Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion

Nicolas Galtier, Camille Roux, Marjolaine Rousselle, Jonathan Romiguier, Emeric Figuet, Sylvain Glemin, Nicolas Bierne, L. Duret

Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion

Nicolas Galtier,*[,1] Camille Roux,[1,2,3] Marjolaine Rousselle,[1] Jonathan Romiguier,[1,2] Emeric Figuet,[1] Sylvain Glémin,[1,4] Nicolas Bierne,[1] and Laurent Duret[5]

[1]UMR5554, Institut des Sciences de l'Evolution, University Montpellier, CNRS, IRD, EPHE, Montpellier, France

[2]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

[3]UMR 8198 – Evo-Eco-Paleo, CNRS, Université de Lille—Sciences et Technologies, Villeneuve d'Ascq, France

[4]Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[5]Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, CNRS, Université de Lyon, Université Lyon 1, Villeurbanne, France

***Corresponding author:** E-mail: nicolas.galtier@univ-montp2.fr.

Associate editor: Nadia Singh

Abstract

Selection on codon usage bias is well documented in a number of microorganisms. Whether codon usage is also generally shaped by natural selection in large organisms, despite their relatively small effective population size (N_e), is unclear. In animals, the population genetics of codon usage bias has only been studied in a handful of model organisms so far, and can be affected by confounding, nonadaptive processes such as GC-biased gene conversion and experimental artefacts. Using population transcriptomics data, we analyzed the relationship between codon usage, gene expression, allele frequency distribution, and recombination rate in 30 nonmodel species of animals, each from a different family, covering a wide range of effective population sizes. We disentangled the effects of translational selection and GC-biased gene conversion on codon usage by separately analyzing GC-conservative and GC-changing mutations. We report evidence for effective translational selection on codon usage in large-N_e species of animals, but not in small-N_e ones, in agreement with the nearly neutral theory of molecular evolution. C- and T-ending codons tend to be preferred over synonymous G- and A-ending ones, for reasons that remain to be determined. In contrast, we uncovered a conspicuous effect of GC-biased gene conversion, which is widespread in animals and the main force determining the fate of AT↔GC mutations. Intriguingly, the strength of its effect was uncorrelated with N_e.

Key words: **synonymous codon usage, GC-content, recombination, evolution, gene expression, nonmodel organisms.**

Article

Introduction

The reasons why synonymous codons do not occur at equal frequencies in protein coding sequences have been puzzling molecular evolutionary researchers for decades (Duret 2002; Hershberg and Petrov 2008). One fascinating aspect is the early discovery that, in various microbial genomes, codon usage responds to natural selection. In *Escherichia coli* and *Saccharomyces cerevisiae*, for instance, the codons most commonly observed in highly expressed genes match the most abundant tRNAs in the cell (Ikemura 1985), strongly suggesting that codon usage and tRNA content have coevolved in a way that optimizes translation—hence the term "translational selection." These observations promoted codon usage bias as a textbook example of a weak selective pressure operating at molecular level, detectable from patterns of coding sequence variation, but difficult to apprehend experimentally.

Because selection on codon usage is presumably weak, other evolutionary forces might contribute to explain its variation across genes and genomes (Sharp and Li 1986). In particular, it is expected that, in small populations, the random fluctuations of allele frequencies due to genetic drift should decrease the efficiency of natural selection and lessen the efficiency of selection on codon usage. The theory predicts that if N_e, the effective population size, is sufficiently small such that the $4N_e s$ product is much <1, s being the selection coefficient in favor of optimal codons, then the effect of selection should be negligible. One question of interest, therefore, is whether selection shapes codon usage in large organisms, such as animals, the same way as in microbes, despite their presumably smaller N_e.

Evidence for selection on codon usage has been reported in fruit flies (Shields et al. 1988; Akashi 1994; Bierne and Eyre-Walker 2006), in the nematode *Caenorhabditis elegans* (Duret and Mouchiroud 1999), and in the branchiopod *Daphnia pulex* (Lynch et al. 2017). In contrast, codon usage in mammals is primarily governed by within-genome variation in GC-content, and only weakly, if at all, correlated to gene expression and tRNA content (Semon et al. 2006; Rudolph et al. 2016; Pouyet et al. 2017, see also Doherty and McInerney

2013). The effectiveness of selection on codon usage in small-sized invertebrates but not in large-sized vertebrates is superficially in agreement with the hypothesis of a N_e effect. Please note, however, that the population genetic analyses of codon usage bias so far have only been conducted in a relatively small number of species of animals.

Subramanian (2008) analyzed codon usage bias intensity across 20 species of eukaryotes and reported a higher bias in short generation time, presumably large-N_e species than in long generation time, presumably small-N_e ones. This author questioned the N_e hypothesis and rather suggested that s might vary among species. According to his hypothesis, the selective pressure on translation efficiency would be stronger in fast growing species. Interestingly, growth rate seems to be the main determinant of among-species variation in codon usage bias intensity in bacteria (Rocha 2004; Sharp et al. 2005; Vieira-Silva and Rocha 2010). On the other hand, Machado et al. (2017) argued that s is not necessarily constant across synonymous codons and mutations. Comparison of polymorphism and divergence patterns in *Drosophila melanogaster* indeed suggested that both strong ($4N_es \gg 1$) and weak ($4N_es \sim 1$) selection applies on synonymous sites in this species (Lawrie et al. 2013; Machado et al. 2017). It is also well established that in many species selection for optimal translation is stronger in highly expressed than in lowly expressed genes (Gouy and Gautier 1982; Duret and Mouchiroud 1999), and even in humans there is documented evidence for a phenotypic effect of specific synonymous mutations (Sauna and Kimchi-Sarfaty 2011). The question "does codon usage affect translational efficiency in species X" should therefore probably be rephrased as "what fraction of synonymous mutations in species X is effectively selected?," and "how does N_e influence this fraction?"

Besides selection and drift, patterns of codon usage might be influenced by neutral, directional forces such as mutation biases and GC-biased gene conversion (gBGC), a recombination-associated segregation bias that favors G and C over A and T alleles in high recombining regions (Duret and Galtier 2009; Mugal et al. 2015). The existence of gBGC has been experimentally demonstrated in yeast (Mancera et al. 2008; Lesecque et al. 2013), humans (Williams et al. 2015; Halldorsson et al. 2016), flycatcher (Smeds et al. 2016), and *Daphnia* (Keith et al. 2016). gBGC has been identified as the main driver of GC-content evolution in vertebrates (Duret and Galtier 2009; Figuet et al. 2014; Glémin et al. 2015; Bolívar et al. 2016) and several other taxa (Pessia et al. 2012; Glémin et al. 2014; Wallberg et al. 2015). In many respects, the expected impact of gBGC on patterns of sequence variation is similar to that of directional selection. For instance, the expected fate and frequency distribution of an allele promoted by gBGC is identical to that of a favorable allele under codominant selection (Nagylaki 1983). Importantly, the so-called "preferred" synonymous codons—that is, codons more frequently used in high-expressed genes—often end with C or G, *D. melanogaster* being an extreme example in which all the preferred codons are C- or G-ending (Duret and Mouchiroud 1999). This implies that the effects of translational selection and gBGC on synonymous positions can be very difficult to disentangle from coding sequence analysis only (Jackson et al. 2017, but see Clément et al. 2017). gBGC, however, is expected to apply more strongly to highly recombining regions, and to affect non coding, flanking sequences as well as coding ones, unlike translational selection.

