Response to editor and point-by-point response to reviewers' comments on "*Detecting diversifying selection for a trait from within and between-species genotypes and phenotypes*"

Editor:

The majority of comments, including my own, are about clarity and impact statement of this study. Reviewer #1, for instance, mentioned a couple of points that could be tighten up, from the point of population genetics, for better scholarship and clarity of the method presented. Similarly, reviewer #2 pointed out a room to discuss how the proposed method and results are related to an index of neutrality proposed in Lynch 1990 and Lande's LGGD. I have a few additional ones that relate this work to the PCM literature. Addressing those will further strengthen the micro-macro link of this article. Besides, there are few technical details that should be clarified and verified. Please see relevant comments from the reviewers.

But overall, I am extremely excited to see this manuscript and the perspective that this method could bring to the field!

We are grateful for your comments and the one raised by the two reviewers. We give below a point-by-point response to the editor and reviewer comments, including the text modified in the manuscript as a response to each question below. We believe that our manuscript has benefited greatly from this revision. We hope that we have been able to further convey the micro-macro link by making more salient the difference (both strengths and weaknesses) between our methodology and previous work, that the method we employ is now explained with more clarity, and that the assumptions and limitations of our approach are now more adequately described.

Comments to the Author

Heritability

Either at a section presented in lines 130-141 or in discussion, you could mention a meta-analysis of heritability in Hansen et al (2011) Evolutionary Biology 38: 258-277 and in Hansen and Pélabon (2021) Ann Rev Ecol Evol System 52: 153-175. The main message of their complication is that empirical estimates of h^2 are surprisingly stable, falling within the range of 0.2-0.5 in the vast majority of cases (see their Table 1 in these papers). These results could be used to further strengthen your method because a lack of heritability estimates may not introduce much uncertainty in the proposed neutrality index. The bias could be corrected by using an empirically estimated h^2 in the same trait types. The caveats of course is that we never know the actual h^2 until we measure it.

Thanks a lot, this is a really good addition to the manuscript, we changed the paragraph in the section *Applicability to empirical data* from:

"Similarly, using the broad-sense heritability (H^2) instead of narrow-sense heritability (h^2) results in an underestimation of ρ since $h^2 \leq H^2$. to

"Additionally, empirical estimates of h^2 are surprisingly stable across species and fall within the range of 0.2-0.5 in a vast majority of phenotypic traits tested (Hansen, 2011; Hansen and Pélabon, 2021). Alternatively, using the broad-sense heritability (H^2) instead of narrow-sense heritability (h^2) results in an underestimation of ρ since $h^2 \leq H^2$. If available, such prior knowledge on h^2 can be leveraged instead of assuming complete heritability to increase the statistical power to detect diversifying selection."

Add in the *Discussion* we reiterated this argument, and we added:

"If available, any prior knowledge on h^2 can be leveraged instead of assuming complete heritability to increase the statistical power to detect diversifying selection (Hansen, 2011; Hansen and Pélabon, 2021)."

Implication for PCM and further

I see an opportunity to mention further application of this line of thinking to phylogenetic comparative methods. What comes on top of my head is to use the expression presented in eq (8) as part of a prior for a method that detects shifts in adaptive regimes from data (Ingram and Mahler 2013 Methods Ecol Evol, Uyeda and Harmon 2014 Sys Bio, Khabbazian et al 2016 Methods Ecol Evol, Mitov et al 2019 Theo Pop Biol). More generally, I see a room to expand the discussion starting in line 267 to mention that the presented method could be seen as a way to disentangle congruence models (e.g. Louca and Pennell Nature 2020) by allowing researchers to use genomic information. This is an additional approach to model comparison and evaluation of model adequacy (e.g. absolute fit of a model to the data, Pennell et al 2015, Am Nat).

We completely agree that we should mention in the *Discussion* that the genomic information could be used more generally in the context of phylogenetic comparative methods. We restructured the *Discussion* (see comments by reviewer II), and integrated these ideas:

"In the context of phylogenetic comparative methods, modeling mean trait evolution as a function of nucleotide divergence (*d*) instead of time has more general consequences. As an example, trait variation is often modeled as a Brownian process running on a time-calibrated tree, which can produce biases (Litsios & Salamin, 2012). Indeed, for a neutrally evolving trait, trait variation depends directly on the number of generations, which in turn correlates with time. But, since species generation time might vary along the phylogenetic tree, *d*-scaled trees absorbing changes in generation time should be used instead of time-scaled trees. Using nucleotide divergence would also remove the potential effect of model assumptions required to calibrate ancestral node ages (e.g. molecular clocks). We argue that the soundness of studying trait evolution on *d*-scaled trees can be evaluated by the absolute fit of a model to the data (Pennell et al., 2015). More generally, genomic information could potentially be seen as a way to disentangle congruence models (Louca & Pennell, 2020), or as prior for methods that detect shifts in adaptive regimes (Ingram & Mahler, 2013; Uyeda & Harmon, 2014; Khabbazian et al., 2016; Mitov et al., 2019)."

Minor points:

Line 65-68: there is a recent argument that adds to this line of issue. Grabowski et al. (2023) Systematic Biology 72: 955-963

We fully agree and changed the sentence from:

“In other words, the better fit of a Brownian process does not necessarily constitute proof of the neutral model.” to:

“In other words, the better fit of a Brownian process does not necessarily constitute proof of the neutral model. Also, a better fit of a Brownian could be due to a trait evolving with a rate too low compared to the timespan on which it is measured (Grabowski et al., 2023).”

After eq 15: trade -> trait (I suppose)

Yes, sorry for this mistake.

Line 210: similar -> similarity

We modified the sentence and developed also on the consequences (see comments to reviewer I), changing the text from:

“Under stabilizing selection, the variation between species is depleted because the mean trait value is **maintained similar** between different species, which leads to ρ<1.” to:

“Under stabilizing selection, the variation between species is depleted because the mean trait is **maintained toward similar values** between different species, which theoretically leads to ρ<1.”

Reviewer I:

Comments to the Author

In this manuscript, the authors propose a neutrality test for macroevolution of traits along phylogenies, which makes use of between and within-species variation, for phenotypic and DNA sequence data. The main idea is that, by scaling between-species divergence by with species variation, and then phenotypic data by molecular data, the dependence on mutation rates and effective population size can be removed. This leads to a unitless neutrality index, which equals 1 under neutrality, is larger than 1 under divergent selection, and close to 0 under stabilizing selection (with identical optima among species). Furthermore, the method is also robust to temporal changes in mutation rates and effective population size. The authors show by simulation that the test for divergent selection is overall powerful and conservative, while stabilizing selection can be confounded with neutrality. They then apply the method to a dataset of body and brain mass in mammals, and show that brain mass is under divergent selection.

I found this manuscript quite stimulating and well written. It does a nice job of bridging between the micro- and macro-evolutionary literatures (but see comments below), which should appeal to the large readership interested in this interface. And the approach the authors propose can potentially be applied broadly, provided that the relevant data becomes available in more taxa. So overall I really enjoyed reading this ms, but I also found that some omissions and ambiguities should be corrected to improve its clarity.

