A hierarchical Bayesian model for detecting convergent rate changes of conserved 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

Abstract

Conservation of DNA sequence over evolutionary time is a strong indicator of function, and thus gain or loss of sequence conservation can be used to infer putative changes in function across a phylogeny. Changes in evolutionary rates on particular clades or 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 complex 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, which can handle diverse evolutionary patterns and complex patterns of convergence. The model 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 on each branch. Simulations show that PhyloAcc can detect rate shifts in multiple species with greater power and fewer false positives 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 the context of 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 (CNEEs) 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 phylogenetic tree.

Introduction

One of the major revelations of comparative genomics has been the discovery of regions of the genome that exhibit considerable levels of conservation across evolutionary time, often falling well outside protein-coding genes (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., 2012; Marcovitz et al., 2016). Changes of conservation of these elements in one or 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. (2015) discovered 166 bat-accelerated regions overlapping with enhancers in developing mouse limbs.

Phenotypic convergence (in which the same function evolves multiple times independently, mainly because of 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 et al., 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 et al., 2005; Tenaillon et al., 2012; Parker et al., 2013; Storz et al., 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., 2010), 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 to detect 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 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 rates variation among species. Also, TraitRate can only test for an association between rates changes and a given trait, but cannot detect where these transitions happen in a phylogeny.

Here, we introduce PhyloAcc, an alternative Bayesian method to model multiple substitution rate changes on a phylogeny. In our new method, we relaxed the parsimony assumption on the history of rate shifts and developed a model-based method to estimate the conservation state of each branch based on sequences of extant species, allowing each genomic region to have a different substitution shift pattern. By 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 factor. Bayes factor, defined as the ratio of marginal likelihood by integrating over parameters space in each model, is a compelling choice for model selection, which 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.

By simulation, we show that our method has higher sensitivity and specificity to identify genomic regions accelerated in varying lineages on the tree than the leading likelihood ratio test based method. Crucially, our method can identify the correct shift pattern of substitution rates in a variety of models of phenotypic convergence. While our method has some similarity to random local clock models (Drummond & Suchard, 2010) implemented in several software packages (e.g. BEAST2, Bouckaert et al., 2014), which also allow for varying substitution rates across phylogeny, the focus of our method is on detecting shift patterns instead of reconstructing phylogeny. Our simulations demonstrate that BEAST2 is not as accurate as PhyloAcc at reconstructing rate shifts, especially when we simulated multiple shifts of substitution rates across different lineages. Moreover, BEAST2 does not provide a method for selecting elements with evolutionary rate shifts 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 conserved state will have fewer substitutions than under neutrality ($r_{i1} < 1$), while branches in the accelerated state are expected to have more substitutions than those of conserved states ($r_{i2} > r_{i1}$), although this may 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 having same state and similar substitution rates, which is a common assumption in phylogenetics and reasonable with closely-related species in 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 **Z** and substitution rates **r** for each element by using collapsed Gibbs sampling and outputs posterior distribution of **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

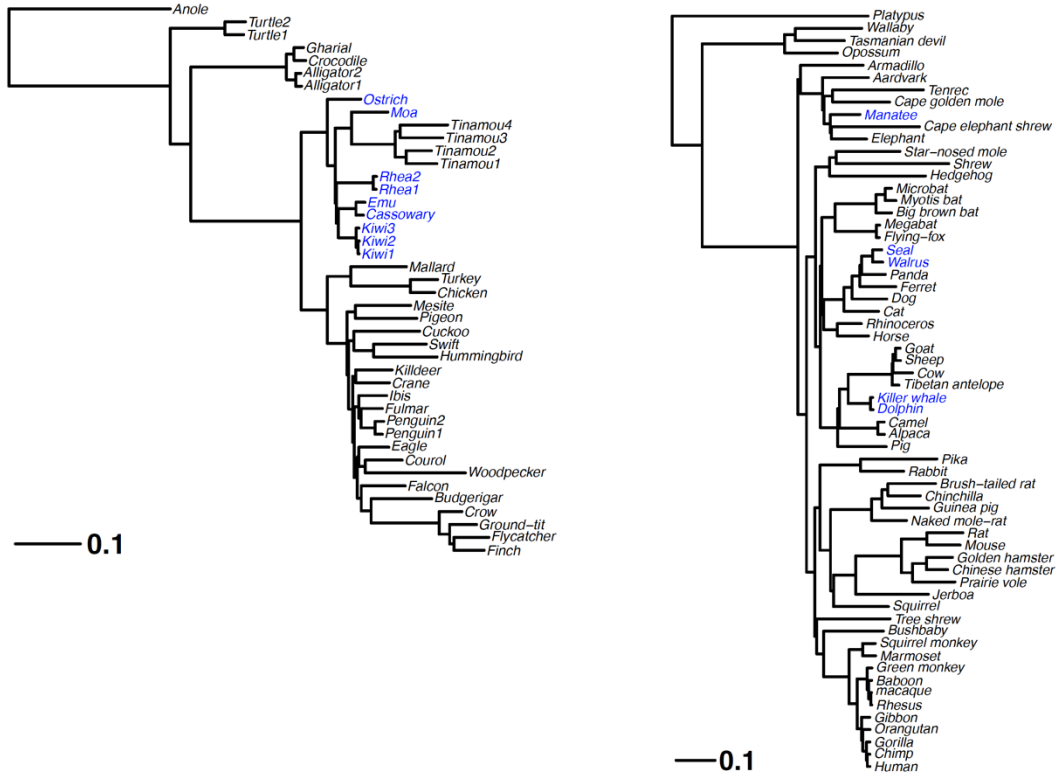
accelerated ($Z_s = 2$) while all other branches must be either neutral or conserved ($Z_s = 0\ or\ 1$); in the full model (M_2), the latent conservation states **Z** can take any configuration across the phylogeny. To compare models, we compute marginal likelihood $P(Y|M_i)$ for each model by integrating out (unobserved) ancestral DNA sequences, latent conservation states **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 be regulatory elements with broader function.

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.

