GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes

Paulina Bolívar, Laurent Guéguen, Laurent Duret, Hans Ellegren, Carina Mugal

Bolívar et al. Genome Biology (2019) 20:5
https://doi.org/10.1186/s13059-018-1613-z

Genome Biology

RESEARCH Open Access

GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes

Paulina Bolívar[1], Laurent Guéguen[2], Laurent Duret[2], Hans Ellegren[1] and Carina F. Mugal[1*]

Abstract

Background: The nearly neutral theory of molecular evolution predicts that the efficacy of natural selection increases with the effective population size. This prediction has been verified by independent observations in diverse taxa, which show that life-history traits are strongly correlated with measures of the efficacy of selection, such as the d_N/d_S ratio. Surprisingly, avian taxa are an exception to this theory because correlations between life-history traits and d_N/d_S are apparently absent. Here we explore the role of GC-biased gene conversion on estimates of substitution rates as a potential driver of these unexpected observations.

Results: We analyze the relationship between d_N/d_S estimated from alignments of 47 avian genomes and several proxies for effective population size. To distinguish the impact of GC-biased gene conversion from selection, we use an approach that accounts for non-stationary base composition and estimate d_N/d_S separately for changes affected or unaffected by GC-biased gene conversion. This analysis shows that the impact of GC-biased gene conversion on substitution rates can explain the lack of correlations between life-history traits and d_N/d_S. Strong correlations between life-history traits and d_N/d_S are recovered after accounting for GC-biased gene conversion. The correlations are robust to variation in base composition and genomic location.

Conclusions: Our study shows that gene sequence evolution across a wide range of avian lineages meets the prediction of the nearly neutral theory, the efficacy of selection increases with effective population size. Moreover, our study illustrates that accounting for GC-biased gene conversion is important to correctly estimate the strength of selection.

Keywords: Nearly neutral theory, Life-history traits, d_N/d_S, GC-biased gene conversion, Base composition, Avian genomes

Background

With his proposal of the neutral theory of molecular evolution in 1968, Kimura introduced a revolutionizing concept to evolutionary biology, which at that time was strongly influenced by the view that evolution is driven by positive Darwinian selection [1]. Instead, Kimura proposed that at the molecular level deleterious mutations are common, while advantageous mutations are rare, and that in a finite population most evolutionary changes are a consequence of the fixation of neutral mutations due to genetic drift. Kimura thus put a new emphasis on the stochasticity of population genetics, and further established the relationship between sequence conservation and functional importance, which is key to many bioinformatics software for the identification of conserved coding as well as non-coding elements [2].

After realizing the importance of nearly neutral (mainly slightly deleterious) mutations in molecular evolution, Tomoko Ohta extended Kimura's theory to the nearly neutral theory of molecular evolution [3, 4]. The nearly neutral theory of molecular evolution states that the effectiveness of selection depends on a balance between the strength of random genetic drift and the selection coefficient (s) of new mutations [5]. A measure of the strength of genetic drift is the effective population size (N_e) [6]. At the population level, aside from the mutation

* Correspondence: carina.mugal@ebc.uu.se
[1]Department of Ecology and Genetics, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden
Full list of author information is available at the end of the article

rate, N_e and s together determine the level of genetic diversity [7, 8]. Over larger evolutionary time scales, N_e and s determine the probability of fixation of new mutations, which ultimately affects the evolutionary rate of change [5, 8]. Therefore, the nearly neutral theory not only forms the basis for a solid null-hypothesis to test for evidence of positive Darwinian selection, but also allows clear predictions to be formulated, which can be tested against data [9]. Specifically, the nearly neutral theory predicts that if N_e is small, the effect of genetic drift is strong, and slightly deleterious mutations are more likely to reach fixation than if N_e is large [5]. Consequently, species with small N_e will accumulate more slightly deleterious substitutions over time than species with large N_e. Here, comparison of the evolutionary rate at non-synonymous and synonymous sites allows assessment of the strength of natural selection acting on protein-coding genes, where low effectiveness of selection results in an elevated (but generally < 1), non-synonymous to synonymous substitution rate ratio (d_N/d_S). Therefore, to test the validity of the nearly neutral theory, the relationship between N_e and d_N/d_S can be explored.

Since measuring N_e in natural populations is a challenging task [10–12], a common alternative is to use species characteristics that are associated with N_e as proxies [13–15]. For example, this includes geographic range, where widely distributed species usually have larger N_e than species with a limited geographic range, such as island species [16]; social organization, such as eusociality versus solitarity, where eusocial species usually have lower N_e than solitary species [17]; and mating system, where selfing species usually have lower N_e than outcrossers [18]. It also includes life-history traits that are associated with growth, survival, and reproductive success and strategy [19]. Common examples of life-history traits include, among others, adult body mass or size, maximum longevity, generation time, age of sexual maturity or age at first reproduction, and fecundity. For instance, species with a small body size tend to have a short generation time, large reproductive output, and usually a large N_e [15, 20].

By empirically testing for a relationship between social organization, mating system or life-history traits and d_N/d_S, several studies have confirmed the prediction of the nearly neutral theory. For example, it has been observed that eusocial insects show higher d_N/d_S ratios than solitary species [21]. Another study reported a general trend for higher d_N/d_S ratios in the chloroplasts of selfers than of outcrossing species of plants [22]. Moreover, several studies in mammalian taxa have reported strong positive correlations between generation time, body mass, or longevity and d_N/d_S in mitochondrial and nuclear genes [23–29]. Similar correlations were also found for non-avian reptiles [30].

However, not all observations are in accordance with the prediction of the nearly neutral theory. An intriguing example is the repeated failure to find a positive correlation between life-history traits and d_N/d_S in birds [25, 30, 31]. Since the scenario that the nearly neutral theory does not hold in birds is rather unlikely, there must be a methodological reason or a biological phenomenon that conceals this relationship. To explore the lack of a positive correlation in more detail, the ratio of radical to conservative amino acid substitutions (K_r/K_c) has been used as an alternative proxy for the efficacy of selection in avian lineages [25, 31]. This revealed a positive correlation between body mass and K_r/K_c in birds. This observation questions whether d_N/d_S is an appropriate measure for the efficacy of selection in birds. However, the use of the K_r/K_c ratio as a measure of the efficacy of selection is not straightforward either given that radical and conservative mutations can be both affected by effectively neutral mutations and hence vary with N_e [30]. Later, Figuet et al. [30] explored the lack of correlations between d_N/d_S and life-history traits in birds from another angle. Specifically, they first investigated the hypotheses that life-history traits are poor proxies for N_e and that variation in life-history traits is narrow in birds. They further explored the possibility that a peculiar distribution of fitness effects of new mutations in birds could be responsible for low power to detect a correlation. However, they found evidence to reject all these hypotheses and could thus not explain the lack of a positive correlation. Botero-Castro et al. [32] recently found evidence that sequence alignment quality affects the correlation between life-history traits and d_N/d_S; after removing putative alignment errors, a positive correlation between d_N/d_S and longevity, but not between d_N/d_S and body mass, could be recovered. Moreover, they found stronger correlations when previously undetected GC-rich orthologs were included in the analysis. Taken together, observations from previous studies suggest that d_N/d_S might not be an appropriate measure for the efficacy of selection in birds, potentially because there is some force that biases estimates of d_N/d_S and thereby obscures the correlation between d_N/d_S and life-history traits in birds.

