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 a simple theory based on 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 the widely-used OU model for stabilizing selection. We confirm the predictions empirically using additive molecular phenotypes calculated from ancestral reconstructions of putatively unconstrained DNA sequences in primate genomes. We show that the OU model is favoured by inference software even when applied to GC content of unconstrained sequences or simulations of DNA evolution. 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 explicitly formulated based on empirical estimates from neutral sites [19] or, more recently, biophysics of protein folding (e.g.,[21]).

Brownian Motion (also referred to as a "random walk"[22]) 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, phenotypes are not expected to change over time[24]. On the other hand, it has long been appreciated that in even in the absence of selection, mutation is expected to be (at least) 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.

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 [28]–[30], and that biased mutation can drive the phenotype away from the optimum when selection is not strong[31]. 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 prediction[32]. Perhaps most intriguingly, a recent empirical study that estimated the effects of mutations on gene expression levels for 10 genes found asymmetric distributions (inconstant with Brownian Motion) and used simulations to predict approximately linear change over time in the absence of selection for most genes [33].

Here we aimed to develop a neutral model of phenotype evolution that could be compared directly to DNA-based molecular phenotypes. We work out the dynamics of additive phenotypes in the weak-mutation regime[34], the regime of most relevance for molecular evolutionary studies. Under Felsenstein's 1981 model of DNA evolution[35], 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]. We find remarkable agreement with this prediction in observations of molecular phenotypes computed from reconstructed ancestors of putatively neutral sequences in primate genomes, even when the phenotypes are not strictly additive. Our results are inconsistent with the Brownian Motion null hypothesis used in phylogenetic comparative analysis, and consistent with this, an inference software package rejects this null hypothesis for our neutral phenotype data.

Results

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

Our goal is to compare a neutral model of phenotype evolution 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 under a standard model of DNA substitution. Let Z(X) be the phenotype as a function of the genotype X, and a_{nm} be the contribution of the m-th allele at the n-th locus. For example, for a DNA-based phenotype with n=10 contributing loci, a represents a 4 x 10 matrix. In order to compare with empirical phenotype data, we allow arbitrary values for a_{nm}, thus making no assumptions about the distribution of effect sizes. We assume the haploid loci are in linkage disequilibrium, and use a so-called "one-hot" encoding to represent the genotype, so that $X_{nm} = 1$ if the haploid genotype is the m-th allele at the n-the locus and zero otherwise. Thus, for n=10 loci, the genotype, X, is also a 4 x 10 matrix. If the phenotype is additive, we have:

$$Z(X) = \sum_n \sum_m a_{nm} X_{nm} = \sum_n a_n \cdot X_n$$

where $a_n \cdot X_n$ is a dot product between vectors of length 4 for DNA. Given a starting phenotype Z_0, because evolution is a random process, we aim to calculate the mean phenotype over "replicate" populations as a function of time, *t*.

$$E[Z(t)|Z_0] = \sum_{X\ s.t.\ Z(X)=Z_0} P(X) \sum_{Y} P(Y|X,t)Z(Y)$$

Where P(*x*) is the probability density of the random variable *x* and E[*x*] denotes the expectation (or mean), $E[x] \stackrel{\text{def}}{=} \sum_X P(x)x$, $P(Y|X,t)$ is the probability of the initial genotype *X* mutating into another genotype *Y* after evolutionary time *t*, and $P(X)$ is the probability of observing the initial genotype. In assuming that an initial genotype mutates into another genotype we are neglecting the possibility of polymorphism, competition and recombination between genotypes and other population genetic processes. This is the so-called "weak mutation" assumption commonly used to model molecular evolution at the timescale of inter species divergence[34].

To make progress, we first consider the dynamics of the mean phenotype given an initial genotype *X*:

$$E[Z(t)|X] = \sum_{Y} P(Y|X,t)Z(Y) = \sum_{n} \sum_{Y_n} P(Y_n|X_n,t)a_n \cdot Y_n$$

Where the last equality is achieved using the linearity of the expectation and the assumption of the independence of genotype evolution at each locus. Under Felsenstein's 1981 model (F81) of DNA evolution [35] the substitution probabilities are,

$$P(Y_n|X_n,t) = \pi_{Y_n}(1-e^{-ut})\ for\ Y_n \neq X_n$$

$$P(Y_n|X_n,t) = e^{-ut} + \pi_{Y_n}(1-e^{-ut})\ for\ Y_n = X_n$$

where $u \stackrel{\text{def}}{=} \frac{1}{1-\sum_m {\pi_m}^2}$ is the DNA substitution rate, scaled to that evolutionary distance, *ut*, is measured in substitutions per site and π is the stationary distribution that parameterizes the continuous time markov process and can be interpreted as the long-term probabilities of each allele[35]. For notational convenience we define $\pi_{Y_n} \stackrel{\text{def}}{=} \pi \cdot Y_n$ to represent the (scalar) long-term probability of the allele that is found at the n-th locus. Note that this model assumes that the loci evolve independently and identically, so there is no dependence on the locus for π_{Y_n}. Using straightforward algebra (see Appendix), we can show that

$$E[Z(t)|X] = e^{-ut}Z_0 + (1-e^{-ut})Z_{eq} = e^{-ut}(Z_0 - Z_{eq}) + Z_{eq}$$

Where $Z_{eq} \stackrel{\text{def}}{=} \sum_n \pi \cdot a_n$ and can be interpreted as the phenotype expected after a long-time of neutral evolution, or the phenotype at "mutational equilibrium."

Remarkably, our formula for the dynamics of the mean of an additive phenotype is *independent* of the starting genotype, X, and depends only on the starting phenotype Z_0. In other words, the dynamics of the phenotype are *the same for all genotypes* that encode that phenotype. We conjecture that the F81 model is the most realistic DNA substitution model for which this can be

true (for example, it is not true of models with transition-transversion rate bias[36], see Discussion). This removes the need to average over all $X\ such\ that\ \ Z(X) = Z_0$, leading to our main theoretical result, which we refer to as the F81 model for mean phenotypes:

$$E[Z(t)|Z_0] = E[Z(t)|X] = e^{-ut}(Z_0 - Z_{eq}) + Z_{eq}$$

Thus, under the F81 DNA substitution model, the dynamics of the mean phenotype depend only on the initial phenotype, the mutational equilibrium phenotype and the evolutionary rate. The dynamics of the mean phenotype are *independent* of the number of alleles, loci, and the distribution of allelic effect sizes, a. The dynamics of the mean phenotype are similar to those expected under the (not DNA-based) house-of-cards mutation model[22], if the mean of the phenotype effect distribution in the house-of-cards model is chosen to be the "mutational equilibrium" phenotype, Z_{eq}, defined above.

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

$$E[Z(ut \ll 1)|Z_0] \approx ut\big(Z_{eq} - Z_0\big) + Z_0$$

This says that the mean phenotype is driven towards the equilibrium with a force simply proportional to the distance of the initial phenotype from the mean. Near the 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].

Another remarkable feature of our result is that the dynamics of the mean phenotype match exactly to the mean of the well-studied 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). The OU process is a Gaussian process, and therefore the dynamics are fully specified by the mean and variance. The parameters of the OU process are the restoring force (which must be u based on the form of the mean dynamics we predict above) and the fluctuation size (σ_Z^2), which doesn't affect the mean dynamics. If the F81 neutral phenotype dynamics matched an OU process exactly, the variance would be independent of the starting phenotype, and at short evolutionary distance would increase proportional to time

$$V[Z(t)|Z_0] = V[Z(t)] = \frac{\sigma_Z^2}{2u}(1 - e^{-2ut}) \approx \frac{\sigma_Z^2}{2u}2ut = \sigma_Z^2 t$$

Where $V[x]$ is the variance of the random variable x. To figure out the fluctuation size, σ_Z^2, we consider that the long time variance should go to $\frac{\sigma_Z^2}{2u} = V[Z(t \to \infty)]$. So under an OU model with the only force being that of mutation, $\sigma_Z^2 = 2uV[Z(t \to \infty)]$ and

$$V[Z(ut \ll 1)] \approx 2utV[Z(t \to \infty)]$$

This means that, if the phenotypic dynamics in the absence of selection follow an OU process, the variance of the phenotype will increase proportionally to evolutionary distance, at the rate of twice the long-time or equilibrium variance. Since this prediction is the same as the prediction of the Brownian-Motion model[37], the variance dynamics at short times cannot be used to distinguish the Brownian-Motion and OU models.

Molecular phenotypes from ERVs in primate genomes

Since our goal is to test empirically the predictions of the model, 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[38], [39], 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 number over the human genome[40]. 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., [41]. 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 $u \stackrel{\text{def}}{=} \frac{1}{1-\sum_m {\pi_m}^2} = 1.37$. 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 $a_n =$

$(a_A, a_C, a_G, a_T) = \left(0, \frac{1}{L}, \frac{1}{L}, 0\right)$ for all n. From these we compute the other key parameters for GC content

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

$$V[Z(t \to \infty)] = \frac{1}{L}(\pi_C + \pi_G)(1 - \pi_C - \pi_G) = 0.002304$$

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 mis-specification 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 an OU process and the observed increase in variance for GC content. To rule out possible circularity due to the use of probabilistic models of evolution in ancestral genome reconstruction [42] 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.

