

Dear authors,

Two external reviewers and myself have carefully read and then evaluated your manuscript. All the three of us had mixed feelings about the current version of this review. Although, we all believe this is an important and timely topic, the current version may be too short to cover a large and deep topic of molecular evolution.

> We would like to thank Guillaume Achaz, Benoit Nabholz, and the second reviewer for their comments. This has greatly helped us to edit our manuscript, and we hope this new, longer version is more suitable for acceptance. Briefly, we restructured the manuscript into three main sections: the impact of context on alleles, the non-selective population-level processes that change allele frequencies, and the molecular-level processes that can affect detection of selection. We have broadened the introduction to include some previously existing sections and better target the remainder of the manuscript. Our specific responses follow below after each >.

Sincerely,

Fanny Pouyet and Kimberly Gilbert

Although I cannot recommend the ms in its current version, I truly believe that there is the potential basis for a more satisfying version, with some extra work. Mainly, you will have to face two choices: (1) either elongate significantly the content (I am aware that most journal editors requires to shorten article as much as possible, so that this recommendation may seem weird) and discuss more generally the interference between drift, selection, demography and/or structure (processes mostly acting at the population scale) together with molecular processes (such as BGC or for example meiotic drive). (2) Another possible choice is to refocus your review on a more specific point and to treat it more deeply.

> We reorganized the manuscript into 3 paragraphs that are the “genomic and environmental context”, the “population-level processes” and the “molecular-level processes”. This should help the reader to distinguish the origins of the evolutionary forces we describe. We also reorganized Table 1 accordingly to these new paragraphs and added sub-sections for the population- and molecular-level processes.

On a personal note, I share your interest on BGC but believe that it belongs to a larger class of molecular processes on which you could include mutations (a topic on which several new observations and theory have been brought recently), meiotic drive, among others.

> This is an excellent point and we hope our inclusion of more molecular processes (segregation distortion, CpG hypermutability, and mutational processes as well as transposable elements) to a section dedicated to a range of these processes and their impact on selection inference has improved the manuscript.

If you decide to send a revised version of the manuscript to PCI Evol Biol, please address all the points of the two reviewers and rework your article making an explicit choice between both options I have discussed above. I understand this decision may be a frustrating for you, but be assured that none of the

reviewers nor myself have a negative opinion toward the general objective of your review. It is actually quite the opposite, we all have a strong interest in these questions.

Sincerely yours,

Guillaume Achaz

## Reviews

*Reviewed by Benoit Nabholz, 2019-11-08 21:35*

In this article, the authors propose a point of view on the processes that must be considered to understand polymorphism patterns. They position their perspective in the context of the recent debate on the utility of the neutral theory to explain patterns of molecular evolution (Kern and Hahn 2018; Jensen et al. 2019). They also aim to “clarify the terminology” and “refine our definitions”.

I have a mixed opinion about this perspective. I agree that it is important to stress out that linkages, GC-biased gene conversion, and allele surfing could have an impact on alleles' fate and that these processes are not adaptive. Many previous articles had already indicated that these processes could mimic selection or limit our ability to detect positive selection. So, nothing is new in this manuscript. However, it has the strength to consider all these processes at once. Unfortunately, I found that the article has some approximations and ambiguities that make that it does not live up to its ambition to clarify the terminology. Moreover, the manuscript put too much emphasis on allele surfing compared to other demographic processes such as structure or bottleneck.

> We have shifted the emphasis of the manuscript as not so much a clarification of terminology, but an emphasis of the range of factors impacting inference of selection. We now elaborate further on other demographic processes (population-level processes) and molecular-level processes.

Below, I will elaborate on the points I have addressed in the paragraph above in order to help authors improve their manuscripts. In my opinion, it unfortunately requires a lot of work before to be recommended.

I apologized if my comments refer to page number and quotes of author's sentences but it would have been convenient to have numbering of the lines.

> Submission to Arxiv made us remove line numbering before accepting our ms. We are sorry for this inconvenience.

P2: “... of polymorphisms, but do not fall under the umbrella of selection, into a third category of non-adaptive evolution.” There is a problem with this sentence. Do you mean: “... a second category...”. Otherwise what is the second category? It is drift? But allele surfing is linked to drift.

> We have greatly modified the text in this section and hope that this re-writing has clarified the issue.

