Evolution of synonymous codon usage in metazoans

Laurent Duret

The vast amount of data generated by genome projects and the recent development of population genetics models make comparative sequence analyses a very powerful approach with which to detect the footprints of selection. Studies on synonymous codon usage show that traits with minuscule phenotypic effects can be molded by natural selection. But variations in mutation patterns and processes of biased gene conversion make it difficult to distinguish between selective and neutral evolutionary processes.

Addresses
Laboratoire de Biométrie et Biologie Evolutive UMR CNRS 5558,
Université Claude Bernard Lyon 1 69622, Villeurbanne cedex, France;
e-mail: duret@biomserv.univ-lyon1.fr

Current Opinion in Genetics & Development 2002, **12**:640–649

0959-437X/02/$ – see front matter

Published online 4 October 2002

Abbreviations

BGC	biased gene conversion
EST	expressed sequence tag
GC3	GC content at third codon position
HRi	Hill–Robertson interference
MCB	maximum-likelihood codon bias
r_S	Spearman's rank correlation coefficient
SAGE	serial analysis of gene expression

Introduction

The primary aim of a genome project is to make an inventory of the functional elements, such as genes and regulatory regions, that are embedded within chromosomes. From this inventory emerges insights into genomic organization. Increasingly, we have the data to determine whether the genome is simply a bag of genes or whether there are there some constraints on its structure — for example, the location of genes, variations in base composition, repeated sequences — that are needed for the proper expression of genetic information. In other words, to understand a genome it is necessary to identify all of the features of that genome that are subject to the action of natural selection.

In large eukaryotic (e.g. metazoan) genomes, it is particularly difficult to identify functional elements because they are hidden in a vast amount of noncoding sequence. In addition, the demonstration that a genetic element is functional classically relies on the identification of a mutation that results in a particular phenotype. Because of limitations in the size of populations that can be studied in laboratories, this experimental approach is limited to the identification of mutations with a strong phenotypic impact [1]. This problem is far from negligible; for example, 50% of gene knockouts in yeast have no detectable effect on the phenotype [2]. Notably, it seems that the absence of detectable phenotypic effects is due to the fact that those genes make only a marginal contribution to the fitness of yeasts [3]. The characterization of genomic features (not only genes but also any other functional element) under weak selective pressure in metazoans is therefore a particularly challenging task.

The comparative analysis of homologous sequences (either within or between species) can simplify this task: studying sequence evolution can reveal footprints left by the action of natural selection, which may in turn allow the identification of functional features. The main difficulty in the comparative approach is distinguishing the action of selection from the results of neutral evolutionary processes [4]. In this review, the power and limits of comparative analysis are illustrated by recent work on the evolution of synonymous codon usage in metazoans.

Imagine the following experiment: take the genome of a nematode and replace one codon by a synonymous codon in just 1 of its 19,000 genes. Can this mutation have any impact on the phenotype of this organism? The answer to this question is 'yes'. Although it is probably impossible to detect the impact experimentally, we will see below that natural selection is clearly able to operate on such small differences in fitness.

Biases in synonymous codon usage: selective and neutral models

Although synonymous codons encode the same amino acids, they are not used randomly and some are used more frequently than others. Such codon usage biases occur in most species from all kingdoms of life. They vary according to the genes within a given genome, and also according to the taxa. Classically, two models (a selective and a neutral one) have been proposed to explain the preferential use of a subset of synonymous codons (for review and references to the principal studies, see [5]). The selective model postulates that there is a co-adaptation of synonymous codon usage and abundance of tRNA to optimize translation efficiency ('translational selection'). According to the neutral model, codon biases result from biases in mutational processes ('mutational bias').

These two models make distinct predictions. For example, if there is a selective pressure to improve the efficiency of translation, then this pressure should be stronger for highly expressed genes than for weakly expressed genes. Thus, the model of translational selection predicts a correlation between codon usage bias and patterns of gene expression. Synonymous codons that are used preferentially in highly expressed genes (optimal codons) should also correspond to the most abundant tRNAs. In addition, the action of selection should result in a decrease in the probability of fixing mutations towards non-optimal codons and thus, at equilibrium, in a lower rate of synonymous substitution in highly expressed genes. Such a fixation bias in favour of optimal codons should also be detectable by comparing polymorphisms with substitution patterns and by analyzing the spectrum of allelic frequencies [6,7].

Conversely, the mutational bias model *a priori* does not predict any relationship between codon usage and gene expression (but see below). Such a mutational pressure should affect all positions in a genome, not only synonymous sites but also all other silent sites (e.g. introns and intergenic regions) and, to a lesser extent, non-synonymous codon positions. Thus, this model predicts a correlation between the base composition of synonymous sites and that of neighboring silent sites in the genome. According to the mutational bias model, the probability of fixation should be the same for all synonymous alleles.

Recent findings, however, have indicated that other processes may impact codon bias besides selection and simple mutation. Notably, it has been shown that transcription can be mutagenic [8•], which may induce a correlation between gene expression and base composition (and thus codon usage). In addition, the process of biased gene conversion can lead to differences in the probability of fixing mutations, even though there is no selective difference between alleles. These different points are discussed below.

Neutral and selective models are not mutually exclusive; indeed, codon usage almost certainly reflects a balance between selective and mutational pressures as well as drift [9]. Note also that translational selection is not the only selective pressure that may act on synonymous codon usage. Indeed, there is evidence that in the bacteria *Escherichia coli* — where translational selection has been clearly demonstrated — non-optimal codons can be maintained within genes because of conflicting selective pressures (e.g. selection upon ribosome-binding motifs at the start of genes) [10•]. In eukaryotic genes, it is known that some regulatory elements involved in splicing or mRNA stability are located in exons [11]. The need for appropriate gene regulation would place constraints on sequences that could account for the selective pressure detected at some synonymous codon positions in mammalian genes [12].