We next repeated these analyses for the strength of matches to the TATA box ([43], see Methods) in the first 10 residues of each ERV segment. Because positions in transcription factor binding sites contribute approximately linearly to affinity[44], 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[45]. 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 $V[Z(t = \infty)] = 24.6$, which are obtained directly from the matrix model (see Methods). We found good agreement with the predictions of the model (Figure 2a-c). Once again the linear, constant increase in variance seems to agree reasonably well with the prediction that neutral phenotypes evolve according to an OU process.

The BM null hypothesis is rejected for GC content evolution in the absence of selection

To test whether the observed deviations from the Brownian Motion null hypothesis were strong enough to be detected in a typical analysis, we applied the widely used OUCH package for comparative phylogenetic analysis[39]. We obtained six well-aligned fragments from an alignment of 39 mammals (see Methods) for an unusually ancient ERV that is thought to evolve in the absence of selection[46]. We found that for the five of six segments where the inference algorithm converged, the OU model was strongly supported over the Brownian Motion null

hypothesis (all differences in corrected AIC >50, Table 1), which in the standard interpretation in comparative phylogenetic analysis would be considered strong evidence for stabilizing selection (Figure 3a). We note that these results cannot be explained by noise [38]in our phenotypic measurements, as GC content is computed exactly from sequences.

Although this ERV is thought to evolve in the absence of selection[46], we wanted to ensure that the results were not due to cryptic selection on GC content. We therefore simulated molecular evolution of these ERV sequences under a simple neutral model of DNA substitutions (see Methods) and found the same results using OUCH for inference: in the five cases where the algorithm converged, the data strongly supported the rejection of the BM model in favour of the OU model usually associated with stabilizing selection (all differences in corrected AUC >50, Figure 3b). Qualitatively, the strength of support (as measured by AUC in bootstrap resamplings) was similar to (albeit stronger than) that obtained with the real data, consistent with the idea that ERV sequences are likely evolving in the absence of selection on GC content. Taken together, these results are consistent with the idea that the force of mutation on quantitative molecular traits is sufficient to reject the Brownian Motion null hypothesis in a phylogenetic comparative analysis (see Discussion)

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

Although our predictions of the mean phenotype dynamics make no assumptions about the underlying distribution of the phenotype, this assumption is made in the Brownian Motion and OU process predictions; both GC content and strength of TATA boxes are expected to be approximately Gaussian (GC content is an estimate of a binomial distribution parameter and the strength of a TATA box is the sum of 10 different numbers). Therefore, we next considered a sequence-based trait that we expected to be far from Gaussian: the number of ATG start codons in 100 bp segments. Since the number of ATGs cannot be negative, sequences that start with 0 ATGs will certainly violate the predictions of Brownian Motion and OU process models. 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 4a, 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 4b 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 $a_m = 1$ for the short sequence of interest, and 0 ortherwise. 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 $wu = \frac{w}{1-\sum_m {\pi_m}^2}$. 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

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

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 4c). We note that the agreement between the observations and theory must be highly 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

$$V[Z(t \to \infty)] \approx (L-2)\pi_A\pi_T\pi_G(1-\pi_A\pi_T\pi_G) = 1.773$$

Unlike for GC content and strength of TATA boxes, we found 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 5a). For number of ATGs we found that sequences with more ATGs than the equilibrium (4 ancestral ATGs, triangles in figure 5a) showed a faster increase in variance than sequences with fewer (0 ancestral ATGs, filled squares in figure 5a). We therefore looked at the ancestral bin that was closest to the mutational equilibrium (2 ancestral ATGs, x's in figure 5a) and found that the variance increased in very good agreement with the OU prediction. Taken together, these results for the variance suggest that the true dynamics for molecular phenotypes do not match an OU process when the phenotype is strongly non-Gaussian and/or sufficiently far from the mutational equilibrium.

Variance dynamics of neutral phenotypes are more complicated than OU or BM model predicts

We next explored the dynamics of the variance. Under the assumptions above (weak-mutation regime, full linkage disequilibrium, haploid), the dynamics of the variance of an additive phenotype given an initial genotype, X, is given by

$$V[Z(t)|X] = \sum_n C(X_n) \cdot a_n {a_n}^T$$

Where $a_n$is the vector of effects of alleles at position n, and $C(X_n)$ is the variance-covariance matrix for the current genotype at the n-th locus with entries given by $C_{ij} = E[(Y_i - E[Y_i])(Y_j - E[Y_j])]$, and the multiplication, indicated by $\cdot$ is an inner product between the two matrices. Because the genotype at each locus is a categorical random variable, we know the form of $C(X_n)$, which works out to:

$$C_{mm} = E[Y_m|X_n](1 - E[Y_m|X_n]) = (1 - e^{-ut})(\pi_m(1 - \pi_m) + e^{-ut}(X_{nm} - \pi_m)^2)$$

$$C_{km} = -E[Y_k|X_n]E[Y_m|X_n] = -(1 - e^{-ut})(\pi_k\pi_m + e^{-ut}(\pi_m X_{nk} + X_{nm}\pi_k - \pi_k\pi_m))$$

Where k and m index two different alleles. In general, under the F81 model, the variance appears to depend on the initial genotype, X, and is non-monotonic.

For the simplest phenotype, we can show that 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. Consider the additive phenotype where only one allele contributes at each locus, and the effect size is the same at each locus, say 1 unit. The phenotype is proportional to the counts of the m-th allele over all the loci, and corresponds to a phenotype considered in recent theoretical work [28]. In our notation, $a_{nm} = 1$ for the m-th allele at each position, and 0 otherwise. For each of L identical loci, if the m-th allele is listed first,

$$a_n {a_n}^T = \begin{matrix} 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \\ \vdots & & & \end{matrix}$$

The variance is simply $V[Z(t)|X] = \sum_n C_{mm}$. Since for this simple phenotype $Z_{eq} = L\pi_m$[28], substituting the formula above gives

$$V[Z(t)|X] = (1 - e^{-ut})Z_{eq}\left(1 - \frac{Z_{eq}}{L}\right) + (1 - e^{-ut})e^{-ut}\sum_n (X_{nm} - \pi_m)^2$$

The first term is a single exponential decay to the long-term variance, $Z_{eq}\left(1 - \frac{Z_{eq}}{L}\right)$. The second term is proportional to the squared difference of the starting genotype from the probability of the m-th allele at equilibrium. However, it can be simplified into a genotype independent form in this case: we get exactly $Z_0 = \sum_n X_{nm}$terms of $(1 - \pi_m)^2$ and exactly $L - Z_0$ of ${\pi_m}^2$, so

$$V[Z(t)|Z_0] = (1 - e^{-ut})Z_{eq}\left(1 - \frac{Z_{eq}}{L}\right) + (1 - e^{-ut})e^{-ut}(Z_0(1 - 2\frac{Z_{eq}}{L}) + \frac{{Z_{eq}}^2}{L})$$

Thus, the variance of the simplest additive phenotype is independent of the genotype, but shows non-monotonic, initial phenotype dependent dynamics. This also appears to be consistent with analysis of a simple phenotype under the house-of-cards mutation model, which also predicts non-monotonic dynamics that depend on the starting phenotype[22]. Only in the case where $\frac{Z_{eq}}{L} = \frac{1}{2}$ is the variance dynamics independent of the initial phenotype (as predicted by the OU dynamics), which corresponds to a case where exactly ½ of the sites have the m-th allele at equilibrium. However, even in this situation the dynamics deviate from the OU prediction.

To reconcile these theoretical results with our observations of additive phenotypes showing nearly OU variance dynamics, we consider the case where Z_0 approaches Z_{eq}. After algebra, we have $V\left[Z(t)|Z_0 \to Z_{eq}\right] = (1 - e^{-ut})(1 + e^{-ut})Z_{eq}\left(1 - \frac{Z_{eq}}{L}\right) = (1 - e^{-2ut})Z_{eq}\left(1 - \frac{Z_{eq}}{L}\right)$,

which is exactly the OU dynamics. Hence, at least in this case, the variance shows OU dynamics when the initial phenotype is at equilibrium. We conjecture that more generally, the variance dynamics approach OU dynamics when the phenotype approaches equilibrium. However, for only slightly more complicated phenotypes (such as GC content as defined above) the dynamics are not exactly OU, even if the phenotype is at the mutational equilibrium, owing to covariance between the alleles at each locus (see Appendix). We note that although the prediction matches the data for GC content and TATA boxes in ERVs, the case where the initial phenotype, Z_0, approaches the mutational equilibrium, Z_{eq}, is not a realistic case for a phenotype that has been under selection, where the phenotype is likely very far from that expected under mutation alone (see Discussion).

Because our approximate model for the number of ATGs described above corresponds to this simple counting phenotype, we compared these predictions to that data. At short evolutionary distance, and since for ATGs in 100 residues, $Z_{eq} \ll L$, we obtain a simple approximation for the variance of number of ATGs under the F81 model:

$$V[Z(t \ll 1)|Z_0] \approx ut(Z_{eq} + Z_0)$$

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 5b,c).

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[47] 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[48], [49], and the latter two of these phenotypes are related to the emergence of new protein-encoding genes from random DNA [50], [51]. 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 ([52], 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 6a) 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 6b) 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.