Our first example of phenotypic convergence is loss of flight in birds, which has occurred many times in bird evolution (Roff, 1994; Harshman et al., 2008; Mitchell et al., 2014). Recent phylogenetic work supports the conclusion that the ratites (ostrich, emu, cassowary, kiwi, rheas, and the extinct moa 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 use 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).

(a) *(b)*

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

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, we did our analysis 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). For this study, we use 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).

We will show simulation results based on both phylogenies as a demonstration of the power of our method to identify element conserved in most species considered but accelerated in a target group of phenotypically convergent species. We focus on the ratite and marine mammal example as these represent two extremes of possible topologies of convergent species on a phylogeny. In the ratite case, multiple independent losses are supported 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. In the marine mammal case, the three independent transitions are widely separately on the phylogeny. Then, we apply our method to conserved non-coding elements identified from multiple alignments in both cases and show evidence of enrichment 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. We focus on this topology, as the scenario where convergent lineages are clustered in paraphyletic clade is particularly challenging for existing methods. In our simulation, we set the

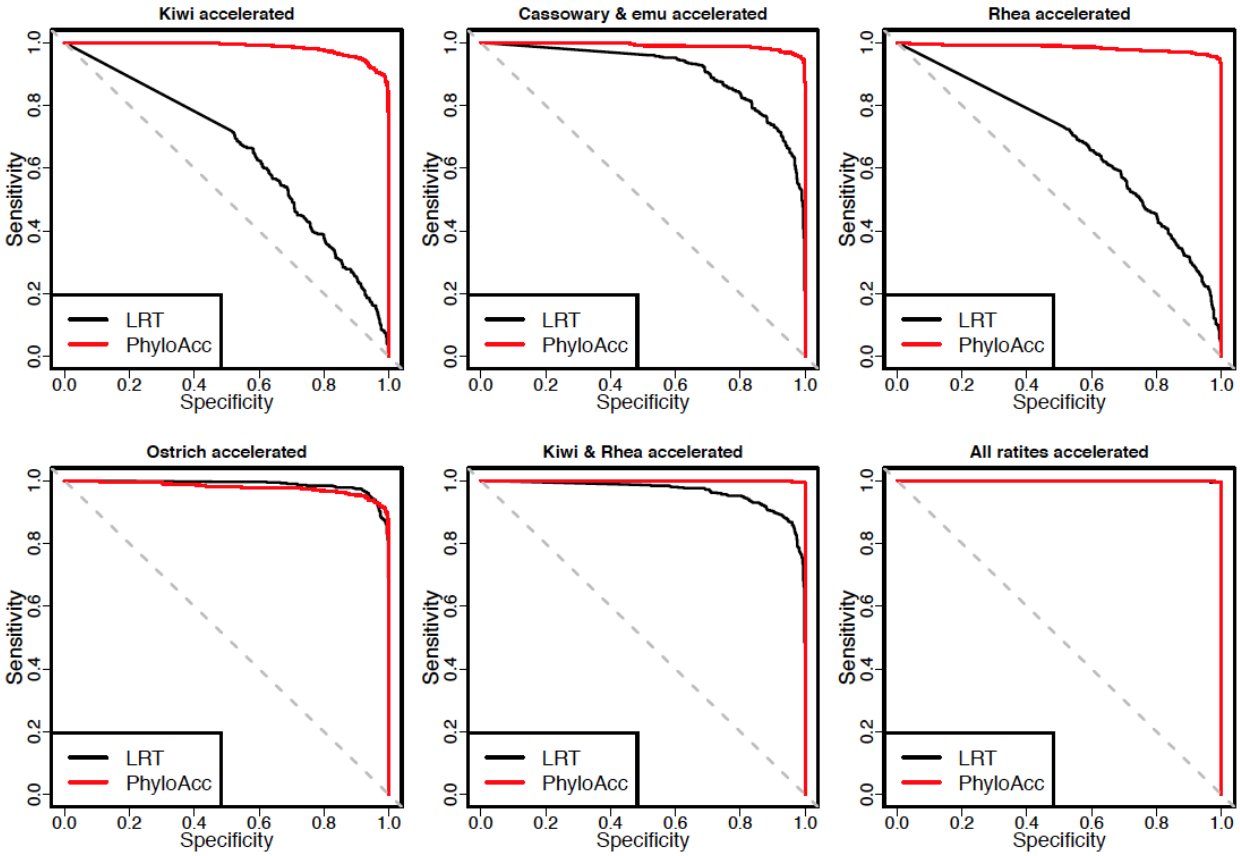
length of element to be 200 bp which is the median length in our real data. We generated 9 scenarios with different levels of convergence either around ratites or other species: 1) all branches are conserved; 2) only kiwi clade accelerated; 3) only emu/cassowary branches accelerated; 4) only rhea clade 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 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 volant tinamou clade resides within ratites, making it hard 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