As mentioned previously, the mutational bias model predicts a correlation between codon usage and base composition in non-coding regions. However, even in absence of selective difference between synonymous codons, the base composition of synonymous sites is not expected to be identical to that of other neutral sites in the genome owing to the selective pressure acting on surrounding non-synonymous codon positions. The substitution pattern differs across regions for many reasons. First, the mutation rate at a given base depends on the nature of the flanking bases (e.g. CpG dinucleotides are mutational hotspots in mammalian genomes). Fedorov *et al.* [13] have shown that 90% of codons in *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana* have a significant context-dependent codon bias. Second, the base composition of introns and intergenic regions may be strongly affected by deletions or insertions (notably of transposable elements), whereas such mutations are strongly counterselected at synonymous sites [14]. Third, transition and transversion substitution pattern at the third codon position differs from that in non-coding regions. This is because of the structure of the genetic code: at twofold degenerate sites (24 of the 61 codons), the transitions are synonymous mutations, but not the transversions. Hence, transversions at third codon positions are under strong selective pressure, while transitions are silent mutations. Conversely, in non-coding regions, both transitions and transverions are silent. Thus, if the GC↔AT mutation pressure is different for transitions and transversion, then sites where selection limits substitution to transitions will have a GC content different from that of neutral sites [15•]. This phenomenon can also affect the base composition at fourfold degenerate sites, because non-synonymous substitutions can change twofold degenerate codons to fourfold degenerate codons [15•].

In summary, to test whether synonymous codon usage biases are due to selection or not, it is necessary to answer the following questions. Is there a relationship between codon usage bias and gene expression? If 'yes', can we exclude the possibility of a transcription-linked mutational process? Is there a relationship between codon usage bias and tRNA abundance? Is there evidence for non-neutral substitution processes at synonymous sites (i.e. do polymorphism and substitution patterns indicate a fixation bias in favour of optimal codons)? If we can reject the null hypothesis that codon usage bias occurs completely as a result of neutral evolutionary processes, then we have to determine both to what extent codon usage bias is shaped by selection and the reasons for this selective pressure (e.g. translation efficiency or something else). It is now possible to address these issues in several model organisms, for which — owing to the success of genome projects — sequences, polymorphism and expression data are available in large quantity.

Synonymous codon usage in *Drosophila* and *C. elegans*

Early studies by Sharp and colleagues [16,17] allowed the identification of a subset of codons in *Drosophila* and *C. elegans* that is used preferentially in genes showing a strong codon usage bias. Furthermore, by using the number of matching expressed sequence tags (ESTs) as a rough estimate of transcriptional activity, myself and Mouchiroud [18] showed directly that the frequency of this subset of codons is correlated positively with the level of gene expression. But ESTs do not provide a perfect estimation of levels of gene expression. Notably, many of the cDNA libraries used in EST projects have been 'normalized' (i.e. prepared using a procedure that tends to reduce the number of clones from over-expressed mRNAs). Hence, the transcription level of highly expressed genes is underestimated by this approach.

Castillo-Davis and Hartl [19] have confirmed the correlation between codon bias and levels of gene expression in *C. elegans* obtained by DNA microarray experiments (Spearman's rank correlation coefficient, $r_S = -0.3$, $P < 10^{-131}$) [19]. Note that in *Drosophila* and *C. elegans* almost all of the optimal codons contain a cytosine or a guanine in the third position. In contrast to the GC content of synonymous sites, the GC content of introns is not positively correlated with levels of gene expression [18]. This latter observation rules out the possibility that the relationship between codon bias and gene expression is due to a transcription-coupled mutational process [8•] and thus shows directly that synonymous codon usage is shaped by natural selection in these two invertebrates.

Few experimental data on the cellular abundances of tRNA in metazoans are available. But it is possible to estimate indirectly the relative abundance of individual tRNAs from the number of copies of the corresponding genes in the genomes that have been sequenced completely. In this manner it has been shown that, in both *D. melanogaster* and *C. elegans*, optimal codons correspond to the most abundant tRNAs [20,21,22••,23•]. These observations clearly support the translation selection hypothesis that synonymous codon usage has been shaped by selection to improve the efficiency of translation.

As predicted by the translational selection model, the rate of synonymous substitution is correlated negatively with codon bias in *C. elegans* [19]; however, the expected negative correlation is not observed in *Drosophila* [24•]. Notably, Kanaya *et al.* [23•] have found that the correspondence between tRNA abundance and optimal codons is less evident in *D. melanogaster* than in *C. elegans*. These observations might be linked to results obtained from population genetics studies that show a sharp decrease in the selective pressure on synonymous codon usage in *D. melanogaster* and, to a lesser extent, in *D. simulans* [25•,26••].

Thus, although there is evidence that translational selection has been acting in the past, this selection is no longer efficient enough to maintain in these two species the same high frequency of optimal codons as in their ancestral lineage. It will be interesting to investigate whether the decline in selection pressure can account for the lack of correlation between synonymous substitution rate and codon bias in *Drosophila* [24•].

Targets of selection on synonymous codon usage

The precise mechanism that drives selection on synonymous codon usage is not clearly known. The use of optimal codons could affect both the speed and the accuracy of translation. If selection on synonymous codon usage acts, at least in part, to improve the accuracy of translation, then it is expected that this selective pressure will be stronger on codons that encode amino acids essential for the function of the protein. In support of this model, two studies have shown that the frequency of optimal codons is higher at codons encoding constrained amino acids than at those encoding non-constrained amino acids in *Drosophila* and *C. elegans* [27,28]. Note, however, that the conclusions of those studies have been questioned by the recent finding that selection at the amino acid level can influence codon bias even in absence of selective difference between synonymous codons [15•].

Another possible target of selection affecting codon bias is the secondary structure of mRNAs. Carlini *et al.* [29] observed that in the *Drosophila* alchohol dehydrogenase genes the potential for secondary structure formation was stronger in weakly expressed mRNAs than in highly expressed mRNAs. They suggest that strong secondary structural elements might be selected for (or against) to decrease (or increase) gene expression. This is a very interesting hypothesis, but one that will be difficult to validate. Indeed, in the absence of selection it is expected that the base composition of a DNA strand will tend to equal frequencies of complementary bases (i.e. A = T and G = C) [30] and so will have a tendency to increase the potential for intrastrand helix formation.

