A hierarchical Bayesian model for detecting convergent rate changes of conserved noncoding elements on phylogenetic trees

Zhirui Hu[1], Timothy B. Sackton[2], Scott V. Edwards[3], Jun S. Liu[1]

[1]Department of Statistics, [2]Informatics and [3]Organismic and Evolutionary Biology, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA

Corresponding to: Zhirui Hu (zhiruihu@g.harvard.edu) or Jun S. Liu (jliu@stat.harvard.edu)

Running title: Hierarchical Bayesian model to detect rate changes

Keywords: Convergence, Phylogenetics, Bayesian model

Abstract

Conservation of DNA sequence over evolutionary time is a strong indicator of function, and gain or loss of sequence conservation can be used to infer changes in function across a phylogeny. Changes in evolutionary rates on particular lineages in phylogeny can indicate shared functional shifts, and thus can be used to detect genomic correlates of phenotypic convergence. However, existing methods do not allow easy detection of patterns of rate variation, which causes challenges for detecting convergent rate shifts or other complex evolutionary scenarios. Here, we introduce PhyloAcc, a new Bayesian method to model substitution rate changes in conserved elements across a phylogeny. The method can handle diverse evolutionary patterns and complex patterns of convergence, assumes a latent conservation state for each branch on the phylogenetic tree, estimates element-wise substitution rates per state, and detects changes of substitution rate as the posterior probability of a state switch. Simulations show that PhyloAcc can detect rate shifts in multiple species better than likelihood ratio based methods, and has higher accuracy to detect complex patterns of substitution rate changes than prevalent Bayesian relaxed clock models. We demonstrate the utility of this method in two classic examples of convergent phenotypes: loss of flight in birds and the transition to marine life in mammals. In each case, our approach reveals numerous examples of conserved non-exonic elements with accelerations specific to the phenotypically convergent lineages. This method is widely applicable to any set of conserved elements where multiple independent rate changes are expected on a phylogeny.

Introduction

One of the major revelations of comparative genomics has been the discovery of regions of the genome falling well outside protein-coding genes that nonetheless exhibit considerable levels of conservation across evolutionary time (Bejerano et al., 2004; Siepel et al., 2005; Woolfe et al., 2005;

Venkatesh et al., 2006; Lindblad-Toh et al., 2011). Conserved non-exonic elements (CNEEs) are likely regulatory in function (Capra et al., 2013a) and are of particular interest because of the likely role that changes in regulation play in phenotypic differences between species (King and Wilson 1975; Pollard et al., 2006; Mclean et al., 2011; Hiller et al., 2012a; Marcovitz et al., 2016). Changes of conservation of these elements in a subset of lineages is thus often associated with altered regulatory activity and ultimately phenotypic divergence (Mclean et al., 2011). Numerous studies have used changes in sequence conservation of conserved elements as means to identify regulatory regions which may be of particular importance for lineage-specific phenotypes. For example, Pollard et al. (2006) identified 202 regions accelerated in the human genome but conserved in other vertebrates, some of which are RNA genes and tissue-specific enhancers. Outside of humans, Holloway et al. (2016) identified 4,797 regions accelerated at the base of therian mammals, many of which are noncoding and close to developmental transcription factors, and Booker et al. (2016) discovered 166 bat-accelerated regions overlapping with enhancers in developing mouse limbs.

Phenotypic convergence, in which the same function evolves multiple times independently, often due to adaption to similar environmental changes, is usually assumed to be one of the strongest signals of natural selection (Kishida et al., 2007; Brawand et al., 2008; Stern, 2013; Meredith et al., 2014). However, we generally do not have a robust understanding of the genomic changes underlying phenotypic convergence (Wray, 2013; Rosenblum et al., 2014). Do convergent phenotypes arise from repeated use of the same underlying genetic elements, or do they arise via independent genetic pathways (Orr, 2005; Tenaillon et al., 2012; Parker et al., 2013; Storz, 2016)? Convergence at the molecular level can arise because of identical substitutions, or via consistent shift of substitution rates in genomic regions encoding particular traits that are altered among these species due to changes of selection pressure (Chikina et al., 2016; Partha et al., 2017).

A variety of methods exist to test for an association between substitution rates and convergent phenotypes, which is predicted if the same genetic elements are associated with a phenotype in multiple species. The Forward Genomics method (Hiller et al., 2012b; Prudent et al., 2016), tests the significance of Pearson correlation between normalized substitutions and hypothetical phenotypic state on each branch by assuming a linear relationship. Chikina et al. (2016), studying protein-coding genes with convergent shifts in marine mammals, performed a Wilcoxon rank sum test of relative substitution rates over "terrestrial" and "marine" branches. Finally, the PHAST method (Hubisz et al., 2011), tests the model allowing substitution rates shift in a subset of branches against null model with constant rate for all branches using likelihood ratio.

However, these methods for detecting genomic regions with parallel substitution rate changes in diverse lineages are generally limited to test a single pre-specified shift pattern on a phylogeny, in which conservation states of ancestral regions are usually inferred using Dollo parsimony based on phenotypes of extant species. These methods also do not always distinguish among strong acceleration in a single tip branch (Supplementary Figure 1A), weaker acceleration across multiple clades containing that branch (Supplementary Figure 1B), and acceleration beyond specific subset of branches (Supplementary Figure 1C). The only likelihood-based method we are aware of that considers multiple patterns of rate/character transitions is TraitRate (Mayrose & Otto, 2011; Karin et al., 2017). However, this method requires an ultrametric species tree as input, which means it cannot consider substitution rate variation among species. Also, TraitRate can only test for an association between rates changes and a given trait, but cannot detect where these transitions occur on a phylogeny.

Here, we introduce PhyloAcc, an alternative Bayesian method to model multiple substitution rate changes on a phylogeny. In our new method, we relax the parsimony assumption on the history of rate shifts and develop a model-based method to estimate the conservation state

of each branch based on sequences of extant species. Our method allows each genomic region to have a different pattern of shifts of substitution rate. Using MCMC to sample from the posterior distribution, our model yields the most probable evolutionary pattern as well as its uncertainty for each genomic region. It also outputs the posterior distribution of the substitution rate for each genetic unit, as an indicator of the age of rate shift or magnitude of selection change. To increase the accuracy of rate estimates, we pool information across elements and shrink substitution rates towards a common prior. Our method also evaluates the strength of the association between rate shifts at a genomic region and phenotypic change using Bayes factors. A Bayes factor, defined as the ratio of marginal likelihoods obtained by integrating over parameter space in competing models, is a compelling choice for model selection, and has a natural interpretation as a measure of evidence from the data supporting one model over another. Unlike previous methods using maximum likelihood estimators of substitution rates and a single evolutionary pattern, PhyloAcc considers the uncertainty of estimated substitution rates and all possible evolutionary histories of conservation states given the phenotypes of extant species. While our method has some similarity to various relaxed clock models (Drummond & Suchard, 2010; Heath et al., 2012), which also allow for varying substitution rates across phylogeny, the focus of our method is on detecting patterns in the shift in rates instead of providing estimates of ages of nodes. Moreover, these methods do not have as a goal detecting genomic regions with evolutionary shifts in rate that are correlated with phenotype change.

To demonstrate the power of PhyloAcc on real data, we applied our new method to two classic examples of phenotypic convergence: loss of flight in birds (Sackton et al., 2018) and transition to marine life in mammals (McGowen et al., 2014; Foote et al., 2015). In both cases we identify conserved elements with specific convergent rate shifts associated with our target phenotype, revealing novel, putative regulatory regions which may be repeatedly associated with these evolutionary transitions.

Results

Hierarchical Bayesian Phylogenetic model

The goal of our model is to identify branches on a phylogeny where particular genome elements are evolving with a different substitution rate. We take as input a neutral phylogenetic tree, with branch lengths representing the expected number of substitutions along that branch under neutrality, and assume that the substitution process follows a standard continuous time Markov Process. To model rate variation, we introduce a relative substitution rate, *r,* such that the expected number of substitutions along rate-varied branch will be scaled by *r* (see Methods). In this model, $r = 1$ for sequences evolving neutrally, with $r < 1$ or $r > 1$ indicating evolutionary departures from neutral rates. We consider mostly conserved genomic regions (average $r < 1$), and allow selection to vary among lineages, so that the relative substitution rate varies across phylogeny.

Because many genomic elements of interest are relatively conserved and short in length, with few substitutions, estimating substitution rates per branch is implausible. Instead, we assume that, for each element, a limited number of discrete rate classes occur on the phylogeny, allowing us to estimate *r* (per class) jointly from all branches sharing similar evolutionary rates. We define $\boldsymbol{Z_i} = (Z_{i1}, Z_{i2}, \dots, Z_{in})$ to denote the latent *conservation state* on each of *n* branches for element i; the substitution rate on a branch then takes discrete value which depends on its latent conservation state. Our model will infer the latent state on each branch and identify when the transitions of states occur. In our applications below, we assumed three conservation states (i.e. $\mathrm{Z_{is}} \in \{0,1,2\}, s = 1,2,\dots,n$), which are: a neutral state with no selection, a conserved state implying purifying selection, and an accelerated state implying relaxed or positive selection. However, our model is extensible to additional states (e.g., to allow multiple acceleration states to model a scenario where independent clades lost conservation at different times and thus have different rates, or to allow a loss of conservation state and a positive selection state with r >> 1). Each state has its own

substitution rate relative to the neutral rate, denoted by $\mathbf{r_i} = (r_0 = 1, r_{i1}, r_{i2})$ respectively for element *i*. Branches in the conserved state will have fewer substitutions than under neutrality ($r_{i1} < 1$), whereas branches in the accelerated state are expected to experience more substitutions than those of conserved states ($r_{i2} > r_{i1}$), although this number can be less than, equal to, or greater than neutrality. While we will refer to these as substitution rates in the following, they are defined relative to the neutral rate.

To model how latent conservation state changes along the phylogeny, we start by assuming that each element is in the neutral state at the root of the tree. We assume that Dollo's irreversible evolution hypothesis (Gould, 1970) holds for transitions from conserved to accelerated states, so that along each lineage Z_{is} can transit from a neutral to a conserved state, and then to an accelerated state but not in reverse. The transition probability of latent states encourages nearby branches to have same state and similar substitution rates, a common assumption in phylogenetics (e.g., autocorrelated rate models; Drummond et al. 2006) and reasonable with closely-related species in a phylogeny. The prior of element-wise substitution rates (r_{i1} and r_{i2}) provides a soft bound on substitution rates for each latent class, and also pools information from all elements to make estimates of substitution rates and latent states more reliable. This is especially useful for the common case where only a few branches are accelerated and/or few substitutions occur. Our method iteratively updates unobserved DNA sequences of ancestral species, latent states $\mathbf{Z}$ and substitution rates $\mathbf{r}$ for each element by using collapsed Gibbs sampling and outputs posterior distribution of $\mathbf{Z}$ which gives the evolutionary pattern, the number of independent accelerations of a particular element as well as the uncertainty of when accelerations occur.

