



Detecting diversifying selection for a trait from within and between-species genotypes and phenotypes

T. Latrille¹ , M. Bastian², T. Gaboriau¹, N. Salamin¹ 

¹Department of Computational Biology, Université de Lausanne, Lausanne, Switzerland

²Laboratoire de Biométrie et Biologie Evolutive, UMR5558, Université Lyon 1, Villeurbanne, France

Corresponding author: T. Latrille, Department of Computational Biology, Université de Lausanne, Lausanne, Switzerland.

Email: thibault.latrille@ens-lyon.org

Abstract

To quantify selection acting on a trait, methods have been developed using either within or between-species variation. However, methods using within-species variation do not integrate the changes at the macro-evolutionary scale. Conversely, current methods using between-species variation usually discard within-species variation, thus not accounting for processes at the micro-evolutionary scale. The main goal of this study is to define a neutrality index for a quantitative trait, by combining within- and between-species variation. This neutrality index integrates nucleotide polymorphism and divergence for normalizing trait variation. As such, it does not require estimation of population size nor of time of speciation for normalization. Our index can be used to seek deviation from the null model of neutral evolution, and test for diversifying selection. Applied to brain mass and body mass at the mammalian scale, we show that brain mass is under diversifying selection. Finally, we show that our test is not sensitive to the assumption that population sizes, mutation rates and generation time are constant across the phylogeny, and automatically adjust for it.

Keywords: quantitative genetics; trait evolution; selection; phylogenetics; population genetics

Introduction

Determining whether a trait is under a particular regime of selection has been a long-standing goal in evolutionary biology. Fundamentally, distinguishing neutral evolution from selection requires determining which selective regime is supported by the observed variation of traits or sequences. The variation of phenotypes (traits) and genotypes (sequences) can be observed at different scales, across different development stages at the individual level, across different individuals and populations at the species level, and finally across different species at the phylogenetic level. All these systems require different assumptions and methodologies, and the endeavor to determine the selective regime for a given trait has thus incorporated theories, methods, and developments across various fields of evolutionary biology such as quantitative genetics, population genetics, phylogenetics and comparative genomics (Lynch & Walsh, 1998; Walsh & Lynch, 2018).

Leveraging individual variations within the same species, genome-wide association studies (GWAS) in humans have shown that traits are mostly polygenic (many loci associated with a given trait) and under stabilizing selection, while the loci affecting those traits are mostly pleiotropic (many traits associated with a given locus) with additive effects (Sella & Barton, 2019; Simons *et al.*, 2018). Given this genetic architecture of traits, from two diverging populations, it is possible to distinguish which traits have evolved under natural selection in controlled experimental settings, by performing genetic cross between individuals (Fraser, 2020). 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 (Crnokrak & Merilä, 2001; Leinonen *et al.*, 2008). Q_{ST} higher than F_{ST} is interpreted as a signature of diversifying selection due to adaptation to 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 (Edelaar *et al.*, 2011; Ovaskainen *et al.*, 2011; Pujol *et al.*, 2008), 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. 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 (Felsenstein, 1988; Lande, 1980a; Turelli, 1984). Empirically, this effect can be observed for genes with larger within-species variation in gene expression level, which exhibits a faster accumulation of divergence in mean expres-

Received July 10, 2024; revised June 14, 2024; accepted July 10, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the European Society of Evolutionary Biology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

sion level (Khaltovich *et al.*, 2004). As an analogy, in the context of protein-coding DNA sequences, leveraging within species variation and divergence to a sister species is the crux of the McDonald and Kreitman (1991) test. In such a test, inflation of divergence to the sister species is compared to polymorphism within species, while neutral markers (usually synonymous sites) are used to determine the neutral expectation and thus are used for normalization. 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).

Alternatively, by accounting for the underlying relationships between several species, the selective regime for a quantitative trait can also be tested at the phylogenetic scale (Felsenstein, 1985). Under neutral evolution, the change in mean trait value along a given branch of the tree is normally distributed, with a variance proportional to divergence time (Felsenstein, 1985, 1988; Hansen & Martins, 1996). As a result, the mean trait value can be modeled as a Brownian process branching at every node of the tree (Harmon, 2018; Hansen & Martins, 1996). Reconstructing the trait variation along the whole phylogeny as a Brownian process can thus constitute a null model of neutral trait evolution. Deviations from the assumptions of the Brownian process are however well known. When trait variation is constrained because of optimum mean trait values across or between species, the pattern of evolution can be modeled by the Ornstein–Uhlenbeck processes, which is often interpreted as a signature of stabilizing selection (Catalán *et al.*, 2019; Hansen, 1997). Alternatively, a trend in the Brownian process (the tendency of a trait to evolve in a certain direction without fixed optimum) is interpreted as a signature of directional selection at the phylogenetic scale (Silvestro *et al.*, 2019). However, studies have shown that such comparative approaches are subject to different biases (Harmon, 2018). 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 (Hansen & Martins, 1996). In other words, the better fit of a Brownian process does not necessarily constitute proof of the neutral model. Second, 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), and third, even for a trait evolving under a neutral regime, the Ornstein–Uhlenbeck process might sometimes be statistically preferred over a Brownian process due to sampling artifacts (Cooper *et al.*, 2016; Price *et al.*, 2022; Silvestro *et al.*, 2015). Those limitations, altogether with the use of mean trait estimates leaving out the variance in traits between individuals, easily generate misclassification of selection from methods at the phylogenetic scale.

At the frontier between micro and macro-evolution, comparative methods at the phylogenetic scale have acknowledged the importance of modeling within-species variation together with changes in mean trait value to either describe measurement errors (Hansen & Bartoszek, 2012; Lynch, 1991), incorporate values for individuals (Felsenstein, 2008) or to scale the rate of change in mean trait value (Gaboriau *et al.*, 2020, 2023; Kostikova *et al.*, 2016). Across many species, within-species variation has also been used to infer diversifying selection by estimating the ratio of between to within species variation of many traits and test for deviation from the

average ratio across traits (Rohlf & Nielsen, 2015; Rohlf *et al.*, 2014). Here, our goal was again to use both variances between and within species to determine the selective regime of a quantitative trait. We build a novel framework that integrates trait variation at the phylogenetic and population scales together with estimates of nucleotide sequence variations at both scales. It allowed us to define an expected ratio of normalized variance between and within species while setting the threshold of this ratio for neutral, stabilizing, and diversifying selection. The ratio that we propose can be considered as a neutrality index for any quantitative trait (Lynch, 1990), while articulating trait and nucleotide variation within and between species. Importantly, our neutrality index also leverages nucleotide divergence and polymorphism to normalize trait variation at both scales, such that it does not require estimating population size (within-species) or speciation time (between species). 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 (Ovaskainen *et al.*, 2011; Pujol *et al.*, 2008), our study can be seen as their macro-evolutionary analog to account for phylogenetic relationships between species. From the field of phylogenetics, our study can be seen as an alternative to the EVE model (Rohlf & Nielsen, 2015; Rohlf *et al.*, 2014) for a single trait, where we set a threshold for neutral evolution by leveraging species nucleotide polymorphism and divergence.

Materials and methods

Neutrality index for a quantitative trait

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. 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. Here we denote these normalized trait variations as respectively σ_W^2 for within species and as σ_B^2 for between species.

Within-species trait variations

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, 1980b), V_M is a function of the mutation rate per locus per generation (μ) and the genetic architecture of the trait as

$$V_M = 2\mu \cdot L \cdot E[a^2]. \quad (1)$$

While in an infinitesimal model mutations supply new genetic variants, random genetic drift depletes standing variation (Barton *et al.*, 2017; Sella & Barton, 2019; Turelli, 2017). For a neutral trait at equilibrium between mutation and drift (Lynch *et al.*, 1998), the additive genetic variance in

a species (V_A) is a function of the mutational variance (V_M) and the effective number of individuals in the population (N_e):

$$V_A = 2N_e \cdot V_M, \quad (2)$$

$$= 4N_e \cdot \mu \cdot L \cdot \mathbb{E} [a^2] \text{ from Equation 1.} \quad (3)$$

