Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds

Fidel Botero-Castro, Emeric Figuet, Marie-Ka Tilak, Nicolas Galtier, Benoit Nabholz

Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds

Fidel Botero-Castro,* Emeric Figuet, Marie-Ka Tilak, Benoit Nabholz, and Nicolas Galtier

Institut des Sciences de l'Evolution, UMR5554, Universitè de Montpellier, CNRS, IRD, EPHE, Montpellier, 34095, France

*Corresponding author: E-mail: fidel.boteroc@gmail.com.

Associate editor: Emma Teeling

Abstract

According to current assemblies, avian genomes differ from those of the other lineages of amniotes in 1) containing a lower number of genes; 2) displaying a high stability of karyotype and recombination map; and 3) lacking any correlation between evolutionary rates (d*N*/d*S*) and life-history traits, unlike mammals and nonavian reptiles. We question the reality of the bird missing genes and investigate whether insufficient representation of bird gene content might have biased previous evolutionary analyses. Mining RNAseq data, we show that the vast majority of the genes missing from avian genome assemblies are actually present in most species of birds. These mainly correspond to the GC-rich fraction of the bird genome, which is the most difficult to sequence, assemble and annotate. With the inclusion of these genes in a phylogenomic analysis of high-quality alignments, we uncover a positive and significant correlation between the ratio of nonsynonymous to synonymous substitution rate (d*N*/d*S*) and life-history traits in Neoaves. We report a strong effect of GC-biased gene conversion on the d*N*/d*S* ratio in birds and a peculiar behavior of Palaeognathae (ostrich and allies) and Galloanserae (chickens, ducks and allies). Avian genomes do not contain fewer genes than mammals or nonavian reptiles. Previous analyses have overlooked ~15% of the bird gene complement. GC-rich regions, which are the most difficult to access, are a key component of amniote genomes. They experience peculiar molecular processes and must be included for unbiased functional and comparative genomic analyses in birds.

Key words: **aves, missing genes, evolutionary rate, life-history traits, recombination, GC content, biased gene conversion.**

Background

Avian genomes differ in many respects from those of the other major lineages of amniotes, that is, mammals and nonavian reptiles. First, birds have a smaller genome size mainly due to a lower fraction of transposable elements and repetitive DNA (Hughes and Piontkivska 2005; Organ et al. 2007) as well as shorter introns (Hughes and Hughes 1995). Interestingly, several studies have also suggested that a massive loss of genes, likely associated with chromosomal rearrangements (Hillier et al. 2004; Lovell et al. 2014; Zhang et al. 2014), may have occurred in birds. The analysis of genome-wide data in 48 species of birds estimated that the total number of genes in avian genomes was around 70% of those present in humans (Zhang et al. 2014). Lovell et al. (2014) analyzed genome assemblies in birds, turtle, lizard and humans and identified 274 genes in synthenic blocks predicted to have been lost in the bird lineage. Hron et al. (2015), however, showed that some of these genes are actually present in chicken and in the Tibetan ground tit *Pseudopodoces humilis*, and that they tend to have a high GC content and carry long GC stretches. The question of how many genes are truly missing from avian genomes versus not properly assembled or annotated thus remains open. Whether missing or hidden, these genes have been absent in birds comparative genomic studies.

Compared with mammals, another remarkable feature of avian genomes is the high stability of karyotype, the conserved landscape of recombination rates and the high level of synteny. The typical avian karyotype, which consists in 14–16 macrochromosomes and numerous microchromosomes, has not varied much since the common ancestor of birds (Burt 2002; Ellegren 2010). Besides a conspicuous difference in size, microchromosomes have been shown to harbour higher gene-density, GC content and recombination rates than macrochromosomes (Smith et al. 2000; Burt 2002). The way recombination operates also differs between mammals and birds. In the former, recombination takes place in hotspots that are alternatively turned on and off as determined by the *PRDM9* protein and the genomic distribution of the DNA motif it recognizes (Baudat et al. 2010). In contrast, *PRDM9* is absent in birds (Oliver et al. 2009) and recombination, rather than taking place at variable-in-time chromosome locations, is highly localized in the neighborhood of transcription start/termination sites and CpG islands (Singhal et al. 2015).

Recombination is tightly related to the phenomenon of GC-biased gene conversion (gBGC), which preferentially repairs mismatches resulting from double-strand breaks

Article

towards G/C. Consequently, GC-content is also evolving differently in birds versus mammals. Whereas in mammals GC-content evolution appears erratic across lineages and genomic regions (Romiguier et al. 2010), GC-content in birds has been increasing in time, especially in high-recombining regions, and is still far from its equilibrium (Webster et al. 2006; Nabholz et al. 2011; Bolívar et al. 2016). Given the particular evolutionary dynamics of GC-rich genes, the high levels of GC-content and gene density are in the mostly unassembled or unannotated microchromosomes, and the apparent trend for GC-rich genes to be reported as missing, it is legitimate to investigate if there is a relationship between the two phenomena: 1) how many of the genes, reported as missing, are really absent in avian genomes? And 2) for those present, how many were unassembled because of extreme GC-content?

