Analytical Biases Associated with GC-Content in Molecular Evolution

Jonathan Romiguier* and Camille Roux

Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

Molecular evolution is being revolutionized by high-throughput sequencing allowing an increased amount of genome-wide data available for multiple species. While base composition summarized by GC-content is one of the first metrics measured in genomes, its genomic distribution is a frequently neglected feature in downstream analyses based on DNA sequence comparisons. Here, we show how base composition heterogeneity among loci and taxa can bias common molecular evolution analyses such as phylogenetic tree reconstruction, detection of natural selection and estimation of codon usage. We then discuss the biological, technical and methodological causes of these GC-associated biases and suggest approaches to overcome them.

Keywords: GC-content, positive selection, biased gene conversion, codon usage bias, phylogeny, methodological biases

INTRODUCTION

GC-content is shaped by a complex balance among mutation, selection, recombination, and genetic drift (Bulmer, 1991; Eyre-Walker and Hurst, 2001; Duret et al., 2002). As a consequence of variation in this subtle balance, it has been observed that GC-content varies considerably at two levels: (i) among genomes from different species and (ii) along chromosomes of a single species (Bernardi et al., 1985). Among species, the average genomic GC-content ranges from 13 to 75% (Pagani et al., 2011). Within the same genome, large chromosomal regions can also greatly differ in their nucleotide composition as first described in humans (Bernardi et al., 1985). For instance, GC-content is distributed across the human genome over successive long stretches of >100 kb that can be either GC-rich (with a GC-content ~60%) or GC-poor (with a GC-content ~35%; International Human Genome Sequencing Consortium, 2001).

After several years of debate among neutral or selective hypotheses [reviewed in Duret and Galtier (2009)], it is now widely accepted that one of the major drivers of base composition heterogeneity is GC-biased gene conversion (gBGC), a repair bias that favors GC over AT alleles during meiotic recombination (Eyre-Walker, 1993; Galtier et al., 2001; Montoya-Burgos et al., 2003; Duret and Arndt, 2008; Kent et al., 2012; Arbeithuber et al., 2015; Mugal et al., 2015). As a result of this link between GC and recombination, local GC-content increases faster in genomic hotspots of recombination (Spencer, 2006) while genome-wide GC-content increases faster in species with higher recombination rates per time unit (Romiguier et al., 2010, 2013b; Figuet et al., 2014; Weber et al., 2014). By conferring a higher transmission probability of GC alleles over AT in heterozygotes, gBGC mimics natural selection but is frequently overlooked in molecular evolution studies. Here, we revisit how much intra-genomic and inter-specific variations in base composition have a strong power to bias popular analyses in molecular evolution such as phylogenetic tree reconstruction, detection of natural selection and estimation of codon usage bias (Figure 1).
**PHYLOGENETIC TREE RECONSTRUCTION**

Reconstructions of phylogenetic trees from molecular datasets are central in evolutionary biology. While initially limited to a handful of loci with limited power to resolve difficult phylogenetic relationships, phylogenetic tree reconstruction is no longer restricted by the number of genetic markers. However, some phylogenetic relationships of the Tree of Life remain unresolved (Philippe et al., 2011). This difficulty stems from the mosaic nature of genomes gathering alternative and conflicting gene trees (Degnan and Rosenberg, 2009), where some but not all loci support the true species genealogy. Determining which loci are reliable phylogenetic markers is thus one of the biggest challenges in phylogenomics. While mixed historical signals along genomes are likely to have different natures, we will here focus to issues related to base composition.

A recent phylogenomic study reported that base composition is a relevant criterion to select markers carrying unambiguous phylogenetic signal: gene GC-content (average GC% of the sequences of an alignment) and GC-heterogeneity (variance of GC% among sequences of an alignment) were proved to bias species tree reconstructions (Romiguier et al., 2013a).