Thanks a lot, we indeed hope that the relevant data becomes available in more taxa. This is initially the reason why we started to investigate the relationship between tests of selection at different scales and whether they are congruent (theoretically and empirically).

First regarding the objectives of the method, the authors state at the end of their intro (85-87): “From the field of population genetics, our study can be seen as the macro-evolutionary generalization of Q_{ST}–F_{ST} methods to account for phylogenetic relationships between species”. This is correct, but one key citation that is missing here is Ovaskainen et al (2011 Genetics), which accounted for differentiation between pairs of populations in a way that is quite similar to what you do here for divergence between species. In addition, the method dealt with multiple correlated traits, as you describe in lines 116-118 and the appendix, so I think this method should be taken as the closest point of reference and referred to in some detail in the ms.

We apologize for missing this key citation in the manuscript. Ovaskainen *et al.* (2011) accounted for differentiation between pairs of populations using co-ancestry between any individuals, while also modeling trait and nucleotide differentiation as random variables originating from a random process (drift in the case of neutral evolution). Accordingly, we referred to it in some detail the *Introduction*, changed from:

“Across several populations, by contrasting both trait and genetic differentiation, Q_{ST}–F_{ST} methods have been used to determine the selective regime and to quantify the strength of selection acting on a trait (Leinonen et al., 2008; Merila & Crnokrak, 2001). A trait differentiation (Q_{ST}) higher than genetic differentiation (F_{ST}) is interpreted as a signature of diversifying selection due to adaptation in different optimum trait value in the different populations (Lamy et al., 2012). Contrarily, Q_{ST} lower than F_{ST} is interpreted as a signature of stabilizing selection. However, Q_{ST}–F_{ST} methods have been found to require many populations (O’Hara & Merila, 2005), and that various factors can generate a spurious signal of selection (Edelaar et al., 2011; Pujol et al., 2008). Moreover, the test for diversifying selection is limited to recent local adaptation since the test is based on the variation observed within a single species.” to:

“Across several populations, by contrasting both trait differentiation (Q_{ST}) and genetic differentiation (F_{ST}), so-called Q_{ST}–F_{ST} methods have been used to determine the selective regime and to quantify the strength of selection acting on a trait (Merila & Crnokrak, 2001; Leinonen et al., 2008; Ovaskainen et al., 2011). Q_{ST} higher than F_{ST} is interpreted as a signature of diversifying selection due to adaptation in different optimum trait values in the different populations. Contrarily, Q_{ST} lower than F_{ST} is interpreted as a signature of stabilizing selection (Lamy et al., 2012). Other frameworks explicitly model genetic drift as a random process generating both trait and genetic differences between individuals and populations. This integrated framework can discriminate between selection and genetic drift as a cause of trait differentiation between populations of the same species **(Ovaskainen et al., 2011).** However, regardless of the strengths and weaknesses of each method (Pujol et al., 2008; Edelaar et al., 2011; **Ovaskainen et al., 2011)**, tests of trait differentiation between populations are ultimately limited to recent local adaptation since they are based on the variation observed within a single species.”

And later on, in the last paragraph of the *Introduction*:

“From the field of population genetics, our study can be seen as the macro-evolutionary generalization of Q_{ST}–F_{ST} methods to account for phylogenetic relationships between species.” to:

“From the field of population genetics, while Q_{ST}–F_{ST} methods and their derivatives ultimately seek trait differentiation among different populations from the same species **(Leinonen et al., 2008; Ovaskainen et al., 2011),** our study can be seen as their macro-evolutionary generalization to account for phylogenetic relationships between species.”

And finally we modified the *Discussion* from

“The main novelty of our study was to use the nucleotide divergence and polymorphism to normalize trait variation between and within species. In the context of within species variation, Q_{ST}–F_{ST} tests have been developed to compare trait and sequence across several populations to test for selection (Leinonen et al., 2013; Martin et al., 2008). Our neutrality index also used the genetic sequences from which nucleotide divergence and polymorphism are estimated.” to

“As such, the main novelty of our study was to use the nucleotide divergence and polymorphism to normalize trait variation between and within species. In this context, our test bears many similarities to Q_{ST}–F_{ST} tests (and their derivatives) that have been developed to test for selection of a trait across several populations while also leveraging genetic differentiation (Martin et al., 2008; Leinonen et al., 2013) or co-ancestry between individuals **(Ovaskainen et al., 2011).** Our method can be seen as an extension at the phylogenetic scale [...]”

Similarly, I think the authors should refer much earlier in the paper (starting from the introduction and the methods) to the McDonald-Kreitman test, or MK test (1991 Nature). At the moment they just cite it once in the discussion (line 308), but I think this does not give justice to how what they do here relates to this classic approach in molecular evolution. Equation (20) relates most directly to the MK test, the neutrality index is the same, except that the synonym/non-synonym distinction is replaced by one between quantitative trait vs (neutral) genomic sequence. And when they mention in the discussion that their “test ultimately bear analogy to the codon-based test of selection, where the ratio of non-synonymous to synonymous substitutions (ω) is compared to 1" (310-312), actually this is even more directly related to the α parameter in the MK test (https://en.wikipedia.org/wiki/McDonald%E2%80%93Kreitman_test).

We completely agree that we should mention the MK test in the introduction. Because the MK test has been designed for a pair of species (not across a phylogeny), it also helps smooth the transition from within species variations (i.e.$Q_{ST}–F_{ST}$ methods) and the phylogenetic comparative method (Browian process in a phylogenetic tree). Accordingly, we modified the introduction from:

"To disentangle neutral evolution and selection, trait evolution can also be observed at a larger time scale. [....]. Altogether, both the trait variance and the evolution in mean value can be used to test for trait selection in a pair of species (Walsh & Lynch, 2018)." to:

"To disentangle neutral evolution and selection, trait evolution can also be observed at a larger time scale. [....]. As an analogy, in the context of protein-coding DNA sequences, leveraging both within species diversity and divergence to a sister species is the crux of the McDonald and Kreitman (1991) test. In such a test, inflation of nucleotide divergence to the sister species is compared to polymorphism within species, while neutral makers (usually synonymous sites) are used to determine the neutral expectation and are used for normalization. Altogether, testing for selection in a pair of species can leverages both the within and between species variations (Walsh & Lynch, 2018). "

In the discussion, we have chosen to compare our neutrality index to the ratio of non-synonymous to synonymous substitutions (ω) instead of the α parameter in the MK test for several reasons. First, at the phylogenetic scale, only ω methods can leverage many species and an underlying tree topology, while MK tests can only be made on a pair of species (one of each must have p_N and p_S estimates). Second, it is easier to discuss the literature with the analogy of our index ρ with ω. Indeed ω=1 can happen under a mix of adaptation and purifying selection (Nielsen, 2005), and strong purifying selection can result in ω<1 even though those genes and sites are under adaptation (Latrille et al., 2023). These properties of ω are not as well described in the literature for α. As such we would advocate for keeping the analogy between ω and ρ. In the *Discussion*, we agree that our wording was misleading, and we changed the paragraph to state more explicitly the reasons to compare our neutrality index ρ to ω, from:

"First, we acknowledge that our test took inspiration from the McDonald and Kreitman (1991) test devised for protein-coding DNA sequences, where synonymous mutations were used to determine the neutral expectation, and the inflation of divergence was compared to polymorphism within species. Second, because ρ was compared to 1, our test ultimately bear analogy to the codon-based test of selection, where the ratio of non-synonymous to synonymous substitutions (ω) is compared to 1 (Goldman & Yang, 1994; Muse & Gaut, 1994)." to:

"First, we acknowledge that our test took inspiration from the McDonald and Kreitman (1991) test devised for protein-coding DNA sequences in a pair of species, except that the non-synonymous versus synonymous distinction is replaced by the comparison between quantitative trait and neutral genomic sequence. Second, at the phylogenetic scale, when comparison is done across several species, our test also bears analogy to codon-based test of selection, where the ratio of non-synonymous to synonymous substitutions (ω) is compared to 1 (Goldman & Yang, 1994; Muse & Gaut, 1994)."

