Update

Genome Analysis

GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates

Nicolas Galtier[1], Laurent Duret[2], Sylvain Glémin[1] and Vincent Ranwez[1]

[1] Université Montpellier 2 – CNRS UMR 5554 – Institut des Sciences de l'Evolution, Place E. Bataillon – CC64 – 34095 Montpellier, France
[2] Université de Lyon, Université Lyon 1, CNRS, UMR5558 – Laboratoire de Biométrie et Biologie Évolutive, F-69622, Villeurbanne, France

GC-biased gene conversion (gBGC) is a recently discovered, recombination-associated segregation distortion, which influences GC-content dynamics in the mammalian genome. We scanned the primate proteome for examples of exon-specific, lineage-specific accelerated amino acid evolution. Here, we show that such episodes are frequently accompanied by an increase in GC-content, which extends to synonymous and intronic positions. This demonstrates that gBGC has substantially (negatively) impacted the evolutionary trajectory of human proteins by promoting the fixation of deleterious AT→GC mutations.

Biased gene conversion and mammalian genome evolution

The analysis of genome variation patterns in mammals has demonstrated the evolutionary importance of GC-biased gene conversion (gBGC), a recombination-associated segregation distortion favoring G and C over A and T bases [1]. A large body of evidence indicates that gBGC largely determines the dynamics of GC-content at neutral sites in mammals [2–7]. However, it is not clear whether, and how, gBGC affects selected genomic components. Theoretically, strong gBGC could overcome natural selection, promote the segregation of deleterious AT→GC mutations or impede the fixation of advantageous GC→AT mutations. Empirically, very little is known about the impact of gBGC on selected sequences, although we previously suggested [8] that gBGC could explain the sudden accelerated evolution in human of otherwise conserved noncoding elements [9]. To investigate the functional consequences of gBGC, we searched the primate proteome for events of lineage-specific accelerated amino acid evolution and asked whether such episodes are caused by variations in mutation rate, selective regime or gBGC.

Accelerated, GC-biased episodes of amino acid evolution

We scanned 12 452 primate exons >200 base pairs (bp) [10] for elevated amino acid evolution in four short branches of the hominid tree: the human, chimpanzee, human–chimpanzee ancestral and human–chimpanzee–orangutan ancestral lineages. We used macaque, lemur and galago as external references. Lineage-specific accelerations were detected by comparing the shape of exon-specific trees with that of a reference tree, reconstructed from 1000 concatenated exons (Figure 1, Box 1 and supplementary material online). We detected 622 episodes of significant branch-specific accelerated amino acid evolution (p-value <0.01), 379 of which involved three amino acid changes or more (see supplementary material online). Some of the episodes were highly significant: 30 tests yielded a p-value $<10^{-5}$ (0.5 expected).

The four lineages examined did not yield equal numbers of accelerated episodes: 53 were found in human, 84 in chimpanzee, 103 in the human–chimpanzee ancestral branch and 139 in the human–chimpanzee–orangutan ancestral branch. This is consistent with the recent report of a greater number of positively selected genes in chimpanzee than in human [11]. Bakewell *et al.* [11] attributed this to a greater long-term effective population size in chimpanzee, convincingly excluding the hypothesis of a differential sequencing error rate between the two genomes. We note, however, that this pattern could also result from gBGC, which, like natural selection, is more efficient in large populations [12]. Eighteen exons yielded two distinct significant episodes in distinct lineages (~3 expected by chance), indicating that some exons have a propensity for episodic accelerated evolution in the long term.

We analyzed the nucleotide substitution pattern underlying the episodes of accelerated amino acid evolution. For these exons, the percentage of AT→GC changes was significantly greater in amino acid accelerated (43.6%) than in the non-accelerated (39.6%) branches. 33 episodes were significantly GC-biased at the 5% level (19 expected under the null hypothesis) and 19 at the 1% level (four would be expected). Only four episodes were significantly AT-biased at the 5% level and none at the 1% level. Accelerated amino acid evolution, therefore, is frequently and sometimes strongly accompanied by a GC-biased nucleotide substitution process, indicating that gBGC could be involved.