We also repeated the OUCH analysis on the longest ORF in ancient mammalian ERV segments to test if the mutational force would be strong enough to reject the BM null hypothesis. Indeed, we find significant support for the OU process model for the longest ORF (all 4 of 6 segments where the inference converged show corrected AIC differences > 50, Table 1). We note that under the current interpretation of this as evidence for stabilizing selection, this would lead to the (interesting) inference that ancient ERVs are under selection to retain coding capacity. However, we can find the similar evidence for the OU process in simulated versions of these segments (see methods) where we know there is no selection for coding capacity (all 3 of 6 segments where the inference converged show corrected AIC differences > 20, Table 1), consistent with the idea that mutation alone (or model mis-specification due to mutation, see Discussion) appears sufficient to create a detectable evolutionary restoring force on molecular phenotypes of current research interest.

Discussion

Our results show that mutation is expected to aid the creation of molecular phenotypes de novo and relate to several areas of molecular evolution. For example, in agreement with simulation results [53], 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[51], the force of mutation is expected on average to increase their tendency for folding[50]. Remarkably, but consistent with a recent report [33], 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[29], [31], 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[28], [31]. 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, which is a specific instance of the F81 model where all $\pi_m=1/4$ ($u=4/3$) 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.

Interestingly, for GC content we found agreement with the predictions of the variance based on the OU process, even though we can show that this is only approximate (See Appendix). We believe this is because GC content in ERVs is close enough to mutational equilibrium, where we find the dynamics to be well approximated by the OU process. However, more generally, our F81 phenotype evolution model is meant to be used as a null hypothesis for phenotypes that may be under selection (such as gene expression levels), and these phenotypes are unlikely to be close to mutational equilibrium. This is consistent with observations of directional evolution of phenotypes in mutation accumulation studies[26], where selection is removed, but phenotypes are far from their mutational equilibriums.

We note that in practice even neutral evolution of GC content may still be problematic if analyzed with the standard OU process models[41] because the rate of increase of variance depends on the number of loci (See Appendix). Thus, although for a fixed sequence size GC content appears to evolve according to an OU process, for real sequences (that change in length over the phylogeny) although the variance in GC content is expected to increase approximately linearly for reasonably short evolutionary times, it will increase at a slower rate for longer sequences. This reduction in variance could be mis-interpreted in the OU framework as increased selection intensity.

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]). Indeed, a widely used software implementation [13], [39] infers strong evidence for selection on GC content and length of longest ORF in our ERV sequences and in DNA-based simulations with no selection. Also, our results complicate whether statistical evidence for multi-optimum OU models represents evidence for adaptive shifts, which is a widely adopted assumption in comparative phylogenetic analysis[10]. Shifts in optima could be expected even in the absence of selection, due to changes in the mutational equilibrium over phylogenies (for example, simply because of changes in GC content.) 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[54] to the previously reported challenges with phylogenetic inference of OU models[38], [39], [55]. Developing tests based only on the mean (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[55].

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 phentoypes[11], [28], 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 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[56], [57], 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 [58], [59]. 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[40]) or simulations at the sequence level[60]. The simulation approach has been applied successfully to molecular traits that can be computed directly from intrinsically disordered protein sequences[61], [62]. 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[63], 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[42] from the 100 vertebrate alignment tree from Ensembl, we extracted the species in the aligned segments using the Ensembl compara REST API[64]. As a control for possible circularity due to probabilistic models of sequence evolution used in ancestral genome reconstruction [42], we also performed analysis on ancestral sequences reconstructed maximum parsimony as implemented in the phangorn package[65].

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 [43].

a =

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

We assumed a position independent background model, as above, $\pi = (0.32,0.18,0.18,0.32)$. 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[66]. To compute the other parameters for the strength of TATA boxes, which have width w=10, we used

$$Z_{eq} = \sum_{n=1}^{w} \sum_{m} \pi_m \log \frac{p(X_{nm} = 1|motif)}{\pi_m} \equiv \sum_{n=1}^{w} Z_{neq}$$

$$V[Z(t \to \infty)] = \sum_{n=1}^{w} \pi_m \left(\sum_{m} \log \frac{p(X_{nm} = 1|motif)}{\pi_m} - Z_{neq} \right)^2$$

Where $p(X_{nm} = 1|motif)$ is the probability of observing the mth allele at the nth position in the transcription factor binding probability matrix. 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[66], $Z(X) = \sum_{n=1}^{L-w+1} \frac{1}{1+e^{-S_n - R}}$, where each S_n represents the standard likelihood ratio score [66] for the subsequence of length w starting at position n, , 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 and simulation of ancient mammalian ERVs

we retrieved the human ERV sequence [46] and used BLAST on the Ensembl database[67] 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) [68]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 [60] to evolve the reconstructed sequences along the mammalian phylogeny. For each set of real or simulated sequences we computed the quantitative molecular phenotypes and analyzed them with OUCH[39]. Tree was drawn using plot function from the APE package for R [69]

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 [52]. 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 –the BM null hypothesis is rejected for GC content in ERVs

a) shows the inferred phylogenetic tree and associated GC content measurements (GC%) for the segment 320 – 722 for the species used in the phylogenetic comparative analysis of GC content. b) shows the difference in corrected AIC between Brownian Motion (BM) and OU models for

six segments of an alignment of an ancient ERV. The average over 200 bootstrap samples for each segment where the algorithm converged, and the standard deviation is indicated by error bars. Filled bars represent observed ERV segments (Real), and unfilled represent results of a simulation with no selection on GC content (Simulated).

Figure 4 - 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 5 – 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 6 – 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 and BM from OUCH analysis

Segment	GC	simulated segment GC	Longest ORF	Simulated segment longest ORF
10270_10563	62.27	77.4	Convergence error	49.17
11935_12275	194.76	255.55	252.13	Convergence error
11196_11705	Convergence error	223.01	Convergence error	210.91
320_722	161.05	184.27	229.07	Convergence error
7916_8695	54.65	59.3	56.76	22.84
8938_9248	259	Convergence error	219.14	237.68

Each cell represents the corrected AIC estimated by OUCH [39] comparing Brownian Motion to a single optimum OU model. When the algorithm reported errors in convergence, this is indicated by "Convergence error" in the table.

Supplementary figures

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

Figure S2 – as in figure 4 and 5, 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.

[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] M. Lynch, "Phylogenetic divergence of cell biological features," *eLife*, vol. 7, 21 2018, doi: 10.7554/eLife.34820.

[29] 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).

[30] 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.

[31] 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).

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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).

[44] 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).

[45] 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.

[46] 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.

[47] 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).

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

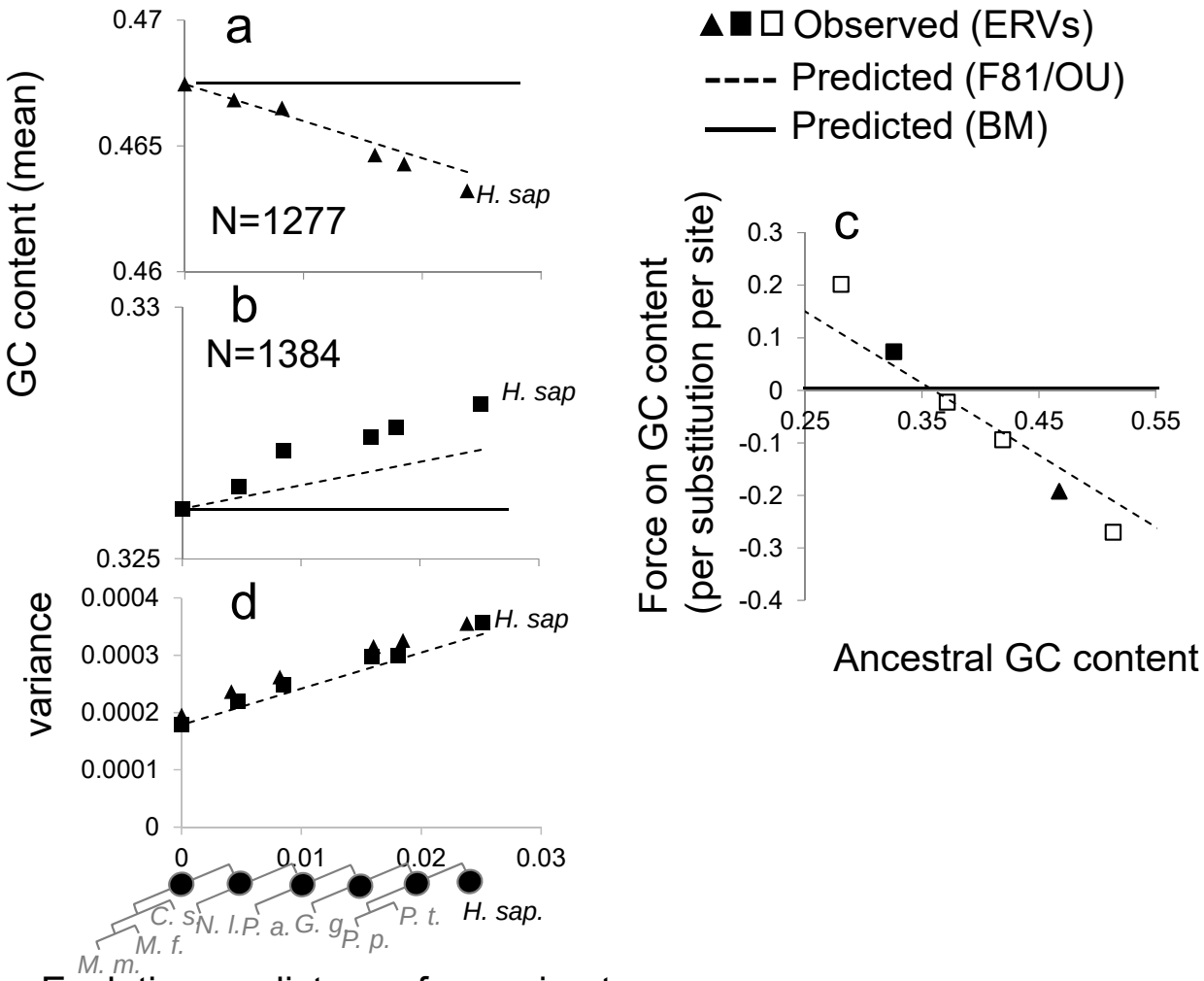
[65] K. P. Schliep, "phangorn: phylogenetic analysis in R," *Bioinformatics*, vol. 27, no. 4, pp. 592–593, Feb. 2011, doi: 10.1093/bioinformatics/btq706.