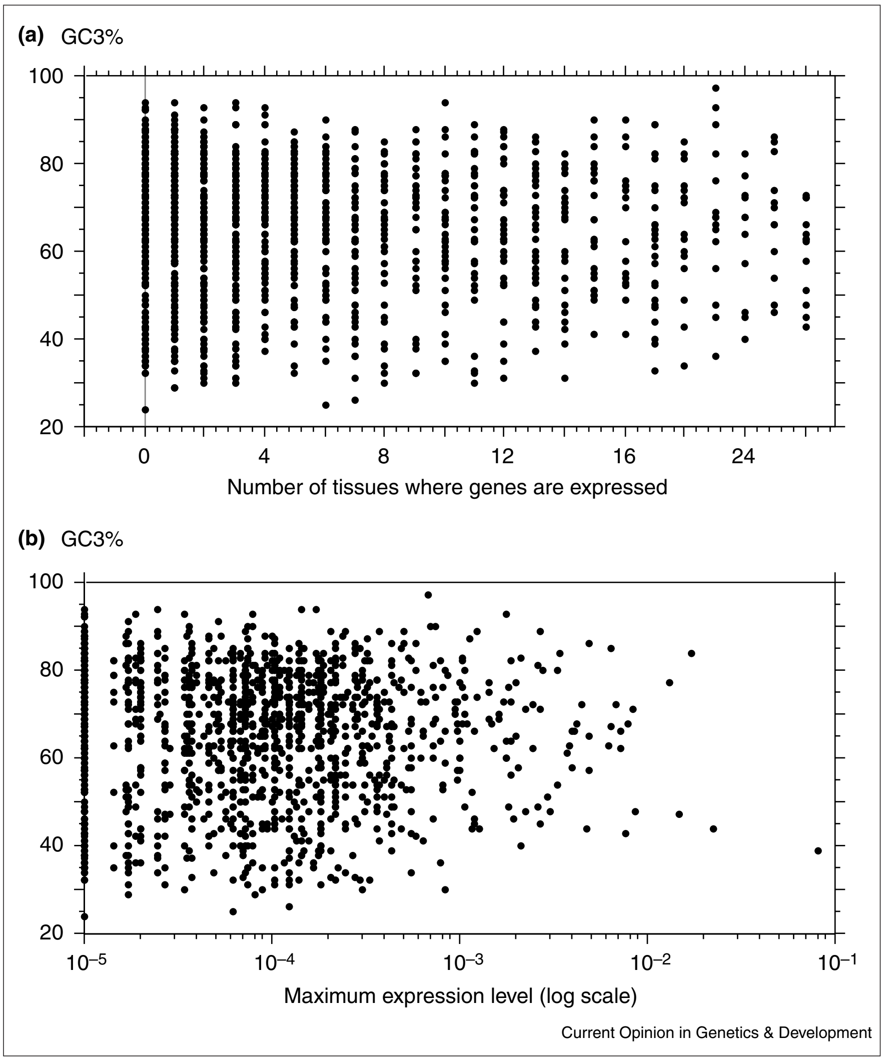
A few years ago, different groups showed in eukaryotes that, where translational selection occurs, the frequency of optimal codons is correlated negatively with the length of the encoded protein [18,31,32]. The effect of protein length on codon usage is almost as strong as the effect of gene expression [18]. This very puzzling observation is not predicted by any of the current models for selection on synonymous codon usage and at present remains unexplained [28].

Synonymous codon usage in vertebrates

Selection on synonymous codon usage has been shown to occur in very diverse eukaryotes including plants, fungi and invertebrates. But what about selection on synonymous codon usage in vertebrates? In these species, multivariate analyses reveal that the variability in codon usage is reflected essentially by a single major trend that is correlated strongly with the GC content at the third codon position (GC3) [23•,33••].

Recently, several groups have examined the relationship between GC3 and the levels of gene expression measured by different techniques in vertebrates [23•,33••,34]. In human and *Xenopus*, Kanaya *et al.* [23•] found no difference in codon usage between ribosomal protein genes (that are known to be expressed at very high levels) and other genes. In agreement with this analysis, ESTs reveal no positive correlation between GC3 and the breadth of tissue distribution, nor between GC3 and the expression level of human genes — there is in fact a significant negative

Figure 1



Gene expression pattern and GC content at the third codon (GC3) position in the human genome. Data on the base composition of 1396 complete human protein-encoding genes were taken from [14]. Expression patterns were measured in 22 tissues using EST data, as described in [53]. Only normal tissues with more than 10,000 ESTs were selected. **(a)** Correlation between GC3 and the number of tissues where at least one matching EST has been detected ($r_S = -0.09$, $P = 0.0008$). **(b)** Correlation between GC3 and the maximum expression level among the 22 tissues ($r_S = -0.06$, $P = 0.03$). The expression level of a gene in a given tissue is measured by the number of matching ESTs divided by the total number of ESTs sampled in that tissue. Both measures of expression (a,b) are negatively correlated with GC3. Although statistically significantly different from zero, these correlations are extremely weak and may have no biological meaning. At the least, we can conclude that there is an absence of a positive relationship between the GC3 content and the expression level or the breadth of the tissue distribution of genes (L Duret, D Mouchiroud, unpublished data).

correlation, albeit an extremely weak one (Figure 1). In *Xenopus* also, no correlation was found between GC3 and gene expression (based on ESTs) [33••].

In contradiction with these results, Konu and Li [34] reported a positive correlation beween GC3 and the levels of gene expression in rodents. Note, however, that this correlation depends on the technique used to measure expression: whereas GC3 was correlated positively with data obtained by serial analysis of gene expression (SAGE), it was not correlated with data obtained by cDNA microarrays. Konu and Li [34] attributed this lack of correlation with microarray data to the fact that the relationship between GC3 and expression is not linear and reaches a plateau for very high levels of gene expression. Indeed, after removing highly expressed genes from the analysis, they found a positive correlation between GC3 and cDNA microarray expression data [34]. This latter correlation should, however, be interpreted with caution because the significance of the statistical test cannot be assessed properly after having removed points.

