

# Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution

Nicolas Galtier<sup>1</sup> and Laurent Duret<sup>2</sup>

<sup>1</sup> CNRS UMR 5554 – Institut des Sciences de l'Evolution, Université Montpellier 2 – CC64, Place E. Bataillon, 34095 Montpellier Cedex, France

<sup>2</sup> Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS; Université Lyon 1, 43 Bd du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

**The analysis of evolutionary rates is a popular approach to characterizing the effect of natural selection at the molecular level. Sequences contributing to species adaptation are expected to evolve faster than nonfunctional sequences because favourable mutations have a higher fixation probability than neutral ones. Such an accelerated rate of evolution might be due to factors other than natural selection, in particular GC-biased gene conversion. This is true of neutral sequences, but also of constrained sequences, which can be illustrated using the mouse *Fxy* gene. Several criteria can discriminate between the natural selection and biased gene conversion models. These criteria suggest that the recently reported human accelerated regions are most likely the result of biased gene conversion. We argue that these regions, far from contributing to human adaptation, might represent the Achilles' heel of our genome.**

## Introduction

Having characterized the complete sequence of genomes, the next step is to decipher the functional information embedded within these sequences. This task is particularly arduous in large eukaryotic genomes, where chromosomes are replete with neutral sequences (i.e. functionless sequences that can be mutated without affecting the fitness of organisms). A popular approach to identifying functional regions is based on the analysis of evolutionary rates. Most functional sequences are subject to negative (purifying) selection (see Glossary) and hence can be recognized because they evolve less rapidly than neutral sequences. However, this approach only identifies functional elements that are conserved among species, whereas there is a major interest in detecting the elements that evolved toward new functions in some specific lineages and contributed to species adaptation. The evolution of such sites is driven by positive (adaptive) selection. In contrast to purifying selection, adaptation promotes the fixation of advantageous mutations, resulting in a higher substitution rate than with neutral evolution. This rationale has been widely applied to the analysis of protein-coding genes: adaptation is invoked

when the nonsynonymous substitution rate exceeds the (presumably neutral) synonymous one [1].

## The quest for targets of positive selection: human accelerated regions

Recently, Pollard *et al.* [2] proposed an elegant approach, relying on substitution rates to detect candidate targets of positive selection in non-coding sequences. They first sought 35 000 non-coding regions showing a high level of conservation in mammals, excluding humans. Among these, they discovered 49 regions [100 base pairs (bp) to 500 bp] in which the substitution rate had been significantly increased in humans. These regions were called human accelerated regions (HARs). The 118 bp HAR1, for instance, showed only two differences between chimpanzees and chickens (~310 million years of divergence) but 18 between chimpanzees and humans (~6 million years of divergence). The strong conservation of HARs among non-human vertebrates clearly indicates that they are under functional constraints. The sudden burst of substitutions in the human lineage cannot be explained simply by loss of function because the amount of divergence is much higher than expected for a neutral sequence – the average human–chimpanzee divergence in non-coding sequences is less than 2%. HARs are therefore good candidates for having evolved under positive selection in humans. Interestingly, the HAR1 region is part of a non-translated RNA gene (*HAR1F*). This gene is expressed during brain development, which suggests that it might have contributed to the evolution of human-specific brain features [2].

## Glossary

**Genetic hitch-hiking:** a process by which population selective effects affect loci genetically linked to the selected locus.

**Negative (purifying) selection:** removal of deleterious mutations from populations, resulting from the lower reproductive success of individuals carrying the mutations.

**Meiotic drive:** process by which one allele is over-represented in the gamete pool after meiosis.

**Positive (adaptive) selection:** fixation of advantageous mutations in populations, resulting from the higher reproductive success of individuals carrying the mutations.

**Selective sweep:** rapid fixation of a favourable allele through adaptive evolution, resulting in a sudden drop of genetic diversity at the corresponding locus.

Corresponding author: Galtier, N. (galtier@univ-montp2.fr).  
Available online 5 April 2007.