It is well documented that avian genomes are strongly impacted by GC-biased gene conversion (gBGC) [33–36]. gBGC is a process associated with meiotic recombination that leads to the preferential fixation of GC (strong: S) over AT (weak: W) alleles in GC/AT heterozygous sites close to recombination-initiating double-strand breaks [37, 38]. It acts in a manner similar to directional selection, increasing the probability of fixation of S over W alleles [39]. Thus, like the strength of natural selection, also the strength of gBGC increases with N_e and recombination rate, which both vary along the

genome and among species. gBGC can therefore interfere with natural selection and bias inferences of the strength and efficacy of selection based on divergence and/or diversity data [40–47]. The net impact of gBGC on substitution rates also depends on the relative contribution of substitutions from AT to GC (W-to-S) and from GC to AT (S-to-W), which is reflected in ΔGC, the difference between the equilibrium GC content (GC*) and the ancestral GC content [41, 48]. If ΔGC differs between synonymous and non-synonymous substitutions, gBGC may increase or decrease estimates of d_N/d_S depending on the relative impact of gBGC on synonymous and non-synonymous substitution rates.

Here, we revisit the relationship between d_N/d_S and life-history traits in birds, and investigate if the impact of gBGC on d_N/d_S conceals the correlation between life-history traits and d_N/d_S in the avian clade. In order to distinguish the impact of gBGC on lineage-specific substitution rates from the impact of selection, we estimate d_N/d_S separately for different substitution categories, namely W-to-S and S-to-W changes, which are both affected by gBGC, and GC-conservative changes, i.e., S-to-S and W-to-W, which are not affected by gBGC. To do so, we adapt a recently implemented approach that explicitly allows for non-stationary base composition [49]. This analysis provides clear evidence that gBGC conceals the correlation between life-history traits and d_N/d_S in birds. Moreover, it stresses the importance of accounting for gBGC to correctly estimate the strength of selection in comparative genomics studies, which seems particularly important when the impact of gBGC varies among lineages. Finally, we propose a new statistic, d_N/d_S based on GC-conservative changes, together with a program to perform estimates of the strength of selection after accounting for the impact of gBGC.

Results

gBGC conceals the correlation between dN/dS and life-history traits in birds

We explored the relationship between estimates of d_N/d_S based on publicly available coding sequence alignments of 7986 genes in 47 avian species and three life-history traits (body mass, longevity, and age of sexual maturity). To estimate non-synonymous and synonymous substitution rates, we applied methods implemented in the bio++ libraries [50, 51]. First, we fitted a non-stationary homogeneous YN98 substitution model by maximum likelihood, to retrieve the most likely branch lengths, codon frequencies at the root, and model parameters. Second, we applied a recently developed approach based on substitution mapping to estimate d_N/d_S [49]. We further modified this approach to split non-synonymous (and synonymous) counts in three categories (W-to-S, S-to-W, and GC-conservative changes, i.e., S-to-S plus W-to-W), such that d_N/d_S was estimated separately for each of those categories. Briefly, for each of these categories, we used substitution mapping [52] to estimate the expected number of substitutions on each branch given the distribution of the scenarios provided by the previously optimized model and tree. We normalized these counts by the expected numbers of such substitutions for the same scenarios under a neutral model (i.e., $\omega = 1$) [49] (see Methods section).

Estimates of d_N/d_S based on all substitutions together were on average smaller than estimates based on only GC-conservative changes, but differences between d_N/d_S based on all substitutions and d_N/d_S based on GC-conservative changes ranged from negative to positive values (Additional file 1: Table S1). Data for GC-conservative changes showed very strong positive correlations between d_N/d_S and all life-history traits, most prominently for body mass followed by the age of sexual maturity and longevity (Table 1). In contrast, none of the relationships between d_N/d_S based on all substitutions together and life-history traits were statistically significant (Fig. 1 and Table 1). Data for the two substitution categories influenced by gBGC showed weak correlations with opposite directions to each other. Since correlation analysis can be affected by phylogenetic relationships of species and their life-history traits, we further performed correlation analysis after accounting for the phylogenetic relationships of species. The conclusion remained the same after accounting for the phylogenetic relationships of species when estimating the correlation between d_N/d_S and body mass (Additional file 2: Table S2). For the other two life-history traits, longevity and age of sexual maturity, correlations for d_N/d_S based on GC-conservative changes lost their statistical significance, but were still larger than correlations for d_N/d_S based on all substitution categories. Note that correction for phylogenetic relationships is the correct practice to perform correlation analysis among related species. Here, uncorrected correlations are reported in the main text and corrected ones in the Supplement in order to allow for a more direct comparison between this and previous studies.

To test the robustness of the results obtained from substitution mapping, we repeated the analyses using two other substitution models implemented in the bio++ libraries, model T92X3 and model L95X3 (for details see Methods). Estimates of d_N/d_S and consequently the strength of the correlations with life-history traits were highly similar to that obtained using the YN98 model (Additional file 2: Figures S1 and S2). This suggests that our inferences are robust to the underlying codon substitution model.

Table 1 Pearson correlation coefficients (R) and their statistical significance between d_N/d_S for different substitution categories and different proxies for N_e. Significant correlations are highlighted in italics

Life history trait	Total		GC-conservative		S-to-W		W-to-S	
	R	p value	R	p value	R	p value	R	p value
Body mass	0.08	6.09×10^{-1}	*0.57*	*3.15×10^{-5}*	*−0.33*	*2.25×10^{-2}*	*0.33*	*2.28×10^{-2}*
Longevity	0.08	6.25×10^{-1}	*0.32*	*4.63×10^{-2}*	−0.22	1.64×10^{-1}	0.27	9.51×10^{-2}
Age of sexual maturity	0.08	5.85×10^{-1}	*0.45*	*1.97×10^{-3}*	−0.19	1.99×10^{-1}	0.28	6.41×10^{-2}

S-to-W: strong to weak, W-to-S: weak to strong

Estimates of substitution rates have been shown to be biased for genes that have gene trees discordant to the species tree if substitution rates are estimated based on the species tree [53]. To examine the possible influence of gene tree discordance on our observations, we repeated the analysis using a gene-by-gene approach, where the number of non-synonymous and synonymous substitutions and sites were estimated for each gene separately with their respective gene tree. In order to calculate average species-specific substitution rates for the different substitution categories, we divided the sum of the substitution counts over all genes of the respective category by the sum of the expected number of substitutions over all genes of the same category. Additional file 2: Tables S3 and S4 show that the lack of a positive correlation between life-history traits and d_N/d_S based on all substitutions as well as the presence of a significant positive correlation between life-history traits and d_N/d_S based on GC-conservative changes are robust to gene tree heterogeneity. However, compared

