	3	4	2	6.152
SSPROSDSN.PE1	BTAB4647.PE1	OAR133642.PE1	0.136	0.424	88.4	0	3	4	4.774
SSOXTRA.PE1	AF101724.PE1	OSOXYTREC.PE1	0.050	0.398	87.6	2	8	4	4.310
AF120326.PE1	AF074854	S44612.PE1	0.063	0.174	87.0	4	4	4	2.184
SSBLACMR.PE1	BTLGB.PE1	OALGB.PE1	0.087	0.316	85.0	3	3	2	2.853
SSCNP.CNP	BTCNP1.PE1	AF037467.CNP	0.055	0.207	84.7	0	3	0	2.146
SS12574.MYOD	AF093675	OAMYOD1.PE1	0.073	0.473	84.2	0	3	1	3.555
AF159382	S82652.PE1	AF034842	0.061	0.322	84.2	0	2	0	2.760
SSU59924.NOS	BTNOS.PE1	AF223471	0.108	0.326	84.1	6	4	7	2.698
U68482.G-CSF	AF092533.GCSF	OOCSFGR	0.106	0.225	83.0	0	5	2	2.000
SS14406.PE1	AF177290	OAPPCHY.PE1	0.073	0.241	82.9	2	5	4	2.119
SSGLUTP.PE1	BTGLUTI.GLUT-I	OAU89029.GLUT-1	0.100	0.247	82.3	7	11	8	2.056
U66254.OB	BT43943	OAU84247	0.073	0.215	81.1	0	3	2	1.834
AF064077	BTACTHA.PE1	AF116874.PE1	0.119	0.362	78.6	1	2	1	2.127
SSA005521	BTBRRIBO.PE1	S81745	0.047	0.384	78.4	0	2	0	2.225
SSTNFAB.PE2	AF011926.TNFA	OATNFA.PE1	0.037	0.339	78.4	1	1	1	2.099
SSJ001201.PE1	BTEP3B	AF035417	0.065	0.236	77.5	0	4	0	1.689
SS53020.PE1	BTY17260.STAR	S80098	0.089	0.224	77.1	2	3	0	1.621
SSIFNA1.PE1	BTIFNAA.PE1	OVU77908.PE1	0.078	0.332	76.7	4	1	0	1.906
SSMOTSA.PE1	AF068196.PE1	AF022771.PE1	0.042	0.249	76.5	0	2	0	1.692
SSINTL10A.IL-10	BT799.PE1	OA11421.IL-10	0.090	0.270	75.4	3	0	0	1.658
S96211.PE1	BTTIM.PE1	S67450.TIMP-1	0.069	0.236	75.2	0	5	0	1.577
						38	86	58	
			0.082	**0.310**	**82.8**				**3.22**

[a] Genes having a minimum $GC_3 > 75\%$ and represented by sequences in at least three cetartiodactyl species are shown in order of decreasing GC_3. The distance between the ingroup species (Distance 1) and the average distance between the ingroup species and the outgroup (Distance 2) were estimated by maximum likelihood using PAML (Yang, 2002). The table reports the GC_3 level of each gene in each species (indicated by its EMBL/GenBank name with ACNUC extension; retrieved from http://pbil.univ-lyon1.fr) as well as the mean GC_3 level across the three species. The expected ratio of (GC → AT)/(AT → GC) substitutions, at compositional equilibrium, was determined by the formulae of Eyre-Walker (1998). In calculating this expected value, we have, however, followed tradition in assuming constant substitution rates along a gene, and at CpG and non-CpG sites, although this first-order approximation tends to underestimate the expected value (see text). Genes and in-/outgroup species were taken from the supplementary table of Duret et al. (2002). A|B|C, number of synonymous substitutions from A or T to G or C | from G or C to A or T at non-CpG sites | from G or C to A or T at CpG sites. Bold figures in the bottom row are column averages. To calculate the average of the expected ratios, each gene is weighted by the number of substitutions that it contributes to the total $(\sum s_i \mathrm{Exp}(ratio_i))/(\sum s_i)$. Here, $\mathrm{Exp}(ratio_i)$ is the expected ratio for gene i, and s_i is the total number of substitutions in gene i for which parsimony inferred a direction GC → AT or AT → GC. The corresponding table for 13 very GC-rich primate genes is given as Web Table 1A in the online supplement.