Birds, finally, are peculiar with respect to the relationship between evolutionary rates and life-history traits (LHT). A positive correlation between the ratio of nonsynonymous to synonymous substitution rates (d*N*/d*S*), and body mass, longevity or age of sexual maturity has been reported in mammals and nonavian reptiles (Romiguier et al. 2013; Figuet et al. 2014; Figuet et al. 2016). This correlation is interpreted as reflecting an effect of the effective population size (Ne): large/long-lived species would have, on average, a smaller Ne than small/short lived ones, and therefore experience less efficient purifying selection, leading to an increased rate of fixation of slightly deleterious nonsynonymous mutations, inflating the d*N*/d*S* ratio. Interestingly, no such relationships have been detected in birds, which, as far as d*N*/d*S* is concerned, apparently depart the nearly neutral theory (Weber et al. 2014; Figuet et al. 2016). The exact reasons for this paradox are unexplained although different hypotheses including the action of gBGC, the lack of correlation between life-history traits and historical population size and unreliable estimates of d*N*/d*S* have been suggested (Lartillot 2013; Bolívar et al. 2016; Figuet et al. 2016; Hua and Bromham 2017). gBGC and GC-content are known to affect the d*N*/d*S* ratio (Berglund et al. 2009; Galtier et al. 2009; Bolívar et al. 2016), so that, again, the impact of hidden/missing genes on rates versus traits relationships in birds is worth investigating.

We here mined transcriptome data and coding sequences predicted by an automated pipeline of annotation in 78 species of birds. We report that a vast majority of the purportedly missing genes are actually present in a large number of species, and present a higher GC-content, on average, than those present in current genomic assemblies. Then, we revisit the correlation between d*N*/d*S* and LHT using this new set of genes and uncover a positive relationship with longevity for all avian species and for body mass for Neoaves only. Interestingly, paleognaths and galloanseres show d*N*/d*S* values that deviate from what would be expected given their LHT.

Results

Newly Annotated Bird Orthologs

Querying the ENSEMBL database, we identified 2,454 genes annotated as one-to-one orthologs between human (*Homo sapiens*) and the Chinese soft-shell turtle (*Pelodiscus sinensis*), but missing in chicken (*Gallus gallus*) and collared flycatcher (*Ficedula albicollis*). These were reciprocally BLASTed to *de novo* transcriptome assemblies and/or NCBI-predicted coding sequences from 79 species of birds. We recovered between 519 and 1,775 (average: 1,200) orthologous copies of these genes in each of the targeted species, such that 2,132 (86.9%) of the 2,454 purportedly missing genes were found in at least one species of birds (fig. 1, supplementary fig. S2, Supplementary Material online). Of the 1,574 sequences recovered this way in *Gallus gallus*, 474 were found to be present in the phylogenomic data set of Jarvis et al (2014), 585 more were also annotated in chicken genome assembly GALgal_v4, and 515 represent newly annotated orthologs. The species for which both transcriptomic data and NCBI-predicted coding sequences were available produced higher numbers of new orthologs than those for which data consisted of only one of the two sources of data (supplementary fig. S1, Supplementary Material online). 75.5% of the purportedly missing genes were recovered in ten distinct species or more, and 51.5% were found in 40 species or more (supplementary fig. S2, Supplementary Material online).

The new set of genes exhibited substantially higher GC and GC at third codon position (GC3) values, on average, than previously annotated genes (fig. 2), in agreement with previous suggestions that the GC-rich fraction of the bird genome is badly annotated. Moreover, we found a negative correlation ($r = -0.42$, $P < 2.2e\text{-}16$) between mean GC content of genes and the number of species for which a given gene was recovered (fig. 3), this is, genes with high GC-content tend to be harder to recover than those with low or intermediate GC. We examined the GC-content in turtle of the 353 human-turtle orthologs that were not recovered in any bird species in our analysis. We found that the distribution of GC-content for these still-missing genes is highly skewed towards elevated GC values, compared with the whole gene complement of *P. sinensis* (supplementary fig. S3, Supplementary Material online). This either implies that GC-rich genes have a higher probability than AT-rich genes to have been lost in birds, or that these extremely GC-rich genes are actually present in birds but so far unreachable, even with the help of RNAseq data.

d*N*/d*S* Ratio and Correlations with LHT and GC-Content

Focusing on 44 species analyzed by Jarvis et al. (Jarvis et al. 2014), we created a phylogenomic data set including the 1,077 genes of Figuet et al. (2016) and 1,245 genes newly identified here, for which orthology relationships were predicted. Ambiguously aligned sites were removed with the HMMcleaner program (Amemiya et al. 2013; Philippe et al. 2017). Based on this data set, significantly positive correlations between d*N*/d*S* ratio and each of body mass ($r = 0.59$, $P = 7.0 \times 10^{-5}$) and longevity ($r = 0.62$, $P = 1.2 \times 10^{-4}$) were recovered (fig. 4 and table 1). Interestingly, the five species of Palaeognathae (ostrich, tinamou) or Galloanserae (duck, chicken, turkey) had a lower d*N*/d*S* ratio than expected given their body mass (fig. 4*a*) or longevity (fig. 4*b*) and the