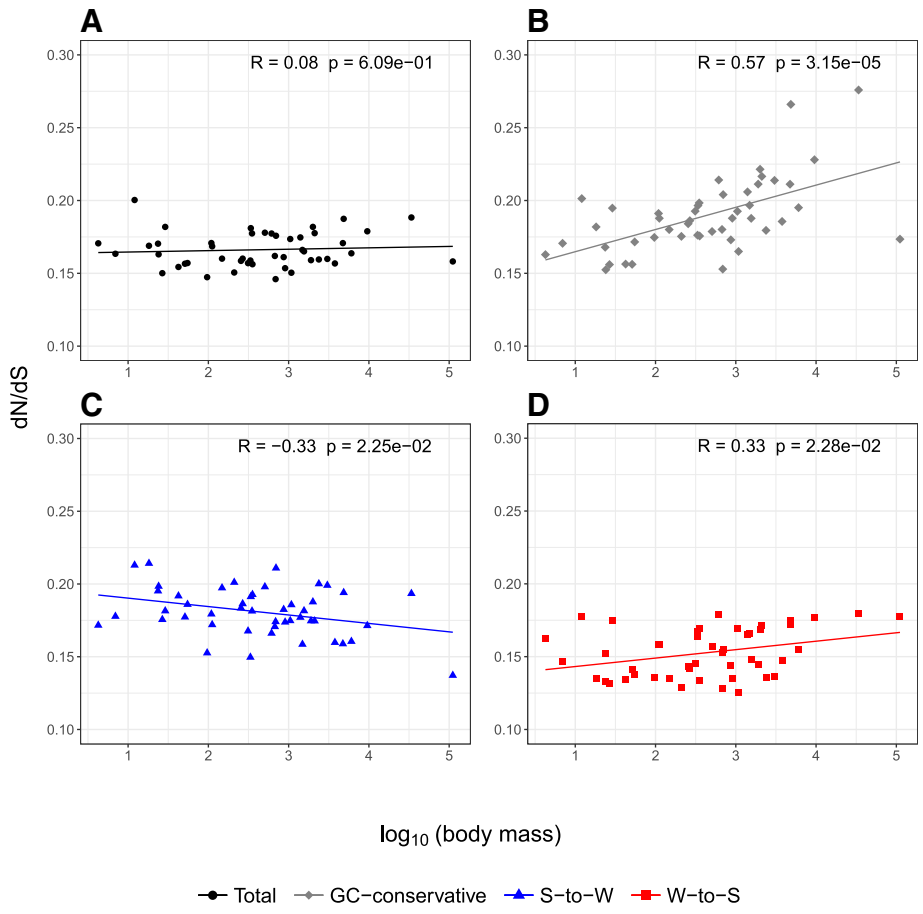

Fig. 1 Relationship between d_N/d_S and log-transformed body mass for different substitution categories. **a** All substitution categories together (total; black circles), **b** GC-conservative substitutions (gray diamonds), **c** S-to-W substitutions (blue triangles), and **d** W-to-S substitutions (red squares). Each dot corresponds to a terminal branch of the avian phylogeny. Regression lines for each substitution category are shown using the same color scheme. Pearson correlation coefficients and their statistical significance are provided in the right upper corner of each panel

to the analysis based on the species tree, correlations between life-history traits and d_N/d_S based on W-to-S substitutions lost significance for the gene-by-gene approach.

The impact of gBGC on d_N/d_S varies among lineages

The overall impact of gBGC on total substitution rates (and consequently on d_N/d_S based on all substitution categories) not only depends on the strength of gBGC but also on the relative contribution of S-to-W and W-to-S substitutions to the total number of substitutions. While the strength of gBGC is reflected in the equilibrium GC content (GC*) [36, 54], the relative contribution of S-to-W and W-to-S to the total number of substitutions in a certain lineage is governed by its ancestral GC content (i.e., the GC content at its most recent ancestral node). For each lineage, we computed the difference between the equilibrium GC content of the substitution process (GC*) and the ancestral GC content (denoted as ΔGC) (see Methods) to summarize the dynamics of base composition, separately for synonymous and non-synonymous substitutions. We observed that ΔGC differed between synonymous and non-synonymous substitution rates and varied among lineages. Interestingly, in the majority of the avian lineages analyzed, ΔGC synonymous was larger than ΔGC non-synonymous (mean ± standard deviation; ΔGC synonymous = 0.076 ± 0.071 and ΔGC non-synonymous = 0.026 ± 0.038), which indicates that in avian lineages neutrally evolving sites are on average further away from their equilibrium GC content than sites evolving under selection. This would suggest that on average the impact of gBGC is lower on d_N than on d_S, which will influence the net impact of gBGC on the d_N/d_S ratio [41, 48] and could explain why d_N/d_S based on all substitution categories together is on average decreased (compared to d_N/d_S based on GC-conservative changes). ΔGC synonymous and ΔGC non-synonymous showed a negative relationship with body mass (Fig. 2; ΔGC synonymous: $R = -0.33$, p value $= 2.36 \times 10^{-2}$; and ΔGC non-synonymous: $R = -0.34$, p value $= 1.81 \times 10^{-2}$), indicating that species with large N_e and thus higher strength of gBGC generally are further away from their equilibrium than species with small N_e.

Comparison with and re-analysis of previously investigated bird datasets

Several previous studies have addressed the relationship between d_N/d_S and life-history traits in birds [25, 30–32]. Table 2 provides an overview of the relationship between d_N/d_S and body mass among studies (including our study), which illustrates the consistent lack of a significant positive correlation if gBGC is not accounted for. Although all these studies, like us, used a substitution mapping approach to estimate branch-specific d_N/d_S, our study differs from previous work in three key aspects.

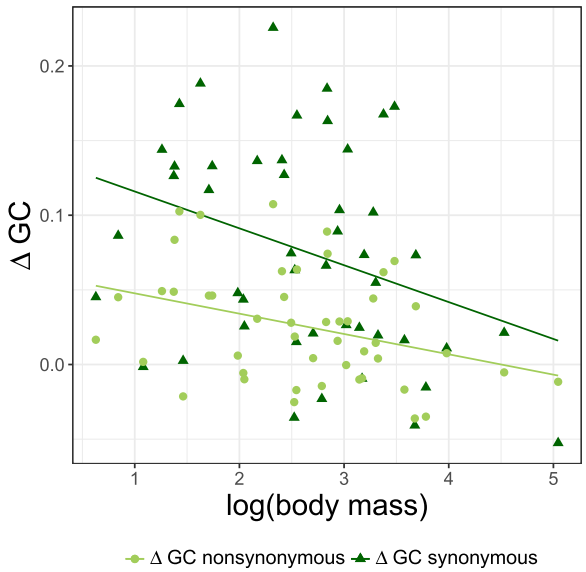

Fig. 2 Relationship between ΔGC and body mass. ΔGC non-synonymous (light green circles) and ΔGC synonymous (dark green triangles). Regression lines for each substitution category are shown using the same color scheme

First, the underlying substitution model that we used explicitly allows for non-stationary base composition, in contrast to the assumption of stationary base composition used in previous studies. Second, the subset of genes analyzed and the alignment procedures varied between studies (Table 2). Finally, only in our study, d_N/d_S was separately estimated for S-to-W, W-to-S, and GC-conservative substitution categories. Comparison of the relationships between d_N/d_S and life-history traits among studies is therefore limited to estimates of d_N/d_S based on all substitutions together (Table 2).