Regarding the method per se, I have two main comments. The first methodological point is that I find it misleading to write that "This ratio allows removing the effect of N_e and μ" (94 for within species variation, and similar statements for between species divergence). This is only valid if μ is the same between the different categories, but this is almost certainly not true. The mutation rate of a quantitative trait is determined by an unknown number of loci, with an unknown number of base pairs contributing to the trait (=the mutational target of the trait), so it cannot be considered by default to be the same as the mutation rate per bp of randomly picked (or neutral) sequences in the genome. In practical terms, in a typical mutation accumulation experiment, the mutational variance of traits can be estimated, as well as the (per bp) mutation rate of DNA sequence, but in no way can the ratio of these quantities remove all influence of mutation rate. So I think the ratio of mutation rates for quantitative traits over neutral sequence should remain in eq. 6-7, and similarly for eqs. 17-18 below. Then this ratio disappears in the neutrality index in eq. 20, so your results are actually more general, since they allow the mutation rate of quantitative traits to differ from that of genomic sequence, as it should.

This is a very good point. We concur that a quantitative trait is determined by an unknown number of loci, with an unknown number of base pairs contributing to the trait. Thus, we agree that the mutation rate per base in neutral sequence does not equate to the mutation rate per loci underlying the quantitative trait. However, theoretically, given that we know the number of base pairs contributing to the trait for each loci, both mutation rates become comparable when scaling by the number of loci (L), and the effect of each loci (a). This is implicitly our reasoning but we agree that it is convoluted and it would be more straightforward to explicitly acknowledge different mutation rates. Additionally, we agree that "allows removing the effect of N_e and μ" is misleading and inaccurate since what the normalization really does is removing the effect of N_e and the number of generations.

We thus have made several changes to the *Materials and Methods* and across the manuscript:

- We now use "μ" for the mutation rate per generation of loci underlying the quantitative trait, and "*u*" for the mutation rate per generation of nucleotide sites.
- We kept "μ" and "*u*" in equations until the definition of the neutrality index, where they all cancel out in the ratio of σ^2_B over σ^2_W.
- We changed the text from "removing the effect of N_e and μ" to removing the effect of N_e and t (number of generations)"
- We explicitly state that since the number of generations is the ratio of time divided by generation time (average time between two consecutive generations), canceling the effect of the number of generations also cancels the effect of both time and generation time.
- We removed σ^2_M (see comments by reviewer 2) and kept V_M until they cancel out.
- We have updated all equations 1-24 and definitions to reflect these changes.

The second methodological point is that I haven't been quite able to figure out what nucleotide divergence is used in the denominator of eq. (16). The variance in mean phenotypes in the numerator of this equation is generally obtained by comparing multiple species (by which I mean more than 2), with different values of nucleotide divergence between them (one per pair of species). For each species pair, the phenotypic divergence (difference in mean phenotypes) should be proportional to their molecular divergence, so the variance in mean phenotypes among species should relate to the variance in molecular divergence. More precisely, there should be a covariance between phenotypic and molecular divergence among groups of species, as explicitly addressed by the method by Ovaskainen et al (2011) cited above, which seems like it could also apply here to between species comparisons. In any case, I think the scaling by nucleotide divergence should be explained in more detail, as it is currently difficult to understand what is actually implemented in the method.
We completely agree that the method is not clearly explained, and that the jump from the formalization to the estimate should be more explicit.

In the section *Between-species trait variations,* we had chosen to adopt the formalism of Luke Harmon's book *Phylogenetic Comparative Methods* (lukejharmon.github.io/pcm/chapters/). In this formalism, starting from the same ancestral population, divergent lineages accumulate changes in mean phenotypes, hence one can write the variance in mean phenotypes (Var[$\overline{P}_t$]) across the different lineages has a function of the number of generations (t). Divergent lineages will also accumulate nucleotide changes reaching fixation in the population, generating different nucleotide divergence (d).

We agree that this formalism is not the best suited and is confusing in our case since d is also a random variable (contrarily to t). Given a phylogenetic tree, the formalism from Hansen & Martins (1996) is more appropriate since it refers to a pair of species, and it is the manuscript that we are citing anyway. Hence we performed the following substitutions in notations:

- Var[$\overline{P}_t$] → cov($\overline{P}_i$, $\overline{P}_j$) is the covariance between mean phenotype in species *i* and mean phenotype in species *j*.
- t → $t_{i,j}$ is the number of generations between the root of the tree and the most recent common ancestor of taxa i and j.
- d → $d_{i,j}$ is the nucleotide divergence between the root of the tree and the most recent common ancestor of taxa i and j. In other words, $d_{i,j}$ is the number of observed substitutions per nucleotide site during the $t_{i,j}$ generations.

The generalization from a pair of species to a phylogeny requires accounting for different nucleotide divergence between species, where one can leverage the comparative framework and the Brownian process (Felsenstein, 1985; O'Meara et al., 2006). Generally, \hat(σ)$^{2}_{B}$ can thus be seen as an estimate of the rate of the evolution of the quantitative trait along a phylogeny, when the tree is measured in units of 4*d*. As such, any phylogenetic comparative methods that allow the estimation of phenotypic rates of evolution on a tree scaled by 4*d*, (instead of time as is usually the case) can be used to estimate σ^{2}_{B}.

Thus we added a paragraph at the beginning of the section *Estimate* to clarify this point:
"[...] On the other hand, σ^2_B such as as defined in eq. 18 only refers to a pair of species, and thus must be generalized to account for different species divergence, as is done in the comparative framework (Felsenstein, 1985; O'Meara et al., 2006). Generally, \hat(σ)^2_B can thus be seen as an estimate of the rate of the evolution of the quantitative trait along a phylogeny, when the tree is measured in units of 4*d* (*d* is the nucleotide divergence). As such, any phylogenetic comparative methods that allow the estimation of phenotypic rates of evolution on a tree scaled by 4*d*, instead of time as is usually the case, can be used to estimate σ^2_B."

And later on, in the section *Multivariate Brownian process*, we reiterated this argument:
"The branch lengths of the tree used to model the Brownian process is measured in units of 4*d* (*d* being the nucleotide divergence). The off-diagonal elements of Σ are the covariance between traits, and the diagonal elements are the variance of each trait when measured in 4*d* units, and thus equate to σ^2_B."