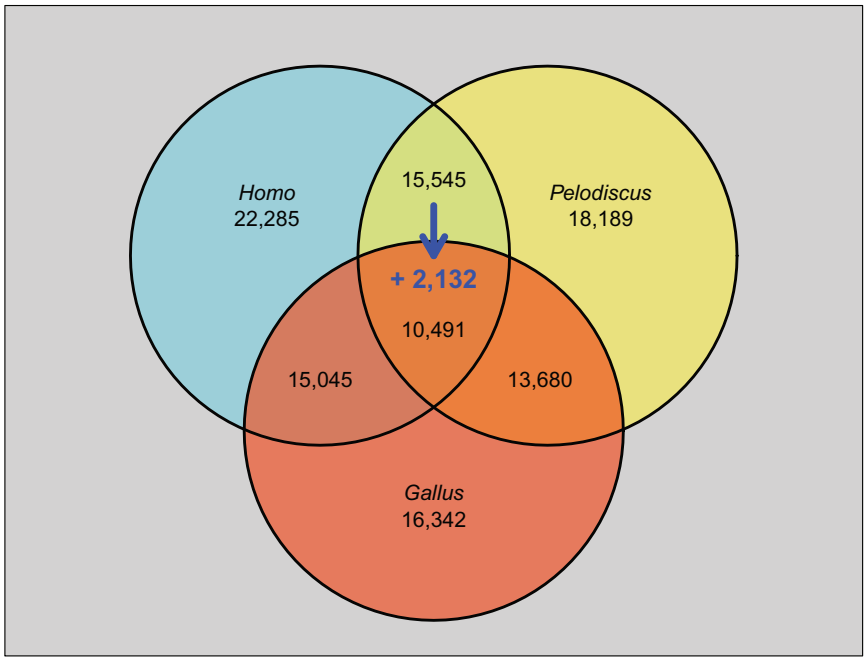

FIG. 1. Number of available annotations in human (*Homo*), Chinese soft-shell turtle (*Pelodiscus*) and chicken (*Gallus*). Numbers in intersections correspond to the amount of 1-to-1 orthologs available for each pair of species and, in blue, the total number of new sequences corresponding to orthologs missing in *Gallus* annotations that were recovered in at least one avian species.

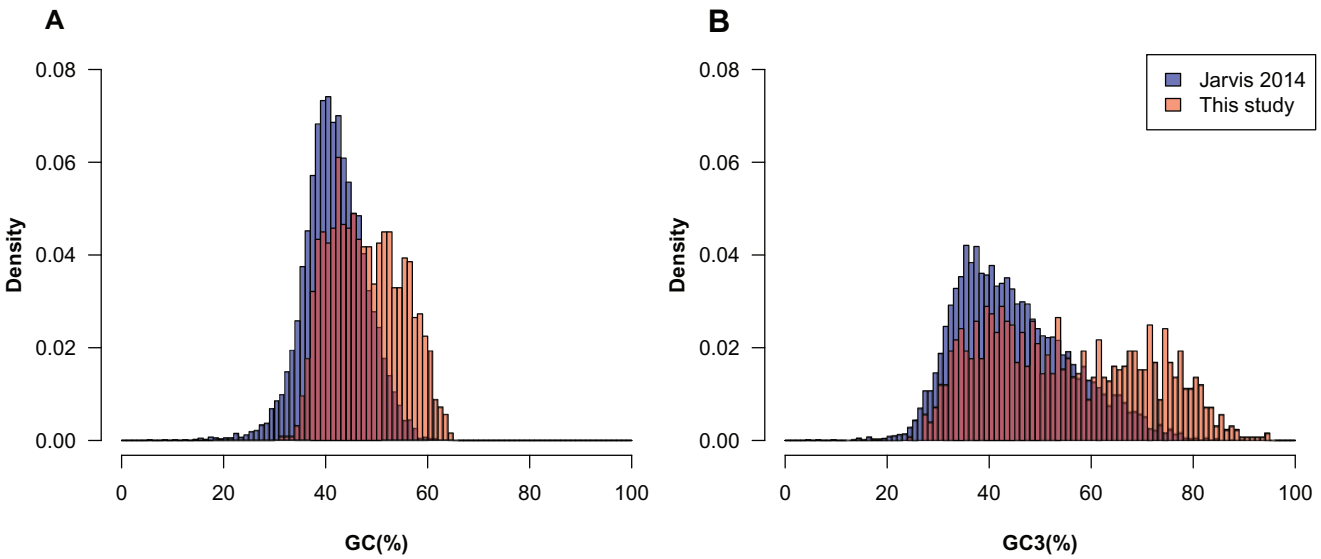

FIG. 2. Comparison of (*a*) GC total and (*b*) GC3 for the set of 8,235 genes produced in Jarvis et al (2014), in blue, and the sequences retrieved in this study, in red.

correlation with body mass is not significant when these species are included ($r = 0.23$, $P = 0.12$).

Genes were split in three bins according to GC-content and the analysis was reconducted for Neoaves species only. The three categories of genes yielded a significant, positive relationship between d*N*/d*S* and LHT. The average d*N*/d*S*, however, was strikingly higher in GC-poor genes than in the other two categories (fig. 5). To analyze this effect more deeply we separately measured the rate of nonsynonymous and of synonymous evolution across for each gene (total tree length) and we evaluated their variation as functions of gene mean GC-content. We found that the synonymous rate is consistently, positively correlated to GC-content ($r = 0.16$, $P = 4.67\text{e-}14$), whereas the nonsynonymous rate is slightly negatively related to GC-content ($r = -0.12$, $P = 2.24\text{e-}09$).

Discussion

Missing Genes in Birds?

The existence of a missing fraction of genes in avian genomes was pointed since the genome of chicken (*Gallus gallus*), the first avian genome sequenced, was published (Hillier et al.