Testing a priori evolutionary patterns

If phenotypic convergence is associated with convergence at the molecular level, we predict that changes in substitution rate will be associated with lineages displaying the convergent

phenotype. To test this association, given the sequence alignments and a pre-specified set of target convergent lineages, we compare marginal likelihoods between a null model assuming no acceleration in any lineage, and alternate models allowing either shifts only in lineages associated with the convergent trait or allowing shifts in arbitrary lineages. In the null model (M_0), all branches are either neutral or conserved ($Z_s = 0\ or\ 1$); in the lineage-specific model (M_1), substitution rates on the branches leading to target species with the trait of interest can be accelerated ($Z_s = 2$) while all other branches must be either neutral or conserved ($Z_s = 0\ or\ 1$); in the full model (M_2), the latent conservation states $\mathbf{Z}$ can take any configuration across the phylogeny. To compare models, we compute the marginal likelihood $P(Y|M_i)$ for each model by integrating out (unobserved) ancestral DNA sequences, latent conservation states $\mathbf{Z}$, and substitution rates (r_1 and r_2). We then compute two Bayes factors, $BF1 = \frac{P(Y|M_1)}{P(Y|M_0)}$ and $BF2 = \frac{P(Y|M_1)}{P(Y|M_2)}$, as criteria to identify DNA elements accelerated exclusively in target lineages. BF1 distinguishes elements accelerated in target species from those with no acceleration, while BF2 distinguishes elements that are specifically accelerated in target species from those that lost conservation in other lineages. If both Bayes factors are large, we might conclude that M_1 is better fitted by the data and this element is exclusively accelerated in our target species. Including BF2 to identify elements with a specific evolutionary pattern is crucial to exclude elements accelerated in species not associated with the target phenotypic change, which might include regulatory elements with broader functions.

Applications of PhyloAcc to examples of phenotypic convergence

To demonstrate the power of PhyloAcc, we focus on two classic cases of convergent evolution: loss of flight in palaeognath birds (Mitchell et al., 2014; Sackton et al., 2018) and the transition to marine environments in mammals (Foote et al., 2015; Chikina et al., 2016). We start by simulating data under the phylogenetic model for birds or mammals to verify the performance of

our method, and then test for convergently accelerated non-coding elements in real data. Here, we begin by introducing the two datasets that form the basis for the remainder of our work.

Flightlessness in palaeognathous birds - Our first example of phenotypic convergence is loss of flight in birds, which has occurred multiple times in bird evolution (Roff, 1994; Harshman et al., 2008; Mitchell et al., 2014). Recent phylogenetic work supports the conclusion that the ratites (including species of ostrich, emu, cassowary, kiwi, rheas, and the extinct moas and elephant bird) are paraphyletic, implying convergent loss of flight in this classic group (Harshman et al., 2008; Baker et al., 2014; Mitchell et al., 2014; Yonezawa et al., 2017; Sackton et al., 2018). We used a set of 284,001 CNEEs identified in a recent study (Sackton et al., 2018) and aligned in 43 species of birds and non-avian reptiles, including 23 neognath birds, 9 flightless ratites (moa, ostrich, 2 rheas, 3 kiwis, emu and cassowary), 4 volant tinamous and 7 non-avian reptiles as outgroup (Figure 1A).

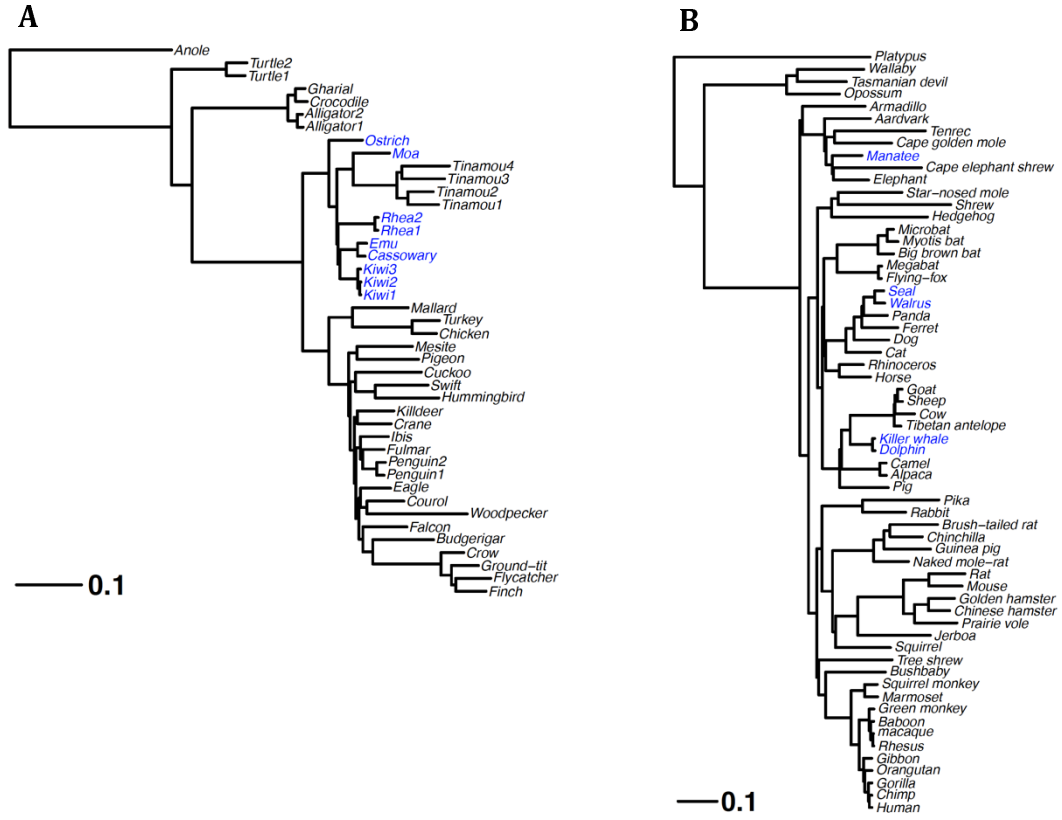


Figure 1: (a) Phylogeny for avian data. Palaeognaths consist of the flightless ratites and volant tinamous. Branch lengths are estimated from phyloFit in the phast package Ratites are shown in blue. (b) Phylogeny of mammalian data. 5 marine mammals are shown in blue.

Convergence in marine mammals - Another classic example of convergent trait is transition to a marine habitat in mammals, which originated three times independently in cetaceans, pinnipeds and sirens. For this example, our analysis was based on the phylogeny of 62 mammals including 5 marine mammals (Figure 1B): two cetacean species (bottlenose dolphin and killer whale), two pinnipeds (Weddell seal, walrus) and one siren (West Indian manatee). Several groups have studied convergence in protein-coding genes (Foote et al., 2015; Chikina et al., 2016), but few focused on non-coding regions. For this study, we used a set of 148,567 CNEEs extracted from the 100-way vertebrate alignment and phastCons conserved elements downloaded from the UCSC genome browser (see Methods).

Our simulation results based on both phylogenies demonstrate the ability of our method to identify elements conserved in most species in a phylogeny but accelerated in a target group of phenotypically convergent species. The ratite and marine mammal example studied here represent two extremes of possible topologies of convergent species on a phylogeny. In the ratite case, multiple independent losses are suggested by paraphyly of the target clade (inclusion of the volant tinamous within the ratites), and the convergent lineages are clustered in one region of the tree. By contrast, in the marine mammal case, the three independent transitions are widely separately on the phylogeny. We then apply our method to conserved non-coding elements identified from multiple alignments in both cases and show evidence of convergent evolution in CNEEs as well as functional enrichment of genes potentially regulated by these CNEEs.

Simulation study: avian topology

To verify our ability to detect the correct evolutionary pattern, we simulated DNA elements with different evolutionary patterns (i.e. different **Z**s) using a tree mirroring the inferred avian phylogeny. This scenario, in which convergent lineages are clustered in paraphyletic clade, is particularly challenging for existing methods. In our simulation, we set the length of element to be 200 bp which is the median length in our real data. We generated 9 scenarios with different levels of convergence either around ratites or other species: 1) all branches are conserved; 2) only kiwi clade accelerated; 3) only emu/cassowary branches accelerated; 4) only rhea clade accelerated; 5) only ostrich accelerated; 6) all ratites accelerated except ostrich and moa; 7) all ratites accelerated; 8) both ratites and volant tinamous accelerated; 9) 5 random species across all neognath birds accelerated (Supplementary Figure 2). Simulations 1), 8) and 9) are either negative or positive controls which should not be selected as ratite-specific accelerated elements whereas in all other cases one or more ratite lineages are accelerated. Since the volant tinamou clade resides within the ratite clade, making it difficult to distinguish elements accelerated from the ancestor of both tinamous and ratites from those only accelerated in ratites, we designed scenario 8) to demonstrate the specificity of our method. In each case, we simulated 500 elements with different conserved and accelerated rates.

Sensitivity and specificity of identifying accelerated elements in different scenarios

To test the sensitivity and specificity of our method, we mixed elements from (2)–(7) with background elements (1) separately and show receiver operating characteristic (ROC) curves for each experiment, where the area under the curve reflects the performance of each method to select ratite-specific accelerated elements from background conserved ones (Figure 2A). We labeled elements with $BF2 < 0$ as negative as it indicates species other than ratites might be accelerated, and then ranked other elements and plotted ROC curve by varying the threshold based on $BF1$. We

compared our method with phyloP in phast (denoted as LRT in the following), which tests for clade specific acceleration using a likelihood ratio test. For the LRT, we obtained the ROC curve using the test statistic output by phyloP. Not surprisingly, both methods achieve higher sensitivity as the number of accelerated ratite lineages increases. However, our method has consistently higher sensitivity in detecting elements accelerated among ratites based on BF1, which is much larger for ratite-specific accelerated elements than for conserved ones (Supplementary Figure 3). Thus, under a variety of evolutionary scenarios, including ones that are challenging for LRT methods (e.g., simulations 2 and 4), PhyloAcc has high power to detect lineage-specific rate shifts in conserved elements.

Second, our method has low false discovery rate. We mixed 100 elements from (2)–(9) together and with 5000 background conserved elements from (1), which imitates the small proportion of ratite-specific accelerated elements in real data. We then computed false discovery rate (FDR) from PhyloAcc using varying criteria based on Bayes factors, and compared this with the log-likelihood ratio statistic in phyloP (Figure 2B). The LRT fails to distinguish ratite-specific acceleration from the other scenarios, because FDR is still quite high, even though the likelihood ratio is large (cyan curve in Figure 2B); in contrast, when using BFs in PhyloAcc, the FDR drops below 5% at reasonable cutoffs (e.g., BF1 > 0, BF2 > 0). In real applications, the null distributions of BFs for elements not specifically accelerated in ratites are complicated and depend on the prior as well as the evolutionary pattern, making it difficult to specify a threshold of BFs controlling FDR. Nevertheless, we found that elements with $BF1 > 20$ and $BF2 > 0$ show strong evidence of ratite-specific acceleration in real data.