In order to improve comparability between results and to be able to distinguish between the effects of the underlying methodological differences (related to the stationarity assumption) and gene dataset, we re-analyzed two datasets with our revised approach, the cleaned alignments of Figuet et al. [30] (referred to as Figuet+HMMclean) and the alignments of Botero-Castro et al. [32]. Note that these two datasets are subsets of the avian dataset published by Jarvis et al. [57]. We addressed two different aspects when re-analyzing these two data sets. First, we applied our modified (non-stationary) model to analyze all substitution categories together. So, instead of a method that assumes stationarity in base composition as implemented in the original publications, we used a method that allows for non-stationarity. Second, we re-analyzed previous datasets also using the modified (non-stationary) model but this

Table 2 Description and Pearson correlation coefficients (*R*) and their statistical significance between d_N/d_S (based on all substitution categories together) and body mass reported in different studies. Significant correlations are highlighted in italics

Study (dataset)	Number of genes	Description	*Neornithes*		*Neoaves*	
			R	*p* value	*R*	*p* value
Weber et al. [31]	921	Alignments from Jarvis et al. [57] present in all 48 species	*− 0.43*	*2.70×10^{-3}*	NA	NA
Figuet et al. [30]	1077	Alignments from Jarvis et al. [57] for orthologs that are also present in mammals and non-avian sauropsids. Max missing data of 6 species	0.13	3.20×10^{-1}	NA	NA
Botero-Castro et al. [32] (Figuet+HMMclean)	1077	Dataset from Figuet et al. 2016 + alignment filtering based on HMMclean	0.17	2.80×10^{-1}	*0.42*	*4.00×10^{-3}*
Botero-Castro et al. [32] (Botero-Castro)	1077 + 1245	Figuet+HMMclean dataset + 1245 previously non-annotated GC-rich genes present in at least 10 species	0.23	1.30×10^{-1}	*0.59*	*7.34×10^{-5}*
This study	7986	Alignments from Jarvis et al. [57] Max missing data of 6 species	0.08	6.09×10^{-1}	0.19	2.19×10^{-1}

time limited to only GC-conservative changes in order to investigate the impact of gBGC.

Similar to the original publications, correlations between life-history traits and d_N/d_S based on all substitution categories together (which are affected by gBGC) were only marginally or non-significant (Table 3). This suggests that the relaxation of the stationarity assumption has only a minor impact on the results. On the other hand, correlations between life-history traits and d_N/d_S based on GC-conservative changes only (which are not affected by gBGC), were strong and significant for all datasets, which suggests that gBGC has a major impact on the results regardless of the data set analyzed. In addition, as observed in the original studies, Paleognathae (ratites and tinamous) and Galloanserae (gamefowl and waterfowl) appeared to be outliers in the Figuet+HMMclean and the Botero-Castro et al. datasets (Additional file 2: Figures S3-S5); estimates of d_N/d_S were lower than expected by their respective body mass. We therefore followed Botero-Castro et al. [32] and repeated the correlation analysis within the group of Neoaves only. Again in agreement with previous observations [32], correlations between d_N/d_S and life-history traits were stronger within Neoaves than within Neornithes (Table 3 and Additional file 2: Figures S3-S5,). However, after accounting for the phylogenetic relationships of species, the difference in the strength of correlations within Neoaves and within Neornithes reduced to only a minor effect (Additional file 2: Table S5). This suggests that correction for phylogenetic relationships of species is important and indicates that contrary to previous suggestions [32], life-history traits do not seem to lose power to properly describe variation in N_e over large evolutionary distances. Nevertheless, our results provide evidence that gBGC weakens the correlation between d_N/d_S and life-history traits in birds and that this observation is robust to the underlying gene dataset.

Local GC content and chromosome size affect estimates of d_N/d_S but not its correlations with life-history traits

Previous studies have observed differences in estimates of d_N/d_S between GC-rich and GC-poor genes [32]. In order to test if our results are robust to local variation in GC content, we split the dataset into two groups according to their GC content, i.e., high GC content and low GC content. We then estimated d_N/d_S separately for the two datasets and investigated the correlation between body mass and d_N/d_S for different substitution categories. The correlation between body mass and d_N/d_S based on GC-conservative changes was significantly positive both for GC-rich and GC-poor genes, but not significant when analyzing all substitutions together. Interestingly, correlations were similar between datasets, but estimates of d_N/d_S were consistently lower for GC-rich genes than for GC-poor genes, irrespective of the substitution category (Fig. 3, Table 4, and Additional file 2: Table S5, for correlations after correction for phylogenetic relationships of species). The difference in d_N/d_S between

Table 3 Pearson correlation coefficients (*R*) and their statistical significance between d_N/d_S and body mass for different datasets re-analyzed in the present study. Significant correlations are highlighted in italics

Dataset	*Neornithes*		*Neoaves*		*Neornithes* GC-conservative		*Neoaves* GC-conservative	
	R	*p* value	*R*	*p* value	*R*	*p* value	*R*	*p* value
Figuet+HMMclean	0.11	4.66×10^{-1}	0.3	6.00×10^{-1}	*0.51*	*3.53×10^{-4}*	*0.75*	*3.68×10^{-8}*
Botero-Castro	*0.33*	*2.68×10^{-2}*	*0.69*	*1.41×10^{-6}*	*0.55*	*1.23×10^{-4}*	*0.83*	*5.86×10^{-11}*
Original dataset of our study	0.08	6.09×10^{-1}	0.19	2.19×10^{-1}	*0.57*	*3.15×10^{-5}*	*0.72*	*7.34×10^{-8}*

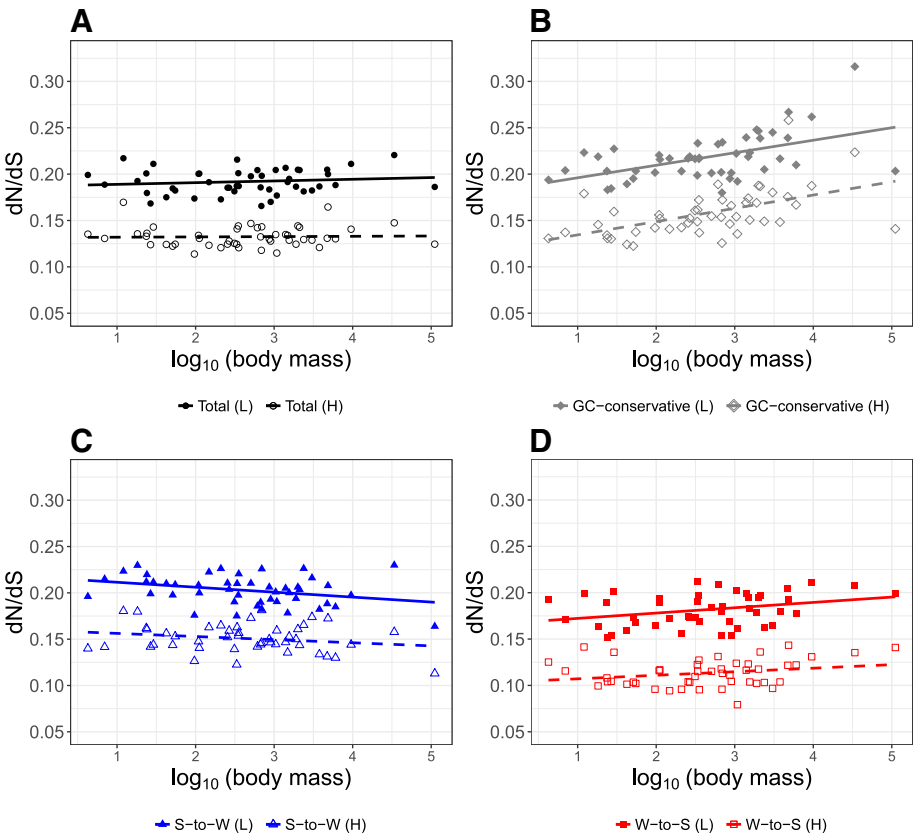

Fig. 3 Relationships between d_N/d_S and body mass separately for GC-poor genes (low GC content) and GC-rich genes (high GC content). Low GC content (L; filled shapes) and high GC content (H; hollow shapes). **a** All substitution categories together (total; black circles), **b** GC-conservative substitutions (gray diamonds), **c** S-to-W substitutions (blue triangles), and **d** W-to-S substitutions (red squares). Each dot corresponds to a terminal branch of the avian phylogeny. Regression lines for each substitution category are shown using the same color scheme. Solid lines correspond to GC-poor genes, dashed lines to GC-rich genes

GC-rich and GC-poor genes appears to be independent of the effect of gBGC on d_N/d_S, since all substitution categories were affected in the same way. Instead, the difference in d_N/d_S seems to be a consequence of both higher d_S and lower d_N in GC-rich genes as compared to GC-poor genes (relative difference in d_S and d_N for GC-conservative changes, respectively, significance based on a *t*-test; $\Delta d_S = 12.57$, p value $< 2.2 \times 10^{-16}$, $\Delta d_N = -6.51$, p value $= 4.89 \times 10^{-8}$; for all substitution categories together, $\Delta d_S = 31.02$, p value $< 2.22 \times 10^{-16}$, $\Delta d_N = -13.41$, p value $< 2.22 \times 10^{-16}$; Additional file 2: Figure S6).