To test further this hypothesis, we focused on the 19 episodes showing significantly GC-biased acceleration at the 1% level. These exons tend to be located in high-recombining regions; their median crossover rate was 2.33 cM Mb^{-1}, versus 1.15 cM Mb^{-1} for the full set of analyzed exons (Kruskal-Wallis rank sum test: p-value <0.001; crossover rates from HapMap [http://www.hapmap.org], 1 Mb scale; 1–5 Mb scales gave similar results; see Table S1). Interestingly, this effect is specific to crossovers occurring in the male germline; the median male crossover

Corresponding author: Galtier, N. (nicolas.galtier@univ-montp2.fr).

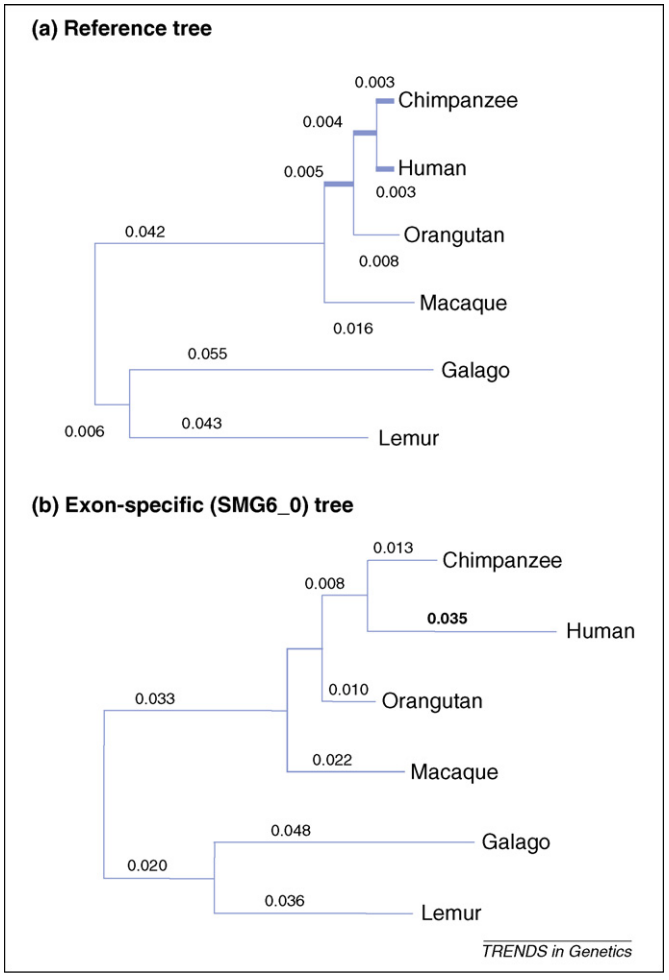

Figure 1. Diagram measuring accelerated amino acid evolution. **(a)** A reference tree, built from a large proteomic dataset. The four branches in which accelerated evolution was tested are in bold type. **(b)** An exon-specific tree, built from exon SMG6_0 (Table 1). An episode of accelerated amino acid evolution is detected in the human lineage: the human branch contributes 17% of the total tree length in the exon-specific tree, which is significantly higher than in the reference tree (1.6%). Branch lengths units are given in average number of amino acid substitutions per site.

rate in accelerated exons is 4.8 times greater than in the whole dataset (2.91 versus 0.61 cM Mb^{-1}; p-value <0.001; crossover rates from deCODE [http://www.decode.com/genotyping/], 1 Mb scale), but the female crossover rate is similar between accelerated (0.97 cM Mb^{-1}) and non-accelerated (1.47 cM Mb^{-1}; p-value = 0.35) exons. We analyzed the synonymous changes and found that they were more numerous than expected and strongly biased towards GC during episodes of accelerated, GC-biased amino acid evolution (sum over all 19 episodes: 91 AT→GC synonymous changes; 22 GC→AT synonymous changes; p-value $<10^{-6}$). This is again consistent with the idea that gBGC, not positive selection, is the underlying process because all codon positions, including the (presumably neutral) synonymous ones, are affected.