### Substitution hotspots: possible neutral explanations

Forces other than natural selection, however, can lead to an increased nucleotide substitution rate [3]. First, the rate of mutation can vary along genomes [4]. Such variations are observed at different scales [kilobases (kb) and megabases]. However, variations that have been reported so far are relatively limited (~2-fold increase) compared with the 20-fold increase in substitution rate in HAR1 [2]. To our knowledge, there is currently no direct evidence for strong variations in mutation rates at the 100 bp scale, although the existence of such mutation hotspots has not been excluded. The second possible explanation is biased gene conversion (BGC), a neutral process associated with a recombination by which an AT/GC heterozygote will produce a greater number of gametes carrying G or C than A or T, presumably through the GC-biased repair of A:C and G:T mismatches in heteroduplexed recombination intermediates. Several arguments support the notion that BGC has influenced GC-content evolution in mammalian genomes [5]. In short: (i) the base excision repair process is biased towards GC in mammals [6,7]; (ii) highly recombining chromosomes or chromosomal regions are becoming GC enriched [8–11]; (iii) duplicated genes undergoing frequent gene conversion between paralogues are GC enriched [12–14]; (iv) G and C allele frequencies are increased in single nucleotide polymorphisms located close to recombination hotspots [15]; (v) GC content is higher near recombination hotspots [16].

### How to distinguish selection from BGC?

From a population genetics point of view, the BGC meiotic drive is essentially equivalent to directional selection [17]. Under the BGC model, AT → GC mutations have a higher probability to be transmitted to the next generation, and eventually fixed, than is the case for other mutations (GC → AT, AT → TA or GC → CG). Similarly to an episode of adaptive evolution, an episode of BGC should therefore result in an increased substitution rate, especially during the non-equilibrium stage, when base composition has not yet reached its GC-enriched steady state [18]. Given that recombination occurs predominantly within hotspots (i.e. at the ~1 kb scale [19]), BGC is expected to result in substitution hotspots. Hence, the signature of BGC on