To further elaborate on the relationship between local GC content and estimates of d_N/d_S, we explored the association between chromosome size and estimates of d_N/d_S. Avian genomes are characterized by a very large variation in chromosome size, with macrochromosomes (here defined as chromosomes larger than 100 Mb) having lower overall recombination rate and lower GC

Table 4 Pearson correlation coefficients and their statistical significance between d_N/d_S for different substitution categories and body mass for genes with low and high GC content, and for genes located in microchromosomes or macrochromosomes. Significant correlations are highlighted in italics

	Total		GC-conservative		S-to-W		W-to-S	
	R	*p* value	*R*	*p* value	*R*	*p* value	*R*	*p* value
High GC content	0.03	8.54×10^{-1}	*0.53*	*1.22×10^{-4}*	−0.21	1.53×10^{-1}	0.25	9.43×10^{-2}
Low GC content	0.13	3.83×10^{-1}	*0.52*	*1.93×10^{-4}*	*−0.32*	*2.73×10^{-2}*	*0.31*	*3.16×10^{-2}*
Microchromosomes	0.14	3.45×10^{-1}	*0.55*	*7.49×10^{-5}*	−0.26	7.79×10^{-2}	*0.40*	*5.42×10^{-3}*
Macrochromosomes	0.01	9.27×10^{-1}	*0.47*	*7.65×10^{-4}*	−0.27	6.18×10^{-2}	0.17	2.61×10^{-1}

S-to-W: strong to weak, W-to-S: weak to strong

content than microchromosomes (here smaller than 16 Mb). Hence, GC-rich genes are more often found on microchromosomes. Consistent with what has been described above, we observed lower d_N/d_S estimates for microchromosomes than for macrochromosomes (Table 4 and Additional file 2: Figure S7). Also similar to above, the difference in d_N/d_S seemed to be a consequence of both higher d_S and lower d_N in microchromosomes than in macrochromosomes (relative difference in d_S and d_N for GC-conservative changes, respectively, significance based on a *t*-test; $\Delta d_S = 6.83$, p value = 1.64×10^{-8}, $\Delta d_N = -8.91$, p value = 1.42×10^{-11}; for all substitution categories together, $\Delta d_S = 11.15$, p value = 1.14×10^{-14}, $\Delta d_N = -10.62$, p value = 5.75×10^{-14}; Additional file 2: Figure S8). Taken together, these results indicate that local GC content and chromosome size affect estimates of d_N/d_S. Since GC content and chromosome size are correlated, it is difficult to know to what extent the effect of chromosome size simply is an effect of GC content, or vice versa. The correlation between life-history traits and d_N/d_S based on GC-conservative changes is on the other hand robust to variation in local GC content and chromosome size.

Discussion

The lack of a positive relationship between d_N/d_S and life-history traits in avian taxa has been puzzling and has therefore been explored from various angles [25, 30–32]. Our results offer a solution to the puzzle by providing clear evidence that gBGC distorts d_N/d_S such that the ratio does not reflect the variation in the efficacy of selection among avian lineages. Strong positive correlations were recovered after accounting for gBGC. In contrast, and in agreement with earlier observations, life-history traits and d_N/d_S were not positively correlated when the estimation of d_N/d_S was based on all substitution categories together without accounting for gBGC. Thus, our study clearly illustrates that accounting for gBGC is crucial to correctly estimate the strength of selection in comparative genomic studies across taxa that are affected by gBGC.

gBGC has been observed in a wide range of taxa [38]. Both theoretical investigations and empirical observations suggest that gBGC shows a strong impact on d_N/d_S [41, 42, 45, 46, 48]. Interestingly, while gBGC was found to increase estimates of d_N/d_S in mammals and fishes [55, 56], previous work in birds (chicken and flycatchers) suggested that d_N/d_S was decreased by gBGC [41, 55]. However, our avian analyses indicate a more nuanced picture. While we observe that the impact of gBGC in the majority of avian lineages indeed results in an underestimation of d_N/d_S, in some it results in an overestimation (Fig. 1, Additional file 1: Table S1). The net effect of gBGC on d_N/d_S depends on N_e and recombination rate, but also builds upon the relative contribution of W-to-S and S-to-W substitutions to synonymous versus non-synonymous substitution rates [41]. All these parameters are reflected in the dynamics of base composition. Since we found that the dynamics of base composition differs between neutrally evolving and selected sites, and varies substantially among lineages, this might very well explain why gBGC shows contrasting effects on d_N/d_S. However, more specific models are needed to fully understand how gBGC alters d_N/d_S.

Besides gBGC, deviations of individual gene trees from the species tree as a consequence of incomplete lineage sorting (ILS) could impact estimates of d_N/d_S if estimation is based on the species tree [53]. Specifically, it has been suggested that gene tree discordance may lead to artificially higher substitution rates and also to an increase in d_N/d_S, especially when levels of ILS are high. Since ILS has been suggested to be abundant in birds [57], it is possible that gene tree heterogeneity could contribute to the observed correlations between life-history traits and d_N/d_S if the estimation of d_N/d_S is based on the species tree. On the other hand, it has been suggested that gBGC increases the error rate in tree inference [29, 34]. As a consequence, the estimation of d_N/d_S based on individual gene trees could be affected by a high error rate in tree inference, particularly so for genes located in regions strongly affected by gBGC. However, comparison of correlations between life-history traits and d_N/d_S, one based on the species tree and one based on individual gene trees, suggests that our conclusion is robust to gene tree heterogeneity. For both approaches, we observe no significant correlations between life-history traits and d_N/d_S based on all substitution categories, but significant positive correlations if d_N/d_S is based on GC-conservative changes only (Fig. 1, Table 1, and Additional file 2: Table S3).

Yet other factors that could impact estimates of d_N/d_S and consequently the correlation between life-history traits and d_N/d_S in birds are alignment quality and/or missing genes [32]. After accounting for alignment errors (and inclusion of previously undetected genes), Botero-Castro et al. [32] found a significant positive correlation between d_N/d_S and longevity. They also found that d_N/d_S was significantly correlated with body mass in the Neoaves clade, i.e., all birds except Paleognathae (ratites and tinamous) and Galloanserae (gamefowl and waterfowl). The authors suggested that the latter group of birds might be outliers, since for those birds current life-history traits might not properly reflect long-term N_e (see also ref. [25]). However, after correction for phylogenetic relationships of species, we no longer observe striking differences between correlations for their dataset. Moreover, in the larger dataset analyzed in this study, we do not observe that Paleognathae and Galloanserae deviate from the overall pattern. Taken together,

the hypothesis that gBGC conceals the prediction of the nearly neutral theory appears robust to the set of genes analyzed as well as to alignment quality. d_N/d_S based on GC-conservative changes showed strong positive correlations with all life-history traits analyzed in this study in all datasets, while correlations were weak when analyzing all substitutions together. This demonstrates that gene set and alignment quality alone do not explain the lack of a positive correlation between life-history traits and d_N/d_S in birds.