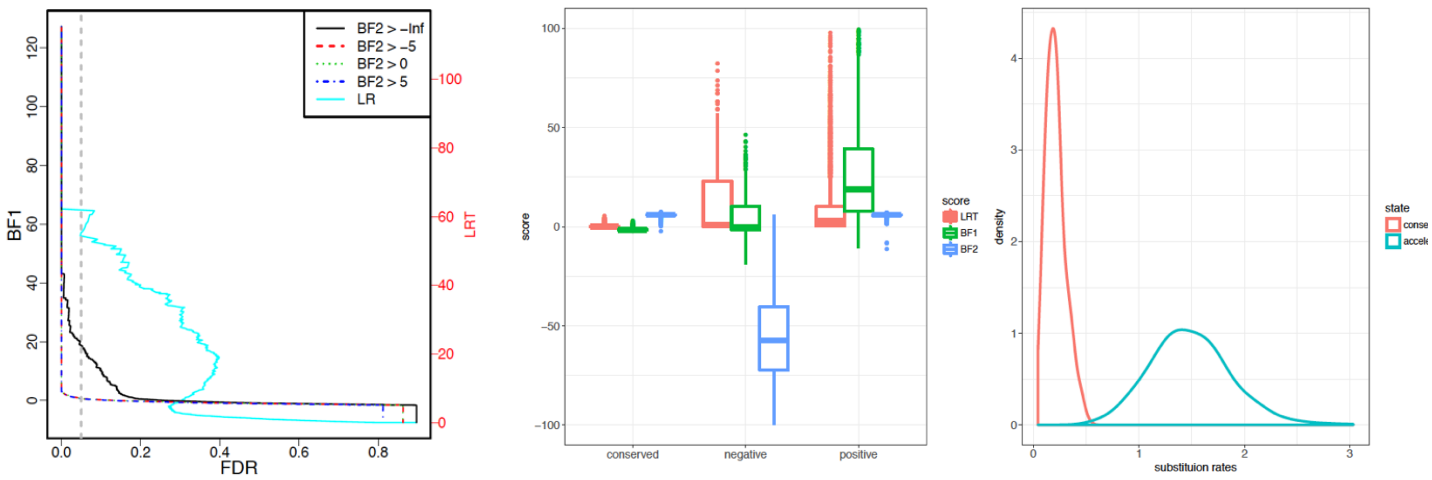
To test the sensitivity and specificity of our method, we mixed elements from (2)–(7) with background elements (1) separately and show receiver operating characteristic (ROC) curve in each experiment, which the area under the curve reflects the performance of each method to select ratite-specific accelerated 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 program 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 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, we showed that 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 this compared with log-likelihood ratio statistic in phyloP (Figure 2B). The LRT fails to distinguish ratite-specific acceleration from the rest as FDR is still quite high even though the likelihood ratio is large (cyan curve in Figure 2B); in contrast, when using BFs in PhyloAcc, FDR drops below 5% at reasonable cutoffs (e.g., BF1 > 0, BF2 > 0). In real application, the null distributions of BFs for elements not specifically accelerated in ratites are complicated and depend on the prior as well as evolutionary pattern, making it hard 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.


Kiwi accelerated
Cassowary & emu accelerated
Rhea accelerated
Ostrich accelerated
Kiwi & Rhea accelerated
All ratites accelerated
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
LRT
PhyloAcc

(a)



(b) (c) (d)

Figure 2: (a) ROC curve for PhyloAcc and LRT in different acceleration scenarios within ratites. We mixed elements with different patterns of acceleration with all conserved elements separately and compared sensitivity and specificity of our method to phyloP. Figure (1)–(6) are ROC curves treating elements from scenario (2)–(7) respectively as positive and those from scenario (1) as negative. (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 (7-8: accelerated in non-ratite species) and positive (1-6: 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) where acceleration happens at the 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 that 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). In contrast, almost no element has BF2 less than 0 in case (1)-(7) since Bayes factor favors 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 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). Clearly, in this realistic scenario PhyloAcc has much better performance than phyloP. 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 acceleration pattern of individual genomic element

Finally, we validated that our method can recover the true acceleration pattern (latent states) for individual genomic element. For each simulated element, we compared the posterior of **Z** output from our method with true simulated pattern, and defined the result as "correct" if the posterior probability of the corresponding true latent state is greater than 70% 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.

In order to investigate the impact of accelerated and conserved rate ratio on the ability of our method to recover the true pattern of acceleration, we ordered and divided the simulated elements into 10 equal-sized groups according to the ratio (the quantiles of r_2/r_1 in each group are shown in Supplementary Figure 4A). As phyloP can't produce the acceleration pattern directly, we compare the performance with BEAST2 (Bouckaert et al., 2014). Figure 3 shows the accuracy of our method compared to BEAST2 as well as BF1/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 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 the simulated ground truth. 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 limited informative sites (Supplementary Figure 6 and Supplementary Figure 4B).