Analysis of the GC_3 intermediate class ($57\% < GC_3 < 75\%$) leads to a similar observation: the excess GC → AT substitutions can readily be explained by the parsimony artefact (data not shown).

As mentioned before, even though expected ratios of (GC → AT)/(AT → GC) substitutions were calculated in Duret et al. (2002), they found figures that differ markedly from the observed ratio and from the expected ratios that we present in this paper. A possible answer for the discrepancy between our estimates and those obtained by Duret et al. is the very different values we obtained for the synonymous distances. Where as our calculations give an average distance between ingroup and outgroup is 0.31 substitutions per site for the GC_3-rich cetartiodactyl data set shown in Table 1, Duret et al. found it to be only 0.11 substitutions per site, i.e., roughly three times smaller. We observed a similar yet less pronounced difference for primates: the average distance between the ingroup and the outgroup in the GC_3-rich primate data set to be 0.11 substitutions per site (web Table 1A) while Duret et al. found it to be 0.07 substitutions per site. The discrepancies cannot be attributed to the method that we used (Maximum Likelihood using three parameters to estimate codon frequencies, Yang, 2002) for estimating synonymous distances, because other methods such as those of Nei and Gojobori (1986), Li (1993), and especially Bielawski et al. (2000) all yielded distances that were about three times as high as those presented by Duret et al., or higher, in the case of cetartiodactyls (values not shown here). The point is that if the distances are underestimated, so will be the expected ratio of (GC → AT)/(AT → GC) substitutions.

2.2. Second-order artefacts accumulate instead of cancelling out, and may explain the apparent "vanishing GC" effect also in primates

The ratios between inferred synonymous GC → AT and AT → GC substitutions in very GC_3-rich cetartiodactyl genes correspond to the first-order expectations at equilibrium, if the parsimony artefact is taken into account. Indeed, the observed ratio for this group of sequences is around 3.2 (according to Duret et al. or our own analysis), which is just the expected ratio (Table 1). In primates we did not find such agreement (first-order expectation 1.8, observed ratio 3.7), yet as we will argue, second-order artefacts can accumulate to give an expected ratio, at equilibrium, that comes close to the observed ratio. Any still remaining difference between observed and expected values is likely to be small, and its significance would need to be tested.

Noise from alignment and other errors is systematically biased in the same direction as the parsimony artefact. In other words, when there is noise in GC-rich sequences, GC → AT substitutions are erroneously inferred more frequently than AT → GC substitutions. This means that errors or artefacts from different sources will tend to accumulate, instead of cancelling each other. The principle can be illustrated by considering a correct outgroup site that is replaced by another, randomly chosen site from the same gene in the same outgroup species (a scenario that could correspond, for example, to a tiny mistake in an alignment). If a third-position site is informative for the question at hand, it will have, say, an AT base pair in chimp and a GC base pair in human. It is important to realize that, by construction, the ingroup's informative sites will always have 50% GC when there is compositional equilibrium, regardless of the GC or GC_3 of the entire sequence. By contrast, a random outgroup site will not. Instead, a random outgroup site will tend to be GC rather than AT (with probability equal to the GC of the entire coding sequence). The tendency is, therefore, that the site will appear to have experienced a GC → AT substitution, even when it did not. Sensitivity to noise was not included in previous calculations or simulations (Eyre-Walker, 1998; Galtier and Gouy, 1998) of the parsimony artefact. Where noise is present, the apparent (GC → AT)/(AT → GC) ratios expected at equilibrium, as calculated to first-order accuracy in those earlier studies or in Table 1, will therefore tend to be underestimates.