As illustrated with mammalian genomes, phylogenetic trees of genes located in GC-rich regions produce five times more contradicting topologies than GC-poor genes, leading to important reconstruction biases and a poor resolution for both accepted and controversial nodes. This negative analytical effects of GC-content on tree reconstructions is widespread across the tree of life, as reported in basal eukaryote lineages (Rodríguez-Ezepeleta et al., 2007), yeasts (Collins et al., 2005), beetles (Sheffield et al., 2009), bees (Romiguier et al., 2016), hexapods (Delsuc, 2003), fishes (Li and Orti, 2007; Betancur-R et al., 2013), birds (Nabholz et al., 2011), and bats (Teeling et al., 2000). Despite the accumulating empirical evidence demonstrating the pervasiveness of base composition issues in phylogeny, the reasons underlying such strong biases are unexplored. Here, we suggest three non-mutually exclusive hypotheses to explain this negative GC-effect in phylogenomics studies.

First, some aspects of the GC-bias are likely to be due to model misspecifications. Probabilistic methods for phylogenetic reconstruction (maximum likelihood or Bayesian inference) are indeed generally based on models of sequence evolution that assume a homogeneous base composition along the tree. However, this assumption is often violated (Phillips et al., 2004). Indeed, average GC-content of an alignment correlates strongly with GC-heterogeneity among sequences as a result of variation in the dynamic of gBGC among sampled species (Romiguier et al., 2013a). Such departures from the assumption of base composition homogeneity can lead to severe biases by incorrectly grouping distantly related taxa that converge in extreme nucleotide composition on a given locus (Phillips et al., 2004). This type of issues can be, however, easily solved by model-based solutions (see last paragraph of this section for more details).

The second hypothesis proposes that incomplete lineage sorting (ILS) is more important in GC-rich than GC-poor regions. ILS is known to produce conflicts among gene trees and the species tree because of the retention of ancestral polymorphisms (Degnan and Rosenberg, 2009). At the scale of the whole genome, the amount of incompletely sorted genes increases when the time of divergence is small relative to the average effective population sizes (Clark, 1997). Genomic variation in ILS was also empirically reported to be associated to GC-content in hominid genomes (Hobolth et al., 2011). This indirect (Charlesworth et al., 1993) relationship...
between GC-content and ILS can be explained by the dual effects of local recombination rates on base composition and linkage disequilibrium. High local recombination rates increase GC-content through gBGC, but also decrease the effect of genetic interferences, i.e., background selection (Charlesworth et al., 1993) and hitchhiking (Smith and Haigh, 1974). By being less affected by linked selective processes, GC-rich regions are thus expected to have relatively high effective population sizes than GC-poor regions, leading to an extended retention time of ancestral polymorphism.

The third hypothesis is that gBGC is associated to saturation in multiple substitutions. Following the rapid birth and death of local recombination hotspots, gBGC is expected to occur in short, intense episodes (Duret and Galtier, 2009) where deleterious GC substitutions are likely to occur (Necula et al., 2011). Following a gBGC episode, natural selection is likely to revert such deleterious substitutions through AT replacement (Galtier et al., 2009). This toggling between GC deleterious and AT compensatory substitutions at the same nucleotide site is expected to lead to homoplasy, a direct consequence of multiple substitutions causing spurious similarity not due to common ancestry (Philippe et al., 2011). This type of AT/GC toggling is expected to be particularly fast and difficult to track because of the short-life of gBGC episodes that depends on the self-destructive nature of recombination hotspots (Coop and Myers, 2007). Even at very short evolutionary scales such as the Denisovan/Modern human divergence (0.4–0.8 Myrs), local recombination hotspots are not conserved (Leseque et al., 2014), which could imply a complete loss of phylogenetic signals due to multiple turnovers between gBGC and natural selection at larger evolutionary scales. Although genomically small (1–2 kb), these short-lived recombination hotspots tend to arise and disappear in the same genomic regions of 1–2 Mb (Duret and Galtier, 2009) exhibit homoplasy issues. Common in fast-evolving sequences, homoplasy is also at the origin of the so-called and undesired “long branch attraction artifact” (Felsenstein, 1978). Reinforcing the idea that GC-rich genes might be affected by such biases, GC-rich and GC-heterogeneous genes have fast rates of evolution (Romiguier et al., 2013a, 2016). These abnormally fast-evolving genes are then likely to cause long-branch attraction artifacts, but also more general issues related to heterotachy-driven biases (Philippe et al., 2005). Even if long-branch attraction is generally considered as a minor problem in likelihood-based phylogenetics compared to parsimony, maximum likelihood methods using GC-rich genes can have a biased support toward topologies grouping long branches together (Romiguier et al., 2013a, 2016).