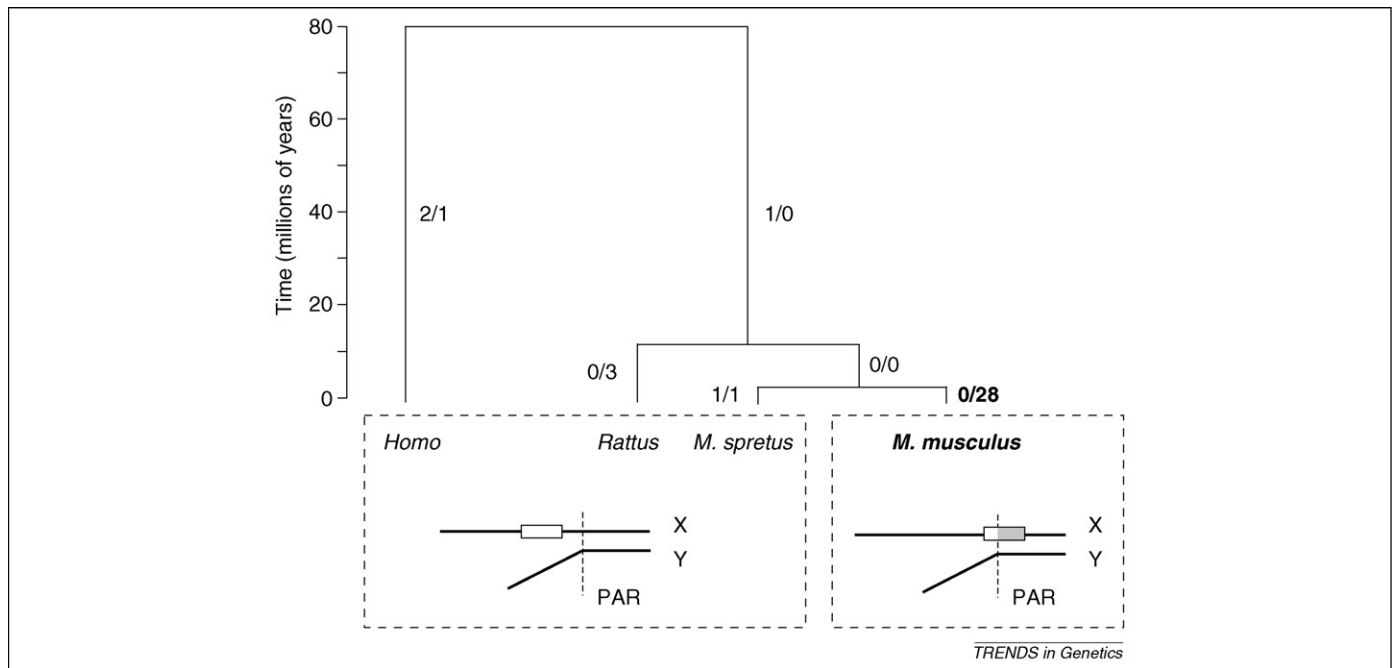
sequence evolution is similar to that of selection. There are, however, several features that might help to distinguish BGC from selection (Table 1). First, in the case of BGC, it is expected that the favoured alleles are always G or C, whereas when sequence evolution is driven by selection, there is *a priori* no reason why the advantageous alleles should systematically be G or C. Functional sequences are not typically GC rich; the average GC content in the second codon position of human protein-coding sequences is 43%, and the average GC content of the 35 000 conserved functional non-coding sequences identified by Pollard *et al.* [2] is 41%. Second, BGC is linked directly to the process of recombination. Hence, BGC would be expected to be more prevalent in regions of high recombination. Third, directional selection leads to an increased evolutionary rate only at functional sites, whereas BGC might affect any region of the genome, functional or not. Finally, directional selection affects the pattern of polymorphism at linked neutral sites, and, as noted by Pollard *et al.* [3], the width of the genomic window affected by such selective sweeps can be large, whereas the impact of BGC should be limited to the length of the conversion track (several hundred bp, on average).

### The BGC molecular drive might affect functional sequences

Most of the existing literature about BGC discusses its impact on neutral sequences. It should be stressed that functional sequences under selection can also be affected by BGC; if strong enough, BGC can theoretically overcome purifying selection and lead to the fixation of deleterious AT → GC mutations. A spectacular example of the influence of BGC on sequence evolution is provided by the *Fxy* (also known as *Mid1*) gene in the mouse. This gene is X-linked in the human, rat and short-tailed mouse *Mus spretus* but was recently translocated in the house mouse *Mus musculus*, in which it now overlaps the boundary between the X-specific region and the pseudoautosomal region (PAR); the 5' end of the gene (from exon 1–3) is located in the X-specific region, whereas the 3' end (exon 4–10) is within the PAR [20] (Figure 1). The PAR, a short segment of homology between the X and Y chromosomes, is intrinsically a highly recombining region [21]. By moving

**Table 1. Predictions of the different models that might explain substitution hotspots**

Criterion	Mutation hotspot	Model predictions		Test	Caveats and comments
		BGC	Adaptation		
Target sites	All	All	Functional sites only	Reject selection if nonfunctional sites are involved	The functional or nonfunctional status is not always obvious
Substitution pattern	Variable; AT biased in most eukaryotes	GC biased	No systematic bias	Reject selection in case of strong bias; reject BGC in case of AT bias	Selection can occasionally lead to a biased process
Relationship with recombination	Unclear	Strong	Weak	Favour BGC if substitution hotspots are concentrated in regions of high recombination	Recombination hotspots have a short lifetime
Selective sweep	No	Yes, but no hitch-hiking	Yes, with hitch-hiking	Reject mutation hotspots and BGC if coalescence-based neutrality tests applied to flanking regions are positive	Power is weak for old sweeps