The contradictions between these results obtained with different measures of expression is probably due to a methodological artifact of the SAGE technique. SAGE relies upon the generation of short (10 bp) tags, specific for each cDNA. The thermal stability of such short sequences depends on their GC-content, and hence, during the preparation of SAGE libraries, AT-rich tags are lost more frequently than GC-rich tags [35•]. A consequence of this is that SAGE tends to underestimate the expression level of GC-poor genes. Thus, this methodological bias induces an artificial correlation between GC content and the number of SAGE tags [35•]. It should also be noted that the analysis of cDNA microarray data reported in [34] is complicated by the fact that the intensity of the hybridization signal is dependent on the number of radioactive cytosine nucleotides that are incorporated in the cDNA probes and thus on their GC content. Although some biases arise from using ESTs to estimate levels of gene expression (see above) [18], they are, *a priori*, not dependent on the GC content of mRNA. In conclusion, there is currently no evidence for a significant relationship between the GC content of a gene and its expression pattern (i.e. level or breadth of expression) in vertebrates.

This is not surprising because in vertebrates the GC3 of a gene is correlated strongly with the GC content of the isochore in which it is located (for a review, see [36,37••]). Note that it has been proposed that most housekeeping genes should be located in GC-rich isochores, whereas tissue-specific or developmentally regulated genes should be located in GC-poor isochores [36]. As shown in Figure 1, however, EST data do not support this hypothesis. Although the evolutionary forces acting on isochores are under debate [36,37••], it is clear that the evolutionary forces responsible for variations in base composition (and thus codon usage) along the genome act not only on synonymous sites but also on introns and intergenic regions. This observation strongly argues against the model of translational selection as a principal determinant of codon usage in vertebrates. In addition, there is no detectable relationship between codon usage and the number of genes encoding tRNAs in the human genome [23•,38••]. But a question still remains: is it possible to detect some traces of translational selection after the impact of isochores has been taken into account?

Iida and Akashi [39] have proposed a very elegant test that relies on the analysis of alternatively spliced genes. Because codons located in constitutively expressed exons are translated more often than those in alternatively spliced exons, the translational selection model predicts a stronger codon bias in the former than in the latter. This method controls for both regional differences in base composition and transcription-coupled mutational processes. They found that GC-ending codons are more abundant in constitutive than alternatively spliced exons in both *Drosophila* and human. Unfortunately, the comparison is complicated by a gradient in GC content along genes [40,41], and more data will be necessary to determine whether the excess of major codons observed in constitutive exons is caused by this polarity in base composition.

Urrutia and Hurst [42•] have proposed a new measure of codon bias, called the maximum-likelihood codon bias (MCB), that corrects for local variations in base composition. The basic principle of MCB lies in measuring the deviation between the codon usage observed for a given amino acid and the codon usage expected according to the nucleotide composition at the third codon position of all other codons that have the same degree of degeneracy within a gene, under the assumption that they should all be subject to the same mutational biases. (Note, however, that context-dependent biases [13] are not counted by this method.) This measure also has the advantage of being less sensitive than other methods to biases arising from the proportion of fourfold degenerate amino acids in a gene [42•].

Urrutia and Hurst [42•] found a positive correlation between MCB and the breadth of expression for human genes (correlation coefficient $r = 0.18$, $P < 0.001$). It should be stressed that this correlation is very weak (only 3% of the variance of MCB is explained by gene expression). Although it is clearly statistically significant, this correlation should be interpreted cautiously because it could be induced by any slight methodological artefact. Indeed, Urrutia and Hurst show that this correlation is due to the fact that the MCB measure is affected by gene length and that, in human, the length of genes is correlated weakly with their breadth of expression. Thus, it is again concluded that there is an absence of evidence for translational selection in the human genome.

Another possible approach to analyse variations of codon usage, independently of variations of isochore GC content, consists of using multivariate analyses [23•,33••]. This technique allows the identification of axes (principal factors) that reflect the maximum of variability in the dataset. Because successive axes are under orthogonality constraint, the variability revealed by each axis is independent of the variability revealed by other axes. As mentioned previously, the first axis identified by multivariate analyses (i.e. the factor that reflects the maximum of variability in the dataset) is the GC3. By analysing the second principal axis, Kanaya *et al.* [23•] found that in the human genome the second main factor that contributes to codon usage diversity is the number of CpG-containing codons. In mammals, the density of CpG dinucleotides within genes depends on their expression pattern: genes that are expressed in early stages of development or in the germline (including housekeeping genes) almost always contain a CpG island in their promoter region [43]. CpG islands often overlap with the first exon or exons of genes and thus affect codon usage. It is generally thought that the high density of CpG dinucleotides within CpG islands is a consequence of a reduction in the mutation rate of cytosine owing to the reduction in methylation in these promoter regions. It is not known whether the presence of CpG motif *per se* is subject to selective pressure.

By using the same approach, Musto *et al.* [33••] found that in *Xenopus* the second principal axis (which accounts for 6.5% of the total variance in codon usage) is correlated with the level of gene expression measured with ESTs. Highly expressed genes are characterized by a high frequency of pyrimidines, notably thymines, at the third codon position. This result is important: it is the first time that a significant correlation between codon usage and gene expression has been demonstrated in a vertebrate. Note, however, that this observation is not sufficient to show conclusively that selection acts on codon usage in *Xenopus*. Indeed, it has been shown in bacteria that transcription increases the frequency of cytosine to thymine mutations on the sense strand of genes. This is probably because this strand remains single during transcription, which increases the deamination rate of cytosine [8•].

It is possible that the same effect occurs in eukaryotes. Indeed, myself and Mouchiroud have found that in human genes with a wide breadth of tissue distribution (i.e. housekeeping genes, which are presumably also transcribed in the germline), there is an excess of thymine, both at the third position of fourfold degenerate codons and in introns (L Duret, D Mouchiroud, unpublished data; Figure 2). Thus, it would be interesting to analyse the base composition of introns in *Xenopus* to determine whether the relationship between codon usage and gene expression observed by Musto *et al.* [33••] is due to translational selection or to a transcription-coupled mutational bias.