We analyzed the nine GC-biased, accelerated exons that were >600 bp in more detail (Table 1). A maximum-likelihood analysis [13] of the lineage-specific non-synonymous/synonymous (d_N/d_S) rate ratio revealed that the d_N/d_S ratio was increased in the accelerated branch (this trend was found in all nine exons and was significant in four and highly significant when the exons were considered collectively). These episodes, therefore, cannot be explained by a sudden burst of AT→GC mutations; if only mutation was involved, we would not expect any change in the d_N/d_S ratio. This pattern, however, would be expected under the gBGC model; if strong enough, it can overcome purifying selection and lead to the 'undesired' accumulation of deleterious mutations, thus, increasing the d_N/d_S ratio (Box 2). Even highly constrained genes can be substantially perturbed. Exon SORBS2_4, for instance, which is under strong purifying selection in nonhuman

Box 1. Detecting lineage-specific accelerated evolution

To detect exon-specific episodes of accelerated amino acid evolution, we searched for disproportionately long branches in the ([*Otolemur*, *Microcebus*], *Macaca*, [*Pongo*, *Homo*, *Pan*]) primate tree, considered as the true topology. Reference branch lengths were estimated by the maximum-likelihood method from a dataset made of 1000 randomly chosen exons (Figure 1a). Then, exons were considered separately and exon-specific branch lengths were calculated using the same procedure (Figure 1b). To assess whether a specific exon used to undergo significantly accelerated evolution in a specific branch, we first measured the relative evolutionary rate of the focal exon, r_e, excluding the focal branch:

$$r_e = (L_e - l_e)/(L_{ref} - l_{ref}) \quad (1)$$

where l_{ref} and l_e are the lengths of the focal branch in the reference and exon-specific trees, respectively, and where L_{ref} and L_e are the total lengths (sum of branch lengths) of the two trees. Then, we defined the expected number of amino acid changes in the focal branch for this exon, $E(m_e)$, as:

$$E(m_e) = l_{ref} \cdot r_e \cdot n_e \quad (2)$$

where n_e is the length of the considered exon. Equation (2) expresses that, if the exon-specific rate was unchanged in the focal lineage, then the focal exonic branch length would be equally proportional to the focal reference branch length. We qualified the focal branch as accelerated when the estimated number of amino acid changes, m_e, was significantly greater than its expectation, $E(m_e)$. In its basic version, the p-value is written as:

$$p - value\_basic_{[acceleration]} = Pr[P(E(m_e)) \geq m_e] \quad (3)$$

where $P(\times)$, the Poisson distribution of mean $\times$, is the distribution of the number of events in a Markov process. Equation (3), however, considers the expected number of changes in the focal branch, $E(m_e)$, as a known quantity when it was actually estimated from finite data. We need to account for the sampling variance around this number, which linearly depends on exon length. Approximating the posterior distribution of the estimated expected number of changes by a Gaussian variable, we have:

$$p - value_{[acceleration]} = \int_0^\infty g(x) Pr(P(x) > m_e) dx \quad (4)$$

where g is the probability density of a (truncated in zero) Gaussian distribution of mean $E(m_e)$ and variance $E(m_e)\ l_{ref}\ /(L_{ref} - l_{ref})$. This sampling variance was derived from equations (1) and (2) by assuming that branch lengths in the reference tree have zero variance, whereas any branch length of the exon tree, b_e, has variance b_e/n_e. Equation (4) accounts for the variable amount of information available about the expected rate in short versus long exons.

For each exon, the test defined by equation (4) was successively applied to the four branches indicated in Figure 1 (top): the human branch (H), the chimpanzee branch (C), the human–chimpanzee ancestral branch (HC), and the human–chimpanzee–orangutan ancestral branch (HCO). These are short branches of the tree, corresponding to periods of time <10 million years (My), in which episodes of accelerated evolution are more easily detected and interpreted.

Table 1. GC-biased, amino acid accelerated exons.