One solution to cope with base composition issues is the use of models of sequence evolution that takes into account heterogeneity in GC-content (Galtier and Gouy, 1998; Foster, 2004; Blanquart and Lartillot, 2006; Boussau and Gouy, 2006; Gowri-Shankar and Rajatray, 2007; Dutheil and Boussau, 2008). However, these so-called non-homogeneous models are computationally costly. Albeit useful to alleviate GC-heterogeneity issues in phylogeny, empirical studies illustrate their limits to retrieve high bootstrap supports in the most GC-heterogeneous sequences (Betancur-R et al., 2013; Romiguier et al., 2016), shedding light on other GC-dependent biases in phylogeny such as ILS and gBGC-driven homoplasy. To date, the best practice recommended to discard noisy signals in sequences is the use of non-homogeneous models and/or the use of GC-poor phylogenetic markers. In this regard, it is noteworthy that coding sequences tend to be clustered in recombination hotspots and GC-rich regions (Duret and Galtier, 2009). Consequently, the use of the rare phylogenetic markers located in AT-rich regions is recommended. This is the case of ultra-conserved non-coding elements (UCE) that have the advantage to be AT-rich and evolve particularly slowly (McCormack et al., 2012). Compared to these non-coding AT-rich markers, clusters of AT-rich coding genes in low-recombining regions could undergo a higher rate of background selection, decreasing the effective population size and then, the amount of ILS. It is noteworthy that UCE and AT-rich genes both support the same topology for the controversial rooting of placental mammals (McCormack et al., 2012; Romiguier et al., 2013a), highlighting relevance of these markers to overcome GC-biases. Other strategies might involve to compare these markers with markers that cannot be affected by recombination and gBGC, such as mitochondrial genes. Further methodological improvements could come from coalescent-based supertree methods (Liu et al., 2009) that account for ILS. By weighting the confidence in each gene tree according to the GC-content of an alignment, they may allow the integration of most of the available information and alleviate the spurious signal inherent to GC-rich markers. To date, methods computing the exact likelihood of alternative topologies are restricted to relatively simple models neglecting direct and indirect effects of background selection, selective sweep, gBGC and ILS on phylogenetic reconstruction. But these processes are now implemented in recent simulators (Haller and Messer, 2017), allowing them to be treated as nuisance parameters during computational evolution of sequences. Although such highly complex models are currently intractable by maximum likelihood approaches, the possibility to simulate them within an approximate Bayesian computation (ABC) framework (Beaumont et al., 2002; Csillery et al., 2010a,b; Pudlo et al., 2016) could bring new methodological perspectives in phylogenetic reconstruction. ABC has been proved to be a powerful framework to compare complex evolutionary scenarios in phylogenetic reconstruction. It has been used to explore the recent improvements made in flexible machine learning algorithms. Applied to phylogenetic reconstructions, efficient computational tools like SLiM 2 are already available to simulate models with gBGC episodes and multiple substitutions along a branch as well as statistical packages to compute the probabilities of alternative scenarios (Csillery et al., 2012; Pudlo et al., 2016). Altogether, current available softwares already provide stimulating leads for future developments in phylogeny.

