Evolutionary dynamics of neutral phenotypes under DNA substitution models

Shadi Zabad[1] and Alan M Moses[1,2,3,4]

Departments of Computer Science[1], Cell & Systems Biology[2], Ecology and Evolutionary Biology[3], and Centre for Analysis of Genome Evolution and Function[4], University of Toronto, Canada

Abstract

We study the evolution of quantitative molecular traits in the absence of selection. Using Felsenstein's 1981 DNA substitution model, we predict a linear restoring force on the mean of an additive phenotype. Remarkably, the mean dynamics are independent of the effect sizes and genotype and are similar to those predicted by the widely-used OU model for stabilizing selection and the house-of-cards model for phenotype evolution. We confirm the predictions empirically using additive molecular phenotypes calculated from ancestral reconstructions of putatively unconstrained DNA sequences in primate genomes. We predict and confirm empirically that the dynamics of the variance are more complicated than those predicted by the OU model, and show that our results for the restoring force of mutation hold even for non-additive phenotypes, such as number of transcription factor binding sites, longest encoded peptide and folding propensity of the encoded peptide. Our results have implications for efforts to infer selection based on quantitative phenotype dynamics as well as to understand long-term trends in evolution of quantitative molecular traits.

Introduction

With the increasing availability of high-throughput data about molecular and cellular function from many species, a key challenge is detect selection on these quantitative molecular traits [1]–[9]. The models used for selection in the comparative phylogenetic approach have rich theoretical grounding [10], [11] and software packages are available for data analysis under a variety of inference frameworks [12]–[14]. However, despite evidence that molecular phenotypes may evolve with little selective constraint [15]–[17] and classical work in the area[18], relatively little attention has been paid to improving the realism of the "neutral" or "null hypotheses" for phenotype evolution, as compared to protein coding sequences[19], [20] where the null hypothesis can be explicitly formulated based on empirical estimates from neutral sites [19] or, more recently, the biophysics of protein folding (e.g.,[21]).

The widely-adopted Brownian Motion (also referred to as a "random walk"[22]) model is an appealing null hypothesis for phenotype evolution in comparative analysis of quantitative molecular traits[8]. It is mathematically convenient [23] and is referred to as a model of "pure drift" [1], [13] because it matches quantitative trait evolution in theory under genetic drift alone[23]. A key prediction of the Brownian Motion model is that changes from the ancestral phenotype are symmetric, and therefore, on average, mean phenotypes are not expected to change over time[24]. On the other hand, it has long been appreciated that even in the absence of selection, mutation is expected to be a weak directional force on phenotype evolution[22], [25] and this has been observed empirically in at least one recent mutation accumulation study[26].

This view is captured in the so-called house-of-cards model [27], which was meant to capture the idea that the force of random mutation would be expected to break down what had been built up by selection. More recently, it has been suggested that mutation could be a constructive force in evolution, leading to non-adaptive increases in biological complexity[28], [29].

With the increasing interest in studying evolution of molecular and cellular traits over long evolutionary time-scales, several theoretical studies have begun to take a closer look at the expected effects of mutation on the phenotype, finding that the equilibrium phenotype is the result of a trade-off between the directional effects of mutation and selection [30]–[32], and that biased mutation can drive the phenotype away from the optimum when selection is not strong[33]. To our knowledge, these predictions have not yet been compared directly with data, but global gene expression variation in mutation accumulation studies appears more consistent with the house-of-cards model[34]. Perhaps most intriguingly, a recent empirical study that estimated the effects of mutations on gene expression levels for 10 genes found asymmetric distributions (inconsistent with Brownian Motion) and used simulations to predict approximately linear change over time in the absence of selection for most genes [35].

In order to gain insight into the appropriate null hypothesis for use in comparative studies of quantitative molecular phenotypes, we first develop a neutral model of phenotype evolution that could be compared directly to observations. We work out the dynamics of additive phenotypes in the weak-mutation regime[36], the regime of most relevance for molecular evolutionary studies. Under Felsenstein's 1981 model of DNA evolution[37], we predict that mutation acts as a linear restoring force on the mean phenotype, similar to the OU models currently used for stabilizing selection[10] and to the house-of-cards model for phenotype evolution[22]. Next, we develop a paradigm based on ancestral genome reconstruction to directly compare the predictions of these models to observations of molecular phenotype evolution. We find remarkable agreement with observations of molecular phenotypes computed from reconstructed ancestors of putatively neutral sequences in primate genomes, even when the phenotypes are not strictly additive.

Results

Neutral dynamics of the mean of an additive phenotype under a DNA substitution model

Our goal is to develop a neutral model of phenotype evolution with sufficient realism so that it can be directly compared to observations of molecular phenotypes obtained from closely related extant species. We therefore sought to derive the dynamics of the moments of the phenotype distribution starting with a standard model of DNA substitution. Thus, we model phenotype evolution as a consequence of genetic changes due to mutation and neutral fixation alone.

We denote the mean phenotype in a population or species with Z and we assume that it linearly depends on the underlying genotype matrix X. In our model, X is an L x K binary matrix, where L is the number of contributing loci and K is the number of distinct alleles at each locus. The binary coding of the genotype matrix follows from our assumption of a monomorphic haploid population, so that $X_{jk} = 1$ if the k-th allele is fixed at the j-the locus and zero otherwise. Next, we use β_{jk} to represent the effect size of the k-th allele at the j-th locus. With this, we can write the mean phenotype in a population or species as:

$$Z = \sum_j \sum_k X_{jk}\beta_{jk} = \sum_j X_j \cdot \beta_j$$

For example, for a DNA-based phenotype with L=10 contributing loci, X and β are 10 x 4 matrices. In order to compare with empirical phenotype data, we allow arbitrary values for β_{jk}, thus making no assumptions about the distribution of effect sizes.

We are interested in examining the dynamics of the mean phenotype as a function of time Z(t). If we assume that the effect sizes remain constant, the time-dependent variable in our model is the underlying genotype matrix X(t):

$$Z(t) = \sum_j \sum_k X_{jk}(t)\beta_{jk} = \sum_j X_j(t) \cdot \beta_j$$

Characterizing the evolution of the genotype matrix in response to mutation, selection, and drift has been a central focus of population genetic theory for the past century[38]. In our model, we make the simplifying assumptions that the only operating evolutionary forces are mutation and drift and that the loci evolve independently according to a continuous-time Markov chain [20] with K states corresponding to the K different alleles. In assuming that a monomorphic initial genotype mutates into another monomorphic genotype we are neglecting the possibility of polymorphism, competition and linkage between genotypes and other population genetic processes. This regime has been referred to as the sequential model [39] and is equivalent to the so-called "weak mutation" assumption commonly used to model molecular evolution at the timescale of inter species divergence [36].

In this work, we use Felsenstein's 1981 model (F81) of DNA evolution [37] to model the evolution of the individual loci. For each allele k, the model defines the following transition probabilities

$$P\big(X_j(t) = k \big| X_j(0) = k, t\big) = e^{-ut} + \pi_k(1 - e^{-ut})$$

$$P\big(X_j(t) = k \big| X_j(0) \neq k, t\big) = \pi_k(1 - e^{-ut})$$

Where $X_j(t) = k$ denotes that the k-th allele is fixed at the j-th locus at time t and $u \stackrel{\text{def}}{=} \frac{1}{1-\sum_m \pi_m{}^2}$ is the DNA substitution rate, scaled so that evolutionary distance, *ut*, is measured in substitutions per site. Given an ancestral population with mean phenotype Z(0), mutation and drift will fix some alleles while driving others to extinction. Assuming that we have an ensemble of replicate populations emanating from that ancestral population, our goal is to provide expressions for the ensemble mean E[Z(t)] and variance, V[Z(t)] of the mean phenotype over time.

In the appendix, we provide detailed derivations that show that the ensemble mean has the following form

$$E[Z(t)] = e^{-ut}Z(0) + (1 - e^{-ut})Z_{eq} = e^{-ut}(Z(0) - Z_{eq}) + Z_{eq}$$

Since we measure evolutionary distance in units of substitutions per site, for evolutionary times of interest in the primate phylogeny considered below, $ut \ll 1$, so $e^{-ut} \approx 1 - ut$. Hence

$$E[Z(t)] \approx ut\big(Z_{eq} - Z(0)\big) + Z(0)$$

Where $Z_{eq} \stackrel{\text{def}}{=} \sum_j \sum_k \pi_k \beta_{jk}$ and can be interpreted as the phenotype expected after a long-time of neutral evolution, or the phenotype at "mutational equilibrium." Thus, under the F81 model, the dynamics of the mean phenotype depend only on the initial phenotype, the phenotype at its mutational equilibrium and the evolutionary distance. Remarkably, we find that the dynamics of the ensemble mean of an additive phenotype is predicted to be *independent* of the starting genotype, X. In other words, the dynamics of the mean phenotype are *the same for all population genotypes* that encode that phenotype. We conjecture that the F81 model is the most realistic DNA substitution model[20] for which this can be true (for example, it is not true of models with transition-transversion rate bias[40], see Discussion). The dynamics of the mean phenotype are also *independent* of the number of alleles, loci, and the distribution of allelic effect sizes, β.

Our model predicts that mean phenotype is driven towards the equilibrium with a force simply proportional to the distance of the initial phenotype from its mutational equilibrium, $Z(0) - Z_{eq}$. Thus, near the mutational equilibrium, mutation can be neglected, but the further the phenotype is from equilibrium, the stronger the directional force of mutation becomes. The appearance of a directional force on the mean phenotype in the absence of selection is contradictory to the Brownian Motion null hypothesis for phenotype evolution[23], which argues that in the absence of selection, the mean phenotype is not expected on average to change in one direction or the other[23]. Importantly, the dynamics of the mean phenotype match exactly the dynamics of the mean of the well-studied Ornstein-Uhlenbeck (OU) process, a stochastic model that is widely used to capture the restoring force of stabilizing selection in phylogenetic comparative analysis[10], [11], [13]. If our prediction of OU-like mean dynamics in the absence of selection is correct, naïve use of the OU process to model selection may be misleading (see discussion).

In the appendix, we provide detailed derivations that show that the ensemble variance of the mean phenotype has the following form:

$$V[Z(t)] = (1 - e^{-ut})(e^{-ut}(\Psi(\mathrm{X}) - \sigma_{eq}^2) + \sigma_{eq}^2)$$

Once again, for evolutionary times of interest in the primate phylogeny considered below, $ut \ll 1$, so $e^{-ut} \approx 1 - ut$. We have

$$V[Z(t)] \approx ut\Psi(\mathrm{X})$$

In these expressions, $\sigma_{eq}^2 \stackrel{\text{def}}{=} \sum_j \sum_{k,m} \big(\beta_{jk} - \beta_{jm}\big)^2 \pi_k \pi_m$, which is the ensemble variance at equilibrium, and $\Psi(\mathrm{X}) \stackrel{\text{def}}{=} \sum_j \sum_{k,m} \big(\beta_{jk} - \beta_{jm}\big)^2 \left(X_{jk}(0)\pi_m + X_{jm}(0)\pi_k\right)$ is a new, genotype-dependent quantity whose meaning is not entirely clear, but it captures something about the dispersion of the evolutionary process (see Appendix). In the sums, k, m index the unique pairs of alleles. The expressions for the variance make immediately clear that at long times the variance reaches an equilibrium value, σ_{eq}^2, and at short times it increases proportional to

evolutionary distance. Importantly, unlike the mean phenotype, whose dynamics depend only on phenotypic quantities, in general, the variance of the mean phenotype *does* depend on the initial genotype through the quantity $\Psi(\mathrm{X})$. In other words, populations with initially identical mean phenotypes, but different genotypes, are predicted to show *identical mean* phenotype evolution, but will potentially show *different variance* evolution. Furthermore, the variance does not always increase monotonically to its equilibrium value. These predictions differ qualitatively from both the Brownian Motion and OU process models typically used in comparative phylogenetic analysis.

Under the assumptions that the genotypes in the population are initially at mutational equilibrium (not likely to be the case for real populations) or that the phenotype is encoded by bi-allelic loci with uniform effect sizes and evolves with symmetric mutation rates (see Appendix) we find that $\Psi(\mathrm{X}) = 2\sigma_{eq}^2$ and $V[Z(t)] = \sigma_{eq}^2(1 - e^{-2ut}) \approx 2ut\sigma_{eq}^2$ (with the approximation being for $ut \ll 1$ as above.) Thus, under these assumptions, we do obtain genotype-independent dynamics, whose short-time behavior agrees with the predictions of the Brownian-Motion model [41].

Although in general we predict some differences in the variance dynamics between the F81 and the OU models, we wondered under what circumstances the OU process might provide a good approximation to the variance dynamics of the F81 model. Under the OU model, the variance is

$$V[Z(t)] = \frac{\sigma_Z^2}{2u}(1 - e^{-2ut}) \approx \frac{\sigma_Z^2}{2u}2ut = \sigma_Z^2 t$$

To figure out the fluctuation size, σ_Z^2, we note that the long-time variance should go to $\frac{\sigma_Z^2}{2u} = \sigma_{eq}^2$. Therefore, under an OU model with only the force of mutation, $\sigma_Z^2 = 2u\sigma_{eq}^2$ and

$$V[Z(ut \ll 1)] \approx 2ut\sigma_{eq}^2$$

Thus, the OU model (with mutation alone) also matches the predictions of the F81 model exactly under the assumptions of initial genetic equilibrium, or symmetric biallelic mutation with constant effect size. We note, however, that in the F81 model, the force of mutation, u, is never separate from the time, t, implying that these cannot be inferred separately[37]. In the OU model, these appear separately, which could in principle lead additional complexity in inference.

Molecular phenotypes from ERVs in primate genomes

Since our goal is to test empirically which neutral model should be used in practice, we considered phenotypes that could be computed directly from DNA sequences: these represent so-called quantitative molecular traits[5] with known genetic architecture and no measurement noise. This experimental set-up is attractive for several reasons: we can be sure that the phenotypes are truly "additive" (therefore not violating the assumptions of the model), we can rule out that stochastic effects we observe are intrinsic to the evolutionary process and not due to measurement or sampling errors[42], [43], and we can compare the observations from real DNA sequences to comparable observations from forward simulations under the DNA substitution models to determine the effects (if any) of mis-specification of the substitution model.

To obtain neutral phenotypes, we analyzed alignments of endogenous retroviruses (ERVs, see Methods), which are well-annotated ancient pseudogenized copies of retroviruses that are no longer replicating, but are easily identifiable and distributed in large numbers over the human genome[44]. We treat these sequences as independent "replicates" of the evolutionary process and compute the mean and variance of quantitative molecular phenotypes derived from these sequences. We use reconstructed ancestral DNA sequences (based on alignments of closely related primates) along the lineage leading to human, and infer the ancestral phenotypes from these. This allows us to directly study the forward evolution of phenotypes in an evolutionary ensemble without relying on model assumptions for parameter inference.

Dynamics of GC content and TATA box strength confirm the directional force of mutation

We first considered the "simplest" possible DNA-based phenotype: GC content, a phenotype that has been studied using comparative phylogenetic methods, e.g., [45]. We computed GC content in 100 nucleotide segments extracted from ERVs (see Methods) binned by the GC content of the inferred ancestral sequence. As expected for a simple restoring force of mutation, qualitatively, sequences with relatively high GC content (Figure 1a, triangles) show a decrease over evolutionary time, while sequences that start with a low GC content show an increase in GC content over time (figure 1b, filled squares). As predicted, the decrease at these short evolutionary distances appears linear.

To compare quantitatively, we express GC content in the notation of our theory. The GC content of the ancestral sequences in the bin represents $Z(0)$ and we used $\pi = (\pi_A, \pi_C, \pi_G, \pi_T) = (0.32, 0.18, 0.18, 0.32)$ which gives $u \stackrel{\text{def}}{=} \frac{1}{1-\sum_k \pi_k{}^2} = 1.37$. For GC content in a sequence of length L, each locus contributes 1/L if it is G or C, and 0 otherwise. This means $\beta_j = (\beta_A, \beta_C, \beta_G, \beta_T) = \left(0, \frac{1}{L}, \frac{1}{L}, 0\right)$ for all j. From these we compute the other key parameters for GC content

$$Z_{eq} \stackrel{\text{def}}{=} \sum_{j=1}^{L} \pi \cdot \beta_j = \pi_C + \pi_G = 0.36$$

$$\sigma_{eq}^2 = \frac{1}{L}(\pi_C + \pi_G)(1 - \pi_C - \pi_G) = 0.002304$$

In the Appendix, we show that in the case of GC content, $\Psi(\mathrm{X})$ takes a genotype independent form, i.e., $\Psi(\mathrm{X}) = \Psi\big(Z_{eq}, Z(0)\big) = \frac{1}{L}(Z(0) - 2Z(0)Z_{eq} + Z_{eq})$, so the dynamics of the variance of GC content can also be predicted from the quantities above.

We found good agreement between the theory based on the F81 DNA substitution model (dashed lines in Figure 1a,b) and the observed changes in GC content. We also used simple linear regression to infer the change in GC content over time as an estimate of the evolutionary force (see Methods). Once again, we found good agreement between the theory and the observations, (Figure 1c) with some slight deviations for sequences with ancestral GC content much different than equilibrium. We believe these deviations are due to the misspecification of our DNA substitution model which does not include CpG bias in the mutation rate or transition-

transversion rate bias. Finally, we note good agreement between the variance predicted by the F81 neutral model and the observed increase in variance for GC content (Figure 1d). Given the range of observed GC content and the parameters values above, the rate of increase in variance predicted by the F81 model is within 10% of the prediction of the OU process ($2\sigma_{eq}^2$), so we lack power to distinguish between these models with these data. This suggests that the OU process (with restoring force set to u as defined above) provides a good approximation to the neutral evolution of GC content.

To rule out possible circularity due to the use of probabilistic models of evolution in ancestral genome reconstruction [46], we repeated this analysis using ancestral sequences reconstructed using maximum-parsimony, which makes no assumptions about the relative likelihood of substitutions between different nucleotides. We found qualitatively similar results (Supplementary Figure S1).

Next, we repeated these analyses for the strength of matches to the TATA box ([47], see Methods) in the first 10 residues of each ERV segment. Because positions in transcription factor binding sites contribute approximately linearly to affinity[48], this represents a simple biochemical phenotype whose evolution is well studied[4]. For clarity we note that in our case none of these sequences are expected to have functioning TATA boxes: these are sequences with similarity to the TATA box that arise by chance[49]. The strength of matches to the TATA box is computed using a 4 x 10 weight matrix model, which corresponds to a in our model (see Methods). We bin the sequences by the average strength of TATA box in the inferred ancestor, $Z(0)$, and again we used $\pi = (0.32, 0.18, 0.18, 0.32)$ which implies $u = 1.37$. The other parameters are $Z_{eq} = -9.223$ and $\sigma_{eq}^2 = 24.6$, which are obtained directly from the matrix model (see Methods). We found good agreement with the predictions of the model (Figure 2a-c). In this case, we cannot obtain a genotype independent form for $\Psi(\mathrm{X})$, so we do not have a simple prediction for the variance. Nevertheless, once again the linear, constant increase in variance seems to agree reasonably well with the prediction of the OU process (Figure 2d).

Dynamics of ATGs show the predicted mutational force, but no longer match an OU process

We next considered a sequence-based trait where we expected to find example sequences that are far from equilibrium: the number of ATG start codons in 100 bp segments. Sequences can only encode a natural number, 0, 1, 2 … ATGs. Hence even a single point mutation can have a large effect on the phenotype. Once again, we emphasize that we do not expect these ERV sequences to encode functional proteins, so these are simply ATGs that occur by chance or may have functioned in the ancestral viral proteins.

As for GC content and strength of TATA boxes, we found that the number of ATGs in neutral sequences shows a clear mean-reverting force. For example, in ancestral sequences that have exactly 4 ATGs, we see the number of ATGs decreasing over time (Figure 3a, triangles). This is intuitive as we expect random mutation to destroy these "informative" signals over time. Less intuitive is the analogous result for ancestral sequences that start with no ATGs: in these sequences we see (on average) the accumulation of ATGs over evolutionary time (figure 3b filled squares). Evolution appears to be creating start codons, such that overall 142 ATGs are

found in human ERV segments that had none in the primate ancestor. Again, we emphasize that the creation of ATGs is not the result of any selection or biological function, rather, in these sequences ATGs are simply being created by random mutations faster than they are being destroyed. We note that the temptation to create adaptive explanations for the appearance of ATGs in these sequences is very strong (see Discussion), but we confirmed that the dynamics of ATGs are similar in simulations where we can be certain there is no selection or function.

To predict the dynamics of the number of exact matches to a short sequence, which is not strictly an additive phenotype, we treat the DNA sequences as a series of overlapping w-mers, which we assume are independent; in the case of ATGs, $w = 3$. This means that we imagine 4^w possible alleles at each locus, and $\beta_j = 1$ for the short sequence of interest, and 0 otherwise. If we assume that each DNA letter still evolves independently and at the same per-site rate, the effective evolutionary rate between these alleles will be $wu = \frac{w}{1-\sum_k {\pi_k}^2}$. Furthermore, with $4^w \gg 1$, as long as the mutation process is not too biased, ${\pi_m}^2 \to 0$, so $u \to 1$ and the mutational restoring force is simply w. As before, the number of ATGs in the ancestral sequences in the bin represents $Z(0)$. The equilibrium phenotype in this model can be computed exactly

$$Z_{eq} \stackrel{\text{def}}{=} \sum_{n\epsilon 1, L-w+1} \sum_{m} \pi_m \beta_m = (L - w + 1)\beta \cdot \pi = (L-2)\pi_A \pi_T \pi_G = 1.806$$

As with GC content and strength of TATA boxes, we find very good agreement between the theory and the mean dynamics (dashed lines in figures 4a,b,c) and clear evidence for the linear restoring force with strength simply equal to 3 start codons per substitution per 100nt sites (Figure 3c). We note that the agreement between the observations and theory must be approximate in this case, because the exact dynamics of the number of ATGs is strongly genotype dependent: initial sequences with many "one-off" sequences (e.g., ATC, ATT, ATA, AAG, etc.) are much more likely to increase their number of ATGs than sequences with few of these "one-off" sequences. We believe that because we are measuring the average over a large number of initial sequences, these effects are averaged out (see Discussion).