Exon[a]	Exon length	Lineage[b]	Amino acid changes		Synonymous changes		GC-bias[c]	ω_0[d]	ω_1[e]
			exp.	obs.	exp.	obs.			
DMRT3_1	660	C	0.3	15.2	1.9	14.3	22/0 (8/14)	0.066	0.544**
KCNV2_0	822	C	1.0	5.2	10.2	7.7	9/2 (4/7)	0.042	0.119*
KIAA1430_0	657	C	1.7	7.6	1.8	6.4	11/1 (6/6)	0.401	0.970
LRRC25_0	750	HCO	5.4	14.9	6.0	9.8	12/9 (7/14)	0.387	0.699
LRRC33_1	1923	C	2.6	13.2	8.3	13.7	13/6 (7/12)	0.134	0.219
LYSMD4_1	603	HC	1.9	9.6	2.0	7.8	17/0 (8/9)	0.404	0.533
MKL1_5	630	H	0.2	4.0	3.1	6.1	8/1 (3/6)	0.028	0.199**
SMG6_0	1530	H	1.9	17.7	3.2	15.2	25/4 (12/17)	0.255	0.507
SORBS2_4	606	H	0.7	3.9	3.9	3.5	5/0 (2/3)	0.077	0.314*

[a]Gene short name and exon number (ORTHOMAM [http://www.orthomam.univ-montp2.fr] pipeline). Abbreviations: DMRT3, double sex and mab-3 related transcription factor; KCNV2, potassium channel subfamily V member 2; KIAA1430, un-annotated gene; LRRC25, LRRC33, leucine-rich repeat-containing protein 25/33 precursors; LYSMD4, LysM and putative peptidoglycan-binding domain-containing protein 4; MKL1, myocardin-related transcription factor A; SMG6, telomerase binding protein EST1A; SORBS2, sorbin isoform 2.

[b]Branch of the primate tree in which exon evolution was accelerated. Abbreviations: H, human; C, chimpanzee; HC, human-chimpanzee ancestral; HCO, human–chimpanzee–orangutan ancestral.

[c]Estimated number of AT→GC/GC→AT changes (expected numbers of changes knowing exon GC-content are given in parentheses).

[d]Estimated d_N/d_S ratio in non-accelerated branches.

[e]Estimated d_N/d_S ratio in accelerated branches (*ω_1 significantly higher than ω_0, 5% level; **1% level).

branches ($d_N/d_S = 0.08$), has undergone a substantial increase in d_N/d_S ratio in the human lineage ($d_N/d_S = 0.31$).

Being independent of the coding nature of DNA, gBGC should also affect noncoding sequences. Therefore, we extracted the non-coding sequences (500 bp, introns or untranslated regions) flanking each of these nine exons from ENSEMBL (http://www.ensembl.org). Among the 15 available flanking regions, six showed significantly accelerated and GC-biased evolution during episodes of accelerated and GC-biased exonic evolution. The estimated number of nucleotide changes in flanking regions during these episodes was 116 (the sum over the 15 flanking regions), when 51 would have been expected under non-accelerated evolution (p-value $<10^{-6}$). The GC-bias was even more spectacular: 81 AT→GC; 11 GC→AT; p-value $<10^{-6}$. This demonstrates that the amino acid pattern in these exons was determined by genomic location, not coding property and, therefore, was probably caused by gBGC, not selection.

Proteic Achilles' heel

Most analyses of molecular evolutionary rates rely on two basic principles: (i) what is conserved is functional; and (ii) what evolves quickly is adaptive. These two rules underlie current approaches to genome annotation and positive selection detection. Here, we have shown that another player, gBGC, can substantially influence protein evolutionary rates in primates. gBGC leads to sudden episodes of accelerated evolution in otherwise conserved exons by promoting the fixation of AT→GC mutations. The effect can be extremely strong. Exon DMRT3_1, for instance, has experienced 29 nucleotide changes in the chimpanzee branch, among which, 15 were non-synonymous (0.3 expected, given the rate in other branches), 22 from AT→GC and none from GC→AT (Table 1). Strong gBGC counteracts purifying selection and leads to the fixation of deleterious amino acid mutations, as revealed by d_N/d_S analyses. The proteins in primates have been accumulating an undesired amount of deleterious substitutions as a consequence of a seemingly unimportant characteristic of the recombinatory machinery: biased DNA repair of heteroduplexed intermediates. The episodic nature of this process is consistent with the short life span of recombination hotspots in primates [14]. The gBGC episodes we detected might, therefore, reflect the existence of former recombination hotspots in the immediate neighborhood of the target exons. Consistent with this idea, most of the accelerated GC-biased exons we detected are located in high-recombining large-scale regions, in which the probability of birth of a hotspot is increased [15].