For any neutral genomic region of interest, 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 nucleotide sites (Tajima, 1989). Any segregating mutations will eventually reach fixation or extinction due to random genetic drift and π is also at a balance between mutations and drift. As shown in Tajima (1989), π is a function of the mutation rate (u , per nucleotide site per generation) and the effective population size (N_e):

$$\pi = 4N_e \cdot u. \quad (4)$$

To remove the effect of N_e , we define σ_W^2 as the ratio of additive genetic variance of the trait (V_A) over π of any neutral genomic region of interest. After simplification, σ_W^2 is then solely a function of the underlying genetic architecture as

$$\sigma_W^2 \stackrel{\text{def}}{=} \frac{V_A}{\pi}, \quad (5)$$

$$= \frac{4N_e \cdot \mu \cdot L \cdot \mathbb{E} [a^2]}{4N_e \cdot u} \text{ from Equations 1 and 4,} \quad (6)$$

$$= \frac{\mu \cdot L \cdot \mathbb{E} [a^2]}{u}. \quad (7)$$

If V_A is not empirically accessible, it can be related to the observed phenotypic variance (V_P), multiplied by narrow-sense heritability of the trait (h^2), as (Hill et al., 2008)

$$V_A = h^2 \cdot V_P. \quad (8)$$

Which leads to σ_W^2 being a function of V_P and h^2 instead of V_A as

$$\sigma_W^2 = \frac{h^2 \cdot V_P}{\pi} \text{ from definition Equations 5 and 8,} \quad (9)$$

$$= \frac{\mu \cdot L \cdot \mathbb{E} [a^2]}{u} \text{ from Equation 7} \quad (10)$$

Between-species trait variations

For a given species i , we denote by \bar{P}_i the mean value of the trait across the individuals. If the trait is neutral and encoded by many loci as assumed by the infinitesimal model, \bar{P}_i evolves as a Brownian process (Felsenstein, 1985; Hansen & Martins, 1996). Given a phylogenetic tree, for a pair of species i and j from this tree, we denote as $t_{i,j}$ the number of generations between the root of the tree and the most recent common ancestor of taxa i and j . Then, the covariance between \bar{P}_i and \bar{P}_j depends on $t_{i,j}$ as given by Hansen & Martins (1996)

$$\text{cov}(\bar{P}_i, \bar{P}_j) = \frac{V_A}{N_e} \cdot t_{i,j} \quad (11)$$

$$= 4t_{i,j} \cdot \mu \cdot L \cdot \mathbb{E} [a^2], \text{ from Equation 3.} \quad (12)$$

Moreover, for any genomic region under neutral evolution, some mutations will eventually reach fixation due to random

genetic drift, resulting in a substitution of a nucleotide at the species level. The probability of fixation (\mathbb{P}_{fix}) of a neutral mutation is $1/2N_e$ (Kimura, 1962). We can derive the substitution rate (q , per nucleotide site per generation) as the number of newly arisen mutations ($2N_e \cdot u$) multiplied by the probability of fixation for each newly arisen mutations \mathbb{P}_{fix} (Kimura, 1968), giving:

$$q = 2N_e \cdot u \cdot \mathbb{P}_{\text{fix}}, \quad (13)$$

$$= 2N_e \cdot u \cdot \frac{1}{2N_e}, \quad (14)$$

$$= u. \quad (15)$$

That is, if mutations are neutral, the rate of substitution per generation within a genomic region equals the rate at which new mutations arise per generation for the same genomic region, reviewed by McCandlish and Stoltzfus (2014).

Next, we denote $d_{i,j}$ as 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 expected number of substitutions per nucleotide site during the $t_{i,j}$ generations. Assuming that no multiple substitutions occurred at the same site, $d_{i,j}$ is the number of generations ($t_{i,j}$) multiplied by the nucleotide substitution rate per generation (q):

$$d_{i,j} = t_{i,j} \cdot q \quad (16)$$

$$= t_{i,j} \cdot u \text{ from Equation 15.} \quad (17)$$

To remove the effect of the number of generations ($t_{i,j}$) first, and to also equate to σ_W^2 (Equation 7), we define σ_B^2 as the covariance in the mean trait value ($\text{cov}(\bar{P}_i, \bar{P}_j)$) normalized by 4 times the nucleotide divergence of any neutral genomic region ($4d_{i,j}$). After simplification, σ_B^2 is also solely a function of the underlying genetic architecture as

$$\sigma_B^2 \stackrel{\text{def}}{=} \frac{\text{cov}(\bar{P}_i, \bar{P}_j)}{4d_{i,j}}, \quad (18)$$

$$= \frac{4t_{i,j} \cdot \mu \cdot L \cdot \mathbb{E} [a^2]}{4t_{i,j} \cdot u} \text{ from Equations 12 and 17,} \quad (19)$$

$$= \frac{\mu \cdot L \cdot \mathbb{E} [a^2]}{u}. \quad (20)$$

In Equation 20, we show that the covariance in mean trait value between a pair of species ($\text{cov}(\bar{P}_i, \bar{P}_j)$) does increase linearly with shared nucleotide divergence ($d_{i,j}$), if the trait and sequences are neutrally evolving and the genetic architecture of the trait has not changed. Importantly, since the number of generations is the ratio of time divided by generation time (average time between two consecutive generations), removing the effect of the number of generations in Equation 20 also removes the effect of both time and generation time.

Neutrality index

The variability between either individuals or species can be obtained for both quantitative traits and genomic sequences. At the population level, the variability of the trait between individuals can be combined with the nucleotide diversity of any neutrally evolving genomic region to obtain σ_W^2 . At the phylogenetic level, the variability of the mean trait value between species can be combined with the nucleotide divergence of any neutrally evolving genomic region to obtain σ_B^2 .

If the trait is neutrally evolving and the genetic architecture of the trait has not changed along the phylogenetic tree, we thus have

$$\frac{\sigma_B^2}{\sigma_W^2} = \frac{\text{cov}(\bar{P}_i, \bar{P}_j)}{4d_{i,j}} \cdot \frac{\pi}{h^2 \cdot V_P} \text{ by definition and} \\ \text{Equations 9 and 18,} \quad (21)$$

$$= \frac{\mu \cdot L \cdot \mathbb{E}[a^2]}{u} \cdot \frac{u}{\mu \cdot L \cdot \mathbb{E}[a^2]} \text{ from Equations 10} \\ \text{and 20,} \quad (22)$$

$$= 1. \quad (23)$$

We define a neutrality index ρ as

$$\rho \stackrel{\text{def}}{=} \frac{\sigma_B^2}{\sigma_W^2}, \quad (24)$$

which will equal to 1 for a trait evolving neutrally. Both σ_B^2 and σ_W^2 can be estimated using quantitative trait and genomic sequences within and between species, while neither the mutation rates (μ and u), nor the effective population size (N_e), generation time or time of divergence ($t_{i,j}$) need to be estimated. Moreover, the nucleotide sequence from which π and $d_{i,j}$ are obtained should be neutrally evolving, but they are not necessarily linked to the quantitative trait under study.

Estimation

We hereby seek to obtain point estimates of σ_B^2 , σ_W^2 and ultimately ρ . For each species with data available, σ_W^2 as defined in Equation 9 can be seen as a replicate sample. Thus, σ_W^2 can be obtained by averaging out across all the sampled species. On the other hand, σ_B^2 such as as defined in Equation 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, σ_B^2 can thus be seen as an estimate of the rate of evolution of the quantitative trait along a phylogenetic tree, when the tree is measured in units of $4d$ (d being the nucleotide divergence). As such, any phylogenetic comparative methods that allow the estimation of phenotypic rates of evolution on a tree scaled by $4d$, instead of time as is usually the case, can be used to estimate σ_W^2 . We provide a maximum likelihood estimate for ρ as well as a Bayesian estimate to derive posterior probabilities that the null model of neutrality (i.e. $\rho = 1$) is rejected.

Maximum likelihood estimate

At the phylogenetic scale, for n taxa in the tree, \mathbf{D} ($n \times n$) is the symmetric distance matrix computed from the branch lengths and the topology of the phylogenetic tree. The diagonal $\mathbf{D}_{i,i}$ represents the total nucleotide divergence from the root of the tree to each taxon (i). The off-diagonal elements ($\mathbf{D}_{i,j} = d_{i,j}$) are the distances between the root and the most recent common ancestor of taxa i and j , as in Equation 17. The mean trait value at the root of the tree (ϕ) can be estimated from the $n \times 1$ vector of mean trait values $\bar{\mathbf{P}}$ at the tips of the tree using maximum likelihood (O'Meara *et al.*, 2006):

$$\phi = \left(\mathbf{1}^\top \times \mathbf{D}^{-1} \times \mathbf{1} \right)^{-1} \cdot \left(\mathbf{1}^\top \times \mathbf{D}^{-1} \times \bar{\mathbf{P}} \right), \quad (25)$$

where $\mathbf{1}$ is an $n \times 1$ column vector of ones.