Since the number of exact matches is approximately binomial, the long-time variance will be approximately

$$\sigma_{eq}^2 \approx (L-2)\pi_A \pi_T \pi_G (1 - \pi_A \pi_T \pi_G) = 1.773$$

Unlike for GC content and strength of TATA boxes, we observed that the variance in number of ATGs was not independent of the starting phenotype $Z(0)$, and therefore inconsistent with the OU prediction (dashed line in figure 4a). We found that sequences with more ATGs than the equilibrium (4 ancestral ATGs, triangles in figure 4a) showed a faster increase in variance than sequences with fewer (0 ancestral ATGs, filled squares in figure 4a). We therefore looked at the ancestral bin that was closest to the mutational equilibrium (2 ancestral ATGs, x's in figure 4a) and found that the variance increased in very good agreement with the OU prediction, consistent with our theoretical prediction that the F81 model matches the OU dynamics when the phenotype is at equilibrium (Appendix).

Our approximate model for the number of ATGs described above corresponds to a simple phenotype, which is just the counts of the k-th allele [30]. For this phenotype, like for GC content, the F81 neutral dynamics of the variance are independent of the initial genotype, X, but they are still more complicated that the OU model predicts (see Appendix). In this case, $\Psi(X) \approx Z_{eq} + Z(0)$, with the approximation valid for $\pi_k \ll 1$. At short evolutionary distance, and since for ATGs, $\pi_k \ll 1$, we obtain a simple approximation for the variance of number of ATGs under the F81 model:

$$V[Z(ut \ll 1)] \approx ut(Z_{eq} + Z(0))$$

To our knowledge this simple prediction for the rate of increase of variance based on the initial phenotype has not been reported under the house-of-cards model[22]. Comparing this to our observations for the variance of ATGs (with u=3) gives remarkably good agreement, and better predicts the linear increase in variance than the OU model (Figure 4b,c). Taken together, these results confirm the prediction above that the true dynamics for molecular phenotypes do not match an OU process when the phenotype is strongly sufficiently far from the mutational equilibrium.

Neutral dynamics of more complex molecular phenotypes also show a linear restoring force

Many of the molecular phenotypes of interest (e.g., gene expression level) are much more complicated than GC content or strength of binding and are hard to express as additive phenotypes. Nevertheless, a recent study of the mutational effects on gene expression used simulations parameterized by empirical measurements to show simple linear dynamics of gene expression phenotypes[50] even though the traits are unlikely to be linear. To test the generality of our finding of a mutational force on the mean phenotype, we next empirically studied the neutral dynamics of non-additive phenotypes that can be computed from sequence. We chose three phenotypes that are of more biological interest: the number of TATA boxes in a 100 bp sequence, the length of the longest encoded peptide, and the intrinsic disorder in the longest encoded peptide. The first of these relates to how non-coding sequences evolve so-called homotypic clusters of binding sites[51], [52], and the latter two of these phenotypes are related to the emergence of new protein-encoding genes from random DNA [53], [54]. To obtain these, we computed the lengths of six-frame translations from our 100-basepair ERV segments, and the propensity of the peptide to fold ([55], see methods). We emphasize that these phenotypes are highly non-linear in the DNA sequence genotype, and therefore strongly violate the assumptions used to derive our results above.

Remarkably, we found that these phenotypes also showed simple linear mean dynamics, albeit with stronger evolutionary forces pushing the phenotypes to their mutational equilibrium (Figure 5a) than the additive phenotypes considered above. The force appears to be different for each phenotype, but we have no theory to predict it. To a large extent, the evolutionary force appears simply proportional to the distance of the ancestral sequence from the equilibrium (Figure 5b) and the quantitative strength of the force is less than 10 units per 100 bp DNA sequence. Thus, non-additive phenotypes also appear to show a simple force of mutation proportional to the distance of the ancestral phenotype from the mutational equilibrium.

The BM null hypothesis cannot always be rejected using standard inference approaches on GC content in ancient ERVs

To test whether the predicted and observed deviations from the Brownian Motion (BM) null hypothesis were strong enough to be detected in a typical comparative phylogenetic analysis, we implemented standard Gaussian maximum likelihood inference ([13], described in the Appendix; experiments using the OUCH package [43] for inference did not converge reliably, and led to errors in estimation). We obtained six well-aligned fragments from an alignment of 39 mammals (see Methods, supplementary Figure S2) for an unusually ancient ERV that is thought to evolve in the absence of selection[56]. We found that corrected AIC was the lower for BM than for OU for all 6 segments (Table 1) because the likelihood of the data under BM and OU was very similar, so the simpler model was chosen. We also implemented maximum-likelihood inference for the F81 neutral model (See Appendix and Methods), and compared it (with parameters set to match the GC content of the ancient ERVs) to BM. We found that it was favoured for two of the 6 segments, BM favored in another two, and the other two showing similar corrected AICs. These results suggest that despite theoretical and empirical evidence for a mutational force, in an inference set up typical of phylogenetic comparative analysis, the BM model, which does not include this force, fits the data as well as models (OU and F81) that do include this force. We note that these results cannot be explained by noise [42]in our phenotypic measurements, as GC content is computed exactly from sequences.

Although this ERV is thought to evolve in the absence of selection[56], our prediction of the dynamics of the mean and variance of GC content is based on a simple DNA substitution model, which only captures some features of sequence evolution. We wondered if mis-specification of the neutral model could explain the lack of evidence for mutational force in the ancient ERVs. We therefore simulated molecular evolution of these ERV sequences under a simple neutral model of DNA substitutions (see Methods) and found much clearer support for models that do include the force of mutation: for the simulated data, 4/6 segments favour OU over BM, and 5/6 favour the F81 neutral model over BM (Table 1). These results confirm that simple substitution models do produce a measureable force of mutation, and suggest that there are other aspects of the neutral evolutionary processes in the real ERV sequences that is not accounted for by either OU or F81 models, but is better captured by Brownian Motion (see Discussion).

Discussion

Our results show that mutation is expected to aid the creation of molecular phenotypes *de novo*[28] and relate to several areas of research in molecular evolution. For example, in agreement with simulation results [57], we find that mutation alone is sufficient to create transcription factor binding sites in DNA sequences, and we quantify this effect: 100 nt sequences that have no strong matches to the TATA box are expected to accumulate them at a rate of approximately 0.5 TATA box matches per substitution per site. Similarly, in random 100 nt sequences with no open reading frames, mutation will tend to create open reading frames at a rate of >50 codons per substitution per site. Even if the encoded peptides start out as strongly disordered[54], the force of mutation is expected on average to increase their tendency for folding[53]. Remarkably, but consistent with a recent report [35], even though all of these

molecular phenotypes are non-additive, they show simple linear change over time, following our predictions for additive phenotypes. We believe that this is because our ERV sequences are relatively near the mutational equilibrium for these phenotypes, and speculate that the evolutionary dynamics are approximately linear near the mutational equilibrium. It will be of great interest to extend the theoretical work to explain these observations.

Unlike previous theoretical work which focused on quantifying the force of mutation on phenotype evolution by including mutational bias[33], [58], here we used the F81 substitution model, which is formulated in terms of the (possibly non-uniform) equilibrium frequencies for the DNA bases (alleles). Our results suggest that the mutational equilibrium phenotype, rather than the mutation rate bias, plays the key role in determining the neutral dynamics of phenotypes, consistent with recent theoretical results[30], [33]. Although the F81 model does have mutation bias (the rate bias from allele m to allele k is π_m/π_k; there is no transition-transversion bias or CpG hypermutation), we also considered the Jukes-Cantor model[20], which is a specific instance of the F81 model where all π_m=1/4 (u=4/3) for DNA, and therefore shows no mutation bias. Importantly, even under this simplification, we still predict a restoring force proportional to the distance from the mutational equilibrium, although some simplification of the dynamics is obtained (e.g., the neutral variance dynamics of GC content is no longer genotype dependent). Even simplified bi-allelic (non-DNA based) models with no mutation bias show a restoring force of mutation when far from mutational equilibrium (Appendix), strongly suggesting that the force of mutation in phenotype evolution is not due to mutation rate bias, but is a more fundamental result of the mapping between the discrete genotype space and the continuous (or ordinal) phenotype. We did not pursue these simplified models further because the predictions of the dynamics were not close to our observations from ERV sequences, presumably because the predicted mutational equilibrium is too far from the real data (e.g., equilibrium GC content of 50%, rather than 36%). This highlights the need to develop simpler (approximate) theory that can still be quantitatively compared to data.

The dynamics of the mean phenotype in the F81 model are similar to those expected under the house-of-cards mutation model[22], a model defined and derived using only phenotypic quantities. The similarity is most obvious if the mean of the phenotype effect distribution in the house-of-cards model is taken to be the "mutational equilibrium" phenotype, Z_{eq}, defined above. To understand how this similarity arises, we note that if the F81 DNA substitution model is viewed as a (4-dimensional) phenotype model, the dynamics of the genotype probabilities match the mean dynamics of a house-of-cards mutation model. Since the mean of an additive phenotype is a linear combination of the genotype probabilities, it follows the same dynamics.

For GC content and strength of matches to TATA boxes we found agreement with the predictions of the variance dynamics based on the OU process, even though we can show that this is only approximate (See Appendix). We believe this is because we measured these phenotypes in ERVs that are close enough to mutational equilibrium, where we expect the dynamics to be well approximated by the OU process. However, neutral phenotype evolution models (such as the F81 model developed here) should accommodate phenotypes that may have been under strong selection pressure in the past, making it unlikely that the initial ancestral

phenotype is close to mutational equilibrium at the time selection is relaxed. This scenario is consistent with observations from mutation accumulation studies[26], where once selection is weakened, phenotypes evolve in a directional manner.

Despite the clear empirical evidence for the force of mutation on GC content at short evolutionary distances when we averaged over thousands of sequences, we found inconclusive results when comparing models using a standard maximum likelihood/AIC inference analysis of GC content in ancient ERVs. In our experiments Brownian Motion, which does not include the force of mutation, fit the data as well as models that include the force of mutation. We speculate that a considerable portion of the variance in GC content at longer evolutionary distances is due to neutral processes not included in simple DNA substitution models. Understanding these sources of variance is an area for further research.

More generally, the similarity of the dynamics of neutral phenotypes to an OU process suggests that at time-scales considered here, similarity of phenotype evolution to OU process should not be used as evidence for purifying selection (as suggested in[1]). Further, since even neutral phenotypes show more complicated variance dynamics (genotype-dependent and non-monotonic) than the OU model predicts, our results add the possibility of misspecification of the underlying models[59] to the previously reported challenges with phylogenetic inference of OU models[42], [43], [60]. Developing tests based only on the mean phenotype (which does seem to match the OU process assumptions remarkably well) is an area of promise, though it remains unclear how much information about mean dynamics is contained in the observations at the tips of the tree[60]. On the positive side, our results are consistent with the idea that mutation and selection act in a simple additive way on the evolution of mean phenotypes[11], [30], both leading to OU-like dynamics. Further, our results limit the quantitative range of restoring force expected due to mutation alone: if the restoring force estimated in an OU model exceeds the predicted value, we believe this can be interpreted as stabilizing selection. Similarly, large changes in the optima in multi-optimum OU models [10] are unlikely to result from mutation alone.

Finally, our approach of using reconstructed ancestral sequences to empirically study the forward evolution of phenotypes is applicable whenever the genetic architecture of a trait is known. With the increasing power of association and other high-throughput studies to determine the loci and their effect sizes for many phenotypes[61], [62], additive models of traits can be inferred. Furthermore, machine learning methods trained on genome-scale data and massively parallel assays may be able to predict non-additive molecular phenotypes from sequences [63], [64]. Once a genotype to phenotype model is defined, observed phenotype evolution can be compared directly to neutral phenotypes obtained from neutral sites (such as ERVs[44]) or simulations at the sequence level[65]. The simulation approach has been applied successfully to molecular traits that can be computed directly from intrinsically disordered protein sequences[66], [67]. If applied more generally to quantitative characters, this could be viewed as an extension to practice of inferring phylogenetic trees from genotypes, even when studying phenotypes [18], [24]: the null hypotheses for phenotype dynamics can also be obtained using our understanding

of genotype evolution, further cementing the inevitable merger of systematics and evolutionary genetics[18].

Methods

Unconstrained sequences in primate genomes

We obtained 6362 coordinates for predicted ERVs in the human genome from the gEVE database[68], corresponding to a subset of ERV segments that span more than 100 nucleotide residues and contain no ambiguous letters. We obtained alignments of these along with their reconstructed ancestors[46] from the 100 vertebrate alignment tree from Ensembl, we extracted the species in the aligned segments using the Ensembl compara REST API[69]. As a control for possible circularity due to probabilistic models of sequence evolution used in ancestral genome reconstruction [46], we also performed analysis on ancestral sequences reconstructed using the maximum parsimony criterion as implemented in the phangorn package[70].

We used the extant human sequence, as well as the following reconstructed ancestors using the naming convention from ensembl: Hsap-Ppan-Ptro[3], Ggor-Hsap-Ppan-Ptro[4], Ggor-Hsap-Pabe-Ppan-Ptro[5], Ggor-Hsap-Nleu-Pabe-Ppan-Ptro[6], and Csab-Ggor-Hsap-Mfas-Mmul-Nleu-Pabe-Panu-Ppan-Ptro-Tgel[11] (square brackets in these names refer to the ancestor number in the ensembl tree are not references). For each ERV alignment we extracted segments of at length 100 with <10% gapped columns, required all sequences to be present in each alignment, yielding 8386 segments. Csab-Ggor-Hsap-Mfas-Mmul-Nleu-Pabe-Panu-Ppan-Ptro-Tgel[11] was considered the "primate ancestor" and the value of the phenotypes computed based on this sequence was used as the ancestral phenotype, Z(0). Binning was done based on this value, and each of the descendent sequence phenotype values were assigned to this bin. Phenotype mean and variance were calculated for all sequences in the bin, and these sample sizes are reported as N for bins shown in figures 1, 2 and 4.

For each "descendent" sequence, we computed the evolutionary distance in substitutions per site from the ancestor using the Juke-Cantor model and used the average of these as the evolutionary distance for that phenotype bin. The common ancestor was always assigned a distance of zero.

To estimate the mutational restoring force and rate of variance increase, we used simple linear regression to estimate the slope of the regression of the mean phenotype and variance of the phenotype on time (in substitutions per site). Because the ancestral value was determined by the binning process, and the evolutionary time for the ancestor was zero by construction, to avoid potentially biasing the estimate of the slope, we excluded the common ancestor from the regression and used the remaining 5 datapoints. R^2 for these was typically >0.9.

Strength and number of TATA boxes

We obtained the matrix model of TBP from the Jaspar database [47].

$p(X_{jk} = 1|motif) =$

0.157760814 0.371501272 0.389312977 0.081424936

0.043256997	0.119592875	0.048346056	0.788804071
0.89821883	0.002544529	0.007633588	0.091603053
0.010178117	0.027989822	0.007633588	0.954198473
0.903307888	0.002544529	0.015267176	0.078880407
0.684478372	0.002544529	0.002544529	0.31043257
0.942558747	0.010443864	0.028720627	0.018276762
0.567430025	0.007633588	0.114503817	0.31043257
0.396946565	0.114503817	0.402035623	0.086513995
0.145038168	0.34605598	0.384223919	0.124681934

Where $p(X_{jk} = 1|motif)$ is the probability of observing the kth allele at the jth position in the transcription factor binding probability matrix.We assumed a position independent background model, as above, $\pi = (0.32,0.18,0.18,0.32)$, so we define $\beta_{jk} = \log\frac{p(X_{jk}=1|motif)}{\pi_k}$. For the strength of TATA box analysis, we used the strength of the match in the first 10 bp of each ERV alignment and used the standard log-likelihood ratio score, S, as a measure of binding strength[71]. To compute the other parameters for the strength of TATA boxes, which have width w=10, we used

$$Z_{eq} = \sum_{j=1}^{w}\sum_{k}\pi_k\beta_{jk} \equiv \sum_{j=1}^{w} Z_{j,eq}$$

$$\sigma_{eq}^2 = \sum_{j=1}^{w}\sum_{k}\pi_k\left(\beta_{jk} - Z_{j,eq}\right)^2$$

For the expected number of TATA boxes in the entire sequence, we used the sum of the posterior probability of each position being a match to the motif[71], $Z(X) = \sum_{j=1}^{L-w+1}\frac{1}{1+e^{-S_j-R}}$, where each S_j represents the standard likelihood ratio score [71] for the subsequence of length w starting at position j, , and R is the (position independent) log odds of the prior probabilities of observing a match to the motif or not at each position. $R = \log\frac{p(motif)}{1-p(motif)}$. We used 1/100 for this prior, so $R = \log\frac{1}{99}$..

Alignment, phylogeny, simulation and inference for ancient mammalian ERVs

we retrieved the human ERV sequence [56] and used BLAST on the Ensembl database[72] to obtain an alignment with all orthologous mammalian sequences therein. After filtering the results for quality, we were left with orthologous matches in a total of 39 other mammalian species. We manually extracted 6 segments with relatively few gaps and most of the species present (supplementary data).

For each of the 6 segments, we used PAML (v4.9) [73]to infer the branch lengths of the evolutionary tree and reconstruct the corresponding ancestral sequence of the 40 mammalian species included in the study. Then, we employed a DNA substitution simulator [65] to evolve the reconstructed sequences along the mammalian phylogeny. For each set of real or simulated sequences we computed the quantitative molecular phenotypes. Tree was drawn using plot function from the APE package for R [74].

We implemented maximum likelihood inference (described in detail in the Appendix) for the three models of phenotype evolution considered here (BM, OU, F81) in python using scipy [75]. To numerically optimize the log-likelihoods ($\log L$), we used Broyden–Fletcher–Goldfarb–Shanno (L-BFGS-B) starting with random initializations of parameters. We confirmed the inference results comparing the likelihoods of BM and OU models to those obtained with the Geiger package[76]. Models were compared using corrected AIC (AICc), which is defined as $AICc = 2\left[-\log L + k + \frac{k(k+1)}{n-k-1}\right]$, where k is the number of parameters and n is the number of species for which we have a mean phenotype measurement.. We adopted the widely-used assumption for inference using OU models in comparative phylogenetic analysis that the OU process is initially at equilibrium[76]. Models with corrected AIC ~2 units below the competing model was deemed better fitting.

Alignments, scripts and evolutionary trees are available at https://github.com/shz9/neutral-phenotype-model.

Foldedness of longest peptides

We considered all 6-frame translations within the 100bp sequences starting with ATG, and selected the longest. Using these peptides, we measured the propensity to fold using the method described in [55]. In this scale, positive values indicate folded, while disordered regions obtain negative values.

Acknowledgements

We acknowledge Dr. Amin Zia, Gavin Douglas, Caressa Tsai, and other members of the Moses Lab who contributed ideas and valiant attempts to this research over many years. We acknowledge Prof. Michael Laessig for hosting SZ during part of this research. We thank Prof. Michael Lynch for helpful comments on the manuscript. AMM and SZ were supported by NSERC discovery and CFI grants to AMM.

Figure Legends

Figure 1 – neutral dynamics of GC content

a) observed mean GC content in inferred ancestral primate ERV sequences with GC content in the common ancestor between 0.45 and 0.5 (filled symbols) compared to predictions of the model considered here (F81, dashed line) and the Brownian Motion model (BM, solid line). *H.*

sap indicates the extant human sequences. b) as in a), but GC content in the common ancestor between 0.3 and 0.35 (filled symbols) c) inferred strength of evolutionary restoring force estimated by simple linear regression on the GC content for each bin of ancestral GC content (symbols) compared to predictions of the model considered here (F81, dashed line) and the Brownian Motion model (BM, solid line). d) observed variance of GC content in the two bins of ancestral GC content shown in a) and b) (filled triangles and squares, respectively) compared to predictions of an OU model with mean restoring force given by our theory (OU, dashed line). Filled circles below the graphs represent the phylogenetic positions of the reconstructed primate ancestors used in the analysis. *H. sap* indicates the phenotype of the extant human sequences.

Figure 2 – neutral dynamics of TATA box strength

As in figure 1, but for the strength of the match to the TATA box in the first 10 nucleotides of the ERV sequences. Ancestral strength in a) is between 0 and 2, while in b) between -14 and -12.

Figure 3 - neutral dynamics of mean number of ATGs

As in figure 1, but for the number of ATGs in the ERV sequences. Ancestral number of ATGs in a) is 4, while in b) it is 0.

Figure 4 – neutral dynamics of the variance of ATGs.

a-b) variance of number of ATGs in 100 bp ERV segments that in the primate ancestor are inferred to have no ATGs (filled squares), 2 ATGs (x's) or 4 ATGs (filled triangles). Predictions based on an OU process model and the F81 model developed here are shown as dashed lines in panels a) and b), respectively. c) the rate of increase of the variance (inferred using simple linear regression) as a function of the ancestral ATG number (symbols) compared to the predictions of the OU model or the F81 model developed here.

Figure 5 – mean dynamics of non-additive molecular phenotypes show mutational restoring force

a) three molecular phenotypes computed from 100 bp ERV segments in reconstructed sequences, binned by their values in the primate ancestor. Colours indicate different ancestral values. See text for phenotype details. b) inferred mutational restoring force as a function of ancestral phenotype, estimated through simple linear regression.

Table 1 – differences in corrected AUC between OU, F81 and BM

	Observed ERV segments		Simulated ERV segments	
Segment	OU	F81	OU	F81
10270_10563	2.35	-2.27	2.15	-0.80
11935_12275	2.46	0.63	-29.0	-3.94
11196_11705	2.39	-1.94	-6.57	-4.74
320_722	2.36	2.57	2.40	-9.90
7916_8695	2.36	0.57	-13.8	-3.38

8938_9248	2.37	2.60	-18.5	-6.67

Each cell represents the corrected AIC comparison vs. the Brownian Motion (BM) null hypothesis. AIC is inversely proportional to likelihood, so lower corrected AICs represent better fitting models.

Supplementary figures

Figure S1 – as in figure 1a-c, but with ancestral reconstruction done using maximum parsimony

Figure S2 – example phylogeny and GC content in ancient ERVs. The inferred phylogenetic tree and associated GC content measurements (GC%) for the segment 320 – 722 for the species used in the Maximum Likelihood inference analysis of GC content.

Figure S3 – as in figure 3 and 4, but with ancestral reconstruction done using maximum parsimony

References

[1] T. Bedford and D. L. Hartl, "Optimization of gene expression by natural selection," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 4, pp. 1133–1138, Jan. 2009, doi: 10.1073/pnas.0812009106.

[2] P. Khaitovich *et al.*, "A Neutral Model of Transcriptome Evolution," *PLoS Biol.*, vol. 2, no. 5, May 2004, doi: 10.1371/journal.pbio.0020132.

[3] R. A. Studer *et al.*, "Evolution of protein phosphorylation across 18 fungal species," *Science*, vol. 354, no. 6309, pp. 229–232, 14 2016, doi: 10.1126/science.aaf2144.