DETECTION OF POSITIVE SELECTION

Identifying candidate loci for natural selection is a central goal explored by two traditional approaches in adaptation-genomics: top-down (GWA and QTL) and bottom-up (genomic scan) approaches. With the advent of high-throughput sequencing,
genomic scans became a popular approach to detect candidate target of selection. Such scans have the merit to identify candidates without the a priori expectation of a candidate gene approach (Ellegren, 2014). However, they have various limitations with false-positive issues (Mallick et al., 2009; Bierne et al., 2011), narrow signatures of balancing selection (Roux et al., 2012), and over-interpretation of outlier loci (Pavlidis et al., 2012). Here, we detail how GC-content can lead to important additional bias during genome scans for detecting natural selection.

Genome scans of positive selection often rely on methods that look for lineage-specific accelerations in the protein rate of evolution. Such accelerations are classically measured through \( dN/dS \), which calculates the excess of amino-acid substitutions (\( dN \): non-synonymous mutation rate per site) relative to \( dS \), the substitution rate per site used as a proxy of the neutral clock. This \( dN/dS \) ratio is generally smaller than 1, reflecting the pervasiveness of purifying selection that eliminates non-synonymous mutations to preserve the protein structure. Conversely, a \( dN/dS \) ratio greater than 1 is considered as a signature of positive selection that favors the fixation of beneficial non-synonymous mutations. From a population genetics point of view, gBGC mimics positive selection by favoring the fixation of AT-\( > \)GC mutations, regardless of their beneficial or deleterious status (Nagylaki, 1983). Because GC alleles are actively selected by the repair systems of meiotic recombination, they are over-represented in the gamete pool and benefit of increased transmission to the next generation in a similar way than beneficial mutations subject to positive selection. Consequently, many accelerations of the substitution rate attributed to positive selection during genome scans are actually due to gBGC episodes (Galtier and Duret, 2007; Berglund et al., 2009; Galtier et al., 2009; Ratnakumar et al., 2010; Kostka et al., 2012). When a mutation toward GC is deleterious, gBGC can counteract positive selection and maintain or fix deleterious alleles. High fixation rates of non-synonymous mutations at a locus should thus be systematically interpreted as being beneficial for the fitness of the individual, particularly when considering that gBGC has been proved to be able to maintain deleterious mutations associated to human diseases (Neczulea et al., 2011; Capra et al., 2013; Lachance and Tishkoff, 2014).

Confusion between positive selection and gBGC could be avoided through two different ways. The first is by filtering the results of classical tests of positive selection and consider with caution positive selection signatures in GC-rich regions. This is particularly true for selection tests that rely more on overall evolutionary rate rather than \( dN/dS \) (Pollard et al., 2006; Kostka et al., 2012). Even if gBGC can increase \( dN/dS \) in some conditions (Galtier et al., 2009; Bolivar et al., 2015), AT-\( > \)GC mutations are more likely to happen in synonymous sites, which limits the effect of gBGC on \( dN/dS \) compared to the evolutionary rate. Several criterions can be used in both cases to differentiate gBGC from positive selection, such as the number of mutations toward GC in the surrounding non-coding regions (Galtier and Duret, 2007). The second would be to develop methods that restrict \( dN/dS \) estimations to GC-conservative substitutions in the context of codon-models aimed to detect positive selection events (Yang and Nielsen, 2002; Lartillot, 2013).

**CODON USAGE BIAS**

Popular analytical methods in molecular evolution rely on a strong assumption: synonymous mutations are neutral. GC-content at synonymous positions is frequently claimed to be exposed only to the mutation/drift equilibrium. However, natural selection was proposed to be superimposed to these two evolutionary forces at synonymous codons (Urrutia, 2003; Comeron, 2004; Plotkin et al., 2004). Although initially challenged (Williamson et al., 2005), natural selection acting on standing synonymous variation was found to be associated to gene expression level, the most expressed genes using a set of preferred codons (Comeron, 2004). This association is explained by selection for increased translational efficiency. The analysis of >1,000 genes in Drosophila demonstrated that the most used synonymous codons corresponded to the most available tRNAs in the genome (Moriyama and Powell, 1997). Translational efficiency would then be optimized by increasing the usage of the preferred synonymous codons. Such a process can be tested in coding sequences by measuring the effective number of codons (ENc) in a given gene. ENc takes a value of 61 when all codons of the genetic code (minus the three stop codons) are used without bias, and decreases to 20 (the number of amino-acids) for the most biased genes. In agreement with the hypothesis of selection for translational efficiency, population genetics analyses in Drosophila described signatures of selection on synonymous mutations (Akashi, 1995; Akashi and Schaeffer, 1997).