Preferred codons are usually defined as codons used more frequently in high-expressed than in low-expressed genes (Duret and Mouchiroud 1999). This could be problematic in practice because DNA or cDNA libraries generated for high throughput sequencing are known to be biased with respect to sequence base composition (Dohm et al. 2008; Aird et al. 2011). The GC-richest and GC-poorest fractions of target DNA are typically underrepresented, and sequences of medium GC-content overrepresented, in Illumina data (Choudhari and Grigoriev 2017). This experimental bias, if not properly taken into account, could corrupt the definition of preferred codons in generating false correlations between codon usage and sequencing coverage. The bias apparently varies between experiments and libraries, which makes it difficult to model and correct for (Benjamini and Speed 2012). GC-content is therefore a potential confounder of analyses of selection on codon usage, both biologically and methodologically.

Current knowledge on codon usage biases in animals is therefore limited in at least two respects. First, published analyses have so far focused on a relatively small number of species—mainly model organisms—and an even smaller number of taxa—mainly drosophilids and vertebrates. Secondly, the confounding effects of gBGC and experimental biases have not always been taken into account. We therefore lack a global picture of the relative impact of selection, drift, and gBGC on codon usage evolution in animals. Here, we analyzed a data set covering 30 nonmodel species of animals. In each species, the transcriptomes of five to eleven diploid individuals plus one outgroup have been previously sequenced (Romiguier, Gayral, et al. 2014). In principle transcriptome-based population, genomic data are ideal for codon usage bias analysis in providing access to codon usage tables, gene expression level, allele frequencies at polymorphic positions, and flanking UTR sequences. We focused our analysis of translational selection on GC-conservative pairs of synonymous codons—that is, codons differing by a G↔C or an A↔T substitution—and separately analyzed the effect of gBGC. We found that translational selection on codon usage is only detectable in short-lived, large-N_e species, whereas gBGC is widespread across animals and of strength apparently independent of N_e.

Results

GC-Changing versus GC-Conservative Mutations

In each of the 30 focal species, we correlated GC12, GC3, and GC_UTR across genes. The three measures of GC-content were strongly correlated with each other in nearly all species (supplementary table S1, Supplementary Material online). Analyzing polymorphism data, we calculated the mean allele frequency of the G or C allele at synonymous, AT↔GC SNPs.

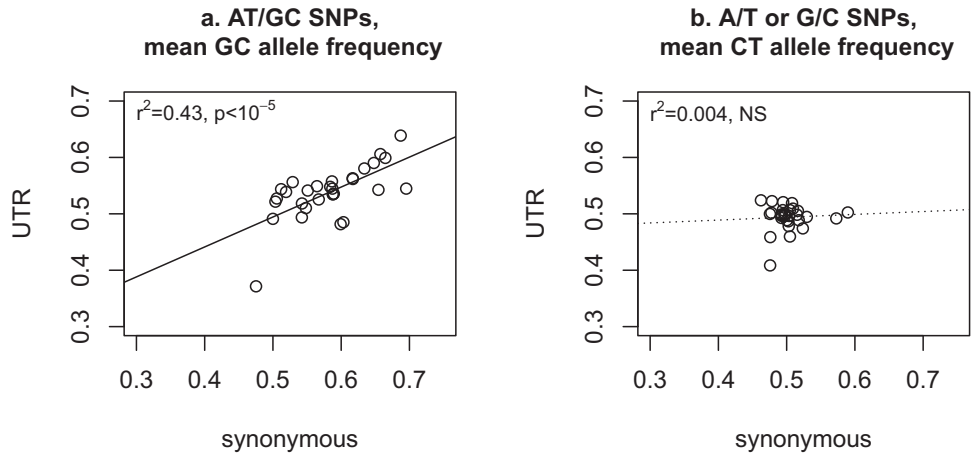

Fig. 1. Mean allele frequency at GC-changing and GC-conservative SNPs. Each dot represents a species. *X* axis: synonymous SNPs; *Y* axis: flanking noncoding SNPs. (*a*) Mean allele frequency of G or C alleles at AT versus GC SNPs. (*b*) Mean allele frequency of T or C alleles at A versus T and G versus C SNPs.

The mean frequency of GC alleles was >0.5 in a majority of species, which is not expected from sequences at mutational equilibrium (Glémin et al. 2015). We applied the exact same analysis to UTR sequences and found that the mean allele frequency of GC alleles in UTRs was strongly correlated to the mean allele frequency of GC alleles at synonymous sites across species (fig. 1*a*). This shows that the evolutionary fate of AT→GC and GC→AT mutations at third codon positions is primarily governed by forces similarly impacting noncoding DNA, that is, independent of selection on codon usage. We reproduced the analysis using G↔C and A↔T polymorphisms instead of AT↔GC ones and obtained a very different picture (fig. 1*b*): the frequency of C and T alleles showed no significant correlation between UTRs and third codon positions. These observations are consistent with the hypothesis that the GC-content of genes can be affected by gBGC, both in UTRs and at synonymous sites and hence that this process has to be taken into account to investigate signatures of selection on codon usage.

Furthermore, one key point for investigating translational selection is to correctly characterize preferred codons. The classical approach consists in identifying codons whose frequency increases with expression level. However, in a majority of species, we observed significant relationships, either positive or negative, between expression level (inferred from sequencing depth) and GC_UTR (supplementary table S1, Supplementary Material online). The hypothesis of selection on codon usage bias does not predict any relationship between GC_UTR and expression. As mentioned in Introduction, these correlations might reflect artefacts, resulting from the well documented GC-bias in Illumina libraries (Benjamini and Speed 2012). It is also possible that these correlations reflect covariations between expression level and recombination rate (and hence intensity of gBGC). For instance, in humans, intragenic recombination rates (and GC3) correlate negatively with expression level in meiotic cells (Pouyet et al. 2017), whereas in vertebrate species lacking PRDM9, on the contrary, recombination hotspots tend to be associated with promoters of active genes (Baker et al. 2017). In any case, be it artefactual or real, the correlation between GC-content and expression can confound the identification of preferred codons.

Preferred Synonymous Codons

We therefore decided to focus our analysis of codon usage on synonymous codon pairs of the form XYA/XYT or XYG/XYC. There are 17 such GC-conservative pairs of synonymous codons in the standard genetic code—two per 4- or 6-fold degenerate amino acid, and the isoleucine-coding ATA/ATT pair. For each pair of GC-conservative codon and each CDS, we calculated the relative frequency of the C- or T-ending codon as

$$f_{XYC}/(f_{XYC} + f_{XYG})$$

or

$$f_{XYT}/(f_{XYT} + f_{XYA}),$$

where f_{XYZ} is the frequency of occurrence of codon XYZ in the considered CDS (X, Y, and Z in {A; C; G; T}). We correlated these with gene expression level. The corresponding correlation coefficient, hereafter called r_{PYR} (for pyrimidine), were taken as measures of preference for codons XYC or XYT, compared with XYG or XYA. This was done independently in 17 GC-conservative codon pairs and 30 species, that is, 510 correlation analyses. Supplementary figure S1, Supplementary Material online, shows a heatmap of the distribution of r_{PYR} across codon pairs and species.