**Figure 1.** Evolutionary history of *Fxy* in mammals. The *Fxy* gene was translocated into *M. musculus* from an X-linked position to a new position, in which it overlaps the pseudoautosomal boundary (inset boxes). The timescale is given in millions of years. For each branch, the numbers of amino acid changes that have occurred in the 5' and 3' ends of the gene, respectively, are given. A strong increase in amino acid substitution rate occurred in the *M. musculus* lineage for the translocated fragment only. For comparison, the estimated numbers of synonymous substitutions in the *Rattus*, *M. spretus* and *M. musculus* branches are 12, 2 and 1 (respectively) for the 5' end of the gene, and 20, 0 and 163 (respectively) for the 3' end. GenBank accession numbers: *Homo sapiens* (Y13667), *Rattus norvegicus* (AF186461), *M. spretus* (AF186460), *M. musculus* (Y14848).

from the X-specific region to the PAR, the 3' *Fxy* sequence therefore experienced a sudden increase in recombination rate. This translocation was followed by a dramatic increase in GC content and in substitution rate at third codon positions and in introns of the 3' end of *Fxy* [20,22]. This spectacular acceleration was not observed in the 5' end of the gene, which is still X-linked in *M. musculus*.

Let us now examine the consequences of this move at the protein level. The 667 amino acid *Fxy* protein is highly conserved between mammals (excluding *M. musculus*). The human and *M. spretus* sequences, which diverged ~80 million years ago, differ by just six amino acid changes. The house mouse sequence, however, has accumulated as many as 28 amino acid substitutions since the divergence with *M. spretus*, 1–3 million years ago. This corresponds to a >100-fold increase in amino acid substitution rate in the *M. musculus* lineage (Figure 1). All of these substitutions have occurred in the PAR-linked 3' end of the gene, and all of them were caused by AT → GC nucleotide changes. This elevated amino acid substitution rate obviously has nothing to do with adaptation; if directional selection had been acting on the protein sequence, then silent sites (third codon positions and introns) should have remained unaffected. One could speculate that this rapid evolution reflects selection for higher GC content, to increase gene stability [23] or expression level [24]. However, if such selective pressures were at work, they should have affected the whole gene, and not only its 3' end. Overall, the fact that this increased rate of evolution affects both coding and silent sites, involves only AT → GC changes and is restricted to the highly recombining part of the gene clearly demonstrates that it is the result of BGC and not of selection. The high rate of nonsynonymous

substitution in *M. musculus* compared with other mammals illustrates how BGC, when applying to functional sequences, can promote the fixation of otherwise counterselected AT → GC mutations.

#### HARs: a consequence of BGC?

The *Fxy* example demonstrates that BGC, similarly to adaptation, can result in a sudden increase in substitution rate in nonfunctional, but also in functional, regions, leading to a pattern similar to the HAR pattern. Furthermore, several aspects of the evolution of HARs seem to be consistent with the BGC model, as discussed by Pollard *et al.* [3]. First, the substitutions that have accumulated in the human lineage are mostly AT → GC changes (and, notably, the 18 substitutions observed in HAR1 are all AT → GC changes) [2,3]. This is expected under BGC but not under an adaptive model. It should be noted that HARs are involved in various functions; some of them are part of untranslated structured RNAs (e.g. HAR1), whereas others are probably DNA regulatory elements [2,3]. It was observed that the same GC bias affects HARs, irrespective of their transcriptional status; the proportion of AT → GC substitutions in untranscribed HARs (74%;  $n = 25$ ) is not significantly different from that in transcribed HARs (70%;  $n = 24$ ). Hence, selective pressures increasing RNA stability or expression level cannot explain the general GC bias observed in HARs. Second, HARs tend to be located in highly recombining regions [2,3]. Again, this is expected under BGC but not under a selective scenario. Third, there is no clear evidence of selective sweeps around the five HARs for which population data have been collected [2,3]. Fourth, as far as the well-characterized HAR1 is concerned, the region of rapid and

GC-biased substitution is not restricted to the 118 bp HAR1 conserved element but extends to >1 kb of flanking sequences.