[4] V. Mustonen, J. Kinney, C. G. Callan, and M. Lässig, "Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 34, pp. 12376–12381, Aug. 2008, doi: 10.1073/pnas.0805909105.

[5] A. Nourmohammad, T. Held, and M. Lässig, "Universality and predictability in molecular quantitative genetics," *Curr. Opin. Genet. Dev.*, vol. 23, no. 6, pp. 684–693, Dec. 2013, doi: 10.1016/j.gde.2013.11.001.

[6] A. Nourmohammad, J. Rambeau, T. Held, V. Kovacova, J. Berg, and M. Lässig, "Adaptive Evolution of Gene Expression in Drosophila," *Cell Rep.*, vol. 20, no. 6, pp. 1385–1395, 08 2017, doi: 10.1016/j.celrep.2017.07.033.

[7] D. Brawand *et al.*, "The evolution of gene expression levels in mammalian organs," *Nature*, vol. 478, no. 7369, pp. 343–348, Oct. 2011, doi: 10.1038/nature10532.

[8] J. Chen *et al.*, "A quantitative framework for characterizing the evolutionary history of mammalian gene expression," *Genome Res.*, vol. 29, no. 1, pp. 53–63, 2019, doi: 10.1101/gr.237636.118.

[9] J. G. Schraiber, Y. Mostovoy, T. Y. Hsu, and R. B. Brem, "Inferring Evolutionary Histories of Pathway Regulation from Transcriptional Profiling Data," *PLOS Comput. Biol.*, vol. 9, no. 10, p. e1003255, Oct. 2013, doi: 10.1371/journal.pcbi.1003255.

[10] T. F. Hansen, "Stabilizing Selection and the Comparative Analysis of Adaptation," *Evolution*, vol. 51, no. 5, pp. 1341–1351, 1997, doi: 10.2307/2411186.

[11] R. Lande, "Natural Selection and Random Genetic Drift in Phenotypic Evolution," *Evolution*, vol. 30, no. 2, pp. 314–334, Jun. 1976, doi: 10.2307/2407703.

[12] J. M. Beaulieu, D.-C. Jhwueng, C. Boettiger, and B. C. O'Meara, "Modeling stabilizing selection: expanding the ornstein–uhlenbeck model of adaptive evolution," *Evolution*, vol. 66, no. 8, pp. 2369–2383, Aug. 2012, doi: 10.1111/j.1558-5646.2012.01619.x.

[13] M. A. Butler and A. A. King, “Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution.,” *Am. Nat.*, vol. 164, no. 6, pp. 683–695, Dec. 2004, doi: 10.1086/426002.
[14] J. C. Uyeda and J. E. and L. Harmon, *bayou: Bayesian Fitting of Ornstein-Uhlenbeck Models to Phylogenies*. 2018. [Online]. Available: https://cran.r-project.org/web/packages/bayou/index.html
[15] W.-C. Ho, Y. Ohya, and J. Zhang, “Testing the neutral hypothesis of phenotypic evolution,” *Proc. Natl. Acad. Sci.*, vol. 114, no. 46, pp. 12219–12224, Nov. 2017, doi: 10.1073/pnas.1710351114.
[16] A. Whitehead and D. L. Crawford, “Neutral and adaptive variation in gene expression,” *Proc. Natl. Acad. Sci.*, vol. 103, no. 14, pp. 5425–5430, Apr. 2006, doi: 10.1073/pnas.0507648103.
[17] J. Zhang, “Neutral Theory and Phenotypic Evolution,” *Mol. Biol. Evol.*, vol. 35, no. 6, pp. 1327–1331, Jun. 2018, doi: 10.1093/molbev/msy065.
[18] J. Felsenstein, “Phylogenies and quantitative characters,” *Annu. Rev. Ecol. Syst.*, vol. 19, no. 1, pp. 445–471, Nov. 1988, doi: 10.1146/annurev.es.19.110188.002305.
[19] M. Arenas, “Trends in substitution models of molecular evolution,” *Front. Genet.*, vol. 6, Oct. 2015, doi: 10.3389/fgene.2015.00319.
[20] Z. Yang, *Computational Molecular Evolution*. Oxford, New York: Oxford University Press, 2006.
[21] R. A. Goldstein and D. D. Pollock, “Sequence entropy of folding and the absolute rate of amino acid substitutions,” *Nat. Ecol. Evol.*, vol. 1, no. 12, pp. 1923–1930, Dec. 2017, doi: 10.1038/s41559-017-0338-9.
[22] Z. B. Zeng and C. C. Cockerham, “Mutation models and quantitative genetic variation.,” *Genetics*, vol. 133, no. 3, pp. 729–736, Mar. 1993.
[23] J. Felsenstein, “Maximum-likelihood estimation of evolutionary trees from continuous characters.,” *Am. J. Hum. Genet.*, vol. 25, no. 5, pp. 471–492, Sep. 1973.
[24] J. Felsenstein, “Phylogenies and the Comparative Method,” *Am. Nat.*, vol. 125, no. 1, pp. 1–15, 1985.
[25] R. Chakraborty and M. Nei, “Genetic differentiation of quantitative characters between populations or species: I. Mutation and random genetic drift,” *Genet. Res.*, vol. 39, no. 3, pp. 303–314, Jun. 1982, doi: 10.1017/S0016672300020978.
[26] S. K. Davies, A. Leroi, A. Burt, J. G. Bundy, and C. F. Baer, “THE MUTATIONAL STRUCTURE OF METABOLISM IN CAENORHABDITIS ELEGANS,” *Evol. Int. J. Org. Evol.*, vol. 70, no. 10, pp. 2239–2246, Oct. 2016, doi: 10.1111/evo.13020.
[27] J. F. C. Kingman, “A Simple Model for the Balance between Selection and Mutation,” *J. Appl. Probab.*, vol. 15, no. 1, pp. 1–12, 1978, doi: 10.2307/3213231.
[28] A. Stoltzfus, “On the possibility of constructive neutral evolution,” *J. Mol. Evol.*, vol. 49, no. 2, pp. 169–181, Aug. 1999, doi: 10.1007/pl00006540.
[29] M. Lynch, “The frailty of adaptive hypotheses for the origins of organismal complexity,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104 Suppl 1, pp. 8597–8604, May 2007, doi: 10.1073/pnas.0702207104.
[30] M. Lynch, “Phylogenetic divergence of cell biological features,” *eLife*, vol. 7, 21 2018, doi: 10.7554/eLife.34820.
[31] A. Nourmohammad, S. Schiffels, and M. Lässig, “Evolution of molecular phenotypes under stabilizing selection,” *J. Stat. Mech. Theory Exp.*, vol. 2013, no. 01, p. P01012, Jan. 2013, doi: 10.1088/1742-5468/2013/01/P01012.
[32] B. Charlesworth, “Stabilizing selection, purifying selection, and mutational bias in finite populations,” *Genetics*, vol. 194, no. 4, pp. 955–971, Aug. 2013, doi: 10.1534/genetics.113.151555.
[33] L. M, “The Evolutionary Scaling of Cellular Traits Imposed by the Drift Barrier,” *Proceedings of the National Academy of Sciences of the United States of America*, May 12, 2020. https://pubmed.ncbi.nlm.nih.gov/32345718/ (accessed Jul. 03, 2020).

[34] A. Hodgins-Davis, D. P. Rice, and J. P. Townsend, “Gene Expression Evolves under a House-of-Cards Model of Stabilizing Selection,” *Mol. Biol. Evol.*, vol. 32, no. 8, pp. 2130–2140, Aug. 2015, doi: 10.1093/molbev/msv094.
[35] A. Hodgins-Davis, F. Duveau, E. A. Walker, and P. J. Wittkopp, “Empirical measures of mutational effects define neutral models of regulatory evolution in Saccharomyces cerevisiae,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 42, pp. 21085–21093, 15 2019, doi: 10.1073/pnas.1902823116.
[36] B. Golding and J. Felsenstein, “A maximum likelihood approach to the detection of selection from a phylogeny,” *J. Mol. Evol.*, vol. 31, no. 6, pp. 511–523, Dec. 1990, doi: 10.1007/BF02102078.
[37] J. Felsenstein, “Evolutionary trees from DNA sequences: A maximum likelihood approach,” *J. Mol. Evol.*, vol. 17, no. 6, pp. 368–376, Nov. 1981, doi: 10.1007/BF01734359.
[38] W. J. Ewens, *Mathematical Population Genetics 1: Theoretical Introduction*, 2nd ed. New York: Springer-Verlag, 2004. doi: 10.1007/978-0-387-21822-9.
[39] M. Lynch, “The evolutionary scaling of cellular traits imposed by the drift barrier,” *Proc. Natl. Acad. Sci.*, vol. 117, no. 19, pp. 10435–10444, May 2020, doi: 10.1073/pnas.2000446117.
[40] M. Kimura, “A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences,” *J. Mol. Evol.*, vol. 16, no. 2, pp. 111–120, Dec. 1980, doi: 10.1007/BF01731581.
[41] M. Lynch and W. G. Hill, “PHENOTYPIC EVOLUTION BY NEUTRAL MUTATION,” *Evol. Int. J. Org. Evol.*, vol. 40, no. 5, pp. 915–935, Sep. 1986, doi: 10.1111/j.1558-5646.1986.tb00561.x.
[42] N. Cooper, G. H. Thomas, C. Venditti, A. Meade, and R. P. Freckleton, “A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies,” *Biol. J. Linn. Soc.*, vol. 118, no. 1, pp. 64–77, May 2016, doi: 10.1111/bij.12701.
[43] C. E. Cressler, M. A. Butler, and A. A. King, “Detecting Adaptive Evolution in Phylogenetic Comparative Analysis Using the Ornstein–Uhlenbeck Model,” *Syst. Biol.*, vol. 64, no. 6, pp. 953–968, Nov. 2015, doi: 10.1093/sysbio/syv043.
[44] T. J. Meyer, J. L. Rosenkrantz, L. Carbone, and S. L. Chavez, “Endogenous Retroviruses: With Us and against Us,” *Front. Chem.*, vol. 5, 2017, doi: 10.3389/fchem.2017.00023.
[45] P. Trávníček, M. Čertner, J. Ponert, Z. Chumová, J. Jersáková, and J. Suda, “Diversity in genome size and GC content shows adaptive potential in orchids and is closely linked to partial endoreplication, plant life-history traits and climatic conditions,” *New Phytol.*, vol. 224, no. 4, pp. 1642–1656, Dec. 2019, doi: 10.1111/nph.15996.
[46] B. Paten *et al.*, “Genome-wide nucleotide-level mammalian ancestor reconstruction,” *Genome Res.*, vol. 18, no. 11, pp. 1829–1843, Nov. 2008, doi: 10.1101/gr.076521.108.
[47] K. A *et al.*, “JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and Its Web Framework,” *Nucleic acids research*, Jan. 04, 2018. https://pubmed.ncbi.nlm.nih.gov/29140473/ (accessed Jul. 11, 2020).
[48] B. Og and von H. Ph, “Selection of DNA Binding Sites by Regulatory Proteins. Statistical-mechanical Theory and Application to Operators and Promoters,” *Journal of molecular biology*, Feb. 20, 1987. https://pubmed.ncbi.nlm.nih.gov/3612791/ (accessed Jul. 11, 2020).
[49] W. W. Wasserman and A. Sandelin, “Applied bioinformatics for the identification of regulatory elements,” *Nat. Rev. Genet.*, vol. 5, no. 4, pp. 276–287, Apr. 2004, doi: 10.1038/nrg1315.
[50] H.-D. A, D. F, W. Ea, and W. Pj, “Empirical Measures of Mutational Effects Define Neutral Models of Regulatory Evolution in Saccharomyces cerevisiae,” *Proceedings of the National Academy of Sciences of the United States of America*, Oct. 15, 2019. https://pubmed.ncbi.nlm.nih.gov/31570626/ (accessed Jul. 10, 2020).
[51] R. W. Lusk and M. B. Eisen, “Evolutionary Mirages: Selection on Binding Site Composition Creates the Illusion of Conserved Grammars in Drosophila Enhancers,” *PLOS Genet.*, vol. 6, no. 1, p. e1000829, Jan. 2010, doi: 10.1371/journal.pgen.1000829.

[52] X. He, T. S. P. C. Duque, and S. Sinha, “Evolutionary Origins of Transcription Factor Binding Site Clusters,” *Mol. Biol. Evol.*, vol. 29, no. 3, pp. 1059–1070, Mar. 2012, doi: 10.1093/molbev/msr277.

[53] B. A. Wilson, S. G. Foy, R. Neme, and J. Masel, “Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth,” *Nat. Ecol. Evol.*, vol. 1, no. 6, p. 0146, Jun. 2017, doi: 10.1038/s41559-017-0146.

[54] W. Basile, O. Sachenkova, S. Light, and A. Elofsson, “High GC content causes orphan proteins to be intrinsically disordered,” *PLoS Comput. Biol.*, vol. 13, no. 3, p. e1005375, 2017, doi: 10.1371/journal.pcbi.1005375.

[55] J. Prilusky *et al.*, “FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded,” *Bioinforma. Oxf. Engl.*, vol. 21, no. 16, pp. 3435–3438, Aug. 2005, doi: 10.1093/bioinformatics/bti537.

[56] A. Lee, A. Nolan, J. Watson, and M. Tristem, “Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 368, no. 1626, Sep. 2013, doi: 10.1098/rstb.2012.0503.

[57] J. R. Stone and G. A. Wray, “Rapid evolution of cis-regulatory sequences via local point mutations,” *Mol. Biol. Evol.*, vol. 18, no. 9, pp. 1764–1770, Sep. 2001, doi: 10.1093/oxfordjournals.molbev.a003964.

[58] C. B, “Stabilizing Selection, Purifying Selection, and Mutational Bias in Finite Populations,” *Genetics*, Aug. 2013. https://pubmed.ncbi.nlm.nih.gov/23709636/ (accessed Jul. 10, 2020).

[59] Z. Yang and T. Zhu, “Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 8, pp. 1854–1859, Feb. 2018, doi: 10.1073/pnas.1712673115.

[60] L. S. T. Ho and C. Ané, “Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models,” *Methods Ecol. Evol.*, vol. 5, no. 11, pp. 1133–1146, 2014, doi: 10.1111/2041-210X.12285.

[61] B. P. McEvoy and P. M. Visscher, “Genetics of human height,” *Econ. Hum. Biol.*, vol. 7, no. 3, pp. 294–306, Dec. 2009, doi: 10.1016/j.ehb.2009.09.005.

[62] C. M. Jakobson and D. F. Jarosz, “Molecular Origins of Complex Heritability in Natural Genotype-to-Phenotype Relationships,” *Cell Syst.*, vol. 8, no. 5, pp. 363-379.e3, 22 2019, doi: 10.1016/j.cels.2019.04.002.

[63] C. G. de Boer, E. D. Vaishnav, R. Sadeh, E. L. Abeyta, N. Friedman, and A. Regev, “Deciphering eukaryotic gene-regulatory logic with 100 million random promoters,” *Nat. Biotechnol.*, vol. 38, no. 1, pp. 56–65, 2020, doi: 10.1038/s41587-019-0315-8.

[64] D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek, “Sequential regulatory activity prediction across chromosomes with convolutional neural networks,” *Genome Res.*, vol. 28, no. 5, pp. 739–750, 2018, doi: 10.1101/gr.227819.117.

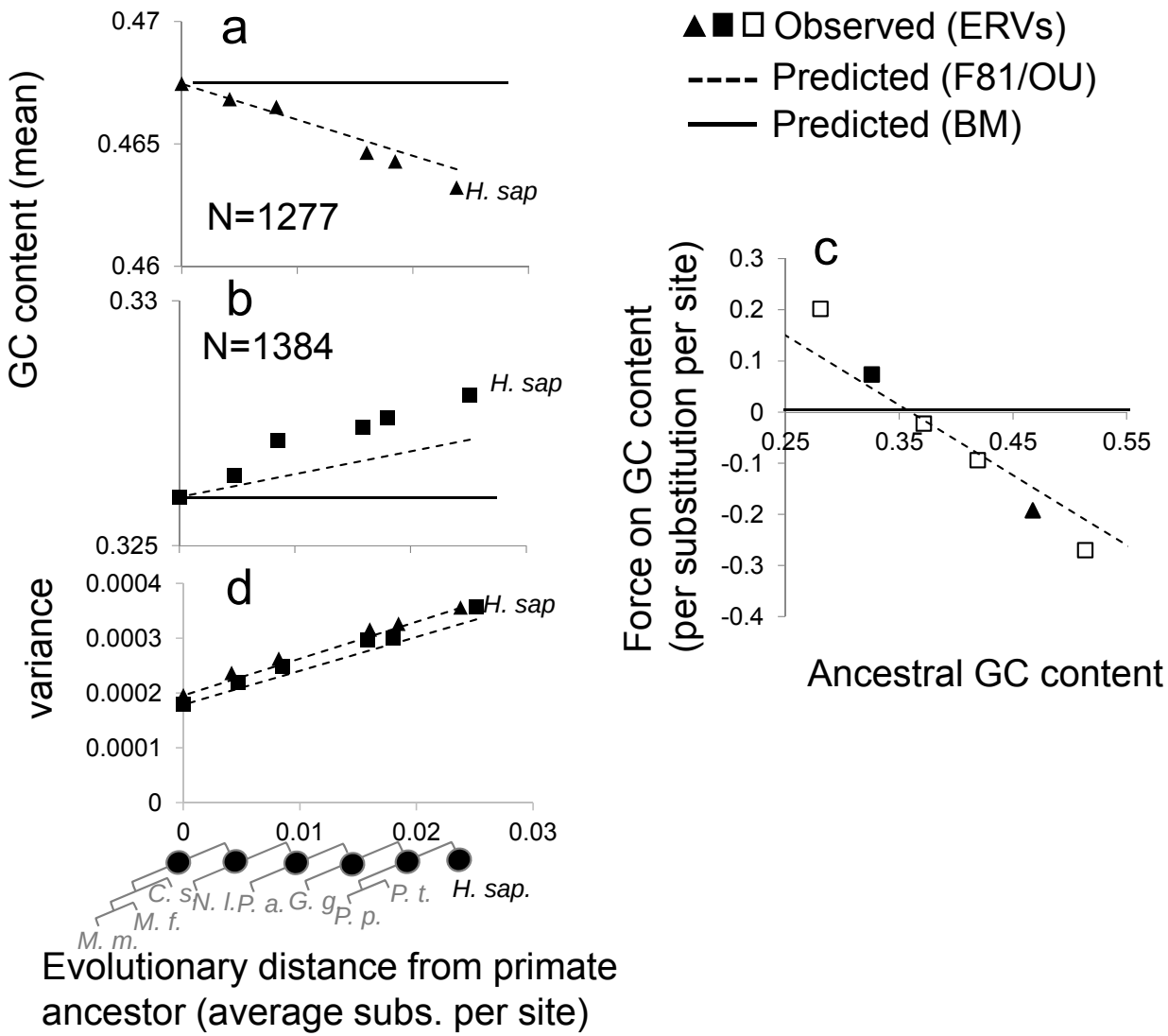
[65] C. L. Strope, K. Abel, S. D. Scott, and E. N. Moriyama, “Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0,” *Mol. Biol. Evol.*, vol. 26, no. 11, pp. 2581–2593, Nov. 2009, doi: 10.1093/molbev/msp174.

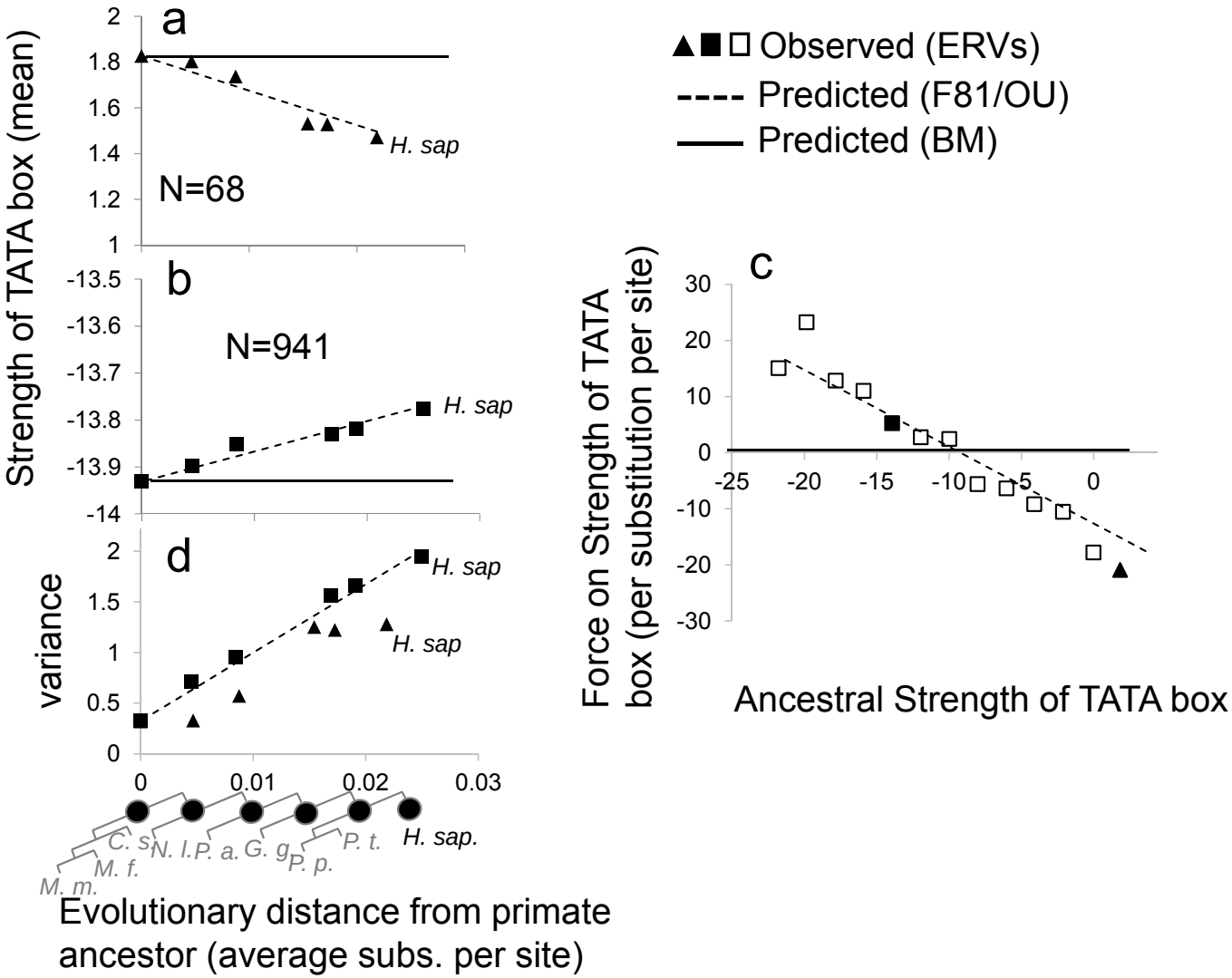
[66] T. Zarin, C. N. Tsai, A. N. Nguyen Ba, and A. M. Moses, “Selection maintains signaling function of a highly diverged intrinsically disordered region,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 8, pp. E1450–E1459, 21 2017, doi: 10.1073/pnas.1614787114.