Detecting selection on silent sites: troubles with biased gene conversion

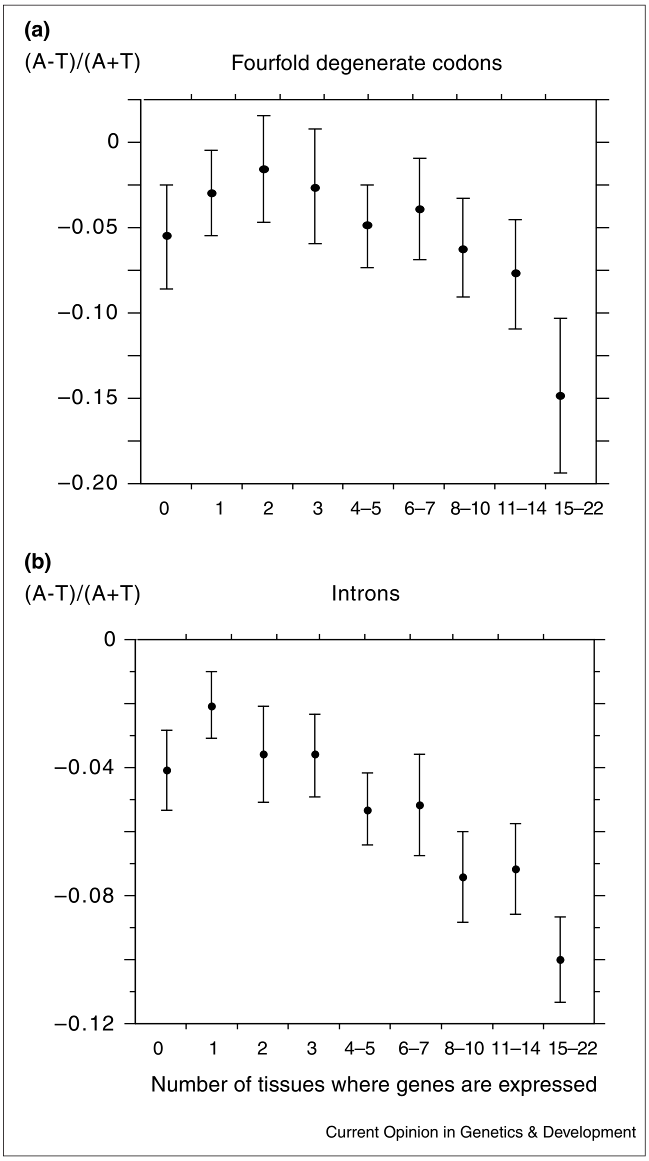
According to the mutational bias model, codon biases result only because some mutations are more frequent than others. This neutral model makes a strong prediction: because there is no selective difference between frequent (major) and rare (minor) codons, any synonymous mutation, towards a minor or major codon, should have the same probability of fixation (i.e. no fixation bias). It is possible to test this prediction by analysing the spectrum of allelic frequencies of polymorphic codons in a population. In the absence of selection, frequency distributions should be the same for major and minor polymorphic codons. But if major codons are selectively advantageous, then on average they should segregate in populations at higher frequencies than minor codons [7,26••].

A second prediction (which is a direct consequence of the first) is that at equilibrium substitution patterns should be identical to mutation patterns. This can be tested by comparing patterns of polymorphism and divergence [6,26••]. Such tests have been used to infer the parameters of selection that act on optimal codons in different *Drosophila* species; notably, these tests have shown a recent decrease in the strength of translational selection in *D. melanogaster* and *D. simulans* [6,7,26••]. With the growing amount of polymorphism data, this approach can now be applied to the human genome. Notably, it has been shown that there is a fixation bias in favour of GC alleles at synonymous sites, which is not compatible with the mutational bias model for the origin of isochores in mammals [44,45•].

The trouble is that selection is not the only possible interpretation of a fixation bias: the alternative neutral model of biased gene conversion (BGC) can also generate fixation biases. BGC is linked to the process of meiotic recombination (Figure 3). Like selection, the impact of BGC on the fixation process depends on the population size [46]. It also depends on the rate of recombination and on the bias in the repair of DNA mismatches. This model is not new (e.g. see [46]), but it has been put forward recently to explain the evolution of base composition [37••,47•,48••].

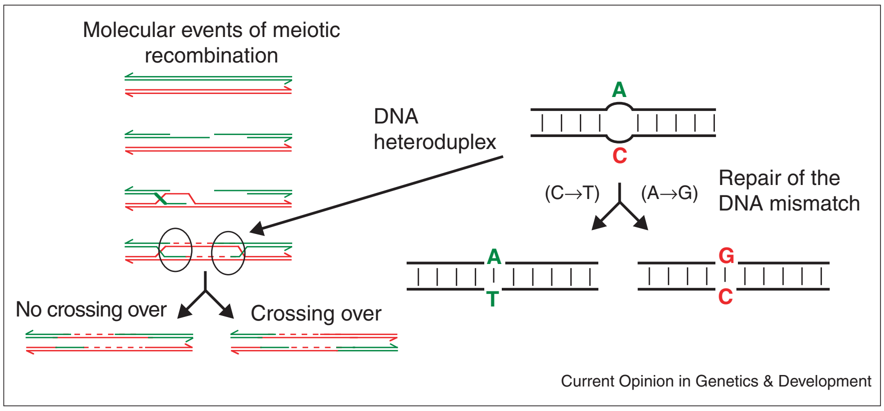
Several arguments support the hypothesis that BGC is a widespread phenomenon [47•,48••]. First, experimental evidence indicates that the repair of DNA mismatches is biased in favour of GC in many taxa (bacteria, archaea and eukaryotes) including mammals, birds, *Xenopus* and yeast. This GC bias in mismatch repair might correspond to an evolutionary adaptation to a universal mutational bias towards AT [48••]. Second, as predicted by this model, there is a positive correlation between recombination rates and GC content, not only in mammals and yeast, but also in *C. elegans* and *Drosophila* [47•,48••]. In addition, gene families that frequently undergo ectopic gene conversions

Figure 2



Gene expression pattern and AT skew in human. The AT skew [(A–T)/(A+T)] was measured at the third codon positions of fourfold degenerate codons **(a)** and in introns **(b)** from 1396 complete human protein-encoding genes (containing introns) taken from [14]. In both cases (a,b), there is a relative excess of thymine in genes that have a broad tissue distribution. Expression patterns were measured as in Figure 1. Genes were divided into nine groups of equal sample size according to the number of tissues where they were expressed. Error bars represent the 95% confidence interval (L Duret, D Mouchiroud, unpublished data).