Of the 510 estimated r_{PYR}, 441 (86%) were positive, indicating a general preference for C over G and for T over A at synonymous positions in animals. We arbitrarily called "preferred" those codons for which r_{PYR} was >0.05 (or below −0.05) and significantly different from zero with a *P* value <0.001. We identified 117 preferred codons out of 510 codon pairs, of which 64 were C-ending, 43 were T-ending, seven were A-ending and three were G-ending. Pyrimidines tend to

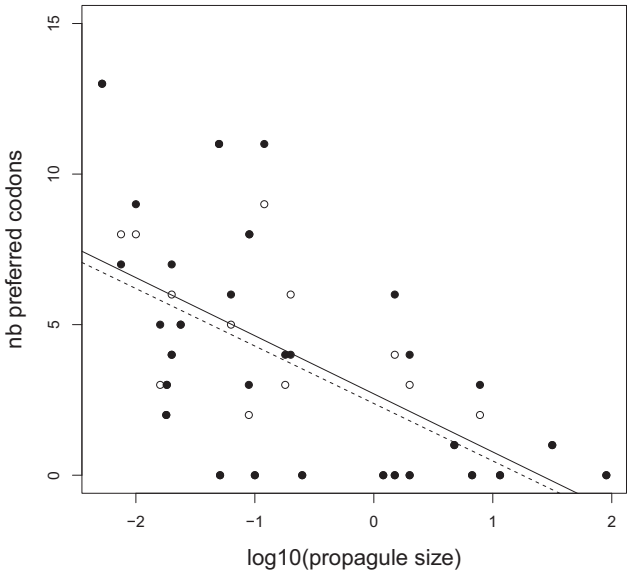

FIG. 2. Number of preferred codons among 17 GC-conservative codon pairs. Each dot is for a species. Codons are called preferred when their prevalence is correlated with gene expression with correlation coefficient >0.05 and *P* value <0.001. Plain dots: gene expression = sequencing depth of coverage. Open dots: gene expression = residual of the regression of sequencing depth of coverage on GC12.

be preferred over purines as far as GC-conservative synonymous codons are concerned in animals. The number of preferred codons varied across species from zero (in nine different species) to 13 (in *C. brenneri*), and was negatively correlated to species propagule size ($n = 30$, $r = -0.59$, $P = 0.00054$; fig. 2), suggesting that translational selection on codon usage is more efficient in large N_e species. This correlation was robust to a control for phylogenetic independence ($P = 0.006$).

We performed a similar analysis this time focusing on the 29 synonymous codon pairs of the form XYG/XYA or XYC/XYT. For each species and codon pair, we calculated the correlation coefficient between the frequency of occurrence of the G- or C-ending codon and gene expression level, r_{GC}, 80% of which were positive. The number of preferred codons varied between zero and 26 across species and was not correlated with species propagule size ($n = 30$, $r = -0.16$, NS). Large numbers of preferred XYG/XYA or XYC/XYT codons were found in Sauropsids *Chelonoidis nigra* (giant Galapagos turtle), *Aptenodytes patagonicus* (penguin), and *Cyanistes caeruleus* (great tit), a surprising result at odds with current knowledge on coding sequence evolution in vertebrates (Rao et al. 2011; Figuet et al. 2014).

In an attempt to explicitly account for the confounding effect of GC-content, we computed the residual of the regression of sequencing read depth on gene GC-content at first and second codon positions (GC12), here taken as a measure of gene expression level independent of GC-bias. We then correlated codon frequency to this measure of gene expression level to identify preferred codons, as described earlier. As far as GC-conservative codon pairs were concerned, the GC12-corrected measure of codon bias was very similar to the uncorrected one. The corrected and uncorrected numbers of preferred codons were almost perfectly correlated across species ($n = 30$; $r = 0.97$; $P < 10^{-15}$), and both were significantly correlated with species propagule size (fig. 2 and supplementary fig. S2, left, Supplementary Material online). XYG/XYA and XYC/XYT codon pairs, however, behaved quite differently. GC12-corrected and uncorrected numbers of preferred codons were less strongly correlated with each other ($n = 30$; $r = 0.73$; $P < 10^{-5}$), and not significantly related to propagule size (supplementary fig. S2, right, Supplementary Material online), confirming the confounding effect of GC-content. Given the sensitivity of XYG/XYA and XYC/XYT codon usage to GC-biases, we below restricted our analysis of codon preference on GC-conservative codon pairs.

Codon Usage Population Genetics

For each pair of GC-conservative, synonymous codons, we extracted the corresponding biallelic SNPs and recorded allele frequencies, independently in the 30 species. We first focused on the 117 preferred codons as defined above, and found that their average allele frequency, 0.546, was significantly >0.5 (standard error of the mean estimate: 0.015). In contrast, the mean allele frequency of codons weakly correlated to gene expression ($\text{abs}(r_{PYR}) < 0.05$ or $P > 0.001$) was 0.508 and not significantly different from 0.5.

In order to extract information from the entire data set, we recorded *MAJ*, the binary variable equal to one when the frequency of the C or T allele was >0.5 and zero when it was <0.5, SNPs at which allele frequency was exactly 0.5 being disregarded. This was done independently in the 30 species. Then we pooled the data across SNPs, codons, and species and performed a logistic regression of *MAJ* on the correlation between codon usage and expression, r_{PYR}. In this analysis, the number of data points equalled the total (across species and GC-conservative codon pairs) number of SNPs at which allele frequency was different from 0.5. Detailed data are provided in supplementary table S3, Supplementary Material online. The logistic regression is appropriate here because it accounts for unequal SNP sample size among species and codon pairs.

We found a significant, positive effect of r_{PYR} on *MAJ* ($P < 10^{-15}$), indicating that synonymous codons more commonly used in high-expressed genes tend to segregate at high population frequency, consistent with the hypothesis of translational selection on codon usage. The data set was split in four bins of species defined on the basis of propagule size, a variable negatively correlated to N_e (Romiguier, Gayral, et al. 2014), and the logistic regression was separately applied to the four bins. The relationship was strongly significant as far as the small propagule size bin was concerned ($P < 10^{-15}$), less strongly so in the second bin ($P = 6 \times 10^{-4}$), and not significant in the large propagule size bins. Translational selection on codon usage, therefore, is apparently affected by variations in N_e among species. When species were analyzed separately, a significant relationship between *MAJ* and r_{PYR} was detected in twelve species (supplementary table S1, Supplementary Material online). The mean propagule size across these twelve species was 0.69 mm, whereas the mean propagule size across the 18 species for which no significant effect was detected was

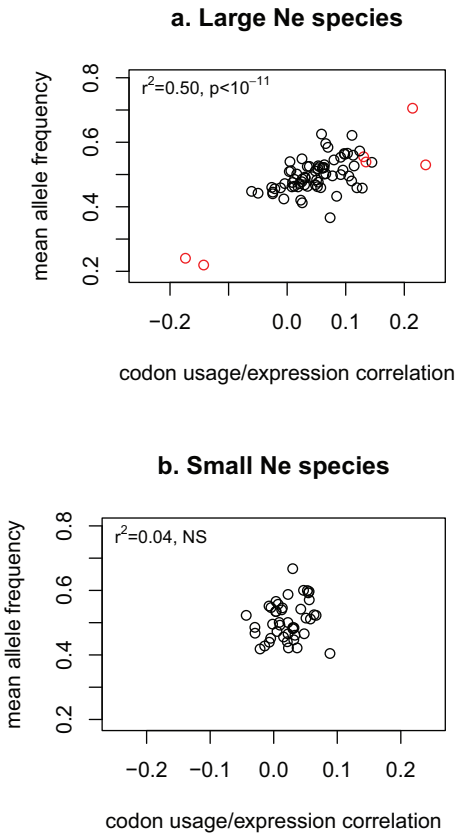

FIG. 3. Intensity of selection on codon usage bias. Each dot represents a particular pair of synonymous codon in a particular species. Only GC-conservative pairs of synonymous codons for which a minimum of 20 SNPS are available are considered. *X* axis: r_{PYR}, the correlation coefficient between C or T usage and gene expression. *Y* axis: mean frequency of the C or T allele. (*a*) Species in which propagule size is <0.2 mm. (*b*) Species in which propagule size is >2 mm. Red/darker dots correspond to the nematode *Caenorhabditis brenneri*.

8.5 mm. The two groups of species were also markedly different in terms of average longevity (7.0 vs. 36 years) and average π_N/π_S (0.089 vs. 0.16).