And finally, in the section *Discussion*, we extend on the consequence of using tree scaled by 4d, as also suggested by the editor:
"In the context of phylogenetic comparative methods, modeling mean trait evolution as a function of nucleotide divergence (*d*) instead of time has more general consequences. As an example, trait variation is often modeled as a Brownian process running on a time-calibrated tree, which can produce biases (Litsios & Salamin, 2012). Indeed, for a neutrally evolving trait, trait variation depends directly on the number of generations, which in turn correlates with time. But, since species generation time might vary along the phylogenetic tree, *d*-scaled trees absorbing changes in generation time should be used instead of time-scaled trees. Using nucleotide divergence would also remove the potential effect of model assumptions required to calibrate ancestral node ages (e.g. molecular clocks). We argue that the soundness of studying trait evolution on *d*-scaled trees can be evaluated by the absolute fit of a model to the data (Pennell et al., 2015). More generally, genomic information could potentially be seen as a way to disentangle congruence models (Louca & Pennell, 2020), or as prior for methods that detect shifts in adaptive regimes (Ingram & Mahler, 2013; Uyeda & Harmon, 2014; Khabbazian et al., 2016; Mitov et al., 2019)."

Minor points

32 : "adaptation TO different optimum trait valueS" rather than "in different optimum trait value"
Completely agreed, sorry for these mistakes

34-36: This sentence seems grammatically incorrect: "However, Q_{ST}–F_{ST} methods have been found to require many populations (O'Hara & Merila, 2005), and that various factors can generate ..." Perhaps clearer to write something like "However, it has been found that Q_{ST}–F_{ST} methods require many populations (O'Hara & Merila, 2005), and that various factors can generate …". Alternatively, just remove "that" from your sentence.
This is true, accordingly we changed the sentence from:
"However, Q_{ST}–F_{ST} methods have been found to require many populations (O'Hara & Merila, 2005), and that various factors can generate" to:
"However, **it has been found that** Q_{ST}–F_{ST} methods require many populations (O'Hara & Merila, 2005), and that various factors can generate"

38-40: “change in mean trait value accumulates linearly with time of divergence from a sister species, and also proportionally to the trait variance (Lande, 1980a; Turelli, 1984).” This is not quite correct, and can be misleading. The average change in mean trait value is zero and does not accumulate under pure drift (+mutation). The variance in the mean trait values across divergent lineages (isolated populations or species) does increase linearly with time, but each individual path is noisy and without any trend.

We completely agree that this sentence is inaccurate and misleading, thanks for pointing this out. This ambiguity also reflects your main comments about the variance in phenotype proportional to time of divergence (or nucleotide divergence) that we did not explicitly state. Accordingly, we corrected the sentence from:

“To disentangle selection from neutral evolution, trait variation can also be observed at a larger time scale. For example, change in mean trait value accumulates linearly with time of divergence from a sister species, and also proportionally to the trait variance (Lande, 1980a; Turelli, 1984).” to :

“To disentangle selection from neutral evolution, trait variation can also be observed at a larger time scale. For example, starting from the same ancestral population, divergent lineages accumulate phenotypic changes that will reach fixation in the population. These changes ultimately result in different mean trait values across lineages. Theoretically, the variance in mean trait value (between lineages) does increase linearly with time of divergence, and also proportionally to the trait variance at the population scale (Lande, 1980a; Turelli, 1984).”

44-49: this is repeated twice

Sorry for this error, the first erroneous occurrence is now deleted.

57: typo: “When trait variation is constrainED”

Completely agreed

63: “the optimal trait value is also evolving as a Brownian process”. Not sure that “evolving” is the right term here, as the optimum is likely to change for reasons not related to evolution, such as environmental change.

Completely agreed, we changed from:

“First, a trait under stabilizing selection for which the optimal trait value is also **evolving** as a Brownian process will not deviate from a Brownian process, and thus be wrongly classified as neutral” to

“First, a trait under stabilizing selection for which the optimal trait value is also **changing** as a Brownian process will not deviate from a Brownian process, and thus be wrongly classified as neutral”

76: please write for completeness: “between to within-species variation”

Completely agreed

79-80: “trait variation at the phylogenetic and population scales together with estimates of molecular divergence at both scales.” Perhaps better to write variation, since divergence applies only to the between-species scale

Completely agreed, we changed from:

“trait variation at the phylogenetic and population scales together with estimates of **molecular divergence** at both scales.” to

“trait variation at the phylogenetic and population scales together with estimates of **nucleotide sequence variations** at both scales.”

94: “the average effect of a mutation on the trait is…”. This doesn’t seem correct, as the average effect is assumed to be 0. What are your focusing on here is the variance of mutation effects per haploid genome (hence the multiplication by 2 in eq. (1) below to account for diploidy). You should also state that the expression for the mutational variance that follows assumes no bias in mutation effects, otherwise the expectation of the square would include an additional term.

Completely agreed, we have re-worded this whole section (see main comments), and this paragraph changed from:

“New mutations are generating trait variance and the average effect of a mutation on the trait is σ^2_M”. At the individual level, the mutational variance (V_M) is the rate at which new mutations contribute to the trait variance per generation.” to

“For a given trait, the genetic architecture is mainly defined by the number of loci encoding the trait (L) and the random additive effect of a mutation on the trait (a). For a diploid individual, the mutational variance (V_M) is the rate at which new mutations contribute to the trait variance per generation. As shown in Lande (1979, 1980), V_M is a function of the mutation rate per locus per generation (μ) and the genetic architecture of the trait as:

$$V_M=2\mu\times L\times E[a^2].$$

94: “the nucleotide diversity, pi, is measured as the number of mutations segregating in the population divided by the length of the region”. Not really: pi is the average number of differences between pairs of sequences drawn at random, which is also equal to the sum of expected heterozygosities 2p(1-p) over all loci (Tajima 1989). What you describe here (number of mutations segregating in the population) can only provide an estimator of 4 N_e mu (as per your eq. 4) after being scaled by a factor that depends on sample size, as shown by Watterson (1975): https://en.wikipedia.org/wiki/Watterson_estimator

Entirely true and many thanks for spotting it. We indeed used the pairwise diversity (π) as an estimator of θ and not the Watterson estimator (number of segregating sites divided by the (k-1)th harmonic number). Our mistake originates from the fact that in the Zoonomia dataset, we only have access to heterozygosity for a single individual per species, and the pairwise diversity and Watterson’s estimator are thus confounded. We apologize and changed the sentence from:

“...the nucleotide diversity, π, is measured as the number of mutations segregating in the population divided by the length of the region.” to

“...the nucleotide diversity, π, is the average number of differences between pairs of sequences drawn at random, which is also equal to the sum of expected heterozygosities over all loci (Tajima, 1989).”

