Peer Community In Genomics

A systematic approach to the study of GC-biased gene conversion in mammals

Carina Farah Mugal *based on peer reviews by* ***David Castellano***, ***Fanny Pouyet*** *and 1 anonymous reviewer*

Nicolas Galtier (2021) Fine-scale quantification of GC-biased gene conversion intensity in mammals. Missing preprint_server, ver. Missing article_version, peer-reviewed and recommended by Peer Community in Genomics.
https://doi.org/10.1101/2021.05.05.442789

Submitted: 25 May 2021, Recommended: 07 October 2021

Cite this recommendation as:
Mugal, C. (2021) A systematic approach to the study of GC-biased gene conversion in mammals. *Peer Community in Genomics*, 100012. https://doi.org/10.24072/pci.genomics.100012

Published: 07 October 2021

The role of GC-biased gene conversion (gBGC) in molecular evolution has interested scientists for the last two decades since its discovery in 1999 (Eyre-Walker 1999; Galtier et al. 2001). gBGC is a process that is associated with meiotic recombination, and is characterized by a transmission distortion in favor of G and C over A and T alleles at GC/AT heterozygous sites that occur in the vicinity of recombination-inducing double-strand breaks (Duret and Galtier 2009; Mugal et al. 2015). This transmission distortion results in a fixation bias of G and C alleles, equivalent to directional selection for G and C (Nagylaki 1983). The fixation bias subsequently leads to a correlation between recombination rate and GC content across the genome, which has served as indirect evidence for the prevalence of gBGC in many organisms. The fixation bias also produces shifts in the allele frequency spectrum (AFS) towards higher frequencies of G and C alleles.

These molecular signatures of gBGC provide a means to quantify the strength of gBGC and study its variation among species and across the genome. Following this idea, first Lartillot (2013) and Capra et al. (2013) developed phylogenetic methodology to quantify gBGC based on substitutions, and De Maio et al. (2013) combined information on polymorphism into a phylogenetic setting. Complementary to the phylogenetic methods, later Glemin et al. (2015) developed a method that draws information solely from polymorphism data and the shape of the AFS. Application of these methods to primates (Capra et al. 2013; De Maio et al. 2013; Glemin et al. 2015) and mammals (Lartillot 2013) supported the notion that variation in the strength of gBGC across the genome reflects the dynamics of the recombination landscape, while variation among species correlates with proxies of the effective population size. However, application of the polymorphism-based method by Glemin et al. (2015) to distantly related Metazoa did not confirm the correlation with effective population size (Galtier et al. 2018).

Here, Galtier (2021) introduces a novel phylogenetic approach applicable to the study of closely related species. Specifically, Galtier introduces a statistical framework that enables the systematic study of variation in the strength of gBGC among species and among genes. In addition, Galtier assesses fine-scale variation of gBGC across the genome by means of spatial autocorrelation analysis. This puts Galtier in a position to study variation in the strength of gBGC at three different scales, i) among species, ii) among genes, and iii) within genes. Galtier applies his method to four families of mammals, Hominidae, Cercopithecidae, Bovidae, and Muridae and provides a thorough discussion of his findings and methodology.

Galtier found that the strength of gBGC correlates with proxies of the effective population size (Ne), but that the slope of the relationship differs among the four families of mammals. Given the relationship between the population-scaled strength of gBGC B = 4Neb, this finding suggests that the conversion bias (b) could vary among mammalian species. Variation in b could either result from differences in the strength of the transmission distortion (Galtier et al. 2018) or evolutionary changes in the rate of recombination (Boman et al. 2021). Alternatively, Galtier suggests that also systematic variation in proxies of Ne could lead to similar observations. Finally, the present study reports intriguing inter-species differences between the extent of variation in the strength of gBGC among and within genes, which are interpreted in consideration of the recombination dynamics in mammals.