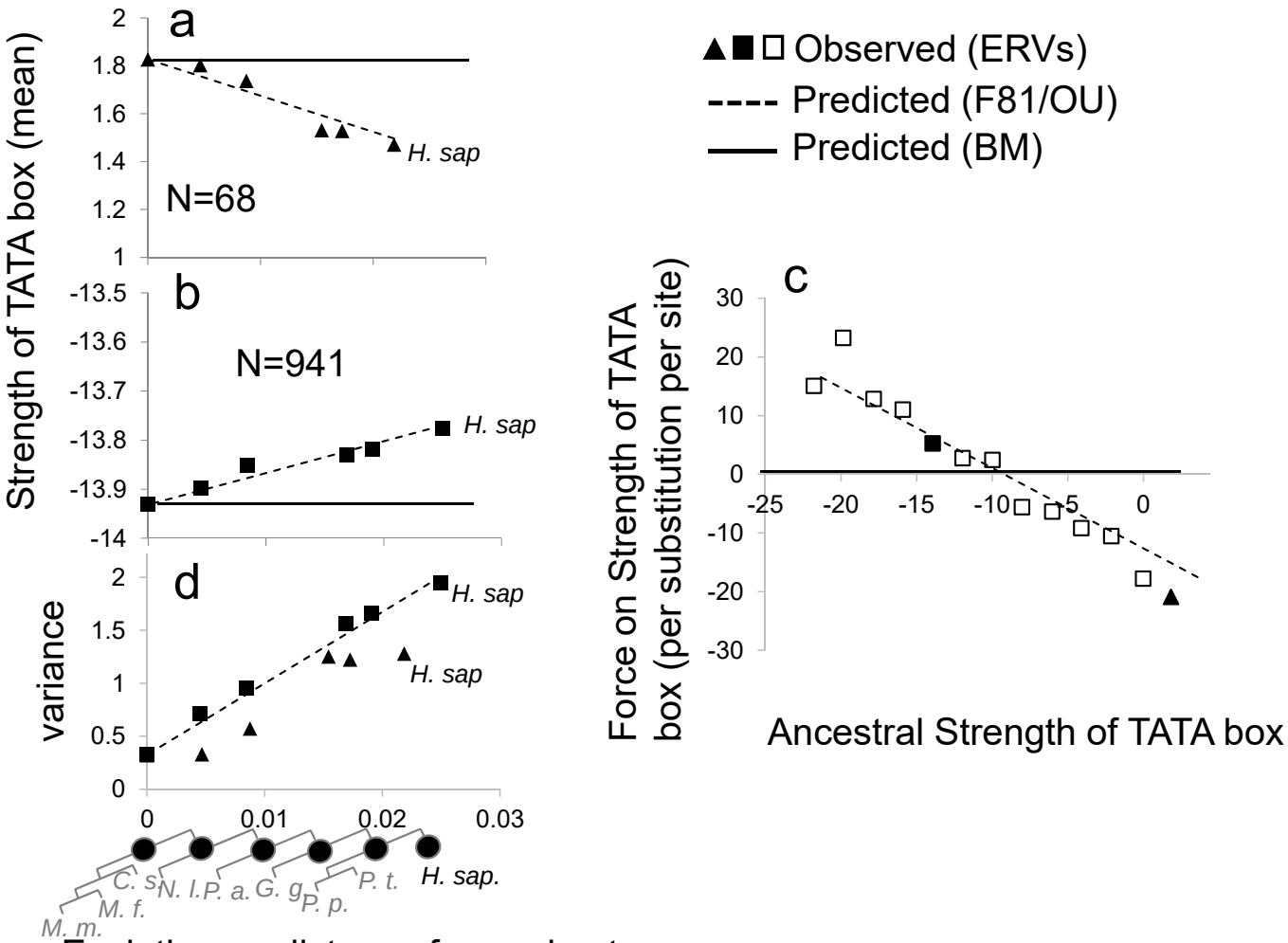
[66] A. Moses and S. Sinha, "Regulatory Motif Analysis," in *Bioinformatics: Tools and Applications*, 2009, pp. 137–163.

[67] D. R. Zerbino *et al.*, "Ensembl 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D754–D761, 04 2018, doi: 10.1093/nar/gkx1098.

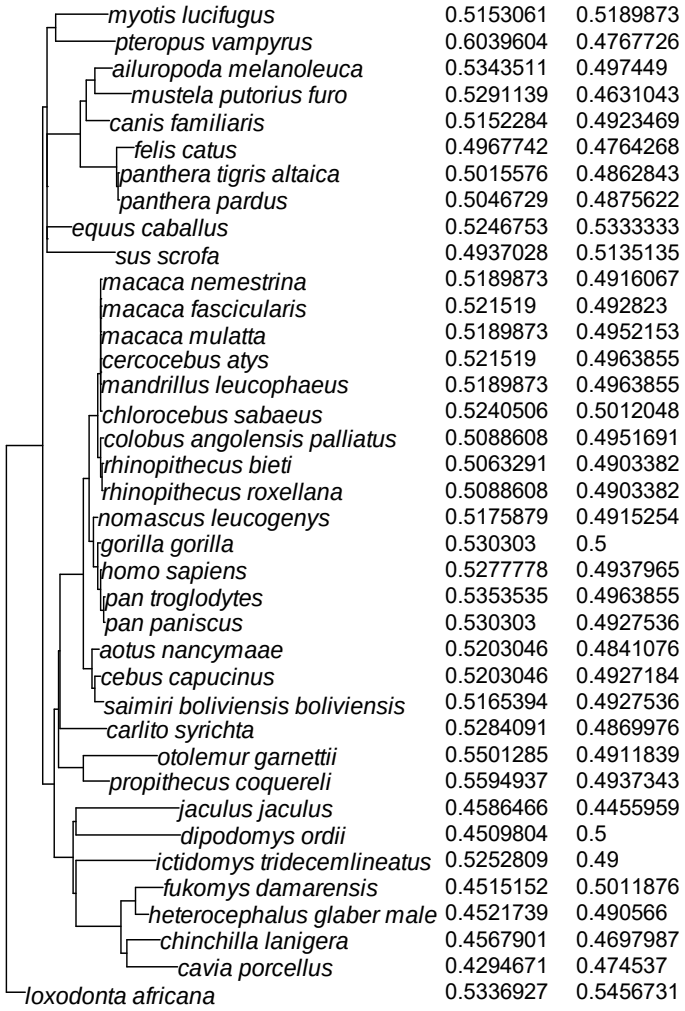
[68] Z. Yang, "PAML 4: phylogenetic analysis by maximum likelihood," *Mol. Biol. Evol.*, vol. 24, no. 8, pp. 1586–1591, Aug. 2007, doi: 10.1093/molbev/msm088.

[69] 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.