The N_e effect on codon usage bias is illustrated in figure 3*a*, in which we plotted the average allele frequency of XYC or XYT codons against r_{PYR}, separately in low propagule size (top) and high propagule size (bottom) species. Codon pairs for which <20 SNPs were available in the considered species were here excluded. Figure 3*a* shows a strong, positive correlation between codon usage bias and allele frequencies in low propagule size, large-N_e species. Red dots in figure 3*a* correspond to the nematode *C. brenneri*, in which codon usage bias is particularly pronounced (see supplementary fig. S1, Supplementary Material online). In contrast, no such relationship was uncovered in high propagule size, small-N_e species (fig. 3*b*).

Different tissues have been used for RNA extraction in distinct species of our sample, and this might affect our results. In particular, our estimate of gene expression level could be less relevant in species where RNA has been extracted out of a single tissue, compared with multiple tissues. To control for this problem, we focused on the subset of 20 species in which RNA had been extracted from at least four distinct tissues or the whole animal body (supplementary table S1, Supplementary Material online). We reproduced the above analyses and obtained very similar results (supplementary fig. S3, Supplementary Material online), indicating that our report of a link between propagule size, codon usage, and gene expression is not affected by tissue choice.

Insect Data Analysis

It has been suggested that growth rate, not N_e, could be the main driver of codon usage bias intensity across species (Subramanian 2008). The two effects are not easy to disentangle since N_e was found to be strongly correlated to life history traits related to growth rate, such as fecundity, longevity, and propagule size, in animals (Romiguier, Gayral, et al. 2014). To address this problem, we focused on insects and compared eusocial with solitary species. Eusocial species are characterized by a dramatic reduction in N_e, compared with solitary insects (Romiguier, Lourenço, et al. 2014). Eusocial and solitary insects, however, share similar cellular and developmental processes, so that the selective pressure for efficient protein translation can be assumed to be similar in the two groups of species.

We downloaded transcriptome data from 20 eusocial and 30 solitary insects and calculated in each species the 17 r_{PYR}, our measure of preference for codons XYC (respectively, XYT). About 691 (81%) of the 850 estimated correlation coefficients were positive, similarly to our main data set. We considered the 850 r_{PYR} as independent data points and tested the effect of eusociality on the absolute value of this variable (fig. 4*a*). We found that translational selection on codon usage is significantly stronger in solitary than in eusocial insects (*t*-test, $P < 10^{-15}$). Eusocial insects belong either to Hymenoptera (ants, eusocial bees, and eusocial wasps) or to Blattodea (termites). To control, for taxonomy, the data set was split in five categories: eusocial Hymenoptera, solitary Hymenoptera, eusocial Blattodea, solitary Blattodea, and other solitary insects (fig. 4*b*). The effect of eusociality on r_{PYR} was significant both within Hymenoptera (*t*-test, $P < 10^{-3}$) and within Blattodea (*t*-test, $P < 10^{-3}$).

GC-Biased Gene Conversion Analysis

In each species, synonymous SNPs were oriented by assuming that the outgroup species carries the ancestral allele. We focused on the 28 species for which at least 500 oriented synonymous SNPs were available. We calculated the frequency of the derived alleles and found that in 27 species out of 28, the average frequency of alleles resulting from an AT→GC mutation was above the average frequency of alleles resulting from a GC→AT mutation (supplementary table S1, Supplementary Material online), in agreement with the gBGC hypothesis. The unfolded SFS for AT→GC SNPs, GC→AT SNPs, and GC-conservative SNPs, respectively, were built in each species using all contigs. A mutation/drift/gBGC model was fitted to the three SFS (Glémin et al. 2015). A significantly positive segregation bias in favor of GC alleles was detected in 19 species out of 28 (supplementary table S1, Supplementary Material online and fig. 5*a*). These

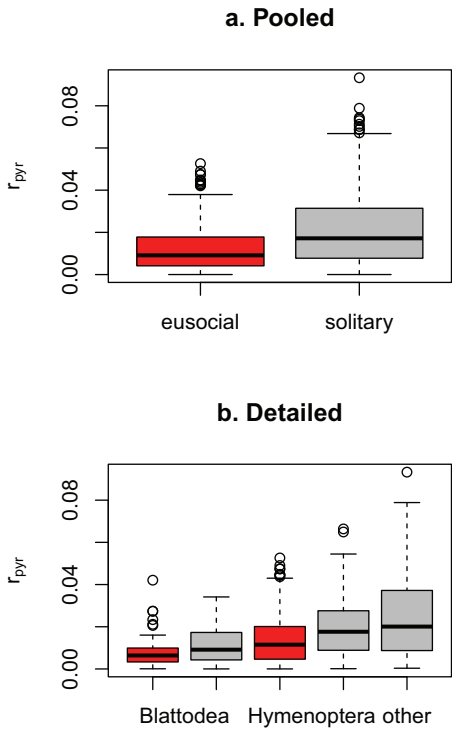

FIG. 4. Codon usage bias in eusocial versus solitary insects. (*a*) Distribution of the absolute value of r_{PYR}, the correlation coefficient between C- or T-ending codon frequency and gene expression, in 20 species of eusocial insects versus 30 species of solitary insects. (*b*) Eusocial species are split in Hymenoptera versus Blattodea; solitary species are split in Hymenoptera versus Blattodea versus other.

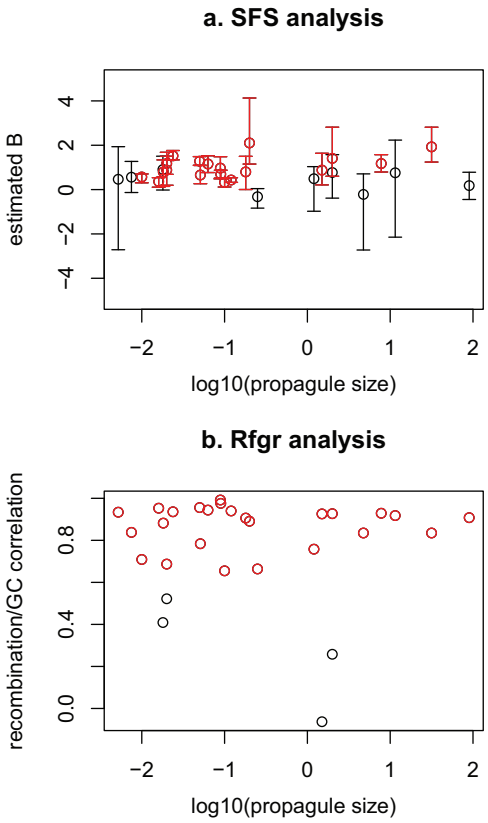

FIG. 5. Evidence for GC-biased gene conversion. Each dot represents a species. Top: SFS analysis; *X* axis: propagule size (log scale); *Y* axis: estimated scaled gBGC coefficient *B*; red/darker dots: significant, positive *B*; other dots: *B* not significantly different from zero. Error bars reflect 95% confidence intervals obtained by bootstrapping SNPs; Bottom: GC/recombination relationship; *X* axis: propagule size (log scale); *Y* axis: correlation coefficient between bins of R_{fgr} and bins of GC; red/darker dots: significant, positive correlation coefficient; other dots: correlation coefficient not significantly different from zero.

species belong to six of the eight metazoan phyla that were sampled. The estimated scaled gBGC coefficient, *B*, was not significantly correlated with π_N/π_S, propagule size, or longevity across species. SNP sample size was limiting in some species, as illustrated by error bars around estimated *B* in figure 5a. Still, even when only species with a narrow confidence interval are considered, no relationship between propagule size, and *B* is detectable.