Figure 3


Molecular events of meiotic recombination
DNA heteroduplex
A
C
(C→T)
(A→G)
Repair of the DNA mismatch
A
T
G
C
No crossing over
Crossing over
Current Opinion in Genetics & Development

Biased gene conversion. During the process of meiotic recombination (which may or may not lead to a crossing-over), stretches of heteroduplex DNA are formed with one strand coming from the maternal chromosome and the other coming from the paternal chromosome. When a heterozygous site is involved in such a heteroduplex, this results in a DNA mismatch that may be repaired by one of the cellular DNA repair systems. This repair will result in the loss (i.e. conversion) of one of the two alleles. This process is said to be biased if the probability of conversion is not symmetric – for example, if GC alleles convert AT alleles more frequently than the reverse. Biased gene conversion increases the probability of fixation of the allele that is favoured by the bias in the repair of mismatches. Like selection, the impact of biased gene conversion on the fixation process depends on the population size. It also depends on the rate of recombination (more precisely on the rate of formation of DNA heteroduplex), the size of DNA heteroduplex and the bias in the repair of mismatches [46,47•].

are GC-rich ([47•]; e.g. see the codon usage of histone genes in various eukaryotes in Figure 1 of [23•]). Third and last, calculations show that the range of parameters for which BGC does affect the evolution of base composition is compatible with known biological and population parameters [47•].

Are there alternative explanations for the correlation between GC content and recombination? It has been shown that recombination can be mutagenic, and my co-workers and I have postulated [49••] that such recombination-linked mutations might be biased towards GC. But experimental data contradict this hypothesis [48••]. Population genetics models also predict a positive correlation between selection efficacy (notably on codon usage) and the recombination rate. This is because the action of selection at a given locus interferes with the action of selection at other genetically linked loci (Hill–Robertson interference [HRi]). Such an effect can complicate the distinction between selective and neutral models: both HRi and BGC models predict a correlation between GC3 and recombination because optimal codons often end in guanine or cytosine.

But the impact of HRi on codon usage seems to be limited. As mentioned previously, the BGC model predicts a positive correlation between recombination rate and GC-content. Conversely, the HRi model predicts a positive correlation between recombination rate and the frequency of optimal codons, whatever their GC-content (i.e. HRi predicts a positive correlation with recombination, both for AU-ending and GC-ending optimal codons). Both in yeast and *C. elegans*, observations are in contradiction with the predictions of the HRi model. In yeast, recombination is correlated with GC3, but not with the frequency of optimal codons [48••]. In *C. elegans*, recombination is correlated positively with the frequency of GC-ending optimal codons, but negatively with the frequency of AU-ending optimal codons [49••,50•]. Thus, there is no evidence for an impact of HRi on codon usage in these two species, and BGC seems to be the only explanation for the observed correlation between recombination rate and GC3 content. HRi seems to have an effect in *Drosophila* [32,51], but it is limited to a small subset of genes (only 4% of the total) that are subject to strong translational selection and located in regions of very low recombination [50•,52•].

Note that, again, the two models are not exclusive: a genome can be subject to both BGC and HRi. But HRi should affect only sites that are under selective pressure, whereas BGC should affect all positions in the genome. In agreement with this latter prediction of the BGC model, the GC content of noncoding regions (e.g. introns and flanking regions) is positively correlated with recombination rate in both *Drosophila* and *C. elegans* [49••,50•]. Thus, it seems clear that BGC also occurs in *Drosophila*. This means that the fixation biases in favour of GC synonymous sites observed in *Drosophila* [26••] do not necessarily reflect the action of selection totally, in particular for the genes that are located in regions of high recombination rate.

Conclusions

Studies on synonymous codon usage show that even minuscule phenotypic effects can be subject to natural selection in metazoan species such as *C. elegans* or *Drosophila*. In vertebrates, the main determinant of codon usage is the GC content of the isochore where the gene is located, and not translational selection. After taking into account the effect of isochores on base composition, it is possible to detect a significant relationship between codon usage and gene expression [33••]; however, it is not yet established whether this relationship is a consequence of translational selection or a consequence of a transcription-coupled mutational bias.

These analyses illustrate the power and complexity of comparative sequence analysis. But the detection of tiny selective effects is hindered by two problems. First, the enormous volume of data allows the detection of even very weak correlations, such results are highly sensitive to any methodological artefact or sampling bias. In addition, because the effects are weak, the risk of indirect correlations increases with the number of parameters that have to be considered. Thus, one must be careful not to overinterpret weak correlations. Second, all inferences of selection are based on the rejection of the null hypothesis of neutral evolution. As we have seen above, even in the absence of selection many factors affect the evolution of sequences. For example, mutation patterns vary within a genome, within a gene, and according to local context (e.g. dinucleotides) and the level of transcription. In addition, the central prediction of the neutral theory — that is, at equilibrium, substitution and mutation patterns should be identical — can be confounded by recent changes in mutational properties and by the process of biased gene conversion. As we have seen, the impact of BGC on genome evolution seems to be widespread. In conclusion, identifying all of the footprints of selection will require both a careful analysis of mutational processes and the impact of BGC to be taken into account in models of population genetics.

Acknowledgements

I thank Dominique Mouchiroud, Gabriel Marais, Nicolas Galtier, Gwenael Piganeau and Manolo Gouy for their friendly and fruitful collaboration. This work is supported by the Centre National de la Recherche Scientifique (CNRS).

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Tautz D: **A genetic uncertainty problem.** *Trends Genet* 2000, **16**:475-477.

2. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H *et al.*: **Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis.** *Science* 1999, **285**:901-906.

3. Thatcher JW, Shaw JM, Dickinson WJ: **Marginal fitness contributions of nonessential genes in yeast.** *Proc Natl Acad Sci USA* 1998, **95**:253-257.

4. Otto SP: **Detecting the form of selection from DNA sequence data.** *Trends Genet* 2000, **16**:526-529.

5. Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF: **DNA sequence evolution: the sounds of silence.** *Philos Trans R Soc Lond Ser B* 1995, **349**:241-247.

6. Akashi H: **Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA.** *Genetics* 1995, **139**:1067-1076.

7. Akashi H, Schaeffer SW: **Natural selection and the frequency distributions of 'silent' DNA polymorphism in *Drosophila*.** *Genetics* 1997, **146**:295-307.