Finally, between-species variation σ_B^2 is estimated as (O'Meara *et al.*, 2006):

$$\sigma_B^2 = \frac{1}{4} \frac{(\bar{\mathbf{P}} - \phi \cdot \mathbf{1})^\top \times \mathbf{D}^{-1} \times (\bar{\mathbf{P}} - \phi \cdot \mathbf{1})}{n-1}. \quad (26)$$

For a given species i with inter-individual data available, additive genetic variance of a trait ($V_{A,i}$) is the product of heritability (h_i^2) and phenotypic variance ($V_{P,i}$). The ratio of $V_{A,i}$ over nucleotide diversity of neutrally evolving sequences (π_i) is a sample estimate of σ_W^2 . Averaged across all species, we obtain the estimate σ_W^2 as

$$\sigma_W^2 = \frac{1}{n} \sum_{i=1}^n \frac{V_{A,i}}{\pi_i} = \frac{1}{n} \sum_{i=1}^n \frac{V_{P,i} \cdot h_i^2}{\pi_i}. \quad (27)$$

As depicted in Figure 1, the neutrality index is estimated as

$$\rho = \frac{\sigma_B^2}{\sigma_W^2}. \quad (28)$$

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, since simultaneously estimating all σ_B^2 allows improving their estimation Adams & Collyer (2018). More specifically, trait variation along the phylogenetic tree is modeled as a K -dimensional Brownian process \mathcal{B} ($1 \times K$) starting at the root and branching along the tree topology (Huelsenbeck & Rannala, 2003; Lartillot & Poujol, 2011; Lartillot & Delsuc, 2012; Latrille *et al.*, 2021). The rate of change of the Brownian process is determined by the positive semi-definite and symmetric covariance matrix between traits Σ ($K \times K$). The branch lengths of the tree used to model the Brownian process runs is measured in units of $4d$ (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 $4d$ units, and thus equate to σ_B^2 (see online supplementary material Section S2.1). Of note, modeling trait evolution as a multi-dimensional process is reliable only if $K \ll n$, meaning that the number of species is largely superior to the number of traits (Adams & Collyer, 2018). Thus, relying on a K -dimensional process should be reserved for a handful of allometric traits (e.g., brain mass and body mass). If K is large, the traits are better tested independently each with a 1-dimensional Brownian process, which is a specific case of the multi-dimensional process.

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). With an inverse Wishart distribution as the prior on the covariance matrix, the posterior on Σ , conditional on \mathcal{B} is also an invert Wishart distribution (see online supplementary material Section S2.2). We used Metropolis-Hastings algorithm to sample \mathcal{B} , while the posterior distribution of Σ is sampled using Gibbs sampling. For each trait and each species, the prior on heritability (h^2) for each species is set as a uniform distribution with user-defined boundaries. Heritability and phenotypic variance for each trait are combined with nucleotide

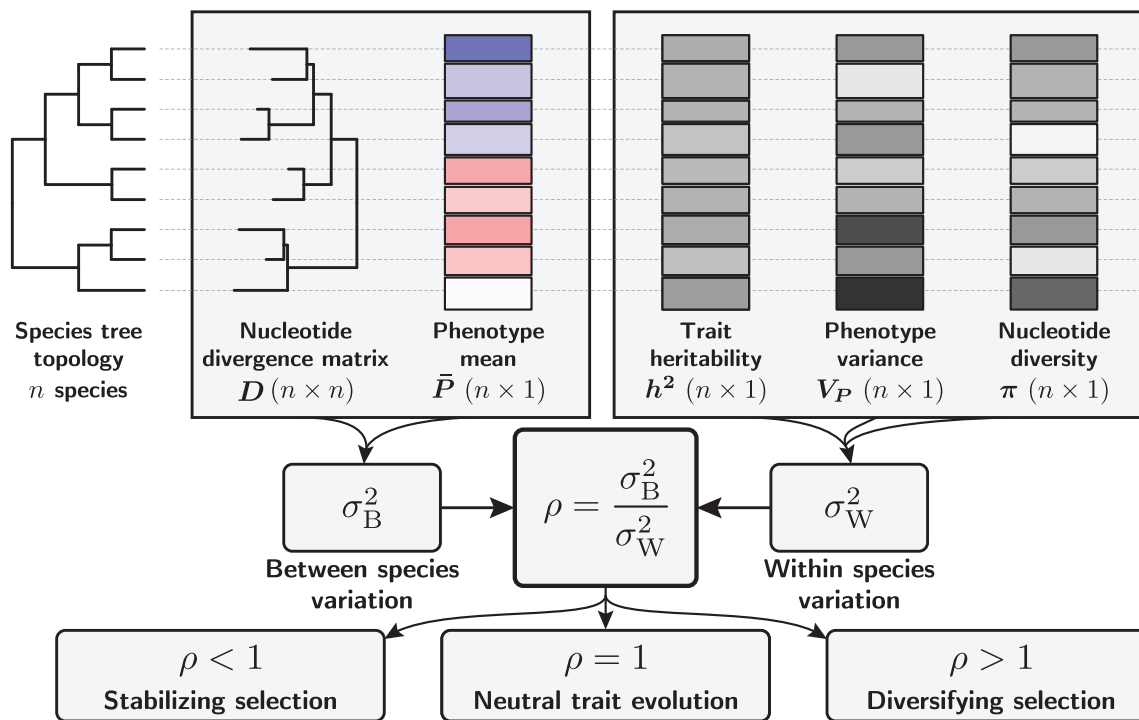


Figure 1. Between species, the change along the phylogeny of the mean phenotypic trait allows the estimation of between-species trait variation, σ_B^2 , which is normalized by nucleotide divergence. Within species, the genetic variance allows the estimation of within-species trait variation, σ_W^2 , which is normalized by nucleotide diversity. ρ is the ratio of σ_B^2 over σ_W^2 . Under neutral evolution, ρ is expected to be equal to one. Under diversifying selection, the trait is heterogeneous between species, but homogeneous within species, leading to ρ greater than one. Under stabilizing selection, the trait is homogeneous between species, leading to ρ smaller than one. Importantly, the sequence from which nucleotide diversity and divergence are estimated should be neutrally evolving, but they are not necessarily linked to the quantitative trait under study, they allow for discarding the confounding effect on mutation rate diversity, population size and divergence time.

diversity to compute σ_W^2 for each species before being averaged across species (as in Equation 27). From σ_W^2 estimated independently for each trait and the diagonal elements of Σ (i.e., the σ_B^2 for each trait), the posterior distribution of ρ (as in Equation 28) is obtained for each trait. The posterior distribution of ρ thus allows testing for deviation from neutrality (Figure 1), for example, by computing $\mathbb{P}[\rho > 1]$ to test for evidence of diversifying selection and $\mathbb{P}[\rho < 1]$ to test for evidence of stabilizing selection.

Applicability to empirical data

Our method assumes that the narrow-sense heritability (h^2) of a trait is known such as to estimate additive genetic variance (V_A) from phenotypic variance (V_P) as $V_A = h^2 \cdot V_P$. Fortunately, if heritability is not known, the test for diversifying selection can still be performed, although it is underpowered. Indeed, if the additive genetic variance is substituted by phenotypic variance, it is equivalent to assuming complete heritability ($h^2 = 1$). Because $h^2 \leq 1$ by definition, we overestimate the within-species variation and thus underestimate ρ . It is, however, possible to test for diversifying selection because testing for $\rho > 1$ while using phenotypic variance instead of additive genetic variance means that knowing the additive genetic variance would have only increased the evidence for diversifying selection. Similarly, using the broad-sense heritability (H^2) instead of narrow-sense heritability (h^2) results in an underestimation of ρ since $h^2 \leq H^2$ and thus can be used to detect diversifying selection if h^2 is not available. 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 et al., 2011; Hansen & Pélabon, 2021). Thus, 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.

In contrast to the test of diversifying selection, the test for stabilizing selection is invalid if ρ is underestimated. Several assumptions made by our test might not hold on empirical data and their consequences on the neutrality index and the test that can be performed are shown in Table 2.

Simulation

We tested the performance of our neutrality index (ρ) to detect selection on a quantitative trait using simulations. We performed simulations under different selective regimes (neutral, stabilizing, diversifying), different demographic histories (constant or fluctuating population size) and different evolution of the mutation rate (constant or fluctuating). Simulations were individual-based and followed a Wright-Fisher model with mutation, selection and drift for a diploid population including speciation along a predefined ultrametric phylogenetic tree (Figure 2). Each individual phenotypic value was the sum of genotypic value and an environmental effect. The environmental effect was normally distributed with variance V_E . We assumed that the genotypic value was encoded by $L = 5,000$ loci, with each locus contributing an additive effect that was normally distributed with standard deviation $a = 1$