A study of codon usage bias in Caenorhabditis elegans, Drosophila melanogaster, and Arabidopsis thaliana has shed light on the over-expression of genes featuring codon preference, with a large predominance of preferred codons ending with G or C (Duret and Mouchiroud, 1999). However, the GC-content at third coding positions (GC3) is also correlated to the GC-content of the surrounding non-coding regions (Kliman and Hey, 1994; Akashi et al., 1998), which suggests the action of gBGC shaping local base compositions. By locally increasing GC-content, gBGC mechanically restricts the number of used codons and reduces the measured ENc independently of selection for translational efficiency. The measured ENc is thus biased by gBGC and must be corrected with local background nucleotide compositions. In addition, variation in GC-content also impacts measures of gene expression. With the advent of high-throughput sequencing technologies, it is now a standard practice to approximate gene expression levels by counting the number of reads mapping a target in ChIP-seq or RNA-seq analysis. However, sequencing biases artificially over-represent genomic regions with intermediate levels of GC-content (50%), which in turn bias the estimates of gene expression levels (Chouvarine et al., 2016). Testing selection for translational efficiency by
measuring the correlation between ENc and gene expression levels therefore requires the use of both GC-corrected ENc and GC-corrected expression levels. ENc estimates can be corrected by GC-content of neighborhood regions (Novembre, 2002), while GC-corrected expression levels can be obtained by applying local LOESS regression (Miller et al., 2011; Benjamini and Speed, 2012; Chandrananda et al., 2014) or quantile normalization-methods (Risso et al., 2011), i.e., by normalizing the raw number of mapped reads by the local GC-content.

The ongoing surge of transcriptomic data will permit measurement of GC-content heterogeneity, preferred codons usage and expression levels across a large number of loci and species. This type of large-scale analysis could open the door to a better understanding of the relationship linking effective population sizes (Ne) and codon usage. As theoretically predicted (Bulmer, 1991), selection on synonymous codons might be stronger in species with large Ne. While the Ne-hypothesis to explain variation in selection on codon usage remains untested by empirical studies, a descriptive study of the Ne-effect on variation in gBGc will be necessary to avoid entangling the two effects. Future projects aiming to test these hypotheses are expected to be strongly biased if GC-content biases are naively neglected regarding estimates of gene expression levels or codon usage.

CONCLUSION

GC-content is associated to multiple biases of different nature (Figure 1). Whether through technological reasons (sequencing technologies biases), biological reasons (GC-biased gene conversion) or methodological reasons (models of sequence evolution limitations), all these biases affect the results of downstream analyses. With the surge of genomic data from various non-model species, comparative genomics have the opportunity to solve many unresolved questions in evolution. However, one should be aware of the methodological challenges associated to the GC-content heterogeneity inherent to large scale studies, whether it be for a large number of species or loci.

AUTHOR CONTRIBUTIONS

JR had the idea of the project. JR and CR wrote the article.

FUNDING

This work was supported by a Federation of European Biochemical Societies (FEBS) long-term fellowship to JR.
Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H., and Mailund, T. (2011). Recombination and mammalian genome evolution. Proc. R. Soc. B Biol. Sci. 278, 1837–1847.