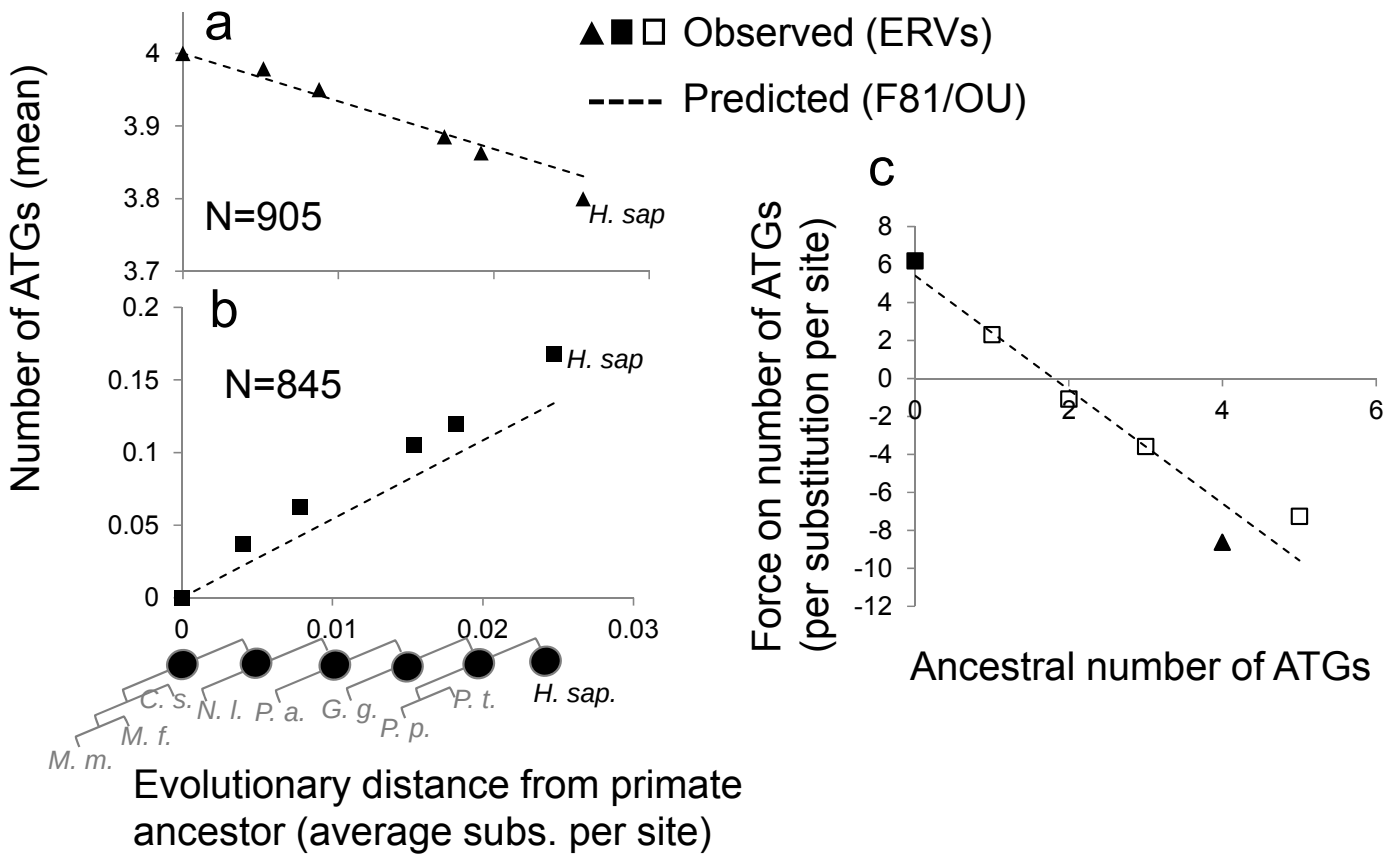
[67] T. Zarin, B. Strome, A. N. Nguyen Ba, S. Alberti, J. D. Forman-Kay, and A. M. Moses, “Proteome-wide signatures of function in highly diverged intrinsically disordered regions,” *eLife*, vol. 8, Jul. 2019, doi: 10.7554/eLife.46883.

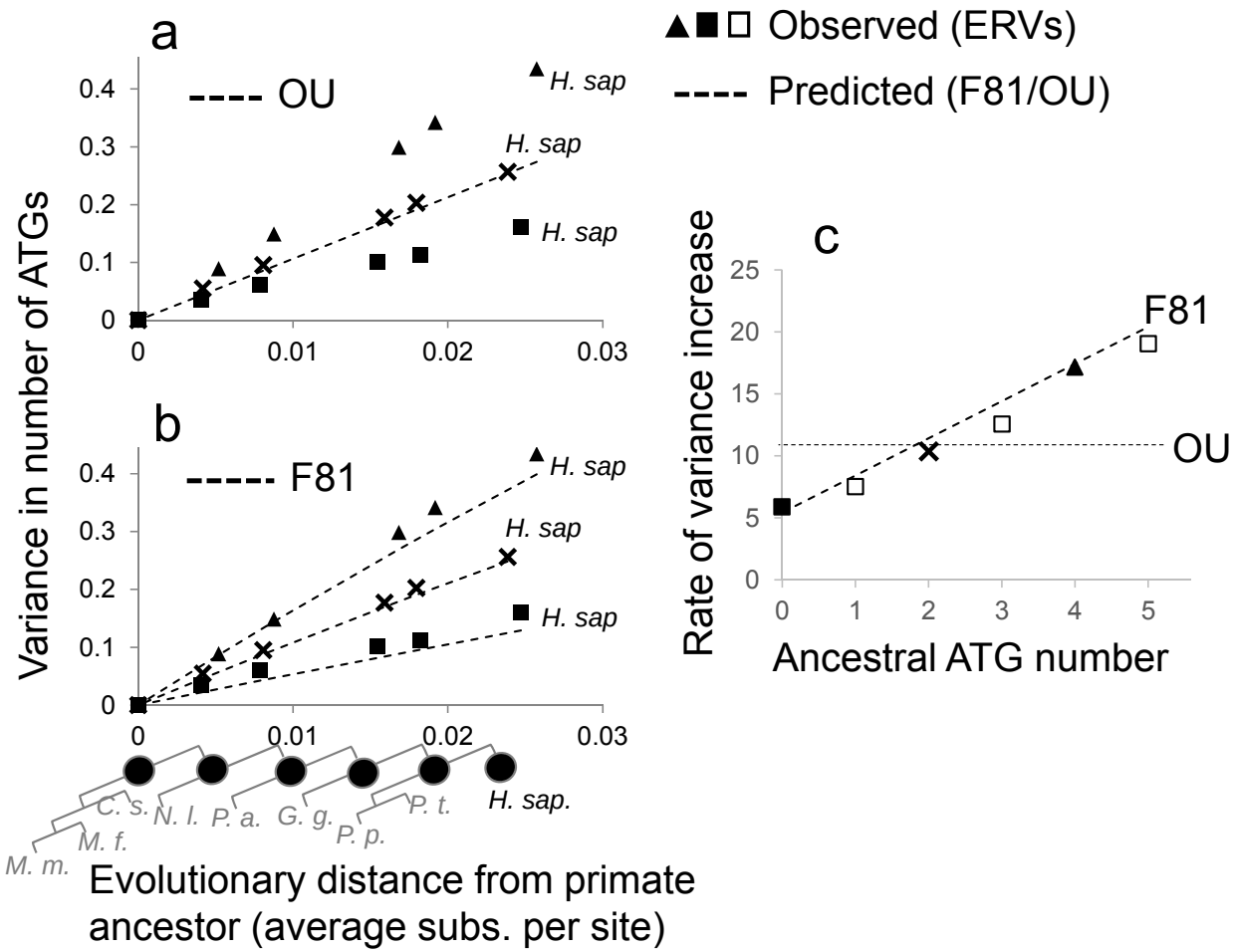
[68] S. Nakagawa and M. U. Takahashi, “gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes,” *Database J. Biol. Databases Curation*, vol. 2016, 2016, doi: 10.1093/database/baw087.

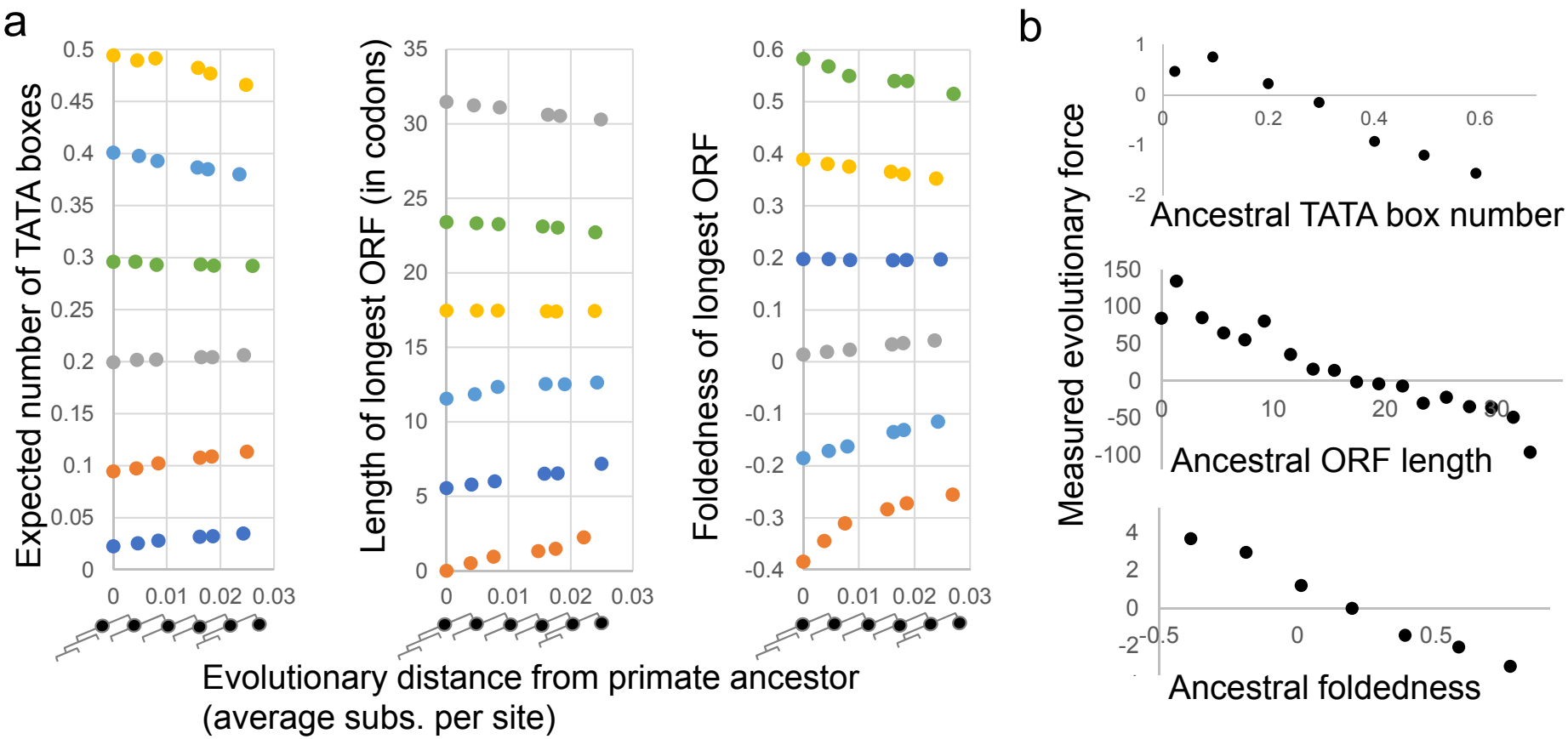
[69] A. Yates *et al.*, “The Ensembl REST API: Ensembl Data for Any Language,” *Bioinformatics*, vol. 31, no. 1, pp. 143–145, Jan. 2015, doi: 10.1093/bioinformatics/btu613.
[70] K. P. Schliep, “phangorn: phylogenetic analysis in R,” *Bioinformatics*, vol. 27, no. 4, pp. 592–593, Feb. 2011, doi: 10.1093/bioinformatics/btq706.
[71] A. Moses and S. Sinha, “Regulatory Motif Analysis,” in *Bioinformatics: Tools and Applications*, 2009, pp. 137–163. doi: 10.1007/978-0-387-92738-1_7.
[72] D. R. Zerbino *et al.*, “Ensembl 2018,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D754–D761, 04 2018, doi: 10.1093/nar/gkx1098.
[73] Z. Yang, “PAML 4: phylogenetic analysis by maximum likelihood,” *Mol. Biol. Evol.*, vol. 24, no. 8, pp. 1586–1591, Aug. 2007, doi: 10.1093/molbev/msm088.
[74] E. Paradis and K. Schliep, “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R,” *Bioinforma. Oxf. Engl.*, vol. 35, no. 3, pp. 526–528, 01 2019, doi: 10.1093/bioinformatics/bty633.
[75] P. Virtanen *et al.*, “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.
[76] M. W. Pennell *et al.*, “geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees,” *Bioinformatics*, vol. 30, no. 15, pp. 2216–2218, Aug. 2014, doi: 10.1093/bioinformatics/btu181.


a
b
c
d
GC content (mean)
0.47
0.465
0.46
N=1277
H. sap
0.33
0.325
N=1384
H. sap
variance
0.0004
0.0003
0.0002
0.0001
0
0
0.01
0.02
0.03
H. sap
C. s
N. l.
P. a.
G. g
P. p.
P. t.
H. sap.
M. f.
M. m.
Evolutionary distance from primate ancestor (average subs. per site)
Observed (ERVs)
Predicted (F81/OU)
Predicted (BM)
Force on GC content (per substitution per site)
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0.25
0.35
0.45
0.55
Ancestral GC content

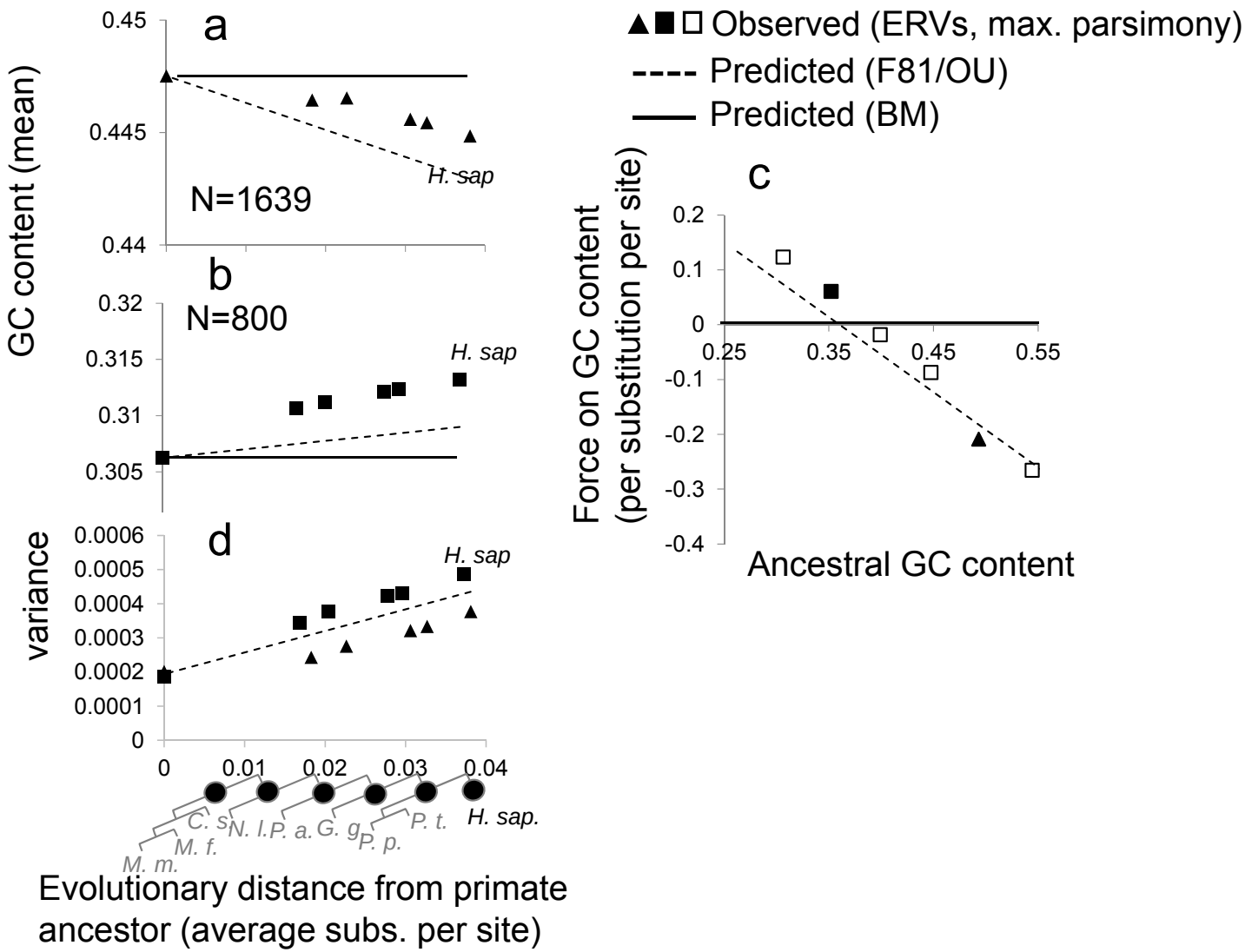

a
b
c
d
N=68
N=941
H. sap
Observed (ERVs)
Predicted (F81/OU)
Predicted (BM)
Strength of TATA box (mean)
variance
Force on Strength of TATA box (per substitution per site)
Ancestral Strength of TATA box
Evolutionary distance from primate ancestor (average subs. per site)
M. m.
M. f.
C. s.
N. l.
P. a.
G. g.
P. p.
P. t.
H. sap.

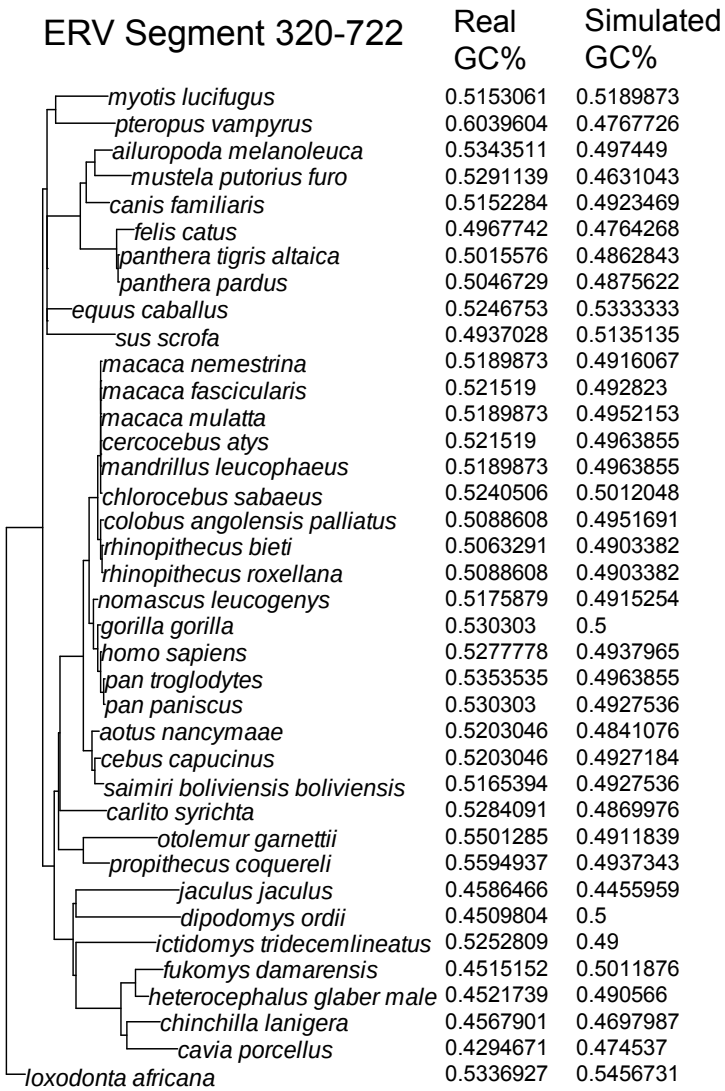

a
b
c
Observed (ERVs)
Predicted (F81/OU)
Number of ATGs (mean)
N=905
N=845
H. sap
4
3.9
3.8
3.7
0.2
0.15
0.1
0.05
0
0.01
0.02
0.03
C. s.
N. l.
P. a.
G. g.
P. p.
P. t.
H. sap.
M. f.
M. m.
Evolutionary distance from primate ancestor (average subs. per site)
Force on number of ATGs (per substitution per site)
8
6
4
2
0
-2
-4
-6
-8
-10
-12
Ancestral number of ATGs