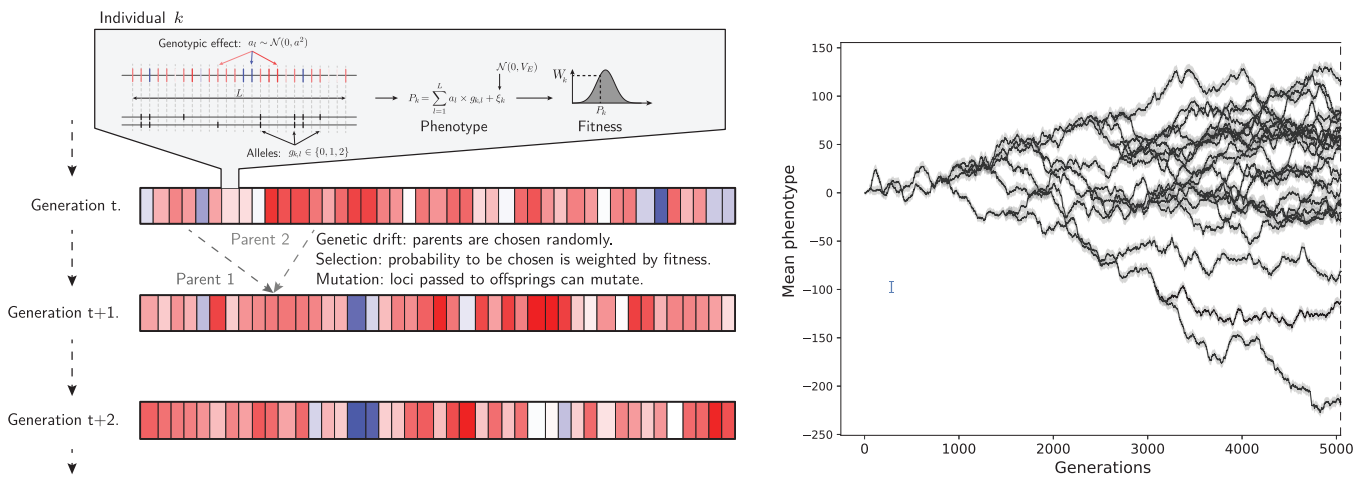


Figure 2. Wright–Fisher simulations with mutation, selection and drift. Left panel: For a given individual, the trait phenotypic value is the sum of genotypic value and an environmental effect (standard deviation V_E). The trait’s genotypic value is encoded by L independent loci (meaning no linkage), with each locus contributing additively to the genotypic value. Parents are selected for reproduction to the next generation according to their phenotypic value, with a probability proportional to their fitness. Mutations are drawn from a Poisson distribution, with each locus having a probability μ to mutate. Drift is modeled by the resampling of parents. Right panel: examples of a trait evolving along a phylogenetic tree, with the mean phenotype (black line) and the variance of the trait genotypic value (gray area).

(Figure 2 and for the theoretical formulation see online supplementary material Section S1.1 and Figure S1). We assumed a trait with a narrow-sense heritability of $h^2 = 0.2$ and computed the theoretical V_E accordingly (see online supplementary material Section S1.1). Assuming a diploid panmictic population of size $N_e = 50$ at the root of the tree, and with non-overlapping generations, we simulated explicitly each generation along an ultrametric phylogenetic tree. For each offspring, the number of mutations was drawn from a Poisson distribution with mean $2 \cdot \mu \cdot L$, with the mutation rate per locus per generation μ . From the empirical mammalian dataset (see next section), we computed an average nucleotide divergence from the root to leaves of 0.18 and average genetic diversity of 0.00276. We scaled parameters in our simulations to fit plausible values for mammals. We thus used a nucleotide mutation rate of $u = 0.00276/4N_e = 1.38 \times 10^{-5}$ per site per generation and a total of $0.18/1.38 \times 10^{-5} = 13,500$ generations from root to leaves, and the number of generations along each branch was proportional to the branch length. We set $\mu = u$ without loss in generality since the genetic architecture (L and u) is assumed constant in the simulator.

The changes in μ and N_e along the lineages were both modeled by a Brownian process on the log scale ($\log-\mu$ and $\log-N_e$), leading to geometric Brownian motion on the linear scale (μ and N_e). These processes are parameterized as $\mathcal{B}(0, \sigma_\mu = 0.0086)$ and $\mathcal{B}(0, \sigma_{N_e} = 0.0086)$, 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 $\log-N_e$ and $\log-\mu$ between the extant species and the root is 1.0. An Ornstein–Uhlenbeck process was overlaid to the instant value of $\log-N_e$ provided by the geometric Brownian process to account for short-term changes between generations ($\text{OU}(0, \sigma_{N_e} = 0.1, \theta_{N_e} = 0.9)$). The geometric Brownian motion accounted for long-term fluctuations (low rate of changes σ_{N_e} but unbounded), while

the Ornstein–Uhlenbeck introduced short-term fluctuations (high rate of changes σ_{N_e} but bounded and mean-reverting). The simulation started from an initial sequence at equilibrium at the root of the tree and, at each node, the process was split until it finally reached the leaves of the tree. From a speciation process perspective, this was equivalent to an allopatric speciation over one generation.

At each generation, parents were randomly sampled with a weight proportional to their fitness (W). Selection was modeled as a one-dimensional Fisher’s geometric landscape, with the fitness of an individual being a monotonously decreasing function of the distance between the individual and the optimal phenotype (Blanquart & Bataillon, 2016; Tenaillon, 2014). More specifically, the fitness of an individual was given by $W = e^{-(P-\lambda)^2/\alpha}$, where P was the trait value of the individual, $\lambda = 0.0$ was the optimal trait value, and $\alpha = 0.02$ was the strength of selection. Mutations were considered as a displacement of the phenotype in the multidimensional space. Beneficial mutations moved the phenotype closer to the optimum, while deleterious mutations moved it further away. Stabilizing selection was implemented by fixing the optimum phenotype to a single value ($\lambda = 0.0$). Diversifying selection was implemented by allowing the optimum phenotype to move along the phylogenetic tree as a geometric Brownian process (Hansen, 1997) ($\lambda \sim \mathcal{B}(0, \sigma_\lambda = 1.0)$). Neutral evolution was implemented by flattening the fitness landscape ($W = 1$), which meant that each individual had the same probability of being sampled at each generation.

Nucleotide diversity (π) was measured as the heterozygosity of neutral markers that were simulated along the phylogenetic tree but not linked to the trait simulated. Nucleotide divergence (d) was measured as the number of substitutions per site of neutral markers along the branches of the phylogenetic tree. The additive genetic variance was measured

as phenotypic variance multiplied by heritability. Heritability was estimated from the slopes of the regression of offspring's phenotypic trait values on parental phenotypic trait values (Lynch & Walsh, 1998) averaged over the last 10 simulated generations. Heritability was thus not a given parameter of the simulations, but rather measured as it would be in empirical data.

Empirical dataset

We analyzed a dataset of body and brain masses from mammals. The log-transformed values of body and brain masses were taken from Tsuboi *et al.* (2018). We removed individuals not marked as adults and split the data into males and females due to sexual dimorphism in body and brain masses. We discarded species with only one representative since phenotypic variance cannot be estimated. 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).

We also analyzed a dataset of primate species, with the nucleotide variation obtained from Kuderna *et al.* (2023) and the quantitative trait variation also from Tsuboi *et al.* (2018), using the same filtering as for the mammalian dataset. 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).

Results

Neutrality index

For a neutral trait, the genetic architecture, meaning the number of loci encoding the trait and the average effect of a mutation on the trait, is formally related to both within and between-species variation of the trait. We defined the neutrality index as $\rho = \sigma_B^2 / \sigma_W^2$, which equals 1 for a neutral trait (see *Materials and methods*), suggesting that traits for which this relationship was not verified were putatively under selection. Under stabilizing selection, the variation between species is depleted because the mean trait is maintained toward similar values between different species, which leads to $\rho < 1$. In contrast, under diversifying selection, the variation between species is inflated because species will have potentially different trait values (Hansen, 1997), which leads to $\rho > 1$. Our neutrality index for a quantitative trait leveraged the data for any number of species, and took advantage of the signal over the whole phylogenetic tree, at the same time taking into account phylogenetic inertia and addressing the non-independence between species (Figure 1). This statistic was obtained as a maximum likelihood estimate from Equations 27 and 26. We also devised a Bayesian estimate to obtain the posterior distribution of the neutrality index, and test for diversifying selection as $\mathbb{P}[\rho > 1]$, and stabilizing selection as $\mathbb{P}[\rho < 1]$.

Our neutrality index made a series of assumptions that we described in details in *Material and methods*. Table 2 summarized these assumptions and outlined possible consequences for the neutrality test that we proposed.

Results against simulations

The inference framework was first tested on independently simulated datasets matching an empirically relevant mammalian empirical regime (see *Materials and methods*). Under constant population size (N_e) and constant mutation rates (μ and u) across the phylogenetic tree (Figure 3, top row), we found no false negative for simulations of stabilizing ($\mathbb{P}[\rho < 1] > 0.975$; blue in Figure 3) or diversifying ($\mathbb{P}[\rho > 1] > 0.975$; red in Figure 3) selection. For simulations under neutral evolution, 77% of those were correctly identified ($0.025 \leq \mathbb{P}[\rho > 1] \leq 0.975$; yellow in Figure 3), while 21% and 2% were wrongly detected as stabilizing or diversifying selection, respectively. Once we introduced fluctuating N_e , μ and u (Figure 3, bottom row), our ability to identify simulations under either diversifying or stabilizing selection remained the same with all cases detected correctly. For simulations under neutral evolution, 51% of the simulations were correctly detected ($0.025 \leq \mathbb{P}[\rho > 1] \leq 0.975$), while 49% were detected as stabilizing selection ($\mathbb{P}[\rho < 1] > 0.975$) and none as diversifying selection.