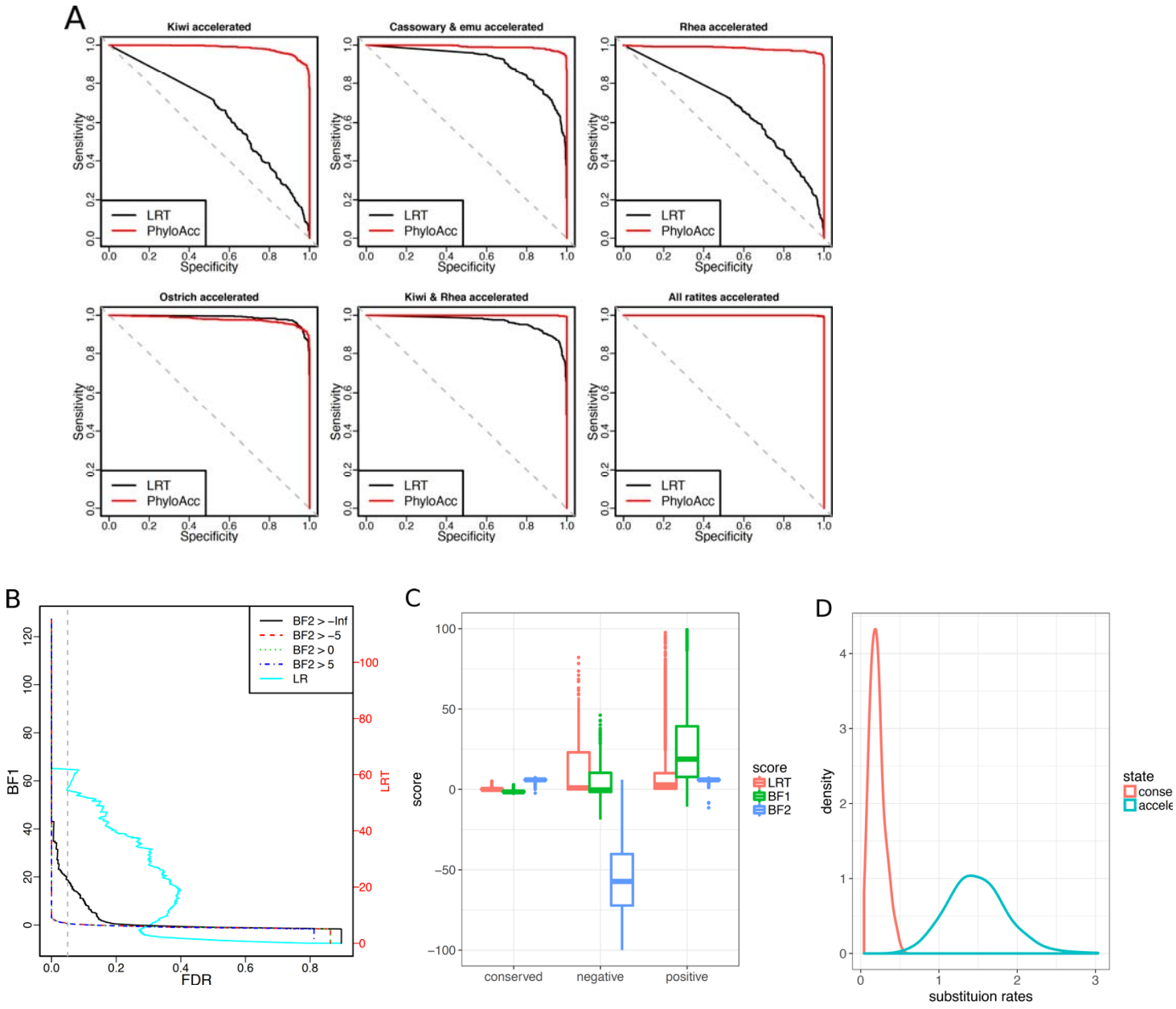


Figure 2: (a) ROC curve for PhyloAcc and LRT in different acceleration scenarios within ratite birds. We treated elements with each acceleration pattern (scenario (2)–(7) separately) as positive and all conserved elements (scenario (1)) as negative, and compared sensitivity and specificity of our method to phyloP. (b) FDR under different cutoffs of BF1 and BF2 for PhyloAcc (left axis) and log-likelihood ratio (LR) for phyloP (cyan curve, right axis); different cutoffs of BF2 are shown as different curves, and each curve represents FDR varying cutoffs of BF1. 5% FDR is shown as vertical gray line. (c) Boxplot of scores (BF1, BF2 for PhyloAcc and LRT for phyloP)

for conserved, negative (8-9: accelerated in non-ratite species) and positive (2-7: ratite-specific acceleration) scenarios. (d) shows the distribution of conserved and accelerated rate in the simulations.

The main reason for the superior performance of our method in terms of controlling false positive rate is that our method will not select elements with accelerated rates in species other than ratites (e.g. case (8), specifically in situations where acceleration occurs in ancestors of ratites and tinamous). phyloP, however, is not designed to control for this case, and will typically select these situations as false positives based on the log-likelihood ratio statistic, which can be even larger in (8) than in some positive cases (Figure 2C). In case (8), BF2 is less than -10 for 95% elements, because only the full model (M2), which allows arbitrary branches to experience rate shifts, fits the data adequately. Thus, almost all elements are labeled as negative (not ratite specific). By contrast, very few elements have BF2 less than 0 in cases (1)-(7), since the Bayes factors favor the simpler model if both models fit the data equally well (Supplementary Figure 3). Thus, our method achieves high specificity using BF2 as a filtering criterion.

At 5% FDR, phyloP could only identify few true ratite-specific accelerated elements since elements in scenario (8) have a larger likelihood ratio than ratite-specific accelerated elements, leading to low power. In contrast, PhyloAcc will not select elements accelerated outside of the target lineages, and successfully identified almost all the ratite-specific accelerated elements across all simulated substitution rate shift patterns at low false discovery rate (Supplementary Table 1). However, the alternative model in the LRT includes scenario (8), so it may be fairer to treat elements from (8) as positives for this method. The LRT method still loses power when only some ratite branches have accelerated (shown as LRT1 in Supplementary Table 1), especially when the accelerated branches are relatively short or the acceleration is recent. Overall, our method has lower false positive rate and higher power in identifying elements with substitution rate shift within a set of species, and thus is well suited as a screening tool for either shared or independent genetic changes.

Inferring the pattern of acceleration of individual genomic elements

Finally, we validated that our method can recover the true pattern of acceleration (latent states) for individual genomic elements. For each simulated element, we compared the posterior of **Z** output from our method with the true simulated pattern, and defined the result as "correct" if the posterior probability of the corresponding true latent state is greater than 0.7 for all branches. The accuracy of recovering the true substitution shift pattern increases with the difference between accelerated and conserved rates, and is mainly limited by lower accuracy on short branches. In our simulation, the ratio between accelerated and conserved rates is typically around $5 \sim 10$, and the accuracy is above 60% in all scenarios.

To investigate the impact of the ratio of accelerated and conserved rates on the ability of PhyloAcc to recover the true pattern of acceleration, we ordered and divided the simulated elements into 10 equal-sized groups according to the rate ratio (the quantiles of r_2/r_1 in each group are shown in Supplementary Figure 4A). Because phyloP cannot deduce the pattern of acceleration directly, we compare the performance with BEAST2 (Bouckaert et al., 2014). We defined that the detected pattern is correct if BEAST2 outputs the posterior probability of rate shift on the true state transition branches greater than 0.7 and less than 0.3 for others. Figure 3 shows the accuracy of our method compared to BEAST2 as well as the values of BF1 and BF2 for each group in various rate shifts scenarios. As might be expected, BF1 increases as the ratio increases and more species are simulated to have Z=2 (accelerated state). Moreover, in our positive control scenario (8) where acceleration is not specific to ratites, BF2 stays below zero and decreases as r_2/r_1 grows, since as this ratio increases the simulated (true) model diverges further from ratite-specific acceleration model. The accuracy of our method also increases as the rate ratio increases, since the conservation state of short internal branches is easier to determine when we observe more substitutions, which will tend to occur when accelerated rates are high relative to conserved rates.

BEAST2 has comparable accuracy in some cases, but performs worse in cases with multiple independent rate shifts (e.g. scenario (6), (7), (9)) or where rate shifts are confined to short branches (e.g., scenario (2)). The model implemented in BEAST2 allows transitions between conserved and accelerated rates in both directions. Under some circumstances in our simulations (e.g., when one clade originating from a common ancestor is accelerated), BEAST2 tends to place the origin of acceleration at a deeper node and then infer a regain of conservation in the conserved clade. To be more favorable to BEAST2, we still count this inference as correct even though it means that the conservation states of some internal branches are different from those simulated. Overall, PhyloAcc has higher accuracy for recovering convergent rate shifts in multiple lineages. We also measured "correctness" as the proportion of branches correctly labeled for each element (Supplementary Figure 4A). Our model can recover the true conservation state with high certainty (posterior of true latent state is around 1) for most branches, though accuracy is not as high for short, internal branches due to the limited number of informative sites (Supplementary Figure 6 and Supplementary Figure 4B).