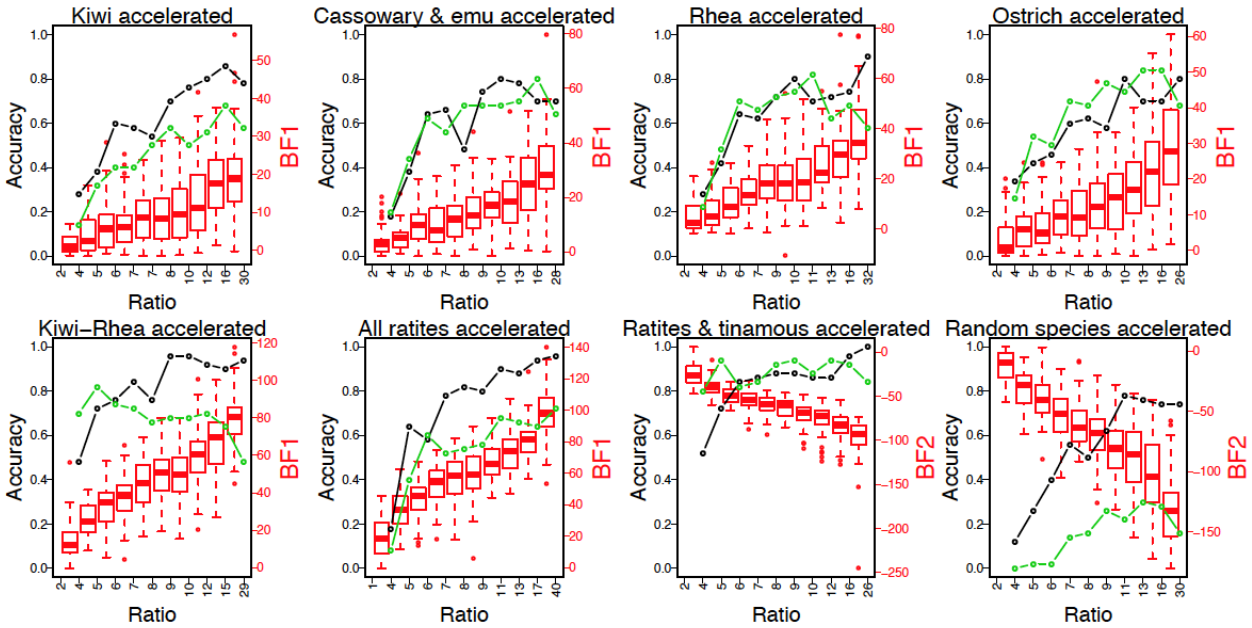


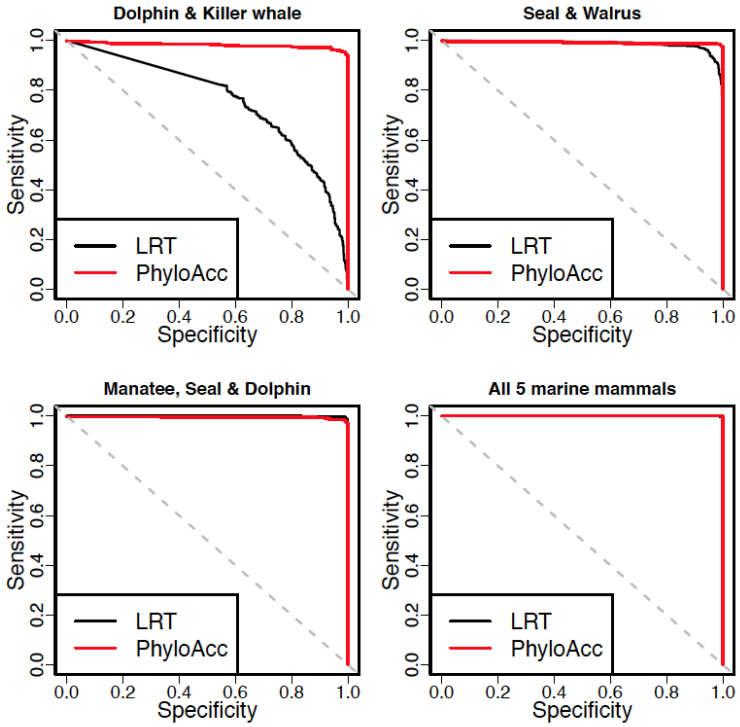
Figure 3: Comparison of accuracy of substitution rate shift patterns between BEAST2 and PhyloAcc in each simulation scenario. We divided the simulated elements into 10 equal-sized groups according to the ratio of accelerated and conserved rate for scenarios (2)-(9). X-axis shows the boundary of ratio in each group; black curve is the accuracy using PhyloAcc and green curve is for BEAST2. 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, 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 proportion of indels into the simulated data, approximately the observed range in the real data. The accuracy of identifying the shift patterns of substitution rates, as well as BF1, is relatively

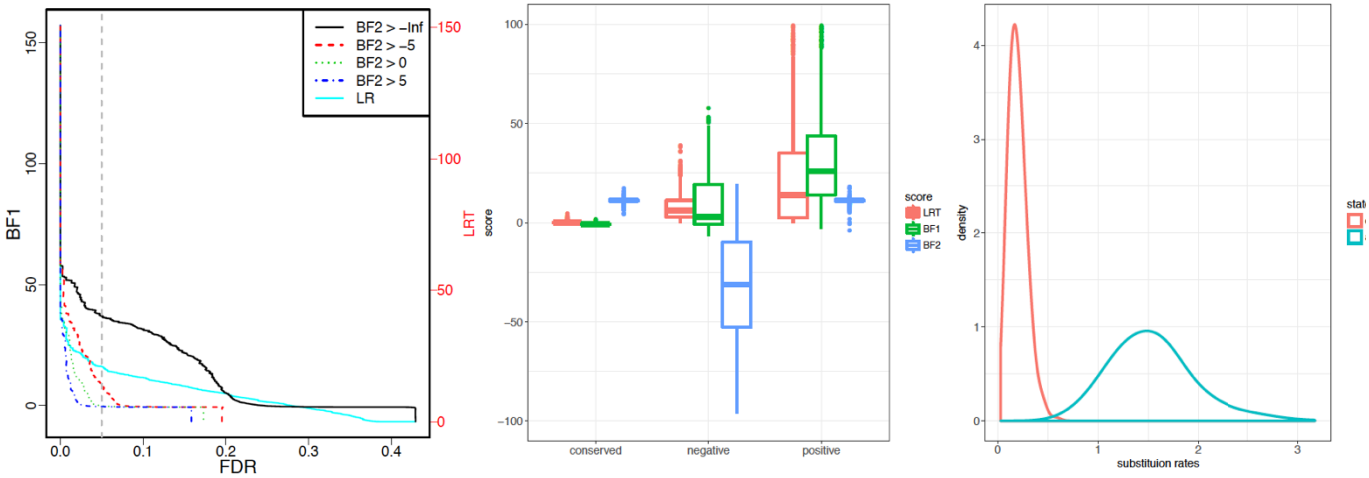
insensitive to indels (Supplementary Figure 5). In short, this simulation demonstrates that our method is capable of discovering multiple shifts of substitution rates.

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 phylogeny of 62 mammalian species evolving under different patterns of substitution rates variation. We compared our method with phyloP in various scenarios: 1) all 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 under common ancestor of cat and pinnipeds (Supplementary Figure 7). 2)- 5) are marine mammal specific accelerated cases, while 6)-7) are negative controls with non-specific acceleration. We demonstrated that our method has higher sensitive and specificity for identifying substitution rates accelerated exclusively in marine mammals than phyloP (Figure 4). Figure 4A shows that PhyloAcc has higher sensitivity to detect genomic elements accelerated within marine mammals than phyloP; (b) and (c) show that PhyloAcc will exclude elements accelerated in species other than marine mammals by criterion on BF2 and FDR drops below 5% when selecting elements with BF2>0 and BF1>0. These results suggest that the sensitivity and specificity of PhlyoAcc is expected to be high under a range of evolutionary scenarios.



(a)



(b) (c) (d)

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 CNEEs accelerated 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). We identified 820 CNEEs with strong evidence for ratite-specific acceleration (BF1 > 20 and BF2 > 0). Rhea clades are 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 (Methods). Many of these CNEEs have experienced multiple independent losses: 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).