Results on empirical data

For mammalian body and brain mass, we obtained male (σ) and female (φ) trait variations. Combined with nucleotide diversity and divergence, we estimated ρ and posterior probabilities of diversifying selection under different assumptions for trait heritability as shown in Table 1. For body mass, assuming complete heritability led to zero posterior probabilities of diversifying selection for both males and females ($\mathbb{P}[\rho > 1] = 0.0$). If we assumed that heritability (h^2) of body mass was uniformly distributed between 20% and 40% (Hu *et al.*, 2022), posterior probabilities of diversifying selection became 0.635 for males and 0.324 for females. Mammalian brain mass was found to be under diversifying selection with posterior probabilities of 0.877 for males and 0.972 for females when complete heritability was assumed. Assuming a uniform distribution between 20% and 40% for heritability led to posterior probabilities of diversifying selection of 1.0 for both males and females.

We also analyzed a similar dataset for body mass focusing this time only at Primates (Table 1). For primates body mass, assuming complete heritability led to zero posterior probabilities of diversifying selection for both males and females, exactly as in the mammal dataset. However, we found posterior probabilities of diversifying selection of 1.0 for males and 0.914 for females when assuming a uniform distribution for the heritability of body mass between 20% and 40%. For brain mass, assuming complete heritability or not (between 20% and 40%) did not change the posterior probability of diversifying selection, which was 1.0. Evidence for diversifying selection on both brain and body mass was therefore more pronounced in Primates than in mammals. However, the genetic markers used to normalize trait variance with nucleotide divergence were not necessarily neutral, which could create spurious false positives by artificially inflating ρ (Table 2 and *Material and methods*).

Discussion

In this study, we proposed a neutrality index for a quantitative trait that can be used within a statistical framework to test for selection. Our neutrality index for a trait, ρ , is calculated as the ratio of the normalized within- to between-species

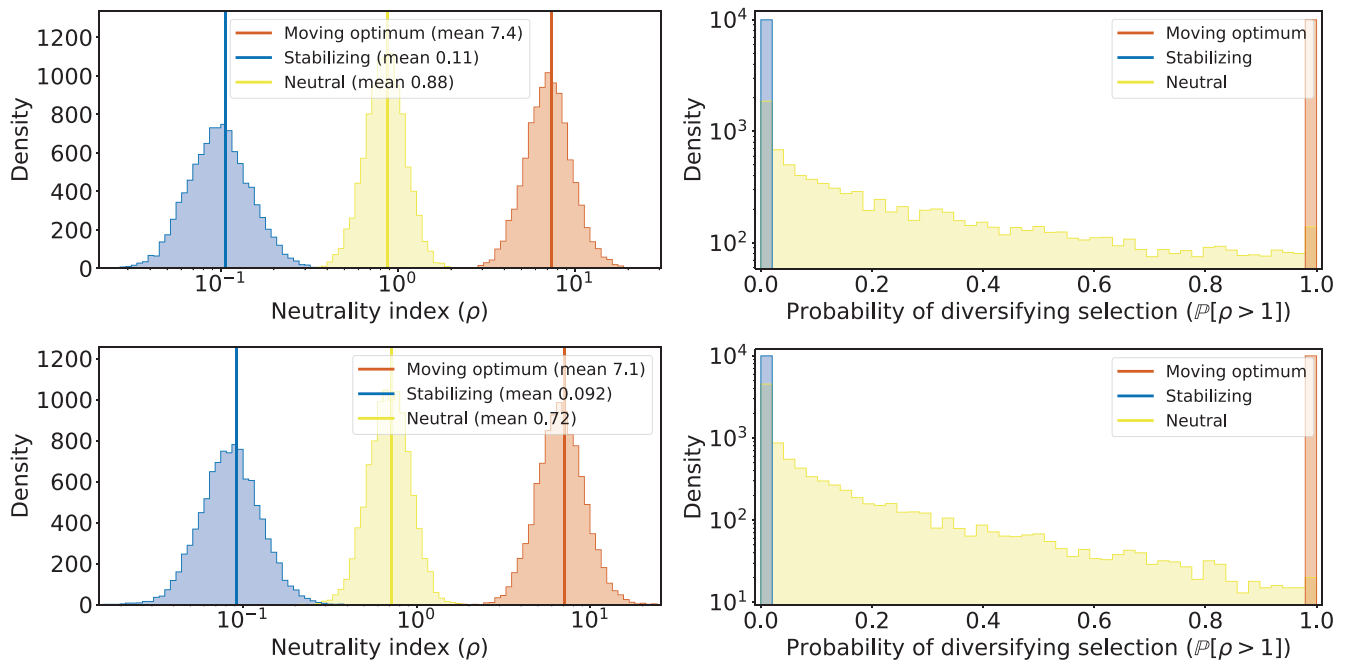


Figure 3. 10,000 simulations of trait evolution along a phylogenetic tree under different selection regimes. Traits simulated under stabilizing selection (blue), under a neutral evolution (yellow), and under a moving optimum (red). Histogram of ratio of between-species trait variation (σ_B^2) over within-species trait variation σ_W^2 with $\rho = \sigma_B^2/\sigma_W^2$ estimated from each simulated data (left) and probabilities of ρ being greater than 1 (right). Effective population size (N_e) and mutation rates (μ and u) were either constant (top row), or fluctuating as a Brownian process along the phylogenetic tree (bottom row).

Table 1. Test of diversifying selection on a mammal and a primate dataset, by splitting males (σ) and females (φ). Traits considered were body mass or brain mass (log-transformed). Heritability (h^2) was either assumed complete ($h^2 = 1.0$) or uniformly distributed between 20% and 40% ($h^2 \sim \mathcal{U}(0.2, 0.4)$). n was the number of species in the dataset. ρ was the posterior estimate of our neutrality index, with the 95% credible interval (CI) for ρ also computed. $\mathbb{P}[\rho > 1]$ was the estimated posterior probability of diversifying selection.

Dataset	Trait	h^2	Sex	n	ρ	95% CI for ρ	$\mathbb{P}[\rho > 1]$
Mammals	Body mass	1.0	σ	36	0.340	0.217-0.523	0.000
Mammals	Body mass	1.0	φ	26	0.277	0.160-0.490	0.000
Mammals	Body mass	$\mathcal{U}(0.2, 0.4)$	σ	36	1.124	0.721-1.754	0.635
Mammals	Body mass	$\mathcal{U}(0.2, 0.4)$	φ	26	0.936	0.523-1.715	0.324
Mammals	Brain mass	1.0	σ	36	1.351	0.851-2.173	0.877
Mammals	Brain mass	1.0	φ	26	1.727	0.991-2.938	0.972
Mammals	Brain mass	$\mathcal{U}(0.2, 0.4)$	σ	36	4.527	2.831-7.091	1.000
Mammals	Brain mass	$\mathcal{U}(0.2, 0.4)$	φ	26	6.001	3.288-10.941	1.000
Primates	Body mass	1.0	σ	71	0.558	0.401-0.784	0.000
Primates	Body mass	1.0	φ	65	0.389	0.278-0.547	0.000
Primates	Body mass	$\mathcal{U}(0.2, 0.4)$	σ	71	1.875	1.288-2.695	1.000
Primates	Body mass	$\mathcal{U}(0.2, 0.4)$	φ	65	1.296	0.899-1.821	0.914
Primates	Brain mass	1.0	σ	71	1.929	1.395-2.616	1.000
Primates	Brain mass	1.0	φ	65	1.950	1.399-2.790	1.000
Primates	Brain mass	$\mathcal{U}(0.2, 0.4)$	σ	71	6.479	4.658-8.944	1.000
Primates	Brain mass	$\mathcal{U}(0.2, 0.4)$	φ	65	6.522	4.664-9.294	1.000

variation and it allowed the identification of the evolutionary regime of a quantitative trait. At the phylogenetic scale, trait variation between species was normalized by sequence divergence obtained from a neutral set of markers. Similarly, trait variation within species was normalized by sequence polymorphism obtained also from a neutral set of markers. Our estimate of ρ could be tested for deviation from the value of 1.0

expected under the null hypothesis of neutrality. Technically, the neutrality index can be estimated either as a maximum likelihood point estimate, or as a mean posterior estimate from a Bayesian implementation (see online [supplementary material Section S3](#)). The latter also enabled the estimation of the posterior credible interval to test for departure from a neutrally evolving trait (e.g., $\mathbb{P}[\rho > 1]$). We tested our statistical

Table 2. Assumptions breaks and their consequences on the estimation of within-species variation (σ_W^2), between-species variation (σ_B^2), and on the neutrality index $\rho = \sigma_B^2/\sigma_W^2$. The last two columns indicate whether the test for diversifying selection ($\rho > 1$) and for stabilizing selection $\rho < 1$ are conservative or invalid due to violated assumptions.