We investigated whether the GC-bias likely reflects the action of gBGC by analyzing the impact of recombination rate. We binned contigs according to their GC-content, and measured the average recombination rate (R_{fgr}) within each bin. We then correlated average R_{fgr} to average GC-content across bins in 29 species for which at least 800 loci were eligible for recombinant haplotype detection. The correlation coefficient was positive in 28 species, >0.75 in 21 species and significantly positive ($P < 0.05$) in 25 species from eight distinct phyla (supplementary table S1, Supplementary Material online and fig. 5*b*). In one species (harvester ant *M. capitatus*), the correlation coefficient was negative but not significantly different from zero. The correlation coefficient between R_{fgr} and GC-content was not significantly correlated with propagule size (fig. 5*b*), longevity or π_N/π_S across species. Finally, we calculated the average derived allele frequency (DAF) of AT→GC, GC→AT, and GC-conservative mutations in distinct bins of contigs. We found that the DAF of AT→GC mutations increased with contig GC-content, again in agreement with the gBGC hypothesis (supplementary fig. S4, Supplementary Material online). The DAF of GC→AT mutations showed the reverse pattern—although less markedly—and the DAF of GC-conservative mutations was unaffected by contig GC-content.

Discussion

Selection versus gBGC: Methodological Aspects

Disentangling the effect of natural selection from that of neutral forces such as gBGC is a difficult task (Ratnakumar et al. 2010). Clément et al. (2017) introduced a model of codon usage accounting both for gBGC, which is assumed to affect AT↔GC mutations, and selection, which is assumed to affect mutations between preferred and non preferred codons. Their approach was successful in discriminating between the two evolutionary forces in eleven species of plants (Clément et al. 2017). Here, we faced the additional problem that, in a substantial number of species, gene GC content was correlated to gene expression irrespective of codon usage, so that even defining preferred codons was challenging. We

therefore addressed the problem by restricting our analysis of selection on codon usage to GC-conservative pairs of synonymous codons, which are supposedly unaffected by gBGC. One consequence is that we did not analyze 2-fold degenerate codons, and among 4- and 6-fold did not compare all synonymous codons with each other, thus potentially missing a part of the signal. Of note, the classical approaches, which focus on the comparison between preferred and nonpreferred codons, also have their limitations in that, as far as 4- and 6-fold codons are concerned, the nonpreferred category is a mixture of several codons between which no distinction is made. For instance, published analyses of codon usage bias in *Drosophila* have hardly considered the preferences between T-ending over A-ending codons at 4-fold sites, since in this group all such codons fall in the nonpreferred category (but see Zeng 2010).

Selection on Codon Usage in Animals: A Global Picture

We report a significant effect of translational selection on codon usage in animals. Codons showing a higher prevalence in highly expressed genes tend to segregate at higher population frequency. We found that C-ending codons tend to be preferred over G-ending codons, and T-ending codons over A-ending ones. This is, to our knowledge, the first report of a general preference for pyrimidines over purines at third codon positions in animals. We checked from previously published data (Duret and Mouchiroud 1999; Lynch et al. 2017) that the trend is also found in *C. elegans* (with the same exceptions as in *C. brenneri*), *D. melanogaster*, and *D. pulex*. We would expect codon preference to be quite stable over evolutionary times since switching to a new preferred codon should impose a high genetic load by simultaneously modifying the selection coefficient at many synonymous positions. This does not explain why pyrimidines would be preferred over purines, though. In *C. elegans* and perhaps more generally, C-ending and T-ending synonymous codons are translated by the same tRNA—the so-called wobble effect—whereas each A-ending and G-ending codon has its specific tRNA (Duret 2000; Percudani 2001). For this reason, one should probably expect correlated preferences for C and T at third codon positions—that is, a frequent usage of both C- and T-ending codons when their shared tRNA is abundant, infrequent usage otherwise. Again, we see no obvious reason why the existence of a shared tRNA for C- and T-ending codons would explain that these are generally favored over A- and G-ending codons.

The effect of translational selection on GC-conservative codons is significant but weak, and only detectable in a subset of species and codon pairs. This is perhaps surprising knowing that strong phenotypic effects of codon usage on expression levels of single genes have been experimentally reported in various systems, including fruit flies (Carlini et al. 2001; Carlini and Stephan 2003). Our approach, however, relies on polymorphic sites and can only detect relatively weak effects—sufficiently weak such that deleterious alleles are segregating in natural populations. Our results are indeed consistent with the existence of a broad distribution of fitness effect of synonymous mutations. In *C. brenneri*, for instance, no biased usage or skewed allele frequency distribution was detected for the proline-coding CCC versus CCG pair, whereas strong effects were detected for, for example, GCT versus GCA (Ala) and CGT versus CGA (Arg). Besides such differences between synonymous codon pairs, the effect of a particular type of synonymous mutation should also vary depending on which gene and which position is affected (Akashi 1994; Zhou et al. 2009; Machado et al. 2017), whereas we are here measuring the average strength of selection across sites, for any given pair of synonymous codons.

Our analysis specifically targets selective effects on codon usage that are related to gene expression and the efficacy of translation—either fidelity or speed. It is important to keep in mind that third codon positions can be affected by translation-independent selective pressures, for example, owing to mRNA splicing (Wu and Hurst 2015). These could in principle be distinguished by separately analyzing codons near versus far away from intron/exon boundaries. However, information on the position of introns is lacking in most of the species we have analyzed.

gBGC Is Widespread across the Metazoan Phylogeny

We detected a significant effect of gBGC in a majority of species of the data set. gBGC manifested itself via a higher average allele frequency of GC over AT alleles both at synonymous and flanking regions, a significant difference between AT→GC and GC→AT oriented SFSs, and a correlation between the long-term recombination rate and GC-content. Here, scaled recombination rate was approached at contig level using an approximate method derived from the four-gamete rule. The approach is suboptimal in several respects. First, we analyze spliced sequences and have no information on intron length, so that normalization by contig length is inexact. Secondly, we analyzed unphased data, thus loosing power compared with data sets consisting in experimentally phased haplotypes. Thirdly, the calculation only partially accounts for allele frequencies and the probability of detecting recombinant haplotypes when they exist—for example, R_{fgr} can only increase as sample size increases. Despite these many approximations, a strong and significant correlation between R_{fgr} and GC-content was identified in >80% of the species we sampled, which is indicative of a prominent and widespread effect of gBGC in animals. SFS analysis corroborated this finding in uncovering a significant segregation bias in favor of G and C alleles in a majority of species. Among the 29 species for which sufficient polymorphism data was available, 28 yielded evidence for gBGC in either the SFS or the recombination rate analysis—only in the oyster Ostrea *edulis* did both approaches fail to identify a significant signal. Of note, our SFS analysis captures the effect of both gBGC and, potentially, selection on GC-ending versus AT-ending codon usage. UTR sequence analysis demonstrates the impact of gBGC (fig. 1), but selection on GC-changing synonymous mutations might also be at work in some or many of the analyzed species.

In animals, gBGC had so far been identified in vertebrates (Figuet et al. 2014), bees, and ants (Kent et al. 2012;

Wallberg et al. 2015), and *Daphnia* (Keith et al. 2016), but not in *D. melanogaster* (Robinson et al. 2014), albeit on the X chromosome (Galtier et al. 2006; Haddrill and Charlesworth 2008). We here considerably expand the range of species and taxa in which gBGC is documented, adding annelids, echinoderms, tunicates, nemertians, cnidarians, lepidopterans, gastropod and bivalve molluscs, decapod and isopod crustaceans. gBGC is obviously widespread among animals. It significantly impacts the population frequency and fixation probability of AT↔GC mutations in a majority of species and should be considered as a potential confounder of molecular evolutionary studies, particularly studies of molecular adaptation, not only in mammals and vertebrates (Ratnakumar et al. 2010; Corcoran et al. 2017) but more generally in Metazoa. This study adds to the growing evidence that gBGC is a nearly universal process affecting a wide range of organisms (Pessia et al. 2012; Long et al. 2018).