One interesting feature of HAR1 is that most substitutions that have occurred in humans are pairs of compensatory changes, by which the A:T pair of an RNA stem evolves into G:C. Such changes, conservative with respect to the RNA structure, are usually not considered as adaptive when observed, for example, in ribosomal RNAs. Compensatory substitutions are, however, consistent with the hypothesis that HAR1 has been subject both to negative selective pressure (tending to preserve the RNA structure) and BGC (favouring the fixation of AT → GC mutations). Strong BGC can overcome purifying selection and lead to the fixation of, for example, a T → C deleterious mutation (resulting in a C:A bulge in the RNA stem). A compensatory A → G mutation occurring at the interacting site would then be favoured by both BGC and selection, and have a good chance to fix. In this example, BGC could therefore help to solve the main theoretical problem posed by compensatory substitutions in RNA (i.e. how the first of the two mutations can reach fixation even though it disrupts the Watson–Crick interaction and is therefore deleterious).

Why is this peculiar pattern of substitution not observed in the same regions of other mammals? It has been shown that, despite a strong sequence similarity between humans and chimpanzees, there is little or no conservation of the location of recombination hotspots in the two genomes [25,26]. This indicates that the location of hotspots changes rapidly during evolution. Hence, episodes of BGC would be expected to occur at independent locations in the genome of different species. In other words, HARs might correspond to functional regions under negative selective pressure that have at one time been submitted to strong BGC pressure because of the activity of recombination hotspots appearing in the human lineage after the human–chimpanzee split, ~6 million years ago. Given the short life span of recombination hotspots in primates, the location of HARs does not necessarily coincide with present recombination hotspots in the human genome.

### BGC substitution hotspots: genomic Achilles' heel

The  $F_{xy}$  case, similar in many aspects to that of the HARs, instantiates the theoretical notion that BGC can lead to lineage-specific increases in substitution rate in functional sequences in the absence of adaptation. Moreover, several features of the HARs seem to be more consistent with the BGC model than with selective scenarios. This does not imply that every HAR is explained solely by BGC. Adaptation might well have promoted some of the human-specific changes detected by Pollard and co-workers. To be confident that adaptation has been the driving force, however, one should be able to reject the BGC model, as well as the classical neutral model, thus extending the null hypothesis of molecular evolution. The alternative BGC explanation makes a substantial difference from a functional point of view. Under the BGC model, HARs, far from contributing to human adaptation, would represent weak points of our genome, whose function needed to be preserved, in spite of the 'undesired' fixation of numerous deleterious mutations.

To understand better the impact of BGC on genome evolution, it will be important to study recombination hotspots. For example, how long do these hotspots remain active? Are some hotspots more evolutionarily stable than others? Which factors determine the distribution of recombination hotspots along chromosomes? It would also be useful to have direct experimental evaluations of the BGC process; up until now, out of 15 crossover hotspots analysed in humans by sperm typing, at least two have shown evidence of meiotic drive [27]. This is a minimum estimate because only a small number of men have been studied. Hence, meiotic drive within human crossover hotspots seems to be a relatively common phenomenon [27]. However, more data would be needed to estimate precisely the frequency of recombination hotspots showing meiotic drive, and to assess the extent to which this process is biased towards GC alleles.

### Acknowledgements

We thank Katie Pollard for providing the sequences and human genomic locations of vertebrate conserved elements. This work was supported by the Agence Nationale de la Recherche and by the Centre National de la Recherche Scientifique. This is publication number ISEM 2007-023.