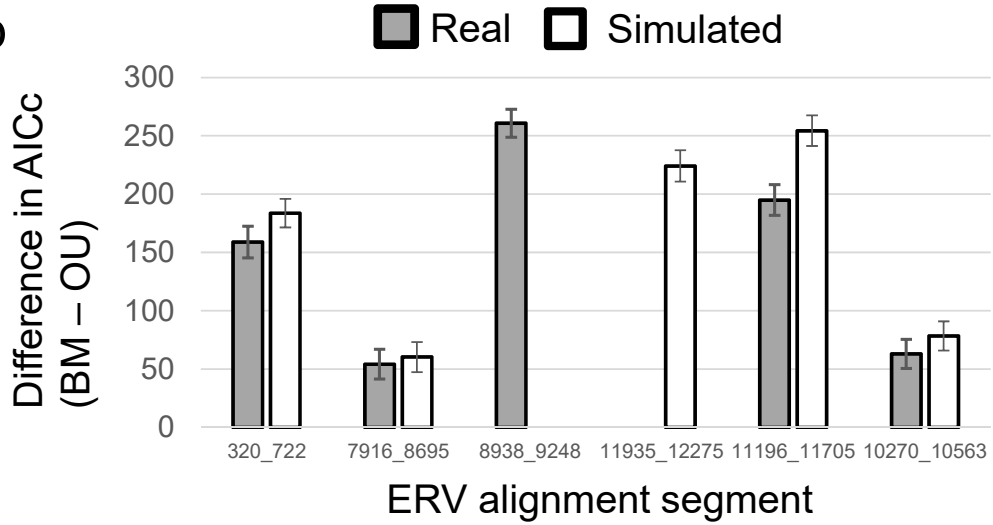


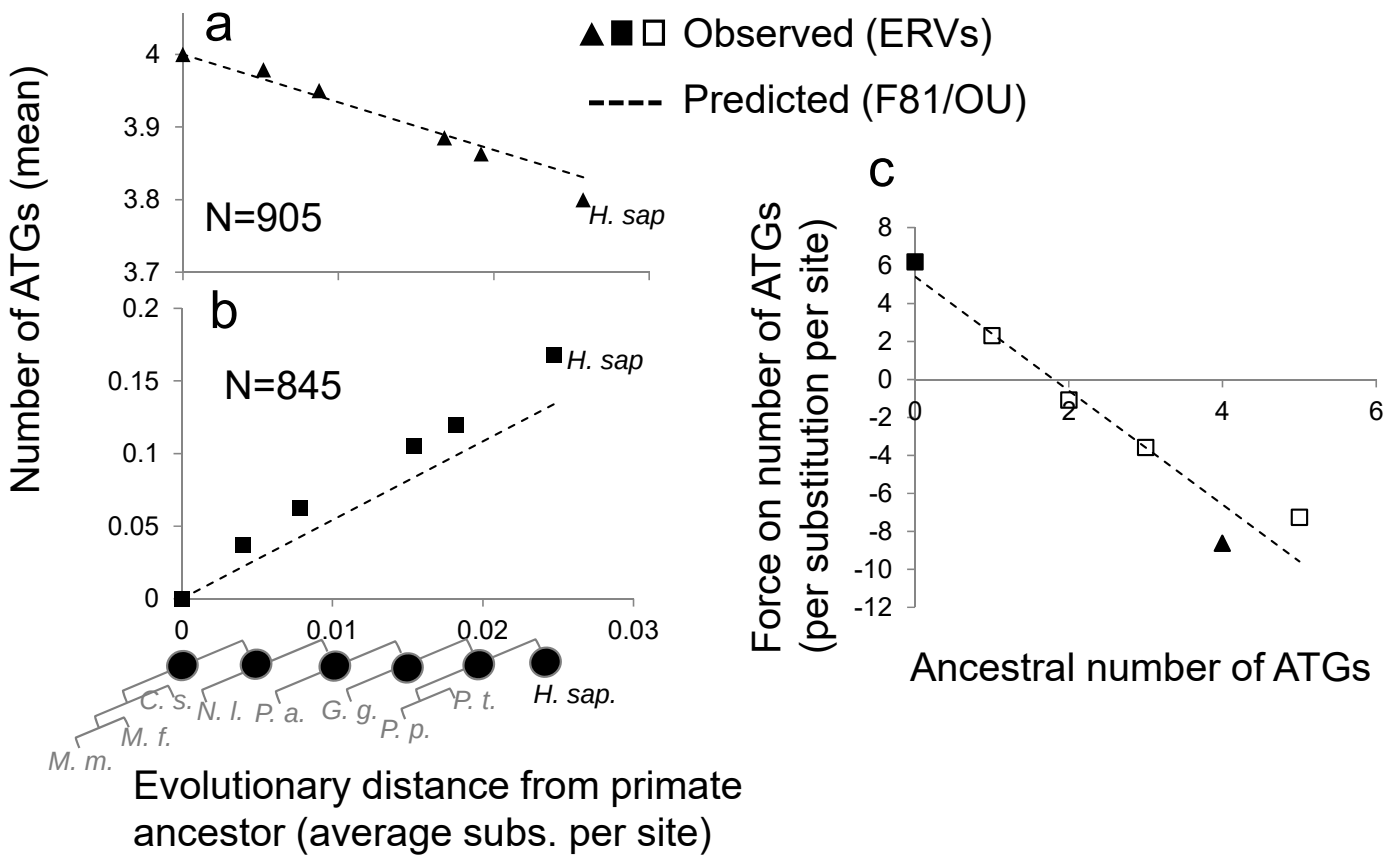


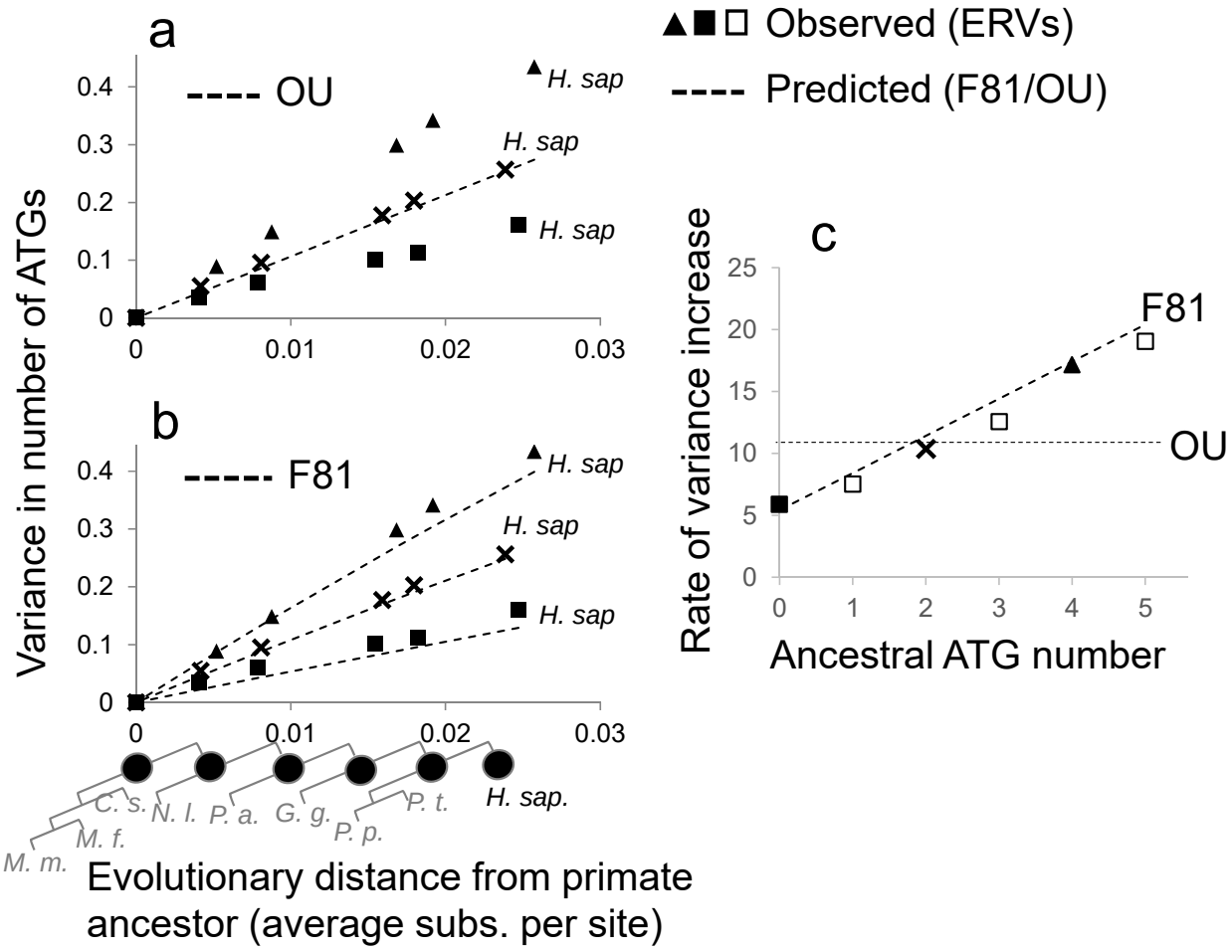
a

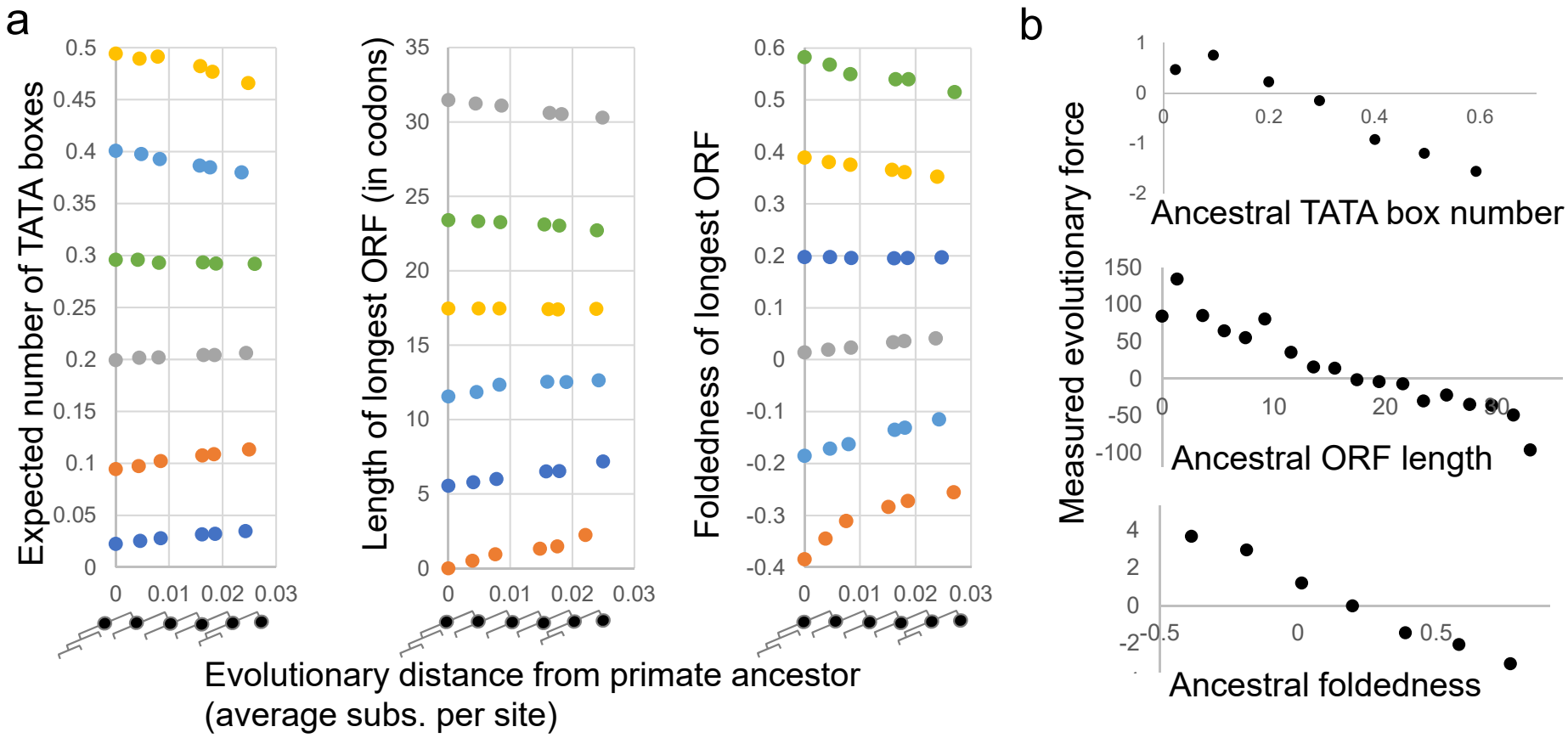


b