Dutheil, J., and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. U.S.A. 96, 4482–4487. doi: 10.1073/pnas.96.8.4482

Lesecque, Y., Glémín, S., Lartillot, N., Mouchiroud, D., and Durét, L. (2014). The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. PLoS Genet. 10:e1004790. doi: 10.1371/journal.pgen.1004790

Philippe, H., Brinkmann, H., Lavriv, D. V., Littlewood, D. T. J., Manel, M., Worheide, G., et al. (2011). Resolving difficult phylogenetic questions: why
more sequences are not enough. *PLoS Biol.* 9:e1000602. doi: 10.1371/journal.phbio.1000602

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.* 5:50. doi: 10.1186/1471-2148-5-50

Phillips, M. J., Delsuc, F., and Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* 21, 1455–1458. doi: 10.1093/molbev/msi137

Plotkin, J. B., Robins, H., and Levine, A. J. (2004). Tissue-specific codon usage and the expression of human genes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12588–12591. doi: 10.1073/pnas.0404957101

Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 445, 167–172. doi: 10.1038/nature05113

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics* 32, 859–866. doi: 10.1093/bioinformatics/btv684

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., et al. (2010). Detecting positive selection within genomes: the problem of biased gene conversion. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 2571–2580. doi: 10.1098/rstb.2010.0007

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data. *BMC Bioinformatics* 12:480. doi: 10.1186/1471-2105-12-480

Rodriguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B. F., and Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst. Biol.* 56, 389–399. doi: 10.1080/10635150701397643

Romiguier, J., Cameron, S. A., Woodard, S. H., Fischman, B. J., Keller, L., and Praz, C. J. (2016). Phylogenomics controlling for base compositional bias reveals a single origin of eusociality in corbiculate bees. *Mol. Biol. Evol.* 33, 670–678. doi: 10.1093/molbev/msv258

Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., and Douzery, E. J. P. (2013a). Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol. Biol. Evol.* 30, 2134–2144. doi: 10.1093/molbev/mss116

Romiguier, J., Ranwez, V., Douzery, E. J. P., and Galtier, N. (2010). Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res.* 20, 1001–1009. doi: 10.1101/gr.104372.109

Romiguier, J., Ranwez, V., Douzery, E. J. P., and Galtier, N. (2013b). Genomic evidence for large, long-lived ancestors to placental mammals. *Mol. Biol. Evol.* 30, 5–13. doi: 10.1093/molbev/mss211

Roux, C., Fraise, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biol.* 14:e2000234. doi: 10.1371/journal.pbio.2000234

Roux, C., Puwels, M., Ruggiero, M.-V., Charlesworth, D., Castric, V., and Vekemans, X. (2012). Recent and ancient Signature of balancing selection around the S-Locus in *Arabidopsis thaliana* and *A. lyrata*. *Mol. Biol. Evol.* 30, 435–447. doi: 10.1093/molbev/mss246

Sheffield, N. C., Song, H., Cameron, S. L., and Whiting, M. F. (2009). Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics. *Syst. Biol.* 58, 381–394. doi: 10.1093/sysbio/syp037

Smith, J. M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35. doi: 10.1017/S0016672300014634

Spencer, C. A. A. (2006). Human polymorphism around recombination hotspots. *Biochem. Soc. Trans.* 34, 535–536. doi: 10.1042/BST0340535

Teeling, E. C., Scally, M., Kao, D. J., Romagnoli, M. L., Springer, M. S., and Stanhope, M. J. (2000). Molecular evidence regarding the origin of echolocation and flight in bats. *Nature* 403, 188–192. doi: 10.1038/35003188

Urrutia, A. O. (2003). The signature of selection mediated by expression on human genes. *Genome Res.* 13, 2260–2264. doi: 10.1101/gr.41103

Weber, C. C., Boussau, B., Romiguier, J., Jarvis, E. D., and Ellegren, H. (2014). Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol.* 15:549. doi: 10.1186/s13059-014-0549-1

Williamson, S. H., Hernandez, R., Fleckel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C. D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7882–7887. doi: 10.1073/pnas.0502300102

Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19, 908–917. doi: 10.1093/oxfordjournals.molbev.a004148

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2017 Romiguier and Roux. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*