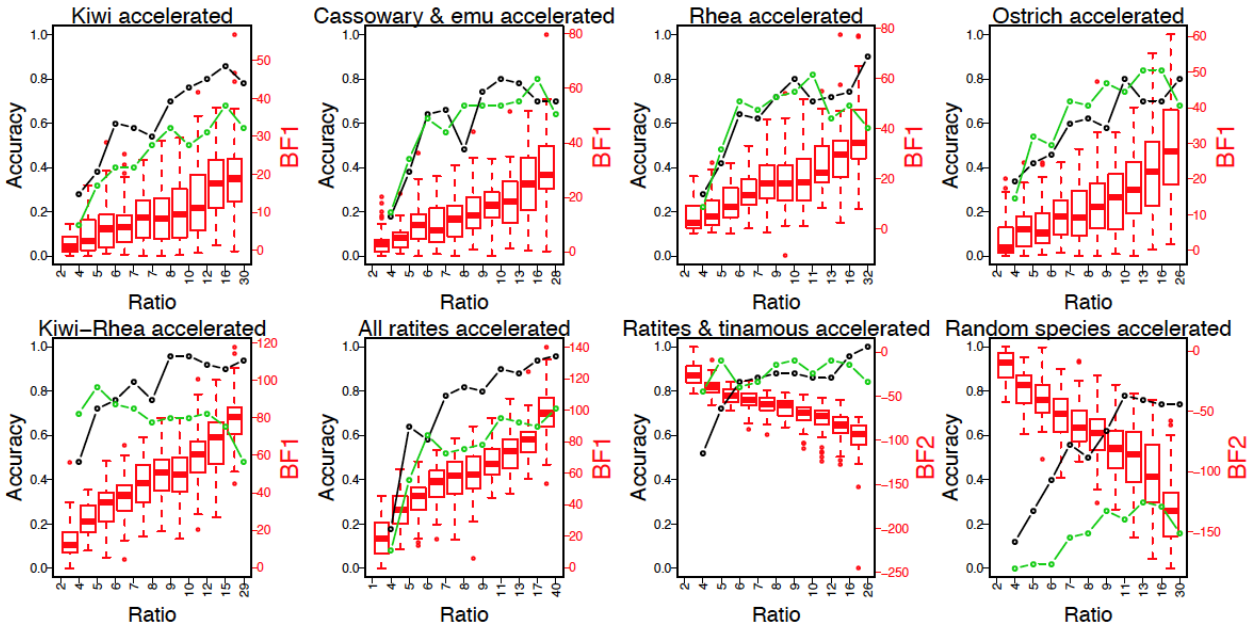


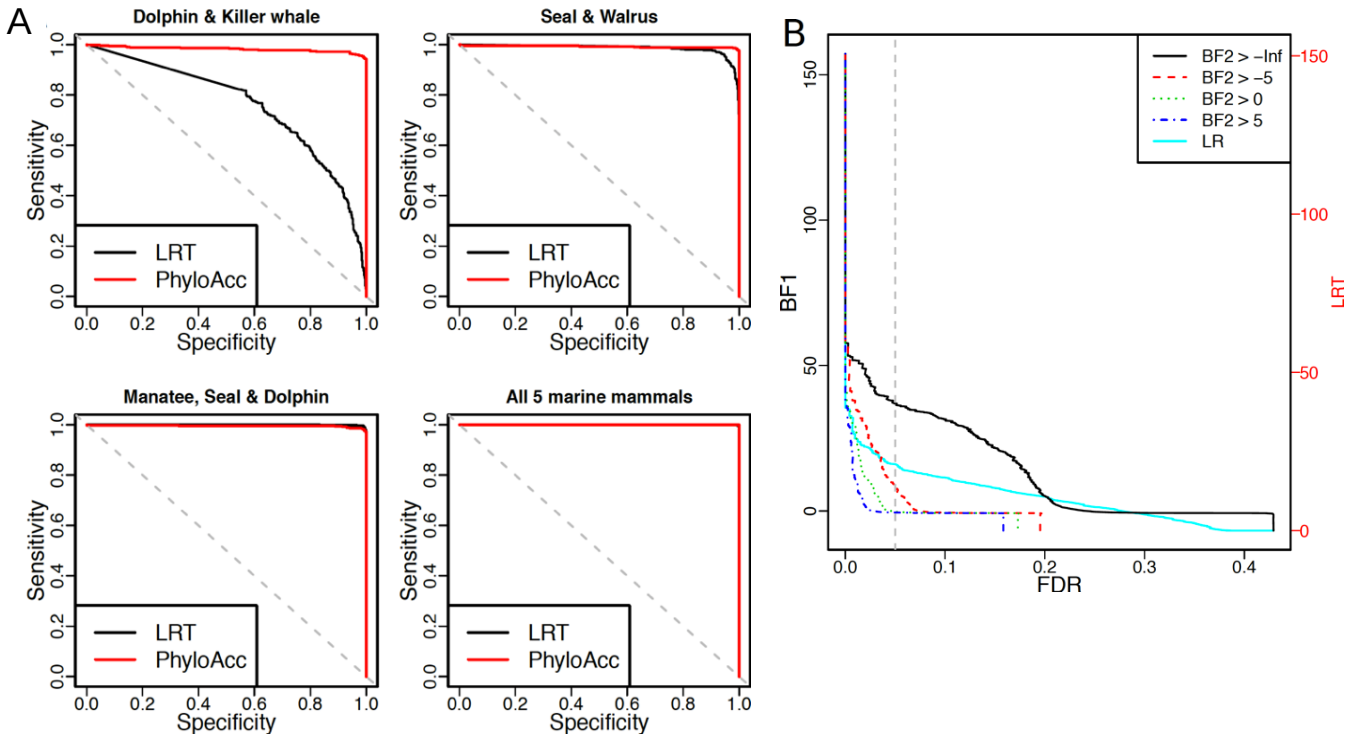
Figure 3: Comparison of accuracy of substitution rate shift patterns between BEAST2 and PhyloAcc in each simulation scenario. We divided the simulated elements into 10 equal-sized groups according to the ratio of accelerated and conserved rate for scenarios (2)-(9). X-axis shows the boundary of ratio in each group; black curve is the accuracy using PhyloAcc and green curve is for BEAST2, which is the proportion of simulated elements whose acceleration patterns are correctly detected by each method. Red box shows the quantiles of BF1 in (2)-(7), and the quantiles of BF2 in (8)-(9) for each rate ratio group.

We also examined the impact of indels on the performance of our method. Since it is hard to model indels explicitly, we sampled indels from their empirical joint distribution across species in the real avian data set and added them into the simulated multiple alignments (see Methods). We added different proportions of indels into the simulated data in approximately the observed range as found in the real data. The accuracy of identifying the shift patterns of substitution rates, as well as BF1, is relatively insensitive to indels (Supplementary Figure 5). In short, this simulation demonstrates that our method is capable of discovering multiple shifts of substitution rates.

Simulation study: mammalian phylogeny

We next sought to validate our method in a second simulation study, this time focusing on the common scenario where a convergent phenotype arises in multiple, distantly separately lineages on a phylogeny. We use the transition to marine habit in mammals as a model, simulated DNA elements based on the phylogeny of 62 mammalian species evolving under different patterns of substitution rates variation. We compared our method with phyloP in various scenarios: 1) all lineages conserved; 2) cetaceans (dolphin and killer whales) accelerated; 3) pinnipeds (seal and walrus) accelerated; 4) manatee, seal and dolphin accelerated, i.e. one species from each of the three independent lineages; 5) all 5 marine mammals accelerated; 6) pinnipeds and panda (sister lineage of pinnipeds) accelerated; 7) species descending from the common ancestor of cat and

pinnipeds (Supplementary Figure 7). Scenarios 2 through 5 are marine mammal-specific accelerated cases, whereas 6 and 7 are negative controls with non-specific acceleration. Our results show that our method has higher sensitive and specificity for identifying substitution rates accelerated exclusively in marine mammals than phyloP (Figure 4). PhyloAcc has higher sensitivity to detect genomic elements accelerated within marine mammals than phyloP (Fig. 4A), will exclude elements accelerated in species other than marine mammals by criterion on BF2, and the FDR drops below 5% when selecting elements with BF2>0 and BF1>0 (Fig. 4B and C). These results suggest that the sensitivity and specificity of PhlyoAcc is expected to be high under a range of evolutionary scenarios.



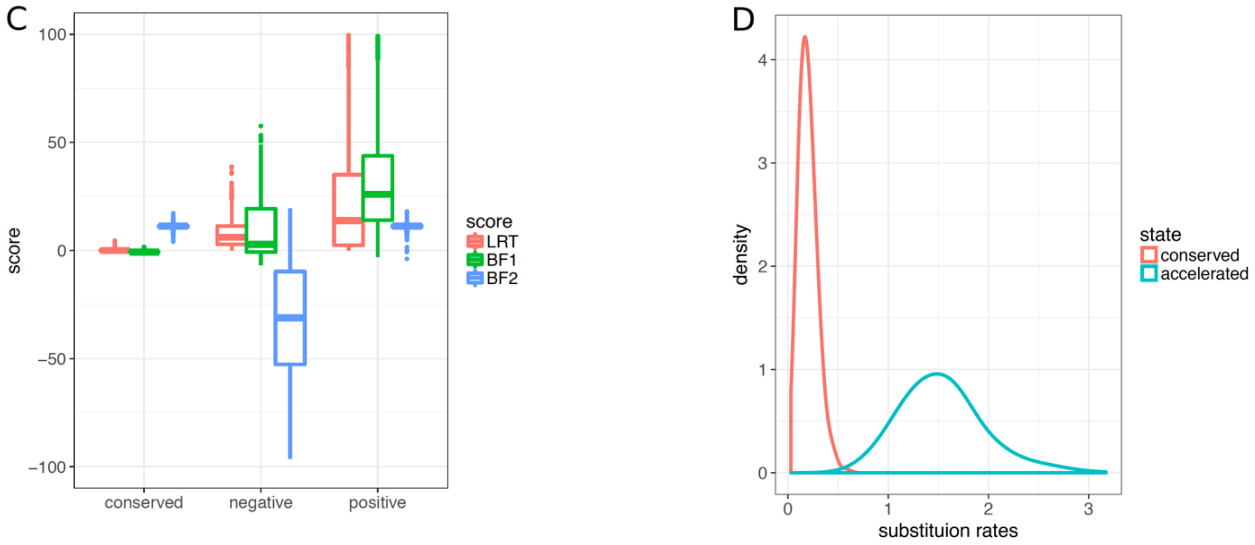


Figure 4: (a) ROC curve for PhyloAcc and LRT in different acceleration scenarios within marine mammal. We mixed elements with different patterns of acceleration with all conserved elements separately and compared sensitivity and specificity of our method to phyloP. Each figure shows ROC curve to identify acceleration happened in different lineages: (1) cetacean clade, (2) pinnipeds clade, (3) manatee, seal and dolphin, i.e. one species from three independent evolutionary origins, (4) all marine mammals. (b) FDR under different cutoffs of BF1 and BF2 for PhyloAcc (left axis) and log-likelihood ratio (LR) for phyloP (right axis); different cutoffs of BF2 are shown as different curves, and each curve represents FDR varying cutoffs of BF1. 5% FDR is shown as vertical gray line. FDR is computed by mixing elements simulated from all scenarios: positive samples are marine-specific acceleration; negative samples include conserved and other species accelerated elements. (c) Boxplot of scores (BF1, BF2 for our method and LRT for phyloP) for conserved, 2 negative (other species acceleration) and 4 positive (marine-specific acceleration) scenarios. (d) shows the distribution of conserved and accelerated rate in the simulations.

Detecting accelerated CNEEs in real data: avian case

We next applied our method to detect ratite-accelerated conserved non-coding regions based on a set of 284,001 CNEEs identified in birds (Sackton et al., 2018). Using PhyloAcc, we

identified 820 CNEEs with strong evidence for ratite-specific acceleration (BF1 > 20 and BF2 > 0). The rhea clade is the most likely to be accelerated among the 820 ratite-specific accelerated CNEEs, followed by kiwis, with the ostrich branch less likely to be accelerated among all ratites (Figure 5). The model outputs the posterior probability of the latent state of the substitution rate on each branch, which we used to infer how many ratites are accelerated for each element as well as how many independent accelerations occurred within ratites (see Methods). Many of these CNEEs have experienced multiple independent accelerations: 64 (8%) CNEEs have four or more expected independent losses; 234 (29%) have been accelerated 2-3 times; and 447 (58%) have been lost 1-2 times (Supplementary Table 2; Sackton et al., 2018).

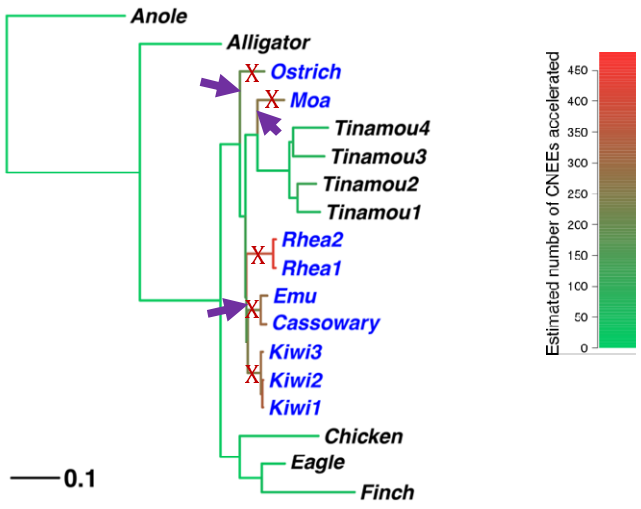


Figure 5: Number of CNEEs being accelerated per branch. Phylogeny for avian data set (only some species in neognathae and reptiles are shown for illustration). Palaeognaths consist of the flightless ratites and volant tinamous. Ratites are shown in blue. Potential losses of flight in ratites are shown as arrows or crosses: at least three independent losses (purple arrow), or five independent losses (red cross) are suggested by biogeographic history, or any pattern in between. The gradient of the color indicates expected number of elements being

accelerated on that branch among 820 ratite-specific accelerated CNEEs (from 0 to 482 grading from green to red).

Figure 6 shows the evolutionary patterns of some ratite-accelerated CNEEs with the largest Bayes factor (BF1). Although all of them show strong evidence of acceleration in ratites, they have different patterns of acceleration. A few of them are accelerated in a single species (e.g. mCE190953 accelerated only in cassowary, Figure 6D); many of them are accelerated in a single clade with one loss (e.g. mCE1389154 accelerated in both kiwis and emu/cassowary shown in Figure 6A; mCE1022564 accelerated only in rheas shown in Figure 6B; mCE600387 accelerated only in kiwis shown in Figure 6C); others are accelerated in more than one clades with multiple losses (e.g. mCE1217964 accelerated in both rheas and kiwis shown in Figure 6E and mCE114824 accelerated in both rheas and emu/cassowary shown in Figure 6F). Longer and redder branch indicates acceleration happened at earlier age while shorter and greener one means later or no acceleration. These are interesting candidate regulatory regions for further functional study (e.g. Sackton et al., 2018).