The correlations between life-history traits and d_N/d_S based on GC-conservative changes were robust to variation in local GC content and chromosome location. We observed significant positive correlations for GC-rich and GC-poor genes, as well as for genes located in macrochromosomes or in microchromosomes. Estimates of d_N/d_S were lower for genes with higher GC content and for genes located in microchromosomes. This reduction appears to be unrelated to gBGC, as it is observed for all substitution categories including GC-conservative substitutions. We found that the reduction in d_N/d_S is a result of lower d_N values at the same time as d_S values are increased. Higher d_S values for microchromosomes than for macrochromosomes have been reported before and were suggested to be a result of higher mutation rates on microchromosomes than on macrochromosomes [58]. Still, d_N was reduced for GC-rich genes and genes located on microchromosomes. In birds, recombination rate is positively correlated with GC content and negatively with chromosome size [59, 60]. Thus, the reduction in d_N could be a result of a lower interference between selected sites (Hill-Robertson interference) due to higher recombination rates [61]. Alternatively, more rapid sequence saturation in GC-rich regions relative to GC-poor regions could influence estimation of d_N/d_S. Weber et al. [31] investigated this particular concern in avian genomes and found that sequence saturation affects synonymous sites more strongly than non-synonymous sites, leading to a greater underestimation of d_S relative to d_N [31]. As a consequence, the d_N/d_S ratio would be increased by sequence saturation in GC-rich regions relative to GC-poor regions. Given that we observe lower d_N/d_S in GC-rich regions than in GC-poor regions, we believe that sequence saturation should not be of concern for our conclusions.

Conclusions

We have explored the role of gBGC behind the apparent lack of a positive correlation between life-history traits and d_N/d_S in birds, a correlation that would be expected based on the prediction of the nearly neutral theory of molecular evolution. By estimating nucleotide substitution rates separately for different substitution categories, we observe strong positive correlations between three life-history traits (body mass, age of sexual maturity and longevity) and d_N/d_S when the estimation of d_N/d_S is based on GC-conservative substitutions only. No significant correlations were observed when estimation of d_N/d_S was based on all substitutions together, which can be ascribed to the influence of gBGC on S-to-W or W-to-S substitutions. gBGC thus impacts the correlation between life-history traits and d_N/d_S, and after accounting for gBGC, we find that the efficacy of selection increases with proxies for N_e, as predicted by the nearly neutral theory. The impact of gBGC on d_N/d_S varies substantially among lineages, where it increases the d_N/d_S ratio in some lineages, but decreases it in others. Our analysis suggests that these contrasting effects are related to differences in the dynamics of base composition between non-synonymous and synonymous substitutions. Moreover, altogether our study clearly illustrates that gBGC interferes with natural selection and that accounting for gBGC is a crucial step to correctly infer measures of the efficacy of natural selection such as the d_N/d_S ratio in comparative genomic studies. In light of this observation, we suggest that conclusions of previous studies that ignored the impact of gBGC on molecular evolutionary rates might need a careful re-evaluation. We here provide a protocol to do this.

Methods

Multiple sequence alignments

We downloaded publicly available coding sequence alignments of 8253 orthologous genes from 48 avian genomes and their inferred phylogenetic tree from Jarvis et al. [57]. We excluded the white-tailed eagle (*Haliaeetus albicilla*), since the branch length between this species and another eagle species included in the data set (bald eagle, *Haliaeetus leucocephalus*), was too short to reliably estimate d_N/d_S ratios [62]. We further excluded Z-linked genes according to the chicken annotation from Ensembl v90 (Gallus_gallus-5.0) [63, 64], since the efficacy of selection differs between sex-linked and autosomal genes [65, 66]. This resulted in a dataset of 7986 orthologous genes of 47 bird species. In addition, following previous approaches [30], only codons represented in at least 41 species were retained.

Two additional sets of multiple sequence alignments were retrieved from Botero-Castro et al. [32]. The first of these, which we refer to as the Figuet+HMMclean dataset, is a set of 1077 avian genes analyzed by Figuet et al. [30]. This dataset is a subset of the Jarvis et al. [57] data, which are used in the present study. The dataset consisted of genes shared by sauropsids and mammals based on orthology prediction between chicken, green anole, and human. Misaligned sites in these alignments were subsequently filtered using HMMclean [32]. The second set of alignments, which

we refer to as Botero-Castro dataset, included all genes in the Figuet+HMMclean dataset plus a set of 1245 previously undetected avian GC-rich genes [32]. For further description on the gene sets and cleaning methods, we refer to the original publications.

Concatenation of sequence alignments

Concatenation of gene-sequence alignments into a single alignment caused problems with large computational memory for the subsequent estimation of substitution rates.

Therefore, we randomly concatenated gene alignments into 20 different bins of roughly 400 genes each and estimated substitution rates for each bin. The average d_N and d_S across all bins were computed to obtain genome-wide averages of d_N/d_S. Concatenated alignments were used if estimation of substitution rates was based on the species tree. This approach was used to increase the signal-to-noise ratio for individual estimates. We repeated the analyses for 20 bins grouped according to their GC content. To do so, we first calculated the average GC content per species for each gene. Then, we calculated the mean GC content per gene over all species and ranked the genes accordingly. In order to separate genes into two classes of GC-rich and GC-poor genes, we split the 20 bins into two groups, each containing 10 bins with the highest and lowest GC content, respectively. We further repeated the analyses by binning genes according to their genomic location; one bin of genes located in microchromosomes and one bin of genes located in macrochromosomes, respectively. To avoid analyzing genes located in intermediate size chromosomes, we included genes located in chromosomes smaller than 16 Mb (chromosomes 14–28; a total of 1369 genes) in the category of microchromosomes and genes located in chromosomes larger than 100 Mb (chromosomes 1–3; a total of 1321 genes) in the category of macrochromosomes. We excluded genes from chromosome 1 if their location in the zebra finch was chromosome 1A, representing a well-known chromosomal rearrangement, or unknown (according to Ensemble v.90; [64]).

For the Figuet+HMM and Botero-Castro datasets, the same binning procedure as for the main analyses was performed. We randomly concatenated gene alignments into bins of roughly 400 genes each and estimated substitution rates for each bin. Subsequently, the average d_N and d_S estimates across all bins were computed to obtain genome-wide averages of d_N/d_S.