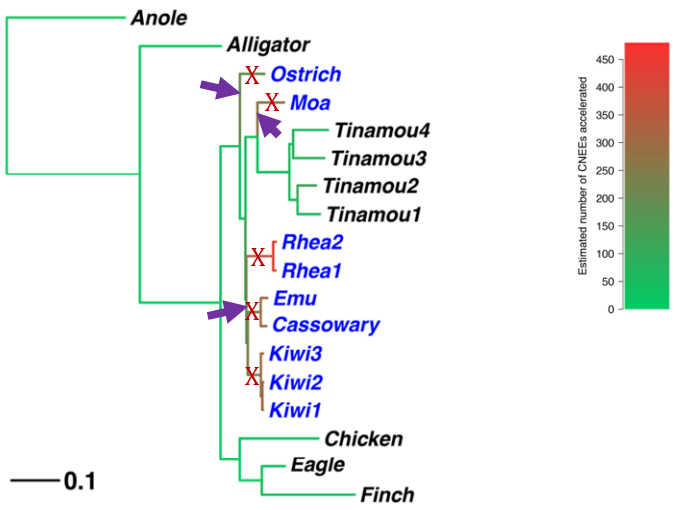


Figure 5: Number of CNEEs being accelerated per branch. Phylogeny for avian data set (only some species in neognathae and reptiles are shown for illustration). Palaeognaths consist of the flightless ratites and volant tinamous. Ratites are shown in blue. Potential losses of flight in ratites are shown 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 largest Bayes factor (BF1). Although all of them show strong evidence of acceleration in ratites, they have different acceleration patterns. A few of them are accelerated in a single species (e.g. middle right element accelerated only in cassowary); many of them are accelerated in a single clade with one loss (e.g. top left element accelerated in both kiwis and emu/cassowary, top right one accelerated only in rheas, middle left one accelerated only in kiwis); others are accelerated in more than one clades with multiple losses (e.g. the bottom left one accelerated in both rheas and kiwis and the bottom right one accelerated in both rheas and emu/cassowary). 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).

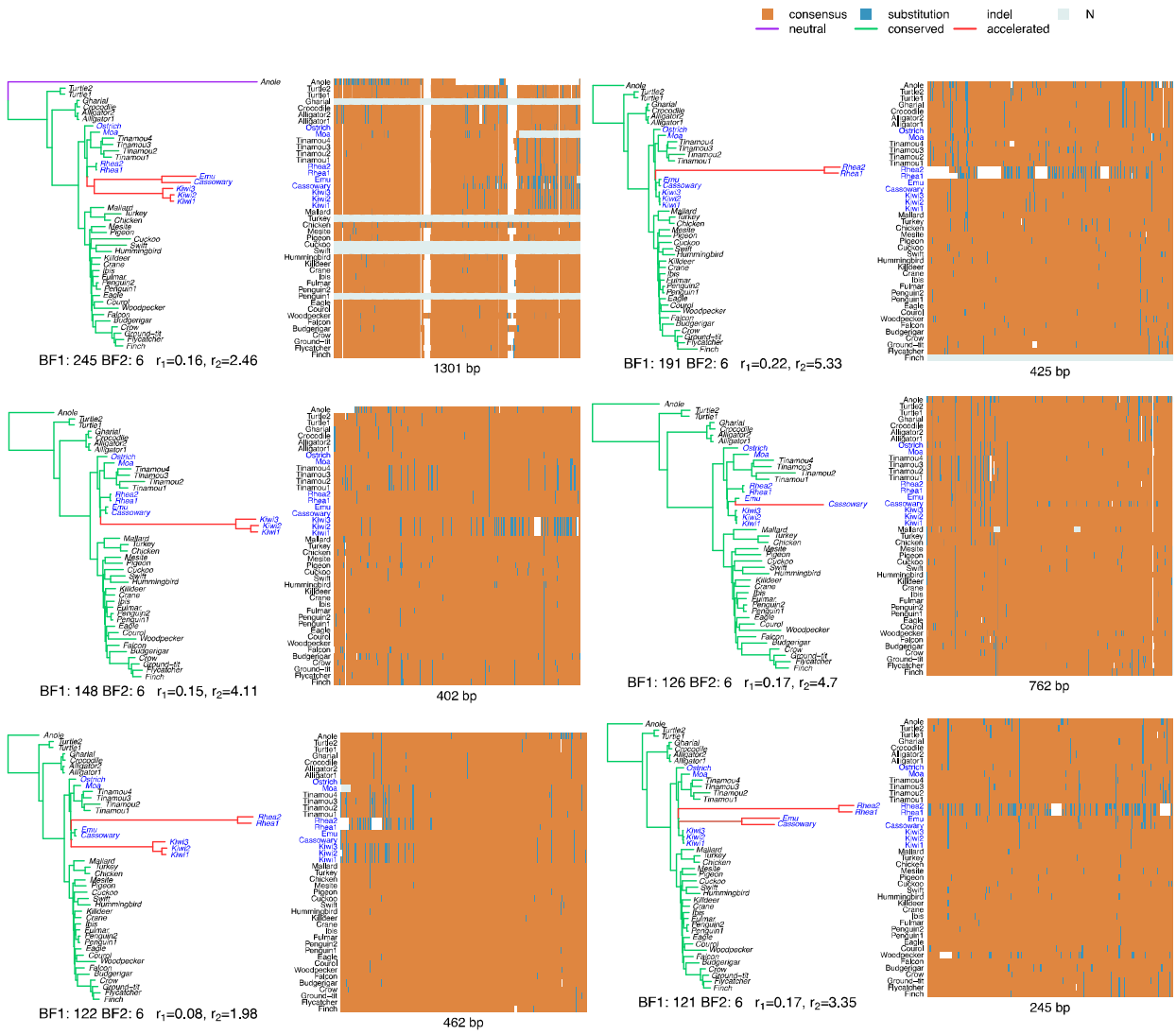


Figure 6: Examples of ratite-accelerated elements. For each element, shift pattern of substitution rates is shown on the left represented by a phylogenetic tree with branch lengths proportional to posterior substitution rate and colored by posterior mean of Z, along with the sequence alignments shown on the right. For the phylogenetic tree, green: conserved; red: accelerated; purple: 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 CNEEs accelerated 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 didn't find significant convergent changes at specific amino acid sites beyond that could explained by neutral evolution model (Foote et al., 2015); while others claimed convergence on the basis of shifts of substitution rates at gene level (Chikina et al., 2016). Among marine-accelerated protein-coding genes, they found evidence of adaptive evolution in skin and lung genes as well as loss of functions 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 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 phenotype 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 (2106 for marine-accelerated vs. 1472 for control-accelerated elements with BF1>5 and BF2 > 5) show substitution rate shifts in marine mammals than control species. Furthermore, we found

larger Bayes factor between lineage-specific and null model for marine-accelerated elements, indicating more dramatic change of substitution rates affecting more species in marine mammals (Supplementary Figure 8). In addition, more marine-accelerated CNEEs show parallel shifts in multiple lineages than controls: 696 (33%) of marine-accelerated elements show acceleration in more than 3 species 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, 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 8).