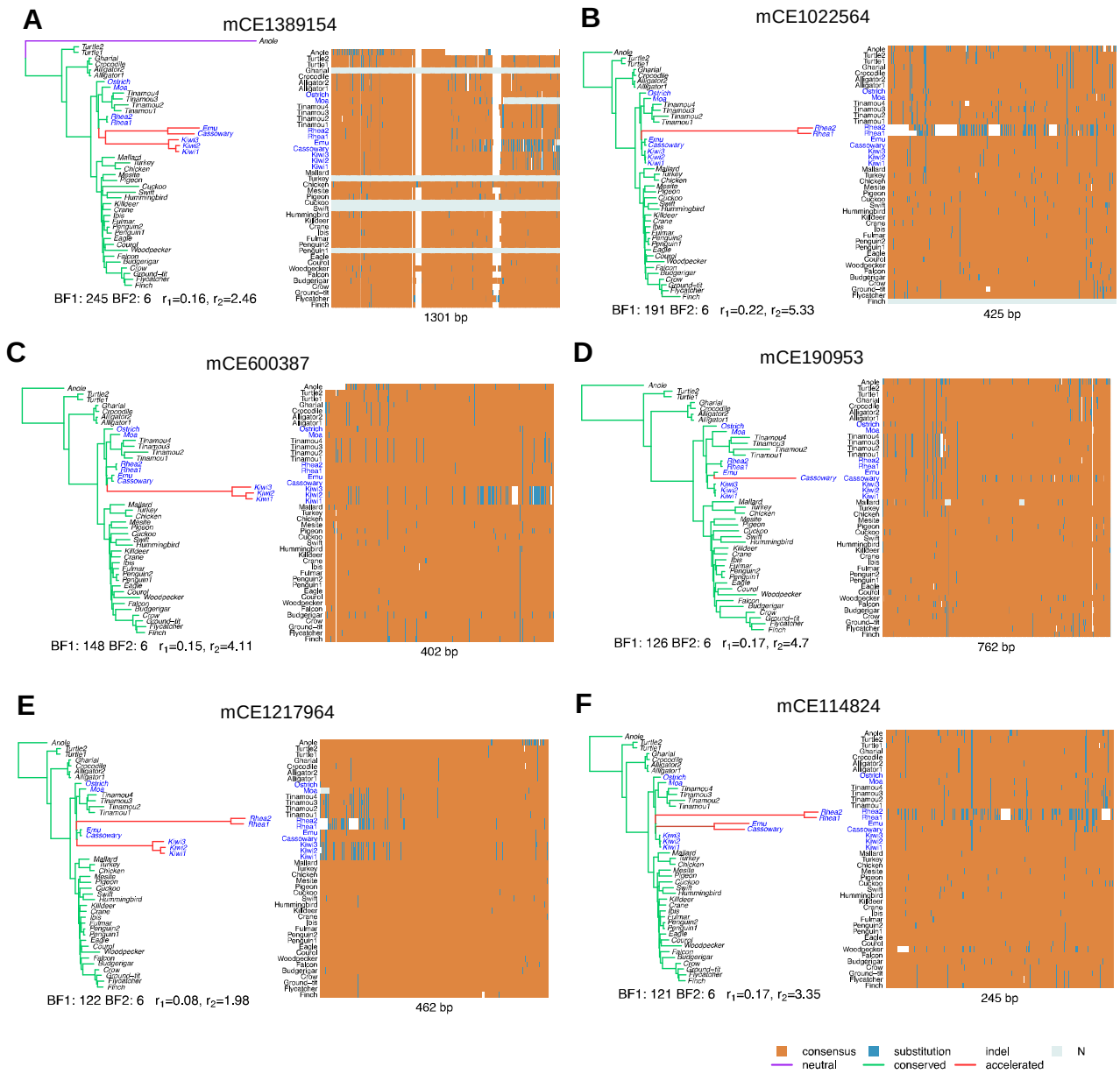


Figure 6: Examples of ratite-accelerated elements. For each element, shift pattern of substitution rates is shown on the left represented by a phylogenetic tree with branch lengths proportional to posterior substitution rate and colored by posterior mean of Z, along with the sequence alignments shown on the right. For the phylogenetic tree, green is conserved, red is the accelerated and purple is the neutral state. Below the tree shows two BFs plus conserved (r_1) and accelerated rate (r_2). In the sequence alignment heatmap, each column is one position, each row is a species and the element length is shown below. For each position, the majority nucleotide (T, C, G, A)

among all species is labeled as ''consensus'' and colored as orange; others are labeled as "substitution'' and colored as blue; unknown sequence is labeled as ''N'' and colored as gray; indels are shown as white space.

Detecting accelerated CNEEs in real data: mammalian case

As a second case study, we examined CNEEs accelerated in marine mammals. Though these mammals exhibit similar phenotypes upon transition to marine environment, the extent of molecular convergence in this system has been controversial, and largely focused on protein-coding genes. Some studies do not find significant evidence of convergent changes at specific amino acid sites beyond that what can explained by neutral evolution models (Foote et al., 2015); others claim convergence on the basis of shifts of substitution rates in protein-coding genes (Chikina et al., 2016). Among marine-accelerated protein-coding genes, Chikina et al. (2016) found evidence of adaptive evolution in skin and lung genes as well as loss of function in gustatory and olfactory genes. However, most of them are physiological and structural genes, with little evidence for convergent evolution in protein-coding genes controlling morphological adaptations, which may typically involve regulatory regions (Carroll, 2008).

We applied our method on 148,567 CNEEs identified from a whole genome alignment of 62 mammalian species, and identified 864 elements showing evidence of substitution rate shifts specifically in marine mammals. To test the hypothesis of convergent evolution in conserved noncoding regions underlying organism-level convergent phenotypes in marine mammals, we compared the number of parallel rate shifts in marine mammals with another group of mammals (aardvark, alpaca, camel, microbat, and David's myotis bat) with no obvious shared characters but which match the topology on the phylogeny with those five marine mammals (Figure 7A). More CNEEs show substitution rate shifts in marine mammals than in control species (2106 for marine-accelerated vs. 1472 for control-accelerated elements with BF1>5 and BF2 > 5). Furthermore, we found larger Bayes factor between the lineage-specific and null models for marine-accelerated

elements, indicating more dramatic changes of substitution rates affecting more species in marine mammals (Supplementary Figure 9). In addition, more marine-accelerated CNEEs show parallel shifts in multiple lineages than controls: 696 (33%) of marine-accelerated elements show acceleration in 3 lineages or more compared to 374 (25%) for control-accelerated elements (Supplementary Table 3); 93 (4.4%) of marine-accelerated elements show more than 2 independent losses compared to 33 (2.2%) for control-accelerated elements (Supplementary Table 4). To control for the artifact that marine-accelerated elements are generally accelerated in more species, we compared the number of non-specific accelerations (that is, the number of accelerated non-target species) in each marine-accelerated CNEE with controls and observed rare and fewer non-specific acceleration in marine-accelerated CNEEs (Supplementary Figure 9).

Finally, we tested for functional enrichment of genes near to marine-accelerated CNEEs in mammalian genomes using GREAT (McLean et al., 2010). Marine-accelerated CNEEs are predicted to regulate genes related to nervous and immune system including protein polyglutamylation, cerebellum morphogenesis, complement activation, and hindbrain morphogensis, etc.; these genes are also enriched in the corresponding mammalian phenotype terms such as olfactory bulb granule cell layer morphology, hippocampus layer morphology, and subplate morphology (Figure 7B). Many of the enriched functional terms are related to morphological traits, which reveals molecular adaptations overlooked by previous studies that focused primarily on protein-coding genes. Checking individual genes associated with these enriched functional annotations, we found that a handful of top marine-accelerated CNEEs are close to several genes, including TTLL3, a beta-tublin polyglutamylase modifying microtubules and highly expressed in nervous system (Ikegami et al., 2006); PROX1, a member of the homeobox transcription factor family, associated with cerebellum morphogenesis; C8B, one component of the membrane attack complex, and in the complement pathway as part of the body's immune response; DAB1, a key regulator of Reelin signaling pathway, playing an important role for neurogenesis; KLF7, a transcription factor, crucial for neuronal

morphogenesis in olfactory and visual systems, the cerebral cortex, and the hippocampus (Laub et al. 2005); FOXG1, a transcription repressor, essential for brain development, especially for the region controlling sensory perception, learning and memory (Martynoga et al. 2005); and GAS1 and GLI2, which function as transcription regulators in the hedgehog (Hh) pathway, important for embryogenesis (Martinelli and Fan, 2007).

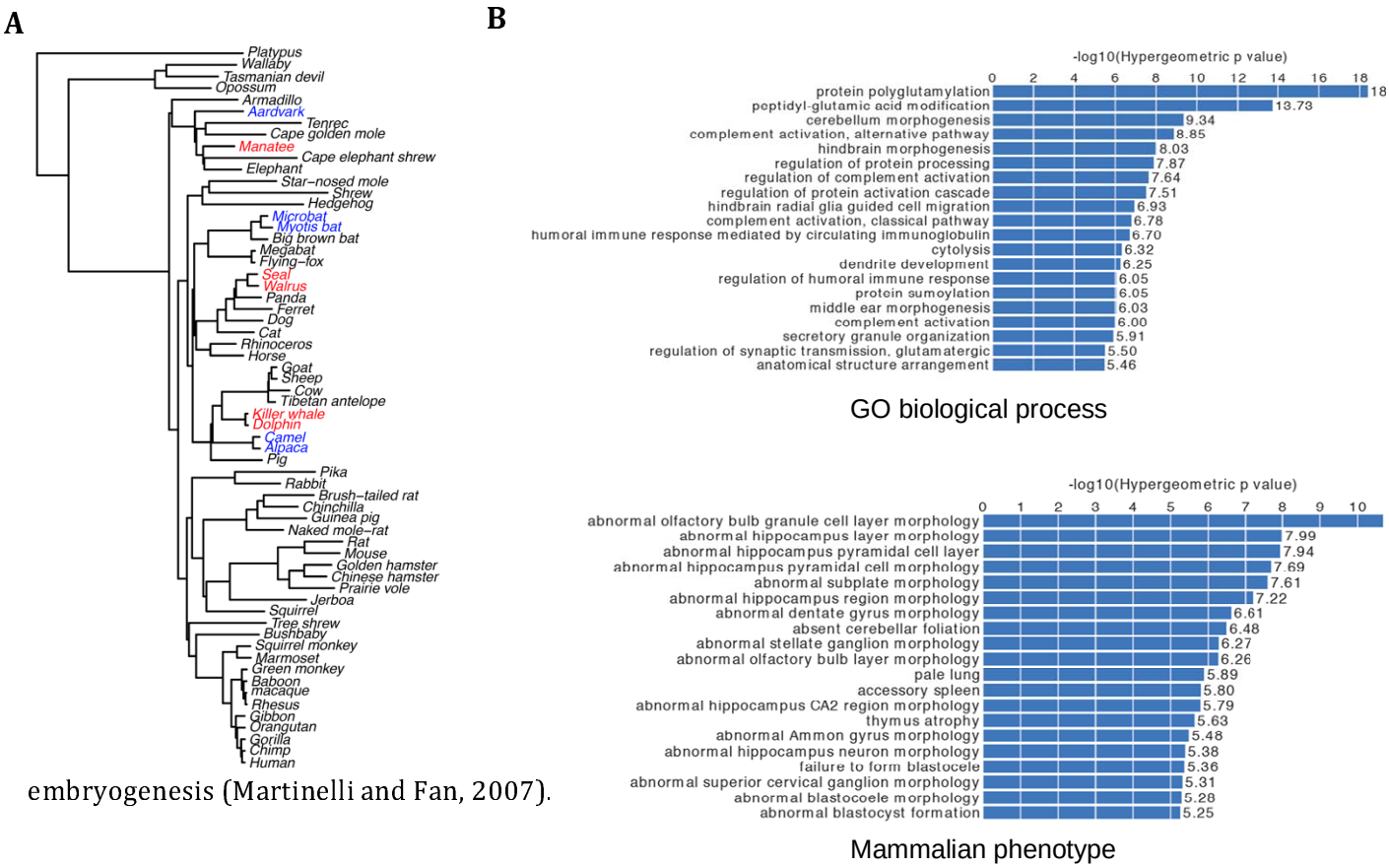


Figure 7: (a) Phylogeny of mammalian data. 5 marines are shown in red and 5 control species are shown in blue. (b) Enriched GO terms and mammalian phenotypes of genes near marine-accelerated CNEEs. Only shown top 20 terms (all of them with FDR <0.01).