95, eq (11-13): it would make more sense to cite the original reference by Kimura here rather than this historical review by McCandlish & Stoltzfus (2014)

We changed from:

"probability of fixation for each newly arisen mutations Pfix (McCandlish & Stoltzfus, 2014)" to

"probability of fixation for each newly arisen mutations Pfix **(Kimura, 1968)**"

An next from:

"the rate of substitution within a genomic region equals the rate at which new mutations arise per generation for the same genomic region (Kimura, 1968)" to

"the rate of substitution within a genomic region equals the rate at which new mutations arise per generation for the same genomic region **(McCandlish & Stoltzfus, 2014, review)**"

162: "The changes in log-μ and log-Ne along the lineages were both modeled by a geometric Brownian process". This is slightly misleading. I think what the authors mean is Brownian motion on the log scale (log-μ and log-Ne), leading to geometric Brownian motion on the linear scale (μ and N_e)

SYNC

Completely agreed, we changed from:

"The changes in log-μ and log-N_e along the lineages were both modeled by a geometric Brownian process" to

"The changes in μ and N_e along the lineages were both modeled by a Brownian process on the log scale (log-μ and log-N_e), leading to geometric Brownian motion on the linear scale (μ and N_e)."

Figure 2 legend, explanation of the simulation model: What about recombination? That is, how do you produce the offspring genotype from the parental ones? Should be explained here and/or in the main text.

Completely agreed, we changed from:

"The trait's genotypic value is encoded by L loci, with each locus contributing additively...." to

"The trait's genotypic value is encoded by L **independent loci (meaning no linkage)** with each locus contributing additively..."

264-266: "We thus argue that our method should be used to detect diversifying selection, but that it had low accuracy to detect stabilizing selection due to false positives." Any idea why this occurs? This would be interesting to the reader

Very good point. An assumption in our test is that the neutral phenotypic trait is evolving as a Brownian process and is unbounded. But if the phenotype is constrained between some bounds, for example due to its genetic architecture, phenotypic divergence should plateau at some point. We performed several simulations to test this effect, and added a section in the *Supplementary Materials* (section S4 and figure S2).

We discuss this possibility in more details now, we changed the text in the *Discussion* from:

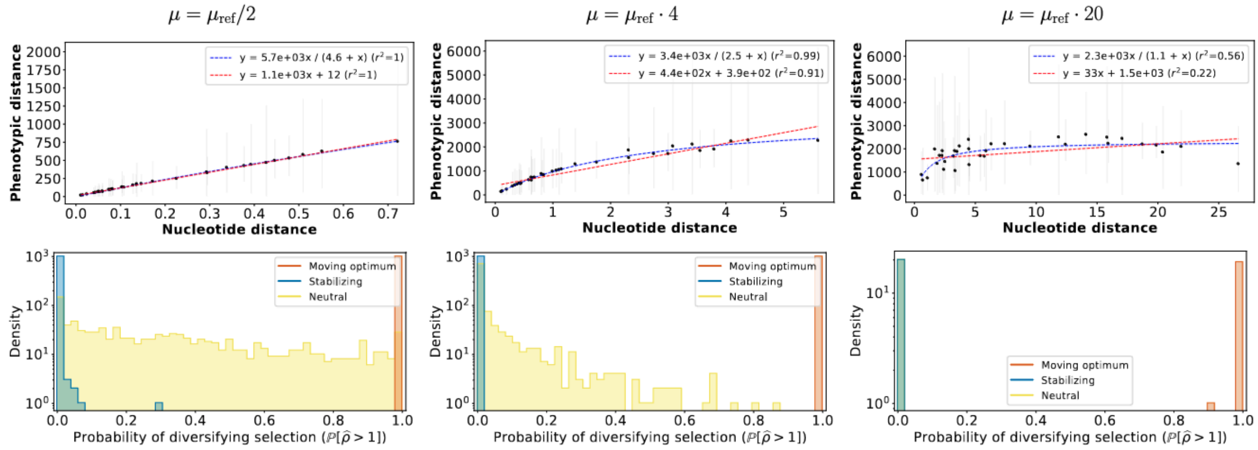
"However, our test detected a spurious signal of stabilizing selection (ρ<1) when we simulated the evolution of a neutral trait." to

"However, our test detected a spurious signal of stabilizing selection (ρ<1) when we simulated the evolution of a neutral trait. An assumption of our test is that the neutral phenotypic trait is evolving as a Brownian process and is, therefore, unbounded. However, the phenotype is encoded by a genetic architecture, and is thus ultimately bounded, even more so if the trait is encoded by a few loci (see table 2). At the macro-evolutionary scale, phenotypic divergence should plateau at some point, resulting in a reduced between-species trait variation. We argue that this effect can result in a spurious signal of stabilizing selection (ρ<1), especially for deeper phylogeny (figure S2 and section S4).

And we added section S4 and figure S2 in the *Supplementary Materials*:

The phenotype is encoded by a genetic architecture and is thus ultimately bounded. At the macro-evolutionary scale, phenotypic divergence should plateau at some point, ultimately resulting in a reduced σ_B^2.

Figure S2: Phenotypic saturation.



- Left column: 1,000 simulations with low divergence between species (half that of mammals).
- Middle column: 1,000 simulations with high divergence between species (4 times that of mammals).
- Right column: 200 simulations with very high divergence between species (20 times that of mammals).
- Top row: Simulation of neutral trait, phenotypic divergence between species as a function of the nucleotide divergence. Phenotypic divergence is computed between two species as the covariance between the trait (cov($\overline{P}_i$, $\overline{P}_j$)), and the nucleotide divergence is computed as the number of substitutions per site shared between the two species ($d_{i,j}$). Each point is a pair of species, and the bounds of the intervals in gray are the 2.5% and 97.5% quantiles across the replicates. The blue line is the linear regression and the red line is the saturation model (y = α · x/(β + x)).
- Bottom row: Traits simulated under stabilizing selection (blue), under a neutral evolution (yellow), and under a moving optimum (red). Histogram of probabilities of ρ being greater than 1.

Under a model of neutral trait evolution, when mutation rate increases, or equivalently the divergence between species increases, the phenotypic divergence between species saturates faster than the nucleotide divergence. This saturation effect can result in a spurious signal of stabilizing selection (ρ < 1) for deeper phylogeny when the trait is evolving neutrally.”

273: typo “which is a clear an improvement”

Sorry, we removed the additional “an”

591: “independent contrast C_j defined as change in trait along the branch normalized by sqrt(d_j)”. This square root seems to relate to my main comment above on how the variance in mean phenotypes relates to substitution rates at the molecular level.

We hope that the answer to the main comment (l 260-330 above) allows us to better clarify how the changes in mean phenotype is related to substitution rates at the nucleotide level. As shown in equation 18, by taking i=j, we have Var(P_i)=d_i, meaning the variance in mean phenotypes is proportional to nucleotide divergence (*d*). Hence, by taking the square root, the difference in difference in mean phenotypes is proportional to $\sqrt{d}$.

Appendix 3.1 on data formatting: It would help the reader to state more explicitly that the values of molecular divergence to be used in the approach come from the branch length in the tree (if I understood well)..