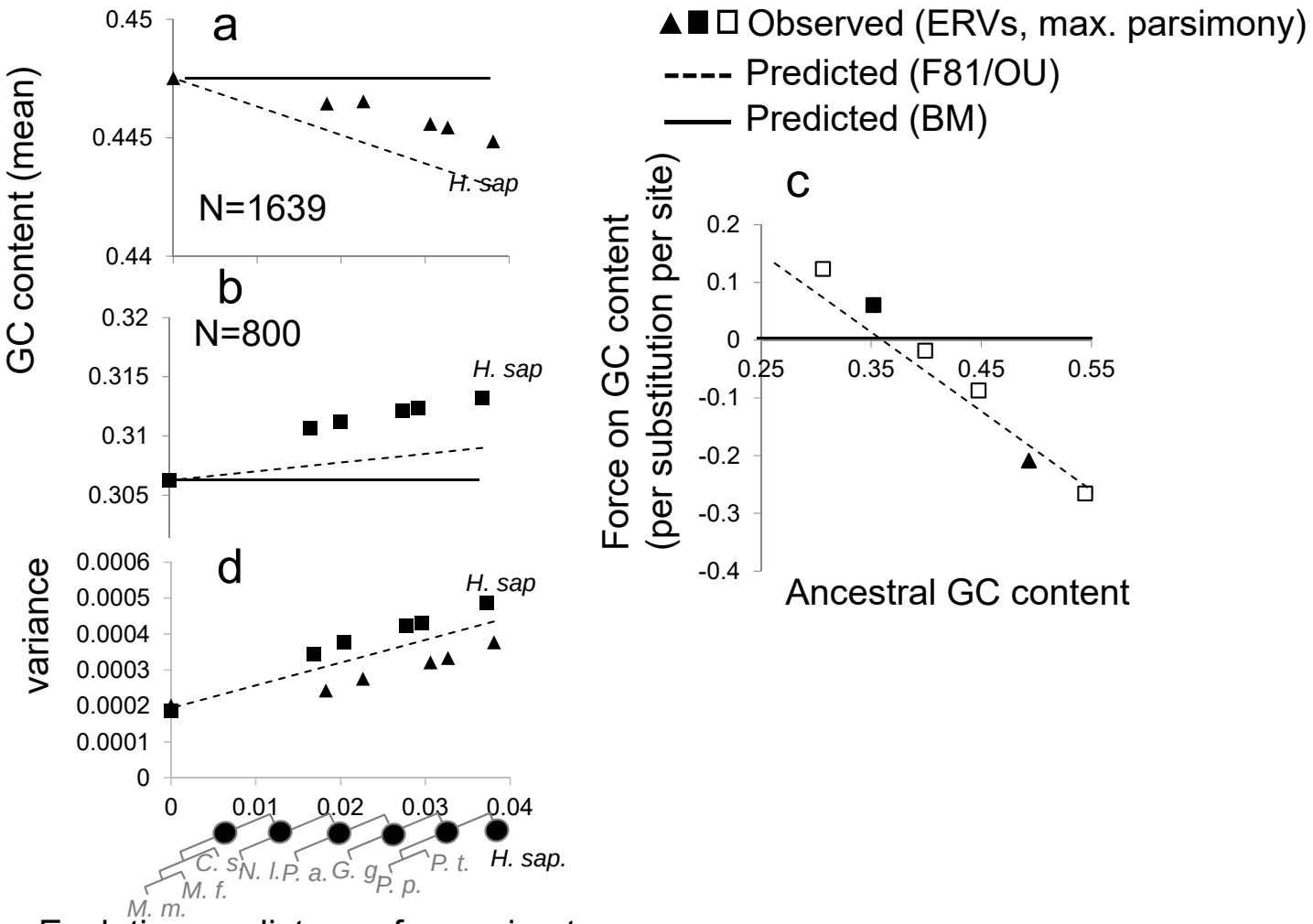


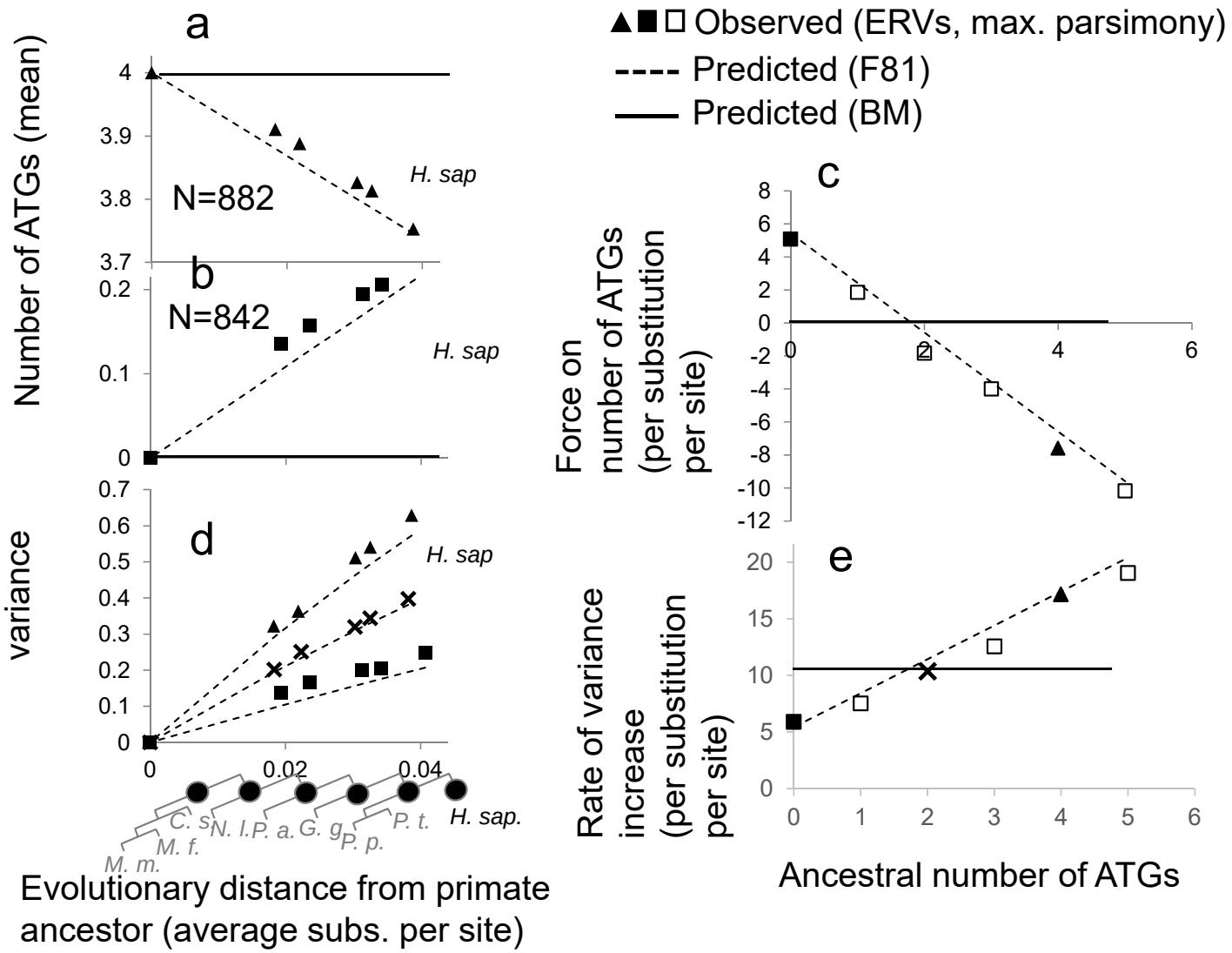






Supplementary figures





Appendix - Evolutionary dynamics of neutral phenotypes under DNA substitution models