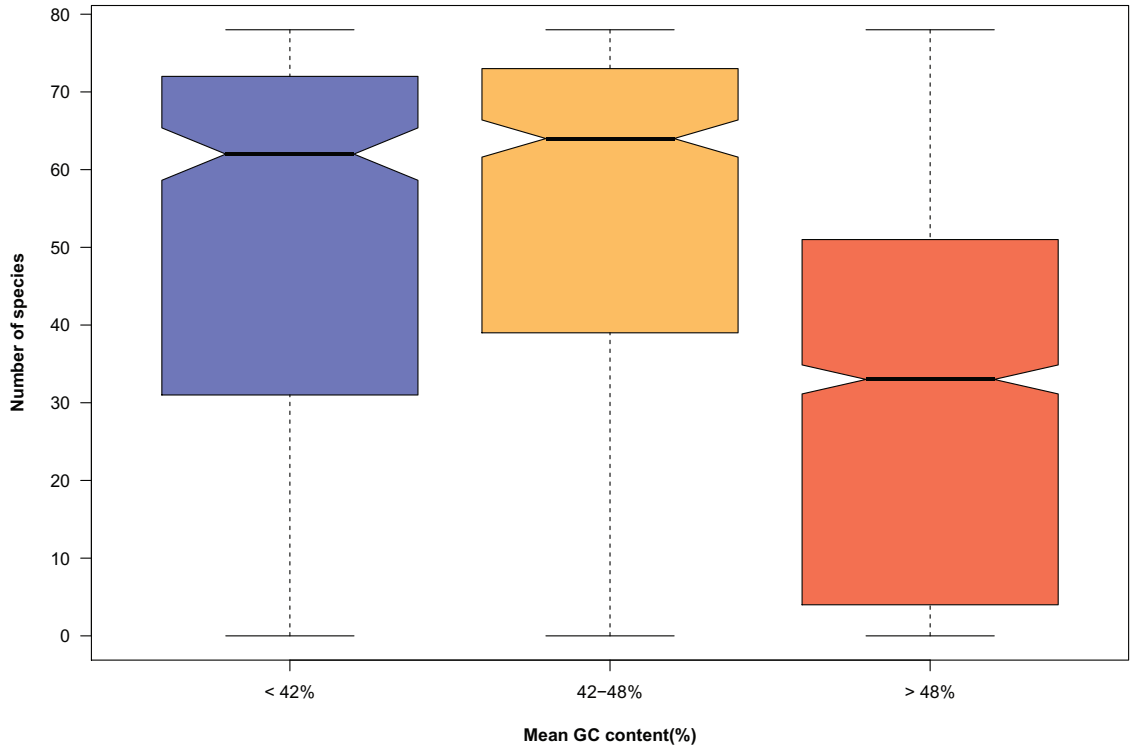

FIG. 3. Number of genes recovered as a function of GC content. Genes with high GC content (red) are less often recovered than those with low (blue) and intermediate (orange). *r*: −0.42; $P < 2.26E\text{-}16$.

2004). The authors explicitly mentioned that a number of genes present in RNA and EST data were only partially or fragmentarily recovered in the assembly and attributed the problem to the elevated GC-content of these genes. The chicken genome became the natural reference for annotation of subsequent avian genomes (Dalloul et al. 2010; Warren et al. 2010) and this fraction of unannotated genes has been systematically absent since then in bird comparative genomic studies.

Mining transcriptome data via BLAST or the Gnomon pipeline, we here demonstrate that the vast majority of these genes are actually present in the genome of many, if not all, bird species. We therefore confirm the usefulness of RNA-Seq in identifying genes missing from draft genome assemblies thus improving assessment of gene content (e.g., Denton et al. 2014). Our analysis retrieved 91 of the 273 genes reported as missing by Lovell et al. (2014) and predicted 515 new chicken-turtle orthologs. Importantly, human-turtle orthologs still missing in birds after our analysis are strongly GC-enriched in turtle (supplementary fig. S3, Supplementary Material online). Given the strong impact of GC-content on the probability for a gene to be missed in birds (fig. 3), this strongly suggests that many, if not all, of these still-missing genes are actually present in avian genomes. Indeed, our analysis confirms that GC content is a major cause for the fragmentary nature of genomic assemblies. Both the GC-poor (AT-rich) and the GC-rich (AT-poor) extremes have been documented to hinder assembly and annotation in several groups of organisms including the honeybee (Elsik et al. 2014) and the sand rat (*Psammomys obesus*; Hargreaves et al. 2017). In both cases, resequencing and incorporating transcriptomic data improved the assembly and gene annotation.

The application of single-molecule real-time (SMRT) sequencing technologies that do not rely on PCR has greatly improved gene annotation in birds (Damas et al. 2017; Warren et al. 2017; Korlach et al. 2017, http://biorxiv.org/content/early/2017/02/02/103911.abstract, last accessed April 14, 2017). As an illustration, the most recent version of chicken genome (GALgalv5, available from Ensemble v87) that was obtained using SMRT sequencing technology in combination with finished BACs and improved physical maps, provided chromosome-level scaffolds for four additional microchromosomes as well as annotations for >4,000 additional genes including 1,911 protein-coding genes, bringing the number of available coding DNA sequences (CDS) above 18,000 (Warren et al. 2017) compared with the ca. 16,000 CDS in the previous version.

The Correlation between dN/dS and LHT

Although strong positive correlations have been documented between dN/dS and body mass or longevity in mammals and nonavian reptiles (Romiguier et al. 2013; Figuet et al. 2016), such relationships have been reported to be absent in birds (Weber et al. 2014; Figuet et al. 2016). This is paradoxical because other molecular markers of the effective population size, such as within-species diversity and the ratio of non-synonymous to synonymous polymorphisms, do correlate with life-history traits in birds (Figuet et al. 2016). The two studies that reported an absence of the correlation (Weber et al. 2014; Figuet et al. 2016) analyzed gene sets derived from the alignment of Jarvis et al. (Jarvis et al. 2014), which includes

all orthologs recovered from genomic assemblies in 48 species of birds.