A number of problems will thus influence inferred (GC → AT)/(AT → GC) ratios. Any one of them, in isolation, may not be able to explain the high primate substitution ratios inferred by parsimony. On the other hand, their cumulative effect is likely to be large. The first problem is the well-characterized artefact, described above, which allows a first-order calculation of expected ratios when one properly aligns sequences of bona fide orthologous genes that have evolved under essentially identical conditions (no transpositions, no relaxation of selective pressures by the appearance of paralogs in one species but not in the others, etc.), and in which substitution rates are the same in all parts of the gene.

A second problem is that not all the alignments may be reliable. One of the genes belonging to the GC-richest primate sequences is for involucrin (M13903 in human; see legend to Table 1), an extremely repetitive protein for which the alignment is riddled with indels and therefore open to question, but this gene appears to be an exception in the data set of Duret et al. Alignment uncertainties may, however, affect the alignments of contig DNA described by Smith et al. (2002) and Webster et al. (2003), which were not available in print or on a web site and which we did not analyze. We only point out the well-known difficulties in aligning non-coding DNA, and that occasional, accidentally misaligned nucleotides may be inevitable, since the total length of aligned DNA analyzed by these authors is very large (1.8 Mb). If the yield of informative sites is small (due to short evolutionary distances between human and chimp, and/or ambiguous alignments in some regions), even rare misaligned sites could represent a non-negligible proportion of the total counts.

A third problem is hypermutability at CpG sites. Again, such sites may not, alone, explain the observed counts, but they could contribute. CpG sites preferentially undergo deamination; as a result, synonymous substitution rates or multiple-hit rates are higher in CpG sites than elsewhere, and the time to saturation or complete loss of information can be short. On the other hand, where a CpG has hypermutated via deamination to a TpG or CpA in two (rather than only one) of three lineages, parsimony could erroneously infer AT → GC instead of GC → AT. The expected ratios of the inferred (GC → AT)/(AT → GC) substitutions are therefore not easily calculated at sites containing CpGs. Although Duret et al. give counts showing that the number of informative synonymous sites involving CpGs is, in general, not high (with one or two exceptions in the artiodactyl and primate data; see Table 1), the traces of some ancestral CpG sites may have been eroded beyond recognition, it is difficult to speculate on the magnitude of the CpG effect.

A fourth problem is that in at least some genes, not all (non-CpG) synonymous positions in the gene evolve at the same rate. Such intragenic variability in synonymous rates has been repeatedly reported in mammalian genes (Alvarez-Valin et al., 1998; Hurst and Pal, 2001). Thus, apparently not all of the synonymous sites can be freely "chosen" for substitutions, i.e., the effective substitution rate is higher than the overall rate: one should actually be dividing by the effective number of available sites, not by the total number of sites. In genes exhibiting such intragenic variability, rates calculated

routinely to first-order accuracy are underestimates and yield underestimates of the ratio in which we are interested. Indeed, the expected (GC → AT)/(AT → GC) ratio, as inferred by parsimony, increases dramatically as a function of the rate.

A fifth potential problem, which does not appear to affect the GC-richest primate sequences of Duret et al., is that paralogs can occasionally masquerade as orthologs: if the paralogs differ in GC, then one would not be monitoring substitutions but simply inter-paralog differences within a genome. A similar problem could arise, for example, if a GC-rich gene were translocated into a GC-poor isochore in one of two species (GC-poor DNA being much more frequent than GC-rich DNA in mammalian genomes). In that case, one would not be monitoring the vanishing GC of an isochore, but simply an intragenic change of third positions in a gene following its transfer to a different isochore.