References:

Boman J, Mugal CF, Backström N (2021) The Effects of GC-Biased Gene Conversion on Patterns of Genetic Diversity among and across Butterfly Genomes. Genome Biology and Evolution, 13. https://doi.org/10.1093/gbe/evab064

Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A (2013) A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. PLOS Genetics, 9, e1003684. https://doi.org/10.1371/journal.pgen.1003684

De Maio N, Schlötterer C, Kosiol C (2013) Linking Great Apes Genome Evolution across Time Scales Using Polymorphism-Aware Phylogenetic Models. Molecular Biology and Evolution, 30, 2249–2262. https://doi.org/10.1093/molbev/mst131

Duret L, Galtier N (2009) Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annual Review of Genomics and Human Genetics, 10, 285–311. https://doi.org/10.1146/annurev-genom-082908-150001

Eyre-Walker A (1999) Evidence of Selection on Silent Site Base Composition in Mammals: Potential Implications for the Evolution of Isochores and Junk DNA. Genetics, 152, 675–683. https://doi.org/10.1093/genetics/152.2.675

Galtier N (2021) Fine-scale quantification of GC-biased gene conversion intensity in mammals. bioRxiv, 2021.05.05.442789, ver. 5 peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2021.05.05.442789

Galtier N, Piganeau G, Mouchiroud D, Duret L (2001) GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics, 159, 907–911. https://doi.org/10.1093/genetics/159.2.907

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L (2018) Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. Molecular Biology and Evolution, 35, 1092–1103. https://doi.org/10.1093/molbev/msy015

Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L (2015) Quantification of GC-biased gene conversion in the human genome. Genome Research, 25, 1215–1228. https://doi.org/10.1101/gr.185488.114

Lartillot N (2013) Phylogenetic Patterns of GC-Biased Gene Conversion in Placental Mammals and the Evolutionary Dynamics of Recombination Landscapes. Molecular Biology and Evolution, 30, 489–502. https://doi.org/10.1093/molbev/mss239

Mugal CF, Weber CC, Ellegren H (2015) GC-biased gene conversion links the recombination landscape and demography to genomic base composition. BioEssays, 37, 1317–1326. https://doi.org/10.1002/bies.201500058

Nagylaki T (1983) Evolution of a finite population under gene conversion. Proceedings of the National Academy of Sciences, 80, 6278–6281. https://doi.org/10.1073/pnas.80.20.6278

Reviews

Evaluation round #2

DOI or URL of the preprint: https://www.biorxiv.org/content/10.1101/2021.05.05.442789v4
Version of the preprint: 4

Authors' reply, 20 September 2021

Thanks for a second round of reviewing and the positive outcome. There were three more comments:

- Line 88-90: "A total of 1,104,917 third codon position synonymous substitutions and 514,552 first or second codon position non-synonymous substitutions were called." Why there are more third codon positions than first+second codon positions?

-> Here we are talking about substitutions not positions. There indeed are more synonymous than non-synonymous substitutions in this (and most) data set(s). I did not modify the ms based on this suggestion.

- Line 88-90: "A total of 1,104,917 third codon position synonymous substitutions and 514,552 first or second codon position non-synonymous substitutions were called." Why there are more third codon positions than first+second codon positions?

-> Yes, trying to fit a formal model to the clustering pattern sounds like an interesting perspective, in part (although not very explicitely) covered by the discussion. Thanks for this. I did not modify the ms based on this suggestion.

- Line 283-285: "That said, not only Ne influences the variation of π: the mutation rate also matters. Among-species differences in per generation mutation rate, if any, should be taken into account for a better assessment of the b vs. Ne relationship". Maybe the author wants to back up this argument using and citing some of the results reported here: https://doi.org/10.1093/gbe/evab150

-> Good suggestion; this interesting paper is now cited.

Best regards,
Nicolas Galtier

Decision by Carina Farah Mugal, posted 15 September 2021

minor revisons

Dear Nicolas Galtier,

I am pleased to inform you that all three reviewers and I found that your revisions address all the earlier concerns and that your revised manuscript is in principle ready for recommendation. Only one reviewer has some very minor suggestions, which I think will be straight forward to address.

Best wishes,

Carina Farah Mugal

Reviewed by Fanny Pouyet, 08 September 2021

I have reviewed the paper entitled ”Fine-scale quantification of GC-biased gene conversion intensity in mammals.” by Nicolas Galtier.

The revised version of the manuscript encompasses all of the remarks I had and even more. I especially enjoyed the rewriting of the methods and the explanation of the weighted AIC which help the reader to fully understand the study. The discussion has been extended regarding very interesting suggestions from other reviewers.

I have no further comments.
Fanny Pouyet
It is my standard policy to sign my reviews (see round 1 for motivation)

Reviewed by anonymous reviewer 1, 24 August 2021

The authors did a good job at improving the manuscript. My comments have been addressed and I think the manuscript can be accepted in the current version.
Specifically, the authors have:
- Clarified the values for certain parameters used and added further description on the simulation methods sections.
- Added new analyses for life-history traits.
- Improved the discussion.
- Addressed editing suggestions in figures and text.
- Added a supplementary table which summarised the results and model parametrisation (as suggested by another reviewer).