Finally, we tested for functional enrichment of genes nearby marine-accelerated CNEEs 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 uncovers molecular adaptations overlooked by previous studies in searching protein-coding genes. Checking individual genes associated with these enriched functional annotations, we found a handful of top marine-accelerated CNEEs are close to: 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, plays 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); GAS1 and GLI2, function as transcription regulator in the hedgehog (Hh) pathway which is important for embryogenesis (Martinelli et al., 2007).

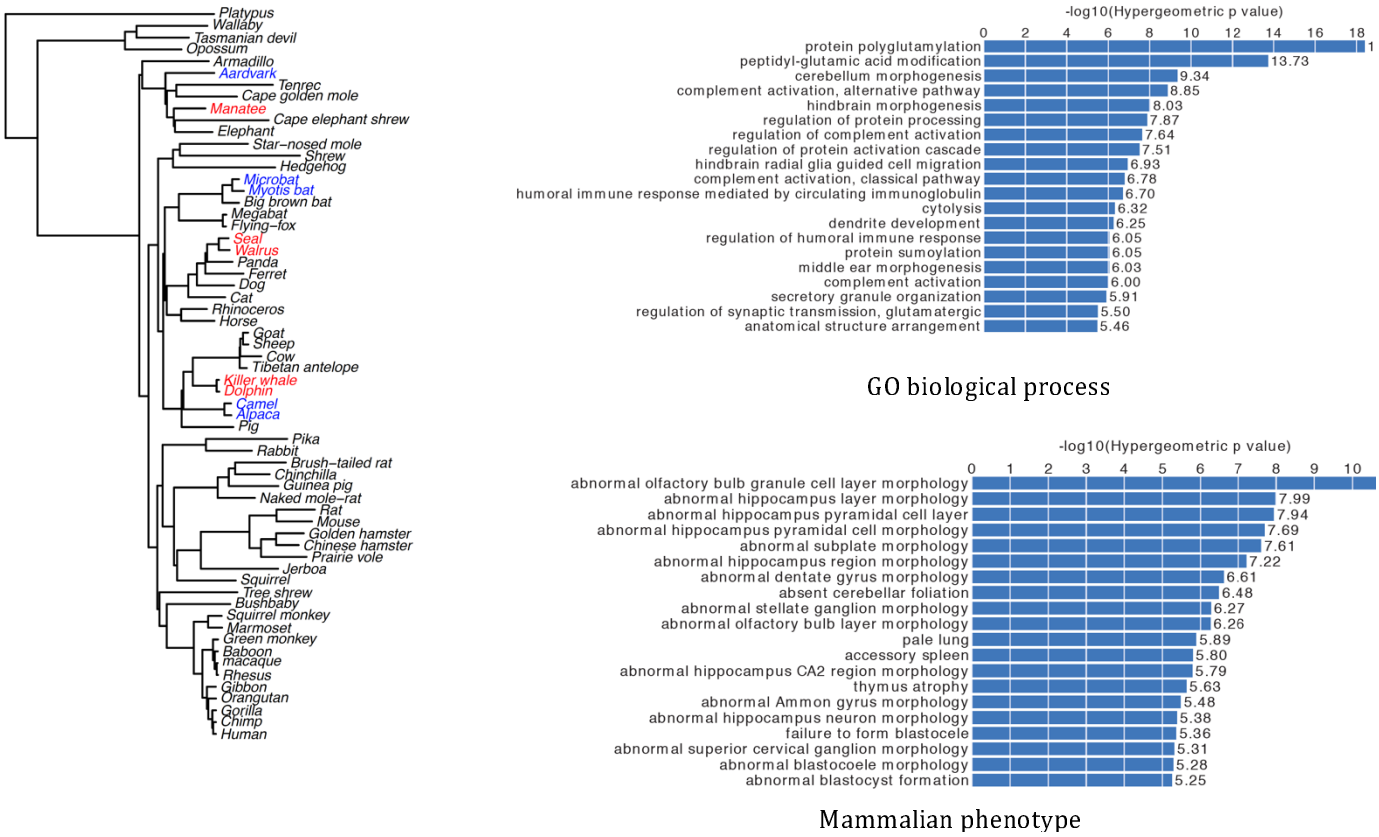


Figure 7: (a) Phylogeny of mammalian data. 5 marines are shown in red and 5 control species are shown in blue. (b) Enriched GO terms and mammalian phenotypes of genes near marine-accelerated CNEEs. Only shown top 20 terms (all of them with FDR <0.01).

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 9).

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; and 3) a bed file with the position of each individual element in the coordinate of concatenated alignment file (0-based). 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) under null, lineage-specific and full model 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). It can not only identify DNA elements showing changes of substitution rate in the species of interest, but also infer on which branches, either single or multiple lineages, changes of substitution rate happen, which facilitates 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 flight loss 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 gene or non-coding region with repeated losses or gains of a given trait in multiple independent lineages in

order to gain insight of molecular basis for phenotype differences has first proposed as "Forward Genomics" by Hiller and his colleagues (Hiller et al, 2012b). Since then, it has been used in various species (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 lower false positive rate by comparing marginal likelihood of model either allowing or prohibiting acceleration in species without phenotype change. Moreover, using Bayes factor to select specifically accelerated element doesn't rely on the asymptotic normal assumption required for the likelihood ratio test, and is more robust for some extreme cases (e.g. a few branches have many substitutions) where maximum likelihood is unbounded.