In this study, we focused on long, conserved exons (6% of primate exons) on four short branches of the primate tree and on strongly accelerated exons. Given that gBGC must also influence short exons, long branches and less spectacular episodes, the number of gBGC-induced events we detected must be an underestimate of the true long-term impact of gBGC on proteome evolution. Among the accelerated exons we analyzed, 10% showed a significant GC-bias. This does not mean, however, that the remaining 90% have been immune from gBGC; only for episodes involving a large number of substitutions can we detect a potential GC-bias. Note that an episode of gBGC leading to the fixation of deleterious alleles is likely to be followed by compensatory, positively selected substitutions, restoring the function of the protein. Additional modeling efforts will be needed to quantify the respective effects (and relationship) of positive selection, relaxed purifying selection and gBGC on amino evolution in primates.

gBGC: pathological implications?

GC-content is highly variable across genes and regions of the human genome [16]. Several adaptive scenarios have been proposed to explain the existence of local GC-enrichment [17], albeit without any empirical support. Our results indicate that genes located in high-recombining (GC-rich) regions, far from providing an advantage, could be weak points of the primate genome – our proteic Achilles' heel. We suggest that the local accumulation of G and C bases in various regions of the mammalian genome was probably not adaptive and perhaps even costly, because it resulted from a genetic process that counteracts the action of natural selection. Having impacted the long-term evolution of proteins, gBGC must also currently apply within human populations. It is

tempting to suggest that some of the genetic diseases observed at unexpectedly high prevalence could correspond to AT→GC mutations sustained by gBGC despite their deleterious effect.

The population genetic theory indicates that recombination increases the efficiency of multi-locus natural selection by generating new combinations of alleles, therefore increasing the genetic variance within populations [18]. This property has been invoked to explain various patterns of DNA sequence variation within and between species [19,20]. Here, we suggest that recombination can locally decrease the efficiency of purifying selection on protein coding genes through gBGC. The large-scale (genetic exchanges) and short-scale (gBGC) effects of recombination on selective efficiency are, therefore, opposite and recombination is not always favorable to species adaptation. Remarkably, the effect we detect is specifically correlated to the male recombination rate, as previously noticed for non-coding sequences [4,6,7]. The reason for such male-driven gBGC is not yet clearly established; it might reflect a mechanistic difference between the male and female meiosis.

Box 2. How does gBGC affect the d_N/d_S ratio?

Here, we present a simple model showing that gBGC can increase the d_N/d_S ratio. We consider a genic sequence with independent bi-allelic, AT vs. GC sites, either synonymous or non-synonymous. We assume that GC is the advantageous state for half of the non-synonymous sites and that back mutations can occur. We note $S = 4N_e|s|$, where N_e is the effective population size and s the selection coefficient (positive for advantageous mutations, negative for deleterious mutations). gBGC affects both synonymous and nonsynonymous sites, and we note $B = 4N_e b$, where b is the gBGC coefficient. We also note N, the population size, u and v the mutation rates from GC to AT and from AT to GC, respectively, and p_{GC^0}, p_{GC^+}, *and* p_{GC^-}, the proportions of sites fixed for neutral, advantageous, and deleterious GC alleles, respectively. Under these assumptions and assuming the same selection coefficient for all sites, we have:

$$d_s = 2Nu\,p_{GC^0}F_{GC^0\to AT^0} + 2Nv(1 - p_{GC^0})F_{AT^0\to GC^0} \quad (1)$$

and

$$d_N = \frac{1}{2}d_N(GC\ \ advantageous) + \frac{1}{2}d_N(GC\ \ deleterious)$$

$$d_N = \frac{1}{2}(2Nu\,p_{GC^+\to AT^-} + 2Nv(1 - p_{GC^+})F_{AT^-\to GC^+}) \quad (2)$$

$$+\frac{1}{2}(2Nu\,p_{GC^-}F_{GC^-\to AT^+} + 2Nv(1 - p_{GC^-})F_{AT^+\to GC^-})$$

where the $F_{X\to Y}$ terms are the fixation probabilities of $X\to Y$ mutations, the superscripts 0, +, and − holding for neutral, advantageous, and deleterious, respectively. $F_{X\to Y}$ terms can be obtained from Nagylaki's model [12] (see supplementary material online).