Why a N_e Effect on Codon Usage Bias but Not on gBGC?

We detected evidence for translational selection on codon usage only in the small propagule size, large-N_e fraction of the species we sampled. The estimated efficiency of selection was strongest in the nematode *C. brenneri*, whereas no evidence for translational selection on codon usage was found in large mammals, birds, reptiles. This is in agreement with the nearly neutral theory of molecular evolution (Ohta and Gillespie 1996). Of note, the distinction between small-N_e and large-N_e species does not perfectly fit the vertebrates/invertebrates contrast: we detected significant evidence for translational selection on codon usage in common vole *Microtus arvalis*, but not in cuttlefish *Sepia officinalis*, for instance.

It has been suggested that among taxa variation in codon usage bias intensity is determined by variation in selective pressure, not N_e, with short generation time, high growth rate species having stronger requirement for efficient protein synthesis (Subramanian 2008). To distinguish between the two hypothesis, we compared codon usage bias in eusocial versus solitary insects, which differ in terms of N_e but share similar developmental processes. We found a strong effect of eusociality on codon usage bias, strongly suggesting that N_e, not variable selective pressure, explains the among-taxa variation we detect. Subramanian (2008) noted that under the N_e hypothesis one would expect a step-like relationship between codon usage intensity and N_e, since the theory predicts no biased usage for every N_e well below $1/4\,s$ and almost perfect codon usage (i.e., ~100% of preferred codons) for every N_e well above $1/4\,s$, s being the selection coefficient in favor of preferred codons. This rationale, however, implicitly assumes that a constant selection coefficient applies to every synonymous mutation, which is unlikely to be true. If one rather assumes a distribution of s across synonymous mutations then a gradual effect of N_e on the intensity of codon usage bias is expected, consistent with our and Subramanian's (2008) results.

In contrast, no relationship was detected between the intensity of gBGC and N_e in our analysis. Significant gBGC was detected in the presumably small-N_e *Lepus granatensis* (hare) and *Abatus cordatus* (brooding sea urchin), for instance, whereas the detected effect was weaker in the presumably large-N_e *M. galloprovincialis* (mussel) and *Ciona intestinalis* (tunicate) despite large numbers of SNPs available in the latter two species (supplementary table S1, Supplementary Material online). A similar pattern was recently reported in plants, based on a data set of eleven species (Clément et al. 2017). This result is somewhat surprising in that, just like selection, gBGC should only be effective if of magnitude well above that of drift. The intensity of the signal for gBGC is expected to be determined by the product of four parameters, namely N_e, the effective population size, r, the per base recombination rate, l, the length of gene conversion tracts, and b_0, the repair bias in favor of GC. The rlb_0 product is often denoted as b (Glémin et al. 2015). Our results rule out the hypothesis that b is constant—or a N_e effect should be detected. r, l and/or b_0 must therefore vary substantially across species, and/or be inversely related to N_e.

The average per base recombination rate is known to vary among metazoans, from ~0.1 cM/Mb to >15 cM/Mb (Wilfert et al. 2007). There is also evidence that the length of gene conversion tracts (l) varies across species. For instance, in mammals, l is of the order of 400 bp for crossovers and 50 bp for noncrossover recombination events (Cole et al. 2014), whereas l is ~2,000 bp for both type of events in budding yeast (Mancera et al. 2008). In drosophila, noncrossover gene conversion tracts are on an average 440 bp long (Miller et al. 2016), that is, ~8 times longer than in mammals. More data are crucially needed to characterize more thoroughly the variation in l among animals. Repair bias b_0, finally, was estimated to be 0.014 in yeast (Mancera et al. 2008), 0.12 in *Daphnia* (Keith et al. 2016), 0.18 in flycatcher (Smeds et al. 2016), and up to 0.36 in humans (Halldorsson et al. 2016). This so far limited sample suggests that b_0 varies substantially among species and could be inversely correlated with N_e, perhaps explaining the absence of a N_e effect on gBGC intensity in our analysis.

We can think of two possible reasons why b_0 would scale inversely with N_e. First, gBGC is generally speaking a deleterious process in that it promotes G and C alleles irrespective of their effect on fitness (Galtier et al. 2009; Glémin 2010; Necşulea et al. 2011; Lachance and Tishkoff 2014). It might be that the molecular machinery involved in recombination is more efficiently selected to minimize b in large N_e species. A formal model would be required to validate this verbal hypothesis, though. Secondly, Lesecque et al. (2013) demonstrated that in yeast, when several SNPs are part of the same conversion tract, these are most often converted in the same direction—same donor and same recipient chromosomes—the direction only being influenced by SNPs located at the extremities of tracts. This implies a mechanical decay of the average GC bias as the number of SNPs per tract increases, since AT versus GC SNPs located in the middle of a conversion tract are converted in either direction with probability 0.5. This mechanism, if effective in animals too, might contribute to explaining the lack of a N_e effect on gBGC intensity, SNP density being positively correlated with N_e. Of note, the evolution of genomic GC-content has been

associated with traits related to N_e in mammals and birds (Romiguier et al. 2010; Weber et al. 2014). This might be explained by b being fairly homogeneous within groups, but much more variable across distantly related taxa, so that B would only respond to N_e at a relatively small time scale. It might also be the case that the relationship between GC-content dynamics and life-history traits in mammals and birds is not (entirely) mediated by N_e—but rather by, for example, the mutation rate in a nonequilibrium situation (Romiguier et al. 2010; Bolívar et al. 2016).

Conclusions

Translational selection is a significant determinant of codon usage patterns in large-N_e species of animals, but is weak or absent in small-N_e ones, such as large vertebrates and social insects. In contrast, gBGC is widespread across animals and of strength independent of N_e. gBGC is therefore a major confounder that must be seriously taken into account in any analysis of codon usage bias. This study uncovered two unexpected results that remain to be elucidated, that is, a general preference for C- and T-ending codons over G- and A-ending ones, respectively, and an inverse relationship between the recombination-associated GC repair bias and N_e.

Materials and Methods

Species Sampling

We used recently published Illumina transcriptome data from population samples of non model animals (Romiguier, Gayral, et al. 2014; Rousselle et al. 2016; Ballenghien et al. 2017; Romiguier et al. 2017), which covered 32 distinct families of Metazoa. In each family, we selected the species with the largest number of individuals, provided this number was five or more. Mosquito *Culex pipiens* (Culicidae) and trumpet worm *Pectinaria koreni* (Pectinariidae) were excluded because transcriptome assembly in these species yielded a small number of very short contigs (Romiguier, Gayral, et al. 2014). Harvester ant *Messor barbarus* (Formicidae) was excluded because of its peculiar mating system, which dramatically departs the Hardy–Weinberg assumption (Romiguier et al. 2017). Its sister species *Messor capitatus* was rather included, despite a lower number of sampled individuals. The final data set included 30 species, of which seven vertebrates, six insects, five molluscs, three crustaceans, three echinoderms, two tunicates, one annelid, one nematode, one nemertian, and one cnidarian (supplementary table S1, Supplementary Material online). Five to eleven individuals per species were analyzed. For each of these focal species, one outgroup from the same family was selected. The tissues from which RNA has been extracted, which differ across species, are provided in supplementary table S1, Supplementary Material online.

Romiguier, Gayral, et al. (2014) reported significant correlations between life history traits, such as species longevity, fecundity, and propagule size, and population genomic variables theoretically related to N_e, such as the synonymous diversity, π_S, and the ratio of nonsynonymous over synonymous heterozygosity, π_N/π_S. In this study, we used longevity, propagule size, and π_N/π_S as markers of the long-term N_e of the analyzed species. π_N/π_S is expected to be negatively correlated with N_e due to the decreased efficiency of purifying selection against slightly deleterious nonsynonymous alleles in small populations (Lanfear et al. 2014).