Although in our examples we focus on loss of conservation with more rapid substitution rate, and do not attempt to distinguish among possible evolutionary models that can produce this pattern, our software is able to detect a change of substitution rate. It will output the posterior of r_2 for each element, from which the direction of change could be inferred. To identify elements with adaptive selection, our software provides 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 in order to identify elements more conserved in a group of species than others. Currently, we assume the same substitution rate for all accelerated branches, 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. This extension could also allow for a model that distinguishes loss of conservation from selection, by introducing additional latent states. In addition, by Dollo's irreversible assumption, our model allows at most two shifts (e.g. at first gain and then lose conservation) for each lineage on the phylogeny, which may not be powerful to detect elements which regain conservation after adaptive selection long

time ago. We could adjust the transition probability matrix of conservation state (Z) to allow small probability to transit from accelerated to conserved state. However, in many scenarios the parsimony assumption is helpful, since the sequence data of extant species doesn't 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 model.

The local substitution rate and phylogenetic tree are likely not constant under neutral evolution across the genome, which may impact our method (and all previous methods). One way to tackle this is to estimate substitution rate and phylogenetic tree under neutral evolution for different segments of the genome. However, this introduces some degree of arbitrariness in the decision of how to segment the genome, and at least variation in the neutral rate is not a serious issue for our model as we already model variation of 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, only few outgroup lineages are in typically neutral state so the actual value of the neutral rate has relatively little direct impact. 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.

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). 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 ~30% ratite-accelerated elements also showing GC-biased in the accelerated lineages, indicating that gBGC is one of the force prompting loss of conservation for DNA elements released from purifying selection. gBGC can be intertwined with adaptive selection, e.g. many human-accelerated regions are partly caused by gBGC which may not be positively selected to increase fitness (Pollard et al., 2006). Though for our purpose, we didn't disentangle gBGC from other molecular mechanisms associated with accumulating substitutions and possible function loss, our software also provides an extension version which could distinguish the effect of gBGC from selection.

Identifying function of regulatory region is still a challenging task and linking sequence evolution (e.g. substitution rate) from a wide range of species with organism-level evolutionary phenomena (e.g. convergent evolution) has the potential to shed light upon regulatory function of conserved non-coding regions. Our method could provide the probability of match between evolutionary profile of genetic elements with presence/absent patterns of hundreds of traits to predict phenotype-genotype pairs that might be functional linked, as an extension for the "Reverse Genomics" method (Marcovitz et al., 2016). Based on parsimony, their method didn't consider the uncertainty of conservation pattern estimated from sequencing data and the probability of each match just by chance. It's also not easy to deal with missing phenotype/genotype information without specifying a probability model. In addition to identify genotype-phenotype link, our model can also be extended to cluster genomic regions (e.g. using Dirichlet process as prior for Z) based on similar sequence evolutionary pattern to discover novel functional group of genomic loci that may or may not link to 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 birds 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 region in the genome is called by PhastCons in Phast package, and then 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 defined as for birds (conserved regions identified by PHAST, at least 50 bp, not overlapping any exons). 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 $\boldsymbol{X}_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ 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 species 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). The standard DNA substitution model depicts that substitution on one branch follows continuous time Markov Process with rate matrix Q and stationary distribution π, then transition probability matrix along a branch with

length t and relative substitution rate r is e^{rtQ}. We estimated Q matrix and branch length from neutral sequence (fourfold degenerate sites) precisely using phyloFit in phast package. As the total length of neutral sequence is long, we could neglect the uncertainty and treat them as fix 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 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 plugging in estimates of them 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 state to be larger than r_{i1} for conserved state.

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 for each element once it is accelerated on a particular branch, the downstream branches cannot regain conservation. Therefore, the transition matrix has a simplified form: $\Phi = \begin{bmatrix} 1-\alpha & \alpha & 0 \\ 0 & 1-\beta & \beta \\ 0 & 0 & 1 \end{bmatrix}$. It is worth noting that 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. 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 $Z_s = 1$. On the other hand, 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 could 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 $P\left(Z_{pa(s)} = 1, Z_s = 2 \middle| \boldsymbol{Y}\right) = P(\, Z_s = 2 | \boldsymbol{Y}) - \, P\left(\, Z_s = 2, Z_{pa(s)} = 2 \middle| \boldsymbol{Y}\right) =$

$P(Z_s = 2|\boldsymbol{Y}) - P(Z_{pa(s)} = 2|\boldsymbol{Y})$ (the second equality is because once accelerated it will retain in accelerated state). Then the expected number of accelerated species (N_1) within phenotypically convergent species (S_0) is: $EN_1 = \sum_{s \in S_0} P(Z_s = 2|\boldsymbol{Y})$ and the expected number of independent loss (N_2) within ancestors of S_0 is the sum of posterior probability of loss on each branch towards S_0:

$$EN_2 = \sum_{s \in S_1} P(Z_{pa(s)} = 1, Z_s = 2|\boldsymbol{Y}) = \sum_{s \in S_1} P(Z_s = 2|\boldsymbol{Y}) - P(Z_{pa(s)} = 2|\boldsymbol{Y})$$

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 model. Given $\mathbf{Z}$ and $\mathbf{r}$, each site j in an element is independent. Thus, we could 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 whole sequences of that element. Given sequences and conservation state, i.e. $(\mathbf{H}, \boldsymbol{Z})$, substitution probability on each branch could be easily computed by standard DNA substitution model. Posterior sampling of $\mathbf{Z}$ with biology 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 $\mathbf{H}$) can't be directly sampled, so we use adaptive Metropolis-Hasting algorithm within our Gibbs sampling scheme. We use a gamma

distribution centered around current r as proposal and adaptively tune the variance of proposal distribution based on the acceptance rate of previous MCMC steps.