The synonymous and non-synonymous GC-contents can be computed at equilibrium. In human, however, GC-content is typically not at equilibrium. In particular, the GC-content of GC-rich genes is decreasing [5]. Even if gBGC is not currently active, GC3 can still be high because of the slow return to equilibrium through mutation, whereas GC12 should return faster to the mutation-selection-drift equilibrium. As an example, we consider the effect of a gBGC episode on a GC-rich gene, using $p_{GC^0} = 0.7$ (~GC3) and computing p_{GC^+} *and* p_{GC^-} at mutation-selection-drift equilibrium, assuming an AT-biased mutation process ($v = 2u$). We used $B = 8.7$, an estimate obtained from the HapMap project [15,21].

Figure Ia shows the distribution of d_N as a function of S, before or during the gBGC episode. gBGC strongly affects the substitution rate of weakly selected sites by promoting the fixation of AT→GC mutations. The effect is maximal for sites at which GC is weakly deleterious – these sites were slow-evolving and mostly fixed to AT before the episode. By overcoming weak selection, gBGC increases the proportion of 'evolvable' amino acid sites, which substantially increases d_N. In non-equilibrium conditions (this example), gBGC can also increase d_S. The net effect, however, is an increase of the d_N/d_S ratio, which can even reach values >1. In this numerical example, if, according to Eyre-Walker *et al.* [22], we assume a γ distribution of selection coefficients with mean $|S| = 1300$ and shape parameter = 0.23, resulting in GC12 = 0.47, then the d_N/d_S ratio increases from 0.18 (before the gBGC episode) to 0.43 (during the gBGC episode), which is consistent with our empirical results (Table 1). More generally, using the same parameters, Figure Ib shows that d_N/d_S monotonically increases with B. The qualitative predictions presented here are robust to other evolutionary scenarios, such as variations in GC-content, proportion of GC deleterious sites, distribution of fitness effect and inclusion of A↔T and C↔G mutations (see supplementary material online).

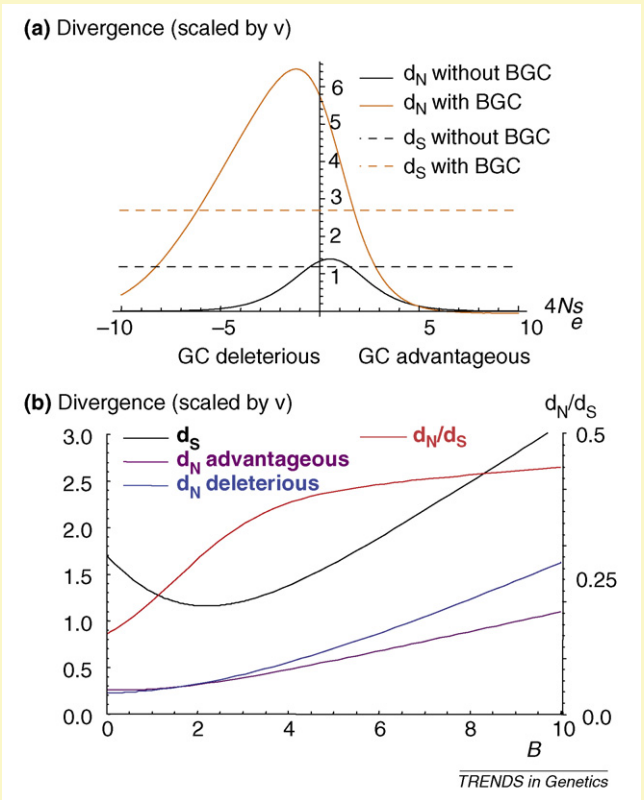

Figure I. gBGC and coding sequence evolutionary rate.

Concluding remarks

This work demonstrates that gBGC, like adaptive evolution, can lead to strong and sudden accelerated evolution of functional elements and impact the d_N/d_S ratio, thus, potentially corrupting positive selection detection tests. Therefore, one must be cautious when interpreting fast evolution in adaptive terms [8]. Selective hypothesis should be invoked only after both the neutral and the gBGC models have been rejected – the extended null hypothesis of molecular evolution.

Acknowledgements

This work was supported by the Centre National de la Recherche Scientifique and the Agence Nationale de la Recherche (GIP ANR-08-GENO-003–01). This is contribution ISEM 2008_097 from Institut des Sciences de l'Evolution de Montpellier.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2008.10.011.