Reviewed by David Castellano, 01 September 2021

In this second version of the manuscript, Galtier has addressed all my questions and comments. The manuscript was already good but now it clarifies some parts that were a bit obscure in the original version. I like the discussion, it proposes multiple future lines of investigation. I just have three minor comments:

> Line 88-90: ”A total of 1,104,917 third codon position synonymous substitutions and 514,552 first or second codon position non-synonymous substitutions were called.” Why there are more third codon positions than first+second codon positions?

> I find that ABC might be a more formal (but also computationally intensive) way of assessing the amount of clustering needed to replicate Moran's I statistic. However, given the simplicity of the simulations I believe the current approach might be already yielding accurate estimates.

> Line 283-285: "That said, not only Ne influences the variation of π: the mutation rate also matters. Among-species differences in per generation mutation rate, if any, should be taken into account for a better assessment of the b vs. Ne relationship". Maybe the author wants to back up this argument using and citing some of the results reported here: https://doi.org/10.1093/gbe/evab150

Evaluation round #1

DOI or URL of the preprint: https://www.biorxiv.org/content/10.1101/2021.05.05.442789v3
Version of the preprint: 3

Authors' reply, 23 July 2021

Download author's reply

Decision by Carina Farah Mugal, posted 11 July 2021

moderate revision

Dear Nicolas Galtier,

Three reviewers have now read your manuscript "Fine-scale quantification of GC-biased gene conversion intensity in mammals". All three reviewers and I find the manuscript well-written, and results and method-contribution relevant and interesting. Nevertheless, all three reviewers also provide valuable suggestions for improvement, which I think will help to improve the quality and clarity of the study. Briefly,

(1) All three reviewers ask for more clarity of Method and Figure description. You will find specific comments and suggestions in the respective review reports.

(2) Reviewer #3 suggests an alternative hypothesis to explain the weak clustering of WS substitutions within Hominadae genes, which I find very interesting. This reviewer also suggests an hypothesis to explain observations of the RSD analysis, which I think could be worth exploring.

(3) Personally, I am wondering how the contribution of ancestral and lineage-specific polymorphisms could bias the estimation of substitution rates in closely related species (see e.g. doi: 10.1093/molbev/msz203)?

Besides, the author might find this reference on DSB evolution in mice useful in relation to their own observations in mice (doi: 10.1093/molbev/mst154).

I look forward to receiving a revised version of the manuscript.
Best regards,
Carina Farah Mugal

Reviewed by Fanny Pouyet, 14 June 2021

In the manuscript entitled "Fine-scale quantification of GC-biased gene conversion intensity in mammals." by Nicolas Galtier (DOI https://www.biorxiv.org/content/10.1101/2021.05.05.442789v3), the author investigates how to measure gBGC strength in 4 clades of mammals (Hominidae, Cercopithecidae, Bovidae and Muridae).

The study is well designed as it compares observed statistics such as Moran's I or substitution density to simulations with nested/increasing number of parameters. The study uses a maximum-likelihood approach to reject the simplest models as it should be done. I don't have comments concerning the experimental design. The analysis and interpretation of results are clear as well. The flaws of this manuscript rely on the text or figure's legend which do not always help the reader to properly understand what was done and why. The presence of the scripts to make figures helped me doing the review.

Here are the specific comments:

1. Fig1: Please explicit in the legend what are the green branches (I assumed it is the studied branches but I'm unsure).
2. Fig2: I missed out why you need branches with at least 100 genes with each of them having at least 3 substitutions of each type. I would have expected this criterion will emphesize any signal of clustering. If you are willing to test whether there is a cluster, shouldn't you keep branches where there is a minimal number of substitution to have a signel (here, 300) regardeless they are well dispersed accross genes?
3. In several figures you wrote there is 1 dot per species. I thought you were also looking at internal branches. Do you calculate Moran's I solely at tips or in all the branches suitable for the study? Specifically, sup fig 1: you said 1 dot = 1 species but there are 8 dots in Bovidae and only 7 species. Moreover, I see only 7 blue dots and 8 red ones. Is it possible 2 blues dots are overlapping on the right? You could add a darker perimeter for the circles such that we see if 2 overlaps or not.
4. Supp Fig 1: Why is there differences between line blue and red in the simulations? Given the equations, I thought we would estimate the exact same value of B for both SW and WS substitutions.
5. Table S1: Please explicit what is the number of genes_cleaned. Given the explanation in the mat and methods (X->Y substitutions : All the descendants must be X on one side of the tree and Y on the rest), I don't get why we don't have the same number of genes within a clade.
6. In general, many sentences are long which does not help the reader. Please rewrite at least the sentence line 349-352. It is too long and the idea is complex (I had to re-read it 6 times before getting to the point).
7. Do you have an idea why there is an outlier in Bovidae at 3.86. If I get it right it is the capra-ovis ovarie branch (see supp table). It has a huge sd (~6). Do you think it is because there is no info, few substitution and so it is difficult to estimate a signal or do you think it is a biological signal, like a huge hotspot of gBGC intensity along that branch? I would have been interested to read about it in the discussion part (like if it's biological, is it realted to anything know about their evolutionary history?).
8. Models are complex and while they are well explained in the methods, I think you could help the reader by making a table summarizing the models. I had to dig to understand why f or z in models' names. In the same topic, could you make a table summarizing how many branches are rejected per model ?
9. Which branch is not rejected using M3sh compared to M3h ? Is it in humans where there is small gBGC or is it somewhere else ? Do you have any comments to make on this branch ?
10. What do you mean by averaging the Akaike Information Criterium of each model in practice? I couldn't find it in the scripts and I am interested in understanding what you did there.

10b. In equation 16 and 17. What is « k »: the genes or the AIC values ?

11. Fig4 : I think you represent the correlation of B and dN/dS using a log transformed scale (but the y axis and x axis values are still the values of B and dN/dS). The title is misleading.
12. Fig5 : The second sentence in the legend is unclear : « in for ».

I sign my review to increase the transparency of that process.

Reviewed by anonymous reviewer 1, 23 June 2021

In this work, N. Galtier estimated the strength of gBGC and investigated its relationship with Ne acrosss 4 different families of mammals. To do this, he analysed nucleotide substitution patterns in coding sequences of

40 mammalian lineages using a maximum likelihood approach. The results of the study suggest that gBGC is prevalent in these mammalian families, estimating that large proportion of WS synonymous substitution can be attributed to this process and that its strenght varies across lineages and genes depending on Ne and the dynamics of recombination hotspots.

This work joins a large body of literature that demonstrates that gBGC is a major force shaping patterns of molecular evolution. The article is well writen and I enjoyed reading it. The potencial limitation that came to mind while I reviewd the study where properly discussed, such as the fact that these results are dependent on assuming a constant mutational process across species. So, I do not have major comments for improving this manuscript.

In terms of novelty of the work, other studies have tried to estimate B accross mammalian lineages. However, most studies have estimated B from site frequency spectra. Few studies, which are cited in this manuscript, have already tried to estiamate B using substitution patterns. Specifically, Lartillot (2012) proposed an integrated Bayesian model for reconstructing the evolutionary history of gBGC, and for estimating its correlation with life-history and karyotypic traits. Nonetheless, this maximum-likelohood framework is an anlternative model that confirmed many previous studies and seems very valuable for the research field and community.

Minor comments:

line 102: I suggest editing: ”As far as SW substitutions were concerned, ” to ”The centered Moran’s I for WS substitutions ”

line 110: Why is there a discrepancy between the the bp scales used by the author when calculating Moran’s I (400 bp) and the one used in the simulation (40 and 500 bp?) If the aim is to assess the amount of clustering needed to explain the observed values, real and simulatated data should have the same bp scale.

Figure 4: The sample size was here too small to investigate the within-family relationships. To further investigate the relationship of B and Ne, the author could show if there is a correlation betnween B and Ne-related life history traits and assess this within-family relationship. This would help strengthening the argument given that even this relationship (putative correlation between B and Ne) as judged by the correlation between B and dN/dS is weakly convincing as there are few data points within each family. Moreover, the family with the largest number of lineages is the one that shows no significant relationship.

Supplementary Figure 1: It was difficult to understand this plot.
It is not clear if numbers in red are shared between Bovidae and Muridae or if they are missing for Bovidae panel (same for the two upper panels).
The last sentence of the legend: ”accounting for substitutions that were lost because appearing within introns or flanking regions.” It is not clear in the methods how these were accounted for.