Estimation of lineage-specific d_N/d_S

We estimated d_N/d_S for each bin by using the bio++ libraries [50, 51] and the total evidence nucleotide tree (TENT) species tree from Jarvis et al. [57]. As a first step, we used a non-stationary homogeneous codon model of molecular evolution implemented in bppml to retrieve the most likely branch lengths, codon frequencies at the root, and substitution model parameters. We implemented three different substitution models, which all allow for different GC content dynamics between codon positions, YN98 (F3X4) [67], T92X3 [68], and L95X3 [69]. As a second step, we used MapNH for substitution mapping to estimate d_N and d_S, and then d_N/d_S [49, 52, 70]. The idea of the second step is to compute, on any branch, the expectation of any random variable over all the possible histories provided by the model and tree optimized in the first step. For example, d_S is the ratio of the number of synonymous substitutions performed by the model divided by the number of synonymous substitutions that would have been performed by a similar model, but set as neutral (i.e., $\omega = 1$ for YN98 model). So, the program computes the expectation of both numbers and returns the ratio as d_S. In an analogous way, d_N is computed for non-synonymous substitutions. Since it is possible to consider any kind of event, we also computed the ratios of the expected numbers of non-synonymous and synonymous substitutions restricted to specific substitution categories, such as S-to-S, S-to-W, W-to-S, and W-to-W. For example, d_S for the W-to-S substitution category is the expected number of W-to-S synonymous substitutions, divided the number of such substitutions that would have been performed by a similar neutral model. To combine S-to-S and W-to-W substitution rates into one category, i.e., the GC-conservative category, we weighed the S-to-S and W-to-W substitution rates of each species according to the GC content at the most recent internal node of each particular tip of the tree. To reconstruct the most likely ancestral sequence at each node, we used the bppancestor program [51, 71].

To account for gene tree heterogeneity, we estimated d_N/d_S with an additional approach, where the number of substitutions was estimated separately for each gene (instead of concatenated bins) using gene-specific phylogenetic trees (based on the first and second codon positions) from Jarvis et al. [57]. In order to obtain genome-wide averages of d_N, d_S, and d_N/d_S, the sum of number of substitutions for each substitution category over all genes was computed and normalized by the sum of all the expected substitutions for the same category.

Estimation of GC content

Current and ancestral GC content were estimated for 0- and 4-fold degenerate sites as proxies for non-synonymous and synonymous GC content, respectively. Current GC content was defined as the sum of G and C nucleotides in the respective lineage. As described above, ancestral GC content refers to the GC content at the most recent internal node of a particular tip of the tree. In

addition, we computed equilibrium GC content for non-synonymous and synonymous changes (GC*) as W-to-S/(W-to-S + S-to-W) substitution rates separately for each lineage. Lineage-specific ΔGC was estimated as the difference between lineage-specific GC* and the lineage-specific ancestral GC content.

Life-history traits

Estimates of body mass (grams), maximum longevity (years), and age of sexual maturity (days) were retrieved from Figuet et al. [30]. Briefly, the authors retrieved body mass estimates from the CRC Handbook of Avian Body Masses [72], while longevity and age of sexual maturity were retrieved from the literature (for detailed information see supplementary tables of the original publication [30]). All estimates were log-transformed to the base of 10.

Statistical analyses

All statistical analyses were performed in R v. 3.2.2 (R core team 2015). Pearson's correlation coefficient was used to test for correlations among variables. To account for phylogenetic relationships of species in correlation analysis, we used the ape package in R in order to compute phylogenetically independent contrasts following the method described by Felsenstein [73].

Additional files

Additional file 1: Table S1. Species specific dN/dS estimates for different substitution categories. (XLSB 56 kb)

Additional file 2: Supplementary Tables S2-S6, and Supplementary Figures S1-S8. (DOCX 608 kb)

Acknowledgements

The authors acknowledge three anonymous reviewers for constructive comments. Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

Funding

This work was supported by the Swedish Research Council (2013–8271 to HE) and the Knut and Alice Wallenberg Foundation (2014/0044 to HE).

Availability of data and materials

Coding sequence alignments and gene-specific phylogenetic trees of 48 avian species from Jarvis et al. [57] were retrieved at GigaDB [74], https://doi.org/10.5524/101041. Two additional sets of multiple sequence alignments from Botero-Castro et al. [32] were retrieved at https://doi.org/10.6084/m9.figshare.5202853. The Bio++ source code and custom scripts used in this study have been deposited at the Zenodo research data repository and are available at https://doi.org/10.5281/zenodo.2149686. The Bio++ source code and a tutorial on the computation of d_N and d_S based on stochastic mapping, taking into account gBGC, have also been deposited in a public repository under the link: https://github.com/BioPP/supp-mat/tree/master/mapdNdS.

Authors' contributions

CFM conceived the study and supervised its execution. PB performed statistical and bioinformatics analysis. LG implemented the Bio++ source code. All authors reviewed the results and discussed the analysis. CFM and PB wrote the manuscript with contributions from all other authors. All authors approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Department of Ecology and Genetics, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden. [2]Laboratoire de Biologie et Biométrie Évolutive CNRS UMR 5558, Université Claude Bernard Lyon 1, Lyon, France.

Received: 12 January 2018 Accepted: 17 December 2018
Published online: 07 January 2019

References

1. Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217:624–6.
2. Siepel A, Bejerano G, Pedersen JS, Hinrichs A, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.
3. Ohta T. Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor Popul Biol. 1976;10:254–75.
4. Ohta T. Extension to the neutral mutation random drift hypothesis. In *Molecular Evolution and Polymorphism*. Edited by Kimura M. Mishima: National Institute of Genetics Publications; 1977.
5. Ohta T. The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst. 1992;23:263–86.
6. Wright S. Evolution in Mendelian populations. Genetics. 1931;16:0097–159.
7. Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1983.
8. Akashi H, Osada N, Ohta T. Weak selection and protein evolution. Genetics. 2012;192:15–31.
9. Hughes AL. Near neutrality: leading edge of the neutral theory of molecular evolution. Ann N Y Acad Sci. 2008;1133:162–79.
10. Caballero A. Developments in the prediction of effective population-size. Heredity. 1994;73:657–79.
11. Gilbert KJ, Whitlock MC. Evaluating methods for estimating local effective population size with and without migration. Evolution. 2015;69:2154–66.
12. Wang J, Santiago E, Caballero A. Prediction and estimation of effective population size. Heredity. 2016;117:193–206.
13. Ellegren H, Galtier N. Determinants of genetic diversity. Nat Rev Genet. 2016;17:422–33.
14. Nabholz B, Mauffrey JF, Bazin E, Galtier N, Glemin S. Determination of mitochondrial genetic diversity in mammals. Genetics. 2008;178:351–61.
15. Waples RS, Luikart G, Faulkner JR, Tallmon DA. Simple life-history traits explain key effective population size ratios across diverse taxa. Proc Royal Soc B-Biol Sci. 2013;280:20131339.
16. Woolfit M, Bromham L. Population size and molecular evolution on islands. Proc Royal Soc B-Biol Sci. 2005;272:2277–82.
17. Nomura T, Takahashi J. Effective population size in eusocial Hymenoptera with worker-produced males. Heredity. 2012;109:261–8.
18. Jarne P. Mating system, bottlenecks and genetic-polymorphism in hermaphroditic animals. Genet Res. 1995;65:193–207.
19. Hill K, Kaplan H. Life history traits in humans: theory and empirical studies. Annu Rev Anthropol. 1999;28:397–430.
20. White EP, Morgan Ernest SK, Kerkhoff AJ, Enquist BJ. Relationships between body size and abundance in ecology. Trends Ecol Evol. 2007;22:323–30.
21. Romiguier J, Lourenco J, Gayral P, Faivre N, Weinert LA, Ravel S, Ballenghien M, Cahais V, Bernard A, Loire E, et al. Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. J Evol Biol. 2014;27:593–603.