Broken assumption	Consequences	σ_W^2	σ_B^2	Test $\rho > 1$	Test $\rho < 1$
Trait encoded by few loci	Between-species trait variation is underestimated	–	Underestimated	Conservative	Invalid
Sexual dimorphism	Within-species trait variation is overestimated	Overestimated	–	Conservative	Invalid
Phenotypic plasticity	Trait responding to individual environments	Overestimated	–	Conservative	Invalid
Inbreeding	Nucleotide diversity (π) is underestimated	Overestimated	–	Conservative	Invalid
Markers for polymorphism are negatively selected	Nucleotide diversity (π) is underestimated	Overestimated	–	Conservative	Invalid
Markers for polymorphism are positively selected	Nucleotide diversity (π) is underestimated	Overestimated	–	Conservative	Invalid
Markers for divergence are positively selected	Nucleotide divergence (d) is overestimated	–	Underestimated	Conservative	Invalid
Markers for polymorphism under balanced selection	Nucleotide diversity (π) is overestimated	Underestimated	–	Invalid	Conservative
Markers for divergence are negatively selected	Nucleotide divergence (d) is underestimated	–	Overestimated	Invalid	Conservative
Multiple nucleotide substitutions at the same locus	Nucleotide divergence (d) is underestimated	–	Overestimated	Invalid	Conservative

procedure against simulated data and showed that our test was able to correctly detect simulations under diversifying selection (test of $\rho > 1$) or under stabilizing selection (test of $\rho < 1$). However, our test detected a spurious signal of stabilizing selection ($\rho < 1$) 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 may be bounded by what the genetic architecture can produce, and this could cause a slowdown of phenotypic divergence over time due to the erosion of possible phenotypic changes at the underlying loci. Typically, such an effect depends on the number of alleles per locus, whether new mutations are generating new alleles or instead reverting to previous alleles. Altogether, in our simulation setting under a constant genetic architecture with a fixed number of loci, such a slowdown of phenotypic divergence can result in a spurious signal of stabilizing selection ($\rho < 1$), especially for deeper phylogeny (see online [supplementary material Figure S2](#) and [Section S4](#)). 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.

Our results showed that our method significantly improved over currently available methods to detect selection acting on a trait at the phylogenetic scale. Current methods relying on evolution of the mean trait value between species also tend to statistically prefer a model of stabilizing selection over a Brownian process when the trait is neutral ([Cooper et al., 2016](#); [Price et al., 2022](#); [Silvestro et al., 2015](#)). Our approach could in theory be applied to detect stabilizing selection at the phylogenetic scale, but we showed that it did not have the statistical power to identify those cases. In contrast, we showed that our method was able to identify correctly cases

of diversifying selection, which is a clear improvement over current methods that model only mean trait value. Indeed, under diversifying selection, mean trait value will not deviate from a Brownian process, and thus cannot be distinguished from neutral evolution ([Hansen & Martins, 1996](#); [Harmon, 2018](#)). For example, testing the selective regime in the expression level of the majority of genes led to the selection of a Brownian process as the preferred model and the interpretation that the expression was evolving neutrally ([Catalán et al., 2019](#)). 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 ([Rohlf 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 outlier genes ([Gillard et al., 2021](#); [Rohlf & 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 ([Lande, 1979](#); [Lemos et al., 2001, 2005](#); [Lynch & Crease, 1990](#); [Porto et al., 2015](#); [Weaver et al., 2007](#)). Specifically, by leveraging estimates of effective population size (N_e) and 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 (Machado *et al.*, 2022; Schroeder & von Cramon-Taubadel, 2017) 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 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).

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 sequence variation (Leinonen *et al.*, 2013; Martin *et al.*, 2008) or co-ancestry between individuals (Ovaskainen *et al.*, 2011). Our method can be seen as an analog 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 ρ .

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; Khabbazian *et al.*, 2016; Mitov *et al.*, 2020; Uyeda & Harmon, 2014).

Even though our test was developed for a quantitative trait, analogies with other tests of selection developed for molecular sequences also provided insight into its behavior. 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). As $\omega < 1$ is interpreted as purifying selection acting on the protein, $\rho < 1$ is interpreted as stabilizing selection acting on the trait. Similarly, the interpretation of adaptation for $\omega > 1$ is analogous to diversifying selection for $\rho > 1$. With this analogy in mind, we could leverage the vast literature discussing and interpreting the results of these tests and their pitfalls (Anisimova & Kosiol, 2009; Jensen *et al.*, 2019; Nielsen, 2005). First, not rejecting the neutral null model of $\rho = 1$ did not necessarily imply that the trait was effectively neutral, since diversifying and stabilizing selection could compensate each other resulting in $\rho = 1$, analogously to $\omega = 1$ under a mix of adaptation and purifying selection (Nielsen, 2005). Second, empirical evidence for $\rho < 1$ did not rule out diversifying selection, but rather that this diversifying selection was not strong enough to overcome the stabilizing selection, similarly to strong purifying selection resulting $\omega < 1$ even though those genes and sites are under adaptation (Latrille *et al.*, 2023). By explicitly modeling stabilizing selection as a moving optimum, it would theoretically be possible to tease apart the effect of diversifying and stabilizing selection in the context of quantitative traits to obtain a statistically more powerful test.

In the context of detecting diversifying selection on a trait, we argue that the main drawback of our method is that the additive genetic variance of the trait is required instead of the phenotypic variance. If phenotypic variance was used instead of additive genetic variance to estimate ρ , meaning that we assumed complete heritability, the neutrality index ρ was ultimately underestimated. Similarly, using broad-sense heritability instead of narrow-sense heritability would result in underestimated ρ . In such context, the test of stabilizing selection ($\rho < 1$) would be statistically invalid. However, the test of diversifying selection ($\rho > 1$) was underpowered although not invalidated, meaning that absence of evidence would not be evidence of absence. As an example, even though we assumed complete heritability for brain mass, we uncovered diversifying selection in mammals since $\rho > 1$. If available, any prior knowledge on heritability can be leveraged instead of assuming complete heritability to increase the statistical power to detect diversifying selection (Hansen & Pélabon, 2021; Hansen *et al.*, 2011). Additionally, phenotypic plasticity also affects the genotype-phenotype relationship with intricate consequences for our test of selection. First, at the level of within species variation, individuals might occupy different patches with different environments. Responding to these individual environmental conditions, phenotypic plasticity would then result in increased trait variation within species. In this scenario, as hypothesized in Rohlf & Nielsen (2015), phenotypic plasticity then leads to a reduced ratio of between to within species variations, thus ultimately leading to our tests of diversifying selection being underpowered although not invalid. Alternatively, it is also possible that different species are experiencing different macro-environments, for example with species spread along a latitudinal or elevation gradient, with different temperatures or precipitation. These species could thus have different mean phenotypes solely because of phenotypic plasticity, while such changes are not encoded in their genome (Schraiber & Edge, 2023; Stamp & Hadfield, 2020). Such an effect can lead to $\rho > 1$ erroneously interpreting diversifying selection. The test of $\rho > 1$

would however be correct that the changes in mean phenotypes across species is due to change in environment, albeit in such a case not encoded by the genotype of individuals but due to phenotypic plasticity.

The development of our neutrality index was also based on several assumptions that could be relaxed in future studies. First, we cannot predict the behavior of our test in the context of population structures, gene flow and introgression. These factors should be thoroughly investigated using simulations. Second, loci were assumed to contribute additively to the phenotype. Although the effects of dominance and epistasis is typically weak compared to the additive effects on the quantitative traits, their influence should be assessed (Crow, 2010; Hill *et al.*, 2008). Third, the genetic architecture of the trait was assumed to be constant across the phylogenetic tree, whereas it might actually be variable among individuals and species (Huber *et al.*, 2015; Tung *et al.*, 2015). Such an assumption can theoretically be relaxed and changes in genetic architecture along the phylogenetic tree could jointly be estimated (Arnold *et al.*, 2008; Gaboriau *et al.*, 2020; Hohenlohe & Arnold, 2008; Kostikova *et al.*, 2016). Finally, from a statistical perspective, our Bayesian estimation could integrate uncertainty from the estimation of genetic variation, using sequences as input instead of estimated values of nucleotide diversity and divergence.