Shadi Zabad and Alan M Moses

1. *Mean phenotype dynamics under the F81 model*

Starting with the formula given in the main text, we show that the mean phenotype dynamics are independent of the genotype.

$$E[Z(t)|X] = \sum_{Y} P(Y|X,t)Z(Y) = \sum_{n}\sum_{Y_n} P(Y_n|X_n,t)Z_n(Y_n)$$

We note that there are two possibilities, either the genotype at the n-th locus has stayed the same or it has changed into one of the other genotypes:

$$= \sum_{n} P(X_n|X_n,t)Z_n(X_n) + \sum_{Y_n \neq X_n} P(Y_n|X_n,t)Z_n(Y_n)$$

Substituting in the F81 model,

$$P(Y_n|X_n,t) = \pi_{Y_n}(1-e^{-ut})\ for\ Y_n \neq X_n$$

$$P(Y_n|X_n,t) = e^{-ut} + \pi_{Y_n}(1-e^{-ut})\ for\ Y_n = X_n$$

we have:

$$E[Z(t)|X] = \sum_{n} e^{-ut}Z_n(X_n) + \pi_{X_n}(1-e^{-ut})Z_n(X_n) + \sum_{Y_n \neq X_n} \pi_{Y_n}(1-e^{-ut})Z_n(Y_n)$$

Noting that the middle term is exactly the term for $Y_n = X_n$ in the last sum, we have

$$E[Z(t)|X] = \sum_{n} e^{-ut}Z_n(X_n) + \sum_{Y_n} \pi_{Y_n}(1-e^{-ut})Z_n(Y_n)$$

Now we can factor out terms that are independent of n

$$E[Z(t)|X] = e^{-ut}\sum_{n} Z_n(X_n) + (1-e^{-ut})\sum_{n}\sum_{Y_n} \pi_{Y_n}Z_n(Y_n)$$

The sum over n in the first term is the definition of Z_0, the initial phenotype, and the double sum in the second term is the definition of Z_{eq}, the long-time mutational equilibrium. This gives

$$E[Z(t)|X] = e^{-ut}Z_0 + (1-e^{-ut})Z_{eq}$$

Which depends only on Z_0 and Z_{eq} as claimed in the main text.

2. *A Biallelic model*

We envisage L independent and identical loci evolving in an unconstrained fashion, subject only to mutation and drift. These L loci, it is further assumed, are in linkage equilibrium (i.e. evolving in a free recombination regime) and evolve independently. This last assumption will enable us to derive expressions for the evolutionary dynamics of the mean phenotype as a linear function of the single locus statistics. Using the notation commonly employed in modern GWAS studies, we have:

$$z = X\beta + \epsilon = \sum_{n=1}^{L} X_n\beta_n + \epsilon$$

Where z denotes the phenotype , $X\beta$ is the genetic contribution to the phenotype, β are the effect sizes at each locus, and ϵ captures the contribution of the environment and other non-linear effects. We assume evolution proceeds in the weak mutation regime, where at any point in time the population is monomorphic. X is a 1×L binary vector indicating whether that population (or species) has the reference or alternative allele at each locus.

$$X_n = \begin{cases} 0 & \text{(if reference allele at locus n)} \\ 1 & \text{(if alternative allele at locus n)} \end{cases}$$

The main quantity of interest in our analysis is the mean over an ensemble of replicate evolutionary processes, $E[Z(t)]$, given an initial starting phenotype $Z(t=0) \equiv Z_0$. The environmental contribution is omitted because we follow the standard assumption that $E[\epsilon] = 0$. Assuming that the effect sizes are fixed and the genetic architecture remains constant, we use the linearity of the expectation to obtain:

$$E[Z(t)] = E[X(t)]\beta = \sum_{n=1}^{L} E[X_n(t)]\beta_n$$

The weak-mutation assumption means that instead of modeling the full trajectory of the genotype, $X_n(t)$, from introduction to fixation or extinction, we model the state of the locus as a jump process between 2 states (reference vs. alternative), governed by a Continuous Time Markov Chain (CTMC). Under this model, the reference allele is expected to be substituted by the alternative allele with an infinitesimal rate, λ, while the reverse happens with rate $\lambda\kappa$, where $\kappa > 0$ is a parameter that quantifies the possible bias in mutation rates. We note that in this model (in contrast to Jukes- Cantor or F81) rates are not correctly scaled so that evolutionary distance is measured in substitutions per site. The transition probabilities in this model are given by

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

To find the probability that the alternative allele is fixed at any point in time, we marginalize over the ancestral state $P(X(t)=1) = \sum_{i\epsilon 0,1} P(X(0)=i)\, P_{i1}(t)$. We have

$$P(X(t)=1) = P(X(0)=1)e^{-\lambda(1+\kappa)t} + \frac{1}{1+\kappa}(1 - e^{-\lambda(1+\kappa)t})$$

This system has a stationary distribution, $\begin{bmatrix} 1-\pi \\ \pi \end{bmatrix} = \begin{bmatrix} P(X(t\to\infty)=0) \\ P(X(t\to\infty)=1) \end{bmatrix} = \frac{1}{1+\kappa}\begin{bmatrix} \kappa \\ 1 \end{bmatrix}$, so we can also write

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

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 at time t as $E[X(t)] = P(X(t)=1)$. Substituting this in,

$$E[Z(t)] = E[X(t)]\beta = \sum_{n=1}^{L} P(X(0)=1)e^{-\lambda(1+\kappa)t}\beta_n + \pi(1-e^{-\lambda(1+\kappa)t})\beta_n$$

Factoring out the locus independent terms,

$$E[Z(t)] = e^{-\lambda(1+\kappa)t}\sum_{n=1}^{L} P(X(0)=1)\beta_n + (1-e^{-\lambda(1+\kappa)t})\sum_{n=1}^{L}\pi\beta_n$$

We note that $\sum_{n=1}^{L} P(X(0)=1)\beta_n = E[Z(t=0)]$ and $\sum_{n=1}^{L}\pi\beta_n = E[Z(t\to\infty)]$.Since $E[Z(t\to\infty)] \equiv Z_{eq}$ and since we assume the population is monomorphic, $E[Z(t=0)] = Z_0$.

$$E[Z(t)] = e^{-\lambda(1+\kappa)t}Z_0 + (1-e^{-\lambda(1+\kappa)t})Z_{eq}$$

The above expression for the mean has the exact mean-reverting form as the mean of the Ornstein-Uhlenbeck model, as claimed in the main text. 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.

A second quantity that we can use 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 different species that evolved completely independently. The first species evolved for s time units and the second species evolved for r time units. Since the loci are assumed to be independent, the phylogenetic covariance then reduces to the sum of the covariances between the state of locus j at times (t + s) and (t + r).

$$Cov[Z(t+s), Z(t+r)] = \sum_{n} {\beta_n}^2 Cov(X(t+s), X(t+r))$$

$$= \sum_{n} {\beta_n}^2 (E[X(t+s)X(t+r)] - E[X(t+s)]E[X(t+r)])$$

Substituting these in and simplifying gives

$$Cov[Z(t+s), Z(t+r)] = e^{-\lambda(1+\kappa)(r+s+2t)} \sum_n {\beta_n}^2 \Big[(2\pi - 1)\big(\pi - P(X(0)=1)\big)e^{\lambda(1+\kappa)t} - \big(\pi - P(X(0)=1)\big)^2 + \pi(1-\pi)e^{\lambda(1+\kappa)2t}\Big]$$

Note that the covariance cannot be expressed in terms of phenotypic quantities alone, and therefore depends on the genotype. Under certain assumptions we can obtain genotype independent forms. The simplest is when we assume the initial genotype is at its long term equilibrium, $P(X(0)=1) = \pi$.

$$Cov[Z(t+s), Z(t+r)] = e^{-\lambda(1+\kappa)(r+s)} \sum_n {\beta_n}^2 \pi(1-\pi) = e^{-\lambda(1+\kappa)(r+s)} V[Z(t\to\infty)]$$

Where we recognized $\sum_n {\beta_n}^2 \pi(1-\pi) = \sum_n \big(E\big[{Z_n}^2 | t\to\infty\big] - E[Z_n | t\to\infty]^2\big) = V[Z(t\to\infty)]$ as the long-term phenotypic variance. When t and s are small, the covariance is nearly equal to the variance (so the phenotypes are totally correlated) but at long divergence time (t and s large) the covariance decays. While the form of this is attractive, it is inconsistent with the weak-mutation assumption (and empirical observation) that genotypes are not at their equilibriums in finite populations, but rather close to monomorphic. For example, this formula predicts that the variance in the population is always at its mutational equilibrium.