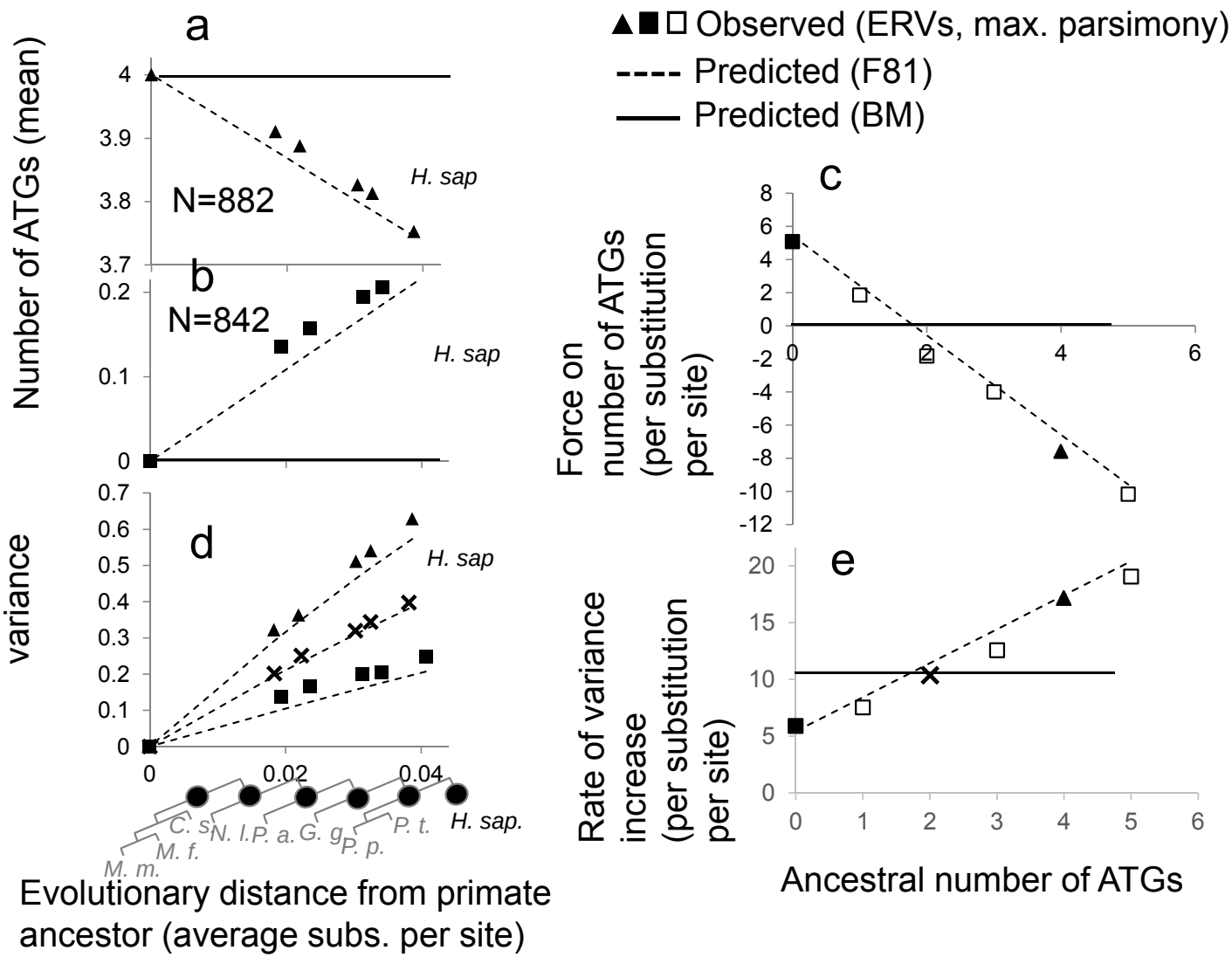

a
---- OU
H. sap
H. sap
H. sap
0.4
0.3
0.2
0.1
0
0
0.01
0.02
0.03
b
----- F81
H. sap
H. sap
H. sap
Variance in number of ATGs
C. s.
N. l.
P. a.
G. g.
P. p.
P. t.
H. sap.
M. f.
M. m.
Evolutionary distance from primate ancestor (average subs. per site)
▲■□ Observed (ERVs)
---- Predicted (F81/OU)
c
Rate of variance increase
25
20
15
10
5
0
F81
OU
0
1
2
3
4
5
Ancestral ATG number


a
Expected number of TATA boxes
Length of longest ORF (in codons)
Foldedness of longest ORF
Evolutionary distance from primate ancestor
(average subs. per site)
b
Measured evolutionary force
Ancestral TATA box number
Ancestral ORF length
Ancestral foldedness

Supplementary figures


a
b
c
d
GC content (mean)
0.45
0.445
0.44
N=1639
H. sap
0.32
0.315
0.31
0.305
N=800
H. sap
variance
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0
0
0.01
0.02
0.03
0.04
H. sap
H. sap.
C. s.
M. f.
M. m.
N. l.
P. a.
G. g.
P. p.
P. t.
Evolutionary distance from primate ancestor (average subs. per site)
Observed (ERVs, max. parsimony)
Predicted (F81/OU)
Predicted (BM)
Force on GC content (per substitution per site)
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0.25
0.35
0.45
0.55
Ancestral GC content


ERV Segment 320-722
Real GC%
Simulated GC%
myotis lucifugus 0.5153061 0.5189873
pteropus vampyrus 0.6039604 0.4767726
ailuropoda melanoleuca 0.5343511 0.497449
mustela putorius furo 0.5291139 0.4631043
canis familiaris 0.5152284 0.4923469
felis catus 0.4967742 0.4764268
panthera tigris altaica 0.5015576 0.4862843
panthera pardus 0.5046729 0.4875622
equus caballus 0.5246753 0.5333333
sus scrofa 0.4937028 0.5135135
macaca nemestrina 0.5189873 0.4916067
macaca fascicularis 0.521519 0.492823
macaca mulatta 0.5189873 0.4952153
cercocebus atys 0.521519 0.4963855
mandrillus leucophaeus 0.5189873 0.4963855
chlorocebus sabaeus 0.5240506 0.5012048
colobus angolensis palliatus 0.5088608 0.4951691
rhinopithecus bieti 0.5063291 0.4903382
rhinopithecus roxellana 0.5088608 0.4903382
nomascus leucogenys 0.5175879 0.4915254
gorilla gorilla 0.530303 0.5
homo sapiens 0.5277778 0.4937965
pan troglodytes 0.5353535 0.4963855
pan paniscus 0.530303 0.4927536
aotus nancymaae 0.5203046 0.4841076
cebus capucinus 0.5203046 0.4927184
saimiri boliviensis boliviensis 0.5165394 0.4927536
carlito syrichta 0.5284091 0.4869976
otolemur garnettii 0.5501285 0.4911839
propithecus coquereli 0.5594937 0.4937343
jaculus jaculus 0.4586466 0.4455959
dipodomys ordii 0.4509804 0.5
ictidomys tridecemlineatus 0.5252809 0.49
fukomys damarensis 0.4515152 0.5011876
heterocephalus glaber male 0.4521739 0.490566
chinchilla lanigera 0.4567901 0.4697987
cavia porcellus 0.4294671 0.474537
loxodonta africana 0.5336927 0.5456731


a
b
c
d
e
Observed (ERVs, max. parsimony)
Predicted (F81)
Predicted (BM)
Number of ATGs (mean)
N=882
N=842
H. sap
variance
Force on number of ATGs (per substitution per site)
Rate of variance increase (per substitution per site)
Ancestral number of ATGs
Evolutionary distance from primate ancestor (average subs. per site)
M. m.
M. f.
C. s.
N. l.
P. a.
G. g.
P. p.
P. t.
H. sap.

Appendix
"Evolutionary dynamics of neutral phenotypes under DNA substitution models"

Shadi Zabad and Alan M Moses

Contents

1 A bi-allelic model **2**
1.1 Single locus statistics . . . 3
1.2 Mean phenotype statistics . . . 7
1.2.1 Phylogenetic mean of mean phenotype . . . 7
1.2.2 Phylogenetic covariance of mean phenotype . . . 9
1.2.3 Phylogenetic variance of mean phenotype . . . 10

2 K alleles **11**
2.1 Single locus statistics . . . 11
2.2 Mean phenotype statistics . . . 13
2.2.1 Phylogenetic mean of mean phenotype . . . 13
2.2.2 Phylogenetic covariance of mean phenotype . . . 14
2.2.3 Phylogenetic variance of mean phenotype . . . 15

3 Simplifying and extending the model **16**
3.1 Lumping alleles and the grouped model . . . 16
3.2 Abstracting away the genetic architecture from the covariance . . . 18
3.3 Phylogenetic divergence of mean phenotype . . . 19
3.4 Linear approximation for the variance at small time . . . 20
3.5 Simplifying covariance in the bi-allelic model . . . 20

4 Phylogenetic inference under the model **22**
4.1 The likelihood . . . 22
4.2 Comparison with Brownian motion and the Ornstein-Uhlenbeck process . . . 23

5 Examples and applications **25**
5.1 Dynamics of GC content under the F-81 model . . . 25
5.1.1 Dynamics of the mean and variance . . . 25
5.1.2 Linear approximation for the variance of GC Content under F-81 . . . 26
5.2 Dynamics of GC content under the JC-69 model . . . 27
5.2.1 Dynamics of the mean and variance . . . 27
5.2.2 Linear approximation for the variance of GC Content under JC-69 . . . 28
5.3 Dynamics of ATGs under the F-81 model . . . 28
5.3.1 Dynamics of the mean and variance . . . 28
5.3.2 Linear approximation for the variance of ATGs under F-81 . . . 29

1 A bi-allelic model

To understand the evolutionary dynamics of mean phenotypes, we start with a linear additive trait that is expressed in a haploid organism with a bi-allelic genetic architecture. Assuming that this trait is underpinned by L bi-allelic loci, using the notation commonly employed in modern GWAS literature, we can express the dependence of the phenotype on the underlying genetics as follows:

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i = \beta_0 + \sum_{j=1}^{L} X_{ij}\beta_j + \epsilon_i \tag{1}$$

Here, y_i denotes the phenotype for individual i, $\boldsymbol{X}_i\boldsymbol{\beta}$ is the genetic contribution to the phenotype and ϵ_i captures the contribution of the environment and other non-linear effects. $\boldsymbol{X}_i$ is a $1 \times (L+1)$ binary vector indicating whether individual i has the reference or alternative alleles at each locus j:

$$X_{ij} = \begin{cases} 0 & \text{(if individual } i \text{ has reference allele at locus } j\text{)} \\ 1 & \text{(if individual } i \text{ has alternative allele at locus } j\text{)} \end{cases}$$

$\boldsymbol{\beta}$ is an $(L+1) \times 1$ vector of effect sizes. We assume, without loss of generality, that at each locus, the effect sizes for the reference and alternative alleles are encoded by β_{j1} and β_{j2}, respectively. This implies that $\beta_j = \beta_{j2} - \beta_{j1}$ and from this, it follows that the intercept term β_0 is the sum of the effect sizes of the reference alleles at each locus: $\beta_0 = \sum_{j=1}^{L} \beta_{j1}$ (e.g. individuals with no mutations at any site will have a genotypic value of β_0). Modern GWAS methodologies typically dispense with the intercept term by standardizing the phenotype vector and genotype matrix. In our case, it is a constant that will not affect the dynamics of phenotypic evolution and in fact has a value of zero for many of the traits that we will consider in the main text.

The main quantity of interest in our analysis is **the mean phenotype in a population or species** $Z := \mathbb{E}[y]$. If we assume that the effect sizes are fixed, we obtain:

$$Z := \mathbb{E}[y] = \mathbb{E}[\boldsymbol{X}]\boldsymbol{\beta} + \mathbb{E}[\epsilon] = \beta_0 + \sum_{j}^{L} \mathbb{E}[X_j]\beta_j \tag{2}$$

The environmental contribution is omitted because we follow the standard assumption that $\mathbb{E}[\epsilon] = 0$. Furthermore, in the context of this analysis, most of the traits that we are interested in are expressed at the molecular and cellular level, often comprising aggregate statistics of the underlying sequence, e.g. GC content. Therefore, the environmental contribution is assumed to be nil. Note that within a given population or species, $\mathbb{E}[X_j]$ is simply the allele frequency f_j at the jth locus.

In Phylogenetic Comparative Methods [7, 6], one of the central aims is to analyze the evolution of the mean phenotype over time $Z(t)$, given an initial starting point $Z_0 := Z(0)$. In particular, given an initial mean phenotype, the goal is to characterize its evolution over an ensemble of different hypothetical populations or experimental replicates. Assuming that the effect sizes (β_js) and the genetic architecture of the trait remain constant, the only time-dependent parameter is the genotype at each locus:

$$Z(t) = \boldsymbol{X}(t)\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{L} f_j(t)\beta_j$$

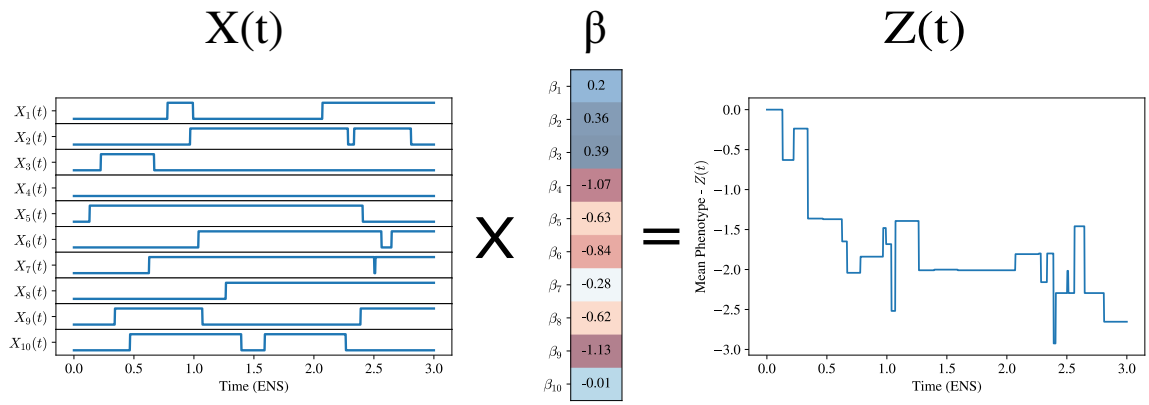


Figure 1: An illustration of the bi-allelic model for the evolution of mean phenotype in a single population. In this example, we show the allelic state of $L = 10$ genetic loci as a function of time $\boldsymbol{X}(t)$, measured in units of Expected Number of Substitutions (ENS), and the corresponding effect sizes $\boldsymbol{\beta}$. In this simulation, all loci start with the reference allele and then, through mutation and drift, alleles can undergo fixation or extinction. We model this evolutionary process with a substitution model where loci evolve according to a Continuous-Time Markov Chain (CTMC). The mean phenotype as a function of time $Z(t)$ is obtained by multiplying the state of the locus at time t with the effect sizes according to Equation 3.

Characterizing the evolution of the genotype matrix $\boldsymbol{X}(t)$ over time in response to mutation, selection, and drift has been a central focus of population genetics theory for the past century [6]. Much of the recent effort in this area made use of the diffusion approximation, first introduced in the works of Wright and Fisher and then later extended by Kimura, among others [4]. In this work, we are interested in characterizing the evolution of the genotype matrix in response to mutation and drift alone and over phylogenetic timescales. In this context, we make use of 2 simplifying assumptions: (1) The loci evolve independently and (2) Evolution proceeds in the weak mutation regime, where at any point in time the population is largely monomorphic. In practical terms, this means that instead of modeling the full trajectory of $f_j(t)$ from introduction to fixation or extinction, we model the state of the locus as a jump process between 2 states (reference vs. alternative allele), governed by a Continuous Time Markov Chain (CTMC) [14].

At the risk of overloading notation, for simplicity we replace the continuous random variables $\mathbb{E}[X_j(t)]$ and $f_j(t)$ with a binary random variable $X_j(t)$ that captures the state of the locus j in a given population at any point in time t:

$$Z(t) = \boldsymbol{X}(t)\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{L} X_j(t)\beta_j \tag{3}$$

This model is illustrated in Figure 1. Since the phenotype model is linear with respect to the individual loci, most of the quantities that we need to study the evolutionary dynamics of mean phenotypes can be derived as simple functions of single-locus statistics, which we turn to in the next section.

1.1 Single locus statistics

As explained in the previous section, we model the genotype $X_j(t)$ at a given locus j with a Continuous Time Markov Chain (CTMC) that jumps between 2 states corresponding

to the 2 different alleles. Because of the way we setup the phenotypic model, the states of the system in this case are $X_j(t) = 0$ if the reference allele is fixed at time t and $X_j(t) = 1$ if the alternative allele is fixed. Under this model, the reference allele is expected to be substituted by the alternative allele with an infinitesimal rate λ and the converse happens with rate $\kappa\lambda$, where $\kappa > 0$ is a parameter that quantifies the extent of mutational bias. The generator matrix Q of this substitution process has the following form:

$$Q = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} -\lambda & \lambda \\ \kappa\lambda & -\kappa\lambda \end{pmatrix} \end{array}$$

From the above generator matrix, it is straightforward to obtain expressions for the transition probabilities in a time interval t:

$$\begin{aligned}
P(X_j(t) = 0|X_j(0) = 0) = P_{00}(t) &= \frac{\kappa}{1+\kappa} + \frac{1}{1+\kappa}e^{-\lambda(1+\kappa)t} \\
P(X_j(t) = 1|X_j(0) = 0) = P_{01}(t) &= \frac{1}{1+\kappa} - \frac{1}{1+\kappa}e^{-\lambda(1+\kappa)t} \\
P(X_j(t) = 1|X_j(0) = 1) = P_{11}(t) &= \frac{1}{1+\kappa} + \frac{\kappa}{1+\kappa}e^{-\lambda(1+\kappa)t} \\
P(X_j(t) = 0|X_j(0) = 1) = P_{10}(t) &= \frac{\kappa}{1+\kappa} - \frac{\kappa}{1+\kappa}e^{-\lambda(1+\kappa)t}
\end{aligned}$$

The transition probabilities defined above provide the probability of transitioning from one state to the other in a given time interval. To obtain an expression for the probability that the alternative allele is fixed at any point in time, we marginalize over the ancestral state:

$$\begin{aligned}
P\left(X_j(t) = 1\right) &= P(X_j(0) = 0)P_{01}(t) + P(X_j(0) = 1)P_{11}(t) \\
&= P(X_j(0) = 1)e^{-\lambda(1+\kappa)t} + \frac{1}{1+\kappa}(1 - e^{-\lambda(1+\kappa)t})
\end{aligned}$$

The complement of this, the probability that the reference allele is fixed at any point in time is $P(X_j(t) = 0) = 1 - P(X_j(t) = 1)$. If we take the limit as $t \to \infty$, we see that this system has a stationary distribution with $\pi = \frac{1}{1+\kappa}$ and $1 - \pi = \frac{\kappa}{1+\kappa}$. We therefore write the probability of observing the alternative allele at time t as:

$$P(X_j(t) = 1) = P(X_j(0) = 1)e^{-\lambda(1+\kappa)t} + \pi(1 - e^{-\lambda(1+\kappa)t}) \tag{4}$$

Before proceeding, it is worth highlighting that in this context we use probability in the frequentist sense: we envisage a large ensemble of populations emanating from a single founder population, as in a phylogeny or a set of genes that evolve independently after a duplication event [6]. Each population may lose or fix alleles independently and we wish to characterize the dynamics of mean phenotype over this hypothetical ensemble (see Figure 2). This is in line with the tradition in the Evolutionary Quantitative Genetics literature where researchers model the phylogenetic (i.e. cross-population) mean and variance of the mean phenotype.

Next, in order to derive expressions for the evolutionary dynamics of mean phenotypes, we need to derive expressions for the moments of allelic state at each locus. These

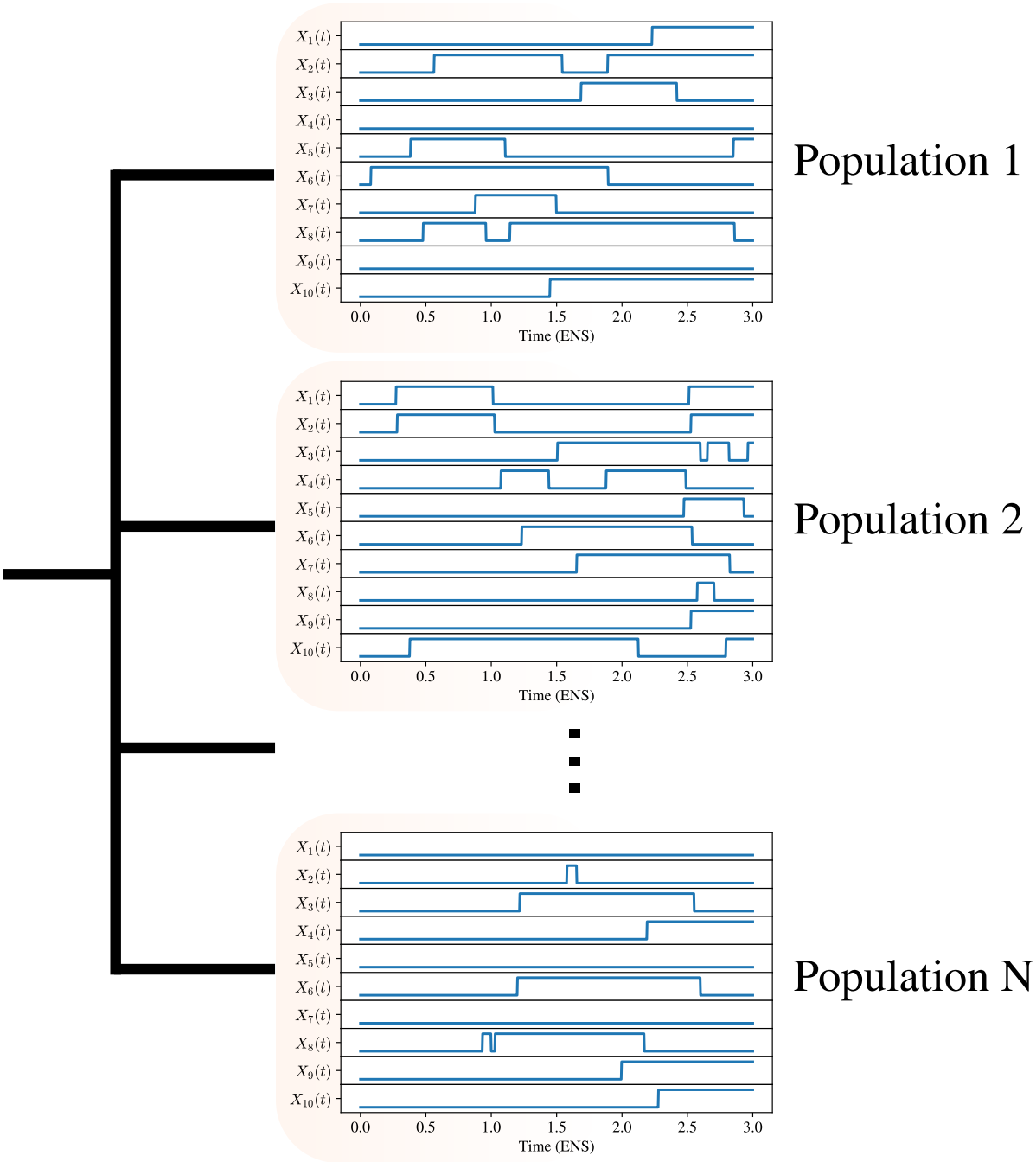


Figure 2: An illustration of the evolutionary process in an ensemble of *independent* populations or experimental replicates. At time $t = 0$, a single founder population gives rise to an ensemble of N separate populations that are then presumed to evolve completely independently. In each population, the (bi-allelic) loci are subject to mutation and drift, which can result in the fixation or extinction of the ancestral allele. This evolutionary process is modeled with a Continuous Time Markov Chain (CTMC) that jumps between two states at each locus.

quantities will be used in later sections to derive the phylogenetic (i.e. ensemble) mean and phylogenetic covariance for the mean phenotype.