From an empirical point of view, our method required integrating genomic and trait variation, which could reduce the possible datasets to be used. 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, both showing signals of diversifying selection. As such, this result corroborates studies relying solely on changes in mean trait values across mammals, showing strong statistical support for several distinct evolutionary regimes for body- and brain mass (Mitov *et al.*, 2019). Interestingly, our strongest signal is for brain mass, corroborating studies in hominids where skull size (related to brain mass) is the only trait that exceeded the expected rate of phenotypic evolution under a neutral model (Lynch, 1990). Hence, one first interpretation here is that brain mass might be an exceptional case among many phenotypic traits (e.g., dental and skeletal measures). Second, from a macro-evolutionary perspective, the consensus is that empirical rates of evolution calculated on phylogenetic trees and the fossil record are far inferior to the expected under drift (Lynch & Crease, 1990; Uyeda *et al.*, 2011), where such methods assume constancy of N_e , generation time and mutation rates. 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 results and interpretations. Because our test was also based on several assumptions that might not hold on empirical data, we also provided a table containing the main assumptions and their consequences on the neutrality index and the test that can be performed (Table 2). For example, at the primate scale, the evidence for $\rho > 1$ does not necessarily imply that the brain mass was evolving under diversifying selection since the markers used for nucleotide divergences were not neutral,

which can lead to a spurious $\rho > 1$. In conclusion, our study provided a statistical framework to test for diversifying selection acting on a quantitative trait while integrating the trove of genomic data available both within and between species, and we believe that our new approach is a promising tool to investigate the evolution of quantitative traits.

Funding

The work has been funded by Université de Lausanne and the Swiss National Science Foundation (315230_219757).

Acknowledgments

We gratefully acknowledge the help of Nicolas Lartillot, Philippe Veber, Isabela Jeronimo do Ó, Anna Marcionetti, Julien Clavel, and Daniele Silvestro for their insightful discussions and Julien Joseph for his advice and reviews concerning this manuscript.

Conflicts of interest

None declared.

Data availability

The data and code that support the findings of this study are openly available in Zenodo at <https://doi.org/10.5281/zenodo.12666506>. Snakemake pipeline, analysis scripts and documentation are available in the repository to replicate the study.

References

- Adams, D. C. & Collyer, M. L. 2018. Multivariate phylogenetic comparative methods: Evaluations, comparisons, and recommendations. *Systematic Biology*, 67(1): 14–31.
- Anisimova, M. & Kosiol, C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution*, 26(2): 255–271.
- Arnold, S. J., Bürger, R., Hohenlohe, P. A., . . . , Jones, A. G. 2008. Understanding the evolution and stability of the G-matrix. *Evolution*, 62(10): 2451–2461.
- Barton, N. H., Etheridge, A. M., & Véber, A. 2017. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, 118: 50–73.
- Bergeron, L. A., Besenbacher, S., Zheng, J., . . . , Zhang, G. 2023. Evolution of the germline mutation rate across vertebrates. *Nature*, 615(7951): 285–291.
- Blanquart, F. & Bataillon, T. 2016. Epistasis and the structure of fitness landscapes: Are experimental fitness landscapes compatible with fisher's geometric model? *Genetics*, 203(2): 847–862.
- Catalán, A., Briscoe, A. D., & Höhna, S. 2019. Drift and directional selection are the evolutionary forces driving gene expression divergence in eye and brain tissue of *Heliconius* butterflies. *Genetics*, 213(2): 581–594.
- Cooper, N., Thomas, G. H., Venditti, C., . . . , Freckleton, R. P. 2016. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biological Journal of the Linnean Society*, 118(1): 64–77.
- Crow, J. F. 2010. On epistasis: Why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544): 1241–1244.
- Edelaar, P., Burraco, P., & Gomez-Mestre, I. 2011. Comparisons between QST and FST—how wrong have we been? *Molecular Ecology*, 20(23): 4830–4839.

- Felsenstein, J. 1985. Phylogenies and the Comparative Method. *The American Naturalist*, 125(1): 1–15.
- Felsenstein, J. 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, 19(1): 445–471.
- Felsenstein, J. 2008. Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. *The American Naturalist*, 171(6): 713–725.
- Foley, N. M., Mason, V. C., Harris, A. J., . . . , Murphy, W. J. 2023. A genomic timescale for placental mammal evolution. *Science*, 380(6643): eabl8189.
- Fraser, H. B. 2020. Detecting selection with a genetic cross. *Proceedings of the National Academy of Sciences United States of America*, 117(36): 22323–22330.
- Gaboriau, T., Mendes, F. K., Joly, S., . . . , Salamin, N. 2020. A multi-platform package for the analysis of intra- and interspecific trait evolution. *Methods in Ecology and Evolution*, 11(11): 1439–1447.
- Gaboriau, T., Tobias, J. A., Silvestro, D., & Salamin, N. 2023. Exploring the Macroevolutionary Signature of Asymmetric Inheritance at Speciation.
- Geneux, D. P., Serres, A., Armstrong, J., . . . , Zoonomia Consortium 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature*, 587(7833): 240–245.
- Gillard, G. B., Grønvold, L., Røsaeg, L. L., . . . , Hvidsten, T. R. 2021. Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biology*, 22(1): 103.
- Goldman, N. & Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5): 725–736.
- Grabowski, M., Pienaar, J., Voje, K. L., . . . , Hansen, T. F. 2023. A Cautionary Note on “A Cautionary Note on the Use of Ornstein Uhlenbeck Models in Macroevolutionary Studies”. *Systematic Biology*, 72(4): 955–963.
- Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5): 1341–1351.
- Hansen, T. F. & Bartoszek, K. 2012. Interpreting the evolutionary regression: The interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology*, 61(3): 413–425.
- Hansen, T. F. & Martins, E. P. 1996. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*, 50(4): 1404–1417.
- Hansen, T. F. & Pélabon, C. 2021. Evolvability: A Quantitative-Genetics Perspective. *Annual Review of Ecology, Evolution, and Systematics*, 52(1): 153–175.
- Hansen, T. F., Pélabon, C., & Houle, D. 2011. Heritability is not Evolvability. *Evolutionary Biology*, 38(3): 258–277.
- Harmon, L. 2018. Phylogenetic comparative methods: Learning from trees.
- Hill, W. G., Goddard, M. E., & Visscher, P. M. 2008. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genetics*, 4(2): e1000008.
- Hohenlohe, P. A. & Arnold, S. J. 2008. MIPoD: A hypothesis-testing framework for microevolutionary inference from patterns of divergence. *The American Naturalist*, 171(3): 366–385.
- Hu, Z.-L., Park, C. A., & Reecy, J. M. 2022. Bringing the Animal QTLdb and CorrDB into the future: Meeting new challenges and providing updated services. *Nucleic Acids Research*, 50(D1): D956–D961.
- Huber, B., Whibley, A., Poul, Y. L., . . . , Joron, M. 2015. Conservatism and novelty in the genetic architecture of adaptation in *Heliconius* butterflies. *Heredity*, 114(5): 515–524.
- Huelsenbeck, J. P. & Rannala, B. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution*, 57(6): 1237–1247.
- Ingram, T. & Mahler, D. 2013. SURFACE: Detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. *Methods in Ecology and Evolution*, 4(5): 416–425.
- Jensen, J. D., Payseur, B. A., Stephan, W., . . . , Charlesworth, B. 2019. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. *Evolution*, 73(1): 111–114.
- Khazzazian, M., Kriebel, R., Rohe, K., & Ané, C. 2016. Fast and accurate detection of evolutionary shifts in Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution*, 7(7): 811–824.
- Khaitovich, P., Weiss, G., Lachmann, M., . . . , Pääbo, S. 2004. A neutral model of transcriptome evolution. *PLoS Biology*, 2(5): e132.
- Kimura, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6): 713–719.
- Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature*, 217(5129): 624–626.
- Kostikova, A., Silvestro, D., Pearman, P. B., & Salamin, N. 2016. Bridging inter- and intraspecific trait evolution with a hierarchical bayesian approach. *Systematic Biology*, 65(3): 417–431.
- Kuderna, L. F. K., Gao, H., Janiak, M. C., . . . , Marques Bonet, T. 2023. A global catalog of whole-genome diversity from 233 primate species. *Science*, 380(6648): 906–913.
- Lamy, J.-B., Plomion, C., Kremer, A., & Delzon, S. 2012. QST < FST As a signature of canalization. *Molecular Ecology*, 21(23): 5646–5655.
- Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain: Body size allometry. *Evolution*, 33(1): 402–416.
- Lande, R. 1980a. Genetic variation and phenotypic evolution during allopatric speciation. *The American Naturalist*, 116(4): 463–479.
- Lande, R. 1980b. Sexual dimorphism, sexual selection, and adaptation in polygenic characters. *Evolution*, 34(2): 292–305.
- Lartillot, N. & Delsuc, F. 2012. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution*, 66(6): 1773–1787.
- Lartillot, N. & Poujol, R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular Biology and Evolution*, 28(1): 729–744.
- Latrille, T., Lanore, V., & Lartillot, N. 2021. Inferring long-term effective population size with mutation-selection models. *Molecular Biology and Evolution*, 38(10): 4573–4587.
- Latrille, T., Rodrigue, N., & Lartillot, N. 2023. Genes and sites under adaptation at the phylogenetic scale also exhibit adaptation at the population-genetic scale. *Proceedings of the National Academy of Sciences of the United States of America*, 120(11): e2214977120.
- Leinonen, T., O’Hara, R. B., Cano, J. M., & Merilä, J. 2008. Comparative studies of quantitative trait and neutral marker divergence: A meta-analysis. *Journal of Evolutionary Biology*, 21(1): 1–17.
- Leinonen, T., McCairns, R. J. S., O’Hara, R. B., & Merilä, J. 2013. QST–FST comparisons: Evolutionary and ecological insights from genomic heterogeneity. *Nature Reviews Genetics*, 14(3): 179–190.
- Lemos, B., Marroig, G., & Cerqueira, R. 2001. Evolutionary rates and stabilizing selection in large-bodied opossum skulls (Didelphimorphia: Didelphidae). *Journal of Zoology*, 255(2): 181–189.
- Lemos, B., Meiklejohn, C. D., Cáceres, M., & Hartl, D. L. 2005. Rates of Divergence in Gene Expression Profiles of Primates, Mice, and Flies: Stabilizing Selection and Variability Among Functional Categories. *Evolution*, 59(1): 126–137.
- Litsios, G. & Salamin, N. 2012. Effects of Phylogenetic Signal on Ancestral State Reconstruction. *Systematic Biology*, 61(3): 533–538.
- Louca, S. & Pennell, M. W. 2020. Extant timetrees are consistent with a myriad of diversification histories. *Nature*, 580(7804): 502–505.
- Lynch, M. 1990. The Rate of Morphological Evolution in Mammals from the Standpoint of the Neutral Expectation. *The American Naturalist*, 136(6): 727–741.
- Lynch, M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution*, 45(5): 1065–1080.
- Lynch, M. & Crease, T. J. 1990. The analysis of population survey data on DNA sequence variation. *Molecular Biology and Evolution*, 7(4): 377–394.