Exactly, we changed from:

“A phylogenetic tree in newick format, with branch lengths in number of substitutions per site (neutral markers).” to

“A phylogenetic tree in newick format, with branch lengths in number of substitutions per site (neutral markers)**, from which the values of nucleotide divergence (d) used in denominator of eq. 18 are used.”**

Reviewer II:

Comments:

Comments to the Author

I've read and reviewed the paper entitled "Detecting diversifying selection for a trait from within and between-species genotypes and phenotypes". This paper presents an exciting perspective for analyzing macroevolutionary patterns of evolution through the lens of neutral theory. I find the unification of population and quantitative genetics very promising, and I have no major reservations about the publication of this manuscript. I myself long thought that using tree branch lengths proportional to genetic neutral divergence would be a better approximation for the neutral expectation than absolute times. Still, I was never capable of articulating the argument as well as the authors could. For that, I commend the authors.

That being said, I find the text unnecessarily complicated at some points, which really makes it difficult to follow the analytical argument being made. Below I list these minor concerns. I list the comments on the sections (numbers) they are associated to on the borders of the text, which do not seem to be related to lines necessarily.

Thanks a lot for these nice words. In this revision we focused our efforts on readability of the manuscript. Moreover, in the *Discussion* we now also develop on the consequences and the potential perspective of modeling trait evolution on tree branch lengths proportional to genetic neutral divergence instead of absolute times.

94- I think this portion of the text is both the most important one and the most difficult to follow. The definition of σ^2_M takes a bit to fully understand, and I am still not entirely sure it is properly defined. According to the text, σ^2_M is the "average effect [generation of trait variance] of a mutation on the trait". However, that definition only amounts to $E[a^2]$. σ^2_M, as defined here, is the average variance mutational effect per loci ($E[a^2]$) times the number of loci encoding a trait (L), which would be the, maybe, the maximum mutational variance added if all loci had a mutation? It is unclear to me what that means if anything. To add to the confusion, equation 1 defines V_m as the mutational variance, which, to my knowledge, is usually referred to as σ^2_M. Since the authors do not cite anyone in the definition of σ^2_M, I assume this variable is a novel definition. If that is the case I would suggest the authors to 1- evaluate if the definition of this σ^2_M is necessary (as far as I understand from the algebra, it is not necessary), 2- if it is, better define it and provide an intuition for what that means (e.g. maximum mutation potential, mutational variance divided by 2 times the mutation rate- and what would that entail), 3- if is not, re-work the algebra to make σ^2_W and σ^2_B as a function of $L*E[a^2]$ (but see a similar issue with σ^2_W and σ^2_B below).

We agree and concur that σ^2_M is neither well defined in the text, nor necessary. It is solely used as a shortcut to avoid writing $L\times E[a^2]$ every time. We agree this is doing us a disservice and is confusing. As a result, we have re-written this whole section to remove σ^2_M and kept the V_M until they canceled out on eq. 22. We have updated all equations 1-24 and definitions to reflect these changes.

Moreover, we also have made several changes to this section related to the main comments of reviewer I:

- We now use “μ” for the mutation rate per generation of loci underlying the quantitative trait, and “*u*” for the mutation rate per generation of nucleotides.
- We kept “μ” and “*u*” in equations until the definition of the neutrality index, where they all cancel out in the ratio of σ^2_B over σ^2_W.
- We changed from removing the effect of “N_e and μ” to “N_e and t (number of generations)”

Still in this section, equations 5-7 require maybe a bit more of the intuitive meaning for equating equation 6 to σ^2_M. This depends if σ^2_M has any intuitive meaning that can be used to articulate this part of the text. This would help readers understand if σ^2_M is useful by itself (which could allow the use of comparative data to estimate it, for example), or is just one stepping-stone for the test described by the authors.

σ^2_M is indeed used as a stepping stone, and now it has been entirely removed from the manuscript to avoid confusion.

I feel that the sentence leading to the equation 8 could also reinforce how one can estimate σ^2_M using the phenotypic variance and heritabilities.

Completely agreed, we have expanded the equations to clearly state that phenotypic variance and heritabilities can be used instead of additive genetic variance in equations 8-10.

Lastly, for the definition of σ^2_W and σ^2_B (in the next section) can be confusing, especially because both symbols are routinely used to refer to within and between group variances, respectively. In Figure 1 the authors use the hat, which at first I believed it referred to a normalization (as is usually the case in algebra), but from the first paragraph of the discussion it made me believe that the hat is only there to denote an estimate. If that is indeed the case, care should be given to differentiating the authors' σ^2 statistics from their standard, non-normalized versions.

This is true that both symbols (σ^2_W and σ^2_B) are routinely used to refer to within and between group variances. We argue that although it is not equivalent, it has some relationship with our study of trait evolution, since our statistical question can be framed as: Is the variance of means equal to the mean of variances? If not, what causes this deviation? The difficulty in our case is that points (the individuals) are not independent samples, and thus we have to normalize by genetic distance. We have now introduced these symbols at the beginning of the section *Neutrality index for a quantitative trait* such as to avoid ambiguity with our notation, changing the text from:

“Prior to developing our neutrality index, we review theoretical expectations for variations of quantitative traits and genomic sequences under neutral evolution for both within- and between-species variation.” to:

“While observing trait variations across individuals of several species, we ask if the variation within species compared to variation between species is compatible with neutral evolution or not. In statistical terms, this can also be framed as: Is the variance of means equal to the mean of variances? The difficulty in such a study is that individuals are not independent samples, but are from species that diverged at different times, and each species has a putatively different population size. By reviewing theoretical expectations and leveraging nucleotide sequence variations, the goal of this section is thus to obtain normalized trait variation between and within species that are equal if the trait is neutral. We denote such normalized trait variations as respectively σ^2_W for within species and as σ^2_B for between species.

This is completely true that we never had defined that the hat (^) is used a standard symbol for estimates, we added the sentence for this definition at the beginning of the section *Estimation*:

"We denote \hat(ρ), \hat(σ)^2_W and \hat(σ)^2_B the point estimates of respectively ρ, σ^2_W and σ^2_B."

108-109 I appreciate this section quite a bit, but I think it would be good to reinforce to the reader that hat(σ^2_B) can be estimated through any phylogenetic comparative methods that allow the estimation of phenotypic rates of evolution. Specifically, I suggest emphasizing that equation 8 is simply the rate of evolution calculated on a d-scaled tree divided by 4.

Completely agreed and it also reflects a comment by reviewer I. We therefore added the following paragraph to emphasize that σ^2_B can be estimated through any phylogenetic comparative methods that allow the estimation of phenotypic rates of evolution, as along the rate of evolution calculated on a d-scaled tree divided by 4 :

"For each species with data available, σ^2_W as defined in eq. 5 can be seen as a replicate sample. Thus, \hat(σ)^2_W can be obtained by averaging out across all the sampled species. On the other hand, σ^2_B such as as defined in eq. 18 only refers to a pair of species, and thus must be generalized to account for different species divergence, as is done in the comparative framework (Felsenstein, 1985; O'Meara et al., 2006). Generally, \hat(σ)^2_B can thus be seen as an estimate of the rate of the evolution of the quantitative trait along a phylogeny, when the tree is measured in units of 4*d* (*d* is the nucleotide divergence). As such, any phylogenetic comparative methods that allow the estimation of phenotypic rates of evolution on a tree scaled by 4*d*, instead of time as is usually the case, can be used to estimate σ^2_B."

112-113 There are plenty of multivariate maximum likelihood methods that can accomplish the same.