22. Glemin S, Muyle A. Mating systems and selection efficacy: a test using Chloroplastic sequence data in angiosperms. J Evol Biol. 2014;27:1386–99.
23. Lartillot N, Delsuc F. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. Evolution. 2012;66:1773–87.
24. Lartillot N, Poujol R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Mol Biol Evol. 2011;28:729–44.
25. Nabholz B, Uwimana N, Lartillot N. Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds. Genome Biology and Evolution. 2013;5:1273–90.
26. Nikolaev SI, Montoya-Burgos JI, Popadin K, Parand L, Margulies EH, Antonarakis SE, Program N. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. Proc Natl Acad Sci U S A. 2007;104:20443–8.
27. Popadin K, Polishchuk LV, Mamirova L, Knorre D, Gunbin K. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. Proc Natl Acad Sci U S A. 2007;104:13390–5.
28. Romiguier J, Figuet E, Galtier N, Douzery EJP, Boussau B, Dutheil JY, Ranwez V. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. PLoS One. 2012;7:e33852.
29. Romiguier J, Ranwez V, Douzery EJP, Galtier N. Genomic evidence for large, long-lived ancestors to placental mammals. Mol Biol Evol. 2013;30:5–13.
30. Figuet E, Nabholz B, Bonneau M, Carrio EM, Nadachowska-Brzyska K, Ellegren H, Galtier N. Life history traits, protein evolution, and the nearly neutral theory in amniotes. Mol Biol Evol. 2016;33:1517–27.
31. Weber CC, Nabholz B, Romiguier J, Ellegren H. Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. Genome Biol. 2014;15:542.
32. Botero-Castro F, Figuet F, Tilak M, Nabholz B, Galtier N. Avian genomes revisited: hidden genes uncovered and the rates versus traits paradox in birds. Mol Biol Evol. 2017;34:3123–31.
33. Mugal CF, Arndt PF, Ellegren H. Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. Mol Biol Evol. 2013;30:1700–12.
34. Nabholz B, Kunstner A, Wang R, Jarvis ED, Ellegren H. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Mol Biol Evol. 2011;28:2197–210.
35. Smeds L, Mugal CF, Qvarnstrom A, Ellegren H. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. PLoS Genet. 2016;12:e1006044.
36. Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. Genome Biol. 2014;15:549.
37. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009;10:285–311.
38. Mugal CF, Weber CC, Ellegren H. GC-biased gene conversion links the recombination landscape and demography to genomic base composition GC-biased gene conversion drives genomic base composition across a wide range of species. BioEssays. 2015;37:1317–26.
39. Nagylaki T. Evolution of a finite population under gene conversion. Proc Natl Acad Sci USA Biol Sci. 1983;80:6278–81.
40. Backstrom N, Zhang Q, Edwards SV. Evidence from a house finch (Haemorhous Mexicanus) spleen transcriptome for adaptive evolution and biased gene conversion in passerine birds. Mol Biol Evol. 2013;30:1046–50.
41. Bolívar P, Mugal CF, Nater A, Ellegren H. Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill-Robertson interference, in an avian system. Mol Biol Evol. 2016;33:216–27.
42. Corcoran P, Gossmann TI, Barton HJ, Consortium GTH, Slate J, Zeng K. Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. Genome Biol Evol. 2017;9:2987–3007.
43. Kostka D, Hubisz MJ, Siepel A, Pollard KS. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. Mol Biol Evol. 2012;29:1047–57.
44. Lartillot N. Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. Mol Biol Evol. 2013;30:356–68.
45. Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR, Doll A, Da Fonseca RR, Fulton TL, et al: Natural selection shaped the rise and fall of passenger pigeon genomic diversity. Science 2017, 358**:** 951–954.
46. Ratnakumar A, Mousset S, Glemin S, Berglund J, Galtier N, Duret L, Webster MT. Detecting positive selection within genomes: the problem of biased gene conversion. Philos Trans Royal Soc B. 2010;365:2571–80.
47. Bolívar P, Mugal CF, Rossi M, Nater A, Wang M, Dutoit L, Ellegren H: Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. Mol Biol Evol. 2018;35:2475–86.
48. Galtier N, Duret L, Glemin S, Ranwez V. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates (Vol 25, Pg 1, 2009). Trends Genet. 2009;25:287.
49. Guéguen L, Duret L: Unbiased estimate of synonymous and non-synonymous substitution rates with non-stationary base composition. Mol Biol Evol. 2017;35:734–42.
50. Dutheil J, Boussau B. Non-homogeneous models of sequence evolution in the bio++ suite of libraries and programs. BMC Evol Biol. 2008;8:255.
51. Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V, et al. Bio++ : efficient extensible libraries and tools for computational molecular evolution. Mol Biol Evol. 2013;30:1745–50.
52. Minin VN, Suchard MA. Fast, accurate and simulation-free stochastic mapping. Philos Trans Royal Soc B. 2008;363:3985–95.
53. Mendes FK, Hahn MW. Gene tree discordance causes apparent substitution rate variation. Syst Biol. 2016;65:711–21.
54. Romiguier J, Ranwez V, Douzery EJP, Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. 2010;20:1001–9.
55. Capra JA, Pollard KS. Substitution patterns are GC-biased in divergent sequences across the metazoans. Genome Biol Evol. 2011;3:516–27.
56. Berglund J, Pollard KS, Webster MT. Hotspots of biased nucleotide substitutions in human genes. PLoS Biol. 2009;7:45–62.
57. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 2014;346:1320–31.
58. Axelsson E, Webster MT, Smith NGC, Burt DW, Ellegren H. Comparison of the chicken and Turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. Genome Res. 2005;15:120–5.
59. Kawakami T, Smeds L, Backstrom N, Husby A, Qvarnstrom A, Mugal CF, Olason P, Ellegren H. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. Mol Ecol. 2014;23:4035–58.
60. Backstrom N, Forstmeier W, Schielzeth H, Mellenius H, Nam K, Bolund E, Webster MT, Ost T, Schneider M, Kempenaers B, Ellegren H. The recombination landscape of the zebra finch Taeniopygia guttata genome. Genome Res. 2010;20:485–95.
61. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res. 1966;8:269–94.
62. Mugal CF, Wolf JBW, Kaj I. Why time matters: codon evolution and the temporal dynamics of dN/dS. Mol Biol Evol. 2014;31:212–31.
63. Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F, et al. A new chicken genome assembly provides insight into avian genome. Structure G3 (Bethesda). 2017;7:109–17.
64. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. Ensembl 2016. Nucleic Acids Res. 2016;44:D710–6.
65. Singh ND, Larracuente AM, Clark AG. Contrasting the efficacy of selection on the X and autosomes in drosophila. Mol Biol Evol. 2008;25:454–67.
66. Rousselle M, Faivre N, Ballenghien M, Galtier N, Nabholz B. Hemizygosity enhances purifying selection: lack of fast-Z evolution in two satyrine butterflies. Genome Biol Evol. 2016;8:3108–19.
67. Yang ZH, Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 1998;46:409–18.
68. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Mol Biol Evol. 1993;10:512–26.
69. Lobry JR. Properties of a general-model of DNA evolution under no-strand-bias conditions. J Mol Evol. 1995;40:326–30.

70. Dutheil JY, Galtier N, Romiguier J, Douzery EJP, Ranwez V, Boussau B. Efficient selection of branch-specific models of sequence evolution. Mol Biol Evol. 2012;29:1861–74.
71. Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, Galtier N, Belkhir K. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. BMC Bioinformatics. 2006;7:188.
72. Dunning Jr JB. CRC handbook of avian body masses. 2nd ed. Boca Raton: CRC Press; 2007.
73. Felsenstein J. Phylogenies and the comparative method. Am Nat. 1985; 125:1–15.
74. Jarvis ED, Mirarab S, Aberer AJ, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, Suh A, et al: (2014): Phylogenomic analyses data of the avian phylogenomics project. GigaScience Database. https://doi.org/10.5524/101041