Compute Bayes Factor for model comparison

To select elements with specific evolutionary pattern, we compare the probability of sequence alignments of extant species under null (M_0), lineage-specific (M_1) and full model (M_2) allowing for increasing configurations of $\mathbf{Z}$. In order to compute Bayes factor, the problem becomes how to compute marginal probability under different constraints of $\mathbf{Z}$. Since it's impossible to sum over all possible configurations of $\mathbf{Z}$, Chib's method is commonly used to compute marginal probability using the ratio of joint and conditional probability 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 though usually it's taken as the posterior mode or MLE. However, both the numerator and denominator can't be computed analytically in our case. We use the posterior distribution of Z output by MCMC to approximate the denominator and construct an upper bound for the numerator similar to variational method (Blei et al., 2016). To reduce the variance of the estimator of $P(Y|C)$, we extended 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. Since given $\mathbf{Z}$, sequence alignments Y and constraints C are independent, joint log-likelihood can be written as: $\log P(Y, \mathbf{Z}|C) = \log P(Y|\mathbf{Z}) +$

$\log P(\mathbf{Z}|C)$, we could calculate the second term by the forward-backward procedure, but we can't compute $\log P(Y|\mathbf{Z})$ directly since it involves integration over r s. However, we provided an upper bound for it:

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

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

$q(\boldsymbol{r}|\gamma)$ 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 $q(\boldsymbol{r}|\gamma)$ to be 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 factor are in 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. 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, there is about 60% of loci with 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 coordinate 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 kept 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 and members in Professor Edwards and Jun's lab for kindly discussion. 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, A. J., Haddrath, O., Mcpherson, J. D. & Cloutier, A. Genomic support for a moa-tinamou clade and adaptive morphological convergence in flightless ratites. *Mol. Biol.Evol.* **31,** 1686–1696 (2014).

Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).

Berger, M. J., Wenger, A. M., Guturu, H. & Bejerano, G. Independent erosion of conserved transcription factor binding sites points to shared hindlimb, vision, and scrotum loss in different mammals. *bioRxiv* (2017). doi: http://biorxiv.org/content/early/2017/10/03/197756.abstract

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational Inference: A Review for Statisticians. (2016). doi:10.1080/01621459.2017.1285773

Booker, B. M. et al. Bat Accelerated Regions Identify a Bat Forelimb Specific Enhancer in the HoxD Locus. *PLoS Genet.* **12**, (2016).

Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **10,** (2014).

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, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. B Biol. Sci.* **368,** 20130025–20130025 (2013a).

Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S. & Siepel, A. A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. *PLoS Genet.* **9**, (2013b).

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

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, A. J. & Suchard, M. A. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* **8,** 114 (2010).

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.

Harshman, J. *et al.* Phylogenomic evidence for multiple losses of flight in ratite birds. *Proc. Natl. Acad. Sci.* **105,** 13462–13467 (2008).

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, A. K., Bruneau, B. G., Sukonnik, T., Rubenstein, J. L. & Pollard, K. S. Accelerated evolution of enhancer hotspots in the mammal ancestor. *Mol. Biol. Evol.* **33,** 1008–1018 (2016).

Hubisz, M. J., Pollard, K. S. & Siepel, A. Phast and Rphast: Phylogenetic analysis with space/time models. *Brief. Bioinform.* **12,** 41–51 (2011).

Ikegami, K. *et al.* TTLL7 is a mammalian β-tubulin polyglutamylase required for growth of MAP2-positive neurites. *J. Biol. Chem.* **281,** 30707–30716 (2006).

Levy Karin, E., Wicke, S., Pupko, T. & Mayrose, I. An Integrated Model of Phenotypic Trait Changes and Site-Specific Sequence Evolution. *Syst. Biol.* **66,** 917–933 (2017).

King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).

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, M. J., Siepel, A. & Pollard, K. S. The role of GC-Biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol. Biol. Evol.* **29**, 1047–1057 (2012).

Laub, F. *et al.* Transcription factor KLF7 is important for neuronal morphogenesis in selected regions of the nervous system. *Mol. Cell. Biol.* **25,** 5699–5711 (2005).

Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–82 (2011).

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

Marcovitz, A. et al. A novel unbiased test for molecular convergent evolution and discoveries in echolocating, aquatic and high-altitude mammals. *bioRxiv* (2017). doi: https://doi.org/10.1101/170985

Martinelli, D. C. & Fan, C. M. Gas1 extends the range of Hedgehog action by facilitating its signaling. *Genes Dev.* **21,** 1231–1243 (2007).

Martynoga, B., Morrison, H., Price, D. J. & Mason, J. O. Foxg1 is required for specification of ventral telencephalon and region-specific regulation of dorsal telencephalic precursor proliferation and apoptosis. *Dev. Biol.* **283,** 113–127 (2005).

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

McGowen, M. R., Gatesy, J. & Wildman, D. E. Molecular evolution tracks macroevolutionary transitions in Cetacea. *Trends in Ecology and Evolution* **29,** 336–346 (2014).

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, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., … Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, **28(5)**, 495–501.

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.

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, H. A. (2005). THE PROBABILITY OF PARALLEL EVOLUTION. *Evolution*, **59(1)**, 216–220.

Parker, J. et al. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228–231 (2013).

Partha, R. *et al.* Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *Elife* **6,** (2017).

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, J. G. & Hiller, M. Controlling for Phylogenetic Relatedness and Evolutionary Rates Improves the Discovery of Associations between Species' Phenotypic and Genomic Differences. *Mol. Biol. Evol.* **33,** 2135–2150 (2016).

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

Roscito, J. G. *et al.* Phenotype loss is associated with widespread divergence of the gene regulatory landscape in evolution. *bioRxiv* (2017). doi: http://biorxiv.org/content/early/2017/12/22/238634.abstract

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

Sackton, T. B. *et al.* Convergent regulatory evolution and the origin of flightlessness in palaeognathous birds. *bioRxiv* (2018). doi: https://doi.org/10.1101/262584

Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).

Storz, J. F. Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews Genetics* **17,** 239–250 (2016).

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

Tenaillon, O., Rodriguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D., & Gaut, B. S. (2012). The Molecular Diversity of Adaptive Convergence. *Science*, **335(6067)**, 457–461.

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. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3,** (2005).

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

Yonezawa, T. *et al.* Phylogenomics and Morphology of Extinct Paleognaths Reveal the Origin and Evolution of the Ratites. *Curr. Biol.* **27,** 68–77 (2017).

SUPPLEMENTAL MATERIALS

Title: Supplementary_Material.pdf

Supplementary text, figures and tables. (pdf file)