### References

- 1 Yang, Z. (2002) Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.* 12, 688–694
- 2 Pollard, K.S. *et al.* (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443, 167–172
- 3 Pollard, K.S. *et al.* (2006) Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* 2, e168
- 4 Ellegren, H. *et al.* (2003) Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* 13, 562–568
- 5 Duret, L. *et al.* (2006) A new perspective on isochore evolution. *Gene* 385, 71–74
- 6 Brown, T.C. and Jiricny, J. (1987) A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell* 50, 945–950
- 7 Bill, C.A. *et al.* (1998) Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese Hamster ovary cells. Competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. *Genetics* 149, 1935–1943
- 8 Galtier, N. *et al.* (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159, 907–911
- 9 Meunier, J. and Duret, L. (2004) Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* 21, 984–990
- 10 Webster, M.T. *et al.* (2005) Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. *Mol. Biol. Evol.* 22, 1468–1474
- 11 Webster, M.T. *et al.* (2006) Strong regional biases in nucleotide substitution in the chicken genome. *Mol. Biol. Evol.* 23, 1203–1216
- 12 Galtier, N. (2003) Gene conversion drives GC-content evolution in mammalian histones. *Trends Genet.* 19, 65–68
- 13 Kudla, G. *et al.* (2004) Gene conversion and GC-content evolution in mammalian Hsp70. *Mol. Biol. Evol.* 21, 1438–1444
- 14 Backstrom, N. *et al.* (2005) Gene conversion drives the evolution of HINTW, an ampliconic gene on the female-specific avian W chromosome. *Mol. Biol. Evol.* 22, 1992–1999
- 15 Spencer, C.C. (2006) Human polymorphism around recombination hotspots. *Biochem. Soc. Trans.* 34, 535–536
- 16 Spencer, C.C. *et al.* (2006) The influence of recombination on human genetic diversity. *PLoS Genet.* 2, e148
- 17 Nagylaki, T. (1983) Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. U. S. A.* 80, 5941–5945
- 18 Galtier, N. (2004) Recombination, GC-content and the human pseudoautosomal boundary paradox. *Trends Genet.* 20, 347–349
- 19 McVean, G.A. *et al.* (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304, 581–584
- 20 Perry, J. and Ashworth, A. (1999) Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* 9, 987–989

- 21 Soriano, P. *et al.* (1987) High rate of recombination and double crossovers in the mouse pseudoautosomal region during male meiosis. *Proc. Natl. Acad. Sci. U. S. A.* 84, 7218–7220
- 22 Montoya-Burgos, J.I. *et al.* (2003) Recombination explains isochores in mammalian genomes. *Trends Genet.* 19, 128–130
- 23 Jabbari, K. and Bernardi, G. (2004) Body temperature and evolutionary genomics of vertebrates: a lesson from the genomes of *Takifugu rubripes* and *Tetraodon nigroviridis*. *Gene* 333, 179–181
- 24 Kudla, G. *et al.* (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* 4, e180
- 25 Ptak, S.E. *et al.* (2005) Fine-scale recombination patterns differ between chimpanzees and humans. *Nat. Genet.* 37, 429–434
- 26 Winckler, W. *et al.* (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* 308, 107–111
- 27 Jeffreys, A.J. and Neumann, R. (2005) Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum. Mol. Genet.* 14, 2277–2287

## Have you contributed to an Elsevier publication? Did you know that you are entitled to a 30% discount on books?

A 30% discount is available to all Elsevier book and journal contributors when ordering books or stand-alone CD-ROMs directly from us.

To take advantage of your discount:

**1. Choose your book(s) from [www.elsevier.com](http://www.elsevier.com) or [www.books.elsevier.com](http://www.books.elsevier.com)**

**2. Place your order**

Americas:

Phone: +1 800 782 4927 for US customers

Phone: +1 800 460 3110 for Canada, South and Central America customers

Fax: +1 314 453 4898

[author.contributor@elsevier.com](mailto:author.contributor@elsevier.com)

All other countries:

Phone: +44 (0)1865 474 010

Fax: +44 (0)1865 474 011

[directorders@elsevier.com](mailto:directorders@elsevier.com)

You'll need to provide the name of the Elsevier book or journal to which you have contributed. Shipping is free on prepaid orders within the US.

If you are faxing your order, please enclose a copy of this page.

**3. Make your payment**

This discount is only available on prepaid orders. Please note that this offer does not apply to multi-volume reference works or Elsevier Health Sciences products.

**For more information, visit [www.books.elsevier.com](http://www.books.elsevier.com)**