In contrast, control-accelerated CNEEs are enriched in few very general GO terms, such as cell fate determination, regulation of transcription and translation (Supplementary Figure 10).

PhyloAcc software

We implemented our method in the program PhyloAcc, which was written in C++ and has been tested on Mac and Linux system. The package can be downloaded from https://github.com/xyz111131/PhyloAcc. It requires: 1) a phylogeny in .mod format (such as one produced by phyloFit in the PHAST package); 2) a multiple alignment file concatenating sequences of all input conserved elements in FASTA format; 3) a bed file with the position of each individual element in the coordinate of concatenated alignment file (0-based); 4) and a parameter file. The .mod file should contain the transition rate matrix Q and the phylogenetic tree in Newick format with branch lengths (in the unit of substitutions per site) under neutral evolution. The parameter file contains the paths for these input files and information of species and parameters for MCMC. PhyloAcc will output the posterior of latent conservation state (Z) for each branch, indicating neutral, conserved or accelerated states under the null, lineage-specific and full models, respectively, and the marginal log-likelihood under each model as well as Bayes factors for each element. Detailed description of usage is available in the GitHub repository. We also provide R scripts to generate figures summarizing the rate shift patterns as in this paper.

Discussion

Our method provides a flexible framework to detect substitution rate changes along phylogenetic trees based on multiple alignments of DNA sequences, conditional on annotated elements of interest (e.g. from PHAST or other tools). The method not only identifies DNA elements exhibiting changes of substitution rate in the lineages of interest, but also determines the branches, containing either single or multiple lineages, experiencing changes of substitution rate, all of which facilitate testing whether phenotypic convergence also involves convergence at the molecular level (e.g. Sackton et al., 2018). We show here that PhyloAcc, our new Bayesian method, outperforms existing methods in simulations. Application to two biological datasets (convergent loss of flight in

ratites and convergent shifts to marine habitat in mammals) revealed a number of noncoding elements accelerated independently on multiple target, phenotypically convergent, lineages, suggesting that molecular convergence in regulatory regions may be commonly associated with phenotypic convergence.

The idea of matching sequence divergence profile of either protein-coding genes or non-coding regions with repeated losses or gains of a given trait in multiple independent lineages to gain insight into the molecular basis of phenotype differences was first proposed as "Forward Genomics" by Hiller and his colleagues (Hiller et al, 2012b). Since then, this approach has been used in various groups of organisms, often yielding important insights into genome evolution (Prudent et al., 2016; Chikina et al., 2016; Partha et al., 2017; Berger et al., 2017; Roscito et al., 2017). Comparing with previous methods, our method can distinguish genomic elements with multiple independent accelerations within phenotypically convergent species from a single strong acceleration in a larger clade. Our method also achieves a lower rate of false positives by comparing the marginal likelihoods of models either allowing or prohibiting acceleration in species without phenotype change. Moreover, using Bayes factors to specifically identify accelerated elements does not rely on the asymptotic normal assumption required for the likelihood ratio test, and therefore is more robust for some extreme cases, such as when only a few branches have many substitutions. In such cases, maximum likelihood function is unbounded and the test may not apply appropriately.

The core utility of our software is its ability to detect a change of substitution rate of a large number of elements on a tree, yielding the posterior of r_2 for each element on each branch, from which the direction of rate change can be inferred. To identify elements with signatures of acceleration, our software provides an option restricting r_2 to be greater than some threshold (e.g. 1). The prior of r_2 could also be adjusted in our software to identify elements more conserved in a group of species than others. Currently, we assume the same substitution rate for all accelerated

branches, although our model could be extended to allow for different acceleration rates for each independently evolving clade. These acceleration rates could either take several discrete values (Supplementary material) or generate from a common distribution for each element. By introducing additional latent states, this extension could also allow for models distinguishing simple loss of conservation from acceleration due to natural selection. In addition, via Dollo's assumption of irreversibility, our model allows at most two shifts on the tree for each lineage on the phylogeny, which may not be efficient for detecting elements that regain conservation after an ancient episode of adaptation. For example, we could adjust the transition probability matrix of conservation states (Z) to allow for a small probability of transition from accelerated to conserved state. However, in many scenarios, the parsimony assumption is helpful, since the sequence data of extant species often does not provide enough information to distinguish multiple rate changes with opposite directions from no change of substitution rates, as illustrated in the simulation section when comparing with BEAST2. Additionally, the marginal likelihood is harder to compute for more complex models.

The local substitution rate along the phylogenetic tree is likely not constant under neutral evolution across different regions of the genome, a pattern that may impact our method (and all previous methods). One way to tackle this genome-wide rate heterogeneity is to estimate substitution rates and branch lengths on the phylogenetic tree under neutral evolution for different segments of the genome. However, this may introduce a degree of arbitrariness in the decision as to how to segment the genome. Additionally, we suspect that genome-wide variation in the local neutral rate is not a serious issue for our model, because our model already includes variation in the conserved substitution rate across elements and define acceleration relative to other branches at the same genomic locus. Though the neutral rate is constant in our model, in our examples, only a few outgroup lineages are in typically in the neutral state, so the actual value of the neutral rate has relatively little direct impact on the estimated non-neutral rates.

Another major issue not addressed by our model is the potential for heterogeneity in the topologies of gene trees across elements and across the genome. Heterogeneity in the topology of gene trees is expected to occur, especially in rapid radiations (Edwards 2009). But it is also the case that mis-specifying the phylogenetic tree on which parameters are estimated can lead to mis-estimation of substitution rates (Hahn and Nakhleh 2016; Mendes and Hahn 2016). To account for phylogenetic uncertainty (due to gene tree error or incomplete lineage sorting), we could average the likelihood over all probable gene trees, then compare the averaged likelihood under each model.

Although in our examples we focus on loss of conservation accompanied by faster substitution rates, we do not attempt to distinguish among different types of mutation or selection that can produce a specific pattern. GC-biased gene conversion (gBGC) is one of the factors that can increase substitution rates in local regions of the genome, and is often a confounding factor to detect adaptive selection (Kostka et al., 2012; Capra et al., 2013b). We observed that ratite-accelerated elements have a higher GC content in ratites (Supplementary Figure 8), which suggests a role for gBGC in acceleration. To demonstrate an approach for accounting for gBGC, we extended our method to jointly model gBGC and selection effect on substitution rates. To do this, we reparametrized our substitution rates in terms of a selection coefficient and gene conversion disparity (Kostka et al., 2012), used another indicator for gBGC on each branch (Supplementary material) and compared the marginal likelihood under lineage-specific model with null model after taking account of gBGC in both models. We found that ~30% ratite-accelerated elements also exhibited evidence for gBGC in accelerated lineages, indicating that gBGC is likely a major force prompting loss of conservation for DNA elements released from purifying selection. gBGC can be intertwined with adaptive selection; for example, many human-accelerated regions are partly caused by gBGC, which may or may not be positively selected to increase fitness (Pollard et al., 2006). Although we did not distinguish gBGC from other molecular or selective mechanisms

associated with accumulating substitutions and loss of function here, PhyloAcc also provides an extended version available online which could distinguish the effect of gBGC from selection.

Identifying the functions of regulatory regions is still a challenging task and linking patterns of sequence evolution from diverse species with organism-level phenotypes has the potential to shed light upon regulatory function of conserved non-coding regions. Our method could be extended to provide the probability of a match between evolutionary profile of genetic elements with presence/absent patterns of hundreds of traits to predict phenotype-genotype pairs, an extension for the "Reverse Genomics" approach (Marcovitz et al., 2016). By using parsimony to reconstruct traits and genome transitions, the method of Marcovitz et al. (2016) does not consider the uncertainty of patterns of conservation estimated from the sequencing data or the probability of a chance match between genome pattern and phenotype. The reverse genomics approach also does not provide a straightforward way to incorporate missing phenotype/genotype information without specifying a probability model. In addition, to identify links between genotype and phenotype, our model can be extended to cluster genomic regions (e.g. using Dirichlet process as prior for Z) based on similar patterns in sequences to discover novel functional group of genomic loci that may or may not influence known physiological and morphological traits. Jointly modeling a group of functional related genomic regions in different species will give a more comprehensive and deeper insight of evolution history and functional interaction of regulatory regions (Marcovitz et al., 2017).

Methods

Data sources for bird and mammal CNEEs

We obtained a whole genome alignment of 42 species (birds and non-avian reptiles) for ratite-accelerated region detection from Sackton et al. (2018). The conserved regions in the genome

alignment were called by PhastCons using the Phast package; 284,001 CNEEs were extracted as DNA regions not overlapping with any exons and at least 50 bp in length. Sequence from the extinct moa was subsequently added to CNEE alignments based on a pairwise moa-emu whole genome alignment (see Sackton et al. 2018 for details). For the mammalian dataset, we started with the UCSC 100-way vertebrate alignment (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/), removed all non-mammalian sequences, and then extracted sequence for 383,185 CNEEs in a fashion similar to that for birds (conserved regions identified by PHAST, at least 50 bp, not overlapping any exons). The list of species is in Supplementary material. We filtered out CNEEs with poor alignment quality in 62 mammal species: elements with alignment gaps longer than 80% total length of the element in more than 50 species, and were left with 148,567 candidate CNEEs. For both phylogenies, we also obtained neutral models from phyloFit (Sackton et al. 2018 or UCSC, respectively).

Bayesian Model for PhyloAcc

Suppose $\boldsymbol{X}_i \in \{\mathrm{A, C, G, T}\}$ is a $l \times n$ matrix of the sequence alignments of element i, where l is the length of the element, $n = 2S - 1$ is the total number of nodes in the tree and S is the number of existing species at the leaves of the tree, whose sequence can be observed; $\mathbf{T} = \{t_1, \ldots, t_n\}$ is the phylogenetic tree (including topology and t_s is the branch length (in unit of substitutions per site) for the branch directly towards node s). Standard DNA substitution models are used so that substitution on one branch follows a continuous time Markov Process with rate matrix Q and stationary distribution π, in which case transition probability matrix along a branch with length t and relative substitution rate r is e^{rtQ}. We estimated the Q matrix and branch length from neutral sequence (fourfold degenerate sites) precisely using phyloFit in the phast package. Because the total length of neutral sequence used to estimate branch lengths is long, we neglect uncertainty here and treat these as fixed parameters in the model. The latent conservation state for branch s,

Z_{is} can be 0, 1, or 2 corresponding to neutral, conserved or accelerated states with relative substitution rates $\mathbf{r_i} = (r_0 = 1, r_{i1}, r_{i2})$ respectively for element i. The prior of r_{i1} and r_{i2} follows a Gamma distribution with hyperparameter (a_c, b_c) and (a_n, b_n) respectively, which controls the degree of pooling across all elements. In our software, we provide a full Bayesian method by assuming a hyperprior for these hyperparameters and also an empirical Bayes method by using their estimates from data (Supplementary Material). The former is computationally demanding, but the latter approach performs well in practice. In this paper, we illustrated our method using the second approach. With reasonable hyperparameters, our prior of substitution rates encourages r_{i2} for accelerated states to be larger than r_{i1} for conserved states.