Transcriptome Assembly and Annotation

Transcriptome assembly, open reading frame (ORF) prediction, orthology prediction, and alignment between focal and outgroup coding sequences were achieved using the Abyss v1.3.4, Cap3 v10/15/07, Trinity_ORF, BLAST, and MACSe v1.02 programs, as previously described (Gayral et al. 2013; Romiguier, Gayral, et al. 2014, http://http://kimura.univ-montp2.fr/PopPhyl). We only retained contigs containing a predicted coding sequences (CDS) longer than 200 bp. The median number of contigs per species was 3,480 (supplementary table S1, Supplementary Material online). Contig expression level was measured as the per base pair read depth, averaged across individuals. For each contig of each species, we calculated GC-content at first and second codon positions (GC12), third codon positions (GC3) and UTR (GC_UTR), and the frequency of the 61 sense codons.

SNP and Genotype Calling

Genotypes and single nucleotide polymorphisms (SNPs) were called using the reads2snp v1.0 program, which was designed for genotyping based on RNAseq data (Tsagkogeorga et al. 2012; Gayral et al. 2013; Ballenghien et al. 2017). This method models read counts at each position as a multinomial distribution determined by allele frequencies, genotype frequencies, sequencing error rate, and cross-contamination rate. Allele frequencies are estimated a priori from read counts across all individuals. Genotype frequencies are assumed to follow the Hardy–Weinberg prior. The method first estimates the error rate by maximum likelihood, then the posterior distribution of genotypes in the empirical Bayesian framework (Tsagkogeorga et al. 2012). Contamination rate (Flickinger et al. 2015; Ballenghien et al. 2017) was here set to 0.2. This parameter likely captures a combination of effects leading to overdispersion of read counts and spurious calls of heterozygote genotypes (Ballenghien et al. 2017). A filter for false SNPs due to hidden paralogy was applied posterior to genotyping.

Site Frequency Spectrum Analysis

Synonymous SNPs were oriented assuming that the state observed at the orthologous position in the outgroup is ancestral. Unfolded site frequency spectra (SFS), that is, the observed distribution of derived allele frequency across SNPs, were built separately for AT→GC, GC→AT, and A↔T or G↔C synonymous mutations. The three SFS's were analyzed in the maximum likelihood framework using model M1 in Glémin et al. (2015), which accounts for the effect of mutation bias, gBGC, and drift and assumes constant gBGC among sites. Two versions of the model were considered, accounting or not for SNP orientation error. We estimated $B = 4\,N_e\,b$, the scaled gBGC coefficient, under both models and used as our point estimate a weighted average between estimates from the two models, weights being derived from Akaike's

Information Criterion as suggested by Posada and Buckley (2004). Confidence intervals around the estimated *B* were obtained by bootstrapping SNPs (1,000 replicates). Estimated *B* was said to be significantly different from zero when zero was outside the bootstrap 95% confidence interval.

Effect of Recombination Rate

We approximated the population-scaled recombination rate of each locus via a calculation based on the four-gamete rule (Hudson and Kaplan 1985). For every pair of SNPs in a locus, haplotypes were identified from individuals homozygous at both SNPs, and from individuals heterozygous at one SNP and homozygous at the other SNP. In these two situations, linkage relationships between alleles can be determined with certainty even from unphased data. Individuals carrying a heterozygous genotype at both SNPs were disregarded here. When the four possible haplotypes were found to be segregating in the sample, a recombination event was inferred. The total number of recombination events per contig was recorded by summing across pairs of SNPs, taking care of only counting once events supported by non independent pairs of SNPs. We defined R_{fgr} (for "four-gamete rule") as the ratio of total number of inferred recombination events by contig length. This was calculated, in each species, for each contig carrying at least one pair of SNPs, excluding singletons, such that four haplotypes or more could be inferred. Contigs departing these conditions were not considered eligible for recombination analysis, and missing data was recorded. Species in which <500 eligible contigs were available were not considered. We performed simulations to assess whether R_{fgr} could be used as a proxy for *Rho*. These simulations showed that R_{fgr} is affected by SNP density. However, for a given level of polymorphism and for the range of *rho* observed in animals, R_{fgr} strongly covaries with *Rho* (see supplementary fig. S5, Supplementary Material online). Thus, given the relatively limited variation in SNP density withing genomes, R_{fgr} appears to be a good indicator for intragenomic variation in recombination rate.

Insect Transcriptome Data

We downloaded Illumina RNAseq reads from 20 eusocial and 30 solitary species of insects from the NCBI SRA database. Specifically, we selected five species of ant, five eusocial bees, five eusocial wasps, five termites, five solitary hymenopterans, five solitary cockroaches (Blattodea, same order as termites), and 20 species from other orders of insects (supplementary table S2, Supplementary Material online). Species sampled in the context of the 1KITE project (http://www.1kite.org/) were favored whenever possible. Transcriptome assemblies were downloaded from NCBI Sequence Set in 43 species. In the remaining seven species, transcriptomes were assembled as above. In each species, reads were mapped to predicted cDNA. Depth of coverage and codon usage were computed for each contig of each species. The species list and accession numbers are provided in supplementary table S2, Supplementary Material online.

Correlation Analyses

Pearson's correlation coefficient and associated *P* values were calculated in R. Phylogenetic control analyses were performed using the independent contrasts method as implemented in the CAPER package.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Thomas Bataillon for helpful suggestions regarding statistics and the Montpellier Bioinformatics and Biodiversity platform for computational resources. This work was supported by European Research Council grant 232971, Swiss National Foundation grant CRSII3_160723, and Agence Nationale de la Recherche grant DaSiRe ANR-15-CE12-0010.

References

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*. 12(2):R18.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136(3):927–935.

Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG, Przeworski M. 2017. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. *Elife* 6:1–58.

Ballenghien M, Faivre N, Galtier N. 2017. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biol*. 15(1):25.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res*. 40(10):e72.

Bierne N, Eyre-Walker A. 2006. Variation in synonymous codon use and DNA polymorphism within the Drosophila genome. *J Evol Biol*. 19(1):1–11.

Bolívar P, Mugal CF, Nater A, Ellegren H. 2016. Recombination rate variation modulates gene sequence evolution mainly via GC-Biased Gene Conversion, not Hill-Robertson interference, in an avian system. *Mol Biol Evol*. 33(1):216–227.

Carlini DB, Chen Y, Stephan W. 2001. 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* 159(2):623–633.

Carlini DB, Stephan W. 2003. *In vivo* introduction of unpreferred synonymous codons into the Drosophila *Adh* gene results in reduced levels of ADH protein. *Genetics* 163(1):239–243.

Choudhari S, Grigoriev A. 2017. Phylogenetic heatmaps highlight composition biases in sequenced reads. *Microorganisms* 5(1):4.

Clément Y, Sarah G, Holtz Y, Homa F, Pointet S, Contreras S, et al. 2017. Evolutionary forces affecting synonymous variations in plant genomes. *PLoS Genet*. 13(5):e1006799.

Cole F, Baudat F, Grey C, Keeney S, de Massy B, Jasin M. 2014. Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. *Nat Genet*. 46(10):1072–1080.

Corcoran P, Gossmann TI, Barton HJ, Great Tit HapMap Consortium, Slate J, Zeng K. 2017. Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. *Genome Biol Evol*. 9:2987–3007.