Here, we analyzed the 1,077 genes selected by Figuet et al. (2016) on the basis of their availability in nonavian reptiles plus the 1,245 new genes identified in this study, badly aligned sites being removed with HMMClean (Amemiya et al. 2013; Philippe et al. 2017). Our analysis recovered a significant, positive relationship between d*N*/d*S* and each of body mass and longevity, thus solving the rates versus traits paradox in birds. Birds are no exception to the nearly neutral theory: coding sequence evolutionary rate in this group appears to be dominated by the balance between purifying selection and genetic drift, body mass and longevity being reasonable predictors of the long-term Ne, consistent with other groups of amniotes. The relationship, however, has been more difficult to verify in birds than in mammals, in which a significant correlation has been reported in many different studies (Lartillot and Poujol 2011; Nabholz et al. 2013; Romiguier et al. 2013; Figuet et al. 2014, 2016; supplementary fig. S5, Supplementary Material online). Along the same lines, a positive correlation between LHT and d*N*/d*S* was obtained using the mitochondrial genome of mammals (Popadin et al. 2007) but not using the mitochondrial genome of birds (Nabholz et al. 2013). A narrower range of variation in LHT in birds could be invoked to explain this difference. Controlling for this effect, however, Figuet et al. (2016) still observed a different behavior in mammals compared with birds.

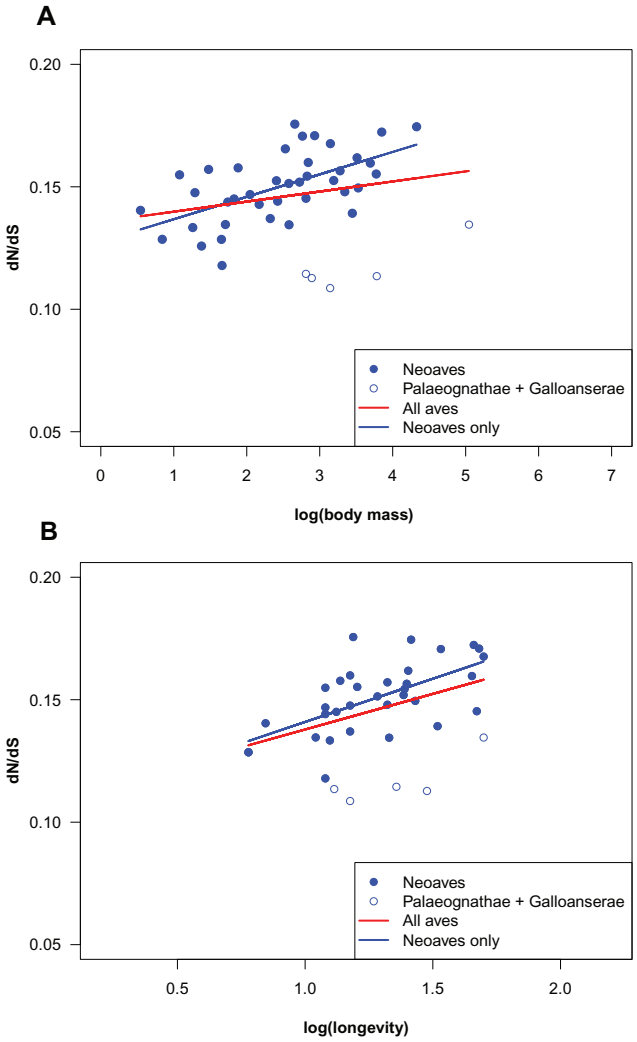

Fig. 4. Correlation between d*N*/d*S* and body mass (*a*) and longevity (*b*) in Aves. Empty dots correspond to Palaeognathae (*Struthio* and *Tinamou*) and Galloanserae (*Anas*, *Gallus*, and *Meleagris*). Solid dots correspond to Neoaves. Traits correspond to linear regressions for all aves in the data set (red) and Neoaves only (blue).

The HMMClean method models alignments of amino-acid sequences thanks to a hidden Markov Model and identifies sequence segments that depart the general pattern (Philippe et al. 2017). The impact of this cleaning step is appreciable: treating the alignments of Figuet et al. (2016) was sufficient to generate a significant correlation between d*N*/d*S* and life-history traits (table 1). Correlation coefficients were further increased and reached levels similar to those obtained in mammals and nonavian reptiles with the inclusion of the new genes identified in this study (table 1). Interestingly, we found that GC content had a major impact on d*N*/d*S* estimation and, thus, in the evaluation of these correlations. Indeed, as illustrated in figure 5, sets of genes differing in GC content produced very different d*N*/d*S* averages and the strength of the correlation with LHT was also impacted, further highlighting the need for an appropriate representation of the avian genome for proper evolutionary inferences.

The d*N*/d*S* versus LHT relationship we report is significant at the whole bird level for longevity and it is markedly stronger in Neoaves-only analyses (table 1) whereas for body mass the correlation is only significant when evaluated within Neoaves. In general, the deeply branching lineages of paleognaths and galloanseres are characterized by a lower d*N*/d*S* ratio than expected from their body masses and longevities (fig. 4), in agreement with the work of Nabholz et al. (2011). In particular, these lineages were also shown to be the exception to the estimated trend of a reduction in body mass during bird

Table 1. Correlations between Log-transformed values of body mass and longevity and d*N*/d*S* for 1) All Species in Data Set and 2) for Neoaves.[a]

Data set	Body mass				Longevity			
	r All aves	*P*	*r Neoaves*	*P*	*r All aves*	*P*	*r Neoaves*	*P*
Figuet et al (2016)	0.12 [−0.18–0.40]	0.44	0.24 [−0.085–0.51]	0.15	0.24 [−0.083–0.52]	0.14	0.32 [−0.027–0.60]	0.07
Figuet et al (2016) + HMMclean	0.17 [−0.14–0.44]	0.28	0.42 [0.12–0.65]	**0.004**	0.34 [0.028–0.060]	0.03	0.48 [0.16–0.71]	**0.005**
This study	0.23 [−0.072–0.49]	0.13	0.59 [0.19–0.69]	**7.344e-05**	0.40 [0.092–0.64]	0.013	0.62 [0.35–0.79]	**1.2e-04**

[a]Values in brackets correspond to 95% CI and *P* values in bold indicate significant correlations.