First, we derive an expression for the expected state of the locus at time t. Given that the state of the locus at any point in time is a Bernoulli random variable with the success probability corresponding to the probability of the alternative allele being fixed, we can write the expected value for the state of locus j at time t as:

$$\begin{aligned}\mathbb{E}[X_j(t)] &= P(X_j(t) = 1) \\ &= P(X_j(0) = 1)e^{-\lambda(1+\kappa)t} + \pi(1 - e^{-\lambda(1+\kappa)t})\end{aligned} \tag{5}$$

Second, we derive an expression for the the covariance in the state of the locus between two descendant lineages that diverged at time t and then evolved completely independently for s and r time units, respectively. This idea is illustrated in Figure 4(a). Here, we use the standard definition of the covariance:

$$Cov\Big(X_j(t+s), X_j(t+r)\Big) = \mathbb{E}[X_j(t+s)X_j(t+r)] - \mathbb{E}[X_j(t+s)]\,\mathbb{E}[X_j(t+r)]$$

The unknown quantity in this case is the joint expectation of the state of the locus $\mathbb{E}[X_j(t+s)X_j(t+r)]$, which we write in terms of the joint probability that the locus is fixed for the alternative allele in both lineages:

$$\mathbb{E}[X_j(t+s)X_j(t+r)] = P\Big(X_j(t+s) = 1, \quad X_j(t+r) = 1\Big)$$

In order to express this in quantities that we have previously defined, we make use of the Law of Total Probability and factor the joint probability in the equation above into the marginal probability of the state of the locus in the ancestral population at time t and the probability of transitions in times s and r (see Figure 4(a)):

$$\begin{aligned}&\mathbb{E}[X_j(t+s)X_j(t+r)] = \\ &\Bigg(P(X_j(t) = 0)P_{01}(s)P_{01}(r) + P(X_j(t) = 1)P_{11}(s)P_{11}(r)\Bigg) \\ &= P(X_j(t) = 1)\Big(P_{11}(s)P_{11}(r) - P_{01}(s)P_{01}(r)\Big) + P_{01}(s)P_{01}(r)\end{aligned}$$

The factorization above assumes that the evolutionary process is independent after time t. As we will discuss later, this quantity will be most relevant in computing phylogenetic covariance of species that split at time t and then evolve for s and r time units completely independently. However, it is general and will be used to compute expressions for the time-dependent variance when we set $r = s = 0$. After substituting and simplifying, we obtain the following expression for the phylogenetic covariance in the state of the locus in the two descendant lineages:

$$
\begin{aligned}
Cov\Big(X_j(t+s), X_j(t+r)\Big) &= e^{-\lambda(1+\kappa)(r+s)} P(X_j(t)=1)\Big(1 - P(X_j(t)=1)\Big) \\
&= e^{-\lambda(1+\kappa)(r+s)}\Bigg[\pi(1-\pi) + e^{-\lambda(1+\kappa)t}(2\pi-1)\Big(\pi - P(X_j(0)=1)\Big) \\
&\qquad - e^{-2\lambda(1+\kappa)t}\Big(\pi - P(X_j(0)=1)\Big)^2\Bigg]
\end{aligned} \tag{6}
$$

The above expression indicates that as s and r become large, the phylogenetic covariance for any given locus j decays exponentially towards zero.

1.2 Mean phenotype statistics

We defined the mean phenotype as a linear function of L bi-allelic loci whose state at any point in time we model with independent CTMCs with 2 states, corresponding to the reference and alternative alleles. The equation for the mean phenotype in a given population is:

$$
Z(t) = \beta_0 + \sum_{j=1}^{L} X_j(t)\beta_j
$$

In this section, we derive expressions for the moments of this mean phenotype as a function of time and in the process we will comment on the dynamics that we observe.

1.2.1 Phylogenetic mean of mean phenotype

In order to obtain expressions for the phylogenetic or ensemble mean of mean phenotype, we take the expectation with respect to all the random variables:

$$
\mathbb{E}[Z(t)] = \beta_0 + \sum_{j=1}^{L} \mathbb{E}[X_j(t)]\beta_j
$$

The above equation follows from the linearity the expectation and the assumption of independence between sites. If we plug in the definition of $\mathbb{E}[X_j(t)]$ from Equation 5 into the equation above, we obtain:

$$
\begin{aligned}
\mathbb{E}[Z(t)] &= \beta_0 + \sum_{j=1}^{L}\Bigg(P(X_j(0)=1)e^{-\lambda(1+\kappa)t} + \pi(1-e^{-\lambda(1+\kappa)t})\Bigg)\beta_j \\
&= \beta_0 + e^{-\lambda(1+\kappa)t}\underbrace{\sum_{j=1}^{L}\beta_j P(X_j(0)=1)}_{\mathbb{E}[Z(0)]-\beta_0} + (1-e^{-\lambda(1+\kappa)t})\underbrace{\sum_{j=1}^{L}\beta_j\pi}_{Z_{eq}-\beta_0} \\
&= Z_0 e^{-\lambda(1+\kappa)t} + Z_{eq}(1-e^{-\lambda(1+\kappa)t})
\end{aligned} \tag{7}
$$

Where $\mathbb{E}[Z(0)]$ is the expected value of the mean phenotype at $t = 0$ and $Z_{eq} := \lim_{t\to\infty}\mathbb{E}[Z(t)]$ is the expected value of the mean phenotype at equilibrium. In phylogenetic modeling, we often assume that there is a single founding population at time $t = 0$,

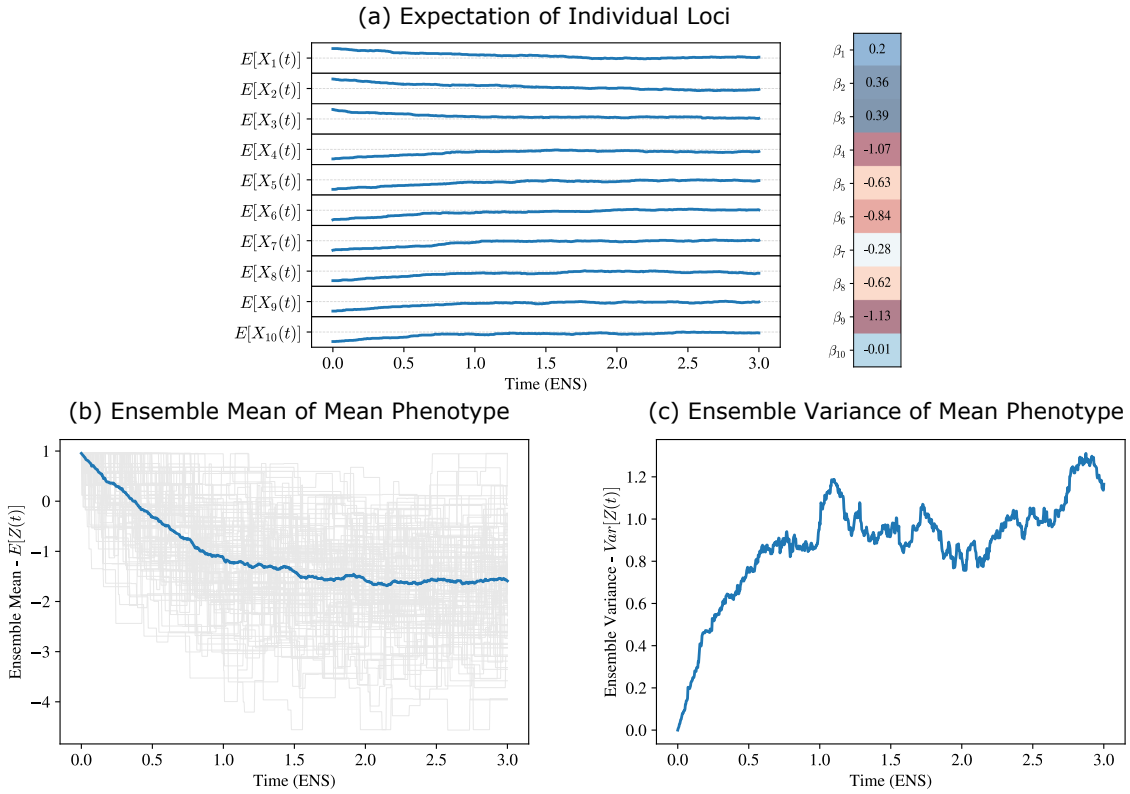


Figure 3: The ensemble mean and variance in mean phenotype as a function of time. Panel **(a)** shows the single locus statistics for $L = 10$ bi-allelic loci, which are obtained by averaging over the state of each locus in $N = 100$ independent replicates. In the ancestral population, the first three loci were fixed for the alternative allele while the remaining loci were fixed for the reference allele. Given the effect sizes $\boldsymbol{\beta}$, this results in an ancestral mean phenotype of $Z_0 \approx 1$. In this simulation we set $\kappa = 1$ for all loci and thus at equilibrium the expected value tends towards $\pi = 0.5$ for each locus. Panel **(b)** shows the mean phenotype in those replicate populations as a function of time (gray lines) as well as the ensemble mean $\mathbb{E}[Z(t)]$ (blue line) that is obtained by averaging over the mean phenotype in all replicates. Panel **(c)** shows the empirical variance in mean phenotype across the different replicates. Note the variance saturates and reaches its equilibrium roughly twice as fast as the ensemble mean.

and we use this assumption to justify dropping the expectation from $\mathbb{E}[Z(0)] = Z_0$. If we were to fit this model to data, both the initial phenotype Z_0 and the equilibrium mean phenotype Z_{eq} would be free parameters.

The above expression for the mean has the exact mean-reverting form as the mean of the Ornstein-Uhlenbeck process, as claimed in the main text and illustrated in Figure 3(b). Formulated in this way, we see that the role of mutation bias is to modulate the strength of the mutational force on the mean phenotype: if the bias is in the direction of the alternative allele, the mutational equilibrium is obtained faster, while if the bias is toward the reference allele, the equilibrium is obtained slower. In practice, since the bias would be on the order of 1, the effect on the mutational force is small, and with no bias, $\kappa = 1$, the force of mutation remains.

1.2.2 Phylogenetic covariance of mean phenotype

The second quantity that we need to characterize the evolutionary dynamics of neutral phenotypes is the covariance, particularly the phylogenetic covariance between two or more species. The scenario that we envision in this case is that of an ancestral population that evolved for t time units and then split into two distinct species that evolved completely independently. The first species evolved for s time units and the second species evolved for r time units (Figure 4). This general characterization is flexible and allows for modeling non-ultrametric phylogenetic trees, which are common in studies of molecular evolution.

We use the definition of the covariance between two random variables and then plug in the definitions and simplify to obtain:

$$
\begin{aligned}
Cov\Big(Z(t+s), Z(t+r)\Big) &= \mathbb{E}[Z(t+s)Z(t+r)] - \mathbb{E}[Z(t+s)]\,\mathbb{E}[Z(t+r)] \\
&= \mathbb{E}\left[(\beta_0 + \sum_{j=1}^{L} X_j(t+s)\beta_j)(\beta_0 + \sum_{j=1}^{L} X_j(t+r)\beta_j)\right] - \\
&\quad \left(\beta_0 + \sum_{j=1}^{L} \mathbb{E}[X_j(t+s)]\beta_j\right)\left(\beta_0 + \sum_{j=1}^{L} \mathbb{E}[X_j(t+r)]\beta_j\right) \\
&= \sum_{j=1}^{L}\sum_{i=1}^{L} \beta_j\beta_i\Big(\mathbb{E}[X_j(t+s)X_i(t+r)] - \mathbb{E}[X_j(t+s)]\,\mathbb{E}[X_i(t+r)]\Big) \\
&= \sum_{j=1}^{L} \beta_j^2\Big(\mathbb{E}[X_j(t+s)X_j(t+r)] - \mathbb{E}[X_j(t+s)]\,\mathbb{E}[X_j(t+r)]\Big) \\
&= \sum_{j=1}^{L} \beta_j^2 Cov\Big(X_j(t+s), X_j(t+r)\Big)
\end{aligned}
$$

The last equality is due to the assumption of independence between sites. The phylogenetic covariance then reduces to the sum of the covariances between the state of each locus j at times $(t+s)$ and $(t+r)$, which we defined in the previous section. If we substitute the definitions and simplify, we obtain:

$$
\begin{aligned}
Cov\Big(Z(t+s), Z(t+r)\Big) &= e^{-\lambda(1+\kappa)(r+s)} \sum_{j=1}^{L} \beta_j^2 P(X_j(t)=1)\Big(1 - P(X_j(t)=1)\Big) \\
&= e^{-\lambda(1+\kappa)(r+s)} \sum_{j=1}^{L} \beta_j^2 \Bigg[(2\pi - 1)(\pi - P(X_j(0)=1))e^{-\lambda(1+\kappa)t} - \\
&\qquad (\pi - P(X_j(0)=1))^2 e^{-\lambda(1+\kappa)2t} + \pi(1-\pi)\Bigg]
\end{aligned} \tag{8}
$$

In this case, the covariance cannot be straightforwardly expressed in terms of phenotypic quantities such as Z_0 and Z_{eq}, and therefore depends on the genotype. In Section 3, we will show how to abstract away these genotypic quantities in order to perform inference at the phenotypic level.

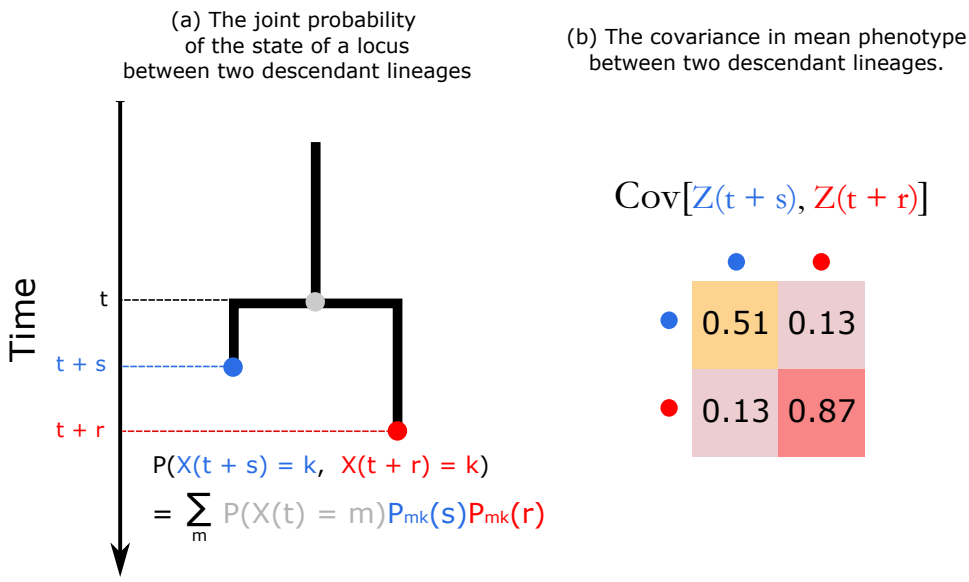


Figure 4: The phylogenetic covariance in mean phenotype as a function of time. In panel **(a)**, we illustrate the decomposition of the joint probability of the state of a locus X_j in two descendant lineages at times $t+s$ and $t+r$ into the marginal probability of the state of the locus in the ancestral population at time t and the probability of *independent* transitions in times s and r. Panel **(b)** shows a numerical example of the covariance matrix of the mean phenotype between those two descendant lineages.

1.2.3 Phylogenetic variance of mean phenotype

We can use the time-dependent covariance equation that we derived in the previous section to provide an explicit form for the phylogenetic variance of mean phenotype as a function of time. By substituting $r = s = 0$ and simplifying, we obtain:

$$\begin{aligned} Var\Big(Z(t)\Big) &= \sum_{j=1}^{L} \beta_j^2 P(X_j(t)=1)\Big(1 - P(X_j(t)=1)\Big) \\ &= \sum_{j=1}^{L} \beta_j^2 \Big((2\pi - 1)(\pi - P(X_j(0)=1))e^{-\lambda(1+\kappa)t} - \\ &\qquad e^{-\lambda(1+\kappa)2t}(\pi - P(X_j(0)=1))^2 + \pi(1-\pi)\Big) \end{aligned} \tag{9}$$

Unlike what the Brownian Motion model of phenotypic evolution predicts, this expression for the variance implies that it tends towards a stable equilibrium in the limit as $t \to \infty$ (see Figure 3(c)):

$$\sigma_{eq}^2 := \lim_{t\to\infty} Var\Big(Z(t)\Big) = \sum_{j=1}^{L} \beta_j^2 \pi(1-\pi)$$

Where we define σ_{eq}^2 as the ensemble equilibrium variance of mean phenotype.

2 K alleles

While the bi-allelic model is useful for the purposes of deriving insights into the evolutionary dynamics of simple traits, in reality many of the the traits that are of interest to molecular biologists are underpinned by multi-allelic genetic architectures. This includes DNA sequence properties, such as GC content, as well as biophysical properties of peptides,such as net charge of a protein. When examining these types of molecular phenotypes on phylogenetic time-scales, accounting for multiple alleles becomes important. In this scenario, we model the phenotype as a linear function of L loci, each of which can have up to K different alleles:

$$y = tr(\boldsymbol{X}^\top \boldsymbol{\beta}) + \epsilon = \sum_{j=1}^{L} \boldsymbol{X}_j \cdot \boldsymbol{\beta}_j + \epsilon = \sum_{j=1}^{L} \sum_{k=1}^{K} X_{jk}\beta_{jk} + \epsilon$$

Where $tr()$ is the trace operator, $\boldsymbol{X}$ is a $L \times K$ genotype matrix with each column indicating which of the K alleles is fixed, and $\boldsymbol{\beta}$ is the corresponding matrix of effect sizes. In this setup, X_{jk} is a binary random variable indicating whether the kth allele is fixed at the jth locus. One main deviation from the bi-allelic model outlined previously is that we no longer maintain an intercept or a "reference allele", instead keeping track of all alleles simultaneously.

We make use of the same assumptions that we employed in deriving the dynamics of the bi-allelic model, giving the following expression for the mean phenotype in a single population as a function of time:

$$Z(t) = tr(\boldsymbol{X}(t)^\top \boldsymbol{\beta}) = \sum_{j=1}^{L} \boldsymbol{X}_j(t) \cdot \boldsymbol{\beta}_j = \sum_{j=1}^{L} \sum_{k=1}^{K} X_{jk}(t)\beta_{jk} \tag{10}$$

2.1 Single locus statistics

In the multi-allelic case, we model the genotype at each locus j with a CTMC that jumps between K discrete states, corresponding to the K different alleles. For simplicity, we make use of a generator matrix that is similar to that of the Felsenstein (1981) model of DNA evolution [5, 14]. This model gives the following transition probabilities for a given allele k:

$$P(X_j(t) = k | X_j(0) = k) = P_{kk}(t) = e^{-ut} + \pi_k(1 - e^{-ut})$$
$$P(X_j(t) = k | X_j(0) \neq k) = P_{*k}(t) = \pi_k(1 - e^{-ut})$$

Here, $X_j(t) = k$ denotes that the kth allele is fixed at the jth locus at time t and π_k is the equilibrium probability of observing the kth allele. The first equation is the conditional probability of the kth allele staying fixed after a time interval t, given that it was fixed in the founding population. The second equation gives the probability of transitioning to the kth allele after a time interval of t, given that any of the other alleles was fixed in the founding population. Unlike in the bi-allelic model, here we measure time in units of expected number of substitutions per site and thus define the scaling parameter u as:

$$u := \frac{1}{1 - \sum_k \pi_k^2}$$

The equations above provide the probability of fixation or extinction in a given time interval conditional on the ancestral state. To get the probability that the kth allele is fixed at any point in time, we marginalize over the ancestral state:

$$\begin{aligned} P(X_j(t) = k) &= P(X_j(0) = k)P_{kk}(t) + \Big(1 - P(X_j(0) = k)\Big)P_{*k} \\ &= P(X_j(0) = k)e^{-ut} + \pi_k(1 - e^{-ut}) \end{aligned}$$

With this in hand, we can derive the single locus statistics that we would need to provide expressions for the dynamics of the mean phenotype. First, the expected value for the quantity $X_{jk}(t)$ becomes:

$$\mathbb{E}[X_{jk}(t)] = P(X_j(t) = k) = P(X_j(0) = k)e^{-ut} + \pi_k(1 - e^{-ut}) \tag{11}$$

In addition to this, we need expressions for the time-dependent covariance between the allelic state of the locus j in two descendant lineages (see Figure 4). However, unlike the bi-allelic model, here we have to consider 2 scenarios: $Cov\Big(X_{jk}(t+s), X_{jk}(t+r)\Big)$ and $Cov\Big(X_{jk}(t+s), X_{jm}(t+r)\Big)$. The first scenario corresponds to the covariance in the state of the allele (fixed vs. extinct) in the two lineages, while the second scenario captures the covariance between the state of 2 different alleles (indexed by k and m). To derive expressions for those covariances, we first start by writing out expressions for the joint expectation. As we did previously in deriving the dynamics of the bi-allelic model, we make use of the Law of Total Probability to decompose the joint probabilities into familiar quantities.