Section 5.4 Need clarification. Contrary to the rest of the manuscript, this part was not clear.

line 344: It is uncelar what does the author mean by ”randomly sample the location of the first substitution ” What substitution? For the first substitution in a 4 species alignment?. For this the location in the hypothetical branch was randomly sampled ? It is also unclear what do these authors mean by ”across genes and exons;” was this done once for genes (incudng introns) and once for exons? (I assume this has to do with my previous question for Supp Fig.1 so some clarification hee is needed).

line 360: ”Two parameters of the simulation procedure were varied among conditions, namely the per third codon position density of substitutions, and the probability pclust for two successive substitutions”. It is not clear in the text what where the values for these parameters accross different similations.

line 399: Empirical estimates of mutation rate in humans were used. It is possible that mutation rates vary between the investigated taxon families. Could a real difference in mutation rates between lineage lead to the observed patterns attributed to differences in B?

Figure 4 and 5 could be placed in the appropiate sections.

Reviewed by David Castellano, 01 July 2021

In this manuscript, Galtier quantifies the strength of GC-biased gene conversion (gBGC) and its impact on protein evolution in 4 families and 32 species of mammals. He founds a substantial impact of gBGC on AT > GC synonymous substitutions (explaining ~60% of the variance). I've divided this revision into 4 sections.

1. Is the science sound, with a logical narrative and well-supported results and conclusions?

The manuscript follows a logical narrative and the methods are sound. The literature context provided in the introduction is very helpful. However, there is a key question regarding the interpretation of an important result that should be addressed before recommendation: I agree that if recombination hotspots are more ephemeral in Hominadae than in other groups then this could explain the weak clustering of WS substitution within genes. However, there is another alternative hypothesis. Could the weak clustering of WS substitutions within Hominadae genes be due to their lower diversity? The more distant the segregating sites are, the less likely would be for gBGC to generate a cluster of substitutions in a given gene. Is there a correlation between heterozygosity at the gene level and Moran's I for WS substitutions? Hence, maybe the substitution clustering within genes is conditional on B intensity + gene heterozygosity + (local & global) Ne. I am also assuming that the gene conversion tract length is not negatively correlated to the Ne. I am not sure there is literature regarding the correlation between gene conversion tract lengths and Ne.

Ideally, genes' heterozygosity and Ne should be decoupled to assess this hypothesis (by comparing genes with different mutation rates within a genome?), which is hard. But maybe this alternative can be further discussed or elaborated by the author.

2. Is there enough info to allow verifying and reproducing the data?

The supplementary information, plus the scripts, are easy to access and rerun.

3. Are there obscure passages that a potential reader can't go through?

So far the paper is easy to follow, but of course, there are always things that can be clarified. For example:

3.1 It would be good to have a table (in the main text?) with all five models (M1, M2, M3z, M3h, and M3sh), their number of parameters, and the average lnL across species. I can not find in supplementary table 1 the info regarding model M3h and the p-value of the LRTs commented in the main text.

3.2 I don't quite understand model M3sh. If q (Is q equivalent to the number of hotspots per gene?) approaches zero then does this mean that there are no hotspots within genes? or that hotspots occur in a very tiny fraction of the gene? Maybe the definition of model M3sh can be extended or rephrased?

3.3 Line 135-136. "These were very similar to estimates obtained by averaging B across 136 the M1, M2, M3z, M3sh, and M3h models, weighting by the AIC of each model." Could it be possible to add to supplementary table 1 the AIC weights too? and have a supplementary figure equivalent to figure 3 but with the AIC weighted parameters? Just to back up this sentence with figures and tables.

4. Potential extra analysis only if interesting enough to the recommender and/or author:

4.1 Regarding the across genes RSD analysis. Is the recombination map in Muridae also more uniform than in Bovidae, Hominidae, and Cercopithecidae? That could explain the results, but I understand that the recombination map for all these groups might not be available.

4.2 Again regarding the clustering of mutations within genes. Would it be possible to assess whether most WS clustering is happening at first exons (the ones closest to CpG islands)? As far as I know, at least in humans, recombination hotspots tend to occur at CpG islands at the starting of genes.

4.3 Relative to the genome-wide excess of WS mutations due to gBGC. Would it be possible to estimate the defect of SW mutations too? In other words, it would be interesting to know the overall impact of gBCG on substitution rate taking into account that the absolute number of WS and SW substitutions might be different? Maybe controlling by GC conservative substitutions across species?