P2: “...environmental context of polymorphisms” The term environmental is ambiguous here because it may refer to the ecological definition of the term. The reader understands much better that it is the genetic environment (i.e., linkage) to which the authors refer here. More generally, I don't like that the authors don't go straight to the point that they want to consider in the perspective : i) linkages, ii) allele

surfing and iii) GC-Biased gene conservation. This leaves ambiguities and some time, it seems that the authors beat around the bush.

> We broadened the introduction to include the definitions of selection and drift and after this begin to discuss the main point of the manuscript – processes biasing selection inference.

P3: "... or cases where selection behaves in a stochastic manner (e.g. in finite populations)." I don't understand this statement. What do the authors mean by "selection that behaves in a stochastic manner"? Is it nearly neutral allele that behaves neutrally in a small population and is selected in a large population? This needs clarification.

> We have removed and rewritten this section.

P3: The whole paragraph on defining natural selection: I understand that this perspective is intended for non-specialists but this paragraph explains very basic concepts that are not useful in my opinion.

> To not skip over the definition, but to shorten it and direct it at a more appropriate audience, we now describe selection and drift more briefly in the introduction.

P3: Toward the end of the paragraph: Sexual selection is not the only form of selection that "are not necessarily advantageous at the population level". Many forms of selfish behaviors selected by positive selection can lead to a fixation of allele not advantageous at the population level (Hamilton 1970).

> We now mention selfish behaviors in the molecular-processes level, paragraph on "segregation distortion".

P4: "...but these other complicating processes to fully understand evolutionary biology and the generation and maintenance of genetic diversity across the genome and across populations and species." This is typically the kind of sentence that is not very clear at first but becomes clearer once you read the rest of the manuscript and understand that the 'complicating processes' are linkages, gBGC, and demography. Here, the authors also refer to variation across species (between-species divergence) but this is not addressed at all by the rest of the manuscript.

> We have removed this sentence while rewriting the introduction. We do not refer to between-species divergence nor to variation across populations anymore.

P4: "... their impacts on genetic diversity are often underappreciated (see [13, 14] and [15] for reply)." The fact that these processes are often acknowledged "by the vast majority of the community" actually reduces the interest of this perspective. I recommend the authors to find more examples of studies where these processes have been underappreciated to support their view that it is worth to make a manuscript. Here, the authors only refer to one example dating back to 2005/2006 that it is not even a genome-wide study (although, I agree that this example is relevant but I want to emphasize that the authors need more examples from genome-wide studies since the authors focus on the field of "evolutionary genomics"). As an example, the author could also consider the case of gBGC in the human HARs (Galtier and Duret 2007). But a literature review would strengthen the manuscript.

> We now include the example on HARs as well as an example from 2017 on codon usage in humans and from 2018 in flycatchers:

N Galtier, L Duret. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics* 23(2007) 273-277

F. Pouyet, D. Mouchiroud, L. Duret, M. Semon, Recombination, meiotic expression and human codon usage., *Elife* 6 (2017) e27344.18

P. Bolivar, C. F. Mugal, M. Rossi, A. Nater, M. Wang, L. Du-toit, H. Ellegren, Biased Inference of Selection Due to GC-Biased Gene Conversion and the Rate of Protein Evolution in Flycatchers When Accounting for It, *Molecular Biology and Evolution* 35 (10) (2018) 2475–2486

P6: “The former may simplify detecting the presence of selection but ...”. I don't understand what the authors are referring to by “The former”. Is it background selection? If so, I don't think that background selection simplify detecting the presence of selection. You should elaborate on that with some references to theoretical or empirical works.

> We have rephrased these sentences to clarify that detecting selection may be simpler when several linked and selected amplify the strength of selection, in contrast to Hill-Robertson interference, where opposing allelic impacts may weaken the strength of selection.

In section 4.2 “The impact of demography” : the authors focus exclusively on gene surfing. One can feel that the authors have expertise on the subject and this section is interesting to read with many references. However, I don't think that this focus is justified. Indeed, other demographic scenario are known to mimic positive selection. This is particularly true for population bottlenecks (Thornton and Jensen 2007; Innan and Stephan 2003), but population structure (Tian et al. 2008) and possibly other intriguing demographic scenario such as sweep stack (Sargsyan and Wakeley 2008) could be considered as well. Similarly, gene flow could also be considered (Bierne et al. 2011) although I don't know if it should be defined as a demographic event. I have the same comment for table 1 where gene surfing has it own line together with genetic drift. However, gene surfing could be considered as a particular case of drift.