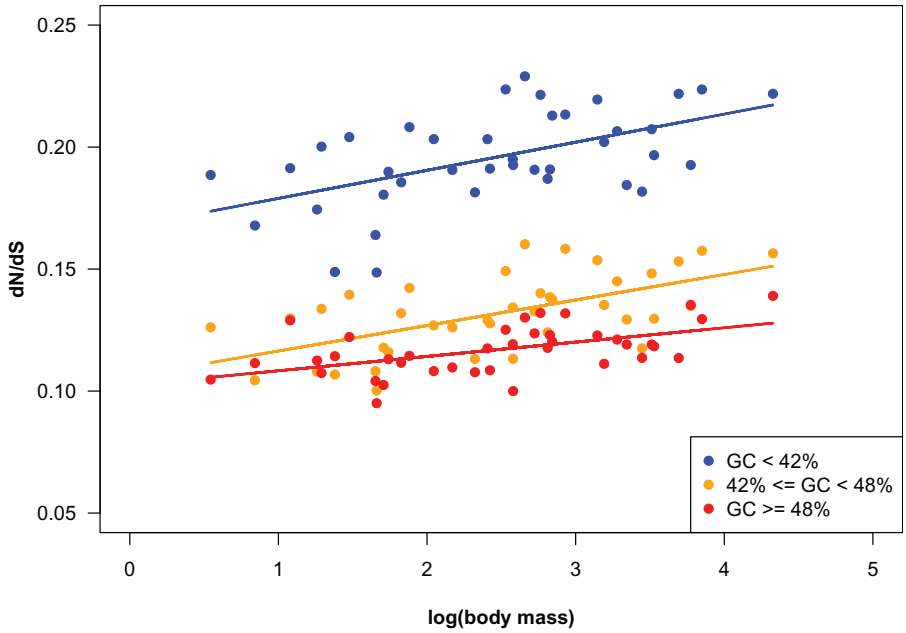

Fig. 5. Correlation between d*N*/d*S* and body mass in Neoaves, for gene sets differing in total GC content.

evolution (Nabholz et al. 2013). Paleognaths, in particular, are supposed to descend from a flying ancestor (Yonezawa et al. 2017), so that current traits might not reflect well the long-term effective population size in this group. The combination of these peculiarities could explain the outlying position of Palaeognathae and Galloanserae in the d*N*/d*S* versus LHT plots (fig. 4).

GC-Biased Gene Conversion Impacts Substitution Rates in Birds

Our report of a strong impact of gene GC-content on the d*N*/d*S* ratio across the bird phylogeny corroborates the analysis of Bolívar et al. (2016), who focused on the flycatcher lineage (*Ficedula albicollis*). Analyzing the genome-wide landscape of divergence with another passerine, the zebra finch, these authors reported a positive effect of GC-content and the recombination rate on the d*N*/d*S* ratio. At first sight, the correlation between d*N*/d*S* and recombination rate might be interpreted as reflecting Hill–Robertson interferences, that is, the negative effect of linkage on the efficiency of multilocus selection. Bolívar et al. (2016), however, argued that the relationship between recombination and d*N*/d*S* is rather mediated by gBGC, as attested by the stronger effect of recombination rate on d*S* than on d*N*, which is unexpected under the Hill–Robertson hypothesis.

Similarly, we here report a strong, positive correlation of d*S* with GC-content, again suggesting that Hill–Robertson effect is not the major force at work. Regarding d*N*, we detected a weaker but significantly negative relationship with GC-content regardless of the data set, oppositely to the results of Bolívar et al. (2016). The two studies analyzed distinct sets of genes and used distinct methods of estimation of d*N* and d*S*. Estimating d*N* and d*S* is particularly tricky in case of heterogeneous base composition (Guéguen L, Duret L, unpublished data, http://biorxiv.org/content/early/2017/04/06/124925.abstract, last accessed April 13, 2017). Additional research appears needed to characterize the complex impact of gBGC on the rate of amino-acid substitution.

Conclusions

Our analysis recovered a majority of the genes annotated as missing in birds according to ENSEMBL and genome assemblies, and suggests that many of the still-missing genes are part of the avian gene complement, contradicting previous claims of massive gene loss in the bird lineage. Based on our results, there is no strong reason to believe that avian genomes would contain fewer genes than mammals or nonavian reptiles. Most of the published analyses of bird genomes, therefore, have overlooked 10–20% of the gene complement, particularly genes carried by microchromosomes. We show that GC-rich regions, which are the most difficult to access, are a key component of the genomes of amniotes, experience peculiar molecular processes and must be included in genomic data sets for unbiased functional and comparative analyses in birds.