The transition probability matrix of $\mathbf{Z}s$ is denoted by Φ. Φ can take any form but by Dollo's irreversible evolution hypothesis, we assume that once an element is accelerated on a particular branch, the downstream branches cannot regain conservation. Therefore, the transition matrix has a simplified form: $\Phi = \begin{bmatrix} 1-\alpha & \alpha & 0 \\ 0 & 1-\beta & \beta \\ 0 & 0 & 1 \end{bmatrix}$. This assumption still allows independent gain and loss of conservation on multiple lineages. We assume that the element is neutral at the root of the phylogeny, then becomes conserved (usually in one but potentially in more than one lineage), and finally may lose conservation in some lineages. Each element might not have all three states. For our data, most of the elements are conserved in all species and only a few are accelerated in some clades, so we assume $\alpha = P_{0\to1}$ is large while $\beta = P_{1\to2}$ is small. In our model, $\boldsymbol{X}_i$ are partially observed (only the sequences of extant species are observed, denoted by $\boldsymbol{Y}_i$) and $\boldsymbol{Z}_i$is unobserved but may have some biological constraints (denoted by C, e.g. we constrained acceleration only happened in some species in lineage-specific model) as prior information, which makes our inference harder. The joint distribution of our model is (to ease notation, we omit the subscript i below):

$$P(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{r} | \boldsymbol{T}, Q, \Phi, C, a_c, b_c, a_n, b_n) = \prod_{j=1}^{l} P\left(\boldsymbol{X_j} \,|\boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{r}, Q\right) \cdot P(\boldsymbol{Z}|\boldsymbol{T}, \Phi, \mathrm{C}) \cdot P(\boldsymbol{r}|a_c, b_c, a_n, b_n)$$

$$= \prod_{j=1}^{l} \prod_{s=1}^{n} P\left(X_{js} \,|\boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{r}, X_{j,pa(s)}, Q\right) \cdot \frac{1}{\eta_C} \prod_{s=1}^{n} P\left(Z_s | \boldsymbol{T}, \Phi, Z_{pa(s)}\right) I(Z \in C) \cdot \mathrm{Gam}(r_1 | a_c, b_c)$$

$$\cdot \mathrm{Gam}(r_2 | a_n, b_n)$$

$$= \prod_{j=1}^{l} \prod_{s=1}^{n-1} \left(P e^{r_{Z_s} t_s \Lambda} P^{-1}\right)_{X_{j,pa(s)} X_{js}} \pi\left(X_{jn}\right) \cdot \frac{1}{\eta_C} \prod_{s=1}^{n-1} \Phi_{Z_{pa(s)} Z_s} I(Z \in C) \cdot \mathrm{Gam}(r_1 | a_c, b_c)$$

$$\cdot \mathrm{Gam}(r_2 | a_n, b_n) \quad (1)$$

η_C is the unknown normalizing constant depending on constraints on configurations of $\mathbf{Z}$. By our assumption, for root n, $Z_n = 0$ and $X_{jn} \sim \pi$ is stationary.

The posterior of $\mathbf{Z}$ indicates the change points of the substitution rate on the tree. Alignment gaps, small indels and unknown base pairs complicate the probability model. We have some heuristics dealing with them and more sophisticated modeling of them is out of scope of this paper. If the sequence alignment of a species is occupied by gaps in e.g. $\geq 80\%$, of the total length of an element, it's unlikely for the element to be conserved in that species. Thus, we assign small probability (e.g. 0.01) of observing long alignment gaps given $Z_s = 1$. In other cases, we integrate out all possible values of X_s for small indels or unknown base pair in species s (Siepel et al. 2005).

From the posterior of $\mathbf{Z}$, we compute the posterior probability of being accelerated on each branch by $P(\, Z_s = 2 | \boldsymbol{Y})$ and the posterior probability of loss conservation on each branch by $P\left(Z_{pa(s)} = 1, Z_s = 2 \middle| \boldsymbol{Y}\right) = P(\, Z_s = 2 | \boldsymbol{Y}) - P\left(\, Z_s = 2, Z_{pa(s)} = 2 \middle| \boldsymbol{Y}\right) =$ $P(\, Z_s = 2 | \boldsymbol{Y}) - P\left(\, Z_{pa(s)} = 2 \middle| \boldsymbol{Y}\right)$ (the second equality is because once accelerated it will remain in the accelerated state). Then the expected number of accelerated species (N_1) within phenotypically convergent species (S_0) is: $\mathrm{EN}_1 = \sum_{s \in S_0} P(Z_s = 2 | \boldsymbol{Y})$ and the expected number of independent

losses of constraint (N_2) within ancestors of S_0 is the sum of posterior probability of loss on each branch towards S_0:

$$\mathrm{EN}_2 = \sum_{s \in S_1} P(Z_{pa(s)} = 1, Z_s = 2|\boldsymbol{Y}) = \sum_{s \in S_1} P(\, Z_s = 2|\boldsymbol{Y}) - P\big(\, Z_{pa(s)} = 2\big|\boldsymbol{Y}\big)$$

S_1 includes S_0 and their common ancestors.

MCMC algorithm for Bayesian inference

Since the posterior is difficult to compute, we use collapsed Gibbs to do inference. We iteratively update DNA sequences of ancestral species $\mathbf{H} := \mathbf{X}_{S+1..n}$ ($\mathbf{Y} := \mathbf{X}_{1..S}$ is the observed DNA sequences), latent states $\mathbf{Z}$ and substitution rates $\mathbf{r}$ for each element by Gibbs sampling.

Sample H: Sampling $\mathbf{H}$ given $\mathbf{Z}$ and other parameters can be done efficiently by forward-backward sampling, a common algorithm for state-space models. Given $\mathbf{Z}$ and $\mathbf{r}$, each site j in an element is independent. Thus, we iterate over every site to sample the unobserved sequences of ancestors.

Sample Z: Sampling $\mathbf{Z}$ given $\mathbf{H}$ and other parameters is also straightforward by forward-backward sampling. Since each site of an element shares the same $\mathbf{Z}$, conditional distribution of $\mathbf{Z}$ depends on the entire sequence of that element. Given sequences and conservation state, i.e. $(\mathbf{H}, \boldsymbol{Z})$, the probability of substitution on each branch can be easily computed by standard DNA substitution models. Posterior sampling of $\mathbf{Z}$ with biological constraints on $\mathbf{Z}$ is similar if we treat these constraints as another kind of observation of $\mathbf{Z}$ (Supplementary Material).

Sample r: The conditional distribution of $\mathbf{r}$ on $\mathbf{Z}$ (integrating out $\mathbf{H}$) cannot be directly sampled, thus we use adaptive Metropolis-Hasting algorithm within our Gibbs sampling scheme. We use a gamma distribution centered around the current r as proposal and adaptively tune the variance of proposal distribution based on the acceptance rate of previous MCMC steps.

Computing Bayes Factors for model comparison

To select elements with a specific evolutionary pattern of constraint and acceleration, we compare the probability of sequence alignments of extant species under the null (M_0), lineage-specific (M_1) and full (M_2) models, allowing for increasing configurations of $\mathbf{Z}$. In order to compute Bayes factors, the problem becomes how to compute marginal probability under different constraints of $\mathbf{Z}$. Because it is not possible to sum over all possible configurations of $\mathbf{Z}$, Chib's method (Chib 1995) is commonly used to compute the marginal probability using the ratio of joint and conditional probabilities of one configuration of $\mathbf{Z}$ (fixed parameters in the conditional probability omitted in the equations below):

$$P(Y|C) = \frac{P(Y, \mathbf{Z}^*|C)}{P(\mathbf{Z}^*|Y, C)} = \frac{\int P(Y, \mathbf{Z}^*|C, \boldsymbol{r})P(\boldsymbol{r})d\boldsymbol{r}}{P(\mathbf{Z}^*|Y, C)}$$
$$\log P(Y|C) = \log P(Y, \mathbf{Z}^*|C) - \log P(\mathbf{Z}^*|Y, C)$$

Here C represents constraints under different models. Any Z^* is valid although usually it is taken as the posterior mode or MLE. However, both the numerator and denominator cannot be computed analytically in our case. We use the posterior distribution of Z yielded by MCMC to approximate the denominator and construct an upper bound for the numerator similar to the variational method (Blei et al., 2016). To reduce the variance of the estimator of $P(Y|C)$, we extended the previous method to be a weighted average over each individual Chib's estimator based on configurations of $\mathbf{Z}$ with high posterior probability $P(\mathbf{Z}|Y, C)$:

$$\log P(Y|C) = \sum_{\mathbf{Z}} (\log \hat{P}(Y, \mathbf{Z}|C) - \log \hat{P}(\mathbf{Z}|Y, C)) * \hat{P}(\mathbf{Z}|Y, C)$$

$\hat{P}(\mathbf{Z}|Y, C)$ is the empirical posterior distribution. Because $\mathbf{Z}$ is given, sequence alignments Y and constraints C are independent, and the joint log-likelihood can be written as: $\log P(Y, \mathbf{Z}|C) = \log P(Y|\mathbf{Z}) + \log P(\mathbf{Z}|C)$. We calculate the second term by the forward-backward procedure, but we

cannot compute $\log P(Y|\mathbf{Z})$ directly because it involves integration over r s. However, we provide an upper bound for it using:

$$\log P(Y|\mathbf{Z}) = \int dP(\boldsymbol{r}|Y,\boldsymbol{Z}) \log P(Y,\boldsymbol{r}|\boldsymbol{Z}) - \int dP(\boldsymbol{r}|Y,\boldsymbol{Z}) \log P(\boldsymbol{r}|Y,\boldsymbol{Z})$$

$$\leq \int dP(\boldsymbol{r}|Y,\boldsymbol{Z}) \log P(Y,\boldsymbol{r}|\boldsymbol{Z}) - \int dP(\boldsymbol{r}|Y,\boldsymbol{Z}) \log q(\boldsymbol{r}|\gamma) \quad (2)$$

$q(\boldsymbol{r}|\gamma)$ can be any distribution and the approximation error of inequality (2) is the KL divergence between posterior distribution of r and $q(\boldsymbol{r}|\gamma)$: $KL(P(\boldsymbol{r}|Y,\mathbf{Z})||q(\boldsymbol{r}|\gamma))$. In order to get a tighter bound, we could find the optimal value of $\hat{\gamma}$ to minimize the KL divergence. $q(\boldsymbol{r}|\gamma)$ is usually taken to be the distribution family where $\hat{\gamma}$ is easy to compute. We let $q(\boldsymbol{r}|\gamma)$ conform to a multivariate Gaussian distribution; then $\hat{\gamma}$ are mean and covariance matrix of the posterior sampling of $\mathbf{r}$ given $\mathbf{Z}$. All the thresholds for Bayes factors are presented on a log-scale in the result section.