> This is a fair point, and we have now broadened our discussion of demographic and population-level processes that can bias the inference of selection.

Section 4.3 “gene conversion” : Once again, this section is well written with abundant literature but too much emphasis is put on dBGC. The process is detailed whereas, to my knowledge, it has never been involved in the false-positive cases of natural selection. In contrast, gBGC has often been responsible for false positive (Duret and Galtier 2009; Ratnakumar Abhirami et al. 2010; Galtier et al. 2009).

> Yes, this is true. We removed the section on dBGC and now write: “To our knowledge dBGC is not expected to be confounded with selection and will not be discussed further herein. On the contrary, gBGC is often responsible for false positives in inferences of selection.”

A much more accurate reference for the existence of gBGC in birds is (Webster et al. 2006) rather than (Weber et al. 2014).

> We thank the reviewer for spotting this error due to a bibtex typo!

## REFERENCES

- Bierne N, Welch J, Loire E, Bonhomme F, David P. 2011. The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Mol Ecol* 20: 2044–2072.
- Duret L, Galtier N. 2009. Comment on " Human-specific gain of function in a developmental enhancer". *Science* 323: 714–714.
- Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *TRENDS Genet* 23: 273–277.
- 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: 1–5.
- Hamilton WD. 1970. Selfish and Spiteful Behaviour in an Evolutionary Model. *Nature* 228: 1218–1220.
- Innan H, Stephan W. 2003. Distinguishing the hitchhiking and background selection models. *Genetics* 165: 2307–2312.
- Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, Charlesworth B. 2019. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. *Evolution* 73: 111–114.
- Kern AD, Hahn MW. 2018. The neutral theory in light of natural selection. *Mol Biol Evol* 35: 1366–1371.
- Ratnakumar Abhirami, Mousset Sylvain, Glémin Sylvain, Berglund Jonas, Galtier Nicolas, Duret Laurent, Webster Matthew T. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc B Biol Sci* 365: 2571–2580.
- Sargsyan O, Wakeley J. 2008. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor Popul Biol* 74: 104–114.
- Thornton KR, Jensen JD. 2007. Controlling the false-positive rate in multilocus genome scans for selection. *Genetics* 175: 737–750.
- Tian C, Gregersen PK, Seldin MF. 2008. Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet* 17: R143–R150.
- 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: 542.
- Webster MT, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol* 23: 1203–1216.

*Reviewed by anonymous reviewer, 2019-11-12 15:13*

In this short review, Pouyet and Gilbert tackle a huge topic that has been a major focus of the field for decades, putting their own spin on the question of the relative importance of drift and selection in molecular evolution. This is an ambitious topic to approach, and the authors highlight some important factors to consider in thinking about this topic, especially focusing on linked selection, demography, and biased gene conversion. Overall, however, I think this review suffers from some flaws that substantially limit its value at the present. First, there are many places throughout the review where it seems like a desire for brevity has

led to some imprecise or potentially confusing statements. For example:

- On page 2, I'm not sure it is correct to say that "whether a majority of polymorphisms in the genome are neutral or selected" is a major point of debate. While there are big arguments about whether the majority of polymorphisms are affected by linked selection, I'm not sure even an ardent selectionist would argue the majority of polymorphisms in the genome themselves have selection coefficients different from 0.

> We have modified the writing to reflect that the dividing view is whether there is a large impact of positive selection on polymorphisms, directly or indirectly.

"A major dividing view on this point is whether a large proportion of polymorphisms is affected by adaptive natural selection directly or indirectly. Identifying variants evolving solely under genetic drift is not straightforward, and there is ample space for additional data across a wider range of species to contribute towards these investigations and our understanding of molecular evolution."

- On page 3, the definition of balancing selection does not encompass negative frequency dependent selection, or other alternate mechanisms that maintain variation such as spatially varying selection.

> We now give a brief general definition of balancing selection, but address other mechanisms such as frequency-dependent selection and spatially varying selection in the section 'Genomic and environmental context':

- Also on page 3, the definition of sexual selection does not really encompass variation in reproductive success; I'm not sure a mutation would be considered subject to sexual selection if it increased survival in males but was neutral in females, yet that is completely consistent with the listed definition.

> We have removed the definition of sexual selection since it is not a focus of our manuscript. "There is also the case of sexual selection that we do not address as its many and complex cases merit a review of their own."