Material and Methods

Building the List of 'Missing' Genes

Lists of 1-to-1 orthologs were created for all pairs among humans (*Homo sapiens*) and Chinese soft-shell turtle (*Pelodiscus sinensis*) as nonavian references, and chicken (*Gallus gallus*) and collared flycatcher (*Ficedula albicollis*) as avian target species from ENSEMBL annotations as of version 86. Comparison of these lists yielded 2,454 protein-coding genes that were annotated in nonavian taxa but absent in avian annotations. The choice of *P. sinensis* as a reference is justified by 1) its place in the phylogeny as a close-relative of birds, and 2) its slower evolutionary rate compared with, for example, crocodilians and lizards (e.g., Chiari et al. 2012),

which maximizes the efficiency of similarity search. We recovered the amino acid sequences available for these genes in *P. sinensis*. When several transcripts were available for a gene, we kept the longest one.

Recovery of Homologous Sequences from Avian Data

We targeted two sources of data to search for homologous sequences: (i) *de novo* predicted cDNAs obtained from transcriptome data available from the SRA database and assembled using *SOAP-DeNovo-Trans* (Xie et al. 2014) with parameters *max_rd_len* and *rd_len_cutof* adjusted for each library, *avg_ins* = 200, *reverse_seq* = 0, *asm_flags* = 3, and *map_len* = 32; 2) the "PREDICTED" sequences available in GenBank, which are generated by the Gnomon pipeline of annotation using a maximum of the available data for a given taxon, or 3) both kinds of data sets. Over the 48 bird species included in the phylogenomic work by Jarvis et al (2014), three species—the Caribbean flamingo (*Phoenicopterus ruber*), the great crested grebe (*Podiceps cristatus*) and the turkey vulture (*Cathartes aura*)—were missing the two sources of data and they were thus discarded from the analysis. Additionally, as there were two representatives of the genus *Haliaeetus*, only *Haliaeetus leucocephalus* was used. Accession numbers and detailed information on the type of data used for each species are provided in supplementary table S1, Supplementary Material online.

TBLASTN analyses were performed between the newly gathered data sets and the amino-acid sequences of reference genes from *Pelodiscus*. The best hit was kept for each gene provided that *e* value was below 1e-10. The rationale of conducting a *TBLASTN* research, instead of BLASTN, is the higher conservation of amino acid sequences, especially in case of base composition biases or high evolutionary rates.

Check for Orthology

In order to verify the orthology among the obtained sequences, a reciprocal TBLASTN search was conducted, this time using the best hits of each species as database and *P. sinensis* protein sequences (ENSEMBL v86) as query. Only those genes for which the best hit corresponded to the original reference were kept for subsequent analyses. An additional filtering of paralogs, which are expected to present a higher proportion of misaligned sites, was indirectly done when cleaning amino acid alignments with HMMCleaner (Amemiya et al. 2013; Philippe et al. 2017; see below).

Sequences Alignment and Cleaning

Coding sequences were aligned using MACSE v1.01 b (Ranwez et al. 2011), warranting conservation of reading frames, with parameters *-gap_op = 3* and *-ext_gap_ratio = 0.75* in order to facilitate gap opening at the beginning and the end of the sequences. The resulting amino-acid alignments were treated with HMMcleaner to identify and mask highly misaligned sites. This software first builds a Hidden Markov Model profile of the alignment minus the target sequence, and then measures the score of the different sequence regions along this profile (Philippe et al. 2017). Every positions diverging more than the estimated score are removed (see Supplementary information for Amemiya et al. 2013). We used a very stringent value of 5 as threshold. Site masking was then propagated to nucleotide alignments. Sequences reduced by >50% in length were discarded. Sites (codons) for which <60% of the species were represented in the alignment were discarded. We selected the set of 1,245 genes in which at least ten species were present, equivalent to ~2, 0 Mb of data. Final alignments are available at https://doi.org/10.6084/m9.figshare.5202853.

GC Content

GC-content was estimated for the whole sequence (GC) and third codon positions only (GC3) for each assembled/recovered gene in order to compare them to mean genomic values of birds and those of representatives of other nonavian lineages (i.e., *Pelodiscus*, *Homo*).

Estimation of dN/dS and Kr/Kc Ratios

In order to assure comparability among both studies we used the same analytical pipeline ans species set as in Figuet et al. (2016). We used the phylogeny made available by Jarvis et al (2014) because 1) it is, to our knowledge, the best available up to present 2) it includes and describes the relationships for the species in our data set (supplementary fig. S4, Supplementary Material online). The ratio of nonsynonymous to synonymous substitutions (dN/dS) was estimated by first optimizing branch lengths under the YN98 model and then mapping each type of substitutions (nonsynonymous and synonymous) along the branches of the avian phylogeny. This method is significantly faster than for example estimations implemented in PAML while providing similar estimations (Romiguier et al. 2012). Counts for each type of substitution were then corrected for the opportunity of mutation using the transition-to-transversion ratio (kappa) as estimated from initial optimization (Figuet et al. 2016).

Life-HistoryTraits

Values for body mass and longevity were taken from Figuet et al. (2016).

Supplementary Material

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

Acknowledgments

We thank Laurent Duret and Sylvain Glémin for fruitful discussions and their constructive suggestions during different stages of the project. This work benefited from support by the Agence Nationale de la Recherche (ANR) under projects ANR Ancestrome (10-BINF-0001) and ANR DaSiRe (ANR-15-CE12-0010) and the computational resources of the Montpellier Bioinformatics Biodiversity platform (MBB). This publication is the contribution No 2017-193 of the Institut des Sciences de l'Evolution de Montpellier.

References

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, MacCallum I, Braasch I, Manousaki T, Schneider I, Rohner N, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* 496(7445): 311–316.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, De Massy B. 2010. *Prdm9* is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327(5967): 836–840.

Berglund J, Pollard KS, Webster MT, Hurst LD. 2009. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol*. 7(1): e1000026.

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.