Completely agreed, we made the mistake of introducing the multivariate and the bayesian extension at the same time while it is true that they are independent. To avoid this confusion, we have restructured the section to first mention the multivariate extension in its own section. Then in another section, we introduce the bayesian extension.

We changed the sentence from:

"The Bayesian framework allows obtaining the posterior distribution of the neutrality index (ρ) for a given trait. Even though ρ is estimated independently for each trait of interest in the maximum likelihood framework (previous section), here we generalize to K traits co-varying along the phylogenetic tree using the *BayesCode* software (Latrille et al., 2021). ..." to

"**Multivariate Brownian process**

In the previous section, ρ is estimated independently for each trait of interest. Here we generalize to K traits co-varying along the phylogenetic tree. Trait variation along the phylogenetic tree is modeled as a K-dimensional Brownian process *B* (1×K) starting at the root and branching along the tree topology [...]

Bayesian estimate

The Bayesian framework allows obtaining the posterior distribution of neutrality index (ρ) for traits of interest. We used the *BayesCode* software to model K-dimensional Brownian processes along a phylogenetic tree (Latrille et al., 2021). [...]"

L118-119 Entries of the diagonal of Σ is 4 times the hat(σ^2_{W}). I guess it is true that they “refer to hat(σ^2_{W})” but that is rather vague.

We agree that this is rather vague, we added a sentence to avoid confusion, changing the text from:

“The off-diagonal elements of Σ are the covariance between traits, and the diagonal elements are the variance of each trait, thus corresponding to σ^2_{B}.” to

“The branch lengths of the tree used to model the Brownian process is measured in units of 4*d* (*d* is the nucleotide divergence). The off-diagonal elements of Σ are the covariance between traits, and the diagonal elements are the variance of each trait when measured in 4*d* units, and thus equate to σ^2_{B}.”

L164-165 I could not fully understand this justification.

We clarified the meaning of the standard deviation from root to leaves, we changed the text from:

“B(0, σ_{Ne}=0.0086), which led to a standard deviation of $0.0086 \cdot \sqrt{13, 500} = 1.0$ in log-space from root to leaves.” to

“B(0, σ_{Ne}=0.0086), which, if counted across 13,500 generations, leads to a standard deviation of $0.0086 \cdot \sqrt{13, 500} = 1.0$. In other words, the deviation in log-N_e between the extant species and the root is 1.0.”

L172-184 This part has some confusing parts. At the beginning the authors state that drift is added when individuals were sampled with a probability proportional to fitness, which is not consistent with any definition of drift I know. But at the end, they state that neutral evolution was achieved by setting fitness to be constant, which would make the first phrase true, but only in that instance. Was the first phrase meant to say simply that parents were sampled according to their fitness?

Yes completely the first phrase meant to say simply that parents were randomly sampled with a weight proportional to their fitness. We changed the sentence from:

“A random genetic drift was introduced by resampling individuals at each generation, with each parent having a probability of being sampled that was proportional to its fitness (W).” to

“At each generation, parents were randomly sampled with a weight proportional to their fitness (W).”

L196-198- This portion is a bit hard to follow. Where was the nucleotide diversity extracted from? Genereux et al. 2020 or Wilder et al 2023?

We completely agree that the references are ambiguous. The mammalian genomic data are gathered from the Zoonomia project (https://zoonomiaproject.org/), presented in Genereux et al. (2020). More specifically, nucleotide divergence is estimated on a set of neutral markers in Foley et al. (2023), and with nucleotide diversity measured as heterozygosity in Wilder et al. (2023). We thus changed the paragraph from:

“The mammalian nucleotide diversity was obtained from the Zoonomia project (Genereux et al., 2020), with nucleotide divergence obtained on a set of neutral markers in Foley et al. (2023), and with nucleotide diversity measured as heterozygosity in Wilder et al. (2023).” to

“The mammalian genomic data are gathered from the Zoonomia project (Genereux et al., 2020). More specifically, nucleotide divergence is estimated on a set of neutral markers in Foley *et al.* (2023), and with nucleotide diversity measured as heterozygosity in Wilder *et al.* (2023).”

L202- What are the implications of using full genomes instead of only neutral markers?

This creates a bias, and the evidence for $\rho>1$ does then not necessarily imply that the trait is evolving under diversifying selection since non-neutral markers for divergence can lead to a spurious $\rho>1$ (see table 2). We clarify this point and changed the text from:

"However, the primate nucleotide divergence was not obtained on a set of neutral markers as for the mammalian dataset, but across the whole genome." to

"However, the primate nucleotide divergence was not obtained on a set of neutral markers as for the mammalian dataset, but across the whole genome. As such, the evidence for $\rho>1$ does not necessarily imply that the trait is evolving under diversifying selection since non-neutral markers included in the estimate of divergence can lead to a spurious $\rho>1$ (see table 2)."

L205-220- This part feels a bit redundant and unnecessary. Maybe the only necessary part refers to table 2, but that could be easily reinforced in the methods section and this whole part can be excluded.

While we completely agree that this part is redundant if the reader has been through the *Materials and Methods*, we however kindly disagree that it should be excluded. We sought to make the paper as readable as possible to readers even though some might not have read the *Materials and Methods* and jump straight to *Results* and *Discussion*. As such we believe that this paragraph allows the reader to start from this section without going through the *Materials and Methods*. Moreover, since the formalization of the neutrality index is in itself a result, we argue that it should be included in the *Results* section.

Table 2- Maybe add a couple of additional lines with the possibility that some of the divergence among species is caused by plasticity.

As stated in Rohlfs et al. (2014) and Rohlfs & Nielsen (2015) in the context of gene expression evolution, plastic traits are responding to individual environmental conditions, thus resulting in traits with conserved mean trait values across species, but high variance within species due to the plasticity (individuals have different phenotypes). From a quantitative genetics point of view, phenotypic plasticity results in increased phenotypic variance (V_P) due to an additional term (V_{ExG}). As such it is similar in our study as using the broad-sense heritability (H^2) instead of narrow-sense heritability (h^2) resulting in an underestimation of ρ since $h^2 \leq H^2$, ultimately leading to overestimated σ^2_W. We thus added the following line:

Broken assumption	Consequences	σ^2_W	σ^2_B	Test $\rho>1$	Test $\rho<1$
Phenotypic plasticity	Trait responding to individual environmental conditions	Overestimated	-	Conservative	Invalid
...	...	...	...	...	...

L278-285- The authors don't cite any work using the Lande's generalized genetic distance (LGGD) method (Schroeder et al. 2017, Machado et al 2022, Machado et al. 2023) and related methods (Lynch 1990, Lemos et al 2001, 20XX, Weaver et al 2007, 2015, Porto et al. 2015). Specifically, for the LGGD, the distance among species is also normalized by the additive genetic variation. In this sense, they are similar to Rohlfs et al., (2014) and Rohlfs & Nielsen (2015). The LGGD was successful in identifying specific instances of divergent selection (Schroeder et al. 2017, Machado et al 2022) and near-drift (Machado et al. 2023).