- Lynch, M. & Walsh, B. 1998. *Genetics and Analysis of Quantitative Traits*, volume 1. Sinauer Sunderland, MA.
- Lynch, M., Latta, L., Hicks, J., and Giorgianni, M. 1998. Mutation, selection, and the maintenance of life-history variation in a natural population. *Evolution*, 52(3): 727–733.
- Machado, F. A., Marroig, G., & Hubbe, A. 2022. The pre-eminent role of directional selection in generating extreme morphological change in glyptodonts (Cingulata; Xenarthra). *Proceedings of the Royal Society B: Biological Sciences*, 289(1967): 20212521.
- Machado, F. A., Mongle, C. S., Slater, G., . . . , Uyeda, J. C. 2023. Using developmental rules to align microevolution with macroevolution.
- Martin, G., Chapuis, E., & Goudet, J. 2008. Multivariate Q_{ST} - F_{ST} comparisons: A neutrality test for the evolution of the G matrix in structured populations. *Genetics*, 180(4): 2135–2149.
- McCandlish, D. M. & Stoltzfus, A. 2014. Modeling evolution using the probability of fixation: History and implications. *Quarterly Review of Biology*, 89(3): 225–252.
- McDonald, J. H. & Kreitman, M. 1991. Adaptive protein evolution at Adh locus in *Drosophila*. *Nature*, 351(6328): 652–654.
- Merilä, J. & Crnokrak, P. 2001. Comparison of genetic differentiation at marker loci and quantitative traits. *Journal of Evolutionary Biology*, 14(6): 892–903.
- Mitov, V., Bartoszek, K., & Stadler, T. 2019. Automatic generation of evolutionary hypotheses using mixed Gaussian phylogenetic models. *Proceedings of the National Academy of Sciences*, 116(34): 16921–16926.
- Mitov, V., Bartoszek, K., Asimomitis, G., & Stadler, T. 2020. Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. *Theoretical Population Biology*, 131: 66–78.
- Muse, S. V. & Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5): 715–724.
- Nielsen, R. 2005. Molecular signatures of natural selection. *Annual Review of Genetics*, 39(1): 197–218.
- O'Meara, B. C., Ané, C., Sanderson, M. J., & Wainwright, P. C. 2006. Testing for different rates of continuous trait evolution using likelihood. *Evolution*, 60(5): 922–933.
- Ovaskainen, O., Karhunen, M., Zheng, C., . . . , Merilä, J. 2011. New method to uncover signatures of divergent and stabilizing selection in quantitative traits. *Genetics*, 189(2): 621–632.
- Pennell, M. W., FitzJohn, R. G., Cornwell, W. K., & Harmon, L. J. 2015. Model Adequacy and the macroevolution of angiosperm functional traits. *The American Naturalist*, 186(2): E33–E50.
- Porto, A., Sebastião, H., Pavan, S. E., . . . , Cheverud, J. M. 2015. Rate of evolutionary change in cranial morphology of the marsupial genus *Monodelphis* is constrained by the availability of additive genetic variation. *Journal of Evolutionary Biology*, 28(4): 973–985.
- Price, P. D., Palmer Drogue, D. H., Taylor, J. A., . . . , Wright, A. E. 2022. Detecting signatures of selection on gene expression. *Nature Ecology & Evolution*, 6(7): 1035–1045.
- Pujol, B., Wilson, A. J., Ross, R. I. C., & Pannell, J. R. 2008. Are Q_{ST} - F_{ST} comparisons for natural populations meaningful? *Molecular Ecology*, 17(22): 4782–4785.
- Rohlf, R. V. & Nielsen, R. 2015. Phylogenetic ANOVA: The expression variance and evolution model for quantitative trait evolution. *Systematic Biology*, 64(5): 695–708.
- Rohlf, R. V., Harrigan, P., & Nielsen, R. 2014. Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Molecular Biology and Evolution*, 31(1): 201–211.
- Rolland, J., Henao-Diaz, L. F., Doebeli, M., . . . , Schluter, D. 2023. Conceptual and empirical bridges between micro- and macroevolution. *Nature Ecology & Evolution*, 7(8): 1181–1193.
- Schraiber, J. G. & Edge, M. D. 2023. Heritability within groups is uninformative about differences among groups: Cases from behavioral, evolutionary, and statistical genetics. *Proceedings of the National Academy of Sciences*, 121(12): e2319496121.
- Schroeder, L. & von Cramon-Taubadel, N. 2017. The evolution of hominoid cranial diversity: A quantitative genetic approach. *Evolution*, 71(11): 2634–2649.
- Sella, G. & Barton, N. H. 2019. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annual Review of Genomics and Human Genetics*, 20(1): 461–493.
- Silvestro, D., Kostikova, A., Litsios, G., . . . , Salamin, N. 2015. Measurement errors should always be incorporated in phylogenetic comparative analysis. *Methods in Ecology and Evolution*, 6(3): 340–346.
- Silvestro, D., Tejedor, M. F., Serrano-Serrano, M. L., . . . , Salamin, N. 2019. Early arrival and climatically-linked geographic expansion of new world monkeys from tiny African ancestors. *Systematic Biology*, 68(1): 78–92.
- Simons, Y. B., Bullaughey, K., Hudson, R. R., & Sella, G. 2018. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biology*, 16(3): e2002985.
- Stamp, M. A. & Hadfield, J. D. 2020. The relative importance of plasticity versus genetic differentiation in explaining between population differences; a meta-analysis. *Ecology Letters*, 23(10): 1432–1441.
- Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3): 585–595.
- Tenaillon, O. 2014. The utility of Fisher's geometric model in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics*, 45(1): 179–201.
- Tsuboi, M., van der Bijl, W., Kopperud, B. T., . . . , Kolm, N. 2018. Breakdown of brain-body allometry and the encephalization of birds and mammals. *Nature Ecology & Evolution*, 2(9): 1492–1500.
- Tung, J., Zhou, X., Alberts, . . . , Gilad, Y. 2015. The genetic architecture of gene expression levels in wild baboons. *eLife*, 4: e04729.
- Turelli, M. 1984. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. *Theoretical Population Biology*, 25(2): 138–193.
- Turelli, M. 2017. Commentary: Fisher's infinitesimal model: A story for the ages. *Theoretical Population Biology*, 118: 46–49.
- Uyeda, J. C. & Harmon, L. J. 2014. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Systematic Biology*, 63(6): 902–918.
- Uyeda, J. C., Hansen, T. F., Arnold, S. J., & Pienaar, J. 2011. The million-year wait for macroevolutionary bursts. *Proceedings of the National Academy of Sciences*, 108(38): 15908–15913.
- Walsh, B. & Lynch, M. 2018. *Evolution and selection of quantitative traits*. Oxford University Press.
- Weaver, T. D., Roseman, C. C., & Stringer, C. B. 2007. Were neandertal and modern human cranial differences produced by natural selection or genetic drift? *Journal of Human Evolution*, 53(2): 135–145.
- Wilder, A. P., Supple, M. A., Subramanian, A., . . . , Shapiro, B. 2023. The contribution of historical processes to contemporary extinction risk in placental mammals. *Science*, 380(6643): eabn5856.