Burt DW. 2002. Origin and evolution of avian microchromosomes. *Cytogenet Genome Res*. 96(1–4): 97–112.

Chiari Y, Cahais V, Galtier N, Delsuc F. 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biol*. 10: 65.

Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Ann Blomberg L, Bouffard P, Burt DW, Crasta O, Crooijmans RPMA. 2010. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biol*. 8: e1000475.

Damas J, O'Connor R, Farré M, Lenis VPE, Martell HJ, Mandawala A, Fowler K, Joseph S, Swain MT, Griffin DK, et al. 2017. Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Res*. 27(5): 875–884.

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol*. 10(12): e1003998.

Ellegren H. 2010. Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol Evol*. 25(5): 283–291.

Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, de Graaf DC, Debyser G, Deng J, Devreese B, et al. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics* 15: 86.

Figuet E, Nabholz B, Bonneau M, Mas Carrio E, Nadachowska-Brzyska K, Ellegren H, Galtier N. 2016. Life-history traits, protein evolution, and the nearly neutral theory in amniotes. *Mol Biol Evol*. 33: 1517–1527.

Figuet E, Romiguier J, Dutheil JY, Galtier N. 2014. Mitochondrial DNA as a tool for reconstructing past life-history traits in mammals. *J Evol Biol*. 27: 899–910.

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.

Hargreaves AD, Zhou L, Christensen J, Marlétaz F, Liu S, Li F, Jansen PG, Spiga E, Hansen MT, Pedersen SVH, et al. 2017. Genome sequence of a diabetes-prone rodent reveals a mutation hotspot around the *ParaHox* gene cluster. *Proc Natl Acad Sci U S A*. 114(29): 7677–7682.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432(7018): 695–716.

Hron T, Pajer P, Pačes J, Bartůněk P, Elleder D. 2015. Hidden genes in birds. *Genome Biol*. 16: 164.

Hua X, Bromham L. 2017. Darwinism for the genomic age: connecting mutation to diversification. *Front Genet*. 8: 12.

Hughes AL, Hughes MK. 1995. Small genomes for better flyers. *Nature* 377(6548): 391.

Hughes AL, Piontkivska H. 2005. DNA repeat arrays in chicken and human genomes and the adaptive evolution of avian genome size. *BMC Evol Biol*. 5: 12.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215): 1320–1331.

Korlach J, Gedman G, Kingan SB, Chin CS, Howard J, Cantin L, Jarvis ED. 2017. GigaScience, gix085, https://doi.org/10.1093//gigascience/gix085.

Lartillot N. 2013. Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. *Mol Biol Evol*. 30(2): 356–368.

Lartillot N, Poujol R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol Biol Evol*. 28(1): 729–744.

Lovell PV, Wirthlin M, Wilhelm L, Minx P, Lazar NH, Carbone L, Warren WC, Mello CV. 2014. Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biol*. 15(12): 565.

Nabholz B, Kunstner A, Wang R, Jarvis ED, Ellegren H. 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. *Mol Biol Evol*. 28(8): 2197–2210.

Nabholz B, Uwimana N, Lartillot N. 2013. 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 Biol Evol*. 5(7): 1273–1290.

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, Beatson SA, Lunter G, Malik HS, Ponting CP. 2009. Accelerated evolution of the *Prdm9* speciation gene across diverse metazoan taxa. *PLoS Genet*. 5(12): e1000753.

Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV. 2007. Origin of avian genome size and structure in non-avian dinosaurs. *Nature* 446(7132): 180–184.

Philippe H, Vienne DM. d, Ranwez V, Roure B, Baurain D, Delsuc F. 2017. Pitfalls in supermatrix phylogenomics. *Eur J Taxon*. 283: 1–25.

Popadin K, Polishchuk LV, Mamirova L, Knorre D, Gunbin K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc Natl Acad Sci U S A*. 104(33): 13390–13395.

Ranwez V, Harispe S, Delsuc F, Douzery EJP, Murphy WJ. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One* 6(9): e22594.

Romiguier J, Figuet E, Galtier N, Douzery EJP, Boussau B, Dutheil JY, Ranwez V, Liberles D. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. *PLoS One* 7(3): e33852.

Romiguier J, Ranwez V, Douzery EJP, 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.

Romiguier J, Ranwez V, Douzery EJP, Galtier N. 2013. Genomic evidence for large, long-lived ancestors to placental mammals. *Mol Biol Evol*. 30(1): 5–13.

Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, Strand AI, Li Q, Raney B, Balakrishnan CN, et al. 2015. Stable recombination hotspots in birds. *Science* 350(6263): 928–932.

Smith J, Bruley CK, Paton IR, Dunn I, Jones CT, Windsor D, Morrice DR, Law AS, Masabanda J, Sazanov A, et al. 2000. Differences in gene density on chicken macrochromosomes and microchromosomes. *Anim Genet*. 31(2): 96–103.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S, et al. 2010. The genome of a songbird. *Nature* 464(7289): 757–762.

Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F, et al. 2017. A new chicken genome assembly provides insight into avian genome structure. *G3* 7: 109–117.

Weber CC, Nabholz B, Romiguier J, Ellegren H. 2014. Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. *Genome Biol*. 15(12): 542–542.

Webster MT, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol*. 23(6): 1203–1216.

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, et al. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30(12): 1660–1666.

Yonezawa T, Segawa T, Mori H, Campos PF, Hongoh Y, Endo H, Akiyoshi A, Kohno N, Nishida S, Wu J, et al. 2017. Phylogenomics and morphology of extinct paleognaths reveal the origin and evolution of the ratites. *Curr Biol*. 27(1): 68–77.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346(6215): 1311–1320.