Another assumption that allows a simple expression for the covariance is that there is no mutational bias, $\kappa = 1$. In this case, $\pi = \frac{1}{2}$.

$$Cov[Z(t+s), Z(t+r)] = e^{-2\lambda(r+s+2t)} \sum_n {\beta_n}^2 \left[-\left(\frac{1}{2} - P(X(0)=1)\right)^2 + \frac{1}{4}e^{4\lambda t}\right]$$

$$= e^{-\lambda 2(r+s+2t)} \sum_n {\beta_n}^2 \left[-\left(\frac{1}{4} - P(X(0)=1)(1-P(X(0)=1))\right) + \frac{1}{4}e^{4\lambda t}\right]$$

$$= e^{-2\lambda(r+s+2t)} \sum_n {\beta_n}^2 \left[V[X(t=0) - \frac{1}{4}(1-e^{4\lambda 2t})\right]$$

We can recognize $\sum_n {\beta_n}^2 V[X(t=0)] = V[Z(t=0)]$ and $\sum_n \frac{1}{4}{\beta_n}^2 = V[Z(t\to\infty)]$, so we have

$$Cov[Z(t+s), Z(t+r)] = e^{-2\lambda(r+s+2t)} \Big[V[Z(t=0)] - (1-e^{4\lambda t})V[Z(t\to\infty)]\Big]$$

This is now genotype independent, and depends only on the initial phenotypic variance and the long-term phenotypic variance. In the weak-mutation regime, where we assume the population is monomorphic, the initial variance is $V[Z(t=0)] = 0$, and we can further simplify:

$$Cov[Z(t+s), Z(t+r)] = e^{-2\lambda(r+s+2t)} V[Z(t\to\infty)] + e^{-2\lambda(r+s)} V[Z(t\to\infty)]$$

Thus, the variance (r=0, s=0) shows OU dynamics with rate = 2λ. This model could be used directly as a null hypothesis in standard phylogenetic comparative inference frameworks.

3. *Variance of GC content under the F81 model*

Starting from the general formula for the variance of an additive phenotype,

$$V[Z(t)|X] = \sum_n C(X) \cdot a_n {a_n}^T$$

For GC content, $a_n = \frac{1}{L}\begin{bmatrix}0\\1\\1\\0\end{bmatrix}$, for all n, so for all n

$${a_n}{a_n}^T = \frac{1}{L^2}\begin{bmatrix}0&0&0&0\\0&1&1&0\\0&1&1&0\\0&0&0&0\end{bmatrix}$$

$$C_{mm} = E[Y_m|X](1 - E[Y_m|X]) = (1 - e^{-ut})(\pi_m(1-\pi_m) + e^{-ut}(X_m - \pi_m)^2)$$

$$C_{km} = -E[Y_k|X]E[Y_m|X] = -(1 - e^{-ut})(\pi_k\pi_m + e^{-ut}(\pi_m X_k + X_m\pi_k - \pi_k\pi_m))$$

$$V[Z(t)|X] = \frac{1}{L^2}\sum_n C_{CC} + C_{GG} + C_{CG} + C_{GC}$$

Because $C_{CG} = C_{GC}$, noting that $\pi_C + \pi_G = Z_{eq}$

$$V[Z(t)|X] = \frac{(1 - e^{-ut})}{L^2}\sum_n \sum_{m\epsilon C,G} (\pi_m(1-\pi_m) + e^{-ut}(X_m - \pi_m)^2) - 2(\pi_C\pi_G + e^{-ut}(\pi_C X_G + X_C\pi_G - \pi_C\pi_G))$$

Defining $N_G + N_C = LZ_0$ to be the numbers of Gs and Cs in the initial genotype,

$$V[Z(t)|X] = \frac{(1 - e^{-ut})}{L^2}(L\sum_{m\epsilon C,G} \pi_m(1-\pi_m) - 2L\pi_C\pi_G + e^{-ut}\left(-2(\pi_C N_G + N_C\pi_G - L\pi_C\pi_G) + \sum_n \sum_{m\epsilon C,G} (X_m - \pi_m)^2\right))$$

$$V[Z(t)|X] = \frac{(1 - e^{-ut})}{L^2}(L\sum_{m\epsilon C,G} \pi_m(1-\pi_m) - 2L\pi_C\pi_G + e^{-ut}(-2(\pi_C N_G + N_C\pi_G - L\pi_C\pi_G) + N_G(1-\pi_C)^2 + N_C{\pi_C}^2 + N_C(1-\pi_G)^2 + N_G{\pi_G}^2))$$

$$V[Z(t)|X] = \frac{(1 - e^{-ut})}{L^2}(L\sum_{m\epsilon C,G} \pi_m(1-\pi_m) - 2(1 - e^{-ut})L\pi_C\pi_G + e^{-ut}(N_C({\pi_C}^2 - 4\pi_G + 1 + {\pi_G}^2) + N_G({\pi_G}^2 - 4\pi_C + 1 + {\pi_C}^2))$$

Which clearly depends on the genotype. If we assume that $\pi_C = \pi_G = \frac{Z_{eq}}{2}$ which is usually the case, we can get a genotype independent result by substituting:

$$V[Z(t)|X] = \frac{(1-e^{-ut})}{L^2}(2L\frac{Z_{eq}}{2} - 2L\left(\frac{Z_{eq}}{2}\right)^2 - 2\left(1-e^{-\beta t}\right)L\left(\frac{Z_{eq}}{2}\right)^2 + e^{-ut}\left(LZ_0(2\left(\frac{Z_{eq}}{2}\right)^2 - 4\frac{Z_{eq}}{2} + 1)\right))$$

This can be simplified

$$V[Z(t)|X] = 2\frac{(1-e^{-ut})}{L}(\frac{Z_{eq}}{2} - \left(\frac{Z_{eq}}{2}\right)^2 - (1-e^{-ut})\left(\frac{Z_{eq}}{2}\right)^2 + e^{-ut}\left(Z_0(\left(\frac{Z_{eq}}{2}\right)^2 - Z_{eq} + \frac{1}{2})\right))$$

$$V[Z(t)|X] = Z_{eq}\frac{(1-e^{-ut})}{L}(1 - Z_{eq} + e^{-ut}\left(Z_{eq} + Z_0(\frac{Z_{eq}}{2} - 2 + \frac{1}{Z_{eq}})\right))$$

But it still looks non-monotonic. If $Z_0 = Z_{eq}$

$$V[Z(t)|X] = Z_{eq}\frac{(1-e^{-ut})}{L}\left((1-Z_{eq})(1+e^{-ut}) + e^{-ut}\frac{{Z_{eq}}^2}{2}\right)$$

$$V[Z(t)|X] = \frac{(1-e^{-2ut})}{L}Z_{eq}(1-Z_{eq}) + \frac{(e^{-ut}-e^{-2ut})}{L}(\frac{{Z_{eq}}^3}{2}) \approx \frac{(1-e^{-2ut})}{L}Z_{eq}(1-Z_{eq})$$

Since ${Z_{eq}}^3$ is small, this is very close to an exponential decay to the equilibrium variance with twice the rate, which is the OU dynamics. So it is very close to an OU process at equilibrium, but not exact. The equilibrium variance is

$$V[Z(t\to\infty)] = \frac{1}{L}Z_{eq}(1-Z_{eq})$$

and is inversely proportional to the number of loci (L). Therefore the rate of increase of the variance will be inversely proportional to the length of the sequence considered.

4. *Dynamics of GC content under the Jukes cantor model*

This means that $\pi_C = \pi_G = \frac{1}{4}$, $u = \frac{4}{3}$ and $Z_{eq} = \frac{1}{2}$. Now

$$E[Z(t)|X] = e^{-\frac{4}{3}t}Z_0 + (1-e^{-\frac{4}{3}t})\frac{1}{2} \approx \frac{4}{3}t(\frac{1}{2} - Z_0) + Z_0$$

Which still shows a linear restoring force proportional to the distance of the initial phenotype from 1/2.

$$V[Z(t)|X] = \frac{\left(1 - e^{-\frac{4}{3}t}\right)}{2L}\left(\frac{1}{2}\left(1 + e^{-\frac{4}{3}t}\right) + e^{-\frac{4}{3}t}Z_0\left(\frac{1}{4}\right)\right) = \frac{\left(1 - e^{-\frac{8}{3}t}\right)}{L}\frac{1}{4} + \frac{e^{-\frac{4}{3}t} - e^{-\frac{8}{3}t}}{8L}Z_0$$

Now there is no genotype dependence on the variance. Since 8L >> 4L, this shows a weak dependence on the initial phenotype, and is weakly non-monotonic. As before, the first term matches exactly the OU variance dynamics, with the equilibrium variance of 1/4L as expected.