Although we did not re-analyze the 1.8 Mb of DNA discussed by Webster et al. (2003, Table 2) we noted an averaging operation that these authors applied when they corrected for the parsimony artefact. The values they observe in the GC-richest primate DNA, namely 67 inferred GC → AT substitutions versus 24 inferred AT → GC substitutions, are reported to have remained almost unchanged after the correction: 67.7 and 23.3, despite the high GC and the high substitution ratio (2.8–2.9). As we have seen, the artefact typically changes the true ratio much more than that, in GC-rich DNA at equilibrium, even if CpGs were excluded (the text mentions that substitutions putatively due to CpG hypermutation were identified). The surprisingly small corrections were obtained when the authors "derived an ML estimate of the numbers of GC → AT and AT → GC substitutions by weighting the frequency of the different [human/chimp/baboon] site patterns by the relative probabilities of the two most likely [human/chimp/ancestor] states." The relative probabilities were calculated using formulae derived from Galtier and Gouy (1998), but apparently as bulk averages, whereas the relative probabilities that would be relevant for the claim under discussion are those found in GC-rich regions, preferably after partitioning the DNA according to known GC or rate differences (bona fide intergenic DNA, first/second/third codon positions of exons, introns, etc.) Bulk averaging over the authors' full sample of alignable primate DNA, of which the majority is likely to be GC-poor (if the sample was biased), would dilute out such potentially important information.

To our knowledge, no significant genome-wide decrease in GC of GC-rich regions or genic silent positions has yet been convincingly demonstrated in primates or cetartiodactyls, although we acknowledge that the evidence in the former case is stronger. We would like to stress that the hypothesis of decreasing GC content can be true for some mammalian groups, such as murids, and perhaps even primates. But as we show here, this hypothesis has no support in cetartiodactyls. As a consequence, the suggestion made by Duret et al. (2002) that this process of GC erosion started before the radiation of mammals, and therefore that it should affect most or all mammalian orders, does not yet have support. Until the idea of "vanishing GC-rich isochores," i.e., compositional disequilibrium, can be backed by a rigorous analysis involving pertinent observed/expected calculations, it is likely to remain an interesting yet speculative hypothesis.

3. Note added in proof

By aligning copies of mammalian repetitive elements, Arndt, Petrov and Hwa (Mol. Biol. Evol. 20, 1887–1896) have recently offered additional evidence for an AT-substitution bias within such elements. Taking into account that these repetitive sequences are assumed to be neutral, the new results are an additional proof that the mutation pattern is AT-biased throughout the genome. It will be interesting to investigate whether similar substitution biases apply to single-copy intergenic and genic DNA (particularly at synonymous positions).

References

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

Alvarez-Valin, F., Lamolle, G., Bernardi, G., 2002. Isochores, GC_3 and mutation biases in the human genome. Gene 300, 161–168.

Bernardi, G., 2001. Misunderstandings about isochores. Part I. Gene 276, 3–13.

Bielawski, J.P., Dunn, K.A., Yang, Z., 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. Genetics 156, 1299–1308.

Diaconis, P., 1996. The cutoff phenomenon in finite Markov chains. Proc. Natl. Acad. Sci. USA 93, 1659–1664.

Douady, C., Carels, N., Clay, O., Catzeflis, F., Bernardi, G., 2000. Diversity and phylogenetic implications of CsCl profiles from rodent DNAs. Mol. Phylogenet. Evol. 17, 219–230.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., Galtier, N., 2002. Vanishing GC-rich isochores in mammalian genomes. Genetics 162, 1837–1847.

Eyre-Walker, A., 1998. Problems with parsimony in sequences of biased base composition. J. Mol. Evol. 47, 686–690.

Galtier, N., Gouy, M., 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15, 871–879.

Hurst, L.D., Pal, C., 2001. Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet. 17, 62–65.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Li, W.-H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36, 96–99.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Smith, N., Webster, M., Ellegren, H., 2002. Deterministic mutation rate variation in the human genome. Genome Res. 12, 1350–1356.

Webster, M., Smith, N., Ellegren, H., 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. Mol. Biol. Evol. 20, 278–286.

Yang, Z., 2002. Phylogenetic Analysis by Maximum Likelihood (PAML), Version 3.12. University College London, England.