Yes, thanks a lot for pointing this out. Moreover we agree that this paragraph mentioning Rohlfs et al., (2014) and Rohlfs & Nielsen (2015) is well suited to also mention LGGD methods and their differences. As such we have extended the text from:

"Our diversity index has the advantage to discriminate the alternative model of diversifying selection from the neutral case by comparing within- and between-species variation which are normalized to remove confounding factors. Our approach is not the first one to normalize between-species variation to detect selection, but this was done by using within-species variations (Rohlfs et al., 2014; Rohlfs & Nielsen, 2015) and not estimates of neutral molecular divergence as done in our study. These studies have further compared their statistic across a pool of traits, which allowed them to identify outlier traits putatively under diversifying selection but without testing for selection on a single trait at a time (Gillard et al., 2021; Rohlfs & Nielsen, 2015). Instead, our procedure can be applied to a single trait, estimating the neutrality index and giving a statistical test for departures from the null model of neutral evolution for a single test. Our diversity index opens new avenues to revisit these studies and better test for the selective regime affecting the quantitative traits, assuming we have access to genomic datasets to estimate nucleotide divergence and polymorphism." to

"Instead, our diversity index has the advantage to discriminate the alternative model of diversifying selection from the neutral case by comparing within- and between-species variation while correctly normalizing them using nucleotide markers. Our approach is not the first one coupling between-species and within-species variations, and those approaches employ different strategies to detect selection. First, one empirical strategy is to compare the ratio of between to within variation across a pool of traits, which allow to identify outlier traits putatively under diversifying selection (Rohlfs et al., 2014). However, this method does not formally allow testing for diversifying selection, and requires many traits such as expression level data to seek outliers genes (Gillard et al., 2021; Rohlfs & Nielsen, 2015). Second, other methods leverage Lande's generalized genetic distance (LGGD), which relate the ratio of between to within variations to population-genetic parameters (Lynch, 1990; Lemos et al., 2001, 2005, Weaver et al., 2007; Porto et al., 2015). Specifically, by leveraging estimates of effective population size (N_e) and the number of generations between species, or alternatively by assuming their constancy, these methods can test for departures from the null model of neutral evolution for a single trait. Such methods have been successful in identifying specific instances of diversifying selection (Schroeder et al., 2017, Machado et al., 2022) and near-drift (Machado et al., 2023). However, N_e and the number of generations are complex parameters to correctly infer and is usually done for a pair of species or only a few species, and ultimately requires large genomic datasets and heavy statistical methods (Wilder et al., 2023). Instead, our diversity index opens new avenues to revisit these studies testing for the selective regime affecting the quantitative traits, by formally incorporating nucleotide divergence and polymorphism, bypassing estimation of N_e, generation time and calibration of ancestral node ages (Machado et al., 2023)."

Since we now mention normalization by effective population size (N_e), generation time and calibration of ancestral node ages in this paragraph, we also updated the next paragraph to accommodate those changes, from:

“The main novelty of our study was to use the nucleotide divergence and polymorphism to normalize trait variation between and within species. In the context of within species variation, Q_{ST}–F_{ST} tests have been developed to compare trait and sequence across several populations to test for selection (Leinonen et al., 2013; Martin et al., 2008). Our neutrality index also used the genetic sequences from which nucleotide divergence and polymorphism are estimated. Although the sequences should be neutrally evolving, they do not have to be necessarily linked to the quantitative trait under study. Nucleotide variation allows normalizing for diversity driven by confounding factors such as population sizes (N_e), mutation rates (μ) and generation time (Hansen & Martins, 1996; Harmon, 2018). Thus our test avoids the estimation of the parameters, which are complex to correctly infer, and it also bypasses the estimation of divergence time, which was necessary in previous approaches (Walsh & Lynch, 2018). But importantly, by normalizing with sequence variation, we also showed using simulated data that our test was not sensitive to the assumption that N_e, μ and generation time were constant across the phylogenetic tree, an unmet assumption empirically (Bergeron et al., 2023; Wilder et al., 2023). Indeed, under the neutral case of evolution, changes in N_e, μ and generation time impacted similarly trait and sequence variation. The normalization by nucleotide divergence and polymorphism automatically absorbed long-term and short-term changes in N_e, μ and generation time, which canceled out in the neutrality index.” to

“As such, the main novelty of our study was to use the nucleotide divergence and polymorphism to normalize trait variation between and within species. In this context, our test bears many similarities to Q_{ST}–F_{ST} tests that have been developed to test for selection of a trait across several populations while also leveraging sequence variation (Martin et al., 2008; Ovaskainen et al., 2011; Leinonen et al., 2013). Our method can be seen as an extension at the phylogenetic scale, where although the sequences used should be neutrally evolving, they can be obtained from different sampled individuals than for the trait. Importantly, by normalizing with sequence variation, we also showed using simulated data that our test was not sensitive to the assumption that N_e and mutation rates were constant across the phylogenetic tree, an unmet assumption empirically (Bergeron et al., 2023; Wilder et al., 2023). Indeed, under the neutral case of evolution, the normalization by nucleotide divergence and polymorphism automatically absorbed long-term and short-term changes in N_e, generation time and mutation rates, which canceled out in the neutrality index.”

L295-297 Machado et al. 2023 suggest a method that relies on estimates of N_e and generation time, which are weaknesses the proposed methodology overcomes.

We completely agree. As mentioned in the previous comment, we have reworked two paragraphs and we now illustrate that the proposed method formally incorporates nucleotide divergence and polymorphism to normalize trait variations instead of estimating N_e and generation time (Machado et al., 2023).

L351-356 The authors can expand on the presence of a signal of diversifying selection on macroevolutionary data. The consensus on macroevolutionary studies is that empirical rates of evolution calculated on phylogenetic trees and the fossil record are far inferior to the expected under drift (Lynch 1990, Uyeda et al. 2008). I wonder if the authors' finding of diversifying selection on body and brain mass could be an argument against that interpretation. Those interpretations, however, assume constancy of N_e and μ and generation time, something that this method does not.

Very good point and thanks for this burning question. We have to admit that right now we have no clear evidence on what could cause this discrepancy between their and our result. One possible line of arguments is that rates of nucleotide evolution also show a tendency for slowing down on a longer timescale (Rolland et al., 2023). As such our normalization by nucleotide divergence could absorb this slowing rate of evolution. We thus have included this comment as an open question requiring further studies in the *Discussion*, and have changed the last paragraph from:

“However, such datasets will become more and more accessible and we showed the applicability of our method by applying it to the illustrative example of mammals brain and body mass.” to

“However, such datasets will become more and more accessible and we showed the applicability of our method by applying it to the illustrative example of mammals' brain and body mass, showing signals of diversifying selection. The consensus on macro-evolutionary studies, assuming constancy of N_e, generation time and mutation rates, is that empirical rates of evolution calculated on phylogenetic trees and the fossil record are far inferior to the expected under drift (Lynch, 1990, Uyeda et al., 2008). Our finding of diversifying selection on body and brain mass could be seen as an argument against that interpretation. In fact, rates of nucleotide evolution also show a tendency for slowing down on a longer timescale (Rolland et al., 2023). One possible interpretation is that normalization by nucleotide divergence could absorb this observed slowing rate of evolution. Altogether, further empirical and theoretical studies are required to disentangle this discrepancy between these different interpretations.”