8. • Francino MP, Ochman H: **Deamination as the basis of strand asymmetric evolution in transcribed *Escherichia coli* sequences.** *Mol Biol Evol* 2001, **18**:1147-1150.

This paper shows that strand-specific mutation patterns occur in transcribed sequences, probably as a consequence of cytosine deamination in single-stranded DNA. It is essential to take into account such transcription-coupled mutational biases when interpreting correlations between codon usage and gene expression.

9. Bulmer M: **The selection–mutation–drift theory of synonymous codon usage.** *Genetics* 1991, **129**:897-907.

10. • Smith NG, Eyre-Walker A: **Why are translationally sub-optimal synonymous codons used in *Escherichia coli?*** *J Mol Evol* 2001, **53**:225-236.

The authors show that three mechanisms contribute to the existence of translationally suboptimal codons in *E. coli*: first, the balance between mutation, selection and drift; second, relaxation of translational selection; and third, translationally suboptimal codons can be favoured by alternative selective pressure.

11. Blencowe BJ: **Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases.** *Trends Biochem Sci* 2000, **25**:106-110.

12. Hurst LD, Pal C: **Evidence for purifying selection acting on silent sites in BRCA1.** *Trends Genet* 2001, **17**:62-65.

13. Fedorov A, Saxonov S, Gilbert W: **Regularities of context-dependent codon bias in eukaryotic genes.** *Nucleic Acids Res* 2002, **30**:1192-1197.

14. Duret L, Hurst LD: **The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution.** *Mol Biol Evol* 2001, **18**:757-762.

15. • Morton BR: **Selection at the amino acid level can influence synonymous codon usage: implications for the study of codon adaptation in plastid genes.** *Genetics* 2001, **159**:347-358.

In most cases, synonymous changes can be only transitions and not transversions. If the GC↔AT mutation pressure is different for transitions and transversions, then sites where selection limits substitution to transitions will have a GC content that differs from that of neutral sites. This phenomenon can affect the base composition not only at twofold degenerate sites, but also at fourfold degenerate sites. Thus, even if there is no selective difference between synonymous codons, synonymous codon usage may vary according to the strength of selection acting at non-synonymous sites.

16. Shields DC, Sharp PM, Higgins DG, Wright F: **'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons.** *Mol Biol Evol* 1988, **5**:704-716.

17. Stenico M, Lloyd AT, Sharp PM: **Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases.** *Nucleic Acids Res* 1994, **22**:2437-2446.

18. Duret L, Mouchiroud D: **Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*.** *Proc Natl Acad Sci USA* 1999, **96**:4482-4487.

19. Castillo-Davis CI, Hartl DL: **Genome evolution and developmental constraint in *Caenorhabditis elegans*.** *Mol Biol Evol* 2002, **19**:728-735.

20. Moriyama EN, Powell JR: **Codon usage bias and tRNA abundance in *Drosophila*.** *J Mol Evol* 1997, **45**:514-523.

21. Duret L: **tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes.** *Trends Genet* 2000, **16**:287-289.

22. •• Percudani R: **Restricted wobble rules for eukaryotic genomes.** *Trends Genet* 2001, **17**:133-135.

The author shows that the 'classical' wobble rules cannot be applied directly in eukaryotes, where the decoding system seems to have evolved towards a more restricted use of wobbling and thus towards an expansion of the number of tRNA species. He proposes an empirical 'parsimony of wobbling' criterion to assign the decoding properties of individual tRNAs, in species for which the genome has been sequenced completely. Correcting the assignments made in [21], he confirms (and indeed reinforces) the observed correlation between tRNA abundance and codon usage bias in *C. elegans*.

23. • Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T: **Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis.** *J Mol Evol* 2001, **53**:290-298.

The number of tRNA genes in a genome can be used to estimate the relative abundance of cellular tRNAs. The authors find a good correspondence between codon bias and tRNA gene numbers in *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae* and *C. elegans*, but not in *D. melanogaster* and human.

24. • Dunn KA, Bielawski JP, Yang Z: **Substitution rates in *Drosophila* nuclear genes: implications for translational selection.** *Genetics* 2001, **157**:295-305.

Contrary to the prediction of the translational selection model, no correlation is found between codon bias and synonymous substitution rate in *Drosophila*. The reasons for this unexpected observation are not yet established.

25. • McVean GA, Vieira J: **Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*.** *Genetics* 2001, **157**:245-257.

The authors develop a combination of population genetics models and likelihood methods to infer the strength of selection on synonymous sites from the pattern of synonymous substitution. They find evidence for considerable variations in selective pressures according to the genes or according to the amino acids. They also reveal both long-term and short-term changes in the strength of selection in the *Drosophila* lineage. Note, however, that their method assumes that there is no variation in the mutation rate between genes or sites.

26. •• Begun DJ: **The frequency distribution of nucleotide variation in *Drosophila simulans*.** *Mol Biol Evol* 2001, **18**:1343-1352.

The author analyses polymorphism and divergence data in *Drosophila*. The frequency distribution of optimal and suboptimal polymorphic codons supports the hypothesis that suboptimal codons are slightly deleterious. But the frequency of optimal codons is not at equilibrium: in *D. simulans*, substitutions toward suboptimal codons have occurred more frequently than the reverse. This indicates a decrease in the strength of selection on codon usage in the *D. simulans* lineage or a decrease in population size.

27. Akashi H: **Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy.** *Genetics* 1994, **136**:927-935.

28. Marais G, Duret L: **Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*.** *J Mol Evol* 2001, **52**:275-280.

29. Carlini DB, Chen Y, Stephan W: **The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the Drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*.** *Genetics* 2001, **159**:623-633.

30. Lobry JR: **Properties of a general model of DNA evolution under no-strand-bias conditions.** *J Mol Evol* 1995, **40**:326-330.

31. Moriyama EN, Powell JR: **Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*.** *Nucleic Acids Res* 1998, **26**:3188-3193.

32. Comeron JM, Kreitman M, Aguade M: **Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*.** *Genetics* 1999, **151**:239-249.

33. •• Musto H, Cruveiller S, D'Onofrio G, Romero H, Bernardi G: **Translational selection on codon usage in *Xenopus laevis*.** *Mol Biol Evol* 2001, **18**:1703-1707.