References

1 Marais, G. (2003) Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* 19, 330–338
2 Galtier, N. *et al.* (2001) GC-content evolution in Mamm. *Genomes: the biased gene conversion hypothesis. Genetics* 159, 907–911
3 Webster, M.T. and Smith, N.G. (2004) Fixation biases affecting human SNPs. *Trends Genet.* 20, 122–126
4 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
5 Duret, L. *et al.* (2006) A new perspective on isochore evolution. *Gene* 385, 71–74
6 Dreszer, T.R. *et al.* (2007) Biased clustered substitutions in the human genome: The footprints of male-driven biased gene conversion. *Genome Res.* 17, 1420–1430
7 Duret, L. and Arndt, P.F. (2008) The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* 4, e1000071
8 Galtier, N. and Duret, L. (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* 23, 273–277
9 Pollard, K.S. *et al.* (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443, 167–172
10 Ranwez, V. *et al.* (2007) OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol. Biol.* 7, 241
11 Bakewell, M.A. *et al.* (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl. Acad. Sci. U. S. A.* 104, 7489–7494
12 Nagylaki, T. (1983) Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. U. S. A.* 80, 6278–6281
13 Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15, 568–573
14 Winckler, W. *et al.* (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* 308, 107–111
15 Myers, S. *et al.* (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310, 321–324
16 Bernardi, G. *et al.* (1985) The mosaic genome of warm-blooded vertebrates. *Science* 228, 953–958
17 Bernardi, G. (2007) The neoselectionist theory of genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 104, 8385–8390
18 Otto, S.P. and Barton, N.H. (1997) The evolution of recombination: removing the limits to natural selection. *Genetics* 147, 879–906
19 Begun, D.J. and Aquadro, C.F. (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356, 519–520
20 Bachtrog, D. and Charlesworth, B. (2002) Reduced adaptation of a non-recombining neo-Y chromosome. *Nature* 416, 323–326
21 Spencer, C.C. *et al.* (2006) The influence of recombination on human genetic diversity. *PLoS Genet.* 2, e148
22 Eyre-Walker, A. *et al.* (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173, 891–900

0168-9525/$ – see front matter
doi:10.1016/j.tig.2008.10.011 Available online 21 November 2008

Genome Analysis

Protein material costs: single atoms can make an evolutionary difference

Jason G. Bragg[1,2] and Andreas Wagner[3,4,5]

[1] Biology Department, University of New Mexico, Albuquerque, NM 87131, USA
[2] Present address: Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3] University of Zurich, Department of Biochemistry, Building Y27, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
[4] The Santa Fe Institute, Santa Fe, NM 87501, USA
[5] The Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

The process of gene expression has material costs caused by the quantities of carbon, nitrogen, sulfur and phosphorus that are needed to make mRNAs and proteins. When any such chemical element is ecologically limiting, mutations increasing these costs can reduce growth. Here, we ask if such mutations are 'visible' to natural selection in the yeast *Saccharomyces cerevisiae*. We find that mutations causing small increases in expression and even single amino acid replacements can be subject to natural selection on the basis of their material costs.

Nutrient limitation can influence protein evolution

Over vast regions of the globe, elemental nutrients including nitrogen, phosphorus and carbon limit the growth of organisms and mediate competition between them. Specific elements often limit growth because they are needed to make important biomolecules [1]. For example, mRNAs contain carbon (C), nitrogen (N) and phosphorus (P) atoms and proteins contain carbon, nitrogen and sulfur (S) atoms. Protein composition is constrained by natural selection because proteins need to perform specific functions, but this is not the only compositional constraint. Natural selection can also influence protein material costs (e.g. see Refs [2–7]). For example, highly expressed proteins can contain fewer atoms of ecologically limiting elements (per amino acid) than other proteins in the same proteomes [4,6,7]. Currently, we lack a quantitative, mechanistic explanation for the adaptive evolution of protein material costs. That is, mutations might affect material costs by directly changing the number of atoms required to make a protein, or by changing gene expression levels. However, we do not know if the resulting changes in

Corresponding author: Wagner, A. (aw@bioc.unizh.ch).