Doherty A, McInerney JO. 2013. Translational selection frequently overcomes genetic drift in shaping synonymous codon usage in vertebrates. *Mol Biol Evol*. 30(10):2263–2267.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*. 36(16):e105.
Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet*. 16(7):287–289.
Duret L. 2002. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev*. 12(6):640–649.
Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*. 10:285–311.
Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci U S A*. 96(8):4482–4487.
Figuet E, Ballenghien M, Romiguier J, Galtier N. 2014. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biol Evol*. 7:240–250.
Flickinger M, Jun G, Abecasis GR, Boehnke M, Kang HM. 2015. Correcting for sample contamination in genotype calling of DNA sequence data. *Am J Hum Genet*. 97(2):284–290.
Galtier N, Bazin E, Bierne N. 2006. GC-biased segregation of noncoding polymorphisms in Drosophila. *Genetics* 172(1):221–228.
Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet*. 25(1):1–5.
Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N, et al. 2013. Reference-free population genomics from Next-Generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genet*. 9(4):e1003457.
Glémin S. 2010. Surprising fitness consequences of GC-biased gene conversion: i. Mutation load and inbreeding depression. *Genetics* 185(3):939–959.
Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. *Genome Res*. 25(8):1215–1228.
Glémin S, Clément Y, David J, Ressayre A. 2014. GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends Genet*. 30(7):263–270.
Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res*. 10(22):7055–7074.
Haddrill PR, Charlesworth B. 2008. Non-neutral processes drive the nucleotide composition of non-coding sequences in Drosophila. *Biol Lett*. 4(4):438–441.
Halldorsson BV, Hardarson MT, Kehr B, Styrkarsdottir U, Gylfason A, Thorleifsson G, Zink F, Jonasdottir A, Jonasdottir A, Sulem P, et al. 2016. The rate of meiotic gene conversion varies by sex and age. *Nat Genet*. 48(11):1377–1384.
Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu Rev Genet*. 42:287–299.
Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111(1):147–164.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*. 2(1):13–34.
Jackson BC, Campos JL, Haddrill PR, Charlesworth B, Zeng K. 2017. Variation in the intensity of selection on codon bias over time causes contrasting patterns of base composition evolution in Drosophila. *Genome Biol Evol*. 9(1):102–123.
Keith N, Tucker AE, Jackson CE, Sung W, Lucas Lledó JI, Schrider DR, Schaack S, Dudycha JL, Ackerman M, Younge AJ, et al. 2016. High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. *Genome Res*. 26(1):60–69.
Kent CF, Minaei S, Harpur BA, Zayed A. 2012. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. *Proc Natl Acad Sci U S A*. 109(44):18012–18017.
Lachance J, Tishkoff SA. 2014. Biased Gene Conversion Skews Allele Frequencies in Human Populations, Increasing the Disease Burden of Recessive Alleles. *Am J Hum Genet*. 95(4):408–420.
Lanfear R, Kokko H, Eyre-Walker A. 2014. Population size and the rate of evolution. *Trends Ecol Evol*. 29(1):33–41.
Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet*. 9(5):e1003527.
Lesecque Y, Mouchiroud D, Duret L. 2013. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Mol Biol Evol*. 30(6):1409–1419.
Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C, Strauss C, Stone C, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol*. 2(2):237–240.
Lynch M, Gutenkunst R, Ackerman M, Spitze K, Ye Z, Maruki T, Jia Z. 2017. Population genomics of *Daphnia pulex*. *Genetics* 206(1):315–332.
Machado HE, Lawrie DS, Petrov DA. 2017. Strong selection at the level of codon usage bias: evidence against the Li-Bulmer model. *biorxiv* Available from: https://doi.org/10.1101/106476.
Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454(7203):479–485.
Miller DE, Smith CB, Kazemi NY, Cockrell AJ, Arvanitakas AV, Blumenstiel JP, Jaspersen SL, Hawley RS. 2016. Whole-genome analysis of individual meiotic events in *Drosophila melanogaster* reveals that noncrossover gene conversions are insensitive to interference and the centromere effect. *Genetics* 203(1):159–171.
Mugal CF, Weber CC, Ellegren H. 2015. GC-biased gene conversion links the recombination landscape and demography to genomic base composition. *Bioessays* 37(12):1317–1326.
Nagylaki T. 1983. Evolution of a finite population under gene conversion. *Proc Natl Acad Sci U S A*. 80(20):6278–6281.
Necşulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L. 2011. Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum Mutat*. 32(2):198–206.
Ohta T, Gillespie JH. 1996. Development of neutral and nearly neutral theories. *Theor Popul Biol*. 49(2):128–142.
Percudani R. 2001. Restricted wobble rules for eukaryotic genomes. *Trends Genet*. 17(3):133–135.
Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GA. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol*. 4(7):675–682.
Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol*. 53(5):793–808.
Pouyet F, Mouchiroud D, Duret L, Sémon M. 2017. Recombination, meiotic expression and human codon usage. *Elife* 6:e27344.
Rao Y, Wu G, Wang Z, Chai X, Nie Q, Zhang X. 2011. Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. *DNA Res*. 18(6):499–512.
Ratnakumar A, Mousset S, Glemin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc Lond B Biol Sci*. 365(1552):2571–2580.
Robinson MC, Stone EA, Singh ND. 2014. Population genomic analysis reveals no evidence for GC-biased gene conversion in *Drosophila melanogaster*. *Mol Biol Evol*. 31(2):425–433.
Rocha EP. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res*. 14(11):2279–2286.
Romiguier J, Fournier A, Yek SH, Keller L. 2017. Convergent evolution of social hybridogenesis in Messor harvester ants. *Mol Ecol*. 26(4):1108–1117.
Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* 515(7526):261–263.
Romiguier J, Lourenço JM, Gayral P, Faivre N, Weinert LA, Ravel S, et al. 2014. Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. *J Evol Biol*. 27:593–603.

Romiguier J, Ranwez V, Douzery EJ, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res.* 20(8):1001–1009.

Rousselle M, Faivre N, Ballenghien M, Galtier N, Nabholz B. 2016. Hemizygosity enhances purifying selection: lack of fast-Z evolution in two satyrine butterflies. *Genome Biol Evol.* 8(10):3108–3119.

Rudolph KLM, Schmitt BM, Villar D, White RJ, Marioni JC, Kutter C, Odom DT, Galtier N. 2016. Codon-Driven Translational efficiency is stable across diverse mammalian cell states. *PLoS Genet.* 12(5):e1006024.

Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet.* 12(10):683–691.

Semon M, Lobry JR, Duret L. 2006. No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Mol Biol Evol.* 23(3):523–529.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* 33(4):1141–1153.

Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol.* 24(1–2):28–38.

Shields DC, Sharp PM, Higgins DG, Wright F. 1988. "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol.* 5(6):704–716.

Smeds L, Mugal CF, Qvarnström A, Ellegren H, Henderson I. 2016. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS Genet.* 12(5):e1006044.

Subramanian S. 2008. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. *Genetics* 178(4):2429–2432.

Tsagkogeorga G, Cahais V, Galtier N. 2012. The population genomics of a fast evolver: high levels of diversity, functional constraint and molecular adaptation in the tunicate *Ciona intestinalis*. *Genome Biol Evol.* 4(8):852–859.

Vieira-Silva S, Rocha EP. 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* 6(1):e1000808.

Wallberg A, Glémin S, Webster MT, Barton NH. 2015. Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLoS Genet.* 11(4):e1005189.

Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol.* 15(12):549.

Wilfert L, Gadau J, Schmid-Hempel P. 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity* 98(4):189–197.

Williams AL, Genovese G, Dyer T, Altemose N, Truax K, Jun G. 2015. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *Elife* 4:e04637.

Wu X, Hurst LD. 2015. Why selection may be stronger when populations are small: intron size and density predict within and between species usage of exonic splice associated cis-motifs. *Mol Biol Evol.* 32(7):1847–1861.

Zeng K. 2010. A simple multiallele model and its application to identifying preferred-unpreferred codons using polymorphism data. *Mol Biol Evol.* 27(6):1327–1337.

Zhou T, Weems M, Wilke CO. 2009. Translational optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol.* 26(7):1572–1580.