In vertebrates, multivariate analyses of codon usage show that the most variation in codon bias is explained by GC content at the third codon position (see also [23•]), and thus by the GC content of the isochore where the gene is located. This study shows that in *Xenopus* the second axis of correspondence analysis is correlated with the level of gene expression. This is the first demonstration that codon usage is correlated with gene expression in a vertebrate (after the effect of the isochore has been taken into account). It is, however, not yet established whether this relationship is a consequence of translational selection (as proposed by the authors) or of a transcription-coupled mutational bias.

34. Konu O, Li MD: **Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents.** *J Mol Evol* 2002, **54**:35-41.

35. • Margulies EH, Kardia SL, Innis JW: **Identification and prevention of a GC content bias in SAGE libraries.** *Nucleic Acids Res* 2001, **29**:E60.

This paper shows that the serial analysis of gene expression (SAGE) technique tends to underestimate the expression level of AT-rich genes, because the melting temperature of the short oligonucleotide tags depends on their base composition. This methodological bias should be taken into account when investigating the relationship between codon usage and gene expression.

36. Bernardi G: **Isochores and the evolutionary genomics of vertebrates.** *Gene* 2000, **241**:3-17.

37. •• Eyre-Walker A, Hurst LD: **The evolution of isochores.** *Nat Rev Genet* 2001, **2**:549-555.

An excellent synthesis of the strengths and weaknesses of the different models that have been proposed for the evolution of isochores in vertebrates.

38. •• International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.

A fascinating work and an excellent overview of the organization and molecular evolution of the human genome. Among many other results, there is a good description of the variability in base composition along the genome. Notably, the analysis of substitution patterns in defective transposons suggests that the base composition of the human genome is not at equilibrium. The authors find no obvious relationship between codon usage and the number of tRNA genes in the genome.

39. Iida K, Akashi H: **A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes.** *Gene* 2000, **261**:93-105.

40. Kliman RM, Eyre-Walker A: **Patterns of base composition within the genes of *Drosophila melanogaster*.** *J Mol Evol* 1998, **46**:534-541.

41. Wong GK, Wang J, Tao L, Tan J, Zhang J, Passey DA, Yu J: **Compositional gradients in Gramineae genes.** *Genome Res* 2002, **12**:851-856.

42. • Urrutia AO, Hurst LD: **Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection.** *Genetics* 2001, **159**:1191-1199.

The authors propose a new measure of codon bias, MCB, that takes into account the variations of base composition that occur along genomes (e.g. isochores). In the human genome, they find a correlation between MCB and patterns of gene expression. But the variance of the MCB measure depends on the length of genes, and the authors show that this methodological artefact is responsible for the weak correlation that they detect. They conclude that there is an absence of evidence for translational selection in humans.

43. Ponger L, Duret L, Mouchiroud D: **Determinants of CpG islands: expression in early embryo and isochore structure.** *Genome Res* 2001, **11**:1854-1860.

44. Eyre-Walker A: **Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA.** *Genetics* 1999, **152**:675-683.

45. • Smith NG, Eyre-Walker A: **Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans.** *Mol Biol Evol* 2001, **18**:982-986.

The analysis of polymorphisms at synonymous sites indicates that in GC-rich genes, GC→AT mutations are more frequent than are AT→GC mutations, whereas interspecies comparisons suggest that AT↔GC substitutions are at equilibrium. Taken together, these data indicate that GC mutations have a higher probability of fixation than have AT mutations. This result can be explained by selection or biased gene conversion but is not compatible with the equilibrium mutational bias model for the origin of isochores. Note, however, that the assumption of equilibrium might be incorrect (see [38••]).

46. Nagylaki T: **Evolution of a finite population under gene conversion.** *Proc Natl Acad Sci USA* 1983, **80**:6278-6281.

47. • Galtier N, Piganeau G, Mouchiroud D, Duret L: **GC-content evolution in mammalian genomes: the biased gene conversion hypothesis.** *Genetics* 2001, **159**:907-911.

See annotation [48••].

48. •• Birdsell JA: **Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution.** *Mol Biol Evol* 2002, **19**:1181-1197.

This paper shows that in yeast the GC content at silent sites is correlated positively with recombination rate. Neither selection nor mutation can explain this correlation. In agreement with a similar analysis in mammals [47•], the author concludes that this correlation results from biased gene conversion (BGC). In many taxa (including yeast and mammals), the repair of DNA mismatches is biased toward GC, which probably corresponds to an evolutionary adaptation to a universal mutational bias towards AT. In recombining genomes, this bias in mismatch repair will lead to gene conversion biases. Together with [47•], this paper highlights the importance of taking into account BGC in models of population genetics.

49. •• Marais G, Mouchiroud D, Duret L: **Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes.** *Proc Natl Acad Sci USA* 2001, **98**:5688-5692.

Population genetics models show that the efficacy of selection should increase with recombination because of Hill–Robertson interference (HRi). In agreement with that prediction, the frequency of optimal codons is correlated positively with recombination both in nematode and *Drosophila*. In this paper, however, my co-workers and I show that in both species this correlation is essentially due to a relationship between recombination rate and GC content at silent sites, such as introns and flanking regions (note that we had proposed that this was a consequence of mutation patterns associated with recombination but biased gene conversion seems to be a more likely explanation [48••]). When the effect of recombination on local GC content is taken into account, there is no evidence for an impact of HRi on codon usage in the nematode [50•]. In *Drosophila*, HRi seems to have some impact on codon usage but is limited to a small subset of genes (only 4% of the total) that are subject to strong translational selection and located in regions of very low recombination [50•,52•].

50. Marais G, Piganeau G: **Hill–Robertson interference is a minor**
• **determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genomes.** *Mol Biol Evol* 2002, **19**:1399-1406.

See annotation [49••].

51. Kliman RM, Hey J: **Reduced natural selection associated with low recombination in *Drosophila melanogaster*.** *Mol Biol Evol* 1993, **10**:1239-1258.

52. Hey J, Kliman RM: **Interactions between natural selection,**
• **recombination and gene density in the genes of *Drosophila*.** *Genetics* 2002, **160**:595-608.

See annotation [49••].

53. Duret L, Mouchiroud D: **Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate.** *Mol Biol Evol* 2000, **17**:68-74.