For the first covariance term, we decompose the joint probability into the marginal probability of the state of the locus in the ancestral population at time t and the probability of independent transitions to the kth allele in times s and r:

$$\begin{aligned} \mathbb{E}[X_{jk}(t+s)X_{jk}(t+r)] &= P\Big(X_j(t+s) = k, \quad X_j(t+r) = k\Big) \\ &= P(X_j(t) = k)P_{kk}(s)P_{kk}(r) + \Big(1 - P(X_j(t) = k)\Big)P_{*k}(s)P_{*k}(r) \end{aligned}$$

The decomposition for the second covariance term is similar. However, in the ancestral population at time t there are three possible cases, mainly that the locus was fixed for the kth allele, for the mth allele, and for neither allele:

$$\begin{aligned} \mathbb{E}[X_{jk}(t+s)X_{jm}(t+r)] &= P\Big(X_j(t+s) = k, \quad X_j(t+r) = m\Big) \\ &= P(X_j(t) = k)P_{kk}(s)P_{*m}(r) + \\ &\quad P(X_j(t) = m)P_{*k}(s)P_{mm}(r) + \\ &\quad \Big(1 - P(X_j(t) = k) - P(X_j(t) = m)\Big)P_{*k}(s)P_{*m}(r) \end{aligned}$$

With this in hand, after plugging in the definitions and simplifying the algebra, we can write the single-locus covariances as:

$$
\begin{aligned}
&Cov\Big(X_{jk}(t+s), X_{jk}(t+r)\Big) = \\
&\quad \mathbb{E}[X_{jk}(t+s)X_{jk}(t+r)] - \mathbb{E}[X_{jk}(t+s)]\,\mathbb{E}[X_{jk}(t+r)] \\
&\quad = e^{-u(r+s)} P(X_j(t)=k)\Big(1 - p(X_j(t)=k)\Big)
\end{aligned} \tag{12}
$$

Which we can expand to obtain:

$$
\begin{aligned}
&Cov\Big(X_{jk}(t+s), X_{jk}(t+r)\Big) = \\
&\quad e^{-u(r+s)}\Bigg[\pi_k(1-\pi_k) + e^{-ut}(2\pi_k - 1)\Big(\pi_k - P(X_j(0)=k)\Big) \\
&\qquad - e^{-2ut}\Big(\pi_k - P(X_j(0)=k)\Big)^2\Bigg]
\end{aligned}
$$

And the covariance of two different alleles can be written as:

$$
\begin{aligned}
&Cov\Big(X_{jk}(t+s), X_{jm}(t+r)\Big) = \\
&\quad \mathbb{E}[X_{jk}(t+s)X_{jm}(t+r)] - \mathbb{E}[X_{jk}(t+s)]\,\mathbb{E}[X_{jm}(t+r)] \\
&\quad = -e^{-u(r+s)} P(X_j(t)=k)P(X_j(t)=m)
\end{aligned} \tag{13}
$$

Naturally, the first covariance term resembles the expression for the single locus covariance that we derived for the bi-allelic model. The second covariance term is unique, since it captures the covariance between 2 distinct alleles. With these quantities in hand, and remembering that we still maintain the assumption of independence between sites, we can derive the mean phenotype statistics, which we turn to in the next section.

2.2 Mean phenotype statistics

Recall that we defined the mean phenotype as a linear combination of L loci whose state at any point in time we model with independent CTMCs with K states corresponding to the K distinct alleles. The equation for the mean phenotype in a given population is given by:

$$
Z(t) = tr(\boldsymbol{X}(t)^{\top}\boldsymbol{\beta}) = \sum_{j=1}^{L} \boldsymbol{X}_j(t)\cdot\boldsymbol{\beta}_j = \sum_{j=1}^{L}\sum_{k=1}^{K} X_{jk}(t)\beta_{jk}
$$

In this section, we derive expressions for the moments of this mean phenotype as a function of time and in the process we will comment on the dynamics that we observe.

2.2.1 Phylogenetic mean of mean phenotype

In order to obtain expressions for the phylogenetic or ensemble mean of the mean phenotype, we take the expectation with respect all random components of this model:

$$
\mathbb{E}[Z(t)] = \sum_{j=1}^{L}\sum_{k=1}^{K} \mathbb{E}[X_{jk}(t)]\beta_{jk}
$$

The above equation follows from the linearity the expectation and the assumption of independence between sites. If we plug in the definition of $\mathbb{E}[X_{jk}(t)]$ from above, we obtain:

$$\begin{aligned}
\mathbb{E}[Z(t)] &= \sum_{j=1}^{L}\sum_{k=1}^{K}\Big[P(X_j(0)=k)e^{-ut} + (1-e^{-ut})\pi_k\Big]\beta_{jk} \\
&= e^{-ut}\sum_{j=1}^{L}\sum_{k=1}^{K}P(X_j(0)=k)\beta_{jk} + (1-e^{-ut})\sum_{j=1}^{L}\sum_{k=1}^{K}\pi_k\beta_{jk} \\
&= e^{-ut}Z_0 + (1-e^{-ut})Z_{eq}
\end{aligned} \tag{14}$$

Where we define $Z_0 := \sum_{j=1}^{L}\sum_{k=1}^{K}P(X_j(0)=k)\beta_{jk}$ and $Z_{eq} := \sum_{j=1}^{L}\sum_{k=1}^{K}\pi_k\beta_{jk}$ as the mean phenotype in the founding population and the ensemble mean at equilibrium, respectively. Similar to what we saw with the bi-allelic model, the expresion for the mean has the exact form of the mean dynamics of the Ornstein-Uhlenbeck process.

2.2.2 Phylogenetic covariance of mean phenotype

To derive an expression for the covariance of the multi-allelic model, we use the definition of the covariance and simplify:

$$\begin{aligned}
Cov\Big(Z(t+s), Z(t+r)\Big) &= \mathbb{E}[Z(t+s)Z(t+r)] - \mathbb{E}[Z(t+s)]\,\mathbb{E}[Z(t+r)] \\
&= \mathbb{E}\left[(\sum_{j=1}^{L}\sum_{k=1}^{K}X_{jk}(t+s)\beta_{jk})(\sum_{j=1}^{L}\sum_{k=1}^{K}X_{jk}(t+r)\beta_{jk})\right] - \\
&\quad \left(\sum_{j=1}^{L}\sum_{k=1}^{K}\mathbb{E}[X_{jk}(t+s)]\beta_{jk}\right)\left(\sum_{j=1}^{L}\sum_{k=1}^{K}\mathbb{E}[X_{jk}(t+r)]\beta_{jk}\right) \\
&= \sum_{j=1}^{L}\sum_{i=1}^{L}\sum_{k=1}^{K}\sum_{m=1}^{K}\beta_{ik}\beta_{jm}\Big(\mathbb{E}[X_{ik}(t+s)X_{jm}(t+r)] - \mathbb{E}[X_{ik}(t+s)]\,\mathbb{E}[X_{jm}(t+r)]\Big) \\
&= \sum_{j=1}^{L}\sum_{k=1}^{K}\beta_{jk}^2 Cov\Big(X_{jk}(t+s), X_{jk}(t+r)\Big) + \sum_{m\neq k}^{K}\beta_{jk}\beta_{jm}Cov\Big(X_{jk}(t+s), X_{jm}(t+r)\Big)
\end{aligned}$$

The last equality holds because of the assumption of independence between sites. After plugging in the single-locus covariances that we derived previously, we can write the phylogenetic covariance in mean phenotype as:

$$\begin{aligned}
&Cov\Big(Z(t+s), Z(t+r)\Big) = \\
&e^{-u(r+s)}\Bigg[\sum_{j=1}^{L}\sum_{k=1}^{K}\beta_{jk}^2 P(X_j(t)=k)\Big(1-P(X_j(t)=k)\Big) \\
&\qquad - \sum_{m\neq k}^{K}\beta_{jk}\beta_{jm}\Big(P(X_j(t)=k)P(X_j(t)=m)\Big)\Bigg]
\end{aligned} \tag{15}$$

2.2.3 Phylogenetic variance of mean phenotype

To obtain an expression for the phylogenetic variance of mean phenotype as a function of time, we simply set $s = r = 0$ in the covariance equation we obtained above, giving:

$$\begin{aligned} &Var\Big(Z(t)\Big) = \\ &\sum_{j=1}^{L}\sum_{k=1}^{K}\beta_{jk}^2 P(X_j(t)=k)\Bigg(1 - P(X_j(t)=k)\Bigg) \\ &\quad - \sum_{m\neq k}^{K}\beta_{jk}\beta_{jm}\Bigg(P(X_j(t)=k)P(X_j(t)=m)\Bigg) \end{aligned} \tag{16}$$

We note that the expression for the variance corresponds to the variance of a categorical random variable with K outcomes, as expected. In general, the variance appears to depend on the initial genotype and is non-monotonic. Furthermore, in line with the results from the bi-allelic model, the variance in this model is bounded. The ensemble equilibrium variance can be obtained by taking the limit as $t \to \infty$, giving:

$$\sigma_{eq}^2 := \sum_{j=1}^{L}\sum_{k=1}^{K}\beta_{jk}^2\pi_k(1-\pi_k) - \sum_{m\neq k}^{K}\beta_{jk}\beta_{jm}\pi_k\pi_m$$

3 Simplifying and extending the model

3.1 Lumping alleles and the grouped model

In this section, we will show that the multi-allelic models can be vastly simplified for a wide-range of traits by only tracking the state of a number of alleles that is equivalent to the *unique* number of effect sizes. For example, for counting-based traits such as GC Content, instead of modeling the contribution of all four DNA letters to the dynamics, it suffices to keep track of two "lumped" or grouped alleles, corresponding to the G/C and non-G/C state. This construction will help simplify analyzing the evolutionary dynamics of many molecular and cellular traits.

In particular, we will show that the ensemble mean and covariance dynamics under the grouped model are almost identical to the dynamics under the full model, with a an important difference in the scaling factor that we will comment on shortly. To illustrate, we assume that out of the K alleles that characterize the underlying sequence, an arbitrary subset, denoted by S, have an effect size of β_1 while the remaining alleles have an effect size of β_2:

$$\beta_k = \begin{cases} \beta_1 & k \in S \\ \beta_2 & k \notin S \end{cases}$$

Furthermore, since the fixation of alleles in a given population constitutes disjoint events by construction, we can write the probability of observing any of the alleles in the subset S at time t as:

$$P(X_j(t) \in S) = \sum_{k \in S} P(X_j(t) = k)$$

With the complement of this being $P(X_j(t) \notin S) = 1 - P(X_j(t) \in S)$. And because this is a bi-allelic model where the alleles are the states S and its complement, we observe that the single locus covariance can be written as:

$$Cov(X_j(t+s) \in S, X_j(t+r) \in S) = e^{-u(r+s)} P(X_j(t) \in S)\Big(1 - P(X_j(t) \in S)\Big)$$

Phylogenetic Mean of Mean Phenotype Given our construction above, we can write the contribution of a single locus j to the phylogenetic mean under the grouped model as:

$$\begin{aligned} \mathbb{E}[Z(t)] &= \sum_{j=1}^{L} \beta_1 P(X_j(t) \in S) + \beta_2 P(X_j(t) \notin S) \\ &= \sum_{j=1}^{L} \beta_1 \sum_{k \in S} P(X_j(t) = k) + \beta_2 \sum_{k \notin S} P(X_j(t) = k) \end{aligned}$$

Thus, the dynamics of the phylogenetic mean under the full and grouped models are identical. In the case of the mean, it is straightforward to generalize this to any arbitrary grouping of alleles by their effect size.

Phylogenetic Covariance of Mean Phenotype Showing that the equivalence holds for the covariance requires more work. In order to show this, it helps to re-write the covariance in the full model as follows and then simplify:

$$
\begin{aligned}
&Cov\Big(Z(t+s), Z(t+r)\Big) \\
&= e^{-u(r+s)} \sum_{j=1}^{L} \sum_{k=1}^{K} \beta_{jk}^2 P(X_j(t)=k)\Big(1 - P(X_j(t)=k)\Big) - \sum_{m \neq k}^{K} \beta_{jk}\beta_{jm} P(X_j(t)=k)P(X_j(t)=m) \\
&= e^{-u(r+s)} \sum_{j=1}^{L} \sum_{k=1}^{K} \sum_{m \neq k}^{K} \beta_{jk}^2 P(X_j(t)=k)P(X_j(t)=m) - \beta_{jk}\beta_{jm} P(X_j(t)=k)P(X_j(t)=m) \\
&= e^{-u(r+s)} \sum_{j=1}^{L} \sum_{k,m} (\beta_{jk} - \beta_{jm})^2 P(X_j(t)=k)P(X_j(t)=m)
\end{aligned}
$$

Where the sum $\sum_{k,m}$ is over *unique pairs* of alleles j and m. The second equality follows from the fact that $P(X_j(t)=k)\Big(1-P(X_j(t)=k)\Big) = \sum_{m \neq k} P(X_j(t)=k)P(X_j(t)=m)$. When an effect size is shared between a pair of alleles, then $(\beta_{jk}-\beta_{jm})^2 = 0$ and their joint contribution to the covariance drops out of the sum. If there are only two unique effect sizes β_1 and β_2, then the sum above would simplify to:

$$
Cov\Big(Z(t+s), Z(t+r)\Big) = e^{-u(r+s)} \sum_{j=1} (\beta_1 - \beta_2)^2 \sum_{k \in S} \sum_{m \notin S} P(X_j(t)=k)P(X_j(t)=m)
$$

Where S is the subset of alleles with effect size β_1. This expression is exactly what we would obtain if we solved for the covariance in the grouped bi-allelic model and then plugged in the definitions of the probabilities from before:

$$
\begin{aligned}
Cov\Big(Z(t+s), Z(t+r)\Big) &= e^{-u(r+s)} \sum_{j=1}^{L} (\beta_1 - \beta_2)^2 P(X_j(t) \in S)\Big(1 - P(X_j(t) \in S)\Big) \\
&= e^{-u(r+s)} \sum_{j=1}^{L} (\beta_1 - \beta_2)^2 \left[\sum_{k \in S} P(X_j(t)=k)\right]\left[1 - \sum_{k \in S} P(X_j(t)=k)\right] \\
&= e^{-u(r+s)} \sum_{j=1}^{L} (\beta_1 - \beta_2)^2 \sum_{k \in S} \sum_{m \notin S} P(X_j(t)=k)P(X_j(t)=m)
\end{aligned}
$$

Therefore, we have shown that the covariance dynamics under the full and grouped models are identical.

The scaling factor u under the grouped model An important detail that we omitted in the preceding discussion is the scaling factor u and whether it is the same or takes on a different value under the grouped model. A naive interpretation may suggest the u under the grouped model should be defined as:

$$
u = \frac{1}{1 - \sum_S (\sum_{k \in S} \pi_k)^2}
$$

Where the sum in the denominator is over the grouped alleles instead of the individual alleles in the underlying sequence. However, this is the major difference between the full

model and the grouped model: the scaling factor has to be defined in terms of the equilibrium probabilities of all alleles:

$$u = \frac{1}{1 - \sum_k \pi_k^2}$$

To see that this is the case, recall that in our construction of this grouped formulation, we defined the probabilities of observing the grouped allele as a sum of the probabilities of observing the individual alleles. And in these individual probabilities, the scaling factor is defined under the full model. Therefore, under the grouped model, u retains the same value and interpretation as under the full model.

3.2 Abstracting away the genetic architecture from the covariance

In our analysis of the model, we noted that the dynamics of the mean can be expressed in phenotypic quantities, abstracting away the underlying genetic architecture of the trait, including the number of causal loci and their effect sizes. Here, we will show that a similar abstraction can be achieved in the case of the covariance.

The covariance involves many terms that are harder to abstract away. This is mainly the case because of the cross-terms that describe the covariance between different alleles at each site. However, it is possible to write the covariance in terms that completely abstract away from the underlying genetic architecture. To see this, we start with the formulation of the covariance that we derived in the description of the grouped model (Section 3.1):

$$Cov\Big(Z(t+s), Z(t+r)\Big) = e^{-u(r+s)} \sum_{j=1}^{L} \sum_{k,m} (\beta_{jk} - \beta_{jm})^2 P(X_j(t) = k) P(X_j(t) = m)$$

Recalling the fact that $P(X_j(t) = k) = P(X_j(0) = k)e^{-ut} + \pi_k(1 - e^{-ut})$ and the assumption that we have a single ancestral monomorphic population (thus $P(X_j(0) = m)P(X_j(0) = k) = 0$), it is straightforward to show that the above expression simplifies to:

$$\begin{aligned} Cov\Big(Z(t+s), Z(t+r)\Big) =& e^{-u(r+s)}(1 - e^{-ut})\Bigg[\\ & e^{-ut} \sum_{j=1}^{L} \sum_{k,m} (\beta_{jk} - \beta_{jm})^2 \Big(P(X_j(0) = k)\pi_m + P(X_j(0) = m)\pi_k\Big) \\ & +(1 - e^{-ut}) \sum_{j=1}^{L} \sum_{k,m} (\beta_{jk} - \beta_{jm})^2 \pi_k \pi_m \Bigg] \end{aligned}$$

This expression can be further simplified by observing that when we set $r = s = 0$ and take the limit as $t \to \infty$, we get that the sum on the third line equals the ensemble equilibrium variance that we have previously defined: $\sigma_{eq}^2 = \sum_{j=1}^{L} \sum_{k,m} (\beta_{jk} - \beta_{jm})^2 \pi_k \pi_m$. Furthermore, introducing a new quantity that we denote by Ψ, we define:

$$\Psi = \sum_{j=1}^{L} \sum_{k,m} (\beta_{jk} - \beta_{jm})^2 \Big(P(X_j(0) = k)\pi_m + P(X_j(0) = m)\pi_k \Big) \tag{17}$$

The quantity Ψ does not have a straightforward interpretation in general, though it is always positive and we may gain insights into its role by analogy with the Gaussian models of trait evolution. For instance, if the evolutionary process starts at equilibrium (i.e. $P(X_j(0) = k) = \pi_k$), then we see that Ψ takes on the value of $2\sigma_{eq}^2$. By comparison, under the Ornstein-Uhlenbeck model, the equilibrium variance is $\frac{\sigma^2}{2\alpha}$ [1]. Furthermore, as we will show in Section 3.4, the quantity $u\Psi$ determines the rate at which the variance increases at small time. Under the OU model, on the other hand, this behavior is controlled by the parameter σ^2. All of this is suggestive of the fact that the quantity $u\Psi$ acts in similar ways to behavior of the dispersion parameter σ^2 in the Ornstein-Uhlenbeck model.

With this in hand, we can write the covariance as:

$$Cov\Big(Z(t+s), Z(t+r) \Big) = e^{-u(r+s)}(1 - e^{-ut})\Big(e^{-ut}(\Psi - \sigma_{eq}^2) + \sigma_{eq}^2 \Big) \tag{18}$$

Thus, we have arrived at an expression for the covariance in mean phenotype between two descendant lineages that abstracts away the underlying genetic architecture.

3.3 Phylogenetic divergence of mean phenotype

The expressions for phylogenetic mean and covariance derived in previous sections are sufficient to construct a likelihood-based inference framework to fit our model to data. Indeed, these quantities have been the central focus of analysis and extensions in the Phylogenetic Comparative Methods community [3, 7]. However, a more recent theoretical framework inspired by techniques from statistical physics has shown that another useful quantity to differentiate between the dynamics of neutral and adaptive evolution is the phylogenetic divergence of mean phenotype [11, 8, 10]. Defined as the expected squared difference in the mean phenotype in two extant lineages, the divergence can be expressed in terms of quantities that we defined above. Mathematically, we can write the pairwise divergence between the mean phenotypes of the two lineages as:

$$\begin{aligned} D\Big(Z(t+s), Z(t+r) \Big) &= \mathbb{E}\left[\Big(Z(t+s) - Z(t+r) \Big)^2 \right] \\ &= \mathbb{E}[Z(t+s)^2] - 2\,\mathbb{E}[Z(t+s)Z(t+r)] + \mathbb{E}[Z(t+r)^2] \end{aligned}$$

By observing that $\mathbb{E}[Z(t+s)^2] = Var\Big(Z(t+s) \Big) + \mathbb{E}[Z(t+s)]^2$ and that $\mathbb{E}[Z(t+s)Z(t+r)] = Cov\Big(Z(t+s), Z(t+r) \Big) + \mathbb{E}[Z(t+s)]\,\mathbb{E}[Z(t+r)]$, we can see that the phylogenetic divergence can be expressed in terms of the mean, variance, and covariance in the mean phenotype between the two lineages:

$$\begin{aligned} D\Big(Z(t+s), Z(t+r) \Big) =& Var\Big(Z(t+s) \Big) + Var\Big(Z(t+r) \Big) - 2Cov\Big(Z(t+s), Z(t+r) \Big) \\ &+ \left[\mathbb{E}[Z(t+s)] - \mathbb{E}[Z(t+r)] \right]^2 \end{aligned} \tag{19}$$

This is a general expression that can work with any model of phenotypic evolution. Under our model, the expression can be a bit unwieldy, though by examining the dynamics of this expression at equilibrium, we recover some of the general formulations for quantities that were defined in the framework of Nourmohammad et al. [11, 10]:

$$D_{t\to\infty}\Big(Z(t+s), Z(t+r)\Big) = 2\sigma_{eq}^2(1-e^{-u(r+s)})$$

In the limit of long divergence times (as $r+s$ becomes large), we see that this quantity becomes twice the equilibrium variance $2\sigma_{eq}^2$, which in this context is the *multi-allelic* generalization of the trait scale D_0 as defined in [11, 10].

3.4 Linear approximation for the variance at small time

In this section, our aim is to examine the dynamics of the ensemble variance at small time ($ut \ll 1$) and to obtain a linear approximation in this regime. Starting with the compact representation of the covariance in Equation 18 and setting $r = s = 0$, we obtain the following expression for the ensemble variance at time t:

$$Var\Big(Z(t)\Big) = (1-e^{-ut})\Big(e^{-ut}(\Psi - \sigma_{eq}^2) + \sigma_{eq}^2\Big)$$

At small time, we have that $1 - e^{-ut} \approx ut$ and $(1-e^{-ut})e^{-ut} \approx ut$, which gives the following linear expression for the variance at small time:

$$\begin{aligned} Var_{t\ll 1}\Big(Z(t)\Big) &\approx ut\Psi \\ &= ut\sum_{j=1}^{L}\sum_{k,m}(\beta_{jk}-\beta_{jm})^2\Big(P(X_j(0)=k)\pi_m + P(X_j(0)=m)\pi_k\Big) \end{aligned} \tag{20}$$

Thus, the variance dynamics at small time only depend on the quantity $u\Psi$, which we defined in previous sections and shown to be related to the dispersion of the evolutionary process.

3.5 Simplifying covariance in the bi-allelic model

There are a number of simplifying assumptions that we can invoke to gain insights into the behavior of the time-dependent covariance in the bi-allelic model.

Covariance at equilibrium One common assumption states that the evolutionary process starts in an equilibrium state and thus $P(X_j(0) = 1) = \pi$. This simplifies the expression for the covariance in the bi-allelic model to:

$$Cov\Big(Z(t+s), Z(t+r)\Big) = e^{-\lambda(1+\kappa)(r+s)}\underbrace{\sum_{j=1}^{L}\beta_j^2\pi(1-\pi)}_{\sigma_{eq}^2}$$

Where σ_{eq}^2 is the equilibrium variance of the mean phenotype Z. In this case, σ_{eq}^2 would be a free parameter in the model that can be fit to data. While the form of this is attractive, it is inconsistent with the weak-mutation assumption and empirical observations that genotypes are not at their equilibrium in finite populations, but rather

close to monomorphic. For example, this formula predicts that the cross-species variance is always at its mutational equilibrium.