Simulating DNA sequences

We simulated DNA sequences according to the joint likelihood in equation (1) using the same phylogenetic tree and estimated rate matrix Q from sequence alignments either as in the avian or the mammalian data set using our in-house program. For ratite simulation, we simulated 500 elements in scenarios (2)-(8) and 5000 elements in scenario (1) with length 200bp under different configurations of $\mathbf{Z}$; for mammal simulation, we simulated 500 elements (200bp) for each scenario. The conserved rate r_1 was sampled from $Gamma(5,0.04)$ and the accelerated rate r_2 was sampled from $Gamma(15,0.1)$, which are about the range of conserved and accelerated rates from real data. To simulate indels, we uniformly sampled some number of sites from the simulated sequences (i.e. 30%, 50% and 70% of the total simulated data), extracted randomly the same number of loci from all positions with at least one indel across all species in the multiple alignments of the avian data set,

and removed the nucleotides at which deletions occur in the subsampled real sequence alignments. In the real data, about 60% of the loci contain at least one indel, so the proportion of indels in our simulated data is about the same scale as the real data.

Function Prediction of CNEEs using GREAT

To predict the regulatory function of CNEEs in mammalian data set, we first extracted the genomic coordinates of these CNEEs using human (hg19) genome as reference. To associated CNEEs with nearby genes, we used the "Basal plus extension" (up to 500Kb) option in GREAT. Then, we compared genes associated with marine- or control- accelerated CNEEs to genes near all CNEEs (background), and searched for any functional enrichment in GO biological processes and mammalian phenotypes from MGI. We only retained annotation terms containing more than 5 genes in total, including at least 2 genes associated with accelerated CNEEs and at least 1.5-fold enrichment of tested CNEEs over all CNEEs.

Data access

Installation instruction, documentation, as well as example simulation data sets and results are available at https://github.com/xyz111131/PhyloAcc.

Acknowledgements

We thank Professor Ziheng Yang for valuable comments on the manuscript, members in Professor Edwards's lab, especially Alison Cloutier, and Shaoyang Ning and other members in Jun's lab for kindly discussion. SVE and TBS were supported by NSF grant DEB-1355343/EAR-1355292 to SVE and Julia Clarke. The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

Disclosure declaration

The authors declare no competing financial interest.

References

Baker AJ, Haddrath O, Mcpherson JD, Cloutier A. 2014. Genomic support for a moa-tinamou clade and adaptive morphological convergence in flightless ratites. *Mol. Biol.Evol.* **31:** 1686–1696.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321–1325.

Berger MJ, Wenger AM, Guturu H, Bejerano G. 2017. Independent erosion of conserved transcription factor binding sites points to shared hindlimb, vision, and scrotum loss in different mammals. *bioRxiv* doi: https://doi.org/10.1101/197756

Blei DM, Kucukelbir A, McAuliffe JD. 2016. Variational Inference: A Review for Statisticians. doi:10.1080/01621459.2017.1285773

Booker BM, Friedrich T, Mason MK, VanderMeer JE, Zhao J, Eckalbar WL, Logan M, Illing N, Pollard KS, Ahituv N. 2016. Bat Accelerated Regions Identify a Bat Forelimb Specific Enhancer in the HoxD Locus. *PLoS Genet* **12**(3): e1005738. https://doi.org/10.1371/journal.pgen.1005738

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol* **10**.

Brawand D, Wahli W, and Kaessmann H. 2008. Loss of egg yolk genes in mammals and the origin of lactation and placentation. *PLoS Biol.* **6**: 0507-0517.

Capra JA, Erwin GD, McKinsey G, Rubenstein JLR, Pollard KS. 2013a. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc B Biol Sci* **368**: 20130025–20130025. http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2013.0025.

Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. 2013b. A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. *PLoS Genet* **9**.

Carroll SB. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134**: 25–36.

Chib S. 1995. Marginal Likelihood from the Gibbs Output. *J. Am. Stat. Assoc.* **90**: 1313.

Chikina M, Robinson JD, and Clark NL. 2016. Hundreds of Genes Experienced Convergent Shifts in Selective Pressure in Marine Mammals. *Mol. Biol. Evol.* **33**: 2182-2192.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**: 699–710.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biol* **8**: 114. http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-8-114.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* **63**:1-19.

Foote AD, Liu Y, Thomas GWC, Vina T, Alfoldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, et al. 2015. Convergent evolution of the genomes of marine mammals. *Nat. Genet.* **47**: 272-275.

Gould SJ. 1970. Dollo on Dollo ' s Law: Irreversibility and the Status of Evolutionary Laws. *J Hist Biol.* **3**: 189-212.

Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees. *Evolution* **70**:7-17.

Harshman J, Braun EL, Braun MJ, Huddleston CJ, Bowie RCK, Chojnowski JL, Hackett SJ, Han K-L, Kimball RT, Marks BD, et al. 2008. Phylogenomic evidence for multiple losses of flight in ratite birds. *Proc Natl Acad Sci* **105**: 13462–13467. http://www.pnas.org/cgi/doi/10.1073/pnas.0803242105.

Heath TA, Holder MT, Huelsenbeck JP. 2012. A dirichlet process prior for estimating lineage-specific substitution rates. *Mol Biol Evol* **29**: 939–955.

Hiller M, Schaar BT, and Bejerano G. 2012a. Hundreds of conserved non-coding genomic regions are independently lost in mammals. *Nucleic Acids Res.* **40**: 11463-11476.

Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, and Bejerano G. 2012b. A "Forward Genomics" Approach Links Genotype to Phenotype using Independent Phenotypic Losses among Related Species. *Cell Rep.* **2**: 817-823.

Holloway AK, Bruneau BG, Sukonnik T, Rubenstein JL, Pollard KS. 2016. Accelerated evolution of enhancer hotspots in the mammal ancestor. *Mol Biol Evol* **33**: 1008–1018.

Hubisz MJ, Pollard KS, Siepel A. 2011. Phastand Rphast: Phylogenetic analysis with space/time models. *Brief Bioinform* **12**: 41–51.

Ikegami K, Mukai M, Tsuchida JI, Heier RL, MacGregor GR, Setou M. 2006. TTLL7 is a mammalian β-tubulin polyglutamylase required for growth of MAP2-positive neurites. *J Biol Chem* **281**: 30707–30716.

Levy Karin E, Wicke S, Pupko T, Mayrose I. 2017. An Integrated Model of Phenotypic Trait Changes and Site-Specific Sequence Evolution. *Syst Biol* **66**: 917–933. https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syx032.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–116.

Kishida T, Kubota S, Shirayama Y, and Fukami H. 2007. The olfactory receptor gene repertoires in secondary-adapted marine vertebrates: evidence for reduction of the functional proportions in cetaceans. *Biol. Lett.* **3**: 428-430.

Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2012. The role of GC-Biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol Biol Evol* **29**: 1047–1057.

Laub F, Lei L, Sumiyoshi H, Kajimura D, Dragomir C, Smaldone S, Puche AC, Petros TJ, Mason C, Parada LF, et al. 2005. Transcription factor KLF7 is important for neuronal morphogenesis in selected regions of the nervous system. *Mol Cell Biol* **25**: 5699–5711.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**: 476–82. http://dx.doi.org/10.1038/nature10530.

Marcovitz A, Jia R, Bejerano G. 2016. "reverse Genomics" Predicts Function of Human Conserved Noncoding Elements. *Mol Biol Evol* **33**: 1358–1369.

Marcovitz A, Turakhia Y, Gloudemans M, Braun BA, Chen HI, Bejerano G. 2017. A novel unbiased test for molecular convergent evolution and discoveries in echolocating, aquatic and high-altitude mammals. *bioRxiv* doi: https://doi.org/10.1101/170985

Martinelli DC, Fan CM. 2007. Gas1 extends the range of Hedgehog action by facilitating its signaling. *Genes Dev* **21**: 1231–1243.

Martynoga B, Morrison H, Price DJ, Mason JO. 2005. Foxg1 is required for specification of ventral telencephalon and region-specific regulation of dorsal telencephalic precursor proliferation and apoptosis. *Dev Biol* **283**: 113–127.

Mayrose I, Otto SP. 2011. A likelihood method for detecting trait-dependent shifts in the rate of molecular evolution. *Mol Biol Evol* **28**: 759–770.

McGowen MR, Gatesy J, Wildman DE. 2014. Molecular evolution tracks macroevolutionary transitions in Cetacea. *Trends Ecol Evol* **29**: 336–346.

Mclean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, et al.. 2011. evolution of human-specific traits. *Nature* **471**: 216-219.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**: 495–501. http://www.nature.com/doifinder/10.1038/nbt.1630.

Meredith RW, Zhang G, Gilbert MTP, Jarvis ED, and Springer MS. 2014. Evidence for a single loss of mineralized teeth in the common avian ancestor. *Science*, **346**: 1254390-1254390.

Mendes FK, Hahn MW. 2016. Gene Tree Discordance Causes Apparent Substitution Rate Variation. *Syst Biol* **65**:711-721.

Mitchell KJ, Llamas B, Soubrier J, Rawlence NJ, Worthy TH, Wood J, Lee MSY, and Cooper A. 2014. Ancient DNA reveals elephant birds and kiwi are sister taxa and clarifies ratite bird evolution. *Science*, **344**: 898-900.

Orr, HA. 2005. The probability of parallel evolution. *Evolution*, **59(1)**: 216–220.

Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**: 228–231.

Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL. 2017. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *Elife* **6**.

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, et al.. 2006. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2**: 1599-1611.

Prudent X, Parra G, Schwede P, Roscito JG, Hiller M. 2016. Controlling for Phylogenetic Relatedness and Evolutionary Rates Improves the Discovery of Associations between Species' Phenotypic and Genomic Differences. *Mol. Biol. Evol.* **33**: 2135–2150.

Roff DA. 1994. The evolution of flightlessness: Is history important? *Evol. Ecol.* **8**: 639-657.

Roscito JG, Sameith K, Parra G, Langer B, Petzold A, Rodrigues MT, Hiller M. 2017. Phenotype loss is associated with widespread divergence of the gene regulatory landscape in evolution. *bioRxiv* https://doi.org/10.1101/238634.

Rosenblum EB, Parent CE, and Brandt EE. 2014. The Molecular Basis of Phenotypic Convergence. *Annu. Rev. Ecol. Evol. Syst.* **45**: 203-226.

Sackton TB, Grayson P, Cloutier A, Hu Z, Liu JS, Wheeler NE, Gardner PP, Clarke JA, Baker AJ, Clamp M, et al. 2018. Convergent regulatory evolution and the origin of flightlessness in palaeognathous birds. *bioRxiv*. https://doi.org/10.1101/262584

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LDW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Storz, JF. Causes of molecular convergence and parallelism in protein evolution. 2016. *Nature Reviews Genetics* **17**: 239–250.

Stern DL. 2013. The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**: 751-764.

Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS. 2012. The Molecular Diversity of Adaptive Convergence. *Science (80-)* **335**: 457–461. http://www.sciencemag.org/cgi/doi/10.1126/science.1212986.

Venkatesh B, Kirkness EF, Loh Yh, Halpern AL, Lee AP, Johnson J, Dandona N, Viswanathan LD, Tay A, Venter JC, et al.. 2006. Conserved in the Human Genome. *Science* (80-.). **1892**: 2005-2006.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**(1): e7. https://doi.org/10.1371/journal.pbio.0030007

Wray GA. 2013. Genomics and the Evolution of Phenotypic Traits. *Annu. Rev. Ecol. Evol. Syst.* **44**: 51-72.

Yonezawa T, Segawa T, Mori H, Campos PF, Hongoh Y, Endo H, Akiyoshi A, Kohno N, Nishida S, Wu J, et al. 2017. Phylogenomics and Morphology of Extinct Paleognaths Reveal the Origin and Evolution of the Ratites. *Curr Biol* **27**: 68–77.

SUPPLEMENTAL MATERIALS

Title: Supplementary_Material.pdf

Supplementary text, figures and tables. (pdf file)

Mammal_species.txt

The scientific name, common name and UCSC genome assembly of all species on the mammalian phylogeny. (txt file)