- On page 7, I think most definitions of demography would encompass more processes than just changes in population size (e.g., population splits and migration).

> As mentioned in the specific responses to reviewer 1, we now elaborate further on other demographic and population-level processes.

- On page 8, I'm not sure it is precise to say that biased gene conversion has no impact fitness. In the extreme case of a e.g. a recessive lethal mutation, no amount of segregation distortion can force it to increase in frequency too much in the population. The fitness effects of the mutations are still expressed in the next generation. I think the important point is that the transmission advantage can cause 'faster-than-neutral' increase in frequency (and shorter time to fixation with a higher prob of fixation), but only if the phenotypic effect of the mutation is small enough to be overcome by the transmission advantage.

> We have replaced this with the following sentence to be more precise "In some cases when G/C alleles are beneficial compared to A/T, gBGC can amplify the speed of fixation of a beneficial allele in the population while in other cases when G/C alleles are slightly deleterious, gBGC can counteract the effect of selection (if the transmission bias parameter  $b$  is stronger than the selection parameter  $s$ )."

Second, the way the authors discuss linked selection is not always very clear to me. This is most relevant to some statements in the conclusion, in particular the statement: "to infer selection, one must not only disprove genetic drift [...] but also disprove the additional prevalent non- adaptive processes [such as linked selection]." The major thrust of this statement seems to be getting at estimates of

selection coefficients of particular mutations, e.g. in the context of determining the causal locus under selection. But if the goal is more general, such as to detect regions of the genome subject to a selective sweep, most methods rely on the existence of linked selection for their power. E.g., a reduction in diversity around a hard sweep can only be observed if recombination rate is low enough that some nearby region of the genome evolves by linked selection; most methods to detect selection in resequencing data are generally detecting the perturbations to the neutral coalescent induced by selection. Throughout the review, the authors seem to move between focusing on the challenge of identifying the precise sites under selection (where linkage poses a real issue as they discuss on the bottom of page 6), and the more general question of the prevalence of selection in genomes, where linkage can be a major help, in the sense that the impact of selection on linked neutral variants is one of the main ways we can detect it.

> This is a valid point, and we agree in part. There are many methods that function to identify selection by identifying patterns across stretches of the genome. We mention one example of such method (S/HiC, a machine-learning approach) now in our introduction. However, we also come at this review from the point of view of methods that require only neutrally evolving sites (SFS methods of inference, or other examples mentioned in the text), and in this case finding the sites under selection and any that may be linked to them is still a goal, so that they can be disregarded in a given analysis. This becomes particularly tricky when trying to infer demographic history, so as to control for neutral effects when inferring contemporary selection, since the process becomes a bit circular in terms of trying to distinguish which sites are selected or subject to the forces of selection. We hope that our rewriting of the goal and introduction of the manuscript as well as mentioning methods that use information from linkage to infer selection has helped to clarify this important point.

Finally, related to the above point, it is not really clear to me what the major message of this review is intended to be. As discussed above, the focus of the conclusion seems to be on the challenge of identifying causal variants in selection scans. This is a valuable point and one that is worth discussion, especially as in some cases the more robust the signature of selection, the harder it can be to pinpoint the causal mutation (e.g. in regions of low recombination, the region of the genome impacted by a sweep can make selection easier to detect, but harder to disentangle). However, the introduction seems to set up a different question, on the relative importance of drift and selection; this is a greater focus in the sections on demography and biased gene conversion. Overall, jumping back and forth makes the entire piece a little hard to follow and it makes the overall message somewhat muddled.

> We have edited our abstract, introduction, and conclusion to better clarify our goal, which we have now focused on the topic of ‘processes that can bias selection inference’, and not in particular on the importance of drift vs selection.

I think many of these issues could be resolved by a longer manuscript. A piece like this – that is tackling a big topic with a goal of avoiding polemics and trying to bring clarity to the debate – really needs space to breathe, to expand on definitions, to pull apart the different messages with enough text to explore them. At just under 2500 words, this review is trying to tackle too many things in too few words, in my opinion.

> Thank you for this comment; we have lengthened our manuscript in response to it. It is a difficult line to walk since there are still many more topics and points that could be included and discussed at greater length, but hopefully our manuscript in its current state has improved upon the depth of discussion for the most important points relevant for inference of selection and the biases that go along with it.