Covariance with no mutational bias A different simplification is obtained if we assume that the mutation rate is symmetric between the 2 alleles (i.e. $\kappa = 1$) and that there was a single founding population (i.e. $P(X_j(0) = 1) \in \{0, 1\}$). The latter assumption is common in phylogenetic analysis and it implies that the initial variance is zero. Under this scenario, we obtain that $\pi = 0.5$, and therefore the covariance simplifies to:

$$Cov\Big(Z(t+s), Z(t+r)\Big) = e^{-2\lambda(r+s+2t)} \sum_{j=1}^{L} \beta_j^2 \Bigg[P(X_j(0) = 1)\Big(1 - P(X_j(0) = 1)\Big) - \frac{1}{4}(1 - e^{4\lambda t}) \Bigg]$$

By observing that $P(X_j(0) = 1)\Big(1 - P(X_j(0) = 1)\Big) = 0$ and recalling the definition of the equilibrium variance, we can write the covariance under these assumptions as:

$$Cov\Big(Z(t+s), Z(t+r)\Big) = e^{-2\lambda(r+s)}(1 - e^{-4\lambda t})\sigma_{eq}^2$$

By setting $s = r = 0$, we also obtain an expression for the variance under this model, which turns out to match exactly the variance dynamics of the Ornstein-Uhlenbeck process, as alluded to previously:

$$Var\Big(Z(t)\Big) = (1 - e^{-4\lambda t})\sigma_{eq}^2$$

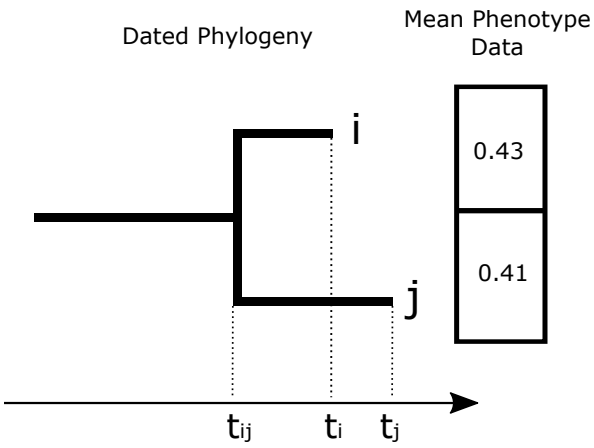


Figure 5: An illustration of the phylogenetic inference setup. In typical applications, we have a dated phylogeny, where time is in units of Expected Number of Substitutions (ENS), as well as mean phenotype measurement in extant species. These inputs are typically combined with a parametric model, e.g. a multivariate Gaussian likelihood, whose parameters are optimized using Maximum Likelihood.

4 Phylogenetic inference under the model

The theoretical model that we outlined in the previous sections provides a general picture of the dynamics of the moments of neutral phenotypes on phylogenetic time-scales. A natural extension of this model would be to use it for phylogenetic inference tasks that are popularly encountered within the realm of Phylogenetic Comparative Methods (PCMs). For example, this model can be used to reconstruct the ancestral mean phenotype, assuming that we have sufficient evidence that the phenotype evolves under little selective constraint. Another important application of this model is to provide a more powerful test for selection, replacing Brownian motion as the null model of phenotypic evolution. For these and other tasks, we need to derive a parametric model for the evolutionary dynamics of mean phenotype that can be fit to data using a Maximum Likelihood framework.

4.1 The likelihood

We assume that we are given a (rooted) phylogeny that describes the relationships between N species $\{S_1, \ldots, S_N\}$ and the divergence times that separate them in units of Expected Number of Substitutions per site (ENS). In this setup, we assume that species i is removed from the root of the phylogeny by t_i time units and that a pair of species i and j diverged at time t_{ij} and then each evolved for $t_i - t_{ij}$ and $t_j - t_{ij}$ time units, respectively. Additionally, assuming that we have measurements of the mean phenotype in each of those species $\boldsymbol{Y} = \{y_1, \ldots, y_N\}$, our aim in this section is to define a likelihood for the data given the parameters of the model $\mathcal{L}(\boldsymbol{Y}|\theta)$. See Figure 5 for an illustration of this setup.

Our theoretical model does not describe an explicit probability distribution. However, assuming that the trait of interest is sufficiently Gaussian, we can approximate its time-dependent dynamics with a Gaussian Process that is characterized by the following mean and covariance terms:

$$\mathcal{L}(\boldsymbol{Y}|\theta) = \mathcal{N}_{MVN}(\boldsymbol{Y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{21}$$

Where $\mathcal{N}_{MVN}$ is a multivariate Gaussian distribution with N coordinates and where the entry in the mean vector for species i is defined as:

$$\mu_i = Z_0 e^{-ut_i} + Z_{eq}(1 - e^{-ut_i}) \tag{22}$$

And the covariance between any pair of species i and j is defined as:

$$\Sigma_{ij} = e^{-u(t_i+t_j-2t_{ij})}(1 - e^{-ut_{ij}})\left(e^{-ut_{ij}}(\Psi - \sigma_{eq}^2) + \sigma_{eq}^2\right) \tag{23}$$

This most general model has a total of 5 parameters: $\boldsymbol{\theta} = (Z_0, Z_{eq}, u, \Psi, \sigma_{eq}^2)$. However, note that for many traits and species of interest, the parameter u can be fixed to reasonable values based on genome-wide estimates of the equilibrium probabilities of the 4 DNA letters. For instance, assuming a Jukes-Cantor (1969) model of DNA evolution [9, 14], u may be preset to 1.3. This results in a neutral null model of phenotypic evolution with 4 parameters to estimate, similar to the standard formulation of the phylogenetic Ornstein-Uhlenbeck model [7]. In some applications, when analyzing simple linear traits, e.g. GC Content, other parameters such as (Z_{eq}, σ_{eq}^2) may also be fixed to reasonable values that can be straightforwardly derived based on the equilibrium probabilities of the 4 DNA letters. An example of this will be shown in Section 5.

While the general model has a relatively large number of parameters, these parameters are easily interpretable and it is straightforward for practitioners to examine their meaning or add reasonable constraints on their values in model fitting procedures. For example, the parameters u can be restricted to the range $(1, 2)$. Additionally, the parameters Ψ and σ_{eq}^2 may be restricted to positive real numbers.

Finally, note that, like the Ornstein-Uhlenbeck model of phenotypic evolution, the covariance matrix predicted by our neutral model has a Generalized 3-Point Structure as identified by [13], making it amenable linear time model fitting procedures.

4.2 Comparison with Brownian motion and the Ornstein-Uhlenbeck process

To compare our model to the standard Gaussian models of phenotypic evolution, we first write out their likelihood.

Brownian motion Under the Brownian motion model, the mean and covariance of the multivariate Gaussian distribution are given by [7]:

$$\mu_i = Z_0$$
$$\Sigma_{ij} = \sigma^2 t_{ij}$$

The Brownian motion model has only two parameters, Z_0 and σ^2, and is thus the simplest Gaussian model for phenotypic evolution. However, it predicts that the phenotypic variance increases linearly with time and does not saturate, violating intuitions and observations. Note that the Brownian motion model also predicts a linear, non-saturating pairwise divergence, as defined in Section 3.3 and in the framework of Nourmohammad et al. [11]. The pairwise divergence according to the Brownian motion model is given by:

$$D\Big(Z(t+s), Z(t+r)\Big) = \sigma^2(s+r)$$

This disagrees with our model predictions as well as with the neutral null model of Nourmohammad et al. [11, 8, 10]. It is well known, however, that the Brownian motion model is a good approximation for the dynamics at small time t.

Ornstein-Uhlenbeck (OU) model The OU model of phenotypic evolution, which is used to model the action of stabilizing selection, defines a Gaussian process likelihood for the data where the mean and covariance are given by [7]:

$$\mu_i = Z_0 e^{-ut_i} + (1 - e^{-ut_i}) Z_{eq}$$
$$\Sigma_{ij} = e^{-\alpha(t_i + t_j - 2t_{ij})}(1 - e^{-2\alpha t_{ij}})\frac{\sigma^2}{2\alpha}$$

Similar to our model with fixed scale u, the OU model has a total of 4 parameters $\boldsymbol{\theta} = (Z_0, Z_{eq}, \sigma^2, \alpha)$. However, in standard implementations, as in the `R` packages `ouch` [1] and `geiger` [12], practitioners assume that the ancestral phenotype is the same as the equilibrium phenotype ($Z_0 = Z_{eq}$), thus reducing the number of parameters to 3. As noted elsewhere in the Appendix and the main text, our model shares the exact form of the mean with the OU model, though the two models diverge in the form of the covariance.

At equilibrium, the OU model predicts a constrained pairwise divergence that is modulated by the parameter α, as we would intuitively expect and consistent with the action of the parameter c in the models of Nourmohammad et al. [11, 8, 10]:

$$D_{t\to\infty}\Big(Z(t+s), Z(t+r)\Big) = \frac{\sigma^2}{\alpha}(1 - e^{-\alpha(r+s)})$$

However, note that in standard applications of the OU model, the parameter α is often difficult to estimate accurately and there are debates about its meaning and interpretation in this context [2].

5 Examples and applications

In this section, we provide three examples that illustrate the utility and insights that can be derived from this modeling approach. In the first example, we derive explicit expressions for the time-dependent dynamics of the phylogenetic mean and variance of GC Content under the Felsenstein (1981) [5] model of DNA evolution. In the second example, we examine these same dynamics under the Jukes-Cantor (1969) [9] model of DNA evolution. Finally, in the third example we show that this approach can tackle more complex genetic architectures by examining dynamics of the number of ATG codons in a DNA sequence.

5.1 Dynamics of GC content under the F-81 model

Under the modeling framework that we outlined, our genotype matrix $\boldsymbol{X}$ and the corresponding matrix of effect sizes $\boldsymbol{\beta}$ are of dimension $L \times 4$ for a sequence of length L. For a trait like GC Content, each row of the matrix of effect sizes has the following form:

$$\boldsymbol{\beta}_j = \frac{1}{L}\begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix}$$

Under the F-81 substitution model, the four alleles have a stationary distribution specified by: $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$. These stationary probabilities can be inferred from genome-wide data or from a set of sequences that are of interest to the analyst. Additionally, the scaling parameter u, which modulates the convergence rate to the equilibrium state, is given by [14]:

$$u = \frac{1}{1 - \pi_A^2 - \pi_C^2 - \pi_G^2 - \pi_T^2}$$

5.1.1 Dynamics of the mean and variance

Dynamics of the Phylogenetic Mean The dynamics of the mean for GC Content can be obtained straightforwardly from Equation 14 by plugging in the effect sizes:

$$E[Z(t)] = e^{-ut}\frac{1}{L}\sum_{j=1}^{L}\Big(P(X_j(0) = C) + P(X_j(0) = G)\Big) + (1 - e^{-ut})(\pi_C + \pi_G)$$

If we assume that there is a single founding population, then the quantity $Z_0 := \frac{1}{L}\sum_{j=1}^{L}\Big(P(X_j(0) = C) + P(X_j(0) = G)\Big)$ becomes simply the fraction of G/Cs in the ancestral sequence. In addition, as can be intuitively gleaned, the expected mean phenotype at equilibrium under this model is shown to be $Z_{eq} := \pi_C + \pi_G$. This example shows that the phylogenetic dynamics of the mean for some molecular traits can be explored analytically.

Dynamics of the Phylogenetic Variance Similarly, we can obtain the dynamics of the variance for GC Content by plugging in the effect sizes into Equation 16 and simplifying:

$$
\begin{aligned}
Var\Big(Z(t)\Big) = \frac{1}{L^2}\sum_{j=1}^{L} \pi_C + \pi_G - (\pi_C + \pi_G)^2 \\
+ e^{-ut}\Big(2(\pi_c + \pi_G) - 1\Big)\Big(\pi_C + \pi_G - P(X_j(0) = C) - P(X_j(0) = G)\Big) \\
- e^{-2ut}\Big(\pi_C + \pi_G - P(X_j(0) = C) - P(X_j(0) = G)\Big)^2
\end{aligned}
$$

An interesting feature of the equation above is that it resembles the variance expression for the bi-allelic model, as if in effect we only have 2 alleles corresponding to the A/T and G/C states. This suggests that, in similar scenarios and for traits that have similar distributions of effect sizes, it makes sense to "lump" the alleles that share the same effect size into a single allele, thus reducing the number of states that we need to keep track of. Indeed, this supposition is shown to be true for any arbitrary grouping of alleles by their effect sizes in Section 3.1.

If we use the definitions of Z_0 and Z_{eq} from above, we can further simplify the expression for the variance as follows:

$$
\begin{aligned}
Var\Big(Z(t)\Big) = \frac{1}{L}\Bigg[Z_{eq}(1 - Z_{eq}) \\
+ e^{-ut}\Big(2Z_{eq} - 1\Big)\Big(Z_{eq} - Z_0\Big) \\
- e^{-2ut}\Big(Z_{eq}^2 + Z_0(1 - 2Z_{eq})\Big)\Bigg]
\end{aligned}
$$

In this case, the phylogenetic variance is independent of the genotype. If we take the limit as $t \to \infty$, we obtain the following expression for the equilibrium variance:

$$
\sigma_{eq}^2 = \frac{1}{L} Z_{eq}(1 - Z_{eq}) = \frac{1}{L}(\pi_C + \pi_G)(1 - \pi_C - \pi_G)
$$

It is important to highlight here that the variance is inversely proportional to the number of loci L, as would be expected for traits that are expressed as a fraction of the length of the sequence.

5.1.2 Linear approximation for the variance of GC Content under F-81

In the main text, we analyze the variance of GC Content at small phylogenetic timescales, on the order of the divergence of modern primates. Here, we derive the model predictions for the variance dynamics of GC Content in the regime when $ut \ll 1$. In Section 3.4, we have shown that the variance at small time only depends on the parameter Ψ (see Equation 20):

$$
Var_{t \ll 1}\Big(Z(t)\Big) \approx ut\Psi
$$

For GC Content, by noting that there are only 4 unique combinations of alleles that do not share effect sizes $(A/C, A/G, T/C, T/G)$, it is straightforward to write Ψ in terms of other quantities, mainly Z_0 and Z_{eq}:

$$\Psi = \frac{1}{L^2}\sum_{j=1}^{L}(\pi_A + \pi_T)\Big(P(X_j(0) = C) + P(X_j(0) = G)\Big) +$$
$$(\pi_C + \pi_G)\Big(P(X_j(0) = A) + P(X_j(0) = T)\Big)$$
$$= \frac{1}{L}\Big[(1 - Z_{eq})Z_0 + Z_{eq}(1 - Z_0)\Big]$$

Which gives the following linear approximation for the variance of GC Content at small time:

$$Var_{t \ll 1}\Big(Z(t)\Big) \approx \frac{ut}{L}\Big[(1 - Z_{eq})Z_0 + Z_{eq}(1 - Z_0)\Big]$$

5.2 Dynamics of GC content under the JC-69 model

A special case of the example that we outlined in the previous section is obtained when we examine sequences evolving according to Jukes-Cantor model, which is specified by the following stationary distribution: $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. From this, we can see that the scaling parameter becomes $u = \frac{4}{3}$.

5.2.1 Dynamics of the mean and variance

Dynamics of the Phylogenetic Mean Under this model, the phylogenetic mean has a simple form:

$$E[Z(t)] = e^{-\frac{4}{3}t}\frac{1}{L}\sum_{j=1}^{L}\Big(P(X_j(0) = C) + P(X_j(0) = G)\Big) + (1 - e^{-\frac{4}{3}t})\frac{1}{2}$$
$$= e^{-\frac{4}{3}t}Z_0 + (1 - e^{-\frac{4}{3}t})\frac{1}{2}$$

Where $Z_{eq} = \pi_C + \pi_G = \frac{1}{2}$ and Z_0 again corresponds to the fraction of G/Cs in the ancestral sequence.

Dynamics of the Phylogenetic Variance To obtain an expression for the phylogenetic variance under the Jukes-Cantor model, we simply plug in $Z_{eq} = \frac{1}{2}$ in the expression we derived for the F-81 model:

$$Var\Big(Z(t)\Big) = \frac{1}{4L}(1 - e^{-\frac{8}{3}t})$$

Because of the symmetry of the JC69 model, the expression above has the form of the variance as that of the Ornstein-Uhlenbeck model (i.e. a single exponential decay towards the equilibrium variance at a rate that is twice as fast as the rate of the mean reverting to its equilibrium value). The equilibrium variance under this model is given by:

$$\sigma_{eq}^2 = \frac{1}{4L}$$

5.2.2 Linear approximation for the variance of GC Content under JC-69

Using the expression we derived in section 5.1.2 and noting the $u = \frac{4}{3}$ and $Z_{eq} = \frac{1}{2}$, we can write the linear approximation for the variance dynamics at small time under the Jukes-Cantor model as:

$$Var_{t \ll 1}\Big(Z(t)\Big) \approx \frac{2}{3L} t$$

Similar to what we observed for the full-scale expression for the variance dynamics, the short timescale variance behavior is also independent of the initial phenotype Z_0.

5.3 Dynamics of ATGs under the F-81 model

To showcase the generality of the framework we outlined above, here we analyze a more complex case involving codon counts in a DNA sequence. The trait of interest here is the number of ATG codons in a DNA sequence of length L. In this context, there are 64 alleles corresponding to the 64 distinct codons. Therefore, the scaling factor u is defined as:

$$u := \frac{1}{1 - \sum_k^{64} \pi_k^2}$$

Where π_k is the equilibrium probability of codon k, which can be derived from the equilibrium probabilities of the individual DNA letters. Since we are defining a counting-based trait, the effect sizes for all codons are 0 except for the ATG codon, which has an effect size of 1.

5.3.1 Dynamics of the mean and variance

Dynamics of the Phylogenetic Mean The dynamics of the mean for the number of ATGs can be obtained straightforwardly from Equation 14 by plugging in the effect sizes:

$$\begin{aligned} E[Z(t)] &= e^{-ut} \sum_{j=1}^{L} P(X_j(0) = ATG) + (1 - e^{-ut}) \sum_{j=1}^{L} \pi_{ATG} \\ &= e^{-ut} \sum_{j=1}^{L} P(X_j(0) = ATG) + (1 - e^{-ut}) L \pi_{ATG} \end{aligned}$$

Assuming a single founder population, we can see that $Z_0 := \sum_{j=1}^{L} P(X_j(0) = ATG)$ is the number of ATG codons in the ancestral sequence and $Z_{eq} := L\pi_{ATG}$ is the expected number of ATGs at equilibrium.

Dynamics of the Phylogenetic Variance Similarly, we can obtain the dynamics of the variance for GC Content by plugging in the effect sizes into Equation 16 and simplifying (it may be useful to refer to the results in Section 3.1 here):

$$Var\Big(Z(t)\Big) = (1-e^{-ut})\Bigg[e^{-ut}\Big(\sum_{j=1}^{L}\pi_{ATG} + P(X_j(0)=ATG) - 2P(X_j(0)=ATG)\pi_{ATG}\Big) + (1-e^{-ut})L\pi_{ATG}(1-\pi_{ATG})\Bigg]$$

This expression can be further simplified by using the definitions for the quantities $Z_{eq} := L\pi_{ATG}$ and $Z_0 := \sum_{j=1}^{L} P(X_j(0)=ATG)$, which results in the following form for the variance of ATGs:

$$Var\Big(Z(t)\Big) = (1-e^{-ut})\Bigg[e^{-ut}\Big(Z_{eq} + Z_0(1-2\pi_{ATG})\Big) + (1-e^{-ut})Z_{eq}(1-\pi_{ATG})\Bigg]$$

5.3.2 Linear approximation for the variance of ATGs under F-81

In Section 3.4, we have shown that the variance at small time only depends on the parameter Ψ (see Equation 20):

$$Var_{t\ll 1}\Big(Z(t)\Big) \approx ut\Psi$$

For this trait, it is straightforward to derive an analytical expression for Ψ and write it in terms of phenotypic quantities Z_{eq} and Z_0:

$$\begin{aligned}\Psi &= \sum_{j=1}^{L}\pi_{ATG} + P(X_j(0)=ATG) - 2P(X_j(0)=ATG)\pi_{ATG} \\ &= Z_{eq} + Z_0(1-2\pi_{ATG})\end{aligned}$$

Note that assuming that the length of the sequence is known, π_{ATG} can also be written in terms of phenotypic quantities, e.g. $\pi_{ATG} = \frac{Z_{eq}}{L}$. Given this expression for Ψ, we can write the linear approximation for the variance at short timescales as:

$$Var_{t\ll 1}\Big(Z(t)\Big) \approx ut\Big(Z_{eq} + Z_0(1-2\pi_{ATG})\Big)$$

References

[1] Marguerite A. Butler and Aaron A. King. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *American Naturalist*, 164:683–695, 2004.

[2] Natalie Cooper, Gavin H. Thomas, Chris Venditti, Andrew Meade, and Rob P. Freckleton. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biological Journal of the Linnean Society*, 118(1), 2016.

[3] Will Cornwell and Shinichi Nakagawa. Phylogenetic comparative methods, 2017.

[4] Warren J. Ewens. *Mathematical Population Genetics: I. Theoretical Introduction.* 2004.

[5] Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6), 1981.

[6] Joseph Felsenstein. *Theoretical evolutionary genetics.* 2019.

[7] Luke J. Harmon. *Phylogenetic Comparative Methods: Learning from Trees.* 2018.

[8] Torsten Held, Armita Nourmohammad, and Michael Lässig. Adaptive evolution of molecular phenotypes. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(9), 2014.

[9] T H Jukes and C R Cantor. Evolution of protein molecules BT - Mammalian protein metabolism. In *Mammalian protein metabolism*, volume III. 1969.

[10] Armita Nourmohammad, Joachim Rambeau, Torsten Held, Viera Kovacova, Johannes Berg, and Michael Lässig. Adaptive Evolution of Gene Expression in Drosophila. *Cell Reports*, 20(6), 2017.

[11] Armita Nourmohammad, Stephan Schiffels, and Michael Lässig. Evolution of molecular phenotypes under stabilizing selection. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(1), 2013.

[12] Matthew W. Pennell, Jonathan M. Eastman, Graham J. Slater, Joseph W. Brown, Josef C. Uyeda, Richard G. Fitzjohn, Michael E. Alfaro, and Luke J. Harmon. Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, 30(15), 2014.

[13] Lam Si Tung Ho and Cécile Ané. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Systematic Biology*, 63(3), 2014.

[14] Ziheng Yang. *Computational Molecular Evolution.* 2010.